نمونه‌گیری

روشها و کاربردها

پل اس. لِه‌وی

استنلی لِمی‌شو

مترجم

گیتی مختاری امیرمجدی

پژوهشکدهٔ آمار

فهرست‌نویسی پیش از انتشار کتابخانهٔ ملّی ایران

لِوی، پل Levy, Paul S.
نمونه‌گیری: روشها و کاربردها / پل اس. لِه‌وی، استنلی لِمی‌شو؛ مترجم گیتی مختاری امیرمجدی
-- تهران: مرکز آمار ایران، پژوهشکدهٔ آمار، ۱۳۸۱.
۵۸۵ ص. : جدول، نمودار.
ISBN 964-365-150-9 : ۳۰۰۰۰ ریال
فهرست‌نویسی بر اساس اطلاعات فیپا.
عنوان اصلی: Sampling of Populations: Methods and Applications.
کتاب‌نامه.
نمایه
۱. جمعیت – روشهای آماری. ۲. آمارگیری نمونه‌ای. الف. لِمِشو، استنلی Lemeshow, Stanley
ب. مختاری امیرمجدی، گیتی، ۱۳۲۳- ، مترجم. ج. مرکز آمار ایران. پژوهشکدهٔ آمار. د. عنوان.
HB۸۴۹/۴۹/ل۹ن۸ ۳۰۴/۶۰۱۵۱۹۵۲
۱۳۸۱
کتابخانهٔ ملی ایران ۸۱-۴۰۶۴۷م

مدیریت تولید : **گروه پژوهشی طرحهای فنی و روشهای آماری**
ویراستار علمی و ادبی : **دکتر علی عمیدی**
حروف‌نگاری، نمونه‌خوانی، صفحه‌بندی، صفحه‌آرایی : **محبوبه کاظمی**
ویراستار هنری : **فرشید خان‌زاده**
طراحی جلد : **نیما دانش‌پرور**
مدیر فنی : **علی اصغر حائری مهریزی**
امور فنی و چاپ : **مؤسسهٔ انتشارات ستایش**

شمارهٔ ۵۲، خیابان شهید فکوری، خیابان باباطاهر، خیابان دکتر فاطمی
تهران ۱۴۱۳۷۱۷۹۱۱، ایران
URL: http://www.src.ac.ir
e-mail: src@src.ac.ir
تلفن: ۸۹۵۹۰۲۹ دورنگار: ۸۰۰۷۹۸۹

حروف‌نگاری شده با قلم‌های فارسی لوتوس و تیتر و میترا، و قلم لاتین Times New Roman.

چاپ و صحافی شده در ایران.
چاپ یکم
شمارگان: ۶۰۰

پیشنهاد برای نحوهٔ نقل مطلب، جدول یا نمودار از این کتاب، به صورت زیر است:
مختاری امیرمجدی، گیتی (۱۳۸۱). نمونه‌گیری: روشها و کاربردها. لِه‌وی، پل اس.؛ لِمی‌شو، استنلی. ترجمه از انگلیسی به فارسی. تهران: پژوهشکدهٔ آمار.

شابک ۹-۱۵۰-۳۶۵-۹۶۴
ISBN 964-365-150-9

بها: سی و پنج هزار ریال

قسمت ۱

مفاهیم پایه

فصل ۱

موارد استفاده از آمارگیریهای نمونه‌ای

۱.۱ چرا از آمارگیریهای نمونه‌ای استفاده می‌شود؟

اطلاعات مربوط به مشخصه‌های جامعه‌ها پیوسته موردنیاز سیاستمداران، بخشهای بازاریابی شرکتها، مأموران دولتی مسئول برنامه‌ریزی خدمات بهداشتی و اجتماعی و سایرین است. به دلایل مربوط به وقت و هزینه، این اطلاعات غالباً با استفاده از آمارگیریهای نمونه‌ای به دست می‌آیند. این قبیل آمارگیریها موضوع این کتاب‌اند.

در زیر مثالی از یک آمارگیری نمونه‌ای ارائه می‌شود که برای به دست آوردن اطلاعاتی دربارهٔ مشخصهٔ بهداشت در جامعه‌ای خاص اجرا شده است. ادارهٔ بهداشت یک ایالت بزرگ علاقمند است نسبت کودکان دبستانی آن ایالت را که در برابر بیماریهای عفونی اطفال (مانند فلج اطفال، دیفتری، کـــزاز، سیاه‌سرفه و غیره) ایمن‌سازی شده‌اند، تعیین کند. این کار به دلایل اداری باید فقط در مدت یک ماه به انجام برسد.

این کار که مستلزم هماهنگی دقیق تعداد زیادی از کارکنانی است که می‌کوشند اطلاعات را یا از طریق والدین و یا از طریق سوابق مصون‌سازی مدرسه‌ها در مورد یکایک کودکان دبستانی ساکن آن ایالت جمع‌آوری کنند در بدو امر بسیار مشکل به نظر می‌رسد. واضح است که بودجهٔ لازم برای چنین مسئولیتی به دلیل صرف وقت، هزینه‌های رفت و آمد و تعداد کودکان مشمول، هنگفت خواهد بود. حتی با به کارگیری تعداد قابل توجهی از کارکنان نیز انجام چنین تعهدی در چارچوب زمانی تعیین شده مشکل است.

این کتاب، برای پرداختن به مشکلاتی از قبیل آنچه در بالا رئوس آن مشخص شد، انواعی از روشها را برای انتخاب زیرمجموعه‌ای (*نمونه*) از مجموعهٔ اصلی همهٔ اندازه‌های (*جامعهٔ*) مورد علاقهٔ پژوهشگران ارائه خواهد کرد. این، اعضای نمونه هستند که مورد مصاحبه، مطالعه، یا اندازه‌گیری قرار می‌گیرند. مثلاً، در مورد مسأله‌ای که در بالا بیان شد، اثر ویژهٔ این‌گونه روشها این خواهد بود که برآوردهـای معتبـر و قابل‌اعتماد از نسبت کودکانی که در برابر این امراض مصون‌سازی شده‌اند در چارچوب زمانی تعیین شده و با کسری از هزینهٔ لازم برای به دست آوردن اطلاعات مربوط به یکایک کودکان دبستانی آن ایالت به دست خواهند آمد.

به بیان رسمیتر، *آمارگیری نمـونه‌ای* را می‌توان به عنوان مطالعـه‌ای مشتـمل بر زیرمجمـوعه (یا نمونه)ای از افراد تعریف کرد که از جامعه‌ای بزرگتر انتخاب می‌شوند. متغیرها یا مشخصه‌های موردنظر برای یکایک افراد نمونه‌گیری شده، مشاهده یا اندازه‌گیری می‌شوند. سپس این اندازه‌ها برای کل افراد نمونه جمع می‌شوند تا *آماره‌های* (برای مثال میانگینها، نسبتها و مجموعها) مربوط به نمونه به دست آیند. از این آماره‌هاست که می‌توان در رابطه با جامعه برون‌یابیهایی به عمل آورد. اعتبار و قابل اعتماد بودن این برون‌یابیها به این بستگی دارد که نمونه چقدر خوب انتخاب شده و اندازه‌گیریها چقدر خوب انجام شده باشند. این موارد، موضوع کتاب درسی حاضر را تشکیل می‌دهند.

وقتی همهٔ افراد جامعه برای اندازه‌گیری انتخاب شوند، بررسی را *سرشماری* می‌نامند. چون در سرشماری یکایک اعضای جامعه اندازه‌گیری می‌شوند، آماره‌های به دست آمده از سرشماری، برون‌یابی محسوب نمی‌شوند، اما اعتبار آماره‌های حاصل بستگی دارد به این که اندازه‌گیریها چقدر خوب انجام شده باشند. آمارگیریهای نمونه‌ای مزایای عمده‌ای که بر سرشماریها دارند بر هزینهٔ کمتر و سرعت بیشتر متکی است که با اندازه‌گیری از یک زیرمجموعه به جای اندازه‌گیری از کل جامعه میسر می‌شود. به علاوه، بررسیهای شامل مطالب پیچیده‌ای که مستلزم شیوه‌های اندازه‌گیریهای دقیق‌اند غالباً فقط درصورتی امکان‌پذیرند که نمونه‌ای از جامعه برای اندازه‌گیری انتخاب شود، زیرا در صورتی می‌توان منابع محدود را به اندازه‌گیریهای مشروح اختصاص داد که تعداد افراد مورد اندازه‌گیری بیش از اندازه زیاد نباشد.

در آمریکا، مانند بسیاری از کشورهای دیگر، به سازمانهای دولتی اختیار داده شده است تا برنامه‌هایی تهیه و اجرا کنند که طی آن از آمارگیریهای نمونه‌ای برای جمع‌آوری داده‌های مربوط به وضعیت اقتصادی، اجتماعی و بهداشتی مردم استفاده شود و از این داده‌ها برای مقاصد پژوهشی و نیز برای تصمیم‌گیری در سیاست‌گذاریها استفاده به عمل آید. مثلاً، مرکز ملی آمار بهداشتی[1] که مرکزی در

[1] National Center for Health Statistics (NCHS)

داخل وزارت بهداشت و خدمات انسانی آمریکا[1]ست طبق قانون اختیار دارد تا برنامهٔ آمارگیریهای ادواری و جاریی را به اجرا درآورد که برای به دست آوردن اطلاعات دربارهٔ بیماری، ازکارافتادگی و استفاده از خدمات مراقبتهای بهداشتی در آمریکا طراحی می‌شود [۱۲]. به همین ترتیب نمایندگیها، مراکز، یا دفاتری در داخل سایر وزارتخانه‌ها هستند (مانند دفتر آمار کار[2] در داخل وزارت کار[3]، مرکز ملی آمار آموزشی[4] در داخل وزارت آموزش[5]) که داده‌های مربوط به مأموریت اداره‌های خود را طی یک برنامهٔ آمارگیریهای نمونه‌ای جمع‌آوری می‌کنند. کارهای میدانی این آمارگیریها را غالباً دفتر سرشماری آمریکا[6] انجام می‌دهد که برنامه‌های آمارگیری خود را نیز دارد.

آمارگیریهایی که توسط این قبیل سازمانهای دولتی تنظیم می‌شوند غالباً طراحی بسیار پیچیده‌ای دارند و به کارکنان بسیار زیادی با مهارتهای عالی (و لذا، به بودجه‌های بسیار زیاد) برای اجرا نیازمندند. در حالی که نفس کار این سازمانهای دولتی ـ که تهیهٔ آماره‌های معتبر و قابل‌اعتماد در مورد انواع گوناگون پارامترها برای سراسر ایالات متحده و زیرگروههای گوناگون آن است ـ این بودجه‌های هنگفت را توجیه می‌کند. چنین هزینه‌هایی برای بیشتر نهادهایی که از آمارگیریهای نمونه‌ای استفاده می‌کنند به ندرت قابل توجیه یا کلاً امکان‌پذیرند. نیازهای اطلاعاتی بیشتر کاربران بالقوهٔ آمارگیریهای نمونه‌ای از لحاظ وسعت دامنهٔ آن، به مراتب محدودتر و بیشتر متمرکز بر مجموعهٔ نسبتاً کوچکی از پرسشهای خاص است. از این‌رو، انواع آمارگیریهایی که خارج از محدودهٔ دولت فدرال اجرا می‌شوند معمولاً از لحاظ طراحی ساده‌تر و بیشتر «یکبار مصرف» اند تا جاری. ما در این کتاب توجه خود را بر انواع این آمارگیریها متمرکز خواهیم کرد. ولی بحثهایی را نیز به آمارگیریهای نمونه‌ای پیچیده‌تر اختصاص خواهیم داد، به خصوص در فصل ۱۲ که روشهایی را برای برآورد کردن واریانس مورد بحث قرار می‌دهد که اصولاً برای تامین نیازهای آمارگیریهای دولتی بسیار پیچیده تهیه شده‌اند.

آمارگیریهای نمونه‌ای به ردهٔ وسیعتری از مطالعات غیرآزمایشی تعلق دارند که عموماً در نوشته‌های علوم اجتماعی یا بهداشتی به نام «مطالعات مشاهداتی» خوانده می‌شوند. بیشتر آمارگیریهای نمونه‌ای را می‌توان در ردهٔ مطالعات مشاهداتی که به عنوان «مطالعات مقطعی» شناخته شده‌اند جای داد. سایر انواع مطالعات مشاهداتی شامل مطالعات همگروهی و مطالعات مورد ـ شاهدند.

مطالعات مقطعی، «تصاویر کلی» از جامعه در یک لحظه از زمان است و هدفهای آن یا برآورد کردن میزان شیوع یا میانگین سطح برخی مشخصه‌ها در جامعه است و یا اندازه‌گیری رابطهٔ بین دو یا

[1] United States Department of Health and Human Services
[2] Bureau of Labor Statistics
[3] Department of Labor
[4] National Center for Educational Statistics
[5] Department of Education
[6] U.S. Bureau of the Census

چند متغیر است که در همان لحظه از زمان اندازه‌گیری می‌شوند.

مطالعات همگروهی و مورد ـ شاهد بیشتر به منظور تحلیل به کار می‌روند تا توصیف. مثلاً، این‌گونه مطالعات در شناخت بیماریهای واگیردار برای آزمونهای فرض مربوط به پیوند بین قرارگرفتن در معرض عوامل مخاطره و وقوع بیماریهای خاص به کار می‌روند.

از این طرحهای مطالعاتی در حد وسیعی برای شناخت روابط استفاده می‌شود. برای مثال، در دنیای کسب و کار ممکن است نمونه‌ای از حسابهای معوقه (مثلاً «مورد») همراه با نمونه‌ای از حسابهایی که عقب افتادگی ندارند (مثلاً «شاهد») اختیار شوند و مشخصه‌های هر گروه برای تعیین آن دسته از عواملی که با عقب افتادگی ارتباط دارند مقایسه شوند. مثالهای متعددی از این طرحهای مطالعاتی را می‌توان در زمینه‌های دیگر ارائه کرد.

به طوری که در بالا گفته شد، مطالعات همگروهی و مورد ـ شاهد با این هدف طراحی می‌شوند که حکمی (*یا فرضی*) در مورد مجموعه‌ای از متغیرهای مستقل (مانند عوامل مظنون به مخاطره) و متغیرهای وابسته (مانند وقوع بیماری) به آزمون گذاشته شود. این قبیل مطالعات در عین حال که اهمیت بسیار زیادی دارند موضوع این کتاب را تشکیل نمی‌دهند. نوع مطالعهٔ موردنظر این متن، غالباً به نام *آمارگیری توصیفی* شناخته شده، که هدف عمدهٔ آن، *برآورد* کردن سطح مجموعه‌ای از متغیرها در جامعه‌ای است که تعریف شده است. مثلاً در مثال فرضی که در آغاز این فصل ارائه شد، هدف عمده، با استفاده از یک نمونه، برآورد نسبت همهٔ کودکان دبستانی است که در مقابل بیماریهای اطفال ایمن‌سازی شده‌اند. در آمارگیریهای توصیفی توجه زیادی به انتخاب نمونه مبذول می‌شود زیرا برونیابی از نمونه به جامعه صورت می‌گیرد. با این که فرضها را می‌توان بر مبنای داده‌های جمع‌آوری شده از این گونه آمارگیریهای توصیفی به آزمون گذاشت ولی در این قبیل آمارگیریها این به طورکلی یک هدف ثانوی است. برآورد کردن تقریباً همیشه هدف اولیه است.

۲.۱ طراحی آمارگیریهای نمونه‌ای

در این بخش به بحث در مورد چهار مؤلفهٔ اصلی در طراحی آمارگیریهای نمونه‌ای می‌پردازیم. اینها طرح نمونه، اندازه‌گیریهای آمارگیری، عملیات آمارگیری، تحلیل آماری و تهیهٔ گزارش‌اند.

۱.۲.۱ طرح نمونه

در آمارگیری نمونه‌ای، مؤلفه‌های آماری عمده را *طرح نمونه* می‌نامند که هم برنامهٔ نمونه‌گیری و هم شیوهٔ برآورد را شامل می‌شود. *برنامهٔ نمونه‌گیری*، روش‌شناسیی است که برای انتخاب نمونه از جامعه مورد استفاده قرار می‌گیرد. *شیوه‌های برآورد*، الگوریتمها یا فرمولهایی هستند که برای به دست آوردن

برآوردهایی از مقدارهای جامعه از روی داده‌های نمونه و نیز برای برآورد کردن قابلیت اعتماد این برآوردهای جامعه‌ای به کار می‌روند.

انتخاب یک طرح نمونه‌ای خاص باید تلاشی دستجمعی باشد که در آن آماردانی که آمارگیری را طراحی خواهد کرد، اشخاصی که در اجرای آمارگیری شرکت خواهند نمود و کسانی که داده‌های حاصل از آمارگیری را مورد استفاده قرار خواهند داد همگی مشارکت داشته باشند. کاربران داده‌ها باید مشخص کنند که چه متغیرهایی باید اندازه‌گیری شوند، چه برآوردهایی موردنیازند، چه سطوحی از قابلیت اعتماد و اعتبار برای برآوردها لازم‌اند و چه محدودیتهایی از لحاظ به موقع بودن و هزینه‌ها بر آمارگیری مترتب است. افرادی که در اجرای آمارگیری مشارکت دارند باید اطلاعاتی دربارهٔ هزینهٔ کارکنان، زمان و مواد لازم و نیز اطلاعاتی دربارهٔ امکان‌پذیر بودن جایگزینی شیوه‌های دیگر نمونه‌گیری و شیوه‌های اندازه‌گیری فراهم کنند. آمارشناس با دریافت این اطلاعات می‌تواند طرحی نمونه‌ای پیشنهاد کند که ویژگیهای موردنیاز کاربران را با پایینترین هزینهٔ ممکن، تأمین نماید.

۲.۲.۱ اندازه‌گیریهای آمارگیری

درست به همان ترتیبی که نمونه‌گیری و برآورد در طرح آمارگیری نمونه‌ای مسئولیت آمارشناس است، مسئولیت انتخاب اندازه‌گیریهایی که باید به عمل آید و شیوه‌های انجام این اندازه‌گیریها به عهدهٔ افرادی است که در موضوع آمارگیری و در علوم اندازه‌گیری تخصص دارند. دستهٔ اول (که غالباً «کارشناس موضوعی» نامیده می‌شوند) اطلاعات اولیه را در تعیین اندازه‌گیریهایی که برای رسیدن به اهداف آمارگیری موردنیاز است فراهم می‌کنند. به محض اینکه اندازه‌گیریها مشخص شد، خبرگان اندازه‌گیری ـ که غالباً روانشناسان یا جامعه شناسانی هستند که در زمینهٔ تحقیقات آمارگیری آموزش خاص دیده‌اند و مهارت دارند ـ شروع به طراحی پرسشنامه‌ها یا فرمهایی می‌کنند که باید در گرفتن داده‌ها از افراد نمونه مورد استفاده واقع شوند. طراحی پرسشنامه یا سایر ابزار آمارگیری که برای جمع‌آوری داده‌های معتبر و قابل‌اعتماد مناسب باشند غالباً وظیفه‌ای بسیار پیچیده است. این کار مستلزم دقت قابل ملاحظه و گاهی اوقات مقداری مطالعهٔ اولیه است، به خصوص اگر برخی از متغیرهایی که قرار است اندازه‌گیری شوند قبلاً هرگز در هیچ فرایند آمارگیری دیگری اندازه‌گیری نشده باشند.

به محض اینکه پیش‌نویس ابزار آمارگیری تهیه شد، آمارشناس سهم خود را با توجه به شیوه‌های مورد استفاده برای ارزشیابی و اطمینان از کیفیت داده‌ها ارائه می‌کند. به علاوه، آمارشناس باید مطمئن شود که داده‌ها می‌توانند به آسانی کدگذاری و برای تحلیل‌آماری پردازش شوند و اطلاعات لازم برای روشهای برآورد را نیز فراهم کنند.

۳.۲.۱ عملیات آمارگیری

همین که نمونه انتخاب شد و ابزار اندازه‌گیری یا پیش‌نویس پرسشنامه تهیه شد، پیش‌آزمون شد، و اجرا شد، کار میدانی آمارگیری شامل جمع‌آوری داده‌ها را می‌توان آغاز کرد. ولی پیش از آنکه جمع‌آوری داده‌ها آغاز شود باید یک اجرای تمرینی یا *آمارگیری مقدماتی* از یک نمونهٔ کوچک با هدف به آزمون گذاشتن ابزار اندازه‌گیری و از میان بردن نقصهای قابل تشخیص در شیوه‌های آمارگیری انجام گیرد.

برای اینکه برآوردهای حاصل از آمارگیری معتبر و قابل‌اعتماد باشند اهمیت دارد که داده‌ها مطابق با طرح آمارگیری جمع‌آوری شوند و این وظیفهٔ افراد مسئول عملیات آمارگیری است که بر شیوه‌های جمع‌آوری داده‌ها نظارت و سرپرستی کنند. ماهیت کارکنان عملیات آمارگیری بستگی به اندازه و گسترهٔ آمارگیری نمونه‌ای، پیچیدگی اندازه‌گیریها و ماهیت آمارگیری (مثلاً آمارگیری یک‌بار مصرف در برابر آمارگیری جاری) دارد. برای مثال، آمارگیری برای معاینهٔ بهداشت ملی و تغذیه(NHANES)[1]، که یک سری آمارگیریهای نمونه‌ای پیچیده در کل کشور است که توسط مرکز ملی آمار بهداشتی به وسیلهٔ معاینات جسمی و مصاحبه اجرا می‌شود، کارکنان عملیاتی بسیار زیاد و بودجهٔ عملیاتی بسیار سنگینی دارد[۱۲،۱۱]. از سوی دیگر، یک آمارگیری نمونه‌ای که مجموعهٔ هدفهای محدودی دارد می‌تواند با کارکنان عملیاتی کمی اجرا شود.

٤.۲.۱ تحلیل آماری و تهیهٔ گزارش

پس از جمع‌آوری، کدگذاری، بازبینی و پردازش می‌توان داده‌ها را از نظر آماری تحلیل و یافته‌ها را در گزارش نهایی ارائه کرد. مانند تمام مؤلفه‌های آمارگیری نمونه‌ای، در تفسیر یافته‌های آمارگیری نیز دقت قابل توجهی باید مبذول شود. این یافته‌ها به صورت مشخصه‌های برآورد شدهٔ جامعه‌ای است که نمونه از آن گرفته شده است. ولی این برآوردها در معرض دو خطای نمونه‌گیری و اندازه‌گیری قرار دارند و هر تفسیری از این یافته‌ها باید این خطاها را نیز در نظر بگیرد. در بسیاری از پروژه‌هایی که مستلزم آمارگیری نمونه‌ای است، وقت و منابع مکفی به طراحی نمونه، تهیهٔ ابزار اندازه‌گیری یا پرسشنامه و عملیات آمارگیری اختصاص داده می‌شود، ولی برای تحلیل آماری نهایی و گزارش نویسی وقت و منابع بسیار کمی اختصاص می‌یابد. این وضعیت باعث تأسف است زیرا تأثیر یافته‌ها غالباً به‌خاطر عدم تلاش در این مرحلهٔ نهایی از دست می‌رود. در این کتاب، فصل ١٦ به تحلیل داده‌های ناشی از آمارگیری نمونه‌ای اختصاص یافته است.

[1] National Health and Nutrition Examination Survey

۳.۱ برنامه‌ریزی اولیهٔ آمارگیری نمونه‌ای

در بخش قبل به بحث در مورد مؤلفه‌های اصلی در آمارگیری نمونه‌ای پرداختیم. از آن مبحث باید آشکار شده باشد که آمارگیری نمونه‌ای می‌تواند تعهد سنگینی باشد که مستلزم صرف وقت و منابع بسیار زیادی چه از نظر مادی و چه از نظر انسانی است. همچنین باید آشکار شده باشد که در تصمیم‌گیری برای اجرا یا عدم اجرای آمارگیری نمونه ای باید توجه جدی مبذول شود و همین که تصمیم بر اجرا گرفته شد باید پیش از آن که کار روی طرح نمونه آغاز شود جداً در مورد فرمولبندی اهداف و ویژگیهای آمارگیری اندیشید.

کسانی که به فکر آمارگیری نمونه‌ای هستند در مرحلهٔ برنامه‌ریزی اولیه باید اهداف آمارگیری پیشنهادی را تنظیم کنند. این اهداف باید شامل ویژگیهای اطلاعاتی باشد که قرار است جمع‌آوری شوند و نیز شامل جامعه‌ای باشد که یافته‌های آمارگیری برای آن برونیابی خواهد شد. جایگزینهای آمارگیری، از قبیل تحلیل ثانوی داده‌هایی که قبلاً جمع‌آوری شده و هم‌اکنون موجود است نیزباید مورد بحث قرارگیرد. به موارد استفاده از داده‌های جمع‌آوری شده از آمارگیری پیشنهادی باید توجه دقیقی مبذول شود، به خصوص به تصمیمهایی که بر مبنای یافته‌های آمارگیری اتخاذ می‌شوند. در این مرحله مشخص خواهد شد که آمارگیری اصلاً ارزش اجرا دارد یا نه و اگر دارد برآوردهای حاصل تا چه حد باید دقیق باشند.

در مرحلهٔ برنامه‌ریزی اولیه باید به زیرحوزه‌هایی از جامعه (مانند گروههای سنی، گروههای جنسی، گروههای نژادی) که برآوردهایی برای آنها موردنیاز است و به سطح درستی موردنیاز برای این برآوردها اندیشید. دربارهٔ منابع موجود از نظر بودجه و کارکنان و نیز چارچوب زمانی داده‌های مورد نیاز باید بیشتر فکر کرد. حل این مسایل کمک می‌کند تا معلوم شود که برنامه‌ریزی و اجرای آمارگیری نمونه‌ای امکان‌پذیر هست یا نه و درصورتی که امکان‌پذیر تشخیص داده شود به تعیین پیکربندی لازم برای آمارگیری برحسب مؤلفه‌های آن کمک خواهد کرد.

تمرین

۱.۱ میانگینها، نسبتها و مجموعها مثالهایی هستند از:

الف. آماره‌های خلاصه

ب. افراد نمونه

پ. گزارشهای آمارگیری

ت. عناصر پایگاه اطلاعاتی

۲.۱ آمارگیریهای نمونه‌ای کمترین توجه را به این مورد دارند:

الف. تولید آماره‌های خلاصه

ب. تولید برآوردهای معتبر و قابل‌اعتماد

پ. توصیف مشخصه‌های جامعه

ت. آزمودن فرضها

۳.۱ دربارهٔ آنچه در طرح نمونه گنجانده می‌شود، کدام یک از موارد زیر از همه صحیحتر است؟

الف. برنامهٔ نمونه‌گیری و گزارشهای آماری

ب. برنامهٔ نمونه‌گیری و شیوه‌های برآورد

پ. برنامهٔ نمونه‌گیری و برآوردهای هزینه

ت. شیوه‌های برآورد کردن و اقدامات کنترل کیفیت

٤.۱ آمارگیریهای نمونه‌ای به ردهٔ بزرگتری از مطالعات تعلق دارند به نام:

الف. مطالعات همگروهی

ب. مطالعات مشاهداتی

پ. مطالعات مورد ـ شاهد

ت. شبه آزمایشها

۵.۱ کدام یک از موارد زیر برآوردهای حاصل از سرشماری را بهتر از همه توصیف می‌کند؟

الف. هم خطاهای نمونه‌گیری دارد و هم خطاهای اندازه‌گیری

ب. خطاهای نمونه‌گیری دارد ولی خطاهای اندازه‌گیری ندارد

پ. خطاهای اندازه‌گیری دارد ولی خطاهای نمونه‌گیری ندارد

ت. نه خطاهای نمونه‌گیری دارد و نه خطاهای اندازه‌گیری

٦.۱ پس‌خور از بررسی مقدماتی معمولاً کدام یک از مزایای زیر را به دست می‌دهد؟

الف. پایین آوردن خطاهای اندازه‌گیری

ب. پایین آوردن خطاهای نمونه‌گیری

پ. کاهش هزینه‌ها

ت. همهٔ موارد بالا

۷.۱ شما مقام اجرایی اصلی یک بیمارستان هستید و می‌خواهید در مدت بسیار کوتاهی از نسبت پذیرش همهٔ بیماران بستری تحت پوشش طرف سومی غیر از بیمهٔ مراقبتهای پزشکی[1] و بیمهٔ کمکهای پزشکی[2] در طول سال ۱۹۹۸ آگاه شوید. برای تعیین این نسبت چگونه اقدام خواهید کرد؟

۸.۱ شما به عنوان همان مقام اجرایی اصلی در تمرین ۷.۱ می‌خواهید میانگین هزینه‌های نقدی به ازای پذیرش هر بیمار بستری را برآورد کنید. برای تعیین این برآورد چه خواهید کرد؟

کتابشناسی

The following general texts in sampling theory have been used for many years and can give the reader additional perspectives on sampling. The first two have been reissued in the Wiley "Classics" Series.

1. Cochran, W. G., *Sampling Techniques.* 3rd ed., Wiley, New York, 1977.
2. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods and Theory,* Vols. 1 and 2, Wiley, New York, 1 953.
3. Kish, L., *Survey Sampling*, Wiley, New York, l 965.
4. Mendenhall, W., Ott, L., and Scheaffer, R. L., *Elementary Survey Sampling,* Duxbury Press, Belmont, Calif., 197 l .

The text by Cochran [1] emphasizes the theoretical development of sampling methodology, but at the same time gives the reader a sense of how the methods are used. The text by Hansen et al. [2] uses the notation and approaches taken historically by the major government agencies involved in sample surveys (e.g., Bureau of the Census, Bureau of Labor Statistics, etc.). The book by *Kish [3] is comprehensive and contains a wide variety of advanced sampling techniques. Mendenhall et al. [4] give a* very *readable presentation of basic sampling techniques.*

The following more recent texts are much more specialized. The text by Sudman [5] addresses a wide variety of issues that are often encountered in the planning of sample surveys. Hedayat and Sinha [6] have written a book that emphasizes theoretical issues. The recent books by Thompson [7] and by Thompson and Seber [8] discuss in detail methods of sampling that are used in the estimation of wildlife populations and in other areas of ecology. The book by Kalton [9] gives a highly readable introduction to sampling that can be valuable to nonstatisticians as well as statisticians. It reflects the author's wide experiences as a sampling statistician. The very *recent text by Lehtonen and Pahkinen [10] places a high emphasis on methods for analysis of data from complex sample surveys.*

5. Sudman, S., *Applied Sampling,* Academic Press, New York, 1976.
6. Hedayat, A. S., and Sinha, B. K., *Design and Inference in Finite Population Sampling,* Wiley, New York, 1991.

[1] Medicare
[2] Medicaid

7. Thompson, S. K., *Sampling,* Wiley, New York, 1992.
8. Thompson, S. K., and Seber, G., *Adaptive Sampling,* Wiley, New York, l996.
9. Kalton, G., *Introduction to Survey Sampling,* Sage University Paper 35: Quantitative Applications in the Social Sciences, 07-035. Sage Publications, Newbury Park, Calif., 1989.
10. Lehtonen, R., and Pahkinen, E. J. *Practical Methods for Design and Analysis of Complex Surveys,* Rev. Ed., Wiley, Chichester, U.K., l 994.

The following relate to the major surveys conducted by the National Center for Health Statistics (NCHS).

11. National Center for Health Statistics, *Catalog of Publications 1980-1987,* U.S. Department of Health and Human Services, Hyattsville, Md., 1988.
12. National Center for Health Statistics, *Plan and Operation of the Health and Nutrition Examination Survey. United States, 1971-1973,* Vital and Hea1th Statistics, Series 1, No. 10a. DHEW Publication No. (HRA) 76-1310, U.S. Government Printing Office, Washington, D.C., 1973.

Both the Encyclopedia of Statistical Sciences [*13] and the more recent* Encyclopedia of Biostatistics [*14] contain expository articles on many topics in survey sampling. Volume 6 in the* Handbook of Statistics *series [15] deals exclusively with sampling methodology and contains 24 chapters authored by experts on various aspects of sampling. Chapter 1*, w*ritten by D. R. Bellhouse [16], gives an excellent historical overview of sampling methodology.*

13. *Kotz,* S., and Johnson, N. L., *Encyclopedia of Statistical Sciences,* Vol. 76, Wiley, New York, 1986.
14. Armitage, P., and Coltons T., Eds., The *Encyclopedia of Biostatistics,* Wiley, Chichester, U.K., l 998.
15. Krishnaiah, P. R., and Rao, C. R., Eds. *Handbook of Statistics,* Vol. 6, *Sampling,* Elsevier, Amsterdam and New York, 1988.
16. Bellhouse, D. R., A brief history of random sampling methods. In *Handbook of Statistics,* Vol. 6, *Sampling,* Krishnaiah, P. R., and Rao, C. R., Eds., Elsevier, Amsterdam and New York, pp. 1-14, 1988.

فصل ۲

جامعه و نمونه

در فصل قبل دربارهٔ آمارگیریهای نمونه‌ای به عنوان مطالعاتی بحث کردیم که توزیع و سطوح مشخصه‌های جامعه را به وسیلهٔ اندازه‌گیریهایی از زیرمجموعهٔ افراد انتخاب شده از آن جامعه برآورد می‌کنند. در این فصل ابتدا با تعریف مؤلفه‌های جامعه برحسب اینکه نسبت به گرفتن نمونه از آن با معنا باشند به گسترش بنیانهای روش‌شناسی نمونه‌گیری می‌پردازیم. پس از این که این قبیل خواص جامعه تعریف و از آنها بحث شد بسط روش‌شناسی نمونه‌گیری را آغاز می‌کنیم.

۱.۲ جامعه

جامعه (یا جمعیت هدف)، کل مجموعهٔ افرادی است که یافته‌های آمارگیری قرار است برای آن برونیابی شوند. در این کتاب از واژه‌های جامعه و جامعهٔ هدف به جای یکدیگر و به یک معنی استفاده می‌کنیم.

هر یک از اعضای جامعه‌ای که قرار است مشخصه‌های آن اندازه‌گیری شود *واحدهای اولیه* یا *عناصر جامعه* نامیده می‌شوند. برای مثال، اگر در حال اجرای یک آمارگیری نمونه‌ای هستیم که اهداف آن برآورد کردن تعداد اشخاص ساکن ایلی‌نوی است که هیچ وقت به دندانپزشک مراجعه نکرده‌اند، همهٔ اشخاص ساکن ایلی‌نوی جامعه را تشکیل می‌دهند و هر یک از اشخاص ساکن ایلی‌نوی یک واحد اولیه یا یک عنصر است. اگر یک آمارگیری نمونه‌ای دربارهٔ سوابق پزشکی بیمارستانی اجرا

می‌کنیم که هدف از آن برآورد تعداد مرخص‌شدگان با بیماری خاص از بیمارستان در یک سال معین باشد، هر یک از مرخص‌شدگان در طی سال یک عنصر است و کل مرخص‌شدگان از این دست، جامعه را تشکیل می‌دهند.

در اجرای آمارگیریهای نمونه‌ای غالباً این امکان وجود ندارد که از واحدهای اولیه مستقیماً نمونه گرفته شود. زیرا فهرستهای واحدهای اولیه که بتوان از آن نمونه گرفت غالباً به راحتی در دسترس نیستند و فقط با صرف هزینهٔ قابل ملاحظه‌ای فراهم می‌شوند. ولی غالباً می‌توان واحدهای اولیه را با سایر انواع واحدهایی مرتبط ساخت که یا فهرست آنها موجود است یا می‌تواند برای مقاصد نمونه‌گیری به آسانی ساخته شود. این نوع واحدها را *واحدهای شمارش* یا *واحدهای فهرست‌برداری* می‌نامند. یک واحد شمارش یا واحد فهرست‌برداری ممکن است شامل یک یا چند واحد اولیه باشد و می‌تواند پیش از انتخاب نمونه مشخص شود. مثلاً، فرض کنید یک آمارگیری نمونه‌ای در غرب ماساچوست در دست برنامه‌ریزی است که هدف از آن تعیین تعداد اشخاص ساکن در ناحیهٔ همپشایر است که در برابر سرخک ایمن‌سازی شده‌اند. در این مورد، جامعهٔ آماری شامل همهٔ اشخاص ساکن ناحیه است و واحدهای اولیه نیز همهٔ اشخاص ساکن ناحیه‌اند. بسیار نامحتمل است که فهرست دقیق و روزآمدی از تمام اشخاص ساکن این ناحیه موجود یا به راحتی قابل تهیه باشد. اگر بود، با استفاده از آن فهرست می‌توان نمونه‌ای انتخاب کرد. ولی قابل تصور است که فهرستی از همهٔ خانوارهای ناحیه موجود یا حداقل بدون زحمت یا هزینهٔ زیاد قابل تهیه باشد. اگر چنین فهرستی موجود باشد می‌توان نمونه‌ای از خانوارها را انتخاب کرد و اشخاصی را که در این خانوارها زندگی می‌کنند به عنوان واحدهای اولیهٔ نمونه در نظر گرفت. خانوارها خود واحدهای شمارش‌اند.

اگر قرار است نمونه از یک فهرست واحدهای شمارش انتخاب شود لازم است واحدهای اولیه‌ای که در نظر است با هر واحد شمارش مرتبط شوند با الگوریتمی مشخص شوند. چنین الگوریتمی را *قاعدهٔ شمارش* یا *قاعدهٔ شمردن* می‌نامند. در آمارگیری فوق‌الذکر خانوارها برای برآورد کردن تعداد اشخاص ساکن در ناحیهٔ همپشایر که در برابر سرخک ایمن‌سازی شده‌اند، اشخاص ساکن در ناحیه، واحدهای اولیه را تشکیل می‌دهند و خانوارها واحدهای شمارش‌اند. قاعدهٔ شمردن در این مثال ممکن است مشخص کند که همهٔ افراد ساکن در یک خانوار خاص با آن خانوار مرتبط‌اند.

در مثال مربوط به ناحیهٔ همپشایر، قاعدهٔ شمردن واضح و سرراست است. ولی گاهی اوقات قاعدهٔ شمردن که واحدهای اولیه را با واحدهای شمارش مرتبط می‌سازد چندان آشکار و واضح نیست. مثلاً، فرض کنید می‌خواهیم تعداد افرادی را برآورد کنیم که در کالیفرنیا به بیماری نسبتاً جدی و نادری

همچون زخم آکله بافت‌ها (SLE)[1] دچار شده‌اند. در این مورد، معقول آن است که از تأمین‌کنندگان مراقبتهای بهداشتی (مانند پزشکان و بیمارستانها) یک آمارگیری نمونه‌ای به عمل آوریم و اطلاعات مربوط به افرادی را که به خاطر این بیماری خاص تحت مراقبت آنها قرار گرفته‌اند به دست آوریم. از آنجا که شخص مبتلا به بیماری ممکن است از سوی چند منبع تحت مراقبت قرار گرفته باشد یا بگیرد، بیش از یک راه منطقی برای مرتبط ساختن موارد (واحدهای اولیه) با منابع (واحدهای شمارش) وجود دارد. مثلاً، یک قاعدهٔ شمردن ممکن است اجازه دهد که یک مورد به تأمین‌کنندهٔ مراقبتهای بهداشتی (منبع) مرتبط شود که در زمان آمارگیری، مسئولیت مراقبت اصلی از آن شخص را به عهده داشته است. قاعدهٔ شمارش دوم ممکن است این اجازه را بدهد که مورد به تمام منابعی که تا آن هنگام به مداوای آن شخص پرداخته‌اند مرتبط گردد. قاعدهٔ سوم ممکن است مورد را به منبعی مرتبط سازد که اول از همه بیماری را در آن شخص تشخیص داده است. در دو دههٔ گذشته حجم فزاینده‌ای از کارها بر این تحقق استوار بوده است که انتخاب یک قاعدهٔ شمارش مناسب، می‌تواند قابلیت اعتماد برآوردها را در آمارگیری نمونه‌ای به صورتی قابل ملاحظه بهبود بخشد [۳].

هدف اولیهٔ تقریباً هر آمارگیری نمونه‌ای، برآورد مقادیر خاص مربوط به توزیع مشخصه‌های ویژهٔ یک جامعه است. این مقادیر بیشتر اوقات به صورت میانگینها، مجموعها یا مجموع تجمعی، و نسبتها یا درصدها هستند. همچنین ممکن است این مقادیر به صورت صدکها، انحراف معیارها، یا سایر پارامترهای توزیع باشند. مثلاً در یک آمارگیری خانوار که در آن نمونه‌ای از ساکنان شیکاگو گرفته می‌شود ممکن است بخواهیم تعداد متوسط موارد وقوع بیماریهای حاد به‌ازای هر فرد (میانگین جامعه)، کل روزهای کار یا تحصیل که به خاطر بیماری حاد در بین همهٔ اعضای جامعه تلف شده است (مجموع جامعه یا مجموع تجمعی)، و نسبت اشخاصی که در سال گذشته دو یا چند بیماری حاد داشته‌اند (نسبت جامعه‌ای) را برآورد کنیم. همچنین ممکن است بخواهیم میانهٔ هزینهٔ سالانهٔ خانوار را برای مراقبتهای بهداشتی و انحراف معیار توزیع روزهای کاری تلف شده به دلیل بیماری حاد را برآورد کنیم. در مبحثی که متعاقباً خواهد آمد، نمادهای رسمی را برای بحث در مورد مفاهیم بالا شرح خواهیم داد.

۱.۱.۲ واحدهای اولیه

تعداد واحدهای اولیه در جامعه با حرف N نمایش داده می‌شود و هر یک از واحدهای اولیه با برچسبی به شکل یک شماره از ۱ تا N شناسایی خواهد شد.هر *مشخصه* (یا *متغیر*) با یک حرف

[1] Systemic Lupus Erythematosis

مانند x یا y نشان داده خواهد شد. مقدار یک مشخصه x در واحد اولیهٔ iام را به صورت X_i نشان می‌دهند. مثلاً، در آمارگیری مربوط به مرخص‌شدگان بیمارستانها که در یک بیمارستان برای سال خاصی اجرا می‌شود، جامعهٔ آماری ممکن است مجموعهٔ مرخص‌شدگان از بیمارستان در طی مدت زمان موردنظر باشد و هر مرخص شده یک واحد اولیه خواهد بود. اگر در آن سال ۲۰۰۰ بیمار مرخص شده باشند، در آن صورت N برابر با ۲۰۰۰ خواهد بود و به هر مورد مرخص شدن (برای مقاصد آمارگیری) یک برچسب شناسایی به شکل یک شماره از ۱ تا ۲۰۰۰ داده خواهد شد. فرض کنیم به توزیع تعداد روزهای بستری بودن (طول مدت اقامت) در بین مرخص‌شدگان علاقه‌مند باشیم. در این صورت X_1 نشان‌دهندهٔ طول مدت اقامت برای مرخص‌شدهٔ دارای برچسب «۱» از بیمارستان و X_2 طول مدت اقامت برای مرخص‌شدهٔ دارای برچسب «۲» از بیمارستان و همین طور الی‌آخر خواهد بود.

۲.۱.۲ پارامترهای جامعه

گفتیم که اهداف آمارگیری شامل برآورد کردن *مقادیر* معینی از توزیع یک متغیر یا مشخصهٔ تعیین شده در جامعه است. این مقادیر برای جامعه پارامتر محسوب می‌شوند و پارامتر برای یک جامعهٔ معین، ثابت است. تعاریف آن دسته از پارامترهایی که بیشتر اوقات مایل به برآورد آن هستیم در مبحث زیر آمده است.

مجموع کل جامعه

مجموع کل جامعه با مشخصهٔ x معمولاً با حرف X نشان داده می‌شود و مجموع مقادیر آن مشخصه برای کل عناصر جامعه است. مجموع جامعه از فرمول زیر به دست می‌آید:

$$X = \sum_{i=1}^{N} X_i \qquad (۱.۲)$$

میانگین جامعه

میانگین جامعه نسبت به مشخصهٔ x از فرمول زیر به دست می‌آید:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (۲.۲)$$

نسبت جامعه

هرگاه مشخصهٔ مورد اندازه‌گیری نمایانگر بودن یا نبودن نوعی صفت کیفی *دوحالتی* باشد غالباً مطلوب آن است که نسبت واحدهای اولیهٔ دارای آن صفت در جامعهٔ آماری برآورد شود. اگر صفت کیفی با

حرف x نشان داده شود و X تعداد کل واحدهای اولیهٔ دارای آن صفت در جامعه باشد، در آن صورت P_x نشان‌دهندهٔ *نسبت جامعه‌ای* عناصر دارای آن صفت خواهد بود و از فرمول زیر محاسبه می‌شود:

$$P_x = \frac{X}{N} \qquad (۳.۲)$$

باید توجه داشت که برای وضعیت خاصی که در آن متغیر x با استفاده از فرمول زیر به دست می‌آید نسبت جامعه‌ای، یک میانگین جامعه‌ای است:

$$X_i = \begin{cases} ۱ & \text{اگر صفت } x \text{ در عنصر } i \text{ وجود داشته باشد} \\ ۰ & \text{اگر صفت } x \text{ در عنصر } i \text{ وجود نداشته باشد} \end{cases}$$

به این ترتیب $X = \sum_{i=1}^{N} X_i$ نمایانگر کل تعداد عناصر دارای آن صفت خواهد بود.

واریانس و انحراف معیار جامعه

واریانس و انحراف معیار توزیع یک مشخصه در جامعه به این دلیل مورد توجه‌اند که *پراکندگی* توزیع را اندازه می‌گیرند. *واریانس جامعه* با مشخصهٔ x با نماد σ_x^2 نشان داده می‌شود و از فرمول زیر به دست می‌آید:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} \qquad (٤.۲)$$

انحراف معیار جامعه که با نماد σ_x نشان داده می‌شود صرفاً ریشهٔ دوم واریانس است و از فرمول زیر به دست می‌آید:

$$\sigma_x = \left(\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} \right)^{1/2} \qquad (٥.۲)$$

وقتی مشخصهٔ موردنظر یک صفت کیفی دوحالتی است می‌توان نشان داد که واریانس جامعه به صورتی که در بالا تعریف شد به فرمول زیر تبدیل می‌شود:

$$\sigma_x^2 = P_x(۱ - P_x) \qquad (٦.۲)$$

برای راحتی کار در ارجاعهای بعدی، همهٔ فرمولهای بالا در تابلوی ۱.۲ خلاصه شده‌اند.

حال ببینیم از این فرمولها در عمل چگونه استفاده می‌شود.

مثال تشریحی: فرض کنید می‌خواهیم توزیع ویزیت بیماران را در خانوار توسط پزشکان در یک سال معین به دست آوریم. به ۲۵ پزشک محل از ۱ تا ۲۵ شماره داده شده و تعداد مراجعهٔ هر پزشک به خانوار بیماران در جدول ۱.۲ نشان داده شده است.

در این مورد، واحدهای اولیه را پزشکان تشکیل می‌دهند و تعداد آنها ۲۵ نفر است. به عبارت دیگر N=۲۵ . اگر X_i برابر با تعداد ویزیتهای پزشک i باشد، میانگین جامعه، مجموع کل، واریانس، و انحراف معیار به شرح زیر خواهند بود (فرمولها را در تابلوی ۱.۲ ببینید):

$\overline{X}$ = ۵/۰۸ ویزیت $\qquad$ σ_x^2 = ۶۷/۹۱ (ویزیت)^2

X = ۱۲۷ ویزیت $\qquad$ σ_x = ۸/۲۴ ویزیت

اگر y را نمایانگر صفت کیفی انجام یک یا چند ویزیت در طی مدت زمان تعیین شده بگیریم، خواهیم داشت:

$$P_y = \frac{۱۴}{۲۵} = ۰/۵۶$$

که در آن P_y نسبت پزشکان جامعهٔ آماری است که در طول زمان موردنظر یک یا چند بیمار را در خانوار بیمار ویزیت کرده‌اند. همچنین داریم

$$\sigma_y^2 = (۰/۵۶) \times (۱ - ۰/۵۶) = ۰/۲۴۶ \qquad \sigma_y = \sqrt{۰/۲۴۶} = ۰/۴۹۶$$

یک پارامتر دیگر جامعه که در نظریهٔ نمونه‌گیری حایز اهمیت است *ضریب تغییرات* است که با نماد V_x نشان داده می‌شود. ضریب تغییرات یک توزیع عبارت است از نسبت انحراف معیار آن توزیع به میانگین توزیع:

$$V_x = \frac{\sigma_x}{\overline{X}} \qquad (۷.۲)$$

ضریب تغییرات نمایانگر پراکندگی توزیع نسبت به میانگین توزیع است. در مورد توزیع ویزیت از خانوارها در مثال تشریحی بالا، ضریب تغییرات از فرمول زیر به دست می‌آید:

$$V_x = \frac{۸/۲۴}{۵/۰۸} = ۱/۶۲$$

می‌توان نشان داد که ضریب تغییرات V_y برای صفت دو حالتی y از فرمول زیر به دست می‌آید:

$$V_y = \left(\frac{۱ - P_y}{P_y}\right)^{۱/۲} \qquad (۸.۲)$$

برای توزیع پزشکان نسبت به صفت کیفی اجرای یک یا چند ویزیت در خانوار بیمار طی مدت زمان مشخص (مثال تشریحی را ببینید)، ضریب تغییرات به صورت زیر است:

$$V_y = \left(\frac{۰/۴۴}{۰/۵۶}\right)^{۱/۲} = ۰/۸۸۶$$

تابلوی ۱.۲ پارامترهای جامعه

مجموع

$$X = \sum_{i=1}^{N} X_i \qquad (۱.۲)$$

میانگین

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (۲.۲)$$

نسبت

$$P_x = \frac{X}{N} \qquad (۳.۲)$$

واریانس

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} \qquad (٤.۲)$$

انحراف معیار

$$\sigma_x = \left(\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} \right)^{1/2} \qquad (۵.۲)$$

واریانس، صفت کیفی دوحالتی

$$\sigma_x^2 = P_x(1 - P_x) \qquad (٦.۲)$$

در این تعریفها N، تعداد واحدهای اولیه در جامعه و X_i، مقدار واحد اولیهٔ i ام است.

مربع ضریب تغییرات، V_x^2 است که به عنوان واریانس نسبی شناخته می‌شود و پارامتری است که در روش‌شناسی نمونه‌گیری از آن استفاده گسترده‌ای می‌شود.

جدول ۱.۲ تعداد ویزیت بیماران در خانوار توسط پزشک در سالی مشخص

پزشک	تعداد ویزیت	پزشک	تعداد ویزیت
۱	۵	۱٤	٤
۲	۰	۱۵	۸
۳	۱	۱٦	۰
٤	٤	۱۷	۷
۵	۷	۱۸	۰
٦	۰	۱۹	۳۷
۷	۱۲	۲۰	۰
۸	۰	۲۱	۸
۹	۰	۲۲	۰
۱۰	۲۲	۲۳	۰
۱۱	۰	۲٤	۱
۱۲	۵	۲۵	۰
۱۳	٦		

□

مثال تشریحی: برای نشان دادن استفاده از ضریب تغییرات به‌عنوان یک آمارهٔ توصیفی، فرض می‌کنیم می‌خواهیم شناختی پیدا کنیم که آیا سطوح کلسترول در یک جامعه، از متغیرهای فشار خون سیستولی در همان جامعه متغیرتر است یا نه. فرض می‌کنیم میانگین سطح فشار خون سیستولی در جامعه ۱۳۰ میلی‌متر Hg (میلی‌متر جیوه) و انحراف معیار ۱۵ میلی‌متر Hg باشد. همچنین فرض می‌کنیم میانگین سطح کلسترول ۲۰۰ میلی‌گرم بر ۱۰۰ میلی‌متر و انحراف معیار ٤۰ میلی‌گرم بر۱۰۰ میلی‌متر است. مشاهدهٔ انحراف معیارهای مربوط به آنها به شکلی با معنا مشخص نمی‌کند که کدام مشخصه در جامعه تغییرپذیری بیشتری دارد، زیرا این انحراف معیارها بر حسب واحدهای متفاوتی اندازه‌گیری شده‌اند (در این مورد میلی‌متر جیوه در مقابل میلی‌گرم بر۱۰۰ میلی‌متر). ولی این دو متغیر را می‌توان با مشاهدهٔ ضریب تغییرات مربوط به هر یک از آنها مقایسه کرد، یعنی $\frac{۱۵}{۱۳۰}$ یا ۰/۱۱۵ برای فشار خون سیستولی در مقابل $\frac{۴۰}{۲۰۰}$ یا ۰/۲۰۰ برای کلسترول. ضرایب تغییرات را می‌توان مقایسه کرد زیرا آنها بدون بعدند. به این ترتیب، چون ضریب تغییرات سطح کلسترول بیشتر از ضریب تغییرات فشار خون سیستولی است می‌توان نتیجه گرفت که در این جامعه کلسترول نسبت به فشار خون سیستولی تغییرپذیری بیشتری دارد.

□

در مثال بالا نمی‌توان از انحراف معیار برای مقایسهٔ تغییرپذیری دو متغیر استفاده کرد زیرا متغیرها با واحدهای اندازه‌گیری یکسان اندازه‌گیری نشده‌اند. حالا دو متغیر را درنظر می‌گیریم که با واحدهای اندازه‌گیری یکسان اندازه‌گیری شده باشند مثل فشار خون سیستولی و فشار خون دیاستولی. فرض کنیم میانگین فشار خون دیاستولی در جامعه ۶۰ میلی‌متر جیوه و انحراف معیار آن برابر با ۸ میلی‌متر جیوه و میانگین و انحراف معیار فشار خون سیستولی همان مقدار داده شده در مثال قبلی باشد (یعنی $\overline{X} = ۱۳۰$ میلی‌متر جیوه و $\sigma_x = ۱۵$ میلی‌متر جیوه). پس به‌طور مطلق، فشار خون سیستولی متغیرتر از فشار خون دیاستولی است ($\sigma_x = ۱۵$ میلی‌متر جیوه برای سیستولی در مقابل ۸ میلی‌متر جیوه برای دیاستولی). ولی به‌طور نسبی، همانطور که اندازه‌گیری ضریب تغییرات نشان می‌دهد، تغییرپذیری فشار خون دیاستولی بیشتر است. ضریب تغییرات برای دیاستولی $\frac{۸}{۶۰}$ یا ۰/۱۳۳ و برای سیستولی $\frac{۱۵}{۱۳۰}$ یا ۰/۱۱۵ است. در طراحی آمارگیری‌های نمونه‌ای، غالباً تغییرات نسبی بیش از تغییرات مطلق مورد توجه‌اند ـ از این رو اهمیت ضریب تغییرات آشکار می‌شود.

۲.۲ نمونه

در بخش ۱.۲ مفاهیم مربوط به جامعه یا جامعهٔ هدف را ارائه کردیم. تأکید می‌کنیم که پارامترهای جامعه‌ای بحث شده در بخش ۱.۲ ـ از قبیل میانگین سطح یک مشخصه، مقدار کل یک مشخصه در جامعه، یا نسبت عناصر موجود در جامعه با برخی صفات کیفی مشخص شده ـ تقریباً همیشه ناشناخته‌اند. از این‌رو، اهداف اولیهٔ آمارگیری نمونه‌ای، گرفتن نمونه‌ای از جامعه و برآورد پارامترهای جامعه از روی آن نمونه است. در این بخش، مفاهیم خاصی را در رابطه با نمونه‌ها ارائه می‌کنیم و در مورد چگونگی برآورد پارامترهای جامعه از روی نمونه بحث می‌کنیم.

۱.۲.۲ نمونه‌گیری احتمالاتی و غیراحتمالاتی

آمارگیریهای نمونه‌ای را می‌توان بر اساس چگونگی انتخاب نمونه در دو ردهٔ بسیار وسیع رسته‌بندی کرد، یعنی نمونه‌های احتمالاتی و نمونه‌های غیراحتمالاتی. نمونهٔ احتمالاتی این مشخصه را دارد که هر یک از عناصر جامعه دارای احتمالی معلوم و غیرصفر برای گنجانیده شدن در نمونه است. نمونهٔ غیراحتمالاتی نمونه‌ای است مبتنی بر برنامهٔ نمونه‌گیری که این خصیصه را ندارد. در نمونه‌گیری احتمالاتی، چون هر یک از عناصر شانس معلومی برای انتخاب شدن دارد، برآوردهای نااریب پارامترهای جامعه را که تابعهایی خطی از مشاهدات‌اند (مانند میانگینهای جامعه‌ای، مجموعها و نسبتها) می‌توان از روی داده‌های نمونه ساخت. همچنین، تحت این شرط که احتمالهای شمول مرتبه دوم

(یعنی احتمال توأم گنجاندن هر یک از واحدهای دو شمارشی) معلوم باشند، خطاهای معیار این بر آوردها را نیز می‌توان برآورد کرد. این به کاربران برآوردهای آمارگیری شناخت می‌دهد که تا چه اندازه می توانند روی ارزش برآوردها حساب کنند. از سوی دیگر، نمونه‌گیری غیراحتمالاتی این خصیصه را ندارد و کاربران هیچ روش مطمئنی برای ارزشیابی اعتبار یا قابل اعتماد بودن برآوردهای حاصل ندارند. این موارد و مفاهیم بعداً در همین فصل مورد بحث قرار خواهند گرفت.

نمونه‌های غیراحتمالاتی به خصوص در تحقیقات بازار و آمارگیری از آرای عمومی بسیار فراوان به کار می‌روند. از نمونه‌های غیراحتمالاتی به این دلیل استفاده می‌شود که نمونه‌گیری احتمالاتی غالباً شیوه‌ای وقت‌گیر و پرهزینه است و در واقع ممکن است در بسیاری از وضعیتها امکان‌پذیر نباشد. مثالی از نمونه‌گیری غیراحتمالاتی، آمارگیری نمونه‌ای به اصطلاح سهمیه‌ای است که در آن به مصاحبه‌گران گفته می‌شود که با تعدادی معین از افراد در زیرگروههای جمعیتی معین تماس بگیرند و مصاحبه کنند. مثلاً ممکن است به یک مصاحبه‌گر گفته شود که با پنج مرد سیاه‌پوست، پنج زن سیاه‌پوست، ده مرد سفیدپوست و ده زن سفیدپوست مصاحبه کند و انتخاب افراد خاص در داخل هر یک از این رده‌ها به عهدهٔ مصاحبه‌گر گذاشته شود. بسیار محتمل است که چنین روش انتخاب نمونه به برآوردهایی با اریبی بسیار زیاد منجر شود. برای مثال، مصاحبه‌گر ممکن است برای راحتی کار خود هر پنج مرد سیاه‌پوست و هر پنج زن سیاه‌پوست را از محله‌هایی که از نظر اقتصادی و اجتماعی در سطح بالاتری هستند انتخاب کند که ممکن است نمایندهٔ همهٔ اقلیت سیاه‌پوست نباشند.

نوع دیگری از نمونه‌گیری غیراحتمالاتی که گاهی اوقات به کار می‌رود به نام نمونه‌گیری *قصدی* یا مبتنی بر *داوری* مشهور است. در این نوع نمونه‌گیری، افرادی انتخاب می‌شوند که به نظر می‌رسند بیشترین نمایندگی در کل جامعه را دارند. برای مثال، ممکن است بخواهیم تعداد کل نمونه‌های خون را که طی یک سال معین در یک درمانگاه بیماران سرپایی گرفته شده است با انتخاب چند روز «عادی» و بررسی سوابق درمانگاه در همان روزهای نمونه‌گیری شده برآورد کنیم. اگر منابعی در اختیار داشتیم که می‌توانستیم فقط چند روز را نمونه‌گیری کنیم، این روش به برآوردهای معتبرتر و قابل اعتمادتری منجر می‌شد تا اینکه از روش نمونهٔ تصادفی روزها استفاده شود، زیرا در رهیافت مبتنی بر داوری محتمل بود که از گنجاندن روزهای غیرعادی اجتناب شود (برای مثال، روزهایی که بار مراجعهٔ بیمار به طور غیر عادی زیاد، کم، یا به صورتی دیگر غیر عادی است). عیب نمونه‌گیری مبتنی بر داوری این است که از نظر ریاضی در رابطه با قابلیت اعتماد برآوردهای حاصل هیچ شناختی به دست نمی‌آید.

در این کتاب فقط به نمونه‌های احتمالاتی می‌پردازیم زیرا قویاً احساس می‌کنیم که آمارگیریهای نمونه‌ای باید برآوردهایی به دست دهند که از نظر آماری بتوانند نسبت به مقادیر مورد انتظار و خطاهای معیارشان ارزشیابی شوند.

۲.۲.۲ چارچوبهای نمونه‌گیری، واحدهای نمونه‌گیری، و واحدهای شمارش

در نمونه‌گیری احتمالاتی باید احتمال ظاهر شدن هر عنصر در نمونه معلوم باشد. برای تحقق این امر باید «فهرستی» موجود باشد که بتوان نمونه را از آن انتخاب کرد. به چنین فهرستی *چارچوب نمونه‌گیری* می‌گویند و باید این خاصیت را داشته باشد که هر روشی برای انتخاب عناصر از چارچوب نمونه‌گیری به کار گرفته شود یکایک عناصر در جامعه شانسی برای انتخاب شدن در نمونه داشته باشند. چارچوب نمونه‌گیری اجباری ندارد که همهٔ عناصر جامعه را فهرست کند. برای مثال، اگر برای یک آمارگیری نمونه‌ای که عناصر آن ساکنان شهرند از دفتر راهنمای شهر به عنوان چارچوب نمونه‌گیری استفاده شود، واضح است که در این صورت همهٔ عناصر به هیچ وجه در چارچوب نمونه‌گیری که در این مورد فهرست (فرضاً) تمام خانوارهای شهر است فهرست نخواهند شد. ولی چون این یک آمارگیری احتمالاتی است، اگر چارچوب نمونه‌گیری در واقع شامل تمام خانوارهای شهر باشد هر یک از عناصر، شانسی برای انتخاب شدن در نمونه دارند.

غالباً یک طرح نمونه‌گیری خاص مشخص می سازد که نمونه‌گیری در دو یا سه مرحله اجرا شود. این طرح را *طرح نمونه‌گیری چندمرحله‌ای* می‌نامند. برای مثال، یک آمارگیری از خانوارها که در ایالتی بزرگ اجرا می‌شود ممکن است طرح نمونه‌گیریی داشته باشد که مشخص کند نمونه‌ای از ناحیه‌ها در داخل ایالت انتخاب شود و در داخل هر ناحیه‌ای که برای نمونه انتخاب می‌شود نمونه‌ای از تقسیمات شهری کوچکتر (شهرکها) گرفته شود و در داخل هر یک از تقسیمات شهری کوچکتر نمونه‌ای از خانوارها گرفته شود. در نمونه‌گیری چندمرحله‌ای، در هر مرحله از نمونه‌گیری از چارچوب نمونه‌گیری متفاوتی استفاده می‌شود. واحدهای فهرست شده در چارچوب را عموماً *واحدهای نمونه‌گیری* می‌نامند. در مثال بالا، چارچوب نمونه‌گیری برای مرحلهٔ اول فهرست ناحیه‌های داخل ایالت است و هر ناحیه برای این مرحله یک واحد نمونه‌گیری است. فهرست شهرکها در داخل هر ناحیه‌ای که در مرحلهٔ اول برای نمونه انتخاب شده‌اند، چارچوب نمونه‌گیری برای مرحلهٔ دوم و هر شهرک یک واحد نمونه‌گیری برای این مرحله است. بالاخره، فهرست خانوارهای داخل هر شهرک که در مرحلهٔ دوم برای نمونه انتخاب شده‌اند، چارچوب نمونه‌گیری برای مرحلهٔ سوم و نهایی است و هر خانوار یک واحد نمونه‌گیری برای این مرحله است. واحدهای نمونه‌گیری مرحلهٔ اول را عموماً *واحدهای نمونه‌گیری اولیه* (PSUs) می‌نامند. واحدهای نمونه‌گیری آخرین مرحله از طرح

نمونه‌گیری چند مرحله‌ای *واحدهای شمارش* یا *واحدهای فهرست برداری* نامیده می‌شوند. دربارهٔ این موارد قبلاً بحث کرده‌ایم.

۳.۲.۲ اندازه‌گیریهای نمونه‌گیری و آماره‌های خلاصه

فرض کنیم که به طریقی نمونه‌ای با n عنصر از جامعه‌ای دارای N عنصر اختیار کرده و هر عنصر نمونه را با توجه به متغیر x اندازه می‌گیریم. برای راحتی کار عناصر نمونه را از ۱ تا n *شماره‌گذاری* می‌کنیم. (مهم نیست که شمارهٔ اصلی آنها در جامعه چه بوده است). فرض می کنیم x_1 نشان‌دهندهٔ مقدار x برای عنصر شمارهٔ «۱»، x_2 نشان‌دهندهٔ مقدار x برای عنصر شمارهٔ «۲»، و همین طور تا آخر باشد. پس از این که نمونه را گرفتیم می‌توانیم درست همان طور که برای جامعه انجام می‌دادیم کمیتهایی همچون مجموعها، میانگینها، نسبتها، و انحراف معیارها را محاسبه کنیم. ولی وقتی این کمیتها برای نمونه محاسبه می‌شوند، به معنای واقعی کلمه، پارامتر نیستند چون در معرض تغییرپذیری نمونه‌گیری قرار دارند (یک پارامتر واقعی، مقداری ثابت است). در عوض به این مقادیر نمونه عموماً به عنوان *آماره‌ها*، یا *آماره‌های خلاصه* یا *آماره‌های توصیفی* اشاره می‌شود. تعریف برخی از آماره‌ها که در بسیاری از طرحهای نمونه‌ای، چه به منظور توصیف و چه در فرمولها برای برآوردهای جامعه‌ای به کار می‌روند در مبحث زیر ارائه شده است. تعریفهای بقیه هر گاه نیاز باشد ارائه خواهند شد.

مجموع نمونه‌ای: *مجموع نمونه‌ای* یک مشخصه، معمولاً با حرف x نشان داده می‌شود و مجموع مقادیر آن مشخصه در کل عناصر نمونه است:

$$x = \sum_{i=1}^{n} x_i \qquad (۹.۲)$$

میانگین نمونه: *میانگین نمونه* با توجه به مشخصهٔ x معمولاً با $\bar{x}$ نشان داده می‌شود و از فرمول زیر به دست می‌آید:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (۱۰.۲)$$

نسبت نمونه‌ای: هرگاه مشخصهٔ مورد اندازه‌گیری x معرف بود یا نبود صفت کیفی *دوحالتی* باشد، نسبت نمونه‌ای را معمولاً با P_x نشان می‌دهند که از فرمول زیر به دست می‌آید:

$$P_x = \frac{x}{n} \qquad (۱۱.۲)$$

که در آن x تعداد عناصری از نمونه است که آن صفت را دارند.

واریانس نمونه و انحراف معیار: *واریانس نمونه*، s_x^2، برای هر مشخصهٔ x از فرمول زیر به دست می‌آید:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (۱۲.۲)$$

هر گاه مشخصهٔ x یک صفت کیفی دوحالتی باشد، واریانس نمونه، s_x^2، که به صورت بالا تعریف شد به شکل زیر تبدیل می‌شود:

$$s_x^2 = \frac{np_x(1-p_x)}{n-1} \qquad (۱۳.۲)$$

اگر اندازهٔ نمونه، n، بزرگ باشد (مثلاً بیشتر از ۲۰) می‌توانیم از تقریب زیر استفاده کنیم:

$$s_x^2 \approx p_x(1-p_x)$$

انحراف معیار نمونه صرفاً، ریشهٔ دوم واریانس نمونه است:

$$s_x = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right)^{1/2} \qquad (۱۴.۲)$$

برای راحتی کار در ارجاعات بعدی، این فرمولها در تابلوی ۲.۲ خلاصه شده‌اند.

۴.۲.۲ برآورد کردن مشخصه‌های جامعه

برآورد مجموع کل، X، را می‌توان از مجموع نمونه، x، با فرمول زیر به دست آورد:

$$x' = \left[\frac{N}{n}\right](x) \qquad (۱۵.۲)$$

به عبارت دیگر اگر مجموع نمونه، x، را در نسبت تعداد عناصر در جامعه به تعداد عناصر در نمونه ضرب کنیم آمارهٔ به دست آمده x' را می‌توانیم به عنوان برآورد جامعهٔ کل X به کار بریم.

$\hat{\sigma}_x^2$ که برآورد واریانس جامعه، σ_x^2، است از فرمول زیر به دست می‌آید:

$$\hat{\sigma}_x^2 = \left(\frac{N-1}{N}\right)\left(s_x^2\right) \qquad (۱۶.۲)$$

اگر تعداد عناصر N در جامعه نسبتا زیاد باشد در آن صورت عبارت $\frac{N-1}{N}$ در معادلهٔ فوق از نظر عددی نزدیک به یک خواهد بود و می‌توانیم از تقریب زیر استفاده کنیم:

$$\hat{\sigma}_x^2 \approx s_x^2$$

در اینجا تأکید می‌کنیم که آماره‌های نمونه‌ای و برآوردهای مشخصه‌های جامعه که در بالا ارائه شدند برای هر طرح نمونه‌ای مورد استفاده قرار نمی‌گیرند. ضمن بحث در مورد طرحهای نمونه‌ای مشخص در فصلهای بعدی، روشهای برآورد کردن مشخصه‌های جامعه را که ویژهٔ طرح مشخص مورد بحث است ارائه خواهیم کرد.

تابلوی ۲.۲ آماره‌های نمونه‌ای

مجموع

$$x = \sum_{i=1}^{n} x_i \qquad (۹.۲)$$

میانگین

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n} \qquad (۱۰.۲)$$

نسبت

$$p_x = \frac{x}{n} \qquad (۱۱.۲)$$

که در معادلهٔ (۱۱.۲) x، تعداد عناصری از نمونه است که دارای صفت کیفی دوحالتی است.

واریانس

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (۱۲.۲)$$

واریانس، صفت کیفی دوحالتی

$$s_x^2 = \frac{np_x(1-p_x)}{n-1} \qquad (۱۳.۲)$$

انحراف معیار

$$s_x = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right)^{1/2} \qquad (۱۴.۲)$$

در این تعاریف، n، تعداد عناصر موجود در نمونه و x_i، مقدار عنصر i ام است.

مثال تشریحی: حال این مفاهیم را با استفاده از داده‌های مربوط به ویزیت پزشکان از خانوارها که در جدول ۱.۲ ارائه شده است نشان می‌دهیم. فرض کنید، به طریقی، نمونه‌ای از نه پزشک در این جامعهٔ

آماری انتخاب کرده‌ایم و فرض کنید نمونهٔ ما شامل نه پزشکی است که در جدول ۲.۲ فهرست شده‌اند.

میانگیـــــن نمونه، مجمـــوع نمونه و واریانس نمونه برای این داده‌ها به شرح زیر است (فرمولهـــا را در تابلوی ۲.۲ ببینید):

جدول ۲.۲ داده‌های نمونه برای تعداد ویزیت از خانوارها

پزشک	تعداد ویزیت	پزشک	تعداد ویزیت
۱	۵	۱۷	۷
٦	۰	۱۹	۳۷
۷	۱۲	۲۱	۸
۱۲	۵	۲۵	۰
۱۳	٦		

$s_x^2 = ۱۲۵/۱۱$ $\quad \bar{x} = ۸/۸۹$ $\quad x = ۵/۸۰$

نسبت نمونه با یک یا چند ویزیت از خانوار عبارت است از: $p_y = ۰/۷۸$

واریانس نمونه، نسبت به انجام یک یا چند ویزیت از خانوار عبارت است از: $s_x^2 = ۰/۱۹٤٤$

اگر بخواهیم از این نمونه، X ، مجموع کل ویزیت بیماران خانوار توسط پزشکان در جامعه را برای مدت زمان مشخصی برآورد کنیم، ابتدا مجموع نمونه، x، ویزیت در خانوار را برای نه پزشک نمونه به دست می‌آوریم. سپس x' را با ضرب کردن x در $\frac{N}{n}$ پیدا می‌کنیم، که در آن $N=۲۵$ و $n=۹$ است. خلاصه‌ای از پارامترهای جامعه و برآوردهای نمونه‌ای (بر اساس فرمولهای بالا و فرمولهای جداول ۱.۲ و ۲.۲) به شرح زیر ارائه می‌شوند.

پارامتر جامعه	برآورد از نمونه
$X = ۱۲۷$	$x = (\frac{۲۵}{۹})\ (۸۰) = ۲۲۲/۲۲$
$\bar{X} = ۵/۰۸$	$\bar{x} = ۸/۸۹$
$\sigma_x^2 = ٦۷/۹۱$	$\hat{\sigma}_x^2 = (\frac{۲۴}{۲۵})\ (۱۲۵/۱۱) = ۱۲۰/۱۱$
$P_y = ۰/۵٦$	$p_y = (\frac{۷}{۹}) = ۰/۷۸$
$\sigma^2 = ۰/۲٤٦$	$\hat{\sigma}^2 = (\frac{۲۴}{۲۵})\ (۰/۱۹٤٤) = ۰/۱۸٦٦$

□

از خلاصه محاسبات در مثال بالا متوجه می‌شویم که برآوردهای پارامترهای جامعه از نمونهٔ گرفته شدهٔ خاص، نه با پارامترهای جامعهٔ مورد برآورد برابرند و نه حتی به مقادیر واقعی این پارامترها خیلی نزدیک‌اند. اگر نمونهٔ دیگری گرفته بودیم، برآوردهایی متفاوت با این پارامترها به دست می‌آوردیم که ممکن بود یا به مقادیر واقعی پارامترها نزدیکتر یا باز هم دورتر باشند. چون هیچ وقت حقیقتاً مقادیر واقعی پارامترهای جامعه را که از روی نمونه برآورد می‌کنیم نمی‌دانیم، هیچ وقت درست نمی‌دانیم که برآوردهای جامعه‌ای ما واقعاً چقدر مناسب یا نامناسب‌اند. ولی اگر برنامهٔ نمونه‌گیری ما از نمونه‌گیری احتمالاتی استفاده کند در آن صورت می‌توانیم از نظر ریاضی این شناخت را پیدا کنیم که برآوردهای ما احتمال دارد که از مقادیر واقعی نامعلوم چقدر فاصله داشته باشند. برای انجام این کار باید مطالبی دربارهٔ توزیع برآوردهای جامعه‌ای در تمام نمونه‌های ممکن بدانیم که می‌تواند از برنامهٔ نمونه‌گیری خاصی که به کار رفته است حاصل شود. در بخش بعد به شرح و بسط روش‌شناسی به دست آوردن این قبیل اطلاعات از نمونه‌های احتمالاتی می‌پردازیم.

۳.۲ توزیعهای نمونه‌گیری

در بخش قبل دربارهٔ برآورد پارامترهای جامعه از روی نمونه بحث کردیم. در این بخش به توزیع این برآوردهای پارامترهای جامعه در همهٔ نمونه‌های ممکن می‌پردازیم که می‌توانند با استفاده از یک برنامهٔ نمونه‌گیری خاص ایجاد شوند.

فرض کنیم یک برنامهٔ نمونه‌گیری و شیوهٔ برآورد کردن خاص بتواند به T نمونهٔ ممکن از یک جامعهٔ معین بینجامد و یک نمونهٔ خاص، برآورد $\hat{d}$ را از پارامتر جامعه‌ای d به دست دهد. توزیع فراوانی نسبی $\hat{d}$ در T نمونهٔ ممکن را *توزیع نمونه‌گیری* $\hat{d}$ با توجه به برنامهٔ نمونه‌گیری خاص و شیوهٔ برآورد می‌نامند.

برای نشان دادن توزیع نمونه‌گیری، مثال زیر را بررسی می‌کنیم.

مثال تشریحی: فرض کنید یک محلهٔ فرضی شش مدرسه دارد. این جامعهٔ آماری متشکل از شش مدرسه، در جدول ۳.۲ ارائه شده است. حال فرض کنید می‌خواهیم برای برآورد کل تعداد دانش‌آموزان ایمن‌سازی نشده در برابر سرخک در این شش مدرسهٔ محله، نمونه‌ای دو مدرسه‌ای بگیریم. فرض کنید یک برنامهٔ نمونه‌گیری تعیین کنیم که شش پاکت مشابه را بگیریم و در هر یک کارتی بگذاریم که از ۱ تا ٦ شماره‌گذاری شده‌اند و در پاکتها را ببندیم. بعد پاکتها را داخل یک کلاه گذاشته و خوب آنها را بر بزنیم. سپس دو پاکت را از درون کلاه برداشته و دو مدرسه‌ای را که شمارهٔ آنها روی کارتهای داخل پاکتهای انتخابی است در نمونه جای دهیم. فرض کنید اطلاعات مربوط به وضعیت ایمن‌سازی در

برابر سرخک را از هر کودک در مدرسه‌های نمونه به دست آورده‌ایم. تعداد کل دانش‌آموزان ایمن‌سازی نشده در مقابل سرخک با روی هم ریختن مجموعهای نمونه‌ای در دو مدرسهٔ نمونه از طریق نسبت $\frac{N}{n}$ به دست می‌آید که در اینجا N=٦ و n=٢ است. این شیوه، ۱۵ نمونهٔ ممکن به دست می‌دهد که همه شانس برابر و یکسانی برای انتخاب شدن دارند. برحسب نمادگذاری توصیف شده در تعریف توزیع نمونه‌گیری، داریم T=۱۵ و $d = X$=۳۰ و هر $\hat{d}$ برابر با x' است که X و x' به ترتیب مجموع کل جامعه و مجموع برآورد شدهٔ جامعهٔ آماری هستند. ۱۵ نمونهٔ ممکن که از این برنامهٔ نمونه‌گیری به دست می‌آیند همراه با مقادیر x' در جدول ٤.٢ فهرست شده‌اند.

جدول ۳.۲ داده‌های مربوط به تعداد دانش‌آموزان ایمن سازی نشده در برابر سرخک در شش مدرسهٔ محله

مدرسه	تعداد دانش‌آموزان	دانش‌آموزان ایمن‌سازی نشده در برابر سرخک	
		مجموع	نسبت
۱	٥٩	٤	۰/۰٦۸
۲	۲۸	٥	۰/۱۷۹
۳	۹۰	۳	۰/۰۳۳
٤	٤٤	۳	۰/۰٦۸
٥	۳٦	۷	۰/۱۹٤
٦	٥۷	۸	۰/۱٤۰
جمع	۳۱٤	۳۰	۰/۰۹٦

چون هر یک از ۱۵ نمونهٔ فهرست شده در جدول ٤.٢ شانس یکسان $\left(\frac{۱}{۱۵}\right)$ برای انتخاب شدن دارند، می‌توانیم توزیع فراوانی x' را به دست آوریم. جدول ۵.۲ توزیع نمونه‌گیری مجموع برآورد شده، x'، را نشان می‌دهد.

فراوانیهای نسبی در آخرین ستون جدول ۵.۲ نشان‌دهندهٔ کسر همهٔ نمونه‌هایی است که مقادیر متناظر x' را می‌پذیرند. با استفاده از این فراوانیهای نسبی می‌توانیم تصویری از توزیع نمونه‌گیری x' رسم کنیم که در شکل ۱.۲ نشان داده شده است.

توزیعهای نمونه‌گیری را می‌توان با مشخصه‌های معینی توصیف کرد. برای مقاصد ما میانگین و واریانس (یا ریشهٔ دوم آن یعنی انحراف معیار) دو مشخصه‌ای هستند که از همه بیشتر اهمیت دارند و بعداً تعریف می‌شوند.

جدول ٤.٢ نمونه‌های ممکن و مقادیر x'

نمونه	مدرسه‌های نمونه	x'
١	٢ , ١	٢٧
٢	٣ , ١	٢١
٣	٤ , ١	٢١
٤	٥ , ١	٣٣
٥	٦ , ١	٣٦
٦	٣ , ٢	٢٤
٧	٤ , ٢	٢٤
٨	٥ , ٢	٣٦
٩	٦ , ٢	٣٩
١٠	٤ , ٣	١٨
١١	٥ , ٣	٣٠
١٢	٦ , ٣	٣٣
١٣	٥ , ٤	٣٠
١٤	٦ , ٤	٣٣
١٥	٦ , ٥	٤٥

جدول ٥.٢ توزیعهای نمونه‌گیری برای داده‌های جدول ٤.٢

x'	فراوانی	فراوانی نسبی
١٨	١	$\frac{١}{١٥}$
٢١	٢	$\frac{٢}{١٥}$
٢٤	٢	$\frac{٢}{١٥}$
٢٧	١	$\frac{١}{١٥}$
٣٠	٢	$\frac{٢}{١٥}$
٣٣	٣	$\frac{٣}{١٥}$
٣٦	٢	$\frac{٢}{١٥}$
٣٩	١	$\frac{١}{١٥}$
٤٥	١	$\frac{١}{١٥}$

میانگین توزیع نمونه‌گیری یک *پارامتر* برآورد شده $\hat{d}$ با توجه به یک برنامهٔ نمونه‌گیری خاص که T نمونهٔ ممکن به دست می‌دهد و به C مقدار ممکن برای $\hat{d}$ می‌انجامد به نام *مقدار مورد انتظار* نیز موسوم است که آن را با $E(\hat{d})$ نشان می‌دهند و به صورت زیر تعریف می‌شود :

$$E(\hat{d}) = \sum_{i=1}^{c} \hat{d}_i \pi_i \qquad (۱۷.۲)$$

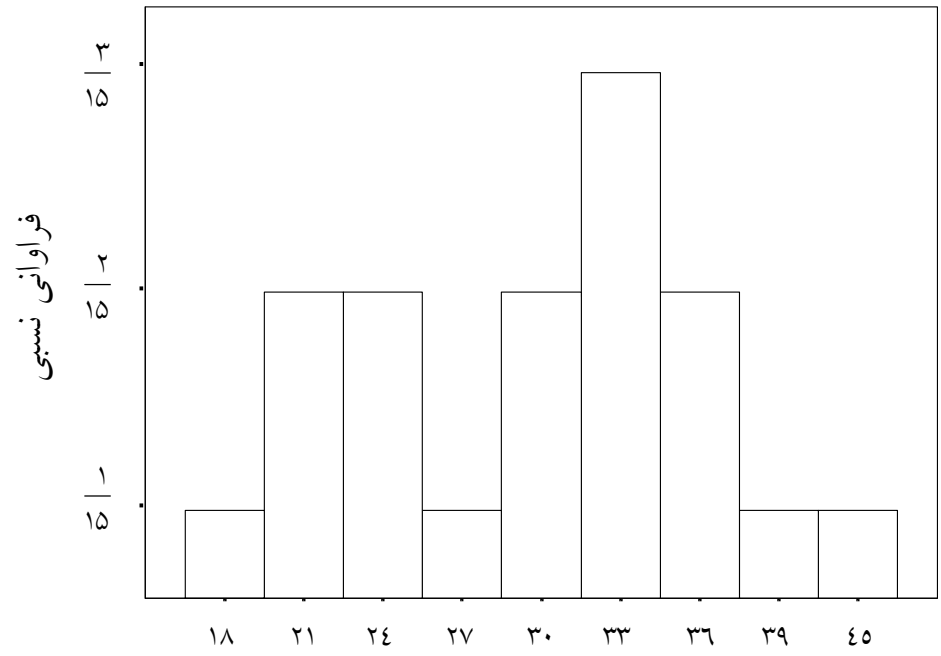
که در آن $\hat{d}_i$ مقدار خاصی از $\hat{d}$ است و π_i احتمال به دست آوردن آن مقدار خاص $\hat{d}$ است. (توجه کنید که اگر احتمال انتخاب هر نمونه برابر باشد، آنگاه $\pi_i = \frac{f_i}{T}$، که f_i تعداد دفعاتی است که مقدار خاص d_i از $\hat{d}$ روی می‌دهد).

واریانس توزیع نمونه‌گیری پارامتر برآورد شده $\hat{d}$ یعنی $Var(\hat{d})$ با توجه به یک طرح نمونه‌گیری خاص از فرمول زیر به دست می‌آید:

$$Var(\hat{d}) = \sum_{i=1}^{c} [\hat{d}_i - E(\hat{d})]^2 \pi_i \qquad (۱۸.۲)$$

هم ارز جبری این معادله که می‌تواند برای محاسبات به کار رود عبارت است از:

$$Var(\hat{d}) = \sum_{i=1}^{c} \hat{d}_i^2 \pi_i - E^2(\hat{d}) \qquad (۱۹.۲)$$



برآورد تعداد دانش‌آموزان ایمن‌سازی نشده

شکل ۱.۲ توزیع فراوانی نسبی توزیع نمونه‌گیری x'

انحراف معیار $SE(\hat{d})$ *توزیع نمونه‌گیری* پارامتر برآورد شده $\hat{d}$ بیشتر به نام *خطای معیار* $\hat{d}$ مشهور است، و عبارت است از ریشهٔ دوم واریانس $Var(\hat{d})$ توزیع نمونه‌گیری $\hat{d}$:

$$SE(\hat{d}) = \left[Var(\hat{d})\right]^{1/2} \qquad (۲۰.۲)$$

با استفاده از این معادله‌های کلی می‌توانیم فرمولهایی برای میانگین و واریانس (و از این طریق برای خطای معیار) توزیع نمونه‌گیری میانگینها، مجموعها و نسبتها به دست آوریم. برای راحتی کار در ارجاعات بعدی، این فرمولها در دو تابلوی بعدی خلاصه شده‌اند.

□

وقتی هر نمونه، احتمال یکسان برای انتخاب شدن نداشته باشد یا در صورتی که بخواهیم میانگینها و واریانسهای توزیع نمونه‌گیری را مستقیماً از یک توزیع نمونه‌گیری مانند توزیع ارائه شده در جدول ۵.۲ محاسبه کنیم، باید از فرمولهای تابلوی ٤.۲ استفاده کنیم.

حال ببینیم چگونه می‌توان از بعضی از این فرمولها در عمل استفاده کرد. برای این منظور از داده‌های مثال قبل استفاده می‌کنیم.

مثال تشریحی: در مثال قبل $C = ۹, T = ۱۵$,$\hat{d} = x'$ و π_i ها فراوانیهای نسبی وابسته به هر یک از مقدارهای خاص x' است. میانگین $E(x')$، توزیع نمونه‌گیری x'، برای این مثال (با استفاده از فرمولهای تابلوی ٤.۲) چنین است:

$$E(x') = \sum_{i=۱}^{۹} x_i'(\frac{f_i}{۱۵}) = ۱۸\ (\frac{۱}{۱۵}) + ۲۱\ (\frac{۲}{۱۵}) + ۲۴\ (\frac{۲}{۱۵}) + ۲۷\ (\frac{۱}{۱۵}) + ۳۰(\frac{۲}{۱۵})$$
$$+ ۳۳\ (\frac{۳}{۱۵}) + ۳۶\ (\frac{۲}{۱۵}) + ۳۹\ (\frac{۱}{۱۵}) + ۴۵\ (\frac{۱}{۱۵}) = ۳۰$$

واریانس $Var(x')$ توزیع نمونه‌گیری x' چنین است:

$$Var(x') = (۱۸ - ۳۰)^۲(\frac{۱}{۱۵}) + (۲۱ - ۳۰)^۲(\frac{۲}{۱۵})$$
$$+ (۲۴ - ۳۰)^۲(\frac{۲}{۱۵}) + (۲۷ - ۳۰)^۲(\frac{۱}{۱۵}) + (۳۰ - ۳۰)^۲(\frac{۲}{۱۵}) + (۳۳ - ۳۰)^۲(\frac{۳}{۱۵})$$
$$+ (۳۶ - ۳۰)^۲(\frac{۲}{۱۵}) + (۳۹ - ۳۰)^۲(\frac{۱}{۱۵}) + (۴۵ - ۳۰)^۲(\frac{۱}{۱۵}) = ۵۲/۸$$

خطای معیار x' عبارت است از:

$$SE(x') = \sqrt{۵۲/۸} = ۷/۲۸$$

□

تابلوی ۳.۲ میانگین و واریانس توزیع نمونه‌گیری در صورتی که هر نمونه دارای احتمال یکسان $\frac{۱}{T}$ برای انتخاب شدن باشد

برای مجموعها

$$E(x') = \frac{\sum_{i=۱}^{T} x'_i}{T} \qquad Var(x') = \frac{\sum_{i=۱}^{T} [x'_i - E(x')]^۲}{T} \qquad (۲۱.۲)$$

که در آنها، x'_i، مجموع محاسبه شدهٔ i امین نمونهٔ ممکن است که از جامعه انتخاب شده است. توجه کنید که برخی از مقدارهای x'_i ممکن است از نمونه‌ای به نمونهٔ دیگر یکسان باشد، ولی همهٔ مقادیر تحقق یافته در جمع منظور می‌شوند حتی اگر تکراری باشند.

برای میانگینها

$$E(\bar{x}) = \frac{\sum_{i=۱}^{T} \bar{x}_i}{T} \qquad Var(\bar{x}) = \frac{\sum_{i=۱}^{T} [\bar{x}_i - E(\bar{x})]^۲}{T} \qquad (۲۲.۲)$$

برای نسبتها

$$E(p_y) = \frac{\sum_{i=۱}^{T} p_{yi}}{T} \qquad Var(p_y) = \frac{\sum_{i=۱}^{T} [p_{yi} - E(p_y)]^۲}{T} \qquad (۲۳.۲)$$

در این فرمولها T، تعداد نمونه‌های ممکن است.

تابلوی ٤.۲ میانگین و واریانس توزیع نمونه‌گیری در صورتی که هر نمونه دارای احتمال یکسان برای انتخاب شدن نباشد*.

برای مجموعها

$$E(x') = \sum_{i=۱}^{C} x'_i \pi_i \qquad Var(x') = \sum_{i=۱}^{C} [x'_i - E(x')]^۲ \pi_i \qquad (۲٤.۲)$$

برای میانگینها

$$E(\bar{x}) = \sum_{i=۱}^{C} \bar{x}_i \pi_i \qquad Var(\bar{x}) = \sum_{i=۱}^{C} [\bar{x}_i - E(\bar{x})]^۲ \pi_i \qquad (۲۵.۲)$$

برای نسبتها

$$E(p_y) = \sum_{i=۱}^{C} p_{yi} \pi_i \qquad Var(p_y) = \sum_{i=۱}^{C} [p_{yi} - E(p_y)]^۲ \pi_i \qquad (۲٦.۲)$$

در این فرمولها C، تعداد مقدارهای یکتای ممکن آماره، و $\pi_i = \frac{f_i}{T}$، نسبت همان مقدار i امین مقدار یکتای حاصل و f_i، فراوانی رویدادها در توزیع نمونه‌گیری i امین تحقق و T تعداد نمونه‌های ممکن است.

* یا در صورتی که مقادیر مجموعها، میانگینها یا نسبتها با فراوانی نسبی یکسان در تمام نمونه‌های ممکن روی نداده باشند.

٤.۲ مشخصه‌های برآوردهای پارامترهای جامعه

در بخش قبل، مفهوم توزیع نمونه‌گیری برآورد یک پارامتر جامعه را با توجه به یک برنامهٔ خاص نمونه‌گیری شرح دادیم. همچنین مفاهیم میانگین توزیع نمونه‌گیری یک برآورد جامعه‌ای و خطای معیار برآورد را نیز ارائه دادیم. اکنون می‌توانیم خواص معینی از برآوردهای جامعه‌ای را با توجه به این مفاهیم به بحث بگذاریم. از نظر شهودی واضح است که یک ویژگی مطلوب برای یک برنامهٔ نمونه‌گیری و شیوهٔ برآورد کردن، آن است که برآوردهایی از پارامترهای جامعه به دست دهد که میانگین توزیع نمونه‌گیری آن برابر با پارامتر نامعلوم واقعی، یا حداقل نزدیک به آن باشد و خطای معیار آن نیز بسیار کم باشد. در واقع دقت یک پارامتر جامعه‌ای برآورد شده با توجه به این دو مشخصه ارزشیابی می‌شود. در این بخش مفاهیمی را معرفی می‌کنیم که در ارزیابی طرحهای نمونه‌ای به کار می‌روند.

۱.٤.۲ اریبی

$B(\hat{d})$، اریبی برآورد $\hat{d}$ از پارامتر جامعه‌ای d به صورت تفاوت بین میانگین توزیع نمونه‌گیری $\hat{d}$ یعنی $E(\hat{d})$ و مقدار واقعی پارامتر نامعلوم d تعریف می‌شود. به عبارت دیگر:

$$B(\hat{d}) = E(\hat{d}) - d \qquad (۲۷.۲)$$

اگر $B(\hat{d}) = ۰$ باشد می‌گوییم برآوردگر $\hat{d}$ ناریب است. به عبارت دیگر، در صورتی که میانگین توزیع نمونه‌گیری $\hat{d}$ برابر با d باشد $\hat{d}$ یک برآوردگر ناریب است.

در مثال موردنظر در بخش قبل، برآورد x' از کل تعداد کودکان ایمن‌سازی نشده در برابر سرخک، یک برآوردگر ناریب برای کل جمعیت واقعی X است، زیرا معلوم شد که $E(x') = ۳۰ = X$.

تأکید می‌کنیم که همان شیوهٔ برآورد که در یک برنامهٔ نمونه‌گیری، ناریب است می‌تواند برای یک برنامهٔ نمونه‌گیری دیگر اریب باشد. مثال بعدی این ایده را روشنتر بیان می‌کند.

مثال تشریحی: همان جامعهٔ متشکل از شش مدرسهٔ مثال قبل را در نظر می‌گیریم ولی این بار از برنامهٔ نمونه‌گیری زیر استفاده می‌کنیم. کارتهایی با شماره‌های ۱ تا ۱۰ در پاکت می‌گذاریم، در پاکتها را می‌چسبانیم و آنها را در کلاهی قرار می‌دهیم. یک پاکت را از داخل کلاه برمی‌داریم و مدرسهٔ نمونه را مطابق شیوهٔ نشان داده شده در جدول ٦.۲ انتخاب می‌کنیم. این شیوه، یک نمونهٔ احتمالاتی است، زیرا هر مدرسه دارای یک احتمال معلوم و غیرصفر برای انتخاب شدن در نمونه است. توزیع مجموعهای برآورد شده در صورت استفاده از این شیوه در جدول ۷.۲ نشان داده شده است.

چون هر یک از ۱۰ نمونهٔ ممکن دارای احتمال یکسان‌اند، پس توزیع x' را داریم که در جدول ۸.۲ نشان داده شده است. در توزیع جدول ۸.۲ ملاحظه می‌کنیم که $E(x')$، میانگین توزیع نمونه‌گیری x' در این برنامهٔ نمونه‌گیری به شرح زیر به دست می‌آید:

$$E(x') = ۱۸\ (\frac{۱}{۱۰}) + ۲۱\ \ (\frac{۲}{۱۰}) + ۲۴\ \ (\frac{۲}{۱۰}) + ۲۷\ \ (\frac{۱}{۱۰}) + ۳۳(\frac{۱}{۱۰}) + ۳۶(\frac{۲}{۱۰}) + ۳۹(\frac{۱}{۱۰}) = ۲۷/۹$$

به این ترتیب چون مجموع جامعه، X، برابر است با ۳۰، مجموع برآورد شده، x'، در این برنامهٔ نمونه‌گیری یک برآورد ناااریب از مجموع جامعه، X نیست.

جدول ۶.۲ شیوهٔ نمونه‌گیری برای جامعهٔ متشکل از شش مدرسه

شمارهٔ انتخاب شده	مدرسه‌های انتخاب شده در نمونه	شمارهٔ انتخاب شده	مدرسه‌های انتخاب شده در نمونه
۱	۲,۱	۶	۳,۲
۲	۳,۱	۷	۴,۲
۳	۴,۱	۸	۵,۲
۴	۵,۱	۹	۶,۲
۵	۶,۱	۱۰	۴,۳

جدول ۷.۲ نمونه‌های ممکن و مقادیر x'

نمونه	مدرسه‌های نمونه	x'
۱	۲,۱	۲۷
۲	۳,۱	۲۱
۳	۴,۱	۲۱
۴	۵,۱	۳۳
۵	۶,۱	۳۶
۶	۳,۲	۲۴
۷	۴,۲	۲۴
۸	۵,۲	۳۶
۹	۶,۲	۳۹
۱۰	۴,۳	۱۸

جدول ۸.۲ توزیع نمونه‌گیری x'

x'	π
۱۸	$\frac{۱}{۱۰}$
۲۱	$\frac{۲}{۱۰}$
۲٤	$\frac{۲}{۱۰}$
۲۷	$\frac{۱}{۱۰}$
۳۳	$\frac{۱}{۱۰}$
۳٦	$\frac{۲}{۱۰}$
۳۹	$\frac{۱}{۱۰}$
مجموع	۱

۲.٤.۲ میانگین توان دوم خطا

میانگین توان دوم خطا برای برآورد جامعه‌ای $\hat{d}$ به صورت $MSE(\hat{d})$ نشان داده می‌شود و تعریف آن عبارت از میانگین توان دوم تفاوتهای بین مقادیر برآورد و مقدار واقعی d برای یک پارامتر نامعلوم در تمام نمونه‌های ممکن است. با توجه به نمادهایی که در بخش آخر ارائه شدند، میانگین توان دوم خطا با رابطهٔ زیر تعریف می‌شود:

$$MSE(\hat{d}) = \sum_{i=۱}^{C} (\hat{d}_i - d)^۲ \pi_i \qquad (۲۸.۲)$$

به تفاوت بین میانگین توان دوم خطای برآورد و واریانس برآورد توجه کنید. میانگین توان دوم خطای برآورد، مقدار میانگین توان دوم انحرافها حول مقدار واقعی پارامتر مورد برآورد است. واریانس یک برآورد، مقدار میانگین توان دوم انحرافها حول مقدار میانگین توزیع نمونه‌گیری برآورد است. اگر برآورد نااریب باشد ـ یا به عبارت دیگر اگر میانگین توزیع نمونه‌گیری برآورد برابر با مقدار واقعی پارامتر باشد ـ آن‌گاه میانگین توان دوم خطای برآورد برابر با واریانس برآورد خواهد بود، زیرا انحرافها حول همان نهاد در نظر گرفته خواهد شد. به طور کلی میانگین توان دوم خطای برآورد با رابطهٔ زیر به اریبی و واریانس آن ارتباط پیدا می‌کند:

$$MSE(\hat{d}) = Var(\hat{d}) = B^۲(\hat{d}) \qquad (۲۹.۲)$$

به عبارت دیگر، میانگین توان دوم خطای برآورد جامعه‌ای برابر با واریانس آن برآورد به اضافهٔ توان دوم اریبی آن است.

□

مثال تشریحی: در مثال مربوط به شش مدرسه، اولین برنامهٔ نمونه‌گیری که مورد بحث قرار گرفت برآورد نااریبی از کل جامعه به دست داد. میانگین توان دوم خطای این برآورد با محاسبهٔ زیر به دست می‌آید:

$$MSE(x') = ۵۲/۸ + ۰^۲ = ۵۲/۸$$

به عبارت دیگر، میانگین توان دوم خطا برابر با واریانس برآورد است.

در مثال مربوط به شش مدرسه و برنامهٔ نمونه‌گیری که برآورد اریب x' را برای X به دست داد، واریانس x' (از معادلهٔ (۲٤.۲)) به صورت زیر است:

$$Var(x') = (۱۸ - ۲۷/۹)^۲(\tfrac{۱}{۱۰}) + (۲۱ - ۲۷/۹)^۲(\tfrac{۲}{۱۰})$$
$$+ (۲۴ - ۲۷/۹)^۲(\tfrac{۲}{۱۰}) + (۲۷ - ۲۷/۹)^۲(\tfrac{۱}{۱۰}) + (۳۳ - ۲۷/۹)^۲(\tfrac{۱}{۱۰})$$
$$+ (۳۶ - ۲۷/۹)^۲(\tfrac{۲}{۱۰}) + (۳۹ - ۲۷/۹)^۲(\tfrac{۱}{۱۰}) = ۵۰/۴۹$$

به این ترتیب میانگین توان دوم خطا برای x' برابر است با:

$$MSE(x') = ۵۰/۴۹ + (۳۰ - ۲۷/۹)^۲ = ۵۴/۹$$

توجه کنید که برنامهٔ نمونه‌گیری دوم، برآوردگر x' را به صورتی به دست می‌دهد که واریانس آن کمتر ولی میانگین توان دوم خطای آن در مجموع بیشتر از برآوردی است که از برنامهٔ نمونه‌گیری اول به دست می‌آید.

□

مثال تشریحی: به مثال دیگری از کاربرد میانگین توان دوم خطا نگاهی می‌اندازیم. فرض کنید برای ارزیابی شدت سوختگیهایی که در آمریکا روی می‌دهند یک آمارگیری طرحریزی می‌شود. در ارتباط با این آمارگیری سه دانشجو در ستاد آمارگیری دربارهٔ ارزیابی بدن بیماری که از سوختگی درجهٔ سه رنج می‌برد آموزش داده می‌شوند (به این نوع سوختگی، سوختگی کامل نیز می‌گویند).

یک جراح ارشد سوختگی برای ارزیابی پیشرفت دانشجویان از مجموعه‌ای عکس استفاده می‌کند که از ده بیمار گرفته شده است. هر یک از بیماران دچار ۳۷ درصد سوختگی کامل‌اند. مجموع درصد سوختگی در سراسر بدن این بیماران یکسان است ولی در اجزای بدن یکسان نیست. این بیماران از میان تعداد زیادی از بیمارانی که جراح معاینه کرده است انتخاب شده‌اند.

فرض می‌کنیم اسامی این دانشجویان **A**، **B** و **D** است. جدول ۹.۲ میانگینها و واریانسهای برآوردهای حاصل توسط هر یک از دانشجویان را نشان می‌دهد. حال هر یک از این برآوردها را ارزیابی می‌کنیم.

اتفاقاً متوسط سوختگی که **A** ارزیابی کرده است برابر با متوسط واقعی تصاویر مورد بررسی است (یعنی ۰= ۳۷ − ۳۷ = اریبی). ولی تغییرپذیری اندازه‌گیریهای او بسیار زیاد و در نتیجه میانگین توان دوم خطا نیز زیاد است (یعنی ٦٤ = ٦٤ + ۰^{۲} = MSE).

B به بیش برآورد کردن مقدار واقعی سوختگی کامل گرایش دارد (یعنی۵=۳۷−٤۲= اریبی)، ولی سازگار عمل کرده است. درنتیجه میانگین توان دوم خطای محاسبات او (MSE =۵^{۲}+۹=۳٤) از نتیجهٔ محاسبه شدهٔ **A** کمتر است. به بیان دیگر، میانگین مربع انحرافهای هر یک از ارزیابیهای سوختگی **B** حول سوختگیهای واقعی کمتر از **A** است.

ارزیابیهای **D** مقدار سوختگی را بیش از مقدار واقعی برآورد کرده بودند (یعنی ۱۳=۳۷−۵۰= اریبی)، ولی این ارزیابیها دارای تغییرپذیری کمی بودند. در نتیجه، میانگین توان دوم خطا از همه بیشتر است (MSE =۱۳^{۲}+۹=۱۷۸).

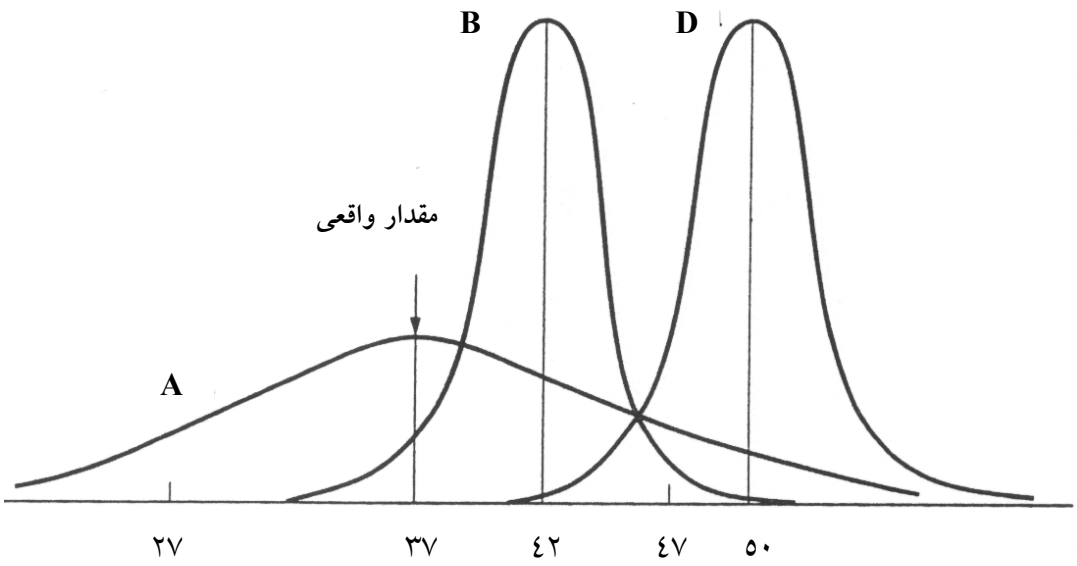
رابطهٔ بین اریبی، تغییرپذیری و میانگین توان دوم خطا از نظر تصویری در شکل ۲.۲ ارائه شده است. در این شکل فرض شده است که اندازه‌گیریها با میانگینها و واریانسهای بیان شده به صورت نرمال توزیع شده‌اند. توجه کنید در حالی که میانگین توزیع **A** برابر با مقدار واقعی است، غیرعادی نخواهد بود که هر یک از ارزیابیهای او با حاشیهٔ زیادی از مقدار واقعی فاصله داشته باشد (پراکندگی گستردهٔ توزیع او نیز حاکی از همین امر است). برای مثال فرض کنید $P_r(\chi > C)$ نشانهٔ این احتمال باشد که تحقق مقدار خاصی از متغیر تصادفی χ از مقدار C بیشتر باشد. در این صورت احتمال این که **A** از مقدار واقعی سوختگی تا بیش از ده نقطه درصدی دور شود از فرمول زیر به دست می‌آید:

$$P_r(\chi > ۴۷) + P_r(\chi < ۲۷) = P_r\left(z > \frac{۴۷ - ۳۷}{\sqrt{۶۴}} = ۱/۲۵\right) + P_r(z < -۱/۲۵) = ۰/۲۱$$

که در آن z، انحراف نرمال استاندارد است.

جدول ۹.۲ داده‌های مربوط به برآوردهای سوختگی

دانشجو	میانگین (درصد)	واریانس^{۲} (٪)
A	۳۷	٦٤
B	٤۲	۹
D	۵۰	۹



شکل ۲.۲ رابطه بین اریبی، تغییرپذیری و میانگین توان دوم خطا برای داده‌های جدول ۹.۲

مقادیر اندازه‌گیریهای **B** معمولاً زیاد است، ولی غالباً طوری نیست که بیش از ده نقطه درصدی با مقدار واقعی فاصله داشته باشد. در مورد **B** :

$$P_r(\chi > ۴۷) + P_r(\chi < ۲۷) = P_r(z > \frac{۴۷-۴۲}{\sqrt{۹}} = ۱/۶۷) + P_r(z < -۵/۰۰) = ۰/۰۵$$

بالاخره، همان طور که مقدار زیاد میانگین توان دوم خطا نشان می‌دهد، احتمال اینکه **D** مقدار واقعی سوختگی را با فاصله‌ای بیش از ده نقطه درصدی ارائه ندهد بسیار زیاد است:

$$P_r(\chi > ۴۷) + P_r(\chi < ۲۷) = P_r(z > \frac{۴۷-۵۰}{\sqrt{۹}} = -۱) + P_r\left(z < -۷/۶۷\right) = ۰/۸۴$$

در این مثال دیدیم که هنگام ارزشیابی برآوردهای خاص بسیار اهمیت دارد که اریبی و واریانس هر دو امتحان شوند. هر دوی اینها در تعیین اندازهٔ میانگین توان دوم خطا نقش مهمی ایفا می‌کنند. □

۳.۴.۲ اعتبار، قابلیت اعتماد و درستی

در بخشهای پیشین از مطلوبیت استفاده از طرحهای نمونه‌ای که برآوردهای معتبر و قابل اعتماد به دست می‌دهند صحبت کردیم. ولی هیچ وقت تعریف نکردیم که اصطلاحات «معتبر» و «قابل اعتماد» از نظر مشخصه‌های برآوردها چه معنی می‌دهند. اکنون به آن اندازهٔ کافی، مفاهیم و

نمادهای مربوط به برآوردها را بسط داده‌ایم که بتوانیم این دو اصطلاح به اضافه اصطلاح سوم، «درستی» یک برآورد را تعریف کنیم که سومی به طوری که خواهیم دید از اعتبار و قابلیت اعتماد نتیجه می‌شود.

قابلیت اعتماد برآورد یک مشخصهٔ جامعه‌ای به این اشاره دارد که برآوردگر در تکرارهای فرایند تولیـد کنندهٔ برآوردگر تا چه اندازه قابل تکثیر است. اگر فرض کنیم که در آمارگیری هیچ خطای اندازه‌گیری نبوده است آن‌گاه قابلیت اعتماد برآوردگر را می‌توان برحسب واریانس نمونه‌گیری یا هم‌ارز آن، برحسب خطای معیار آن بیان کرد. هر چه خطای معیار یک برآوردگر کمتر باشد قابلیت اعتماد آن بیشتر است.

اعتبار برآورد یک مشخصهٔ جامعه بر این اشاره دارد که میانگین برآوردگر در تکرارهای فرایند تولید کنندهٔ آن برآوردگر چقدر با مقدار واقعی پارامتر مورد برآورد تفاوت دارد. باز اگر فرض کنیم که در آمارگیری هیچ خطای اندازه‌گیری نبوده است، اعتبار برآوردگر را می‌توان با امتحان کردن اریبی برآوردگر ارزشیابی کرد. هر چه اریبی کمتر باشد اعتبار بیشتر است.

درستی برآوردگر به این اشاره دارد که مقدار خاصی از یک برآورد به طور متوسط چقدر از مقدار واقعی پارامتر مورد اندازه‌گیری به دور است. درستی برآوردگر را معمولاً بر مبنای میانگین توان دوم خطا یا هم‌ارز آن، بر اساس ریشهٔ دوم میانگین توان دوم خطا (که با علامت RMSE نشان داده، آن را «ریشهٔ میانگین توان دوم خطا» می‌خوانند) ارزشیابی می‌کنند. هر چه میانگین توان دوم خطا کوچکتـر باشد، درستی برآورد بیشتر است.

۵.۲ ملاکهای طرح نمونه‌ای خوب

در بقیهٔ این کتاب به بحث طرحهای نمونه‌ای گوناگون خواهیم پرداخت و نشان خواهیم داد که چگونگی دستکاری در طرح نمونه‌گیری، شیوهٔ برآورد، یا قاعدهٔ شمارش، بر قابلیت اعتماد و اعتبار برآوردهای حاصل می‌تواند اثر بگذارد. بحث خود را به نمونه‌های احتمالاتی محدود خواهیم کرد، زیرا اینها تنها طرحهای نمونه‌گیری هستند که ارزشیابی قابلیت اعتماد برآوردها را از روی داده‌های جمع‌آوری شده در آمارگیری میسر می‌سازند. چون درستی یک برآورد، هم با قابلیت اعتماد و هم با اعتبار مربوط است و چون درستی با استفاده از میانگین توان دوم خطا اندازه‌گیری می‌شود، یکی از ملاکهای ما برای انتخاب یک طرح نمونه‌ای، اندازهٔ میانگین توان دوم خطاست که برای برآوردهای حاصل انتظار داریم.

علاوه بر ملاک میانگین توان دوم خطا، هزینهٔ لازم برای اجرای آمارگیری طبق یک طرح نمونه‌ای خاص نیز به عنوان ملاکی برای ارزشیابی آن طرح نمونه‌ای خاص به کار خواهد رفت. ملاکهای هزینه و درستی را می‌توان در یک ملاک مرکب ترکیب کرد به این ترتیب که ابتدا در مورد کل هزینه‌ای که قرار است به آمارگیری اختصاص یابد تصمیم‌گیری می‌شود و سپس طرحی نمونه‌ای انتخاب می‌شود که برآوردهایی به دست دهد که با توجه به هزینهٔ خاص، کمترین میانگین توان دوم خطا را داشته باشند. برعکس، می‌توانیم ویژگیهای میانگین توان دوم خطای برآوردها را تعیین کنیم و طرحی نمونه‌ای را برگزینیم که با کمترین هزینهٔ ممکن بتواند برآوردهایی به دست دهد که ویژگیهای تعیین شدهٔ میانگین توان دوم خطا را تأمین کند.

بالاخره، علاوه بر درستی و هزینه، ملاک سومی که به کار می‌بریم، شدنی بودن اجرای یک طرح نمونه‌ای خاص است. هر قدر طرح خاصی از نظر هزینه مقرون به صرفه باشد اگر اجرای آن شدنی نباشد هیچ فایده‌ای نخواهد داشت.

٦.٢ خلاصه

در این فصل به بسط مفاهیم مربوط به جامعه‌ها، نمونه‌ها، و برآوردها پرداختیم.

تعریف کردیم که منظور ما از جامعه چیست و گفتیم که چگونه عناصر جامعه غالباً برای مقاصد نمونه‌گیری در واحدهای شمارش یا واحدهای فهرست‌برداری گروه‌بندی می‌شوند. نشان دادیم که چگونه واحدهای اولیه و واحدهای شمارش غالباً با یک قاعدهٔ شمردن با هم پیوند پیدا می‌کنند. پارامترهای خاصی را که توزیع متغیرها در جامعه را مشخص می‌کنند تعریف کردیم. پارامترهای مورد بحث شامل میانگین جامعه، مجموع یا مجموع تجمعی، نسبت، واریانس، انحراف معیار، ضریب تغییرات، و واریانس نسبی بودند.

مفهوم گرفتن نمونه از یک جامعه را مورد بحث قرار دادیم و میان نمونه‌های احتمالاتی و نمونه‌های غیراحتمالاتی تمایز قایل شدیم. در مورد استفاده از چارچوبهای نمونه‌گیری در انتخاب نمونه‌هایی از جامعه و استفاده از انتخاب چند مرحله‌ای نمونه بحث کردیم. آماره‌های خلاصهٔ مربوط به نمونه‌ها از قبیل میانگین نمونه، مجموع نمونه، نسبت نمونه، و واریانس نمونه معرفی شدند و استفاده ازآنها را در تهیهٔ برآوردهای مشخصه‌ها یا متغیرهای جامعه مورد بحث قرار دادیم.

مفاهیم مربوط به توزیع نمونه‌گیری برآوردها، از جمله میانگین توزیع نمونه‌گیری برآورد، واریانس و خطای معیار آن، اریبی آن، و میانگین توان دوم خطای آن شرح داده شد. سپس مفاهیم قابلیت

اعتماد، اعتبار و درستی برآورد یک مشخصهٔ جامعه در رابطه با مفاهیم واریانس، اریبی و میانگین توان دوم خطا تعریف شدند.

تمرین

۱.۲ جدول زیر جامعه‌ای متشکل از پنج بیمارستان را ارائه می‌دهد که با نشانه‌های A، B، C، D و E نامگذاری شده‌اند. تعداد کل تختهای هر بیمارستان نیز ارائه شده است.

بیمارستان	تعداد تخت
A	۱۶۰
B	۲۲۰
C	۸۵۰
D	۵۱۰
E	۱۱۰

الف. میانگین تعداد تختها، $\bar{x}$، و انحراف معیار توزیع تعداد تختها، σ_x، را در جامعهٔ متشکل از پنج بیمارستان حساب کنید.

ب. از این جامعهٔ متشکل از پنج بیمارستان، چند نمونهٔ متشکل از ۲ بیمارستان می‌توان انتخاب کرد؟

پ. هر یک از نمونه‌های ممکن متشکل از دو بیمارستان را فهرست کنید و میانگین تعداد تختها را برای هر نمونه به ازای هر بیمارستان محاسبه کنید.

ت. با فرض این که هر یک از نمونه‌های فهرست شده در بخش پ احتمال برابر برای انتخاب شدن داشته باشند، میانگین $E(\bar{x})$ و $Var(\bar{x})$، واریانس میانگینهای توزیع نمونه‌گیری $\bar{x}$ را حساب کنید. مقایسهٔ $E(\bar{x})$ با میانگین جامعه‌ای $\bar{X}$ چگونه است؟

ث. خطای معیار $\bar{x}$ را حساب کنید. مقایسهٔ $SE(\bar{x})$ با انحراف معیار جامعه، σ_x، از چه قرار است؟

ج. چند نمونهٔ مختلف متشکل از چهار بیمارستان می‌توان از این جامعه به دست آورد؟ هر یک از این نمونه‌ها را همراه با میانگین نمونه، $\bar{x}$، برای هر نمونه فهرست کنید.

چ. $E(\bar{x})$ و $Var(\bar{x})$ را برای میانگینهای نمونهٔ فهرست شده در بخش ج حساب کنید. مقایسهٔ این مقادیر با $E(\bar{x})$ و $Var(\bar{x})$ که در بخش ت به دست آمده‌اند از چه قرار است؟

۲.۲ مشخـص کنید که برای هر یک از مسئله‌های زیر چگونه باید یک نمونه‌گیری را اجـــرا کرد. موارد زیر را تعیین کنید:

الف. جامعه

ب. متغیر(ها)

پ. واحد اولیه

ت. چارچوب

ث. واحد شمارش

۱. فرض کنید می‌خواهیم متوسط هزینهٔ عمل جراحی آپاندیس را در یک ایالت مشخص برآورد کنیم. این ایالت ۲۷ بیمارستان دارد.

۲. فرض کنید می‌خواهیم یک آمارگیری تغذیه انجام دهیم تا مقدار متوسط فیبر مصرف شده توسط افراد را در یک شهر برآورد کنیم. فرض کنید هیچ فهرستی از خانواده‌های این شهر موجود نیست ولی نقشه‌ای هست که هر بلوک شهر را با جزییات کامل نشان می‌دهد.

۳. فرض کنید می‌خواهیم نسبت گوساله‌هایی را به دست آوریم که تا قبل از رسیدن به یک سالگی در تمام دامداریهای پرورش‌دهندهٔ گاو شیری در یک ایالت خاص تلف می‌شوند.

۳.۲ فرض کنید یک آمارگیری برای برآورد کردن متوسط تعداد ساعتهای ورزش روزانهٔ بزرگسالان (۱۸ ساله به بالا) در یک محلهٔ خاص در دست برنامه‌ریزی است. فهرستی از افراد ساکـــن در این شهر موجود نیست، ولی فهرستی از تمام خانوارها در دفتر ثبت احوال شهر موجود است. برای سهولت کار فرض کنید که این فهرست شامل ۹ خانوار است و اگر به خانوارها مراجعه می‌کردید اطلاعاتی به شرح جدول زیر به دست می‌آوردید.

الف. موارد زیر را تعریف کنید:

۱. جامعه

۲. واحد اولیه

۳. واحد شمارش

٤. چارچوب

۵. متغیر

خانوار	تعداد بزرگسالان	متوسط تعداد ساعتهای ورزش روزانهٔ همهٔ بزرگسالان
۱	۲	۱
۲	۳	۳
۳	۲	۷
٤	۵	۸
۵	۳	٤
٦	۱	۰
۷	۲	۱
۸	۳	۲
۹	۲	۰

ب. برای برآورد زمان متوسطی که روزانه صرف ورزش توسط بزرگسالان ساکن در محله می‌شود یک طرح نمونه‌گیری تهیه کنید.

پ. با استفاده از این چارچوب، تمام نمونه‌های متشکل از دو خانوار را انتخاب و متوسط زمان ورزش را برای هر یک از افراد هر نمونه حساب کنید.

ت. مقدار مورد انتظار برآورد خود را حساب و آن را با پارامتر واقعی جامعه مقایسه کنید.

ث. خطای معیار برآورد خود را حساب کنید.

٤.۲ به عنوان بخشی از یک برنامهٔ آموزش ایدز، به ۱۲۰ نفر از مصرف‌کنندگان مواد مخدر تزریقی که آزمایش خون آنها در اولین غربالگری برای HIV [1] (ویروس کاهش ایمنی انسان) منفی بود گفته شد که سوزنهای خود را ضدعفونی و «روابط جنسی سالم» برقرار کنند. یک سال پس از شروع برنامه، نمونه‌ای متشکل از ۳۰ نفر از این آزمودنیها گرفته شد به این ترتیب که به شرکت‌کنندگان از ۱ تا ۱۲۰ شماره داده شد و همهٔ کسانی که شمارهٔ آنها به چهار قابل قسمت بود انتخاب شدند (برای مثال ٤، ۸، ۱۲ و الخ).

الف. شانس هر یک از افراد برای انتخاب شدن در نمونه چقدر است؟

ب. اگر نتیجهٔ آزمایش خون آزمودنیهای شمارهٔ ۱، ۳، ٤، ۸، ۲۹ و ٦۵برای HIV مثبت باشد، نسبت تغییر وضع آزمایش خون آزمودنیها برای این جامعه چقدر است؟

[1] Human Immunodeficiency Virus

پ. نسبت آزمودنیهای با تغییر وضعیت در نمونه چقدر است؟ آیا این برآورد نسبت جامعه‌ای نااریب است؟

۵.۲ به عنوان بخشی از یک برنامهٔ بازاریابی، یک بلوک شهری دارای ٤ خانوار انتخاب شده و از آن نمونـه‌ای متشکل از ۳ خانوار به صورت زیر گرفته شد: آقای *J*، دستیار تحقیق، خانوارها را شناسایی و هر یک را از ۱ تا ٤ شماره‌گذاری کرد. سپس تمام ترکیبهای دوتایی از ٤ خانوار را فهرست کرد. این ترکیبها چنین‌اند:

۱ , ۲	۱ , ۳	۱ , ٤
۲ , ۳	۲ , ٤	۳ , ٤

متاسفانه، آقای *J* بسیار بی‌دقت بود و این کار را به این دلیل به دست آورده بود که با رئیس اداره نسبتی داشت. اوترکیب ۳ , ٤ را فراموش کرد. یک شمارهٔ تصادفی را بین ۱ و ۵ انتخاب کرد (که تصادفاً ٤ درآمد) وترکیب ۲, ۳ برای نمونه انتخاب شد. متغیر موردنظر، هزینه‌های درمانی نقدی بود که هرخانوار متقبل می‌شد. این هزینه‌ها برای ٤ خانوار به شرح زیر بودند:

خانوار	هزینه‌ها (به دلار)
۱	۳٤۵/–
۲	۱۲٦/–
۳	٤۹۲/–
٤	۹٦۲/–

الف. بر اساس شیوهٔ نمونه‌گیری *J*، میانگین، خطای معیار و میانگین تـوان دوم خطای برآورد میانگین هزینه‌های درمانی نقدی چقدر است؟

ب. آیا برآورد حاصل از این شیوهٔ نمونه‌گیری، نااریب است؟

پ. آیا هر خانوار شانس یکسانی برای انتخاب شدن در نمونه دارد؟ چرا آری؟ چرا نه؟

٦.۲ مایل‌اند برای کنترل کیفیت داده‌های آزمایشگاهی، از یـک آزمایش بالینی بزرگ به منظـور برآورد کردن نسبت مقادیر آزمایشگاهی غیر معتبرموجود در پایگاه اطلاعاتی بررسیی انجام دهند. دراین آزمایش بالینی ۳۹٤ بیمار حضور دارند که برای هر یک از ٦۰ تا ۲۰۰ تشخیص آزمایشگاهی در طی دورهٔ آزمایش به دست آمده است. برنامهٔ نمونه‌گیری انتخابی، انتخاب یک نمونهٔ تصادفی متشکل از ۱۰ بیمار بود و برای هر فرد نمونه ۱۰ تشخیص آزمایشگاهی به

صورت نمونهٔ تصادفی انتخاب شد که بعداً بررسی درستی آن با سوابق درمانی بیمار چک می‌شد.

الف. واحدهای اولیه در این طرح نمونه‌ای کدام‌اند؟

ب واحدهای نمونه‌گیری در این طرح نمونه‌ای کدام‌اند؟

پ. آیا هر یک از افراد جامعه شانس برابر برای انتخاب شدن در نمونه را دارند؟

ت. آیا هر تشخیص آزمایشگاهی در این جامعه، شانس برابر برای انتخاب شدن در نمونه را دارد؟

۷.۲ داده‌های حاصل از نمونهٔ توصیف شده در تمرین ٦.٢ به شرح زیر است:

آزمودنی	کل مقادیر نامعتبر آزمایشگاهی در ۱۰ نمونه
۱	۱
۲	۰
۳	۲
٤	۰
٥	۱
٦	۰
۷	۰
۸	۰
۹	۰
۱۰	۱

بر اساس این داده‌ها، برآورد نسبت نتایج آزمایشگاهی نامعتبر چقدر است؟

۸.۲ جدول زیر تعداد کل تشخیصهای آزمایشگاهی برای ۱۰ بیمار نمونهٔ توصیف شده در تمـــرین ٦.٢ همراه با تعداد تشخیصهای نامعتبر بین ۱۰ تشخیص نمونه را نشان می‌دهد:

بر اساس این داده‌ها، نسبت مقادیر آزمایشگاهی نامعتبر را با استفاده از همهٔ داده‌های جدول بالا برآورد کنید.

آزمودنی	مجموع مقادیر آزمایشگاهی	مجموع مقادیر نامعتبر آزمایشگاهی در ۱۰ نمونه
۱	۱۰۳	۱
۲	۱۲۳	۰
۳	۹۳	۲
٤	۲۰۰	۰
٥	۱۲۸	۱
٦	۱٦٥	۰
۷	۱۳۲	۰
۸	۱۸۹	۰
۹	۱۷٦	۰
۱۰	۱۸۰	۱

۹.۲ اصطلاح ستون ۱ را با مناسبترین اصطلاح ستون ۲ جور کنید.

ستون ۱	ستون ۲	
۱- میانگین توان دوم خطا	الف.	$\sum_{i=1}^{n} x_i$
۲- قابلیت اعتماد	ب.	جامعه
۳- اریبی	پ.	درستی
٤- جامعه	ت.	$\sum_{i=1}^{N} X_i$
٥- مجموع نمونه	ث.	اعتبار
٦- مجموع جامعه	ج.	واریانس

کتابشناسی

The sampling texts cited in Chapter 1 [1-16] all develop the concepts discussed in this chapter, each in its own way. *The following sampling text by Hajek [1] presents in Chapter 1 an especially good discussion of the concepts developed in this chapter.*

1. Hajek, J. H., *Sampling from a Finite Population,* Marcel Dekker, New York and Basel, 198 1 .

The concept of counting rule will be developed further in a later chapter when we discuss the topic of network sampling. The following articles give examples of the use of counting rules in sample surveys.

2. Sirken, M. G., Household surveys with multiplicity. *Journal of the American Statistical Association,* 65: 257, 1970.
3. Sirken, M. G., and Levy, P. S., Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association,* 69: 68, 1974.

4. Czaja, R., and Blair, J., Using network sampling in victimization surveys. *Journal of Quantitative Criminology,* 6: 186. 1990.
5. Sirken, M. G., Network sampling. In *The Encyclopedia of Bioslatistics.* Armitage, P. A., and Colton, T., Eds. Wiley, Chichester, U.K., 1998.

The following expository articles appearing in The Encyclopedia of Biostatistics *provide a more complete discussion of some of the topics discussed in this chapter.*

6. King, B., Quota, representative, and other methods of purposive sampling. In *The Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T., Eds. Wiley, Chichester, U.K., 1998.
7. Warnecke, R. B., Sampling frames. In *The Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T., Eds. Wiley, Chichester, U.K., 1998.
8. Xia, Z., Probability sampling. In *The Encyclopedia of Biostatistics.* Armitage, P. A., and Colton, T., Eds. Wiley, Chichester, U.K., 1998.

قسمت ۲

طرحهای نمونه‌گیری عمده و شیوه‌های برآورد کردن

فصل ۳

نمونه‌گیری تصادفی ساده

در فصل ۲ با ارائهٔ مفاهیم کلی دربارهٔ جامعهٔ عام یا هدف، نمونه، توزیعهای نمونه‌گیری، و خاصیتهای مطلوب برآوردهای جامعه‌ای، بحث نمونه‌گیری را پایه‌ریزی کردیم. با شروع این فصل، از طریق ربط دادن این مفاهیم با برنامه‌های نمونه‌گیری خاص و شیوه‌های برآورد که به صورتی متداول در آمارگیریهای نمونه‌ای به کار می‌روند این پایه را تقویت می‌کنیم. این فرایند را با بحث دربارهٔ آن دسته از برنامه‌های نمونه‌گیری آغاز می‌کنیم که در آنها واحدهای اولیه، خود به عنوان واحدهای نمونه‌گیری به کار گرفته می‌شوند. این قبیل شیوه‌های نمونه‌گیری را گاهی اوقات *نمونه‌گیری از عناصر* می‌نامند. در این فصل از یک روش نمونه‌گیری از عناصر ـ یعنی نمونه‌گیری تصادفی ساده ـ بحث می‌کنیم که نه از لحاظ استفاده وسیع آن در آمارگیریهای نمونه‌ای واقعی، بلکه از این نظر حایز اهمیت است که پایه‌ای فراهم می‌سازد که نظریهٔ آماری نمونه‌گیری بر آن پایه بنا شود. لذا، هرچند ممکن است در عمل هرگز تصمیم نگیرید که طرحی مبتنی بر نمونه‌گیری تصادفی ساده انتخاب کنید (از علت این امر در بخش ۸.۳ بحث می‌کنیم)، ولی، فصل حاضر را به خاطر سهمی که نمونه‌گیری تصادفی ساده در نظریهٔ نمونه‌گیری دارد به بحث دربارهٔ آن اختصاص می‌دهیم.

۱.۳ نمونهٔ تصادفی ساده چیست؟

فـرض کنید جامعه‌ای با N عنصر داریم و می‌خواهیم یک نمونهٔ nتایی از این عناصر انتخاب کنیم. از نظـریهٔ ریاضـی جایگشتها و ترکیبها می‌توان نشان داد که T، تعداد نمونه‌های ممکن n عنصری از یک جامعهٔ N عنصری با $\binom{N}{n}$ نشان داده می‌شود و از فرمول زیر به دست می‌آید:

$$T=\binom{N}{n}=\frac{N!}{n!(N-n)!} \qquad (۱.۳)$$

که در آن $n!=n(n-۱)(n-۲)\dots(۱)$ و $۱=!۰$. مثلاً اگر جامعه‌ای دارای ۲۵ عنصر باشد و بخواهیم یک نمونهٔ ۵ عنصری انتخاب کنیم، در آن صورت از رابطهٔ (۱.۳) وقتی $N=۲۵$ و $n=۵$، خواهیم داشت:

$$T=\frac{۲۵!}{(۵!)(۲۵-۵)!}=۵۳۱۳۰$$

به این ترتیب ۵۳۱۳۰ نمونهٔ ممکن ۵ عنصری از یک جامعهٔ ۲۵ عنصری وجود دارند.

با در نظر داشتن این رابطه، تعریف زیر را از نمونه‌گیری تصادفی ساده خواهیم داشت:

یـک *نمونـهٔ تصـادفی سـادهٔ* n *عنصـری* از یک جامعهٔ N عنصری، نمونه‌ای است که در آن هر یک از $\binom{N}{n}$ نمونهٔ ممکن n عنصری احتمال برابر برای انتخاب شدن داشته باشند، یعنی احتمال $۱/\binom{N}{n}$.

باید متذکر شد که نوع نمونه‌گیریی که در بالا و در سراسر این کتاب مورد بحث قرار گرفته است با نام *نمونه‌گیری بدون جایگذاری* شناخته می‌شود. در نمونه‌گیری بدون جایگذاری، هر عنصر مشخص فقط یک بار می‌تواند در یک نمونه ظاهر شود، در حالی که اگر نمونه‌گیری با جایگذاری باشد یک عنصر معین می‌تواند بیش از یک بار در یک نمونهٔ خاص ظاهر شود. در عمل، تقریباً تمام نمونه‌گیریها بدون جایگذاری انجام می‌گیرند. نمونه‌گیری با جایگذاری فقط به دلایل نظری مورد توجه است، زیرا نظریهٔ ریاضی این‌گونه نمونه‌گیری بهتر از نظریهٔ نمونه‌گیری بدون جایگذاری اداره کردنی است. نتایج حاصل از فرضهای نمونه‌گیری با جایگذاری در شرایطی معین تقریباً، حتی اگر نمونه‌گیری بدون جایگذاری انجام گیرد درست‌اند. برای بحثهای مربوط به نمونه‌گیـــری با جایگذاری، در صورت تمایل می‌توانیـــد به متون نمونه‌گیریی که بیشتر جنبهٔ ریاضی دارند نظیر آثار کوکران [۱]، هنسن و همکاران [۲]، یا کیش [۳] رجوع کنید.

۱.۱.۳ چگونگی گرفتن نمونهٔ تصادفی ساده

اولین گام در راه گرفتن نمونهٔ تصادفی ساده، تخصیص شماره به عناصر جامعه به ترتیب از ۱ تا N است. گام بعدی، انتخاب یک نمونهٔ n تایی از این شماره‌ها با استفاده از نوعی فرایند تصادفی از قبیل

جدول اعداد تصادفی، رایانه، یا ماشین حسابی با مولد اعداد تصادفی است. هر شیوه‌ای که به کار گرفته می‌شود باید اطمینان دهد که شماره‌های انتخاب شده همه با هم تفاوت دارند و هیچکدام بزرگتر از N نیستند. به محض انتخاب شماره‌ها، عناصری از جامعه که با آن شماره‌ها مطابقت دارند به عنوان نمونه اختیار می‌شوند.

برای مثال اگر بخواهیم شش پزشک را به صورت نمونهٔ تصادفی از فهرست متشکل از ۲۵ پزشک در جدول ۱.۲ انتخاب کنیم، می‌توانیم از جدول اعداد تصادفی پیوست (جدول پ ـ۱) استفاده کنیم. با استفاده از اعداد دورقمی، از یک نقطهٔ دلخواه شروع می‌کنیم (برای مثال ردیف ۷، دو رقم اول ستون۱) و به طرف پایین ستون پیش می‌رویم تا شش شمارهٔ مختلف بین ۰۱ تا ۲۵ انتخاب شوند. تمام شماره‌های برابر با ۰۰ یا بیشتر از ۲۵ را کنار می‌گذاریم. برای این مثال، اعدادی که به دست می‌آوریم عبارت‌اند از ۰۹، ۱۰، ۰۷، ۰۲، ۰۱ و ۰۵ این اعداد تصادفی، افرادی را که باید در نمونهٔ مورد مطالعه قرار بگیرند مشخص می‌کنند. (اگر این مثال را با مراجعه به جدول پ ـ۱ انجام دهید ملاحظه می‌کنید که با شماره‌های ۰۷ و ۰۲ دو بار مواجه می‌شوید. چون بدون جایگذاری نمونه‌گیری می‌کنیم بار دوم که به این شماره‌ها برمی‌خوریم آنها را کنار می‌گذاریم. همچنین توجه کنید که وقتی به انتهای ستون خاصی رسیدیم از بالاترین سطر ستون بعدی در همان صفحه به کار ادامه می‌دهیم. تا وقتی که ستونهای صفحهٔ فعلی تمام نشده است به صفحهٔ بعد اعداد تصادفی نمی‌رویم).

۲.۱.۳ احتمال انتخاب شدن یک عنصر

در نمونه‌گیری تصادفی ساده، $\binom{N}{n}$ نمونهٔ ممکن n عنصری از یک جامعهٔ متشکل از N عنصر وجود دارند و هر نمونه، احتمال $1/\binom{N}{n}$ برای انتخاب شدن دارد. همچنین در نمونه‌گیری تصادفی ساده احتمال انتخاب شدن هر عنصر برابر با $\frac{n}{N}$، یعنی نسبت اندازهٔ نمونه به اندازهٔ جامعه است. این موضوع را می‌توان به راحتی با استدلال زیر نشان داد. تعداد نمونه های n عنصری که فاقد عنصر خاصی است برابر با $\binom{N-1}{n}$ است و این، تعداد راههایی است که $N-1$ عنصر دیگر می‌توانند در گروههایی با اندازهٔ n ترکیب شوند. بنابراین، احتمال این که هر عنصری در نمونه گنجانده نشود، برابر است با $\frac{\binom{N-1}{n}}{\binom{N}{n}}$، که خود برابر است با $\frac{N-n}{N}$. لذا احتمال گنجانیده شدن هر عنصر برابر است با $1-\frac{N-n}{N}$ یا $\frac{n}{N}$.

عملاً در تمام برنامه‌های نرم‌افزاری آماری استاندارد، شیوه‌های مبتنی بر مولدهای اعداد تصادفی برای انتخاب نمونه‌های تصادفی ساده، با یا بدون جایگذاری، موجودند. در صورت امکان، استفاده از رایانه برای تولید نمونه‌ها می‌تواند بخش عمده‌ای از زحمت کار فرایند نمونه‌گیری را حذف کند. ولی کاربر باید دقت کند تا اطمینان حاصل نماید که روش به کار رفته برای تولید رایانه‌ای نمونه با طرح نمونه‌گیری تعیین شده مطابقت داشته باشد.

یکی از مفیدترین برنامه‌های نمونه‌گیری مدول SAMPLE [۷] است که برای Version5 SYSTAT نوشته شده و یک نرم‌افزار آماری برای مقاصد کلی با کاربری وسیع است. در حال حاضر به نظر می‌رسد که این تنها مدول نمونه‌گیری است که با یک برنامهٔ آماری مبتنی بر PC (رایانهٔ شخصی) برای مقاصد کلی تلفیق شده و دارای قابلیت نمونه گرفتن از انواع گوناگونی از برنامه‌های نمونه‌گیری متداول در آمارگیریهای نمونه‌ای است. (مانند نمونه‌گیری تصادفی ساده، نمونه‌گیری سیستماتیک، نمونه‌گیری تصادفی طبقه‌بندی شده، و نمونه‌گیری با احتمال متناسب با اندازه).

۲.۳ برآورد مشخصه‌های جامعه تحت نمونه‌گیری تصادفی ساده

۱.۲.۳ فرمولهای برآورد

در نمونه‌گیری تصادفی ساده، مقادیر x'، $\bar{x}$، p_y و s_x^2 که در زیر آمده‌اند به ترتیب، برآوردهای مجموعها، میانگینها، نسبتها و واریانسهای جامعه‌ای هستند که بدون استثنا در طرحهای نمونه‌گیری تصادفی ساده به کار می‌روند:

$$x' = \frac{N\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$p_y = \frac{\sum_{i=1}^{n} y_i}{n}$$

(در این فرمولها y_i متغیری دو حالتی با مقادیر ۰ یا ۱ است)

$$\hat{\sigma}_x^2 = \left(\frac{N-1}{N}\right) s_x^2$$

که در آن

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

تابلوی ۱.۳ برآورد مجموعها، میانگینها، نسبتها و واریانسها، تحت نمونه‌گیری تصادفی ساده، و برآورد واریانسها و خطاهای معیار این برآوردها

برآورد	واریانس برآورد شده و خطای معیار برآورد
مجموع	
$x' = \frac{N}{n}\sum_{i=1}^{n} x_i$	$\hat{Var}(x') = N^2\left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right)$ $\hat{SE}(x') = N\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$
میانگین	
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\hat{Var}(x') = \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right)$ $\hat{SE}(x') = \sqrt{\frac{N-n}{N}}\,\frac{s_x}{\sqrt{n}}$
نسبت	
$p_y = \frac{\sum_{i=1}^{n} y_i}{n}$	$\hat{Var}(p_y) = \left(\frac{N-n}{N}\right)\frac{p_y(1-p_y)}{n-1}$ $\hat{SE}(p_y) = \sqrt{\frac{N-n}{N}}\sqrt{\frac{p_y(1-p_y)}{n-1}}$

که در آن y، متغیر دوحالتی با مقدار صفر یا یک است.

به طوری که در بخش بعد بحث خواهیم کرد برآوردهای مزبور به این دلیل به کار می‌روند که برآوردهای نااریبی از پارامترهای جامعه‌ای متناظر هستند. برآورد واریانسهای x'، $\bar{x}$ و p_y در تابلوی ۱.۳ نشان داده شده‌اند.

۲.۲.۳ محاسبهٔ عددی برآوردها و خطای معیار آنها

از لحاظ تاریخی، تولید برآوردها و خطای معیار آنها از روی داده‌های آمارگیریهای نمونه‌ای مستلزم استفاده از برنامه‌هایی برای مقاصد خاص یا ماکرو بوده است و به طور روزمره با استفادهٔ مستقیم از مدولهای موجود در نرم‌افزارهای آماری استاندارد برای مقاصد کلی (از قبیل SAS، SPSS و BMDP) قابل اجرا نبوده است. حتی برای نمونه‌گیری تصادفی ساده نیز برآورد خطای معیار یک میانگین یا نسبت نمونه‌ای که از معمولیترین نرم‌افزارهای آماری به دست می‌آید خطای معیار واقعی را با ضریب

$\sqrt{\frac{N}{N-n}}$ بیشتر از اندازهٔ واقعی برآورد خواهد کرد. به طور کلی، محاسبهٔ برآوردهای مناسب و خطـای معـیار آنها از روی داده‌های آمارگـیری نمونه‌ای مستـلزم استفـاده از برنامه‌هـایی نظـیر WESVAR، SUDAAN و PC-CARP بوده است که نرم‌افزارهایی برای مقاصد کلی نیستند بلکه هدف اصلی از نگارش آنها اجرای تحلیل داده‌های حاصل از آمارگیریهای نمونه‌ای بوده است.

ولی اخیراً تولیدکنندگان نرم‌افزارهای آماری برای مقاصد کلی به گسترش این قبیل مدولها توجه کرده‌اند. یکی از نرم‌افزارهای آماری کلی که کاربردی وسیع پیدا کرده STATA است که مجموعه‌ای از فرمانها را در Version5.0 تلفیق کرده است که می‌توانند برآوردها و خطای معیار آنها را برای انواع گوناگونی از طرحهای نمونه ای محاسبه کنند و محتمل است که سایر نرم‌افزارهای عمده نیز در آیندهٔ نزدیک تواناییهایی مشابه پیدا کنند. چون افزایش زیادی در دسترسی به این گونه نرم‌افزارها را پیش‌بینی می‌کنیم، در بسیاری از مثالهای تشریحی عددی، دربارهٔ چگونگی محاسبهٔ برآوردها و خطـای معیار آنها با استفاده از STATA و یا SUDAAN بحث خواهیم کرد. STATA یک نرم‌افزار آماری برای مقاصد کلی است که اخیراً مدولهایی برای تحلیل داده‌های آمارگیری نمونه‌ای به آن اضافه شده و به صورتی گسترده مورد استفاده قرار گرفته است. انتخاب این دو نرم‌افزار از جانب ما به مفهوم تأیید ترجیح آنها نسبت به بقیهٔ نرم‌افزارها نیست، بلکه یه این دلیل آنها را انتخاب کرده‌ایم که فعلاً آشنایی بیشتری با آنها داریم.

مثال تشریحی: در ایلی‌نوی نمونه‌ای از بیمارستانها گرفته شد که هدف از آن انجام بررسی مربوط به غربالگری نوزادان و مادران به دلیل پادگن سطحی هپاتیت B (HBsAG) بود که نشانه‌ای برای ابتلا به ویروس هپاتیت B است. این یک آمارگیری نسبتاً بزرگ بود که طرح نمونه‌گیری پیچیده‌ای داشت و در سراسر این کتاب برای نشان دادن شیوه‌ها و مفاهیم عمده مورد استفاده قرار خواهد گرفت. برای این مثال خاص، کاربرد STATA را برای محاسبهٔ برآوردها و خطاهای معیار نشان خواهیم داد و فرض خواهیم کرد که نمونهٔ تصادفی ساده‌ای متشکل از ۲۵ گزارش تولد داریم که از مجموع ۷۷۳ مورد تولد که در طول سال در یکی از بیمارستانهای نمونه روی داده گرفته شده‌اند.

داده‌های مربوط در یک پروندهٔ حاوی ۲۵ گزارش تنظیم شده است که هر گزارش دارای متغیرهای زیر است:

hospno : شمارهٔ بیمارستان: شمارهٔ شناسایی بیمارستان (در این مورد hospno=۱۳)

birth : تولد: کل تعداد N تولد که در طول سال در بیمارستان روی داده است (N =۷۷۳)

weight1 : وزن ۱ : معکوس کسر نمونه‌گیری ($\frac{N}{n} = \frac{۷۷۳}{۲۵} =$ ۳۰/۹۲)

momsag : پادگن مادر: آیا وضعیت پادگنی مادر با توجه به (HBsAG) پادگن سطحی هپاتیت B، در گزارش نوزاد وارد شده است یا خیر (اگر بلی momsag =۱ و اگر نه momsag =۰).

پرونـدهٔ حاوی ۲۵ گزارش همراه با متغیـــــرها به ترتیب (شمارهٔ بیمارستان، تولد، وزن ۱ و پادگن مادر) در زیر نشان داده شده است.

hospno	birth	Weight1	momsag
13	773	30.92	0
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	0
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1
13	773	30.92	1

برای این که STATA بتواند داده‌ها را برای طرح نمونه گیری خاصی پردازش کند هر گزارش باید دارای اطلاعاتی مشخص باشد که برای آن طرح موردنیاز است. برای یک نمونهٔ تصادفی ساده، دو متغیر زیر موردنیاز است:

۱. وزن نمونه‌گیری، $\frac{N}{n}$. در این پرونده این وزن به صورت متغیر وزن۱ ارائه شده و برابر با $\frac{۷۷۳}{۲۵}$ یا ۳۰/۹۲ برای هر گزارش است.

۲. تعداد واحدهای شمارش N در جامعه. این اطلاع برای محاسبهٔ مقدار $\frac{(N-n)}{N}$ موردنیاز است که در فرمول خطای معیار برآورد یک ضریب محسوب می‌شود. اطلاع مزبور در این پرونده در متغیر تولد قرار داده شده و برای هر گزارش برابر با ۷۷۳ است.

سپس داده‌ها باید در یک فایل داده‌ای STATA قرار داده شود که دارای انشعاب *dta* باشد. برای انجام این کار راههای گوناگونی وجود دارند و برای اطلاعات بیشتر می‌توان به راهنمای [۹]STATA رجوع کرد. پروندهٔ حاصل را *momsag.dta* (پادگن مادر) می‌نامیم.

اکنون می‌توان داده‌ها را با مجموعه فرمانهای زیر پردازش کرد:

```
. use a:\momsag
. svyset pweight weight1
. svyset fpc birth
. svymean momsag
. svytotal momsag
```

فرمان اول حاکی از آن است که باید از مجموعهٔ *momsag.dta* (داده‌های پادگن مادر) که در فلاپی دیسکت در درایو A قرار دارد استفاده شود. دو فرمان بعدی دو پارامتر لازم برای نمونهٔ تصادفی ساده را پردازش می‌کنند. فرمان svyset pweight weight1 نشان می‌دهد که وزن نمونه‌گیری $\frac{N}{n}$ که STATA با واژهٔ *pweight* به آن رجوع می کند در متغیر موسوم به *weight1* قرار دارد. فرمان svyset fpc birth پارامتر N را که در ضریب $\frac{(N-n)}{N}$ مربوط به فرمولهای خطای معیار برآورد به کار می‌رود تنظیم می‌کند. در این مورد فرمان مزبور نشان می‌دهد که این ضریب در متغیر موسوم به *birth* (تولد) قرار دارد. بالاخره، دو فرمان آخر به برنامه دستور می‌دهند که میانگین و مجموع متغیر پادگن مادر *momsag* را برآورد کند. STATA خروجی زیر را تولید می‌کند.

برآورد مجموع، که در بالا نشان داده شده است و برآورد خطای معیار آن دقیقاً همانهایی هستند که از فرمولهای مناسب ارائه شده در تابلوی ۱.۳ به دست می‌آیند. میانگین برآورد شده در خروجی رایانه عبارت است از میانگین یک متغیر دو حالتی با مقدار ۰ یا ۱ و به عنوان یک نسبت تفسیر می‌شود. به همین ترتیب، این متغیر وخطای معیار آن دقیقاً همانها هستند که از فرمولهای مناسب ارائه شده در تابلوی ۱.۳ به دست می‌آیند.

.svymean momsag

Survey mean estimation

pweight: weight1 Number of obs = 25
Strata: <one> Number of strata = 1
PSU: <observations> Number of PSUs = 25
FPC: birth Population size = 773

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
momsag	.92	.0544746	.8075699	1.03243	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

. svytotal momsag

Survey total estimation

Pweight: weight1 Number of obs = 25
Strata: <one> Number of strata = 1
PSU: <observations> Number of PSUs = 25
FPC: birth Population size = 773

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
momsag	711.16	42.10889	642.2515	798.0685	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

□

عبارت *PSU* [1] که در خروجی نشان داده شد به مفهوم *واحد نمونه‌گیری اولیه* است و در مورد نمونه‌گیری تصادفی ساده همان واحد شمارش است (که در این مثال همان گزارش بیمارستان است). در سایر طرحهای نمونه‌گیری ممکن است واحد نمونه‌گیری اولیه و واحد شمارش یکسان نباشند. عبارت *Deff* که در خروجی نشان داده شده است اصطلاحی است که به *اثر طرح* موسوم است و نسبت واریانس یک برآورد حاصل از طرح نمونه‌گیری خاص به واریانس یک برآورد حاصل از یک نمونهٔ تصادفی ساده با همان تعداد واحدهای شمارش است. این اثر در بخشهای بعدی این کتاب مورد بحث قرار خواهد گرفت.

دربارهٔ دلایل اینکه چرا از برآوردهایی که در تابلوی ۱.۳ نشان داده شده‌اند و از واریانسها و خطاهای معیار برآورد شدهٔ آنها استفاده می‌کنیم در بخش ۳.۳ بحث خواهیم کرد.

[1] Primary Sampling Unit

۳.۳ توزیعهای نمونه‌گیری مشخصه‌های برآورد شدهٔ جامعه

برآوردهای حاصل از نمونهٔ تصادفی ساده، برآوردهای نااریب از پارامترهای متناظر در جامعه هستند. این ایده را با یک مثال نشان می دهیم.

مثال تشریحی: به عنوان یک مثال ساده، همان جامعهٔ متشکل از شش مدرسه را که در جدول ۳.۲ نشان داده شد در نظر می‌گیریم و فرض می‌کنیم که می‌خواهیم از روی یک نمونهٔ تصادفی ساده متشکل از سه مدرسه، کل تعداد کودکان شش مدرسه را که در برابر سرخک ایمن‌سازی نشده‌اند برآورد کنیم. با استفاده از برآورد مجموع $x' = \left(\frac{N}{n}\right)x = \left(\frac{۶}{۳}\right)x$، برآوردها را در جدول ۱.۳ برای $\binom{۶}{۳}$ یا ۲۰ نمونه از سه مدرسه در جامعه‌ای متشکل از ٦ مدرسه نشان داده‌ایم.

توجه داشته باشید که هر مدرسه در ۱۰ نمونه از ۲۰ نمونهٔ ممکن قرار می‌گیرد. به این ترتیب امکان این که هر مدرسهٔ خاص در نمونه قرار بگیرد $\frac{۱۰}{۲۰}$ است که با $\frac{n}{N}$ یا $\frac{۳}{۶}$ برابر است.

توزیع نمونه‌گیری برآورد مجموع کل x' (با n=۳) برای داده‌های جدول ۱.۳ در جدول ۲.۳ ارائه شده است. در جدول ۲.۳ می‌بینیم که ۲۰ نمونه‌ای که دارای احتمال وقوع برابرند ۱۱ مقدار متفاوت x' را به بار می‌آورند.

میانگین، واریانس، و انحراف معیار توزیع نمونه‌گیری x' به شرح زیرند:

جدول ۱.۳ نمونه‌های ممکن متشکل از سه مدرسه و مقادیر x'

مدرسه‌ها در نمونه	x'	مدرسه‌ها در نمونه	x'
۱ , ۲ , ۳	۲٤	۲ , ۳ , ٤	۲۲
۱ , ۲ , ٤	۲٤	۲ , ۳ , ٥	۳۰
۱ , ۲ , ٥	۳۲	۲ , ۳ , ٦	۳۲
۱ , ۲ , ٦	۳٤	۲ , ٤ , ٥	۳۰
۱ , ۳ , ٤	۲۰	۲ , ٤ , ٦	۳۲
۱ , ۳ , ٥	۲۸	۲ , ٥ , ٦	٤۰
۱ , ۳ , ٦	۳۰	۳ , ٤ , ٥	۲٦
۱ , ٤ , ٥	۲۸	۳ , ٤ , ٦	۲۸
۱ , ٤ , ٦	۳۰	۳ , ٥ , ٦	۳٦
۱ , ٥ , ٦	۳۸	٤ , ٥ , ٦	۳٦

جدول ۲.۳ توزیع نمونه‌گیری x' در جدول ۱.۳

x'	f	$\pi = \frac{f}{t}$
۲۰	۱	۰/۰۵
۲۲	۱	۰/۰۵
۲٤	۲	۰/۱۰
۲٦	۱	۰/۰۵
۲۸	۳	۰/۱۵
۳۰	٤	۰/۲۰
۳۲	۳	۰/۱۵
۳٤	۱	۰/۰۵
۳٦	۲	۰/۱۰
۳۸	۱	۰/۰۵
٤۰	۱	۰/۰۵
مجموع	۲۰	۱/۰۰

$$E(x') = \sum_{i=۱}^{۱۱} x_i' \pi_i = ۲۰(۰/۰۵) + ۲۲(۰/۰۵) + \cdots + ۴۰(۰/۰۵) = ۳۰$$

$$V(x') = \sum_{i=۱}^{۱۱} [x_i' - E(x')]^۲ \pi_i = (۲۰ - ۳۰)^۲ (۰/۰۵) + (۲۲ - ۳۰)^۲ (۰/۰۵) + \cdots + (۴۰ - ۳۰)^۲ (۰/۰۵) = ۲۶/۴$$

$$SE(x') = \sqrt{Var(x')} = \sqrt{۲۶/۴} = ۵/۱۳۸$$

متوجه می‌شویم که پارامترهای جامعه‌ای X، $\sigma_x^۲$ و σ_x برای شش مدرسهٔ داده شده در جدول ۳.۲ عبارت‌اند از

$$X = ۳۰ \qquad \sigma_x^۲ = ۳/۶۷ \qquad \sigma_x = ۱/۹۱۵$$

و باز متوجه می‌شویم که برای توزیع نمونه‌گیری x' که در جدول ۲.۳ نشان داده شده است

$$E(x') = X = ۳۰$$

یعنی به نظر می‌رسد که برآورد مجموع در نمونه‌گیری تصادفی ساده، برآورد نااریب مجموع کل جامعهٔ واقعی است.

□

می‌توان نشان داد که واریانس $Var(x')$ و انحراف معیار $SE(x')$ در توزیع نمونه‌گیری x'، از فرمولهای زیر به دست می‌آیند:

$$Var(x') = \left(\frac{N^2}{n}\right)(\sigma_x^2)\left(\frac{N-n}{N-1}\right)$$

و

$$SE(x') = \left(\frac{N}{\sqrt{n}}\right)(\sigma_x)\left(\frac{N-n}{N-1}\right)^{1/2}$$

در مورد مثال بالا

$$Var(x') = \left(\frac{6^2}{3}\right)(3/67)\left(\frac{6-3}{6-1}\right) = 26/4$$

که این همان نتیجه‌ای است که از طریق محاسبهٔ مستقیم به دست آمد.

بـه ایـن تـرتیب، مـی‌بیـنیم کـه برآورد مجموع جامعه، x'، از روی نمونه‌گیری تصادفی ساده، یک بـرآورد نـااریب از مجمـوع جامعـه، X، اسـت. خطـای معیـار x' که از معادلهٔ بالا به دست مـی‌آید با انحراف معیار σ_x توزیع x در جامعه نسبت مستقیم و با ریشهٔ دوم اندازهٔ نمونه n نسبت معکوس دارد. خطـای معیـار بـه ریشه دوم ضریب $\left(\frac{N-n}{N-1}\right)$ که به نام *تصحیح جامعهٔ متناهی* موسوم است و غالباً با نماد fpc[1] نشان داده می‌شود بستگی دارد.

بـرای بـه دسـت آوردن شـناختی از نقش fpc می‌توانیم مقدار آن را برای یک جامعهٔ فرضی شامل N=۱۰۰۰ عنصـر و بـرای انـدازه‌های نمونه‌ایی که در جدول ۳.۳ ارائه شده است بررسی کنیم. از این جـدول پی می‌بریم که اگر اندازهٔ نمونه، n، بسیار کمتر از اندازهٔ جامعه، N، باشد، fpc بسیار نزدیک به یـک اسـت و از ایـن رو تأثیر بسیار کمی بر مقدار عددی خطای معیار $SE(x')$ برآورد مجموع کل x'، خواهـد داشـت. از سـوی دیگر، وقتی که n به N نزدیکتر می‌شود، بزرگی fpc کاهش می‌یابد و به این ترتیب کاهش در مقدار $SE(x')$ را موجب می‌شود.

اگر به جای fpc، $\sqrt{\text{fpc}}$ را در نظر بگیریم

$$\sqrt{fpc} = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{N}{N-1}}\sqrt{1-\frac{n}{N}}$$

در این صورت می‌بینیم که مقـدارfpc به جای بستگی به بزرگی مطلـق n و N منحصـراً به نسبت n بـه N بستـگی دارد. معقول به نظر می‌رسد که fpc در مقدار $SE(X')$ به صورت یک ضریب است زیـرا، نـزدیک بـودن n بـه N حاکی از آن است که بیشتر عناصر جامعه در نمونه قرار دارند. به علاوه نتیجه می‌شود که تعداد نسبتاً کمی از عناصر که در نمونه قرار نگرفته‌اند تأثیر بسیار اندکی بر توزیع x' دارند.

[1] finite population correction

جدول ۳.۳ مقادیر fpc $(N = ۱۰۰۰)$

اندازهٔ نمونه، n،	$\text{fpc} = \sqrt{\frac{N-n}{N-1}}$
۱	۱/۰۰۰۰
۱۰	۰/۹۹۹۵
۱۰۰	۰/۹۹۵۰
۵۰۰	۰/۹۷۴۷
۱۰۰۰	۰/۹۴۸۷
۵۰۰۰	۰/۷۰۷۱
۹۰۰۰	۰/۳۱۶۲

روشی نظیر آنچه که در بالا برای ارائهٔ ویژگیهای مجموع نمونه تحت نمونه‌گیری تصادفی ساده داده شد می‌توان برای ویژگیهای توزیع نمونه‌گیری میانگینهای نمونه‌ای $\bar{x}$ و نسبتهای نمونه‌ای p_yی صفت y نیز ارائه داد. از هر توزیع نمونه‌گیری، جداگانه بحث نخواهیم کرد ولی ویژگیهای آنها را در تابلوی ۲.۳ خلاصه می‌کنیم. با توجه به تابلوی ۲.۳ پی می‌بریم که مجموعهای نمونه‌ای، میانگینها، و نسبتها تحت نمونه‌گیری تصادفی ساده، برآوردهای ناااریب از پارامترهای جامعه‌ای نظیر هستند.

۴.۳ ضریب تغییرات پارامترهای برآورد شدهٔ جامعه

در فصل قبل، ضریب تغییرات V_x توزیع متغیر x را در یک جامعه به صورت انحراف معیار توزیع x تقسیم بر مقدار میانگین x در آن جامعه تعریف کردیم. به همین ترتیب *ضریب تغییرات* $V(\hat{d})$ را، که $\hat{d}$ برآورد پارامتر dی جامعه است، به عنوان خطای معیار $SE(\hat{d})$ آن تقسیم بر مقدار واقعی dی پارامتر برآورد شده تعریف می کنیم. به عبارت دیگر:

$$V(\hat{d}) = \frac{SE(\hat{d})}{d} \qquad (۵.۳)$$

توان دوم ضریب تغییرات، یعنی $V^۲(\hat{d})$ یک پارامتر برآورد شده، به نام *واریانس نسبی پارامتر برآورد شده* موسوم است.

ضریب تغییرات یک پارامتر برآورد شده، تغییرپذیری نمونه‌ای برآورد را نسبت به مقدار پارامتر برآورد شده اندازه‌گیری می‌کند و در ارزیابی قابلیت اعتماد برآورد به کار می‌رود. به طوری که بعداً خواهیم دید این ضریب به عنوان یک عامل در بسیاری از نتایج بسیار مهم در نظریهٔ نمونه‌گیری ظاهر می‌شود.

تابلوی ۲.۳ برآوردهای جامعه‌ای، میانگینها و خطاهای معیار برآوردهای جامعه‌ای تحت نمونه‌گیری تصادفی ساده

مجموع x'

$$x' = \frac{N\sum_{i=1}^{n} x_i}{n} = \left(\frac{N}{n}\right)x \qquad E(x') = X$$

$$SE(x') = \left(\frac{N}{\sqrt{n}}\right)(\sigma_x)\sqrt{\frac{N-n}{N-1}} \qquad (۲.۳)$$

میانگین $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n} \qquad E(\bar{x}) = \bar{X}$$

$$SE(x') = \left(\frac{\sigma_x}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}} \qquad (۳.۳)$$

نسبت، p_y

$$p_y = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y}{n} \qquad E(p_y) = P_y$$

$$SE(x') = \left(\frac{N}{\sqrt{n}}\right)(\sigma_x)\sqrt{\frac{N-n}{N-1}} \qquad (٤.۳)$$

نمادهای مورد استفاده در این فرمولها در تابلوهای ۱.۲ و ۲.۲ تعریف شده‌اند. n، تعداد عناصر موجود در نمونه‌ای است که از جامعه‌ای متشکل از N عنصر گرفته شده است.

تابلوی ۳.۳، ضریب تغییرات را برای برآوردهای مجموعها، میانگینها و نسبتها، در یک نمونهٔ تصادفی ساده، ارائه می‌دهد.

نتایجی که در تابلوی ۳.۳ نشان داده شده‌اند، با جایگزین کردن مقادیر خطاهای معیار نشان داده شده در تابلوی ۲.۳ در فرمولهای مربوط به ضریب تغییرات یک برآورد [معادلهٔ (۵.۳)] به دست آمده‌اند. برای مثال، ضریب تغییرات یک برآورد مجموع x' بنابر تابلوی ۲.۳ و معادلهٔ (۵.۳) عبارت است از:

$$V(x') = \frac{\left(\frac{N}{\sqrt{n}}\right)(\sigma_x)\sqrt{\frac{N-n}{N-1}}}{X}$$

ولی $X = N\bar{X}$ و بنابراین:

$$V(x') = \left(\frac{1}{\sqrt{n}}\right)\left(\frac{\sigma_x}{\bar{X}}\right)\sqrt{\frac{N-n}{N-1}}$$

چون بنا بر تعریف، $\frac{\sigma_x}{\bar{X}}$، ضریب تغییرات توزیع x در جامعه است، داریم:

$$V(x') = \left(\frac{V_x}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

سایر نتایج نشان داده شده در تابلوی ۳.۳ را نیز می‌توان به همین ترتیب به دست آورد. توجه کنید که ضریب تغییرات برآورد میانگین $\bar{x}$، تحت نمونه‌گیری تصادفی ساده، با ضریب تغییرات مجموع برآورد شده یکسان است.

تابلوی ۳.۳ ضرایب تغییرات برآوردهای جامعه تحت نمونه‌گیری تصادفی ساده

مجموع x'

$$V(x') = \left(\frac{V_x}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}} \qquad (٦.٣)$$

میانگین $\bar{x}$

$$V(\bar{x}) = \left(\frac{V_x}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}} \qquad (٧.٣)$$

نسبت، p_y

$$V(p_y) = \sqrt{\frac{(1-p_y)}{np_y}}\sqrt{\frac{N-n}{N-1}} \qquad (٨.٣)$$

در این فرمولها V_x، ضریب تغییرات متغیر x، p_y نسبت جامعه‌ای، n، تعداد عناصر موجود در نمونه، و N تعداد عناصر موجود در جامعه است.

۵.۳ قابلیت اعتماد برآوردها

خطای معیار یک برآورد،اندازه‌ای از تغییرپذیری برآورد نمونه‌گیری روی همهٔ نمونه‌های ممکن است. با این فرض که خطای اندازه‌گیری وجود ندارد یا قابل چشم‌پوشی است، *قابلیت اعتماد یک برآورد* را می‌توان از روی اندازهٔ خطای معیار قضاوت کرد. هر چه خطای معیار بیشتر باشد، قابلیت اعتماد برآورد کمتر است (بخش ٤.٢ را ببینید).

اگر فرض کنیم که برآوردهای مورد بحث در بخش قبل برای مقادیر nی که به صورتی معقول بزرگ‌اند (مثلاً بیشتر از ۲۰) توزیعهایی دارند که به توزیع گاوسی یا نرمال نزدیک‌اند، در آن صورت می‌توانیم از نظریهٔ نرمال برای به دست آوردن بازهٔ اطمینان تقریبی پارامترهای مجهول جامعه‌ای مورد

برآورد استفاده کنیم. مثلاً، بازه‌های اطمینان تقریبی ٪$(1-\alpha)$۱۰۰ برای مجموع جامعه از فرمول زیر به دست می‌آید:

$$x' \pm z_{1-(\alpha/2)}\left(\frac{N}{\sqrt{n}}\right)(\sigma_x)\sqrt{\frac{N-n}{N-1}}$$

که در آن $z_{1-(\alpha/2)}$، صدک $[1-(\alpha/2)]$۱۰۰ توزیع نرمال استاندارد است. مثلاً برای بازهٔ اطمینان ۹۵٪، $\alpha =$ ۰/۰۵ و $z_{1-(\alpha/2)} = z_{0/975} =$ ۱/۹۶.

چون σ_x یک پارامتر جامعه‌ای نامعلوم است باید از روی نمونه برآورد شود. اگر به جای σ_x در فرمول بالا قرار دهیم،

$$\hat{\sigma}_x = \sqrt{\left(\frac{N-1}{N}\right)(s_x^2)}$$

بازهٔ اطمینان تقریبی زیر را برای مجموع X به دست می‌آوریم:

$$x' \pm z_{1-(\alpha/2)}(N)\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$$

مثال تشریحی: فرض کنید یک نمونهٔ تصادفی سادهٔ ۹ تایی از ۲۵ پزشکی که در جدول ۱.۲ فهرست شده‌اند به این منظور گرفته شود که مجموع X ویزیت پزشکان موجود از خانوارها در جامعه برآورد شود. فرض کنیم که پزشکان انتخاب شده شماره‌های ۱۳، ۳، ۱۷، ۱، ۱٤، ۱۲، ۷، ۱۸ و ٤ هستند. داده‌های نمونه در جدول ٤.۳ ارائه شده‌اند. آماره‌های نمونه (با استفاده از فرمولهای تابلوهای ۲.۲ و ۲.۳) عبارت‌اند از:

$$x = ۴۴ \qquad s_x = ۳/۴۸ \qquad x' = \left(\frac{۲۵}{۲۹}\right)(۴۴) = ۱۲۲/۲۲$$

جدول ٤.۳ داده‌های نمونه برای تعداد دفعات ویزیت از خانوار

پزشک	تعداد ویزیت	پزشک	تعداد ویزیت
۱	۵	۱۳	٦
۳	۱	۱٤	٤
٤	٤	۱۷	۷
۷	۱۲	۱۸	۰
۱۲	۵		

از فرمول قبلی حد بالای بازهٔ اطمینان ۹۵٪ برای X با توجه به $z_{۱-(\alpha/۲)} = z_{۰/۹۷۵} = ۱/۹۶$ به صورت زیر به دست می‌آید

$$۱۲۲/۲۲ + ۱/۹۶(۲۵)\sqrt{\frac{۲۵-۹}{۲۵}}\left(\frac{۳/۴۸}{\sqrt{۹}}\right) = ۱۲۲/۲۲ + ۴۵/۴۷ = ۱۶۷/۶۹$$

به همین ترتیب حد پایین بازهٔ اطمینان ۹۵٪ برای X عبارت است از

$$۱۲۲/۲۲ - ٤٥/٤٧ = ٧٦/٧٥$$

توجه کنید که مجموع واقعی جامعه X =۱۲۷ در این بازهٔ اطمینان قرار دارد.

این بازه‌های اطمینان ۹۵ درصدی دارای تفسیرهای معمولی زیرند: اگر قرار می‌شد به طور مکرر نمونه‌ای n عنصری از یک جامعه طبق یک طرح نمونه‌گیری انتخاب کنیم، و اگر برای هر نمونه، بازه‌های اطمینان محاسبه می‌شد، ۹۵ درصد این قبیل بازه‌های اطمینان شامل پارامتر مجهول جامعه می‌بودند.

□

تابلوی ٤.۳ واریانسهای برآورد شدهٔ توزیع نمونه‌گیری برای مجموع، میانگین و نسبت را خلاصه و شکل بازه‌های اطمینان متناظر را ارائه می‌کند.

باید تأکید کرد که بازه‌های اطمینان به دست آمده با استفاده از معادله‌های تابلوی ٤.۳ بر این فرض استوارند که برآورد (مثلاً x') به طور نرمال توزیع شده است. میزان تخطی از این فرض به ملاحظاتی از قبیل طبیعت توزیع متغیر در جامعه و اندازهٔ نمونه بستگی دارد. اگر متغیر دارای توزیع تقریباً متقارن بوده و اندازه‌های نمونه‌ای کوچک نباشد، آنگاه ضرایب اطمینان که در بازه‌های اطمینان بیان می‌شوند تقریباً صحیح خواهند بود. ولی اگر داده‌ها بسیار چوله باشند و اندازهٔ نمونه هم کوچک باشد، ضرایب اطمینان ممکن است گمراه کننده باشند (تمرین ۱.۳ وضعیت را با استفاده از داده‌های جدول ۱.۲ نشـــان می‌دهد). برای برآوردهای خطی از قبیل آنهایی که در این فصل مورد بحث قرار گرفتنـــــد، *قضیهٔ حدی مرکزی* آمار، تکیه‌گاهی نظری را برای فرض نرمال بودن ارائه می‌کند. این قضیه در واقع می‌گوید که اگر آماره‌ها از قبیل میانگینها، مجموعها و نسبتها بر اندازه‌های نمونه‌ای به قدر کافی بزرگ متـــکی باشند توزیعهای نمونه‌گیری آنها، با هر طبیعت توزیع اساسی مشاهدات اصلی، گرایش به نرمال بودن خواهند داشت.

بازه‌های اطمینان را می‌توان برای میانگینها و نسبتهای جامعه‌ای از روی برآوردهای نمونه‌ای مناسب به روشی مشابه روش بالا برای مجموعها به دست آورد.

تابلوی ۴.۳ واریانسهای برآورد شده و $(۱-\alpha)$٪ بازه‌های اطمینان تحت نمونه‌گیری تصادفی ساده

مجموع x'

$$\hat{Var}(x') = N^۲\left(\frac{N-n}{N}\right)\left(\frac{s_x^۲}{n}\right) \qquad (۹.۳)$$

$$x' \pm z_{۱-(\alpha/۲)}(N)\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$$

میانگین $\bar{x}$

$$\hat{Var}(\bar{x}) = \left(\frac{N-n}{N}\right)\left(\frac{s_x^۲}{n}\right) \qquad (۱۰.۳)$$

$$\bar{x} \pm z_{۱-(\alpha/۲)}\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$$

نسبت p_y

$$\hat{Var}(p_y) = \left(\frac{N-n}{N}\right)\left(\frac{p_y(۱-p_y)}{n-۱}\right) \qquad (۱۱.۳)$$

$$p_y \pm z_{۱-(\alpha/۲)}\sqrt{\frac{N-n}{N}}\sqrt{\frac{p_y(۱-p_y)}{n-۱}}$$

در این معادله‌ها n، تعداد عناصر موجود در نمونه، N، تعداد عناصر در جامعه، $s_x^۲$، واریانس نمونه و $z_{۱-(\alpha/۲)}$ مقدار $(۱-\alpha/۲)$۱۰۰ امین صدک توزیع نرمال استانداردند. توجه کنید که در فرمولهای مربوط به $\hat{Var}(x')$ و $\hat{Var}(\bar{x})$ می‌توانیم به جای $s_x^۲$ از رابطهٔ شامل $\hat{\sigma}_x^۲$ استفاده کنیم [ن. ک. معادلهٔ (۱۶.۲)]. همچنین توجه داشته باشید که چون $\hat{SE}(\hat{d}) = \sqrt{\hat{Var}(\hat{d})}$ بازهٔ اطمینان را می‌توان به صورت خطای معیار برآورد شده نوشت. یعنی می‌توانیم بازه‌های اطمینان را به صورت زیر بنویسیم .

$$x' \pm z_{۱-(\alpha/۲)}\left[\hat{SE}(x')\right] \qquad \bar{x} \pm z_{۱-(\alpha/۲)}\left[\hat{SE}(\bar{x})\right] \qquad p_y \pm z_{۱-(\alpha/۲)}\left[\hat{SE}(p_y)\right]$$

۶.۳ برآورد کردن پارامترها برای زیرحوزه‌ها

اهداف آمارگیریهای نمونه‌ای غالباً برآورد کردن پارامترها را نه تنها برای کل جامعه بلکه برای برخی *زیرحوزه‌ها* (زیرگروهها یا زیرمجموعه‌ها)ی جامعه نیز شامل می شود. برای مثال، یک آمارگیری از بهداشت خانوارها در سراسر کشور ممکن است به برآوردهایی برای کل کشور و در عین حال برای گروههایی نیاز داشته باشد که بر حسب سن، جنس، نژاد، ناحیهٔ جغرافیایی، یا ترکیبهایی از اینها تعریف

شده‌اند. غالباً زیرحوزه‌ها قبل از گرفتن نمونه تعیین می‌شوند و نمونه، جداگانه در داخل هر زیرحوزه گرفته می‌شود. این نوع برنامهٔ نمونه‌گیری در مبحث مربوط به نمونه‌گیری طبقه‌بندی شده در فصـــلهای ۵ و ۶ بررسی خواهد شد. ولی گاهی اوقات یک نمونهٔ تصادفی ساده از کل جامعه گرفته می‌شود و برآوردها برای هر زیرحوزهٔ موردنظر، جداگانه بیان می‌شوند. در این بخش به این نوع طرح نمونه‌گیری می‌پردازیم. این شیوه را با استفاده از یک مثال نشان می‌دهیم.

مثال تشریحی: جامعه‌ای متشکل از شش خانواده را در نظر می‌گیریم که در یک بلوک شهری زندگی می‌کنند و در جدول ۵.۳ نشان داده شده‌اند. با استفاده از این داده‌ها پی می‌بریم که متوسط (یا میانگین) هزینه‌های درمانی نقدی به ازای هر خانواده ۲۹۶/۶۷ دلار برای کل جامعه، ۳۱۶/۶۷ دلار برای خانواده‌های سفیدپوست و ۲۷۶/۶۷ دلار برای خانواده‌های سیاه‌پوست است.

جدول ۵.۳ نژاد و هزینه‌های درمانی نقدی برای شش خانواده (۱۹۸۷)

خانواده	نژاد	هزینه‌های درمانی نقدی (دلار)
۱	سفید	۵۰۰
۲	سیاه	۳۵۰
۳	سیاه	۴۳۰
۴	سفید	۲۸۰
۵	سفید	۱۷۰
۶	سیاه	۵۰

فرض کنیم که می‌خواهیم یک نمونهٔ تصادفی ساده متشکل از چهار خانوار از جامعهٔ شش خانواری بگیریم تا متوسط هزینه‌های درمانی را به ازای هر خانوار برای کل جامعه و برای هر نژاد، جداگانه برآورد کنیم. شیوهٔ به دست آوردن چنین برآوردی برای کل جامعه قبلاً شرح داده شد. بنابراین، در مثال فعلی فقط برآوردهای مربوط به زیرحوزه‌های موردنظر را بررسی می‌کنیم.

می‌توانیم برآوردهای مربوط به یک زیرحوزهٔ خاص (مثلاً جامعهٔ سیاه‌پوست) را به روش زیر تهیه کنیم. فرض کنید

$$Y_i = \begin{cases} ۱ & \text{اگر خانوادهٔ } i \text{ ام سیاه پوست باشد} \\ ۰ & \text{در غیر آن صورت} \end{cases}$$

$X_i =$ هزینه‌های نقدی خانواده i ام

$$Z_i = X_i Y_i$$

برای هر شش خانوادهٔ جامعهٔ موردنظر، مقدار X_i، Y_i و Z_i را داریم که در جدول ٦.٣ نشان داده شده‌اند.

اگر Y و Z را معرف مجموع جامعه، و y و z را طبق معمول مجموعهای نمونه‌ای فرض کنیم، می‌بینیم که $Z = ۸۳۰$ و $Y = ۳$ دلار و متوسط هزینه‌های درمانی نقدی به ازای هر خانواده از خانواده‌های سیاه‌پوست برابر است با

$$\frac{Z}{Y} = \frac{۸۳۰}{۳} = ۲۷۶/۶۷ \text{ دلار}.$$

اینک $\frac{Z}{Y}$ را از روی نمونه، به وسیلهٔ $\frac{z}{y}$ که از عناصر نمونه ساخته می‌شود برآورد می‌کنیم. توزیع $\frac{z}{y}$ برای نمونه‌هایی متشکل از $n = ۴$ عنصر در جدول ۷.۳ داده شده است.

اگر میانگین $E\left(\frac{z}{y}\right)$ توزیع نمونه‌گیری $\frac{z}{y}$، و خطای معیار $SE\left(\frac{z}{y}\right)$ را محاسبه کنیم، خواهیم داشت

$$SE\left(\frac{z}{y}\right) = ۹٦/۷۷ \text{ دلار} \qquad \text{و} \qquad E\left(\frac{z}{y}\right) = ۲۷٦/٦۷ \text{ دلار} = \frac{Z}{Y}$$

به این ترتیب می‌بینیم که میانگین نمونهٔ زیرحـــوزهٔ $\frac{z}{y}$ در این مورد یک برآورد نااریب از میانگین جامعهٔ $\frac{Z}{Y}$ است.

جدول٦.٣ داده‌های مربوط به زیرحوزه بر مبنای خانواده‌های فهرست شده در جدول ۵.۳

خانواده	نژاد	X_i	Y_i	Z_i
۱	سفید	۵۰۰	۰	۰
۲	سیاه	۳۵۰	۱	۳۵۰
۳	سیاه	٤۳۰	۱	٤۳۰
٤	سفید	۲۸۰	۰	۰
۵	سفید	۱۷۰	۰	۰
٦	سیاه	۵۰	۱	۵۰

جدول ۷.۳ توزیع نمونه‌گیری $\frac{z}{y}$

عناصر نمونه	z	y	$\frac{z}{y}$
۱,۲,۳,٤	۷۸۰	۲	۳۹۰
۱,۲,۳,٥	۷۸۰	۲	۳۹۰
۱,۲,۳,٦	۸۳۰	۳	۲۷٦/٦۷
۱,۲,٤,٥	۳٥۰	۱	۳٥۰
۱,۲,٤,٦	٤۰۰	۲	۲۰۰
۱,۲,٥,٦	٤۰۰	۲	۲۰۰
۱,۳,٤,٥	٤۳۰	۱	٤۳۰
۱,۳,٤,٦	٤۸۰	۲	۲٤۰
۱,۳,٥,٦	٤۸۰	۲	۲٤۰
۱,٤,٥,٦	٥۰	۱	٥۰
۲,۳,٤,٥	۷۸۰	۲	۳۹۰
۲,۳,٤,٦	۸۳۰	۳	۲۷٦/٦۷
۲,۳,٥,٦	۸۳۰	۳	۲۷٦/٦۷
۲,٤,٥,٦	٤۰۰	۲	۲۰۰
۳,٤,٥,٦	٤۸۰	۲	۲٤۰

□

نسبت $\frac{z}{y}$ یک حالت خاص از برآورد نسبتی است که در فصل ۷ به تفصیل شرح داده خواهد شد. به طور کلی، برآوردهای نسبتی، نااریب نیستند، گرچه بزرگی اریبی آنها غالباً اندک است. اما، به طوری که در مورد بالا شرح داده شد، وقتی مخرج کسر عبارت از تعداد واحدهای اولیه است و از روش نمونه‌گیری تصادفی ساده استفاده می‌شود این نسبت نااریب است. به عبارت دیگر، آنچه به صورت تجربی در بالا نشان داده‌ایم به طور کلی برای نمونه‌گیری تصادفی ساده درست است، از جمله این که میانگین نمونهٔ یک متغیر در میان اعضای زیرگروه، یک برآوردگر نااریب از میانگین جامعه برای آن زیرگروه است (مشروط بر این که اندازهٔ نمونه در زیرگروه صفر نباشد).

فرمول دقیق ساده‌ای برای *خطای معیار یک میانگین برآورد شدهٔ یک زیرگروه* وجود ندارد. ولی اگر $E(y)$ تعداد مورد انتظار عناصری از زیرحوزه که در نمونه واقع می‌شوند بزرگتر از ۲۰ یا برابر با آن باشد تقریب زیر معتبر است.

$$SE\left(\frac{z}{y}\right)=\left[\frac{\sigma_z}{\sqrt{E(y)}}\right]\times\sqrt{\frac{Y-E(y)}{Y-1}} \qquad (۱۲.۳)$$

در این فرمول، $\frac{z}{y}$ عبارت از میانگین سطح z در میان عناصر y در نمونه، متعلق به زیرحوزه است، و σ_z انحراف معیار توزیع متغیر Z در میان اعضای زیرگروه جامعه است که از فرمول زیر به دست می‌آید:

$$\sigma_z=\left[\frac{\sum_{i=1}^{Y}(Z_i-\bar{Z})^2}{Y}\right]^{1/2}$$

چون Y، $E(y)$ و σ_z معمولاً مجهول‌اند به طور کلی به وسیلهٔ y' (که برابر است با $y\left(\frac{N}{n}\right)$)، y و فرمول زیر برآورد می شوند

$$\hat{\sigma}_z=\left[\frac{y'-1}{y'}\right]^{1/2}\times\sqrt{\frac{\sum_{i=1}^{y}(z_i-\bar{z})^2}{(y-1)}}$$

که در آن z_i معرف مقدار متغیر Z برای i امین عنصر نمونه و $\bar{z}$، میانگین نمونهٔ z_i است. با جایگزین کردن برآوردهای y، y' و $\hat{\sigma}_z$ در معادلهٔ (۱۲.۳)، *برآورد خطای معیار* $\hat{SE}(\frac{z}{y})$ میانگین برآورد شده برای زیرگروه را خواهیم داشت:

$$\hat{SE}(\frac{z}{y})=\left(\frac{\hat{\sigma}_z}{\sqrt{y}}\right)\sqrt{\frac{y'-y}{y'-1}} \qquad (۱۳.۳)$$

توجه کنید که تقریبی که با معادلهٔ (۱۲.۳) داده شده است نمی‌تواند برای داده‌های جدول ۵.۳ به کار رود، زیرا متوسط تعداد آن عناصر زیرگروه که در نمونه قرار می‌گیرند برابر با ۲ است [یعنی $E(y)=\left(\frac{n}{N}\right)y$ در تمام نمونه‌های ممکن]، که به صورتی قابل ملاحظه کوچکتر از ۲۰ است.

حالا نشان می‌دهیم که چگونه با استفاده از معادله‌های پیشین، میانگین زیرگروه و خطای معیار آن را برآورد می‌کنیم.

مثال تشریحی: داده‌های جدول ۸.۳، حاصل از یک نمونهٔ تصادفی ٤٠ تایی را که از یک جامعهٔ با ۱۲۰۰ کارگر گرفته شده است، در نظر می‌گیریم.

اگر بخواهیم میانگین ظرفیت حیاتی تحت فشار ریه (fvc[1]) را برای کارگرانی برآورد کنیم که به شدت در معرض مواد مضر برای ریه‌ها هستند، محاسبات زیر را خواهیم داشت:

$$y = ۲۸ \qquad n = ۴۰ \qquad N = ۱۲۰۰$$

$$y' = \left(\frac{N}{n}\right)y = \left(\frac{۱۲۰۰}{۴۰}\right)(۲۸) = ۸۴۰$$

$$z = \sum_{i=۱}^{n} z_i = ۸۱ + ۶۴ + ۰۰۰ + ۸۴ = ۲۲۱۵$$

$$\frac{z}{y} = \frac{۲۲۱۵}{۲۸} = ۷۹/۱۱$$

$$\hat{\sigma}_z = \sqrt{\frac{y'-۱}{y'}}\left(\frac{\sum_{i=۱}^{y}(z_i - \bar{z})^۲}{y-۱}\right)^{۱/۲}$$

$$= \sqrt{\frac{۸۴۰-۱}{۸۴۰}}\sqrt{\frac{(۸۱-۷۹/۱۱)^۲ + ۰۰۰ + (۸۴-۷۹/۱۱)^۲}{۲۸-۱}} = ۱۲/۵۵$$

$$\hat{SE}\left(\frac{z}{y}\right) = \left(\frac{\hat{\sigma}_z}{\sqrt{y}}\right)\sqrt{\frac{y'-y}{y'-۱}} = \left(\frac{۱۲/۵۵}{\sqrt{۲۸}}\right)\sqrt{\frac{۸۴۰-۲۸}{۸۴۰-۱}} = ۲/۳۳$$

به این ترتیب از نمونهٔ مزبور چنین برآورد می‌شود که در کل ۸۴۰ کارگری که در معرض مواد مضر برای ریه قرار دارند، متوسط ظرفیت حیاتی تحت فشار ریه، fvc، برابر ۷۹/۱۱٪ است و بازهٔ اطمینان ۹۵٪ برای میانگین برآورد شده عبارت است از:

$$\frac{z}{y} - (۱/۹۶)\left[\hat{SE}\left(\frac{z}{y}\right)\right] \le \frac{Z}{Y} \le \frac{z}{y} + (۱/۹۶)\left[\hat{SE}\left(\frac{z}{y}\right)\right]$$

$$۷۹/۱۱ - (۱/۹۶)\,(۲/۳۳) \le \frac{Z}{Y} \le ۷۹/۱۱ + (۱/۹۶)\,(۲/۳۳)$$

$$۷۴/۵۴ \le \frac{Z}{Y} \le ۸۳/۶۹$$

احکـامی که برای برآورد ظرفیت حیاتی اجباری هر یک از گروههای رده‌بندی شده برحسب سطح در معـرض مـواد مضـر بودن به کار می‌رود از پروندهٔ داده‌های STATA در زیر نشان داده شده‌اند، که داده‌های کارگران، حاوی اطلاعاتی است که در جدول ۸.۳ نشان داده شده است:

[1] forced vital capacity

جدول ۸.۳ سطح قرار گرفتن در معرض مواد مضر برای ریه و ظرفیت حیاتی تحت فشار ریه برای نمونهٔ کارگرانی که از کارخانه‌ای با ۱۲۰۰ کارگر شاغل گرفته شده است

کارگر	در معرض مواد مضر*	(fvc) ظرفیت حیاتی تحت فشار*	اندازهٔ جامعه	وزن۱
۱	۳	۸۱	۱۲۰۰	۳۰
۲	۳	٦٤	۱۲۰۰	۳۰
۳	۲	۸۵	۱۲۰۰	۳۰
٤	۲	۹۱	۱۲۰۰	۳۰
۵	۳	٦۰	۱۲۰۰	۳۰
٦	۱	۹۷	۱۲۰۰	۳۰
۷	۱	۸۲	۱۲۰۰	۳۰
۸	۱	۹۹	۱۲۰۰	۳۰
۹	۳	۹٦	۱۲۰۰	۳۰
۱۰	۳	۹۱	۱۲۰۰	۳۰
۱۱	۱	۷۱	۱۲۰۰	۳۰
۱۲	۳	۸۸	۱۲۰۰	۳۰
۱۳	۲	۸٤	۱۲۰۰	۳۰
۱٤	۳	۸۵	۱۲۰۰	۳۰
۱۵	۳	۷۷	۱۲۰۰	۳۰
۱٦	۳	۷٦	۱۲۰۰	۳۰
۱۷	۳	٦۲	۱۲۰۰	۳۰
۱۸	۳	٦۷	۱۲۰۰	۳۰
۱۹	۳	۹۱	۱۲۰۰	۳۰
۲۰	۲	۹۹	۱۲۰۰	۳۰
۲۱	۲	۷۰	۱۲۰۰	۳۰
۲۲	۱	٦٤	۱۲۰۰	۳۰
۲۳	۳	۷۲	۱۲۰۰	۳۰
۲٤	۲	۷۲	۱۲۰۰	۳۰
۲۵	۳	۹۵	۱۲۰۰	۳۰
۲٦	۳	۹٦	۱۲۰۰	۳۰
۲۷	۳	٦۲	۱۲۰۰	۳۰
۲۸	۳	٦۷	۱۲۰۰	۳۰
۲۹	۳	۹۵	۱۲۰۰	۳۰
۳۰	۱	۸۷	۱۲۰۰	۳۰
۳۱	۳	۸٤	۱۲۰۰	۳۰
۳۲	۳	۸۹	۱۲۰۰	۳۰
۳۳	۳	۸۹	۱۲۰۰	۳۰
۳٤	۳	٦۵	۱۲۰۰	۳۰
۳۵	۳	٦۷	۱۲۰۰	۳۰
۳٦	۳	٦۹	۱۲۰۰	۳۰
۳۷	۳	۸۰	۱۲۰۰	۳۰
۳۸	۳	۹۸	۱۲۰۰	۳۰
۳۹	۳	٦۵	۱۲۰۰	۳۰
٤۰	۳	۸٤	۱۲۰۰	۳۰

* درصد مقدار مورد انتظار بر اساس سن، جنس، و قد

* ۱ = کم ۲= متوسط ۳ = زیاد

```
. use “a:\ workers.dta”, clear
. svyset fpc popsize
. svyset pweight wt1
. svymean fpc, by(exposure)
```

توجه کنید که STATA برای نمونهٔ تصادفی ساده تنها به دو پارامتر نیاز دارد که عبارت‌اند از کل تعداد واحدهای شمارش، N، (که برای هر گزارش مربوط به متغیر موسوم به *popsize* (اندازهٔ جامعه) وجود دارد)، و وزن نمونه‌گیری $\frac{N}{n}$ (که برای هر گزارش مربوط به متغیر موسوم به *Wt1* (وزن۱) موجود است). حکم svymean fpc, by(exposure)، نشان می‌دهد که برآورد سطح میانگین متغیرfvc ظرفیت حیاتی تحت فشار باید برای هر یک از سطوح در معرض بودن، جداگانه اجرا شود. تفاوت بین بازه‌های اطمینان که در زیر نشان داده شده و آنهایی که قبلاً ارائه شدند حاصل این واقعیت است که بازهٔ اطمینان را بر مبنای توزیع نرمال در نظر گرفته‌ایم (همان‌گونه که به طور سنتی در نمونه‌گیری از جامعهٔ متناهی اجرا می‌شود)، در حالی که در STATA بازه‌های اطمینان بر مبنای توزیع t استیودنت با $n-۱$ درجهٔ آزادی قرار دارند (در این مورد که درجهٔ آزادی ۳۹ است صدک ۹۷/۵ در مقایسه با ۱/۹۶ مربوط به توزیع نرمال، برابر با ۲/۰۲ است).

خروجی حاصل از STATA در زیر نشان داده شده است:

Survey mean estimation

Pweight:	wt1	Number of obs	= 40
Strata:	<one>	Number of strata	= 1
PSU:	<observations>	Number of PSUs	= 40
FPC:	popsize	Population size	= 1200

Mean Subpop	Estimate	Std.Err.	[95% Conf.	Interval]	Deff
fvc					
exposure = = 1	83.33333	5.177446	72.86096	93.80571	1
exposure = = 2	83.5	4.110476	75.18578	91.81422	1
exposure = = 3	79.10714	2.321226	74.41202	83.80227	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

۷.۳ نمونهٔ موردنیاز چقدر باید بزرگ باشد؟

یکی از مهمترین مسائل در طرح نمونه، تعیین این است که نمونه چقدر باید بزرگ باشد تا برآوردهای به دست آمده در آمارگیری نمونه‌ای به اندازهٔ کافی، برای تأمین اهداف آمارگیری قابل اعتماد باشند. در

این بخش، مشکل مزبور را به صورتی بسیار کلی فرمولبندی خواهیم کرد و سپس نشان خواهیم داد که این مشکل را برای طرح نمونهٔ خاص، مثلاً، برای نمونه‌گیری تصادفی ساده از واحدهای اولیه، چگونه می‌توان حل کرد.

اولین قدم در تعیین اندازهٔ نمونه عبارت است از مشخص کردن سطح قابلیت اعتماد موردنیاز برای برآوردهای حاصل (ن. ک. بخش ٤.٢). به طور کلی، هر چه نمونه بزرگتر باشد قابلیت اعتماد برآوردهای به دست آمده بیشتر خواهد بود. از سوی دیگر، اعتبار، به جای اینکه تابعی از اندازهٔ نمونه باشد تابعی از فرایند اندازه‌گیری است، و به طور کلی با افزایش اندازهٔ نمونه بهبود نخواهد یافت. بهبود اعتبار مستلزم بهبود فرایند اندازه‌گیری است.

برای تعیین سطح قابلیت اعتماد موردنیاز برای برآوردها باید آماردانان و کارشناسان موضوعی، اهداف آمارگیری را مطالعه کنند. برای مثال، فرض کنید از بیمارستانی که سالانه ۲۰۰۰۰ بیمار می‌پـــذیرد، قرار است برای هدف تعیین نسبتی از ۲۰۰۰۰ بیمار که طبق تعریف استانداردهای توصیف شده، مراقبت بهینه دریافت کرده‌اند بیمارانی اختیار شوند. کمیتهٔ بررسی کیفیت مراقبت که برنامه‌ریزی آمارگیری را به عهده دارد ممکن است احساس کند که اگر کمتر از ۸۰ درصد بیماران مراقبت بهینه دریافت کرده باشند لازم است برخی اقدامات اصلاحی به عمل آید. در این مورد، کمیته نگران بیش برآورد نسبت واقعی خواهد بود، ولی اگر نسبت برآورد شده ۸۰ درصد باشد در حالی که نسبت واقعی ۷۵ درصد است چندان نگرانی نخواهد داشت. آماردان این موضوع را به این ترتیب فرمولبندی خواهد کرد که بگوید کاربر مایل است «واقعاً مطمئن» شود که نسبت برآورد شده با نسبت واقعی بیش از $[۷۵ / (۸۰-۷۵)]۱۰۰$٪ یا ٦/٦٧٪ نسبت واقعی تفاوت ندارد.

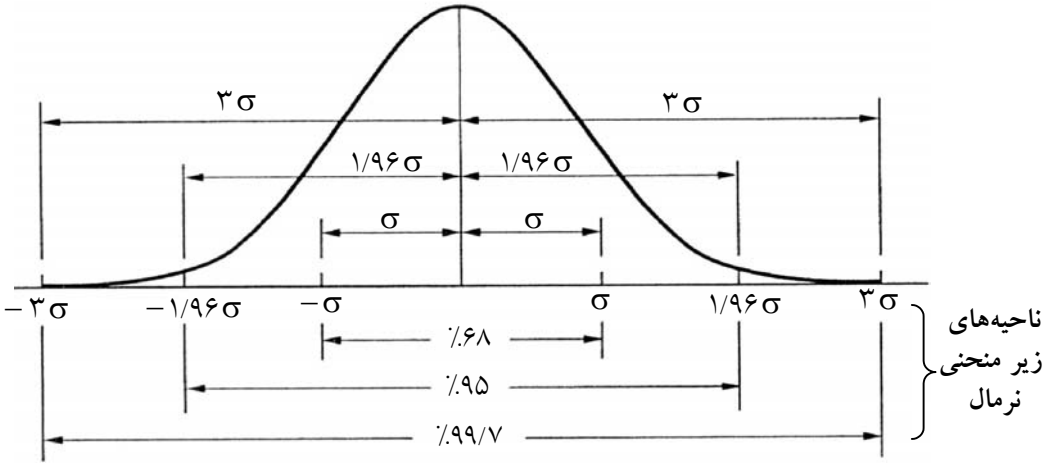
حال ببینیم منظور از «اطمینان واقعی» چیست؟ از لحاظ قابلیت اعتماد، اگر فرض کنیم برآوردهای نمونه‌ای با میانگینی برابر با برآورد جامعه‌ای نامعلوم به طور نرمال توزیع شده‌اند، می‌دانیم که پارامتر واقعی جامعه برای تقریباً ۹۹۷ نمونه از هر هزار نمونه داخل سه خطای معیار برآورد قرار خواهد داشت (ن. ک. شکل ۱.۳). در مورد مثال فوق‌الذکر، اگر p_y نسبت برآورد شده، P_y نسبت واقعی نامعلوم جامعه و $SE(p_y)$ خطای معیار p_y باشند می‌توانیم واقعاً مطمئن باشیم که P_y بزرگتر از $p_y - ۳ \times SE(p_y)$ و کوچکتر از $p_y + ۳ \times SE(p_y)$ است. پس مشکل آماردان این است که نمونه‌ای به اندازهٔ کافی بزرگ انتخاب کند به نحوی که هرگاه P_y تقریباً برابر با ۰/۸۰ باشد:

$$۳ \times SE(p_y) \leq ۰/۰۶۶۷ \times P_y$$

(ن. ک. مثال بعدی).

«اطمینان واقعی» به صورتی که در اینجا به کار رفت چیزی نیست مگر یکی از چند سطح بالقوهٔ اطمینان که می‌تواند در ساختن بازه‌های اطمینان و تعیین اندازه‌های نمونهٔ لازم به کار گرفته شود. این

سطح خاص اطمینان لزومی ندارد که همیشه به کار گرفته شود. ولی، «اطمینان واقعی» همراه با «۹۵٪ اطمینان» سطوحی هستند که با فراوانی زیاد در تعیین اندازه‌های نمونه‌ای برای آمارگیریهای نمونه‌ای مشخص می‌شوند.



شکل ۱.۳ ناحیهٔ زیر منحنی نرمال در داخل ± 1، $\pm 1/96$ ، ± 3 خطای معیار میانگین

مثال تشریحی: فرض کنید در طرح، از یک نمونهٔ تصادفی ساده، از سوابق بیمارستانی استفاده کنیم که پذیرش سالانهٔ آن بیمارستان ۲۰۰۰۰ بیمار است. با استفاده از تابلوی ۲.۳ می‌بینیم که $3 \times SE(p_y)$ چنین به دست می‌آید:

$$3 \times SE(p_y) = 3 \times \sqrt{\frac{P_y(1-P_y)}{n}} \sqrt{\frac{N-n}{N-1}}$$

می‌خواهیم نمونهٔ ما به اندازهٔ کافی بزرگ باشد (مبحث بالا را ببینید) به طوری که:

$$3 \times SE(p_y) \le 0/0667 P_y$$

یا

$$3 \times SE(p_y) = 3 \times \sqrt{\frac{P_y(1-P_y)}{n}} \sqrt{\frac{N-n}{N-1}} \le 0/0667\, P_y$$

با حل این رابطه برای n خواهیم داشت:

$$n \ge \frac{9 N P_y (1-P_y)}{(N-1)(0/0667)^2 P_y^2 + 9 P_y (1-P_y)}$$

با قرار دادن $P_y = 0/80$ و $N = 20000$، خواهیم داشت $n \ge 494$. (باید توجه داشت که در همهٔ محاسبات مربوط به اندازهٔ نمونه، مقدار به دست آمده برای n به نزدیکترین عدد صحیح گرد می‌شود). به این

ترتیب، برای این که واقعاً مطمئن شویم برآوردهایی که به دست می‌آیند قابلیت اعتماد لازم برای اهداف آمارگیری را دارند به نمونه‌ای با اندازهٔ ٤٩٤ بیمار یا بیشتر نیاز داریم.

□

مـثال بـالا را می‌توان به هر نسبت جامعه‌ای P_y تحت نمونه‌گیری تصادفی ساده تعمیم داد. اگر بخواهیم واقعـاً مطمـئن شـویم که برآورد نمونه‌ای p_y از نظر قدر مطلق با نسبت نامعلوم واقعی P_y بیش از εP_y تفاوت ندارد، در آن صورت باید *اندازهٔ نمونه*، n، در رابطهٔ زیر صدق کند:

$$n \geq \frac{۹NP_y(۱-P_y)}{(N-۱)\varepsilon^{۲}P_y^{۲}+۹P_y(۱-P_y)}$$

مقدار ε توسط محققان تعیین می‌شود تا اهداف آمارگیری را بازتاب دهد. اگر اندازهٔ جامعه، N، خیلی بیشتر از n، اندازهٔ نمونهٔ موردنیاز باشد، رابطهٔ فوق را می‌توان با فرمول زیر تقریب زد:

$$n \geq \frac{۹(۱-P_y)}{\varepsilon^{۲}P_y}$$

این رابطـــه نشان می‌دهد که اندازهٔ نمونه‌ای موردنیاز به سه عامل بستگی دارد: ۹، $\varepsilon^{۲}$، و $\frac{(۱-P_y)}{P_y}$.

عامل ۹، عامل اطمینان واقعی و عامل $\varepsilon^{۲}$ عامل نشان‌دهندهٔ مجموعهٔ ویژگیهایی است که بر حسب حداکثر تفاوت نسبی مجاز بین برآورد و نسبت واقعی نامعلوم جامعه‌ای P_y برای برآورد تعیین شده است. سومین عامل، $\frac{(۱-P_y)}{P_y}$، توان دوم ضریب تغییرات (یا واریانس نسبی) متغیر دو حالتـــی است (در این مورد، دریافت یا عدم دریافت مراقبت بهینه) که نسبت P_y بر آن متکی است. از فرمول تقریب برای n پی می‌بریم که هر چه ε را جدیتر (یعنی کوچکتر) بگیریم، اندازهٔ نمونه، n، بزرگتر خواهد شد.

اندازه‌های نمونه‌ای موردنیاز تحت نمونه‌گیری تصادفی ساده برای برآورد کردن میانگینها و مجموعهای جامعه‌ای را می‌توان با روشی مشابه و به صورتهایی شبیه آنچه در بالا برای نسبتها نشان داده شد به دست آورد. اندازه‌های نمونه‌ای موردنیاز برای مجموعها، میانگینها، و نسبتها در تابلوی ۵.۳ خلاصه شده‌اند. در این خلاصه، z عبارت است از *ضریب قابلیت اعتماد* که بر این فرض استوار است که توزیع نمونه‌گیری برآورد خـاص تحت بررسی، نرمـال است. یعنی برای اطمیـنان واقعـی $z = ۳$ است. اگر ۹۵٪ اطمینان مطلوب باشد z برابر ۱/۹٦ می‌شود.

با توجه به معادله‌های تابلوی ۵.۳ توجه می‌کنیم که پارامترهایی از قبیل V_x^2 واریانس نسبی در توزیع متغیر x یک جامعه، عموماً *نامعلوم*اند. از این رو آماردانان معمولاً باید برای محاسبهٔ اندازه‌های نمونه‌ای موردنیاز به حدسهای قریب به یقین متوسل شوند.

تابلوی ۵.۳ اندازه‌های دقیق و تقریبی نمونه‌ای موردنیاز، تحت نمونه‌گیری تصادفی ساده

	تقریبی	دقیق
مجموع، x'		
(۱۴.۳)	$n \geq \dfrac{z^2 V_x^2}{\varepsilon^2}$	$n \geq \dfrac{z^2 N V_x^2}{z^2 V_x^2 + (N-1)\varepsilon^2}$
میانگین، $\bar{x}$		
(۱۵.۳)	$n \geq \dfrac{z^2 V_x^2}{\varepsilon^2}$	$n \geq \dfrac{z^2 N V_x^2}{z^2 V_x^2 + (N-1)\varepsilon^2}$
نسبت، p_y		
(۱۶.۳)	$n \geq \dfrac{z^2 (1-P_y)}{\varepsilon^2 P_y}$	$n \geq \dfrac{z^2 N P_y (1-P_y)}{(N-1)\varepsilon^2 P_y^2 + z^2 P_y (1-P_y)}$

در این معادله‌ها z، ضریب قابلیت اطمینان (برای مثال، $z = 3$ برای اطمینان واقعی و $z = 1/96$ برای اطمینان در سطح ۹۵٪)، N، اندازهٔ جامعه، V_x^2، واریانس نسبی متغیر x ، ε، مقدار تعیین شده توسط محقق (یعنی این که برآورد نمونه‌ای $\hat{d}$ از نظر قدر مطلق نباید بیش از εd با d، پارامتر واقعی نامعلوم جامعه، تفاوت داشته باشد)، و P_y نسبت نامعلوم جامعه‌ای است.

مثال تشریحی: به عنوان مثالی دیگر از چگونگی تعیین اندازهٔ نمونه، فرض می‌کنیم قرار است در ایالتی که دارای ۲۵۰۰ داروخانه است یک آمارگیری نمونه‌ای از داروخانه‌ها اجرا شود. هدف از آمارگیری برآورد کردن متوسط قیمت خرده‌فروشی ۲۰ قرص رگ‌گشاست که به طور متداول مصرف می‌شود.

برآوردی موردنیاز است که در حدود ۱۰٪ با مقدار واقعی متوسط قیمت خرده‌فروشی در ایالت متفاوت باشد. فهرستی از همهٔ داروخانه‌ها موجود است و یک نمونهٔ تصادفی ساده از این فهرست گرفته خواهد شد. یک آمارگیری تلفنی از ۲۰ داروخانه، منتخب از N=۱۰۰۰ داروخانه در ایالتی دیگر، متوسط قیمت ۲۰ قرص را ۷/۰۰ دلار با انحراف معیار ۱/٤۰ دلار نشان داده است.

برای مطالعهٔ پیشنهادی می‌توانیم از اطلاعات حاصل از آمارگیری تلفنی به منظور برآورد اندازهٔ نمونه استفاده کنیم. V_x^2 را به صورت زیر برآورد می‌کنیم*:

$$V_x^2 = \frac{\sigma_x^2}{\bar{x}^2}\frac{\left[\frac{(N-1)}{N}\right]s_x^2}{\bar{X}^2} = \frac{\left[\frac{(999\ )}{1000}\right](1/40)^2}{(7/00)^2} = 0/04$$

با $\varepsilon = 0/1$ و $N = 2500$، از فرمول دقیق (۱۵.۳) به دست می‌آوریم:

$$n = \frac{9\ (2500)\ (0/04)}{9\ (0/04) + 2499\ (0/1\ )^2} = 35/5 \approx 36$$

به این ترتیب، برای مقاصد این بررسی، به نمونه‌ای متشکل از ۳٦ داروخانه نیاز است. توجه کنید که تقریب $\frac{9V_x^2}{\varepsilon^2}$ نیز همین مقدار را برای n به دست خواهد داد.

□

در کاربرد واقعی، به ندرت ممکن است یک متغیر به تنهایی به عنوان پایه‌ای برای محاسبهٔ اندازهٔ نمونه انتخاب شود. معمولاً چند تا از مهمترین متغیرها انتخاب می‌شوند و اندازه‌های نمونه‌ای برای هر یک از این متغیرها محاسبه می‌شود. اندازهٔ نهایی نمونه‌ای منتخب ممکن است بزرگترین اندازهٔ محاسبه شده برای نمونه باشد. اگر برای اجرای آمارگیری با استفاده از بزرگترین اندازهٔ نمونه، بودجهٔ کافی در اختیار نباشد، آنگاه، به عنوان یک اقدام بینابین، میانه یا میانگین nهای محاسبه شده انتخاب می شود.

---

* برآورد V_x^2 به شرح زیر به دست می‌آید: $V_x^2 = \frac{\sigma_x^2}{\bar{X}^2}$ از معادلهٔ (۷.۲). پس $\hat{V}_x^2 = \frac{\hat{\sigma}_x^2}{\hat{\bar{X}}^2}$. ولی از معادلهٔ (۱٦.۲) $\hat{\sigma}_x^2 = \left[\frac{(N-1)}{N}\right]s_x^2$ و برآورد $\bar{X}$ عبارت است از $\bar{x}$. پس نتیجه می‌گیریم که

$$V_x^2 = \frac{\sigma_x^2}{\hat{\bar{X}}^2} = \frac{\left[\frac{(N-1)}{N}\right]s_x^2}{\hat{\bar{x}}^2}\ .$$

همچنین تأکید می‌کنیم که اندازهٔ نمونهٔ انتخاب شده در تأمین ویژگی‌های قابلیت اعتماد یک مشخصهٔ برآورد شده در کل جامعه، برای برآورد آن مشخصه در یک زیرحوزهٔ جامعه، به قدر کافی بزرگ نخواهد بود. برای مثال، ممکن است اندازهٔ نمونهٔ ۵۰۰ نفر برای حصول اطمینان واقعی از اینکه برآورد میزان شیوع فشار خون بالا در کل جامعه با میزان شیوع واقعی ۱۰٪ متفاوت باشد کفایت کند. ولی برآورد میزان شیوع فشار خون در مردان بر مبنای یافته‌های این نمونه احتمال ندارد که به همان اندازه قابل اعتماد باشد، زیرا این برآورد بر مبنای اندازهٔ نمونه‌ای است که به مراتب کوچکتر از ۵۰۰ است. بنابراین، در محاسبهٔ اندازه‌های نمونه‌ای باید زیرگروهها یا زیرحوزه‌های ویژه‌ای که برآوردهایی برای آنها موردنیاز است تعیین شوند و سطوح قابلیت اعتماد مطلوب برای برآوردها در هر زیرحوزه هم مشخص شوند.

۸.۳ چرا از نمونه‌گیری تصادفی ساده به ندرت استفاده می‌شود؟

نمونه‌گیری تصادفی ساده به عنوان پایهٔ بسط نظریهٔ نمونه‌گیری، بسیار حایز اهمیت است. تحت نمونه‌گیری تصادفی ساده، هر نمونهٔ خاص n عنصری از یک جامعهٔ N عنصری را می‌توان انتخاب کرد و به علاوه هر نمونه به اندازهٔ هر نمونهٔ دیگر احتمال انتخاب شدن دارد. با این تعبیر، نمونه‌گیری مزبور از نظر درک، ساده‌ترین روش ممکن است و از این رو تنها روشی است که تمام روشهای دیگر با آن مقایسه می‌شوند.

هر چند نمونه‌گیری تصادفی ساده از نظر درک ساده است، ولی می‌تواند پرهزینه و غالباً در عمل ناشدنی باشد زیرا مستلزم آن است که همهٔ عناصر، پیش از نمونه‌گیری شناسایی و شماره‌گذاری شوند. اغلب اوقات، این شناسایی پیشین امکان‌پذیر نیست و درنتیجه نمی‌توان نمونهٔ تصادفی ساده از عناصر انتخاب کرد.

همچنین، چون برنامهٔ نمونه‌گیری تصادفی ساده به هر نمونهٔ ممکن n واحدی شانس مساوی برای انتخاب شدن می‌دهد، ممکن است به انتخاب نمونه‌هایی منتهی شود که در ناحیهٔ جغرافیایی وسیعی پراکنده‌اند. این‌گونه توزیع جغرافیایی نمونه برای اجرا، در شرایطی که مصاحبه با خانوار موردنظر است، بسیار پرهزینه خواهد بود.

این واقعیت که هر عنصر، شانس برابر برای انتخاب شدن در نمونه دارد احتمال دارد به انتخاب نمونه‌هایی منجر شود که در زیرحوزه‌ها نمایندگانی متناسب با توزیع آنها در جامعه داشته باشند. در حالی که این مورد، برای بعضی از انواع آمارگیریها ممکن است مناسب باشد، برای آن دسته از آمارگیریهایی که در آنها توجه بر زیرگروه‌هایی متمرکز است که از نسبت کوچکی از جامعه تشکیل

شده‌اند خوب نیست. برای مثال، یک نمونهٔ تصادفی از خانوارهای مقیم شیکاگو احتمال ندارد که طرحی کارا برای برآورد کردن رفتار بهداشتی ساکنان کره‌ای تبار شیکاگو باشد.

به عبارت دیگر، حتی اگر انتخاب نمونهٔ تصادفی ساده از عناصر امکان‌پذیر باشد، روشهای نمونه‌گیری دیگری هستند که ممکن است برای رسیدن به اهداف یک آمارگیری خاص مناسبتر باشند. در مابقی این کتاب به بحث در مورد روشهای دیگر می‌پردازیم.

۹.۳ خلاصه

در این فصل، مفهوم نمونهٔ تصادفی ساده را ارائه کردیم و اهمیت آن را به عنوان بنیان نظریهٔ نمونه‌گیری متذکر شدیم. فرمولهای کلی برای میانگین، واریانس، خطای معیار، ضریب تغییرات، و واریانس نسبی برآورد مجموعها، میانگینها و نسبتها را تحت نمونه‌گیری تصادفی ساده معرفی کردیم. استفاده از بازه‌های اطمینان را برای مقاصد ارزیابی قابلیت اعتماد برآوردها در طرحهای نمونه‌ای به طور اعم مورد بحث قرار دادیم، و استفاده از آنها را در نمونه‌گیری تصادفی ساده به طور اخص نشان دادیم. چگونگی برآورد کردن مشخصه‌ها را برای زیرحوزه‌ها وقتی که نمونهٔ تصادفی ساده از کل جامعه گرفته شده باشد نشان دادیم. روش شناسی محاسبهٔ اندازهٔ نمونهٔ موردنیاز را تحت نمونه‌گیری تصادفی ساده برای برآورد مشخصه‌های جامعه با قابلیت اعتماد تعیین شده ارائه کردیم. بالاخره، دربارهٔ معایب نمونه‌گیری تصادفی ساده بحث کردیم که هرگاه برآورد مشخصه‌های زیرحوزه‌هایی موردنیاز است که از نسبت تقریباً کوچکی از کل جامعه تشکیل شده‌اند این نوع نمونه‌گیری امکان‌پذیر، مقرون به صرفه یا مناسب نیست.

تمرین

۱.۳ از روی داده‌های جدول ۱.۲ با استفاده از نمونه‌گیری تصادفی ساده، ده نمونهٔ مختلف متشکل از شش پزشک انتخاب کنید. برای هر نمونه بازهٔ اطمینان تقریبی ۹۵ درصد را برای تعداد متوسط دفعات ویزیت هر پزشک از خانوارها محاسبه کنید. میانگین واقعی جامعه $\overline{X}$ برای چند تا از این بازه‌های اطمینان محاسبه شدهٔ ۹۵ درصدی در داخل کرانهای بازه‌ها واقع شده است؟ اگر این عدد خیلی بیشتر یا کمتر از حد انتظار است چگونه آن را توجیه می کنید؟ (از اولین عدد تصادفی در گوشهٔ سمت چپ بالای جدول پ. ۱ در پیوست شروع کنید و ستونها را از بالا به پایین بخوانید ولی از سطر ۳۵ جلوتر نروید).

۲.۳ از یک جامعهٔ ٦۵ عنصری چند نمونهٔ تصادفی سادهٔ ۱۵ عنصری می‌توان گرفت؟

۳.۳ از روی داده‌های جدول ۸.۳ نسبت همهٔ کارگرانی از کارخانه را که دارای ظرفیت حیاتی تحت فشار fvc ی کمتر از ۷۰٪ مورد انتظارند براساس سن، جنس، و قد برآورد کنید. برای این نسبت برآورد شده بازهٔ اطمینان ۹۵ درصدی را ارائه دهید.

٤.۳ از روی داده‌های جدول ۸.۳ دربارهٔ کارگرانی که به میزان کم یا متوسط در معرض مواد مضر برای ریه هستند، نسبت کسانی را که ظرفیت حیاتی تحت فشار fvcی آنان کمتر از ۹۰٪ مورد انتظار است بر مبنای سن، جنس، و قد برآورد کنید. برای این نسبت بازهٔ اطمینان ۹۵ درصد به دست آورید.

۵.۳ از روی داده‌های جدول ۸.۳ نسبت کارگرانی از کارخانه را که در سطح کم یا متوسط در معرض مواد مضر برای ریه هستند برآورد کنید. برای نسبتهای برآورد شده، بازهٔ اطمینان ۹۵ درصدی به دست آورید.

٦.۳ در کارخانه‌ای بزرگ که محصولات آن مشابه محصولات کارخانه‌ای است که داده‌های جدول ۸.۳ از آن گرفته شده قرار است دربارهٔ کارگران آمارگیری به عمل آید. اهداف آمارگیری عبارت‌اند از برآورد (الف) نسبت همهٔ کارگرانی که دارای ظرفیت حیاتی تحت فشار، fvc، کمتر از ۷۰٪ هستند و (ب) میانگین ظرفیت حیاتی تحت فشار، fvc، همهٔ کارگران. برآوردها در فاصلهٔ ۵٪ از مقدار واقعی پارامتر برآورد شده، موردنیازند. اندازهٔ نمونهٔ کارگران چقدر باید باشد؟ این کارخانه دارای ۵۰۰۰ کارگر است.

۷.۳ محله‌ای از یک شهر دارای ۳۰۰۰ خانوار و ۱۰۰۰۰ نفر جمعیت است. برای برنامه‌ریزی یک شعبه از بخش بهداشت محلی، مایل‌اند کل تعداد دفعات مراجعهٔ اعضای جامعه به پزشک را در طول یک سال تقویمی برآورد کنند. برای این که اطلاعات مزبور سودمند باشد باید کمتر از ۱۰٪ مقدار واقعی با تعداد واقعی تفاوت داشته باشد. نمونه‌گیری مقدماتی کوچکی که از ۱۰ خانوار برای جمع‌آوری اطلاعات اولیه صورت گرفته است، داده‌های زیر را دربارهٔ مراجعه به پزشک طی سال گذشته نتیجه داده است. با استفاده از این داده‌ها به عنوان اطلاعات اولیه، اندازهٔ نمونهٔ موردنیاز برای تأمین ویژگیهای آمارگیری را تعیین کنید.

خانوار	تعداد اعضای خانوار	تعداد دفعات مراجعه به پزشک به ازای هر نفر در طول سال گذشته
۱	۳	٤/–
۲	٦	٤/٥
۳	۲	۸/–
٤	٥	۳/٤
٥	۲	۰/٥
٦	۳	۷/–
۷	٤	۸/٥
۸	۲	٦/–
۹	٦	٤/–
۱۰	٤	۷/٥

۸.۳ بخشی از یک جدول اعداد تصادفی ساده به صورت زیر است:

۷۷	۳۷	۹۷	۰٦
۸۱	۳۹	۰۰	۰۸
۰۱	٥۸	۰۸	۱٤
۱۹	۲٤	۱۷	۲۲
۷۹	۱۲	۷۳	۷٥
٥۳	۳۲	٥۹	٦۹
٤٤	٤۸	۰۳	٥٤

الف. از اولین عدد تصادفی در گوشهٔ سمت چپ بالای این جدول شروع کنید و اعداد را به سمت پایین بخوانید و نمونه‌ای ٦ تایی از ۲۵ پزشک فهرست شده در جدول ۱.۲ انتخاب کنید.

ب. $\overline{X}$، میانگین تعداد ویزیت خانوار توسط پزشک را در جامعه برآورد کنید و یک بازهٔ اطمینان ۹۰ درصدی برای $\overline{X}$ بسازید.

پ. X مجموع تعداد ویزیت خانوار توسط پزشک را در جامعه برآورد کنید و یک بازهٔ اطمینان ۹۵ درصدی برای X بسازید. این بازهٔ اطمینان چگونه با بازهٔ اطمینان مثال تشریحی بخش ۵.۳ مقایسه می‌شود؟

ت. نسبت پزشکان جامعه را که دو یا چند ویزیت در سال داشته‌اند با اطمینان واقعی برآورد کنید.

۹.۳ بخشی در ناحیهٔ خلیج سان فرانسیسکو شامل تقریباً ۵۰۰۰۰ نفر است، که تقریباً ٤٠ درصد آنها قفقازی، ۲۵ درصد آمریکایی افریقایی تبار، ۲۰ درصد ایبریایی و ۱۵ درصد آسیایی هستند. مطلوب است برآورد نسبت اشخاصی که در این بخش تحت پوشش نوعی بیمهٔ بهداشتی نیستند. لازم است ۹۵٪ اطمینان داشته باشیم که این برآورد در پیرامون ۱۵٪ از نسبت واقعی است که تصور می‌شود مقداری بین۱۰٪ و۲۰٪ از کل جمعیت باشد. به فرض استفاده از نمونه‌گیری تصادفی ساده، اندازهٔ نمونه چقدر باید باشد؟

۱۰.۳ در مثال قبلی، اگر قرار باشد یک نمونهٔ تصادفی ساده از کل جمعیت گرفته شود، بزرگی نمونه چقدر باید باشد تا ۹۵٪ مطمئن باشیم که برآورد نسبت جمعیت آسیایی که تحت پوشش نوعی از بیمه قرار ندارد در محدودهٔ ۱۰ درصدی از مقدار واقعی است (باز به فرض این که مقدار واقعی مقداری بین ۱۰٪ و ۲۰٪ باشد).

۱۱.۳ شهری دارای ۲۰ کلینیک بهداشتی است و مایل‌اند نمونه‌ای شامل ۵ کلینیک برای برآورد کردن تعداد کل اشخاصی بگیرند که از همهٔ این کلینیکها در طول ۱۲ ماه گذشته نسخه‌هایی برای یک داروی ضد افسردگی تازه تأیید شده گرفته‌اند. اگر فرض کنیم میانگین تعداد بیمارانی که در این کلینیکها معاینه می‌شوند ۱۵۰۰ نفر در سال بوده و انحراف معیار این توزیع در میان کلینیکها ۳۰۰ باشد و تقریباً به ۵ درصد از همهٔ بیماران بدون در نظر گرفتن کلینیک مورد مراجعه داروی مزبور تجویز شده باشد، آیا نمونهٔ تصادفی ساده متشکل از ۵ کلینیک احتمال دارد که برآوردی به دست دهد که پیرامون ۲۰ درصد از مقدار واقعی باشد؟

۱۲.۳ شرکتی با ۷۰۰ کارمند در نظر دارد کارمندان خود را به خاطر کار کردن تحت تأثیر داروهای غیرمجاز (برای مثال، حشیش، کوکایین) مورد آزمایش قرار دهد. در طول سال، سه روز به طور نمونه برای آزمایش انتخاب می شود و در هر یک از این روز ها نمونهٔ تصادفی ساده‌ای متشکل از ۵۰ کارمند برای آزمایش انتخاب خواهد شد. هیچیک از کارکنان در نوبت دوم یا سوم، حتی اگر در نوبتهای قبلی مورد آزمایش قرار گرفته باشند، از آزمایش حذف نخواهند شد. اگر کارمندی همیشه کوکایین مصرف کند احتمال این که در این برنامهٔ آزمایش شناسایی شود چقدر

است (فرض کنید هر کسی که یکی از مواد مخدر را مصرف کرده و برای نمونه انتخاب شده باشد نتیجهٔ آزمایش او برای آن دارو مثبت خواهد بود)؟

۱۳.۳ در تمرین قبلی، اگر کارمند فقط ۱۰ درصد اوقات ضمن کار ماده مخدر «مصرف کرده باشد» احتمال شناسایی او چقدر است؟

کتابشناسی

The "classic" sampling texts listed below all give very complete discussions of simple random sampling.

1. Cochran W. G., *Sampling Techniques,* 3rd ed., Wiley, New York, 1977.
2. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods and Theory,* Vols. 1 and 2, Wiley, New York, 1953.
3. Kish, L., *Survey Sampling,* Wiley, New York, 1965.

The other sampling texts referenced in the previous chapters all develop the concepts of simple random sampling, each in its own way. The following articles in the Encyclopedia of Biostatistics *also treat various aspects of simple random sampling.*

4. Xia, Z., Sampling with and without replacement. In *The Encyclopedia of Biostatistics,* P. A. Armitage and T. Colton, Eds., Wiley, Chichester, U.K., 1998.
5. Levy, P. S., Simple random sampling. In *The Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.
6. Levy, P. S., Design effect. In *The Encyclopedia of Biostatistics,* Armitage, P. A., and Colton. T., Eds., Wiley, Chichester, U.K., 1998.

The software package, SAMPLE, originally developed as a module for Version 5.0 of the general-purpose statistical package, SYSTAT, is a very *useful program for selecting samples from a wide variety of commonly used designs, including simple random sample. Later versions of SYSTA T, however, no longer support this module.*

7. Frankel, M. R., and Spencer, B.D., *SAMPLE: A Supplementary Module for SYSTAT,* SYSTAT Inc., Evanston, I11., 1990.

Within the SAS system, algorithms are proposed for taking simple random samples with and without replacement. These are discussed in Chapter 12 of the following reference.

8. SAS Institute Inc., *SAS® language and Procedures: Usage 2, Version 6,* 1st ed., *SAS* Institute Inc., Cary, N.C., 1991.

In our discussion of software for obtaining estimates and their standard errors from sample survey data, we *mentioned the software packages*: *STATA*, *SUDAAN, WESVAR, and PC CARP. References to these are listed below..*

9. STATA Corporation, *STATA Technical Bulletin STB-31,* STATA Corporation, College Station, Tex., 1996.
10. Shah, B. V., Barnwell, B. G., and Bieler. G.S., *SUDAAN User's Manual, Version 6.4,* 2nd Ed., Research Triangle Institute, Research Triangle Park, N.C., 1996.
11. Brick, J. M., Broene, P., James, P., and Severynse, J., A *User's Guide to WesVarPC,* Westat, Inc., Rockville, Md., 1996.
12. Fuller, W. A., Kennedy. W., Schell, D., Sullivan, G., and Park, H. J., *PC CARP,* Statistical Laboratory, Iowa State University, Ames, Iowa, 1989.

EPI INFO, a PC software package in the public domain, was w*ritten for applied epidemiologists and has word processing, database, and statistical analysis modules. It has a module, CSAMPLE, that has the capability of producing estimates and their standard errors from survey data having a variety of sample designs. At the time of this writing, it is in Version 6.*

13. Dean, A. G., Dean, J. A., Coulombier, D., Brendel, K. A., Smith, D. C., Burton, A. H., Dicker, R. C., Sullivan, K., Fagan. R, F., and Arner, T G., *Epi Info, Version 6: A Word Processing, Database, and Statistics Program for Epidemiology on Microcomputers,* Centers for Disease Control and Prevention, Atlanta, Ga., 1994.

The article listed below reviews several software packages that can analyze data from sample surveys, including the programs listed above.

14. Carlson, B. L., Software for the statistical analysis of sample survey data. In *The Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. Eds., Wiley, Chichester, U.K., 1998.

The books listed below treat sample-size requirements for a wide variety of estimation and hypothesis testing scenarios, including finite-population sampling.

15. Lemeshow, S., Hosmer, D. W., Klar, J., and Lwanga, S. K., *Adequacy of Sample Size in Health Studies,* Wiley, Chichester, U.K., 1990.
16. Lwanga. S. K., and Lemeshow, S., *Sample Size Determination in Health Studies,* World Health Organization, Geneva, Switzerland, 1991.

The software program listed below includes algorithms for estimation of sample size for simple random sampling from a finite population.

1 7. Elashoff, J. D., *nQuery Advisor^{R} Version 2.0 User's Guide,* Statistical Solutions Ltd., Cork, Ireland, l997.

فصل ۴

نمونه‌گیری سیستماتیک

در فصل قبل، مفهوم نمونه‌گیری تصادفی ساده از عناصر را ارائه کردیم و اهمیت آن به عنوان ساده‌ترین نوع نمونه‌گیری از لحاظ «درک» را مورد بحث قرار دادیم، زیرا که در آن هر ترکیب ممکن متشکل از n عنصر از یک جامعهٔ N عنصری، شانس برابر برای انتخاب شدن دارد. از بعضی مشکلات همراه با نمونه‌گیری تصادفی ساده نیز بحث کردیم، از جمله مشکلاتی را ذکر کردیم که ممکن است مانع از انتخاب نمونه به روش مزبور شوند. دراین فصل به بحث در مورد نوعی نمونه‌گیری موسوم به نمونه‌گیری سیستماتیک می‌پردازیم. نمونه‌گیری سیستماتیک در عمل استفاده گسترده‌ای دارد زیرا کاربرد آن آسان است و می‌تواند به آسانی به افرادی که آشنایی چندانی با روش‌شناسی نمونه‌گیری ندارند آموزش داده شود. در واقع، نمونه‌گیری سیستماتیک چه به تنهایی و چه در ترکیب با برخی روشهای دیگر می‌تواند متداولترین روش نمونه‌گیری باشد.

۱.٤ چگونگی انتخاب نمونهٔ سیستماتیک

شاید بهترین راه برای توصیف شیوهٔ نمونه‌گیری سیستماتیک از طریق مثال باشد. مثال زیر به عنوان مقدمه‌ای بر این روش مهم نمونه‌گیری به کار می‌رود.

مثال تشریحی: فرض می‌کنیم به عنوان بخشی از یک برنامهٔ بررسی کیفیت مراقبتها و جلوگیری از هزینه‌ها، نمونه‌ای از سوابق درمانی بیماران بستری به طور جاری برای یک حسابرسی تفصیلی انتخاب

شود. احتمالاً تعداد کل پرونده‌های موجود در جامعه پیش از نمونه‌گیری مشخص نیست زیرا قرار است سوابق به صورتی که به پیش می‌رویم نمونه‌گیری شوند و بنابراین امکان استفاده از نمونه‌گیری تصادفی ساده برای انتخاب سوابق وجود ندارد. ولی این امکان وجود دارد که تعداد تقریبی گزارشهایی که در دوره‌ای زمانی برای انتخاب فراهم خواهد شد حدس زده شود و همچنان که گزارشها فراهم می‌شوند یک سابقه از هر k گزارش انتخاب شود. k عبارت است از یک عدد صحیح با مقداری خاص که در رابطه با تأمین نیازهای مطالعه انتخاب می‌شود.

مثلاً فرض کنید پیش‌بینی می‌شود که روزی ده پروندهٔ ترخیصی جدید آماده شود و نمونهٔ کل مطلوب، متشکل از ۳۰۰ پرونده در سال است. پس برآورد می‌شود که کل سوابقی که در سال در دسترس قرار می‌گیرد ۳۶۵۰=۳۶۵×۱۰ خواهد بود. برای به دست آوردن چیزی حول و حوش ۳۰۰ پرونده در سال برای نمونه، باید k بزرگترین عدد صحیح در خارج قسمت کسر $\frac{۳۶۵۰}{۳۰۰}$ باشد. چون خارج قسمت کسر ۱۲/۱۷ می‌شود k برابر با ۱۲ خواهد بود. این مقدا k را *بازهٔ نمونه‌گیری* می‌نامند. به این ترتیب باید از هر ۱۲ گزارش یک نمونه انتخاب شود.

یکی از راههای اجرای این روش آن است که هر پرونده به محض این که تشکیل شد با یک شمارهٔ مسلسل که از یک شروع می شود مشخص شود. (ابزار مهرزنی موجود این کار را به راحتی انجام می‌دهند). در شروع مطالعه، یک عدد تصادفی بین ۱ تا ۱۲ به عنوان نقطهٔ شروع انتخاب می‌شود، سپس، همان گزارش و گزارشهای پس از آن به فاصلهٔ ۱۲ تا ۱۲ تا انتخاب خواهند شد. برای مثال، اگر عدد تصادفی ٤ باشد، پرونده‌هایی که برای نمونه انتخاب می‌شوند ٤، ١٦، ۲۸، ٤۰، ۵۲ و الخ خواهند بود.

□

برای تعمیم موضوع، نمونهٔ سیستماتیک به این ترتیب گرفته می‌شود که ابتدا بازهٔ kی مطلوب نمونه‌گیری را تعیین کنیم، یک عدد تصادفی j را بین ۱ تا k انتخاب کنیم، سپس عناصر دارای شماره‌های j، $j+k$، $j+۲k$، $j+۳k$ و ... را انتخاب کنیم. توجه کنید که کسر نمونه‌گیری برای این‌گونه آمارگیری $\frac{۱}{k}$ است.

اگر عناصری که قرار است نمونه‌گیری شوند پیشاپیش به ترتیب شماره‌گذاری شوند و اگر انتخاب نمونه عملاً توسط اشخاص غیر ماهر انجام شود، اندکی تعدیل در این روش بر اساس دستگاه دهدهی به خصوص سودمند است. اگر نمونه‌ای از یک واحد در هر k واحد تعیین شده باشد (که در آن k مثلاً عدد دورقمی است)، می‌توانیم یک عدد تصادفی دورقمی را بین ۰۱ و k انتخاب کنیم. اگر شمارهٔ منتخب j باشد، آن‌گاه شماره‌های دورقمی j، $j+k$، $j+۲k$ و الی آخر، انتخاب می‌شوند تا به یک

عدد سه رقمی برسیم. پس از آن، همهٔ عناصری که به آن اعداد منتخب دورقمی ختم می‌شوند در نمونه گنجانیده خواهند شد. مثلاً، اگر $k=$ ۱۲، یک عدد تصادفی دورقمی بین ۰۱ تا ۱۲ (مثلاً ۰۷) انتخاب می‌شود. سپس شماره‌های نمونهٔ دورقمی تعیین می‌شوند (مانند ۰۷، ۱۹، ۳۱، ٤۳، ۵۵، ٦۷، ۷۹ و ۹۱)، و همهٔ گزارشهایی که به این دورقمی‌ها ختم می‌شوند در نمونه قرار می‌گیرند (مانند ۰۷، ۱۹، ۳۱، ٤۳، ۵۵، ٦۷، ۷۹، ۹۱، ۱۰۷، ۱۱۹، ۱۳۱، ۱٤۳، ۱۵۵، ۱٦۷، ۱۷۹، ۱۹۱، ۲۰۷، ۲۱۹، ۲۳۱ و غیره). استفاده از این شیوه گاهی اوقات، بخصوص اگر افرادی که نمونه را انتخاب می‌کنند غیر ماهر باشند، از شیوهٔ اول آسانتر است، زیرا به آنها آموزش داده می‌شود که فقط پرونده‌هایی را انتخاب کنند که به رقمهای تعیین شده ختم می‌شوند. ولی، بحث ما در مورد خواص برآوردهای حاصل از نمونه‌های تصادفی سیستماتیک مبتنی بر روش اول خواهد بود.

نمونه‌های سیستماتیک را می‌توان به راحتی از پرونده‌های موجود در رایانه به دست آورد. مدول *SAMPLE* در نرم‌افزار *SYSTAT* می‌تواند نمونهٔ سیستماتیک را یا با یک تعداد ثابت n و یا با یک نسبت ثابت r از پرونده‌ها با استفاده از شیوهٔ نمونه‌گیریی به دست آورد که تعدیلی از روش توصیف شدهٔ بالاست.

۲.٤ برآورد کردن مشخصه‌های جامعه

اگر نمونه‌گیری سیستماتیک یکی از هر k عنصر، نمونه‌ای متشکل از n عنصر را نتیجه دهد، در آن صورت *مجموع نمونه*، x'، *میانگین نمونه* $\bar{x}$ *و نسبت نمونه*، p، از روی نمونه به همان صورتی که در نمونه‌گیری تصادفی ساده گفته شد محاسبه می‌شوند (ن. ک. تابلوی ۱.۳). *برآورد مجموع جامعه*، x'، برای نمونه‌گیری سیستماتیک از فرمول زیر به دست می‌آید:

$$x' = \left(\frac{N}{n}\right)x \qquad (۱.٤)$$

و وقتی که $k=\dfrac{N}{n}$ یک عدد صحیح است، داریم:

$$x' = kx$$

لازم نیست که N پیش از نمونه‌گیری مشخص باشد (همان طور که زمان ورود هر k امین شخص به اتاق فوریتهای پزشکی چنین است)، N را می‌توان با شمارش تعداد عناصر باقیمانده پس از انتخاب آخرین عنصر و افزودن این باقیمانده به nk تعیین کرد.

بهتر است این ایده را با مثال نشان دهیم.

مثال تشریحی: می‌خواهیم یک نفر از هر شش پزشک را به صورت نمونهٔ تصادفی سیستماتیک از فهرست ارائه شده در جدول ۱.۲ انتخاب کنیم. برای این منظور ابتدا یک عدد تصادفی از ۱ تا ٦

انتخاب می‌کنیم: مثلاً ۵ را انتخاب می‌کنیم. پس پزشکانی که برای نمونه انتخاب می‌شوند شماره‌های ۵، ۱۱، ۱۷ و ۲۳ هستند. مقادیر متناظر x تعداد ویزیت پزشک از خانوارها در جدول ۱.٤ نشان داده شده است.

برآورد میانگین تعداد $\bar{x}$ ویزیت از خانوارها به ازای هر پزشک عبارت است از:

$$\bar{x} = \frac{x}{n} = \frac{۱۴}{۴} = ۳/۵$$

برآورد مجموع تعداد، x' ویزیت انجام شده توسط همهٔ پزشکان در جامعه، عبارت است از:

$$x' = \left(\frac{N}{n}\right)x = \left(\frac{۲۵}{۱۴}\right)۱۴ = ۸۷/۵$$

برآورد نسبت P_y پزشکانی که یک ویزیت یا بیشتر از خانوار به عمل آورده‌اند عبارت است از:

$$P_y = \frac{y}{n} = \frac{۲}{۴} = ۰/۵۰$$

جدول ۱.٤ نمونهٔ سیستماتیک متشکل از یک پزشک از هر شش پزشک (از جدول ۱.۲)

شمارهٔ نمونه (i)	شمارهٔ پزشک	تعداد ویزیت (x_i)
۱	۵	۷
۲	۱۱	۰
۳	۱۷	۷
٤	۲۳	۰
مجموع		۱٤

□

برآورد مجموعها، میانگینها و نسبتها همراه با برآوردهای واریانسها و خطاهای معیار برای نمونه‌گیری سیستماتیک در تابلوی ۱.٤ ارائه شده‌اند. اگر N در جامعه معلوم باشد، در آن صورت این فرمولها با فرمولهای ارائه شده برای نمونه‌گیری تصادفی ساده تفاوتی نخواهد داشت (ن. ک. تابلوی ۱.۳). در دو بخش بعدی، از خواص این برآوردها تحت نمونه‌گیری سیستماتیک بحث خواهد شد. خواهیم دید که این فرمولها فقط در صورتی مناسب‌اند که هیچ رابطه‌ای بین متغیر مورد برآورد و ترتیب چارچوب نمونه‌گیریی که نمونهٔ سیستماتیک از آن انتخاب می شود وجود نداشته باشد.

۳.٤ توزیع نمونه‌گیری برآوردها

در یک نمونه‌گیری تصادفی سادهٔ n عنصری از جامعه‌ای شامل N عنصر، هر یک از $\binom{N}{n}$ کل نمونه‌ها نه تنها شانس انتخاب شدن دارند بلکه شانس انتخاب شدن همهٔ آنها *یکسان* است. در یک نمونهٔ سیستماتیک یکی از k عنصر، فقط k نمونهٔ ممکن وجود دارند و انتخاب شدن یک نمونهٔ خاص بستگی به عدد تصادفیی دارد که در ابتدا انتخاب می‌شود.

تابلوی ۱.٤ برآورد مجموعها، میانگینها و واریانسها تحت نمونه‌گیری سیستماتیک و برآورد واریانسها و خطای معیار این برآوردها*

	برآورد	برآورد واریانس و خطای معیار برآورد
مجموع	x'	$\hat{Var}(x') = N^2\left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right)$ $\hat{SE}(x') = N\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$
میانگین	$\bar{x}$	$\hat{Var}(\bar{x}) = \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right)$ $\hat{SE}(\bar{x}) = \sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)$
نسبت	P_y	$\hat{Var}(P_y) = \left(\frac{N-n}{N}\right)\frac{P_y(1-P_y)}{n-1}$ $\hat{SE}(P_y) = \sqrt{\frac{N-n}{N}}\sqrt{\frac{P_y(1-P_y)}{n-1}}$

که در آن y یک متغیر دو حالتی دارای مقادیر صفر یا یک است.

بازهٔ نمونه‌گیری k را می‌توان جایگزین $\frac{N}{n}$ کرد و kn را می‌توان در شرایطی که N، یعنی تعداد عناصر موجود در جامعه، معلوم نباشد به جای N قرار داد.

* تذکر: واریانسها و خطاهای معیار برآورد شده ممکن است تحت شرایط خاص بالا به صورتی جدی اریب باشند.

این دو روش نمونه‌گیری را در یک مثال خاص مقایسه می‌کنیم.

مثال تشریحی: فرض کنید از فهرست جدول ۱.۲ نمونه ای متشکل از یک پزشک از هر پنج پزشک مطلوب است. پنج نمونهٔ ممکن همراه با برآوردهای $\bar{x}$ ، x' و p_y برای پارامترهای جامعه‌ای $\bar{X}$ ، X و P_y در جدول ۲.٤ فهرست شده‌اند. چون هر یک از پنج نمونهٔ ممکن که در جدول ۲.٤ فهرست شده‌اند شانس یکسان، یعنی $\frac{۱}{۵}$، برای انتخاب شدن دارند، میانگین $E(\bar{x})$ توزیع نمونه‌گیری میانگین برآورد شده عبارت است از (ن. ک. تابلوی ۳.۲):

$$E(\bar{x}) = \left(\frac{۱}{۵}\right)(۲/۶ + ۴/۸ + ۱/۴ + ۹/۲ + ۷/۴) = ۵/۰۸ = \bar{X}$$

به عبارت دیگر $\bar{x}$ یک برآورد ناااریب از $\bar{X}$ است. به همین ترتیب

$$E(x') = \left(\frac{۱}{۵}\right)(۶۵ + ۱۲۰ + ۳۵ + ۲۳۰ + ۱۸۵) = ۱۲۷ = X$$

$$E(p_y) = \left(\frac{۱}{۵}\right)(۰/۴ + ۰/۶ + ۰/۴ + ۰/۸ + ۰/۶) = ۰/۵۶ = P_y$$

به این ترتیب می‌بینیم که میانگین، مجموع و نسبت برآورد شده در این مورد، برآوردهایی ناااریب برای پارامترهای جامعه‌ای متناظرند.

جدول ۲.٤ پنج نمونهٔ ممکن متشکل از یک پزشک انتخاب شده از هر پنج پزشک جدول ۱.۲

عدد تصادفی انتخاب شده (شمارهٔ پزشک در نمونه)	میانگین (مجموع) برآورد شده [نسبت]
۱ (۱، ٦، ۱۱، ۱٦، ۲۱)	۲/٦ (٦۵) [۰/٤]
۲ (۲، ۷، ۱۲، ۱۷، ۲۲)	٤/۸ (۱۲۰) [۰/٦]
۳ (۳، ۸، ۱۳، ۱۸، ۲۳)	۱/٤ (۳۵) [۰/٤]
٤ (٤، ۹، ۱٤، ۱۹، ۲٤)	۹/۲ (۲۳۰) [۰/۸]
۵ (۵، ۱۰، ۱۵، ۲۰، ۲۵)	۷/٤ (۱۸۵) [۰/٦]

سپس خطاهای معیار $\bar{x}$ ، x' و p_y در زیر داده شده‌اند (ن. ک. تابلوی ۳.۲).

$$SE(\bar{x}) = \sqrt{\frac{۱}{۵}\sum[\bar{x} - E(\bar{x})]^۲}$$

$$= \left\{\frac{۱}{۵}[(۲/۶ - ۵/۰۸)^۲ + (۴/۸ - ۵/۰۸)^۲ + (۱/۴ - ۵/۰۸)^۲ + (۹/۲ - ۵/۰۸)^۲ + (۷/۴ - ۵/۰۸)^۲]\right\}^{۱/۲} = ۲/۹۰$$

$$SE(x') = \sqrt{\frac{۱}{۵}\sum[x' - E(x')]^۲}$$

$$= \left\{\frac{۱}{۵}[(۶۵ - ۱۲۷)^۲ + (۱۲۰ - ۱۲۷)^۲ + (۳۵ - ۱۲۷)^۲ + (۲۳۰ - ۱۲۷)^۲ + (۱۸۵ - ۱۲۷)^۲]\right\}^{۱/۲} = ۷۲/۵۷$$

$$SE(p_y) = \sqrt{\frac{۱}{۵}\sum[p_y - E(p_y)]^۲}$$

$$= \left\{\frac{۱}{۵}[(۰/۴ - ۰/۵۶)^۲ + (۰/۶ - ۰/۵۶)^۲ + (۰/۴ - ۰/۵۶)^۲ + (۰/۸ - ۰/۵۶)^۲ + (۰/۶ - ۰/۵۶)^۲]\right\}^{۱/۲} = ۰/۱۵$$

میانگینها و خطاهای معیار $\bar{x}$ ، x' و p_y در این مثال با میانگینها و خطاهای معیاری که تحت نمونه‌گیری تصادفی ساده به دست آمده‌اند در جدول ۳.٤ مقایسه شده‌اند. با بررسی این جدول می‌بینیم که برآوردها در هر دو طرح نمونه‌گیری، ناااریب‌اند. ولی خطاهای معیار برآوردها یکسان نیستند. در واقع، در این مثال، خطاهای معیار نمونه‌گیری سیستماتیک کمتر از مقادیر متناظر از نمونه‌گیری تصادفی ساده‌اند.

جدول ۳.٤ مقایسهٔ میانگینها و خطاهای معیار برای نمونه‌گیری تصادفی ساده و نمونه‌گیری سیستماتیک*

	برآورد	$\bar{x}$	x'	p_y
میانگین	نمونه‌گیری تصادفی ساده	۵/۰۸	۱۲۷	۰/۵٦
	نمونه‌گیری سیستماتیک	۵/۰۸	۱۲۷	۰/۵٦
خطای معیار	نمونه‌گیری تصادفی ساده	۳/۳٦	۸٤/۱۱	۰/۲۰
	نمونه‌گیری سیستماتیک	۲/۹۰	۷۲/۵۷	۰/۱۵

* در همهٔ موارد $n = ۵$.

□

به طور کلی، میانگین، مجموع و نسبت برآورد شده تحت نمونه‌گیری سیستماتیک به صورتی که در بالا توصیف شدند تنها در صورتی برآوردهایی ناااریب از پارامترهای جامعه‌ای متناظرند که نسبت $\frac{N}{k}$ یک عدد صحیح باشد. در مثال پیشین، $N = ۲۵$، $k = ۵$ و لذا $\frac{N}{k} = \frac{۲۵}{۵}$، که یک عدد صحیح است. پس همانطور که در بالا نشان داده شد، $\bar{x}$ ، x' و p_y برآوردهای ناااریب‌اند.

از سوی دیگر اگر $\frac{N}{k}$ عدد صحیح نباشد، برآوردهای نمونهٔ سیستماتیک ممکن است اریب باشند. این وضعیت را نیز در یک مثال بررسی می‌کنیم.

مثال تشریحی: فرض کنید از پزشکان فهرست شده در جدول ۱.۲ نمونه ای سیستماتیک متشکل از یک پزشک از هر شش پزشک انتخاب کنیم. در آن صورت، نمونه‌های ممکن به صورتی خواهند بود که در جدول ٤.٤ نشان داده شده‌اند. چون احتمال انتخاب هر یک از شش نمونهٔ ممکن، برابر است، میانگینهای توزیع‌های $\bar{x}$ ، x' و p_y عبارت‌اند از:

$$E(\bar{x}) = \frac{۱}{۶}(۱۲ + ۱ + ۴/۲۵ + ۶/۵ + ۳/۵ + ۱/۵) = ۴/۷۹ \neq \bar{X}$$

$$E(x') = \frac{۱}{۶}(۳۰۰ + ۲۵ + ۱۰۶/۲۵ + ۱۶۲/۵ + ۸۷/۵ + ۳۷/۵) = ۱۱۹/۷۹ \neq X$$

$$E(p_y) = \frac{۱}{۶}(۰/۸ + ۰/۲۵ + ۰/۷۵ + ۰/۵ + ۰/۵ + ۰/۵) = ۰/۵۵ \neq p_y$$

جدول ٤.٤ نمونه‌های ممکن متشکل از یک پزشک انتخاب شده از هر شش پزشک از جدول ۱.۲

عدد تصادفی انتخاب شده (شمارهٔ پزشک در نمونه)	میانگین (مجموع) برآورد شده [نسبت]
۱ (۱، ۷، ۱۳، ۱۹، ۲۵)	۱۲ (۳۰۰) [۰/۸]
۲ (۲، ۸، ۱٤، ۲۰)	۱ (۲۵) [۰/۲۵]
۳ (۳، ۹، ۱۵، ۲۱)	٤/۲۵ (۱۰٦/۲۵) [۰/۷۵]
٤ (٤، ۱۰، ۱٦، ۲۲)	٦/۵ (۱٦۲/۵) [۰/۵]
۵ (۵، ۱۱، ۱۷، ۲۳)	۳/۵ (۸۷/۵) [۰/۵]
٦ (٦، ۱۲، ۱۸، ۲٤)	۱/۵ (۳۷/۵) [۰/۵]

به این ترتیب می‌بینیم که نمونه‌گیری سیستماتیک در این مورد به برآوردهای نااریب منتهی نمی شود. □

دلیل این که برآوردهای مثال قبل ناریب نیستند آن است که اگرچه هر یک از عناصر شانس یکسان (مثلاً $\frac{۱}{k}$) برای انتخاب شدن دارند تأثیری که بر برآوردها می‌گذارند برای هر یک از عناصر یکسان نیست. مثلاً تأثیر پزشک ۱ کمتر از پزشک ۲ است زیرا پزشک ۱ با چهار پزشک دیگر در نمونه واقع می‌شود در حالی که پزشک ۲ فقط با سه پزشک دیگر در نمونه قرار می‌گیرد. به این ترتیب، اندازه‌گیریهای مربوط به پزشک ۱ کمتر از اندازه‌گیریهای پزشک ۲ در به دست آوردن برآوردها دخالت دارد. در مثال قبلی، هنگامی که $\frac{N}{k}$ یک عدد صحیح بود، هر پزشک با همان تعداد پزشک دیگر در نمونه‌ها ظاهر می‌شد و تمام پزشکان تأثیری یکسان بر برآوردهای حاصل داشتند.

اگر N، تعداد عناصر جامعه زیاد باشد، اریبیهای برآوردهای حاصل از نمونه‌های سیستماتیک به طور کلی نسبتاً ناچیز خواهند بود و جای نگرانی چندانی نخواهد بود. اندکی تعدیل در روش انتخاب عدد تصادفی آغازین، به برآوردهایی منجر خواهد شد که ناریب‌اند. این تعدیل بعداً در همین فصل (بخش ۵.٤) مورد بحث قرار خواهد گرفت.

٤.٤ واریانس برآوردها

اکنون به واریانسهای برآوردهای حاصل از نمونه‌گیری سیستماتیک می‌پردازیم. همان طور که قبلاً بحث شد، واریانس یک برآورد به این علت اهمیت دارد که معیار قابلیت اعتماد برآورد است.

در مبحث واریانسهای برآوردهای حاصل از نمونه‌گیری سیستماتیک، برای سهولت کار فرض می‌کنیم که $\frac{N}{k}$ یک عدد صحیح است که با حرف n نشان داده می‌شود. پس، نمونه‌گیری سیستماتیک یکی از هر k عنصر، جمعاً k نمونهٔ ممکن را نتیجه خواهد داد که هر یک دارای $\frac{N}{k}$ عنصر خواهند بود. نمونه‌های ممکن در جدول ۵.٤ نشان داده شده‌اند.

با بررسی نمونه‌های سیستماتیکی که در جدول ۵.٤ نشان داده‌ایم می‌بینیم که هر نمونه «خوشه»ای با n (برابر با $\frac{N}{k}$) عنصر است که این عناصر k «واحد» با یکدیگر فاصله دارند. به این ترتیب، نمونه‌گیری سیستماتیک از نظر عملیاتی عبارت است از گروه‌بندی N عنصر در k خوشه که هر یک دارای $\frac{N}{k}$ عنصر است که در روی فهرست، k واحد با یکدیگر فاصله دارند و سپس انتخاب تصادفی یکی از این خوشه‌ها. برای نشان دادن این ایده به یک مثال نگاهی می‌اندازیم.

مثال تشریحی: فرض کنیم که یک نمونهٔ سیستماتیک متشکل از یک نفر از هر پنج کارگر از روی جدول ۸.۳ انتخاب می‌شود. برای تشریح موضوع، ٤۰ کارگر جدول ۸.۳ را به جای نمونه‌ای از یک

جامعهٔ بزرگتر، کلاً یک جامعه در نظر می‌گیریم. در آن صورت $N=$ ٤٠، $k=$ ٥ و $\frac{N}{k}=n=\frac{۴۰}{۵}=۸$ خواهد بود. پنج خوشه‌ای که توسط طرح نمونه‌گیری تعریف شده‌اند در جدول ٦.٤ نشان داده شده‌اند. این خوشه‌ها پنج نمونهٔ ممکن متشکل از یک نفر از هر پنج کارگر را نشان می‌دهند که از روی فهرست مندرج در جدول ۸.۳ انتخاب شده‌اند.

□

جدول ٥.٤ نمونه‌های ممکن حاصل از یکی از هر k عنصر ($\frac{N}{k}$ یک عدد صحیح است)

عدد تصادفی انتخاب شده	شمارهٔ عناصر در نمونه	مقدار متغیر x
۱	$۱,۱+k,۱+۲k,\ldots,۱+(n-۱)k$	$X_{۱}, X_{۱+k}, X_{۱+۲k}, \ldots, X_{۱+(n-۱)k}$
۲	$۲, ۲+k, ۲+۲k, \ldots, ۲+(n-۱)k$	$X_{۲}, X_{۲+k}, X_{۲+۲k}, \ldots, X_{۲+(n-۱)k}$
⋮		
j	$j, j+k, j+۲k, \ldots, j+(n-۱)k$	$X_{j}, X_{j+k}, X_{j+۲k}, \ldots, X_{j+(n-۱)k}$
k	$k, ۲k, ۳k, \ldots, nk$	$X_{k}, X_{۲k}, X_{۳k}, \ldots, X_{nk}$

جدول ٦.٤ نمونه‌های خوشه‌ای بر مبنای داده‌های جدول ۸.۳

خوشه	شمارهٔ کارگران در خوشه	ظرفیت حیاتی تحت فشار کارگران در خوشه
۱	۳٦، ۳۱، ۲٦، ۲۱، ۱٦، ۱۱، ٦، ۱	٦۹، ۸٤، ۹٦، ۷۰، ۷٦، ۷۱، ۹۷، ۸۱
۲	۳۷، ۳۲، ۲۷، ۲۲، ۱۷، ۱۲، ۷، ۲	۸۰، ۸۹، ٦۲، ٦٤، ٦۲، ۸۸، ۸۲، ٦٤
۳	۳۸، ۳۳، ۲۸، ۲۳، ۱۸، ۱۳، ۸، ۳	۹۸، ۸۹، ٦۷، ۷۲، ٦۷، ۸٤، ۹۹، ۸٥
٤	۳۹، ۳٤، ۲۹، ۲٤، ۱۹، ۱٤، ۹، ٤	٦٥، ٦٥، ۹٥، ۷۲، ۹۱، ۸٥، ۹٦، ۹۱
٥	٤۰، ۳٥، ۳۰، ۲٥، ۲۰، ۱٥، ۱۰، ٥	۸٤، ٦۷، ۸۷، ۹٥، ۹۹، ۷۷، ۹۱، ٦۰

اگر عناصر را با اندیسهای دوگانه برای نشان دادن خوشهٔ خاص برچسب بزنیم واریانس برآوردهای حاصل از نمونه‌گیری سیستماتیک را می‌توان آسانتر فهمید. برای مثال، عناصر موجود در خوشهٔ اول، یعنی عناصر ۱، $۱+k$ ، $۱+۲k$ ، ... ، $۱+(n-۱)k$ مجدداً با اندیسهای دوگانه به شرح زیر شماره‌گذاری خواهند شد:

شمارهٔ اصلی	برچسب جدید	مقدار متغیر x
۱	۱,۱	$X_{۱۱}$
$۱+k$	۱,۲	$X_{۱۲}$
$۱+۲k$	۱,۳	$X_{۱۳}$
$۱+(n-۱)k$	۱, n	$X_{۱n}$

عناصر موجود در خوشهٔ jام نیز به همین ترتیب شماره‌گذاری خواهند شد:

شمارهٔ اصلی	برچسب جدید	مقدار متغیر x
j	j ,۱	X_{j1}
$j+k$	j ,۲	X_{j2}
$j+2k$	j ,۳	X_{j3}
$j+(n-1)k$	j , n	X_{jn}

واریانسهای میانگینها، مجموعها و نسبتهای به دست آمده از نمونه‌گیری سیستماتیک با این اندیسگذاری جدید، در تابلوی ۲.۴ ارائه شده‌اند.

عبارتهایی که در تابلوی ۲.۴ نشان داده شده‌اند، برابرند با عبارتهای مربوط به برآوردهای حاصل از نمونه‌گیری تصادفی ساده ضرب در عامل $[1+\delta_x(n-1)]$. پارامتر δ_x به نام *ضریب همبستگی درون ـ رده‌ای* خوانده می‌شود و معیاری برای همگنی عناصر *در داخل* k نمونهٔ سیستماتیک ممکن یا خوشه است که از لحاظ نظری می‌تواند از جامعه ای متشکل از $N=nk$ عنصر انتخاب شود. محاسبهٔ δ_x برای هر خوشه مستلزم تشکیل $\binom{n}{2}$ جفت از مقادیر (X_{ij}, X_{il}) است که همان نقشی را به عهده دارند که جفتهای (X,Y) در ضریب همبستگی گشتاور حاصل‌ضربی معمولی ایفا می‌کنند. تفاوت در این است که انحرافها حول $(\overline{X},\overline{X})$ محاسبه می شوند که $\overline{X}$ عبارت از میانگین X_{ij} برای همهٔ مشاهدات در کلیهٔ نمونه‌های سیستماتیک ممکن (یعنی میانگین جامعه) است. مجموعیابی، در صورت کسر رابطهٔ (۵.۴) مستلزم ضرب متقاطع این انحرافها در همهٔ زوجهای نقطه‌ها برای k نمونهٔ سیستماتیک ممکن است. این ضریب همبستگی درون ـ رده‌ای می‌تواند از مقادیر بسیار کوچک منفی، وقتی عناصر داخل هر خوشه به تنوع بسیار زیاد یا نمایندگی عناصر جامعه گرایش دارند (که این حالت را «ناهمگنی» می‌خوانند) تا حداکثر برابر با ۱، وقتی عناصر داخل هر خوشه مشابه یکدیگرند ولی با عناصر سایر خوشه‌ها تفاوت دارند (که این حالت را «همگنی» می‌گویند) تغییر کند.

از عبارتهای (۲.۴)، (۳.۴) و (۴.۴) واضح است که هرگاه δ_x بزرگ باشد، واریانس مجموع، میانگین، یا نسبت نیز زیاد خواهد بود. این زمانی روی می‌دهد که عناصر جامعه که در فهرستی منظم شده‌اند درجهٔ بالایی از دوره‌ای بودن را نشان دهند. هرگاه δ_x با صفر برابر شود، واریانسهای حاصل همانند واریانسهای حاصل از نمونه‌گیری تصادفی ساده‌اند. این، معمولاً زمانی روی می‌دهد که عناصر جامعه با توجه به متغیر تحت بررسی به ترتیبی تصادفی مرتب شده باشند.

تابلوی ۲.٤ واریانسهای برآوردهای جامعه‌ای تحت نمونه‌گیری سیستماتیک

مجموع، x'

$$Var(x') = \left(\frac{N^{۲}\sigma_x^{۲}}{n}\right)[۱ + \delta_x(n-۱)] \qquad (۲.٤)$$

میانگین، $\bar{x}$

$$Var(\bar{x}) = \left(\frac{\sigma_x^{۲}}{n}\right)[۱ + \delta_x(n-۱)] \qquad (۳.٤)$$

نسبت، P_y

$$Var(p_y) = \frac{p_y(۱-p_y)}{n}[۱ + \delta_x(n-۱)] \qquad (٤.٤)$$

که در آن $n = \frac{n}{k}$، و

$$\delta_x = \frac{۲\sum_{i=۱}^{k}\sum_{j=۱}^{n}\sum_{i<j}(X_{ij}-\bar{X})(X_{il}-\bar{X})}{nk(n-۱)\sigma_x^{۲}} \qquad (۵.٤)$$

در این معادله‌ها، N، اندازهٔ جامعه، n، اندازهٔ نمونه، k، بازهٔ نمونه‌گیری، X_{ij}، i امین عنصر از خوشهٔ jام، X_{il}، یک عنصر دیگر از خوشهٔ i ام $(l \neq j)$، و نمادگذاری جامعه به صورتی است که در تابلوی ۱.۲ تعریف شده‌اند.

سرانجام، هرگاه δ_x کوچک و منفی باشد واریانسهای به دست آمده کوچکتر از واریانسهای حاصل از نمونه‌گیری تصادفی ساده خواهند بود. این حالت می‌تواند زمانی روی دهد که عناصر جامعه با توجه به متغیر تحت بررسی مرتب شده باشند.

حال با بررسی یک مثال ببینیم همهٔ اینها چه معنی می‌دهند.

مثال تشریحی: فرض کنید فهرست وقت ملاقاتهای یک پرستار متخصص را در اختیار داریم و می‌خواهیم نمونه‌ای متشکل از یک بیمار از هر چهار بیماری را که این پرستار در روزی معین دیده است برای برآورد متوسط زمان صرف شده به ازای هر بیمار، انتخاب کنیم. فرض کنید این پرستار در روزی که برای گرفتن نمونه انتخاب شده است ۱۲ بیمار را به ترتیبی که در جدول ۷.٤ نشان داده‌ایم ویزیت کرده باشد.

چون نمونه‌ای متشکل از یک نفر از چهار بیمار را تعیین کرده‌ایم، چهار نمونهٔ ممکن در جدول ۸.٤ نشان داده شده‌اند.

میانگین زمان صرف شده، $\overline{X}$، برای ۱۲ بیمار ۲۹/۵۸۳ دقیقه و واریانس آن σ_x^2=۱۵۳/۰۸ است. واریانس میانگین زمان برآورد شده برای دیدار از هر بیمار برای همهٔ نمونه‌های ممکن عبارت است از (تابلوی ۳.۲ را ببینید):

$$Var(\bar{x}) = \frac{۱}{۴}\left[(۱۷ - ۲۹/۵۸۳)^۲ + (۳۲/۳۳ - ۲۹/۵۸۳)^۲ + (۳۹ - ۲۹/۵۸۳)^۲ + (۳۰ - ۲۹/۵۸۳)^۲\right] = ۶۳/۳$$

جدول ۷.۴ داده‌های مربوط به ویزیتهای پرستار متخصص (فهرست مرتب نشده)

ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)	ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)
۱	۱۵	۷	۴۹
۲	۳۴	۸	۴۰
۳	۳۵	۹	۲۵
۴	۳۶	۱۰	۴۶
۵	۱۱	۱۱	۳۳
۶	۱۷	۱۲	۱۴

جدول ۸.۴ چهار نمونهٔ ممکن برای داده‌های جدول ۷.۴

بیمار	زمان صرف شده (دقیقه)	بیمار	زمان صرف شده (دقیقه)	بیمار	زمان صرف شده (دقیقه)	بیمار	زمان صرف شده (دقیقه)
۱	۱۵	۲	۳۴	۳	۳۵	۴	۳۶
۵	۱۱	۶	۱۷	۷	۴۹	۸	۴۰
۹	۲۵	۱۰	۴۶	۱۱	۳۳	۱۲	۱۴
مجموع	۵۱		۹۷		۱۱۷		۹۰
میانگین	۱۷		۳۲/۳۳		۳۹		۳۰

برای این که نشان دهیم این با عبارت داده شده در رابطهٔ (۳.۴) برابر است، پارامترهای لازم را محاسبه خواهیم کرد. ضریب همبستگی درون ـ رده‌ای عبارت است از (تابلوی ۲.۴):

$$\delta_x = \frac{\begin{array}{l}۲[(۱۵ - ۲۹/۵۸۳)(۱۱ - ۲۹/۵۸۳) + (۱۵ - ۲۹/۵۸۳)(۲۵ - ۲۹/۵۸۳) + (۱۱ - ۲۹/۵۸۳)(۲۵ - ۲۹/۵۸۳) \\ + (۳۴ - ۲۹/۵۸۳)(۱۷ - ۲۹/۵۸۳) + \ldots + (۴۰ - ۲۹/۵۸۳)(۱۴ - ۲۹/۵۸۳)]\end{array}}{۳ \times ۴ \times (۳ - ۱) \times ۱۵۳/۰۸}$$

$$= ۰/۱۲۴۱$$

پس با σ_x^2=۱۵۳/۰۸، n=۳ و δ_x=۰/۱۲۴۱، از روی معادلهٔ (۳.۴) داریم:

$$Var(\bar{x}) = \left(\frac{۱۵۳/۰۸}{۳}\right)[۱ + ۰/۱۲۴۱\ (۳ - ۱)] = ۶۳/۶$$

که با آنچه در عمل به دست آمد توافق دارد.

حالا فرض کنید که این پرستار قرار ملاقاتهای خود را طوری زمانبندی کند که بیماران بدقلقی را که به زمان بیشتری نیاز دارند اول از همه ببیند و بیماران بی‌تکلفتر را که وقت کمتری می‌گیرند در اواخر روز ویزیت کند. فهرست وقت ملاقات برای همان بیمارانی که قبلاً بحث شد ممکن است به صورتی که در جدول ۹.٤ نشان داده شده است درآید.

اگر قرار می‌بود یک نمونهٔ سیستماتیک متشکل از یک بیمار از هر چهار بیمار را از فهرست وقت ملاقات مندرج در جدول ۹.٤ انتخاب کنیم، چهار نمونهٔ ممکن به شرح جدول ۱۰.٤ به دست می‌آمد.

جدول ۹.٤ داده‌های مربوط به ویزیتهای پرستار متخصص (فهرست یکنوا ـ مرتب)

ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)	ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)
۱	٤۹	۷	۳۳
۲	٤٦	۸	۲۵
۳	٤۰	۹	۱۷
٤	۳٦	۱۰	۱۵
۵	۳۵	۱۱	۱٤
٦	۳٤	۱۲	۱۱

جدول ۱۰.٤ چهار نمونهٔ ممکن برای داده‌های جدول ۹.٤

نمونهٔ ۱		نمونهٔ۲		نمونهٔ۳		نمونهٔ ٤	
بیمار	**زمان صرف شده (دقیقه)**	**بیمار**	**زمان صرف شده (دقیقه)**	**بیمار**	**زمان صرف شده (دقیقه)**	**بیمار**	**زمان صرف شده (دقیقه)**
۱	٤۹	۲	٤٦	۳	٤۰	٤	۳٦
۵	۳۵	٦	۳٤	۷	۳۳	۸	۲۵
۹	۱۷	۱۰	۱۵	۱۱	۱٤	۱۲	۱۱
مجموع	۱۰۱		۹۵		۸۷		۷۲
میانگین	۳۳/٦۷		۳۱/٦۷		۲۹		۲٤

چون شیوه‌ای که نمونه با آن انتخاب شده است نمونهٔ سیستماتیک یکی از هر چهار عنصر بوده است، و چون $\frac{N}{k} = \frac{۱۲}{۴} = ۳$ یک عدد صحیح است، میانگین نمونه، مانند قبل، برآوردی ناااریب از میانگین جامعه است، یعنی $E(\bar{x}) = \bar{X} = ۲۹/۵۸۳$. واریانس $\bar{x}$، میانگین برآورد شده، عبارت است از:

$$Var(\bar{x}) = \left(\frac{۱}{۴}\right)[(۳۳/۶۷ - ۲۹/۵۸۳)^۲ + (۳۱/۶۷ - ۲۹/۵۸۳)^۲ + (۲۹ - ۲۹/۵۸۳)^۲ + (۲۴ - ۲۹/۵۸۳)^۲] = ۱۳/۱۳$$

به این ترتیب می بینیم که در این وضعیت ـ وقتی فهرست وقت ملاقات مرتب شود ـ از شیوهٔ نمونه‌گیری سیستماتیک، یک برآورد میانگین به دست می‌آید که واریانس آن کمتر از واریانسی است که از همان شیوهٔ نمونه‌گیری سیستماتیک از فهرست نامرتب به دست می‌آید. این موضوع از نظر شهودی هم معقول است، زیرا فرایند نمونه‌گیری سیستماتیک از فهرستی که مطابق با سطح متغیر مورد اندازه‌گیری مرتب شده است تضمین می‌کند که هر نمونه دارای تعدادی از عناصر با مقدار زیاد، تعدادی با مقدار کم و تعدادی با مقدار متوسط است. به عبارت دیگر، امکان نخواهد داشت که مقادیری که به طور نامعمول زیاد یا به طور نامعمول کم‌اند در یک نمونهٔ خاص تمرکز پیدا کنند.

ضریب همبستگی درون رده‌ای در این مورد برابر است با ۰/۳۷۱- که یک عدد منفی و به صورتی قابل ملاحظه کمتر از آن است که هنگام نمونه‌گیری از فهرست نامرتب به دست می‌آمد.

حالا فرض می‌کنیم که این پرستار زمانبندی خود را طوری تنظیم کند که بعد از یک بیمار آرام، دو بیمار نسبتاً بدقلق و سپس یک بیمار بسیار بدقلق (از لحاظ زمان موردنیاز) ویزیت شوند. در این صورت فهرست وقت ملاقات برای همان ۱۲ بیمار ممکن است به شکلی درآید که در جدول ۱۱.٤ نشان داده شده است.

اگر قرار بود که به طور سیستماتیک یکی از هر چهار بیمار را از فهرست جدول ۱۱.٤ برای نمونه بگیریم، چهار نمونهٔ ممکنی که در جدول ۱۲.٤ فهرست شده‌اند به دست می‌آمدند.

جدول ۱۱.٤ داده‌های مربوط به ویزیتهای پرستار متخصص (دوره‌ای بودن در فهرست)

ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)	ترتیب ویزیت	زمان صرف شده با بیمار (دقیقه)
۱	۱۱	۷	۳۵
۲	۱۷	۸	٤٦
۳	۳٦	۹	۱۵
٤	٤۹	۱۰	۲۵
۵	۱٤	۱۱	۳۳
٦	۳٤	۱۲	٤۰

جدول ۱۲.٤ چهار نمونه ممکن برای داده‌های جدول ۱۱.٤

بیمار	زمان صرف شده	بیمار	زمان صرف شده	بیمار	زمان صرف شده	بیمار	زمان صرف شده
۱	۱۱	۲	۱۷	۳	۳٦	٤	٤۹
۵	۱٤	٦	۳٤	۷	۳۵	۸	٤٦
۹	۱۵	۱۰	۲۵	۱۱	۳۳	۱۲	٤۰
جمع	٤۰		۷٦		۱۰٤		۱۳۵
میانگین	۱۳/۳۳		۲۵/۳۳		۳٤/٦۷		٤۵

باز هم میانگین برآورد شده، یک برآورد نااریب از میانگین جامعه است. واریانس توزیع میانگینهای برآورد شده عبارت است از:

$$Var(\bar{x}) = \left(\frac{۱}{۴}\right)[(۱۳/۳۳ - ۲۹/۵۸۳)^۲ + (۲۵/۳۳ - ۲۹/۵۸۳)^۲ + (۳۴/۶۷ - ۲۹/۵۸۳)^۲ + (۴۵ - ۲۹/۵۸۳)^۲]$$

$$= ۱۳۶/۴۱$$

ضریب همبستگی درون ـ رده‌ای δ_x در این وضعیت برابر است با ۰/۸۳۷. به این ترتیب، واریانس میانگین برآورد شده در این حالت خیلی بیشتر از قبل است. نتایجی که به دست آورده‌ایم درجدول ۱۳.٤ خلاصه شده‌اند.

جدول ۱۳.٤ خلاصهٔ نتایج به دست آمده از چهار نوع نمونه‌گیری

طرح نمونه	$Var(\bar{x})$	ضریب همبستگی درون ـ رده‌ای
نمونه‌گیری سیستماتیک با فهرست نامرتب:	٦۳/٦	۰/۱۲٤
نمونه‌گیری سیستماتیک با فهرست یکنوا ـ مرتب:	۱۳/۱	-۰/۳۷۱
نمونه‌گیری سیستماتیک با دوره‌ای بودن فهرست:	۱۳٦/٤	۰/۸۳۷
نمونه‌گیری تصادفی ساده:	٤۱/۷	ـــــ

در سومین حالت نمونه‌گیری سیستماتیک ـ یعنی حالتی که در آن پرستار بعد از یک بیمار آرام، ویزیت دو بیمار نسبتاً بدقلق و سپس یک بیمار بسیار بدقلق را به صورت الگویی مداوم زمانبندی کرده است ـ واریانس میانگین برآورد شده بسیار زیاد است. علت این واریانس زیاد آن است که الگوی زمانبندی پرستار مستلزم دوره‌ای بودن ۱ در ٤ است که با دوره‌ای بودن در انتخاب نمونه منطبق می‌شود و در نتیجه تمام بیماران بدقلق در یک نمونه و تمام بیماران آرام نیز در یک نمونه قرارمی‌گیرند. از این‌رو، تغییرپذیری نمونه‌گیری میانگینهای برآورد شده بسیار بالا است. این واریانس

بالا با واریانس پایینی که هنگام تنظیم قرار ملاقاتها به ترتیب افزایش بدقلقی به دست آمد در تضاد است. همان‌طور که قبلاً تذکر داده شد، نمونه‌گیری سیستماتیک از این فهرست مرتب به ترتیب افزایش بدقلقی منجر به آن شد که هر نمونهٔ ممکن سهمی از بیماران بدقلق، متوسط و آرام داشته باشد به نحوی که نمونه‌ها از لحاظ توزیع متغیر مورد اندازه‌گیری چندان تفاوتی با یکدیگر نداشتند.
□

برای تعمیم مطلب، نمونه‌گیری سیستماتیک از فهرستی که از نظر متغیر مورد اندازه‌گیری مرتب شده است غالباً به برآوردهایی منجر می‌شود که واریانس نمونه‌گیری کمی دارند. از سوی دیگر، اگر فهرست، دارای دوره‌ای بودنی است که با دوره‌ای بودن نمونه‌گیری مطابقت دارد برآوردهای به دست آمده می‌توانند واریانسهای نمونه‌گیری بسیار زیادی داشته باشند و برآوردهای آمارگیری رابطهٔ چندانی با پارامترهای جامعهٔ نظیر نخواهند داشت.

وقتی فهرستی دارای «ترتیب تصادفی» باشد ـ یعنی، دوره‌ای بودن یا ترتیبی خاص در فهرست نباشد ـ واریانس مجموع، میانگین، یا نسبت برآورد شده که از نمونه‌گیری سیستماتیک به دست می‌آید تقریباً همان است که برای نمونه‌گیری تصادفی ساده مناسب است. به عبارت دیگر

$$Var(x') \approx \left(\frac{N^2}{n}\right)(\sigma_x^2)\left(\frac{N-n}{N-1}\right)$$

$$Var(\bar{x}) \approx \left(\frac{\sigma_x^2}{n}\right)\left(\frac{N-n}{N-1}\right)$$

$$Var(p_y) \approx \left[\frac{p_y(1-p_y)}{n}\right]\left(\frac{N-n}{N-1}\right)$$

توجه کنید که در مثال قبل، ضریب همبستگی درون ـ رده‌ای δ_x، در وضعیتی که فهرست دوره‌ای باشد بیشترین مقدار بود. این نتیجه با توجه به این حقیقت که δ_x همگنی عناصر را با توجه به متغیر مورد اندازه‌گیری می‌سنجد قابل درک است. هنگامی که دوره‌ای بودن موجود در فهرست وقت ملاقات پرستار با دوره‌ای بودن بازهٔ نمونه‌گیری منطبق شد (یعنی هنگامی که $k=4$)، تمام بیماران بدقلق در یک نمونه و تمام بیماران آرام در نمونهٔ دیگر قرار گرفتند. این وضعیت، همگنی بیماران در داخل یک خوشه، تفاوتهای زیاد بین بیماران در نمونه‌های مختلف، ضریب همبستگی درون ـ رده‌ای بالا، و واریانس زیاد برای برآورد میانگین را ایجاد کرد.

۵.٤ تعدیلی که همیشه برآوردهای نااریب را نتیجه می‌دهد

قبلاً نشان دادیم که نمونه‌گیری سیستماتیک یک عنصر از هر k عنصر، وقتی نسبت $\frac{N}{k}$ یک عدد صحیح نباشد برآوردهای نااریب به بار نمی‌آورد، هرچند در صورتی که N و k به صورتی معقول

بزرگ باشند احتمالاً اریبی کم است. در این بخش، روشی را برای گرفتن نمونهٔ سیستماتیک یک عنصری از هر k عنصر ارائه می‌دهیم که همیشه به برآوردهای نااریب از میانگین، مجموع و نسبتهای جامعه می‌انجامد. ولی این روش مستلزم شناخت قبلی اندازهٔ جامعه است و بنابراین تنها در وضعیتهایی سودمند است که N معلوم باشد. این روش را با یک مثال شرح می‌دهیم.

مثال تشریحی: فرض کنید می‌خواهیم یک نفر از هر ٦ نفر را از فهرست ۲۵ پزشک جدول ۱.۲ نمونه بگیریم و پیش از نمونه‌گیری می‌دانیم که ۲۵ پزشک در فهرست وجود دارند. به جای این که یک عدد تصادفی از ۱ تا ٦ برای شروع نمونه‌گیری سیستماتیک انتخاب کنیم یک عدد تصادفی j از ۱ تا ۲۵ انتخاب می‌کنیم. سپس j را به ٦ قسمت تقسیم می‌کنیم و باقیمانده را به دست می‌آوریم. اگر برای مثال، عدد تصادفی منتخب ۹ باشد پس $۱\frac{۳}{۶}=\frac{۹}{۶}$. باقیمانده ۳ است و از سومین پزشک شروع می‌کنیم. اگر باقیمانده ۱ بود با اولین پزشک شروع می‌کردیم و قس علی هذا. اگر باقیمانده‌ای وجود نداشته باشد با عنصر k شروع می‌کنیم (برای مثال در این مورد $k=$ ٦).

حالا توزیع باقیمانده‌ها و نمونه‌ها را (که در جدول ۱٤.٤ نشان داده شده است) برای ۲۵ عدد تصادفی ممکن بررسی می‌کنیم. واضح است که نمونه‌ها احتمال یکسانی برای انتخاب شدن ندارند. مثلاً، شانس وقوع باقیماندهٔ ۱ برابر $\frac{۵}{۲۵}$ است زیرا وقتی رخ می‌دهد که اعداد تصادفی ۱، ۷، ۱۳، ۱۹ و ۲۵ انتخاب شوند. شانس رخداد سایر باقیمانده‌ها $\frac{۴}{۲۵}$ است (مثلاً باقیماندهٔ ۵ هنگامی رخ می‌دهد که اعداد ۵، ۱۱، ۱۷ و ۲۳ انتخاب شوند).

با در نظر داشتن این موضوع، توزیع میانگینهای برآورد شده را که در جدول ۱۵.٤ نشان داده شده است بررسی می‌کنیم.

$E(\bar{x})$ میانگین توزیع میانگین برآورد شدهٔ ویزیتهای از خانوارها در این برنامهٔ تعدیل یافتهٔ نمونه‌گیری سیستماتیک، عبارت است از (ن. ک. تابلوی ٤.۲).

$$E(\bar{x}) = (۱۲)\times(\frac{۵}{۲۵}) + (۱)\times(\frac{۴}{۲۵}) + (۴/۲۵)\times(\frac{۴}{۲۵}) + (۶/۵)\times(\frac{۴}{۲۵})$$
$$+ (۳/۵)\times(\frac{۴}{۲۵}) + (۱/۵)\times(\frac{۴}{۲۵}) = ۵/۰۸ = \overline{X}$$

به این ترتیب می‌بینیم که این روش برخلاف روشی که قبلاً توصیف شد، حتی وقتی که $\frac{N}{k}$ یک عدد صحیح نباشد، به برآوردی نااریب از میانگین جامعه منجر می‌شود. به همین ترتیب به برآوردی نااریب از مجموع کل و نسبتهای جامعه نیز می‌انجامد.

□

جدول ١٤.٤ توزیع باقیمانده‌ها و نمونه‌ها برای ۲۵ عدد تصادفی ممکن

عدد تصادفی j	j/k	باقیمانده	عناصر نمونه
۱	۱/۶	۱	۱، ۷، ۱۳، ۱۹، ۲۵
۲	۲/۶	۲	۲، ۸، ۱٤، ۲۰
۳	۳/۶	۳	۳، ۹، ۱۵، ۲۱
٤	۴/۶	٤	٤، ۱۰، ۱٦، ۲۲
۵	۵/۶	۵	۵، ۱۱، ۱۷، ۲۳
٦	۶/۶	۰	٦، ۱۲، ۱۸، ۲٤
۷	۷/۶	۱	۱، ۷، ۱۳، ۱۹، ۲۵
۸	۸/۶	۲	۲، ۸، ۱٤، ۲۰
۹	۹/۶	۳	۳، ۹، ۱۵، ۲۱
۱۰	۱۰/۶	٤	٤، ۱۰، ۱٦، ۲۲
۱۱	۱۱/۶	۵	۵، ۱۱، ۱۷، ۲۳
۱۲	۱۲/۶	۰	٦، ۱۲، ۱۸، ۲٤
۱۳	۱۳/۶	۱	۱، ۷، ۱۳، ۱۹، ۲۵
۱٤	۱۴/۶	۲	۲، ۸، ۱٤، ۲۰
۱۵	۱۵/۶	۳	۳، ۹، ۱۵، ۲۱
۱٦	۱۶/۶	٤	٤، ۱۰، ۱٦، ۲۲
۱۷	۱۷/۶	۵	۵، ۱۱، ۱۷، ۲۳
۱۸	۱۸/۶	۰	٦، ۱۲، ۱۸، ۲٤
۱۹	۱۹/۶	۱	۱، ۷، ۱۳، ۱۹، ۲۵
۲۰	۲۰/۶	۲	۲، ۸، ۱٤، ۲۰
۲۱	۲۱/۶	۳	۳، ۹، ۱۵، ۲۱
۲۲	۲۲/۶	٤	٤، ۱۰، ۱٦، ۲۲
۲۳	۲۳/۶	۵	۵، ۱۱، ۱۷، ۲۳
۲٤	۲۴/۶	۰	٦، ۱۲، ۱۸، ۲٤
۲۵	۲۵/۶	۱	۱، ۷، ۱۳، ۱۹، ۲۵

برای تعمیم این مثال، می‌توان روش تعدیل یافته را به شرح زیر به کار بست :

۱. انتخاب یک عدد تصادفی بین ۱ و N که در آن N عبارت است از تعداد عناصر در جامعه.

۲. محاسبه خارج قسمت $\frac{j}{k}$ که در آن j عدد تصادفی انتخاب شده و k بازهٔ نمونه‌گیری است. بیان این خارج قسمت به صورت یک عدد صحیح و یک باقیمانده (برای مثال، $۳\frac{۵}{۶} = \frac{۲۳}{۶}$ و باقیمانده ۵ است).

۳. اگر باقیمانده صفر باشد یک نمونهٔ سیستماتیک یکی از هر k عنصر به روش معمول با شروع از عنصر k انتخاب می‌شود. اگر باقیمانده صفر نباشد (مثلاً m)، یک نمونهٔ سیستماتیک یکی از هر k عنصر با شروع از عنصر m انتخاب می‌شود.

جدول ۱۵.٤ توزیع نمونه‌گیری میانگینهای برآورد شده از روی داده‌های جداول ۱٤.٤ و ۱.۲

عناصر نمونه	میانگین تعداد ویزیت‌های برآورد شده $\bar{x}$	شانس انتخاب شدن π
۲۵، ۱۹، ۱۳، ۷، ۱	۱۲/۰۰	$\frac{۵}{۲۵}$
۲۰، ۱٤، ۸، ۲	۱/۰۰	$\frac{۴}{۲۵}$
۲۱، ۱۵، ۹، ۳	٤/۲۵	$\frac{۴}{۲۵}$
۲۲، ۱٦، ۱۰، ٤	٦/۵۰	$\frac{۴}{۲۵}$
۲۳، ۱۷، ۱۱، ۵	۳/۵۰	$\frac{۴}{۲۵}$
۲٤، ۱۸، ۱۲، ٦	۱/۵۰	$\frac{۴}{۲۵}$

٦.٤ برآورد کردن واریانسها

مانند همهٔ روشهای نمونه‌گیری، در اینجا نیز برای ایجاد بازه‌های اطمینان به برآوردهایی از خطاهای معیار پارامترهای جامعه‌ای برآورد شده نیاز است. در این بخش نشان می‌دهیم که برآوردهای خطاهای معیار در نمونه‌گیری سیستماتیک عملاً چگونه به دست می‌آیند.

واریانس $\bar{x}$، میانگین برآورد شدهٔ یک نمونهٔ سیستماتیک یک عنصر از هر k عنصر (با فرض این که $\frac{N}{k}$ یک عدد صحیح باشد) عبارت از متوسط توان دوم انحراف میانگین برآورد شدهٔ k نمونهٔ

ممکـن نسبت به میانگین واقعی جامعه است (زیرا وقتی $\frac{N}{k}$ یک عدد صحیـــــح باشد برآورد نـااریب اسـت). به عبـارت دیگـر، اگـر $\bar{x}_i$ نشـان‌دهنـدهٔ میانگـین مقـدار χ بـرای عناصـر i، $i+k$، $i+۲k$، ...، $i+(n-۱)k$ باشد، در آن صـورت *واریانـس* $\bar{x}$، *برآورد میانگـین* در نمـونهٔ سیستماتیک یک عنصر از هر k عنصر، از فرمول زیر به دست می‌آید:

$$Var(\bar{x}) = \frac{\sum_{i=۱}^{k} (\bar{x}_i - \bar{X})^۲}{k} \qquad (٦.٤)$$

این عبارت را با واریانس میانگین برآورد شده تحت نمونه‌گیری تصادفی ساده که به صورت زیر تعریف می‌شود مقایسه می‌کنیم:

$$Var(\bar{x}) = \frac{\sum_{i=۱}^{M} (\bar{x}_i - \bar{X})^۲}{M}$$

که در آن M تعداد نمونه‌های ممکن است. این یکی از خواص جالب نمونه‌گیری تصادفی ساده است که این واریانس را می‌توان به سادگی برحسب $\sigma_x^۲$، واریانس مشاهدات اصلی، با استفاده از فرمول زیر بیان کرد:

$$Var(\bar{x}) = \left(\frac{N-n}{N-۱}\right)\left(\frac{\sigma_x^۲}{n}\right)$$

آنچه برای آمارشناس از اهمیت بیشتری برخوردار است آن است که به وسیلهٔ برآورد کردن واریانس $\sigma_x^۲$ با آمارهٔ $s_x^۲$ که از مشاهدات نمونه محاسبه می‌شود، می‌توان واریانس $\bar{x}$ را به شرح زیر برآورد کرد:

$$\hat{Var}(\bar{x}) = \left(\frac{N-n}{N}\right)\left(\frac{s_x^۲}{n}\right)$$

به این ترتیب، فقط با دانستن واریانس مشاهدات در یک نمونهٔ *بخصوص*، می‌توانیم واریانس توزیع نمونه‌گیری میانگین برآورد شده را تحت نمونه‌گیری تصادفی ساده برآورد کنیم.

متأسفانه، این مطلب را در مورد نمونه‌گیری سیستماتیک نمی‌توان گفت. برای برآورد واریانس ارائه شده در معادلهٔ (٦.٤)، ضروری است دو یا چند تا از این $\bar{x}_i$ در دسترس ما باشند. ولی در نمونهٔ سیستماتیک ما، میانگین برآورد شده $\bar{x}$ صرفاً یکی از این $\bar{x}_i$ است و به خصوص همان است که به عدد تصادفی برای شروع نمونه‌گیری بستگی دارد. ما از نمونهٔ خود در مورد تغییرپذیری میانگینهای برآورد شده در کل نمونه‌های ممکن هیچ اطلاعی نداریم.

در عمل، اگر بتوانیم فرض کنیم فهرستی که نمونهٔ سیستماتیک از آن گرفته شده است ترتیب تصادفی عناصر را نسبت به متغیر مورد اندازه‌گیری ارائه می‌دهد، در آن صورت می‌توانیم فرض کنیم که نمونهٔ سیستماتیک هم‌ارز با نمونهٔ تصادفی ساده است. بنابراین، از شیوه‌های تهیه شده برای برآورد

واریانسهای برآوردهای حاصل از نمونه‌های تصادفی ساده می‌توان استفاده کرد. به عبارت دیگر، σ_x^2، واریانس جامعه را به وسیله $\hat{\sigma}_x^2$ برآورد می‌کنیم که از فرمول زیر به دست می‌آید:

$$\hat{\sigma}_x^2 = \left(\frac{N-1}{N}\right) s_x^2$$

که در آن

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

و x_iها مشاهدات نمونه بوده و $n = \frac{N}{k}$. پس *برآورد واریانس* $\bar{x}$، یعنی برآورد میانگین حاصل از نمونهٔ سیستماتیک از فرمول زیر به دست می‌آید:

$$\hat{Var}(\bar{x}) = \left(\frac{\hat{\sigma}_x^2}{n}\right)\left(\frac{N-n}{N-1}\right) \qquad (۷.٤)$$

پس با استفاده از این عبارت می‌توانیم بازه‌های اطمینان را برای $\bar{X}$، میانگین جامعه به روش معمول به دست آوریم.

مثال تشریحی: فرض کنید یک نمونهٔ سیستماتیک متشکل از یک پزشک از هر پنج پزشک را از فهرستی که در جدول ۱.۲ ارائه شده است انتخاب کنیم و اولین عدد تصادفی انتخاب شده نیز۳ باشد. جدول ١٦.٤ پزشکان نمونه را همراه با تعداد ویزیتهای آنها از خانوار فهرست می‌کند.

با استفاده از داده‌های جدول ١٦.٤ محاسبات زیر را داریم (تابلوی ۲.۲ و معادله‌های بالا را ببینید):

جدول ١٦.٤ نمونهٔ سیستماتیک یک از پنج تا، انتخاب شده از جدول ۱.۲

پزشک نمونه	تعداد ویزیتها x_i
۳	۱
۸	۰
۱۳	٦
۱۸	۰
۲۳	۰

$N = ۲۵$ $\quad n = ۵$ $\quad \bar{x} = ۱/٤$

$$s_x^2 = \frac{1}{4}[(1 - 1/4)^2 + (0 - 1/4)^2 + (6 - 1/4)^2 + (0 - 1/4)^2 + (0 - 1/4)^2] = 6/8$$

$$\hat{\sigma}_x^2 = \frac{24}{25} \times 6/8 = 6/528$$

$$\hat{Var}(\bar{x}) = \left(\frac{6/528}{5}\right)\left(\frac{25 - 5}{24}\right) = 1/088$$

کران بالاو کران پایین بازه‌های اطمینان برای یک بازهٔ اطمینان ۹۵ درصدی به شرح زیر به دست می‌آید:

$$\bar{x} + (1/96)\sqrt{\hat{Var}(\bar{x})} = 1/4 + (1/96)\sqrt{1/088} = 3/44$$

و

$$\bar{x} - (1/96)\sqrt{\hat{Var}(\bar{x})} = 1/4 - (1/96)\sqrt{1/088} = -0/64 = 0$$

(برای کران پایین، صفر را به کار می‌بریم، زیرا اعداد منفی در این مثال هیچ مفهومی ندارند.) به این ترتیب بازهٔ اطمینان میانگین تعداد ویزیتها بین صفر و۳/٤٤ است .

□

در واقع، اگر فهرست به ترتیب تصادفی نبود، فرض ترتیب تصادفی، به واریانس برآورد شده‌ای از برآوردها منجر می‌شد که یا بیش از حد کم یا بیش از حد زیاد بودند و در نتیجه بازه‌های اطمینان حاصل می‌توانستند گمراه کننده باشند.

۷.٤ نمونه‌گیری سیستماتیک مکرر

در بخشهای قبل شرح دادیم که واریانسها و خطاهای معیار آماره‌هایی که با فرمولهای تابلوی ۱.٤ برآورد می‌شوند در صورتی که رابطه‌ای بین سطح متغیر خاص و موقعیت آن در چارچوب نمونه‌گیری وجود داشته باشد چگونه می‌توانند در نمونه‌گیری سیستماتیک بیش از حد کم یا بیش از اندازه زیاد باشند. همچنین دیدیم که داده‌های نمونه‌گیری خود نمی‌توانند هیچگونه شناختی از ماهیت هر ترتیب یا دوره‌ای بودن در چارچوب نمونه‌گیری ارائه کنند. ولی این امکان وجود دارد که با رهیافت تعـــــدیل یافته‌ای از نمونه‌گیری سیستماتیک بتوان برآوردهایی از واریانـسهای مجموعها، میانگینها و نسبتهای برآورد شده تهیه کرد که علی‌رغم نوع ترتیب یا دوره‌ای بودن موجود در چارچوبی که نمونه از آن گرفته شده است نااریب باشند. این تعدیل به *نمونه‌گیری سیستماتیک مکرر* موسوم است و در مثال زیر نشان داده شده است.

مثال تشریحی: برای این مثال از داده‌های جدول ۱۷.٤ استفاده می‌کنیم.

فرض کنید می‌خواهیم یک نمونهٔ سیستماتیک متشکل از تقریباً ۱۸ کارگر را از فهرست شامل ۱۶۲ کارگر برای برآورد میانگین تعداد روزهای غیبت از کار به علت بیماری حاد به ازای هر کارگر، انتخاب کنیم. چون $n=۱۸$ و $N=۱۶۲$، یک نمونهٔ سیستماتیک یک نفر از هر ۹ کارگر این منظور را تأمین می‌کند. ولی ۶×۳=۱۸، و بنابراین با انتخاب ۶ نمونهٔ سیستماتیک که هرکدام دارای سه کارگر باشند می‌توانیم نمونه‌ای متشکل از ۱۸ کارگر به دست آوریم. در این مورد، بازهٔ نمونه‌گیری $\frac{۱۶۲}{۳}=۵۴$، و می‌توانیم ۶ نمونهٔ سیستماتیک متشکل از یک نفر از هر ۵۴ کارگر بگیریم. برای انجام این کار ابتدا ۶ عدد تصادفی بین ۱ و ۵۴ انتخاب می‌کنیم (مثلاً ۲، ۳۱، ۴۶، ۱۳، ۳۴، ۵۳) و سپس نمونه‌های سیستماتیک یک نفر از ۵۴ نفر را هر بار با شروع از یکی از اعداد تصادفی انتخاب می‌کنیم. نمونه‌های شش گانهٔ ما در جدول ۱۸.۴ نشان داده شده‌اند.

اگر $\bar{x}_i$ میانگین تعداد روزهای غیبت از کار به علت بیماری حاد از نمونهٔ iام، و m تعداد نمونه‌های انتخاب شده باشد، در آن صورت *میانگین برآورد شدهٔ* ما به شرح زیر خواهد بود:

$$\bar{\bar{x}}=\frac{\sum_{i=۱}^{m}\bar{x}_i}{m}=\frac{۵/۰۰+۴/۳۳+۴/۰۰+۳/۳۳+۷/۰۰+۳/۳۳}{۶}=۴/۵$$

چون میانگین برآورد شده $\bar{\bar{x}}$ با انتخاب یک نمونهٔ تصادفی ساده از $m=۶$ میانگین $\bar{x}_i$ از جامعه‌ای متشکل از $M=۵۴$ میانگین، حاصل شده است می‌توانیم از نظریهٔ نمونه‌گیری تصادفی ساده برای به دست آوردن $\hat{Var}(\bar{\bar{x}})$، برآورد واریانس $\bar{\bar{x}}$ استفاده کنیم:

$$\hat{Var}(\bar{\bar{x}})=\left(\frac{۱}{m}\right)\times\frac{\sum_{i=۱}^{m}(\bar{x}_i-\bar{\bar{x}})^۲}{m-۱}\times\left(\frac{M-m}{M}\right)$$

که در آن، m، تعداد نمونه‌های گرفته شده و M، تعداد کل نمونه‌های سیستماتیک ممکن است. در مورد این مثال داریم :

$$\frac{\sum_{i=۱}^{m}(\bar{x}_i-\bar{\bar{x}})^۲}{m-۱}=\frac{۱}{۵}[(۵-۴/۵)^۲+(۴/۳۳-۴/۵)^۲+(۴-۴/۵)^۲+(۳/۳۳-۴/۵)^۲+$$
$$(۷-۴/۵)^۲+(۳/۳۳-۴/۵)^۲]=۱/۹$$

و

$$\hat{Var}(\bar{\bar{x}})=\left(\frac{۱}{۶}\right)(۱/۹)\left(\frac{۵۴-۶}{۵۴}\right)=۰/۲۸۱۴$$

جدول ۱۷.٤ روزهای غیبت از کار در طول یک سال به علت بیماری حاد در میان ۱٦۲ کارگر یک کارخانه

شمارهٔ شناسایی کارگر	روزهای غیبت	شمارهٔ شناسایی کارگر	روزهای غیبت	شمارهٔ شناسایی کارگر	روزهای غیبت	شمارهٔ شناسایی کارگر	روزهای غیبت
۱	۷	٤۲	۵	۸۳	۳	۱۲٤	۳
۲	٦	٤۳	۳	۸٤	۵	۱۲۵	۹
۳	۱۰	٤٤	٦	۸۵	٤	۱۲٦	۹
٤	۱۱	٤۵	۱۱	۸٦	۰	۱۲۷	٦
۵	۳	٤٦	٦	۸۷	۱۱	۱۲۸	۵
٦	۸	٤۷	۵	۸۸	۳	۱۲۹	٤
۷	۰	٤۸	۵	۸۹	٤	۱۳۰	۱
۸	۵	٤۹	۰	۹۰	۱۱	۱۳۱	۱
۹	۸	۵۰	۸	۹۱	۰	۱۳۲	۱۱
۱۰	٤	۵۱	۱	۹۲	٦	۱۳۳	۳
۱۱	۷	۵۲	۱۰	۹۳	۱	۱۳٤	۵
۱۲	۱۳	۵۳	۷	۹٤	۹	۱۳۵	۹
۱۳	٤	۵٤	۹	۹۵	٦	۱۳٦	۵
۱٤	۵	۵۵	۸	۹٦	۰	۱۳۷	۱
۱۵	۲	۵٦	۲	۹۷	۳	۱۳۸	۱۵
۱٦	۰	۵۷	۹	۹۸	٦	۱۳۹	۲
۱۷	۷	۵۸	۹	۹۹	۰	۱٤۰	۱۰
۱۸	۱۷	۵۹	۸	۱۰۰	۱۲	۱٤۱	۸
۱۹	۵	٦۰	٦	۱۰۱	۱۱	۱٤۲	۲
۲۰	٦	٦۱	۵	۱۰۲	٦	۱٤۳	٦
۲۱	۱	٦۲	۳	۱۰۳	۱	۱٤٤	۱٤
۲۲	۷	٦۳	۹	۱۰٤	۳	۱٤۵	۱۰
۲۳	۹	٦٤	٦	۱۰۵	۲	۱٤٦	۸
۲٤	۳	٦۵	۳	۱۰٦	۵	۱٤۷	۷
۲۵	۸	٦٦	۳	۱۰۷	۳	۱٤۸	۹
۲٦	۹	٦۷	٤	۱۰۸	۱۲	۱٤۹	۱
۲۷	٤	٦۸	۹	۱۰۹	۱	۱۵۰	۲
۲۸	۸	٦۹	۵	۱۱۰	۷	۱۵۱	٦
۲۹	٤	۷۰	۸	۱۱۱	۹	۱۵۲	٤
۳۰	۱۷	۷۱	۵	۱۱۲	٦	۱۵۳	٦
۳۱	٦	۷۲	۱۱	۱۱۳	٦	۱۵٤	۳
۳۲	۹	۷۳	۵	۱۱٤	۳	۱۵۵	۱
۳۳	۹	۷٤	۹	۱۱۵	٤	۱۵٦	۸
۳٤	۵	۷۵	۸	۱۱٦	۲	۱۵۷	۰
۳۵	۸	۷٦	۷	۱۱۷	۵	۱۵۸	۳
۳٦	۵	۷۷	٦	۱۱۸	۱۰	۱۵۹	۲
۳۷	۸	۷۸	٤	۱۱۹	۱۰	۱٦۰	۸
۳۸	۵	۷۹	۳	۱۲۰	۱۵	۱٦۱	۰
۳۹	۸	۸۰	۹	۱۲۱	۵	۱٦۲	۱۵
٤۰	۰	۸۱	۵	۱۲۲	۵		
٤۱	۳	۸۲	۵	۱۲۳	٦		

جدول ۱۸.٤ داده‌های مربوط به شش نمونهٔ سیستماتیک گرفته شده از جدول ۱۷.٤

عدد تصادفی	عناصرنمونه	روزهای غیبت	$\bar{x}$، میانگین برآورد شده
۲	۲	٦	
	٥٦	۲	٥/۰۰
	۱۱۰	۷	
۱۳	۱۳	٤	
	٦۷	٤	٤/۳۳
	۱۲۱	٥	
۳۱	۳۱	٦	
	۸٥	٤	٤/۰۰
	۱۳۹	۲	
۳٤	۳٤	٥	
	۸۸	۳	۳/۳۳
	۱٤۲	۲	
٤٦	٤٦	٦	
	۱۰۰	۱۲	۷/۰۰
	۱٥٤	۳	
٥۳	٥۳	۷	
	۱۰۷	۳	۳/۳۳
	۱٦۱	۰	

بازهٔ اطمینان ۹٥ درصدی عبارت است از:

$$\bar{\bar{x}} - ۱/۹۶\sqrt{\hat{Var}(\bar{\bar{x}})} \le \bar{X} \le \bar{\bar{x}} + ۱/۹۶\sqrt{\hat{Var}(\bar{\bar{x}})}$$

$$۴/۵ - (۱/۹۶)\sqrt{۰/۲۸۱۴} \le \bar{X} \le ۴/۵ + (۱/۹۶)\sqrt{۰/۲۸۱۴}$$

$$۳/۴۶ \le \bar{X} \le ۵/۵۴$$

□

خلاصه‌ای از فرمولهای موردنیاز برای شیوه‌های برآورد کردن میانگین نمونه تحت نمونه‌گیری سیستماتیک مکرر در تابلوی ۳.٤ نشان داده شده است. عباراتی مشابه آنچه را که در تابلوی ۳.٤ درج شده است می‌توان برای مجموعهای جامعه‌ای و نسبتها نیز ساخت. در این قبیل عبارتها $\bar{x}_i$ و $\bar{\bar{x}}$ در معادلهٔ (۹.٤) به ترتیب جای خود را به x'_i و $\bar{x}'$ یا به p_{yi} و $\bar{p}_y$ می‌دهند.

برتری نمونه‌گیری سیستماتیک مکرر نسبت به نمونه‌گیری سیستماتیک آن است که واریانس و خطاهای معیار برآوردها را می‌توان مستقیماً از داده‌ها برآورد کرد. عیب آن این است که فهرست را باید بیش از یک بار مرور کرد، در حالی که در نمونه‌گیری سیستماتیک، نمونه با یک بار مرور در فهرست انتخاب می‌شود. همچنین، در بیشتر موارد دوره‌ای بودن در داده‌ها وجود نخواهد داشت و به همین سبب نمونه‌گیری تصادفی ساده بازه‌ها ی اطمینان مناسب را فراهم خواهد کرد.

تابلوی ۳.٤ شیوه‌های برآورد کردن میانگینهای جامعه‌ای تحت نمونه‌گیری سیستماتیک مکرر

برآورد نقطه‌ای میانگین جامعه

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m} \qquad (۸.٤)$$

برآورد واریانس

$$\hat{Var}(\bar{\bar{x}}) = \left(\frac{1}{m}\right) \times \frac{\sum_{i=1}^{m} (\bar{x}_i - \bar{\bar{x}})^2}{m-1} \times \left(\frac{M-m}{M}\right) \qquad (۹.٤)$$

بازهٔ اطمینان (α-۱)۱۰۰ درصدی

$$\bar{\bar{x}} - z_{(1-\alpha/2)} \sqrt{\hat{Var}(\bar{\bar{x}})} \le \bar{X} \le \bar{\bar{x}} + z_{(1-\alpha/2)} \sqrt{\hat{Var}(\bar{\bar{x}})} \qquad (۱۰.٤)$$

در این معادله‌ها m، تعداد نمونه‌های سیستماتیک گرفته شده، هر یک با اندازهٔ n'، M، تعداد کل نمونه‌های سیستماتیک ممکن که برابر است با $\frac{N}{n'}$ و در آن N اندازهٔ جامعه است، $\bar{x}_i$، میانگین مشاهدات در i امین نمونهٔ سیستماتیک، $\bar{X}$، میانگین جامعه، و $z_{(1-\alpha/2)}$، $[1-\alpha/2]$۱۰۰ صدک توزیع نرمال استاندارد است.

استفاده از STATA برای برآورد کردن در نمونه‌گیری سیستماتیک مکرر: تحلیلی را که در بالا ارائه شده است می‌توان با استفاده از STATA اجرا کرد. همان‌طور که بحث شد، نمونهٔ سیستماتیک مکـــرری که منجـــر به انتخاب ۱۸ کارگر نمونـــهٔ بالا شـــد در واقـــع از یک نمونهٔ تصادفـی ساده از «خوشه‌های» یک جامعهٔ متشکل از ٥٤ خوشه به دست آمد که هر خوشه شامل ۳ کارگر بود که توسط ٥٤ کارگر دیگر در فهرست از یکدیگر جدا شده بودند. پروندهٔ داده‌های STATA یعنـی *workloss.dta*. شامل ٦ گزارش است یعنی یک گزارش برای هر خوشه، و محتوی داده‌های زیر است:

cluster	xi	wt1	xibar	M
2	15	9	5	54
13	13	9	4.33	54
31	12	9	4.00	54
34	10	9	3.33	54
46	21	9	7	54
53	10	9	3.33	54

متغیر *cluster* عددی است از ۱ تا ٥٤ که موضع اولین عنصر در خوشهٔ نمونهٔ خاص را در روی چارچوب نمونه‌گیری مشخص می‌کند.

متغیر *xi* عبارت است از x_i مجموع روزهای غیبت از کار برای سه کارگر خوشهٔ i ام.

متغیر *wt1*، نسبت کل خوشه‌ها، M، به خوشه‌های نمونه، m، و برابر است با $\frac{۵۴}{۶} = ۹$.

متغیر *xibar*، میانگین تعداد روزهای غیبت از کار، $\bar{x}_i$، برای سه کارگر خوشهٔ i ام است.

متغیر M، کل تعداد خوشه‌های متشکل از سه کارگر در جامعهٔ ۱٦۲ نفری کارگران است که برابر است با $\frac{۱۶۲}{۳} = ۵۴$.

برای برآورد کردن میانگین تعداد روزهای غیبت از کار و خطای معیار میانگین برآورد شده، از فرمانهای زیر در STATA استفاده می‌شود :

```
. use "A:workloss.dta", clear
. svyset pweight wt1
. svyset fpc M
. svymean xibar
```

اولین فرمان، پروندهٔ داده‌های STATA را که باید مورد استفاده قرار بگیرد مشخص می‌کند (یعنی داده‌های *workloss.dta* که در درایو A جای دارند).

فرمان دوم نشان می‌دهد که وزن نمونه‌گیری در متغیر *wt1* قرار داده شده است . چون این طرح سیستماتیک مکرر هم‌ارز با یک نمونهٔ تصادفی ساده متشکل از ٦ خوشه از یک جامعهٔ ٥٤ خوشه‌ای است، متغیر *wt1* برای هر یک از خوشه‌های نمونه برابر ۹ خواهد بود.

فرمان سوم نشان می‌دهد که کل تعداد خوشه‌های جامعه (در این مورد برابر با ٥٤) در متغیری قرار دارد که M نامیده شده است. تصحیح جامعهٔ متناهی (fpc) از روی این متغیر محاسبه می‌شود.

فرمان چهارم نشان می‌دهد که میانگین متغیر، *xibar*، قرار است برآورد شود. متغیر *xibar*، میانگین تعداد روزهای غیبت از کار در میان افراد داخل هر خوشه است. میانگین این متغیر، برآورد میانگین تعداد روزهای غیبت از کار به ازای هر فرد است.

خروجی STATA برای این مثال تشریحی در زیر نشان داده شده است :

Survey mean estimation

pweight:	wt1	Number of obs	= 6
Strata:	<one>	Number of strata	= 1
PSU:	<observations>	Number of PSUs	= 6
FPC:	M	Population size	= 54

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
xibar	4.5	.5305483	3.136182	5.863818	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

برآورد مجموع جامعـه و برآورد خطـای معیـار آن را می‌توان با فرمان اضافی svytotal xi به دست آورد که خروجی زیر را تولید می‌کند :

Survey mean estimation

pweight:	wt1	Number of obs	= 6
Strata:	<one>	Number of strata	= 1
PSU:	<observations>	Number of PSUs	= 6
FPC:	M	Population size	= 54

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
xibar	729	85.94882	508.0615	949.9385	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

استفاده از SUDAAN برای برآورد کردن در نمونه‌گیری سیستماتیک مکرر :

SUDAAN (که ترکیب سرواژه‌های *Survey data analysis* است) بسته‌ای نرم‌افزاری است که ابتدا در دههٔ ۱۹۷۰ در مؤسسهٔ تحقیقاتی Research Triangle Institute با نام دیگری تهیه شد. برخلاف STATA که یک نرم‌افزار آماری برای مقاصد کلی است، SUDAAN صرفاً برای تحلیل داده‌های حاصل از آمارگیریهای نمونه‌ای پیچیده تهیه شده است. یک ویرایش از SUDAAN برای PC DOS از اواسط دههٔ ۱۹۸۰ در اختیار قرار گرفته که کاربرد وسیعی در میان پژوهشگران حرفه‌ای آمارگیریها پیدا کرده است. در حال حاضر نظر ما این است که یادگیری و استفاده از SUDAAN سخت‌تر از فرمانهـای STATA است ولی می‌تواند انواع بیشتری از طرحهای نمونه‌گیری و روشهـای برآورد را

رو به راه کند. احساس می‌کنیم که هر دو، ابزارهای مهمی‌اند و همان طور که قبلاً اشاره شد در سراسر این کتاب فرمانها و خروجیهای هر دو نرم‌افزار STATA و SUDAAN را نشان خواهیم داد. صورتهای خاص دیگری از SUDAAN که مورد استفادهٔ ما در این کتاب بوده‌اند عبارت‌اند از DOS Version 6.40 و WINDOWS Version 7.50 (که در جریان بازنگری ما در دسترس قرار گرفت). ارائهٔ آنها در این کتاب برای هر دو صورت مناسب است.

در مثال تشریحی نمونه‌گیری سیستماتیک مکرر که در بالا نشان داده شد و در آن از STATA برای محاسبات استفاده کردیم، پروندهٔ داده‌ها از یک گزارش برای هر خوشهٔ نمونه (در مجموع ٦ گزارش) تشکیل می‌شد، و متغیرهای موردنظر، مجموعه‌هایی برای هر یک از خوشه‌های نمونه بودند. برعکس، SUDAAN می‌تواند داده‌ها را برای نمونهٔ سیستماتیک مکرر با استفاده از یک پروندهٔ داده‌ها شامل گزارشهای نمونهٔ اصلی تحلیل کند. به عبارت دیگر، نیازی نیست که یک پروندهٔ داده‌ها شامل داده‌های انباشته ایجاد کرد.

پروندهٔ داده‌های مورد استفادهٔ SUDAAN برای این مثال تشریحی در زیر نشان داده شده است:

record	cluster	element	M	WT1	Xi
1	2	2	54	9	6
2	2	56	54	9	2
3	2	110	54	9	7
4	13	13	54	9	4
5	13	67	54	9	4
6	13	121	54	9	5
7	31	31	54	9	6
8	31	85	54	9	4
9	31	139	54	9	2
10	34	34	54	9	5
11	34	88	54	9	3
12	34	142	54	9	2
13	46	46	54	9	6
14	46	100	54	9	12
15	46	154	54	9	3
16	53	53	54	9	7
17	53	107	54	9	3
18	53	161	54	9	0

توجه کنید که داده‌های مندرج در ۱۸ گزارش نمونه‌ای شامل ٦ متغیر است: شمارهٔ گزارش *record*، شمارهٔ شناسایی خوشه، *cluster*، موضع گزارش در چارچوب نمونه‌گیری، *element*، تعداد خوشه‌ها در چارچوب نمونه‌گیری، *M*، وزن نمونه‌گیری، *WT1*، و تعداد روزهای غیبت از کار در مورد هر فرد نمونه، *Xi*.

برآورد کردن میانگین و کل تعداد روزهای غیبت از کار و خطای معیار آنها را با نرم‌افزار SUDAAN می‌توان با مجموعهٔ فرمانهای زیر انجام داد:

```
1   PROC DESCRIPT DATA=WLOSS2 FILETYPE=SAS DESIGN=WOR
    MEANS  TOTALS;
2   NEST_ONE_CLUSTER;
3   WEIGHT WT1;
4   TOTCNT M_ZERO_;
5   VAR X1;
6   SETENV COLWIDTH=15;
7   SETENV DESWIDTH=3;
```

فرمان اول *PROC DESCRIPT* را فرا می‌خواند که یک مدول SUDAAN است که اساساً برای برآورد کردن میانگینها و مجموعها و خطاهای معیار آنها به کار می‌رود. این فرمان نشان می‌دهد که داده‌ها در پروندهٔ داده‌های SAS هستند که *WLOSS2.SSD* (شامل ۱۸ گزارش نمونه‌ای نشان‌داده شده در بالا) نامـــگذاری شده است و طرح نمونه از نوعی است که SUDAAN آن را به عنوان *WOR* (برای *بدون جایگذاری*) می‌شناسد. این اساساً به آن معناست که مرحلهٔ اول نمونه‌گیری، نمونه‌گیری تصادفی ساده بدون جایگذاری است. بالاخره، این فرمان نشان می‌دهد که میانگینها و مجموعها قرار است برآورد شوند.

فرمان دوم، حکم "nest" است. در این حکم، واژهٔ _ONE_ نشان می‌دهد که طبقه‌بندی وجود ندارد و اصطلاح CLUSTER نشان می‌دهد که واحدهای اولیه یا واحدهای نمونه‌گیری مرحلهٔ اول با متغیر *cluster* شناسایی می‌شوند.

فرمان سوم نشان می‌دهد که وزن نمونه‌گیری در متغیر *WT1* جای داده شده است که در این مورد برابر است با تعداد M خوشه در جامعه تقسیم بر تعداد m خوشه در نمونه یا $\frac{۵۴}{۶} = ۹$.

در فرمان چهارم، اصطلاح *TOTCNT* نشان می‌دهد که کل تعداد خوشه‌های جامعه در متغیر M یافت می‌شود و اصطلاح *_ZERO_* نشانهٔ آن است که پس از نمونه‌گیری مرحلهٔ اول، نمونه‌گیری فرعی بیشتری وجود نخواهد داشت (یعنی همهٔ عناصر در داخل هر خوشهٔ نمونه انتخاب شده‌اند و در این طرح هیچ مؤلفهٔ واریانسی به علت تغییرات بین عناصر در داخل همان خوشه وجود ندارد).

فرمان پنجم نشان می‌دهد که برآورد کردن قرار است برای واریانس *Xi* (که نشانهٔ روزهای غیبت از کار است) اجرا شود .

فرمانهای ششم و هفتم نشانهٔ پهنای ستون وتعداد ارقام اعشاریی است که باید در خروجی ارائه شود.

مجموعهٔ فرمانهایی که در بالا شرح داده شد خروجی SUDAAN را به شرح زیر ارائه خواهد کرد:

Number of observations read : 18 Weighted count : 162
Number of observations skipped : 0
(WEIGHT variable nonpositive)
Denominator degrees of freedom : 5

Date : 07-10-97 Research Triangle Institute Page:1

Time : 14 : 25 : 90 Table:1

by : Variable, One.

Variable		One 1
XI	Sample Size	18.000
	Weighted Size	162.000
	Total	729.000
	SE Total	85.949
	Mean	4.500
	SE Mean	0.531

۸.٤ نمونهٔ موردنیاز چقدر باید بزرگ باشد؟

اگر فرض کنیم فهرستی که نمونهٔ سیستماتیک از آن گرفته می‌شود ترتیب تصادفی داشته باشد، می‌توانیم فرض‌کنیم که وضعیت تقریباً همان وضعیت نمونه‌گیری تصادفی ساده است. در چنین مواردی، می‌توان از همان روشهایی که در فصل ۳ برای تعیین اندازهٔ نمونه شرح داده شد استفاده‌کرد.

ولی اگر فرض‌کنیم که فهرست ترتیب تصادفی دارد، در آن صورت تعیین اندازهٔ نمونه مسئله‌ای بسیار مشکل خواهد شد. علت این امر آن است که واریانس یک برآورد حاصل از نمونه‌گیری سیستماتیک بستگی به بازهٔ نمونه‌گیری دارد. برای مثال، یک نمونهٔ سیستماتیک ۱ از ٤ از فهرستی شامل ۱۰۰ واحد شمارش، نمونه‌ای متشکل از ۲۵ واحد شمارش به دست می‌دهد در حالی که نمونهٔ سیستماتیک ۱ از ۵ از همان فهرست، نمونه‌ای متشکل از ۲۰ واحد شمارش را نتیجه می‌دهد. ولی، دوره‌ای بودن در فهرست ممکن است چنان باشد که برآوردهای حاصل از نمونهٔ اول دارای واریانسی بیشتر از برآوردهای حاصل از نمونهٔ دوم باشند هرچند که واحدهای شمارش در نمونهٔ اول بیشتر از نمونهٔ دوم است. چون معمولاً تا قبل از انتخاب نمونه از مشخصه‌های فهرست اطلاعی نداریم، تعیین یک بازهٔ نمونه‌گیری مناسب و از آن رو تعیین اندازهٔ نمونهٔ مناسب برای ما مشکل خواهد بود .

در نمونه‌گیری سیستماتیک تکراری این امکان وجود دارد که تصوری از اندازه نمونهٔ موردنیاز به دست آید به این ترتیب که یک مجموعهٔ مقدماتی از m' نمونهٔ سیستماتیک تکراری از روی فهرست

انتخاب می‌شود، پارامترها بر اساس نمونهٔ مقدماتی برآورد می‌شوند و m، تعداد نمونه‌های موردنیاز بر مبنای فرمول زیر تعیین می‌شود:

$$m = \frac{z^{۲}_{۱-(\alpha/۲)} \times \frac{\sum_{i=۱}^{m'} \frac{(\bar{x}_i - \bar{\bar{x}})^{۲}}{(m'-۱)}}{\bar{\bar{x}}^{۲}} \times \left(\frac{N}{n}\right)}{\left(\frac{N}{n}-۱\right) \times \varepsilon^{۲} + z^{۲}_{۱-(\alpha/۲)} \times \frac{\sum_{i=۱}^{m'} \frac{(\bar{x}_i - \bar{\bar{x}})^{۲}}{(m'-۱)}}{\bar{\bar{x}}^{۲}}} \qquad (۱۱.٤)$$

در این فرمول ε، $z_{۱-(\alpha/۲)}$، N و n همانند آنهایی هستند که قبلاً تعریف شده‌اند.

مثال تشریحی: فرض کنید می‌خواهیم از فهرست کارگران که در جدول ۱۷.٤ نشان داده شده است برای مقاصد برآورد میانگین تعداد روزهای غیبت از کار به علت بیماری نمونه بگیریم. برای این منظور نمونه‌های سیستماتیک تکراری یک نفر از هر ۸۱ کارگر را انتخاب خواهیم کرد. باز هم فرض‌کنید که می‌خواهیم واقعاً مطمئن شویم که میانگین تعداد روزهای غیبت از کار را با حداکثر تفاوت ۲۰ درصد از مقدار واقعی برآورد کنیم و می‌خواهیم تعیین‌کنیم که به چند نمونهٔ سیستماتیک m یک نفر از هر ۸۱ کارگر نیاز داریم. یک نمونهٔ مقدماتی متشکل از ٦ نمونه از این نوع به منظور برآورد m انتخاب می‌کنیم. ابتدا شش عدد تصادفی را بین ۱ و ۸۱ انتخاب می‌کنیم (مثلاً ۲۲، ٤۸، ۲۷، ٦۱، ۵۳، ۱۰) و شش نمونهٔ مندرج در جدول ۱۹.٤ را به دست می‌آوریم.

جدول ۱۹.٤ شش نمونهٔ گرفته شده از جدول ۱۷.٤

عدد تصادفی	کارگران نمونه	$\bar{x}_i$، میانگین روزهای غیبت برآورد شده
۱۰	۹۱,۱۰	۲/۰
۲۲	۱۰۳,۲۲	٤/۰
۲۷	۱۰۸ ,۲۷	۸/۰
٤۸	۱۲۹,٤۸	٤/۵
۵۳	۱۳٤,۵۳	٦/۰
٦۱	۱٤۲ , ٦۱	۳/۵

پس محاسبات زیر را خواهیم داشت:

$$\bar{\bar{x}} = \frac{۲+۴+۸+۴/۵+ ۶+۳/۵}{۶} = ۴/۶۷$$

$$\frac{\sum_{i=1}^{m'}(\bar{x}_i-\bar{\bar{x}})^2}{m'-1}=\left(\frac{1}{6-1}\right)[(2-4/67)^2+(4-4/67)^2+(8-4/67)^2+(4/5-4/67)^2$$
$$+(6-4/67)^2+(3/5-4/67)^2]=4/367$$

$$\varepsilon=0/2 \qquad N=162 \qquad n=2 \qquad \frac{N}{n}=81$$

از رابطهٔ ۱۱.٤ داریم:

$$m=\frac{9\times\left[\frac{4/367}{(4/67)^2}\right]\times 81}{(81-1)\times(0/2)^2+9\times\left[\frac{4/367}{(4/67)^2}\right]}=29/18\approx 30$$

به این ترتیب برای تأمین خواسته‌های مسئله تقریباً به ۳۰ نمونهٔ یک نفر از هر ۸۱ کارگر از روی فهرست نیاز خواهیم داشت.

□

فرمول مربوط به اندازهٔ نمونهٔ موردنیاز در نمونه‌گیری سیستماتیک تکراری بر این واقعیت استوار است که نمونه‌گیری سیستماتیک تکراری ۱ عنصر از هر $\frac{N}{n}$ عنصر فهرست هم‌ارز با نمونه‌گیری تصادفی ساده از جامعه‌ای با $\frac{N}{n}$ عنصر است که در آن برآوردهای به دست آمده از هر $\frac{N}{n}$ نمونه به عنوان متغیرهای اصلی در نظر گرفته می‌شوند.

۹.٤ استفاده از چارچوبهایی که فهرست نیستند

تا اینجا نمونه‌گیری سیستماتیک را منحصراً از دیدگاه انتخاب یک نمونهٔ ۱ از k، از فهرستی از عناصر، مورد بحث قرار دادیم. این فهرست ممکن است از قبل وجود داشته باشد یا در طی فرایند نمونه‌گیری ایجاد شود. مثالی از نوع اخیر فهرست، وضعیتی است که طی آن ممکن است یکی از هر پنج بیماری را که وارد اتاق فوریتهای پزشکی بیمارستان می‌شوند برای نمونه انتخاب کنیم. در این حالت هیچ فهرستی از بیماران نمی‌تواند از قبل موجود باشد. یکی از مزایای نمونه‌گیری سیستماتیک نسبت به سایر طرحهای نمونه‌گیری آن است که نمونه‌گیری را می‌توان ضمن ساخت چارچوب انجام داد.

نمونه‌گیری سیستماتیک را می‌توان در مواقعی که هیچ فهرستی در اختیار نیست نیز انجام داد. برای مثال، اگر مجموعه‌ای از پرونده‌ها در ۱۲ کشوی بایگانی باشند و هر کشو ۲٦ اینچ عمق داشته باشد و نمونه‌ای متشکل از ۱۰۰ سابقه موردنیاز باشد، می‌توانیم با روش زیر، یک نمونهٔ سیستماتیک بگیریم:

۱. کل طول پرونده‌های بایگانی شده را محاسبه می‌کنیم (مثلاً اینچ ۳۱۲= ۲٦×۱۲).

۲. کل طول پرونده‌ها را به کل تعداد پرونده‌هایی که باید نمونه‌گیری شوند تقسیم می‌کنیم و نسبت را با حرف k نشان می‌دهیم (مثلاً اینچ $k = \frac{۳۱۲}{۱۰۰} = ۳/۱۲$).

۳. یک عدد تصادفی j از ۱ تا k انتخاب می‌کنیم (مثلاً بین ۰/۰۰ و ۳/۱۲. فرض کنید شمارهٔ انتخابی ۱/۱۹ است).

٤. با استفاده از نوعی ابزار اندازه‌گیری، پرونده‌ای را که j واحد از جلوی اولین کشوی بایگانی فاصله دارد برمی‌داریم. سپس پرونده‌ای را که با طول k واحد از پروندهٔ اول فاصله دارد برمی‌داریم (شاید لازم باشد برای انتخاب پرونده به کشوی بعدی برویم). این فرایند را ادامه می‌دهیم تا به انتهای آخرین کشوی بایگانی برسیم.

(ما در مثال خود ابتدا پرونده‌ای را برمی‌داریم که ۱/۱۹ اینچ از جلوی کشوی اول فاصله دارد. سپس پرونده‌ای را برمی‌داریم که ۳/۱۲ اینچ از پروندهٔ اول فاصله دارد و این کار را با برداشتن هر پرونده در فاصله ۳/۱۲ اینچی پروندهٔ قبلی ادامه می‌دهیم تا به انتهای آخرین کشو برسیم).

مورد دیگری که طی آن می‌توان نمونه‌گیری سیستماتیک را بدون وجود فهرست به کار برد در نمونه‌گیری نواحی جغرافیایی از روی نقشه است. مثلاً، فرض کنید می‌خواهیم متوسط تراکم فسفات را در رودخانه‌ای که ۲۵۰ مایل طول دارد با گرفتن نمونه‌ای متشکل از ۱۰۰ واحد نمونه از رودخانه برآورد کنیم. یک روش ساده آن است که طول رودخانه (۲۵۰ مایل) را به تعداد واحدهای نمونهٔ موردنیاز (۱۰۰) تقسیم کنیم تا بازهٔ نمونه‌گیری (۲/۵ مایل) به دست آید. سپس می‌توانیم یک عدد تصادفی بین صفر و ۲/۵ (مثلاً ۲/۱) انتخاب و اولین نقطهٔ نمونه‌گیری را در ۲/۱ مایلی مبدأ رودخانه مکان‌یابی کنیم.(یک عدد تصادفی دوم نیز مطابق با هر نقطه می‌توان انتخاب کرد که نشان دهندهٔ فاصلهٔ نقطهٔ انتخاب واحد نمونه تا ساحل رودخانه باشد). محل واحد دوم ٤/٦ مایل دورتر از مبدأ رودخانه خواهد بود (۲/۱+۲/۵). نقطهٔ سوم در فاصلهٔ ۷/۱ مایلی از مبدأ رودخانه قرار خواهد داشت و همین طور الی آخر. این شیوه را می‌توان به آسانی با استفاده از یک نقشهٔ خوب انجام داد. یک مزیت به طور شهودی جالب توجه نمونه‌گیری سیستماتیک در مورد این مثال آن است که به واحدهای نمونه‌ایی منجر می‌شود که از سراسر طول رودخانه انتخاب می‌شوند به نحوی که مشکل است نقاطی که تراکم مقدار فسفات در آنها به طور غیرمعمول زیاد (یا کم) است نادیده بمانند.

۱۰.٤ خلاصه

در این فصل، نمونه‌گیری سیستماتیک را مورد بحث قرار دادیم که متداولترین شیوه از شیوه‌های مورد استفاده در نمونه‌گیری است. نمونه‌گیری سیستماتیک برخلاف بیشتر شیوه‌های نمونه‌گیری، مستلزم

دانستن کل تعداد واحدهای نمونه‌گیری جامعه نیست و به همین دلیل نمونه‌گیری را می‌توان همزمان با ایجاد چارچوب نمونه‌گیری اجرا کرد.

از سه روش نمونه‌گیری سیستماتیک بحث شد. یک روش تنها هنگامی به برآوردهای نااریب منجر می‌شود که نسبت $\frac{N}{k}$ یعنی نسبت تعداد عناصر جامعه به بازهٔ نمونه‌گیری یک عدد صحیح باشد. روش دوم همیشه به برآوردهای نااریب منتهی می‌شود ولی استفاده از آن محدود است زیرا مستلزم آن است که N، تعداد عناصر جامعه از قبل معلوم باشد. روش سوم که نمونه‌گیری سیستماتیک تکراری است این امکان را فراهم می‌سازد که بتوان واریانسهای برآوردها و در نتیجه، بازه‌های اطمینان را به دست آورد.

روش‌شناسی برای به دست آوردن اندازهٔ نمونهٔ موردنیاز شرح داده شد و موارد مربوط به استفاده از نمونه‌گیری سیستماتیک، وقتی فهرستهایی در اختیار نیستند توضیح داده شد.

تمرین

۱.٤ فرض‌کنید شورای پیشگیری از مسمومیت سرب در کودکان[1] در یک منطقهٔ کلان شهر در غرب تنسی[2] مسئولیت تعیین نسبت خانه‌های بدون ایمنی از لحاظ سطح سرب را در یک منطقهٔ نوساز ویژه با ۱۲۰ خانه به عهده گرفته است. به دلیل هزینهٔ بسیار سنگین اجرای آزمایشهای طیف‌سنجی از دیوارهای داخلی، سقفها، کف اتاقها، پاچینها، کابینتها و سایر خطرات آشکار سرب از قبیل میله‌های تخت بچه و نیز روکشهای بیرون ساختمان، ایوانها و نرده‌های ایوانها، تصمیم گرفته می‌شود که نمونه‌ای از خانه‌های تحت بررسی انتخاب شود. برای مقاصد نمونه‌گیری، یک چارچوب روزآمد خوب موجود است. این چارچوب، یک فرم فهرست‌برداری از خیابانهاست که شامل آدرس و مشخصات مالک هر خانه در خیابانهای ناحیهٔ موردنظر است. تصمیم بر این است که یک خانه از هر ۳ خانه برای نمونه انتخاب شود. فرض می‌کنیم که تنها خانه‌هایی که واقعاً دارای مشکلات جدی مربوط به خطر سرب هستند خانه‌های با شمارهٔ ۲٦، ۲۷، ۲۸ و ۲۹ در روی فهرست باشند.

الف. فرض کنید عدد تصادفی ۲ برای شروع دنباله انتخاب شود. نسبت خانه‌های دارای خطر سرب را از روی نمونه برآورد کنید.

[1] Childhood Lead Poisoning Prevention Council
[2] Tennessee

ب. یک بازهٔ اطمینان ۹۵ درصدی برای نسبت خانه‌های دارای خطر سرب به دست آورید. فرضهای شما چه بوده است؟

پ. واریانس واقعی توزیع نسبت برآورد شدهٔ خانه‌های دارای خطر سرب چقدر است؟ این واریانس با واریانس برآورد شده در قسمت (ب) چگونه مقایسه می‌شود؟

ت. فرض کنید به جای این کار، ٤٠ خانه به صورت نمونهٔ تصادفی ساده انتخاب شود. واریانس توزیع نسبت برآورد شدهٔ خانه‌های دارای خطر سرب در این مورد چقدر است؟ این مقدار با واریانس حاصل از نمونهٔ سیستماتیک ۱ از ۳ چگونه مقایسه می‌شود؟

۲.٤ فرض کنید از ۱۲۰ خانهٔ تمرین ۱، یک نمونهٔ سیستماتیک ۱ از ۵ انتخاب شود، و فرض کنید که عدد تصادفی آغازین ۵ است.

الف. نسبت خانه‌های دارای خطر سرب را از روی این نمونه برآورد کنید.

ب. یک بازهٔ اطمینان ۹۵ درصدی برای نسبت خانه‌های دارای خطر سرب به دست آورید.

پ. واریانس واقعی توزیع نسبت برآورد شدهٔ خانه‌های دارای خطر سرب چقدر است؟ این نتیجه را با واریانس برآورد شده که در قسمت (ب) به کار بردید مقایسه کنید.

ت. فرض کنید به جای یک نمونهٔ سیستماتیک ۱ از ۵ یک نمونهٔ تصادفی ساده با همین تعداد از خانه‌ها گرفته شود. واریانس توزیع نمونه‌گیری نسبت برآورد شدهٔ خانه‌های دارای خطر سرب از روی این طرح نمونه‌گیری چقدر است؟ این مقدار با مقدار به دست آمده از نمونهٔ سیستماتیک ۱ از ۵ چگونه مقایسه می‌شود؟

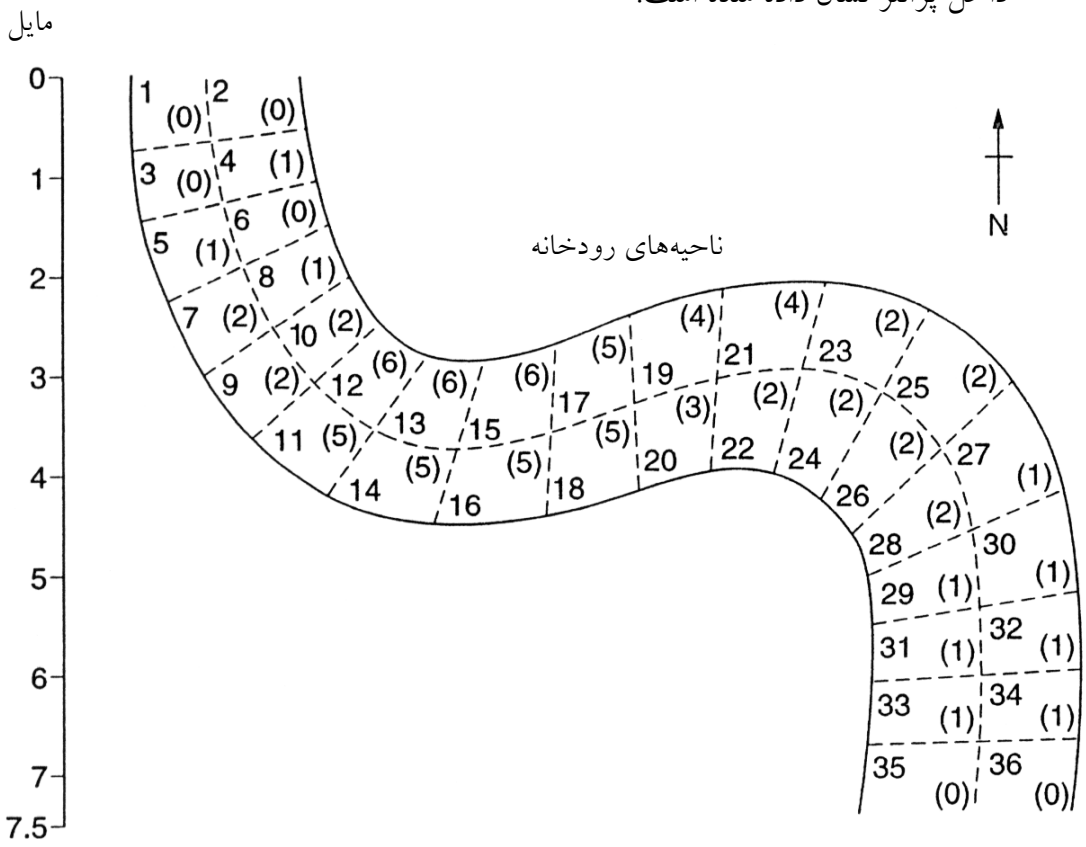
۳.٤ دوباره به تمرین ۱.٤ باز می‌گردیم. فرض کنید به جای یک نمونهٔ سیستماتیک ۱ از ۵، در مجموع ٢٤ خانه، نمونه‌ای از طریق نمونه‌گیری سیستماتیک مکرر ۱ از ٤٠ خانه به دست آمده است.

الف. فرض کنید اعداد تصادفی انتخاب شده ۳، ۷، ۱۲، ٢٦، ۳۱، ۳۳، ۳۸ و ٤٠ باشند. نسبت خانه‌های دارای خطر سرب را برآورد کنید و یک بازهٔ اطمینان برای نسبت برآورد شده به دست آورید.

ب. واریانس برآورد شدهٔ توزیع نسبت برآورد شدهٔ خانه‌های دارای خطر سرب که از این هشت نمونهٔ سیستماتیک یکی از هر ٤٠ خانوار به دست می‌آید چقدر است؟ این مقدار را با واریانس واقعی که از نمونهٔ سیستماتیک ۱ از ۵ به دست می‌آید و در قسمت (پ) تمرین ۲.٤ محاسبه کرده‌اید مقایسه کنید.

٤.٤ از فهرست ١٦٢ کارگر در جدول ١٧.٤، با استفاده از نمونه‌گیری سیستماتیک مکرر جمعاً ١٨ کارگر نمونه را برای برآورد کل تعداد روزهای غیبت از کار به علت بیماری سخت همهٔ کارگران انتخاب کنید و نسبت کارگرانی را که هشت روز یا بیشتر به علت بیماری شدید غیبت داشته‌اند برآورد نمایید. برای هر یک از این برآوردها بازهٔ ۹۵ درصدی اطمینان را به دست آورید (فرض کنید اعداد تصادفی ٤، ٤٤، ٢٩، ٢٠، ٢٧ و ٥ انتخاب شده باشند).

٥.٤ فرض کنید مطالعه‌ای در مورد میزان سم ضدآفات دی‌الدرین که تصور می‌شود سرطان‌زا باشد در ۷/۵ مایل از طول رودخانهٔ خاصی در دست انجام است. برای تضمین نماینده بودن نمونه‌ها، نقشه‌ای از رودخانه را به ٣٦ ناحیه تقسیم می‌کنیم (شکل زیر را ببینید) و یک نمونهٔ سیستماتیک ١ از ٤ از این ناحیه‌ها انتخاب می‌کنیم. برای گرفتن نمونه‌های آب با قایق به مرکز جغرافیایی ناحیهٔ تعیین شده می‌رویم و از عمق چندین سانتیمتری پایینتر از سطح آب مقداری آب برمی‌داریم. مقدار دی‌الدرین برحسب میکروگرم در لیتر، برای هر یک از این نواحی در داخل پرانتز نشان داده شده است.



الف. بازهٔ ۹۵ درصدی اطمینان را برای متوسط سطح دی‌الدرین در این بخش از رودخانه حساب کنید.

ب. برای این روش نمونه‌گیری از رودخانه چه برتریهایی را نسبت به نمونه‌گیری تصادفی ساده می‌توانید مشخص کنید؟

٦.٤ در طول یک سال خاص ۲۰۰ مورد عمل سوند قلبی روی اشخاص ۷۰ ساله به بالا در یک بیمارستان دانشگاهی بزرگ انجام شده است. از روی فهرستی از این بیماران ٤ نمونهٔ سیستماتیک یک بیمار از هر ۵۰ بیمار نتایج زیر را در رابطه با فشار سرخرگ ریوی نتیجه داده است:

نمونه	شمارهٔ فرد	میانگین فشار ریوی سرخرگ
۱	۳۰	١٦
	۸۰	۲۵
	۱۳۰	۱۵
	۱۸۰	۱۷
۲	۱۷	١٤
	٦٧	۱۹
	۱۱۷	۱۸
	١٦٧	۲۰
۳	۲۲	۱۹
	۷۲	۲۷
	۱۲۲	۳۰
	۱۷۲	۱۱
٤	٤۳	۱۰
	۹۳	۳۳
	١٤۳	۱۷
	۱۹۳	۱۳

براساس این داده‌های اولیه، برای برآورد میانگین فشار سرخرگ ریوی در این جامعه حول ۱۰ درصد از مقدار واقعی، چند نمونهٔ سیستماتیک مکرر متشکل از یک نفر از هر ۵۰ نفر باید انتخاب شود؟

۷.٤ جدول زیر، برنامهٔ ویزیتهای هفتگی یک دندانپزشک را در یک مرکز بهداشت محلی کوچک برحسب فعالیت اصلی نشان می‌دهد (B= مراقبت مقدماتی، C= مراقبت کامل):

هفته ۱	دوشنبه	سه‌شنبه	چهارشنبه	پنج‌شنبه	جمعه
۹/۰۰	B	B	B	B	B
۱۰/۰۰	B	C	C	C	B
۱۱/۰۰	B	C	B	C	B
۱/۰۰	C	B	C	C	C
۲/۰۰	B	C	B	C	C
۳/۰۰	B	B	B	B	B
٤/۰۰	B	B	B	B	B

مطلوب است برآورد نسبت ویزیتهای اختصاص یافته به مراقبت کامل طی یک دورهٔ ۵۲ هفته‌ای خاص با استفاده از برنامهٔ نمونه‌گیری سیستماتیک. یک نمونهٔ سیستماتیک یکی از هـــر۷ مورد پیشنهاد می‌شود. آیا این برنامهٔ نمونه‌گیـری برای این شـــرایط مناسب است؟ چرا آری؟ چرا نه؟

۸.٤ در وضعیتی که برای تمرین ۷.٤ توصیف شد، طی دوره‌ای ۵۲ هفته‌ای، ۱۸۲۰ (۵۲×۳۵) وقت ملاقات وجود دارد. قرار است برای برآورد نسبت ملاقاتهای اختصاص یافته به مراقبت کامل در طی ۵۲ هفته، از طرح نمونه‌گیری سیستماتیک مکرر یکی از هر ۲٦ قرار ملاقات استفاده شود. **در** یک بررسی مقدماتی با استفاده از سه نمونهٔ متشکل از یکی از هر ۲٦ قرار ملاقات، نتایج زیر به دست آمده‌اند:

نمونه	نسبت مراقبت کامل
۱	$\frac{۱۲}{۷۰}$
۲	$\frac{۲۳}{۷۰}$
۳	$\frac{۱۸}{۷۰}$

براساس نتایج بالا چند نمونهٔ تکراری متشکل از یکی از هر ۲٦ قرار ملاقات موردنیاز است تا این نسبت با ۹۵ درصد اطمینان حول ۱۰ درصد مقدار واقعی آن برآورد شود؟

۹.٤ کدام یک از احکام زیر دربارهٔ نمونه‌گیری سیستماتیک درست *نیست*؟

الف. وقتی نسبت نمونه‌گیری با دوره‌ای بودن در چارچوب منطبق شود واریانسهای برآوردها زیاد خواهد بود.

ب. واریانسهای برآوردها به اندازهٔ ضریب همبستگی درون ـ رده‌ای وابسته‌اند.

پ. برخلاف نمونه‌گیری تصادفی ساده، برای اجرای نمونه‌گیری سیستماتیک نیازی به دانستن تعداد عناصر جامعه نیست.

ت. در نمونه‌گیری سیستماتیک، میانگینها، مجموعها و نسبتهای برآورد شده همیشه نااریب‌اند.

۱۰.٤ نمونه‌گیری سیستماتیک در کدام یک از وضعیتهای زیر بهترین کارکرد را دارد؟

الف. چارچوب (یا فهرست) نمونه‌گیری نسبت به متغیری مرتب شده باشد که مستقیماً با متغیر موردنظر وابسته بوده و با آن همبستگی زیاد داشته باشد.

ب. چارچوب نمونه‌گیری نه مرتب شده است و نه با توجه به متغیر موردنظر دوره‌ای است.

پ. چارچوب نمونه‌گیری دوره‌ای بودنی دارد که با کسر نمونه‌گیری همنهشت است.

ت. چارچوب نمونه‌گیری، یک فهرست نیست.

۱۱.٤ فرض کنید که در وضعیت تمرین ۱.٤ تصمیم بر این است که ۱۸ خانهٔ نمونه براساس طرح نمونه‌گیری سیستماتیک مکرر انتخاب شود، که ٦ تکرار از یک نمونهٔ سیستماتیک سه‌خانه‌ای را مشخص می‌کند.

الف. از روی نمادهایی که در این فصل برای نمونه‌گیری سیستماتیک مکرر تعیین شده است N، M، n و m را مشخص کنید. مقدار عددی وزن نمونه‌گیری چقدر است؟

ب. فرض کنید شش عدد تصادفی که برای طرح نمونه‌گیری قسمت الف انتخاب شده‌اند ۳، ۱۵، ۲۰، ۲۷، ۳٤ و ۳۹ باشند. چه خانوارهایی در نمونه ظاهر می‌شوند؟

پ. (برای کسانی که به SUDAAN یا STATA دسترسی دارند) از روی نمونه‌ای که در قسمت ب گرفته شده است SUDAAN یا STATA را برای برآورد کردن تعداد و نسبت خانه‌های دارای خطر سرب به کار برید و خطاهای معیار این برآوردها را برآورد کنید.

۱۲.٤ یک پروندهٔ داده‌ها دارای ۱۰۰۰۰۰ گزارش در مورد ۲۰۱۲ نفر از اعضای یک سازمان حفظ بهداشت (HMO) است. گزارشها از «۱» تا «۱۰۰۰۰۰» شماره‌گذاری شده‌اند، و یک عدد تصادفی بین ۱ و ۱۰۰۰ انتخاب شده است. عدد تصادفی انتخاب شده ۲۵۳ است و همهٔ

گزارشهایی که شمارهٔ آنها به ۲۵۳ ختم می‌شود به عنوان واحدهای نمـونه انتخاب شده‌اند (برای مثال، ۲۵۳، ۱۲۵۳، ۲۲۵۳، ۳۲۵۳ و الخ). برای هر یک از گزارشهای انتخاب شده شخصی که گزارش نمونه متعلق به اوست در مورد عوامل مخاطره برای بیماری کرونر قلب مورد سؤال قرار می‌گیرد. سپس با استفاده از نسبت نمونه (تعداد اشخاص در معرض مخاطرهٔ زیاد تقسیم بر تعداد اشخاص در نمونه) برآوردی از نسبت اشخاصی که در معرض مخاطرهٔ زیاد برای ابتلا به بیماری کرونر قلبی هستند تهیه می‌شود.

الف. نسبت نمونه که در بالا توصیف شد یک برآورد اریب از نسبت کل جامعه است. نشان دهید چرا چنین چیزی درست است.

ب. یک برآورد نااریب برای این طرح نمونه تهیه کنید.

۱۳.٤ نمونه‌ای که در تمرین ۱۲.٤ توصیف شد در مورد ۱۰۰ نفر از ثبت‌نام شدگان سازمان حفظ بهداشت که به روش بالا انتخاب شده‌اند داده‌های زیر را به دست داده است. از روی این داده‌ها، نسبت اشخاصی را که در جامعهٔ سازمان حفظ بهداشت در معرض مخاطرهٔ زیاد برای ابتلا به بیماری کرونر قلبی هستند برآورد کنید.

شمارهٔ گزارش	تعداد گزارشها برای هر عضو سازمان	مخاطرهٔ زیاد برای ابتلا به بیماری کرونر قلبی	شمارهٔ گزارش	تعداد گزارشها برای هر عضو سازمان	مخاطرهٔ زیاد برای ابتلا به بیماری کرونر قلبی
۲۵۳	۵۰	۰	۵۰۲۵۳	۵۰	۰
۱۲۵۳	۳٦	۰	۵۱۲۵۳	٤۷	۰
۲۲۵۳	٤۵	۰	۵۲۲۵۳	۵۵	۰
۳۲۵۳	٦۱	۰	۵۳۲۵۳	۵٦	۰
٤۲۵۳	۵۰	۰	۵٤۲۵۳	٤٤	۰
۵۲۵۳	۷۷	۰	۵۵۲۵۳	۳٦	۰
٦۲۵۳	۷۰	۰	۵٦۲۵۳	٤٦	۰
۷۲۵۳	٤۱	۰	۵۷۲۵۳	٦٤	۰
۸۲۵۳	٦۳	۰	۵۸۲۵۳	۵۹	۰
۹۲۵۳	٤۳	۰	۵۹۲۵۳	۵۵	۰
۱۰۲۵۳	۵۳	۰	٦۰۲۵۳	۱۰	۱
۱۱۲۵۳	۱۷	۱	٦۱۲۵۳	۱۲	۱
۱۲۲۵۳	٤٤	۰	٦۲۲۵۳	٦۲	۰
۱۳۲۵۳	۵٤	۰	٦۳۲۵۳	۳٤	۰
۱٤۲۵۳	۷۰	۰	٦٤۲۵۳	٤۸	۰

١٥٢٥٣	٤٦	٠	٦٥٢٥٣	٢٧	١
١٦٢٥٣	٦٢	٠	٦٦٢٥٣	٤٣	٠
١٧٢٥٣	٤٨	٠	٦٧٢٥٣	٢٣	١
١٨٢٥٣	٣٧	٠	٦٨٢٥٣	٣٦	٠
١٩٢٥٣	٥٣	٠	٦٩٢٥٣	٥٢	٠
٢٠٢٥٣	٣٨	٠	٧٠٢٥٣	٥٥	٠
٢١٢٥٣	٣٩	٠	٧١٢٥٣	١٩	١
٢٢٢٥٣	٥٣	٠	٧٢٢٥٣	٣٥	٠
٢٣٢٥٣	٢٥	١	٧٣٢٥٣	١٩	١
٢٤٢٥٣	٥٤	٠	٧٤٢٥٣	٥٠	٠
٢٥٢٥٣	٣٩	٠	٧٥٢٥٣	٤٠	٠
٢٦٢٥٣	٤١	٠	٧٦٢٥٣	٣٧	٠
٢٧٢٥٣	٥٦	٠	٧٧٢٥٣	٧٠	٠
٢٨٢٥٣	٢٤	١	٧٨٢٥٣	٥٦	٠
٢٩٢٥٣	٥٩	٠	٧٩٢٥٣	٣٣	٠
٣٠٢٥٣	٦٨	٠	٨٠٢٥٣	٢٨	١
٣١٢٥٣	٦٥	٠	٨١٢٥٣	٣٦	٠
٣٢٢٥٣	٦٣	٠	٨٢٢٥٣	٥٤	٠
٣٣٢٥٣	٦٣	٠	٨٣٢٥٣	٤٨	٠
٣٤٢٥٣	٧٠	٠	٨٤٢٥٣	٥٧	٠
٣٥٢٥٣	٤٥	٠	٨٥٢٥٣	٥٥	٠
٣٦٢٥٣	١٨	١	٨٦٢٥٣	٤١	٠
٣٧٢٥٣	٢١	١	٨٧٢٥٣	٥٠	٠
٣٨٢٥٣	٤٦	٠	٨٨٢٥٣	٦٤	٠
٣٩٢٥٣	٤٨	٠	٨٩٢٥٣	٥٤	٠
٤٠٢٥٣	٥٧	٠	٩٠٢٥٣	٣٣	٠
٤١٢٥٣	٤٠	٠	٩١٢٥٣	٦٣	٠
٤٢٢٥٣	٥٧	٠	٩٢٢٥٣	٥١	٠
٤٣٢٥٣	٢٧	١	٩٣٢٥٣	٥٥	٠
٤٤٢٥٣	٤٨	٠	٩٤٢٥٣	٣٦	٠
٤٥٢٥٣	٥١	٠	٩٥٢٥٣	٣٦	٠
٤٦٢٥٣	٥٩	٠	٩٦٢٥٣	٥١	٠
٤٧٢٥٣	٤٨	٠	٩٧٢٥٣	٣١	٠
٤٨٢٥٣	٧٤	٠	٩٨٢٥٣	٢٨	١
٤٩٢٥٣	٤٦	٠	٩٩٢٥٣	٣٥	٠

کتابشناسی

The sampling texts cited in Chapters l and 2 all develop the concepts of systematic sampling discussed in this chapter, each in its own way. As *had been stated earlier, systematic sampling is widely used in practice since it is considerably easier to sample systematically from a list than to take a simple random sample.*
The module, SAMPLE, written for the statistical package, SYSTAT, gives several methods for taking systematic samples.

1. Frankel, M. R., and Spencer, B. D., *SAMPLE:* A *Supplementary Module for SYSTAT,* SYSTAT, Inc, Evanston, I11. 1990.

Bellhouse has written two review articles on systematic sampling that contain many valuable cross references to methodological articles and substantive applications.

2. Bellhouse, D. R., Systematic sampling. In *Sampling, Handbook of Statistics,* Vol. 6, Krishniah, P. R., and Rao, C. R., Eds., Amsterdam, Elsevier, 1988.
3. Bellhouse, D. R., Systematic sampling methods. In *Encyclopedia of Biostatistics,* Armitage, P. A. and Colton, T., Eds., Wiley, Chichester, U.K., 1998.

فصل ۵

طبقه‌بندی و نمونه‌گیری تصادفی طبقه‌بندی شده

در دو نوع نمونه‌گیری که تا اینجا مورد بحث قرار گرفتند یعنی در نمونه‌گیری تصادفی ساده و نمونه‌گیری سیستماتیک، لازم است که نمونه‌ای از جامعه به صورت یک کل گرفته شود و در هیچ یک از آنها لازم نیست که زیرحوزه‌ها یا زیرگروههای جامعه قبل از گرفتن نمونه شناسایی شوند. ولی گاهی اوقات می‌توان چارچوب نمونه‌گیری را به گروهها یا *طبقه‌هایی* افراز و نمونه‌گیری را جداگانه در داخل هر طبقه اجرا کرد. این طرح نمونه‌گیری را *نمونه‌گیری طبقه‌بندی شده* می‌نامند. اگر برای انتخاب نمونه در داخل هر طبقه از نمونه‌گیری تصادفی ساده استفاده شود طرح نمونه را *نمونه‌گیری تصادفی طبقه‌بندی شده* می‌نامند.

در این فصل برخی «مبانی» طبقه‌بندی و نمونه‌گیری تصادفی طبقه‌بندی شده را معرفی می‌کنیم. اینها شامل بحث دربارهٔ روشهای گرفتن نمونه، تعریف اصطلاحها و نمادگذاریهای مشخصه‌های جامعه و نمونه که در بحث طرحهای نمونه با طبقه‌بندی به کار رفته‌اند، و برآوردهای بنیادی مشخصه‌های جامعه خواهند بود. در فصل بعد با جزئیات بیشتری به روشهای استفاده از طرحهای طبقه‌بندی از راههایی که احتمال تولید قابل اعتمادترین برآوردها بیشتر است می پردازیم:

۱.۵ نمونهٔ تصادفی طبقه‌بندی شده چیست؟

همان طور که در بالا اشاره شد، *نمونهٔ تصادفی طبقه‌بندی شده* یک برنامهٔ نمونه‌گیری است که در آن جامعه به L طبقهٔ دو به دو ناسازگار و جامع طبقه‌بندی می‌شود، و یک نمونهٔ تصادفی ساده متشکل از n_h عنصر از داخل هر طبقهٔ h گرفته می‌شود. نمونه‌گیری در داخل هر طبقه به طور مستقل اجرا می‌شود. در واقع می‌توانیم طرح نمونه‌گیری تصادفی ساده را به صورت L نمونهٔ تصادفی سادهٔ مجزا تصور کنیم.

از نظر عملیاتی، نمونهٔ تصادفی طبقه‌بندی شده به همان طریق نمونهٔ تصادفی ساده گرفته می‌شود ولی نمونه‌گیری در داخل هر طبقه به طور جداگانه و مستقل انجام می‌گیرد. اگر $N_1, N_2, \ldots, N_L$ معرف تعداد واحدهای نمونه‌گیری در داخل هر طبقه و $n_1, n_2, \ldots, n_L$ معرف تعداد واحدهای نمونه‌گیری انتخاب شده به طور تصادفی در داخل هر طبقه باشند، آنگاه، تعداد کل نمونه‌های تصادفی طبقه‌بندی شدهٔ ممکن برابر خواهد بود با:

$$\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \cdots \times \binom{N_L}{n_L}$$

که از $\binom{N}{n}$، کل تعداد نمونه‌های تصادفی سادهٔ ممکن، کوچکتر بوده و یا با آن برابر است.

برای مثال، اگر سه طبقه داشته باشیم و $N_1 = ۳$، $N_2 = ۵$ و $N_3 = ٦$ باشد تعداد کل نمونه‌های ممکن متشکل از $n_1 = ۱$ عنصر از طبقهٔ ۱، $n_2 = ۲$ عنصر از طبقهٔ ۲ و $n_3 = ٤$ عنصر از طبقهٔ ۳ چنین است:

$$\binom{۳}{۱} \times \binom{۵}{۲} \times \binom{۶}{۴} = ۴۵۰$$

تعداد کل نمونه‌های تصادفی ساده متشکل از ۷ عنصر از ۱٤ عنصر جامعه عبارت است از

$$\binom{۱۴}{۷} = ۳۴۳۲$$

احتمال انتخاب شدن یک عنصر در نمونه بستگی دارد به طبقهٔ خاصی که آن عنصر در آن گروه‌بندی شده است و می‌توان نشان داد که برابر است با $\frac{n_h}{N_h}$ (در صورتی که آن عنصر در طبقهٔ h باشد). در مثالی که هم اکنون شرح دادیم، احتمال انتخاب شدن یک عنصر برابر است با $\frac{۱}{۳}$ برای عناصر طبقهٔ ۱، $\frac{۲}{۵}$ برای عناصر طبقهٔ ۲ و $\frac{۴}{۶}$ برای عناصر طبقهٔ ۳.

۲.۵ چگونگی گرفتن نمونهٔ تصادفی طبقه‌بندی شده

همان طور که در بالا شرح داده شد، نمونهٔ تصادفی طبقه‌بندی شده صرفاً با گرفتن نمونه‌های تصادفی ساده متشکل از n_h عنصر ($h = ۱,\ldots,L$) به طور مستقل در داخل هر طبقه تهیه می‌شود. انتخاب عملی نمونه، با استفاده از هر یک از روشهایی که قبلاً برای گرفتن نمونهٔ تصادفی ساده شرح داده شد اجرا می‌شود.

مثلاً، اگر کسی بخواهد از گزارشهای ذخیره شده در پروندهٔ اطلاعاتی رایانه نمونه‌ای بگیرد می‌تواند از یک الگوریتم برای گرفتن نمونهٔ تصادفی طبقه‌بندی شده که در کتاب راهنمای SAS [۷] توصیف شده است استفاده کند. همچنین مدول تکمیلی SYSTAT، SAMPLE نیز می‌تواند برای گرفتن نمونهٔ تصادفی طبقه‌بندی شده مورد استفاده قرار گیرد.

۳.۵ چرا نمونه‌گیری طبقه‌بندی شده؟

نمونه‌گیری طبقه‌بندی شده به این دلیل در انواع خاصی از آمارگیریها به کار می‌رود که آسانی درک نمونه‌گیری تصادفی ساده را با افزایش قابل توجهی در قابلیت اعتماد بالقوه ترکیب می‌کند. هر گاه بخواهیم برآوردهای جداگانه‌ای از پارامترهای جامعه برای هر یک از زیرحوزه‌ها در داخل کل جامعه به دست آوریم و علاوه بر آن بخواهیم مطمئن باشیم که نمونهٔ ما نمایندهٔ جامعه است استفاده از این فن راحت است. مثلاً، فرض کنید می‌خواهیم تعداد کل تختها را در بیمارستانهای یک ایالت مشخص برآورد کنیم. می‌دانیم که اکثر بیمارستانها، اندازه‌های کوچک یا متوسط دارند و تنها چندتایی بیمارستانهای بزرگ موجودند. می‌دانیم که این بیمارستانهای بزرگ بخش قابل توجهی از کل تعداد تختها را دارند.

حالا فرض کنید تصمیم بر این است که یک نمونهٔ تصادفی ساده از بیمارستانهای این ایالت انتخاب کنیم، تعداد تختها را در هر یک از نمونه‌هایی که به این ترتیب انتخاب شده‌اند تعیین، و با استفاده از روشهای فصل ۳، کل تعداد تختها را در تمام بیمارستانهای سراسر ایالت برآورد کنیم. مشکل این شیوه آن است که شانس زیادی وجود دارد که نمونهٔ ما تعداد بسیار زیاد یا بسیار کمی از بیمارستانهای بسیار بزرگ را در خود جای دهد. در نتیجه ممکن است نمونه به صورتی مناسب معرف جامعه نباشد.

راه حل ما برای این مشکل آن است که واحدهای نمونه‌گیری (بیمارستانها) را قبل از نمونه‌گیری بر اساس اندازه (یعنی کوچک، متوسط، بزرگ) در سه گروه طبقه‌بندی کنیم و سپس با استفاده از فنون نمونه‌گیری تصادفی ساده، از هر یک از گروههای سه‌گانه تعداد معینی از بیمارستانها را انتخاب نماییم.

سپس برآورد تعداد کل تختها را می‌توان از ترکیب نتایج سه طبقه به دست آورد. این، اساس نمونه‌گیری طبقه‌بندی شده است.

برای نشان دادن ایده‌ها و مزایای طبقه‌بندی به یک مثال نگاهی می‌اندازیم.

مثال تشریحی: فرض کنید جاده‌ای با ۲٤ مایل طول از مناطقی عبور می‌کند که می‌توانند به صورت شهری و روستایی طبقه‌بندی شوند و این جاده به هشت قطعه تقسیم شده است که هر یک دارای طولی برابر ۳ مایل است. نمونه‌ای شامل سه قطعه گرفته شده و در هر قطعهٔ نمونه‌گیری شده تجهیزات ویژه‌ای نصب گردیده است تا مجموع مایل‌های (مسافتهای) طی شده توسط وسایل نقلیه، اتومبیل و کامیون، در آن بخش طی یک سال خاص شمارش شود. به علاوه، گزارشی از تمام حوادثی که در هر قطعهٔ نمونه اتفاق می‌افتد نیز ثبت شود.

مقدار مسافت طی شده توسط کامیونها و تعداد حوادثی که در آن کامیون دخالت داشته است طی یک دورهٔ معین برای هر یک از قطعه‌های هشتگانهٔ جامعه در جدول ۱.۵ ارائه شده است.

فرض کنید یک نمونهٔ تصادفی ساده متشکل از سه قطعه را به منظور برآورد کل تعداد کامیون ـ مایل طی شده در جاده گرفته‌ایم. برای یک جامعهٔ متشکل از هشت قطعه، ٥٦ نمونهٔ ممکن سه قطعه‌ای وجود دارند. توزیع نمونه‌گیری تعداد کامیون ـ مایل طی شده در جاده در جدول ۲.۵ ارائه شده است.

جدول ۱.۵ کامیون ـ مایل و تعداد حوادثی که در آن کامیون مداخله داشته است بر حسب قطعه‌های جاده

قطعه	نوع	تعداد کامیون ـ مایل (×۱۰۰)	تعداد حوادث با مداخلهٔ کامیون
۱	شهری	٦۳۲۷	۸
۲	روستایی	۲٥٥٥	٥
۳	شهری	۸٦۹۱	۹
٤	شهری	۷۸۳٤	۹
٥	روستایی	۱٥۸٦	٥
٦	روستای	۲۰۳٤	۱
۷	روستایی	۲۰۱٥	۹
۸	روستایی	۳۰۱۲	٤

جدول ۲.۵ توزیع نمونه‌گیری x' برای ٥٦ نمونهٔ ممکن سه قطعه‌ای

قطعه‌ها در نمونه	x'	قطعه‌ها در نمونه	x'
(۱,۲,۳)	٤٦٨٦١/٣٣	(۲,٤,۷)	٣٣٠٧٧/٣٣
(۱,۲,٤)	٤٤٥٧٦	(۲,٤,۸)	٣٥٧٣٦
(۱,۲,٥)	٢٧٩١٤/٦٧	(۲,٥,٦)	١٦٤٦٦/٦٧
(۱,۲,٦)	٢٩١٠٩/٣٣	(۲,٥,۷)	١٦٤١٦
(۱,۲,۷)	٢٩٠٥٨/٦٧	(۲,٥,۸)	١٩٠٧٤/٦٧
(۱,۲,۸)	٣١٧١٧/٣٣	(۲,٦,۷)	١٧٦١٠/٦٧
(۱,۳,٤)	٦٠٩٣٨/٦٧	(۲,٦,۸)	٢٠٢٦٩/٣٣
(۱,۳,٥)	٤٤٢٧٧/٣٣	(۲,۷,۸)	٢٠٢١٨/٦٧
(۱,۳,٦)	٤٥٤٧٢	(۳,٤,٥)	٤٨٢٩٦
(۱,۳,۷)	٤٥٤٢١/٣٣	(۳,٤,٦)	٤٩٤٩٠/٦٧
(۱,۳,۸)	٤٨٠٨٠	(۳,٤,۷)	٤٩٤٤٠
(۱,٤,٥)	٤١٩٩٢	(۳,٤,۸)	٥٢٠٩٨/٦٧
(۱,٤,٦)	٤٣١٨٦/٦٧	(۳,٥,٦)	٣٢٨٢٩/٣٣
(۱,٤,۷)	٤٣١٣٦	(۳,٥,۷)	٣٢٧٧٨/٦٧
(۱,٤,۸)	٤٥٧٩٤/٦٧	(۳,٥,۸)	٣٥٤٣٧/٣٣
(۱,٥,٦)	٢٦٥٢٥/٣٣	(۳,٦,۷)	٣٣٩٧٣/٣٣
(۱,٥,۷)	٢٦٤٧٤/٦٧	(۳,٦,۸)	٣٦٦٣٢
(۱,٥,۸)	٢٩١٣٣/٣٣	(۳,۷,۸)	٣٦٥٨١/٣٣
(۱,٦,۷)	٢٧٦٦٩/٣٣	(٤,٥,٦)	٣٠٥٤٤
(۱,٦,۸)	٣٠٣٢٨	(٤,٥,۷)	٣٠٤٩٣/٣٣
(۱,۷,۸)	٣٠٢٧٧/٣٣	(٤,٥,۸)	٣٣١٥٢
(۲,۳,٤)	٥٠٨٨٠	(٤,٦,۷)	٣١٦٨٨
(۲,۳,٥)	٣٤٢١٨/٦٧	(٤,٦,۸)	٣٤٣٤٦/٦٧
(۲,۳,٦)	٣٥٤١٣/٣٣	(٤,۷,۸)	٣٤٢٩٦
(۲,۳,۷)	٣٥٣٦٢/٦٧	(٥,٦,۷)	١٥٠٢٦/٦٧
(۲,۳,۸)	٣٨٠٢١/٣٣	(٥,٦,۸)	١٧٦٨٥/٣٣
(۲,٤,٥)	٣١٩٣٣/٣٣	(٥,۷,۸)	١٧٦٣٤/٦٧
(۲,٤,٦)	٣٣١٢٨	(٦,۷,۸)	١٨٨٢٩/٣٣

جدول ۳.۵ دو طبقه برای داده‌های جدول ۱.۵

طبقهٔ ۱ (قطعات شهری)		طبقهٔ ۲ (قطعات روستایی)	
قطعه	۱۰۰۰× تعداد کامیون ـ مایل	قطعه	۱۰۰۰× تعداد کامیون ـ مایل
۱	٦۳۲۷	۲	۲۵۵۵
۳	۸٦۹۱	۵	۱۵۸٦
٤	۷۸۳٤	٦	۲۰۳٤
		۷	۲۰۱۵
		۸	۳۰۱۲

برآوردهای کل تعداد کامیون ـ مایل که از این طریق به دست می‌آید از ۱۵۰۲٦/٦۷ تا ٦۰۹۳۸/٦۷ تغییر می‌کنند و میانگین توزیع نمونه‌گیری x' برابر است با ۳٤۰۵٤ یعنی مجموع جامعه، و خطای معیار x' برابر است با ۱۰۵۳٦/۹.

حالا فرض کنید که به جای گرفتن یک نمونهٔ تصادفی ساده متشکل از سه قطعه از جامعه‌ای با هشت قطعه، اول قطعه‌ها را به دو طبقه تقسیم کنیم، یکی شامل قطعات شهری، دیگری شامل قطعات روستایی، به صورتی که در جدول ۳.۵ نشان داده شده است.

حالا می‌توانیم یک نمونه متشکل از یک قطعه از طبقهٔ ۱ و دو قطعه از طبقهٔ ۲ انتخاب کنیم و کل تعداد کامیون ـ مایل را به وسیلهٔ برآورد x'_{str} به صورت زیر برآورد کنیم:

$$x'_{str} = x'_{۱} + x'_{۲}$$

که در آن

$x'_{۱}$ = تعداد کامیون ـ مایل برآورد شده در سه قطعهٔ تشکیل دهندهٔ طبقهٔ ۱

$x'_{۲}$ = تعداد کامیون ـ مایل برآورد شده در پنج قطعهٔ تشکیل دهندهٔ طبقهٔ ۲

از نظر مفهومی، هر طبقه را یک زیرجامعه در نظر می‌گیریم، در داخل هر زیرجامعه به طور مستقل نمونه می‌گیریم، و برآوردی برای کل جامعه با انبوه سازی برآورد تک تک طبقه‌ها در همهٔ زیرجامعه‌ها یا طبقات به دست می‌آوریم. بنابراین، اگر یک نمونهٔ تصادفی ساده از یک قطعه از طبقهٔ ۱ و دو قطعه از طبقهٔ ۲ بگیریم در مجموع $\binom{۳}{۱}\times\binom{۵}{۲} = ۳۰$ نمونهٔ ممکن به دست می‌آوریم. پس از آن اگر از شیوهٔ برآورد که در بالا با معادلهٔ مربوط به x'_{str} ارائه شد استفاده کنیم توزیع نمونه‌گیری x'_{str} را به دست می‌آوریم که در جدول ٤.۵ نشان داده شده است.

$E(x'_{str})$، میانگین توزیع نمونه‌گیری x'_{str} برابر است با ۳٤۰۵٤ (تابلوی ۳.۲ را ببینید) که همان مجموع واقعی جامعه است. $SE(x'_{str})$، خطای معیار کل برآورد شدهٔ x'_{str} برابر است با ۳۲۹۷/٦ که

جدول ٤.٥ توزیع نمونه‌گیری x'_{str} برای ۳۰ نمونهٔ ممکن از سه قطعه

طبقهٔ ۱	طبقهٔ ۲	$x'_۱(=۳\bar{x}_۱)$	$x'_۲(=۵\bar{x}_۲)$	$x'_{str}=(x'_۱+x'_۲)$
۱	(۲,۵)	۱۸۹۸۱	۱۰۳۵۲/۵	۲۹۳۳۳/۵
۱	(۲,٦)	۱۸۹۸۱	۱۱٤۷۲/۵	۳۰٤۵۳/۵
۱	(۲,۷)	۱۸۹۸۱	۱۱٤۲۵	۳۰٤۰٦
۱	(۲,۸)	۱۸۹۸۱	۱۳۹۱۷/۵	۳۲۸۹۸/۵
۱	(۵,٦)	۱۸۹۸۱	۹۰۵۰	۲۸۰۳۱
۱	(۵,۷)	۱۸۹۸۱	۹۰۰۲/۵	۲۷۹۸۳/۵
۱	(۵,۸)	۱۸۹۸۱	۱۱٤۹۵	۳۰٤۷٦
۱	(٦,۷)	۱۸۹۸۱	۱۰۱۲۲/۵	۲۹۱۰۳/۵
۱	(٦,۸)	۱۸۹۸۱	۱۲٦۱۵	۳۱۵۹٦
۱	(۷,۸)	۱۸۹۸۱	۱۲۵٦۷/۵	۳۱۵٤۸/۵
۳	(۲,۵)	۲٦۰۷۳	۱۰۳۵۲/۵	۳٦٤۲۵/۵
۳	(۲,٦)	۲٦۰۷۳	۱۱٤۷۲/۵	۳۷۵٤۵/۵
۳	(۲,۷)	۲٦۰۷۳	۱۱٤۲۵	۳۷٤۹۸
۳	(۲,۸)	۲٦۰۷۳	۱۳۹۱۷/۵	۳۹۹۹۰/۵
۳	(۵,٦)	۲٦۰۷۳	۹۰۵۰	۳۵۱۲۳
۳	(۵,۷)	۲٦۰۷۳	۹۰۰۲/۵	۳۵۰۷۵/۵
۳	(۵,۸)	۲٦۰۷۳	۱۱٤۹۵	۳۷۵٦۸
۳	(٦,۷)	۲٦۰۷۳	۱۰۱۲۲/۵	۳٦۱۹۵/۵
۳	(٦,۸)	۲٦۰۷۳	۱۲٦۱۵	۳۸٦۸۸
۳	(۷,۸)	۲٦۰۷۳	۱۲۵٦۷/۵	۳۸٦٤۰/۵
٤	(۲,۵)	۲۳۵۰۲	۱۰۳۵۲/۵	۳۳۸۵٤/۵
٤	(۲,٦)	۲۳۵۰۲	۱۱٤۷۲/۵	۳٤۹۷٤/۵
٤	(۲,۷)	۲۳۵۰۲	۱۱٤۲۵	۳٤۹۲۷
٤	(۲,۸)	۲۳۵۰۲	۱۳۹۱۷/۵	۳۷٤۱۹/۵
٤	(۵,٦)	۲۳۵۰۲	۹۰۵۰	۳۲۵۵۲
٤	(۵,۷)	۲۳۵۰۲	۹۰۰۲/۵	۳۲۵۰٤/۵
٤	(۵,۸)	۲۳۵۰۲	۱۱٤۹۵	۳٤۹۹۷
٤	(٦,۷)	۲۳۵۰۲	۱۰۱۲۲/۵	۳۳٦۲٤/۵
٤	(٦,۸)	۲۳۵۰۲	۱۲٦۱۵	۳٦۱۱۷
٤	(۷,۸)	۲۳۵۰۲	۱۲۵٦۷/۵	۳٦۰٦۹/۵

خیلی کمتر از خطای معیار x'، یعنی برآورد معادل کل جامعهٔ تحت نمونه‌گیری تصادفی ساده است. این نتایج در جدول ۵.۵ خلاصه شده‌اند.

به صورت شهودی می‌توان دید که چرا برنامهٔ نمونه‌گیری مبتنی بر طبقه‌بندی برآوردی را به دست می‌دهد که خطای معیار آن کمتر از برآورد نظیر بر مبنای نمونه‌گیری تصادفی ساده است. برای x'، برآورد مجموع، با استفاده از نمونه‌گیری تصادفی ساده ٥٦ مقدار ممکن به دست می‌آید در حالی که از طبقه‌بندی ۳۰ مقدار برای x'_{str} به دست می‌آید. همچنین، امتحان کردن دامنهٔ تغییرات دو توزیع نشان می‌دهد که طبقه‌بندی، آن دسته از نمونه‌ها را که منجر به برآوردهای بسیار زیاد یا بسیار کم برای مجموع می‌شد حذف کرده است. سه قطعهٔ شهری دارای مقادیر زیاد و پنج قطعهٔ روستایی دارای مقادیر کم مشخصه‌های مورد اندازه‌گیری (کامیون – مایل) بودند. طبقه‌بندی اطمینان می‌دهد که حداقل یک قطعهٔ شهری و یک قطعهٔ روستایی در نمونه انتخاب می‌شوند و به این ترتیب امکان برآوردهای بسیار زیاد و بسیار کم مجموع را از میان می‌برد.

جدول ۵.۵ مقایسهٔ نتایج نمونه‌گیری تصادفی ساده و طبقه‌بندی

	طرح طبقه‌بندی	
	نمونه‌گیری تصادفی ساده	طبقه‌بندی
تعداد عناصر در نمونه	۳	*۳
تعداد نمونه‌های ممکن	٥٦	۳۰
میانگین توزیع برآورد مجموع	۳٤۰٥٤†	۳٤۰٥٤†
خطای معیار برآورد مجموع	۱۰٥۳٦/۹	۳۲۹۷/٦
دامنهٔ تغییرات توزیعهای برآورد مجموعها	٤٥۹۱۲	۱۲۰۰۷

* یک عنصر از طبقهٔ ۱ و دو عنصر از طبقهٔ ۲.

† این مقدار همان مجموع جامعه است.

□

طبقه‌بندی نسبت به نمونه‌گیری تصادفی ساده، سه امتیاز عمده دارد.

۱. با شرایط معین مفروض، دقت آن نسبت به نمونه‌گیری تصادفی ساده ممکن است افزایش پیدا کند (یعنی ممکن است خطاهای معیار به دست آمده از این شیوهٔ برآورد، کمتر باشند).
۲. امکان به دست آوردن برآوردهایی با دقت مشخص برای هر یک از طبقه‌ها وجود دارد.
۳. ممکن است جمع‌آوری اطلاعات برای نمونهٔ طبقه‌بندی شده نیز به دلایل سیاسی یا اداری به همان آسانی جمع‌آوری اطلاعات برای یک نمونهٔ تصادفی ساده باشد. در این صورت، با

گرفتن نمونهٔ طبقه‌بندی شده چیزی را از دست نمی‌دهیم زیرا، خطاهای معیار حاصل از این طرح به ندرت ممکن است از خطاهای معیار نمونه‌گیری تصادفی ساده بیشتر شود.

راهبرد مورد استفاده برای ساختن طبقات مستلزم دو گام است. ابتدا پارامتر مورد نظر جامعه برای برآورد کردن را تعیین می‌کنیم. سپس جامعه را نسبت به متغیر دیگری که تصور می‌شـــود با متغیر مورد نظر مربوط باشد طبقه‌بندی می‌کنیم. اگر فرض ما راجع به این پیوند درست باشد این گام دوم تضمین می‌کند که طبقه‌ها نسبت به متغیر تحت بررسی همگن‌اند.

برای مثال، اگر بخواهیم تعداد تختها را در تمام بیمارستانهای ایالت برآورد کنیم ممکن است بخواهیم بیمارستانها را از لحاظ زیربنا (که اطلاعی است که به خاطر مالیات به راحتی قابل دسترسی است) با این استدلال طبقه‌بندی کنیم که بیمارستانهایی که فضای بیشتری دارند تختهای بیشتری خواهند داشت. از سوی دیگر، اگر به متوسط هزینهٔ روزانهٔ هر تخت بیمارستانی به ازای هر بیمار علاقه‌مندیم می‌توانیم طبقه‌بندی بیمارستانها را بر مبنای محل جغرافیایی آنان با این استدلال انجام دهیم که بیمارستانهای مناطقی از ایالت که دارای اقتصاد پر رونقی‌اند می‌توانند هزینه‌های سنگینتری از بیماران خود مطالبه کنند تا بیمارستانهایی که در مناطق با اقتصاد کساد واقع شده‌اند.

در بیشتر موقعیتهای عملی، طبقه‌بندی جامعه نسبت به متغیر تحت بررسی صرفاً به دلایل مربوط به هزینه و عملی بودن، کاری مشکل است. در بسیاری از موارد جامعه به راحت‌ترین طریق با استفاده از معیارهای اداری (مثل نواحی رأی‌گیری)، معیارهای جغرافیایی (مثل شمال، جنوب، شرق یا غرب) یا سایر معیارهای طبیعی (مثل جنس یا سن) طبقه‌بندی می‌شود. طبقه‌بندی بر اساس راحتی غیرمنطقی نیست زیرا برآورد یک تک پارامتر در آمارگیری مدرن متداول نیست. در عوض، اطلاعات زیادی در مورد هر واحد نمونه‌گیری جمع‌آوری می‌شود و پارامترهای زیادی مورد توجه‌اند. واضح است که آن چه برای یک متغیر می‌تواند راهبرد طبقه‌بندی بهینه‌ای برای تهیه طبقات نسبتاً همگن باشد ممکن است نسبت به یک متغیر دیگر، طبقات بسیار ناهمگن فراهم نماید. اهمیت دارد که آماردان قبل از تصمیم‌گیری در مورد ملاک مناسب برای طبقه‌بندی، دامنهٔ داده‌هایی را که قرار است جمع‌آوری شوند مورد توجه قرار دهد. این قبیل مطالب گاهی اوقات در گزارشهای آمارگیریهای بهداشتی که از طبقه‌بندی استفاده کرده‌اند مورد بحث قرار می‌گیرند.

عیب عمدهٔ نمونه‌گیری طبقه‌بندی شده آن است که نیازمند شناسایی یکایک واحدهای شمارش بر حسب طبقه پیش از اجرای نمونه‌گیری است. اگر چنین اطلاعاتی به سهولت فراهم نباشد این روش به ندرت ممکن است عملی باشد.

ممکن است بخواهید طبقه‌بندی را با راهبرد نمونه‌گیری خوشه‌ای (فصلهای ۸ تا ۱۱) مقایسه کنید که در آن ممکن است صرفه‌جوییهایی عمده در وقت و هزینه امکان‌پذیر شود. مع‌هذا، به دلیل مزایایی که در بالا شرح داده شد، طبقه بندی فن بسیار توانایی است که در سطح وسیعی مورد استفاده قرار می‌گیرد.

٤.٥ پارامترهای جامعه برای طبقات

پارامترهای جامعه برای طبقات را می‌توان با همان نمادهایی تعریف کرد که برای تعریف پارامترهای جامعه به طور کلی مورد استفاده قرار دادیم. در این بخش نمادهایی را معرفی می‌کنیم که در سراسر بحث از برنامه‌های نمونه‌گیری بر مبنای طبقه‌بندی به کار خواهند رفت.

جامعه‌ای را در نظر می‌گیریم که دارای N واحد اولیه است که در L طبقهٔ دو به دو ناسازگار و جامع به گونه‌ای گروه‌بندی شده‌اند که طبقهٔ ۱ دارای N_1 واحد اولیه، طبقهٔ ۲ دارای N_2 واحد اولیه، ... ، و طبقهٔ L دارای N_L واحد اولیه است. به عبارت دیگر، $N = \sum_{h=1}^{L} N_h$ *اندازهٔ جامعه* است.

فرض کنید یک متغیر یا مشخصهٔ x را در جامعه در نظر گرفته‌ایم. از نماد $X_{h,i}$ برای نشان دادن *مقدار مشخصهٔ* x برای iامین واحد اولیه در داخل طبقهٔ h استفاده می‌کنیم. به عبارت دیگر، واحدهای اولیه داخل هر طبقهٔ ویژهٔ h از ۱ تا N_h شماره‌گذاری می‌شوند.

مثال تشریحی: فرض کنید می‌خواهیم متوسط هزینهٔ روزانهٔ دارو به ازای هر بیمار را در یک بیمارستان برآورد کنیم. تصمیم می‌گیریم بیمارستان را بر حسب خدمات (درمانی، جراحی، زنان و زایمان، و همهٔ خدمات دیگر با هم) طبقه‌بندی کنیم و واحدهای اولیه را به صورت بیماران در یک روز معین تعریف می‌کنیم.

فرض کنید که در یک روز تعیین شده ۲۵۰ بیمار در بیمارستان است که ۱۰۰ نفر برای درمان، ۷۵ نفر برای جراحی، ۵۰ نفر زنان و زایمان و ۲۵ نفر برای سایر خدمات مراجعه کرده‌اند. پس با استفاده از نمادهایی که در بالا معرفی شدند داریم:

$$N = ۲۵۰ \qquad N_1 = ۱۰۰ \qquad N_2 = ۷۵ \qquad N_3 = ۵۰ \qquad N_4 = ۲۵$$

اگر $X_{h,i}$ نشان‌دهندهٔ مقدار متغیر x برای iامین واحد اولیه در داخل طبقهٔ h باشد، آنگاه مثلاً برای طبقهٔ ۲ داریم:

$X_{2,1}$ = مقدار متغیر x برای عنصر ۱ در داخل طبقهٔ ۲

$X_{2,2}$ = مقدار متغیر x برای عنصر ۲ در داخل طبقهٔ ۲

و به همین ترتیب تا می‌رسیم به

$X_{۲,۷۵}$ = مقدار متغیر x برای عنصر ۷۵ در داخل طبقهٔ ۲

عناصر داخل سایر طبقه‌ها نیزبه همین ترتیب تعریف می‌شوند.

□

پارامترهای جامعه برای طبقات را به روشی تعریف می‌کنیم که پارامترهای جامعهٔ کل را تعریف کردیم.

مجموع یا مقدار تجمعی متغیر x *در داخل یک طبقهٔ* h را با نماد X_{h+} نشان می‌دهیم که از فرمول زیر به دست می‌آید:

$$X_{h+} = \sum_{i=۱}^{N_h} X_{h,i}$$

مجموع کل جامعه با جمع کردن مجموع طبقه‌ها به دست می‌آید، یا

$$X = \sum_{h=۱}^{L}\sum_{i=۱}^{N_h} X_{h,i} = \sum_{h=۱}^{L} X_{h+}$$

سطح *میانگین* یک مشخصهٔ x برای طبقهٔ h با نماد $\overline{X}_h$ نشان داده می‌شود و از فرمول زیر به دست می‌آید:

$$\overline{X}_h = \frac{\sum_{i=۱}^{N_h} X_{h,i}}{N_h} = \frac{X_{h+}}{N_h}$$

میانگین $\overline{X}$ متغیر x برای تمام جامعه از فرمول زیر

$$\overline{X} = \frac{X}{N}$$

یا معادل جبری آن،

$$\overline{X} = \frac{\sum_{h=۱}^{L} N_h \overline{X}_h}{N} = \sum_{h=۱}^{L} W_h \overline{X}_h$$

به دست می‌آید، که در آن،

$$W_h = \frac{N_h}{N}$$

به عبارت دیگر، میانگین $\overline{X}$ برای تمام جامعه عبارت است از *متوسط موزون* میانگینهای تکی طبقه‌ها، $\overline{X}_h$، با وزنهای ($W_h = \frac{N_h}{N}$) متناسب با تعداد عناصر در هر طبقه.

واریانس $\sigma^{۲}_{hx}$ توزیع متغیر x در داخل یک طبقهٔ خاص h به صورت متوسط توان دوم انحرافها حول میانگین طبقه تعریف و با فرمول زیر نشان داده می‌شود:

$$\sigma^{۲}_{hx} = \frac{\sum_{i=۱}^{N_h} (X_{h,i} - \overline{X}_h)^{۲}}{N_h}$$

ضرایب تغییرات و واریانسهای نسبی برای هر طبقه به همان صورتی تعریف می‌شوند که برای جامعه‌ای که به طبقات گروه‌بندی نشده است تعریف شد. *ضریب تغییرات* برای توزیع در داخل یک طبقهٔ خاص از فرمول زیر به دست می‌آید:

$$V_{hx} = \frac{\sigma_{hx}}{\overline{X}_h}$$

و واریانس نسبی صرفاً عبارت است از توان دوم ضریب تغییرات.

بـرای راحتی کار در ارجاعهای بعدی، پارامترهای جامعه برای نمونه‌گیری طبقه‌بندی شده در تابلوی ۱.۵ خلاصه شده‌اند. حالا ببینیم این فرمولها در عمل چگونه به کار می‌روند.

مثال تشریحی: جامعه‌ای متشکل از ١٤ خانواده را در نظر بگیرید که در سه بلوک شهری زندگی می‌کنند. اگر خانواده‌ها را واحدهای اولیه، بلوکها را به عنوان طبقه‌ها، و اندازهٔ خانواده‌ها را مشخصهٔ x در نظر بگیریم، وضعیتی خواهیم داشت که در جدول ٦.٥ نشان داده شده است.

تابلوی ۱.۵ پارامترهای طبقه‌ها و جامعه برای نمونه‌گیری طبقه‌بندی شده

	تمام جامعه	در داخل طبقه
مجموع		
(۱.۵)	$X = \sum_{h=1}^{L}\sum_{i=1}^{N_h} X_{h,i} = \sum_{h=1}^{L} X_{h+}$	$X_{h+} = \sum_{i=1}^{N_h} X_{h,i}$
میانگین		
(۲.۵)	$\overline{X} = \frac{\sum_{h=1}^{L} N_h \overline{X}_h}{N} = \sum_{h=1}^{L} W_h \overline{X}_h = \frac{X}{N}$	$\overline{X}_h = \frac{\sum_{i=1}^{N_h} X_{h,i}}{N_h} = \frac{X_{h+}}{N_h}$
نسبت		
(۳.۵)	$P_y = \frac{\sum_{h=1}^{L} N_h P_{hy}}{N} = \sum_{h=1}^{L} W_h P_{hy}$	$P_{hy} = \frac{\sum_{i=1}^{N_h} Y_{h,i}}{N_h}$
واریانس		
(٤.۵)		$\sigma_{hx}^2 = \frac{\sum_{i=1}^{N_h} (X_{h,i} - \overline{X}_h)^2}{N_h}$
واریانس نسبی		
(۵.۵)		$V_{hx}^2 = \frac{\sigma_{hx}^2}{\overline{X}_h^2}$

در این تعریفها L تعداد طبقه‌ها، N_h تعداد عناصر در طبقهٔ h، N تعداد کل عناصر، $X_{h,i}$ مقدار x برای iامین عنصر در طبقهٔ h، $y_{h,i}$ نشانهٔ بود یا نبود یک صفت کیفی دوحالتی y برای iامین عنصر در طبقهٔ h، و $W_h = \frac{N_h}{N}$ نسبتی از کل جامعه که متعلق به طبقهٔ h است.

جدول ٦.٥ طبقه‌ها برای جامعهٔ متشکل از ١٤ خانواده

بلوک	خانواده	اندازهٔ خانواده
١	١	٤
	٢	٣
	٣	٤
٢	١	٤
	٢	٦
	٣	٤
	٤	٨
	٥	٨
٣	١	٢
	٢	٣
	٣	٢
	٤	٢
	٥	٢
	٦	٣

بر حسب نمادهایی که در بالا معرفی شدند، مقادیر x برای هر طبقه به شرح زیرند:

طبقهٔ ١	طبقهٔ ٢	طبقهٔ ٣
$(N_1 = ٣)$	$(N_2 = ٥)$	$(N_3 = ٦)$
$X_{1,1} = ۴$	$X_{2,1} = ۴$	$X_{3,1} = ٢$
$X_{1,2} = ٣$	$X_{2,2} = ۶$	$X_{3,2} = ٣$
$X_{1,3} = ۴$	$X_{2,3} = ۴$	$X_{3,3} = ٢$
	$X_{2,4} = ٧$	$X_{3,4} = ٢$
	$X_{2,5} = ٨$	$X_{3,5} = ٢$
		$X_{3,6} = ٣$

مجموع متغیر x در داخل هر طبقهٔ h با استفاده از معادلهٔ (١.٥) به شرح زیر است:

$$X_{۱+} = ۴ + ۳ + ۴ = ۱۱$$
$$X_{۲+} = ۴ + ۶ + ۴ + ۷ + ۸ = ۲۹$$
$$X_{۳+} = ۲ + ۳ + ۲ + ۲ + ۲ + ۳ = ۱۴$$

مجموع برای تمام جامعه دوباره با استفاده از معادلهٔ (۱.۵) عبارت است از:

$$X = ۱۱ + ۲۹ + ۱۴ = ۵۴$$

میانگینهای جامعه‌ای برای طبقات با استفاده از معادلهٔ (۲.۵) به صورت زیرند:

$$\overline{X}_{۱} = \frac{۱۱}{۳} = ۳/۶۷ \qquad \overline{X}_{۲} = \frac{۲۹}{۵} = ۵/۸ \qquad \overline{X}_{۳} = \frac{۱۴}{۶} = ۲/۳۳$$

میانگین جامعهٔ کل از معادلهٔ (۲.۵) به صورت زیر به دست می‌آید:

$$\overline{X} = \frac{۳}{۱۴} \times ۳/۶۷ + \frac{۵}{۱۴} \times ۵/۸ + \frac{۶}{۱۴} \times ۲/۳۳ = ۳/۸۵۷$$

واریانسهای طبقات با استفاده از معادلهٔ (٤.۵) عبارت‌اند از:

$$\sigma_{۱x}^{۲} = \frac{(۴ - ۳/۶۷)^{۲} + (۳ - ۳/۶۷)^{۲} + (۴ - ۳/۶۷)^{۲}}{۳} = ۰/۲۲۲$$

$$\sigma_{۲x}^{۲} = \frac{(۴ - ۵/۸)^{۲} + (۶ - ۵/۸)^{۲} + \cdots + (۸ - ۵/۸)^{۲}}{۵} = ۲/۵۶$$

$$\sigma_{۳x}^{۲} = \frac{(۲ - ۲/۳۳)^{۲} + (۳ - ۲/۳۳)^{۲} + \cdots + (۳ - ۲/۳۳)^{۲}}{۶} = ۰/۲۲۲$$

واریانسهای نسبی برای طبقات از معادلهٔ (۵.۵) به شرح زیر به دست می‌آیند:

$$V_{۱x}^{۲} = \frac{۰/۲۲۲}{(۳/۶۷)^{۲}} = ۰/۰۱۶۵$$
$$V_{۲x}^{۲} = \frac{۲/۵۶}{(۵/۸)^{۲}} = ۰/۰۷۶۱$$
$$V_{۳x}^{۲} = \frac{۰/۲۲۲}{(۲/۳۳)^{۲}} = ۰/۰۴۰۹$$

□

۵.۵ آماره‌های نمونه برای طبقات

فرض کنید که در داخل یک طبقهٔ خاص h به نحوی یک نمونهٔ n_h عنصری از N_h عنصر موجود در طبقه انتخاب و هر عنصر نمونه را نسبت به نوعی متغیر x اندازه‌گیری کرده‌ایم. برای راحتی کار، عناصر

نمونه را از ۱ تا n_h شماره‌گذاری می‌کنیم و قرار می‌گذاریم که $x_{h,1}$ مقدار متغیر x برای عنصر شمارهٔ «۱» نمونه و $x_{h,2}$ مقدار متغیر x برای عنصر شمارهٔ «۲» نمونه و $x_{h.n_h}$ مقدار x برای عنصر شمارهٔ «n_h» نمونه باشد. اگر نمونه‌ای متشکل از n_h عنصر در داخل هر طبقهٔ h گرفته شود، در آن صورت کل اندازهٔ نمونهٔ n از راه زیر به دست می‌آید:

$$n = \sum_{h=1}^{L} n_h$$

که در آن L، تعداد طبقاتی است که عناصر جامعه در آنها گروه‌بندی شده‌اند.

مثلاً، در مثال قبل اگر یک عنصر از طبقهٔ ۱، دو عنصر از طبقهٔ ۲ و چهار عنصر از طبقهٔ ۳، انتخاب کنیم خواهیم داشت

$$n_1 = 1 \quad , \quad n_2 = 2 \quad , \quad n_3 = 4$$

و

$$n = 1+2+4 = 7$$

اگر مقادیر چهار عنصر انتخاب شده از طبقهٔ ۳ عبارت باشند از $X_{3,1}$، $X_{3,3}$، $X_{3,4}$ و $X_{3,6}$ در آن صورت خواهیم داشت:

$$x_{3,1} = X_{3,1} = 2 \qquad x_{3,3} = X_{3,4} = 2$$
$$x_{3,2} = X_{3,3} = 2 \qquad x_{3,4} = X_{3,6} = 3$$

مجموع نمونه‌ای و میانگین *نمونه‌ای برای یک طبقهٔ خاص* از فرمول زیر به دست می‌آیند:

$$x_{h+} = \sum_{i=1}^{n_h} x_{h,i} \qquad \text{و} \qquad \bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{h,i}}{n_h} = \frac{x_{h+}}{n_h}$$

۶.۵ برآورد پارامترهای جامعه از روی نمونه‌گیری تصادفی طبقه‌بندی شده

اگر انتخاب عناصر نمونه به طور مستقل در داخل هر طبقه با استفاده از نمونه‌گیری تصادفی ساده انجام پذیرفته باشد، در آن صورت این طرح را *طرح نمونه‌گیری تصادفی طبقه‌بندی شده* می‌نامند و برآوردهای مناسب پارامترهای جامعه همراه با خطاهای معیار آنها در تابلوی ۲.۵ نشان داده شده‌اند.

در این تعاریف، L تعداد طبقات، N_h تعداد عناصر در طبقهٔ h، N کل تعداد عناصر، $x_{h,i}$ مقدار x برای iامین عنصر در طبقهٔ h و $y_{h,i}$ معرف بود یا نبود یک صفت کیفی دو حالتی y برای iامین عنصر در طبقهٔ h است.

همچنین $S_{hx}^{۲}$، برآورد واریانس توزیع x برای طبقهٔ h است که از فـرمول زیر به دست می‌آید:

$$S_{hx}^{۲} = \frac{\sum_{i=۱}^{n_h}(x_{h,i} - \bar{x}_h)^{۲}}{n_h - ۱} \qquad (۹.۵)$$

تابلوی ۲.۵ برآوردهای پارامترهای جامعه و خطاهای معیار این برآوردها در نمونه‌گیری طبقه‌بندی شده

پارامتر	خطای معیار برآورد برای تمام جامعه	برآورد برای تمام جامعه	برآورد برای طبقه
مجموع			
(۶.۵)	$\hat{SE}(x'_{\text{str}}) = \sqrt{\sum_{h=۱}^{L} \frac{N_h^{۲} S_{hx}^{۲}}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$	$x'_{\text{str}} = \sum_{h=۱}^{L} x'_h$	$x'_h = N_h \bar{x}_h$
میانگین			
(۷.۵)	$\hat{SE}(\bar{x}_{\text{str}}) = \sqrt{\sum_{h=۱}^{L}\left(\frac{N_h}{N}\right)^{۲} \frac{S_{hx}^{۲}}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$	$\bar{x}_{\text{str}} = \frac{\sum_{h=۱}^{L} N_h \bar{x}_h}{N}$	$\bar{x}_h = \frac{\sum_{i=۱}^{n_h} x_{h,i}}{n_h}$
نسبت			
(۸.۵)	$\hat{SE}(p_{y,\text{str}}) = \sqrt{\sum_{h=۱}^{L}\left(\frac{N_h}{N}\right)^{۲} \frac{p_{hy}(۱-p_{hy})}{n_h - ۱}\left(\frac{N_h - n_h}{N_h}\right)}$	$p_{y,\text{str}} = \frac{\sum_{h=۱}^{L} y_{hi}}{N}$	$p_{hy} = \frac{\sum_{i=۱}^{n_h} y_{h,i}}{n_h}$

مثال تشریحی: با استفاده از STATA یکصد و پنجاه و هشت (۱۵۸) بیمارستان در ناحیه‌ای بر حسب سطح خدمات زایمان در سه طبقه گروه‌بندی شده‌اند که با متغیر *oblevel* نشان داده می‌شود. از این ۱۵۸ بیمارستان ناحیه، ٤۲ بیمارستان در طبقهٔ ۱ (تخصصی)، ۹۹ بیمارستان در طبقهٔ ۲ (متوسط) و ۱۷ بیمارستان در طبقـــهٔ ۳ (درجهٔ سه) قرار دارند. یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۱۵ بیمارستان از این جامعه گرفته شده است: ٤ بیمارستان در طبقهٔ ۱؛ ۵ بیمارستان در طبقهٔ ۲؛ و ٦ بیمارستان در طبقهٔ ۳. این نمونهٔ خاص متشکل از ۱۵ بیمارستان همراه با متغیرهای مربوط برای برآورد، در پروندهٔ اطلاعاتی STATA به صورت *hospsamp.dta* ثبت شده‌اند. این پرونده در صفحهٔ بعد نشان داده شده است.

Hospno	oblevel	weighta	tothosp	births
15	1	10.50	42	480
80	1	10.50	42	426
86	1	10.50	42	342
136	1	10.50	42	174
7	2	19.80	99	2022
26	2	19.80	99	576
62	2	19.80	99	1999
90	2	19.80	99	482
101	2	19.80	99	836
28	3	2.83	17	3108
34	3	2.83	17	4674
39	3	2.83	17	2539
102	3	2.83	17	1610
119	3	2.83	17	4618
149	3	2.83	17	1781

معانی متغیرها به شرح زیرند:

hospno = کد شناسایی برای بیمارستان خاص.

oblevel = سطح خدمات زایمان برای هر بیمارستان.

weighta = وزن نمونه‌گیری $\frac{N_h}{n_h}$ برای بیمارستان خاص.

tothosp = تعداد کل جامعهٔ بیمارستانها در طبقه‌ای که بیمارستان نمونه قرار دارد.

births = تعداد رویدادهای تولد در بیمارستان در طی سال قبل.

از فرمانهای زیر می‌توان برای به دست آوردن برآوردهای کل تعداد تولدها و خطاهای معیار برآوردها استفاده کرد.

```
. use "a:\ hospsamp.dta", clear
. svyset pweight weighta
. svyset strata oblevel
. svyset fpc tothosp
. svytotal births
. svytotal births, by(oblevel)
```

فرمان اول نشان می‌دهد که داده‌ها در فایل اطلاعاتی STATA به نام *hospsamp.dta* قرار دارند.

فرمان دوم نشان می‌دهد که وزن نمونه‌گیری برای بیمارستان در متغیر *weighta* قرار دارد.

فرمان سوم نشان می‌دهد که متغیر نشانگر طبقه در متغیر *oblevel* قرار دارد.

فرمان چهارم نشان می‌دهد که N_h، تعداد بیمارستانهای طبقه در متغیر *tothosp* قرار دارد. در اینجا نیز این فرمان برای گزارشهای متعلق به یک طبقه یکسان است ولی برای گزارشهای مربوط به بیمارستانها در طبقه‌های مختلف یکسان نیست. تصحیح جامعهٔ متناهی از روی این متغیر محاسبه می‌شود.

فرمان پنجم نشان می‌دهد که کل تعداد تولدها باید از روی داده‌های نمونه برآورد شود.

فرمان ششم نشان می‌دهد که کل تعداد تولدها برای هر طبقه باید از این داده‌ها برآورد شود.

خروجی STATA که از فرمانهای بالا تولید شده به شرح زیر است:

Survey total estimation

pweight:	weighta	Number of obs	= 15
Strata:	oblevel	Number of strata	= 3
PSU:	<observations>	Number of PSUs	= 15
FPC:	tothosp	Population size	= 158

Total	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
births	183983	34014.35	109872.1	258093.9	.7035476

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

Survey total estimation

pweight:	weighta	Number of obs	= 15
Strata:	oblevel	Number of strata	= 3
PSU:	<observations>	Number of PSUs	= 15
FPC:	tothosp	Population size	= 158

Total Subpop	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
births					
oblevel = = 1	14931	2669.857	9113.882	20748.12	.1564799
oblevel = = 2	117117	33067.68	45068.7	189165.3	1.089406
oblevel = = 3	51935	7508.403	35575.6	68294.4	.0330073

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

مثال تشریحی با استفاده از SUDAAN: دوباره مثالی را که در آن ۱۵۸ بیمارستان در یک ناحیه بر حسب سطح خدمات زایمان در سه طبقه گروه‌بندی شده‌اند در نظر می‌گیریم، که در آن یک نمونهٔ تصادفی طبقه‌بندی شده گرفتیم و ٤ بیمارستان از طبقهٔ ۱، ۵ بیمارستان از طبقهٔ ۲ و ٦ بیمارستان از طبقهٔ ۳ انتخاب کردیم. این نمونهٔ خاص متشکل از ۱۵ بیمارستان در بحث استفاده از STATA برای تحلیل این داده‌ها در بالا نشان داده شد. حالا فرض می‌کنیم که پروندهٔ اطلاعاتی *hospsamp.dta* یک

پروندهٔ اطلاعاتی SAS PC است (*hospsamp.ssd*). در زیر نشان خواهیم داد که چگونه نرم‌افزار آماری SUDAAN می‌تواند برآوردها و خطاهای معیار را برای این طرح نمونه‌گیری تصادفی طبقه‌بندی شده تولید کند.

دوباره با فرض این که *hospsamp* به پروندهٔ اطلاعاتی SAS ارسال شده است، می‌توانیم از فرمانهای زیر در به دست آوردن برآوردهای کل تولدها برای هر طبقه و برای ۱۵۸ بیمارستانی که جامعه را تشکیل می‌دهند، استفاده کنیم.

```
PROC DESCRIPT DATA = HOSPSAMP FILETYPE = SAS  DESIGN = WOR TOTALS;
NEST OBLEVEL;
WEIGHT WEIGHTA;
TOTCNT TOTHOSP;
VAR BIRTHS;
SUBGROUP OBLEVEL;
LEVELS 3;
SETENV COLWIDTH=20;
SETENV DESWIDTH=3;
```

سطر اول اطلاعاتی دربارهٔ پروندهٔ اطلاعاتی ارائه می‌دهد و حاکی از آن است که نمونه‌گیری بدون جایگذاری است و قرار است برآورد مجموعها را به دست آوریم.

سطر دوم نشان می‌دهد که گزارشهای نمونه در داخل متغیر طبقه *oblevel* جای داده شده‌اند.

سطر سوم متغیری را که قرار است به عنوان وزن نمونه‌گیری به کار رود مشخص می‌سازد که در این مورد *weighta* نامیده می‌شود و وارون کسر نمونه‌گیری خاص طبقه است.

سطر چهارم نشان می‌دهد که متغیر *tothosp* شامل کل تعداد بیمارستانهای جامعهٔ آماری است، (٤٢ بیمارستان برای طبقهٔ ۱، ۹۹ بیمارستان برای طبقهٔ ۲ و ۱۷ بیمارستان برای طبقهٔ ۳).

سطر پنجم نشان می‌دهد که کل تعداد تولدها قرار است برآورد شود.

سطرهای ششم و هفتم نشان می‌دهند که کل تعداد تولدها برای هر طبقه قرار است برآورد شود.

سطرهای ۸ و ۹ فورمت خروجی را تعیین می‌کنند.

خروجی زیر از فرمانهای فوق‌الذکر SUDAAN تولید شده است.

برآوردهای حاصل از SUDAAN و STATA با یکدیگر مطابقت دارند و خواننده می‌تواند تحقیق کند که با برآوردهای حاصل از فرمولهای تابلوی ۲.۵ نیز مطابقت دارند.

□

۷.۵ خلاصه

در این فصل مروری بر طبقه‌بندی داشتیم. مفهوم طبقه‌بندی را معرفی و تعریف کردیم که منظور از نمونهٔ تصادفی طبقه‌بندی شده چیست. مواردی را به بحث گذاشتیم که ممکن بود طبقه‌بندی در آنها

مناسب باشد و مزایای طبقه‌بندی را نسبت به نمونه‌گیری تصادفی ساده و نیز معایب عمدهٔ آن را توضیح دادیم. عبارتهایی را که برای مشخص کردن پارامترهای جامعه به کار می‌روند و آماره‌های تحت نمونه‌گیری تصادفی ساده را ارائه کردیم. سرانجام نشان دادیم که چگونه می‌توان برآوردهایی از طریق نمونه‌گیری تصادفی طبقه‌بندی شده، همراه با خطاهای معیار آنها، با استفاده از دو نرم‌افزار SUDAAN و STATA به دست آورد.

```
1  PROC DESCRIPT DATA = HOSPSAMP FILETYPE = SAS  DESIGN = WOR
   TOTALS;
2  NEST OBLEVEL;
3  WEIGHT WEIGHTA;
4  TOTCNT TOTHOSP;
5  VAR BIRTHS;
6  SUBGROUP OBLEVEL;
7  LEVELS 3;
8  SETENV COLWIDTH=20;
9  SETENV DECWIDTH=3;

Number of observations read : 15 Weighted count: 158
Number of observations skipped : 0
(WEIGHT variable nonpositive)
denominator degrees of freedom : 12

by: Variable, OBLEVEL.
```

Variable		OBLEVEL Total	1
BIRTHS	Sample Size	15.000	4.000
	Weighted Size	158.000	42.000
	Total	183982.904	14931.000
	SE Total	34014.329	2669.857
	Mean	1164.449	355.500
	SE Mean	215.281	63.568

by: Variable. OBLEVEL.

Variable		OBLEVEL 2	3
BIRTHS	Sample Size	5.000	6.000
	Weighted Size	99.000	17.000
	Total	117116.928	51934.977
	SE Total	33067.664	7508.399
	Mean	1183.000	3055.000
	SE Mean	334.017	441.671

تمرین

۱.۵ در یک نمونهٔ تصادفی ساده در مورد۱۰۰۰ دونده از میان ۱۰۰۰۰ دونده‌ای که در مسابقهٔ ماراتون ۱۹۹۵ شیکاگو شرکت کرده بودند، نتیجهٔ آزمایش خون ۳۵ نفر برای مصرف استرویید و سایر داروهایی که فعالیت را زیاد می‌کنند مثبت بود. وقتی نتایج بر حسب زمان تکمیل مسابقه دسته‌بندی شدند جدول زیر به دست آمد.

زمان تکمیل مسابقه (ساعت)	تعداد در نمونه	تعداد نتیجهٔ مثبت برای مصرف دارو	درصد مثبت
کمتر از ۲/۵	۱۰۰	۲۵	۲۵/۰۰
۲/۵– ٤	۵۰۰	۷	۱/٤۰
بیشتر از ٤	٤۰۰	۳	۰/۷۵
جمع	۱۰۰۰	۳۵	۳/۵۰

الف. خطای معیار نسبتی که برای مصرف داروی فعالیت‌زا مثبت برآورد شده چقدر است؟

ب. با بررسی (بدون محاسبه) نرخهای ارائه شده در بالا فکر می‌کنید طبقه‌بندی می‌توانست منجر به برآورد اساساً بهتری شود؟ چرا آری یا چرا نه؟

۲.۵ فرض کنید که در تمرین ۱.۵ از ۱۰۰۰۰ نفری که مسابقهٔ ماراتون را به پایان رساندند ۲۰۰۰ نفر در زمانی کمتر از ۲/۵ ساعت، ٦۰۰۰ نفر بین ۲/۵ تا ٤ ساعت و ۲۰۰۰ نفر در زمانی بیشتر از ٤ ساعت مسابقه را تکمیل کرده باشند. در مورد نتایج نمونه‌گیری به صورتی که در تمرین ۱.۵ ارائه شد اظهار نظر کنید.

۳.۵ اگر یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۳۳۳ نفر در هر یک از گروههای سه‌گانه گرفته می‌شد، برآورد درصد مثبت که از این طریق به دست می‌آمد احتمالاً چقدر می‌شد؟

٤.۵ تعداد ۲۰۰۰ دوندهٔ دیگر نیز وارد بازی شدند ولی آن را به پایان نرساندند. از میان این عده به تصادف برای ٦۰۰ نفر پرسشنامه‌هایی از طریق پست ارسال شد که ۵۰۰ نفر آن را پر کردند. یکی از اقلام این پرسشنامه از پاسخگو می‌خواست که متوسط تعداد مایلهایی را که در هفته طی مدت ۸ هفته قبل از مسابقهٔ ماراتون پیموده است برآورد کند. میانگین، $\bar{x}$، تعداد مایلهای طـــی شده ۳۲/٤ با انحراف معیار S_x معادل ۷/۳ بود. یک نمونهٔ تصادفی ساده متشکل از ۵۰۰ نفر از ۱۰۰۰۰ دونده‌ای که مسابقهٔ ماراتون را تکمیل کرده بودند ٤۰۰ پاسخگو داشت که

متوسط تعداد مایلهای هفتگی آنها ٤٦/٨ با انحراف معیاری معادل ٦/٢ مایل بود. متوسط مایلهای طی شده در هفته طی مدت مورد نظر برای همهٔ کسانی که وارد مسابقهٔ ماراتون شدند چقدر است؟

۵.۵ داده‌های زیر مربوط به ٦ سازمان حفظ بهداشت برای سال ۱۹۸۸ در یک شهر با اندازهٔ متوسط است:

تعداد کارکنان تأمین کنندهٔ مراقبت از بیماران و تعداد رویارویی با بیمار طی سال ۱۹۸۸ توسط ٦ سازمان

HMO: سازمان حفظ بهداشت	تعداد پزشکان تأمین کنندهٔ مراقبت ازبیماران	تعداد رویارویی با بیمار
۱	۱۰	۲۲۰۰۰
۲	٦	۱٤۰۰۰
۳	٤	۱۰۲۰۰
٤	۳۰	۷۰۰۰۰
٥	۷	۱٥۰۰۰
٦	۳	٥۰۰۰

چون این طرح مستلزم استفادهٔ نوبتی از هزینه و زمان است پیشنهاد می‌شود که یک نمونهٔ دوتایی از این سازمانهای حفظ بهداشت برای برآورد کل تعداد رویارویی با بیماران در این شش سازمان طی سال ۱۹۸۸ گرفته شود.

الف. همهٔ نمونه‌های تصادفی سادهٔ دوتایی را شمارش و کل تعداد رویارویی با بیمار در سال ۱۹۸۸ را برای ٦ سازمان برآورد کنید.

ب. خطای معیار برآورد مجموع از روی طرح نمونهٔ تعیین شده در بالا چقدر است؟

پ. سازمانهای شمارهٔ ۱، ۲، ۳، ٥ و ٦ را در یک طبقه و سازمان شمارهٔ ٤ را (به تنهایی) در یک طبقهٔ دیگر گروه‌بندی کنید. از روی این دو طبقه، تمام نمونه‌های تصادفی طبقه‌بندی شدهٔ دوتایی را شمارش کنید. خطای معیار برآورد مجموع حاصل از این برنامهٔ نمونه‌گیری چقدر است؟

ت. در مورد مطلوبیت استفاده از طبقه‌بندی در این وضعیت در مقابل نمونه‌گیری تصادفی ساده اظهار نظر کنید.

۶.۵ فرض کنید که در وضعیت ارائه شده در تمرین ۵.۵، برآورد تعداد رویارویی در سال ۱۹۸۸ به ازای هر پزشک مورد نظر است. قسمتهای (الف) تا (ت) تمرین ۵.۵ را برای این وضعیت برآورد تکرار کنید. آیا در این وضعیت نمونه‌گیری تصادفی طبقه‌بندی شده بهتر از نمونه‌گیری تصادفی ساده است؟ چرا بهتر است و چرا نیست؟

۷.۵ در درمانگاه بزرگی واقع در یک بیمارستان مرکز شهر، ۵۶ بیماری که دارای عفونت بدون علامتهای ظاهری ایدز هستند با یک داروی آزمایشی تحت درمان قرار گرفته‌اند که تصور می‌شود توانایی بازدهی برخی از کارکردهای سیستم دفاعی بدن را که با عفونت HIV همراه است دارند. از این بیماران شمارش گلبولی CD4 برای ۱۲ نفر در ویزیت اول کمتر از ۲۵۰ بود، برای ۲۰ نفر بین ۲۵۰ تا ۴۰۰ بود و برای ۲۴ نفر بیشتر از ۴۰۰ بود. (هر چه این شمارش کمتر باشد وضعیت بیمار وخیمتر است). می‌خواهیم یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۳۰ بیمار یعنی ۱۰ نفر از هر یک از گروههای سه‌گانهٔ بالا بگیریم با این هدف که وقوع پیشامدهای ناشی از ایدز را در یک دورهٔ ۱۲ ماهه در میان این بیماران برآورد کنیم.

الف. چند نمونهٔ متشکل از ۳۰ بیمار به شرحی که در بالا گذشت امکان پذیر است؟

ب. با فرض این که بیمار می‌تواند بیش از یک پیشامد حاکی از ایدز طی این مدت داشته باشد به صورت جبری نشان دهید که چگونه وقوع پیشامدهای حاکی از ایدز را از روی نمونه برای دورهٔ ۱۲ ماهه برآورد می‌کنید.

پ. چند نمونهٔ ۳۰ تایی تحت نمونه‌گیری تصادفی ساده بدون طبقه‌بندی امکان پذیر است؟

۸.۵ در زیر نتایج آمارگیری نمونه‌ای که در تمرین ۷.۵ توصیف شد ارائه شده است:

	تعداد افراد		
تعداد پیشامدها	$CD4 < ۲۵۰$	$۲۵۰ \le CD4 < ۴۰۰$	$CD4 \ge ۴۰۰$
۰	۵	۷	۹
۱	۱	۲	۱
۲	۲	۱	۰
۳	۲	۰	۰

از روی این داده‌ها وقوع پیشامدهای حاکی از ایدز را در جامعهٔ هدف برآورد کنید. خطای معیار برآورد نرخ وقوع چقدر است؟ (باید فرمول خطای معیار را از «اصول اولیه» به دست آورید).

۹.۵ از روی داده‌های تمرین ۸.۵ نسبت بیماران دارای یک یا چند پیشامد حاکی از ایدز را برآورد کنید. خطای معیار این نسبت برآورد شده چقدر است؟

۱۰.۵ یک نمونهٔ تصادفی ساده (بدون طبقه‌بندی) متشکل از ۳۰ نفر از این ٥٦ بیمار، داده‌های زیر را به دست داده است:

	تعداد افراد		
تعداد پیشامدها	CD 4 < ۲۵۰	۲۵۰ ≤ CD 4 < ٤۰۰	CD 4 ≥ ٤۰۰
۰	۳	۸	۱۲
۱	۱	۲	۰
۲	۰	۱	۱
۳	۲	۰	۰

وقوع پیشامدهای حاکی از ایدز را از روی این داده‌ها برآورد کنید و خطای معیار این برآورد را به دست آورید. این نتایج را با نتایج حاصل از طرح نمونه‌گیری طبقه‌بندی شده مقایسه کنید. کدام طرح، برآورد قابل اعتمادتری را تولید می‌کند؟ چرا؟

کتابشناسی

The following articles present sample surveys in which stratification is used.

1. Hemphill, F. M., A sample survey of home injuries, *Public Health Reports,* 67: 1026, 1952.
2. Horvitz, D. G., Sampling and field procedures of the Pittsburgh morbidity survey, *Public Health Reports,* 67: 1003, 1952.
3. Goldberg, J., Levy, P. S., Mullner, R., Gelfand, H., Iverson, N., Lemeshow, S., and Rothrock, J., Factors affecting trauma center utilization in Illinois, *Medical Care* 19: 547, 1981.
4. Stasny, E. A., Toomey, B. G., and First, R. J., Estimating the rate of rural homelessness: A study of nonurban Ohio, *Survey Methodology,* 20: 87, 1994.
5. Barner, B. M., and Levy, P. S., State-wide shoulder belt usage by type of roadway and posted speed limit: A three year comparison, *38th Annual Proceedings Association for the Advancement of Automotive Medicine,* Association for the Advancement of Automotive Medicine, Des Plaines Ill., 1994.

The following software programs have the capability of taking stratified samples.

6. Frankel, M. R., and Spencer, B. D., *SAMPLE. A Supplementary Module for SYSTAT,* SYSTAT, Inc., Evanston, Ill., 1990.
7. SAS Institute Inc., *SAS Language and Procedures,* Usage 2, Version 6, 1st ed., Cary, N.C., SAS Institute Inc., 1991, pp. 649.

In addition, there is material on stratification in virtually every text on sampling theory and survey methodology, including those listed in the bibliography sections of earlier chapters.

فصل ۶

مطالبی بیشتر دربارهٔ نمونه‌گیری تصادفی طبقه‌بندی شده

در فصل قبل مفهوم طبقه‌بندی را معرفی کردیم و شرح دادیم که چرا طبقه‌بندی به عنوان راهبردی در طراحی آمارگیریهای نمونه‌ای به کار می‌رود. همچنین نمادهایی را که آمارشناسان به طور متداول در بحث از مشخصه‌های جامعه به کار می‌برند و شیوه‌های برآورد کردن مناسب برای نمونه‌گیری طبقه‌بندی شده را ارائه کردیم. در این فصل دربارهٔ نوعی از نمونه‌گیری طبقه‌بندی شده به طور مشروح بحث می‌کنیم که نمونه‌گیری تصادفی طبقه‌بندی شده نامیده می‌شود.

۱.٦ برآورد کردن پارامترهای جامعه

برآوردهای میانگینها، نسبتها، و مجموعهای جامعه‌ای در نمونه‌گیری تصادفی طبقه‌بندی شده با محاسبهٔ متوسطهای موزون برآوردهای تکی ویژهٔ هر طبقه و جمع زدن آنها برای تمام طبقات به دست می‌آید. فرمولهای این برآوردها در تابلوی ۲.۵ ارائه شده‌اند. حالا با نگاهی به یک مثال می‌بینیم که چگونه می‌توان از این فرمولها استفاده کرد.

مثال تشریحی: جامعهٔ متشکل از ١٤ خانواده را که در جدول ٦.٥ نشان داده شده است در نظر بگیرید. فرض کنید تصمیم گرفته‌ایم نمونه‌ای متشکل از دو خانواده از طبقهٔ ۱، دو خانواده از طبقهٔ ۲، و چهار خانواده از طبقهٔ ۳ انتخاب کنیم. در آن صورت خواهیم داشت

$n_۱ = ۲,\quad n_۲ = ۲,\quad n_۳ = ٤,\quad N_۱ = ۳,\quad N_۲ = ۵,\quad N_۳ = ٦\quad$ و $\quad N = ۱٤$

فرض کنید عناصر $X_{۱,۲}$ و $X_{۱,۳}$ را از طبقهٔ ۱ (یعنی $x_{۱,۱} = ۳$ و $x_{۱,۲} = ٤$)، $X_{۲,۲}$ و $X_{۲,۵}$ را از طبقهٔ ۲ (یعنی $x_{۲,۱} = ٦$ و $x_{۲,۲} = ۸$) و $X_{۳,۱}$، $X_{۳,۲}$، $X_{۳,۴}$ و $X_{۳,۶}$ را از طبقهٔ ۳ (یعنی $x_{۳,۱} = ۲$، $x_{۳,۲} = ۳$، $x_{۳,۳} = ۲$ و $x_{۳,۴} = ۳$) انتخاب کرده‌ایم.

با استفاده از معادلهٔ (٦.۵) برآورد کل تعداد افراد را در بلوک ۱ به دست می‌آوریم :

$$x'_۱ = \frac{۳\times(۳+۴)}{۲} = ۱۰/۵$$

به همین ترتیب $x'_۲ = ۳۵$ و $x'_۳ = ۱۵$ به دست می‌آیند. سپس با استفاده از معادلهٔ (٦.۵) برآورد کل تعداد افراد در جامعه به شرح زیر به دست می‌آید.

$$x'_{str} = ۱۰/۵ + ۳۵ + ۱۵ = ۶۰/۵$$

با استفاده از معادلهٔ (۷.۵) برآورد میانگین تعداد افراد در هر خانوادهٔ بلوک ۱،

$$\bar{x}_۱ = \frac{۳+۴}{۲} = ۳/۵$$

به همین ترتیب $\bar{x}_۲ = ۷$ و $\bar{x}_۳ = ۲/۵$. پس برآورد میانگین تعداد افراد در هر خانواده از معادلهٔ (۷.۵) عبارت است از

$$\bar{x}_{str} = \frac{۳\times ۳/۵}{۱۴} + \frac{۵\times ۷}{۱۴} + \frac{۶\times ۲/۵}{۱۴} = ۴/۳۲$$

با استفاده از معادلهٔ (۹.۵) واریانس اندازهٔ خانواده در طبقهٔ ۱ عبارت است از

$$S^۲_{۱x} = \frac{(۳-۳/۵)^۲ + (۴-۳/۵)^۲}{۱} = ۰/۵$$

به همین ترتیب $S^۲_{۲x} = ۲$ و $S^۲_{۳x} = ۰/۳۳$.

اگر فرض کنیم که

$$y_{h,i} = \begin{cases} ۰ & \text{اگر اندازهٔ خانواده } ٤ \text{ یا بیشتر باشد} \\ ۱ & \text{اگر اندازهٔ خانواده کمتر از } ٤ \text{ باشد} \end{cases}$$

در آن صورت از معادلهٔ (۸.۵) خواهیم داشت

$$p_{۱y} = \frac{۰+۱}{۲} = ۰/۵$$

به همین ترتیب $p_{۲y} = ۱$ و $p_{۳y} = ۰$ خواهد بود.

□

۲.٦ توزیعهای نمونه‌گیری برآوردها

چـون نمونـهٔ تصادفی طبقه‌بندی شده از L نمونهٔ تصادفی ساده تشکیل شده است که جداگانه و مستقل در داخـل هر طبقه گرفته شده‌اند و چون برآورد میانگین، مجموع یا نسبت جامعه‌ای، یک ترکیب خطی از برآورد میانگینها، مجموعها یا نسبتهای تکی است که از نمونه به دست آمده‌اند به این نتیجه می‌رسیم که میانگین توزیع نمونه‌گیری هر یک از این مقادیر برآورد شده برابر با ترکیب خطی متناظر پارامترهای جامعـه اسـت. به عبارت دیگر، مجموعها، میانگینها و نسبتهای جامعه‌ای، هرگاه به صورتی برآورد شده باشند که در رابطه‌های (٦.۵)، (۷.۵) و (۸.۵) نشان داده شده است تحت نمونه‌گیری تصادفی طبقه‌بندی شده برآوردهای نااریب میانگینها، مجموعها و نسبتهای جامعه‌ای متناظرند.

میانگینها و خطاهای معیار برآوردهای جامعه‌ای در تابلوی ۱.٦ ارائه شده‌اند.

تابلوی **۱.٦ میانگینها و خطاهای معیار برآوردهای جامعه‌ای در نمونه‌گیری تصادفی طبقه‌بندی شده**

میانگین

$$E(x'_{str}) = \sum_{h=1}^{L} E(x'_h) = \sum_{h=1}^{L} X_{h+} = X$$

$$SE(x'_{str}) = N[SE(\bar{x}_{str})] = \left[\sum_{h=1}^{L}(N_h^2)\left(\frac{\sigma_{hx}^2}{n_h}\right)\left(\frac{N_h - n_h}{N_h - 1}\right)\right]^{1/2} \qquad (1.6)$$

مجموع

$$E(\bar{x}_{str}) = \frac{\sum_{h=1}^{L}(N_h)E(\bar{x}_h)}{N} = \frac{\sum_{h=1}^{L} N_h \bar{X}_h}{N} = \bar{X}$$

$$SE(\bar{x}_{str}) = \left[\frac{1}{N^2}\sum_{h=1}^{L}(N_h^2)\left(\frac{\sigma_{hx}^2}{n_h}\right)\left(\frac{N_h - n_h}{N_h - 1}\right)\right]^{1/2} \qquad (2.6)$$

نسبت

$$E(p_{y,str}) = \frac{\sum_{h=1}^{L} N_h P_{hy}}{N} = \frac{\sum_{h=1}^{L} Y_{h+}}{N} = P_y$$

$$SE(p_{y,str}) = \left\{\frac{1}{N^2}\sum_{h=1}^{L}(N_h^2)\left[\frac{P_{hy}(1-P_{hy})}{n_h}\right]\left(\frac{N_h - n_h}{N_h - 1}\right)\right\}^{1/2} \qquad (3.6)$$

نمادهـای مـورد استفاده در این فرمولها در تابلوی ۱.۵ تعریف شده‌اند و n_h، اندازهٔ نمونه برای طبقهٔ h است.

با نگاهی به یک مثال ببینیم این فرمولها چگونه به کار می‌روند.

مـثال تشـریحی: در مثالی که قبلاً به کار گرفته شد، طبقات، سه بلوک شهری (جدول ٦.٥)، خانواده‌ها واحدهـای اولیـه و مـتغیر x انـدازهٔ خـانواده اسـت. یـک نمونهٔ تصادفی طبقه‌بندی شده متشکل از دو خانواده از طبقهٔ ۱، دو خانواده از طبقهٔ ۲ و چهار خانواده از طبقهٔ ۳ گرفته‌ایم. به این ترتیب داریم

$$n_۱=۲\,, \qquad n_۲=۲, \qquad n_۳=٤, \qquad N_۱=۳, \quad N_۲=٥, \quad N_۳=٦$$

$$\sigma^۲_{۱x}=۰/۲۲۲\,, \quad \sigma^۲_{۲x}=۲/٥۶\,, \quad \sigma^۲_{۳x}=۰/۲۲۲\,, \quad P_۱=۰/۶۷\,, \quad P_۲=۱/۰۰\,, \quad P_۳=۰$$

که P_i، نسبت خانواده‌ها در i امین طبقهٔ با چهار نفر یا بیشتـر است. با اسـتـفاده از معادلهٔ (۲.٦) داریم

$$SE(\bar{x}_{str})=\left\{\frac{۱}{۱۴^۲}\left[۳^۲\times\left(\frac{۰/۲۲۲}{۲}\right)\left(\frac{۳-۲}{۳-۱}\right)+٥^۲\times\left(\frac{۲/٥۶}{۲}\right)\left(\frac{٥-۲}{٥-۱}\right)\right.\right.$$
$$\left.\left.+۶^۲\times\left(\frac{۰/۲۲۲}{۴}\right)\left(\frac{۶-۴}{۶-۱}\right)\right]\right\}^{½}=۰/۳٥۹$$

$$SE(x'_{str})=۱۴\times۰/۳٥۹=٥/۰۳$$

از رابطهٔ (۳.٦) داریم

$$SE(p_{y,str})=\left\{\frac{۱}{۱۴^۲}\left[۳^۲\left(\frac{۰/۶۷\times(۱-۰/۶۷)}{۲}\right)\left(\frac{۳-۲}{۳-۱}\right)\right.\right.$$
$$+٥^۲\times\left(\frac{۱\times(۱-۱)}{۲}\right)\left(\frac{٥-۲}{٥-۱}\right)$$
$$\left.\left.+۶^۲\times\left(\frac{۰\times(۱-۰)}{۴}\right)\left(\frac{۶-۴}{۶-۱}\right)\right]\right\}^{½}=۰/۰٥۰۴$$

در این فرمولها y، صفت کیفی، ٤ نفر یا بیشتر بودن افراد یک خانواده است.

□

۳.٦ برآورد کردن خطاهای معیار

در نمونه‌گیری تصادفی طبقه‌بندی شده می‌توان برآورد خطاهای معیار میانگینها، مجموعها و نسبتها را با جایگـزین کـردن $\hat{\sigma}^۲_{hx}$ به جای $\sigma^۲_{hx}$ یا p_{hy} به جای P_{hy} در رابطه‌های (۱.٦)، (۲.٦) و (۳.٦) به دست آورد، که در آنها

$$\hat{\sigma}^۲_{hx}=\frac{(N_h-۱)s^۲_{hx}}{N_h} \qquad (٤.٦)$$

$$s^۲_{hx}=\frac{\sum_{i=۱}^{n_h}(x_{h,i}-\bar{x}_h)^۲}{n_h-۱} \qquad (٥.٦)$$

و p_{hy}، نسبت عناصر مشاهده شدهٔ دارای صفت کیفی y در طبقه است. خطاهای معیار حاصل در تابلوی ۲.۵ ارائه شدند، و دوباره در تابلوی ۲.٦ تکرار می‌شوند.

حالا ببینیم چگونه می‌توانیم برخی از این فرمولها را به کار ببریم.

تابلوی ۲.٦ خطاهای معیار برآورد شده تحت نمونه‌گیری تصادفی طبقه‌بندی شده

مجموع

$$\hat{SE}(\bar{x}_{str}) = \sqrt{\sum_{h=1}^{L} \frac{N_h^2 s_{hx}^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)} \qquad (٦.٦)$$

میانگین

$$\hat{SE}(\bar{x}_{str}) = \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{s_{hx}^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)} \qquad (٧.٦)$$

نسبت

$$\hat{SE}(p_{y.str}) = \sqrt{\left(\frac{N_h}{N}\right)^2 \frac{p_{hx}(1 - p_{hx})}{n_h - 1}\left(\frac{N_h - n_h}{N_h}\right)} \qquad (٨.٦)$$

در این فرمولها N، اندازهٔ جامعه، N_h، تعداد عناصر در طبقهٔ h، و s_{hx} در معادلهٔ (۹.۵) تعریف شده است؛ و p_{hy}، نسبت عناصر مشاهده شدهٔ با صفت کیفی y در طبقهٔ h است.

مثال تشریحی: برای مثال تشریحی بخش ۱.٦ داشتیم:

$\bar{x}_1 = ۳/۵$	$x'_1 = ۱۰/۵$	$S_{1x}^2 = ۰/۵$	$\hat{\sigma}_{1x}^2 = ۰/۳۳$
$\bar{x}_2 = ۷$	$x'_2 = ۳۵$	$S_{2x}^2 = ۲$	$\hat{\sigma}_{2x}^2 = ۱/۶$
$\bar{x}_3 = ۲/۵$	$x'_3 = ۱۵$	$S_{3x}^2 = ۰/۳۳$	$\hat{\sigma}_{3x}^2 = ۰/۲۷۵$

ستون آخر با استفاده از معادلهٔ (٤.٦) به دست آمده است.

همان طور که در مثال بخش ۱.٦ نشان داده شد، برآورد x'_{str} برای مجموع تعداد افراد، X، در ۳ بلوک، از معادلهٔ (٦.۵) عبارت است از

$$x'_{str} = ۱۰/۵ + ۳۵ + ۱۵ = ۶۰/۵$$

و خطای معیار از معادلهٔ (٦.٦) به صورت زیر برآورد می‌شود

$$\hat{SE}(x'_{str}) = \left[(۳)^2\left(\frac{۰/۵}{۲}\right)\left(\frac{۳-۲}{۳}\right) + (۵)^2\left(\frac{۲}{۲}\right)\left(\frac{۵-۲}{۵}\right) + (۶)^2\left(\frac{۰/۳۳}{۴}\right)\left(\frac{۶-۴}{۶}\right)\right]^{1/2} = ۴/۰۹$$

به این ترتیب بازهٔ اطمینان ۹۵ درصدی برای X، مجموع جامعه، چنین است

$$x'_{str} - ۱/۹۶ \times \hat{SE}(x'_{str}) \le X \le x'_{str} + ۱/۹۶ \times \hat{SE}(x'_{str})$$
$$۶۰/۵ - ۱/۹۶ \times ۴/۰۹ \le X \le ۶۰/۵ + ۱/۹۶ \times ۴/۰۹$$
$$۵۲/۴۸ \le X \le ۶۸/۵۲$$

واضح است که این بازه مجموع واقعی جامعه، $X = ۵۴$، را پوشش می دهد.

□

بـه روشی مشابه مثال بالا می‌توان برآوردهایی برای خطاهای معیار میانگینها و نسبتهای برآورد شده به دست آورد و از آنها برای به دست آوردن بازه‌های اطمینان تقریبی استفاده کرد.

٤.٦ برآورد کردن مشخصه‌های زیرگروهها

در فصـل ۳ نشـان دادیـم کـه در نمونه‌گیری تصادفی ساده، میانگینها، مجموعها و نسبتهای برآورد شده بـرای زیرگـروهها بـرآوردهای ناااریب میانگینها، مجموعها و نسبتهای جامعه‌ای متناظر برای زیرگروهها هسـتند. بـه طـوری کـه در مـثال بعـد نشـان داده شده است این موضوع در مورد نمونه‌گیری تصادفی طبقه‌بندی شده الزاماً درست نیست.

مـثال تشـریحی: داده‌هـای ارائـه شـده در جـدول ۱.٦ را در نظـر مـی‌گیـریم. اگر فـرض کـــنیم $\bar{X}_۱$ معـــرف متـــوسط قیمت در پنـــج داروخانهٔ مستقل واقع در دو جامعهٔ ترکیبی باشـــد، می‌بینیـم که ۱۱/٦۰ دلار $= X_۱$. فرض کنید یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از شش داروخانه از طبقهٔ ۱ و سـه داروخانه از طبقهٔ ۲ را برای برآورد $\bar{X}_۱$ انتخاب کنیم. همچنین فرض کنید که از قبل نمی‌دانیم که آیا یک داروخانهٔ مشخص مستقل است یا وابسته به یک زنجیره از داروخانه‌هاست.

$\bar{x}_{۱,str}$، برآورد ما برای $\bar{X}_۱$ از فرمول زیر به دست می‌آید

$$\bar{x}_{۱,str} = \frac{\sum_{h=۱}^{۲} N_h \bar{x}_{۱h}}{N}$$

کـه در آن، $\bar{x}_{۱h}$، بـرآورد میـانگین بـرای داروخانـه‌های مستقل حاصل از نمونهٔ انتخاب شده در طبقهٔ h است.

بـرای یک نمونهٔ شش‌تایی از داروخانه‌هایی که می‌توان در طبقهٔ ۱ گرفت هفت نمونهٔ ممکن وجود دارنـد و بـرای نمونهٔ سه‌تایی در طبقهٔ ۲ می‌توان چهار نمونهٔ ممکن انتخاب کرد. این نمونه‌ها و برآورد میانگین برای هر نمونه در جدول ۲.٦ فهرست شده‌اند.

تعـداد $\binom{۷}{۶} \times \binom{۴}{۳} = ۲۸$ مقـدار ممکن برای $\bar{x}_{۱,str} = \frac{(۷\bar{x}_{۱۱} + ۴\bar{x}_{۱۲})}{۱۱}$ وجود دارد. توزیع نمونه‌گیری $\bar{x}_{۱,str۱}$ در جدول ۳.٦ نشان داده شده است.

جدول ۱.۶ قیمت خرده‌فروشی ۲۰ کپسول آرامبخش در همهٔ داروخانه‌ها در دو جامعه (طبقه)

جامعه	داروخانه	نوع*	قیمت دارو (دلار)
۱	۱	**C**	۱۰/۰۰
	۲	**I**	۹/۰۰
	۳	**I**	۱۲/۰۰
	۴	**I**	۱۱/۰۰
	۵	**C**	۹/۰۰
	۶	**C**	۹/۵۰
	۷	**C**	۹/۹۰
۲	۱	**I**	۱۳/۵۰
	۲	**I**	۱۲/۵۰
	۳	**C**	۱۲/۰۰
	۴	**C**	۱۱/۰۰

* **I** = مستقل؛ **C** = زنجیره‌ای

جدول ۲.۶ نمونه‌های ممکن برای نمونهٔ تصادفی طبقه‌بندی شده

طبقهٔ ۱		طبقهٔ ۲	
داروخانه‌ها در نمونه	$\bar{x}_{1i}$ (دلار)	داروخانه‌ها در نمونه	$\bar{x}_{2i}$ (دلار)
۱,۲,۳,۴,۵,۶	۱۰/۶۷	۱,۲,۳	۱۳/۰۰
۱,۲,۳,۴,۵,۷	۱۰/۶۷	۱,۲,۴	۱۳/۰۰
۱,۲,۳,۴,۶,۷	۱۰/۶۷	۱,۳,۴	۱۳/۵۰
۱,۲,۳,۵,۶,۷	۱۰/۵۰	۲,۳,۴	۱۲/۵۰
۱,۲,۴,۵,۶,۷	۱۰/۰۰		
۱,۳,۴,۵,۶,۷	۱۱/۵۰		
۲,۳,۴,۵,۶,۷	۱۰/۶۷		

میانگین $E(\bar{x}_{1,str})$ توزیع $\bar{x}_{1,str}$ برای ۲۸ نمونه، برابر است با ۱۱/۵۲ دلار که با مقدار $\bar{X}_1$ میانگین قیمت در پنج داروخانهٔ مستقل در دو جامعه یعنی ۱۱/۶۰ دلار برابر نیست.

□

در مثال بالا به صورت تجربی نشان دادیم که در نمونه‌گیری تصادفی طبقه‌بندی شده، برآورد $\bar{x}_{۱,str}$ از میانگین جامعه $\bar{X}_I$ برای زیرگروهی در داخل جامعه الزاماً یک برآورد نااریب برای $\bar{X}_I$ نیست. علت این امر در این واقعیت نهفته است که به میانگینهای نمونهٔ طبقه‌های تکی $\bar{x}_{I_h}$ برای زیرگروه I در ساخت $\bar{x}_{۱,str}$ برحسب $\frac{N_h}{N}$ که نسبت همهٔ عناصری در جامعه است که به طبقهٔ h تعلق دارند وزن داده می‌شود و نه برحسب نسبت کلیهٔ عناصر در زیرگروه I که متعلق به طبقهٔ h است. اگر N_{I_h} معرف تعداد عناصر در طبقهٔ h متعلق به زیرگروه I باشد و اگر N_I معرف تعداد عناصری از جامعه باشد که متعلق به زیرگروه I است (یعنی $N_I = \sum_{h=۱}^{L} N_{I_h}$)، آنگاه لزوماً درست نیست که $\frac{N_{I_h}}{N_I}$ برابر با $\frac{N_h}{N}$ باشد. یعنی نسبتهای عناصر موجود در زیرگروه I در طبقه‌های گوناگون الزاماً با همان نسبتهایی که عناصر موجود در کل جامعه هستند برابر نیستند. به همین دلیل است که برآورد $\bar{x}_{I,str}$ لزوماً برآورد نااریبی از $\bar{X}_I$ نیست. اگر نسبتهای $\frac{N_{I_h}}{N_I}$ معلوم باشند، آنگاه می‌توان از آنها به عنوان وزنهایی در ساخت برآورد نااریب برای $\bar{X}_I$ استفاده کرد. ولی معمولاً این نسبتها معلوم نیستند.

جدول ۳.٦ توزیع نمونه‌گیری برای $\bar{x}_{I,str}$

$\bar{x}_{I,str}$ (دلار)	f	$\bar{x}_{I,str}$ (دلار)	f
۱۱/۵۱۰	۸	۱۲/۰٤٦	۲
۱۱/٦۹۲	٤	۱۱/۵۹۰	۱
۱۱/۳۳۰	٤	۱۱/۲۲۸	۱
۱۱/٤۱۰	۲	۱۱/۲۷۲	۱
۱۱/۰۹۰	۲	۱۲/۲۲۸	۱
۱۰/۹۱۰	۱	۱۱/۸٦٤	۱
مجموع			۲۸

به همان دلیلی که در بالا ارائه شد، نسبت نمونه‌ای $p_{I_{y,str}}$ و برآورد مجموع، $x'_{I,str}$، به طور کلی برآوردهایی نااریب از p_{I_y} و X_I برای زیرگروههای تحت نمونه‌گیری تصادفی طبقه‌بندی شده نیستند.

برآورد کردن مشخصه‌های زیرگروههای یک جامعه تحت نمونه‌گیری تصادفی طبقه‌بندی شده معمولاً با شیوه‌های *برآورد نسبتی* انجام می‌پذیرد که در فصل بعد دربارهٔ آن بحث می‌کنیم.

۵.٦ انتساب نمونه به طبقات

همین که تصمیم به استفاده از نمونه‌گیری طبقه‌بندی شده گرفتیم و به محض اینکه طبقه‌ها و n، تعداد کل عناصر نمونه را تعیین کردیم، تصمیم مهم بعدی که باید اتخاذ شود موضوع *انتساب* یا تعیین تعداد عناصری است که باید از هر طبقه با این شرط گرفته شود که مجموع عنصرهای منتخب در همهٔ طبقه‌ها n باشد. به طوری که در این بخش خواهیم دید اگر در مورد انتساب به دقت اندیشیده شود خطاهای معیار برآورد پارامترهای جامعه ممکن است تا حدی قابل ملاحظه کاهش یابد.

۱.۵.٦ انتساب برابر

در انتساب برابر، از هر طبقه به تعداد مساوی از عناصر، برای نمونه انتخاب می‌شوند. به عبارت دیگر، برای طبقهٔ h *اندازهٔ نمونه* از فرمول زیر به دست می‌آید

$$n_h = \frac{n}{L}$$

اگر هدف اصلی از آمارگیری نمونه‌ای، آزمون فرضهایی دربارهٔ تفاوتهای بین طبقات نسبت به سطوح متغیرهای موردنظر باشد، با این فرض که واریانسها در داخل طبقه برابرند آنگاه انتساب برابر، انتساب منتخب خواهد بود. اگر چنین فرضی را نتوان اعمال کرد، در آن صورت انتساب منتخب برای آزمون چنین فرضهایی از فرمول زیر به دست خواهد آمد

$$n_h = \frac{\sigma_{hx}}{\sum_{h=1}^{L} \sigma_{hx}} \times n$$

به عبارت دیگر، برای آزمون فرضهای مربوط به تفاوتهای بین طبقات، نسبت به سطوح متغیرها باید اندازه‌های نمونه برای هر طبقه متناسب با انحراف معیار متغیر موردنظر در داخل آن طبقه باشد. اگر بین طبقه‌ها از لحاظ انحراف معیارهای درون طبقه تفاوتی وجود نداشته باشد، آنگاه این حالت به انتساب برابر تبدیل خواهد شد. چون در این متن، هدف اصلی ما بیشتر برآورد کردن است تا آزمون کردن فرض، بیش از این دربارهٔ این نوع انتساب بحث نمی‌کنیم.

۲.۵.٦ انتساب متناسب: نمونه‌های خود- موزون

کسر نمونه‌گیری $\frac{n_h}{N_h}$ درانتساب متناسب طوری تعیین می‌شود که برای هر طبقه یکسان باشد، و بدین معنی است که کسر نمونه‌گیری کلی $\frac{n}{N}$ همان کسری است که از هر طبقه گرفته می‌شود. به عبارت دیگر، *تعداد عناصر* n_h که از هر طبقه گرفته می‌شود از فرمول زیر به دست می‌آید

$$n_h = N_h \times \frac{n}{N} \qquad (۹.٦)$$

تحت انتساب متناسب، برآورد میانگین جامعه، $\bar{x}_{str}$، که از رابطهٔ (۷.۵) به دست می‌آمد به صورت زیر تبدیل می‌شود

$$\bar{x}_{str} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} x_{h,i}}{n} \qquad (۱۰.٦)$$

که این فرمول به مراتب ساده‌تر از فرمولی است که در معادلهٔ (۷.۵) ارائه شده است.

حالا فرمول مربوط به $\bar{x}_{str}$ را تحت انتساب متناسب با فرمول کلی مربوط به برآورد میانگین جامعهٔ تحت نمونه‌گیری تصادفی طبقه‌بندی شده مقایسه می‌کنیم. می‌بینیم که در فرمول کلی [رابطهٔ (۷.۵)] مقدار $x_{h,i}$ مشخصهٔ x برای یک عنصر نمونه در نسبت $\frac{N_h}{(n_h N)}$ ضرب می‌شود که الزاماً برای همهٔ طبقه‌ها یکسان نیست. پس، برای به دست آوردن $\bar{x}_{str}$ ضروری است که رد طبقه‌ای را که هر عنصر به آن تعلق دارد حفظ کنیم. از سوی دیگر، رابطهٔ (۱۰.٦) معرف آن است که برای انتساب متناسب باید هر عنصر نمونه در یک نسبت ثابت $\frac{۱}{n}$ ضرب شود بدون توجه به طبقه‌ای که این عنصر به آن تعلق دارد. این نوع برآوردها به عنوان *برآوردهای خود- موزون* معروف‌اند.

این فن انتساب متناسب تا حد بسیار زیادی حجم کارهای دفترداری را در پردازش داده‌ها آسان می‌سازد و در نتیجه هزینه‌های محاسباتی را کاهش می‌دهد. در آمارگیریهای بزرگ که اطلاعات زیادی در مورد هر فرد نمونه جمع‌آوری می‌شود، غالباً از انتساب متناسب به دلیل آسان بودن آن استفاده می‌شود، حتی اگر از لحاظ دقت برآوردها طرح بهینه نباشد.

متذکر می‌شویم که نمونهٔ طبقه‌بندی شده با انتساب متناسب تنها هنگامی خود- موزون خواهد بود که نسبت افراد نمونه *که پاسخ می‌دهند* در داخل طبقه‌ها یکسان باشد. بی‌پاسخی، به خصوص هنگامی که نرخ بی‌پاسخی از طبقه‌ای به طبقهٔ دیگر متفاوت باشد، می‌تواند به شدت بر اعتبار برآورد ارائه شده در عبارت (۱۰.٦) اثر کند. به همین دلیل روشهای برخورد با بی‌پاسخی که در فصل ۱۳ مورد بحث قرار گرفته‌اند از اهمیت قاطع برخوردارند.

حالا به مثالی در مورد انتساب متناسب نگاهی بیندازیم.

مثال تشریحی: داده‌های جدول ٤.٦ را در نظر بگیرید که تعداد بیمارستانهای عمومی واقع در ٤ طبقه را نشان می‌دهد که یکی از طبقه‌های آن از یک یا چند ناحیه جغرافیایی در ایلی‌نوی تشکیل شده است.

فرض کنید می‌خواهیم یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۵۱ بیمارستان از میان ۲۵۵ بیمارستان موجود در جامعهٔ آماری انتخاب کنیم و می‌خواهیم از انتساب متناسب استفاده کنیم. پس با N=۲۵۵ و n=۵۱ از رابطهٔ (۹.٦) خواهیم داشت:

$$n_۱ = (۴۴)\left(\frac{۵۱}{۲۵۵}\right) = ۸/۸ \approx ۹$$

$$n_۲ = (۱۱۶)\left(\frac{۵۱}{۲۵۵}\right) = ۲۳/۲ \approx ۲۳$$

$$n_۳ = (۴۸)\left(\frac{۵۱}{۲۵۵}\right) = ۹/۶ \approx ۱۰$$

$$n_۴ = (۴۷)\left(\frac{۵۱}{۲۵۵}\right) = ۹/۴ \approx ۹$$

جدول ٤.٦ بیمارستانهای عمومی در ایلی‌نوی برحسب طبقه، ۱۹۷۱

طبقه	تعداد بیمارستانهای عمومی
۱	٤٤
۲	۱۱٦
۳	٤۸
٤	٤۷
مجموع	۲۵۵

بـه این ترتیب، ۹ عنصر (بیمارستان) از طبقهٔ ۱، ۲۳ عنصر از طبقهٔ ۲، ۱۰ عنصر از طبقهٔ ۳ و ۹ عنصر از طبقهٔ ٤ می‌گیریم.

کسر نمونه‌گیری در داخل هر طبقه عبارت است از

$$\frac{n_۱}{N_۱} = \frac{۹}{۴۴} = ۰/۲۰۴۵ \qquad \frac{n_۲}{N_۲} = \frac{۲۳}{۱۱۶} = ۰/۱۹۸۳$$

$$\frac{n_۳}{N_۳} = \frac{۱۰}{۴۸} = ۰/۲۰۸۳ \qquad \frac{n_۴}{N_۴} = \frac{۹}{۴۷} = ۰/۱۹۱۵$$

توجه کنید که $\sum_{h=۱}^{۴} n_h = n$.

تفاوتهـای جـزئی در کسـرهای نمونـه‌گیـری میـان طـبقه‌هـا ناشی از این واقعیت است که انتساب موردنیاز که از رابطهٔ (۹.٦) به دست می‌آید الزاماً اعدادی صحیح را نتیجه نمی‌دهد. پس n_h های حاصل همانهـایی هستند که از رابطهٔ (۹.٦) تعیین و به نزدیکترین عدد صحیح قبل یا بعد از خود گرد شده‌اند. این تفاوتهای ناچیز بین کسرهای نمونه‌گیری عموماً در ایجاد برآوردها نادیده گرفته می‌شوند و با نمونه به طور کلی به گونه‌ای رفتار می‌شود که گویی دقیقاً یک نمونهٔ خود- موزون است.

□

در انتساب متناسب، $Var(\bar{x}_{str})$ برآورد میانگین، $\bar{x}_{str}$، که از رابطهٔ (۱۰.٦) با مجموعهٔ n_h برابر با $N_h\left(\frac{n}{N}\right)$ به دست می‌آمد به صورت زیر درمی‌آید

$$Var(\bar{x}_{str}) = \frac{N-n}{N^2}\sum_{h=1}^{L}\left(\frac{N_h^2}{N_h-1}\right)\left(\frac{\sigma_{hx}^2}{n}\right) \qquad (11.6)$$

اگر همهٔ N_h ها به اندازه‌ای معقول بزرگ باشند این عبارت به تقریب زیر تبدیل می‌شود.

$$Var(\bar{x}_{str}) \approx \left(\frac{\sigma_{wx}^2}{n}\right)\left(\frac{N-n}{N}\right) \qquad (12.6)$$

که در آن

$$\sigma_{wx}^2 = \frac{\sum_{h=1}^{L} N_h\sigma_{hx}^2}{N} \qquad (13.6)$$

برآورد نمونه‌ای $Var(\bar{x}_{str})$ از فرمول زیر به دست می‌آید

$$\hat{Var}(\bar{x}_{str}) = \left(\frac{N-n}{N^2}\right)\sum_{h=1}^{L} N_h\left(\frac{s_{hx}^2}{n}\right)$$

توجه کنید که رابطهٔ (۱۲.٦) به صورتی است که بسیار شبیه به فرمول مربوط به واریانس برآورد میانگین، $\bar{x}$، تحت نمونه‌گیری تصادفی ساده است. فرمول مربوط به خطای معیار در تابلوی ۱.۳ ارائه شده است و توان دوم این مقدار، یعنی واریانس، از فرمول زیر به دست می‌آید

$$Var(\bar{x}) = \left(\frac{\sigma_x^2}{n}\right)\left(\frac{N-n}{N-1}\right) \qquad (14.6)$$

تفاوت بین دو فرمول آن است که برای انتساب متناسب در نمونه‌گیری تصادفی طبقه‌بندی شده، واریانس جامعه، σ_x^2، جای خود را به σ_{wx}^2 داده که میانگین موزون متغیرهای فردی σ_{hx}^2 توزیع x بین عناصر داخل هر طبقه است. وزنهای σ_{wx}^2، متناسب با N_h ها، تعداد عناصر در هر طبقه‌اند.

مقایسهٔ رابطه‌های (۱۲.٦) و (۱٤.٦) معرف این است که نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب برآوردی از میانگین را نتیجه می‌دهد که هرگاه σ_{wx}^2 کمتر از σ_x^2 باشد واریانس کمتری نسبت به میانگین به دست آمده از نمونه‌گیری تصادفی ساده خواهد داشت. ولی توجه کنید همان‌طور که در تحلیل روش‌شناسی واریانس دیده می‌شود، واریانس جامعه‌ای σ_x^2 را می‌توان به دو مؤلفهٔ σ_{bx}^2 و σ_{wx}^2 تفکیک کرد

$$\sigma_x^2 = \sigma_{bx}^2 + \sigma_{wx}^2$$

که در آن

$$\sigma_{bx}^2 = \frac{\sum_{h=1}^{L} N_h (\bar{X}_h - \bar{X})^2}{N} \qquad (۱۵.٦)$$

و σ_{wx}^2 از رابطهٔ (۱۳.٦) به دست می‌آید. به این ترتیب، نسبت واریانس برآورد میانگین، $\bar{x}$، تحت نمونه‌گیری تصادفی ساده به نسبت $\bar{x}_{str}$ که برآورد میانگین تحت نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب است از فرمول زیر به دست می‌آید

$$\frac{Var(\bar{x})}{Var(\bar{x}_{str})} = \frac{\sigma_{bx}^2 + \sigma_{wx}^2}{\sigma_{wx}^2} = 1 + \frac{\sigma_{bx}^2}{\sigma_{wx}^2} \qquad (۱٦.٦)$$

این نسبت همیشه بزرگتر از یا برابر با یک است و میزان تفاوت آن نسبت به یک بستگی به اندازهٔ نسبت $\frac{\sigma_{bx}^2}{\sigma_{wx}}$ دارد. هرگاه این نسبت زیاد باشد برآورد میانگین تحت نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب، دارای واریانس کمتری نسبت به برآورد متناظر تحت نمونه‌گیری تصادفی ساده خواهد بود. مؤلفهٔ σ_{bx}^2 نشان‌دهندهٔ *واریانس بین میانگینهای طبقه‌ای* است، در حالی که مؤلفهٔ σ_{wx}^2 نشان‌دهندهٔ *واریانس بین عناصر* داخل همان طبقه است.

اگر میانگینهای طبقه‌ای، $\bar{X}_h$، دارای مرتبهٔ بزرگی یکسانی باشند، آنگاه استفاده از نمونه‌گیری تصادفی طبقه‌بندی شده نسبت به نمونه‌گیری تصادفی ساده هیچ امتیازی نخواهد داشت یا فایدهٔ آن کم خواهد بود. از سوی دیگر، اگر میانگینهای طبقه‌ای بسیار متفاوت باشند احتمال دارد که کاهش در واریانس برآورد میانگین با استفاده از نمونه‌گیری تصادفی طبقه‌بندی شده نسبت به نمونه‌گیری تصادفی ساده قابل توجه باشد. این از لحاظ شهودی هم درست به نظر می‌رسد زیرا هدف طبقه‌بندی، گروه‌بندی عناصر در طبقه‌ها، پیش از نمونه‌گیری بر اساس شباهتهای آنها از لحاظ مقدارهای یک متغیر یا مجموعه‌ای از متغیرهاست. اگر مقادیر متغیرهای مورد اندازه‌گیری برای عناصر داخل هر طبقه بسیار شبیه یکدیگر باشند، به دست آوردن یک نمونهٔ «بد» مشکل خواهد بود، زیرا هر طبقه در نمونه دارای نماینده است. در آن صورت می‌توان با نمونه‌گیری از تعداد کمی از عناصر داخل هر طبقه، برآوردی قابل اعتماد به دست آورد. از سوی دیگر، اگر میانگینهای طبقه‌ای بسیار شبیه باشند در آن صورت دلیلی برای طبقه‌بندی وجود ندارد و تلاش اضافی موردنیاز برای انتخاب یک نمونهٔ طبقه‌بندی شده به برآورد بهتری منجر نخواهد شد.

بـرای راحـتـی کـار، فرمولهای مربوط به برآوردهای پارامترهای جامعه، تحت انتساب متناسب، در تابلوی ۳.٦ خلاصه شده‌اند.

حـالا بـه مـثالی مـی‌پـردازیم تا بررسی کنیم که آیا این احتمال وجود دارد که نمونه‌گیری تصادفی طبقه‌بندی شده واریانسی کمتر از واریانس حاصل از نمونه‌گیری تصادفی ساده داشته باشد.

مـثال تشـریحی: فـرض کنید می‌خواهیم متوسط تعداد پذیرشهای بیمارستانی را برای بیماریهای جدی روانـی بـه ازای هـر بخـش در بیـن ۸۲ بخش ایالت ایلی‌نوی که بیمارستان عمومی دارند برآورد کنیم. نمونـه‌ای از بخشها گرفته خواهد شد و سوابق پذیرش همهٔ بیمارستانهای موجود در بخشهای نمونه در یافـتن پذیـرش بـرای بیماریهـای جدی روانی مورد بررسی قرار خواهند گرفت. اگر این فرض معقول باشـد که ممکن است همبستگی شدیدی بین تعداد تختهای بیمارستانهای داخل یک بخش و تعداد کل پذیرشـها بـرای بیماریهـای جـدی روانـی وجـود داشته باشد، در آن صورت طبقه‌بندی برحسب تعداد تخـتهای بیمارسـتانی منطقی خواهد بود. بنابراین برنامهٔ نمونه‌گیری انتخاب شده است. در جـدول ۵.٦ بخشـهای ایـلی‌نـوی براساس تعداد تختهای بیمارستانی در دو طبقه گروه‌بندی شده‌اند. طبـقهٔ ۱ شامل آن دسـته از بخشـهایی اسـت کـه دارای ۱ تا ۳۹۹ تخت و طبقهٔ ۲ شامل آنهایی است کـه دارای ٤۰۰ تخت و بیشترند.

از جدول ۵.٦ موارد زیر را حساب می‌کنیم:

$\bar{X}_۱ = ۱۲۳/۹۱$	$\sigma^۲_{۱x} = ۶۱۳۱/۶۳$	$N_۱ = ۶۵$
$\bar{X}_۲ = ۸۷۱/۵۹$	$\sigma^۲_{۲x} = ۷۷۲۸۷/۹۲$	$N_۲ = ۱۷$
$\bar{X} = ۲۷۸/۹۲$	$\sigma^۲_{x} = ۱۱۲۷۵۱/۹۳$	$N = ۸۲$

در این مثال، تعداد تخت $=x$

از رابطه‌های (۱۵.٦) و (۱۳.٦) داریم

$$\sigma^۲_{bx} = \frac{۶۵\times(۱۲۳/۹۱-۲۷۸/۹۲)^۲ + ۱۷\times(۸۷۱/۵۹-۲۷۸/۹۲)^۲}{۸۲} = ۹۱۸۶۸/۳۹$$

$$\sigma^۲_{wx} = \frac{۶۵\times۶۱۳۱/۶۳ + ۱۷\times۷۷۲۸۷/۹۲}{۸۲} = ۲۰۸۸۳/۵۴$$

به این ترتیب، از رابطهٔ (۱٦.٦) نسبت واریانسها، یعنی $\frac{Var(\bar{x})}{Var(\bar{x}_{str})}$ به صورت زیر به دست می‌آید

تابلوی ۳.٦ برآوردهای پارامترهای جامعه تحت نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب

در انتساب متناسب $\left[n_h = \left(\frac{N_h}{N}\right) n \right]$ یعنی از فرمولهای زیر برای برآورد پارامترهای جامعه استفاده می‌شود

مجموع

$$x'_{str} = N \times \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h,i}}{n}$$

$$\hat{Var}(x'_{str}) = N^2 \times \left(\frac{N-n}{N^2}\right) \sum_{h=1}^{L} N_h \times \left(\frac{s_{hx}^2}{n}\right)$$

میانگین

$$\bar{x}_{str} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h,i}}{n}$$

$$\hat{Var}(\bar{x}_{str}) = \left(\frac{N-n}{N^2}\right) \sum_{h=1}^{L} N_h \times \left(\frac{s_{hx}^2}{n}\right)$$

نسبت

$$p_{y,str} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{h,i}}{n}$$

$$\hat{Var}(p_{y,str}) = \left(\frac{N-n}{N^2}\right) \sum_{h=1}^{L} N_h \times \left[\frac{p_{hy}(1-p_{hy})}{n-1}\right]$$

تمام کمیتها در این عبارتها مانند تابلوی ۲.۵ تعریف می‌شوند.

$$\frac{Var(\bar{x})}{Var(\bar{x}_{str})} = 1 + \frac{۹۱۸۶۸/۳۹}{۲۰۸۸۳/۵۴} = ۵/۴۰$$

بنابراین نتیجه می‌گیریم که از لحاظ کاهش واریانس برآورد میانگین، کار طبقه‌بندی احتمال دارد در این وضعیت بسیار سودمند باشد، در صورتی که واقعاً پذیرشها برای بیماری جدی روانی و تعداد تخت به هم بستگی داشته باشند، زیرا واریانس در طبقه‌بندی کمتر از ۲۰ درصد واریانس در نمونه‌گیری تصادفی ساده است.

□

جدول ۵.٦ طبقه‌ها برای تعداد تختهای بیمارستانی برحسب بخش
در بین بخشهای دارای بیمارستانهای عمومی در ایلی‌نوی (به استثنای بخش کوک)

طبقهٔ ۱ (دارای ۱ تا ۳۹۹ تخت)				طبقهٔ ۲ (دارای ٤۰۰ تخت و بیشتر)	
بخش	تعداد تخت	بخش	تعداد تخت	بخش	تعداد تخت
۱	۲۱٦	۳٤	۱۱۳	۱	۸۲۳
۲	۱۷۰	۳۵	٦٤	۲	۱۳٤۳
۳	۲۵۲	۳٦	۱۰۰	۳	۹۰۸
٤	۳۸	۳۷	۵۸	٤	٦٤۸
۵	۱۷۰	۳۸	۵٤	۵	۱۰٤۳
٦	۱۷۹	۳۹	۸۲	٦	۱۳۲۵
۷	۳۱	٤۰	۳۵	۷	۱۱۲۳
۸	٤۰	٤۱	۲۰٤	۸	٦۹۰
۹	۲۹۵	٤۲	٤۲	۹	۵۱۹
۱۰	۳۳٦	٤۳	۷۲	۱۰	۱۱۱۸
۱۱	۱٦٦	٤٤	۳۹	۱۱	۵۲۲
۱۲	۱۲۱	٤۵	۱٤٤	۱۲	۷۱۵
۱۳	۲۸۰	٤٦	۲۱۰	۱۳	۸۵۱
۱٤	۱۸۸	٤۷	۱٦۰	۱٤	۵۵۲
۱۵	۳۵	٤۸	۲۰٤	۱۵	٤۷۰
۱٦	۱۳٤	٤۹	۲۰۰	۱٦	۱۱۸۷
۱۷	٦۳	۵۰	۱۹۵	۱۷	۹۸۰
۱۸	۲۸۰	۵۱	۱٤۰		
۱۹	۷۵	۵۲	۹٦		
۲۰	۱٤۲	۵۳	۱۰۸		
۲۱	۲۹۳	۵٤	۱۲۱		
۲۲	۱۵۲	۵۵	٦۱		
۲۳	۱۰۳	۵٦	٦۳		
۲٤	۲٦۲	۵۷	۱۰٤		
۲۵	۱۰۵	۵۸	۱۵۰		
۲٦	۸۰	۵۹	٤۸		
۲۷	۵٤	٦۰	٤۸		
۲۸	٦٦	٦۱	٦۹		
۲۹	۵٤	٦۲	۷۹		
۳۰	۵۰	٦۳	۷۵		
۳۱	۵۰	٦٤	۳۲		
۳۲	۱٦۵	٦۵	۳۹		
۳۳	۲۰۰				

۳.۵.٦ انتساب بهینه

انتساب متناسب غالباً آن نوع انتسابی نیست که به برآوردی از مجموع، میانگین، یا نسبت منجر شود که کمترین واریانس را در میان تمام راههای ممکن برای انتساب کل نمونهٔ n عنصری در بین N طبقه داشته باشد. می‌توان نشان داد که انتساب n واحد نمونه به هر طبقه که برآورد مجموع، میانگین یا نسبت را با کمترین واریانس برای متغیر x نتیجه می‌دهد از فرمول زیر به دست می‌آید:

$$n_h = \left(\frac{N_h \sigma_{hx}}{\sum_{h=1}^{L} N_h \sigma_{hx}} \right)(n) \qquad (۱۷.٦)$$

مثال تشریحی: از داده‌های جدول ۵.٦ استفاده کرده فرض می‌کنیم رابطهٔ نزدیکی بین تعداد تخت و تعداد پذیرش برای بیماری روانی جدی وجود داشته باشد. انتساب ۲۵ عنصر نمونه از رابطهٔ (۱۷.٦) به شرح زیر برآوردی از میانگین تولید خواهد کرد که کمترین واریانس را خواهد داشت:

$$n_1 = \left(\frac{۶۵\sqrt{۶۱۳۱/۶۳}}{۶۵\sqrt{۶۱۳۱/۶۳} + ۱۷\sqrt{۷۷۲۸۷/۹۲}} \right) \times ۲۵ = ۱۲/۹۶ \approx ۱۳$$

$$n_2 = \left(\frac{۱۷\sqrt{۷۷۲۸۷/۹۲}}{۶۵\sqrt{۶۱۳۱/۶۳} + ۱۷\sqrt{۷۷۲۸۷/۹۲}} \right) \times ۲۵ = ۱۲/۰۴ \approx ۱۲$$

به این ترتیب، انتساب بهینهٔ ۲۵ عنصر نمونه عبارت است از ۱۳ عنصر از طبقهٔ ۱ و ۱۲ عنصر از طبقهٔ ۲. با انتساب متناسب، ۲۰ عنصر برای طبقهٔ ۱ و ۵ عنصر برای طبقهٔ ۲ تعیین شد.

خطای معیار برآورد میانگین تحت نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب بهینه از رابطهٔ (۲.٦) به دست می‌آید، که به طور کلی برای هر نوع انتساب تحت نمونه‌گیری تصادفی طبقه‌بندی شده معتبر است. برای داده‌های جدول ۵.٦، با استفاده از تعداد تخت بیمارستانی به عنوان مشخصهٔ موردنظر، از روی معادلهٔ (۱.٦) برای انتساب بهینه خواهیم داشت

$$SE(\bar{x}_{str}) = \left\{ \left(\frac{۱}{۸۲^2} \right) \left[(۶۵)^2 \left(\frac{۶۱۳۱/۶۳}{۱۳} \right) \left(\frac{۶۵ - ۱۳}{۶۵ - ۱} \right) + (۱۷)^2 \left(\frac{۷۷۲۸۷/۹۲}{۱۲} \right) \left(\frac{۱۷ - ۱۲}{۱۷ - ۱} \right) \right] \right\}^{۱/۲} = ۱۸/۰۹$$

برای انتساب متناسب، با گرفتن ریشهٔ دوم عبارت معادلهٔ (۱۱.٦)، خواهیم داشت

$$SE(\bar{x}_{str}) = \left[\left(\frac{۸۲ - ۲۵}{۸۲^{۲}}\right)\left\{\left(\frac{۶۵^{۲}}{۶۴}\right)\left(\frac{۶۱۳۱/۶۳}{۱۳}\right) + \left(\frac{۱۷^{۲}}{۱۶}\right)\left(\frac{۷۷۲۸۷/۹۲}{۲۵}\right)\right\}\right]^{۱/۲} = ۲۴/۷۱$$

بـه این ترتیب می‌بینیم که برای این داده‌ها، برآورد میانگین، تحت انتساب بهینه دارای خطای معیاری به مراتب کمتر از خطای معیار تحت انتساب متناسب است

□

از رابطـهٔ (۱۷.٦) می‌بینیم که تعداد بهینهٔ عناصر نمونه که باید از یک طبقهٔ معین گرفته شود با N_h، کـل تعـداد عناصـر در طبقهٔ h، و با σ_{hx}، انحراف معیار توزیع x بین همهٔ عناصر در این طبقه متناسب اسـت. ایـن امـر از نظـر شـهودی هم درست است. میانگین جمعیت، $\bar{X}$، که سعی در برآورد آن داریم برابـر است با $\sum_{h=۱}^{L} N_h \bar{X}_h / N$، یا به عبارت دیگر $\bar{X}$، متوسط موزون میانگینهای طبقه‌های تکی یعنی $\bar{X}_h$ هاست که وزن هر طبقه متناسب با کل تعداد عناصر در طبقه است. چون طبقه‌های دارای بیشترین تعـداد عناصر، بیشترین اهمیت را در تعیین میانگین جامعه دارند منطقی است که بیشترین اهمیت را در بـرآورد آن از یک نمونه نیز داشته باشند. اگر توزیع مشخصهٔ x بین عناصر موجود در یک طبقهٔ خاص، دارای انحـراف معیـار کمـی باشد، آنگاه فقط تعداد کمی از عناصر نمونه، نسبت به تعداد مورد نیاز در طـبقه‌ای که در آن توزیع x دارای انحراف معیار زیادی است لازم است، تا برآوردی قابل اعتماد از یک پارامتر طبقه به دست دهد. این واقعیت در فرمول انتساب بهینه مدنظر بوده است، زیرا انتساب نمونه به صـورتی که در رابطهٔ (۱۷.٦) ارائه شده متناسب با اندازهٔ انحراف معیار σ_{hx} است. همچنین توجه کنید کـه اگـر توزیـع x در داخـل طـبقه‌هـا انحـراف معیار برابر داشته باشد، در آن صورت انتساب بهینه به صـورتی کـه در رابطـهٔ (۱۷.٦) ارائـه شده است به انتساب متناسب به صورتی که در رابطهٔ (۹.٦) آمده است تبدیل می‌شود.

٤.٥.٦ انتساب بهینه و اقتصاد

حـالا فـــرض کـنید کـه هزینهٔ نمونه‌گیری یک واحد اولیه برای هر طبقه یکسان نباشد. در آن صورت *کـــل هزینهٔ* C برای گرفتن نمونه‌ای $n_۱$ عنصری از طبقهٔ ۱، $n_۲$ عنصری از طبقهٔ ۲ و همین طور تا آخر، از فرمول زیر به دست می‌آید

$$C = \sum_{h=۱}^{L} n_h C_h$$

که در آن، C_h، هزینهٔ نمونه‌گیری یک واحد اولیه در طبقهٔ h است.

بـرای یک اندازهٔ معین نمونه، n، انتسابی که برآوردی با کمترین واریانس را به ازای واحد هزینه به دست می‌دهد از فرمول زیر نتیجه می‌شود

$$n_h = \frac{\frac{N_h \sigma_{hx}}{\sqrt{C_h}}}{\sum_{h=1}^{L}\left(\frac{N_h \sigma_{hx}}{\sqrt{C_h}}\right)} \times n \qquad (۱۸.٦)$$

همچـنین اگـر کـل هزیـنهٔ گرفـتن نمونـه در حـد C ثابت باشد، انتسابی که برآورد میانگین با کمترین انحراف معیار در سطح هزینهٔ ثابت C را نتیجه می‌دهد از فرمول زیر به دست می‌آید

$$n_h = \frac{\frac{N_h \sigma_{hx}}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h \sigma_{hx} \sqrt{C_h}} \times C \qquad (۱۹.٦)$$

هـر دو رابطـهٔ (۱۸.٦) و (۱۹.٦) نمونـه‌هایی به اندازه‌های n_h انتخاب می‌کنند که با N_h و σ_{hx} نسبت مستقیم و با هزینهٔ C_h نمونه‌گیری از یک عنصر در یک طبقهٔ خاص نسبت معکوس دارند.

اینک مثالی از انتساب بهینه را با توجه به هزینه در نظر می‌گیریم.

مـثال تشریحی: فرض می‌کنیم شرکتی ۲٦۰۰۰۰ پروندهٔ تصادف برای یک دورهٔ زمانی در اختیار دارد و یک آمارگیری نمونه‌ای به منظور برآورد متوسط تعداد روزهای کاری تلف شده به ازای هر تصادف در دسـت تهیـه اسـت. از ایـن ۲٦۰۰۰۰ گـزارش تصـادف، ۱۵۰۰۰۰ گـزارش کدگذاری شده و ۱۱۰۰۰۰ گـزارش کدگذاری نشده‌اند. فرمهای کدگذاری شده را می‌توان مستقیماً با رایانه پردازش کرد، در حالی که فرمهای کدگذاری نشده ابتدا باید پیش از پردازش کدگذاری شوند. برای انتخاب نمونه و کدگذاری و پردازش داده‌ها تقـریباً ۱۰۰۰۰ دلار تأمین شده است. با توجه به این امر، پیدا کردن بهترین راه برای انتساب عناصر نمونه به فرمهای کدگذاری شده و کدگذاری نشده مطلوب است.

براساس نمادگذاری نمونه‌گیری طبقه‌بندی شده، داریم

N_1=۱۵۰۰۰۰ فرم کدگذاری شده (طبقهٔ ۱)

N_2=۱۱۰۰۰۰ فرم کدگذاری نشده (طبقهٔ ۲)

C=۱۰۰۰۰ دلار

فـرض کـنیم هزیـنهٔ نمونـه‌گیـری و پـردازش فـرمهای نمونه به ترتیب برابر با ۰/۳۲ دلار برای هر فرم کدگذاری شده و ۰/۹۸ دلار برای هر فرم کدگذاری نشده باشد، یعنی

C_1 =۰/۳۲ دلار و C_2 =۰/۹۸ دلار

اگـر فـرض کـنیم انحراف معیار توزیع روزهای کاری تلف شده درگزارشهای کدگذاری نشده دو برابر گزارشهای کدگذاری شده است (یعنی $\sigma_{۱x}=\frac{\sigma_{۲x}}{۲}$)، آنگاه از رابطهٔ (۱۹.٦) خواهیم داشت

$$n_۱=\frac{\frac{۱۵۰\cdots\times(\frac{\sigma_{۲x}}{۲})}{\sqrt{۰/۳۲}}}{۱۵۰\cdots\times(\frac{\sigma_{۲x}}{۲})\times\sqrt{۰/۳۲}+۱۱۰\cdots\times\sigma_{۲x}\times\sqrt{۰/۹۸}}\times۱۰\cdots\approx ۸۷۶۲$$

$$n_۲=\frac{\frac{۱۱۰\cdots\times\sigma_{۲x}}{\sqrt{۰/۹۸}}}{۱۵۰\cdots\times(\frac{\sigma_{۲x}}{۲})\times\sqrt{۰/۳۲}+۱۱۰\cdots\times\sigma_{۲x}\times\sqrt{۰/۹۸}}\times۱۰\cdots\approx ۷۳۴۳$$

بـه ایـن ترتیب باید نمونه‌ای متشکل از ۸۷٦۲ گزارش کدگذاری شده و ۷۳٤۳ گزارش کدگذاری نشده انتخاب کنیم.

مـی‌توانیـم بـا جایگـزین کردن مقدارهای مربوط به $C_۱$، $C_۲$، $n_۱$ و $n_۲$ در رابطهٔ مربوط به محاسبهٔ C تحقیق کنیم که کل هزینهٔ نمونه‌گیری برابر با ۱۰۰۰۰ دلار است.

$$C=۸۷۶۲\times۰/۳۲+۷۳۴۳\times۰/۹۸=۱۰\cdots\text{ دلار}$$

□

مـتوجه مـی‌شـویم کـه برا ی به دست آوردن انتساب بهینه، نیازی به دانستن مقادیر واقعی σ_{hx}ها نیسـت. اگـر بتوانیم هر σ_{hx} را به صورت یکی از آنها بیان کنیم (برای مثال σ_{rx}) همان طور که در مثال بـالا بحـث شـد، در آن صـورت σ_{rx} بـه عنوان یک عامل مشترک هم در صورت و هم در مخرج کسر ظاهر می‌شود و بنابراین می‌تواند حذف شود.

مشـکلی کـه غالـباً در انتساب بهینه با آن مواجه می‌شویم، چه هزینه‌ها در نظر گرفته شود یا نشود، ایـن است که اندازهٔ نمونه‌ای بهینهٔ n_h ممکن است از N_h یعنی کل تعداد عناصر موجود در طبقه بیشتر باشد. هرگاه چنین موردی پیش بیاید، برای هر طبقه‌ای که انتساب بهینهٔ آن بیشتر از N_h به دست آمده اسـت n_hرا بـا N_h برابـر مـی‌گیـریم. سپس نمونه‌های باقیمانده را به صورتی که به وسیلهٔ الگوریتم به دست آوردن انتساب نمونه مشخص شده است مجدداً به سایر طبقه‌ها اختصاص می‌دهیم.

برای مثال، خلاصه داده‌هایی را از سه طبقه در نظر می‌گیریم:

طبقه	N_h	σ_{hx}
۱	۱۰۰	۵۰
۲	۱۱۰	۱۰
۳	۱۲۰	۵۰

اگـر بخواهیـم کـل نمونـهٔ ۱٤۰ عنصـری را بـا اسـتفاده از انتساب بهینه به هر طبقه اختصاص دهیم، با استفاده از رابطهٔ (۱۷.٦) خواهیم داشت

$$n_۱ = ۱۰٤ \qquad n_۲ = ۲۳ \qquad n_۳ = ۱۳$$

در آن صـورت $n_۱ = N_۱ = ۱۰۰$ را می‌گیریم و چهار عنصر باقیمانده را طبق رابطهٔ (۱۷.٦) به شرح زیر به طبقه‌های ۲ و ۳ اختصاص می‌دهیم:

$$n_۲ = \left[\frac{۱۱۰\times۱۰}{۱۱۰\times۱۰+۱۲۰\times۵}\right](۴) = ۲/۶ \approx ۳$$

$$n_۳ = \left[\frac{۱۲۰\times۵}{۱۱۰\times۱۰+۱۲۰\times۵}\right](۴) = ۱/۴ \approx ۱$$

به این ترتیب انتساب بهینهٔ نهایی عبارت است از

$$n_۱ = ۱۰۰ \qquad n_۲ = ۲٦ \qquad n_۳ = ۱٤$$

در بـرنامه‌ریـزی آمارگیـری نمونـه‌ای که برای آن نمونه‌گیری تصادفی طبقه‌بندی شده در نظر گرفته شـده اسـت، محاسـبهٔ انتسـاب بهیـنه برای مهمترین متغیرهای آمارگیری غالباً راهبرد خوبی است. اگر انتسـاب بیـن متغیرها متفاوت است باید نوعی مصالحه در انتساب را در نظر گرفت (مانند میانگین n_h بهیـنه بـرای تمـام متغیرهای حایز اهمیت). همچنین، انتساب متناسب نیز باید تا حدودی در نظر گرفته شـود. اگـر خطاهـای معیار پیش‌بینی شده تحت انتساب متناسب از خطاهای معیار پیش‌بینی شده تحت انتسـاب بهینه چندان بیشتر نباشد آنگاه سادگی و راحتی انتساب متناسب می‌تواند اندکی کاهش خطای معیار تحت انتساب بهینه را جبران کند و انتساب متناسب می‌تواند بهترین انتخاب باشد.

مـثال تشـریحی: مطالعـهٔ موردی. این مثال از مطالعه‌ای گرفته شده است که اخیراً در مورد دوقلوهای سـالخورده بـه عمـل آمـده و اسـتفاده از انتسـاب بهینه را در نمونه‌گیری تصادفی طبقه‌بندی شده نشان مـی‌دهـد. هـدف ایـن بررسـی آزمـودن روش شناسایی دوقلوهای سالخورده (٦۵ ساله به بالا) از روی فهرسـتهای بیمـه‌شـدگان زندهٔ خدمات درمانی بود. مطالعات مربوط به دوقلوهای توأمان و ناتوأمان در فـراهم کـردن بینشی نسبت به سهم نسبی عوامل ژنتیکی و غیرژنتیکی مؤثر در سلامت و بیماری بسیار سـودمند اسـت و دوقلوهایی که به این ترتیب شناسایی می‌شدند برای مشارکت احتمالی در بررسیهای پزشکی آینده در یک بایگانی به ثبت می‌رسیدند.

چـون تقریـباً یکـی از هر ۱۰۰ زایمان ممکن است چندقلو باشد تلاش برای غربالگری فهرستهای غیرگزینشـی افـراد بـرای شناسـایی دوقـلوها بـه صورتی بازدارنده پرخرج خواهد بود. ولی می‌توان از مشخصـه‌های دوقلوها به شرح زیر برای به دست آوردن فهرستهای اصلاح شده‌ای که ظاهر شدن تولد دوقلوها در آن بیشتر باشد استفاده کرد:

۱. هـر دو عضـو از یک جفت دوقلو (به استثنای موارد بسیار نادر) در یک محل به دنیا آمده‌اند و تاریخ تولد آنها یکی است.
۲. هر دو نفر از یک نژادند.
۳. هر دو نفر از یک جفت دوقلوی پسرـ پسر دارای یک نام خانوادگی و نامهای مختلف خواهند بود.
٤. فـرض بـر ایـن بـود که به احتمال بسیار زیاد شمارهٔ تأمین اجتماعی (#SS) اعضای یک جفت دوقلو بسیار نزدیک به یکدیگر خواهد بود.

بـا در نظـر گرفـتن ایـن موضوع، مجموعه داده‌های زیر از یک بایگانی که شامل تقریباً ۱۰ میلیون بیمه شدهٔ زندهٔ خدمات درمانی بود استخراج شد:

- از میان بیمه‌شدگان مرد، که در قید حیات بودند جفتهایی در نظر گرفته شدند که شامل افرادی بـا تـاریخ تولـد یکسان، وضعیت تولد یکسان، نام خانوادگی یکسان و نامهای متفاوت بودند. سـپس به جفتهایی که به این ترتیب به دست آمدند شماره‌هایی اختصاص یافت که تفاوت بین شـمارهٔ تـأمین اجـتماعی (#SS) آنـان را نشـان مـی‌داد. این تفاوت (که *تفاوت دنباله‌ای* نامیده می‌شد) با استفاده از الگوریتم پیچیده‌ای به دست آمد و اندازهٔ این تفاوت دنباله‌ای متناسب بود بـا طـول زمانی که تاریخ صدور کارت تأمین اجتماعی برای هر یک از اعضای هر جفت را از هم جدا می‌کرد.
- از میـان بیمـه‌شـدگان زن، جفـتهای زن ـ زن از روی سوابقی ساخته شدند که شامل افرادی با تـاریخ و وضـــعیت تولد یکســـان بودند و هفت رقـم اول کارت تأمین اجتماعی (#SS) آنان نیـــز یکـی بـود. (از نام خانوادگی نمی‌شد مانند مردان استفاده کرد زیرا نام خانوادگی زنان با ازدواج تغییر می‌کند.)
- جفـتهای مـرد ـ زن نیـز بـا همـان الگوریـتم مورد استفاده در تشکیل جفتهای زن ـ زن ساخته شدند.

مجموعـهٔ داده‌هـایی کـه بـه شـرح بـالا بـه وجـود آمد شامل ۲۵۵۸٤۸ گزارش جفت شده بود که برحسـب نـژاد (سفیدپوسـت/ امـریکایی افریقایی‌الاصل) و جنس (مرد ـ مرد، مرد ـ زن، زن ـ زن) در شـش طـبقه رده‌بـندی شـده بودند. هر یک از این شش گروه براساس اندازهٔ تفاوت دنباله‌ای در شمارهٔ تـأمین اجتماعی به سه ردهٔ فرعی تقسیم شده بود. (چار ک اول/ چارکهای دوم و سوم/ چارک چهارم). آن جفـتهایی کـه تفاوتهای دنباله‌ای آنان در چارک اول قرار می‌گرفت دارای شماره‌های تأمین اجتماعی

بودنـد کـه از نظـر زمـانی تقریـباً نزدیک به یکدیگر صادر شده بودند و همین طور تا آخر. جدول ٦.٦ تعداد جفتهایی را نشان می‌دهد که در هر یک از ۱۸ «طبقهٔ» تعریف شدهٔ بالا به دست آمد.

جدول ٦.٦ تعداد جفتهای موجود در هر یک از ۱۸ طبقه
برحسب جنس ـ نژاد ـ چارکهای تفاوت دنباله‌ای

	چندک	جنس		
نژاد	تفاوت دنباله‌ای	مرد- مرد	مرد- زن	زن- زن
سفیدپوست	۱	۳۹۸۷۲	۱۱۲٦۳	۱۰۰۲٤
	۲-۳	۷۹۷۲۷	۲۲۵۰۱	۲۰۰۳۱
	٤	۳۹۸۷۲	۱۱۲٦۳	۱۰۰۲٤
امریکایی افریقایی‌الاصل	۱	۲۵٤٦	۹۹	۱۵۳
	۲-۳	۵۰۷٦	۲۹۹	۳۰۰
	٤	۲۵٤٦	۹۹	۱۵۳

قـرار شـد یـک آمارگیـری مقدمـاتی از تقریباً ۱۰۰۰ جفت به عمل آید که هدف آن برآورد نسبت جفتهای واقعاً دوقلو در هر یک از گروههای شش‌گانهٔ نژاد ـ جنس موجود در پایگاه اطلاعاتی بود. این کـار از این نظر مهم بود که آزمون می‌کرد آیا این روش‌شناسی یک پایگاه اطلاعاتی با تعداد بسیار زیاد از تولـد دوقـلوها را تولیـد خواهـد کـرد یـا نـه. قـرار شـد نمونه‌ای از جفتها گرفته شود و از هر فرد نمونـه‌گیری شده در مورد وضعیت جفت دوقلوی وی سؤالاتی بپرسند. طرح نمونه برای این آمارگیری مقدمـاتی قرار شد نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب بهینه باشد که باید جداگانه در هر یک از گـروههای شـش‌گانـهٔ نـژاد ـ جنس به کار گرفته می‌شد. با در نظر داشتن این موارد، فرمول انتساب نمونه به سه گروه تفاوت دنباله‌ای در شمارهٔ تأمین اجتماعی از معادلهٔ (۱۷.٦) به دست می‌آید که در آن $\sigma_{hx} = \sqrt{p_{hx}(1-p_{hx})}$، که p_{hx}، نسبت دوقلوها در طبقهٔ $h(h=1,2,3)$ است.

نسـبتها، p_{hx} ها، نامعلوم‌اند و برخی حدسهای قریب به یقین دربارهٔ مقادیر آنها ضروری است. نرخ کـلی فـراوانی دوقـلوها در پـروندهٔ اصلی بیمه‌شدگان خدمات درمانی احتمالاً حدود ۱٪ و برابر با نرخ رایـج در کـل جامعـه اسـت. پس انتظار داریم که الگوریتمهای مورداستفاده در ساخت مجموعه جفتها، تولیـد بسـیار بیشتر دوقلوها را نتیجه دهد ــ فرض ایجاد ۱۰ برابر بیشتر را ــ که نرخ ۱۰٪ خواهد بود مـی‌پذیریم. باز هم فرض می‌کنیم که فراوانی نسبی دوقلوها در اولین چارک تفاوت دنباله‌ای چهار برابر فراوانی نسبی دوقلوها در دو چارک میانی و هشت برابر فراوانی نسبی دوقلوها در چارک چهارم است.

بـا ایـن فرضـها و بـا اسـتفاده از جفتهای سفیدپوست مرد ـ مرد به عنوان مثال می‌توانیم فراوانی نسبی دوقلوها را در چارک چهارم از رابطهٔ زیر تعیین کنیم:

$$\frac{\sum_{h=1}^{3} N_h p_{hx}}{\sum_{h=1}^{2} N_h} = ۰/۱۰$$

که در آن، N_h برابر با کل تعداد جفتها در طبقهٔ h است.

با قرار دادن $p_{1x} = 8p_{3x}$ و $p_{2x} = 2p_{3x}$ خواهیم داشت:

$$\frac{8 \times p_{3x} \times ۳۹۸۷۲ + 2 \times p_{3x} \times ۷۹۷۲۷ + p_{3x} \times ۳۹۸۷۲}{۳۹۸۷۲ + ۷۹۷۲۷ + ۳۹۸۷۲} = ۰/۱۰$$

یا

$$p_{3x} = ۰/۰۳۰۸$$

پس نتیجه می‌گیریم که

$$p_{2x} = ۰/۰۶۱۵ \quad و \quad p_{1x} = ۰/۲۴۶۱$$

حالا می‌توانیم σ_{hx} را به دست آوریم

$$\sigma_{3x} = \left[(۰/۰۳۰۸)(۱-۰/۰۳۰۸)\right]^{۰/۵} = ۰/۱۷۲۸$$
$$\sigma_{2x} = \left[(۰/۰۶۱۵)(۱-۰/۰۶۱۵)\right]^{۰/۵} = ۰/۲۴۰۲$$
$$\sigma_{1x} = \left[(۰/۲۴۶۱)(۱-۰/۲۴۶۱)\right]^{۰/۵} = ۰/۴۳۰۷$$

انتساب بهینه بر مبنای رابطهٔ (۱۷.٦) به صورت زیر است:

$$n = \frac{۱۰۰۰}{۶} \approx ۱۶۷ \qquad \left[\frac{\text{کل نمونهٔ ۱۰۰۰ تایی}}{\text{۶ طبقهٔ نژاد ـ جنس}}\right]$$

$$n_1 = \frac{۳۹۸۷۲ \times ۰/۴۳۰۷}{۳۹۸۷۲ \times ۰/۴۳۰۷ + ۷۹۷۲۷ \times ۰/۲۴۰۲ + ۳۹۸۷۲ \times ۰/۱۷۲۸} ۱۶۷ \approx ۶۶$$

$$n_2 = \frac{۷۹۷۲۷ \times ۰/۲۴۰۲}{۳۹۸۷۲ \times ۰/۴۳۰۷ + ۷۹۷۲۷ \times ۰/۲۴۰۲ + ۳۹۸۷۲ \times ۰/۱۷۲۸} ۱۶۷ \approx ۷۴$$

$$n_3 = \frac{۳۹۸۷۲ \times ۰/۱۷۲۸}{۳۹۸۷۲ \times ۰/۴۳۰۷ + ۷۹۷۲۷ \times ۰/۲۴۰۲ + ۳۹۸۷۲ \times ۰/۱۷۲۸} ۱۶۷ \approx ۲۷$$

در حـالی کـه روش بـالا بـرای تعیین انتساب در ۱۸ طبقهٔ ارائه شده در جدول ٦.٦ مورد استفاده قرار گرفـته بـود، نمونـهٔ واقعـی کـه بـه این ترتیب و به شرح جدول ۷.٦ به دست آمد به علت پیشامدهای غیـرقابل پیش‌بینی (مانند ناتوانی در مکان‌یابی جفتهای نمونه‌گیری شده، ملاحظات مالی و غیره) اندکی با انتساب مطلوب تفاوت داشت.

جدول ۷.٦ مجموع زوجهای موجود و مجموع زوجهای نمونه‌گیری شده

نژاد	چارک تفاوت دنباله‌ای	جنس: مرد- مرد جامعه	مرد- مرد نمونه	مرد- زن جامعه	مرد- زن نمونه	زن- زن جامعه	زن- زن نمونه
سفیدپوست	۱	۳۹۸۷۲	۵۲	۱۱۲٦۳	۵۹	۱۰۰۲٤	٦۰
	۲-۳	۷۹۷۲۷	۵۹	۲۲۵۰۱	٦٦	۲۰۰۳۱	٦۷
	٤	۳۹۸۷۲	۲۱	۱۱۲٦۳	۲٤	۱۰۰۲٤	۲٤
سیاه‌پوست	۱	۲۵٤٦	٤۹	۹۹	۵۳	۱۵۳	۵٦
	۲-۳	۵۰۷٦	۵۵	۲۹۹	۵۹	۳۰۰	٦۳
	٤	۲۵٤٦	۲۰	۹۹	۲۱	۱۵۳	۲۳

برآورد کردن فراوانی نسبی دوقلوها و خطای معیار آن برای کل جامعهٔ دوقلوها و برای هر چارک تفاوت دنباله‌ای با استفاده از رابطه‌های (۸.۵) میسر می‌شود که در تابلوی ۲.۵ نشان داده شده‌اند. ذیلاً نشان می‌دهیم که چگونه می‌توان این برآوردها را با استفاده از نرم‌افزار SUDAAN به دست آورد.
داده‌های موردنیاز برای پردازش SUDAAN از سوابق هر یک از ۸۳۱ جفت نمونه به شرح زیر است:

id شمارهٔ یکتایی که به هر جفت داده شده است.

Stratum یکی از ۱۸ چارک تفاوت دنباله‌ای جنس ـ سن فهرست شدهٔ بالا.

npop کل تعداد جفتها در طبقه.

Sampwt کل تعداد جفتها در طبقه تقسیم بر کل تعداد نمونه‌گیری شده در طبقه.

twin متغیر دوحالتی: در صورت دوقلو بودن «۱» و در صورت دوقلو نبودن «۰».

quart 1 چارک بر مبنای تفاوت شمارهٔ تأمین اجتماعی: «۱» اگر در پایینترین چارک است، «۲» اگر در دو چارک میانی است، «۳» اگر در بالاترین چارک است.

داده‌های شامل متغیرهای فهرست شدهٔ بالا در پروندهٔ SAS تحت عنوان JACKTWIN قرار دارند، و برای به دست آوردن برآورد فراوانی نسبی آنها در کل جامعه و در هر یک از گروههای سه‌گانهٔ چارکی بر مبنای تفاوت دنباله‌ای شمارهٔ تأمین اجتماعی از پروندهٔ فرمان SUDAAN به شرح زیر استفاده می‌شود.

```
1.PROC DESCRIPT DATA = JACKTWIN FILETYPE = SAS MEANS TOTALS
     DESIGN = STRWOR;
2. NEST STRATUM;
3. WEIGHT SAMPWT;
4. TOTCNT NPOP;
5. SETENV DECWIDTH = 3;
6. SUBGROUP QUART 1;
7. LEVELS 3;
8. TABLES QUART 1;
9. VARTWIN;
```

پروندهٔ فرمان بالا دارای همان صورت کلی مورد استفاده در مثال تشریحی فصل قبل است و خواننده برای توضیحات مربوط به فرمانهای فردی به آن مثال ارجاع داده می‌شود. خروجی حاصل به صورت زیر خواهد بود.

Variable		Quart 1 Total	1	2	3
Twin	Sample size	831.000	329.000	369.000	133.000
	Weighted size	256,998.000	63,937.000	127,874.000	65,187.000
	Total	26,055.397	19,183.803	6,737.907	133.687
	SE Total	3,791.044	2,661.629	2,696.605	126.744
	Mean	0.101	0.300	0.053	0.002
	SE mean	0.015	0.042	0.021	0.002

یافته‌های عمده، معرف این هستند که برآوردی معادل ۱۰٪ جفتهایی که به شرح بالا ایجاد شده‌اند دوقلو بوده‌اند (در مقابل تقریباً ۱٪ از تمام افراد در کل جامعه). همچنین اگر غربالگری به پایین‌ترین چارک تفاوت دنباله‌ای شمارهٔ تأمین اجتماعی محدود شود، فراوانی نسبی دوقلوها در آن چارک ۳۰٪ خواهد بود (یعنی ۳۰ برابر فراوانی نسبی در میان افراد گزینش نشده). به این ترتیب، آمارگیری مقدماتی نشان می‌دهد که این روش می‌تواند برای شناسایی دوقلوها به صورتی نسبتاً کارآمد مورد استفاده قرار گیرد.

تحلیلی معادل را می‌توان با استفاده از پروندهٔ فرمان زیر و پروندهٔ داده‌های *jacktwin 2.dta* با به کارگیری STATA اجرا کرد.

```
. use "a:\jacktwn 2.dta", clear
. svyset  strata stratum
. svyset pweight sampwt
. svyset fpc npop
. svytotal twin
. svytotal twin, by (quart 1)
```

خروجی STATA به شرح زیر حاصل فرمانهای بالاست:

Survey total estimation

pweight:	sampwt	Number of obs =	831
Strata:	stratum	Number of strata =	18
PSU:	<observations>	Number of PSUs =	831
FPC:	npop	Population size =	256998

Total	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
twin	26055.4	3791.044	18614.01	33496.78	1.988843

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

. svytotal twin, by (quart1)

Survey total estimation

pweight:	sampwt	Number of obs =	831
Strata:	stratum	Number of strata =	18
PSU:	<observations>	Number of PSUs =	831
FPC:	npop	Population size =	256998

Total Subpop	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
twin					
quart1 = = 1	19183.8	2661.629	13959.33	24408.28	1.293026
quart1 = = 2	6737.907	2696.605	1444.778	12031.04	3.590895
quart1 = = 3	133.687	126.7443	-115.0976	382.4715	.3895368

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

□

٦.٦ طبقه‌بندی پس از نمونه‌گیری

طرح نمونه‌ایی که برنامهٔ نمونه‌گیری آن، نمونه‌گیری تصادفی ساده ولی شیوهٔ برآورد کردن آن شیوهٔ مناسب برای نمونه‌گیری تصادفی طبقه‌بندی شده است می‌تواند گاهی اوقات برآوردهایی تولید کند که خطاهای معیار آن چندان بیشتر از خطاهای معیار حاصل از نمونه‌گیری تصادفی طبقه‌بندی شده نباشند. مزیت این طرح آن است که ناممکنی یا دردسر گروه‌بندی عناصر در طبقات را پیشاپیش نمونه‌گیری از میان برمی‌دارد. این نوع طرح را علاوه بر سایر محققان، هنسن و دیگران [۱۰] و کوکران [۹] مورد بررسی قرار داده‌اند. این طرح به عنوان *طبقه‌بندی پس از نمونه‌گیری* یا *پس طبقه‌بندی* نامیده می‌شود.

ممکـن اسـت، برای مثال، برآورد کردن نسبت تولد نوزادان نارس در یک بیمارستان مشخص طی سال گذشـته موردنظـر باشـد. از تجـربه‌هـای قبلی معلوم است که نرخ تولدهای نارس در میان سیاهپوستان بیشـتر از نـرخ متـناظر برای سفیدپوستان است. ولی طبقه‌بندی کل مجموعهٔ سوابق بیمارستان برحسب گروههای نژادی عملی نیست، زیرا گروه نژادی در سوابق ثبت شده است و برای انجام طبقه‌بندی پیش از نمونـه‌گیری باید تمام سوابق بررسی شوند. ولی اگر کل تعداد سیاهپوستان و کل تعداد سفیدپوستانی کـه طـی سـال گذشـته برای زایمان وارد بیمارستان شده‌اند معلوم باشد (که ممکن است برای مدیریت بیمارسـتان به خوبی معلوم باشد) یک نمونهٔ تصادفی ساده می‌تواند پس از نمونه‌گیری برای بهبود دقت برآورد، طبقه‌بندی شود.

$\bar{x}_{pstr}$ و $Var(\bar{x}_{pstr})$ را بـه تـرتیب معـرف *میـانگین نمونـهٔ پـس طبقه‌بندی شده* و *واریانس توزیع نمونه‌گیری آن* بگیرید. پس

$$\bar{x}_{pstr} = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\bar{x}_h$$

$$Var(\bar{x}_{pstr}) = \left(\frac{N-n}{nN}\right)\sum_{h=1}^{L}\frac{N_h}{N}S_{hx}^{2} + \left(\frac{1}{n^{2}}\right)\sum_{h=1}^{L}S_{hx}^{2}\left(\frac{N-n_h}{N}\right) \qquad (۲۰.٦)$$

که در آن

$$S_{hx}^{2} = \frac{\sum_{i=1}^{N_h}\left(X_{h,i}-\bar{X}_h\right)^{2}}{\left(N_h-1\right)}$$

اولیـن عبارت در رابطهٔ (۲۰.٦) تقریباً، واریانس برآورد میانگین تحت نمونه‌گیری طبقه‌بندی شده با انتسـاب متناسـب اسـت.عبارت دوم واریانس را افزایش می‌دهد و این واقعیت را منعکس می‌کنـد که n_h ها در نمونهٔ به دست آمده متغیرهایی تصادفی هستند. هرگاه اندازهٔ نمونه n بزرگ باشد عبارت دوم عموماً کوچک خواهد بود.

با این که S_{hx}^{2} معلوم نیست می‌توان آن را به وسیلهٔ s_{hx}^{2} [رابطهٔ (۹.۵)] برآورد کرد و *برآورد نمونه‌ای* $Var(\bar{x}_{pstr})$ از فرمول زیر به دست می‌آید

$$\hat{Var}(\bar{x}_{pstr}) = \left(\frac{N-n}{nN}\right)\sum_{h=1}^{L}\frac{N_h}{N}s_{hx}^{2} + \left(\frac{1}{n^{2}}\right)\sum_{h=1}^{L}s_{hx}^{2}\left(\frac{N-n_h}{N}\right) \qquad (۲۱.٦)$$

عـبارتهایی شـبیه رابطـهٔ (۲۰.٦) و (۲۱.٦) را مـی‌تـوان برای واریانس برآورد مجموعها و نسبتهای پس طبقه‌بندی شده و نیز برآورد واریانسها از روی اطلاعات نمونه نیز به دست آورد.

حـالا نگـاهی بیـندازیم بـه مـثالی که چگونه پس طبقه‌بندی می‌تواند در کاهش خطای نمونه‌گیری سودمند باشد.

مـثال تشـریحی: یک دامپزشک مایل است هزینهٔ سالانهٔ دامپزشکی مشتریان دائمی خود را (که یا سگ دارنـد یا گربه) بررسی کند. از روی یک سیستم بایگانی جداگانه می‌داند که در حرفهٔ خود ۸۵۰ سگ و ۴۵۰ گـربه را بـه طـور مـنظم مـی‌بیـند (این ارقام مربوط به تعداد حیوانات است نه تعداد ویزیتها). او مـی‌دانـد کـه اطلاعـات مربوط به نوع حیوان (یعنی سگ یا گربه) در سوابق درمانی موجودند ولـــــی جـور کردن سوابق در طبقه‌هایی که برحسب نوع حیوان تعریف شده باشند وقت بسیار زیادی می‌گیرد. بـنابراین تصـمیم مـی‌گیـرد یـک نمونهٔ تصادفی ساده انتخاب کند و سپس پس طبقه‌بندی اجرا کند. او فرایـند پـس طـبقه‌بـندی را ضروری تلقی می‌کند زیرا می‌داند که به طور متوسط سالم نگهداشتن سگ بیشـتر از سـالم نگهداشـتن گربه هزینه دربردارد. او ۵۰ سابقه را برای نمونه انتخاب می‌کند و کل مبلغ پـول هزینه شده توسط صاحبان حیواناتی را که در طول دو سال گذشته دیده است (شامل هزینهٔ دارو) ثبت می‌کند. نتیجهٔ نمونه‌گیری در جدول ۸.۶ ارائه شده است.

حـالا فـرض کـنید که این نمونهٔ متشکل از ۵۰ حیوان قرار است برای برآورد متوسط هزینهٔ سالانهٔ نگهداری سگ یا گربه مورد استفاده قرار گیرد. در این صورت محاسبات زیر را خواهیم داشت (رجوع کنید به تابلوهای ۲.۲ و ۱.۳):

$$\bar{x} = \frac{۴۵/۱۴ + ۵۰/۱۳ + \cdots + ۳۹/۲۶}{۵۰} = ۳۹/۷۳$$

$$s_x^۲ = \frac{(۴۵/۱۴ - ۳۹/۷۳)^۲ + \cdots + (۳۹/۲۶ - ۳۹/۷۳)^۲}{(۵۰ - ۱)} = ۲۵۶/۶۸$$

$$\hat{SE}(\bar{x}) = \left[\left(\frac{۱۳۰۰ - ۵۰}{۱۳۰۰}\right)\left(\frac{۲۵۶/۶۸}{۵۰}\right)\right]^{۱/۲} = \sqrt{۴/۹۳۶} = ۲/۲۲۲$$

از این رو برآورد میانگین جامعه، $\bar{X}$، با بازهٔ اطمینان ۹۵ درصدی از فرمول زیر به دست می‌آید

$$\bar{x} - ۱/۹۶ \times \hat{SE}(\bar{x}) \le \bar{X} \le \bar{x} + ۱/۹۶ \times \hat{SE}(\bar{x})$$

$$۳۹/۷۳ - ۱/۹۶ \times ۲/۲۲۲ \le \bar{X} \le ۳۹/۷۳ + ۱/۹۶ \times ۲/۲۲۲$$

$$۳۵/۳۸ \le \bar{X} \le ۴۴/۰۸$$

اینک از مجموعهای معلوم طبقات در یک فرایند پس طبقه‌بندی استفاده می‌کنیم تا برآورد دقیقتری از $\bar{X}$ به دست آوریم. دامپـــــزشک می‌داند که تعـداد سگها در بایگانی او $N_۱ = ۸۵۰$ و کل تعـــداد گربه‌ها $N_۲ = ۴۵۰$ است. طبقـه‌بندی ۵۰ حیوان نمونه ارائه شده در جدول ۸.۶ برحسب ســگ و گربـــه، نمـونه‌ای متشکل از $n_۱ = ۳۲$ سگ و $n_۲ = ۱۸$ گربه را نتیجه می‌دهد.

جدول ۸.٦ داده‌های نمونه‌ای مربوط به آمارگیری دامپزشک

حیوان نمونه	نوع حیوان	تعداد ویزیت	کل هزینه‌ها(دلار)	حیوان نمونه	نوع حیوان	تعداد ویزیت	کل هزینه‌ها(دلار)
۱	سگ	٤	٤٥/١٤	٢٦	سگ	٤	٤٨/٣٠
٢	سگ	٥	٥٠/١٣	٢٧	سگ	٥	٥٤/٦٤
٣	گربه	٢	٢٧/١٥	٢٨	گربه	٣	٢١/٤٥
٤	سگ	٣	٤٥/٨٠	٢٩	گربه	٣	١٠/٧١
٥	گربه	١	٢٣/٣٩	٣٠	سگ	٤	٦٠/٥٧
٦	گربه	٢	٨/٢٤	٣١	سگ	٦	٥٣/٣٧
٧	سگ	٦	٦١/٢٢	٣٢	سگ	٥	٤٠/٥٢
٨	گربه	٢	٢٩/٩٠	٣٣	سگ	٤	٥٠/٢٦
٩	سگ	٥	٥٦/٥٧	٣٤	گربه	٢	١٥/٢٣
١٠	سگ	٤	٤٢/٣٩	٣٥	سگ	٤	٤٢/٠٢
١١	گربه	٢	٢٧/٢٤	٣٦	سگ	٥	٣٢/٧٨
١٢	گربه	٣	٢٢/١٧	٣٧	گربه	٢	٣٠/٢١
١٣	سگ	٦	٣٩/٦٧	٣٨	گربه	١	٢٧/٥٤
١٤	سگ	٤	٤٠/٥٢	٣٩	سگ	٦	٥٢/٠٣
١٥	سگ	٤	٣٩/٤٨	٤٠	سگ	٥	٥٤/٤٧
١٦	گربه	١	٧/١٤	٤١	سگ	٥	٤٦/٨٨
١٧	سگ	٤	٦١/٨٢	٤٢	گربه	٢	٢٣/٧٧
١٨	گربه	٢	٣٩/٨٨	٤٣	سگ	٣	٥٢/٤٨
١٩	گربه	٢	١٦/٨٩	٤٤	سگ	٢	٦٠/٤٩
٢٠	سگ	٣	٥٥/٣١	٤٥	سگ	٢	٥٣/٧٠
٢١	سگ	٢	٦٣/١٩	٤٦	سگ	٢	٤٦/٣٩
٢٢	سگ	٢	٤٥/١١	٤٧	سگ	٢	٥٣/٢٤
٢٣	سگ	٣	٦٦/٢٠	٤٨	گربه	١	١٤/١٨
٢٤	گربه	٣	١٧/١٦	٤٩	سگ	٣	٤١/٥٢
٢٥	گربه	٣	٢٨/٥٥	٥٠	سگ	٢	٣٩/٢٦

به این ترتیب (رجوع کنید به تابلوی ۲.۲) خواهیم داشت:

$$\bar{x}_{۱} = \frac{۴۵/۱۴ + ۵۰/۱۳ + \ldots + ۳۹/۲۶}{۳۲} = ۴۹/۸۶$$
$$\bar{x}_{۲} = \frac{۲۷/۱۵ + ۲۳/۳۹ + \ldots + ۱۴/۱۸}{۱۸} = ۲۱/۷۱$$

$$s^{۲}_{۱x} = \frac{(۴۵/۴۴ - ۴۹/۸۶)^{۲} + \ldots + (۳۹/۲۶ - ۴۹/۸۶)^{۲}}{۳۱} = ۷۰/۲۲$$
$$s^{۲}_{۲x} = \frac{(۲۷/۱۵ - ۲۱/۷۱)^{۲} + \ldots + (۱۴/۱۸ - ۲۱/۷۱)^{۲}}{۱۷} = ۷۵/۰۰$$

$$\bar{x}_{pstr} = \left(\frac{۸۵۰}{۱۳۰۰}\right) \times ۴۹/۸۶ + \left(\frac{۴۵۰}{۱۳۰۰}\right) \times ۲۱/۷۱ = ۴۰/۱۲$$

$$\hat{Var}(\bar{x}_{pstr}) = \left(\frac{۱۳۰۰ - ۵۰}{۵۰ \times ۱۳۰۰}\right)\left[\frac{۸۵۰}{۱۳۰۰} ۷۰/۲۲ \frac{۴۵۰}{۱۳۰۰} ۷۵/۰۰\right]$$
$$+ \left(\frac{۱}{۵۰^{۲}}\right)\left[\left(\frac{۱۳۰۰ - ۸۵۰}{۱۳۰۰}\right) \times ۷۰/۲۲ + \left(\frac{۱۳۰۰ - ۴۵۰}{۱۳۰۰}\right) \times ۷۵/۰۰\right] = ۱/۴۳۹$$

$$\hat{SE}(\bar{x}_{pstr}) = \sqrt{۱/۴۳۹} = ۱/۲۰$$

از این رو برآورد میانگین جامعه، $\bar{X}$، با بازهٔ اطمینان ۹۵ درصدی با استفاده از پس طبقه‌بندی به صورت زیر محاسبه می‌شود

$$\bar{x}_{pstr} - ۱/۹۶ \times \hat{SE}(\bar{x}_{pstr}) \le \bar{X} \le \bar{x}_{pstr} + ۱/۹۶ \times \hat{SE}(\bar{x}_{pstr})$$
$$۴۰/۱۲ - ۱/۹۶ \times ۱/۲۰ \le \bar{X} \le ۴۰/۱۲ + ۱/۹۶ \times ۱/۲۰$$
$$۳۷/۷۷ \le \bar{X} \le ۴۲/۴۷$$

می‌بینیم که دامپزشک با پس طبقه‌بندی نمونهٔ تصادفی که در ابتدا انتخاب کرده بود بازهٔ اطمینانی به دست آورد که از بازهٔ اطمینان محاسبه شده برای نمونهٔ تصادفی سادهٔ اولیه کوتاهتر بود.

برآوردهای پس طبقه‌بندی و خطاهای معیار آنها را می‌توان با استفاده از SUDAAN به دست آورد. برای مثال تشریحی که در بالا بحث شد می‌توان از مجموعهٔ فرمانهای زیر برای به دست آوردن برآورد پس طبقه‌بندی میانگین هزینه‌های مربوط به سگها و گربه‌ها استفاده کرد.

```
PROC DESCRIPT DATA=DOGSCATS FILETYPE=SAS DESIGN=WOR;
NEST_ONE_;
TOTCNT N;
WEIGHT WEIGHT;
SUBGROUP TYPE;
LEVELS 2;
VAR TOTEXP;
POSTVAR TYPE;
POSTWGT 850 450;
```

پـروندهٔ داده‌هـای *dogscats.ssd* شـــامل ۵۰ حیوان نمونه که برحسب نوع حیـــوان تفکیک شده است (سگ = «۱» و گربه = «۲») در پایین نشان داده شده است:

id	type	totexp	weight	N	id	type	totexp	weight	N
35	1	42.02	26	1300	32	1	40.52	26	1300
33	1	50.26	26	1300	4	1	45.80	26	1300
50	1	39.26	26	1300	39	1	52.03	26	1300
31	1	53.37	26	1300	9	1	56.57	26	1300
10	1	42.39	26	1300	7	1	61.22	26	1300
26	1	48.30	26	1300	17	1	61.82	26	1300
20	1	55.31	26	1300	47	1	53.24	26	1300
27	1	54.64	26	1300	16	2	7.14	26	1300
41	1	46.88	26	1300	29	2	10.71	26	1300
46	1	46.39	26	1300	37	2	30.21	26	1300
15	1	39.48	26	1300	6	2	8.24	26	1300
36	1	32.78	26	1300	34	2	15.23	26	1300
22	1	45.11	26	1300	28	2	21.45	26	1300
21	1	63.19	26	1300	48	2	14.18	26	1300
43	1	52.48	26	1300	38	2	27.54	26	1300
23	1	66.20	26	1300	3	2	27.15	26	1300
45	1	53.70	26	1300	19	2	16.89	26	1300
1	1	45.14	26	1300	18	2	39.88	26	1300
13	1	39.67	26	1300	42	2	23.77	26	1300
2	1	50.13	26	1300	24	2	17.16	26	1300
40	1	54.47	26	1300	12	2	22.17	26	1300
14	1	40.52	26	1300	25	2	28.55	26	1300
30	1	60.57	26	1300	11	2	27.24	26	1300
44	1	60.49	26	1300	5	2	23.39	26	1300
49	1	41.52	26	1300	8	2	29.90	26	1300

هر گزارش شامل پنج متغیر زیر است:

id : شمارهٔشناسایی هر حیوان نمونه.

type : سگ = ۱ ؛ گربه = ۲.

totexp : کل هزینه‌های درمانی دامپزشکی که برای حیوان تقبل شده است.

weight : $\frac{N}{n} = \frac{۱۳۰۰}{۵۰} = ۲۶$

N : کل تعداد حیوانات در جامعه.

توجه کنید که متغیرهای *weight* و *totcnt* که در پردازش مورد استفاده قرار گرفته‌اند برای همهٔ گزارشها یکسان‌اند و برای برآورد کردن از یک طرح نمونه‌گیری تصادفی ساده مناسب‌اند. فرمان *postvar* معرف متغیر رسته‌ای است که به عنوان پایهٔ پس طبقه‌بندی (که در این مورد نوع حیوان است) به کار می‌رود و فرمان *postwgt* معرف مجموعهای جامعه‌ای برای هر رسته است. خروجی زیر با استفاده از این فرمان تولید خواهد شد.

```
1       PROC DESCRIPT DATA = DOGSCATS FILETYPE= SAS DESIGN = WORK;
2       NEST _ONE_;
3       TOTCNT N;
4       WEIGHT WEIGHT;
5       SUBGROUP TYPE;
6       LEVEL 2;
7       VAR TOTEXP;
8       POSTVAR TYPE;
9       P OSTWGT 850 450;

Number of observations read     :     50      Weight count :  1300
Number of observations skipped  :      0
(WEIGHT variable nonpositive)
Denominator degrees of freedom  :     49

                        Research Triangle Institute          Page   : 1
                                                             Table  : 1

Post-stratified estimates
by: Variable, TYPE.
```

Variable		TYPE total	1
TOTEXP	Sample Size	50	32
	Weighted Size	1300.00	850.00
	Total	52149.67	42379.67
	Mean	40.12	49.86
	SE Mean	1.16	1.44

Post-stratified estimates
by: Variable, TYPE.

Variable		TYPE 2
TOTEXP	Sample Size	18
	Weighted Size	450.00
	Total	9770.00
	Mean	21.71
	SE Mean	1.97

باید توجه داشت که برآورد خطای معیار میانگین پس طبقه‌بندی شده، $\bar{x}_{pstr}$،که به وسیلهٔ SUDAAN تولید می‌شود با میانگین به دست آمده از رابطهٔ (۲۱.٦) در این متن (۱/۱٦ محاسبه شده توسط SUDAAN در برابر ۱/۲۰ محاسبه شده از روی فرمول متن) اندکی تفاوت دارد. علت، آن است که فرمول تقریب مورد استفادهٔ SUDAAN اندکی متفاوت است.

در زمان نگارش این متن، برآوردهای پس طبقه‌بندی را نمی‌شد مستقیماً از فرمانهای آمارگیری STATA به دست آورد.

متوجه می‌شویم که پس طبقه‌بندی تنها هنگامی (با توجه به خطاهای معیار کمتر) مقرون به صرفه است که طبقات تعیین شده از لحاظ متغیر موردنظر همگن باشند. به عبارت دیگر، پس طبقه‌بندی هنگامی خوب عمل می‌کند که طبقه‌بندی خوب عمل کند. علاوه بر این، روش مزبور را نمی‌توان اجرا کرد مگر این که مجموعهای طبقات معلوم باشند. ولی هنگامی که نمونه از جوامع انسانی گرفته می‌شود غالباً امکان استفاده از سرشماریهای موجود یا سایر داده‌های جمعیتی وجود دارد تا حدسهایی به قدر کافی مناسب از مجموعهای طبقه‌ای برای استفاده از روش سودمند زده شوند.

۷.٦ نمونهٔ موردنیاز چقدر باید بزرگ باشد؟

فرض کنید می‌خواهیم تعداد موردنیاز عناصر را طوری تعیین کنیم که $(1-\alpha)\times 100$ درصد مطمئن شویم که با استفاده از نمونه‌گیری تصادفی طبقه‌بندی شده، برآوردی از میانگین، $\bar{x}_{str}$، به دست می‌آوریم که با میانگین واقعی $\bar{X}$ بیشتر از $\varepsilon \times 100$ درصد تفاوت ندارد. این فرمولبندی همان فرمولبندیی است که قبلاً برای نمونه‌گیری تصادفی ساده و نمونه‌گیری سیستماتیک مورد بحث قرار گرفت. فرمول (معتبر برای N_h با بزرگی معقول) مربوط به n مورد نیاز به صورت زیر است:

$$n \approx \frac{\left(\frac{z^2_{1-(\alpha/2)}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2\sigma_{hx}^2}{\pi_h \bar{X}^2}\right)}{\varepsilon^2 + \left(\frac{z^2_{1-(\alpha/2)}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h\sigma_{hx}^2}{\bar{X}^2}\right)} \qquad (۲۲.٦)$$

که در این فرمول

$$\pi_h = \frac{n_h}{n}$$

رابطهٔ (۲۲.٦) برای هر نوع انتسابی معتبر است. برای برآورد مجموع یک جامعه نیز معتبر است. فرمول متناظر اندازهٔ نمونه برای برآورد نسبت جامعه‌ای P_y از نمونه‌گیری تصادفی طبقه‌بندی شده به صورت زیر است:

$$n \approx \frac{\left(\frac{z^{۲}_{۱-(\alpha/۲)}}{N^{۲}}\right)\left(\sum_{h=۱}^{L}\frac{N_h^{۲}P_{hy}(۱-P_{hy})}{\pi_h P_y^{۲}}\right)}{\varepsilon^{۲}+\left(\frac{z^{۲}_{۱-(\alpha/۲)}}{N^{۲}}\right)\left(\sum_{h=۱}^{L}\frac{N_h P_{hy}(۱-P_{hy})}{P_y^{۲}}\right)} \qquad (۲۳.٦)$$

با دیدن رابطهٔ (۲۲.٦) پی می‌بریم که استفاده از آن به آگاهیهایی دربارهٔ پارامترهای توزیع نیاز دارد که به مراتب بیشتر از آن است که احتمال دسترسی به آن وجود داشته باشد یا بتواند با هر میزانی از اطمینان حدس زده شود. به همین دلیل احتمال ندارد که رابطهٔ (۲۲.٦) در کاربست عملی چندان کمکی بکند. ولی اگر فرض بر انتساب متناسب باشد در آن صورت رابطهٔ (۲۲.٦) به صورت زیر تبدیل می‌شود

$$n \approx \frac{Nz^{۲}_{۱-(\alpha/۲)}\left(\sigma^{۲}_{wx}/\bar{X}^{۲}\right)}{N\varepsilon^{۲}+z^{۲}_{۱-(\alpha/۲)}\left(\sigma^{۲}_{wx}/\bar{X}^{۲}\right)} \qquad (۲٤.٦)$$

توجه کنید که رابطهٔ (۲٤.٦) شبیه عبارت (۱۵.۳) برای اندازهٔ نمونهٔ مورد نیاز در برآورد کردن میانگین نمونه تحت نمونه‌گیری تصادفی ساده است (تابلوی ٤.۳). تنها تفاوت میان دو عبارت آن است که $V_x^{۲}$ (که برابر است با $\frac{\sigma_x^{۲}}{\bar{X}^{۲}}$) در رابطهٔ (۱۵.۳) با $\frac{\sigma_{wx}^{۲}}{\bar{X}^{۲}}$ در رابطهٔ (۲٤.٦) جایگزین شده است. اگر ایده‌ای از ترتیب بزرگی $V_x^{۲}$ داشته باشیم، که واریانس نسبی توزیع متغیر x در جامعه است، و اگر علاوه بر آن ایده‌ای هم از نسبت $\gamma = \frac{\sigma_{bx}^{۲}}{\sigma_{wx}^{۲}}$ داشته باشیم، آنگاه چون $\sigma_x^{۲} = \sigma_{bx}^{۲} + \sigma_{wx}^{۲}$، رابطهٔ (۲٤.٦) به صورت زیر درمی‌آید

$$n \approx \frac{z^{۲}_{۱-(\alpha/۲)} \times \frac{N}{۱+\gamma} \times V_x^{۲}}{N\varepsilon^{۲}+z^{۲}_{۱-(\alpha/۲)} \times \frac{V_x^{۲}}{۱+\gamma}} \qquad (۲۵.٦)$$

نگاهی به یک مثال بیندازیم تا ببینیم رابطهٔ (۲۵.٦) در عمل چگونه به کار می‌رود.

مثال تشریحی: فرض کنید در نظر داریم برای برآورد کردن متوسط تعداد موارد بستری شدن در بیمارستان به ازای هر فرد، از اعضای یک سازمان حفظ بهداشت (HMO) نمونه بگیریم. نمونه از فهرستهای عضویت انتخاب خواهد شد که براساس سن گروه‌بندی شده‌اند (کمتر از ٤۵ سال، ٤۵ تا ٦٤ سال، ٦۵ سال به بالا). فرض می‌کنیم که توزیعهای مربوط به بستری شدن در بیمارستان از داده‌های ملی قابل حصول بوده (مانند آمارگیری مصاحبه‌ای سازمان بهداشت ملی) و به صورت جدول ۹.٦ باشد.

بـه عـلاوه فـرض کـنید که تعداد اعضای سازمان حفظ بهداشت (**HMO**) در هر یک از گروههای سنی به شرح زیر است:

گروه سنی ۱ : $N_۱ = ٦۰۰$

گروه سنی ۲ : $N_۲ = ۵۰۰$

گروه سنی ۳ : $N_۳ = ٤۰۰$

جدول ۹.٦ توزیع موارد بستری شدن در بیمارستان به ازای هر شخص در سال

	گروه سنی	متوسط تعداد موارد بستری شدن در بیمارستان	واریانس توزیع موارد بستری شدن در بیمارستان
۱	کمتر از ٤۵ سال	۰/۱٦٤	۰/۲٤۵
۲	٤۵ تا ٦٤ سال	۰/۱٦٦	۰/۲۹٦
۳	٦۵ سال به بالا	۰/۲۳٦	۰/٤۳٦

اگـر فـرض کـنیم کـه نـتایج به دست آمده در سطح ملی، احتمال دارد که در مورد اعضای سازمان حفاظت بهداشت نیز درست باشد، آنگاه میانگین پیش‌بینی شدهٔ تعداد موارد بستری شدن در بیمارستان به ازای هر نفر از معادلهٔ (۱.٦) چنین خواهد بود

$$\overline{X} = \frac{۶۰۰\times۰/۱۶۴ + ۵۰۰\times۰/۱۶۶ + ۴۰۰\times۰/۲۳۶}{۱۵۰۰} = ۰/۱۸۴$$

مؤلفهٔ واریانس پیش‌بینی شده، $\sigma^۲_{bx}$، از معادلهٔ (۱۵.٦) عبارت است از

$$\sigma^۲_{bx} = \frac{۶۰۰\times(۰/۱۶۴-۰/۱۸۴)^۲ + ۵۰۰\times(۰/۱۶۶-۰/۱۸۴)^۲ + ۴۰۰\times(۰/۲۳۶-۰/۱۸۴)^۲}{۱۵۰۰} = ۰/۰۰۰۹۸۹۱$$

مؤلفهٔ واریانس پیش‌بینی شده، $\sigma^۲_{wx}$، از معادلهٔ (۱۳.٦) عبارت است از

$$\sigma^۲_{wx} = \frac{۶۰۰\times ۰/۲۴۵ + ۵۰۰\times ۰/۲۹۶ + ۴۰۰\times ۰/۴۳۶}{۱۵۰۰} = ۰/۳۱۲۹۳$$

بالاخره مقادیر پیش‌بینی شده برای $\sigma^۲_x$، $V^۲_x$ و γ عبارت‌اند از

$$\sigma^۲_x = ۰/۰۰۰۹۸۹۱ + ۰/۳۱۲۹۳ = ۰/۳۱۳۹۲$$

$$V^۲_x = \frac{۰/۳۱۳۹۲}{(۰/۱۸۴)^۲} = ۹/۲۷$$

$$\gamma = \frac{۰/۰۰۰۹۸۹۱}{۰/۳۱۲۹۳} = ۰/۰۰۳۱۶$$

حـالا بـا اسـتفاده از رابطـهٔ (۲۵.٦) مـی‌توانیـم برآورد تعداد آزمودنیها را که برای اطمینان واقعی از بـرآورد میانگین تعداد موارد بستری شدن در بیمارستان لازم است در داخل محدودهٔ ۲۰ درصد میانگین واقعی تحت نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب باشد، محاسبه کنیم:

$$n \approx \frac{\left[\frac{(۹\times۱۵۰۰)}{(۱+۰/۰۰۳۱۶)}\right]\times ۹/۲۷}{\left[\frac{(۹\times ۹/۲۷)}{(۱+۰/۰۰۳۱۶)}\right]+۱۵۰۰\times(۰/۲۰)^{۲}} = ۸۷۲$$

پس تعداد n_h متنسب به هر طبقه به صورت زیر خواهد بود:

$$n_۱ = ۶۰۰\times\frac{۸۷۲}{۱۵۰۰} = ۳۴۹$$

$$n_۲ = ۵۰۰\times\frac{۸۷۲}{۱۵۰۰} = ۲۹۱$$

$$n_۳ = ۴۰۰\times\frac{۸۷۲}{۱۵۰۰} = ۲۳۲$$

اگـر بخواهیم انتساب بهینه را با فرض هزینهٔ برابر برای طبقه‌ها به کار ببریم، اندازهٔ نمونهٔ مورد نیاز مـی‌تـــواند بـه ایـن تـرتیب برآورد شود که ابتدا n_h بهینه را از روی برآوردهای ملی تعیین و سپس n مورد نیاز را از رابطهٔ (۲۲.٦) محاسبه کنیم. این روش به شرح زیر اجرا می‌شود.

ابتدا π_h بهینه را (که برابر است با $N_h\sigma_{hx}\Big/\sum_{h=۱}^{L} N_h\sigma_{hx}$) براساس داده‌های ملی محاسبه کنید:

$$\pi_۱ = \frac{۶۰۰\sqrt{۰/۲۴۵}}{۶۰۰\sqrt{۰/۲۴۵}+۵۰۰\sqrt{۰/۲۹۶}+۴۰۰\sqrt{۰/۴۳۶}} = ۰/۳۵۶$$

$$\pi_۲ = \frac{۵۰۰\sqrt{۰/۲۹۶}}{۶۰۰\sqrt{۰/۲۴۵}+۵۰۰\sqrt{۰/۲۹۶}+۴۰۰\sqrt{۰/۴۳۶}} = ۰/۳۲۷$$

$$\pi_۳ = \frac{۴۰۰\sqrt{۰/۴۳۶}}{۶۰۰\sqrt{۰/۲۴۵}+۵۰۰\sqrt{۰/۲۹۶}+۴۰۰\sqrt{۰/۴۳۶}} = ۰/۳۱۷$$

سپس n مورد نیاز را براساس رابطهٔ (۲۲.٦) با $\varepsilon = ۰/۲۰$ حساب کنید:

$$n \approx \frac{\left[\frac{۹}{(۱۵۰۰)^{۲}}\right]\left[\frac{(۶۰۰)^{۲}(۰/۲۴۵)}{(۰/۳۵۶)(۰/۱۸۴)^{۲}}+\frac{(۵۰۰)^{۲}(۰/۲۹۶)}{(۰/۳۲۷)(۰/۱۸۴)^{۲}}+\frac{(۴۰۰)^{۲}(۰/۴۳۶)}{(۰/۳۱۷)(۰/۱۸۴)^{۲}}\right]}{(۰/۲)^{۲}+\left[\frac{۹}{(۱۵۰۰)^{۲}}\right]\left[\frac{۶۰۰\times۰/۲۴۵}{(۰/۱۸۴)^{۲}}+\frac{۵۰۰\times۰/۲۹۶}{(۰/۱۸۴)^{۲}}+\frac{۴۰۰\times۰/۴۳۶}{(۰/۱۸۴)^{۲}}\right]} = ۸۶۰$$

سـرانجام، نمونـهٔ ۸٦۰ تایی لازم برای رسیدن به اطمینان واقعی با $\varepsilon = ۰/۲۰$ با ضرب کردن ۸٦۰ در π_h مناسب به طبقه‌ها تخصیص می‌یابد:

$$n_۱ = n \times \pi_۱ = ۸۶۰ \times ۰/۳۵۶ = ۳۰۶$$
$$n_۲ = n \times \pi_۲ = ۸۶۰ \times ۰/۳۲۷ = ۲۸۱$$
$$n_۳ = n \times \pi_۳ = ۸۶۰ \times ۰/۳۱۷ = ۲۷۳$$

توجه کنید که اندازهٔ نمونهٔ مورد نیاز (۸٦۰) تحت انتساب بهینه کوچکتر از اندازهٔ مورد نیاز تحت انتساب متناسب (۸۷۲) است.

□

باید متذکر شد که تحت انتساب متناسب، اندازهٔ نمونهٔ مورد نیاز برای برآورد یک نسبت [رابطهٔ (۲۳.٦)] به صورت زیر تبدیل می‌شود:

$$n \approx \frac{\left(\frac{z^۲_{۱-(\alpha/۲)}}{N}\right)\left(\sum_{h=۱}^{L}\frac{N_h P_{hy}(۱-P_{hy})}{P_y^۲}\right)}{\varepsilon^۲ + \left(\frac{z^۲_{۱-(\alpha/۲)}}{N}\right)\left(\sum_{h=۱}^{L}\frac{N_h P_{hy}(۱-P_{hy})}{P_y^۲}\right)} \qquad (۲٦.٦)$$

ولی چون $P_{hy}(۱-P_{hy}) \le ۰/۲۵$، رابطهٔ (۲٦.٦) به نابرابری زیر تبدیل می‌شود:

$$n \le \frac{۰/۲۵ \times \frac{z^۲_{۱-(\alpha/۲)}}{P_y^۲}}{\varepsilon^۲ + ۰/۲۵ \times \frac{z^۲_{۱-(\alpha/۲)}}{N \times P_y^۲}} \qquad (۲۷.٦)$$

از رابطهٔ (۲۷.٦) می‌توان برآورد محافظه‌کارانه‌ای از اندازهٔ نمونهٔ مورد نیاز بر مبنای اطلاع از تعداد عناصر موجود در جامعه، N، به دست آورد و نسبت P_y دارای صفت کیفی y در جامعه را «حدس زد».

۸.٦ مرزبندی طبقه‌ها و تعداد طبقه‌های مطلوب

در بسیاری از شرایط، مرزهای طبقات برحسب دسترسی به اطلاعات جامعه‌ای مورد نیاز برای انتخاب نمونهٔ طبقه‌بندی شده و انجام برآورد تعیین می‌شوند. برای مثال اگر هیچ اطلاعات جامعه‌ای در مورد سطح کدپستی در دسترس نباشد و هیچ راهی برای ایجاد چارچوب نمونه‌گیری در محدودهٔ کدپستی نباشد، امکان طبقه‌بندی بر مبنای کدهای پستی در داخل ناحیهٔ جغرافیایی هدف تعریف شده نیز وجود نخواهد داشت.

در وضعیتهای دیگری که تصور می‌شود طبقه‌بندی مطلوب است، شاید انتخاب مرزهای طبقه‌ها امکان‌پذیر باشد. در آن صورت، شخص مایل است مرزهایی را انتخاب کند که قابلیت اعتماد

برآوردهای حاصل را بالا ببرد. مثلاً، ممکن است چارچوب نمونه‌گیری، یک پروندهٔ رایانه‌ای شامل همهٔ مراجعات اعضای یک سازمان حفظ بهداشت **HMO** خاص در طول یک سال خاص به متخصصان بهداشتی باشد. ممکن است طبقه‌بندی براساس یک مشخصهٔ خاص بیمار از قبیل سن یا سطح فشار خون در زمان ثبت‌نام برای عضویت در سازمان حفظ بهداشت، مطلوب تلقی شود. مسئله‌ای که باید مورد توجه قرار گیرد انتخاب ویژهٔ مرزهای طبقه است.

یکی از راهبردهای کلی در انتخاب مرزهای طبقه، انتخاب این مرزها به صورتی است که واریانس برآورد حاصل تحت انتساب بهینه حداقل باشد. دالینیوس [۱۲] معادلاتی را برای تعیین این مرزها تهیه کرده است که، به هر حال، استفاده از آن در عمل به علت وابستگیهای بین مؤلفه‌ها مشکل است. یک روش تقریبی توسط دالینیوس و هاجز [۱۳] تهیه شده است که به نظر می‌رسد در عمل خوب کار می‌کند. این روش تقریبی مستلزم (۱) گروهبندی متغیر طبقه‌بندی، x، به تعدادی از رده‌ها، (۲) تعیین توزیع فراوانی $f(x)$ متغیر x برای هر رده، (۳) محاسبهٔ مقادیر انباشتهٔ ریشهٔ دوم $f(x)$؛ و (٤) تعیین Q، خارج قسمت مجموع کل ریشهٔ دوم $f(x)$ روی همهٔ رده‌ها، و تعداد L طبقهٔ مورد استفاده است. نقاط تقسیم نهایی که به این ترتیب به دست می‌آیند عبارت‌اند از Q، $2Q$، ... ، $(L - 1)Q$. این روش (که از آن به عنوان روش *فراوانی ریشه‌ای* نام می‌بریم) در مثال زیر نشان داده شده است.

مثال تشریحی: قرار است سوابق یک گروه پزشکی بزرگ که بیماران بیمهٔ کمکهای درمانی را معالجه می‌کند حسابرسی شوند. هدف از این حسابرسی برآورد کردن میزان پول اضافی است که این گروه پزشکی طی سال ۱۹۹٦ از کمکهای درمانی مطالبه کرده است. این گروه پزشکی در آن سال ۲۳۸۷ بیمار را درمان کرده است و توزیع فراوانی کل مبلغی که برحسب دلار به ازای هر بیمار در طول سال از بیمهٔ کمکهای درمانی مطالبه کرده در جدول ۱۰.٦ نشان داده شده است (داده‌های گروه‌بندی نشده در پروندهٔ STATA، *medaudit.dta* قرار دارند):

ستون اول در این جدول، دلارهای هزینه شده توسط بیمهٔ کمکهای درمانی را که از طرف هر بیمار هزینه شده است نشان می‌دهد که در ۲۲ رسته گروه‌بندی شده‌اند.
ستون دوم توزیع فراوانی را برای هر یک از این رسته‌ها نشان می‌دهد.
ستون سوم توزیع فراوانی انباشتهٔ آن متغیر را نشان می‌دهد.
ستون چهارم ریشهٔ دوم توزیع فراوانی آن متغیر را ارائه می‌دهد.
ستون پنجم توزیع فراوانی انباشتهٔ ریشهٔ دوم آن متغیر را نشان می‌دهد. توجه کنید که مجموع این فراوانی ریشهٔ انباشته ۱٤۳/۰۸۵۲٦۷ است.

جدول ۱۰.٦ توزیع فراوانی کل مبلغ مطالبه شده از بیمهٔ کمکهای درمانی در سال ۱۹۹٦ توسط یک گروه بزرگ پزشکی برای ۲۳۸۷ بیمار درمان شده

Totpaid ($)	Freq	Cumfreq	Rootfreq	Cumrootf	Stratum
0-49	39	39	6.244997998	6.244998	1
50-99	819	858	28.61817604	34.863174	1
100-149	570	1428	23.87467277	58.7378468	2
150-199	350	1778	18.70828693	77.4461337	3
200-299	314	2092	17.72004515	95.1661789	3
300-399	136	2228	11.66190379	106.828083	4
400-499	71	2299	8.426149773	115.254232	4
500-599	37	2336	6.08276253	121.336995	4
600-699	17	2353	4.123105626	125.460101	4
700-799	11	2364	3.31662479	128.776725	4
800-899	10	2374	3.16227766	131.939003	5
900-999	2	2376	1.414213562	133.353217	5
1000-1099	1	2377	1	134.353217	5
1100-1199	1	2378	1	135.353217	5
1200-1299	3	2381	1.732050808	137.085267	5
1300-1399	1	2382	1	138.085267	5
1400-1499	1	2383	1	139.085267	5
1500-1599	0	2383	0	139.085267	5
1600-1699	1	2384	1	140.085267	5
1700-1799	1	2385	1	141.085267	5
1800-1899	1	2386	1	142.085267	5
1900-1999	1	2387	1	143.085267	5

اگر بخواهیم پنج طبقه داشته باشیم، در آن صورت $Q = \frac{۱۴۳/۰۸۵۲۶۷}{۵} = ۲۸/۶۱۷۰۵۳۵$ و مرزهای فوقانی هر طبقه به وسیلهٔ مقادیر فراوانی ریشهٔ انباشته تعیین می شود: ۲۸/٦۱۷۰۵۳۵، ۵۷/۲۳٤۱۱، ۸۵/۸۵۱۱٦، ۱۱٤/٤٦۸۲ و ۱٤۳/۰۸۵۳. این مقادیر (به طوری که در ستون ٦ جدول ۱۰.٦ نشان داده شده است) با مرزهای طبقات در متغیر *Totpaid* کل مبلغ پرداخت شده به ازای هر بیمار توسط بیمهٔ کمکهای درمانی مطابقت دارد.

انتساب بهینه براساس این طبقه‌بندی، انتساب درصدی را نتیجه می‌دهد که در ستون ۵ جدول ۱۱.٦ نشان داده شده است.

حالا این طبقه‌بندی را با دو راهبرد ظاهراً «معقول» زیر برای ایجاد طبقه‌ها مقایسه می‌کنیم:

۱. طبقه‌بندی براساس تقسیم دامنهٔ توزیع به تعداد طبقه‌ها (روش *دامنهٔ برابر*).

۲. طبقه‌بندی براساس تقسیم توزیع درصدی کل به تعداد مطلوب طبقه‌ها (روش *چندکی*).

مقایسهٔ سه روش طبقه‌بندی *فراوانی ریشه‌ای*، *دامنهٔ برابر* و *چندکی* در جدول ۱۲.٦ ارائه شده است.

جدول ۱۱.٦ انتساب بهینه براساس استفاده از روش *فراوانی ریشه‌ای* برای ایجاد طبقه‌ها

طبقه	کل پرداختها برحسب دلار	تعداد در طبقه (N_h)	انحراف معیار، σ_{hx} کل پرداختها	انتساب بهینه به طبقه($\%n$)
۱	۰–۹۹	۸۵۸	۱٤/۸۹۰	۱٤/٥٦
۲	۱۰۰–۱٤۹	٥۷۰	۱۳/۸٤۷	۹/۰۰
۳	۱٥۰–۲۹۹	٦٦٤	٤۰/٦۰٦	۳۰/۷٦
٤	۳۰۰–۷۹۹	۲۷۲	۱۱٦/٦۱۷	۳٦/۱۹
٥	۸۰۰–۱۹۹۹	۲۳	۳٦۰/۹٤۷	۹/٤۷

جدول ۱۲.٦ نتایج حاصل از سه روش برای ایجاد طبقه‌ها همراه با انتساب بهینه از روی داده‌های مربوط به ۲۳۸۷ بیمار نشان داده شده در جدول ۱۰.٦*

	مرزهای طبقه (تعداد در طبقه، N_h) روش ساختن طبقه‌ها			انحراف معیار درون طبقه (انتساب درصدی به طبقه) روش ساختن طبقه‌ها		
طبقه	فراوانی ریشه‌ای	دامنهٔ برابر	چندکی	فراوانی ریشه‌ای	دامنهٔ برابر	چندکی
۱	۰–۹۹	۰–۳۹۸	۰–۷۱/۲	۱٤/۸۹	۷۸/۹۰	۷/۹۸
	(N_1=۸٥۸)	(N_1=۲۲۲٥)	(N_1=٤۷۷)	(٪۱٤/٥٦)	(٪۹۱/۹۷)	(٪۲/۹۳)
۲	۱۰۰–۱٤۹	۳۹۹–۷۸۰	۷۱/۳–۱۰۷/٥	۱۳/۸٥	۹٥/۷۷	۹/۷٥
	(N_2 =٥۷۰)	(N_2=۱۳٦)	(N_2=٤۷۸)	(٪۹/۰۰)	(٪٦/۸۲)	(٪۳/٥۹)
۳	۱٥۰–۲۹۹	۷۸۱–۱۱٦۲	۱۰۷/٦–۱٥۰/٤	٤۰/٦۱	۹۳/۲۲	۱۱/٥٥
	(N_3 =٦٦٤)	(N_3=۱۷)	(N_3=٤۷۷)	(٪۳۰/۷٦)	(٪۰/۸۳)	(٪٤/۲۳)
٤	۳۰۰–۷۹۹	۱۱٦۳–۱٥٤٤	۱٥۰/٥–۲۲۹/٦	۱۱٦/٦۲	۷۲/٦۲	۲۲/۳٤
	(N_4 =۲۷۲)	(N_4=٥)	(N_4=٤۷۸)	(٪۳٦/۱۹)	(٪۰/۱۹)	(٪۸/۲۲)
٥	۸۰۰–۱۹۹۹	۱٥٤٥–۱۹۲٦	۲۲۹/۷–۱۹۲٦	۳٦۰/۹٥	۹۲/۸۱	۲۲۰/۷۸
	(N_5 =۲۳)	(N_5=٤)	(N_5=٤۷۷)	(٪۹/٤۷)	(٪۰/۱۹)	(٪۸۱/۰٤)

* طبقه‌بندی در مورد متغیر *Totpaid*، کل مبلغ مطالبه شده، از بیمهٔ کمکهای درمانی در طول سال ۱۹۹٦ به ازای هر بیمار است.

روش دامنهٔ برابر. بـرای داده‌هایی از این قبیل که قویاً چوله به راست‌اند، روش *دامنهٔ برابر* به ایجاد طـبقه‌هـایی مـنجر می‌شود که بیشتر عناصر در طبقات با شمار پایینتر و تعداد بسیار کمی از عناصر در طبقات بالاتر قرار می‌گیرند. در مثال حاضر، طبقهٔ ۱ که پایینترین طبقه است شامل ۲۲۲۵ بیمار است در حـالی کـه طـبقهٔ ۵ فقـط ٤ بیمار دارد. چون انحراف معیارهای درون طبقه‌ای، σ_{hx}، بین طبقات با این سـاختمان الگوریتمی، بسیار نزدیک به یکدیگرند، انتساب بهینه منجر به این می‌شود که بیشتر عناصر به پایینـترین طـبقه منتسـب شـوند. در ایـن مـثال، انتساب بهینه تقریباً ۹۲٪ نمونه را به طبقهٔ ۱ اختصاص می‌دهد.

روش چـندکی. روش *چـندکی* بـرای ایجاد طبقات، منجر به این می‌شود که تعداد عناصر در همهٔ طـبقه‌هـا یکسان باشد. به علاوه، در وضعیتهایی نظیر مثال حاضر که در آن متغیر مورد استفاده در ایجاد طـبقات دارای چولگـی مثـبت بسـیار بـالایی است روش چندکی این نتیجه را خواهد داد که انحراف معیارهـای درون طبقه‌ای در بالاترین طبقات دارای بالاترین مقدار باشند (جدول ۱۲.٦ را ببینید). به این تـرتیب، انتسـاب بهیـنه در تـرکیب بـا این روش ساختن طبقات، باعث می‌شود که نسبت بزرگتری از نمونـه در طـبقات بالاتـر قـرار بگیـرد. در مثال حاضر، انتساب بهینه باعث خواهد شد که بیش از ۸۱٪ نمونه به طبقهٔ ۵ اختصاص یابد.

روش فراوانی ریشه‌ای. سـرانجام، روش فراوانی ریشه‌ای گرایش خواهد داشت که از کرانگینیهای نشـان داده شده توسط دو روش دیگر در ساختن طبقات پرهیز شود. یعنی نابرابری در تعداد عناصر در هـر طـبقه کمتر از اعداد به دست آمده از روش دامنهٔ برابر خواهد بود و نابرابری در انحراف معیارهای درون طـبقه‌ای نیز کمتر از انحراف معیارهای به دست آمده از روش چندکی خواهد شد. سپس انتساب بهیـنـــه بـه کـاهش تغییرپذیری بین طبقات نسبت به عناصر نمونهٔ منتسب گرایش خواهـــد داشت. در واقـع، این روش ساختن طبقات دربارهٔ متغیرهای دارای چولگی اندک، به انتساب بهینه‌ای منجر خواهد شـد کـه انـدازه‌هـای طـبقات نمونـهٔ حاصـل از آن، با آنچه که از انتساب برابر به دست می‌آمد تفاوتی قـــابل توجـه نداشته باشد. در وضعیتهایی که آزمودن فرضهایی دربارهٔ تفاوتهای بین طبقات نسبت به سـطوح یـک مـتغیر موردنظـر است مزیت فوق قابل توجه است، زیرا که برای کل اندازهٔ نمـونهٔ تعیین شـده، هـرگاه طـبقات دارای اندازه‌های نمونه‌ای برابر باشند توان آماری عموماً در حد اعلای خود قرار دارد. بالاخره، انتساب بهینه در ترکیب با روش فراوانی ریشه‌ای برای ساختن طبقات عموماً برآوردهایی را نـتیجه مـی‌دهـد کـه در مقایسـه بـا انتساب بهینه در ترکیب با سایر روشهای ساختن طبقات، دارای پایینترین خطاهای معیار است. در مثال حاضر، برای یک نمونهٔ تصادفی طبقه‌بندی شدهٔ فرضی متشکل از n=۵۰۰ بیمـار که اندازه‌های نمونهٔ طبقات آن از انتساب بهینه به دست آمده بود خطای معیار برآورد

سطح میانگین *totpaid* از روش دامنهٔ برابر، معادل ۳/۱۸ دلار، از روش چندکی ۱/۳۳ دلار و از روش فراوانی ریشه ای ۱/۱۹ دلار است (که از همه پایینتر است).

□

یکی از مسائل مربوط به ایجاد مرزهای طبقه‌ای عبارت است از L، تعداد واقعی، طبقاتی که باید ایجاد شوند. این مسئله در متونی توسط کیش [۱۱] و کوکران [۹] مورد بحث قرار گرفته است. به خصوص کوکران از مدلی که ذیلاً نشان داده می‌شود (رابطهٔ ۲۸.٦) استفاده می‌کند که در آن طبقه‌بندی براساس متغیر x است و متغیری که باید برآورد شود y است.

$$Var(\bar{y}_{str}) \approx \frac{\sigma_y^2}{n}\left[\frac{\rho^2}{L^2} + (1-\rho^2)\right] \qquad (۲۸.٦)$$

که در این رابطه ρ، همبستگی بین x و y است. واضح است که عامل $(\rho^2/L^2)+(1-\rho^2)$ نشان‌دهندهٔ کاهشی تقریبی است که واریانس به وسیلهٔ طبقه‌بندی نسبت به نمونه‌گیری تصادفی ساده خواهد داشت. برای همبستگی بین x و y که در بیشتر وضعیتها برقرار است از رابطهٔ (۲۸.٦) می‌توان پی برد که از داشتن بیش از پنج یا شش طبقه فایدهٔ چندانی حاصل نخواهد شد.

۹.٦ خلاصه

در این فصل، مفاهیم طبقه‌بندی را بسط دادیم. به خصوص، دربارهٔ برآورد کردن میانگینها، نسبتها و مجموعهای جامعه‌ای تحت نمونه‌گیری تصادفی طبقه‌بندی شده همراه با روشهای انتساب نمونه به طبقه‌ها بحث کردیم (برای مثال، انتساب برابر، انتساب متناسب، انتساب بهینه با و بدون محدودیتهای هزینه). دربارهٔ پس طبقه‌بندی بحث کردیم که مستلزم استفاده از یک برنامهٔ نمونه‌گیری تصادفی ساده همراه با شیوهٔ برآورد کردن مشابه با شیوه‌ای است که در نمونه‌گیری تصادفی طبقه‌بندی شده به کارمی‌رود. سرانجام، روشهای برآورد کردن اندازهٔ نمونهٔ مورد نیاز را تحت نمونه‌گیری تصادفی طبقه‌بندی شده ارائه کردیم.

تمرین

۱.٦ در جامعه‌ای متشکل از ۱۵۰۰ خانوار قرار است یک آمارگیری نمونه‌ای از خانوارها برای تعیین کل تعداد اشخاص ۱۸ سال به بالای این جامعه که یک یا چند دندان دائمی خود را از دست داده‌اند (غیر از دندانهای عقل) اجرا شود. چون تصور بر این است که این متغیر با سن و درآمد همبستگی دارد، با استفاده از داده‌های جمعیتی موجود، طبقاتی تشکیل شده‌اند که در جدول زیر نشان داده شده‌اند. قرار بر این است که یک نمونهٔ تصادفی طبقه‌بندی شده از ۱۰۰ خانواده گرفته شود.

متغیر	طبقه‌ها ۱	۲	۳	٤
سن				
میانگین	۳۰	۳۲	۲۵	۲۷
انحراف معیار	۱۵	۱۵	۱۰	۱۰
درآمد سالانهٔ خانواده (×۱۰۰۰ دلار)				
میانگین	۱۵	۷	۱۵	۸
انحراف معیار	۵	۳	۳	۲
تعداد خانواده‌ها	۳۰۰	۵۰۰	۱۰۰	٦۰۰

الف. با جزئیات کامل جبری تعیین کنید که چگونه می‌توان کل تعداد اشخاصی را که یک یا چند دندان خود را از دست داده‌اند برآورد کرد.

ب. اگر از انتساب متناسب استفاده شود تعداد خانواده‌هایی را که باید از هر طبقه انتخاب شوند تعیین کنید.

پ. اگر از انتساب بهینه براساس درآمد سالانهٔ خانواده استفاده شود تعداد خانواده‌هایی را که باید از هر طبقه انتخاب شوند تعیین کنید.

ت. اگر از انتساب بهینه براساس سن استفاده شود تعداد خانواده‌هایی را که باید از هر طبقه انتخاب شوند تعیین کنید.

ث. اگر سن و درآمد سالانهٔ خانواده هر دو در نظر گرفته شوند، نمونه را چگونه به طبقات منتسب می‌کنید؟

ج. واریانس توزیع درآمد سالانهٔ خانواده برای کل جامعه چقدر است؟

چ. فرض کنید تعداد اشخاص ۱۸ سال به بالا که دندان خود را در هر خانواده از دست داده‌اند همبستگی زیادی با درآمد خانواده داشته باشد. آیا احتمال دارد که نمونه‌گیری تصادفی طبقه‌بندی شده با انتساب متناسب برآوردی را نتیجه دهد که واریانس آن کمتر از واریانس حاصل از نمونه‌گیری تصادفی سادهٔ همین تعداد خانوار باشد؟

۲.٦ فرض کنیم داده‌های جدول ۸.۳ از یک نمونهٔ تصادفی طبقه‌بندی شده از ۱۲۰۰ کارگر در کارخانه‌ای به دست آمده باشد که نیروی کار آن براساس مواد زیانبخش برای ریه (بالا، متوسط، پایین) طبقه‌بندی شده است و برای انتساب نمونه از انتساب متناسب استفاده شده باشد.

الف. در هر طبقه چند کارگر وجود دارد؟

ب. میانگین ظرفیت حیاتی تحت فشار را در بین کارگران کارخانه برآورد کنید. این برآورد با میانگینی که از انتخاب نمونه از طریق نمونه‌گیری تصادفی ساده به دست می‌آمد چه فرقی دارد؟

پ. بازهٔ اطمینان ۹۵٪ را برای میانگین جامعه به دست آورید.

ت. مزیت نمونه‌گیری تصادفی طبقه‌بندی شده نسبت به آنچه که از نمونه‌گیری تصادفی ساده به دست می‌آمد در چیست؟

۳.٦ فرض کنیم قرار است یک آمارگیری خانوار برای برآورد کردن مشخصه‌های خانواده‌هایی که در آنها سرپرست خانوار زن است انجام شود. چون قبل از آمارگیری معلوم نیست که سرپرست خانوار کدام خانواده‌ها زن است، خانوارهای نمونه غربال خواهند شد و با آن دسته از خانوارهای نمونه که دارای سرپرست زن هستند به تفصیل مصاحبه خواهد شد. پیش‌بینی می‌شود که هزینهٔ غربالگری هر خانوار ۱۰ دلار و هزینهٔ مصاحبه با خانواری که سرپرست زن دارد ۵۰ دلار باشد. جمعیت، براساس اطلاعات آخرین سرشماری در مورد نسبت خانوارهای با سرپرست زن، به سه دسته طبقه‌بندی شده است. این طبقه‌ها در جدول زیر نشان داده شده‌اند. فرض کنید که واریانس مشخصه‌های مورد اندازه‌گیری در هر طبقه یکسان است و بودجه‌ای معادل ۱۰۰۰۰ دلار برای کارهای میدانی تخصیص یافته است. چند خانوار باید در هر طبقه نمونه‌گیری شوند؟

طبقه	تعداد خانوار	درصد خانوارهای دارای سرپرست زن
۱	۱۰۰۰۰	۲۵
۲	۲۰۰۰۰	۱۵
۳	۵۰۰۰	۱۰

٤.٦ تعداد ٤٠ کارگر ارائه شده در جدول ۸.۳ را یک نمونهٔ تصادفی ساده از ۱۲۰۰ کارگر یک کارخانه در نظر بگیرید.

الف. یک بازهٔ اطمینان ۹۰٪ را برای میانگین ظرفیت حیاتی اجباری جامعه محاسبه کنید.

ب. فرض کنید پیش از تحلیل داده‌ها معلوم است که این ۱۲۰۰ کارگر به صورت زیر توزیع شده‌اند:

N_1=۱۰۰۰ (تعداد کارگران در معرض مواد مضر با سطح بالا)

N_2=۱۰۰ (تعداد کارگران در معرض مواد مضر با سطح متوسط)

N_3=۱۰۰ (تعداد کارگران در معرض مواد مضر با سطح پایین)

نمونهٔ اولیهٔ جدول ۸.۳ را پس طبقه‌بندی کنید و بازهٔ اطمینان ۹۰٪ را برای میانگین ظرفیت حیاتی اجباری جامعه بسازید.

پ. بازه‌های بخشهای الف و ب را مقایسه کنید. کدام بیشتر است؟ چرا؟

۵.٦ یک شرکت تحقیقاتی بازاریابی متخصص در فعالیت مراقبتهای بهداشتی دارای پرونده‌ای با تقریباً ۱۵۰۰۰۰۰۰۰ اسم است که برحسب کدپستی تنظیم شده‌اند (این پرونده دارای ٦۵۰۰۰ کدپستی است). قرار است یک نمونهٔ تصادفی طبقه‌بندی شده با انتساب متناسب گرفته شود و کدهای پستی به عنوان طبقه‌ها به کار روند. هدف از این آمارگیری برآورد کردن نسبت اشخاصی است که احتمال دارد یک نوع مسواک برقی جدید را بخرند. اگر در نتیجهٔ این آمارگیری معلوم شود که ۱۵٪ جامعه یا بیشتر تمایل به خرید این کالا نشان داده‌اند، برنامهٔ بازاریابی فشرده‌ای آغاز خواهد شد. اگر برآورد این نسبت با ۹۵٪ اطمینان حول ۵٪ مقدار واقعی آن مطلوب باشد، به ازای هر کدپستی چند اسم باید نمونه‌گیری شود؟

٦.٦ آقای استر، یک متخصص بیماریهای واگیردار و یک بدنساز آماتور موفق است (بدنسازی بدون دارو، شیکاگو، ۱۹۸۹). او برای یک آمارگیری از بدنسازان ناحیهٔ شیکاگو به منظور تعیین نسبت کسانی که از استروییدهای تقویتی استفاده کرده‌اند برنامه‌ریزی می‌کند. چارچوب آماری او شامل فهرستهای عضویت است که از تمام باشگاههای تندرستی مجاز در بخشهای شش‌گانهٔ منطقهٔ کلانشهر شیکاگو به دست آورده است. او باشگاهها را براساس مشتریان دائمی آنها به سه گروه زیر طبقه‌بندی کرده است:

۱. نوکیسه‌های مرکز شهر
۲. کارگری مرکز شهر
۳. حومه

او پیش‌بینی می‌کند که نسبت استفاده از استروییدهای تقویتی در طبقهٔ ۱ که شامل بیشترین حریفان «محض» است دو برابر طبقهٔ ۲، و نسبت طبقهٔ ۳ تقریباً دوسوم نسبت در طبقهٔ ۲ خواهد بود. از فهرستهای عضویت که در اختیار دارد ۸۳٤۵ نفر را در طبقهٔ ۱ و ۵۲۸٦ نفر را در طبقهٔ ۲ و ٦۳۰۰ نفر را در طبقهٔ ۳ می‌شمارد. اگر بتواند ۱۰۰۰ نفر را برای نمونه انتخاب

کـند، بـه هر طبقه باید چند نفر تخصیص یابند؟ (فرض کنید نسبت کل استفاده از استروییـد ۱۲٪ است.)

۷.٦ استر (قهرمان تمرین ٦.٦) می‌تواند از انجمن بدنسازان مخالف با مصرف استروییدهای تقویتی مبـلغ ۲۵۰۰۰ دلار بـرای اجـرای آمارگیری خود بگیرد. این تنها منبع تأمین بودجهٔ او است. او بـرآورد مـی‌کـند که هر مصاحبه در طبقهٔ ۱ و ۲ معادل ۳۰/۰۰ دلار و در طبقهٔ ۳ معادل ٤۰/۰۰ دلار هزیـنه خواهـد داشت. با توجه به این برآوردها، در هر طبقه با چند نفر می‌تواند مصاحبه کند؟

۸.٦ اسـتر (همان استر تمرین ٦.٦ و ۷.٦) احساس می‌کند که آمارگیری موردنظر او تنها در صورتی ارزش اجـرا دارد کـه او ۹۵٪ اطمینان داشته باشد که نرخ برآورد شده حول ۱۵٪ مقدار واقعی خواهـد بـود. او پیـش‌بیـنی می‌کند که نرخ کل واقعی باید حدود ۱۲٪ باشد. با توجه به سطح بودجـهٔ او کـه در تمـرین ۷.٦ به آن اشاره شد آیا بودجهٔ کافی برای تأمین ویژگیهای موردنظر خود در اختیار دارد؟

۹.٦ در یـک آمارگیـری بزرگ جامعه، ۱۵۰۰۰ نفر با عکسبرداری از قفسهٔ سینه غربال شدند. توجه پزشـکان بـه امکـان اتسـاع سرخرگ ریوی در ۲۳۰ نفر از این بیماران جلب شد. عکسبرداری مجـدد، ایـن اتساع را در ۲۰۳ نفر از این ۲۳۰ بیمار تأیید کرد. یک نمونهٔ ۱۷۵ تایی از ۱٤۷۷۰ عکـس قفسـهٔ سینه که هیچ اتساع سرخرگ ریوی در آنها دیده نشده بود ۱۲ مورد را نشان داد که عملاً از لحاظ اتساع سرخرگ ریوی مثبت بودند.

الف. براساس این داده‌ها، میزان شیوع اتساع سرخرگ ریوی در جامعه چقدر است؟

ب. برای این نرخ شیوع، بازهٔ اطمینان ۹۵٪ را به دست آورید.

۱۰.٦ در روز سـوم سـپتامبر سـال ۱۹۸۹ یـک مسابقهٔ ماراتون در یک شهر بزرگ اجرا شد. براساس درخواستهای شرکت در مسابقه، داده‌های زیر به دست آمد.

گروه سنی	N	میانگین تعداد دونده‌هایی که مسابقه را تکمیل کردند ($\overline{X}$)	انحراف معیار σ_x
≤۳۰	۲۳۰۰	۱/۹	۰/٦
۳۰–٤۹	۱٤۷۸	۲/۳	۰/۸
۵۰+	۹۷۸	۳/۱	۰/۷

مطـلوب اسـت انـتخاب نمونـه‌ای از تقریـباً ۵۰۰ نفر از این فهرست برای برآورد کردن تعداد متوسط مایلهایی که در هفته به منظور آماده شدن برای این مسابقهٔ ماراتون دویده‌اند. اگر فرض بـر ایـن باشـد کـه این متوسط ممکن است متناسب با تعداد افرادی باشد که مسابقه را تکمیل کـرده‌انـد، آیـا اسـتفاده از نمونـهٔ تصـادفی طـبقه‌بندی شده با انتساب متناسب نسبت به نمونهٔ تصادفی ساده فایدهٔ بیشتری خواهد داشت؟

۱۱.٦ ایـن تمـرین مـربوط می‌شود به مثال تشریحی در مورد غربالگری دوقلوها که در بخش ۵.٦ از آنهـا بحـث شـد. جفـتهای زن ـ زن سفیدپوست به صورتی که در مثال تشریحی توصیف شد براسـاس تـاریخ تولـد یکسـان و شمارهٔ تأمین اجتماعی آنان که هفت رقم اول آنها با هم جور بودند ساخته شدند. شمارهٔ جفتهایی که به این ترتیب ساخته شدند به شرح زیر است:

چارک براساس تفاوت دنباله‌ای شمارهٔ تأمین اجتماعی	تعداد جفتهای ساخته شده از روی پرونده‌های بیمهٔ مراقبتهای درمانی
چارک اول	۱۰۰۲٤
دو چارک میانی	۲۰۰۳۱
چارک چهارم	۱۰۰۲٤

فـرض کـنید حـدس شـما این است که فراوانی کلی دوقلوها در میان این جفتها ۲۰٪ است و فـراوانی دوقـلوها در چـارک اول دو برابـر چارک دوم و سه برابر چارک سوم است. اگر قرار باشـد یـک نمونـهٔ تصـادفی طـبقه‌بـندی شده متشکل از ۳۰۰ جفت را از این پرونده به منظور بـرآورد فـراوانی دوقـلوها در ایـن پرونده انتخاب کنید، انتساب بهینه در هر یک از این سه تا چارک چقدر خواهد بود؟

فـرض کـنید بـه جـای یـک نمونـهٔ تصادفی طبقه‌بندی شده متشکل از ۳۰۰ جفت تصمیم گرفته‌اند که یک نمونهٔ تصادفی ساده از ۳۰۰ جفت بگیرند و داده‌های زیر به دست آمده‌اند:

چارک براساس تفاوت دنباله‌ای شمارهٔ تأمین اجتماعی	تعداد جفتهای نمونه‌گیری شده	تعداد دوقلوها
چارک اول	۹۰	٤۰
دو چارک میانی	۱۱۰	۱۰
چارک چهارم	۱۰۰	۱

الف. برآورد فراوانی دوقلوها و خطای معیار آن چقدر است؟

ب. برآورد پس طبقه‌بندی فراوانی دوقلوها و خطای معیار آن را پیدا کنید.

کتابشناسی

The following articles present sample surveys in which stratification is used.

1. Hemphill, F. M., A sample survey of home injuries, *Public Health Reports,* 67: 1026, 1952.
2. *Horvitz*, D. G., Sampling and field procedures of the Pittsburgh morbidity survey, *Public Health Reports,* 67: 1003, 1952.
3. Goldberg, J., Levy, P. S., Mullner, R., Gelfand, H., Iverson, N., Lemeshow, S., and Rothrock, J., Factors affecting trauma center utilization in Illinois, *Medical Care,* 19: 547, 1981.
4. Goldberg, J., Miles, T. P., Furner, S., Meyer, J., Hinds, A., Ramakrishnan, V., Lauderdale, D., and Levy, P. S., Identification of a biracial cohort of male and female twins age 65 and above in the United States, *American Journal of Epidemiology,* 145: 175-183, 1997.
5. Barner, B. M., and Levy, P. S., State-wide shoulder belt usage by type of roadway and posted speed limit: A three year comparison, *37th Annual Proceedings Association for the Advancement of Automotive Medicine,* Association for the Advancement of Automotive Medicine, Des Plaines, Ill., 1994.

The following are recent reviews of stratification in sample surveys.

6. Parsons, V., Stratified sampling. In *Encyclopedia of Biostatistics,* Armitage, P. and Colton, T., Eds., Wiley, Chichester, U.K., 1998.
7. Malec, D., Allocation in stratified sampling. In *Encyclopedia of Biostatistics,* Armitage, P., and Colton, T. Eds., Wiley, Chichester, U.K., 1998.
8. Brewer, K. R. W., Stratified designs. In *Encyclopedia of Statistical Sciences,* Johnson, N., and Kotz, S. Eds., Wiley, New York, 1984.

The following books provide an overview of methodology used in the construction of strata as well as a thorough overview of stratification.

9. Cochran, W. G., *Sampling Techniques,* 3rd ed., Wiley, New York, 1977.
10. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods and Theory*, Vols. 1 and 2, Wiley, New York, 1953.
11. Kish, L., *Survey Sampling,* Wiley, New York, 1965.

The following are "classic" articles that present methodology for construction of strata.

12. Dalenius, T., *Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice*, Almqvist and Wicksell, Stockholm, 1957.
13. Dalenius, T., and Hodges, J. L., Jr., Minimum variance stratification, *Journal of the American Statistical Association,* 54: 88-101, 1959.

فصل ۷

برآورد نسبتی

در این فصل مفهوم نسبت $\frac{\bar{x}}{\bar{y}}$ دو میانگین نمونه‌ای $\bar{x}$ و $\bar{y}$ را ارائه می‌کنیم. این نسبت به عنوان برآوردی از نسبت $\frac{\overline{X}}{\overline{Y}}$ میانگینهای دو متغیر x و y در جامعه به کار می‌رود. ولی مهمتر از آن، کاربرد این نسبت به عنوان ابزاری برای به دست آوردن برآورد X، مجموع کل جامعه است که می‌توان آن را دقیقتر از برآورد x' به وسیلهٔ ضرب سادهٔ مجموع نمونه‌ای x در $\frac{N}{n}$ یعنی عکس کسر نمونه‌گیری تعیین کرد. این روش را *برآورد کردن نسبتی* و برآوردهای حاصل از آن را *برآوردهای نسبتی* می‌نامند.

برای شروع این مبحث به مثالی از برآورد نسبتی توجه می‌کنیم.

مثال تشریحی: جامعه‌ای را در نظر می‌گیریم که دارای هشت محله است. فرض کنید می‌خواهیم نسبت R کل هزینه‌های دارویی X را به کل هزینه‌های درمانی Y در میان تمام اشخاص موجود در جامعه برآورد کنیم. برای انجام این کار قرار است نمونهٔ تصادفی ساده‌ای متشکل از دو محله گرفته شود و با یکایک خانوارها در هر محلهٔ نمونه مصاحبه به عمل آید.

عناصر این طرح نمونه‌گیری، محله‌ها هستند و نسبت $\frac{X}{Y}$، کل هزینه‌های دارویی به کل هزینه‌های درمانی، را می‌توان برحسب r برآورد کرد که $r=\frac{x}{y}$ نسبت به دست آمده از نمونه است. فرض کنیم داده‌های مربوط به محله‌ها به صورتی باشند که در جدول ۱.۷ ارائه شده است.

حالا فرض کنید مثلاً محله‌های ۲ و ۵ برای نمونه انتخاب شده‌اند. پس داریم

$$x = ۵۰۰۰۰ + ۱۵۰۰۰۰ = ۲۰۰۰۰۰$$
$$y = ۲۰۰۰۰۰ + ۴۵۰۰۰۰ = ۶۵۰۰۰۰$$

و

$$r = \frac{x}{y} = \frac{۲۰۰۰۰۰}{۶۵۰۰۰۰} = ۰/۳۰۸$$

به این ترتیب، نسبت کل هزینه‌های دارویی به کل هزینه‌های درمانی برابر است با ۰/۳۰۸. □

جدول ۱.۷ هزینه‌های دارویی و کل هزینه‌های درمانی بین همهٔ ساکنان هشت محله

محله	کل هزینه‌های دارویی X (دلار)	کل هزینه‌های درمانی Y (دلار)
۱	۱۰۰۰۰۰	۳۰۰۰۰۰
۲	۵۰۰۰۰	۲۰۰۰۰۰
۳	۷۵۰۰۰	۳۰۰۰۰۰
٤	۲۰۰۰۰۰	٦۰۰۰۰۰
۵	۱۵۰۰۰۰	٤۵۰۰۰۰
٦	۱۷۵۰۰۰	۵۲۰۰۰۰
۷	۱۷۰۰۰۰	٦۸۰۰۰۰
۸	۱۵۰۰۰۰	٤۵۰۰۰۰
مجموع	۱۰۷۰۰۰۰	۳۵۰۰۰۰۰

۱.۷ برآورد کردن نسبتی تحت نمونه‌گیری تصادفی ساده

مثال فوق را می‌توان به شرح زیر تعمیم داد. فرض کنید یک نمونهٔ تصادفی سادهٔ n عنصری از جامعه‌ای N عنصری داریم و می‌خواهیم *نسبت* R دو مجموع کل X و Y جامعه را برآورد کنیم. واضح است که R از فرمول

$$R = \frac{X}{Y}$$

یا از معادل آن

$$R = \frac{\overline{X}}{\overline{Y}}$$

به دست می‌آید.

نسبت R می‌تواند به وسیلهٔ r با استفاده از فرمول زیر

$$r=\frac{x'}{y'}$$

یا از دو فرمول زیر که از نظر جبری معادل‌اند به دست آید

$$r=\frac{x}{y} \qquad \text{و} \qquad r=\frac{\bar{x}}{\bar{y}}$$

برآورد r ، برآورد نسبتی نامیده می‌شود زیرا هم صورت کسر $\bar{x}$ و هم مخرج کسر $\bar{y}$ در معرض تغییرات نمونه‌گیری قرار دارند.

حالا به مثال ارائه شده در مقدمهٔ این فصل برگردیم و ببینیم که آیا نسبت برآورد شده، یک نسبت نااریب است یا نه.

مثال تشریحی: در مثال بالا $\binom{۸}{۲}=۲۸$ نمونهٔ تصادفی سادهٔ ممکن دو عنصری از جامعهٔ متشکل از ۸ عنصر وجود دارند. این نمونه‌های ممکن همراه با مقادیر x، y ، و r به دست آمده از هر نمونه در جدول ۲.۷ فهرست شده‌اند.

توزیـــع نمونه‌گیری r ، برآورد نسبت، دارای میانگین $E(r)$ و خطای معیار $SE(r)$ به شرح زیر است (تابلوی ۳.۲ را ببینید)

$$E(r)=\left(\frac{۱}{۲۸}\right)\times(۰/۳۰۰+۰/۲۹۲+\cdots+۰/۲۸۳)=۰/۳۰۵۴$$

$$SE(r)=\left(\frac{۱}{\sqrt{۲۸}}\right)\times\left[(۰/۳۰۰-۰/۳۰۵۴)^{۲}+(۰/۲۹۲-۰/۳۰۵۴)^{۲}+\cdots+(۰/۲۸۳-۰/۳۰۵۴)^{۲}\right]^{۱/۲}=۰/۰۲۶۸$$

مقدار واقعی R از جدول ۱.۷ عبارت است از

$$R=\frac{۱۰۷۰۰۰۰}{۳۵۰۰۰۰۰}=۰/۳۰۵۷$$

تفاوت بین $E(r)$ میانگین توزیع نمونه‌گیری r ، برآورد نسبت و نسبت واقعی R جامعه ناشی از خطای گرد کردن نیست. به این ترتیب می‌بینیم که r ، برآورد نسبت الزاماً یک برآورد نااریب از نسبت جامعه‌ای R نیست. ولی در بیشتر موارد اریبی r ، برآورد نسبت کم است و برآوردهایی از این دست موارد استفادهٔ زیادی دارند.

□

توجه کنید که برای محاسبهٔ $SE(r)$ در مثال بالا، تولید همهٔ نمونه‌های ممکن و محاسبهٔ خطای معیار توزیع نمونه‌گیری لازم است. البته، در کاربست واقعی، همهٔ توزیع نمونه‌گیری را تولید نمی‌کنیم. ولی برخلاف سایر برآوردهایی که تاکنون در این کتاب بررسی کردیم (مانند میانگینها، مجموعها و نسبتها)، در r ، برآورد نسبت، هم صورت و هم مخرج کسر در معرض تغییرات نمونه‌گیری قرار دارند.

جدول ۲.۷ نمونه‌های ممکن دوعنصری از جامعه‌ای متشکل از هشت عنصر (جدول ۱.۷)

شمارهٔ محله‌ها در نمونه	x	y	r
۱ , ۲	۱۵۰۰۰۰	۵۰۰۰۰۰	۰/۳۰۰
۱ , ۳	۱۷۵۰۰۰	۶۰۰۰۰۰	۰/۲۹۲
۱ , ٤	۳۰۰۰۰۰	۹۰۰۰۰۰	۰/۳۳۳
۱ , ۵	۲۵۰۰۰۰	۷۵۰۰۰۰	۰/۳۳۳
۱ , ۶	۲۷۵۰۰۰	۸۲۰۰۰۰	۰/۳۳۵
۱ , ۷	۲۷۰۰۰۰	۹۸۰۰۰۰	۰/۲۷۶
۱ , ۸	۲۵۰۰۰۰	۷۵۰۰۰۰	۰/۳۳۳
۲ , ۳	۱۲۵۰۰۰	۵۰۰۰۰۰	۰/۲۵۰
۲ , ٤	۲۵۰۰۰۰	۸۰۰۰۰۰	۰/۳۱۳
۲ , ۵	۲۰۰۰۰۰	۶۵۰۰۰۰	۰/۳۰۸
۲ , ۶	۲۲۵۰۰۰	۷۲۰۰۰۰	۰/۳۱۳
۲ , ۷	۲۲۰۰۰۰	۸۸۰۰۰۰	۰/۲۵۰
۲ , ۸	۲۰۰۰۰۰	۶۵۰۰۰۰	۰/۳۰۸
۳ , ٤	۲۷۵۰۰۰	۹۰۰۰۰۰	۰/۳۰۶
۳ , ۵	۲۲۵۰۰۰	۷۵۰۰۰۰	۰/۳۰۰
۳ , ۶	۲۵۰۰۰۰	۸۲۰۰۰۰	۰/۳۰۵
۳ , ۷	۲٤۵۰۰۰	۹۸۰۰۰۰	۰/۲۵۰
۳ , ۸	۲۲۵۰۰۰	۷۵۰۰۰۰	۰/۳۰۰
٤ , ۵	۳۵۰۰۰۰	۱۰۵۰۰۰۰	۰/۳۳۳
٤ , ۶	۳۷۵۰۰۰	۱۱۲۰۰۰۰	۰/۳۳۵
٤ , ۷	۳۷۰۰۰۰	۱۲۸۰۰۰۰	۰/۲۸۹
٤ , ۸	۳۵۰۰۰۰	۱۰۵۰۰۰۰	۰/۳۳۳
۵ , ۶	۳۲۵۰۰۰	۹۷۰۰۰۰	۰/۳۳۵
۵ , ۷	۳۲۰۰۰۰	۱۱۳۰۰۰۰	۰/۲۸۳
۵ , ۸	۳۰۰۰۰۰	۹۰۰۰۰۰	۰/۳۳۳
۶ , ۷	۳٤۵۰۰۰	۱۲۰۰۰۰۰	۰/۲۸۸
۶ , ۸	۳۲۵۰۰۰	۹۷۰۰۰۰	۰/۳۳۵
۷ , ۸	۳۲۰۰۰۰	۱۱۳۰۰۰۰	۰/۲۸۳

از این رو نمی‌توان عبارت دقیقی برای خطای معیار آن به دست آورد. ولی هنسن و همکاران [۱] پیشنهاد کرده‌اند که اگر ضریب تغییرات $V(\bar{y})$ مخرج کسر، $\bar{y}$، یا r (با استفاده از حالت $r=\frac{\bar{x}}{\bar{y}}$) کوچکتر از یا برابر با ۰/۰۵ باشد، آنگاه می‌توان خطای معیار r، یعنی $SE(r)$ را از عبارت زیر تقریب زد.

$$SE(r)\approx\left(\frac{R}{\sqrt{n}}\right)\times\left(V_x^2+V_y^2-2\rho_{xy}V_xV_y\right)^{1/2}\times\sqrt{\frac{N-n}{N-1}} \qquad (۱.۷)$$

که در آن ρ_{xy}، ضریب همبستگی بین x و y است و به صورت زیر تعریف می‌شود

$$\rho_{xy}=\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})/N}{\sigma_x\sigma_y} \qquad (۲.۷)$$

حال بعضی محاسبات را برای مثالی که با آن کار می‌کردیم انجام می‌دهیم.

مثال تشریحی: برای داده‌های جدول ۱.۷ پارامترهای جمعیتی زیر را داریم [رجوع کنید به عبارتهای تابلوی ۱.۲ و معادله‌های (۷.۲) و (۲.۷)]:

$\sigma_x=۴۹۴۱۸/۵$	$\bar{X}=۱۳۳۷۵۰$	$R=۰/۳۰۵۷$
$\sigma_y=۱۵۲۷۰۴/۸$	$\bar{Y}=۴۳۷۵۰۰$	$V_x=۰/۳۶۹۵$
$N=۸$	$\rho_{xy}=۰/۹۲۷۲$	$V_y=۰/۳۴۹۰$

برای داده‌های جدول ۲.۷، ضریب تغییرات مخرج کسر، $\bar{y}$، برای r عبارت است از

$$V(\bar{y})=\left(\frac{1}{\bar{Y}}\right)\times\left(\frac{\sigma_y}{\sqrt{n}}\right)\times\sqrt{\frac{N-n}{N-1}}=\left(\frac{1}{۴۳۷۵۰۰}\right)\times\left(\frac{۱۵۲۷۰۴/۸}{\sqrt{۲}}\right)\times\sqrt{\frac{۸-۲}{۸-۱}}=۰/۲۲۸۵$$

محاسبه با معادلهٔ (۱.۷) مقدار

$$SE(r)\approx\left(\frac{۰/۳۰۵۷}{\sqrt{۲}}\right)\times\left[(۰/۳۶۹۵)^2+(۰/۳۴۹۰)^2-2(۰/۹۲۷۲)(۰/۳۶۹۵)(۰/۳۴۹۰)\right]^{1/2}\times\sqrt{\frac{۸-۲}{۸-۱}}=۰/۰۲۷۷$$

را در مقایسه با مقدار واقعی آن که ۰/۰۲۶۸ است نتیجه می‌دهد. چون ضریب تغییرات $\bar{y}$ بزرگتر از ۰/۰۵ است، تقریب ارائه شده در معادلهٔ (۱.۷) به طور نرمال نباید برای $SE(r)$ مورد استفاده قرار بگیرد. ولی در این مورد به نظر می‌رسد که به صورتی معقول خوب عمل می‌کند.

حالا انتخاب نمونه‌هایی با اندازهٔ $n=۷$ را از جامعهٔ مورد بررسی در نظر بگیرید. توزیع دقیق r برای نمونه‌هایی با $n=۷$ عنصر در جدول ۳.۷ نشان داده شده است.

خطای معیار دقیق $SE(r)$ برای r که با شمارش کلیهٔ نمونه‌های ممکن به دست آمده است ۰/۰۰۶۳ است. محاسبات نشان می‌دهد که

$$V(\bar{y})=\left(\frac{۱}{۴۳۷۵۰۰}\right)\times\left(\frac{۱۵۲۷۰۴/۸}{\sqrt{۷}}\right)\times\sqrt{\frac{۸-۷}{۸-۱}}=۰/۰۴۹۹$$

بنابراین، چون $V(\bar{y})$=۰/۰۴۹۹≤۰/۰۵ انتظار داریم تقریب معادلهٔ (۱.۷) تقریب خوبی از مقدار واقعی $SE(r)$ فراهم کند. این محاسبه به شرح زیر است:

$$SE(r)\approx\left(\frac{۰/۳۰۵۷}{\sqrt{۷}}\right)\times\left[(۰/۳۶۹۵)^{۲}+(۰/۳۴۹۰)^{۲}-۲(۰/۹۲۷۲)(۰/۳۶۹۵)(۰/۳۴۹۰)\right]^{۱/۲}\times\sqrt{\frac{۸-۷}{۸-۱}}=۰/۰۰۶۱$$

به این ترتیب می‌بینیم که تقریب (۱.۷) در این مثال تطابق بسیار نزدیکی با خطای معیار واقعی r دارد.

جدول ۳.۷ نمونه‌های هفت‌تایی از جامعهٔ جدول ۱.۷

نمونه	x	y	r
۱ ,۲ ,۳ ,٤ ,٥ ,٦ ,۷	۹۲۰۰۰۰	۳۰۵۰۰۰۰	۰/۳۰۱٦
۱ ,۲ ,۳ ,٤ ,٥ ,٦ ,۸	۹۰۰۰۰۰	۲۸۲۰۰۰۰	۰/۳۱۹۱
۱ ,۲ ,۳ ,٤ ,٥ ,۷ ,۸	۸۹۵۰۰۰	۲۹۸۰۰۰۰	۰/۳۰۰۳
۱ ,۲ ,۳ ,٤ ,٦ ,۷ ,۸	۹۲۰۰۰۰	۳۰۵۰۰۰۰	۰/۳۰۱٦
۱ ,۲ ,۳ ,٥ ,٦ ,۷ ,۸	۸۷۰۰۰۰	۲۹۰۰۰۰۰	۰/۳۰۰۰
۱ ,۲ ,٤ ,٥ ,٦ ,۷ ,۸	۹۹۵۰۰۰	۳۲۰۰۰۰۰	۰/۳۱۰۹
۱ ,۳ ,٤ ,٥ ,٦ ,۷ ,۸	۱۰۲۰۰۰۰	۳۳۰۰۰۰۰	۰/۳۰۹۱
۲ ,۳ ,٤ ,٥ ,٦ ,۷ ,۸	۹۷۰۰۰۰	۳۲۰۰۰۰۰	۰/۳۰۳۱

□

خـــــطای معیار برآورد یک نسبت را می‌تـــوان از روی این داده‌ها و با جایگزین کردن $\hat{\sigma}_x^{۲}$ به جای $\sigma_x^{۲}$، $\hat{\sigma}_y^{۲}$ به جای $\sigma_y^{۲}$، $\hat{\rho}_{xy}$ به جای ρ_{xy}، $\bar{x}$ به جای $\bar{X}$، $\bar{y}$ به جای $\bar{Y}$ و r به جای R در رابطهٔ (۱.۷) برآورد کرد. برآورد حاصل، $\hat{SE}(r)$، از فرمول زیر به دست می‌آید

$$\hat{SE}(r)=\left(\frac{r}{\sqrt{n}}\right)\times\left(\hat{V}_x^{۲}+\hat{V}_y^{۲}-۲\hat{\rho}_{xy}\hat{V}_x\hat{V}_y\right)^{۱/۲}\times\sqrt{\frac{N-n}{N-۱}}\qquad(۳.۷)$$

که در آن،

$$\hat{V}_x^{۲}=\left(\frac{N-۱}{N}\right)\left(\frac{s_x^{۲}}{\bar{x}^{۲}}\right)$$

$$\hat{V}_y^{۲}=\left(\frac{N-۱}{N}\right)\left(\frac{s_y^{۲}}{\bar{y}^{۲}}\right)$$

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

به طور کلی این تقریب تنها هنگامی مورد استفاده قرار می‌گیرد که نابرابری زیر برقرار باشد:

$$\frac{s_y}{\sqrt{n}\times\bar{y}}\times\sqrt{\frac{N-n}{N}} \le ۰/۰۵$$

برای سهولت کار، فرمولهایی که می‌توانند برای برآورد کردن نسبتی مورد استفاده قرار گیرند در تابلوی ۱.۷ خلاصه شده‌اند.

تابلوی ۱.۷ فرمولهای برآورد کردن نسبتی تحت نمونه‌گیری تصادفی ساده

پارامترهای جامعه

$$R = \frac{X}{Y} = \frac{\bar{X}}{\bar{Y}}$$

$$SE(r) \approx \left(\frac{R}{\sqrt{n}}\right)\times\left(V_x^2 + V_y^2 - 2\rho_{xy}V_x V_y\right)^{1/2}\times\sqrt{\frac{N-n}{N-1}}$$

برآوردهای نمونه‌ای

$$r = \frac{x'}{y'} = \frac{x}{y} = \frac{\bar{x}}{\bar{y}}$$

$$\hat{SE}(r) = \left(\frac{r}{\sqrt{n}}\right)\times\left(\hat{V}_x^2 + \hat{V}_y^2 - 2\hat{\rho}_{xy}\hat{V}_x\hat{V}_y\right)^{1/2}\times\sqrt{\frac{N-n}{N-1}}$$

بازهٔ اطمینان $(1-\alpha)$۱۰۰ درصدی را می‌توان به صورت زیر ساخت

$$r - z_{1-(\alpha/2)}\hat{SE}(r) \le R \le r + z_{1-(\alpha/2)}\hat{SE}(r)$$

V_x و V_y در معادلهٔ (۷.۲) تعریف شده‌اند. X ، $\bar{X}$ ، σ_x ، Y ، $\bar{Y}$ و σ_y به همان صورت تعریف شده در تابلوی ۱.۲ هستند. ρ_{xy} به صورت تعریف شده در معادلهٔ (۲.۷) است. x ، $\bar{x}$ ، x' ، y ، $\bar{y}$ ، y' ، S_x^2 و S_y^2 به صورت تعریف شده در تابلوی ۲.۲ هستند. پس

$$\hat{V}_x^2 = \left(\frac{N-1}{N}\right)\left(\frac{s_x^2}{\bar{x}^2}\right)$$

$$\hat{V}_y^2 = \left(\frac{N-1}{N}\right)\left(\frac{s_y^2}{\bar{y}^2}\right)$$

$$\hat{\rho}_{xy} = \sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

مثال تشریحی: فرض کنید نمونهٔ متشکل از محله‌های ۱، ۲، ۳، ٤، ۵، ٦ و ۸ را از جدول ۱.۷ انتخاب کرده‌ایم. محاسبه می‌کنیم که

$$\frac{s_y}{\sqrt{n\bar{y}}}\times\sqrt{\frac{(N-n)}{N}}=۰/۰۴۶۸ \quad \text{و} \quad \bar{y}=۴۰۲۸۵۷/۱ \quad ، \quad s_y=۱۴۱۰۳۳/۶$$

که کمتر از ۰/۰۵ است. به این ترتیب انتظار داریم که تقریب (۳.۷) برآورد خوبی از خطای معیار نسبت r ، برآورد فراهم نماید. برای این مثال داریم

$$\hat{\rho}_{xy}=۰/۹۹۰۰ \qquad r=۰/۳۱۹۱$$
$$s_y=۱۴۱۰۳۳/۶ \qquad s_x=۵۴۸۲۶/۶$$
$$\bar{y}=۴۰۲۸۵۱/۱ \qquad \bar{x}=۱۲۸۵۷۱/۴$$
$$\hat{V}_y^{۲}=۰/۱۰۷۲ \qquad \hat{V}_x^{۲}=۰/۱۵۹۱$$

پس

$$\hat{SE}(r)=\left(\frac{۰/۳۱۹۱}{\sqrt{۷}}\right)\times\left[۰/۱۵۹۱+۰/۱۰۷۲-۲(۰/۹۹۰۰)\sqrt{۰/۱۵۹۱}\times\sqrt{۰/۱۰۷۲}\right]^{۱/۲}\times\left(\frac{۸-۷}{۸-۱}\right)^{۱/۲}=۰/۰۰۴۰$$

در مقام مقایسه، مقدار واقعی جامعه‌ای $SE(r)$ برابر با ۰/۰۰۶۳ است. این تفاوت هیچ تعجبی ندارد زیرا معلوم شده است که برآورد واریانسها به شدت متغیرند.

□

r ، برآورد نسبت، به طور کلی یک برآورد اریب از R است. ولی هنگامی که اندازه‌های نمونه به صورتی معقول بزرگ باشند اریبی عموماً کوچک است و بازه‌های اطمینان تقریبی برای نسبت R مجهول جامعه را می‌توان با استفاده از برآورد خطای معیار r ، یعنی $\hat{\mathrm{SE}}(r)$ ساخت.

برآوردهای نسبتی در آمارگیریهای نمونه‌ای اهمیت دارند، بخصوص هنگامی که واحدهای شمارش همان واحدهای اولیه نیستند. برای مثال، اگر یک نمونهٔ تصادفی ساده به منظور برآورد میانگین تعداد روزهای غیبت از کار به علت بیماری سخت به ازای هر نفر گرفته شده باشد، یک برآورد نسبتی را به صورتی که در بالا توصیف شد می‌توان به کار برد که صورت کسر آن تعداد روزهای غیبت از کار به ازای هر خانوار و مخرج کسر آن تعداد اشخاص در خانوار خواهد بود. هم صورت و هم مخرج کسر در معرض تغییرپذیری نمونه‌گیری خواهند بود.

در بخش ۵.۳ در برآورد کردن میانگینهای جامعه‌ای برای زیرگروههای جامعهٔ تحت نمونه‌گیری تصادفی ساده بحث کردیم. برآورد مناسب برای این وضعیت، حالتی خاص از نسبت برآورد شده است که مخرج کسر آن تعداد اعضای زیرگروه در نمونه است. اگر برنامهٔ نمونه‌گیری، انتخاب نمونهٔ تصادفی ساده‌ای از عناصر باشد، این صورت برآورد نسبتی، برآوردی نااریب از میانگین زیرگروه مناسب است.

برآورد کردن نسبتها و خطاهای معیار آنها را می‌توان با نرم‌افزارSUDAAN و با استـــفاده از مدول *PROC RATIO* در این نرم‌افزار اجرا کرد. آنچه درپی می‌آید مجموعه فرمانهایی را نشان می‌دهد که برای مثال تشریحی قبل مورد استفاده قرار می‌گیرد.

```
1 PROC RATIO DATA=TAB7PT1   FILETYPE=SAS    DESIGN=STRWOR;
2 TOTCNT TOTCNT;
3 WEIGHT WT1;
4 NEST_ONE_;
5 NUMBER PHARMEXP;
6 DENOM TOTMEDEX;
7 SETENV COLWIDTH = 1;
8 SETENV DECWIDTH = 4;
```

فرمان اول نشان می‌دهد که شیوهٔ *PROC RATIO* قرار است مورد استفاده قرار گیرد و داده‌های نمونه در پروندهٔ داده‌های SAS به نام *TAB7PT1.SSD* قرار دارند. شکل ظاهری این پرونده به صورت زیر است:

AREA	PHARMEXP	TOTMEDEX	TOTCNT	WT1
1	100000	300000	8	1.1428571
2	50000	200000	8	1.1428571
3	75000	300000	8	1.1428571
4	200000	600000	8	1.1428571
5	150000	450000	8	1.1428571
6	175000	520000	8	1.1428571
8	150000	450000	8	1.1428571

پروندهٔ داده‌ها در بالا دارای ۷ سابقه (رکورد)، یعنی یک سابقه برای هر محلهٔ نمونه است. هر سابقه دارای پنج متغیر به شرح زیر است:

area شمارهٔ شناسایی محله.

pharmexp کل هزینه‌های دارویی مربوط به هر محله.

totmedex کل هزینه‌های درمانی مربوط به هر محله.

totcnt کل تعداد N محله‌ها در جامعه.

wt1 وزن نمونه‌گیری، N/n .

فرمان DESIGN=STRWOR همراه با فرمان چهارم NEST=_ONE_ نشان می‌دهند که طرح نمونه‌گیری، نمونه‌گیری تصادفی ساده بدون جایگذاری است. فرمان دوم نشان می‌دهد که N ، اندازهٔ جامعه، به صورت متغیر TOTCNT در هر سابقه قرار دارد و در این مثال برابر با ۸ است. فرمان سوم مشخص می‌سازد که وزن نمونه‌گیری $\frac{۸}{۷}$=۱/۱۴۲۹ در هر سابقهٔ نمونه به صورت متغیر *WT1* ظاهر می‌شود. بالاخره فرمانهای پنجم و ششم، متغیرهای صورت و مخرج کسر را برای محاسبهٔ نسبت

مشخص می‌کنند و فرمان هفتم پهنای ستون و تعداد رقمهای دهدهی را در خروجی حاصل تعیین می‌کند.

خروجی حاصل از فرمانهای بالا در پایین نشان داده شده است. توجه کنید که برآورد خطای معیار، دقیقاً برابر با خطای معیار حاصل از فرمول نشان داده شده در تابلوی ۱.۷ است. جای تعجب نیست، زیرا SUDAAN دقیقاً از همان فرمول استفاده می‌کند.

Variable		One 1
PHARMEXP/TOTMEDEX	Sample Size	7.0000
	Weighted Size	8.0000
	Weighted X-Sum	3222978.0000
	Weighted Y-Sum	1028610.0000
	Ratio Est.	0.3191
	SE Ratio	0.0040

برآورد کردن با استفاده از STATA می‌تواند روی پروندهٔ اطلاعاتی STATA به نام *tab7pt1.dta* با فرمانهای زیر انجام شود:

```
. use a:tab7pt1
. svyset fpc totcnt
. svyset pweight wt1
. svyratio pharmexp/totmedex
```

پروندهٔ *tab7pt1.dta* دارای همان ساختاری است که پروندهٔ *tab7pt1.ssd* در تحلیـــل SUDAAN مورد استفاده قرار داد.

این فرمانها خروجی زیر را تولید می‌کنند.

Survey ratio estimation

Pweight :	wt1	Number of obs =	7
Strata :	<one>	Number of strata =	1
PSU :	<observations>	Number of PSUs =	7
FPC :	totcnt	Population size =	8

Ratio	Estimate	Std.Err.	[95% Conf.	Interval]	Deff
pharmexp/totmedex	.3191489	.0040067	.309345	.3289529	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

۲.۷ برآورد کردن نسبتها در زیرحوزه‌ها، تحت نمونه‌گیری تصادفی ساده

در بخش ۳.٦ به طور نسبتاً غیررسمی از برآورد کردن میانگینها برای زیرحوزه‌ها تحت نمونه‌گیری تصادفی ساده بحث کردیم. متوجه شدیم که میانگین یک زیرحوزه، در واقع یک برآورد نسبتی است که صورت کسر آن برابر است با مجموع نمونه‌ای متغیر مورد برآورد در همهٔ عناصر نمونهٔ متعلق به زیرحوزه و مخرج کسر آن برابر است با تعداد عناصر نمونهٔ متعلق به زیرحوزه. در این بخش، برآورد نسبتها برای زیرحوزه‌ها را به وضعیتهایی تعمیم می‌دهیم که در آنها مخرج کسر نسبت الزاماً تعداد عناصر زیرحوزه در نمونه نیست.

فرض کنید یک نمونهٔ تصادفی سادهٔ n عنصری از جامعه‌ای N عنصری می‌گیریم و می‌خواهیم نسبت $R_{(k)}$ را برآورد کنیم که به صورت زیر تعریف می‌شود:

$$R_{(k)} = \frac{X_{(k)}}{Y_{(k)}}$$

که در آن $X_{(k)}$ و $Y_{(k)}$، به ترتیب مجموع متغیرهای X و Y روی همهٔ عناصر موجود در یک زیرحوزهٔ تعریف شده k هستند. با استفاده از تبدیلی از متغیرهای صورت و مخرج کسر، می‌توانیم از شیوه‌های برآورد که در بخش ۱.۷ شرح دادیم و از فرمولهایی که در تابلوی ۱.۷ نشان دادیم استفاده کنیم. این شیوه در زیر نشان داده می‌شود.

قرار دهید

$$\delta_{ik} = \begin{cases} ۱ & \text{اگر عنصر } i \text{ در زیرحوزهٔ } k \text{ باشد،} \\ ۰ & \text{در غیر صورت بالا،} \end{cases}$$

و

$$r_{(k)} = \frac{x_{(k)}}{y_{(k)}}$$

$r_{(k)}$، برآورد خطای معیار نسبت زیرحوزه را، می‌توان با فرمولی که در تابلوی ۱.۷ نشان دادیم برآورد کرد. این مطلب با مثال زیر نشان داده می‌شود.

مثال تشریحی: فرض کنید در ناحیه‌ای که دارای ۱۰۱ بیمارستان است، نمونه‌ای متشکل از ٥٦ بیمارستان می‌گیریم و برای هر بیمارستان نمونه، پروندهٔ همهٔ نوزادانی را که در سال تقویمی گذشته در آن بیمارستان متولد شده‌اند مرور می‌کنیم تا تعیین کنیم که آیا وضعیت مادر از لحاظ پادگن سطحی هپاتیت **B** (HBsAG) در پروندهٔ پزشکی بیماری نوزاد ثبت شده است یا نه. وجود این اطلاع در پروندهٔ نوزاد از این نظر مهم است که اگر پادگن مادر مثبت باشد باید نوزاد گلوبولین مصون‌سازی دریافت کند تا از بروز هپاتیت **B** پیشگیری شود.

بیمارستانها براساس سطح خدمات بیماریهای زنان و زایمان طبقه‌بندی شده‌اند (۱= خدمات «اولیه»، ۲= «میانی»، ۳= «درجهٔ ۳») و هدف، برآورد کردن نسبت نوزادانی است که وضعیت پادگـــن سطحی هپاتیت **B** مادران آنها در پروندهٔ پزشکی نوزاد ثبت شده است. داده‌های مربوط به این مثال تشریحی در پروندهٔ *BHRATIO.DAT* هستند. برای اهداف مثال، متغیرهای مربوط به ۱۰ سابقهٔ اول آن پرونده در جدول زیر نشان داده شده‌اند.

بیمارستان نمونه	تعداد پرونده‌هایی که پادگن مادر در آنها گزارش شده است (x_i)	تعداد متولد شده‌ها (y_i)	سطح زنان و زایمان بیمارستان	نشانگرمتغیر برای زیرحوزه δ_i	متغیر صورت کسر تبدیل یافته (x_i)	متغیر مخرج کسر تبدیل یافته (y_i)
۱	۲۸۹۸	۲۸۹۸	۲	۱	۲۸۹۸	۲۸۹۸
۲	۱۰۹۵	۱۳۰٤	۲	۱	۱۰۹۵	۱۳۰٤
۳	۱۸٦۰	۲۰۲۲	۲	۱	۱۸٦۰	۲۰۲۲
٤	۱۲۲۷	۱۳۹۵	۲	۱	۱۲۲۷	۱۳۹۵
۵	٦۱۸	۷۷۳	۲	۱	٦۱۸	۷۷۳
٦	۱٤۹۹	۱٦۳۰	۲	۱	۱٤۹۹	۱٦۳۰
۷	۱۳۲۱	۱٤۳٦	۲	۱	۱۳۲۱	۱٤۳٦
۸	۰	۲۵۲۵	۲	۱	۰	۲۵۲۵
۹	۳۵۹۳	٤٦۷٤	۳	۰	۰	۰
۱۰	۱۷۳۲	۲۲۷۹	۲	۱	۱۷۳۲	۲۲۷۹

در بین ۱۰ سابقهٔ اول نمونه که در جدول نشان داده شده‌اند فقط یک سابقه (مربوط به بیمارستان نمونهٔ ۹) تحت تأثیر تبدیل قرار گرفته است، زیرا این سابقه در بین ۱۰ بیمارستان اول از تنها بیمارستانی گرفته شده است که در زیرحوزه موردنظر نیست. در کل نمونهٔ متشکل از ٥٦ بیمارستان، ۱۳ بیمارستان از نظر خدمات زنان و زایمان در سطح «۲» قرار ندارند و تحت تأثیر تبدیل قرار می‌گیرند، ٤۳ بیمارستان دیگر در سطح «۲» قرار دارند. آماره‌های خلاصه برای نمونهٔ متشکل از ٥٦ بیمارستان به شرح زیرند:

$$N = ۱۰۱ \qquad n = ۵۶$$

$$x_{(k)} = \sum_{i=۱}^{۵۶} x_i \delta_{ik} = ۴۳۹۱۰ \qquad y_{(k)} = \sum_{i=۱}^{۵۶} y_i \delta_{ik} = ۵۸۰۸۲$$

$$r_{(k)} = \frac{x_{(k)}}{y_{(k)}} = ۰/۷۵۶۰$$

$$s^{۲}_{x_{(k)}} = ۵۴۷۳۹۳/۰۷۹ \qquad s^{۲}_{y_{(k)}} = ۶۳۹۴۸۴/۹۱۳$$

$$\bar{x}_{(k)} = \frac{x_{(k)}}{۵۶} = ۷۸۴/۱۰۷ \qquad \bar{y}_{(k)} = \frac{y_{(k)}}{۵۶} = ۱۰۳۷/۱۷۹$$

$$\hat{V}_{x_{(k)}} = \sqrt{\frac{۱۰۰}{۱۰۱}} \frac{S_{x_{(k)}}}{\bar{x}_{(k)}} = ۰/۹۳۸۹$$

$$\hat{V}_{y_{(k)}} = ۰/۷۶۷۲ \hat{\rho}_{x_{(k)}, y_{(k)}} = ۰/۸۲۵$$

در این صورت، از فرمول تابلوی ۱.۷ خواهیم داشت

$$\hat{SE}(r_{(k)}) = ۰/۰۳۶۰$$

به این ترتیب برآورد می‌شود که ۷۵/۶٪ همهٔ نوزادان در بیمارستانهای سطح «۲» از نظر خدمات زنان و زایمان، وضعیت پادگن سطحی هپاتیت B مادرشان در پروندهٔ پزشکی آنها ثبت شده و خطای معیار این برآورد ۰/۰۳۶۰ است.

□

۳.۷ برآوردهای نسبتی پس طبقه‌بندی شده تحت نمونه‌گیری تصادفی ساده

فن پس *طبقه‌بندی* که عموماً در روش‌شناسی نمونه‌گیری به کار می‌رود یکی از فنون برآورد کردن است که از اطلاعات معلوم جامعه در مورد زیرگروهها یا زیرحوزه‌ها استفاده می‌کند تا برآوردهایی تولید کند که از بهبود دقت برخوردار باشند. در فصل ٦، بخش ٦.٦، برآوردهای پس طبقه‌بندی شدهٔ میانگینهای جامعه را تحت نمونه‌گیری تصادفی ساده شرح دادیم. مطالب مورد بحث در آن قسمت برای سایر برآوردهای خطی از قبیل مجموعها و نسبتها نیز به همان اندازه قابل اجراست. در این بخش، استفاده از پس طبقه‌بندی را در برآورد کردن نسبتی تحت نمونه‌گیری تصادفی ساده مورد بحث قرار خواهیم داد.

فرض کنیم یک نمونهٔ تصادفی سادهٔ n تایی از جامعه‌ای متشکل از N عنصر داریم و می‌خواهیم نسبت $R = \frac{X}{Y}$ را برآورد کنیم.

برآورد نسبتی، r را که در قسمت ۱.۷ تصریح شد می‌توان به صورت زیر نوشت.

$$r = \frac{x'}{y'} = \frac{\sum_{i=۱}^{n} w_i x_i}{\sum_{i=۱}^{n} w_i y_i}$$

که درآن

$$w_i = \frac{N}{n}$$

اگر k زیرگروه یا زیرحوزهٔ دو به دو ناسازگار و فراگیر وجود داشته باشند آنگاه فرمول زیر برای x' درست است:

$$x' = \sum_{k=1}^{K} \sum_{i=1}^{n} w_i \delta_{ik} x_i = \frac{N}{n} \sum_{k=1}^{K} x_{(k)}$$

که در آن

$$x_{(k)} = \sum_{i=1}^{n} \delta_{ik} x_i$$

و

$$\delta_{ik} = \begin{cases} 1 & \text{اگر عنصر } i \text{ در زیرحوزهٔ } k \text{ باشد} \\ 0 & \text{در غیر صورت بالا} \end{cases}$$

نهاد $x_{(k)}$ به سادگی مجموع نمونه‌ای متغیر X برای همهٔ عناصر نمونه است که در زیرحوزهٔ k قرار دارند. به هیمن ترتیب، نهاد

$$n_{(k)} = \sum_{i=1}^{n} \delta_{ik}$$

تعداد عناصر نمونه است که در زیرحوزهٔ k قرار دارند. نسبت $\frac{x_{(k)}}{n_{(k)}}$ سطح X را به ازای هر عنصر در حوزهٔ k از روی نمونه برآورد می‌کند و اگر کل تعداد عناصر، $N_{(k)}$، در حوزهٔ k ی جامعه معلوم باشد، در آن صورت نسبت $(N_{(k)}/n_{(k)})x_{(k)}$ برآورد «مشخص»تری از کل X در حوزهٔ k خواهد بود. بالاخره، با استفاده از این «تصحیح» برای هر حوزه و نیز برای متغیر Y، برآورد نسبتی پس طبقه‌بندی شدهٔ r_{pstr} را به دست می‌آوریم که در زیر نشان داده شده است:

$$r_{pstr} = \frac{\sum_{k=1}^{K} \frac{N_k}{n_{(k)}} x_{(k)}}{\sum_{k=1}^{K} \frac{N_k}{n_{(k)}} y_{(k)}}$$

که در آن

$$x_{(k)} = \sum_{i=1}^{n} \delta_{ik} x_i$$

$$y_{(k)} = \sum_{i=1}^{n} \delta_{ik} y_i$$

و

$$n_{(k)} = \sum_{i=1}^{n} \delta_{ik}$$

چون نهادهای $x_{(k)}$، $y_{(k)}$ و $n_{(k)}$ همه در معرض تغییرپذیری نمونه‌گیری قرار دارند نتیجه می‌گیریم که برآورد نسبتی پس طبقه‌بندی شدهٔ r_{pstr} در واقع نسبت مجموع برآوردهای نسبتی K شامل $x_{(k)}$ و $n_{(k)}$ تقسیم بر مجموع برآوردهای نسبتی K شامل $y_{(k)}$ و $n_{(k)}$ است. خطای معیار r_{pstr} صورت ساده‌ای ندارد و در اینجا ارائه نخواهد شد ولی می‌توان آن را با استفاده از خطی‌سازی سری تیلور (فصل ۱۲ را ببینید) برآورد کرد که در نرم‌افزار SUDAAN موجود است.

مثال تشریحی: نمونهٔ متشکل از ٥٦ بیمارستان را در نظر می‌گیریم که در بخش ۲.۷ مورد بحث قرار گرفت (داده‌های حاصل از این نمونه در پروندهٔ ASCII، BHRATIO.DAT قرار دارند). استفادهٔ سرراست از فرمولهای برآورد کردن نسبتی تابلوی ۱.۷، برآوردی معادل ۰/۷٥٦ با برآورد انحراف معیار ۰/۰۳٦ را برای نسبت نوزادانی که وضعیت پادگن سطحی هپاتیت **B** مادرانشان در پروندهٔ پزشکی آنها ثبت شده است نتیجه می‌دهد.

به علاوه، می‌دانیم که ۲۹ (٥۱/۸٪) بیمارستان از ٥٦ بیمارستان نمونه دارای سیاستهای مکتوبی هستند مبنی بر اینکه تا زمانی که وضعیت مادر از لحاظ پادگن سطحی هپاتیت **B** در پروندهٔ نوزاد درج نشود نوزاد نباید از بیمارستان مرخص شود. همچنین می‌دانیم که ٦٦ بیمارستان (٦٥/٤٪) از ۱۰۱ بیمارستان موجود در جامعهٔ متناظر که این نمونه از آن گرفته شده است این سیاست مدون را دارند. چون منطقی است که بیمارستانهای پیرو این سیاست دارای نسبت بالاتری از نوزادانی باشند که وضعیت پادگن سطحی هپاتیت **B** مادرانشان در پرونده‌های آنان درج می‌شود، و چون توزیع جامعهٔ بیمارستانهایی که این سیاست را دارند و ندارند معلوم است، شرایطی فراهم شده است که می‌توان به صورتی معقول از برآورد نسبتی پس طبقه‌بندی شده استفاده کرد. زیرحوزهٔ ۱ را زیرحوزهٔ دارای سیاست مدون در نظر می‌گیریم که مستلزم داشتن اطلاعاتی دربارهٔ وضعیت پادگن سطحی هپاتیت **B** مادر در پروندهٔ پزشکی نوزاد قبل از ترخیص از بیمارستان است و زیرحوزهٔ ۲ را زیرحوزهٔ بیمارستانهایی می‌گیریم که چنین سیاستی را دنبال نمی‌کنند. به این ترتیب آماره‌های خلاصهٔ زیر را به دست می‌آوریم.

$n_{(۱)} = ۲۹$ تعداد بیمارستانهای نمونه متعلق به زیرحوزهٔ ۱

$n_{(۲)} = ۲۷$ تعداد بیمارستانهای نمونه متعلق به زیرحوزهٔ ۲

$x_{(۱)} = ۴۲۷۴۹$ کل تعداد نوزادان در زیرحوزهٔ ۱ بیمارستانهای نمونه که وضعیت پادگن سطحی هپاتیت B مادرانشان در پروندهٔ پزشکی ثبت شده است

$x_{(۲)} = ۲۲٥۶۰$ کل تعداد نوزادان در زیرحوزهٔ ۲ بیمارستانهای نمونه که وضعیت پادگن سطحی هپاتیت B مادرانشان در پروندهٔ پزشکی ثبت نشده است

$y_{(1)} = ۴۷۸۵۰$ کل تعداد نوزادان در نمونهٔ زیرحوزهٔ ۱ بیمارستانها

$y_{(2)} = ۳۸۹۲۹$ کل تعداد نوزادان در نمونهٔ زیرحوزهٔ ۲ بیمارستانها

$N_1 = ۶۶$ تعداد جامعهٔ معلوم بیمارستانهای زیرحوزهٔ ۱

$N_2 = ۳۵$ تعداد جامعهٔ معلوم بیمارستانهای زیرحوزهٔ ۲

از روی این آماره‌های خلاصه، برآورد نسبتی پس طبقه‌بندی شده را به شرح زیر می‌سازیم:

$$r_{pstr} = \frac{\frac{۶۶}{۲۹}۴۲۷۴۹ + \frac{۳۵}{۲۷}۲۲۵۶۰}{\frac{۶۶}{۲۹}۴۷۸۵۰ + \frac{۳۵}{۲۷}۳۸۹۲۹} = ۰/۷۹۴$$

در زیر برآورد نسبتی پس طبقه‌بندی شده با برآورد نسبتی معمولی در مورد این مثال مقایسه شده است:

برآورد	برآورد نسبتی	خطای معیار برآورد شده
معمولی	$r = ۰/۷۵۶$	$SE(r) = ۰/۰۳۶$
پس طبقه‌بندی شده	$r_{pstr} = ۰/۷۹۴$	$SE(r_{pstr}) = ۰/۰۲۲$

در این مثال، برآورد نسبتی پس طبقه‌بندی شده آشکارا با برآورد نسبتی معمولی تفاوت دارد و خطای معیار آن به مراتب کمتر است. دلیل این امر آن است که نسبتهای زیرحوزه‌ها تفاوت قابل ملاحظه‌ای دارند و نسبت بیمارستانهای زیرحوزهٔ ۲ در نمونه بیشتر از آن است که در جامعه وجود دارد.

□

٤.٧ برآورد نسبتی مجموعها تحت نمونه‌گیری تصادفی ساده

برآوردهای مجموعها یا انبوهه‌ها که تا اینجا مورد بحث قرار گرفتند از ضرب سادهٔ مجموع نمونه در نسبت $\frac{N}{n}$ به دست می‌آمد. دربارهٔ این واقعیت بحث کردیم که برآوردهای x' از این نوع، برآوردهایی نااریب‌اند و فرمولی برای خطای معیار این قبیل برآوردها و روشی برای ساختن بازه‌های اطمینان تقریبی مجموع نامعلوم X تحت نمونه‌گیری تصادفی ساده شرح دادیم. ولی گاهی اوقات در موقعیتی قرار می‌گیریم که می‌توانیم از اطلاعات اضافی موجود برای مقاصد ساختن برآوردگر دیگری از مجموع کل جامعه استفاده کنیم که بر پایهٔ برآورد نسبتی متکی است، به نحوی که میانگین توان دوم خطای برآوردگر به دست آمده از برآوردگر تورمی ساده x' کمتر است.

این ایده را با نگاهی به یک مثال بررسی می‌کنیم.

مثال تشریحی: فرض کنیم روستایی دارای شش ناحیهٔ سرشماری است که جمعیت سال ۱۹۹۰ آن در جدول ۴.۷ ارائه شده است. میزان ثبت‌نام مدارس این روستا (که پیش از اجرای آمارگیری نامعلوم تلقی می‌شود) نیز در جدول ۴.۷ ارائه شده است.

فرض کنیم که می‌خواهیم کل ثبت‌نام شدگان فعلی مدارس را با انتخاب یک نمونهٔ تصادفی ساده از دو ناحیهٔ سرشماری و معلوم کردن ثبت‌نام شدگان در هر ناحیهٔ نمونه با آمارگیری از هر یک از مدارس موجود در ناحیه، شمارش افراد، و ضرب مجموعهای نمونه‌ای در $\frac{N}{n}$ به طوری که در فصـــل ۳ شرح داده شده است برآورد کنیم. ۱۵=$\binom{۶}{۲}$ نمونهٔ ممکن همراه با مقادیر x'، برآورد مجموع، در جدول ۵.۷ فهرست شده‌اند.

جدول ۴.۷ جمعیت (سرشماری ۱۹۹۰) و ثبت‌نام شدگان فعلی مدارس برحسب ناحیهٔ سرشماری

ناحیه سرشماری	جمعیت سال ۱۹۹۰	ثبت‌نام شدگان فعلی*
۱	۶۶۵۷	۲۲۶۹
۲	۴۰۵۷	۱۳۲۴
۳	۳۶۴۲	۹۵۲
۴	۵۳۲۰	۱۵۵۸
۵	۴۴۸۰	۱۳۵۲
۶	۵۸۸۰	۱۷۹۶
مجموع	۳۰۰۳۶	۹۲۵۱

* تا زمان آمارگیری از مدارس نامعلوم است.

جدول ۵.۷ نمونه‌های ممکن متشکل از دو مدرسه از روی داده‌های جدول ۴.۷

شمارهٔ مدرسه‌ها در نمونه	برآورد ثبت‌نام شدگان در مدارس	شمارهٔ مدرسه‌ها در نمونه	برآورد ثبت‌نام شدگان در مدارس
۱ , ۲	۱۰۷۷۹	۲ , ۶	۹۳۶۰
۱ , ۳	۹۶۶۳	۳ , ۴	۷۵۳۰
۱ , ۴	۱۱۴۸۱	۳ , ۵	۶۹۱۲
۱ , ۵	۱۰۸۶۳	۳ , ۶	۸۲۴۴
۱ , ۶	۱۲۱۹۵	۴ , ۵	۸۷۳۰
۲ , ۳	۶۸۲۸	۴ , ۶	۱۰۰۶۲
۲ , ۴	۸۶۴۶	۵ , ۶	۹۴۴۴
۲ , ۵	۸۰۲۸		

از نتایجی که دربارهٔ نمونه‌گیری تصادفی ساده به دست آوردیم می‌دانیم که برآورد مجموع x' یک برآورد نااریب از کل جمعیت است. خطای معیار x' (که با شمارش همهٔ نمونه‌های فهرست شدهٔ زیر محاسبه می‌شود) بنابر تابلوی ۳.۲ عبارت است از

$$SE(\mathrm{x}')=۱۵۶۸/۴۵۷۷$$

به جای استفاده از برآورد x' که در مثال بالا انجام دادیم می‌توانیم کل ثبت‌نام شدگان X را با استفاده از داده‌های مربوط به جمعیت سال ۱۹۹۰ که برای هر ناحیهٔ سرشماری موجود است برآورد کنیم. هر نمونه‌ای از نواحی سرشماری به ما امکان می‌دهد که نه تنها ثبت‌نام مدارس را از روی نمونه برآورد کنیم بلکه کل جمعیت را نیز از نمونه برآورد کنیم. فرض کنید از نمادهای زیر استفاده شود:

Y تعداد کل افراد در جامعه

Y_i تعداد افراد در ناحیهٔ سرشماری i

y_i تعداد افراد در ناحیهٔ نمونهٔ i

$y=\sum_{i=۱}^{n} y_i$ کل تعداد افراد در نواحی سرشماری نمونه

پس برآورد کل جمعیت، y'، از فرمول زیر به دست می‌آید

$$y'=\left(\frac{N}{n}\right)\times y$$

که در آن،

N = کل تعداد نواحی در جامعه

n = کل تعداد نواحی در نمونه

هر نمونه، برآورد x' را از X و y' را از Y به دست می‌دهد. ولی چون مقدار واقعی Y از روی داده‌های سرشماری برای ما معلوم است، پس معقول خواهد بود که در صورت وجود همبستگی شدید بین میزان ثبت‌نام مدارس و اندازهٔ جمعیت، فرض را بر این بگیریم که برآوردگر x' از X با خود X به همان نسبتی تفاوت خواهد داشت که برآوردگر y' با Y تفاوت دارد. یعنی $\frac{X}{x'}=\frac{Y}{y'}$. این، انگیزهٔ استفاده از برآوردگر x'' برای X می‌شود که از فرمول

$$x''=\left(\frac{x'}{y'}\right)\times Y \qquad (٤.٧)$$

یا از معادل جبری آن

$$x'' = r \times Y \qquad (۵.۷)$$

به دست می‌آید، که در آن r، یک برآورد نسبتی است.

اگر این برآوردگر به صورت زیر نوشته شود شناخت بیشتری از آن آشکار خواهد شد:

$$x'' = \left(\frac{Y}{y'}\right) \times x'$$

اگر برآورد اندازهٔ جمعیت، y'، کمتر از مجموع واقعی معلوم Y باشد، آنگاه y' مقدار Y را کم برآورد می‌کند. در این صورت انتظار خواهیم داشت که برآورد x' که از همین نمونه به دست آمده است نیز تا حد مشابهی مجموع واقعی نامعلوم X را کم برآورد کند، زیرا x و y همبستگی دارند. ولی توجه کنید که نسبت $\frac{Y}{y'}$ در این مورد بیشتر از یک است و x'' از برآورد تورمی سادهٔ x' بزرگتر خواهد بود. هرگاه y' بزرگتر از Y باشد جریان معکوس خواهد شد. به این ترتیب، می‌بینیم که حتی در مورد نمونه‌هایی که به طور خاص بدند (یعنی نمونه‌هایی که مقادیری از x' را نتیجه می‌دهند که از لحاظ قدر مطلق تفاوت فاحشی با X دارند) نسبت $\frac{Y}{y'}$ آن را اصلاح می‌کند و برآوردگر x'' به اندازهٔ x' با X تفاوت نخواهد داشت.

اینک همهٔ این ایده‌ها را با چند مثال بررسی می‌کنیم.

□

مثال تشریحی: برای نمونه‌های $n=۲$ ناحیهٔ سرشماری که از جامعه‌ای که در جدول ۴.۷ نشان داده شده گرفته شده است مقادیر نمونه‌ای x' و x'' را بررسی می‌کنیم. این مقادیر در جدول ۶.۷ فهرست شده‌اند.

از توزیع نمونه‌گیریی که در جدول ۶.۷ نشان داده‌ایم می‌بینیم که میانگین $E(x'')$، خطای معیار $SE(x'')$ و میانگین توان دوم خطا $MSE(x'')$ عبارت‌اند از (رجوع کنید به تابلوی ۳.۲)

$$E(x'') = \frac{۱۰۰۷۳+۹۳۹۴+\cdots+۹۱۲۷}{۱۵} = ۹۲۱۱/۰۷$$

$$SE(x'') = \left[\frac{(۱۰۰۷۳-۹۲۱۱/۰۷)^۲+(۹۳۹۴-۹۲۱۱/۰۷)^۲+\cdots+(۹۱۲۷-۹۲۱۱/۰۷)^۲}{۱۵}\right]^{\frac{۱}{۲}} = ۴۶۶/۳۸$$

$$MSE(x'') = Var(x'') + B^۲(x'') = ۲۱۷۵۱۲/۳۳ + (۹۲۱۱/۰۷-۹۲۵۱)^۲ = ۲۱۹۱۰۶/۷۳$$

جدول ٦.٧ مقادیر x' و x'' برای نمونه‌های جدول ۵.۷

نمونه	x'	x''
۱ , ۲	۱۰۷۷۹	۱۰۰۷۳
۱ , ۳	۹٦٦۳	۹۳۹٤
۱ , ٤	۱۱٤۸۱	۹۵۹۷
۱ , ۵	۱۰۸٦۳	۹۷٦٦
۱ , ٦	۱۲۱۹۵	۹۷۳۹
۲ , ۳	٦۸۲۸	۸۸۷۹
۲ , ٤	۸٦٤٦	۹۲۳۱
۲ , ۵	۸۰۲۸	۹٤۱۵
۲ , ٦	۹۳٦۰	۹٤۳۱
۳ , ٤	۷۵۳۰	۸٤۱۲
۳ , ۵	٦۹۱۲	۸۵۲۰
۳ , ٦	۸۲٤٤	۸٦٦۸
٤ , ۵	۸۷۳۰	۸۹۱۹
٤ , ٦	۱۰۰٦۲	۸۹۹۵
۵ , ٦	۹٤٤٤	۹۱۲۷

چون برآورد تورمی سادهٔ x' برآوردی ناریب از X است میانگین توان دوم خطای MSEی آن برابر با واریانس آن است:

$$MSE(x')=Var(x') = (۱۵۶۸/۴۵۷۷)^۲ =۲۴۶۰۰۵۹/۵۶$$

به این ترتیب، با این که x'' برآوردی اریب برای X است میانگین توان دوم خطای آن در این مثال خیلی کمتر از x' است.

□

مثال تشریحی: از داده‌های جدول ۱.۵ نیز می‌توان برای تشریح برآورد کردن یک مجموع به وسیلهٔ یک نسبت استفاده کرد. اگر بخواهیم کل تعداد X کامیون ـ مایلهای طی شده را از یک نمونهٔ تصادفی ساده از $n=۳$ قسمت برآورد کنیم، می‌توانیم نسبت r کامیون ـ مایلها را به تصادفاتی که در آنها کامیون مداخله داشته است از روی نمونه برآورد کنیم. اگر کل تعداد Y تصادفی که کامیون در آن مداخله داشته است برای کل جاده معلـــوم باشد می‌تـوان برآورد x'' را به دست آورد. برای مثال اگر قسمتهای ۱، ۳ و ٤ به عنوان نمونه انتخاب شوند محاسبـــات زیر را خواهیم داشت (رجوع کنید به تابلوی ۱.۳ و تابلوی ۱.۷):

برآورد تورمی سادهٔ کامیون ـ مایل طی شده $x'=\left(\frac{۸}{۳}\right)\times(۶۳۲۷+۸۶۹۱+۷۸۳۴)=۶۰۹۳۸/۶۷=$

برآورد تورمی سادهٔ تعداد تصادفات با مداخلهٔ کامیون $y'=\left(\frac{۸}{۳}\right)\times(۸+۹+۹)=۶۹/۳۳=$

برآورد نسبت کامیون ـ مایل طی شده به تصادفات با مداخلهٔ کامیون $r=\frac{x'}{y'}=۸۷۸/۹۲۳۰=$

کل تعداد تصادفات با مداخلهٔ کامیون، Y، معلوم است، دقیقاً $Y = ۵۰$

پس

$$x''=\left(\frac{x'}{y'}\right)\times Y=۸۷۸/۹۲۳۰\times ۵۰=۴۳۹۴۶/۱۵$$ کامیون ـ مایل

این رقم به مجموع واقعی $(X=۳۴۰۵۴)$ نزدیکتر است تا برآورد تورمی سادهٔ x'.

□

برآورد x'' به صورت rY است که در آن r، نسبت برآورد شده و Y، مجموع معلوم جامعه است. از این رو نتیجه می‌گیریم که *خطای معیار* $SE(x'')$ برابر است با $Y\times SE(r)$ که می‌توان آن را از فرمول زیر تقریب کرد، به شرطی که $V(\bar{y})$ ضریب تغییرات $\bar{y}$ کمتر از ۰/۰۵ باشد:

$$SE(x'')=\left(\frac{YR}{\sqrt{n}}\right)\times\left(V_x^{۲}+V_y^{۲}-۲\rho_{xy}V_xV_y\right)^{۱/۲}\times\sqrt{\frac{N-n}{N-۱}} \qquad (۶.۷)$$

به همین ترتیب، برآورد $\hat{SE}(x'')$ از $SE(x'')$ را می‌توان از روی داده‌ها باجایگزین کـــردن $\hat{SE}(r)$ [رابطهٔ (۳.۷)] به جای $SE(r)$ در فرمول (۶.۷) به شرح زیر به دست آورد:

$$\hat{SE}(x'')=\left(\frac{Yr}{\sqrt{n}}\right)\times\left(\hat{V}_x^{۲}+\hat{V}_y^{۲}-۲\hat{\rho}_{xy}\hat{V}_x\hat{V}_y\right)^{۱/۲}\times\sqrt{\frac{N-n}{N-۱}} \qquad (۷.۷)$$

از فرمول (۷.۷) می‌توان در به دست آوردن بازه‌های اطمینان تقریبی برای یک مجموع نامعلوم X استفاده کرد، به شرطی که $\hat{V}_{(\bar{y})}$ ضریب برآورد شدهٔ تغییرات $\bar{y}$ کمتر از ۰/۰۵ باشد.

به مثالی دیگر نگاه کنیم.

مثال تشریحی: اگر قسمتهای ۱، ۳ و ۴ از داده‌های جدول ۱.۵ نمونه‌گیری شده باشند داده‌های نمونهٔ جدول ۷.۷ را در اختیار داریم. پس آماره‌های خلاصهٔ زیر را خواهیم داشت:

$n=۳$	$\bar{x}=۷۶۱۷/۳۳$	$s_x=۱۱۹۶/۸۰$	$\hat{V}_x^{۲}=۰/۰۲۱۶$
$N=۸$	$\bar{y}=۸/۶۷$	$s_y=۰/۵۷۷۴$	$\hat{V}_y^{۲}=۰/۰۰۳۸۸۳$
$\hat{\rho}_{xy}=۰/۹۳۳۷$	$Y=۵۰$	$r=۸۷۸/۹۲۳۱$	$\bar{x}=۴۳۹۴۶/۱۵$

برآورد ضریب تغییرات، $\hat{V}(\bar{y})$، عبارت است از[1]

$$\hat{V}(\bar{y}) = \left(\frac{s_y}{\bar{y}\sqrt{n}}\right)\times\sqrt{\frac{N-n}{N}} = \left[\frac{۰/۵۷۷۴}{۸/۶۷\sqrt{۳}}\right]\times\sqrt{\frac{۸-۳}{۸}} = ۰/۰۳۰۴$$

چون $\hat{V}(\bar{y}) < ۰/۰۵$ می‌توان از عبارت (۷.۷) برای برآورد خطای معیار x'' به صورت زیر استفاده کرد:

$$\hat{SE}(x'') = \left(\frac{۵۰\times۸۷۸/۵۸۴۸}{\sqrt{۳}}\right)\times\left[۰/۰۲۴۵ + ۰/۰۰۳۸۸۳ - ۲\times۰/۹۳۳۷ \times\sqrt{۰/۰۲۱۶}\times\sqrt{۰/۰۰۳۸۸۳}\right]^{۱/۲}\times\sqrt{\frac{۸-۳}{۸-۱}} = ۱۹۶۳/۰۴$$

جدول ۷.۷ داده‌های نمونه از جدول ۱.۵

قسمتهای نمونه	تعداد کامیون ـ مایل (x_i)	تعداد تصادفات با مداخلهٔ کامیون (y_i)
۱	٦۳۲۷	۸
۳	۸٦۹۱	۹
٤	۷۸۳٤	۹

در این صورت بازهٔ اطمینان ۹۵٪ را به شرح زیر برای کل تعداد برآورد شدهٔ کامیون ـ مایل طی شده خواهیم داشت:

$$x'' - ۱/۹۶\times\hat{SE}(x'') \le X \le x'' + ۱/۹۶\times\hat{SE}(x'')$$
$$۴۳۹۴۶/۱۵ - ۱/۹۶\times۱۹۶۳/۰۴ \le X \le ۴۳۹۴۶/۱۵ + ۱/۹۶\times۱۹۶۳/۰۴$$
$$۴۰۰۹۸ \le X \le ۴۷۷۹۳/۷۱$$

توجه کنید که در مورد این نمونهٔ بخصوص، بازهٔ اطمینان ۹۵٪، مجموع واقعی جامعه را نمی‌پوشاند. □

[1] فرمول مربوط به $V(\bar{y})$ در معادلهٔ (۷.۳) ارائه شده است. با جایگزین کردن $\hat{V}_y^{۲}$ در فرمول (۷.۳) به صورتی که در تابلوی ۱.۷ ارائه شده نتیجهٔ زیر را به دست می‌آوریم

$$\hat{V}(\bar{y}) = \left(\frac{\hat{V}_y}{\sqrt{n}}\right)\times\sqrt{\frac{N-n}{N-۱}} = \left[\sqrt{\frac{N-۱}{N}}\times\left(\frac{s_y^{۲}}{\bar{y}^{۲}}\right)^{۱/۲}\times\frac{۱}{\sqrt{n}}\right]\times\sqrt{\frac{N-n}{N-۱}} = \left(\frac{s_y}{\bar{y}\sqrt{n}}\right)\times\sqrt{\frac{N-n}{N}}$$

۵.۷ مقایسهٔ برآورد نسبتی با برآورد تورمی ساده

فـــرض کنیـم تقریب خطای معیـار برآورد نسبتی مجموع، تحت نمونه‌گیری تصادفی ساده، که از رابطهٔ (٦.٧) به دست می‌آید معتبر است و اریبی برآورد نسبتی را می‌توان نادیده گرفت. در این صورت می‌توانیم ارزیابی کنیم که آیا برآورد نسبتی x'' احتمال دارد به برآوردی از مجموع منجر شود که نسبت به برآورد تورمی ساده بهبودی بیشتری داشته باشد یا نه. این ارزیابی با بررسی نسبت واریانس x'' به واریانس x' به شرح زیر میسر است:

$$\frac{Var(x'')}{Var(x')} = \frac{V_x^2 + V_y^2 - 2\rho_{xy}V_xV_y}{V_x^2}$$

چون

$$\begin{aligned} Var(x'') &= Y^2 \times Var(r) \\ &= Y^2 \times \left(\frac{R^2}{n}\right) \times (V_x^2 + V_y^2 - 2\rho_{xy}V_xV_y) \times \left(\frac{N-n}{N-1}\right) \end{aligned}$$

و

$$\begin{aligned} Var(x') &= \left(\frac{N^2}{n}\right) \times \sigma_x^2 \times \left(\frac{N-n}{N-1}\right) \\ &= \left(\frac{N^2}{n}\right) \times \left(\bar{X}^2 V_x^2\right) \times \left(\frac{N-n}{N-1}\right) = \left(\frac{X^2}{n}\right) \times V_x^2 \times \left(\frac{N-n}{N-1}\right) \end{aligned}$$

نتیجه می‌گیریم که واریانس x'' وقتی کمتر از واریانس x' خواهد بود که

$$V_x^2 + V_y^2 - 2\rho_{xy}V_xV_y < V_x^2$$

یا به طور هم‌ارز هرگاه

$$\frac{V_y}{V_x} < 2\rho_{xy} \qquad (۸.۷)$$

به این ترتیب برای مقادیر ثابت ضرایب تغییرات x و y در جامعه، هر چه همبستگی بین x و y بیشتر باشد احتمال این که برآورد مجموع، x''، واریانسی کمتر نسبت به برآورد مجموع، x'، یعنی برآورد تورمی ساده، داشته باشد بیشتر است.

٦.٧ تقریب خطای معیار برآورد نسبتی مجموع

هرگاه رابطهٔ بین x و y را بتوان به صورتی نسبتاً دقیق با یک خط راست که از مبدأ عبور کند نشان داد وضعیت خاصی پیش می‌آید که جالب توجه است. به عبارت دیگر، فرض کنید رابطهٔ بین x و y را بتوان با *مدل رگرسیونی* زیر بیان کرد

$$X_i = \beta Y_i + \varepsilon_i \qquad (۹.۷)$$

که در آن β، شیب خط رگرسیونی x و y است و $\sum_{i=1}^{N} \varepsilon_i = 0$.

پس عبارت مربوط به ρ_{xy} به صورت زیر درمی‌آید

$$\rho_{xy} = \frac{\sum_{i=1}^{N} (\beta Y_i + \varepsilon_i - \beta \bar{Y})(Y_i - \bar{Y})/N}{\sigma_x \sigma_y}$$
$$= \frac{\beta \sum_{i=1}^{N} (Y_i - \bar{Y})^2 + \sum_{i=1}^{N} \varepsilon_i (Y_i - \bar{Y})}{N \sigma_x \sigma_y}$$

اگر عبارت $\sum_{i=1}^{N} \varepsilon_i (Y_i - \bar{Y})$ نزدیک به صفر باشد (که اگر مدل بالا به خوبی با داده‌ها برازش داشته باشد همین طور هم خواهد شد) آنگاه ρ_{xy} تقریباً برابر خواهد بود با عبارت

$$\rho_{xy} \approx \frac{\beta \sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N \sigma_x \sigma_y}$$

چون از عبارت (۹.۷) نتیجه می‌گیریم که $\beta = \frac{\bar{X}}{\bar{Y}}$ و چون

$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = N \sigma_y^2,$$

داریم،

$$\rho_{xy} \approx \frac{\bar{X} \sigma_y}{\bar{Y} \sigma_x} = \frac{V_y}{V_x} \qquad (۱۰.۷)$$

این نتیجه، از رابطهٔ (۶.۷) *تقریب خطای معیار* x''، برآورد مجموع حاصل از نسبت را به ما می‌دهد:

$$SE(x'') \approx \left(\frac{YR}{\sqrt{n}}\right) \times (V_x^2 - V_y^2)^{1/2} \sqrt{\frac{N-n}{N-1}} \qquad (۱۱.۷)$$

که می‌تواند از روی داده‌ها با فرمول زیر *برآورد* شود

$$\hat{SE}(x'') \approx \left(\frac{Yr}{\sqrt{n}}\right) \times (\hat{V}_x^2 - \hat{V}_y^2)^{1/2} \sqrt{\frac{N-n}{N-1}} \qquad (۱۲.۷)$$

عبارتهای بالا بخصوص از این جهت سودمندند که شامل ضریب همبستگی به صورت صریح آن نیستند.

در مثال قبلی که در آن سه قسمت به عنوان نمونه انتخاب شدند، با فرض این که رگرسیون، خطی است و از مبدأ می‌گذرد، برآورد خطای معیار یعنی، $\hat{SE}(x'')$ که از تقریب (۱۲.۷) به دست می‌آید عبارت است از

$$\hat{SE}(x'') = \left[\frac{۵۰\times ۸۷۸/۵۸۴۸}{\sqrt{۳}}\right](۰/۰۲۴۵ - ۰/۰۰۴۴۳۵)^{٪}\sqrt{\frac{۸-۳}{۸-۱}} = ۳۰۳۶/۳۳$$

که از خطای معیار حاصل از رابطهٔ (۷.۷) بیشتر است.

۷.۷ تعیین اندازهٔ نمونه

برای این که $(۱-\alpha)\times ۱۰۰$٪ مطمئـــن باشیم که r، برآورد نسبت یا برآورد نسبتی مجموع، x''، در داخل محدودهٔ $\varepsilon\times ۱۰۰$٪ از مقدار واقعی R (یا X) قرار دارد، تقریب *اندازهٔ نمونهٔ* مورد نیاز از فرمول زیر به دست می‌آید

$$n = \frac{z^{۲}_{۱-(\alpha/۲)}\times N\times(V_x^{۲}+V_y^{۲}-۲\rho_{xy}V_xV_y)}{z^{۲}_{۱-(\alpha/۲)}\times(V_x^{۲}+V_y^{۲}-۲\rho_{xy}V_xV_y)+(N-۱)\varepsilon^{۲}} \qquad (۱۳.۷)$$

اگر پارامترهای عبارت (۱۳.۷) را از روی داده‌های اولیه بتوان حدس زد یا برآورد کرد، می‌توان برآورد اندازهٔ نمونه را به دست آورد.

۸.۷ برآورد رگرسیونی مجموعها

برآورد x'' از مجموع X بر مبنای یک نسبت r، موردی خاص از *برآورد رگرسیونی* مجموع است. *برآورد رگرسیونی* x''' دارای صورتی است که از فرمول زیر به دست می‌آید

$$x''' = x' + b(Y - y') \qquad (۱٤.۷)$$

که در آن b، برآورد شیب و x' و y'، برآوردهای تورمی ساده‌اند. برآورد شیب به صورت زیر است

$$b = \hat{\rho}'_{xy}\frac{\hat{SE}(x')}{\hat{SE}(y')} \qquad (۱۵.۷)$$

که در آن $\hat{\rho}'_{xy}$، برآورد ضریب همبستگی بین x' و y'، برآورد مجموعهاست. این فرمول برای نمونه‌گیری تصادفی ساده به صورت زیر تبدیل می‌شود.

$$b = \frac{\sum_{i=۱}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=۱}^{n}(y_i-\bar{y})^{۲}} \qquad (۱۶.۷)$$

برآوردی از واریانس برآورد رگرسیونی x''' که برای نمونه‌های با اندازه‌های بزرگ، تحت نمونه‌گیری تصادفی ساده، معتبر است از فرمول زیر به دست می‌آید

$$\hat{Var}(x''') = \hat{Var}(x')\times(۱-\hat{\rho}^{۲}_{xy}) \qquad (۱۷.۷)$$

این برآورد تحت شرایطی خاص واریانسی کمتر از برآورد نسبتی خواهد داشت. مثال زیر، کاربرد این برآورد را نشان خواهد داد.

مثال تشریحی: شهری دارای چهار محلهٔ تعریف شدهٔ جغرافیایی است. در سال ۱۹۸۵ یک آمارگیری از تأمین‌کنندگان مراقبتهای بهداشتی، از جمله بیمارستانها و تسهیلات مراقبت دراز مدت اجرا شد تا برآوردی از تعداد افرادی به دست آید که در طی سال مذکور به هپاتیت وابسته به استفاده از مواد مخدر تزریقی دچار شده بودند. مایل بودند برآورد مشابهی برای سال ۱۹۹۰ در سراسر شهر به دست آورند، ولی بودجه‌ای فقط برای اجرای بررسی مزبور در نمونه‌ای متشکل از دو محله از چهار محلهٔ شهر موجود بود. فرض کنیم داده‌های واقعی (نامعلوم) جامعه به شرح زیر است:

تعداد افراد دچار هپاتیت وابسته به تزریق مواد مخدر

محله	۱۹۹۰ (X_i)	۱۹۸۵ (Y_i)
۱	٤۰	۳۰
۲	۱۰۳	٦۰
۳	۱۱٥	۷۰
٤	۲۳	۲۱
مجموع	۲۸۱	۱۸۱

اگر محله‌های ۲ و ۳ انتخاب شده باشند، برآورد رگرسیونی x''' به شرح زیر به دست خواهد آمد:

$$x' = ۲\times(۱۰۳+۱۱٥) = ۴۳۶$$
$$y' = ۲\times(۶۰+۷۰) = ۲۶۰$$
$$\bar{x} = \frac{۱۰۳+۱۱٥}{۲} = ۱۰۹$$
$$\bar{y} = \frac{۶۰+۷۰}{۲} = ۶٥$$
$$b = \frac{(۱۰۳-۱۰۹)(۶۰-۶٥)+(۱۱٥-۱۰۹)(۷۰-۶٥)}{(۶۰-۶٥)^۲+(۷۰-۶٥)^۲} = ۱/۲۰$$
$$Y = ۱۸۱$$

و

$$x''' = x' + b(Y-y') = ۴۳۶+۱/۲۰(۱۸۱-۲۶۰) = ۳۴۱/۲$$

برآورد نسبتی x'' به صورت زیر به دست می‌آید

$$x'' = \left(\frac{x'}{y'}\right)\times Y = \left(\frac{۴۳۶}{۲۶۰}\right)\times ۱۸۱ = ۳۰۳/٥$$

توزیعهای برآورد رگرسیونی x'''، برآورد نسبتی x''، و برآورد تورمی ساده x' برای شش نمونهٔ ممکن در پایین نشان داده شده است:

نمونه	برآورد تورمی ساده x'	برآورد نسبتی x''	برآورد رگرسیونی x'''
۱ , ۲	۲۸٦	۲۸۷/۵۹	۲۸۸/۱۰
۱ , ۳	۳۱۰	۲۸۰/۵۵	۲۷٤/۳۸
۱ , ٤	۱۲٦	۲۲۳/۵۹	۲۷۵/۲۲
۲ , ۳	٤۳٦	۳۰۳/۵۲	۳٤۱/۲۰
۲ , ٤	۲۵۲	۲۸۱/۵٦	۲۹۰/۹۷
۳ , ٤	۲۷٦	۲۷٤/٤۸	۲۷٤/۱۲

از توزیعهای بالا، واریانسها، اریبیها، و میانگین توان دوم خطاها (MSEs) برای برآوردهای سه‌گانه عبارت‌اند از :

برآورد	واریانس	^{۲}(اریبی)	میانگین توان دوم خطا
x'	۸۲۹۷	۰	۸۲/۹۷
x''	٦۱۹/۸۳	۳۳/٤۷	٦٤۸/۳۰
x'''	۵۵٦/۳۳	۹۳/٤۳	٦٤۹/۷٦

می‌توان دید که در این مثال هم برآورد رگرسیونی x''' و هم برآورد نسبتی x'' دارای میانگین توان دوم خطایی هستند که به مراتب از میانگین توان دوم خطای برآورد تورمی ساده x' کمتر است. دراین مورد، برآورد مجموع رگرسیونی x''' علی‌رغم دارا بودن بیشترین اریبی، دارای میانگین توان دوم خطایی است که با میانگین توان دوم خطای مجموع حاصل از برآورد نسبتی تقریباً یکسان است. □

برآورد مجموع نسبتی x'' و برآورد مجموع رگرسیونی x''' هر دو از رابطهٔ بین متغیر موردنظر x و یک متغیر کمکی y در به دست آوردن برآوردی دقیقتر برای مجموع واقعی X استفاده می‌کنند. برآورد مجموع رگرسیونی در واقع صورت تعمیم یافتهٔ برآورد مجموع نسبتی است و وقتی ضریب رگرسیونی b برابر با $\frac{x'}{y'}$ باشد به همان صورت برآورد مجموع نسبتی تبدیل می‌شود.

۹.۷ برآورد نسبتی در نمونه‌گیری تصادفی طبقه‌بندی شده

وقتی طرح نمونه‌گیری، نمونه‌گیری تصادفی طبقه‌بندی شده است، برای برآورد نسبتها معمولاً از دو روش زیر استفاده می‌شود، یعنی از برآورد نسبتی ترکیبی r_{strc} و برآورد نسبتی تفکیکی r_{strs}. این دو روش را در زیر شرح می‌دهیم.

برآورد نسبتی ترکیبی، r_{strc}

$$r_{strc} = \frac{\bar{x}_{str}}{\bar{y}_{str}} \qquad (۱۸.۷)$$

که درآن $\bar{x}_{str}$ و $\bar{y}_{str}$، میانگینهای برآورد شده‌ای هستند که برای نمونه‌گیری تصادفی طبقه‌بندی شده که در فصل ٦ شرح داده شد مناسب‌اند. واریانس این برآورد از رابطهٔ زیر تقریب زده می‌شود.

$$Var(r_{strc}) \approx \left(\frac{۱}{N^{۲}\bar{y}^{۲}}\right)\sum_{h=۱}^{L}\frac{N_h^{۲}(N_h - n_h)}{n_h(N_h - ۱)}\sigma_{hz}^{۲} \qquad (۱۹.۷)$$

که در آن،

$$\sigma_{hz}^{۲} = \sigma_{hx}^{۲} + R^{۲}\sigma_{hy}^{۲} - ۲R\rho_{hxy}\sigma_{hx}\sigma_{hy}$$

و $\sigma_{hx}^{۲}$ و $\sigma_{hy}^{۲}$ در داخل محدودهٔ واریانسهای طبقه‌ای x و y قرار دارند و ρ_{hxy}، ضریب همبستگی بین x و y است.

برآور نسبتی تفکیکی r_{strs}

$$r_{strs} = \frac{\sum_{h=۱}^{L} x_h''}{Y} \qquad (۲۰.۷)$$

که درآن

$$x_h'' = Y_h \times \frac{\bar{x}_h}{\bar{y}_h}$$

$\bar{x}_h$ و $\bar{y}_h$ ؛ $h = ۱,\ldots,L$ برآورد میـــانگینهای ویژهٔ طبقه هستند

Y_h؛ $h = ۱,\ldots,L$ مجموعهای معلوم ویژهٔ طبقه برای y هستند

و

$$.Y = \sum_{h=۱}^{L} Y_h$$

واریانس تقریبی آن عبارت است از

$$Var(r_{strs}) \approx \left(\frac{۱}{Y^{۲}}\right)\sum_{h=۱}^{L} Var(x_h'') \qquad (۲۱.۷)$$

که در آن، $Var(x''_h)$ برای $h = ۱,\ldots,L$ ، واریانس برآورد مجموع نسبتی در داخل طبقهٔ h به شرحی است که در بخش ۲.۷ مورد بحث قرار گرفت.

برآورد نسبتی تفکیکی، r_{strs}، برای این که بتواند مورد استفاده قرار گیرد نیازمند شناخت مجموعهای Y_h طبقه است. شرایطی که طی آن یکی از این روشها بر دیگری ترجیح دارد پیچیده است و با جزئیات بیشتر، هنسن و همکاران [۱] و کوکران [۲] از آن بحث کرده‌اند. به طور کلی، هنگامی که نسبتهای درون طبقه‌ای، R_h، بیش از حد متنوع نباشند فایدهٔ چندانی بر استفاده از برآورد نسبتی تفکیکی مترتب نخواهد بود. برآوردهایی از واریانسهای برآورد نسبتی ترکیبی و برآورد نسبتی تفکیکی هر دو را می‌توان با جایگزین کردن آماره‌های نمونه‌ای مناسب برای پارامترهای نشان داده شده در رابطهٔ (۱۹.۷) و رابطهٔ (۲۱.۷) ساخت.

مثال تشریحی: یک دانشگاه بزرگ به کارکنان خود حق انتخاب می‌دهد تا به عنوان بخشی از برنامهٔ مزایای کارکنان به یکی از دو سازمان حفظ بهداشت (HMO) بپیوندند. برای تعیین هزینه‌ای که دانشگاه باید به ازای هر برخورد با بیمار متقبل شود، یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۲۰۰ نفر در داخل هر یک از این دو سازمان HMO گرفته شد و داده‌های زیر به دست آمد:

HMO	N_h	n_h	$\bar{x}_h$	$\hat{\sigma}_{hx}$	$\bar{y}_h$	$\hat{\sigma}_{hy}$	$\hat{\rho}_{hxy}$	Y_h
۱	۱۲۹۵	۲۰۰	۱۳٤٦ دلار	٥٦٤ دلار	۳/۲	۰/۹	۰/٦۹	٦۹۰
۲	۲۳۱۰	۲۰۰	۷۸٥ دلار	۳٤٥ دلار	۲/٥	۰/٦	۰/٦٥	٤۷۲

در این فهرست

$= N_h$ تعداد کارکنان دارای اشتراک HMO_h $(h = ۱,۲)$

$= n_h$ تعداد کارکنان نمونه‌گیری شده در HMO_h $(h = ۱,۲)$

$= \bar{x}_h$ برآورد میانگین هزینهٔ سالانه به ازای هر نفر در HMO_h $(h = ۱,۲)$

$= \hat{\sigma}_{hx}$ برآورد انحراف معیار توزیع میانگین هزینهٔ سالانه به ازای هر نفر در HMO_h $(h = ۱,۲)$

$= \bar{y}_h$ برآورد تعداد ویزیتهای سالانه به ازای هر نفر در HMO_h $(h = ۱,۲)$

$= \hat{\sigma}_{hy}$ برآورد انحراف معیار توزیع تعداد ویزیتها در سال به ازای هر نفر در HMO_h $(h = ۱,۲)$

$= \hat{\rho}_{hxy}$ برآورد همبستگی بین هزینهٔ سالانه به ازای هر نفر و ویزیتهای سالانه به ازای هر نفر در HMO_h $(h = ۱,۲)$

$= Y_h$ کل ویزیتهای سالانه در بین همهٔ افراد در HMO_h $(h = ۱,۲)$

برآورد نسبتی تفکیکی، r_{strs}، به شرح زیر محاسبه می‌شود:

$$r_{strs} = \left(\frac{۱}{۱۱۶۲}\right)\left(\frac{۱۳۴۶}{۳/۲}\times ۶۹۰ + \frac{۷۸۵}{۲/۵}\times ۴۷۲\right) = ۳۷۷/۳۱$$

از رابطه‌های (۱۲.۷) و (۲۱.۷) خطای معیار آن به صورت زیر برآورد می‌شود

$$\hat{SE}(r_{strs}) = ۵/۷۲$$

برآورد نسبتی ترکیبی، r_{strc}، از فرمول زیر به دست می‌آید

$$r_{strc} = \frac{\bar{x}_{str}}{\bar{y}_{str}} = \frac{۹۸۶/۵۲۴۳}{۲/۷۵۱۴۵۶} = ۳۵۸/۵۵$$

و از رابطهٔ (۱۹.۷) خطای معیار آن به صورت زیر برآورد می‌شود

$$\hat{SE}(r_{strc}) = ۱۴/۹۲$$

در این مثال بخصوص، چون نسبتهای تکی طبقه متنوع است ($r_۱ = ۴۲۰/۶۲$, $r_۲ = ۳۱۴/۰۰$) برآورد نسبتی تفکیکی بر برآورد نسبتی ترکیبی ترجیح داده می‌شود و این امر در خطاهای معیار برآورد شدهٔ این دو برآورد بازتاب یافته است.

۱۰.۷ خلاصه

در این فصل، مفاهیم اساسی برآورد نسبتی را شرح و بسط دادیم که فنی است که غالباً برای برآورد نسبتها و نیز مجموعهای جامعه‌ای مورد استفاده قرار می‌گیرد. برآورد کردن نسبتی را می‌توان با هر برنامهٔ نمونه‌گیری به کار برد و هرگاه برای برآورد مجموعهای جامعه‌ای به کار رود غالباً برآوردهایی تولید می‌کند که میانگین توان دوم خطای آن کمتر از آن است که از برآوردهای تورمی ساده تولید می‌شود. در این فصل برآورد نسبتی تحت نمونه‌گیری تصادفی ساده مورد تأکید خاصی قرار گرفت. توزیعهای نمونه‌گیری برآوردهای نسبتی را برای این نوع نمونه‌گیری مورد بحث قرار دادیم و روشهایی برای برآورد کردن خطاهای معیار، برای تعیین اندازه‌های نمونه، و برای ساختن بازه‌های اطمینان ارائه دادیم. بالاخره، برآورد مجموع رگرسیونی را مورد بحث قرار دادیم که از این جهت شبیه برآورد مجموع نسبتی است که از اطلاعات کمکی برای تولید برآورد دقیقتری از مجموع جامعه استفاده می‌کند.

تمرین

۱.۷ یک آمارگیری نمونه‌ای در دست برنامه‌ریزی است که می‌خواهند با آن، متوسط نسبت هزینه‌های درمانی به درآمد خانواده در یک شهر بزرگ با ۲۳٤۷۸۵ خانواده را برآورد کنند. براساس داده‌های همراه با تمرین ۲.۷، اگر برآورد این پارامتر با ۹۵٪ اطمینان در داخل محدودهٔ ۵٪ مقدار واقعی آن موردنظر باشد چند خانواده باید برای نمونه انتخاب شوند؟

۲.۷ جدول همراه با این تمرین که براساس یک نمونهٔ تصادفی ساده متشکل از ۳۳ خانواده از جامعه‌ای شامل ٦۰۰ خانواده گرفته شده است اندازهٔ خانواده، درآمد هفتگی خالص خانواده، و مخارج هفتگی هزینه‌های درمانی شامل دارو را ارائه می‌دهد. این جامعه شامل ۲۷۰۰ نفر است.

الف. متوسط هزینهٔ درمان هفتگی هر خانواده را با بازهٔ اطمینان ۹۵٪ برآورد کنید.

ب. متوسط هزینهٔ درمان هفتگی هر نفر را با بازهٔ اطمینان ۹۵٪ برآورد کنید. روش برآورد کردن را توجیه کنید.

پ. متوسط نسبت درآمد خانواده را که برای مخارج درمانی هزینه شده است با بازهٔ اطمینان ۹۵٪ برآورد کنید.

ت. کل هزینه‌های درمانی هفتگی پرداخت شده توسط جامعه را با بازهٔ اطمینان ۹۵٪ برآورد کنید.

ث. متوسط نسبت درآمد خانواده را که برای مخارج درمانی هزینه شده است با بازهٔ اطمینان ۹۵٪ برای خانواده‌هایی که درآمد خالص هفتگی آنان کمتر از ٤۰۰ دلار است برآورد کنید. همین کار را برای خانواده‌هایی که درآمد خالص هفتگی آنان بیشتر از ٤۰۰ دلار است انجام دهید.

شمارهٔ خانواده	اندازهٔ خانواده	درآمد هفتگی خالص (دلار)	هزینه‌های درمانی در هفته (دلار)
۱	۲	۳۷۲	۲۸/٦۰
۲	۳	۳۷۲	٤۱/٦۰
۳	۳	۵۲۲	٤۵/٤۰
٤	۵	۳۹۰	٦۱/۰۰
۵	٤	۳٤۸	۸۲/٤۰
٦	۷	۵۵۲	۵٦/٤۰

۷	۲	۵۲۸	٤٨/٤٠
۸	٤	۵۷٤	٦٠/٠٠
۹	۲	٤۹۸	٤٨/٤٠
۱۰	۵	۳۷۲	٨٨/٨٠
۱۱	۳	۳۷۸	۲٦/٨٠
۱۲	٦	۳۷۲	۳۹/٦٠
۱۳	٤	۳٦۰	۵٨/٨٠
۱٤	٤	٤۵۰	۵٤/۲۰
۱۵	۲	۵٤۰	٤٤/۲۰
۱٦	۵	٤۵۰	۷۵/٤٠
۱۷	۳	٤۱٤	٤۵/۲۰
۱۸	٤	٤۹۸	۷۲/٠٠
۱۹	۲	۵۱۰	۲۱/۲۰
۲۰	٤	٤۳۸	۵۵/٤٠
۲۱	۲	۳۹٦	۵۱/۹۰
۲۲	۵	۳٤۸	٤٦/٦٠
۲۳	۳	٤٦۲	۷۹/٦٠
۲٤	٤	٤۱٤	۳۳/٦٠
۲۵	۷	۳۹۰	۷۵/٦٠
۲٦	۳	٤٦۲	٦۹/٦٠
۲۷	۳	٤۱٤	۵۷/٤٠
۲۸	٦	۵۷۰	۱۲٦/٠٠
۲۹	۲	٤٦۲	۳۹/۱۰
۳۰	۲	٤۱٤	٤۳/۲۰
۳۱	٦	٤۱٤	۳٦/٤٠
۳۲	٤	٤۰۲	٤۰/۲۰
۳۳	۲	۳۷۸	٤۱/٤٠

۳.۷ برای شهر مذکور در تمرین ۲.۷، می‌خواهند برآورد کل مبلغی را که برای مخارج درمانی هزینه شده است نیز با اطمینان قطعی در محدودهٔ ۱۰ درصد مقدار واقعی آن به دست آورند. اگر قرار باشد این ویژگی تأمین شود، چند خانواده باید برای نمونه انتخاب شوند؟

٤.٧ برای داده‌های جدول همراه با تمرین ۲.۷، برآورد رگرسیونی از کل هزینه‌های درمانی هفتگی پرداخت شده توسط جامعه را به دست آورید. براساس این برآورد رگرسیونی، برای مقدار واقعی این پارامتر بازه‌های اطمینان ۹۵٪ ارائه دهید.

۵.۷ مطلوب است برآورد مجموع X ساعتی که یک پرستار متخصص برای مراقبت مستقیم از بیماران در یک سازمان بزرگ حفظ بهداشت طی سال گذشته صرف کرده است. این کار باید با گرفتن یک نمونهٔ تصادفی ساده از بیماران و تعیین تعداد ساعتهایی که پرستار متخصص برای هر ویزیت در طی سال صرف کرده است انجام شود. می‌دانیم که تعداد اعضای سازمان حفظ بهداشت ۳۵۲٤ نفر و تعداد ویزیتهای انجام شده در آن سال ۸۹۵۰ مورد است. نمونهٔ مقدماتی کوچکی از ۱۰ بیمار، داده‌های زیر را نتیجه داده است:

بیمار	ویزیت	ساعتهای صرف شده توسط پرستار متخصص
۱	۰	۰
۲	۵	۳
۳	۱	۰
٤	۲	٦
۵	۳	۳
٦	۷	۰
۷	۱	۰
۸	۱	۲
۹	۰	۰
۱۰	٤	۳

براساس این داده‌های مقدماتی، برآورد تورمی ساده را به عنوان روش برآورد کردن برای این آمارگیری نمونه‌ای توصیه می‌کنید یا برآورد نسبتی را؟ دلایل خود را برای این توصیه با مدرک ثابت کنید.

٦.٧ براساس داده‌های تمرین ۵.۷، اگر قرار شود برآورد تورمی ساده به کار رود چند بیمار باید نمونه‌گیری شوند؟

۷.۷ براساس داده‌های تمرین ۵.۷، اگر قرار شود برآورد نسبتی به کار رود چند بیمار باید نمونه‌گیری شوند؟

۸.۷ یک آمارگیری نمونه‌ای قرار است اجرا شود که به وسیلهٔ آن می‌خواهند نسبت اشخاص بالاتر از ۷۰ ساله‌ای را که دارای نشانه‌هایی از اختلالات شناختی هستند برآورد کنند. این کار قرار است با گرفتن نمـــونه‌ای از خانوارها و آزمون سـادهٔ کارکرد شناختی همهٔ اعضای بالاتر از ۷۰ سال خانوار انجام شود. یک بررسی مقدماتی از ۲۵ خانوار، به طور متوسط ۱/۲ نفر ۷۰ سال به بالا به ازای هر خانوار با انحراف معیاری برابر ۰/۸ و به طور متوسط ۰/۲٤ نفـــر ۷۰ سال به بالا به ازای هر خانوار با نشانه‌های اختلالات شناختی با انحراف معیـــاری برابر ۰/۷٦ را نتیجه داده است. می‌دانیم که این جامعه دارای ۳۰۵۸ خانوار است و ۲۹٤۹ نفر در آن هستند که بیشتر از ۷۰ سال سن دارند. اگر بخواهیم برآورد نسبت اشخاص ۷۰ سال به بالا که علائمی از اختلالات شناختی از خود نشان داده‌اند با ۹۵٪ اطمینان در محدودهٔ ۲۰٪ مقدار واقعی باشد چند خانوار باید نمونه‌گیری شوند؟

۹.۷ نشان دهید که در نمونه‌گیری تصادفی ساده، همبستگی ρ'_{xy} بین برآورد مجموعها، x'، و y'، از مشخصه‌های x و y برابر است با ρ'_{xy}، همبستگی بین x و y.

۱۰.۷ تمرین زیر در مقاله‌ای که در نشریهٔ گزارشهای بهـداشت عمومی [۳]، به چاپ رسیده پیشنهاد شده است. در یک مرکز انتقال خون واقع در یک شهر بزرگ برای تعیین میزان شیوع تغییرات از لحاظ HIV (ایدز) در سرم خون کسانی که برای اولین بار خون اهدا می‌کنند یک آمارگیری اجرا شد. در طول یک ماه خاص ۱۸۰ نفر برای اولین بار در این مرکز خون دادند. نتیجهٔ آزمایش خون ۱۷۵ نفر از این تعداد منفی بود. نمونه‌ای متشکل از ٦۰ نفر از افرادی که نتیجهٔ اولین خون‌دهی آنان منفی بود انتخاب و طی شش ماه بعد به آنان وقت داده شد تا دوباره برای خون دادن مراجعه کنند. از این اشخاص داده‌های زیر به دست آمد.

تعداد ماهها از		وضعیت از لحاظ HIV	
اولین اهدای خون	تعداد اشخاص	مثبت	منفی
۳	۱۸	۲	۱٦
٤	۱۰	۲	۸
۵	۸	۱	۷
٦	۷	۰	۷
۷	٦	۲	٤
۸	۵	۰	۵
۹	٤	۱	۳
۱۰	۲	۱	۱

از روی این داده‌ها نرخ شیوع سالانهٔ تغییرات سرمی از لحاظ HIV را در میان اهداکنندگان خون که برای اولین بار به مرکز مراجعه می‌کنند برآورد کنید. برای این نرخ شیوع بازهٔ اطمینان ۹۵ درصد ارائه دهید.

۱۱.۷ یک کارخانهٔ بزرگ ۱۰۰۰ کارگر دارد. یک نمونهٔ تصادفی ساده متشکل از ۲۵ کارگر گرفته شده است تا نسبت روزهای غیبت از کار به کل روزهای اشتغال در طول سال تقویمی گذشته به دست آید:

کل روزهای غیبت از کار در میان ۲۵ فرد نمونه‌گیری شده: ۲۵۰

کل روزهای اشتغال در میان ۲۵ فرد نمونه‌گیری شده: ۴۳۷۵

الف. برآورد نسبت روزهای غیبت از کار به کل روزهای اشتغال در بین کارگران این کارخانه چقدر است؟

ب. برای برآورد خطای معیار این برآورد به چه اطلاعات دیگری نیاز خواهید داشت؟

پ. از روی اطلاعات بالا کل تعداد روزهای غیبت از کار را در بین ۱۰۰۰ کارگر این گروه برآورد کنید.

ت. چنین رخ داده است که این ۱۰۰۰ کارگر در طی سال تقویمی گذشته در مجموع ۱۴۰۰۰۰ روز در این کارخانه مشغول به کار بوده‌اند. براساس این اطلاع، دربارهٔ برآورد کل روزهای غیبت از کار که در بخش پ محاسبه شد چه استنباط می‌کنید؟ برای پاسخ خود دلیل بیاورید.

۱۲.۷ در مورد مثال تشریحی قسمت ۳.۷ فرض کنید تعداد معلوم بیمارستانها در زیرحوزهٔ ۱ برابر با ۵۰ و در زیرحوزهٔ ۲ برابر با ۵۱ بیمارستان باشد.

الف. برآورد نسبتی پس طبقه‌بندی شدهٔ نسبت نوزادانی را تعیین کنید که با درج وضعیت پادگن سطحی هپاتیت B مادرانشان در پروندهٔ آنها از بیمارستان مرخص شده‌اند.

ب. دلایلی ارائه دهید که چرا این برآورد از برآورد پس طبقه‌بندی شده که در مثال تشریحی به دست آمد به برآورد معمولی این نرخ نزدیکتر است؟

کتابشناسی

The following texts contain more detailed discussions of ratio and regression estimation.

1. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. *Sample Survey Methods and Theory,* Vols. 1 and 2, Wiley, New York, 1953.
2. Cochran, W. G., *Sampling Techniques,* 2nd ed, Wiley, New York, 1962.

The following article gives an example of the use of a ratio estimate in an HIV seroconversion survey.

3. Petersen, L. R., Dodd, R., and Dondero, T. J., Jr., Methodological approaches to surveillance of HIV infection among blood donors, *Public Health Reports* 105: 153, 1990.

The following three review articles present an overview of ratio and regression estimators along with a long list of references including "classic" articles.

4. Rao, P. S. R. S., Ratio and regression estimators. In *Handbook of Statistics,* Krishnaian, P. R. and Rao, C. R., Eds., Vol. 6, *Sampling,* Chap. 18, pp 449-468, Elsevier, Amsterdam and New York, 1988.
5. Rao, J. N. K., Ratio estimators. In *Encyclopedia of Statistical Sciences*, Kotz, S., and Johnson, N. L., Eds., Vol. 7, pp. 639-646, New York, Wiley, 1986.
6. Cumberland, W. G., Ratio and regression estimators. In *Encyclopedia of Biostatistics*, Armitage, P. and T., Colton, Ed., Wiley, Chichester, U.K., 1998.

فصل ۸

نمونه‌گیری خوشه‌ای: مقدمه و بررسی اجمالی

فنون نمونه‌گیری که در فصلهای قبل مورد بحث قرار گرفتند همه مستلزم چارچوبهای نمونه‌گیریی هستند که واحدهای شمارش تکی (یا واحدهای فهرست‌برداری) را فهرست می‌کنند. ولی گاهی اوقات، بخصوص در آمارگیریهای نمونه‌ای از جوامع انسانی، تدوین چارچوبهایی برای نمونه‌گیری که همهٔ واحدهای شمارش را برای تمام جامعه فهرست کنند عملی و شاید اصلاً امکان‌پذیر نیست. از سوی دیگر، غالباً می‌توان چارچوبهایی برای نمونه‌گیری ساخت که گروهها یا *خوشه*هایی از واحدهای شمارش را شناسایی کنند بدون این که واحدهای شمارش تکی را صریحاً فهرست نمایند. می‌توان از روی این قبیل چارچوبها، با گرفتن نمونه‌ای از خوشه‌ها، به دست آوردن فهرستی از واحدهای شمارش، فقط برای آن خوشه‌هایی که در نمونه انتخاب شده‌اند، و سپس انتخاب نمونه‌ای از واحدهای شمارش، نمونه‌گیری را انجام داد. این طرحهای نمونه به نام *نمونه‌های خوشه‌ای* موسوم‌اند و در عمل کاربرد وسیعی دارند. در این فصل، برخی مفاهیم پایه‌ای نمونه‌گیری خوشه‌ای را معرفی می‌کنیم و در سه فصل بعدی به بحث در مورد برخی فنون خاص نمونه‌گیری خوشه‌ای می‌پردازیم که در سطح گسترده‌ای به کار می‌روند.

برای نشان دادن فرایند نمونه‌گیری خوشه‌ای و مقابلهٔ آن با طرحهای مبتنی بر نمونه‌گیری مستقیم از واحدهای شمارش، مثالی ساده را در نظر می‌گیریم.

مثال تشریحی: فرض کنید می‌خواهیم به منظور بررسی استفاده از خدمات بهداشتی درمیان اهالی یک شهر با اندازهٔ متوسط، نمونه‌ای از خانوارهای آن شهر انتخاب کنیم. اگر راهنمای بهنگام شهر که تمام خانوارهای شهر را فهرست می‌کند در دسترس باشد می‌توان آن را به عنوان چارچوب نمونه‌گیری برای انتخاب نمونه به کار برد. ولی اگر چنین راهنمایی یا هیچ گونه فهرست دیگری از خانوارها موجود نباشد، ساختن چنین چارچوبی برای نمونه‌گیری، با توجه به نفر ـ ساعت لازم، بسیار پرهزینه خواهد بود.

ولی ساختن فهرستی از بلوکهای شهر نسبتاً آسان است. برای بیشتر شهرها، نقشه‌هایی که هر بلوک شهری در آن مشخص و شماره‌گذاری شده است از طریق دفتر سرشماری امریکا در اختیار قرار می‌گیرد. این فهرست بلوکهای شهری را می‌توان به عنوان چارچوب نمونه‌گیری به کار برد. هر بلوک شهری را می‌توان به عنوان خوشه‌ای از خانوارها در نظر گرفت و هر خانوار در شهر با یک بلوک خاص همراه خواهد بود. نمونهٔ خانوارها را می‌توان ابتدا با گرفتن نمونه‌ای از بلوکها و سپس با فهرست‌برداری از یکایک خانوارها در داخل هر بلوکی که در نمونه انتخاب شده است تعیین کرد. نمونهٔ نهایی خانوارها را می‌توان از این فهرست به دست آمده از خانوارها انتخاب نمود. توجه کنید که لازم است فهرست‌برداری فقط از خانوارهای بلوکهایی به عمل آید که در نمونهٔ بلوکهای انتخاب شده قرار دارند.

□

۱.۸ نمونه‌گیری خوشه‌ای چیست؟

اصطلاح *خوشه*، وقتی در روش‌شناسی آمارگیری نمونه‌ای به کار می‌رود، می‌تواند به صورت هر واحد نمونه‌گیری که بتوان یک یا چند واحد فهرست‌برداری را با آن همراه کرد تعریف شود. واحد می‌تواند دارای ماهیت جغرافیایی، زمانی، یا مکانی باشد. چند مثال از خوشه‌هایی که ممکن است در عمل پیش بیاید درجدول ۱.۸ نشان داده شده‌اند.

برای توضیح بیشتر مثال دوم خوشه را که در جدول نشان داده شده است در نظر می‌گیریم. اگر بخواهیم نسبت همهٔ بیماران بستری شده در بیمارستان را که مردهٔ آنها از بیمارستان خارج شده است در یک ایالت خاص، طی یک سال خاص، برآورد کنیم می‌توانیم ابتدا تمام بخشهای آن ایالت را فهرست کنیم و از این فهرست، نمونه‌ای از بخشها بگیریم. سپس می‌توانیم برای هر یک از بخشهای انتخاب شده در نمونه، کلیهٔ بیمارستانهای موجود در آن بخش را فهرست کنیم و از بیمارستانهای تکی،

نمونه‌ای بگیریم. بالاخره، برای هر بیمارستانی که در نمونه انتخاب شده است کل تعداد اشخاصی را که در طول سال پذیرفته شده‌اند و کل افرادی را که مرده ترخیص شده‌اند به دست می‌آوریم. از روی این داده‌ها می‌توان نسبت کسانی را که مرده ترخیص شده‌اند برای سراسر ایالت برآورد کرد.

برای آخرین مثالی که در جدول ۱.۸ ارائه شده است، می‌توان فهرستی شامل ۵۲ هفتهٔ تقویمی تهیه و نمونه‌ای از هفته‌ها را از روی فهرست انتخاب کرد. برای هر یک از هفته‌های انتخاب شده در نمونه، می‌توان نمونه‌ای از روزها را انتخاب کرد و اندازه‌گیری اُزُن را در هر روز نمونه انجام داد.

جدول ۱.۸ برخی مثالهای کاربردی برای خوشه‌ها

خوشه	واحد فهرست‌برداری	واحد اولیه	کاربرد
بلوک شهری	خانوار	شخص	برآورد کردن کل ساکنان شهر که فشار خون دارند
بخش	بیمارستان	بیمار	برآورد کردن نسبت ترخیص‌شدگان مرده در یک ایالت خاص
مدرسه	کلاس	دانش‌آموز	برآورد کردن میانگین دستاوردهای تحصیلی در میان دانش‌آموزان یک ناحیهٔ تحصیلی
بستهٔ سرنگ	سرنگ تکی	سرنگ تکی	برآورد کردن نسبت تمام سرنگهای معیوب
کشو بایگانی	پوشهٔ تکی	حساب	برآورد کردن صورت حسابهایی که به موقع پرداخت نشده‌اند
صفحهٔ متن	سطر متن	واژه	برآورد کردن کل تعداد واژه‌های موجود در یک کتاب
هفته	روز	روز	برآورد کردن کل روزهایی که اُزُن در حداکثر سطح خود و بالاتر از یک سطح تعیین شده است

رابطهٔ بین خوشه‌ها و واحدهای فهرست‌برداری بسیار شبیه به رابطهٔ بین واحدهای فهرست‌برداری و واحدهای اولیه است. یک خوشه دارای واحدهای فهرست‌برداری همراه با خود است، درست به همان ترتیبی که یک واحد فهرست‌برداری دارای واحدهای اولیهٔ همراه با خود است. به طوری که بعداً خواهیم دید، نمونه‌گیری خوشه‌ای نوع سلسله مراتبی نمونه‌گیریی است که در آن، واحدهای اولیه غالباً حداقل دو مرحله با نمونه‌گیری اولیهٔ خوشه‌ها فاصله دارند.

حال که مفهوم خوشه‌ها را تعریف کردیم می‌توانیم *نمونه‌گیری خوشه‌ای* را در سطحی وسیع مانند هر برنامهٔ نمونه‌گیری دیگری که از چارچوب شامل خوشه‌هایی از واحدهای فهرست‌برداری استفاده می‌کند تعریف کنیم. تعریف بالا گسترده‌تر از آن است که در برخی متون دیگر ارائه شده است از این نظر که شامل نمونه‌گیریی است که در بیش از یک مرحله اجرا می‌شود. (یعنی *نمونه‌گیری چندمرحله‌ای*). برخی پژوهشگران اصطلاح *نمونه‌گیری خوشه‌ای* را به آن طرحهای نمونه‌گیری محدود می‌کنند که در آن خوشه‌ها توسط نوعی برنامهٔ نمونه‌گیری انتخاب می‌شوند و سپس هر واحد شمارش در داخل هر خوشه نمونه‌گیری می‌شود. در کتاب حاضر، ما به این قبیل طرحها به عنوان *نمونه‌گیری خوشه‌ای یک‌مرحله‌ای* اشاره می‌کنیم.

با در نظر گرفتن این تعریف گسترده‌تر از نمونه‌گیری خوشه‌ای، برخی جنبه‌های مهم نمونه‌گیری خوشه‌ای را در زیر فهرست می‌کنیم.

۱. فرایندی که طی آن نمونه‌ای از واحدهای فهرست‌برداری انتخاب می‌شود ممکن است گام به گام باشد. برای مثال، اگر بلوکهای شهری خوشه‌ها باشند و خانوارها واحدهای فهرست‌برداری را تشکیل دهند انتخاب خانوارهای نمونه ممکن است در دو گام اجرا شود. گام اول ممکن است مستلزم انتخاب نمونهٔ بلوکها و گام دوم می‌تواند مستلزم انتخاب خانوارهای نمونه در داخل هر یک از بلوکهای انتخاب شده در گام اول باشد. در اصطلاح‌شناسی نمونه‌گیری، این گامها را *مرحله* می‌نامند و برنامه‌های نمونه‌گیری غالباً بر حسب تعداد مرحله‌های به کار رفته مشخص می‌شوند. برای مثال، *نمونهٔ خوشه‌ای یک‌مرحله‌ای* نمونه‌ای است که طی آن نمونه‌گیری فقط در یک گام اجرا می‌شود. یعنی همین که نمونهٔ خوشه‌ها انتخاب شد، همهٔ واحدهای فهرست‌برداری در داخل هر یک از خوشه‌های انتخاب شده در نمونه منظور می‌شوند.

در بسیاری از آمارگیریهایی که مناطق وسیع جغرافیایی را پوشش می‌دهند، غالباً نمونه‌گیری در چندین مرحله انجام می‌شود. برای مثال، آمارگیری مصونیت از کودکان دبستانی در یک ایالت ممکن است مستلزم پنج مرحله باشد. اول، یک نمونه از بخشهای داخل ایالت

می‌گیریم. دوم، یک نمونه از شهرها یا سایر تقسیمات شهری کوچکتر در داخل هر یک از بخشهایی که در مرحلهٔ اول انتخاب شده‌اند می‌گیریم. سوم، یک نمونه از نواحی آموزشی در داخل شهرهایی که درمرحلهٔ دوم انتخاب شده‌اند می‌گیریم. چهارم، یک نمونه از مدرسه‌ها در داخل نواحی آموزشی که در مرحلهٔ سوم انتخاب شده‌اند می‌گیریم. پنجم، یک نمونه از کلاسهای درس در داخل هر یک از مدرسه‌هایی که در مرحلهٔ چهارم انتخاب شده‌اند می‌گیریم. و بالاخره یکایک کودکان را در کلاسهای درسی که در مرحلهٔ پنجم انتخاب شده‌اند آمارگیری می‌کنیم. در این مثال، کودکان واحدهای اولیه محسوب می‌شوند و کلاسهای درس واحدهای فهرست‌برداری هستند. باید پنج مرحلهٔ نمونه‌گیری اجرا شود که مستلزم چهار نوع خوشه‌اند: بخشها، شهرها، نواحی آموزشی، و مدرسه‌ها. در طرحهای نمونه‌ای که شامل دو یا چند مرحله‌اند، خوشه‌های مورد استفاده در مرحلهٔ اول نمونه‌گیری عموماً *واحدهای نمونه‌گیری اولیه* نامیده می‌شوند که به صورت خلاصه شدهٔ PSU[1] به آن اشاره می‌شود.

۲. خوشه‌ها را می‌توان با انواع گوناگونی از فنون نمونه‌گیری انتخاب کرد. برای مثال می‌توانیم نمونه‌ای از خوشه‌ها را با نمونه‌گیری تصادفی ساده یا با نمونه‌گیری سیستماتیک انتخاب کنیم. می‌توانیم خوشه‌ها را در طبقاتی گروه‌بندی کنیم و یک نمونهٔ تصادفی طبقه‌بندی شده از خوشه‌ها بگیریم.

هرگاه خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب شده باشند معمولاً برای توصیف طرح نمونه‌گیری از عبارت *نمونه‌گیری خوشه‌ای ساده* استفاده می‌شود. به طور دقیقتر برای رسته‌بندی طرحهای نمونه‌ای که در آن نمونه‌گیری تنها در یک مرحله اجرا می‌شود و خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب می‌شوند اصطلاح *نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده* به کار می‌رود. به همین ترتیب، برای توصیف طرحهای نمونه‌گیریی که در آن خوشه‌ها در مرحلهٔ اول با نمونه‌گیری تصادفی ساده انتخاب می‌شوند، و واحدهای فهرست‌برداری در مرحلهٔ دوم به طور مستقل در داخل هر خوشهٔ نمونه باز هم با نمونه‌گیری تصادفی ساده انتخاب می‌شوند، و کسر واحدهای فهرست‌برداری که در مرحلهٔ دوم انتخاب شده برای هر خوشهٔ نمونه یکسان است، از اصطلاح *نمونه‌گیری خوشه‌ای دومرحله‌ای ساده* استفاده می‌شود. نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده در فصل ۹ و نمونه‌گیری خوشه‌ای دومرحله‌ای ساده در فصل ۱۰ مورد بحث قرار گرفته‌اند.

[1] Primary Sampling Units

نوع دیگری از نمونه‌گیری خوشه‌ای که غالباً در عمل مورد استفاده قرار می‌گیرد *نمونه‌گیری با احتمال متناسب با اندازه* یا نمونه‌گیری PPS[1] نامیده می‌شود. این نوع نمونه‌گیری که در فصل ۱۱ مورد بحث قرار می‌گیرد، نمونه‌گیریی است که در آن خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب نمی‌شوند.

۳. در این فرایند ممکن است بیش از یک چارچوب نمونه‌گیری به کار رود. برای نشان دادن این وضعیت، به مثال بالا برمی‌گردیم که در آن یک آمارگیری مصونیت با یک شیوهٔ پنج‌مرحله‌ای اجرا می‌شود. چارچوب نمونه‌گیری در مرحلهٔ اول، فهرست تمام بخشهای داخل ایالت است. چارچوب نمونه‌گیری در مرحلهٔ دوم، فهرست شهرها یا سایر تقسیمات کوچکتر شهری در داخل هر یک از بخشهای انتخاب شده در مرحلهٔ اول است. چارچوب نمونه‌گیری در مرحلهٔ سوم، فهرست تمام نواحی آموزشی است که در داخل شهرها یا تقسیمات کوچکتر شهری انتخاب شده در مرحلهٔ دوم واقع شده‌اند. چارچوب نمونه‌گیری در مرحلهٔ چهارم نمونه‌گیری، فهرست مدرسه‌ها در داخل هر یک از نواحی آموزشی مرحلهٔ سوم است. بالاخره، چارچوب نمونه‌گیری مورد استفاده در مرحلهٔ پنجم، فهرست کلاسهای درس در داخل مدرسه‌هایی است که در مرحلهٔ چهارم نمونه‌گیری شده‌اند.

٤. پس از اولین مرحلهٔ نمونه‌گیری، چارچوب نمونه‌گیری فقط برای خوشه‌هایی که در نمونه انتخاب شده‌اند تهیه می‌شود. به محض این که خوشه‌های نمونهٔ مرحلهٔ اول انتخاب شدند، فهرست‌برداری از واحدهای نمونه‌گیری مرحلهٔ دوم فقط برای خوشه‌های نمونه انجام می‌شود. به همین ترتیب، اگر نمونه‌گیری در بیش از دو مرحله اجرا می‌شود واحدهای نمونه‌گیری هر یک از مراحل بعدی فقط در واحدهای نمونه‌گیری که در مرحلهٔ قبلی انتخاب شده‌اند فهرست می‌شوند. چون فهرست‌برداری از خوشه‌ها یا واحدهای فهرست‌برداری غالباً یکی از عملیات میدانی پرهزینه است، معمولاً هزینه‌های فهرست‌برداری برای طرحهای نمونه‌گیری خوشه‌ای به مراتب کمتر از طرحهای دیگر خواهد بود.

۲.۸ چرا نمونه‌گیری خوشه‌ای در سطحی گسترده به کار می‌رود؟

مهمترین دو دلیلی که نمونه‌گیری خوشه‌ای در سطحی چنین گسترده در عمل مورد استفاده قرار می‌گیرد، به خصوص در آمارگیریهای نمونه‌ای از جوامع انسانی و در آمارگیریهایی که مناطق جغرافیایی وسیعی را پوشش می‌دهند *امکان‌پذیر بودن* و *اقتصادی بودن* آن است.

[1] Probability Proportional to Size

نمونه‌گیری خوشه‌ای غالباً تنها روش نمونه‌گیری امکان‌پذیر است، زیرا تنها چارچوبهای نمونه‌گیری که برای جوامع هدف به راحتی در دسترس قرار می‌گیرند فهرستهای خوشه‌ها هستند. این حالت بخصوص در مورد آمارگیریهای جوامع انسانی که در آن از خانوار به عنوان واحد فهرست‌برداری استفاده می‌شود مصداق دارد. تهیهٔ فهرستی از خانوارها برای یک جامعهٔ آماری بزرگ (مانند آمریکا، یک ایالت، یا حتی یک شهر) فقط به منظور اجرای یک آمارگیری با توجه به زمان و منابع لازم، تقریباً هرگز امکان‌پذیر نیست. ولی فهرست بلوکها یا سایر واحدهای جغرافیایی را می‌توان نسبتاً به آسانی تهیه کرد و آنها را به عنوان چارچوب نمونه‌گیری به کار برد.

نمونه‌گیری خوشه‌ای غالباً باصرفه‌ترین شکل نمونه‌گیری است. نه تنها هزینه‌های فهرست‌برداری تقریباً همیشه برای نمونه‌گیری خوشه‌ای از همه کمتر است، بلکه هزینه‌های رفت و آمد هم غالباً از همه کمتر است. برای مثال، اگر خوشه یک واحد جغرافیایی از قبیل یک ناحیهٔ سرشماری باشد، در آن صورت، به محض این که خانوارها در داخل نواحی سرشماری انتخاب شدند هزینه‌های رفت و آمد در داخل نواحی نمونه از خانواری به خانوار دیگر نسبتاً اندک خواهد بود. به این ترتیب، یک نمونهٔ خوشه‌ای از خانوارها در داخل چند ناحیهٔ سرشماری نمونه، مستلزم رفت و آمدی به مراتب کمتر از یک نمونهٔ تصادفی ساده با همان تعداد خانوار است که در تعداد بسیار بیشتری از نواحی سرشماری پراکنده هستند.

نمونه‌گیری خوشه‌ای می‌تواند در آمارگیریهای نهادها از قبیل بیمارستانها سودمند واقع شود. وارد کردن یک نهاد در یک بررسی غالباً بسیار پرخرج و وقت‌گیر است. برای مثال، کسب اجازه از مدیر یک بیمارستان برای بیرون کشیدن نمونه‌ای از پرونده‌های بیماران بیمارستان ممکن است مستلزم تلاش بسیار زیاد، مقداری تخصص در روابط عمومی، و گاهی حتی اعمال نفوذ فراوان باشد. بنابراین، به محض این که دسترسی به پرونده‌های بیمارستان فراهم شد غالباً این ارزش را دارد که به جای چند پرونده تعداد زیادی از پرونده‌ها نمونه‌گیری شوند.

برای توضیح بیشتر در مورد این که نمونه‌گیری خوشه‌ای چگونه می‌تواند هزینه‌های یک آمارگیری نمونه‌ای را پایین بیاورد به بررسی مثال زیر می‌پردازیم.

مثال تشریحی: فرض کنید پنج شهرک مسکونی برای شهروندان سالخورده در سراسر یک ایالت خاص به طور پراکنده وجود دارند و هر یک از این شهرکها دارای ۲۰ آپارتمان است. فرض کنید قرار است نمونه‌ای از ۱۰ آپارتمان انتخاب شود تا کل تعداد شهروندان سالخورده در این شهرکها که نیاز به خدمات پرستار متخصص دارند برآورد شود. علاوه بر این، فرض کنیم که خانوارها واحدهای

فهرست‌برداری هستند و هزینه‌های میدانی عمدهٔ آمارگیری مربوط به کارهای فهرست‌برداری و مصاحبه است.

برای فهرست کردن همهٔ آپارتمانها در یک شهرک مسکونی، فرض می‌کنیم که یکی از اعضای کارکنان میدانی ناچار است به شهرک برود و نامهای خانوادگی را از روی جعبه‌های پستی با سرعت یادداشت کند. فرض کنید رفتن به شهرک ۰/۵ ساعت و فهرست کردن هر خانوار ۳ دقیقه وقت بگیرد، یا فهرست‌برداری از تمام خانوارها در شهرک (دقیقه ٦۰=۲۰×۳) وقت لازم داشته باشد. پس فرض می‌کنیم که برای فهرست کردن سراسر یک شهرک ۱/۵ نفر ساعت (۱ ساعت فهرست‌برداری به اضافهٔ ۰/۵ ساعت سفر) و برای فهرست کردن همهٔ شهرکهای پنج‌گانه ۷/۵ ساعت (۱/۵ ساعت برای هر شهرک ضربدر ۵ شهرک) وقت لازم باشد.

فرض کنید مصاحبه با هر خانوار انتخاب شده در نمونه ۱۵ دقیقه (۰/۲۵ ساعت) وقت بگیرد. پس هزینهٔ مصاحبه از لحاظ نفر ـ ساعت در یک شهرک خاص برابر است با ۰/۲۵ برابر تعداد خانوارهای انتخاب شده در آن شهرک به اضافهٔ هزینهٔ سفر به شهرک (۰/۵ نفر ـ ساعت برای هر سفر).

حالا بیایید دو طرح نمونه‌گیری را با توجه به هزینه‌های فهرست‌برداری و مصاحبه برای نمونه‌ای از ۱۰ خانوار مقایسه کنیم. طرح اول، نمونه‌گیری تصادفی ساده است و فرض می‌کنیم خانوارهایی از چهار شهرک در نمونه انتخاب شده‌اند. طرح دوم، یک نمونهٔ خوشه‌ای دومرحله‌ای ساده از ۱۰ خانوار است یعنی پنج خانوار از هر یک از دو شهرک. هزینه‌های فهرست‌برداری و مصاحبه برای هر طرح در جدول ۲.۸ نشان داده شده است.

جدول ۲.۸ مقایسهٔ هزینه‌ها در دو طرح نمونه‌گیری

	طرح	
هزینه‌ها	نمونه‌گیری تصادفی ساده	نمونه‌گیری خوشه‌ای دومرحله‌ای ساده
هزینه‌های فهرست‌برداری	۱/۵ برای هر شهرک ×۵ شهرک = ۷/۵	۱/۵ برای هر شهرک ×۲ شهرک = ۳/۰
هزینه‌های مصاحبه برحسب نفر ـ ساعت	سفر به ٤ شهرک + مصاحبه ٤×۰/۵+۱۰×۰/۲۵ = ٤/۵	سفر به ۲ شهرک + مصاحبه ۲×۰/۵+۱۰×۰/۲۵ = ۳/۵
کل هزینه‌های میدانی برحسب نفر ـ ساعت	۱۲/۰	٦/۵

از محاسبات جدول ۲.۸ پی می‌بریم که حتی در مورد این نمونهٔ کوچک نیز هزینه‌های میدانی مبتنی بر نمونه‌گیری خوشه‌ای، می‌تواند به مراتب کمتر از هزینه‌های میدانی نمونه‌گیری تصادفی ساده باشد. صرفه‌جوییها بخصوص در هزینه‌های فهرست‌برداری و مصاحبه است.
□

۳.۸ عیب نمونه‌گیری خوشه‌ای: خطاهای معیار زیاد

اکنون که نشان دادیم نمونه‌گیری خوشه‌ای چقدر می‌تواند مقرون به صرفه باشد و غالباً چقدر امکان‌پذیر است، باید خاطرنشان کنیم که خطاهای معیار برآوردهای به دست آمده از طرحهای نمونه‌گیری خوشه‌ای در مقایسه با خطاهای معیار به دست آمده از نمونه‌های دیگری با همین تعداد واحدهای فهرست‌برداری که از طرحهای نمونه‌گیری دیگری انتخاب شده‌اند غالباً بسیار زیاد است. علت این وضعیت آن است که واحدهای فهرست‌برداری در داخل یک خوشه از لحاظ بسیاری از مشخصه‌ها غالباً همگن‌اند. برای مثال، خانوارهای ساکن در یک بلوک غالباً از لحـــاظ وضعیت اجتماعی ـ اقتصادی، قومی، و سایر متغیرها بسیار شبیه‌اند. به علت همگنی میان واحدهای فهرست‌برداری در داخل یک خوشه، انتخاب بیش از یک خانوار در داخل همان خوشه، به صورتی که در نمونه‌گیری خوشه‌ای اجرا می‌شود، به تعبیری، زیاده‌روی محسوب می‌شود. اثر این زیاده‌روی در خطاهای معیار بسیار زیاد برآوردها آشکار می‌شود که غالباً در نمونه‌گیری خوشه‌ای به چشم می‌خورد.

اگر قرار باشد بین نمونه‌گیری خوشه‌ای و هر طرح دیگری صرفاً براساس هزینه و امکان‌پذیر بودن، یکی را انتخاب کنیم، نمونه‌گیری خوشه‌ای به صورتی اجتناب ناپذیر طرح نمونه‌گیری منتخب خواهد بود. از سوی دیگر، اگر قرار باشد طرحی را صرفاً براساس قابل اعتماد بودن برآوردها انتخاب کنیم، در آن صورت نمونه‌گیری خوشه‌ای هرگز نمی‌تواند طرح منتخب باشد. ولی چون امکان گرفتن نمونه‌ای بزرگتر با هزینهٔ ثابت با نمونه‌گیری خوشه‌ای فراهم است، دقت نتایج به مراتب به بیش از آن خواهد رسید که در سایر روشها می‌رسد. معمولاً در انتخاب بین نمونه‌گیری خوشه‌ای وسایر گزینه‌ها، از ملاکی استفاده می‌کنیم که قابل اعتماد بودن و هزینه را مشترکاً در نظر بگیرد. در واقع، عموماً آن طرح نمونه‌گیری را انتخاب می کنیم که کمترین خطای معیار ممکن را با هزینه‌ای مشخص به دست آورد یا، برعکس، آن طرح نمونه‌گیری را انتخاب می‌کنیم که با کمترین هزینه، برآوردهایی را با خطای معیار مشخص نتیجه دهد.

٤.٨ چگونگی بررسی نمونه‌گیری خوشه‌ای در این کتاب

بـه دنـبال بررسی مجمل نمونه‌گیری خوشه‌ای که در این فصل ارائه شد، در سه فصل بعدی به شرح و بسـط طـرحهای مشخص نمونه‌گیری می‌پردازیم و دربارهٔ طرحهای نمونه‌گیری بسیار ساده تا طرحهای نمونـه‌گیـری بسـیار پیچیـده بحـث می‌کنیم. ساده‌ترین نوع نمونه‌گیری خوشه‌ای، *نمونه‌گیری خوشه‌ای یک‌مـرحله‌ای* اسـت کـه در آن خوشـه‌هـا با نمونه‌گیری تصادفی ساده انتخاب می شوند و از این نوع نمونه‌گیری خوشه‌ای در فصل بعد (فصل ۹) بحث شده است. دو فصـل پس از آن (فصلهای ۱۰ و۱۱) هـر دو بـه نمونـه‌گیـری خوشـه‌ای دومـرحله‌ای می‌پردازند. در نمونه‌گیری خوشه‌ای دومرحله‌ای، تهیهٔ بـرآوردهای مشخصـه‌هـای جامعـه و برآوردهای خطاهای معیار آن دسته از آماره‌هایی که برای برآورد مشخصه‌های جامعه به کار می‌روند اگر تعداد واحدهای فهرست‌برداری در هر خوشه در جامعه یکسان باشـد نسـبتاً آسان است. نمونه‌گیری خوشه‌ای دومرحله‌ای برای این وضعیت در فصل ۱۰ بررسی شده اسـت. نمونـه‌گیـری خوشـه‌ای دومـرحله‌ای در پیچیده‌ترین شکل خود هنگامی روی می‌دهد که تعداد واحدهـای فهرسـت‌بـرداری در هـر خوشـه از جامعه یکسان نباشد. متأسفانه، این حالت در بسیاری از وضـعیتهای کاربـردی اتفـاق مـی‌افتد واحساس می‌کنیم که باید در این کتاب به آن بپـــردازیم، گرچه روش مزبور مستلزم استفاده از فرمولهایی است که به مراتب خسته‌کننده‌تر از آنهایی هستند که در سایر بخشـهای این کتاب ارائه شده‌اند. این نوع نمونه‌گیری خوشه‌ای را در فصل ۱۱ شرح و بسط می دهیم. طـی ۲۵ سـال گذشـته، پیشـرفتهایی قابل ملاحظه در توسعهٔ روشها ونرم افزارهای آماری برای برآورد خطاهـای معیـار بـرآوردهای حاصل از حتی پیچیده‌ترین طرحهای نمونه‌گیری خوشه‌ای به دست آمده است. همان طور که در فصلهای قبل انجام شد، نشان خواهیم داد که چگونه دو نرم افزار موجود فعلی یعـنی SUDAAN و STATA را مـی‌توان برای برآورد کردن خطاهای معیار برآورد مشخصه‌های جامعه تحـت طـرحهای نمونه‌گیری خوشه‌ای پیچیده به کار برد. در فصل ۱۲، برخی از روشهای آماری را که کاربـرد وسـیعتری دارنـد بـرای به دست آوردن برآوردهای این خطاهای معیار به صورتی کاملتر مورد بحـث قـرار خواهیـم داد. رفـتار مـا بـا نمونـه‌گیـری خوشه‌ای بسیار شبیه به رفتار ما با سایر طرحهای نمونـه‌گیـری اسـت. بـه خصـوص بحثهایی را به چگونگی انتخاب نمونه‌ها، چگونگی به دست آوردن بـرآوردها، طـبیعت توزیعهـای نمونـه‌گیـری این برآوردها، و روشهای برآورد کردن خطاهای معیار این برآوردها اختصاص می دهیم.

۵.۸ خلاصه

در این فصل، بررسی مجملی از نمونه‌گیری خوشه‌ای ارائه دادیم. مفهوم یک خوشه از واحدهای فهرست‌برداری یا شمارش را معرفی و نمونه‌گیری خوشه‌ای را مانند هر طرح نمونه‌گیری دیگری تعریف کردیم که از یک چارچوب نمونه‌گیری شامل خوشه‌هایی از واحدهای فهرست‌برداری استفاده می‌کند. مثالهایی برای خوشه‌ها ارائه شد و مشخصه‌هایی از نمونه‌گیری خوشه‌ای مورد بحث قرار گرفت از قبیل ماهیت گام به گام دنباله‌ای و این واقعیت که واحدهای فهرست‌برداری لازم است فقط از خوشه‌های نمونه فهرست‌برداری شوند و نه از تمام خوشه‌ها. به مزایای نمونه‌گیری خوشه‌ای (یعنی، امکان‌پذیر بودن و اقتصادی بودن) و نیز به عیب عمدهٔ آن (یعنی، خطاهای معیار بسیار زیاد برآوردها) اشاره کردیم. ولی به علت امکانات بالقوهٔ این طرح برای داشتن مشاهداتی بیشتر از آنچه که با استفاده از سایر طرحها با همان بودجه امکان‌پذیر است غالباً این امکان وجود دارد که خطاهای معیار تا آن حد کاهش یابد که نمونه‌گیری خوشه‌ای روش منتخب باشد.

تمرین

۱.۸ در نمونه‌ای با ۱۰۰ عنصر، کدام یک از طرحهای زیر احتمال دارد برآوردهایی با بیشترین خطاهای معیار به دست دهد؟

الف. نمونه‌گیری سیستماتیک

ب. نمونه‌گیری خوشه‌ای

پ. نمونه‌گیری تصادفی ساده

ت. نمونه‌گیری تصادفی طبقه‌بندی شده

۲.۸ اگر کسی بخواهد از این صفحه در مورد واژه‌هایی که غلط نوشته شده‌اند نمونه بگیرد، خوشهٔ منطقی کدام خواهد بود؟

۳.۸ اگر کسی بخواهد در سرتاسر این کتاب در مورد واژه‌هایی که غلط نوشته شده‌اند براساس یک طرح نمونه‌گیری خوشه‌ای سه‌مرحله‌ای نمونه بگیرد، واحدهای نمونه‌گیری اولیه چه می‌توانند باشند؟ واحدهای نمونه‌گیری ثانویه، واحدهای فهرست‌برداری، و واحدهای مقدماتی کدام خواهند بود؟

کتابشناسی

Cluster sampling is discussed in all of the texts on sampling referenced in previous chapters. References to specific methods and applications of cluster sampling are given in Chapters 9, *10,* 11, *and* 14.

فصل ۹

نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

شـــرح و بسـط خـود را دربـارهٔ بـرخی از متداولـترین طـرحهای نمونـه‌گیـری خوشـه‌ای با بحث از نمـــونـه‌گیـــری خوشـــه‌ای یـک‌مـرحله‌ای سـاده آغـاز مـی‌کـنیم. همـان‌طور که در فصل ۸ ذکر شد، *نمونـه‌گیـری خوشـه‌ای یـک‌مـرحله‌ای سـاده* یک برنامهٔ نمونه‌گیری است که در آن خوشه‌ها به وسیلهٔ نمونـه‌گیری تصادفی ساده انتخاب و در داخل هر خوشهٔ نمونه، همهٔ واحدهای فهرست‌برداری برگزیده می‌شوند.

نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده در وضعیتهایی که هزینهٔ به دست آوردن یکایک واحدهای فهرسـت‌بـرداری در داخـل یـک خوشـه بیشـتر از هزیـنهٔ بـه دسـت آوردن نمونـه‌ای از واحدهـای فهرسـت‌بـرداری نباشد یا فقط اندکی از آن بیشتر باشد فراوان به کار می‌رود. برای مثال، یک آمارگیری نمونه‌ای را در نظر می‌گیریم که در آن، بیمارستانها خوشه‌ها را تشکیل می‌دهند و بیماران بستری شده با تشـخیص یک بیماری خاص، واحدهای فهرست‌برداری هستند. اگر اطلاعات موردنیاز برای آمارگیری را بـتوان از نـتیجه‌هـای چـاپی رایانـه بـه دسـت آورد که تجربهٔ هر بیمار را خلاصه می‌کنند و اگر این اطلاعـات موجـود اسـت یـا مـی‌تـوان آنهـا را به آسانی توسط کتابدار پرونده‌های بیمارستان تهیه کرد، انـتخاب یکـایک بیمـارانی کـه مبـتلا بـه آن بیماری خاص تشخیص داده شده‌اند ارزانتر و راحت‌تر از انتخاب نمونه‌ای از این بیماران خواهد بود. از سوی دیگر اگر داده‌های موردنیاز دربارهٔ هر بیمار باید از

اطلاعـات مـربوط بـه پـروندهٔ پزشکی اصلی بیمارستان برای آن بیمار استخراج شود (که غالباً فرایندی پرخـرج و وقت‌گیر است) آن‌گاه، انتخاب نمونه‌ای از این بیماران باصرفه‌تر از انتخاب همهٔ آنان خواهد بود.

بـه طوری که بعداً در این فصل خواهیم دید، اگر خود خوشه را واحد فهرست‌برداری مؤثر در نظر بگیـریم، شیوه‌های برآورد کردن مربوط به نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده در واقع با شیوه‌های بـرآورد کـردن مـربوط بـه نمونـه‌گیری تصادفی ساده یکسان‌اند. این مفهوم را بعداً در این فصل بیشتر توضـیح خواهیم داد. ولی یکی از معایب این نوع نمونه‌گیری آن است که خطاهای نمونه‌گیری همراه با بـرآوردهای حاصـل از نمونـه‌گیـری خوشـه‌ای یـک‌مرحله‌ای ساده معمولاً به مراتب بیشتر از خطاهای نمونه‌گیری همراه با برآوردهای حاصل از یک نمونهٔ تصادفی ساده با همان تعداد واحد فهرست‌برداری است، بخصوص هنگامی که واحدهای فهرست‌برداری داخل خوشه‌ها نسبت به متغیر مورد اندازه‌گیری همگـن باشـند. فـرض کـنید بـرای مـثال، می‌خواهیم نسبت همهٔ خانواده‌های زیر سطح فقر را در یک جامعـه از روی یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده برآورد کنیم که در آن بلوکهای شهری، خوشه‌ها هسـتند و خانوارهـا واحدهـای فهرسـت‌بـرداری را تشکیل می‌دهند. چون خانوارهای داخل هر بلوک شـهری احـتمال دارد شـامل خانواده‌هایی با درآمد بسیار مشابه باشند، گنجاندن یکایک خانوارهای یک بلوک انتخاب شده در نمونه، راه اسرافکارانه‌ای برای تخصیص واحدهای فهرست‌برداری نمونه خواهد بود و احتمال دارد برآوردهایی با خطاهای نمونه‌گیری زیاد به بار آورد.

۱.۹ چگونگی انتخاب یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده

نمونـهٔ خوشـه‌ای یـک‌مرحله‌ای را به این ترتیب می‌گیریم که اول همهٔ خوشه‌های موجود در جامعه را فهرسـت‌بـرداری می‌کنیم و سپس طبق معمول یک نمونهٔ تصادفی ساده از خوشه‌ها می‌گیریم. سپس در داخـل هـر یـک از خوشـه‌هـایی که در نمونه انتخاب شده‌اند همهٔ واحدهای فهرست‌برداری را منظور می‌کنیم.

بـرای مـثال، فـرض کـنیم کـل تعداد اشخاص ٦٥ سال به بالا و تعداد اشخاص ٦٥ سال به بالا که نیازمـندِ خدمـات یـک پرسـتار متخصـص‌انـد در پنج شهرک مسکونی مورد بحث در فصل ۸ به شرح جدول ۱.۹ باشد. از این داده‌ها برای نشان دادن چگونگی اجرای نمونه‌گیری خوشه‌ای استفاده خواهیم کرد.

فـرض کـنید مـی‌خواهیم یک نمونهٔ خوشه‌ای یک مرحله‌ای ساده متشکل از دو شهرک مسکونی از شهرکهای مسکونی جدول ۱.۹ بگیریم. از بین اعداد ۱ تا ۵ دو عدد تصادفی را انتخاب می‌کنیم و سپس با یکایک خانـــوارها در دو شهرک مسکونی که در نمونه انتخاب شده‌اند مصاحبه می‌کنیم. فرض کنید

جدول ۱.۹ تعداد اشخاص ۶۵ سال به بالا در پنج شهرک مسکونی و تعداد اشخاص ۶۵ سال به بالایی که نیاز به خدمات پرستار متخصص دارند

شهرک ۱			شهرک ۲			شهرک ۳			شهرک ٤			شهرک ٥		
خانوار	تعداد اشخاص ٦٥ سال به بالا	تعداد اشخاص ٦٥ سال به بالا نیازمند خدمات پرستار متخصص	خانوار	تعداد اشخاص ٦٥ سال به بالا	تعداد اشخاص ٦٥ سال به بالا نیازمند خدمات پرستار متخصص	خانوار	تعداد اشخاص ٦٥ سال به بالا	تعداد اشخاص ٦٥ سال به بالا نیازمند خدمات پرستار متخصص	خانوار	تعداد اشخاص ٦٥ سال به بالا	تعداد اشخاص ٦٥ سال به بالا نیازمند خدمات پرستار متخصص	خانوار	تعداد اشخاص ٦٥ سال به بالا	تعداد اشخاص ٦٥ سال به بالا نیازمند خدمات پرستار متخصص
۱	۳	۱	۱	۲	۱	۱	۳	۰	۱	۳	۰	۱	۱	۱
۲	۱	۰	۲	۱	۰	۲	۲	۱	۲	۱	۰	۲	۱	۰
۳	۱	۱	۳	۲	۰	۳	۲	۱	۳	۳	۰	۳	۱	۱
٤	۱	۱	٤	۱	۱	٤	۳	۱	٤	۱	۰	٤	۳	۱
٥	۱	۱	٥	۱	۰	٥	۲	۱	٥	۲	۰	٥	۲	۰
٦	۱	۱	٦	۱	۰	٦	۲	۱	٦	۲	۰	٦	۱	۱
۷	۱	۱	۷	۲	۱	۷	۱	۰	۷	۱	۰	۷	۳	۰
۸	۱	۰	۸	۱	۱	۸	۱	۰	۸	۱	۰	۸	۱	۰
۹	۳	۱	۹	۳	۱	۹	۳	۰	۹	۳	۱	۹	۱	۰
۱۰	۳	۲	۱۰	۱	۱	۱۰	۳	۱	۱۰	۱	۱	۱۰	۳	۲
۱۱	۳	۲	۱۱	۱	۰	۱۱	۲	۰	۱۱	۱	۰	۱۱	۲	۱
۱۲	۱	۱	۱۲	۲	۱	۱۲	۳	۱	۱۲	۱	۰	۱۲	۳	۰
۱۳	۱	۰	۱۳	۱	۰	۱۳	۱	۱	۱۳	۳	۱	۱۳	۱	۱
۱٤	۱	۱	۱٤	۳	۱	۱٤	۱	۱	۱٤	۳	۰	۱٤	۱	۰
۱٥	۱	۰	۱٥	۱	۰	۱٥	۲	۲	۱٥	۱	۰	۱٥	۲	۱
۱٦	۳	۱	۱٦	۳	۲	۱٦	۱	۰	۱٦	۱	۰	۱٦	۲	۱
۱۷	۱	۱	۱۷	۱	۰	۱۷	۱	۱	۱۷	۱	۰	۱۷	۱	۰
۱۸	۳	۲	۱۸	۲	۰	۱۸	۳	۰	۱۸	۱	۱	۱۸	۲	۰
۱۹	۱	۱	۱۹	۱	۱	۱۹	۱	۰	۱۹	۱	۰	۱۹	۲	۱
۲۰	۱	۱	۲۰	۳	۰	۲۰	۲	۲	۲۰	۱	۰	۲۰	۱	۱
مجموع	۳۲	۱۹		۳۳	۱۱		۳۹	۱٤		۳۲	٤		۳٤	۱۲

اعـداد تصـادفی منـتخب ۲ و ۵ باشـند. در ایـن صـورت بـا همهٔ خانوارها در شهرکهای شمارهٔ ۲ و ۵ مصاحبه می‌کنیم و داده‌های مربوط به آنها جمع‌آوری خواهد شد.

۲.۹ برآورد کردن مشخصه‌های جامعه

بـه محـض این‌که خوشه‌های نمونه انتخاب شدند داده‌های لازم از همهٔ واحدهای نمونه‌گیری در داخل خوشـه‌هـای انـتخاب شـده جمـع‌آوری مـی‌شـوند. البته فرمهایی که برای جمع‌آوری این داده‌ها به کار مـی‌رونـد بـه نـوع اطلاعـات موردنیاز و روشی که با آن، داده‌ها جمع‌آوری می‌شوند بستـــگی دارنـد. (مـثـلاً پسـتی، مصـاحبه با شخص، تلفنی، استخراج از سوابق و غیره). سپس برآوردهای مشخصه‌های جامعه از روی داده‌های جمع‌آوری شده محاسبه می‌شوند.

شیوه‌های برآورد کردن را با پرداختن به یک مثال بررسی می‌کنیم.

مـثال تشـریحی: فـرض کـنید کـه از روی نمونـهٔ متشـکل از دو واحـد مسکونی که در بالا اشاره شد می‌خواهیم مشخصه‌های زیر را برای جامعه برآورد کنیم:

۱. کل تعداد اشخاص ٦۵ سال به بالا در پنج شهرک مسکونی؛
۲. کل تعداد اشخاص ٦۵ سال به بالا که به خدمات پرستار متخصص نیاز دارند؛ و
۳. نسبت همهٔ اشخاص ٦۵ سال به بالا که به خدمات پرستار متخصص نیاز دارند؛
٤. میانگین تعداد اشخاص ٦۵ سال به بالا به ازای هر شهرک مسکونی؛
۵. نسـبت همـهٔ خانوارهای دارای حداقل یک نفر ٦۵ سال به بالا که به خدمات پرستار متخصص نیاز دارند.

بـرای بـه دسـت آوردن ایـن بـرآوردها، فرمی که در شکل ۱.۹ نشان داده شده است ممکن است برای جمع‌آوری داده‌ها از طریق مصاحبهٔ شخصی مناسب باشد.

اگـر شـهرکهای مسکونی ۲ و ۵ در نمونه انتخاب شوند در آن صورت فرمی که در شکل ۱.۹ نشان داده شـده اسـت داده‌هـایی را ارائـه مـی‌دهـد کـه از خـانوار شمارهٔ ۱۰ در شهرک ۵ (شهرک نمونهٔ ۲) جمـع‌آوری شـده اسـت. توجـه کـنید کـه در خانوار سه نفر ٦۵ سال به بالا هستند و دو نفر از اینها به خدمـات پرسـتار متخصص نیاز دارند. از این فرمها ۲۰ تا برای این شهرک مسکونی و نیز برای شهرک مسکونی شمارهٔ ۲ (شهرک نمونهٔ ۱) جمع‌آوری خواهد شد.

خلاصـهٔ اطلاعـات از روی ایـن فـرمهای خـانوار بـرای هر شهرک نمونه در قالبی مشابه آن که در جـدول ۲.۹ نشـان داده شـده اسـت جـدول‌بـندی شـده است. سپس برآوردهای مطلوب را از خلاصه داده‌های نشان داده شده در جدول به شرح زیر به دست می‌آوریم.

شهرک مسکونی: *آپارتمانهای روستای اربابی کینگزوود*

شمارهٔ شهرک نمونه: ۲

شمارهٔ خانوار: ۱۰

نام سرپرست خانوار: *جونز، جان*

همهٔ افراد خانوار را فهرست کرده و با پاسخگو شروع کنید:

آیا این شخص به خدمات پرستار متخصص نیاز دارد؟ (فقط برای بیماران ٦٥ سال به بالا با آری یا نه پاسخ دهید)

شخص	سن	آری	نه
۱. لیدیا جونز	٦٧		×
۲. جان جونز	٦٨	×	
۳. سامانتا جونز	۷٥	×	
٤. ادگار جونز	٦٠		
٥.			
٦.			
۷.			

شکل ۱.۹ فرم جمع‌آوری داده‌ها از خانوارهای نمونه در شهرک مسکونی

برای پیدا کردن برآورد کل تعداد اشخاص ٦٥ سال به بالا، ابتدا متوسط تعداد اشخاص بالاتر از ٦٥ سال را به ازای خوشهٔ نمونه محاسبه می‌کنیم. سپس برآورد کل تعداد اشخاص ٦٥ سال به بالا در جامعه با ضرب کردن متوسط این خوشه در کل تعداد خوشه‌های جامعه به دست می‌آید. یعنی

$$\text{متوسط خوشه} = \frac{۶۷}{۲} = ۳۳/۵$$

$$\text{برآورد کل جامعهٔ اشخاص بالاتر از ٦٥ سال} = ۵ \times ۳۳/۵ = ۱۶۷/۵$$

به راه دیگر می‌توان مجموع نمونه را در نسبت کل خوشه‌های جامعه به کل خوشه‌های نمونه (یعنی $\frac{۵}{۲}$) ضرب کرد. در این صورت برآورد کل جامعهٔ بالاتر از ٦٥ سال عبارت است از

$$\frac{۵}{۲} \times ۶۷ = ۱۶۷/۵$$

برای پیدا کردن برآورد کل تعداد اشخاص بالاتر از ٦٥ سال که نیاز به خدمات پرستار متخصص دارند، مجموع کل نمونه را مانند بالا در $\frac{۵}{۲}$ ضرب می‌کنیم:

$$\text{برآورد کل اشخاص بالاتر از ٦٥ سال که نیاز به پرستار دارند} = \frac{۵}{۲}\times ۲۳ = ۵۷/۵$$

بـرای پیـدا کردن نسبت تمام اشخاص بالاتر از ٦٥ سال که نیاز به خدمات پرستار متخصص دارند، نسبت مجموعهای نمونه‌ای مناسب را به دست می‌آوریم:

$$\text{برآورد نسبت تمام اشخاص بالاتر از ٦٥ سال که نیاز به پرستار دارند} = \frac{۲۳}{۶۷} = ۰/۳۴۳۳$$

بـرای پیدا کردن برآورد متوسط تعداد اشخاص بالاتر از ٦٥ سال در هر شهرک مسکونی، کل تعداد اشخاص بالاتر از ٦٥ سال در نمونه را به کل تعداد شهرکهای مسکونی در نمونه تقسیم می‌کنیم

$$\text{برآورد میانگین تعداد اشخاص بالاتر از ٦٥ سال در هر شهرک} = \frac{۶۷}{۲} = ۳۳/۵$$

بـرای پیـدا کـردن بـرآورد نسـبت همـهٔ خانوارهـای دارای حداقل یک نفر بالاتر از ٦٥ سال که نیاز به خدمـات پرستار متخصص دارد کل تعداد خانوارهای موجود در نمونه را که حداقل دارای یک نفر ٦٥ سال به بالا هستند و نیاز به خدمات پرستار متخصص دارند به تعداد خانوارهای نمونه تقسیم می‌کنیم:

$$\text{برآورد نسبت خانوارهای دارای ۱ یا بیشتر از افراد ٦٥ سال به بالا که نیاز به پرستار دارند} = \frac{۲۱}{۴۰} = ۰/۵۲۵$$

بـرآوردهایی کـه در بـالا نشـان داده شـدند شـامل برآورد مجموعهای جامعه‌ای، نسبتها و میانگینها هستند.

جدول ۲.۹ خلاصه داده‌ها برای دو خوشهٔ انتخاب شده در نمونه

شهرک نمونه	تعداد خانوار	تعداد اشخاص بالاتر از ٦٥ سال	تعداد خانوارهای دارای حداقل یک نفر بالاتر از ٦٥ سال که نیاز به پرستار دارد	تعداد اشخاص بالاتر از ٦٥ سال که نیاز به پرستار دارند
۱- شهرک نمونهٔ ۲	۲۰	۳۳	۱۰	۱۱
۲- شهرک نمونهٔ ۵	۲۰	۳٤	۱۱	۱۲
مجموع	٤۰	٦۷	۲۱	۲۳

□

در نمونـه‌گیری خوشه‌ای، سطح میانگین مشخصهٔ x به ازای خوشه عموماً با $\overline{X}$ نشان داده می‌شود و در مقـابل از $\overline{\overline{X}}$ کـه میـانگین سطح x به ازای واحد فهرست‌برداری را نشان می‌دهد استفاده می‌شود. برای تعمیم، نمادهایی را که در تابلوی ۱.۹ نشان داده شده‌اند به کار می‌بریم.

بـا اسـتفاده از نمـادگذاری تابـلوی ۱.۹، فرمولهای مورد استفاده برای محاسبهٔ برآورد مشخصه‌های جامعـه و بـرآورد خطاهای معیار، برآورد این مشخصه‌ها را فهرست می‌کنیم. توجه کنید که اندیس clu بـرای نشـان دادن ایـن کـه بـرآورد مجمـوع جامعه با نمونه‌گیری خوشه‌ای به دست آمده است به کار مـی‌رود. بـرای مثال x'_{clu} نشان می‌دهد که برآورد مجموع جامعه از نمونه‌گیری خوشه‌ای به دست آمده است.

بـا بررسـی عـبارتهای مـربوط بـه بـرآوردها کـه در تابلوی ۲.۹ ارائه شده‌اند به شباهت زیاد آنها با عـبارتهای مـربوط بـه برآوردهای معادل آنها تحت نمونه‌گیری تصادفی ساده پی می‌بریم که در فصل ۳ بسـط داده شـدند. چـون در نمونـه‌گیـری خوشه‌ای یک‌مرحله‌ای ساده، همهٔ واحدهای فهرست‌برداری داخـل هـر خوشـهٔ نمونـه در نمونـه انـتخاب مـی‌شـوند، تغییرپذیـری نمونـه‌گیری در میان واحدهای فهرسـت‌بـرداری داخـل خوشـه‌هـا عامـلی در تعیین تغییرپذیری نمونه‌گیری برآوردها تحت این برنامهٔ نمونـه‌گیـری بـه شـمار نمـی‌رود. از این رو است که مجموعهای خوشه‌ای، بلوکهای ساختاری مؤثری هسـتند کـه فـرمولهای بـرآورد مـندرج در تابلوی ۲.۹ بر آن مبنا ایجاد شده‌اند. در زیر ، استفاده از این فرمولها برای داده‌های جدول ۲.۹ نشان داده شده است.

مثال تشریحی: داده‌های جدول ۲.۹ را در نظر می‌گیریم. پس داریم

m = ۲ شهرک مسکونی در نمونه

M = ۵ شهرک مسکونی در جامعه

N = ۱۰۰ خانوار در جامعه

$N_۱ = N_۲ =$ ۲۰ خانوار در هر شهرک مسکونی $= \overline{N}$

ابـتدا بـه بـرآورد کـردن میـانگین تعـداد اشـخاص بالاتـر از ۶۵ سال به ازای هر شهرک مسکونی می‌پردازیم که نیاز به خدمات پرستار متخصص دارند. محاسبات مربوط به $\bar{x}_{clu}$ عبارت‌اند از

$$x_۱ = ۱۱ \qquad x_۲ = ۱۲ \qquad x = ۲۳$$

$$\bar{x}_{clu} = \frac{۲۳}{۲} = ۱۱/۵$$

محاسبات مربوط به $\hat{SE}(\bar{x}_{clu})$ عبارت‌اند از

$$\frac{\sum_{i=1}^{m}(x_i-\bar{x}_{clu})^2}{m-1}=\frac{(11-11/5)^2+(12-11/5)^2}{(2-1)}=0/5$$

$$\hat{\sigma}_{1x}=\left[\frac{\sum_{i=1}^{m}(x_i-\bar{x}_{clu})^2}{m-1}\right]^{1/2}\sqrt{\frac{M-1}{M}}$$

$$=(0/5)^{1/2}\left(\frac{5-1}{5}\right)^{1/2}=0/6325$$

$$\hat{SE}(\bar{x}_{clu})=\left(\frac{1}{\sqrt{m}}\right)\hat{\sigma}_{1x}\sqrt{\frac{M-m}{M-1}}$$

$$=\left(\frac{1}{\sqrt{2}}\right)\times 0/6325\sqrt{\frac{5-2}{5-1}}=0/3873$$

یک بازهٔ اطمینان ۹۵ درصدی برای $\bar{X}$ چنین است

$$\bar{x}_{clu}-1/96\times\hat{SE}(\bar{x}_{clu})\le\bar{X}\le\bar{x}_{clu}+1/96\times\hat{SE}(\bar{x}_{clu})$$

$$11/5-1/96\times 0/3873\le\bar{X}\le 11/5+1/96\times 0/3873$$

$$10/74\le\bar{X}\le 12/26$$

اینک به برآورد کردن کل تعداد اشخاص بالاتر از ۶۵ سال که به خدمات یک پرستار متخصص نیاز دارند می‌پردازیم. محاسبهٔ x'_{clu} عبارت است از

$$x'_{clu}=\left(\frac{M}{m}\right)x=\left(\frac{5}{2}\right)\times 23=57/5$$

محاسبهٔ $\hat{SE}(x'_{clu})$ چنین است

$$\hat{SE}(x'_{clu})=\left(\frac{M}{\sqrt{m}}\right)\hat{\sigma}_{1x}\sqrt{\frac{M-m}{M-1}}=\left(\frac{5}{\sqrt{2}}\right)\times 0/6325\sqrt{\frac{5-2}{5-1}}=1/94$$

بازهٔ اطمینان ۹۵ درصدی برای X عبارت است از

$$x'_{clu}-1/96\times\hat{SE}(x'_{clu})\le X\le x'_{clu}+1/96\times\hat{SE}(x'_{clu})$$

$$57/5-1/96\times 1/94\le X\le 57/5+1/96\times 1/94$$

$$53/7\le X\le 61/3$$

حال به برآورد کردن میانگین تعداد اشخاص بالاتر از ۶۵ سال که به خدمات پرستار متخصص نیاز دارند به ازای هر خانوار می‌پردازیم. محاسبهٔ $\bar{\bar{x}}_{clu}$ چنین است

$$\bar{\bar{x}}_{clu}=\frac{x}{m\bar{N}}=\frac{23}{2\times 20}=0/5750$$

تابلوی ۱.۹ نمادگذاری مورد استفاده در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

نمادهای کلی

M = تعداد خوشه‌ها در جامعه

m = تعداد خوشه‌ها در نمونه

x_{ij} = سطح مشخصهٔ x برای واحد فهرست‌برداری نمونه‌ای jدر خوشهٔ نمونه‌ای i

y_{ij} = سطح مشخصهٔ y برای واحد فهرست‌برداری نمونه‌ای jدر خوشهٔ نمونه‌ای i

N_i = تعداد واحدهای فهرست‌برداری در خوشهٔ نمونه‌ای i

$x_i = \sum_{j=1}^{N_i} x_{ij}$ = مجموعهٔ مشخصهٔ x برا ی i امین خوشهٔ نمونه

$y_i = \sum_{j=1}^{N_i} y_{ij}$ = مجموعهٔ مشخصهٔ y برا ی i امین خوشهٔ نمونه

$x = \sum_{i=1}^{m} x_i$ = مجموع نمونه‌ای برای مشخصهٔ x

$y = \sum_{i=1}^{m} y_i$ = مجموع نمونه‌ای برای مشخصهٔ y

N = کل تعداد واحدهای فهرست‌برداری در جامعه

$\overline{N} = N/M$ = متوسط تعداد واحدهای فهرست‌برداری به ازای خوشه در جامعه

مشخصه‌های جامعه

X_{ij} = سطح مشخصهٔ x برای واحد فهرست‌برداری جامعه‌ای jدر خوشهٔ جامعه‌ای i

Y_{ij} = سطح مشخصهٔ y برای واحد فهرست‌برداری جامعه‌ای jدر خوشهٔ جامعه‌ای i

$X_i = \sum_{j=1}^{N_i} X_{ij}$ = مجموع مشخصهٔ x برای i امین خوشهٔ جامعه‌ای

$Y_i = \sum_{j=1}^{N_i} Y_{ij}$ = مجموع مشخصهٔ y برای i امین خوشهٔ جامعه‌ای

$X = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}$ = مجموع جامعه برای مشخصهٔ x

$\overline{X} = X/M$ = سطح میانگین x به ازای خوشه

$Y = \sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij}$ = مجموع جامعه برای مشخصهٔ y

$\overline{Y} = Y/M$ = سطح میانگین y به ازای خوشه

$\overline{\overline{X}} = X/N$ = سطح میانگین x به ازای واحد فهرست‌برداری

$\overline{\overline{Y}} = Y/N$ = سطح میانگین y به ازای واحد فهرست‌برداری

$\sigma_{1x}^2 = \sum_{i=1}^{M} (X_i - \overline{X})^2 / M$ = واریانس توزیع x در همهٔ خوشه‌ها

$\sigma_{1y}^2 = \sum_{i=1}^{M} (Y_i - \overline{Y})^2 / M$ = واریانس توزیع y در همهٔ خوشه‌ها

$\sigma_{1xy} = \sum_{i=1}^{M} (X_i - \overline{X})(Y_i - \overline{Y}) / M$ = کوواریانس مجموعهای خوشه‌ای X_i و Y_i در همهٔ خوشه‌ها

تابلوی ۲.۹ برآورد مشخصه‌های جامعه و برآورد خطاهای معیار برای نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

مجموع، X

$$x'_{clu}=\left(\frac{M}{m}\right)x \qquad \hat{SE}(x'_{clu})=\left(\frac{M}{\sqrt{m}}\right)\hat{\sigma}_{۱x}\sqrt{\frac{M-m}{M-۱}}$$

میانگین به ازای خوشه، $\bar{X}$

$$\bar{x}_{clu}=\frac{x}{m} \qquad \hat{SE}(\bar{x}_{clu})=\left(\frac{۱}{\sqrt{m}}\right)\hat{\sigma}_{۱x}\sqrt{\frac{M-m}{M-۱}}$$

میانگین به ازای واحد فهرست‌برداری، $\bar{\bar{X}}$

$$\bar{\bar{x}}_{clu}=\frac{x}{m\bar{N}} \qquad \hat{SE}\left(\bar{\bar{x}}_{clu}\right)=\left(\frac{۱}{\sqrt{m}\bar{N}}\right)\hat{\sigma}_{۱x}\sqrt{\frac{M-m}{M-۱}}$$

نسبت، $R=\dfrac{X}{Y}$

$$r_{clu}=\frac{x}{y}$$

$$\hat{SE}(r_{clu})=r_{clu}\left\{\frac{[\hat{SE}(\bar{x}_{clu})]^{۲}}{\bar{x}^{۲}_{clu}}+\frac{[\hat{SE}(\bar{y}_{clu})]^{۲}}{\bar{y}^{۲}_{clu}}-\frac{۲}{m}\times\left(\frac{M-m}{M}\right)\times\left(\frac{۱}{\bar{x}_{clu}\bar{y}_{clu}}\right)\right.$$
$$\left.\times\left[\frac{\sum_{i=۱}^{m}(x_i-\bar{x}_{clu})(y_i-\bar{y}_{clu})}{m-۱}\right]\right\}^{½}$$

که در آن $\hat{\sigma}_{۱x}$ به صورت زیر تعریف می‌شود

$$\hat{\sigma}_{۱x}=\left[\frac{\sum_{i=۱}^{m}(x_i-\bar{x}_{clu})^{۲}}{m-۱}\right]^{½}\sqrt{\frac{M-۱}{M}}$$

و $\hat{\sigma}_{۱y}$ نیز به همان صورت تعریف می‌شود. همهٔ نمادهای دیگر در تابلوی ۱.۹ تعریف شده‌اند.

محاسبهٔ $\hat{SE}\left(\bar{\bar{x}}_{clu}\right)$ به صورت زیر است

$$\hat{SE}\left(\bar{\bar{x}}_{clu}\right)=\left(\frac{۱}{\sqrt{m}\bar{N}}\right)\hat{\sigma}_{۱x}\sqrt{\frac{M-m}{M-۱}}$$
$$=\left(\frac{۱}{\sqrt{۲}(۲۰)}\right)\times ۰/۶۳۲۵\sqrt{\frac{۵-۲}{۵-۱}}=۰/۰۱۹۴$$

بازهٔ اطمینان ۹۵ درصدی برای $\bar{\bar{X}}$ عبارت است از

$$\bar{\bar{x}}_{clu}-۱/۹۶\times\hat{SE}\left(\bar{\bar{x}}_{clu}\right)\le\bar{\bar{X}}\le\bar{\bar{x}}_{clu}+۱/۹۶\times\hat{SE}\left(\bar{\bar{x}}_{clu}\right)$$
$$۰/۵۷۵۰-۱/۹۶\times ۰/۰۱۹۴\bar{\bar{X}}\le ۰/۵۷۵۰+۱/۹۶\times ۰/۰۱۹۴$$
$$۰/۵۳۷۰\le\bar{\bar{X}}\le ۰/۶۱۳۰$$

اینک به برآورد کردن میانگین تعداد اشخاص بالاتر از ۶۵ سال به ازای هر شهرک مسکونی می‌پردازیم. محاسبات مربوط به $\bar{y}_{clu}$ عبارت‌اند از

$$y_۱=۳۳ \qquad y_۲=۳۴ \qquad y=۶۷$$

$$\bar{y}_{clu}=\frac{y}{m}=\frac{۶۷}{۲}=۳۳/۵$$

محاسبات مربوط به $\hat{SE}(\bar{y}_{clu})$ عبارت‌اند از

$$\frac{\sum_{i=۱}^{m}(y_i-\bar{y}_{clu})^۲}{m-۱}=\frac{(۳۳-۳۳/۵)^۲+(۳۴-۳۳/۵)^۲}{۲-۱}=۰/۵$$

$$\hat{\sigma}_{۱y}=\left(\frac{\sum_{i=۱}^{m}(y_i-\bar{y}_{clu})^۲}{m-۱}\right)^{۱/۲}\times\sqrt{\frac{M-۱}{M}}$$
$$=(۰/۵)^{۱/۲}\sqrt{\frac{۵-۱}{۵}}=۰/۶۳۲۵$$

$$\hat{SE}(\bar{y}_{clu})=\left(\frac{\hat{\sigma}_{۱y}}{\sqrt{m}}\right)\times\sqrt{\frac{M-m}{M-۱}}=\left(\frac{۰/۶۳۲۵}{\sqrt{۲}}\right)\times\sqrt{\frac{۵-۲}{۵-۱}}=۰/۳۸۷۳$$

بازهٔ اطمینان ۹۵ درصدی برای $\bar{Y}$ چنین است

$$\bar{y}_{clu}-۱/۹۶\times\hat{SE}(\bar{y}_{clu})\le\bar{Y}\le\bar{y}_{clu}+۱/۹۶\times\hat{SE}(\bar{y}_{clu})$$
$$۳۳/۵-۱/۹۶\times ۰/۳۸۷۳\le\bar{Y}\le ۳۳/۵+۱/۹۶\times ۰/۳۸۷۳$$
$$۳۲/۷\le\bar{Y}\le ۳۴/۳$$

بالاخـره به برآورد کردن نسبت اشخاص ٦٥ سال به بالا که نیاز به خدمات پرستار متخصص دارند می‌پردازیم. محاسبهٔ r_{clu} عبارت است از

$$r_{clu} = \frac{x}{y} = \frac{۲۳}{۶۷} = ۰/۳۴۳۳$$

محاسبات مربوط به $\hat{SE}(r_{clu})$ عبارت‌اند از

$$\frac{\sum_{i=۱}^{m}(x_i - \bar{x}_{clu})(y_i - \bar{y}_{clu})}{m-۱} = \frac{(۱۱-۱۱/۵)(۳۳-۳۳/۵)+(۱۲-۱۱/۵)(۳۴-۳۳/۵)}{۲-۱} = ۰/۵$$

$$\hat{SE}(r_{clu}) = r_{clu}\left\{\frac{[\hat{SE}(\bar{x}_{clu})]^۲}{\bar{x}_{clu}^۲} + \frac{[\hat{SE}(\bar{y}_{clu})]^۲}{\bar{y}_{clu}^۲} - \frac{۲}{m}\times\left(\frac{M-m}{M}\right)\times\left(\frac{۱}{\bar{x}_{clu}\bar{y}_{clu}}\right)\times\left[\frac{\sum_{i=۱}^{m}(x_i-\bar{x}_{clu})(y_i-\bar{y}_{clu})}{m-۱}\right]\right\}^{۱/۲}$$

$$= ۰/۳۴۳۳ \times \left\{\frac{۰/۳۸۷۳^۲}{۱۱/۵^۲} + \frac{۰/۳۸۷۳^۲}{۳۳/۵^۲} - \frac{۲}{۲}\times\frac{۵-۲}{۵}\times\frac{۱}{۱۱/۵\times۳۳/۵}\times۰/۵\right\}^{۱/۲}$$

$$= ۰/۰۰۷۶$$

بازهٔ اطمینان ۹۵ درصدی برای R چنین است

$$r_{clu} - ۱/۹۶\times\hat{SE}(r_{clu}) \le R \le r_{clu} + ۱/۹۶\times\hat{SE}(r_{clu})$$
$$۰/۳۴۳۳ - ۱/۹۶\times۰/۰۰۷۶ \le R \le ۰/۳۴۳۳ + ۱/۹۶\times۰/۰۰۷۶$$
$$۰/۳۲۸۴ \le R \le ۰/۳۵۸۲$$

□

اسـتفاده از نـرم‌افزار آماری برای برآورد کردن. چگونگی اجرای برآورد کردن را با استفاده از هر دو نـرم‌افـزار STATA و SUDAAN برای مثال تشریحی قبل نشـان خواهیـم داد و ابتـدا به استفـاده از STATA خواهیـم پرداخت. پروندهٔ اطلاعاتی STATA یعنی TAB9_1A.DTA شامل ٤٠ گـزارش (۲۰ خانوار در هر یک از ۲ شهرک مسکونی) به صورتی است که در زیر نشان داده شده است:

توجه کنید که هر گزارش متناظر با یک خانوار نمونهٔ خاص است و شهرک مسکونی که خانوار در آن ساکن است و همچنین وزن نمونه‌گیری برای هر گزارش نشان داده شده است.

devlpmnt	hh	wt1	M	Nhh	nge65	nvstnrs	hhneedvn	nge65dv
2	1	2.5	5	20	2	1	1	40
2	2	2.5	5	20	1	0	0	20
2	3	2.5	5	20	2	0	0	40
2	4	2.5	5	20	1	1	1	20
2	5	2.5	5	20	1	0	0	20
2	6	2.5	5	20	1	0	0	20
2	7	2.5	5	20	2	1	1	40
2	8	2.5	5	20	1	1	1	20
2	9	2.5	5	20	3	1	1	60
2	10	2.5	5	20	1	1	1	20
2	11	2.5	5	20	1	0	0	20
2	12	2.5	5	20	2	1	1	40
2	13	2.5	5	20	1	0	0	20
2	14	2.5	5	20	3	1	1	60
2	15	2.5	5	20	1	0	0	20
2	16	2.5	5	20	3	2	1	60
2	17	2.5	5	20	1	0	0	20
2	18	2.5	5	20	2	0	0	40
2	19	2.5	5	20	1	1	1	20
2	20	2.5	5	20	3	0	0	60
5	1	2.5	5	20	1	1	1	20
5	2	2.5	5	20	1	0	0	20
5	3	2.5	5	20	1	1	1	20
5	4	2.5	5	20	3	1	1	60
5	5	2.5	5	20	2	0	0	40
5	6	2.5	5	20	1	1	1	20
5	7	2.5	5	20	3	0	0	60
5	8	2.5	5	20	1	0	0	20
5	9	2.5	5	20	1	0	0	20
5	10	2.5	5	20	3	2	1	60
5	11	2.5	5	20	2	1	1	40
5	12	2.5	5	20	3	0	0	60
5	13	2.5	5	20	1	1	1	20
5	14	2.5	5	20	1	0	0	20
5	15	2.5	5	20	2	1	1	40
5	16	2.5	5	20	2	1	1	40
5	17	2.5	5	20	1	0	0	20
5	18	2.5	5	20	2	0	0	40
5	19	2.5	5	20	2	1	1	40
5	20	2.5	5	20	1	1	1	20

هر گزارش شامل متغیرهای زیر است

۱. *devlpmnt* ــ شهرک مسکونی (۲ شهرک انتخاب شده از ۵ شهرک جامعه).
۲. *hh* ــ شمارهٔ خانوار در داخل شهرک.
۳. *wt1* ــ وزن نمونه‌گیری (که برابر است با $\frac{M}{m} = \frac{۵}{۲} = ۲/۵$ برای همهٔ خانوارها).
٤. M ــ تعداد شهرکهای مسکونی در جامعه (= ۵).
۵. *Nhh* ــ تعداد خانوارها در داخل هر شهرک مسکونی (= ۲۰).
٦. *nge65* ــ تعداد اشخاص بالاتر از ٦۵ سال در هر خانوار.
۷. *nvstnrs* ــ تعداد اشخاص بالاتر از ٦۵ سال در هر خانوار که نیاز به خدمات پرستار متخصص دارند.
۸. *hhneedvn* ــ مجموعهٔ متغیر ظاهری برابر با «۱» در صورتی که حداقل یک نفر بالاتر از ٦۵ سال در خانوار باشد که نیاز به خدمات پرستار متخصص دارد و «۰» در غیر این صورت.
۹. *nge65dv* ــ این برابر است با 20**nge65* که متغیر تبدیل یافته‌ای است برای برآورد میانگین تعداد اشخاص ٦۵ سال به بالا به ازای هر شهرک.

مجموعه فرمانهای زیر می‌تواند برای انجام برآورد کردن موردنظر در STATA به کار رود.

```
 use " a:\tab9_1a.dta", clear
. svyset  pweight  wt1
. svyset  fpc  M
. svyset  psu  devlpmnt
. svytotal  nvstnrs  nge65
. svymean  nvstnrs  hhneedvn
. svyratio  nvstnrs  nge65
. svymean  nge65dv
```

این فرمانها خروجی زیر را تولید می‌کنند (که مانند یک متن معمولی بر آن حاشیه‌نویسی می‌کنیم).

```
use " a:\tab9_1a.dta", clear
. svyset pweight wt1
. svyset fpc M
. svyset psu devlpmnt
```

سه فرمان بالا، پارامترهای طرح نمونه‌ای موردنیاز را تعیین می‌کنند. وزن نمونه‌گیری با متغیر *wt1* نشان داده شده و برابر با ۲/۵ برای هر خانوار است. تصحیح جامعهٔ متناهی (fpc) براساس تعداد خوشه‌ها ($M = ۵$) در جامعه محاسبه شده است. واحدهای نمونه‌گیری اولیه (یا خوشه‌ها) در متغیر *devlpmnt* تعریف شده‌اند.

```
. svytotal nvstnrs nge65
```

فرمان بالا، کل تعداد اشخاص ٦٥ سال به بالا را در ۵ شهرک و کل تعدادی را که نیاز به خدمات پرستار متخصص دارند برآورد می‌کند. خروجی این فرمان بلافاصله در زیر نشان داده می‌شود.

```
. survey total estimation
pweight:  wt1                    Number of obs    =  40
Strata:   <one>                  Number of strata =   1
PSU:      devlpmnt               Number of PSUs   =   2
FPC:      M                      Population size  = 100
```

Total	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
Nvstnrs	57.5	1.936492	32.89454	82.10546	.0707804
nge65	167.5	1.936492	142.8945	192.1055	.0393542

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

توجه کنید که برای برآورد فوق و همهٔ برآوردهای دیگر، برآوردها و خطاهای معیار با آنهایی که از فرمولهای مناسب نشان داده شده در متن کتاب به دست می‌آیند یکسان‌اند. ولی بازه‌های اطمینان از آنچه در متن نشان داده شده است و سیعترند زیرا بازه‌های اطمینان STATA براساس توزیع t ی استیودنت هستند، در حالی که آنچه در متن آمده بر مبنای توزیع نرمال است. چون در این مثال تشریحی فقط ۲ خوشهٔ نمونه موجودند و لذا تنها ۱ درجهٔ آزادی وجود دارد، تفاوتها نسبت به پهنای بازهٔ اطمینان فاحش است.

```
. svymean nvstnrs hhneedvn
```

فرمان بالا به برآورد کردن کل تعداد اشخاص ٦٥ سال یا بالاتر در ۵ شهرک منتهی می‌شود که نیاز به خدمات پرستار متخصص دارند. همچنین کل تعداد خانوارهای موجود در ۵ شهرک را که دارای یک یا چند نفر ٦٥ سال یا بالاترند که نیاز به خدمات پرستار متخصص دارند برآورد می‌کند. خروجی حاصل در زیر نشان داده شده است:

Survey mean estimation

pweight: wt1 Number of obs = 40
Strata: <one> Number of strata = 1
PSU: devlpmnt Number of PSUs = 2
FPC: M Population size = 100

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
nvstnrs	.575	.0193649	.3289454	.8210546	.0707804
hhneedvn	.525	.0193649	.2789454	.7710546	.0977444

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

. svyratio nvstnrs nge65

فرمان بالا به محاسبهٔ یک برآوردگر نسبتی منجر می‌شود که نسبت اشخاص ٦٥ سال یا بالاتر را که نیاز به خدمات پرستار متخصص دارند برآورد می‌کند. خروجی آن در زیر نشان داده شده است:

Survey ratio estimation

pweight: wt1 Number of obs = 40
Strata: <one> Number of strata = 1
PSU: devlpmnt Number of PSUs = 2
FPC: M Population size = 100

Ratio	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
nvstnrs/nge65	.3432836	.0075924	.2468131	.4397541	.0325067

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

. svymean nge65dv

فرمان بالا به برآورد کردن میانگین تعداد افراد ٦٥ سال یا بالاتر به ازای هر شهرک مسکونی منجر می‌شود. خروجی آن در زیر آمده است:

Survey mean estimation

pweight: wt1 Number of obs = 40
Strata: <one> Number of strata = 1
PSU: devlpmnt Number of PSUs = 2
FPC: M Population size = 100

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
nge65dv	33.5	.3872983	28.57891	38.42109	.0393542

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه انجام می‌گیرد.

اینک نشان خواهیم داد که SUDAAN چگونه می‌تواند برآوردهایی برای مجموعه داده‌های *tab9_1c.SSd* در SAS فراهم کند که دارای همان داده‌ها و متغیرهای مجموعه داده‌های *tab9_1c.dta* در STATA است که در توضیح بالا به کار رفتند.

مجموعه فرمانهای زیر، برآوردهایی از SUDAAN برای کل تعداد اشخاص ٦٥ سال و بالاتر، کل تعداد اشخاص ٦٥ سال و بالاتر که نیاز به خدمات پرستار متخصص دارند، کل تعداد خانوارهای دارای یک یا چند نفر نیازمند به خدمات پرستار متخصص، و نسبت همهٔ اشخاص ٦٥ سال و بالاتر نیازمند به خدمات پرستار متخصص فراهم خواهند کرد.

```
proc descript data = tab9_1c   filetype = sas   design = WOR   means totals;
nest_one_devlpmnt;
totcnt M_zero_;
weight wt1;
var nge65 nvstnrs hhneedvn;
setenv colwidth = 13   decwidth = 5;
proc ratio data = tab9_1c   filetype = sas   design = WOR;
nest_one_devlpmnt;
totcnt M_zero_;
weight wt1;
number nvstnrs;
denom nge65;
setenv colwidth = 13   decwidth = 5;
```

این فرمانها، خروجی نشان داده شده در صفحهٔ بعد را تولید خواهند کرد.

مثال تشریحی: مثال زیر در اینجا به عنوان مثال دیگری از برآورد کردن نسبتی، تحت نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده ارائه می‌شود. این مثال برای نشان دادن سایر مفاهیم نیز در بخشهای دیگر این فصل مورد استفاده قرار خواهد گرفت. محاسبات را با استفاده از شیوهٔ برآورد کردن نسبتی در SUDAAN اجرا خواهیم کرد.

یک ناحیهٔ خاص که دارای ٢٦ دادگاه فدرال است مایل است یک آمارگیری نمونه‌ای از این دادگاهها به عمل آورد به این منظور که برخی اطلاعات معین در مورد افرادی به دست آید که به آنها آزادی مشروط همراه با توصیه‌نامه‌ای برای درمان اعتیاد داده می‌شود. یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ١٠ تایی از این دادگاههای فدرال گرفته‌اند و در داخل هر یک از دادگاههای انتخاب شده، سوابق مربوط به همهٔ پرونده‌های رسیدگی شده در طول سال گذشته را که منجر به آزادی مشروط همراه با توصیه‌نامه برای درمان اعتیاد شده بودند شناسایی شدند. برای هر مورد شناسایی شده، اطلاعاتی جمع‌آوری شد که آیا شخص موردنظر واقعاً تحت درمان اعتیاد قرار گرفته است یا نه. مهمترین هدف، آن است که نسبت افرادی که واقعاً تحت درمان قرار گرفته‌اند برآورد شود.

پیـش از آمارگیـری، هیـچ فهرستی موجود نبود که افراد واجد شرایط برای این بررسی را مشخص کـند. بـنابراین، بـرای هـر دادگـاه فـدرال نمونه‌گیری شده باید موارد واجد شرایط با غربال کردن همهٔ پـرونده‌هـای دادگـاه کـه مـنجر بـه آزادی مشـروط در طول دورهٔ بررسی شده بودند شناسایی می‌شد. داده‌های زیر از نمونهٔ متشکل از ۱۰ دادگاه فدرال به دست آمدند.

```
1   PROC DESCRIPT DATA=TAB9_1C FILETYPE=SAS DESIGN=WOR MEANS
    TOTALS;
2   NEST_ONE_DEVLPMNT;
3   TOTCNT M_ZERO_;
4   WEIGHT WT1;
5   VAR NGE65 NVSTNRS HHNEEDVN;
6   SETENV COLWIDTH=13 DECWIDTH=5;
```

Number of observations read : 40 Weighted count : 100
Number of observations skipped : 0
(WEIGHT variable nonpositive)
Denominator degrees freedom : 1

Table : 1

by: Variable, One.

Variable		One 1
NGE65	Sample Size	40.00000
	Weighted Size	100.00000
	Total	167.50000
	SE Total	1.93649
	Mean	1.67500
	SE Mean	0.01936
NVSTNRS	Sample Size	40.00000
	Weighted Size	100.00000
	Total	57.50000
	SE Total	1.93649
	Mean	0.57500
	SE Mean	0.01936
HHNEEDVN	Sample Size	40.00000
	Weighted Size	100.00000
	Total	52.50000
	SE Total	1.93649
	Mean	0.52500
	SE Mean	0.01936

```
7   PROC RATIO DATA=TAB9_1C FILETYPE=SAS DESIGN=WOR;
8   NEST_ONE_DEVLPMNT;
9   TOTCNT M_ZERO_;
10  WEIGHT WT1;
11  NUMBER NVSTNRS;
12  DENOM NGE65;
13  SETENV COLWIDTH=13 DECWIDTH=5;
```

Number of observations read : 40 Weighted count : 100
Number of observations skipped : 0
(WEIGHT variable nonpositive)
Denominator degrees freedom : 1

by: Variable, One.

Variable		One 1
NVSTNRS	Sample Size	40.00000
	Weighted Size	100.00000
	Weighted X-Sum	167.50000
	Weighted Y-Sum	57.50000
	Ratio Est.	0.34328
	SE Ratio	0.00759

داده‌های حاصل از ۱۰ دادگاه فدرال

دادگاه فدرال	تعداد افراد واجد شرایط	تعدادی که تحت درمان اعتیاد قرار گرفته‌اند
٦	٤٨٦	٧٩
١٠	٢٤٠	٩٤
١٤	٤٢٨	١٧
١٥	٣٤٣	٥٧
١٧	١١٣٠	٦٣
١٩	٩٨٣	١٠
٢٠	٣٣٣	٥٨
٢١	١٣	٠
٢٢	١٥٠٦	١٠١
٢٥	١٧٥٥	٤١١
مجموع	٧٢١٧	٨٩٠

چون تعداد افراد واجد شرایط در نمونه در معرض تغییرپذیری نمونه‌گیری قرار دارند، برآورد نسبت حاصل برای آنهایی که تحت درمان قرار گرفته‌اند یک برآورد نسبتی به صورتی است که در تابلوی ۲.۹ نشان داده شده است. با استفاده از آن نمادگذاری، خلاصه آماره‌ها و برآوردهای زیر را خواهیم داشت:

$$M = ۲۶ \qquad m = ۱۰$$

$$\bar{x}_{clu} = ۸۹ \qquad s_x = ۱۱۸/۳۰۳ \qquad \hat{SE}(\bar{x}_{clu}) = ۲۹/۳۴۷۳$$

$$\bar{y}_{clu} = ۷۲۱/۷۰۰ \qquad s_y = ۵۸۵/۷۹۷ \qquad \hat{SE}(\bar{y}_{clu}) = ۱۴۵/۳۱۸۵$$

$$\frac{\sum_{i=۱}^{m}(x_i - \bar{x}_{clu})(y_i - \bar{y}_{clu})}{m-۱} = ۴۶۵۷۹/۲۲۲$$

$$r_{clu} = \frac{۸۹}{۷۲۱/۷۰۰} = ۰/۱۲۳۳$$

$$\hat{SE}(r_{clu}) = ۰/۰۳۰۲$$

به این ترتیب برآورد می‌شود که ۳/۰۲٪ ± ۱۲/۳۳٪ افراد واجد شرایط تحت درمان اعتیاد قرار گرفته‌اند.

می‌توان این داده‌ها را با استفاده از SUDAAN از روی پرونده‌ای که در آن، خوشه، سابقهٔ اصلی را تشکیل می‌دهد به صورت زیر تحلیل کرد:

سابقه	بخش	واجد شرایط	درمان شده	وزن	تعداد
۱	٦	٤٨٦	۷۹	۲/٦	۲٦
۲	۱۰	۲٤۰	۹٤	۲/٦	۲٦
۳	۱٤	٤۲۸	۱۷	۲/٦	۲٦
٤	۱۵	۳٤۳	۵۷	۲/٦	۲٦
۵	۱۷	۱۱۳۰	٦۳	۲/٦	۲٦
٦	۱۹	۹۸۳	۱۰	۲/٦	۲٦
۷	۲۰	۳۳۳	۵۸	۲/٦	۲٦
۸	۲۱	۱۳	۰	۲/٦	۲٦
۹	۲۲	۱۵۰٦	۱۰۱	۲/٦	۲٦
۱۰	۲۵	۱۷۵۵	٤۱۱	۲/٦	۲٦

پروندهٔ فرمان برای به دست آوردن برآورد نسبتی، r_{clu} ، و برآورد خطای معیار آن در زیر نشان داده شده است:

```
PROC RATIO DATA = PROBBKSM    FILETYPE = SAS    DESIGN = STRWOR;
NEST_ONE_;
TOTCNT  N;
WEIGHT  W;
NUMBER  TREATED;
DENOM   ELIGIBLE;
SETENV  COLWIDTH = 13;
SETENV  DECWIDTH = 3;
```

سطر اول: ضرورت برآورد نسبتی را بیان می‌کند، نام پروندهٔ اطلاعاتی را مشخص می‌کند، نشان می‌دهد که این اطلاعات در پروندهٔ SAS است و طرح SUDAAN را تعیین می‌کند که STRWOR است.

سطر دوم: _NEST_ONE، نشان می‌دهد که فقط یک طبقه وجود دارد. اثر خالص حکم مربوط به طرح در ترکیب با حکم آشیانه، آن است که یک مرحلهٔ نمونه‌گیری و یک طبقه وجود دارد و نمونه‌گیری بدون جایگذاری است.

سطر سوم: نشان می‌دهد که کل تعداد سوابق در جامعه با متغیر N نشان داده می‌شود که در این مورد ۲٦ است.

سطر چهارم: مشخص می‌کند که وزن نمونه‌گیری با متغیر W نشان داده می‌شود که در این مورد ۲/٦ یا $\left(\frac{۲۶}{۱۰}\right)$ است. سه حکم آخر متغیرهای صورت و مخرج کسر و ظاهر خروجی را تعیین می‌کنند. توجه کنید که این پروندهٔ فرمان دقیقاً همان است که برای نمونهٔ تصادفی سادهٔ بدون جایگذاری به کار می‌رفت. ولی، در این مورد، متغیرهایی که پردازش می‌شوند به جای مقدارهای عناصر تکی، مجموعهای خوشه‌ای هستند.

توجه کنید که برای ساختن پروندهٔ اطلاعاتی، همان طور که در مثال تشریحی قبلی انجام شد، می‌توان از واحدهای شمارش (که در این مثال، موارد واجد شرایط هستند) به عنوان سوابق پایه استفاده کرد. ولی این عمل مستلزم کار کردن با یک مجموعه داده‌ها با ۷۲۱۷ سابقه است که کمی پرزحمت خواهد بود. خروجی SUDAAN برای این مثال ذیلاً نشان داده می‌شود:

Variable		One 1
TREATED/ELIGIBLE	Sample Size	10.000
	Weighted Size	26.000
	Weighted X-Sum	18764.199
	Weighted Y-Sum	2314.000
	Ratio Est.	0.123
	SE Ratio	0.030

توجه کنید که نسبت برآورد شده و خطای معیار آن با آنچه که با استفاده از فرمولهای تابلوی ۲.۹ به دست آمد توافق دارند.

□

۳.۹ توزیعهای نمونه‌گیری برآوردها

در نمونه‌گیری خوشه‌ای غالباً خوشه‌ها خود ساختهایی هستند که برای فراهم آوردن یک برنامهٔ نمونه‌گیری کارا به کار می‌روند، و غالباً برآورد کردن مشخصه‌ها بر مبنای هر خوشه از اهمیت اصلی برخوردار نیست. برای به دست آوردن برآوردها بر مبنای هر عنصر از خوشهٔ نمونه باید از برآوردهای نسبتی استفاده کرد. مثلاً، برای برآورد کردن نسبت همهٔ اشخاص نیازمند به خدمات پرستار متخصص، یک برآورد نسبتی ضروری است، زیرا تعداد افراد در نمونه، و در نتیجه تعداد اشخاص برآورد شده در جامعه، یک متغیر تصادفی است که در معرض تغییرپذیری نمونه‌گیری قرار دارد. به این ترتیب، برآوردهای نسبتی در نمونه‌گیری خوشه‌ای بسیار مهم‌اند.

ابتدا به مثالی از توزیعهای نمونه‌گیری برآوردها نگاه می‌کنیم.

مثال تشریحی: در مثال قبلی که در آن یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده از دو شهرک مسکونی از جامعه‌ای متشکل از پنج شهرک مسکونی انتخاب می‌شود ۱۰ نمونهٔ ممکن وجود دارند که همگی شانس برابر برای انتخاب شدن دارند. توزیعهای نمونه‌گیری سه برآورد ـ یعنی کل تعداد افراد ٦۵ سال

به بالا که نیاز به خدمات پرستار متخصص دارند (x'_{clu})، کل تعداد افراد ٦٥ سال به بالا (y'_{clu})، و نسبت افراد ٦٥ سال به بالا که نیاز به خدمات پرستار متخصص دارند (r_{clu}) ـ در جدول ۳.۹ نشان داده شده‌اند.

میانگین‌ها و خطاهای معیار (x'_{clu})، (y'_{clu}) و (r_{clu}) که از شمارش همهٔ نمونه‌ها به دست آمده‌اند همراه با مقادیر واقعی جامعه (X، Y و R) در جدول ٤.۹ نشان داده شده‌اند.

توجه کنید که در این مثال $E(x'_{clu}) = X$ و $E(y'_{clu}) = Y$ و این به طور کلی دربارهٔ برآورد مجموعها از روی نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده مصداق دارد. همچنین توجه کنید در عین حال که برآورد نسبتی (r_{clu}) یک برآوردگر ناریب برای نسبت جامعه‌ای R نیست، بزرگی این اریبی اندک است.

□

عبارتهای مربوط به خطاهای معیار نظری برآورد مجموعها، میانگینها و نسبتها برای نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده برحسب پارامترهای جامعه در تابلوی ۳.۹ نشان داده شده‌اند. عبارتهای X_i و $\overline{X}$ که در تابلوی ۳.۹ دیده می‌شوند در تابلوی ۱.۹ تعریف شده و به صورت زیرند:

$$X_i = \sum_{j=1}^{N_i} X_{ij}$$

که در آن، X_{ij}، سطح مشخصهٔ x برای jامین واحد فهرست‌برداری در خوشهٔ i است

$$\overline{X} = \frac{\sum_{i=1}^{M} X_i}{M}$$

به عبارت دیگر، X_iها مجموعهای خوشه‌ای هستند و $\overline{X}$ میانگین سطح x به ازای خوشه است. این نمادها مانند فصلهای پیشین هرگاه بحث دربارهٔ مشخصه‌های جامعه باشد با حروف بزرگ و هنگامی که بحث دربارهٔ مشخصه‌های نمونه باشد با حروف کوچک نشان داده می‌شوند.

جدول ۳.۹ توزیع نمونه‌گیری سه برآورد

خوشه‌ها در نمونه	x'_{clu}	y'_{clu}	r_{clu}
۱ , ۲	۷۵/۰	۱٦۲/۵	۰/٤٦۲
۱ , ۳	۸۲/۵	۱۷۷/۵	۰/٤٦٥
۱ , ٤	۵۷/٦	۱٦۰/۰	۰/۳۵۹
۱ , ۵	۷۷/۵	۱٦۵/۰	۰/٤۷۰
۲ , ۳	٦۲/۵	۱۸۰/۰	۰/۳٤۷
۲ , ٤	۳۷/۵	۱٦۲/۵	۰/۲۳۱
۲ , ۵	۵۷/۵	۱٦۷/۵	۰/۳٤۳
۳ , ٤	٤۵/۰	۱۷۷/۵	۰/۲۵٤
۳ , ۵	٦۵/۰	۱۸۲/۵	۰/۳۵٦
٤ , ۵	٤۰/۰	۱٦۵/۰	۰/۲٤۲

جدول ٤.٩ میانگینها و خطاهای معیار برآوردها

برآورد	میانگین برآورد	خطای معیار برآورد	مقدار جامعه
x'_{clu}	٦٠	١٤/٨٧	X=٦٠
y'_{clu}	١٧٠	٧/٩٨	Y=١٧٠
r_{clu}	٠/٣٥٢٨٧	٠/٠٨٧	R=٠/٣٥٢٩٤

استفاده از این فرمولها برای جامعهٔ متشکل از ۵ شهرک مسکونی که در جدول ۱.۹ ارائه شد پس از این نشان داده می‌شود.

مثال تشریحی: برای جامعهٔ ارائه شده در جدول ۱.۹، داریم

$$M=۵ \qquad Y_۱=۳۲ \qquad Y_۲=۳۳ \qquad Y_۳=۳۹ \qquad Y_۴=۳۲ \qquad Y_۵=۳۴$$

$$\bar{Y}=۳۴ \qquad \hat{\sigma}^{۲}_{۱y}=\frac{\sum_{i=۱}^{m}(Y_i-\bar{Y})^{۲}}{M}=۶/۸$$

$$X_۱=۱۹ \qquad X_۲=۱۱ \qquad X_۳=۱۴ \qquad X_۴=۴ \qquad X_۵=۱۲$$

$$\bar{X}=۱۲ \qquad \hat{\sigma}^{۲}_{۱x}=\frac{\sum_{i=۱}^{m}(X_i-\bar{X})^{۲}}{M}=۲۳/۶$$

$$\hat{\sigma}_{۱xy}=\frac{\sum_{i=۱}^{m}(X_i-\bar{X})(Y_i-\bar{Y})}{M}=۲/۶ \qquad R=\frac{۱۲}{۳۴}=۰/۳۵۳$$

پس برای نمونه‌های شامل $m=۲$ خوشه، داریم

$$SE(y'_{clu})=\left(\frac{۵}{\sqrt{۲}}\right)\times\sqrt{۶/۸}\times\sqrt{\frac{۵-۲}{۵-۱}}=۷/۹۸$$

$$SE(x'_{clu})=\left(\frac{۵}{\sqrt{۲}}\right)\times\sqrt{۲۳/۶}\times\sqrt{\frac{۵-۲}{۵-۱}}=۱۴/۸۷$$

$$SE(\bar{x}_{clu})=\left(\frac{۱}{\sqrt{۲}}\right)\times\sqrt{۲۳/۶}\times\sqrt{\frac{۵-۲}{۵-۱}}=۲/۹۷$$

$$SE(\bar{\bar{x}}_{clu})=\left(\frac{۱}{۲۰\sqrt{۲}}\right)\times\sqrt{۲۳/۶}\times\sqrt{\frac{۵-۲}{۵-۱}}=۰/۱۴۹$$

$$SE(r_{clu})=\left(\frac{۰/۳۵۳}{\sqrt{۲}}\right)\times\sqrt{\frac{۵-۲}{۵-۱}}\times\sqrt{\frac{۲۳/۶}{(۱۲)^{۲}}+\frac{۶/۸}{(۳۴)^{۲}}-\frac{۲\times۲/۶}{۱۲\times۳۴}}=۰/۰۸۵۷$$

□

توجه کنید که در این مثال، $SE(x'_{clu})$ و $SE(y'_{clu})$ که از فرمولهای ارائه شده در تابلوی ۳.۹ محاسبه شده‌اند دقیقاً همانهایی هستند که قبلاً با شمارش کامل همهٔ نمونه‌ها به دست آمده بودند (جدول ٤.٩). این به طور کلی برای برآورد مجموعها و میانگینهای حاصل از نمونه‌گیری خوشه‌ای ساده مصداق دارد. ولی فرمول مربوط به خطای معیار r_{clu} فقط یک تقریب است.

تابلوی ۳.۹ خطاهای معیار نظری برای برآوردها تحت نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

مجموع، x'_{clu}

$$SE(x'_{clu}) = \left(\frac{M}{\sqrt{m}}\right) \times \left[\frac{\sum_{i=1}^{M}(X_i - \bar{X})^2}{M}\right]^{1/2} \left(\frac{M-m}{M-1}\right)^{1/2}$$

میانگین به ازای خوشه، $\bar{x}_{clu}$

$$SE(\bar{x}_{clu}) = \left(\frac{1}{\sqrt{m}}\right) \times \left[\frac{\sum_{i=1}^{M}(X_i - \bar{X})^2}{M}\right]^{1/2} \left(\frac{M-m}{M-1}\right)^{1/2}$$

میانگین به ازای واحد شمارش، $\bar{\bar{x}}_{clu}$

$$SE(\bar{\bar{x}}_{clu}) = \left(\frac{1}{\bar{N}\sqrt{m}}\right) \times \left[\frac{\sum_{i=1}^{M}(X_i - \bar{X})^2}{M}\right]^{1/2} \left(\frac{M-m}{M-1}\right)^{1/2}$$

نسبت، r_{clu}

$$SE(r_{clu}) \approx \left(\frac{R}{\sqrt{m}}\right) \times \left(\frac{M-m}{M-1}\right)^{1/2} \times \left[\frac{\sum_{i=1}^{M}(X_i - \bar{X})^2}{M\,\bar{X}^2} + \frac{\sum_{i=1}^{M}(Y_i - \bar{Y})^2}{M\,\bar{Y}^2} - \frac{2\sum_{i=1}^{M}(X_i - \bar{X})(Y_i - \bar{Y})}{M\,\bar{X}\,\bar{Y}}\right]^{1/2}$$

نمادگذاری مورد استفاده در اینجا در تابلوی ۱.۹ تعریف شده است. برآوردهای جامعه در تابلوی ۲.۹ تعریف شده‌اند.

عبارت $\frac{\sum_{i=1}^{m}(X_i-\bar{X})^2}{M}$ که در فرمولهای تابلوی ۳.۹ دیده می‌شود در واقع واریانس توزیع مجموعه‌های خوشه، X_i، است. به این عبارت که با نماد σ_{1x}^2 نشان داده می‌شود غالباً به عنوان *واریانس مرحلهٔ اول* اشاره می‌شود و توصیف‌کنندهٔ تغییرات در میان خوشه‌ها نسبت به توزیع سطوح مجموع مشخصهٔ x است. به عبارت دیگر، برای هر مشخصهٔ x

$$\sigma_{1x}^2=\frac{\sum_{i=1}^{m}(X_i-\bar{X})^2}{M} \qquad (1.9)$$

اگر فرمولهای تابلوی ۳.۹ را بررسی کنیم، با جایگزین کردن رابطهٔ (۱.۹) در جاهای مناسب، می‌بینیم که این عبارتها دقیقاً همان عبارتهای مربوط به خطاهای معیار برآوردهای مشابه در نمونه‌گیری تصادفی ساده‌اند با این تفاوت که m به جای n، M به جای N و σ_{1x}^2 به جای σ_x^2 قرار گرفته است. این نتیجه تعجب‌آور نیست، زیرا نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده دقیقاً همان نمونه‌گیری تصادفی ساده است که در آن مجموع همهٔ واحدهای فهرست‌برداری در خوشه به صورت یک واحد شمارش مؤثر به کار می‌رود.

برآورد خطاهای معیار برای میانگینها، مجموعها و نسبتهای برآورد شده که در تابلوی ۲.۹ نشان داده شدند از خطاهای معیار نظری این برآوردها که در تابلوی ۳.۹ نشان داده شده‌اند با جایگزین کردن $\hat{\sigma}_{1x}^2$ به جای σ_{1x}^2 در جاهای مناسب به دست می‌آیند، که $\hat{\sigma}_{1x}^2$، برآورد σ_{1x}^2 از روی داده‌هاست که از فرمول زیر به صورتی که در تابلوی ۲.۹ تعریف شده است به دست می‌آید

$$\hat{\sigma}_{1x}^2=\left[\frac{\sum_{i=1}^{m}(x_i-\bar{x}_{clu})^2}{m-1}\right]\times\left(\frac{M-1}{M}\right)$$

۴.۹ نمونهٔ مورد نیاز چقدر باید بزرگ باشد؟

فرض کنید در صدد استفاده از یک طرح نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده هستیم و می‌خواهیم واقعاً مطمئن شویم برآوردهایی که به دست می‌آوریم به طور نسبی بیش از ε با مقادیر واقعی پارامترهای نامعلوم تفاوت ندارند (که ε به صورتی است که در فصل ۳ تعریف شده است). در این صورت تعداد خوشه‌های نمونهٔ موردنیاز همان است که در تابلوی ۴.۹ ارائه شده است. به یک محاسبه نظری بیندازیم که در آن، از این فرمولها استفاده شده است.

مثال تشریحی: در مثال مبتنی بر داده‌های جدول ۱.۹، فرض کنید می‌خواهیم واقعاً مطمئن شویم (یعنی $z = ۳$) که کل تعداد Y نفر ٦٥ سال به بالا که در پنج شهرک مسکونی زندگی می‌کنند در محدودهٔ ۱۰٪ مقدار واقعی برآورد می‌شوند. با $\sigma_{۱y}^{۲} = ۶/۸$، $M = ۵$، $\varepsilon = ۰/۱۰$، $\bar{Y} = ۳۴$ خواهیم داشت

$$V_{۱y}^{۲} = \frac{\sigma_{۱y}^{۲}}{\bar{Y}^{۲}} = \frac{۶/۸}{۳۴^{۲}} = ۰/۰۰۵۹$$

تابلوی ٤.٩ اندازه‌های دقیق و تقریبی نمونه‌های موردنیاز، تحت نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

دقیق	تقریبی
مجموع	
$m = \frac{z_{۱-(\alpha/۲)}^{۲} M V_{۱x}^{۲}}{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲} + (M-۱)\varepsilon^{۲}}$	$m \approx \frac{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲}}{\varepsilon^{۲}}$
میانگین به ازای خوشه	
$m = \frac{z_{۱-(\alpha/۲)}^{۲} M V_{۱x}^{۲}}{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲} + (M-۱)\varepsilon^{۲}}$	$m \approx \frac{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲}}{\varepsilon^{۲}}$
میانگین به ازای واحد فهرست‌برداری	
$m = \frac{z_{۱-(\alpha/۲)}^{۲} M V_{۱x}^{۲}}{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲} + (M-۱)\varepsilon^{۲}}$	$m \approx \frac{z_{۱-(\alpha/۲)}^{۲} V_{۱x}^{۲}}{\varepsilon^{۲}}$
نسبت	
$m = \frac{z_{۱-(\alpha/۲)}^{۲} M V_{۱R}^{۲}}{z_{۱-(\alpha/۲)}^{۲} V_{۱R}^{۲} + (M-۱)\varepsilon^{۲}}$	$m \approx \frac{z_{۱-(\alpha/۲)}^{۲} V_{۱R}^{۲}}{\varepsilon^{۲}}$

که در آنها

$$V_{۱x}^{۲} = \frac{\sigma_{۱x}^{۲}}{\bar{X}^{۲}} \qquad V_{۱R}^{۲} = \left[\frac{\sigma_{۱x}^{۲}}{\bar{X}^{۲}} + \frac{\sigma_{۱y}^{۲}}{\bar{Y}^{۲}} - ۲ \times \frac{\sigma_{۱xy}}{\bar{X}\bar{Y}}\right]$$

$$\sigma_{۱xy} = \frac{\sum_{i=۱}^{M}(X_i - \bar{X})(Y_i - \bar{Y})}{M}$$

M، تعداد خوشه‌ها در جامعه؛ $z_{۱-(\alpha/۲)}^{۲}$، ضریب قابلیت اعتماد برای $[۱-(\alpha/۲)]$۱۰۰ اطمینان؛ و ε ویژگیهایی را نشان می‌دهد که برحسب ماکسیمم تفاوت نسبی مجاز بین برآوردها و پارامتر نامعلوم جامعه برای برآورد تعیین شده است.

و از تابلوی ٤.۹ داریم

$$m = \frac{z^{۲}_{۱-(\alpha/۲)} M V^{۲}_{۱y}}{z^{۲}_{۱-(\alpha/۲)} V^{۲}_{۱y} + (M-۱)\varepsilon^{۲}} = \frac{۹ \times ۵ \times ۰/۰۰۵۹}{۹ \times ۰/۰۰۵۹ + ۴ \times (۰/۱۰)^{۲}} = ۲/۸۵$$

پس از گرد کردن خواهیم داشت $m = ۳$. به این ترتیب، برای تأمین این ویژگیها به یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از سه خوشه نیاز خواهیم داشت.

□

برای تعیین اندازه‌ای برای نمونه که بتواند ویژگیهای موردنیاز برای قابل اعتماد بودن برآورد را تأمین کند لازم است مقدار $V^{۲}_{۱x}$ معلوم باشد که نسبت $\sigma^{۲}_{۱x}$، یعنی واریانس توزیع مجموعهای خوشه، به $\bar{X}^{۲}$، یعنی توان دوم میانگین سطح مشخصهٔ x به ازای خوشه است. این کمیتها پارامترهای جامعه‌اند که معمولاً نامعلوم‌اند و باید یا از روی مطالعات اولیه برآورد شوند یا با استفاده از تجربهٔ گذشته یا به طور شهودی حدس زده شوند.

۵.۹ قابلیت اعتماد برآوردها و هزینه‌های مربوط

یکی از مزایای عمدهٔ نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده آن است که هزینهٔ به دست آوردن نمونه‌ای از واحدهای فهرست‌برداری با این روش غالباً کمتر از هزینهٔ به دست آوردن نمونه‌ای با همان تعداد واحدهای فهرست‌برداری به وسیلهٔ روشهای دیگر است. این مطلب در مثال تشریحی فصل ۸ نشان داده شد که شامل مقایسهٔ هزینه‌های گرفتن یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده از ۱۰ خانوار در مقابل انتخاب یک نمونهٔ تصادفی ساده از ۱۰ خانوار بود. در این بخش، از یک مثال تشریحی برای نشان دادن اینکه استفاده از نمونه‌گیری خوشه‌ای یک‌مرحله‌ای می‌تواند هزینه‌ها را کاهش دهد استفاده می‌کنیم ولی این کاهش غالباً به قیمت حصول برآوردهایی تمام می‌شود که خطاهای معیار زیادی دارند.

مثال تشریحی: آمارگیری نمونه‌ای دادگاههای فدرال را که قبلاً در همین فصل شرح داده شد در نظر می‌گیریم. حالا فرض می‌کنیم که این نمونه از جامعهٔ اصلی زیر متشکل از ۲٦ دادگاه فدرال انتخاب شده که در جدول ۵.۹ نشان داده شده است (دادگاههای نمونه‌گیری شده با اعداد ایرانیک نشان داده شده‌اند)

اگر تقریباً ۱۵٪ از همهٔ اشخاصی که در این دادگاههای فدرال آزادی مشروط گرفته‌اند توصیه‌نامه‌ای نیز برای درمان اعتیاد داشته‌اند، در این صورت ۱۵۳٦٤ نفر واجد شرایط از میان تقریباً ۱۰۲٤۲۷ نفر کل افراد مورد بررسی یا به طور متوسط ۳۹۳۹ نفر به ازای هر دادگاه غربالگری شده‌اند. اگر هزینه‌های

جدول ۵.۹ برای ۲۶ دادگاه فدرال، تعداد افراد واجد شرایط و تعدادی که تحت درمان اعتیاد قرار گرفته‌اند در میان افرادی که آزادی مشروط یافته‌اند

دادگاه فدرال	تعداد افراد واجد شرایط	تعدادی که تحت درمان اعتیاد قرار گرفته‌اند
۱	۵۱٤	۲۵۳
۲	۲۰۵	۷۸
۳	۳۲۹	۲۱
٤	۵۰۸	۱٤۵
۵	۲۲٤	۱۷
٦	**٤۸٦**	**۷۹**
۷	۲۷۵	۰
۸	۲۲٤	۱۸
۹	۸۳	۱٤
۱۰	**۲٤۰**	**۹٤**
۱۱	۳۰۵	۳۰
۱۲	۸۲	۲۲
۱۳	۷٤	۸
۱٤	**٤۲۸**	**۱۷**
۱۵	**۳٤۳**	**۵۷**
۱٦	۲۳۲	۱۱
۱۷	**۱۱۳۰**	**٦۳**
۱۸	۱۰۰۰	۱٦۵
۱۹	**۹۸۳**	**۱۰**
۲۰	**۳۳۳**	**۵۸**
۲۱	**۱۳**	۰
۲۲	**۱۵۰٦**	**۱۰۱**
۲۳	۱۱۸۱	۲۲۸
۲٤	۱۳۹۸	۱۲
۲۵	**۱۷۵۵**	**٤۱۱**
۲٦	۱۵۱۳	۵۳۲
جمع	۱۵۳٦٤	۲٤٤٤

کارمـندان برای صرف وقت غربالگری این افراد تقریباً ۲۰ دقیقه به ازای هر نفر باشد (که شامل بررسی سـوابق دادگـاه بـرای تعیین واجد شرایط بودن فرد است)، در آن صورت هزینهٔ مورد انتظار غربالگری بـرای آمارگیـری نمونـه‌ای ۱۰ دادگـاه فـدرال تقریـباً ۱۳۱۳۰ نفـر ساعت ($\frac{۲۰}{۶۰}$×۱۰×۳۹۳۹) است. تعداد اشـخاص واجـد شـرایط کـه انـتظار مـی‌رود با این غربالگری در یک آمارگیری نمونه‌ای از ۱۰ دادگاه شناسایی شوند تقریباً ۵۹۰۹ نفر (۰/۰۱۵×۱۰×۳۹۳۹) خواهد بود.

بـرای هر یک از افرادی که واجد شرایط تشخیص داده شود با مأمور مسئول آزادی مشروط آن فرد تمـاس گرفـته و درخواست می‌شود تا از روی سوابق وی تحقیق کند که آن شخص واقعاً تحت درمان اعتیاد قرار گرفته است یا نه. برآورد می‌شود که هزینهٔ زمانی این تحقیق همراه با استخراج و آماده‌سازی داده‌هـا تقریـباً ٤٥ دقیقـه بـه ازای هـر فـرد واجـد شرایط باشد. به این ترتیب، هزینهٔ مورد انتظار برای آمارگیـری نمونـه‌ای، تقریـباً ٤٤۳۲ نفـر ساعت ($\frac{۴۵}{۶۰}$ ×۵۹۰۹) خواهد بود. کل هزینه‌های میدانی برآورد شده برای این آمارگیری نمونه‌ای به شرح زیر برحسب نفر ساعت (p.h.) خلاصه می‌شود:

هزینهٔ مربوط به شناسایی اشخاص واجد شرایط : ۱۳۱۳۰ نفر ساعت

هزینهٔ مربوط به کسب اطلاعات از اشخاص واجد شرایط : ٤٤۳۲ نفر ساعت

کل هزینه : ۱۷٥٦۱ نفر ساعت

اگـر قـرار باشد به جای نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده، یک نمونهٔ تصادفی ساده با همین تعداد افـراد واجـد شـرایط (یعـنی ۵۹۰۹ نفـر) گرفـته شـود ناچـار بـاید همـهٔ موارد آزادی مشروط در همهٔ دادگاههـای ۲٦ گانـه (بـه جای دادگاههای نمونهٔ ۱۰ گانه) از نظر واجد شرایط بودن، غربالگری شوند. ایـن کار مستلزم غربالگری تقریباً ۱۰۲٤۲۷ نفر با هزینهٔ زمانی ۳٤۱٤۲ نفر ساعت ($\frac{۲۰}{۶۰}$×۱۰۲۴۲۷) می‌بود. بـا ایـن کـار تقریـباً ۱٥۳٦٤ نفـر واجـد شـرایط (۰/۱٥×۱۰۲۴۲۷) به دست می‌آمد. برای به دست آوردن اطلاعـات از ۵۹۰۹ نفـر نمونـه (در نمونـه‌گیـری خوشه‌ای یک‌مرحله‌ای ساده) تقریباً ٤٤۳۲ نفر ساعت وقت لازم می‌بود.

هزینه‌های میدانی برآورد شده برای نمونهٔ تصادفی ساده متشکل از ۵۹۰۹ نفر واجد شرایط به شرح زیر بر حسب نفر ساعت خلاصه شده است:

هزینهٔ مربوط به شناسایی اشخاص واجد شرایط : ۳٤۱٤۲ نفر ساعت

هزینهٔ مربوط به کسب اطلاعات از اشخاص واجد شرایط : ٤٤۳۲ نفر ساعت

کل هزینه : ۳۸٥۷٤ نفر ساعت

از مطالب بالا درمی‌یابیم که کل هزینه‌های میدانی آمارگیری نمونه‌ای با طرح نمونهٔ تصادفی ساده (۳۸۵۷٤ نفر ساعت) به مراتب بیشتر از هزینه‌های میدانی این آمارگیری با طرح نمونه‌ای خوشه‌ای یک‌مرحله‌ای (۱۷۵٦۱ نفر ساعت) است.

باز فرض می‌کنیم که مهمترین هدف آمارگیری نمونه‌ای، برآورد کل تعداد افراد واجد شرایطی است که تحت درمان اعتیاد قرار گرفته‌اند. حالا خطاهای معیار این متغیر را برای دو طرح نمونه‌ای مقایسه می‌کنیم. در مورد نمونهٔ خوشه‌ای یک‌مرحله‌ای، برآورد خطای معیار برآورد، x'_{clu}، به طوری که از فرمول تابلوی ۳.۹ با استفاده از $M=۲۶$، $m=۱۰$ و $\sigma_{۱x}=۱۲۸/۸۶$ به دست می‌آید ۸٤۷/٦۰ است. در مورد نمونهٔ تصادفی ساده متشکل از ۵۹۰۹ نفر واجد شرایط، برآورد خطای معیار برای برآورد کل x' با استفاده از $n=۵۹۰۹$ ، $N=۱۵۳۶۴$ و $p=\frac{۲۴۴۴}{۱۵۳۶۴}=۰/۱۵۹۱$ برابر است با ۵۷/۳٦. این بمراتب کمتر از خطای معیار برآورد کل از روی نمونهٔ خوشه‌ای یک‌مرحله‌ای است. این دو طرح برای ۵۹۰۹ مورد واجد شرایط در پایین مقایسه شده‌اند.

طرح نمونه	کل اشخاص غربالگری شده	کل اشخاص نمونه‌گیری شده	کل هزینه‌های میدانی برحسب نفر ساعت	خطای معیار برآورد مجموع افرادی که تحت درمان اعتیاد قرار گرفته‌اند
نمونه‌گیری تصادفی ساده	۱۰۲٤۲۷	۵۹۰۹	۳۸۵۷٤	۵۷/۳۵
نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده (m=۱۰)	۳۹۳۹۰	۵۹۰۹	۱۷۵٦۱	۸٤۷/٦۰

کدام یک از این دو، طرح منتخب است؟ نمونهٔ تصادفی ساده هزینهٔ بیشتری دارد ولی خطای معیار آن بمراتب کمتر از طرح خوشه‌ای یک‌مرحله‌ای با همان تعداد عناصر نمونه است. در بخش بعد، راهبردهایی را برای انتخاب از بین دو طرح نمونه براساس هزینه و قابلیت اعتماد هر دو مورد بحث قرار می‌دهیم.

□

٦.۹ انتخاب طرح نمونه براساس هزینه و قابلیت اعتماد

اساساً دو راهبرد بسیار متداول برای انتخاب از بین چند طرح نمونه‌ای ممکن وجود دارند. اولین راهبرد شامل چهار گام است که ذیلاً نشان داده می‌شود (برای راحتی کار، فرض می‌کنیم که در انتخاب طرح نمونه، فقط یک برآورد قرار است در نظر گرفته شود).

۱. خطای نسبی ε موردنیاز برای برآورد را تعیین کنید.
۲. برای هر طرح نمونه‌ای تحت بررسی، اندازهٔ نمونهٔ لازم برای به دست آوردن برآوردی با اطمینان مشخص شده را تعیین کنید که بتواند ویژگیهای فهرست شده در گام ۱ را تأمین کند.
۳. برای هر طرح نمونه‌ای، هزینه‌های میدانی لازم برای به دست آوردن اندازهٔ نمونهٔ تعیین شده در گام ۲ را برآورد کنید.
٤. طرح نمونه‌ایی را انتخاب کنید که طبق محاسبهٔ گام ۳ کمترین هزینه را به بار آورد.

به عبارت دیگر، این راهبرد از بین طرحهای نمونه‌ای رقیب، آن طرح نمونه‌ای را انتخاب می‌کند که ویژگیهای مورد نیاز را با کمترین هزینهٔ میدانی تأمین کند. با نگاهی به یک مثال، ببینیم چگونه می‌توان از این راهبرد استفاده کرد.

مثال تشریحی: مجدداً آمارگیری نمونه‌ای دادگاههای فدرال و داده‌های جامعه‌ای ارائه شده در جدول ۵.۹ بخش قبل را در نظر می‌گیریم. فرض کنید می‌خواهیم بین نمونه‌گیری تصادفی ساده و نمونه‌گیری خوشه‌ای یک‌مرحله‌ای یکی را انتخاب کنیم و می‌خواهیم ۹۵٪ مطمئن باشیم که کل تعداد افراد واجد شرایط که تحت درمان اعتیاد قرار گرفته‌اند در محدودهٔ ۳۳٪ مقدار واقعی برآورد می‌شوند. از روی فرمول تابلوی ٤.۹ و با $M = ۲۶$، $z_{۱-\alpha/۲} = ۱/۹۶$، $\varepsilon = ۰/۳۳$، $\sigma_{۱x} = ۱۲۸/۸۶۳$ و $\bar{x} = ۹۴$، تعداد مورد نیاز خوشه‌های نمونه یعنی m برابر با ۱۹ است.

یک نمونهٔ خوشه‌ای یک‌مرحله‌ای متشکل از ۱۹ دادگاه فدرال مستلزم غربالگری تقریباً ۷٤۸٤۱ سابقه (۱۹× ۳۹۳۹) با هزینهٔ ۲٤۹٤۷ نفر ساعت ($\frac{۲۰}{۶۰}$×۷۴۸۴۱) است. تقریباً ۱۱۲۲٦ نفر (۱۹×۰/۱۵×۳۹۳۹) باید با هزینهٔ ۸٤۲۰ نفر ساعت ($\frac{۴۵}{۶۰}$×۱۱۲۲۶) مصاحبه شوند. به این ترتیب، کل هزینهٔ میدانی آمارگیری نمونه‌ای با نمونهٔ خوشه‌ای یک‌مرحله‌ای ۳۳۳٦۷ نفر ساعت است.

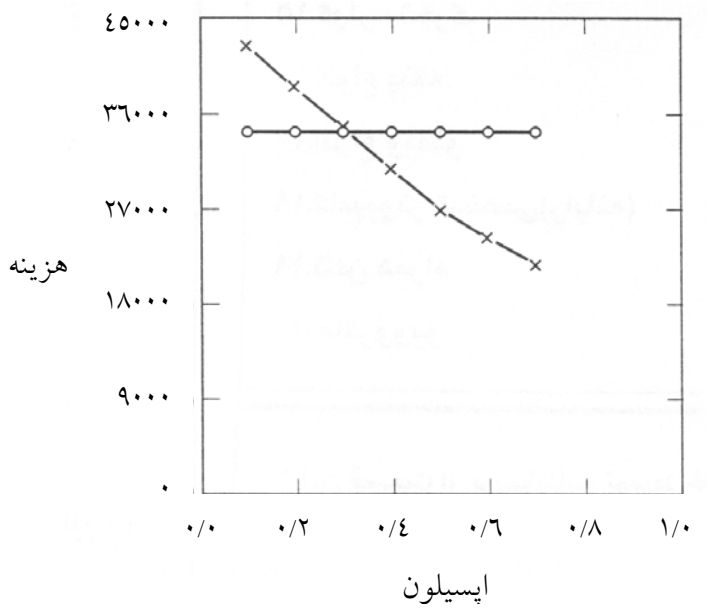
برای تأمین ویژگیهای فوق‌الذکر، نمونه تصادفی ساده به ۱۸۵ آزمودنی نیاز خواهد داشت (تابلوی ۵.۳ با $N = ۱۵۳۶۴$، $p = ۰/۱۵۹۱$، $\varepsilon = ۰/۳۳$ و $z_{۱-\alpha/۲} = ۱/۹۶$). این مستلزم غربالگری همهٔ ۱۰۲٤۲۷ پروندهٔ موجود در سوابق با هزینهٔ ۳٤۱٤۲ نفر ساعت ($\frac{۲۰}{۶۰}$×۱۰۲۴۲۷) است. هزینهٔ به دست آوردن اطلاعات دربارهٔ ۱۸۵ آزمودنی نمونه ۱۳۹ نفر ساعت ($\frac{۴۵}{۶۰}$×۱۸۵) خواهد بود. به این ترتیب کل هزینهٔ میدانی این آمارگیری ۳٤۲۸۱ نفر ساعت خواهد بود که بیشتر از هزینهٔ ۳۳۳٦۷ است که برای طرح نمونهٔ خوشه‌ای یک‌مرحله‌ای تعیین شد.

در بالا دیدیم که طرح نمونهٔ خوشه‌ای یک‌مرحله‌ای، ویژگی قابلیت اعتماد را که با پارامترهای $\varepsilon = ۰/۳۳$ و $z_{۱-\alpha/۲} = ۱/۹۶$ تعیین می‌شود با هزینه‌ای کمتر از طرح نمونهٔ تصادفی ساده تأمین می‌کند. ولی یک ویژگی دیگر می‌تواند با توجه به هزینه‌های نسبی این دو طرح به نتیجه‌گیری دیگری برسد. برای مثال، اگر کسی بخواهد ۹۵٪ مطمئن باشد که برآورد در محدودهٔ ۱۵٪ مقدار واقعی است (یعنی $\varepsilon = ۰/۱۵$ و $z_{۱-\alpha/۲} = ۱/۹۶$)، در آن صورت کل هزینه‌های میدانی طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای بیشتر از هزینهٔ طرح نمونه‌گیری تصادفی ساده (۴۳۹۰۳ در برابر ۳۴۷۸۲) خواهد بود.

رابطهٔ بین کل هزینه‌های میدانی و ε برای هر طرح در شکل ۲.۹ (برای $z_{۱-\alpha/۲} = ۱/۹۶$) نشان داده شده است. در طرح نمونهٔ تصادفی ساده، هزینه‌های عمده مربوط به غربالگری به منظور یافتن موارد واجد شرایط است و این هزینه‌ها مستقل از اندازهٔ نمونهٔ موردنیازند. در طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده، همچنان که ε کاهش می‌یابد، به خوشه‌های نمونه‌ای بیشتری نیاز پیدا می‌شود و این هم هزینهٔ غربالگری و هم هزینه‌های جمع‌آوری داده‌ها را افزایش می‌دهد.

□

دومین راهبرد که درانتخاب یک طرح از بین چندین طرح نمونه‌گیری بسیار متداول است، این است که ابتدا آن طرح نمونه‌ای تعیین شود که برآوردهایی تولید می‌کند که با هزینه‌های میدانی مشخص شده کمترین واریانس را دارند. این کار به این ترتیب انجام می‌شود که ابتدا اندازهٔ نمونه‌ای که می‌توان با هزینه‌های میدانی تعیین شده گرفت محاسبه شود و سپس واریانس برآورد برای آن اندازهٔ نمونه‌ای خاص محاسبه شود.



شکل ۲.۹ هزینه: نمونهٔ تصادفی ساده و نمونه‌گیری خوشه‌ای یک‌مرحله‌ای

X = خوشه؛ ٠= نمونهٔ تصادفی ساده

داشتن یک معادلـــه که اندازهٔ نمونه را به هزینه‌های میدانی مرتبط سازد کمک خواهد کرد که این راهبرد بـه آسـانی مـورد استفاده قرار گیرد. چنین رابطه‌ای را *تابع هزینه* می‌نامند و ابزاری است بسیار سـودمند که به پژوهشگر امکان می‌دهد تا از میان طرحهای نمونه‌ای رقیب، طرحی را انتخاب کند که با هزیـنه‌ای مشـخص دارای کمـترین واریـانس باشـد. در زیر نشان می‌دهیم که چنین تابع هزینهٔ ساده‌ای چگونه می‌تواند به دست آید.

ابـتدا هزیـنه‌هـای میدانی هر یک از دو طرح نمونه‌ای را که مورد بحث قرار دادیم به روشی اندکی مـتفاوت بررسـی مـی‌کـنیم. بـرای طرح نمونه‌گیری تصادفی ساده، هزینهٔ ایجاد چارچوب نمونه‌گیری، مسـتقل از تعداد واحدهای شمارش است که در نمونه انتخاب شده‌اند، زیرا هر واحد شمارش باید قبل از انـتخاب شـدن در نمونـه فهرسـت شده باشد. به علاوه ، ممکن است سایر هزینه‌ها وابسته به تعداد واحدهـای شـمارش انـتخاب شده در نظر گرفته شوند. به این ترتیب، یک برآورد معقول از هزینه‌های میدانی برای طرح نمونهٔ تصادفی ساده ممکن است از رابطهٔ زیر به دست آید

$$C = C_0 + C_1 n \qquad (۲.۹)$$

کـه در آن C_0، مؤلفهٔ هزینهٔ همراه با ایجاد چارچوب نمونه‌گیری به اضافهٔ سایر هزینه‌های میدانی است کـه بـه اندازهٔ نمونه بستگی ندارند، و C_1، هزینهٔ جمع‌آوری داده‌ها به ازای هر واحد شمارش است. در مورد مثال بالا، مقادیر C_0 و C_1 را به صورت زیر در اختیار داریم

C_0 = ۳٤۱٤۲ نفر ساعت $\qquad$ C_1 = ۰/۷۵ نفر ساعت

همـان‌طـور کـه قـبلاً بحـث شد، هزینه، C_0 با ضرب کردن ۱۰۲٤۲۷، تعداد مورد انتظار، کل سابقه‌های مـربوط به موارد آزادی مشروط که باید برای تعیین واجد شرایط بودن غربالگری شوند در زمان برآورد شده، ۲۰ دقیقه یا $\frac{۱}{۳}$ ساعت، که صرف غربالگری سابقهٔ هر مورد می‌شود، به دست می‌آید. هزینه، C_1 ، زمـان بـرآورد شـده، ٤۵ دقیقـه یـا ۰/۷۵ ساعت است، که برای فرایند استخراج سابقهٔ هر نمونه صرف می‌شود.

هزیـنه‌هـای میدانـی بـرای نمونـه‌گیـری خوشـه‌ای یـک‌مرحله‌ای ساده شکل متفاوتی دارند. چون، هزیـنه‌هـای ایجـاد چـارچوب بـه تعداد خوشه‌های انتخاب شده بستگی دارند، هزینه‌های میدانی برای نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده می‌توانند با معادلهٔ زیر توصیف شوند:

$$C = C_1' m + C_2' m\overline{N} \qquad (۳.۹)$$

مؤلفـهٔ هزیـنهٔ C_1' مـلازم بـا خوشـه‌هـا شامل هزینه‌های «پذیرش» خوشه در بررسی و ایجاد چارچوب نمونـه‌گیـری بـرای آن خوشـه و نیز سایر هزینه‌هایی است که به تعداد واحدهای شمارش نمونه‌گیری

شـده بسـتگی ندارنـد. مؤلفهٔ هزینه C_2' شامل همهٔ هزینه‌هایی است که می‌توانند بر مبنای واحد پیش از شمارش (عمدتاً هزینه‌های جمع‌آوری و پردازش داده‌ها) بیان شوند.

در آمارگیری نمونه‌ای دادگاههای فدرال، مؤلفهٔ هزینه، C_1'، اساساً هزینهٔ ایجاد چارچوب نمونه‌گیری بـرای هـر دادگـاه نمونـه اسـت ـــ یعـنی ۲۰ دقیقـه بـرای سابقهٔ هر مورد که برای واجد شرایط بودن غربالگـری شـده است ضرب در ۳۹۳۹/۵ (متوسط تعداد سوابق دعاوی به ازای هر دادگاه فدرال) ـــ و برابر است با ۱۳۱۳/۱۷ نفر ساعت. مؤلفهٔ هزینه، C_2'، هزینهٔ جمع‌آوری داده‌هاست و برابر است با ۰/۷۵ نفر ساعت. به این ترتیب، داریم

$$C_1' = ۱۳۱۳/۱۷ \qquad C_2' = ۰/۷۵$$

تابعهـای هزیـنه کـه در بـالا از آنهـا بحـث شـد تقریبهای خام از هزینه‌های میدانی واقعی‌اند و در انـتخاب یکـی از چـند طرح نمونه‌ای متفاوت سودمندند. از این قبیل تابعهای هزینه در فصلهای بعدی هنگام بحث در مورد کارایی هزینه استفاده خواهد شد.

اکنون به مثالی از این راهبرد دوم با استفاده از همان مثال تشریحی می‌پردازیم.

مـثال تشـریحی: فـرض کـنیم بودجـه‌ای داریـم کـه بیـش از ۲۵۰۰۰ نفر ساعت به هزینه‌های میدانی اختصـاص نـداده اسـت و مـی‌خواهیـم از بیـن نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده و نمونه‌گیری تصادفی ساده یکی را انتخاب کنیم. از معادلهٔ ۳.۹ با

$$C = ۲۵۰۰۰ \qquad C_1' = ۱۳۱۳/۱۷ \qquad C_2' = ۰/۷۵ \qquad \overline{N} = ۵۹۰/۹۲$$

معادلـــــه را بـرای m حـل می‌کنیم و ۱٤/۲ دادگاه فدرال به دست می‌آوریم. به این ترتیب، یک نمونهٔ خوشـه‌ای یـک‌مـرحله‌ای سـاده بـا m =۱٤ دادگاه فدرال محدودیت هزینهٔ ۲۵۰۰۰ نفر ساعتی را تأمین خواهد کرد. اگر برآورد موردنظر ما x'_{clu}، کل تعداد اشخاصی باشد که تحت درمان اعتیاد قرار گرفته‌اند، در آن صورت، خطای معیار این برآورد به صورت زیر است:

$$SE(x'_{clu}) = \left(\frac{M}{\sqrt{m}}\right)\hat{\sigma}_{۱x}\sqrt{\frac{M-m}{M-۱}} = \frac{۲۶}{\sqrt{۱۴}}(۱۲۸/۸۶)\sqrt{\frac{۲۶-۱۴}{۲۶-۱}} = ۶۲۰/۳۷$$

چـون کـل جامعـهٔ افـرادی کـه تحـت درمان اعتیاد قرار گرفته‌اند ۲٤٤٤ نفر است، ضریب تغییرپذیری بـرآورد برابر ۲۵/٤٪ مقدار واقعی خواهد بود. با این هزینه ۲۵۰۰۰ نفر ساعت، می‌توان ۹۵٪ مطمئن بود که برآورد در محدودهٔ ۵۰/۸٪ مقدار واقعی (که ۱/۹٦ برابر ضریب تغییرپذیری برآورد است) قرار دارد.

در مقـابل، اگـر هزیـنه‌های میدانی تخصیص داده شده به آمارگیری نمونه‌ای به ۲۵۰۰۰ نفر ساعت محـدود شـده باشـد نمـی‌تـوان طـرح نمونـه‌گیری تصادفی ساده را به اجرا درآورد. علت آن است که

هزینه‌های ایجاد چارچوب به تنهایی ۳٤۱٤۲ نفر ساعت خواهد بود. به این ترتیب، از میان دو طرح نمونه‌ای، نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده به دلیل نبود طرح دیگر برنده خواهد شد. اگر هزینه‌های میدانی بسیار بیشتری اختصاص داده شود (مثلاً ٤۰۰۰۰ نفر ساعت)، الزاماً چنین موردی پیش نخواهد آمد.

□

در ارتباط با مطالب این بخش و بخش قبل باید به دو نکته اشاره شود. نکتهٔ اول این است که نمونه‌گیری تصادفی ساده و نمونه‌گیری خوشه‌ای یک‌مرحله‌ای معرف دو «کرانگین» هستند. نمونه‌گیری تصادفی ساده برای تعدادی ثابت، n، از عناصر عموماً برآوردهایی تولید می‌کند که در مقایسه با برآوردهای تولید شده از سایر طرحها خطاهای معیار کمی دارد. ولی، هزینه‌های نمونه‌گیری تصادفی ساده به علت هزینه‌های زیاد برای ایجاد چارچوب و هزینه‌های زیاد سفر (برای آمارگیری از طریق مصاحبه حضوری) نسبت به سایر طرحها به طور کلی زیاد است.

در مقابل، نمونه‌گیری خوشه‌ای یک‌مرحله‌ای در مقایسه با سایر طرحها برای تعدادی ثابت از واحدهای شمارش نمونه غالباً هزینه‌های کمتری دارد. علت آن است که هزینه‌های ایجاد چارچوب عموماً کاهش یافته است، زیرا فقط واحدهای شمارش موجود در خوشهٔ نمونه باید فهرست شوند. همچنین، خوشه‌بندی جغرافیایی نیز هزینه‌های سفر را در آمارگیریهای خانوار که با مصاحبهٔ حضوری اجرا می‌شوند کاهش می‌دهد. ولی برآوردهای تولید شده با نمونه‌گیری خوشه‌ای یک‌مرحله‌ای هرگاه با سایر طرحها بر مبنای هر واحد شمارش نمونه مقایسه شود غالباً خطاهای معیار زیادی دارند. علت آن است که شباهتهای موجود میان اعضای یک خوشه می‌تواند به فزونگی منجر شود و اندازهٔ نمونهٔ «مؤثر» بمراتب کمتر از اندازهٔ نمونهٔ واقعی می‌شود. همچنین اگر تغییرپذیری میان خوشه‌ها در کل تعداد واحدهای شمارش به ازای خوشه، یعنی N_i زیاد باشد، برآوردهای مجموعها تحت نمونه‌گیری خوشه‌ای یک‌مرحله‌ای در مقایسه با سایر طرحهایی که همین تعداد واحد شمارش دارند می‌توانند خطاهای معیار بسیار زیاد داشته باشند. در دو فصل بعدی، به بحث در مورد طرحهای نمونه‌گیری خوشه‌ای دیگری می‌پردازیم که از برخی از مشکلات موجود در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای به دورند.

بحث این بخش برای نشان دادن مفاهیم مبتنی بر منظور نمودن هزینه‌ها هنگام انتخاب از بین طرحهای نمونه‌ای، بسیار ساده شده بود. ما در مثال خود فقط براساس یک برآورد انتخاب می‌کردیم (مثلاً کل تعداد افرادی که تحت درمان اعتیاد قرار گرفته بودند). در وضعیتهای موثق‌تر، برآوردهای بسیاری حائز اهمیت تلقی می‌شوند و طرح منتخب طرحی خواهد بود که وقتی همهٔ برآوردها مدنظر قرار گرفته باشند بتواند به مفهومی که در این بخش بحث شد به بهترین شکل عمل کند. همچنین،

تابعهای هزینه که در این بخش شرح داده شدند تابعهای خطی ساده‌ای از خوشه‌ها یا واحدهای شمارش بودند. اگر چه این تابعهای ساده شده غالباً در کاربست مفیدند، گاهی اوقات برای توصیف هزینه‌ها به راهی بامعنا به تابعهای پیچیده‌تری نیاز است. در متون مشهور نمونه‌گیری که توسط هنسن و همکاران [۲] و جِسِن [۳] نگاشته شده، مباحثی فوق‌العاده عالی در مورد تابعهای هزینه مطرح شده است. یک کتاب تازه‌تر توسط گرووز [۱۰] که منحصراً به هزینه‌های آمارگیری و خطاهای آمارگیری می‌پردازد بحثی تفصیلی و در عین حال خواندنی دربارهٔ مؤلفه‌های هزینه و تابعهای هزینه فراهم می‌سازد.

۷.۹ خلاصه

در این فصل، مفاهیم پایه‌ای نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده را مورد بحث قرار دادیم که طرحی است متضمن انتخاب یک نمونهٔ تصادفی ساده از خوشه‌ها و گنجاندن یکایک واحدهای شمارش موجود در داخل هر خوشه در نمونه. روش‌شناسی برآورد کردن میانگینها، مجموعها، و نسبتها را تحت نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده شرح و بسط دادیم و در مورد توزیعهای نمونه‌گیری این برآوردها بحث کردیم. به علاوه، روشهایی را برای برآورد کردن خطاهای معیار این برآوردها از روی داده‌های موجود در نمونه ارائه دادیم.

تعیین اندازهٔ نمونهٔ لازم برای تأمین ویژگیهای موردنظر در قابلیت اعتماد برآوردهای حاصل از نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده را به بحث گذاشتیم و متوجه شدیم فرمولهایی که دراینجا بسط داده‌ایم بسیار شبیه به آنهایی هستند که پیش از این برای نمونه‌گیری تصادفی ساده بسط داده بودیم و دریافتیم که در اینجا مجموعهای خوشه‌ای به صورت مشاهدات مؤثر به کار می‌روند. راهبردهایی را برای انتخاب از میان طرحهای نمونه‌ای مختلف با توجه به هزینه‌ها ارائه کردیم. بالاخره، دربارهٔ مفاهیم مؤلفه‌های هزینه و تابعهای هزینه بحث کردیم.

تمرین

۱.۹ فرض کنید مدارس ابتدایی موجود در یک شهر در ۳۰ ناحیهٔ آموزشی گروه‌بندی شده‌اند و هر ناحیهٔ آموزشی شامل چهار مدرسه است. فرض کنید که یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از سه ناحیهٔ آموزشی گرفته شده است تا تعداد کودکان دبستانی شهر را که (براساس سنجش به وسیلهٔ یک آزمون استاندارد) دچار کوررنگی هستند برآورد کند، و فرض کنید که داده‌های زیر از این نمونه به دست آمده‌اند. کل تعداد کودکان کوررنگ و نسبت همهٔ کودکانی را که کوررنگ‌اند برآورد کنید و بازهٔ اطمینان ۹۵٪ آنها را به دست آورید.

ناحیهٔ آموزشی نمونه	مدرسه	تعداد کودکان	تعداد کودکان کوررنگ
۱	۱	۱۳۰	۲
	۲	۱۵۰	۳
	۳	۱٦۰	۳
	٤	۱۲۰	۵
۲	۱	۱۱۰	۲
	۲	۱۲۰	٤
	۳	۱۰۰	۰
	٤	۱۲۰	۱
۳	۱	۸۹	٤
	۲	۱۳۰	۲
	۳	۱۰۰	۰
	٤	۱۵۰	۲

۲.۹ قرار است از روی سوابق بیماران در یک درمانگاه روانی بیماران سرپایی، نمونه‌ای از بیماران به منظور برآورد کل تعداد بیمارانی که به عنوان بخشی از رژیم درمانی خود داروهای ضد افسردگی تری‌سیکلیک دریافت کرده‌اند گرفته شود. سوابق در کشوهای بایگانی تنظیم شده‌اند که هر کشو محتوی سوابق ۲۰ بیمار است و کشوهای بایگانی جمعاً ٤۰ تا هستند.

الف. فرض کنید می‌خواهیم از نمونه‌گیری تصادفی ساده از سوابق بیماران استفاده کنیم. اگر بخواهیم واقعاً مطمئن باشیم که برآورد کل تعداد افرادی که داروهای ضدافسردگی تری‌سیکلیک به آنها داده شده در محدودهٔ ۱۰٪ مقدار واقعی است و اگر پیش‌بینی کنیم که به تقریباً ۲۰٪ همهٔ بیماران از این داروها تجویز شده است اندازهٔ نمونه باید چقدر باشد؟

ب. هزینه‌های میدانی برای گرفتن نمونهٔ تعیین شده در قسمت (الف) چقدر خواهد بود؟ برای تعیین این هزینه‌های میدانی فرضهایی در مورد مؤلفه های هزینه ارائه کنید.

پ. اگر قرار باشد از طرح نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده [با همان ویژگیهای مندرج در قسمت (الف)] استفاده شود، اندازهٔ نمونهٔ مورد نیاز چقدر است؟

ت. هزینه‌های میدانی برای یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده چقدر خواهد بود؟ باز هم برای تعیین این هزینه‌های میدانی فرضهایی در مورد مؤلفه‌های هزینه ارائه کنید.

ث. کدام یک از دو طرح نمونه‌ای را به کار خواهید برد؟ چرا؟

۳.۹ یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از ۱۰ بیمارستان از یک جامعهٔ شامل ۳۳ بیمارستان در یک ایالت شمال مرکزی گرفته شد که بودجهٔ ایالتی و فدرال دریافت کرده بودند تا سطح خدمات درمانی فوری خود را ارتقا دهند. در هر یک از بیمارستانهایی که در نمونه انتخاب شدند، سوابق همهٔ بیمارانی که در سال تقویمی ۱۹۸۸ به دلیل آسیب‌دیدگیهای سخت بستری شده بودند (یعنی تصادفات، مسمومیتها، خشونت، سوختگی و غیره) مورد بررسی قرار گرفت. تعداد بیمارانی که به علت صدمات بستری شده بودند و تعدادی از آنها که به علت فوت ترخیص شدند در جدول زیر برای هر بیمارستان در نمونه نشان داده شده است.

بیمارستان	کل تعداد بیماران بستری شده به علت صدمه	کل تعداد ترخیص شدگان به علت فوت در میان همهٔ بیماران بستری شده به علت صدمه
۱	۵۶۰	٤
۲	۱۹۰	٤
۳	۲۶۰	۲
٤	۳۷۰	٤
۵	۱۹۰	٤
٦	۱۳۰	۰
۷	۱۷۰	۹
۸	۱۷۰	۲
۹	٦۰	۰
۱۰	۱۱۰	۱

الف. در مورد این نمونه، خوشه‌ها کدام‌اند؟ واحدهای فهرست‌برداری کدام‌اند؟ واحدهای اولیه کدام‌اند؟

ب. کل تعداد اشخاص بستری شده به علت صدمه را در میان ۳۳ بیمارستان برآورد کنید و برای آن یک بازهٔ اطمینان ۹۵ درصدی ارائه دهید.

پ. کل تعداد بیمارانی را که به علت فوت ترخیص شده‌اند در میان همهٔ افراد بستری شده به علت صدمه برآورد کنید و برای آن یک بازهٔ اطمینان ۹۵ درصدی ارائه دهید.

ت. نسبت بیماران ترخیص شده به علت فوت در میان همهٔ آنهایی را که به علت صدمه بستری شده‌اند برآورد کنید و برای آن یک بازهٔ اطمینان ۹۵ درصدی ارائه دهید.

٤.٩ تعداد تختهای هر یک از ده بیمارستان نمونه‌گیری شده در تمرین ۳.۹ در جدول زیر نشـــان داده شده است. مابقی ۲۳ بیمارستان که در نمونه قرار نگرفته‌اند در مجموع ۳٦۸۷ تخت دارند. از این اطلاعات برای به دست آوردن برآوردهای بهبود یافته و بازه‌های اطمینان موارد زیر استفاده کنید:

الف. کل تعداد اشخاص بستری شده به علت صدمه.

ب. کل تعداد اشخاص ترخیص شده به علت فوت در میان همهٔ اشخاص بستری شده به علت صدمه.

بیمارستان	تعداد تخت
۱	۸۲٤
۲	۳۱۲
۳	۳۲۹
٤	٦٤۸
٥	۳٥۸
٦	۲٥۲
۷	۲٥٦
۸	۲٦۳
۹	۱۳۸
۱۰	۱٥۰

٥.٩ یـــک بررسی در شهر بزرگی در چیـن اجرا شـــد که در آن از طـرح نمونه‌گیـــری خوشـــه‌ای یـک‌مرحـلـــــه‌ای سـاده اسـتفاده شـده بـود. خوشه‌ها در این مورد نمادهایی بودند که (به عنوان تـرجمهٔ «jumin xiaozu» چیـنی) «گروههای همسایگی» خوانده می‌شوند و اساساً گروههایی از خانوارهای مجاور یکدیگرند. این به اصطلاح گروههای همسایگی کوچکترین واحدهایی هستند کـه در جمهـوری خلق چین اطلاعات آماری برای آنها موجود است. این شهر شامل شش ناحیه اسـت کـه در ایـن طـرح، طـبقات را تشـکیل می‌دهند. درون هر ناحیه، یک نمونهٔ تصادفی ساده متشـکل از دو گـروه همسـایگی انـتخاب شـد و بـا همـهٔ افراد داخل هر گروه همسایگی دربارهٔ وضـعیت کلی سلامت آنها مصاحبه به عمل آمد. از این آمارگیری داده‌های زیر دربارهٔ تعداد افراد بالاتر از سی سال که دندانهای خود را از دست داده بودند به دست آمد.

ناحیه	گروه همسایگی	تعداد اشخاص بالاتر از سی‌سال	تعداد اشخاص بی‌دندان
۱	۱	۲۸	۷
	۲	۳۵	۹
۲	۱	۲۹	۱۲
	۲	٤۳	۲٦
۳	۱	٦۱	۱۹
	۲	٤۸	۱۲
٤	۱	۱۵	۱۰
	۲	۳۹	۲۸
۵	۱	۲۱	۹
	۲	٤٦	۱۵
٦	۱	۱۲	۰
	۲	۲۵	٤

طبقه‌های ۱، ۲، ٤ و ٦ هر یک شامل ۲۰۰ گروه همسایگی، طبقهٔ ۳ شامل ۱۷۵ گروه همسایگی و طبقهٔ ۵ شامل ۱۵۰ گروه همسایگی است. (۱) کل تعداد افراد سی سال و بالاتر را که دندانهای خود را از دست داده‌اند و (۲) نسبت همهٔ افراد سی سال و بالاتر را که دندانهای خود را از دست داده‌اند برآورد کنید و برای آن یک بازهٔ اطمینان ۹۵ درصدی ارائه دهید.

٦.۹ در تمرین ۵.۹، طی اجرای عملیات میدانی معلوم شد که گروه همسایگی ۲ در ناحیهٔ ۳ به علت توزیع مجدد جمعیت، دیگر وجود خارجی ندارد. با درنظر داشتن این موضوع، برآوردهای اصلاح شدهٔ پارامترهای برآورد شده در تمرین ۵.۹ را به دست آورید.

۷.۹ یک سازمان حفظ بهداشت دهان و دندان دارای ۳٦۸ عضو است و هر عضو ٤ ربع فک دارد (چپ بالا، راست بالا، چپ پایین، راست پایین). مطلوب است اجرای یک آمارگیری نمونه‌ای برای این هدف که کل تعداد ربع فکها در میان اعضا که به نوعی جراحی لثه احتیاج دارند برآورد شود. برنامهٔ نمونه‌گیری مستلزم گرفتن یک نمونهٔ تصادفی ساده از بیماران و ارزیابی وضعیت هر یک از چهار ربع فکهای هر بیمار نمونه خواهد بود. یک بررسی مقدماتی براساس یک نمونه با

داوری شخصی از ۷ بیمار، داده‌های زیر را به دست می‌دهد:

	ربع فک			
بیمار	۱	۲	۳	٤
۱	+	+	+	−
۲	−	−	+	+
۳	−	−	−	−
٤	+	−	−	+
٥	−	−	−	−
٦	+	−	−	−
۷	+	+	+	+

که در آن

۱ = چپ پایین ۲ = راست پایین

۳ = چپ بالا ٤ = راست بالا

+ = جراحی نیاز دارد − = جراحی نیاز ندارد

براساس این داده‌های مقدماتی، اگر بخواهیم کل تعداد ربع فکها را که نیاز به جراحی لثه دارند در میان همهٔ اعضا با ۹۵٪ اطمینان حول ۱۵٪ مقدار واقعی برآورد کنیم به چند بیمار نیاز داریم؟

۸.۹ در گزارشی که برای تمرین ۷.۹ ارائه شد، پیش‌بینی می‌شود که تقریباً ۱۵ دقیقه وقت لازم است تا دندانپزشک هر ربع فک را معاینه کند و به طور تقریبی ۲۰ دقیقه وقت برای کارهای دفتری لازم است تا برای هر بیمار نمونه‌گیری شده وقت ملاقات تنظیم و بیمار برای معاینه آماده شود. اگر هزینه‌های ملازم با آمارگیری فقط همین باشد و اگر مدت زمان دفتری، یک سوم وقت دندانپزشک ارزش داشته باشد بودجهٔ لازم برای تأمین ویژگیهای بیان شده در تمرین ۷.۹ (برحسب نفر ساعت دندانپزشک) چقدر است؟

۹.۹ در بسیاری از وضعیتهای مربوط به علم بهداشت، خوشه از یک جفت عنصر تشکیل شده است (مثلاً در چشم‌پزشکی ممکن است خوشه‌ها را بیماران تشکیل دهند و عناصر، چشمها باشند و از این قبیل). در این وضعیت، عبارت ساده‌ای برای واریانس برآورد مجموع تهیه کنید.

۱۰.۹ در وضعیت توصیف شده در تمرین ۹.۹ یک عبارت ساده شده برای واریانس برآورد نسبت تهیه کنید.

۱۱.۹ فرض کنید ٤۰۰۰۰ نفر ساعت برای هزینه‌های میدانی آمارگیری نمونه‌ای دادگاههای فدرال اختصاص یافته است که در یک مثال تشریحی در این فصل مورد بحث قرار گرفت. با استفاده از تابعهای هزینه و مؤلفه‌های هزینه که برای این مثال شرح و بسط داده شد تعیین کنید کدام یک از دو طرح نمونه‌ای ـ نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده یا نمونه‌گیری تصادفی ساده ـ می‌تواند کل تعداد اشخاصی را که تحت درمان اعتیاد قرار گرفته‌اند با خطای معیار کمتری برآورد کند.

۱۲.۹ در مورد آمارگیری نمونه‌ای دادگاههای فدرال که در یک مثال تشریحی در این فصل مورد بحث قرار گرفت نموداری رسم کنید که طول نقاط آن، کل هزینه‌های ثابت و عرض نقاط آن، ضریب تغییرات کل تعداد برآورد شدهٔ افرادی باشد که تحت درمان اعتیاد قرار گرفته‌اند. نمودار باید دو خط داشته باشد (مانند شکل ۲.۹): یکی برای طرح نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده، دیگری برای طرح نمونه‌گیری تصادفی ساده.

کتابشناسی

The following publications contain more detailed discussions of cluster sampling.

1. Cochran, W. G., *Sampling Techniques,* 3rd ed., Wiley, New York, 1977.
2. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Method and Theory,* Vol. 1, Wiley, New York, 1953.
3. Jessen, R. J., *Statistical Survey Techniques,* Wiley, New York, 1978.
4. Kish, L., *Survey Sampling,* Wiley, New York, 1965.
5. Scheaffer, R. L., Mendenhall, W., and Ott, L., *Elementary Survey Sampling,* 2nd ed., Duxbury Press, Scituate, Mass., 1979.
6. Sudman, S., *Applied Sampling,* Academic Press, New York, 1976.

The following publications report the findings of a survey which used single-stage cluster sampling.

7. Levy, P. S., Yu, E. S. H., Liu, W. T., Zhang, M. Y., Wang, Z. Y., Wong, S., and Katzman, R., A variation on single-stage cluster sampling used in a survey of elderly people in Shanghai, *International Journal of Epidemiology,* 17: 931, 1988.
8. Levy, P. S., Yu, E. S. H., Liu, W. T., Zhang, M., Wang, Z., Wong, S., and Katzman, R. Single stage cluster sampling with a telescopic respondent rule: A variation motivated by a survey of dementia in elderly residents of Shanghai. *Statistics in Medicine,* 8: 1225, 1989.

9. Yu, E. S. H., Liu, W. T., Levy, P. S., Zhang, M. Y., Katzman, R., Lung, C. T., Wong, S., Wang, Z. Y., and Qu, G. Y., Cognitive impairment among elderly adults in Shanghai, China. *Journal of Gerontology: Social Services,* 44: S97, 1989.

The following book gives a very rich discussion of survey costs including determination of cost components.

10. Groves, R. M., Survey Errors and Survey Costs, Wiley, New York, 1989.

The following article gives an overview of cluster sampling using terminology very similar to that used in this chapter.

11. Levy, P. S., Cluster sampling. In *Encyclopedia of Biostatistics.* Armitage, P. A., and Colton, T D., eds., Wiley, New York, 1998.

فصل ۱۰

نمونه‌گیری خوشه‌ای دومرحله‌ای: خوشه‌های نمونه‌گیری شده با احتمال برابر

در فصل قبل دربارهٔ نمونه‌گیری خوشه‌ای یک مرحله‌ای ساده بحث کردیم که مستلزم گرفتن یک نمونهٔ تصادفی ساده از خوشه‌ها و سپس در نمونه قرار دادن همهٔ واحدهای شمارش یا فهرست‌برداری موجود در هر خوشهٔ نمونه بود. در برخی از وضعیتها، اگر نمونه‌گیری در بیش از یک مرحله اجرا شود کارایی بیشتری خواهد داشت. یک وضعیت آشکار آن است که خوشه‌ها بزرگتر از آن‌اند که همهٔ واحدهای آنها به راحتی در نمونه قرار بگیرند. همچنین، هرگاه واحدهای فهرست‌برداری درون خوشه‌ها نسبت به متغیرهایی که باید اندازه‌گیری شوند بسیار همگن باشند در نمونه قرار دادن همهٔ واحدهای فهرست‌برداری موجود در خوشهٔ نمونه باعث فزونگی بسیار زیادی خواهد شد. در چنین وضعیتهایی، غالباً نمونه‌گیری از واحدهای فهرست‌برداری درون خوشهٔ انتخاب شده، بهتر از انتخاب همهٔ آنهاست. به عبارت دیگر، بهترین کار آن است که نمونه در دو مرحله انتخاب شود. به این صورت که مرحلهٔ اول، انتخاب نمونه از خوشه‌ها و مرحلهٔ دوم، انتخاب نمونه‌ای از واحدهای فهرست‌برداری درون هر خوشهٔ منتخب باشد.

در این فصل و فصل بعدی به بحث دربارهٔ طرحهای نمونه‌گیری خوشه‌ای دومرحله‌ای خواهیم پرداخت. طرحهای نمونه‌گیری که در این فصل بررسی می‌شوند شامل آنهایی است که به طور متداول

به عنوان *نمونه‌گیری خوشه‌ای دومرحله‌ای ساده* شناخته شده و در فصل ۸ تعریف شده‌اند. اینها طرحهایی هستند که در آنها خوشه‌ها در مرحلهٔ اول با نمونه‌گیری تصادفی ساده انتخاب می‌شوند؛ واحدهای فهرست‌برداری در مرحلهٔ دوم با نمونه‌گیری تصادفی ساده از داخل هر خوشه‌ای که در مرحلهٔ اول انتخاب شده است گرفته می‌شوند و کسر نمونه‌گیری واحدهای فهرست‌برداری که در مرحلهٔ دوم انتخاب می‌شوند برای هر خوشهٔ نمونه یکسان (یا تقریباً یکسان) است. وضعیت اخیر هنگامی ممکن است روی دهد که تعداد واحدهای فهرست‌برداری موجود برای نمونه‌گیری در داخل هر خوشه یکسان نباشد. مثلاً اگر یک بلوک شهری شامل ۱۰ خانوار و دیگری شامل شش خانوار باشد، گرفتن یک نمونهٔ متشکل از ۳۰٪ خانوارها از هر یک از این دو بلوک امکان نخواهد داشت. ولی اگر راهبرد کلی، داشتن کسرهای نمونه‌گیری مرحله دومی باشد که برای دو بلوک تا حد ممکن برابر باشند، آنگاه سه خانوار از بلوک ده خانواری و دو خانوار از بلوک شش خانواری انتخاب می‌کنیم. طرحهایی از این دست را *نمونه‌گیری خوشه‌ای دومرحله‌ای ساده* می‌نامیم.

در این فصل، فقط نمونه‌گیری خوشه‌ای دومرحله‌ای ساده را بررسی می‌کنیم و ابتدا به بحث در مورد وضعیت خاصی می‌پردازیم که در آن تعداد واحدهای شمارش در همهٔ خوشه‌های جامعه یکسان است؛ و سپس به بحث در مورد وضعیت کلیتری می‌پردازیم که در آن تعداد واحدهای شمارش در همهٔ خوشه‌ها یکسان نیست. در فصل ۱۱، به بحث از وضعیتی پیچیده‌تر خواهیم پرداخت که طی آن خوشه‌ها در مرحلهٔ اول با احتمالهای نابرابر نمونه‌گیری می‌شوند.

۱.۱۰ وضعیتی که در آن N_i، تعداد واحدهای شمارش در همهٔ خوشه‌ها یکسان است

اگرچه در بیشتر وضعیتهایی که مستلزم نمونه‌گیری از جوامع انسانی است تعداد واحدهای شمارش در خوشه‌ها با هم متفاوت‌اند، وضعیتهایی هم برای نمونه‌گیری پیش می‌آیند که در آنها تعداد واحدهای فهرست‌برداری در همهٔ خوشه‌ها یکسان است. ما بحث خود را با این وضعیت خاص شروع می‌کنیم (هر چند که فراوانی رویداد آن کمتر است) زیرا بسیاری از فرمولهای برآورد برای نمونه‌گیری خوشه‌ای دومرحله‌ای بسیار ساده هستند و خواننده می‌تواند بینشی قابل ملاحظه از این فرمولهای با پیچیدگی کمتر به دست آورد. مثالهایی چند از خوشه‌هایی که دارای تعداد واحدهای فهرست‌برداری یکسان‌اند در زیر فهرست می‌شود:

- هفته‌های سال تقویمی ممکن است در اوضاعی خاص به عنوان خوشه‌ها به کار روند و روزها به عنوان واحدهای فهرست‌برداری در نظر گرفته شوند. در چنین وضعیتهایی، خوشه‌ها (هفته‌ها) دارای تعدادی یکسان از واحدهای فهرست‌برداری (روزها) خواهند بود.

- سوابق بهداشتی گاهی روی نوارهای رایانه‌ای به صورت حلقه‌ای تنظیم می‌شوند که هر حلقه محتوی تعداد یکسانی از سابقه‌هاست. از این رو، در یک طرح نمونه‌گیری خوشه‌ای که در آن از حلقه‌ها به عنوان خوشه‌ها و از یکایک سوابق به عنوان واحدهای فهرست‌برداری استفاده می‌شود، هر خوشه (حلقه) تعدادی یکسان از واحدهای فهرست‌برداری (سوابق) خواهد داشت.
- در نمونه‌گیری کنترل کیفیت یا کنترل فرایند، واحدهای تکی یک محصول را غالباً در بسته‌هایی قرار می‌دهند که هر بسته به عنوان یک خوشه و واحدهای تکی به عنوان واحدهای فهرست‌برداری به کار می‌روند. معمولاً بسته‌ها محتوی تعدادی ثابت از واحدها هستند.

۱.۱.۱۰ چگونگی گرفتن یک نمونهٔ خوشه‌ای دومرحله‌ای ساده

برای نشان دادن اینکه چگونه یک نمونهٔ خوشه‌ای دومرحله‌ای ساده را می‌توان گرفت از یک مثال ساده استفاده می‌کنیم.

مثال تشریحی: در این مثال از داده‌های جدول ۱.۱۰ استفاده خواهیم کرد. فرض می‌کنیم که یک ادارهٔ بهداشت محلی، پنج مرکز بهداشت محلی را اداره می‌کند و هر مرکز بهداشت محلی برای مراقبتهای اولیه از سه پرستار متخصص استفاده می‌کند. باز هم فرض می‌کنیم که می‌خواهیم نمونه‌ای متشکل از سه مرکز و در داخل هر مرکز یک زیرنمونهٔ متشکل از دو پرستار متخصص شاغل در آن را انتخاب کنیم. از هر پرستاری که در نمونه انتخاب می‌شود درخواست خواهد شد که دربارهٔ بیمارانی که در طول یک هفتهٔ خاص می‌بیند و از تعداد آنهایی که به پزشک ارجاع می‌دهد گزارش مبسوطی تهیه کند.

برای گرفتن یک نمونهٔ خوشه‌ای دومرحله‌ای ساده متشکل از سه مرکز بهداشت محلی و دو پرستار متخصص از هر مرکز بهداشت که در نمونه انتخاب شده‌اند، ابتدا مراکز بهداشت را از ۱ تا ۵ شماره‌گذاری می‌کنیم و سپس سه شمارهٔ تصادفی را از میان ۱ و ۵ انتخاب می‌کنیم (مثلاً ۱، ۲ و ٤). بعد، در داخل هر یک از این سه مرکز نمونه، هر پرستار متخصص را با یک شماره مشخص می‌کنیم و سپس دو شماره به تصادف انتخاب می‌کنیم. پرستاران متخصص متناظر با این شماره‌ها در نمونه انتخاب می‌شوند (مثلاً پرستاران متخصص ۲ و ۳ از مرکز بهداشت محلی ۱، پرستاران متخصص ۱ و ۳ از مرکز بهداشت محلی ۲ و پرستاران متخصص ۱ و ۲ از مرکز بهداشت ٤).

□

جدول ۱.۱۰ برای پنج مرکز بهداشتی، تعداد بیمارانی که پرستاران متخصص دیده‌اند و تعدادی که به پزشک ارجاع شده‌اند

مرکز بهداشت	پرستار متخصص	تعداد بیماران دیده شده، x	تعداد بیماران ارجاع شده به پزشک، y
۱	۱	۵۸	۵
	۲	٤٤	٦
	۳	۱۸	٦
۲	۱	٤۲	۳
	۲	۵۳	۱۹
	۳	۱۰	۲
۳	۱	۱۳	۱۲
	۲	۱۸	٦
	۳	۳۷	۱۰
٤	۱	۱٦	۵
	۲	۳۲	۱٤
	۳	۱۰	٤
۵	۱	۲۵	۱۷
	۲	۲۳	۹
	۳	۲۳	۱٤

۲.۱.۱۰ برآورد کردن مشخصه‌های جامعه

همین که خوشه‌های نمونه و واحدهای فهرست‌برداری انتخاب شدند، داده‌های موردنیاز از همهٔ واحدهای فهرست‌برداری نمونه جمع‌آوری می‌شوند. همان طور که در فصل ۹ بحث شد، فرمهایی که برای جمع‌آوری این داده‌ها به کار می‌روند به ماهیت داده‌ها و نحوهٔ جمع‌آوری آنها بستگی دارند. سپس با استفاده از داده‌هایی که جمع‌آوری شده‌اند برآوردهای مشخصه‌های جامعه محاسبه می‌شوند.

شیوه‌های برآورد را با توجه به یک مثال بررسی می‌کنیم.

مثال تشریحی: فرض کنید که از روی نمونهٔ متشکل از سه مرکز بهداشت محلی و دو پرستار متخصص منتخب از هر یک از مراکز بهداشت، می‌خواهیم مشخصه‌های جامعه‌ای زیر را برآورد کنیم.

۱. کل تعداد بیماران دیده شده توسط پرستاران متخصص در پنج مرکز بهداشت محلی.

۲. کل تعداد بیماران ارجاع داده شده به پزشکان توسط پرستاران متخصص در پنج مرکز بهداشت محلی.
۳. نسبت ارجاع‌شدگان به پزشک در میان همهٔ کسانی که توسط پرستاران متخصص دیده شده‌اند.
٤. میانگین تعداد اشخاص دیده شده توسط پرستاران متخصص به ازای هر مرکز بهداشت محلی.
۵. میانگین تعداد اشخاص دیده شده به ازای هر پرستار متخصص.

داده‌ها را می‌توان با به دست آوردن کل تعداد بیماران دیده شده و کل تعداد ارجاع‌شدگان به پزشکان از روی سوابق هر پرستار متخصص انتخاب شده برای نمونه، جمع‌آوری کرد. فرض کنید که این داده‌ها در جدول ۲.۱۰ خلاصه شده باشند. در این صورت برآوردهای مطلوب را از خلاصه داده‌های نشان داده شده در جدول ۲.۱۰ به شرح زیر به دست می‌آوریم.

برای یافتن برآورد کل تعداد بیماران دیده شده توسط پرستاران متخصص در پنج مرکز بهداشت محلی، ابتدا کل تعداد بیماران دیده شده توسط همهٔ پرستاران متخصص در نمونه را محاسبه می‌کنیم. سپس این مجموع را به کسر نمونه‌گیری کل تقسیم می‌کنیم. یعنی،

$$\text{کل بیماران دیده شده} = ۱٦۲$$

$$\text{کسر نمونه‌گیری کل} = \frac{۲}{۳} \times \frac{۳}{۵} = \frac{۶}{۱۵} \quad \text{همهٔ پرستاران} = \text{همهٔ مراکز}$$

$$\text{برآورد کل جامعهٔ اشخاص دیده شده توسط پرستاران} = \frac{۱۶۲}{\frac{۶}{۱۵}} = ۴۰۵$$

برای پیدا کردن کل تعداد برآورد شدهٔ بیماران ارجاع شده به پزشکان توسط پرستاران متخصص در پنج مرکز بهداشت محلی، از محاسباتی مشابه محاسبات قبلی استفاده می‌کنیم:

$$\text{کل بیماران ارجاع شده به پزشک} = ۳٦$$

$$\text{کسر نمونه‌گیری کل} = \frac{۶}{۱۵}$$

$$\text{برآورد کل تعداد بیماران ارجاع شده به پزشک} = \frac{۳۶}{\frac{۶}{۱۵}} = ۹۰$$

برای پیدا کردن برآورد نسبت ارجاع شدگان به پزشکان از میان همهٔ بیماران دیده شده توسط پرستاران متخصص، نسبت دو مجموع برآورد شده را که در بالا محاسبه شد حساب می‌کنیم. یعنی،

$$\text{برآورد نسبت ارجاع‌شدگان به پزشک} = \frac{۹۰}{۴۰۵} = ۰/۲۲۲۲$$

جدول ۲.۱۰ خلاصه داده‌های مربوط به سه خوشهٔ انتخاب شده در نمونه

مرکز بهداشت	پرستار متخصص	تعداد بیماران دیده شده، x	تعداد بیماران ارجاع شده به پزشک، y
۱	۲	٤٤	٦
	۳	۱۸	٦
۲	۱	٤۲	۳
	۳	۱۰	۲
٤	۱	۱٦	٥
	۲	۳۲	۱٤
مجموع		۱٦۲	۳٦

بـرای پیـدا کـردن برآورد میانگین تعداد بیماران دیده شده توسط پرستاران متخصص به ازای مرکز بهداشـت محلی، برآورد کل تعداد بیماران دیده شده توسط پرستاران متخصص در همهٔ مراکز بهداشت محلی را به تعداد مراکز بهداشت محلی تقسیم می‌کنیم. یعنی،

$$\text{برآورد میانگین تعداد بیماران دیده شده توسط پرستاران به ازای مرکز بهداشت} = \frac{۴۰۵}{۵} = ۸۱$$

بـرای پیـدا کردن برآورد میانگین تعداد بیماران دیده شده به ازای هر پرستار متخصص، برآورد کل تعـداد بیمـاران دیـده شـده توسـط پرسـتاران متخصـص درهمـهٔ مراکز بهداشت محلی را به کل تعداد پرستاران متخصص تقسیم می‌کنیم. یعنی،

$$\text{برآورد میانگین تعداد بیماران دیده شده به ازای پرستار متخصص} = \frac{۴۰۵}{۱۵} = ۲۷$$

نمـادگذاری مـورد اسـتفاده در نمونـه‌گیـری خوشـه‌ای دومـرحله‌ای سـاده از خوشه‌هایی با تعداد واحدهای فهرست‌برداری برابر، به نمادگذاری مورد استفاده در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده بسیار شبیه است. این نمادگذاری در تابلوی ۱.۱۰ نشان داده شده است.

□

۳.۱.۱۰ برآورد کردن خطای معیارها

با استفاده از نمادگذاری تابلوی ۱.۱۰، فرمولهای جبری مورد استفاده برای محاسبهٔ برآورد مشخصه‌های جامعـه‌ای و بـرآورد خطای معیارهای برآوردگرهای مشخصه‌های جامعه‌ای را در تابلوی ۲.۱۰ فهرست مـی‌کـنیم. بـرآورد واریانسـها را به عنوان *برآوردهای واریانس خوشه‌ای نهایی* می‌شناسند و در کتابهای درسی توسط هنسن، هورویتس، و مادو مورد بحث قرار گرفته‌اند [۷]. اینان نشان داده‌اند که تا زمانـــی

تابلوی ۱.۱۰ نمادگذاری مورد استفاده در نمونه‌گیری خوشه‌ای دومرحله‌ای ساده

M = تعداد خوشه‌ها در جامعه

m = تعداد خوشه‌ها در نمونه

x_{ij} = سطح مشخصهٔ x برای واحد فهرست‌برداری نمونه‌ای jدر خوشهٔ نمونه‌ای i

y_{ij} = سطح مشخصهٔ y برای واحد فهرست‌برداری نمونه‌ای jدر خوشهٔ نمونه‌ای i

$\bar{n}$ = تعداد واحدهای فهرست‌برداری نمونه‌گیری شده از هر خوشه

$n = m\bar{n}$ = کل تعداد واحدهای فهرست‌برداری در نمونه

$x_i = \sum_{j=1}^{\bar{n}} x_{ij}$ = مجموع نمونه‌ای مشخصهٔ x برا ی i امین خوشهٔ نمونه

$y_i = \sum_{j=1}^{\bar{n}} y_{ij}$ = مجموع نمونه‌ای مشخصهٔ y برا ی i امین خوشهٔ نمونه

$x = \sum_{i=1}^{m} x_i$ = مجموع نمونه‌ای برای مشخصهٔ x

$y = \sum_{i=1}^{m} y_i$ = مجموع نمونه‌ای برای مشخصهٔ y

$\bar{N} = \frac{N}{M}$ = متوسط تعداد واحدهای فهرست‌برداری به ازای خوشه در جامعه

$N = M\bar{N}$ = کل تعداد واحدهای فهرست‌برداری در جامعه

$f_1 = \frac{m}{M}$ = کسر نمونه‌گیری مرحلهٔ اول

$f_2 = \frac{\bar{n}}{\bar{N}}$ = کسر نمونه‌گیری مرحلهٔ دوم

$f = f_1 f_2$ = کسر نمونه‌گیری کل

$\bar{x} = \frac{x}{m}$ = متوسط سطح مشخصهٔ x به ازای خوشه در نمونه

$\bar{y} = \frac{y}{m}$ = متوسط سطح مشخصهٔ y به ازای خوشه در نمونه

کـه دو یـا چـند خوشـه در نمونه قرار دارند و نمونه‌گیری مرحلهٔ دوم در یک خوشه به نمونه‌گیری در خوشـه‌هـای دیگر بستگی ندارد، روش مزبور برآوردی سازگار از واریانس را نتیجه می‌دهد. باید توجه داشـت کـه این برآوردهای خوشه‌ای نهایی بر مبنای برزیدن (دستکاری) مجموعها، x_i، روی واحدهای فهرست‌برداری انتخاب شده در هر خوشهٔ نمونه متکی هستند. همچنین، میانگین، $\bar{x}$، که در این فرمولها دیده می‌شود همان $\bar{x}_{clu}$، یعنی میانگین برآورد شده به ازای خوشه، نیست. در واقع، $\bar{x}_{clu} = (\bar{N}/\bar{n})\bar{x}$.

استفاده از این فرمولها با داده‌های ارائه شده در جدول ۲.۱۰ در مثال بعدی نشان داده شده است.

مثال تشریحی: برای استفاده از شیوه‌های برآورد نشان داده شده در تابلوی ۲.۱۰، داده‌های ارائه شده در جدول ۲.۱۰ را در نظر می‌گیریم. اطلاعات زیر را در اختیار داریم:

$m = ۳$ ، تعداد مراکز بهداشت محلی در نمونه

$M = ۵$ ، تعداد مراکز بهداشت محلی در جامعه

$\overline{n} = ۲$ ، تعداد پرستاران متخصص نمونه‌گیری شده از هر مرکز بهداشت منتخب

$\overline{N} = ۳$ ، تعداد پرستاران متخصص شاغل در هر مرکز بهداشت

$f_۱ = \frac{m}{M} = \frac{۳}{۵} = ۰/۶$ ، کسر نمونه‌گیری مرحلهٔ اول

$f_۲ = \frac{\overline{n}}{\overline{N}} = \frac{۲}{۳} = ۰/۶۷$ ، کسر نمونه‌گیری مرحلهٔ دوم

$f = f_۱ f_۲ = \frac{۳}{۵} \times \frac{۲}{۳} = ۰/۴$ ، کسر نمونه‌گیری کل

$n = ۶$ ، تعداد پرستار متخصص در نمونه

$N = ۱۵$ ، تعداد پرستار متخصص در جامعه

اول محاسبهٔ y'_{clu} را بررسی می‌کنیم که کل تعداد بیماران ارجاع شده به پزشک در میان بیمارانی است که پرستاران متخصص در پنج مرکز دیده‌اند. محاسبات برای y'_{clu} عبارت‌اند از

$$y_۱ = ۱۲ \qquad y_۲ = ۵ \qquad y_۳ = ۱۹ \qquad y = ۳۶$$

$$y'_{clu} = \frac{۳۶}{۰/۴} = ۹۰$$

محاسبات برای $\hat{SE}(y'_{clu})$ عبارت‌اند از

$$\bar{y} = \frac{۱۲+۵+۱۹}{۳} = ۱۲$$

$$\frac{\sum_{i=۱}^{m}(y_i - \bar{y})^۲}{m-۱} = \frac{(۱۲-۱۲)^۲+(۵-۱۲)^۲+(۱۹-۱۲)^۲}{۳-۱} = ۴۹$$

$$\hat{SE}(y'_{clu}) = \left[\frac{۵}{\sqrt{۳} \times ۰/۶۷}\right] \times \sqrt{۴۹} \times \sqrt{\frac{۱۵-۶}{۱۵}} = ۲۳/۴۸$$

بازهٔ اطمینان ۹۵٪ برای Y چنین است

$$y'_{clu} - ۱/۹۶ \times \hat{SE}(y'_{clu}) \le Y \le y'_{clu} + ۱/۹۶ \times \hat{SE}(y'_{clu})$$

$$۹۰ - ۱/۹۶ \times ۲۳/۴۸ \le Y \le ۹۰ + ۱/۹۶ \times ۲۳/۴۸$$

$$۴۳/۹۸ \le Y \le ۱۳۶/۰۲$$

تابلوی ۲.۱۰ برآورد مشخصه‌های جامعه‌ای و برآورد خطاهای معیار برای نمونه‌گیری خوشه‌ای دومرحله‌ای ساده

مجموع، X

$$x'_{clu} = \frac{x}{f} \qquad \hat{SE}(x'_{clu}) = \left(\frac{M}{\sqrt{m} f_۲}\right)\left[\frac{\sum_{i=۱}^{m}(x_i - \bar{x})^۲}{m-۱}\right]^{۱/۲}\left(\frac{N-n}{N}\right)^{۱/۲}$$

میانگین به ازای خوشه، $\bar{X}$

$$\bar{x}_{clu} = \frac{x'_{clu}}{M} \qquad \hat{SE}(\bar{x}_{clu}) = \left(\frac{۱}{\sqrt{m} f_۲}\right)\left[\frac{\sum_{i=۱}^{m}(x_i - \bar{x})^۲}{m-۱}\right]^{۱/۲}\left(\frac{N-n}{N}\right)^{۱/۲}$$

میانگین به ازای واحد فهرست‌برداری، $\bar{\bar{X}}$

$$\bar{\bar{x}}_{clu} = \frac{x'_{clu}}{N} \qquad \hat{SE}\left(\bar{\bar{x}}_{clu}\right) = \left(\frac{۱}{\bar{N}\sqrt{m} f_۲}\right)\left[\frac{\sum_{i=۱}^{m}(x_i - \bar{x})^۲}{m-۱}\right]^{۱/۲}\left(\frac{N-n}{N}\right)^{۱/۲}$$

نسبت، R

$$r_{clu} = \frac{x}{y}$$

$$\hat{SE}(r_{clu}) = r_{clu}\sqrt{\frac{N-n}{Nm}} \times \left(\frac{\sum_{i=۱}^{m}(x_i - \bar{x})^۲}{(m-۱)\bar{x}^۲} + \frac{\sum_{i=۱}^{m}(y_i - \bar{y})^۲}{(m-۱)\bar{y}^۲} - ۲\frac{\sum_{i=۱}^{m}(x_i - \bar{x})(y_i - \bar{y})}{(m-۱)\bar{x}\,\bar{y}}\right)^{۱/۲}$$

نمادگذاری مورد استفاده در این فرمولها در تابلوی ۱.۱۰ تعریف شده‌اند.

حالا به بررسی برآورد کردن $\bar{Y}$، میانگین آن تعداد بیمارانی می‌پردازیم که به ازای مرکز بهداشت به وسیلهٔ پرستار متخصص به پزشک ارجاع شده‌اند. محاسبهٔ $\bar{y}_{clu}$، برآورد $\bar{Y}$، عبارت است از

$$\bar{y}_{clu} = \frac{۹۰}{۵} = ۱۸$$

محاسبهٔ $\hat{SE}(\bar{y}_{clu})$ عبارت است از

$$\hat{SE}(\bar{y}_{clu}) = \left(\frac{۱}{\sqrt{۳} \times ۰/۶۷}\right) \times \sqrt{۴۹} \times \sqrt{\frac{۱۵ - ۶}{۱۵}} = ۴/۷۰$$

بازهٔ اطمینان ۹۵ درصدی برای $\bar{Y}$ چنین است

$$\bar{y}_{clu} - ۱/۹۶ \times \hat{SE}(\bar{y}_{clu}) \le \bar{Y} \le \bar{y}_{clu} + ۱/۹۶ \times \hat{SE}(\bar{y}_{clu})$$
$$۱۸ - ۱/۹۶ \times ۴/۷۰ \le \bar{Y} \le ۱۸ + ۱/۹۶ \times ۴/۷۰$$
$$۸/۷۸ \le \bar{Y} \le ۲۷/۲۱$$

حالا به برآورد کردن $\bar{\bar{Y}}$، میانگین تعداد بیماران ارجاع شده به پزشک به ازای پرستار متخصص می‌پردازیم. محاسبهٔ $\bar{\bar{y}}_{clu}$، برآورد $\bar{\bar{Y}}$، عبارت است از

$$\bar{\bar{y}}_{clu} = \frac{۹۰}{۱۵} = ۶$$

محاسبهٔ $\hat{SE}\left(\bar{\bar{y}}_{clu}\right)$ چنین است

$$\hat{SE}\left(\bar{\bar{y}}_{clu}\right) = \left[\frac{۱}{۳\sqrt{۳} \times ۰/۶۷}\right] \times \sqrt{۴۹} \times \sqrt{\frac{۱۵ - ۶}{۱۵}} = ۱/۵۷$$

بازهٔ اطمینان ۹۵ درصدی برای $\bar{\bar{Y}}$ عبارت است از

$$\bar{\bar{y}}_{clu} - ۱/۹۶ \times \hat{SE}\left(\bar{\bar{y}}_{clu}\right) \le \bar{\bar{Y}} \le \bar{\bar{y}}_{clu} + ۱/۹۶ \times \hat{SE}\left(\bar{\bar{y}}_{clu}\right)$$
$$۶ - ۱/۹۶ \times ۱/۵۷ \le \bar{\bar{Y}} \le ۶ + ۱/۹۶ \times ۱/۵۷$$
$$۲/۹۲ \le \bar{\bar{Y}} \le ۹/۰۸$$

بالاخره به بررسی برآورد کردن r_{clu}، نسبت همهٔ بیماران ارجاع شده به پزشک از میان کسانی که پرستاران متخصص آنها را دیده‌اند، می‌پردازیم: محاسبات مربوط به r_{clu} عبارت‌اند از

$$y = ۳۶ \qquad x_۱ = ۶۲ \qquad x_۲ = ۵۲ \qquad x_۳ = ۴۸ \qquad x = ۱۶۲$$

$$r_{clu} = \frac{۳۶}{۱۶۲} = ۰/۲۲۲۲$$

محاسبات مربوط به $\hat{SE}(r_{clu})$ چنین‌اند

$$\bar{x} = \frac{۱۶۲}{۳} = ۵۴ \qquad \bar{y} = \frac{۳۶}{۳} = ۱۲$$

$$\frac{\sum_{i=۱}^{m}(y_i - \bar{y})^۲}{m-۱} = \frac{(۱۲-۱۲)^۲ + (۵-۱۲)^۲ + (۱۹-۱۲)^۲}{۳-۱} = ۴۹$$

$$\frac{\sum_{i=۱}^{m}(x_i - \bar{x})^۲}{m-۱} = \frac{(۶۲-۵۴)^۲ + (۵۲-۵۴)^۲ + (۴۸-۵۴)^۲}{۳-۱} = ۵۲$$

$$\frac{\sum_{i=۱}^{m}(x_i - \bar{x})(y_i - \bar{y})}{m-۱} = \frac{(۶۲-۵۴)(۱۲-۱۲) + (۵۲-۵۴)(۵-۱۲) + (۴۸-۵۴)(۱۹-۱۲)}{۳-۱}$$

$$= -۲۸$$

$$\hat{SE}(r_{clu}) = ۰/۲۲۲۲ \sqrt{\frac{۱۵-۶}{(۱۵)(۳)}\left(\frac{۱۰۴}{(۳-۱)۵۴^۲} + \frac{۹۸}{(۳-۱)۲^۲} - ۲\frac{(-۲۸)}{(۳-۱)(۵۴)(۱۲)}\right)}$$

$$= ۰/۰۶۱۲$$

بازهٔ اطمینان ۹۵٪ برای R عبارت است از

$$r_{clu} - ۱/۹۶ \times \hat{SE}(r_{clu}) \le R \le r_{clu} + ۱/۹۶ \times \hat{SE}(r_{clu})$$

$$۰/۲۲۲۲ - ۱/۹۶ \times ۰/۰۶۱ \le R \le ۰/۲۲۲۲ + ۱/۹۶ \times ۰/۰۶۱$$

$$۰/۱۰۲ \le R \le ۰/۳۴۲$$

□

برآورد کردن مشخصه‌های جامعه و خطای معیارهای آنها را می‌توان با استفاده از نرم‌افزار آماری مناسب اجرا کرد و ما این را برای STATA و SUDAAN در زیر نشان می‌دهیم. هر یک از این دو نرم‌افزار را می‌توان در پروندهٔ اطلاعاتی *IL10PT1.DTA* (برای STATA) و *IL10PT1.SSD* (برای SUDAAN) مورد استفاده قرار داد. این پرونده‌ها از شش سابقه تشکیل شده‌اند (یکی برای هر پرستار نمونه‌گیری شده) و شامل داده‌های زیرند:

CENTER	NURSE	M	NBAR	W	NPATNTS	NREFRRED
1	2	5	3	2.5	44	6
1	3	5	3	2.5	18	6
2	1	5	3	2.5	42	3
2	3	5	3	2.5	10	2
4	1	5	3	2.5	16	5
4	2	5	3	2.5	32	14

هر سابقه شامل متغیرهای زیر است:

CENTER: مرکز بهداشت محلی خاص

NURSE: پرستار متخصص خاص نمونه‌گیری شده

M: کل تعداد مراکز بهداشت محلی در جامعه

NBAR: کل تعداد پرستاران متخصص در هر مرکز بهداشت محلی

W: وزن نمونه‌گیری کل (معکوس کسر نمونه‌گیری کل)

NPATNTS: تعداد بیماران دیده شده توسط هر پرستار متخصص نمونه‌گیری شده

NREFRRED: کل تعداد بیماران ارجاع شده به پزشک توسط هر پرستار متخصص نمونه‌گیری شده

فرمانها برای برآورد کردن با استفاده از STATA

```
Use " a:i110pt1.dta", clear
. svyset psu center
. svyset pweight w
. svytotal nrefrred
. svyratio nrefrred npatnts
```

اولین فرمان، پروندهٔ اطلاعاتی را که قرار است مورد استفاده قرار گیرد شناسایی می‌کند. دو فرمان بعدی نشانهٔ آن‌اند که واحد نمونه‌گیری اولیه در متغیر CENTER تعیین شده و وزن نمونه‌گیری در متغیر W ارائه شده است. فرمان بعدی برای برآورد کردن کل تعداد بیماران ارجاع شده به پزشک، و فرمان آخر برای برآورد کردن نسبت همهٔ بیماران ارجاع شده به پزشک است.

فرمانها برای برآورد کردن با استفاده از SUDAAN

```
PROC DESCRIPT DATA = IL10PT1 FILETYPE = SAS  DESIGN = WOR  MEANS TOTALS;
NEST_ONE_CENTER;
WEIGHT  W;
TOTCNT  M  NBAR;
VAR  NPATNTS  NREFRRED;
SETENV  COLWIDTH = 13;
SETENV  DECWIDTH = 3;

PROC RATIO DATA = IL10PT1 FILETYPE = SAS  DESIGN = WOR;
NEST_ONE_CENTER;
WEIGHT  W;
TOTCNT  M  NBAR;
NUMBER  NREFRRED;
DENOM  NPATNTS;
SETENV  COLWIDTH = 13;
SETENV  DECWIDTH = 3;
```

دو مجموعـه از فـرمانها وجود دارند: یک مجموعه، برآورد کل تعداد بیماران ارجاع شده به پزشک را همـراه بـا برآورد خطای معیار این برآورد تولید می‌کند و مجموعهٔ دوم فرمانها، برآورد نسبتی نسبت افراد ارجاع شده به پزشک (و خطای معیار این برآورد) را تولید می‌کند.

فـرمان اول تعییـن می‌کنـد که مدول PROC DESCRIPT در نرم‌افزار SUDAAN مورد استفـاده قرار خواهد گرفت و قرار است از پروندهٔ IL10PT1.SSD استفاده شود که در پروندهٔ داده‌های SAS اسـت و قـرار است میانگینها و مجموعها تولید شوند؛ و طرح نمونه‌ای خاصی که به کار گرفـته مـی‌شـود WOR نـام دارد (کـه نمـادی اسـت بـرای «بـدون جـایگذاری»). ایـن در اصـطلاح‌شناسـی SUDAAN بـه آن معـنی اسـت کـه در مرحلهٔ اول نمونه‌گیری، خوشه‌ها با احـتمال برابـر بدون جایگذاری نمونه‌گیری می‌شوند و نمونه‌گیری در مراحل بعدی با احتمال برابـر بـا جـایگذاری یا بدون جایگذاری خواهد بود. این فرمان همچنین حاکی از آن است که قرار است میانگینها و مجموعها همراه با خطای معیارهای آنها برآورد شوند.

فرمان دوم نشان می‌دهد که تمام جامعه، یک طبقـهٔ تک در نظر گرفتـه شـــده است و واحـــدهای نمونه‌گیری اولیه در متغیر *CENTER* تعیین شده‌اند.

فرمان سوم مشخص می‌کند که وزن نمونه‌گیری برای هر سابقه در متغیر *W* قرار دارد.

فرمان چهـارم نشان می‌دهـــد که کل تعـــداد خوشـــه‌ها با متغیـر *M* تعییـن شده و کل تعـــداد واحدهای شمارش (در این مورد، پرستاران متخصص) با متغیر *NBAR* مشخص شده است.

فرمـان پنجـم نشان می‌دهـــد کـــه برآوردهایی برای متغیرهای *NPATNTS* و *NREFRRED* قرار است به دست آید.

حکمهای ششم و هفتم نشان‌دهندهٔ شکل خروجی‌اند.

هشـت فـرمان پایـانی مشـخص می‌کنند که برای برآورد کردن نسبت بیماران ارجاع شده به پزشک توسـط پرسـتار متخصص قرار است از برآورد نسبتی استفاده شود. معنی این فرمانها از مبحث قبل روشن می‌شود.

فـرمانهای SUDAAN کـه در بالا نشان داده شدند برای طرحی نمونه‌ای مناسب‌اند که فرض را بر ایـن مـی‌گـذارد کـه نمونـه‌گیـری در هـر مـرحله بـدون جـایگذاری است. اگر فرض بر این باشد که نمونـه‌گیری در هر مرحله با جایگذاری است (یا کسرهای نمونه‌گیری در هر مرحله بسیار کوچک‌اند)، آنگاه می‌توان از مجموعه فرمانهای زیر استفاده کرد:

```
PROC DESCRIPT DATA = IL10PTI FILETYPE = SAS DESIGN = WR MEANS TOTALS;
NEST_ONE_CENTER;
WEIGHT W;
VAR NPATNTS NREFRRED;
SETENV COLWIDTH = 13;
SETENV DECWIDTH = 3;

PROC RATIO DATA = IL10PTI FILETYPE = SAS DESIGN = WR;
NEST_ONE_CENTER;
WEIGHT W;
NUMBER NREFRRED;
DENOM NPATNTS;
SETENV COLWIDTH = 13;
SETENV DECWIDTH = 3;
```

همهٔ شیوه‌هایی که تا اینجا از آنها بحث شد، یعنی فرمولهای تابلوی۲.۱۰، STATA، SUDAAN با فرض نمونه‌گیری بدون جایگذاری در هر مرحله، و SUDAAN با فرض نمونه‌گیری با جایگذاری در هر مرحله، برآوردهای نقطه‌ای یکسانی تولید خواهند کرد. ولی در برآورد خطای معیارهای این برآوردها تفاوتهایی وجود خواهد داشت که ذیلاً نشان داده می‌شوند:

برآورد خطای معیارهای برآوردها

برآورد	فرمولها در تابلوی ۲.۱۰ در متن	فرمولها در تابلوی ۲.۱۰ در متن بدون تصحیح جامعهٔ متناهی (fpc)	STATA	SUDAAN با فرض طرح بدون جایگذاری	SUDAAN با فرض طرح با جایگذاری
کل بیماران ارجاع شده به پزشک	۲۳/٤۸	۳۰/۳۱	۳۰/۳۱	۲۱/٦۸	۳۰/۳۱
نسبت بیماران ارجاع شده به پزشک	۰/۰٦۱۲	۰/۰۸۱	۰/۰۸۱	۰/۰۵۸	۰/۰۸۱

تعجبی ندارد که خطاهای معیار برآورد کل بیماران ارجاع شده به پزشک درSUDAAN با استفاده از طرح نمونه‌گیری با جایگذاری برای STATA و برای برآوردهای خوشه‌ای نهایی بدون تصحیح جامعهٔ متناهی (فرمولهای تابلوی۲.۱۰ بدون عامل fpc ، $\sqrt{(N-n)/N}$) یکسان‌اند. همان طور که انتظار می‌رود، اینها برای طرح SUDAAN بدون جایگذاری و برای برآورد خوشه‌ای نهایی با منظور نمودن تصحیح جامعهٔ متناهی در سطح پایینتری قرار دارند. دو برآورد حاصل از SUDAAN بدون جایگذاری و خوشه‌ای نهایی به یکدیگر نزدیک‌اند ولی یکسان نیستند. این موضوع ناشی از این واقعیت است که

در طرحهایی که نمونه‌گیری را در مرحلهٔ اول بدون جایگذاری فرض می‌کنند، SUDAAN به جای برآورد خوشه‌ای نهایی از برآوردی براساس مؤلفه‌های واریانس برآورد شده در مرحلهٔ اول و دوم استفاده می‌کند.

برای برآورد نسبتی، SUDAAN با جایگذاری، STATA، و برآورد خوشه‌ای نهایی بدون استفاده از عامل تصحیح جامعهٔ متناهی، با هم توافق دارند. برآورد واریانس برآورد نسبتی در SUDAAN بدون جایگذاری و برآورد خوشه‌ای نهایی با عامل تصحیح جامعهٔ متناهی، به دلیل استفاده از یک عامل تصحیح جامعهٔ متناهی،fpc، همان طور که انتظار می‌رود، در سطحی پایینتر نسبت به سایرین قرار دارند. در مورد برآوردهای واریانس برای مجموعها، SUDAAN از یک برآورد خوشه‌ای نهایی واریانس برآورد نسبتی استفاده نمی‌کند؛ از این رو، این مقدار با مقدار به دست آمده با استفاده از فرمولهای تابلوی۲.۱۰ توافق ندارد هر چند که به آن مقدار بسیار نزدیک می‌شود (۰/۰۵۸ در مقابل ۰/٦۱۲).

٤.۱.۱۰ توزیع نمونه‌گیری برآوردها

یک نمونهٔ خوشه‌ای دومرحله‌ای ساده را در نظر بگیرید که در آن m خوشه در مرحلهٔ اول از M خوشهٔ موجود در جامعه انتخاب شده‌اند، و از داخل هر خوشهٔ نمونه، $\overline{n}$ واحد فهرست‌برداری از $\overline{N}$ واحد فهرست‌برداری موجود در خوشه انتخاب شده‌اند. کل تعداد نمونه‌های ممکن از رابطهٔ زیر به دست می‌آید

$$\binom{M}{m}\left(\frac{\overline{N}}{\overline{n}}\right)^m$$

به عنوان مثال به داده‌های جدول ۱.۱۰ مراجعه می‌کنیم. اگر یک نمونهٔ خوشه‌ای دومرحله‌ای ساده از دو مرکز بهداشت را به عنوان نمونهٔ مرحلهٔ اول از پنج مرکز موجود در جامعه انتخاب کنیم و دو پرستار متخصص را از سه پرستار موجود در هر مرکز نمونه به عنوان نمونهٔ مرحلهٔ دوم بگیریم، می‌توانیم $\binom{۵}{۲}\left(\frac{۳}{۲}\right)^۲ = ۱۰\times ۳^۲ = ۹۰$ نمونهٔ ممکن انتخاب کنیم. از روی هر یک از این ۹۰ نمونه می‌توانیم مشخصه‌های جامعه‌ای از قبیل مجموعها، میانگینها، و نسبتها را برآورد کنیم.

در این قسمت، خواص توزیعهای نمونه‌گیری برآورد مشخصه‌های جامعه را مورد بحث قرار می‌دهیم. به خصوص $E(x'_{clu})$، $E(\overline{x}_{clu})$ و $E\left(\overline{\overline{x}}_{clu}\right)$، میانگینهای توزیع نمونه‌گیری مجموعهای برآورد شده، x'_{clu}، میانگینهای برآورد شده به ازای خوشه، $\overline{x}_{clu}$، و میانگینهای برآورد شده به ازای واحد فهرست‌برداری، $\overline{\overline{x}}_{clu}$، با پارامترهای جامعه‌ای متناظر، X، $\overline{X}$ و $\overline{\overline{X}}$، که در تابلوی ۱.۹ تعریف شده‌اند

برابـرند. بـه عـبارت دیگـر، ایـنها برآوردگـرهای نااریب‌اند. از سوی دیگر، نسبتهای برآورد شده، r_{clu}، نـااریب نیسـتند، اما همان طور که قبلاً در مورد سایر طرحهای نمونه‌ای بحث شد، برای نمونه‌هایی که اندازهٔ آنها در حدی معقول بزرگ است عموماً اریبی کوچکی دارند. خطاهای معیار نظری این برآوردها در تابلوی ۳.۱۰ ارائه شده‌اند.

بـه ایـن تـرتیب مـی‌بیـنیم کـه واریانسـهای نمونـه‌گیـری برآورد مجموعها و میانگینهای حاصل از نمونـه‌گیـری خوشـه‌ای دومـرحله‌ای سـاده از دو جمله تشکیل می‌شوند که یکی بستگی دارد به مقدار σ^2_{1x}، واریـانس توزیـع مجموعهـای خوشـه‌ای X_i در میـان خوشـه‌ها و دیگری بستگی دارد به σ^2_{2x}، جملـه‌ای کـه معـرف واریـانس میـان واحدهـای فهرسـت‌بـرداری داخـل همان خوشه نسبت به سطح مشخصـهٔ x است. این جمله‌ها به ترتیب به عنوان مؤلفه‌های *مرحلهٔ اول* و *مرحلهٔ دوم* شناخته می‌شوند. اگر هر خوشه‌ای در نمونه گنجانده شود، یا به عبارت دیگر، اگر $m = M$، آن‌گاه مؤلفه‌های مرحلهٔ اول در عـبارتهای نشـان داده شـده در تابـلوی ۳.۱۰ حـذف مـی‌شوند، زیرا عامل $(M-m)(M-1)$ صفر مـی‌شـود. ایـن بـه طـور شـهودی هـم معقـول است زیرا اگر خوشه‌ای در نمونه گنجانده شود، آن‌گاه، تغییرپذیـری نمونـه‌گیـری ناشـی از تفاوتهـای موجود میان خوشه‌ها نباید عاملی در توزیع نمونه‌گیری بـرآوردها باشـد. پـس بـرآوردهای بـه دسـت آمده با برآوردهای حاصل از نمونه‌گیری طبقه‌بندی شده یکسان خواهند بود. به همین ترتیب، اگر در داخل هر خوشهٔ نمونه، هر واحد فهرست‌برداری در نمونه گـنجانده شـود، یـا بـه عـبارت دیگـر، اگر $\overline{n} = \overline{N}$، آن‌گاه مؤلفه‌های مرحلهٔ دوم در عبارتـــهایی که در تابـلوی ۳.۱۰ دیـده مـی‌شـوند حـذف خواهـند شـد، زیـرا عـامل $(\overline{N}-\overline{n})/(\overline{N}-1)$ به صفر تبـــدیل مـی‌شـــود و واریانسهای برآوردها با واریانسهای حاصل از نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده که در تابـلوی ۳.۹ نشـان داده شده‌اند یکسان می‌شوند. این نیز از نظر شهودی معقول است، زیرا در داخل هـر خوشـهٔ نمونـه، هـر واحـد فهرست‌برداری گنجانده شده است. از این رو، هیچ نمونه‌گیری مرحلهٔ دومی وجود ندارد و فرایند، در واقع، نمونه‌گیری خوشه‌ای یک‌مرحله‌ای است.

اگـر در تابلوی ۳.۱۰ مؤلفه‌های واریانس σ^2_{1x}، σ^2_{2x}، σ^2_{1y}، σ^2_{2y} و غیره، با برآوردگرهای مربوط خود جایگـزین شوند، عبارتهایی برای خطای معیارهای برآورد شدهٔ x'_{clu}، $\bar{x}_{clu}$، $\bar{\bar{x}}_{clu}$، و r_{clu} به دست خواهند آمـد کـه با برآوردگرهای واریانس خوشه‌ای نهایی ارائه شده در تابلوی۲.۱۰ فرق خواهند داشت. اریبی برآوردگـرهای مبتـنی بر مؤلفه‌ها کمتر از برآوردهای خوشه‌ای نهایی خواهد بود، ولی می‌تواند پایداری کمـتری نیز داشته باشد به خصوص هنگامی که $\overline{n}$، تعداد واحدهای فهرست‌برداری نمونه‌گیری شده در داخل خوشه، کم است.

حالا، استفاده از فرمولهای تابلوی ۳.۱۰ را با یک مثال نشان می‌دهیم.

مثال تشریحی: جامعه‌ای را که در جدول ۱.۱۰ نشان داده شده است در نظر می‌گیریم. برخی محاسبات ابتدایی دربارهٔ این جامعه در جدول ۳.۱۰ ارائه شده است. برای همهٔ محاسبات داریم

$$M = ۵ \qquad \overline{N} = ۳ \qquad N = ۱۵$$

براساس اطلاعات جدول ۳.۱۰ و فرمولهای تابلوی ۳.۱۰ می‌توانیم محاسبات زیر را انجام دهیم:

$$\sigma_{۱x}^{۲} = \frac{۲۸۳۷/۲۰}{۵} = ۵۶۷/۴۴ \qquad \sigma_{۱y}^{۲} = \frac{۶۷۷/۲}{۵} = ۱۳۵/۴۴ \qquad \sigma_{۱xy} = \frac{-۸۳۰/۸۰}{۵} = -۱۶۶/۱۶$$

با استفاده از اطلاعات جدول ٤.۱۰، محاسبات زیر را داریم:

$$\sigma_{۲x}^{۲} = \frac{۲۴۰۴/۰۱}{۱۵} = ۱۶۰/۲۷ \qquad \sigma_{۲y}^{۲} = \frac{۵۸۸/۰۱}{۱۵} = ۳۹/۲۰ \qquad \sigma_{۲xy} = \frac{۷۱۱/۶۵}{۱۵} = ۴۷/۴۴$$

با استفاده از فرمولهای ارائه شده در تابلوی ۳.۱۰، می‌توانیم خطای معیار نظری $\hat{SE}(x'_{clu})$ را برای x'_{clu}، کل تعداد بیماران دیده شده توسط ۱۵ پرستار متخصص که از روی نمونهٔ خوشه‌ای دومرحله‌ای متشکل از دو مرکز و دو پرستار در داخل هر مرکز نمونه برآورد شده است، محاسبه کنیم:

$$SE(x'_{clu}) = \left[\frac{۵^{۲}}{۲} \times ۵۶۷/۴۴ \times \left(\frac{۵-۲}{۵-۱}\right) + \frac{۱۵^{۲}}{۴} \times ۱۶۰/۲۷ \times \left(\frac{۳-۲}{۳-۱}\right)\right]^{۱/۲} = ۹۹/۱۳$$

به همین ترتیب، خطای معیارهای نظری $\bar{x}_{clu}$، y'_{clu} و $\bar{\bar{y}}_{clu}$ را می‌توان با جایگزین کردن پارامترهای محاسبه شده در جداول ۳.۱۰ و ٤.۱۰ در عبارتهای مناسب ارائه شده در تابلوی ۳.۱۰ به دست آورد.

□

جدول ۳.۱۰ کار برگه برای محاسبات مربوط به مجموعهای خوشه‌ای

خوشه	X_i	Y_i	$(X_i - \overline{X})^{۲}$	$(Y_i - \overline{Y})^{۲}$	$(X_i - \overline{X}) \times (Y_i - \overline{Y})$
۱	۱۲۰	۱۷	۱۲۶۷/۳۶	۱۷۹/۵۶	-٤۷۷/۰٤
۲	۱۰۵	۲٤	٤۲٤/۳۶	٤۰/۹۶	-۱۳۱/۸٤
۳	۶۸	٤۸	۲۶۸/۹۶	۳۰۹/۷۶	-۲۸۸/۶٤
٤	۵۸	۲۳	۶۹۶/۹۶	۵٤/۷۶	۱۹۵/۳۶
۵	۷۱	٤۰	۱۷۹/۵۶	۹۲/۱۶	-۱۲۸/۶٤
	$\overline{X} = ٤۲۲$	$\overline{Y} = ۱۵۲$	۲۸۳۷/۲۰	۶۷۷/۲۰	-۸۳۰/۸۰
	$\overline{\overline{X}} = ۸٤/٤$	$\overline{\overline{Y}} = ۳۰/٤$			
	$\overline{\overline{\overline{X}}} = ۲۸/۱۳$	$\overline{\overline{\overline{Y}}} = ۱۰/۱۳$			

تابلوی ۳.۱۰ خطاهای معیار برآوردهای جامعه‌ای، تحت نمونه‌گیری خوشه‌ای دومرحله‌ای ساده

مجموع، x'_{clu}

$$SE(x'_{clu}) = \left[\left(\frac{M^{۲}}{m}\right)\sigma^{۲}_{۱x}\left(\frac{M-m}{M-۱}\right)+\left(\frac{N^{۲}}{n}\right)\sigma^{۲}_{۲x}\left(\frac{\overline{N}-\overline{n}}{\overline{N}-۱}\right)\right]^{۱/۲}$$

میانگین به ازای خوشه، $\bar{x}_{clu}$

$$SE(\bar{x}_{clu}) = \left[\frac{\sigma^{۲}_{۱x}}{m}\times\left(\frac{M-m}{M-۱}\right)+\left(\frac{\overline{N}^{۲}}{n}\right)\sigma^{۲}_{۲x}\left(\frac{\overline{N}-\overline{n}}{\overline{N}-۱}\right)\right]^{۱/۲}$$

میانگین به ازای عنصر، $\overline{\overline{x}}_{clu}$

$$SE(\overline{\overline{x}}_{clu}) = \left[\left(\frac{۱}{\overline{N}^{۲}}\right)\times\frac{\sigma^{۲}_{۱x}}{m}\times\left(\frac{M-m}{M-۱}\right)+\left(\frac{\sigma^{۲}_{۲x}}{n}\right)\left(\frac{\overline{N}-\overline{n}}{\overline{N}-۱}\right)\right]^{۱/۲}$$

نسبت، r_{clu}

$$SE(r_{clu}) \approx R\left[\left(\frac{\sigma^{۲}_{۱R}}{m\overline{X}^{۲}}\right)\left(\frac{M-m}{M-۱}\right)+\left(\frac{\sigma^{۲}_{۲R}}{m\overline{n}\,\overline{\overline{X}}^{۲}}\right)\left(\frac{\overline{N}-\overline{n}}{\overline{N}-۱}\right)\right]^{۱/۲}$$

عبارت $\sigma^{۲}_{۱x}$ که در این فرمولها دیده می‌شود واریانس میان مجموعهای خوشه‌ای است که در فصل ۹، معادلهٔ (۱.۹)، به صورت زیر تعریف شده است.

$$\sigma^{۲}_{۱x} = \frac{\sum_{i=۱}^{M}(X_i-\overline{X})^{۲}}{M}$$

عبارت $\sigma^{۲}_{۲x}$ که در این عبارتها دیده می‌شود از فرمول زیر به دست می‌آید

$$\sigma^{۲}_{۲x} = \frac{\sum_{i=۱}^{M}\sum_{j=۱}^{\overline{N}}(X_{ij}-\overline{\overline{X}}_i)^{۲}}{N} \qquad (۱.۱۰)$$

ادامهٔ تابلوی ۳.۱۰

که در آن $\overline{\overline{X}}_i$، میانگین سطح مشخصهٔ x به ازای واحد فهرست‌برداری برای آن دسته از واحدهای فهرست‌برداری است که در خوشهٔ i قرار دارند و از فرمول زیر به دست می‌آید

$$\overline{\overline{X}}_i = \frac{\sum_{j=1}^{\overline{N}} X_{ij}}{\overline{N}}$$

فرمول مربوط به برآورد خطای معیار نسبت برآورد شدهٔ r_{clu} ، یک تقریب است که هرگاه $V(\bar{x}_{clu})$ ، ضریب تغییرات، مخرج کسر نسبت کمتر از ۰/۰۵ باشد معتبر است. پارامترهای σ^2_{1R} و σ^2_{2R} که در عبارت مربوط به $SE(r_{clu})$ دیده می‌شوند از فرمولهای زیر به دست می‌آیند

$$\sigma^2_{1R} = \sigma^2_{1y} + R^2\sigma^2_{1x} - 2R\sigma_{1xy} \qquad (۲.۱۰)$$

$$\sigma^2_{2R} = \sigma^2_{2y} + R^2\sigma^2_{2x} - 2R\sigma_{2xy} \qquad (۳.۱۰)$$

که در آنها عبارتهای σ_{1xy} و σ_{2xy} کوواریانسهای مرحلهٔ اول و مرحلهٔ دوم‌اند که از فرمولهای زیر به دست می‌آیند

$$\sigma_{1xy} = \frac{\sum_{i=1}^{M}(X_i - \overline{X})(Y_i - \overline{Y})}{M} \qquad (٤.۱۰)$$

$$\sigma_{2xy} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{\overline{N}}(X_{ij} - \overline{\overline{X}}_i)(Y_{ij} - \overline{\overline{Y}}_i)}{N} \qquad (۵.۱۰)$$

سایر نمادهای مورد استفاده در این فرمولها در تابلوهای ۱.۹ و ۱.۱۰ تعریف شده‌اند.

۵.۱.۱۰ نمونهٔ مورد نیاز چقدر باید بزرگ باشد؟

در نمونه‌گیری خوشه‌ای دومرحله‌ای ساده، به طوری که در بخش بعدی خواهیم دید، $\overline{n}$، تعداد مطلوب واحدهای فهرست‌برداری که باید در مرحلهٔ دوم از هر خوشهٔ نمونه‌گیری شده در مرحلهٔ اول انتخاب شود، براساس هزینه‌ها و نیز بر مبنای اندازه‌های نسبی مؤلفه‌های واریانس مرحلهٔ اول و دوم (مثلاً σ^2_{1x}، σ^2_{2x}) تعیین می‌شود. همین که تعداد $\overline{n}$ تثبیت شد ممکن است بخواهیم m کل تعداد خوشه‌هایی را تعیین کنیم که باید در مرحلهٔ اول نمونه‌گیری انتخاب شوند تا $(1-\alpha)\times 100\%$ مطمئن

جدول ٤.١٠ کار برگهٔ محاسبات مربوط به واحدهای فهرست‌برداری

خوشه	j	X_{ij}	$\overline{\overline{X}}_i$	$\left(X_{ij}-\overline{\overline{X}}_i\right)^2$	Y_{ij}	$\overline{\overline{Y}}_i$	$\left(Y_{ij}-\overline{\overline{Y}}_i\right)^2$	$\left(X_{ij}-\overline{\overline{X}}_i\right)\left(Y_{ij}-\overline{\overline{Y}}_i\right)$
۱	۱	۵۸		۳۲٤	۵		۰/٤۵	-۱۲/۰۷
	۲	٤٤	٤۰	۱٦	٦	۵/٦۷	۰/۱۱	۱/۳۲
	۳	۱۸		٤۸٤	٦		۰/۱۱	-۷/۲٦
۲	۱	٤۲		٤۹	۳		۲۵	-۳۵
	۲	۵۳	۳۵	۳۲٤	۱۹	۸	۱۲۱	۱۹۸
	۳	۱۰		٦۲۵	۲		۳٦	۱۵۰
۳	۱	۱۳		۹۳/۵۱	۱۲		۱٦	۳۸/٦۸
	۲	۱۸	۲۲/٦۷	۲۱/۸۱	٦	۱٦	۱۰۰	٤٦/۷۰
	۳	۳۷		۲۰۵/۳۵	۳۰		۱۹٦	۲۰۰/٦۲
٤	۱	۱٦		۱۱/۰۹	۵		۷/۱۳	۸/۸۹
	۲	۳۲	۱۹/۳۳	۱٦۰/۵۳	۱٤	۷/٦۷	٤۰/۰۷	۸۰/۲۰
	۳	۱۰		۸۷/۰۵	٤		۱۳/٤۷	۳٤/۲٤
۵	۱	۲۵		۱/۷۷	۱۷		۱۳/٤۷	٤/۸۸
	۲	۲۳	۲۳/٦۷	۰/٤۵	۹	۱۳/۳۳	۱۸/۷۵	۲/۹۰
	۳	۲۳		۰/٤۵	۱٤		۰/٤۵	-۰/٤۵
				۲٤۰٤/۰۱			۵۸۸/۰۱	۷۱۱/٦۵

شویم که برآوردهایی که به دست می‌آوریم بیشتر از ε از مقدار واقعی مشخصهٔ جامعه‌ای مورد برآورد تفاوت ندارند. فرمولهای مورد استفاده برای محاسبهٔ m، تعداد *خوشه‌های نمونه‌گیری شده* که این ویژگیها را تأمین می‌کنند ذیلاً ارائه شده‌اند.

برای برآورد *مجموعها* (x'_{clu}) یا *میانگینها* ($\bar{x}_{clu}$، $\bar{\bar{x}}_{clu}$)

$$m=\frac{\left(\frac{\sigma_{1x}^{2}}{\bar{X}^{2}}\right)\times\left(\frac{M}{M-1}\right)+\left(\frac{1}{\bar{n}}\right)\times\left(\frac{\sigma_{2x}^{2}}{\bar{\bar{X}}^{2}}\right)\times\left(\frac{\bar{N}-\bar{n}}{\bar{N}-1}\right)}{\frac{\varepsilon^{2}}{z_{1-(\alpha/2)}^{2}}+\frac{\sigma_{1x}^{2}}{\bar{X}^{2}(M-1)}} \qquad (6.10)$$

و برای برآورد *نسبتها* ($r_{clu}=y/x$)

$$m=\frac{\left(\frac{\sigma_{1R}^{2}}{\bar{X}^{2}}\right)\times\left(\frac{M}{M-1}\right)+\left(\frac{1}{\bar{n}}\right)\times\left(\frac{\sigma_{2R}^{2}}{\bar{\bar{X}}^{2}}\right)\times\left(\frac{\bar{N}-\bar{n}}{\bar{N}-1}\right)}{\frac{\varepsilon^{2}}{z_{1-(\alpha/2)}^{2}}+\frac{\sigma_{1R}^{2}}{\bar{X}^{2}(M-1)}} \qquad (7.10)$$

مثال تشریحی: فرض کنیم از جامعهٔ پرستاران متخصص که در جدول ۱.۱۰ نشان داده شده است می‌خواهیم یک نمونهٔ خوشه‌ای دومرحله‌ای بگیریم که در مرحلهٔ دوم از هر مرکز بهداشت منتخب در نمونهٔ مرحلهٔ اول دو پرستار انتخاب شود. فرض می‌کنیم که می‌خواهیم ۹۵ درصد مطمئن باشیم که برآورد x'_{clu} که از کل تعداد بیماران دیده شده توسط پرستاران متخصص به دست می‌آید در محدودهٔ ۳۰٪ مقدار واقعی است و برآورد r_{clu}، نسبت بیمارانی که به پزشک ارجاع شده‌اند، نیز در محدودهٔ ۳۰٪ مقدار واقعی است. از محاسبات قبلی داریم

$$\sigma_{1x}^{2}=۵۶۷/۴۴ \qquad \sigma_{2x}^{2}=۱۶۰/۲۷ \qquad \bar{X}=۸۴/۴$$
$$\bar{\bar{X}}=۲۸/۱۳ \qquad M=۵ \qquad \bar{N}=۳$$

چون $\bar{n}=۲$ و $\varepsilon=۰/۳$ ، از رابطهٔ (۶.۱۰) داریم

$$m=\frac{\frac{۵۶۷/۴۴}{۸۴/۴^{2}}\times\frac{۵}{۵-۱}+\frac{۱}{۲}\times\frac{۱۶۰/۲۷}{۲۸/۱۳^{2}}\times\left(\frac{۳-۲}{۳-۱}\right)}{\frac{۰/۳^{2}}{۱/۹۶^{2}}+\frac{۵۶۷/۴۴}{۸۴/۴^{2}(۵-۱)}}=۳/۴۷$$

یعنی چهار خوشه باید انتخاب شوند.

با استفاده از اطلاعات تعیین شده در مثال قبلی، داریم

$$R = \frac{X}{Y} = \frac{۱۵۲}{۴۲۲} = ۰/۳۶۰۲$$

$$\sigma_{۱R}^{۲} = \sigma_{۱y}^{۲} + R^{۲}\sigma_{۱x}^{۲} - ۲R\sigma_{۱xy}$$

$$= ۱۳۵/۴۴ + ۰/۳۶۰۲^{۲}(۵۶۷/۴۴) - ۲(۰/۳۶۰۲)(-۱۶۶/۱۶) = ۳۲۸/۷۶$$

$$\sigma_{۲R}^{۲} = \sigma_{۲y}^{۲} + R^{۲}\sigma_{۲x}^{۲} - ۲R\sigma_{۲xy}$$

$$= ۳۹/۲ + ۰/۳۶۰۲^{۲}(۱۶۰/۲۷) - ۲(۰/۳۶۰۲)(۴۷/۴۴) = ۲۵/۸۲$$

پس، از رابطهٔ (۷.۱۰) داریم

$$m = \frac{\frac{۳۲۸/۷۶}{۸۴/۴^{۲}} \times \frac{۵}{۵-۱} + \frac{۱}{۲} \times \frac{۲۵/۸۲}{۲۸/۱۳^{۲}} \times \left(\frac{۳-۲}{۳-۱}\right)}{\frac{۰/۳^{۲}}{۱/۹۶^{۲}} + \frac{۳۲۸/۷۶}{۸۴/۴^{۲}(۵-۱)}} = ۱/۸۸$$

یعنی دو خوشه باید انتخاب شوند.

به این ترتیب، در مرحلهٔ اول به نمونه‌ای متشکل از چهار مرکز بهداشت نیاز خواهیم داشت تا ۹۵٪ مطمئن باشیم که برآورد کل تعداد بیمارانی که توسط پرستار متخصص دیده شده‌اند (x'_{clu}) ویژگیهای بیان شده را تأمین می‌کند. ولی برای این که ۹۵٪ مطمئن باشیم که برآورد نسبت ارجاع شده به پزشک، ویژگیهای بیان شده را تأمین می‌کند به نمونه‌ای متشکل از فقط دو مرکز بهداشت نیاز خواهیم داشت. بنابراین، اگر بخواهیم که هر دوی این برآوردها، ویژگیهای تعیین شده را تأمین کنند نمونه‌ای از چهار مرکز بهداشت را انتخاب خواهیم کرد.

در عمل، پارامترهای $\sigma_{۱x}^{۲}$، $\sigma_{۲x}^{۲}$ و نظایر آنها، که برای تعیین تعداد مورد نیاز m خوشهٔ نمونه لازم است به ندرت معلوم‌اند و باید یا از روی داده‌های موجود برآورد شوند و یا از روی تجربه یا به طور شهودی حدس زده شوند.

□

۶.۱.۱۰ انتخاب $\bar{n}$، اندازهٔ بهینهٔ خوشه، با در نظر گرفتن هزینه‌ها

فرض کنید می‌خواهیم از میان چند طرح نمونه‌ای، طرحی را انتخاب کنیم که ویژگیهای تعیین شده برای قابلیت اعتماد برآوردها را با کمترین هزینه تأمین کند و فرض کنید می‌خواهیم نمونه‌گیری خوشه‌ای دومرحله‌ای ساده را به عنوان یک طرح ممکن بررسی کنیم. ولی نمونه‌گیری خوشه‌ای دومرحله‌ای ساده یک رده از طرحهاست که با تعداد m خوشهٔ مرحلهٔ اول و تعداد $\bar{n}$ واحدهای فهرست‌برداری مرحلهٔ دوم مشخص می‌شود. پس مشکل عبارت است از بررسی این که چه ترکیبی از

m و $\overline{n}$ ویژگیهای مورد نیاز را با کمترین هزینه تأمین خواهد کرد. این مشکل را با یک مثال تشریح می‌کنیم.

مثال تشریحی: از جامعهٔ متشکل از پنج پروژهٔ خانه‌سازی استفاده می‌کنیم که در جدول ۱.۹ نشان داده شده است. قبلاً اجزای هزینه را که ملازم با این مثال است در مورد نمونه‌گیری تصادفی ساده، نمونه‌گیری تصادفی طبقه‌بندی شده، و نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده مورد بحث قرار داده‌ایم. بخصوص نشان دادیم که در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده، هزینه‌های میدانی تقریبی C با تابع زیر مشخص می‌شود [معادلهٔ (۳.۹)] :

$$C = C_1' m + C_2' m \overline{N}$$

هزینه‌های میدانی برای نمونهٔ خوشه‌ای سادهٔ دومرحله‌ای را می‌توان با تابعی مشابه شکل زیر تقریب زد

$$C = C_1^* m + C_2^* m \overline{n} \qquad (۸.۱۰)$$

که در آن C_1^* از هزینهٔ سفر به هر خوشهٔ نمونه به منظور فهرست‌برداری از $\overline{N}$ واحد نمونه‌گیری، هزینهٔ فهرست‌برداری از این $\overline{N}$ واحدهای فهرست‌برداری، هزینهٔ انتخاب یک نمونه متشکل از $\overline{n}$ واحد از هر فهرست، و هزینهٔ سفر برگشت به خوشه برای انجام مصاحبه تشکیل شده است. C_2^* هزینهٔ مصاحبه با هر یک از واحدهای نمونه‌گیری انتخاب شده است.

در مثال مورد بررسی در فصل ۹، سفر به هر خوشهٔ نمونه ۰/۵ نفر ساعت هزینه داشت. فرض کنید فهرست‌برداری از ۲۰ واحد نمونه‌گیری در خوشه‌های انتخاب شده و سپس انتخاب یک نمونهٔ تصادفی از این واحدهای نمونه‌گیری، ۱ نفر ساعت هزینه داشته باشد، و بازگشت به خوشه به منظور مصاحبه ۰/۵ نفر ساعت هزینه بردارد. پس $C_1^* = ۰/۵۰ + ۱/۰۰ + ۰/۵۰ = ۲/۰۰$. هچنین مانند مثال فصل ۹، هزینهٔ مصاحبه با هر خانوار نمونه‌گیری شده ۰/۲۵ نفر ساعت است. پس $C_2^* = ۰/۲۵$ و تابع هزینهٔ (۸.۱۰) به صورت زیر درمی‌آید

$$C = ۲/۰۰ m + ۰/۲۵ m\overline{n}$$

توجه کنید که عبارت $m\overline{N}$ که در جملهٔ دوم رابطهٔ (۳.۹) آمده کل تعداد واحدهای فهرست‌برداری است که در نمونهٔ طرح نمونه‌گیری خوشه ای یک‌مرحله‌ای ساده گنجانده شده‌اند. به همین قیاس، $m\overline{n}$ که در جملهٔ دوم رابطهٔ (۸.۱۰) آمده، تعداد واحدهای فهرست‌برداری است که در نمونهٔ مربوط به طرح نمونه‌گیری خوشه‌ای دومرحله‌ای ساده گنجانده شده‌اند. به این ترتیب، تابعهای هزینهٔ (۳.۹) و (۸.۱۰) بسیار شبیه به یکدیگرند.

فرض کنید می‌خواهیم کل تعداد افراد ساکن در پنج شهرک مسکونی را برآورد کنیم و واقعاً مطمئن باشیم که این برآورد در محدودهٔ ۲۵٪ مقدار واقعی است. با استفاده از رابطهٔ (٦.١٠) تعداد m خوشهٔ مورد نیاز برای تأمین این ویژگی را برای اندازه‌های گوناگون خوشهٔ مرحلهٔ دوم، $\overline{n}$، بررسی می‌کنیم. سپس از رابطهٔ (٨.١٠) برای تعیین هزینه‌های میدانی در هر یک از این طرحها استفاده می‌کنیم. آن دسته از طرحهایی که این ویژگی را تأمین می‌کنند در جدول ۵.۱۰ فهرست شده‌اند.

برای نشان دادن چگونگی انجام محاسبات جدول ۵.۱۰، یک طرح نمونه‌گیری خوشه‌ای دومرحله‌ای ساده را در نظر می‌گیریم که تعداد خانوارهایی که باید نمونه‌گیری شوند برابر با ۵ منظور شده است (یعنی $\overline{n}$ =۵). از جدول ۱.۹ موارد زیر را محاسبه می‌کنیم:

$$\sigma_{1y}^{2} = ۶/۸ \qquad \overline{Y} = ۳۴ \qquad \sigma_{2y}^{2} = ۰/۶۹۳$$

$$\overline{\overline{Y}} = ۱/۷ \qquad M = ۵ \qquad \overline{N} = ۲۰$$

جدول ۵.۱۰ طرحهایی که ویژگیهای مربوط به مجموع را تأمین می‌کنند

تعداد واحدهای فهرست‌برداری انتخاب شده در مرحلهٔ دوم، $\overline{n}$	m، تعداد خوشه‌های مورد نیاز برای تأمین ویژگی $\varepsilon = ۰/۲۵$	هزینهٔ میدانی (برحسب نفر ساعت) $C = ۲/۰m + ۰/۲۵m\overline{n}$
٦	۵	۱۷/۵
۷	٤	۱۵/۰
۸	٤	۱٦/۰
۹	۳	۱۲/۷۵
۱۰	۳	۱۳/۵
۱۱	۳	۱٤/۲۵
۱۲	۲	۱۰/۰
۱۳	۲	۱۰/۵
۱٤	۲	۱۱/۰
۱۵	۲	۱۱/۵
۱٦	۲	۱۲/۰
۱۷	۲	۱۲/۵
۱۸	۲	۱۳/۰
۱۹	۱	٦/۷۵
۲۰	۱	۷/۰

سپس از رابطهٔ (٦.١٠) داریم

$$m=\frac{\frac{۶/۸}{۳۴^{۲}}\times\frac{۵}{۵-۱}+\frac{۱}{۵}\times\frac{۰/۶۹۳}{۱/۷^{۲}}\times\left(\frac{۲۰-۵}{۲۰-۱}\right)}{\frac{(۰/۲۵)^{۲}}{۳^{۲}}+\frac{۶/۸}{۳۴^{۲}(۵-۱)}}=۵/۲۳$$

یعنی شش خوشه باید انتخاب شوند. به این ترتیب، نمونه‌ای متشکل از شش خوشه برای تأمین ویژگیها لازم خواهد بود. چون فقط پنج خوشه در جامعه موجود است، پس اگر $\overline{n}=۵$ باشد، تأمین ویژگیها امکان‌پذیر نخواهد بود.

حال فرض کنیم $\overline{n}=٦$. از رابطهٔ (٦.١٠) داریم

$$m=\frac{\frac{۶/۸}{۳۴^{۲}}\times\frac{۵}{۵-۱}+\frac{۱}{۶}\times\frac{۰/۶۹۳}{۱/۷^{۲}}\times\left(\frac{۲۰-۶}{۲۰-۱}\right)}{\frac{(۰/۲۵)^{۲}}{۳^{۲}}+\frac{۶/۸}{۳۴^{۲}(۵-۱)}}=۴/۳۷$$

یعنی پنج خوشه باید انتخاب شوند. به این ترتیب اگر $\overline{n}=٦$ ، نمونه‌ای متشکل از $m=۵$ خوشه (یعنی همهٔ خوشه‌های موجود در جامعه) ویژگی مورد نظر را تأمین خواهد کرد. هزینه‌های میدانی چنین طرح نمونه‌ای عبارت است از

$$C=۲/۰۰\times۵+۰/۲۵\times۵\times۶=۱۷/۵ \text{ نفر ساعت}$$

از میان همهٔ طرحهای نمونه‌گیری خوشه‌ای دومرحله‌ای ساده فهرست شده در جدول ۵.۱۰ که ویژگیهای مورد نظر را تأمین می‌کنند، طرحی که با انتخاب مرحلهٔ اول $m=۱$ خوشه و به دنبال آن یک نمونهٔ مرحله دوم $\overline{n}=۱۹$ واحد فهرست‌برداری مشخص شده است کمترین هزینهٔ میدانی ممکن را دارد (یعنی ٦/٧۵ نفر ساعت) و از این رو، طرح انتخابی خواهد بود.

اگر می‌خواستیم یک نسبت را برآورد کنیم ـ مثلاً نسبت R همهٔ اشخاص نیازمند به خدمات پرستار متخصص ـ همین فرایند را طی می‌کردیم، جز این که به جای رابطهٔ (٦.١٠) از رابطهٔ (٧.١٠) استفاده می‌کردیم. اگر باز $\varepsilon=۰/۲۵$را در نظر بگیریم، طرحهای نمونه‌گیری خوشه‌ای دومرحله‌ای فهرست شده در جدول ٦.١٠ می‌توانند ویژگیهای موردنظر برای برآورد نسبت را تأمین کنند.

از داده‌های جدول ۱.۹ می‌توانیم مقادیر زیر را برای تعیین هزینه‌های ارائه شده در جدول ٦.١٠ محاسبه کنیم:

$X = ۶۰$ $Y = ۱۷۰$ $R = ۰/۳۵۲۹۴$

$\sigma_{۱x}^{۲} = ۲۳/۶$ $\sigma_{۱y}^{۲} = ۶/۸$ $\sigma_{۱xy} = ۲/۶$

$\sigma_{۲y}^{۲} = ۰/۶۹۳$ $\sigma_{۲x}^{۲} = ۰/۳۲۱$ $\sigma_{۲xy} = ۰/۱۴۳۵$

$\sigma_{۱R}^{۲} = ۲۲/۶۱$ $\sigma_{۲R}^{۲} = ۰/۳۰۶۰$

بـا بررسـی جدول ٦.۱۰ می‌بینیم که یک نمونهٔ خوشه‌ای دومرحله‌ای که همهٔ پنج شهرک مسکونی، در مرحـلهٔ اول و از هـر شـهرک مسـکونی، ۱۲ خـانوار در مرحـلهٔ دوم نمونه‌گیری شده باشند طرحی خواهـد بـود کـه ویژگیهـای تعییـن شـده بـرای بـرآورد نسـبت R اشخاص نیازمند به خدمات پرستار متخصص را با کمترین هزینهٔ میدانی تأمین خواهد کرد.

□

جدول ٦.۱۰ طرحهایی که ویژگیهای مربوط به نسبت را تأمین می‌کنند

تعداد واحدهای فهرست‌برداری انتخاب شده در مرحلهٔ دوم، $\overline{n}$	m ، تعداد خوشه‌های مورد نیاز برای تأمین ویژگی $\varepsilon = ۰/۲۵$	هزینهٔ میدانی (برحسب نفر ساعت) $C = ۲/۰m + ۰/۲۵m\overline{n}$
۱۲	۵	۲۵/۰۰
۱۳	۵	۲٦/۲۵
۱٤	۵	۲۷/۵۰
۱۵	۵	۲۸/۷۵
۱٦	۵	۳۰/۰۰
۱۷	۵	۳۱/۲۵
۱۸	۵	۳۲/۵۰
۱۹	۵	۳۳/۷۵
۲۰	۵	۳۵/۰۰

۷.۱.۱۰ برخی فرمولهای میان‌بُر برای تعیین تعداد بهینهٔ $\overline{n}$

در بخـش قـبل، مقادیـر $\overline{n}$ و m را که با کمترین هزینهٔ میدانی ممکن ویژگیها را تأمین می‌کنند به روش زیـر انتخاب کردیم. همهٔ ترکیبات ممکن m و $\overline{n}$ را شمردیم و آن ترکیباتی را که ویژگیها را [با استفاده از رابطـه‌هـای (٦.۱۰) و (۷.۱۰)] تأمین نمی‌کردند حذف کردیم. سپس از میان طرحهای باقیمانده، یک طـرح را انـتخاب کردیم که ویژگیها را با کمترین هزینهٔ ممکن تأمین می‌کرد. این فرایند می‌تواند بسیار خسته کننده باشد، بخصوص اگر M و $\overline{N}$ عددهایی بزرگ باشند.

یک فرمول میان‌بر هست که می‌تواند برای انتخاب مقدار بهینهٔ $\bar{n}$ مورد استفاده قرار گیرد. اگر هزینه‌های میدانی را بتوان با تابعی به صورت $C_1^* m + C_2^* m\bar{n}$ تقریب کرد و اگر تعداد خوشه‌ها، M، در جامعه در مقایسه با تعداد خوشه‌های انتخاب شده در نمونه بزرگ باشد، آن‌گاه فرمول میان‌بر مفید خواهد بود. این فرمول (که برای برآورد کردن میانگینها یا مجموعها معتبر است) عبارت است از

$$\bar{n} = \left[\left(\frac{C_1^*}{C_2^*}\right)\left(\frac{1-\delta_x}{\delta_x}\right)\right]^{1/2} \qquad (۹.۱۰)$$

که در آن C_1^* و C_2^* اجزای هزینه‌اند که قبلاً معرفی شده‌اند و δ_x *ضریب همبستگی درون ـ رده‌ای* است که قبلاً در مبحث نمونه‌گیری سیستماتیک [معادلهٔ (۵.٤)] تعریف شد و می‌تواند به شکل زیر بیان شود

$$\delta_x = \frac{(1/\bar{N})\sigma_{1x}^2 - \sigma_x^2}{(\bar{N}-1)\sigma_x^2} \qquad (۱۰.۱۰)$$

که به صورت جبری با عبارت (۵.٤) یکسان است به شرطی که تعداد واحدهای فهرست‌برداری، N_i، برای همهٔ خوشه‌ها یکسان باشد.

به محض اینکه اندازهٔ خوشهٔ بهینه، $\bar{n}$، از رابطهٔ (۹.۱۰) به دست آمد، m، تعداد خوشه‌هایی که باید در مرحلهٔ اول انتخاب شوند از رابطهٔ (٦.۱۰) به دست می‌آید.

رابطه‌های (۹.۱۰) و (۱۰.۱۰) برای برآورد کردن نسبتها به صورت زیر اصلاح می‌شوند:

$$\bar{n} = \left[\left(\frac{C_1^*}{C_2^*}\right)\left(\frac{1-\delta_R}{\delta_R}\right)\right]^{1/2} \qquad (۱۱.۱۰)$$

و

$$\delta_R = \frac{(1/\bar{N})\sigma_{1R}^2 - \sigma_R^2}{(\bar{N}-1)\sigma_R^2} \qquad (۱۲.۱۰)$$

برای نشان دادن این مفاهیم مثال زیر را در نظر می‌گیریم.

مثال تشریحی: فرض کنید قرار است در یک بیمارستان عمومی به منظور برآورد کل مبلغی که براساس صورتحسابها از بیمهٔ مراقبتهای پزشکی برای یک دورهٔ سه‌ماههٔ خاص (۱۳ هفته) مطالبه شده است یک آمارگیری اجرا شود. قرار است یک نمونهٔ خوشه‌ای انتخاب شود به این ترتیب که یک نمونهٔ تصادفی ساده از هفته‌ها و پس از آن یک نمونهٔ تصادفی ساده از روزهای هر یک از هفته‌های انتخاب شده گرفته می‌شود. برای هر روزی که در نمونه انتخاب شود، مقدار صورتحساب مطالبه شده از بیمهٔ مراقبتهای پزشکی با بررسی کلیهٔ اقلام مندرج در صورتحسابهای آن روز به دست خواهد آمد. داده‌های

مربوط به سرتاسر دورهٔ سه‌ماهه (که البته تا پیش از آمارگیری نامعلوم است) برحسب روز در جدول ۷.۱۰ نشان داده شده‌اند.

با در نظر گرفتن هفته‌ها به عنوان خوشه‌ها، روزها به عنوان واحدهای فهرست‌برداری، و X به عنوان مقدار مطالبه شده در صورتحساب از بیمهٔ مراقبتهای پزشکی در یک روز خاص، محاسبات زیر را خواهیم داشت:

$$\sigma_x^2 = \frac{\sum_{i=1}^{13}\sum_{j=1}^{7}(X_{ij} - \overline{\overline{X}})^2}{91} = ۱۳۰۴/۵۸ \qquad \sigma_{1x}^2 = \frac{\sum_{i=1}^{13}(X_i - \overline{X})^2}{13} = ۲۶۶۲۴/۴۴$$

$$\overline{\overline{X}} = ۱۱۰/۴۵ \qquad \overline{X} = ۷۷۳/۱۵ \qquad \sigma_{2x}^2 = ۷۶۱/۲۲$$

$$M = ۱۳ \qquad \overline{N} = ۷ \qquad N = ۹۱$$

$$\delta_x = \frac{(۱/۷)(۲۶۶۲۴/۴۴) - ۱۳۰۴/۵۸}{(۷-۱)(۱۳۰۴/۵۸)} = ۰/۳۱۹۲$$

حالا به بررسی هزینه‌هایی می‌پردازیم که ممکن است با جمع‌آوری داده‌ها همراه باشند. برای هر روز انتخاب شده در نمونه لازم است داده‌های مربوط به مقدار مطالبه شده از بیمهٔ مراقبتهای پزشکی برای هر بیمار بستری در آن روز به دست آید. این کار مستلزم بازیابی و استخراج مقدار قابل توجهی از سوابق است. فرض می‌کنیم که به طور متوسط روزی پنج بیمار در بیمارستان هستند که تحت پوشش بیمهٔ مراقبتهای پزشکی قرار دارند و برای بازیابی و استخراج سوابق مربوط به هر بیمار، توسط یک دستیار تحقیقاتی که ساعتی ۸/۰۰ دلار دریافت می‌کند، ۰/۵ ساعت وقت لازم است. علاوه بر این، فرض می‌کنیم که کدگذاری داده‌ها، آماده‌سازی آن برای رایانه و پردازش آن برای هر بیمار تقریباً ۶/۰۰ دلار هزینه دربردارد. پس کل هزینهٔ C_2^* به ازای واحد فهرست‌برداری (روز) که در نمونه انتخاب شده، عبارت است از

C_2^* = (۵ بیمار در روز) × (۰/۵ × ۸ + ۶) به ازای هر بیمار = ۵۰ دلار در روز

فرض می‌کنیم که سوابق مربوط به هزینهٔ بیماران روی نوارهای رایانه‌ای برحسب هفته تنظیم شده‌اند و برای هر هفته‌ای که در نمونه انتخاب شده است باید سوابق *تمام طول هفته* به وسیلهٔ رایانه فهرست شوند تا هزینه‌های روزانه مشخص گردند و انجام این کارها از نظر صرف وقت رایانه و سایر مخارج پردازش داده‌ها ۱۰۰ دلار هزینه دربردارد. فرض می‌کنیم سایر هزینه‌های ملازم با خوشه‌ها (هفته‌ها) قابل اغماض‌اند به طوری که مؤلفهٔ C_1^* به صورت زیر به دست می‌آید

C_1^* = ۱۰۰ دلار به ازای هر هفتهٔ نمونه

جدول ۷.۱۰ مبلغ مطالبه شده از بیمهٔ مراقبتهای پزشکی برحسب روز و هفته

هفته	مبلغ مطالبه شده (× ۱۰ دلار)							X_i	$\sum (X_{ij} - \bar{X}_i)^2$
	یکشنبه	دوشنبه	سه‌شنبه	چهارشنبه	پنج‌شنبه	جمعه	شنبه		
۱	۱۲۱	۸۸	۱۳۸	۶۹	۸۵	۵۹	۶۳	۶۲۳	۵۴۱۸
۲	۱۵۲	۱۰۵	۱۷۹	۱۲۷	۱۴۵	۱۱۶	۱۶۵	۹۸۹	۴۲۷۳/۴۳
۳	۱۳۴	۶۲	۴۴	۱۲۸	۹۲	۱۲۷	۱۴۱	۷۲۸	۸۸۸۲
۴	۹۴	۱۲۳	۸۱	۱۰۹	۱۱۰	۱۰۱	۳۴	۶۵۲	۵۱۳۴/۸۶
۵	۲۱۰	۱۸۲	۱۲۰	۱۲۰	۱۵۴	۹۸	۱۷۸	۱۰۶۲	۹۹۰۷/۴۳
۶	۱۰۴	۱۰۰	۷۰	۱۱۵	۱۴۲	۱۳۹	۱۳۱	۸۰۱	۳۹۲۹/۷۱
۷	۸۰	۴۳	۵۹	۵۰	۲۰	۶۹	۶۴	۳۸۵	۲۳۱۲
۸	۱۲۳	۵۹	۱۴۰	۱۴۶	۱۴۹	۱۰۹	۱۱۱	۸۳۷	۵۸۴۷/۷۱
۹	۱۳۷	۱۶۸	۸۴	۱۲۰	۱۲۳	۸۷	۱۲۱	۸۴۰	۴۹۸۸
۱۰	۱۱۰	۱۳۲	۱۰۴	۶۵	۹۷	۹۸	۸۴	۶۹۰	۲۶۱۹/۷۱
۱۱	۱۵۸	۱۱۴	۹۳	۱۴۷	۹۱	۱۲۵	۱۳۲	۸۶۰	۳۸۹۰/۸۶
۱۲	۷۹	۱۵۶	۱۴۲	۱۵۴	۱۳۳	۷۴	۸۳	۸۲۱	۸۲۱۹/۴۳
۱۳	۹۷	۹۶	۱۱۰	۸۳	۸۹	۱۴۶	۱۴۲	۷۶۳	۳۸۴۸

چون برای این داده‌ها $\delta_x \approx$ ۰/۳۲، اندازهٔ بهینهٔ خوشه، $\overline{n}$، به شکل زیر به دست می‌آید

$$\overline{n} = \sqrt{\left(\frac{۱۰۰}{۵۰}\right)\left(\frac{۰/۶۸}{۰/۳۲}\right)} \approx ۲$$

سپس، با استفاده از رابطهٔ (۶.۱۰) با $M = ۱۳$، $\overline{n} = ۲$ ، $\sigma_{۱x}^{۲} = ۲۶۶۲۴/۴۴$، $\sigma_{۲x}^{۲} = ۷۶۱/۲۲$، $\overline{\overline{X}} = ۱۱۰/۴۵$، $\overline{X} = ۷۷۳/۱۵$ و $\overline{N} = ۷$، در صورتی که بخواهیم واقعاً مطمئن باشیم که برآورد مقدار کل مبلغ مطالبه شده از بیمهٔ مراقبتهای پزشکی دردورهٔ سه‌ماههٔ موردنظر در محدودهٔ ۲۵٪ مقدار واقعی آن است (یعنی $\varepsilon = ۰/۲۵$)، خواهیم داشت

$$m = \frac{\frac{۲۶۶۲۴/۴۴}{(۷۷۳/۱۵)^{۲}} \times \frac{۱۳}{۱۳-۱} + \frac{۱}{۲} \times \frac{۷۶۱/۲۲}{(۱۱۰/۴۵)^{۲}} \times \left(\frac{۷-۲}{۷-۱}\right)}{\frac{(۰/۲۵)^{۲}}{۳^{۲}} + \frac{۲۶۶۲۴/۴۴}{(۷۷۳/۱۵)^{۲}(۱۳-۱)}} = ۶/۹۷$$

پس باید هفت خوشه انتخاب شوند.

به این ترتیب، با یک نمونهٔ خوشه‌ای دومرحله‌ای $m = ۷$ هفته و $\overline{n} = ۲$ روز می‌توان ویژگیهای موردنظر را با حداقل هزینه نسبت به سایر طرحهای نمونه‌گیری خوشه‌ای دومرحله‌ای تأمین کرد. این موضوع با بررسی هزینه‌های سایر طرحهای نمونه‌گیری خوشه‌ای که ویژگیهای موردنظر را تأمین می‌کنند به صورتی که در جدول ۸.۱۰ نشان داده شده است می‌تواند تأیید شود.

ضریب همبستگی درون رده‌ای δ_x پارامتری است که در نظریهٔ نمونه‌گیری از اهمیت زیادی برخوردار است. همان طور که در بالا نشان داده شد، این ضریب در تعیین تعداد بهینهٔ واحدهای فهرست‌برداری که باید در مرحلهٔ دوم نمونه‌گیری خوشه‌ای انتخاب شوند سودمند است. ضریب مزبور در ربط دادن خطای معیار یک برآورد میانگین یا مجموع از روی یک نمونهٔ خوشه‌ای دومرحله‌ای به خطای معیاری که ممکن بود از یک نمونهٔ تصادفی ساده با همان تعداد واحدهای فهرست‌برداری به دست آید نیز سودمند است. اگر M ، تعداد خوشه‌های موجود در جامعه، در مقایسه با تعداد انتخاب شده در مرحلهٔ اول نمونه‌گیری، زیاد باشد، آن‌گاه رابطه‌های زیر مصداق پیدا می‌کنند:

$$SE(x'_{clu}) = \left(\frac{N\sigma_x}{\sqrt{n}}\right)\sqrt{۱ + \delta_x(\overline{n}-۱)} \qquad (۱۳.۱۰)$$

$$SE(\overline{x}_{clu}) = \left(\frac{N\sigma_x}{\sqrt{Mn}}\right)\sqrt{۱ + \delta_x(\overline{n}-۱)} \qquad (۱۴.۱۰)$$

$$SE\left(\overline{\overline{x}}_{clu}\right) = \left(\frac{\sigma_x}{\sqrt{n}}\right)\sqrt{۱ + \delta_x(\overline{n}-۱)} \qquad (۱۵.۱۰)$$

جدول ۸.۱۰ طرحهایی که ویژگیهای مربوط به مجموع را تأمین می‌کنند

تعداد واحدهای فهرست‌برداری در نمونه $\bar{n}$	m، تعداد خوشه‌های مورد نیاز برای تأمین ویژگی $\varepsilon = ۰/۲۵$	هزینه‌های میدانی (برحسب دلار) $C = ۱۰۰m + ۵۰m\bar{n}$
۱	۱۰	۱۵۰۰
۲	۷	۱۴۰۰
۳	۷	۱۷۵۰
٤	٦	۱۸۰۰
٥	٦	۲۱۰۰
٦	٦	۲٤۰۰
۷	٦	۲۷۰۰

رابطه‌های مزبور از این جهت اهمیت دارند که نشان می‌دهند خطای معیارهای میانگینها و مجموعهای به دست آمده از نمونه‌های خوشه‌ای تقریباً $\sqrt{۱ + \delta_x(\bar{n} - ۱)}$ بار بزرگتر از خطای معیارهای به دست آمده از نمونهٔ تصادفی ساده با همان تعداد واحد فهرست‌برداری هستند. این عامل، $\sqrt{۱ + \delta_x(\bar{n} - ۱)}$، به عنوان *اثر طرح* نامیده می‌شود و بستگی دارد به ضریب همبستگی درون رده‌ای δ_x و تعداد $\bar{n}$ واحد فهرست‌برداری که در هر خوشه نمونه‌گیری شده است.

با عملیات جبری در رابطهٔ (۱۰.۱۰) می‌توان دید که ضریب همبستگی درون رده‌ای δ_x عددی بین $-۱/(\bar{N} - ۱)$ و ۱+ است. این ضریب در واقع مقیاس همگنی واحدهای فهرست‌برداری داخل خوشه‌ها نسبت به سطوح مشخصهٔ x مورد اندازه‌گیری است. اگر واحدهای فهرست‌برداری درون خوشه‌ها از این لحاظ گرایش به همسانی داشته باشند، مقدار δ_x رو به فزونی خواهد گذاشت و اثر طرح، $\sqrt{۱ + \delta_x(\bar{n} - ۱)}$، در صورت بزرگ بودن $\bar{n}$، بالقوه زیاد خواهد بود. این موضوع به طور شهودی هم معقول است زیرا مقدار زیاد δ_x به این مفهوم است که گنجاندن تعداد زیادی از واحدهای فهرست‌برداری در داخل یک خوشهٔ نمونه، شیوهٔ بیهوده‌ای است، زیرا واحدهای فهرست‌برداری داخل خوشه‌ها از نظر سطوح مشخصهٔ x مورد اندازه‌گیری شبیه‌اند.

دانستن مقدار δ_x برای انواع گوناگون خوشه‌ها و متغیرهای مختلف در برنامه‌ریزی آمارگیریها فوق‌العاده سودمند است. در این زمینه، بحثهای مفصلتری در متون دیگر ارائه شده است [۱،۲].

□

۲.۱۰ وضعیتی که در آن، N_i، تعداد واحدهای شمارش در همهٔ خوشه‌ها یکسان نیست

در بخش قبل، وضعیت بسیار محدودی را فرض گرفتیم که طی آن همهٔ خوشه‌های مورد استفاده به عنوان واحدهای نمونه‌گیری اولیه دارای تعداد یکسانی از واحدهای شمارش بودند. این وضعیت به فرمولهای نسبتاً ساده‌ای برای برآورد کردن پارامترهای جامعه و برآورد کردن خطای معیارهای این‌گونه برآوردها منجر شد. ولی، در بیشتر وضعیتهایی که مستلزم نمونه‌گیری از جوامع انسانی است واحدهای نمونه‌گیری اولیه دارای تعداد یکسانی از واحدهای شمارش نیستند. به طوری که در فصل بعدی بحث خواهد شد، راهبرد خوشه‌های نمونه‌گیری با احتمال برابر و با داشتن کسرهای نمونه‌گیری مرحلهٔ دوم تقریباً مساوی در داخل هر خوشهٔ نمونه غالباً به خطاهای نمونه‌گیری بزرگ، به خصوص در برآورد کردن مجموعهای جامعه، منجر خواهد شد. با به خاطر داشتن این موضوع، مابقی این فصل را به برآورد کردن پارامترهای جامعه از طریق طرحهای نمونه‌گیری خوشه‌ای سادهٔ دومرحله‌ای اختصاص می‌دهیم که خوشه‌های آنها از لحاظ تعداد واحدهای شمارش داخل خوشه‌ها با یکدیگر متفاوت‌اند.

۱.۲.۱۰ چگونگی انتخاب یک نمونهٔ خوشه‌ای دومرحله‌ای ساده برای این طرح

مثل قبل، خوشه‌ها با شماره‌گذاری از ۱ تا M و سپس انتخاب m شمارهٔ تصادفی بین ۱ و M انتخاب می‌شوند. خوشه‌های متناظر با شماره‌های تصادفی انتخاب شده در نمونه قرار خواهند گرفت. سپس در داخل هر خوشهٔ نمونه، یک نمونهٔ تصادفی ساده متشکل از n_i واحد شمارش انتخاب می‌کنیم که در آن n_i طوری انتخاب می‌شود که کسر نمونه‌گیری مرحلهٔ دوم n_i/N_i تا حد ممکن به کسر نمونه‌گیری مرحلهٔ دومِ از پیش تعیین شدهٔ f_2 که برای هر خوشهٔ نمونه مشخص شده است نزدیک باشد.

با به خاطر داشتن این موضوع، واحدهای فهرست‌برداری داخل i امین خوشه را که در مرحلهٔ اول انتخاب شده است از ۱ تا N_i شماره‌گذاری می‌کنیم و n_i شمارهٔ تصادفی مختلف انتخاب می‌کنیم. واحدهای فهرست‌برداری متناظر با شماره‌های تصادفی در نمونه قرار خواهند گرفت.

این شیوه را در مثال زیر نشان می‌دهیم.

مثال تشریحی: یکی از تقسیمات کشوری را که دارای ۱۰ بیمارستان است با کل پذیرشها در سال ۱۹۸۷، کل پذیرشهای سال ۱۹۸۷ که دچار بیماری مهلک بوده‌اند و کل پذیرشهای سال ۱۹۸۷ که مرده از بیمارستان ترخیص شده‌اند، به شرح ارائه شده در جدول ۹.۱۰ در نظر می‌گیریم.

فرض کنید می‌خواهیم یک نمونهٔ تصادفی ساده متشکل از سه بیمارستان انتخاب کنیم و در هر بیمارستان انتخاب شده، یک نمونهٔ تصادفی ساده متشکل از تقریباً ۵٪ کل سوابق پذیرش را به منظور

برآورد کردن کل تعداد بیمارانی که در میان همهٔ بیماران پذیرش شده با شرایط مهلک، مرده از بیمارستانها خارج شده‌اند برگزینیم. برای انجام این کار، ابتدا سه شمارهٔ تصادفی را بین ۱ و ۱۰ انتخاب می‌کنیم (مثلاً ۱، ۲ و ۸). سپس هر پذیرش را در داخل هر بیمارستان نمونه با شماره‌ای از ۱ تا N_i مشخص می‌کنیم و با استفاده از اعداد تصادفی، آن n_i پذیرشی را که نسبت n_i/N_i را از همه بیشتر به ۰/۰۵ نزدیک می‌کند انتخاب می‌کنیم. مقدار ۰/۰۵، کسر نمونه‌گیری مطلوب برای مرحلهٔ دوم است. مثلاً، در بیمارستان ۱، $N_1 = ۴۲۸۸$ و $۰/۰۵ \times ۴۲۸۸ = ۲۱۴/۴$. به این ترتیب $n_1 = ۲۱۴$ پذیرش از بیمارستان ۱ انتخاب می‌کنیم. با همین روش $n_2 = ۲۵۲$ پذیرش از بیمارستان ۲ و $n_3 = ۹۱$ پذیرش از بیمارستان ۸ انتخاب می‌کنیم.

جدول ۹.۱۰ کل پذیرشهای با شرایط مهلک و کل پذیرشهای ترخیص شده به صورت مرده از ده بیمارستان، ۱۹۸۷

بیمارستان	کل پذیرشها	کل پذیرشهای با شرایط مهلک	کل ترخیص‌شدگان مرده از میان کسانی که با شرایط مهلک بوده‌اند
۱	٤۲۸۸	۵۰۱	٤۲
۲	۵۰۳٦	۷۸۵	۷۸
۳	۱۱۷۸	۲۱۳	۱۷
٤	٦۳۸	۱۷۳	۹
۵	۲۷۰۱۰	۳٤۰٤	۳۳۸
٦	۱۱۲۲	۲۱۷	۱۷
۷	۲۱۳٤	٤۲٤	۳۷
۸	۱۸۲٤	۲٤٦	۱۸
۹	٤٦۷۲	۷۷۸	٦۸
۱۰	۲۱۵٤	۳٤٦	۲۷

۲.۲.۱۰ برآورد کردن مشخصه‌های جامعه

برآورد مجموعها، میانگینها و نسبتها در این نوع نمونه‌گیری خوشه‌ای که در آن N_iها، تعداد واحدهای فهرست‌برداری در خوشه‌های نمونه نابرابرند در تابلوی ٤.۱۰ تعریف شده‌اند.

توجه داشته باشید که اگر N_i واحد فهرست‌برداری برای همهٔ خوشه‌ها یکی باشد (یعنی $N_i = \overline{N}$) و اگر کسر نمونه‌گیری مرحلهٔ دوم برای همهٔ خوشه‌های نمونه نیز یکسان باشد (یعنی $n_i = \overline{n}$)، آن‌گاه فرمولهای ارائه شده در تابلوی ٤.١٠ به فرمولهای تابلوی ٢.١٠ تبدیل می‌شوند. □

٣.٢.١٠ برآورد کردن خطای معیار برآوردها

برآوردکردن خطای معیار برآوردهای حاصل از نمونه‌گیری خوشه‌هایی که تعداد واحدهای فهرست‌برداری آنها نابرابر است هنگامی که خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب شده باشند کمی دشوار است. ولی اگر کسرهای نمونه‌گیری مرحلهٔ دوم، $f_۲ = n_i / N_i$، برای همهٔ خوشه‌ها یکسان باشد، آن‌گاه می‌توان برآوردهای خوشه‌ای نهایی را به کار برد که در تابلوی ٢.١٠ نشان داده شده‌اند. برای نشان دادن این مطلب به یک مثال نگاه می‌کنیم.

مثال تشریحی: فرض می‌کنیم که یک نمونهٔ تصادفی ساده متشکل از سه بیمارستان از جامعهٔ ١٠ بیمارستان نشان داده شده در جدول ٩.١٠ و یک زیرنمونهٔ ده درصدی (به نزدیکترین عدد صحیح) از بیمارستانهای انتخاب شده در نمونه گرفته شده است. فرض کنید بیمارستانهای ١، ٤ و ١٠ انتخاب شده‌اند و داده‌های مندرج در جدول ١٠.١٠ به دست آمده‌اند. از این داده‌ها و فرمولهای ارائه شده در تابلوهای ٤.١٠ و ٢.١٠ برای برآورد کردن پارامترهای جامعه و خطای معیارها استفاده خواهیم کرد.

ابتدا به برآورد کردن کل تعداد ترخیص‌شدگان مرده در میان کسانی که دچار شرایط مهلک بوده‌اند می‌پردازیم. محاسبات مربوط به x'_{clu} عبارت‌اند از:

$$x_۱ = ۵ \qquad N_۱ = ۴۲۸۸ \qquad n_۱ = ۴۲۹ \qquad \left(\frac{N_۱}{n_۱}\right) x_۱ = ۴۹/۹۸$$

$$x_۲ = ۷ \qquad N_۲ = ۶۳۸ \qquad n_۲ = ۶۴ \qquad \left(\frac{N_۲}{n_۲}\right) x_۲ = ۶۹/۷۸$$

$$x_۳ = ۳ \qquad N_۳ = ۲۱۵۴ \qquad n_۳ = ۲۱۵ \qquad \left(\frac{N_۳}{n_۳}\right) x_۳ = ۳۰/۰۶$$

$$M = ۱۰ \qquad m = ۳$$

$$x'_{clu} = \left(\frac{M}{m}\right) \times \sum_{i=۱}^{m} \left(\frac{N_i}{n_i}\right) x_i = \left(\frac{۱۰}{۳}\right)(۴۹/۹۸ + ۶۹/۷۸ + ۳۰/۰۶) = ۴۹۹/۳۸$$

تابلوی ٤.١٠ برآوردهای مشخصه‌های جامعه تحت نمونه‌گیری خوشه‌ای دومرحله‌ای ساده با تعداد واحدهای فهرست‌برداری نابرابر

مجموع، x'_{clu}

$$x'_{clu} = \left(\frac{M}{m}\right) \times \sum_{i=1}^{m} \left(\frac{N_i}{n_i}\right) \sum_{j=1}^{n_i} x_{ij}$$

میانگین به ازای خوشه، $\bar{x}_{clu}$

$$\bar{x}_{clu} = \left(\frac{1}{m}\right) \times \sum_{i=1}^{m} \left(\frac{N_i}{n_i}\right) \sum_{j=1}^{n_i} x_{ij}$$

میانگین به ازای واحد فهرست‌برداری، $\bar{\bar{x}}_{clu}$

$$\bar{\bar{x}}_{clu} = \frac{x'_{clu}}{N}$$

فرمول بالا هنگامی قابل استفاده است که N، کل تعداد واحدهای فهرست‌برداری قبل از نمونه‌گیری معلوم باشد. هرگاه N معلوم نباشد، میانگین به ازای واحد فهرست‌برداری می‌تواند، با استفاده از برآورد نسبتی که مخرج کسر آن برآوردی حاصل از نمونهٔ متشکل از تعداد واحدهای فهرست‌برداری در جامعه است، برآورد شود. این برآورد نسبتی غالباً واریانس کمتری نسبت به برآورد $\bar{\bar{x}}_{clu}$ دارد، حتی وقتی برآورد اخیر را بتوان به کار برد.

نسبت، r_{clu}

$$r_{clu} = \frac{\bar{x}_{clu}}{\bar{y}_{clu}}$$

که در آن N_i، تعداد واحدهای فهرست‌برداری در هر خوشه، n_i ، تعداد واحدهای فهرست‌برداری نمونه‌گیری شده از هر خوشه است و سایر نمادها به صورتی هستند که در تابلوی ۱.۱۰ تعریف شده‌اند.

محاسبات مربوط به $\hat{SE}(x'_{clu})$ چنین‌اند:

$$\bar{x} = \frac{۵+۷+۳}{۳} = ۵ \qquad f_۲ = ۰/۱۰ \qquad N = ۵۰۰۵۶ \qquad n = ۷۰۸$$

جدول ۱۰.۱۰ خلاصه داده‌های مربوط به نمونهٔ متشکل از سه بیمارستان انتخاب شده از ده بیمارستان جدول ۹.۱۰

بیمارستان	کل پذیرشها N_i	کل پذیرشهای نمونه‌گیری شده n_i	کل بیماران دچار شرایط مهلک y_i	کل بیماران ترخیص‌شدهٔ مرده از میان بیماران دچار شرایط مهلک، x_i
۱	٤۲۸۸	٤۲۹	٤۷	۵
٤	٦۳۸	٦٤	۱۷	۷
۱۰	۲۱۵٤	۲۱۵	۲٤	۳
مجموع	۷۰۸۰	۷۰۸	۸۸	۱۵

$$\frac{\sum_{i=1}^{m}(x_i-\bar{x})^2}{m-1}=\frac{(5-5)^2+(7-5)^2+(3-5)^2}{3-1}=4$$

$$\hat{SE}(x'_{clu})=\left(\frac{10}{\sqrt{3}\times 0.10}\right)\sqrt{4}\sqrt{\frac{5056-708}{5056}}=114.65$$

بازهٔ اطمینان ۹۵ درصدی برای X عبارت است از

$$x'_{clu}-1.96\times \hat{SE}(x'_{clu})\le X\le x'_{clu}+1.96\times \hat{SE}(x'_{clu})$$
$$499.38-1.96\times 114.65\le X\le 499.38+1.96\times 114.65$$
$$274.67\le X\le 724.09$$

بعـد به برآورد کردن میانگین تعداد افراد ترخیص‌شدهٔ مرده از میان دارندگان شرایط مهلک به ازای هر بیمارستان می‌پردازیم. محاسبهٔ $\bar{x}_{clu}$ به صورت زیر است

$$\bar{x}_{clu}=\left(\frac{1}{m}\right)\left[\sum_{i=1}^{m}\left(\frac{N_i}{n_i}\right)x_i\right]=\left(\frac{1}{3}\right)(49.98+69.78+30.06)=49.94$$

محاسبهٔ $\hat{SE}(\bar{x}_{clu})$ عبارت است از

$$\hat{SE}(\bar{x}_{clu})=\left(\frac{1}{(0.10)\sqrt{3}}\right)\sqrt{4}\sqrt{\frac{5056-708}{5056}}=11.47$$

بازهٔ اطمینان ۹۵ درصدی برای $\overline{X}$ چنین است :

$$\bar{x}_{clu} - ۱/۹۶ \times \hat{SE}(\bar{x}_{clu}) \le \overline{X} \le \bar{x}_{clu} + ۱/۹۶ \times \hat{SE}(\bar{x}_{clu})$$
$$۴۹/۹۴ - ۱/۹۶ \times ۱۱/۴۷ \le \overline{X} \le ۴۹/۹۴ + ۱/۹۶ \times ۱۱/۴۷$$
$$۲۷/۴۶ \le \overline{X} \le ۷۲/۴۲$$

سپس به بررسی میانگین تعداد افراد ترخیص‌شدهٔ مرده به ازای افراد پذیرش شده در بیمارستان می‌پردازیم. محاسبهٔ $\bar{\bar{x}}_{clu}$ عبارت است از

$$\bar{\bar{x}}_{clu} = \frac{x'_{clu}}{N} = \frac{۴۹۹/۳۸}{۵۰۰۵۶} = ۰/۰۰۹۸$$

محاسبهٔ $\hat{SE}\left(\bar{\bar{x}}_{clu}\right)$ چنین است :

$$\hat{SE}\left(\bar{\bar{x}}_{clu}\right) = \left(\frac{۱۰}{۵۰۰۵/۶\sqrt{۳}(۰/۱۰)}\right)\sqrt{۴}\sqrt{\frac{۵۰۰۵۶ - ۷۰۸}{۵۰۰۵۶}} = ۰/۰۰۲۳$$

بازهٔ اطمینان ۹۵ درصدی برای $\bar{\bar{\bar{X}}}$ عبارت است از

$$\bar{\bar{x}}_{clu} - ۱/۹۶ \times \hat{SE}\left(\bar{\bar{x}}_{clu}\right) \le \bar{\bar{\bar{X}}} \le \bar{\bar{x}}_{clu} + ۱/۹۶ \times \hat{SE}\left(\bar{\bar{x}}_{clu}\right)$$
$$۰/۰۰۹۸ - ۱/۹۶ \times ۰/۰۰۲۳ \le \bar{\bar{\bar{X}}} \le ۰/۰۰۹۸ + ۱/۹۶ \times ۰/۰۰۲۳$$
$$۰/۰۰۵۳ \le \bar{\bar{\bar{X}}} \le ۰/۰۱۴۳$$

بالاخره به برآورد کردن نسبت افراد ترخیص‌شدهٔ مرده از میان دارندگان شرایط مهلک می‌پردازیم. محاسبات مربوط به r_{clu} عبارت‌اند از

$$\bar{y}_{clu} = \left(\frac{۱}{m}\right)\left[\sum_{i=۱}^{m}\left(\frac{N_i}{n_i}\right)y_i\right] = \left(\frac{۱}{۳}\right)\left[\left(\frac{۴۲۸۸}{۴۲۹}\right)(۴۷) + \left(\frac{۶۳۸}{۶۴}\right)(۱۷) + \left(\frac{۲۱۵۴}{۲۱۵}\right)(۲۴)\right] = ۲۹۳/۲۳$$

$$r_{clu} = \frac{\bar{x}_{clu}}{\bar{y}_{clu}} = \frac{۴۹/۹۴}{۲۹۳/۲۳} = ۰/۱۷۰۳$$

محاسبات مربوط به $\hat{SE}(r_{clu})$ چنین‌اند :

$$\bar{x} = \frac{۵ + ۷ + ۳}{۳} = ۵ \qquad \bar{y} = \frac{۴۷ + ۱۷ + ۲۴}{۳} = ۲۹/۳۳$$

$$\frac{\sum_{i=1}^{m}(y_i-\bar{y})^2}{m-1}=۲۴۶/۳۳$$
$$\frac{\sum_{i=1}^{m}(x_i-\bar{x})^2}{m-1}=۴/۰۰$$
$$\frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{m-1}=-۷$$

$$\hat{SE}(r_{clu})=۰/۱۷۰۳\sqrt{\frac{۵۰۰۵۶-۷۰۸}{(۵۰۰۵۶\times ۳)}\left(\frac{۴}{۵^2}+\frac{۲۴۶/۳۳}{۲۹/۳۳^2}-۲\frac{(-۷)}{(۵\times ۲۹/۳۳)}\right)^{1/2}}=۰/۰۷۱۹$$

بازهٔ اطمینان ۹۵ درصدی برای R عبارت است از

$$r_{clu}-۱/۹۶\times \hat{SE}(r_{clu})\le R\le r_{clu}+۱/۹۶\times \hat{SE}(r_{clu})$$
$$۰/۱۷۰۳-۱/۹۶\times ۰/۰۷۱۹\le R\le ۰/۱۷۰۳+۱/۹۶\times ۰/۰۷۱۹$$
$$۰/۰۲۹۴\le R\le ۰/۳۱۱۲$$

باید تأکید کرد که عبارتهای مربوط به برآورد خطای معیار برآوردها تنها در صورتی کاربرد دارند که خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب شده باشند و کسر نمونه‌گیری واحدهای فهرست‌برداری که در مرحلهٔ دوم به کار می‌رود برای همهٔ خوشه‌هایی که در نمونه انتخاب شده‌اند یکسان باشد. تحت این انتساب، هر واحد فهرست‌برداری موجود در جامعه، شانس یکسان برای انتخاب شدن در نمونه دارد. به عبارت دیگر، همان طور که قبلاً در مورد نمونه‌گیری طبقه‌بندی شده بحث شد، نمونه خود ـ وزن است. در مورد طرحهای نمونه‌ای که خود ـ وزن نیستند، برآورد کردن واریانس به مراتب پیچیده‌تر می‌شود.
□

برآورد کردن پارامترهای جامعه و خطای معیار آنها نیز مانند طرحهای نمونه‌گیری قبلی می‌تواند با استفاده از نرم‌افزارهای آماری مناسب اجرا شود، و ما ذیلاً نشان خواهیم داد که چگونه می‌توان این کار را با استفاده از STATA و SUDAAN انجام داد. پروندهٔ داده‌های مورد نیاز برای داده‌پردازی، شامل ۷۰۸ سابقه محتوی اطلاعات مربوط به بیماران نمونه (واحدهای شمارش) است که در سه بیمارستان نمونه (خوشه‌ها یا واحدهای نمونه‌گیری اولیهٔ PSU) پذیرش شده‌اند. قالب این سوابق و متغیرهای مورد استفاده در محاسبات در زیر نشان داده شده‌اند:

HOSPON	ID	LIFETHRT	DXDEAD	M	NI	W
1	1	1	1	10	4288	33.31779
1	2	1	1	10	4288	33.31779
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	429	0	0	10	4288	33.31779
4	1	1	1	10	638	33.22916
4	2	1	1	10	638	33.22916
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	64	0	0	10	638	33.22916
10	1	1	0	10	2154	33.39535
10	2	1	0	10	2154	33.39535
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	215	0	0	10	2154	33.39535

HOSPNO شمارهٔ شناسایی بیمارستان نمونه‌ای است که بیمار در آن پذیرش شده است.

ID شماره‌ای است که بیمار خاص را شناسایی می‌کند.

LIFETHRT متغیری است که نشان می‌دهد بیمار پذیرش شده دچار شرایط مهلک بوده است یا نه.

DXDEAD متغیری است که نشان می‌دهد بیمار ترخیص شده مرده بوده است یا نه.

M نشانگر تعداد واحدهای نمونه‌گیری اولیه یا خوشه‌های جامعه است که نمونه از آنها انتخاب شده است. این نشانگر، توسط STATA در تعیین تصحیح جامعهٔ متناهی برای نمونه‌گیری مرحلهٔ اول به کار می‌رود.

NI نشانگر کل تعداد واحدهای شمارش (مثلاً بیماران پذیرش شده در بیمارستان) در i امین خوشه (مثلاً بیمارستان) است. توجه کنید که این تعداد برای هر خوشه یکسان نیست. این نشانگر، در تعیین تصحیح جامعهٔ متناهی برای نمونه‌گیری مرحلهٔ دوم توسط SUDAAN به کار می‌رود.

W وزن نمونه‌گیری (عکس کسر نمونه‌گیری کلی) برای هر سابقه است. توجه کنید که این کسر برای هر خوشه دقیقاً یکسان *نیست*، زیرا تفاوتهای موجود بین خوشه‌ها از لحاظ تعداد واحدهای شمارش باعث می‌شود که به دست آوردن یک کسر نمونه‌گیری دقیقاً یکسان برای مرحلهٔ دوم غیرممکن شود.

فرمانهای زیر در STATA برای پروندهٔ داده‌ها، *il10pt2.dta*، مورد استفاده قرار می‌گیرند که دارای ساختاری به شکل نشان داده شده در بالاست.

```
Use " a : \ il10pt2.dta", clear
. svyset psu hospno
. svyset pweight w
. svytotal dxdead
. svyratio dxdead lifethrt
```

دو فرمان اول، متغیرهای خوشه و وزن نمونه‌گیری را نشان می‌دهند، و دو فرمان آخر حاکی از شیوه‌های برآورد است که باید اجرا شوند. همان‌طور که قبلاً بیان شد، STATA در شیوه‌های خود برای برآورد کردن از طریق نمونه‌گیری خوشه‌ای از تصحیح جامعهٔ متناهی استفاده نمی‌کند. خروجی STATA که از این فرمانها تولید می‌شود در زیر نشان داده شده است:

Survey total estimation

pweight:	w	Number of obs	= 708
Strata:	<one>	Number of strata	= 1
PSU:	hospno	Number of PSUs	= 3
		Population size	= 23600

Total	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
Dxdead	499.3792	114.6776	5.961245	992.7971	.8059769

. svyratio dxdead lifethrt

Survey ratio estimation

pweight:	w	Number of obs	= 708
Strata:	<one>	Number of strata	= 1
PSU:	hospno	Number of PSUs	= 3
		Population size	= 23600

Ratio	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
dxdead/lifethrt	.1703017	.072273	-.140664	.4812674	3.247372

فرمانهای زیر در SUDAAN برای برآورد مجموع و نسبت ترخیص‌شدگان مرده از میان افراد دچار شرایط مهلک مورد استفاده قرار می‌گیرند:

```
1  PROC DESCRIPT DATA = IL10PT2 FILETYPE = SAS  DESIGN = WOR  MEANS
   TOTALS;
2  NEST_ONE_HOSPNO;
3  WEIGHT  W;
4  TOTCNT  M  NI;
5  VAR  DXDEAD  LIFETHRT;
6  SETENV  COLWIDTH = 13;
7  SETENV  DECWIDTH = 3;
```

```
8   PROC RATIO DATA = IL10PT2 FILETYPE = SAS  DESIGN = WOR;
9   NEST_ONE_HOSPNO;
10  WEIGHT  W;
11  TOTCNT  M  NI;
12  NUMBER  DXDEAD;
13  DENOM  LIFETHRT;
14  SETENV  COLWIDTH = 13;
15  SETENV  DECWIDTH = 3;
```

خروجی حاصل از SUDAAN در صفحهٔ بعد نشان داده شده است.

برآورد خطای معیارهای حاصل از سه روش مختلف برآورد (متن، STATA وSUDAAN) در زیر نشان داده شده‌اند. برآوردهای نقطه‌ای حاصل از هر سه روش یکسان‌اند. باید به خاطر داشت که STATA در شیوهٔ برآورد کردن خود برای نمونه‌گیری خوشه‌ای، از تصحیح جامعهٔ متناهی استفاده نمی‌کند.

برآورد خطای معیار برآوردها

برآورد	برآورد نقطه‌ای	خطای معیار برآورد حاصل از فرمولهای متن در تابلوی ۵.۱۰	خطای معیار برآورد حاصل از STATA	خطای معیار برآورد حاصل از SUDAAN با فرض طرح بدون جایگذاری	خطای معیار برآورد حاصل از SUDAAN با فرض طرح با جایگذاری
کل تعداد افراد ترخیص شدهٔ مرده	٤٩٩/٣٨	١١٤/٦٥	١١٤/٦٨	١١٦/٠٧	١١٤/٦٨
نسبت بیماران ترخیص‌شدهٔ مرده از میان افراد دچار شرایط مهلک	٠/١٧٠٣	٠/٠٧٢	٠/٠٧٢	٠/٠٦٤	٠/٠٧٢

همان‌طور که در بالا نشان داده شد، برآورد خطای معیارهای حاصل از هر سه روش بسیار به یکدیگر نزدیک‌اند.

٤.٢.١٠ توزیع نمونه‌گیری برآوردها

برای نشان دادن توزیعهای نمونه‌گیری برآوردها در نمونه‌گیری خوشه‌ای دومرحله‌ای ساده از خوشه‌هایی که تعداد واحدهای فهرست‌برداری آنها نابرابر است، به یک مثال نگاهی می‌اندازیم.

Number of observations read: 708 weighted count: 23600
Number of observations skipped: 0
Denominator degrees of freedom: 2

Variable		One 1
DXDEAD	Sample Size	708.000
	Weighted Size	23599.988
	Total	499.379
	SE Total	116.072
	Mean	0.021
	SE Mean	0.011
LIFETHRT	Sample Size	708.000
	Weighted Size	23599.988
	Total	2932.319
	SE Total	773.082
	Mean	0.124
	SE Mean	0.018

by: Variable, One.
Number of observations read: 708 weighted count: 23600
Number of observations skipped: 0
Denominator degrees of freedom: 2

by : Variable, One.

Variable		One 1
DXDEAD/LIFETHRT	Sample Size	708.000
	Weighted Size	23599.988
	Weighted X-Sum	2932.319
	Weighted Y-Sum	499.379
	Ratio Est	0.170
	SE Ratio	0.064

مثال تشریحی: فرض کنید می‌خواهیم X، کل تعداد بیماران مردهٔ ترخیص‌شده و R، نسبت همهٔ پذیرش‌شدگانی را که از ده بیمارستان نشان داده شده در جدول ۹.۱۰ مرده ترخیص شده‌اند برآورد کنیم (برای این مثال فرض خواهیم کرد که فوت فقط در میان افراد دچار شرایط مهلک اتفاق افتاده است). فرض کنید که (به خاطر سهولت کار) این عمل را با استفاده از یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از دو بیمارستان انجام خواهیم داد. توزیعهای نمونه‌گیری برآورد کل ترخیص‌شدگان مرده، x'_{clu}، برآورد کل پذیرش‌شدگان، y'_{clu}، و برآورد نسبت همهٔ بیماران مرده ترخیص‌شده، r_{clu}، در جدول ۱۱.۱۰ نشان داده شده‌اند.

جدول ۱۱.۱۰ توزیع نمونه‌گیری x'_{clu} ، y'_{clu} و r_{clu}

بیمارستانها در نمونه	y'_{clu}	x'_{clu}	r_{clu}
۱ , ۲	٤٦٦٢٠	٦٠٠	۰/۰۱۲۹
۱ , ۳	۲۷۳۳۰	۲۹۵	۰/۰۱۰۸
۱ , ٤	۲٤٦۳۰	۲۵۵	۰/۰۱۰٤
۱ , ۵	۱۵٦٤۹۰	۱۹۰۰	۰/۰۱۲۱
۱ , ٦	۲۷۰۵۰	۲۹۵	۰/۰۱۰۹
۱ , ۷	۳۲۱۱۰	۳۹۵	۰/۰۱۲۳
۱ , ۸	۳۰۵٦۰	۳۰۰	۰/۰۰۹۸
۱ , ۹	٤٤۸۰۰	۵۵۰	۰/۰۱۲۳
۱ , ۱۰	۳۲۲۱۰	۳٤۵	۰/۰۱۰۷
۲ , ۳	۳۱۰۷۰	٤۷۵	۰/۰۱۵۳
۲ , ٤	۲۸۳۷۰	٤۳۵	۰/۰۱۵۳
۲ , ۵	۱٦۰۲۳۰	۲۰۸۰	۰/۰۱۳۰
۲ , ٦	۳۰۷۹۰	٤۷۵	۰/۰۱۵٤
۲ , ۷	۳۵۸۵۰	۵۷۵	۰/۰۱٦۰
۲ , ۸	۳٤۳۰۰	٤۸۰	۰/۰۱٤۰
۲ , ۹	٤۸۵٤۰	۷۳۰	۰/۰۱۵۰
۲ , ۱۰	۳۵۹۵۰	۵۲۵	۰/۰۱٤٦
۳ , ٤	۹۰۸۰	۱۳۰	۰/۰۱٤۳
۳ , ۵	۱٤۰۹٤۰	۱۷۷۵	۰/۰۱۲٦
۳ , ٦	۱۱۵۰۰	۱۷۰	۰/۰۱٤۸
۳ , ۷	۱٦۵٦۰	۲۷۰	۰/۰۱٦۳
۳ , ۸	۱۵۰۱۰	۱۷۵	۰/۰۱۱۷
۳ , ۹	۲۹۲۵۰	٤۲۵	۰/۰۱٤۵
۳ , ۱۰	۱٦٦٦۰	۲۲۰	۰/۰۱۳۲
٤ , ۵	۱۳۸۲٤۰	۱۷۳۵	۰/۰۱۲٦
٤ , ٦	۸۸۰۰	۱۳۰	۰/۰۱٤۸
٤ , ۷	۱۳۸٦۰	۲۳۰	۰/۰۱٦٦
٤ , ۸	۱۲۳۱۰	۱۳۵	۰/۰۱۱۰
٤ , ۹	۲٦۵۵۰	۳۸۵	۰/۰۱٤۵
٤ , ۱۰	۱۳۹٦۰	۱۸۰	۰/۰۱۲۹
۵ , ٦	۱٤۰٦٦۰	۱۷۷۵	۰/۰۱۲٦
۵ , ۷	۱٤۵۷۲۰	۱۸۷۵	۰/۰۱۲۹
۵ , ۸	۱٤٤۱۷۰	۱۷۸۰	۰/۰۱۲۳
۵ , ۹	۱۵۸٤۱۰	۲۰۳۰	۰/۰۱۲۸
۵ , ۱۰	۱٤۵۸۲۰	۱۸۲۵	۰/۰۱۲۵
٦ , ۷	۱٦۲۸۰	۲۷۰	۰/۰۱٦٦
٦ , ۸	۱٤۷۳۰	۱۷۵	۰/۰۱۱۹
٦ , ۹	۲۸۹۷۰	٤۲۵	۰/۰۱٤۷
٦ , ۱۰	۱٦۳۸۰	۲۲۰	۰/۰۱۳٤
۷ , ۸	۱۹۷۹۰	۲۷۵	۰/۰۱۳۹
۷ , ۹	۳٤۰۳۰	۵۲۵	۰/۰۱۵٤
۷ , ۱۰	۲۱٤٤۰	۳۲۰	۰/۰۱٤۹
۸ , ۹	۳۲٤۸۰	٤۳۰	۰/۰۱۳۲
۸ , ۱۰	۱۹۸۹۰	۲۲۵	۰/۰۱۱۳
۹ , ۱۰	۳٤۱۳۰	٤۷۵	۰/۰۱۳۹

میـانگین، خطای معیار، و ضریب تغییرات توزیع x'_{clu} و r_{clu} که از شمارش بیش از ٤٥ نمونهٔ ممکن با استفاده از فنون فصل ۲ به دست آمده‌اند به شرح زیرند:

$E(x'_{clu}) = ۶۵۱$	$E(r_{clu}) = ۰/۰۱۳۴$
$X = ۶۵۱$	$R = ۰/۰۱۳۰$
$SE(x'_{clu}) = ۶۲۳/۳۱۳$	$SE(r_{clu}) = ۰/۰۰۱۷$
$V(x'_{clu}) = ۰/۹۵۷۵$	$V(r_{clu}) = ۰/۱۳۰۸$

توجه کنید که برآورد کل ترخیص‌شدگان مرده، x'_{clu}، دارای خطای معیار بسیار زیاد و ضریب تغییراتی مـتجاوز از ۹۵ درصـد اسـت، در حـالی کـه برآورد نسبت بیمارانی که مرده ترخیص شده‌اند، r_{clu} ، هم خطـای معیـار کمـی دارد و هم ضریب تغییرات آن کم است. این تغییرپذیری زیاد در میان مجموعهای بـرآورد شـدهٔ x'_{clu} بـه وضـوح در توزیع فراوانی آن در ٤٥ نمونه (جدول ۱۲.۱۰) دیده می‌شود. توجه کـنید کـه فقط دو نمونه از ٤٥ نمونهٔ ممکن مقادیری را برای x'_{clu} نتیجه می‌دهند که در بازهٔ ۷۹۹–٦۰۰ واقـع مـی‌شوند، بازه‌ای که مجموع واقعی (X=٦٥۱) در آن قرار دارد. باز هم توجه کنید که ۹ نمونه از ٤٥ نمونـهٔ ممکـن مقادیـری را بـرای x'_{clu} نـتیجه مـی‌دهند که مقدار واقعی را به شدت بیش ـ برآورد می‌کند.
□

الگـوی نشان داده شده برای توزیعهای x'_{clu} و r_{clu} در مثال بالا مثال نوعی از چیزی است که غالباً هـنگام نمونـه‌گیـری تصـادفی سـاده از خوشه‌هایی که از لحاظ تعداد واحدهای فهرست‌برداری دارای تغییرپذیـری زیادی هستند روی می‌دهد. خوشه‌های شامل تعداد زیادی از واحدهای فهرست‌برداری در مقایسـه بـا خوشـه‌هـای شامل تعداد کمتری از واحدهای فهرست‌برداری، شانس بیشتری برای انتخاب شـدن ندارنـد. ایـن شـیوهٔ نمونه‌گیری، تعداد واحدهای فهرست‌برداری را در نظر نمی‌گیرد. در مثال ما، بیمارسـتانها خوشه‌ها را تشکیل می‌دهند و سوابق پزشکی بیمارستانها واحدهای فهرست‌برداری هستند. در میـان ده بیمارسـتان از لحـاظ تعـداد پذیـرش تغییـرات قابل توجهی وجود دارد به طوری که تعداد بیمـارانی که در یک بیمارستـــان (شمـــارهٔ ۵) پذیرش شـــده‌اند بیش از مجمـــوع پذیرش‌شـــدگان در ۹ بیمارسـتان دیگـر اسـت. تعداد بیمارانی که مرده ترخیص شده‌اند همبستگی زیادی با تعداد پذیرشها دارد، و بـه همیـن دلیـل ایـن رقم نیز از بیمارستانی به بیمارستان دیگر بسیار متغیر است، به طوری که فـوت‌شدگان بیمارستان شمارهٔ ۵ بیش از ۵۰ درصد کل فوت‌شدگان در ده بیمارستان است. چون تعداد پذیرشـها در یـک بیمارسـتان معین چه در برنامهٔ نمونه‌گیری و چه در شیوهٔ برآورد کردن در نظر گرفته نمی‌شود توزیع برآورد فوت‌شدگان بیمارستان، تغییرپذیری بسیار زیادی را نشان می‌دهد.

جدول ۱۲.۱۰ توزیع فراوانی برآورد مجموع x'_{clu} در همهٔ نمونه‌های ممکن متشکل از دو بیمارستان

برآورد کل تعداد اشخاص مرده ترخیص شده، x'_{clu}	فراوانی f_i
۰–۱۹۹	۷
۲۰۰–۳۹۹	۱۵
٤۰۰–۵۹۹	۱۲
٦۰۰–۷۹۹	۲
۸۰۰–۹۹۹	۰
۱۰۰۰–۱۱۹۹	۰
۱۲۰۰–۱۳۹۹	۰
۱٤۰۰–۱۵۹۹	۰
۱٦۰۰–۱۷۹۹	٤
۱۸۰۰–۱۹۹۹	۳
۲۰۰۰–۲۱۹۹	۲
مجموع	٤۵

از سوی دیگر، *نسبت* بیمارانی که مرده ترخیص شده‌اند در هر یک از ده بیمارستان تقریباً یکسان است و همبستگی کمی با تعداد پذیرشها دارد. به این ترتیب، برخلاف برآورد مجموع x'_{clu}، برآورد نسبت بیماران مرده ترخیص شده دارای خطای معیار نسبتاً کمی است.

در فصل بعد، روشهای اصلاح شیوهٔ برآورد و یا برنامهٔ نمونه‌گیری را مورد بحث قرار خواهیم داد که وقتی N_i تعداد واحدهای فهرست‌برداری در میان خوشه‌ها بسیار متغیر باشد و سطح مشخصهٔ مورد برآورد در خوشه قویاً به تعداد واحدهای فهرست‌برداری در خوشه بستگی داشته باشد مجموعهای برآورد شده‌ای را نتیجه می‌دهد که خطای معیار کمتری دارند.

خطای معیار نظری برآورد مجموعها، میانگینها، و نسبتها، برای نمونه‌گیری خوشه‌ایی که در آن خوشه‌ها با نمونه‌گیری تصادفی ساده انتخاب می‌شوند و واحدهای فهرست‌برداری نیز در داخل خوشه‌های انتخاب شده با نمونه‌گیری تصادفی ساده انتخاب می‌شوند در تابلوی ۵.۱۰ ارائه شده‌اند.

وقتی همهٔ N_i ها برابر باشند (یعنی $N_i = \overline{N}$)، این عبارتها به عبارتهای فهرست شده در تابلوی ۳.۱۰ تبدیل می‌شوند.

با بررسی فرمول مربوط به خطای معیار یک مجموع برآورد شده، پی می‌بریم که جمله‌ای دارد که به σ^2_{1x}، یعنی واریانس بین خوشه‌ها از لحاظ کل سطح مشخصه‌های مورد اندازه‌گیری در خوشه بستگی

تابـلوی ۵.۱۰ خطـای معیـار نظـری بـرآوردهای جامعـه‌ای بـا استفاده از نمونه‌گیری خوشه‌ای دومرحله‌ای ساده با تعداد نابرابر واحدهای فهرست‌برداری

مجموع، x'_{clu}

$$SE(x'_{clu}) = \left\{\left(\frac{M^{۲}}{m}\right)\sigma^{۲}_{۱x}\left(\frac{M-m}{M-۱}\right) + \left(\frac{M}{m}\right)\left[\sum_{i=۱}^{M}\left(\frac{N_i}{n_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right)\sum_{j=۱}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)^{۲}\right]\right\}^{۱/۲}$$

میانگین به ازای خوشه، $\bar{x}_{clu}$

$$SE(\bar{x}_{clu}) = \left\{\left(\frac{\sigma^{۲}_{۱x}}{m}\right)\left(\frac{M-m}{M-۱}\right) + \left(\frac{۱}{Mm}\right)\left[\sum_{i=۱}^{M}\left(\frac{N_i}{n_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right)\sum_{j=۱}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)^{۲}\right]\right\}^{۱/۲}$$

میانگین به ازای واحد فهرست‌برداری، $\overline{\overline{x}}_{clu}$

$$SE\left(\overline{\overline{x}}_{clu}\right) = \left(\frac{۱}{N}\right)\left\{\left(\frac{M^{۲}}{m}\right)\sigma^{۲}_{۱x}\left(\frac{M-m}{M-۱}\right) + \left(\frac{M}{m}\right)\left[\sum_{i=۱}^{M}\left(\frac{N_i}{n_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right)\sum_{j=۱}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)^{۲}\right]\right\}^{۱/۲}$$

نسبت، r_{clu}

$$SE(r_{clu}) = R\left\{\left(\frac{\sigma^{۲}_{۱R}}{m\overline{X}^{۲}}\right)\left(\frac{M-m}{M-۱}\right) + \left(\frac{۱}{m\overline{X}^{۲}}\right)\sum_{i=۱}^{M}\left(\frac{۱}{N_i n_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right) \times\left[\sum_{j=۱}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)^{۲} + R^{۲}\sum_{j=۱}^{N_i}\left(Y_{ij}-\overline{\overline{Y}}_i\right)^{۲} - ۲R\sum_{j=۱}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)\left(Y_{ij}-\overline{\overline{Y}}_i\right)\right]\right\}^{۱/۲}$$

نمادهای مورد استفاده در اینجا در تابلوهای ۱.۱۰ و ۳.۱۰ تعریف شده‌اند.

دارد و جملهٔ دیگری دارد که به واریانس بین واحدهای فهرست‌برداری داخل یک خوشهٔ معین از لحاظ سطح مشخصه بستگی دارد. و باز اگر مجموعهای خوشه‌ای، X_i، قویاً به N_i، تعداد واحدهای فهرست‌برداری موجود در خوشه همبسته باشند، و اگر در میان خوشه‌ها از لحاظ N_i تنوع زیادی وجود داشته باشد، آن‌گاه $\sigma^{۲}_{۱x}$ می‌تواند بسیار زیاد باشد و از این رو، خطای معیار مجموع برآورد شده نیز (همان‌طور که قبلاً مشاهده کردیم) می‌تواند زیاد باشد.

۵.۲.۱۰ اندازهٔ نمونهٔ مورد نیاز باید چقدر باشد؟

فرض کنید می‌خواهیم یک نمونهٔ خوشه‌ای دومرحله‌ای بگیریم که طی آن یک نمونهٔ تصادفی ساده از خوشه‌ها در مرحلهٔ اول و پس از آن یک نمونهٔ تصادفی ساده متشکل از n_i واحد فهرست‌برداری از N_i واحدهای فهرست‌برداری موجود در داخل هر خوشهٔ نمونه گرفته شود. به عبارت دیگر، فرض می‌کنیم که کسرهای نمونه‌گیری مرحلهٔ دوم قبلاً تعیین شده‌اند. پس می‌خواهیم بدانیم که به چند خوشه، m، در نمونهٔ خود نیاز داریم تا واقعاً مطمئن شویم که تفاوتهای نسبی بین برآوردهای ما و مقادیر واقعی بیش از ε نیست. فرمولهای مربوط به *تعداد m خوشهٔ* مورد نیاز برای تأمین ویژگیهای فوق‌الذکر ذیلاً ارائه می‌شوند.

برای برآورد *مجموعها* ($\bar{x}_{clu}$) یا *میانگینها* ($\bar{\bar{x}}_{clu}$, $\bar{x}_{clu}$):

$$m = \frac{\left(\frac{\sigma^{۲}_{۱x}}{\bar{\bar{X}}^{۲}}\right)\left(\frac{M}{M-۱}\right)+\left(\frac{M}{N^{۲}\bar{\bar{X}}^{۲}}\right)\left[\sum_{i=۱}^{M}\left(\frac{N_i}{n_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right)\sum_{j=۱}^{N_i}\left(X_{ij}-\bar{\bar{X}}_i\right)^{۲}\right]}{\frac{\varepsilon^{۲}}{z^{۲}_{۱-(\alpha/۲)}}+\frac{\sigma^{۲}_{۱x}}{\bar{\bar{X}}^{۲}(M-۱)}} \qquad (۱۶.۱۰)$$

و برای برآورد *نسبتها* ($r_{clu}=\bar{x}_{clu}/\bar{y}_{clu}$) :

$$m = \frac{\left[\begin{array}{c}\left(\frac{\sigma^{۲}_{۱R}}{\bar{\bar{X}}^{۲}}\right)\left(\frac{M}{M-۱}\right)+\left(\frac{M}{N^{۲}\bar{\bar{X}}^{۲}}\right)\sum_{i=۱}^{M}\left(\frac{۱}{n_iN_i}\right)\left(\frac{N_i-n_i}{N_i-۱}\right)\\ \times\left[\sum_{j=۱}^{N_i}\left(X_{ij}-\bar{\bar{X}}_i\right)^{۲}+R^{۲}\sum_{j=۱}^{N_i}\left(Y_{ij}-\bar{\bar{Y}}_i\right)^{۲}-۲R\sum_{j=۱}^{N_i}\left(X_{ij}-\bar{\bar{X}}_i\right)\left(Y_{ij}-\bar{\bar{Y}}_i\right)\right]\end{array}\right]}{\frac{\varepsilon^{۲}}{z^{۲}_{۱-(\alpha/۲)}}+\frac{\sigma^{۲}_{۱R}}{\bar{\bar{X}}^{۲}(M-۱)}}$$

(۱۷.۱۰)

مـثال تشـریحی: فرض کنید می‌خواهیم واقعاً مطمئن باشیم که کل تعداد افراد مرده ترخیص‌شده را در محـدودهٔ ۳۰ درصـد مقدار واقعی و نسبت فوت‌شدگان در میان افراد پذیرش‌شده با شـرایط مهلک را در محـدودهٔ ۲۰ درصـد مقدار واقعی برآورد می‌کنیـم. فرض کنید نمونهٔ مرحلهٔ دوم را ۲۰ درصد (بـه نزدیکـترین عـدد صحیح) پذیرشها در بیمارستانهای انتخاب شده در مرحلهٔ اول می‌گیریم. پس، از داده‌های جامعه در جدول ۹.۱۰ داریم

$$\sigma_{1x}^{2} = ۸۷۴۱/۶۹ \qquad \overline{X} = ۶۵/۱ \qquad \overline{\overline{X}} = ۰/۰۱۳۰$$

$$\sum_{i=1}^{M}\left(\frac{N_i}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)\sum_{j=1}^{N_i}\left(X_{ij} - \overline{\overline{X}}_i\right)^{2} = ۵۷۸۳/۳۲$$

$$\sum_{i=1}^{M}\left(\frac{1}{N_i n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)\left[\sum_{j=1}^{N_i}\left(X_{ij} - \overline{\overline{X}}_i\right)^{2} + R^{2}\sum_{j=1}^{N_i}\left(Y_{ij} - \overline{\overline{Y}}_i\right)^{2} - 2R\sum_{j=1}^{N_i}\left(X_{ij} - \overline{\overline{X}}_i\right)\left(Y_{ij} - \overline{\overline{Y}}_i\right)\right] = ۰/۰۰۰۶۴$$

$$\sigma_{1R}^{2} = ۸۱/۳۹ \qquad N = ۵۰۰۵۶ \qquad M = ۱۰$$

سپس از رابطهٔ (۱٦.۱۰) با $\varepsilon = ۰/۳۰$، داریم

$$m = \frac{\left(\frac{۸۷۴۱/۶۹}{(۶۵/۱)^{2}}\right)\left(\frac{۱۰}{۱۰-۱}\right) + \left(\frac{۱۰}{(۵۰۰۵۶)^{2}(۰/۰۱۳۰)^{2}}\right)(۵۷۸۳/۳۲)}{\frac{(۰/۳۰)^{2}}{۹} + \frac{۸۷۴۱/۶۹}{(۶۵/۱)^{2}(۱۰-۱)}} = ۱۰/۱۵ \approx ۱۰$$

و از رابطهٔ (۱۷.۱۰) با $\varepsilon = ۰/۳۰$ داریم

$$m = \frac{\left(\frac{۸۱/۳۹}{(۶۵/۱)^{2}}\right)\left(\frac{۱۰}{۱۰-۱}\right) + \left(\frac{۱۰}{(۵۰۰۵۶)^{2}(۰/۰۱۳۰)^{2}}\right)(۰/۰۰۰۶۴)}{\frac{(۰/۳۰)^{2}}{۹} + \frac{۸۱/۳۹}{(۶۵/۱)^{2}(۱۰-۱)}} = ۱/۷۶ \approx ۲$$

به این ترتیب، به منظور تأمین ویژگیهای تعیین شده برای برآورد کل ترخیص‌شدگان فوت شده (X_i)، بـه نمونـه‌ای متشـکل از همـهٔ ده بیمارستان نیاز خواهیم داشت، در حالی که در تأمین ویژگیهای تعیین شده برای نسبت (R) فوت‌شدگان مبتلا به شرایط مهلک، تنها به دو خوشه نیاز داریم.

□

۶.۲.۱۰ انتخاب اندازهٔ بهینهٔ خوشه با توجه به هزینه‌ها

باز فرض می‌کنیم که خوشه‌هایی با نمونه‌گیری تصادفی ساده انتخاب می‌شوند و کسر نمونه‌گیری n_i/N_i در داخل هر خوشهٔ نمونه‌گیری مرحلهٔ دوم برای همهٔ خوشه‌های انتخاب شده (در محدوده‌های N_i که قبلاً بحث شد) یکسان باشد. فرض کنید اندازهٔ متوسط خوشه $\overline{n}$ را به صورت زیر نشان می‌دهیم

$$\overline{n} = \frac{\sum_{i=1}^{m} n_i}{m}$$

و فرض کنید *هزینه‌های میدانی* را بتوانیم با تابع زیر تقریب کنیم

$$C = C_1\ m + C_2\ m\overline{n} \qquad (۱۸.۱۰)$$

در این صورت، *متوسط اندازهٔ بهینهٔ خوشه‌ای* $\overline{n}$ که بتواند برآوردی از مجموعها (x'_{clu}) یا میانگینها ($\overline{x}_{clu}$, $\overline{\overline{x}}_{clu}$) را نتیجه دهد که کمترین خطاهای معیار را در مقایسه با همهٔ برآوردهای حاصل از نمونه‌گیری خوشه‌ای دومرحله‌ای با کسر نمونه‌گیری مرحلهٔ دوم ثابت و با همان هزینه‌های میدانی داشته باشد از فرمول زیر به دست می‌آید

$$\overline{n} = \left[\left(\frac{C_1}{C_2}\right)\left(\frac{1-\delta_x}{\delta_x}\right)\right]^{1/2} \qquad (۱۹.۱۰)$$

که در آن δ_x، تعمیم ضریب همبستگی درون رده‌ای است که قبلاً مورد بحث قرار گرفت و از فرمول زیر به دست می‌آید

$$\delta_x = \frac{[M/(M-1)]\sigma_{1x}^2 - \overline{N}\sigma_{2x}^2}{[M/(M-1)]\sigma_{1x}^2 + \overline{N}(\overline{N}-1)\sigma_{2x}^2} \qquad (۲۰.۱۰)$$

و در آن

$$\overline{N} = \frac{\sum_{i=1}^{M} N_i}{M}$$

$$\sigma_{2x}^2 = \left(\frac{1}{N}\right)\sum_{i=1}^{M}\left(\frac{N_i}{N_i-1}\right)\sum_{j=1}^{N_i}\left(X_{ij} - \overline{\overline{X}}_i\right)^2 \qquad (۲۱.۱۰)$$

برای برآورد کردن نسبتها، ($r_{clu} = \overline{x}_{clu}/\overline{y}_{clu}$)، اندازهٔ بهینهٔ خوشه‌ای $\overline{n}$ از فرمول زیر به دست می‌آید

$$\overline{n} = \left[\left(\frac{C_1}{C_2}\right)\left(\frac{1-\delta_R}{\delta_R}\right)\right]^{1/2} \qquad (۲۲.۱۰)$$

که در آن

$$\delta_R = \frac{\left[M/(M-1)\right]\sigma_{1R}^2 - \overline{N}\sigma_{2R}^2}{\left[M/(M-1)\right]\sigma_{1R}^2 + \overline{N}(\overline{N}-1)\sigma_{2R}^2} \qquad (۲۳.۱۰)$$

$$\sigma_{1R}^2 = \sigma_{1x}^2 + R^2\sigma_{1y}^2 - 2R\sigma_{1xy}$$

و

$$\sigma_{2R}^2 = \left(\frac{1}{N}\right)\sum_{i=1}^{M}\left(\frac{N_i}{N_i-1}\right)\left[\sum_{j=1}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)^2 + R^2\sum_{j=1}^{N_i}\left(Y_{ij}-\overline{\overline{Y}}_i\right)^2 - 2R\sum_{j=1}^{N_i}\left(X_{ij}-\overline{\overline{X}}_i\right)\left(Y_{ij}-\overline{\overline{Y}}_i\right)\right] \qquad (۲٤.۱۰)$$

مـثال تشریحی: برای نشان دادن کاربری این فرمولها، فرض کنید می‌خواهیم از جامعهٔ ده بیمارستان که در جـدول ۹.۱۰ نشان داده شده است برای برآورد کردن کل تعداد اشخاصی که مرده ترخیـص شده‌اند (X) و نسـبت فـوت‌شـدگان (R) در میان افرادی که با داشتن شرایط مهلک پذیرش شده‌اند، یک نمونهٔ خوشـه‌ای دومـرحله‌ای سـاده بگیـریم. فـرض کنید هزینه‌های اداری و رفت و آمد برای هر بیمارستان نمـونـه تقریـباً ۵۰۰ دلار و هزیـنـهٔ بررسـی هر سابقهٔ پذیرش تقریباً ۵ دلار باشد. به عبارت دیگـر، $C_1 = ۵۰۰$ و $C_2 = ۵$. از روی جدول ۹.۱۰ محاسبات زیر را انجام می‌دهیم

$$\overline{N} = ۵۰۰۵/۶ \qquad \sigma_{2x}^2 = \frac{۶۴۲/۳۷}{۵۰۰۵۶} = ۰/۰۱۲۸۳ \qquad \sigma_{1x}^2 = ۸۷۴۱/۶۹$$

$$\delta_x = \frac{\frac{۱۰}{۹}(۸۷۴۱/۶۹) - (۵۰۰۵/۶)(۰/۰۱۲۸۳)}{\frac{۱۰}{۹}(۸۷۴۱/۶۹) + (۵۰۰۵/۶)(۵۰۰۴/۶)(۰/۰۱۲۸۳)} = ۰/۰۲۹۱۴$$

بنابراین

$$\overline{n} = \left[\left(\frac{۵۰۰}{۵}\right)\left(\frac{۱-۰/۰۲۹۱۴}{۰/۰۲۹۱۴}\right)\right]^{1/2} = ۵۷/۷۵ \approx ۵۸$$

بـه ایـن تـرتیب، متوسـط انـدازهٔ بهیـنهٔ خوشـه برای برآورد کردن کل تعداد ترخیص‌شدگان مرده، ۵۸ پذیـرش اسـت و کسر بهینهٔ نمونه‌گیری مرحلهٔ دوم $f_2 = ۵۸/۵۰۰۵/۶ = ۰/۰۱۱۵۸۶$ خواهد بود. به محض ایـن‌کـه کسـر بهیـنهٔ نمونـه‌گیـری مرحـلهٔ دوم تعییـن شـد از رابطـهٔ (۱٦.۱۰) استفاده می‌کنیم تا تعداد خوشه‌های لازم برای تأمین ویژگیهایی که برای برآورد منظور شده‌اند تعیین شود.

بـرای نشـان دادن کاربـری رابطهٔ (۲۲.۱۰) در مورد نسبت R، از روی جدول ۹.۱۰ و فرمولهایی که قبلاً ارائه شد، داریم

$$\sigma_{1R}^2 = ۸۱/۳۹$$
$$\sigma_{2R}^2 = \sigma_{2x}^2 + R^2\sigma_{2y}^2 - 2R\sigma_{2xy}$$
$$= ۰/۰۱۲۸۳ + (۰/۰۹۱۸۶)^2(۰/۱۲۰۸۱) - ۲(۰/۰۹۱۸۶)(۰/۰۱۱۱۳) = ۰/۰۱۱۸۰$$
$$\delta_R = \frac{\frac{۱۰}{۹}(۸۱/۳۹) - (۵۰۰۵/۶)(۰/۰۱۱۸۰)}{\frac{۱۰}{۹}(۸۱/۳۹) + (۵۰۰۵/۶)(۵۰۰۴/۶)(۰/۰۱۱۸۰)} = ۰/۰۰۰۱۰۶$$

بنابراین

$$\bar{n} = \left[\left(\frac{۵۰۰}{۵}\right)\left(\frac{۱ - ۰/۰۰۰۱۰۶}{۰/۰۰۰۱۰۶}\right)\right]^{1/2} = ۹۷۱/۲۳۴$$

بـه این ترتیب، کسر بهینهٔ نمونه‌گیری مرحلهٔ دوم برای برآورد کردن نسبت افراد مرده ترخیص شده، به افراد پذیرش شده با شرایط مهلک، عبارت است از

$$f_2 = \frac{۹۷۱/۲۳۴}{۵۰۰۵/۶} = ۰/۱۹۴$$

سـپس می‌توانیم از رابطهٔ (۱۷.۱۰) (با فرض اینکه n_i برابر با $f_2 N_i$ باشد) برای تعیین تعداد خوشه‌هایی استفاده کنیم که باید برای تأمین ویژگیهای مورد نظر برآورد، نمونه‌گیری شوند.
□

غالـباً پیـش مـی‌آیـد کـه انـدازه‌های بهینهٔ خوشه برای یکایک برآوردهای مورد نیاز یکسان نیستند (همـان طـور کـه در ایـن مـثال پیش آمد). در عمل، معمولاً به نوعی مصالحه دربارهٔ اندازهٔ خوشه $\bar{n}$ با اسـتفاده از روشـی همچـون محاسـبهٔ میـانگین بهیـنهٔ $\bar{n}$، برای بیشتر برآوردهای مهمی که از آمارگیری موردنیاز است تن می‌دهیم.

۳.۱۰ نمونه‌گیری سیستماتیک به صورت نمونه‌گیری خوشه‌ای

یـک نمونـهٔ سیسـتماتیک متشکل از یکی از هر M واحد فهرست‌برداری که با یک عدد تصادفی شروع می‌شود، به طوری که در فصل ٤ شرح داده شد، در واقع یک نمونهٔ خوشه‌ای یک‌مرحله‌ای است که در آن فقـط یـک واحـد نمونـه‌گیـری اولیـه نمونـه‌گیـری مـی‌شود. این مطلب هنگامی آشکار می‌شود که واحدهـای فهرسـت‌بـرداری را (کـه بـه تـرتیب درج در فهرسـت، شـماره‌گـذاری شده‌اند) به صورت گروه‌بندی شده در M خوشهٔ زیر در نظر بگیریم:

خوشه	واحدهای فهرست‌برداری
۱	۱ , ۱+ M , ۱+ ۲M , ۱+ ۳M , . . .
۲	۲ , ۲+ M , ۲+ ۲M , ۲+ ۳M , . . .
۳	۳ , ۳+ M , ۳+ ۲M , ۳+ ۳M , . . .
⋮	⋮
M −۱	M −۱, ۲M −۱, ۳M −۱, ٤M −۱, . . .
M	M , ۲M , ۳M , ٤M , . . .

برای سهولت کار، فرض می‌کنیم که N ، تعداد واحدهای فهرست‌برداری دقیقاً قابل قسمت به M است به طوری که هر یک از خوشه‌های بالا دارای $\overline{N} = N/M$ واحد فهرست‌برداری است. در این صورت اگر شمارهٔ انتخاب شده برای شروع نمونهٔ سیستماتیک ۱ باشد، نمونه شامل همهٔ واحدهای فهرست‌برداری در خوشهٔ ۱ خواهد بود، یعنی واحدهای فهرست‌برداری ۱، ۱+M، ۱+۲M، . . . ، ۱+$(\overline{N}-۱)M$. به همین ترتیب اگر شمارهٔ تصادفی ۲ باشد نمونه شامل همهٔ واحدهای موجود در خوشهٔ ۲ خواهد بود و همین‌طور الی آخر. پس، نظریهٔ نمونه‌گیری سیستماتیک با شروع تصادفی، حالتی خاص از نمونه‌گیری خوشه‌ای یک‌مرحله‌ای با m =۱ و $\overline{n} = \overline{N}$ محسوب می‌شود.

٤.۱۰ خلاصه

در این فصل، روشهایی را برای نمونه‌گیری خوشه‌ای دومرحله‌ای شرح و بسط دادیم که در آن، خوشه‌ها در مرحلهٔ اول با احتمال برابر انتخاب می‌شدند و در داخل هر خوشهٔ نمونه، کسرهای نمونه‌گیری مرحلهٔ دوم، یکسان (یا تقریباً یکسان) بودند. هرگاه نمونه‌گیری در هر مرحله توسط نمونه‌گیری تصادفی ساده انجام شود، برنامهٔ نمونه‌گیری را *نمونه‌گیری خوشه‌ای دومرحله‌ای ساده* می‌نامند و ما بحث خود را در این فصل بر همین برنامهٔ نمونه‌گیری متمرکز کردیم. بخش موردنظر به دو زیربخش عمده تفکیک شد: یکی به وضعیتی می‌پرداخت که طی آن همهٔ خوشه‌های موجود در جامعه دارای تعداد واحدهای شمارش یکسان بودند و دیگری به وضعیتی پیچیده‌تر می‌پرداخت که در آن خوشه‌ها از لحاظ تعداد واحدهای شمارش با یکدیگر تفاوتهایی داشتند. برای هر یک از این دو وضعیت، روشهای برآورد کردن مشخصه‌های جامعه و برآورد کردن خطاهای معیار این برآوردها را مورد بحث قرار دادیم. در این فصل، ضریب همبستگی درون رده‌ای را به عنوان شاخصی از همگنی میان واحدهای شمارش در داخل خوشه‌های نمونه با توجه به مشخصهٔ x مورد اندازه‌گیری معرفی

کـردیم. ایـن شـاخص، تعییـن کنندهٔ مهمی برای تعداد بهینهٔ واحدهای فهرست‌برداری است که باید در مرحـلهٔ دوم در داخـل هـر خوشـه انـتخاب شـوند و نیز تعیین کنندهٔ خطاهای نمونه‌گیری برآوردهای حاصل از نمونه‌گیری خوشه‌ای دومرحله‌ای است.

در مـورد کسـرهای نمونـه‌گیـری تعییـن شـده بـرای مرحـلهٔ دوم، شـیوه‌هایی را برای تعیین تعداد خوشـه‌هـایی کـه بـاید در مرحلهٔ اول انتخاب شوند تا تضمین شود که خطای معیار برآوردهای حاصل کمـتراز حد تحمل تعیین شده برای آنهاست مورد بحث قرار دادیم. همچنین در مورد تعیین کسر بهینهٔ نمونـه‌گیـری مرحـلهٔ دوم بر مبنای هزینه و ضریب همبستگی درون رده‌ای بحث کردیم. بالاخره، نشان دادیـم کـه نمونـه‌گیری سیستماتیک در واقع، نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده است که طی آن فقط یک خوشه از جامعه انتخاب می‌شود.

تمرین

۱.۱۰ فرض کنید شیکاگو به ۷۵ ناحیه تقسیم شده و هر ناحیه دارای ۲۰ داروخانهٔ خرده‌فروشی است. فرض کنید می‌خواهید متوسط قیمت مطالبه شده برای برخی داروهای تجویز شدهٔ استاندارد را بـرآورد کـنید و در نظر دارید برای این منظور یک نمونهٔ تصادفی ساده متشکل از هشت ناحیه انـتخاب کـنید و پـس از آن، نمونـهٔ تصـادفی سـاده‌ای متشکل از چهار داروخانه از هر یک از هشت ناحیهٔ انتخاب شده بگیرید.

الف. یک تابع هزینـهٔ ساده برای برآورد کـردن کل هزینـــهٔ آمارگیری، شامل هزینه‌های میدانی تهیه کنید. همهٔ مؤلفه‌های هزینه را توصیف کنید.

ب. اگـر ۱۶ ناحـیه و در هر ناحیـه دو داروخانه انتخـاب می‌شد چه تأثیری بر کل هزینهٔبالا داشت؟

پ. فرض کنید حـدس بر این است که انحراف معیـار σ در همـهٔ فروشگاههای شیکاگو از لحاظ قیمت یک داروی خاص ۱/۵۰ دلار و هزینهٔ متوسط این دارو ۱۰/۰۰ دلار است. با استفاده از تابع هزینه که در بخش (الف) تعیین کرده‌اید، مقدار بهینهٔ $\overline{n}$ را در صورتی که ضریب همبستگی درون رده‌ای برابر با ۰/۳۵ باشد تعیین کنید. در این حالت اگر از شش واحد نمونه‌گیری اولیه (PSU) استفاده می‌شد ضریب تغییرات برآورد چقدر می‌بود؟

۲.۱۰ الف. فرض کنید مدارس ابتدایی یک شهر در ۳۰ ناحیهٔ آموزشی گروه‌بندی شده‌اند و هر ناحیهٔ آموزشی دارای ده مدرسه است. فرض کنید یک نمونهٔ تصادفی ساده متشکل از سه ناحیهٔ

آموزشی گرفته شده و در داخل هر ناحیهٔ آموزشی نمونه، به منظور برآورد تعداد دانش‌آموزان شهر که (براساس سنجش با یک آزمون استاندارد) دچار کوررنگی هستند یک نمونهٔ تصادفی ساده متشکل از چهار مدرسه انتخاب شده است. داده‌های ارائه شده در جدول زیر از این نمونه به دست آمده است. کل تعداد دانش‌آموزان دچار کوررنگی را در سراسر شهر برآورد کنید و بازهٔ اطمینان ۹۵ درصدی برای آن به دست آورید.

ناحیهٔ آموزشی نمونه	مدرسه	تعداد کودکان	تعداد کودکان کوررنگ
۱	۱	۱۳۰	۲
	۲	۱۵۰	۳
	۳	۱۶۰	۳
	٤	۱۲۰	۵
۲	۱	۱۱۰	۲
	۲	۱۲۰	٤
	۳	۱۰۰	۰
	٤	۱۲۰	۱
۳	۱	۸۹	٤
	۲	۱۳۰	۲
	۳	۱۰۰	۰
	٤	۱۵۰	۲

ب. نسبت دانش‌آموزانی را که در سراسر شهر دچار کوررنگی‌اند برآورد کنید و بازهٔ اطمینان ۹۵ درصدی برای آن به دست آورید.

۳.۱۰ یک نمونهٔ تصادفی ساده به ٤۰۰ واحد فهرست‌برداری نیاز دارد تا برآوردهای میانگینها و مجموعها بتوانند ویژگیهای دقت مورد نیاز را تأمین کنند. اگر $\overline{N}$، تعداد واحدهای فهرست‌برداری در همهٔ واحدهای نمونه‌گیری اولیه (PSU ها) یکسان باشد، اندازهٔ موردنیاز برای یک نمونهٔ خوشه‌ای ساده که اندازهٔ خوشهٔ $\overline{n}$ آن برابر با ٤ است چقدر باید باشد تا همان دقتی را که با ضریب همبستگی درون رده‌ای برابر با ۰/٦ به دست می‌آمد تأمین کند؟

٤.۱۰ فرض کنید که در شلوغترین فصل، تعداد گردشگران یک پارک ایالتی و تعداد رویدادهای مصدومیت در میان این گردشگران برحسب هفته و روز به شرح مندرج در جدول زیر است (این پارک جمعه‌ها تعطیل است).

فـرض کنید یک نمونهٔ خوشه‌ای دومرحله‌ای ساده گرفته شده است که هفته‌ها به عنوان خوشـه‌ها، روزها به عنوان واحدهای فهرست‌برداری، $m = ۴$ و $\overline{n} = ۳$ در نظر گرفته شده‌اند. باز هم فرض کنید که در مرحلهٔ اول نمونه‌گیری، خوشه‌های ۲، ٦، ۸ و ۱۰ انتخاب شده‌اند. در مرحـلهٔ دوم نمونـه‌گیـری فـرض مـی‌کنیم که واحدهای فهرست‌برداری ۲، ۳ و ۵ از خوشهٔ ۲؛ واحـدهـای فهـرسـت‌بـرداری ۱، ۳ و ٦ ازخوشـهٔ ٦؛ واحدهـای فهرسـت‌بـرداری ۳، ٤ و ٦ از خوشـــهٔ ۸؛ و واحدهـای فهرسـت‌بـرداری ۲، ٤ و ۵ از خوشهٔ ۱۰ انتخاب شده‌اند. از روی این نمونه، موارد زیر را برآورد کنید و بازهٔ اطمینان ۹۵ درصدی برای هر یک به دست آورید:

هفته	تعداد گردشگران							تعداد مصدومیتها						
	یکشنبه	دوشنبه	سه‌شنبه	چهارشنبه	پنجشنبه	شنبه		یکشنبه	دوشنبه	سه‌شنبه	چهارشنبه	پنجشنبه	شنبه	
۱	۲۰۰	۱۵۰	۱۳۰	۱٤۰	۱۵۰	۱۹۰		۲	۳	۱	٤	۳	۸	
۲	۱۲۰	۱۰۵	۱۱۱	۱۰۳	۱۱۱	۱۳۰		۱	۰	۰	۱	۰	۳	
۳	۳۱۰	۲۰۰	۱۸۰	۱۳۰	۱۲۵	۲۰۸		٤	۰	۱	۰	۱	۳	
٤	۲۰۰	۱۰۷	۱۰۱	۹۵	۱۰۳	۱۳۷		۳	۰	۲	۰	۱	۸	
۵	۱۷۰	۱٦۰	۱۳۰	۱۲۱	۱۰۷	۱۱٤		۳	۰	۰	۱	۰	۵	
٦	۲۵۰	۲۳۷	۲۰۹	۲۱۲	۲۳۱	۱۸۰		۲	۰	۰	۰	۰	۱	
۷	۳۸۰	۳۷۸	۳۲۵	۳۳۰	۳۰٦	۳۳۱		٤	۳	۰	۸	۰	۲	
۸	٤۹۵	٤۰۰	۳۱۵	۳۰۲	۳۵۰	۳۹۵		٤	۰	۳	۲	۲	٤	
۹	۲۰٦	۲۰۰	۱۰۸	۹۵	۱۰۷	۱۹۰		۱	۲	۳	۰	۱	٤	
۱۰	۳۰۸	۳۰۰	۲۹۳	۲۰٦	۲۰۰	۳۰۰		۰	۰	۱	۲	۰	۳	

الف. کل تعداد گردشگران طی شلوغترین فصل

ب. کل تعداد مصدومیتها در شلوغترین فصل

پ. کل تعداد مصدومیتها و گردشگران در هفته

ت. کل تعداد مصدومیتها و گردشگران در روز

ث. تعداد مصدومیتها به ازای هر گردشگر

۵.۱۰ فرض کنید یک نمونهٔ خوشه‌ای یک‌مرحـله‌ای سـاده از جامعـهٔ نشان داده شده در تمرین ٤.۱۰ گـرفتـه شده و خوشـه‌های ۲ و ۸ انتخـاب شـده‌اند. از روی این نمـونه، بازه‌های اطمینـان ۹۵ درصـــدی را برای مشخـــصه‌های ارائه شـده در بخشهای (الف) تا (ث) ی تـــمرین ٤.۱۰ محاسبه کنید.

٦.۱۰ از روی جامعهٔ نشـان داده شـده در تمرین ٤.۱۰، ضریب همبستگی درون رده‌ای را برای تعداد گردشگران پارک محاسبه کنید.

۷.۱۰ با استفاده از ضریب همبستگی درون رده‌ای محاسبه شده در تمرین ٦.۱۰ و با فرض $C_1^* = 2C_2^*$، تعداد بهینهٔ روزها برای نمونه‌گیری در یک نمونهٔ خوشه‌ای دومرحله‌ای ساده که از هفته‌ها به عنوان خوشه‌ها استفاده می‌شود چقدر باید باشد؟

۸.۱۰ سازندهٔ یک وسیلهٔ اورتوپدی مایل است به یک برنامهٔ کنترل کیفیت اقدام کند که در آن نمونه‌ای از این وسیله‌ها نمونه‌گیری شوند و برای عیب‌یابی مورد آزمون قرار گیرند. این وسیله‌ها، پس از خروج از کارگاه در بسته‌هایی حاوی ۲۰ جعبه دسته‌بندی می‌شوند که در هر جعبه ۱۰ دستگاه از این وسیله‌ها قرار دارند. خواستهٔ سازنده، نمونه‌گیری از هر بسته و کنار گذاشتن آن در صورتی است که نسبت برآورد شده برای وسیله‌های معیوب در داخل هر بسته بیشتر از ۵٪ باشد. قبلاً یک بررسی مقدماتی انجام شده است و طی آن همهٔ دستگاههای موجود در یک بسته مورد آزمون قرار گرفته‌اند. نتایج این بررسی مقدماتی ذیلاً ارائه شده است :

جعبه	نسبت وسیله‌های معیوب
۱	۰/۵۰
۲	۰/۰۰
۳	۰/۰۰
٤	۰/٤۰
٥	۰/۰۰
٦	۰/۰۰
۷	۰/۰۰
۸	۰/۰۰
۹	۰/۰۰
۱۰	۰/۰۰
۱۱	۰/۰۰
۱۲	۰/۱۰
۱۳	۰/۱۰
۱٤	۰/۰۰
۱٥	۰/۰۰
۱٦	۰/۰۰
۱۷	۰/۰۰
۱۸	۰/۱۰
۱۹	۰/۰۰
۲۰	۰/۰۰

فـرض کـنید قـرار است از یک برنامهٔ نمونه‌گیری خوشه‌ای دومرحله‌ای استفاده شود که در آن جعـبه‌هـا بـه عـنوان خوشـه‌هـا مـورد اسـتفاده قرار می‌گیرند و از هر جعبه پنج دستگاه از این وسـیله‌هـا انـتخاب می‌شوند. براساس داده‌های بالا، چند جعبه باید برای نمونه انتخاب شود تا ویژگیهای فوق‌الذکر با ۹۵٪ اطمینان تأمین شوند؟

۹.۱۰ اگر در تمـرین ۸.۱۰، هزینـهٔ مربـوط به آزمـون دستـگاهها ۱۵ برابر هزینـهٔ فهرست‌برداری و نمونه‌گیری از جعبه‌ها باشد، در هر جعبهٔ انتخاب شده چند دستگاه باید نمونه‌گیری شود؟

۱۰.۱۰ ضریب همبستگی درون رده‌ای در داده‌های ارائه شده در تمرین۸.۱۰ زیاد است یا کـم؟ برای پاسخ خود دلیل بیاورید و درستی آن را با محاسبهٔ ضریب همبستگی درون رده‌ای تحقیق کنید.

۱۱.۱۰ مطالعه‌ای آغاز شد که هدف اصلی آن برآورد تعـداد بیـمارانی بود که دچـار بیمـاری روانـی شـدید بودنـد و در سال تقویمی ۱۹۷٤ در بیمارستانهایی مورد معاینه قرار گرفته بودند که به عـنوان *مراکـز روانـی* تعییـن نشده بودند، در حالی که در بخشهای شهری ایلی‌نوی بیرون از مـنطقهٔ شیکاگو قرار داشتند که هم دارای مراکز روانی (T) بودند و هم بیمارستانهایی داشتند کـه بـه عنوان مراکز روانی تعیین نشده بودند (NT). داده‌های زیر در اختیار پژوهشگران قرار داده شده بود:

بخش	بیمارستان	تعداد تخت	نوع
۱	۱	۳۱۰	**NT**
	۲	۲۲۹	**T**
	۳	۳٦۷	**NT**
۲	۱	۱۳٤	**T**
	۲	۱۹۸	**NT**
۳	۱	۲٤۲	**T**
	۲	۳۰۰	**NT**
٤	۱	۳٥۸	**T**
	۲	٤۱۰	**NT**
٥	۱	۳۲	**NT**
	۲	۱٥٦	**T**
٦	۱	۲۳۱	**NT**
	۲	۲۰۹	**T**
	۳	٤٤	**NT**

بخش	بیمارستان	تعداد تخت	نوع
۷	۱	۲۲۷	NT
	۲	۱۷۸	T
	۳	٦۱	NT
	٤	٥۹	NT
	٥	۲۲۳	NT
۸	۱	۱٦	NT
	۲	۲٦۳	T
	۳	۲۹٥	NT
۹	۱	۱۸۰	T
	۲	۱٥۲	NT
	۳	۲٥٦	NT
۱۰	۱	۱۰۰	T
	۲	٦٥	NT
۱۱	۱	٥۹٥	NT
	۲	٦٤۸	T
۱۲	۱	۷٦	NT
	۲	۱۳۳	T
	۳	۱۱۷	NT
	٤	۱۱۷	NT
	٥	۲٥٤	NT
۱۳	۱	٥۷٤	NT
	۲	۸۲٤	T
	۳	۳۰٤	NT
۱٤	۱	۳٥۰	NT
	۲	۲٥٦	T
	۳	۲۷٥	NT
	٤	۱٥۰	NT
۱٥	۱	۱۳۳	NT
	۲	۳۱٤	T
	۳	۱۲٤	T
	٤	۱۸۸	NT
	٥	۲۱۲	NT
	٦	۱٤۳	NT
۱٦	۱	٦۰	NT
	۲	٥٥	T

بخش	بیمارستان	تعداد تخت	نوع
۱۷	۱	۱۵۰	T
	۲	۵۰	NT
۱۸	۱	۷۲	NT
	۲	۸۰	T
	۳	۳۸	NT
۱۹	۱	٦٤	NT
	۲	۱۲۵	T
۲۰	۱	۳٦۷	NT
	۲	۳۲۹	T
	۳	۳۱۲	T
	٤	۱۷۸	NT
	۵	۲۸۱	NT

یک طرح نمونه‌گیری خوشه‌ای دومرحله‌ای انتخاب شد که در آن بخشها به عنوان واحدهای نمونه‌گیری اولیه و بیمارستانهای فاقد مرکز روانی به عنوان واحدهای فهرست‌برداری به کار رفتند. شش بخش در مرحلهٔ اول انتخاب شدند و در مرحلهٔ دوم از هر بخش نمونه، یک بیمارستان فاقد مرکز روانی نمونه‌گیری شد. داده‌های زیر از این آمارگیری نمونه‌ای به دست آمدند.

بخش	بیمارستان	موارد بیماری روانی شدید
۱	۱	۱٤
۷	۵	۱۰
۱٤	۱	٦
۱۵	٦	۰
۱۹	۱	۱
۲۰	۱	۲٤

از یک برآورد نسبتی براساس تعداد تختها استفاده کنید و از روی نمونهٔ بالا، کل تعداد موارد بیماری روانی شدید را برای بیمارستانهای این ۲۰ بخش که به عنوان مراکز روانی تعیین نشده‌اند برآورد کنید.

۱۲.۱۰ در شهری که دارای ۲۵ مرکز روانپزشکی است یک آمارگیری از سوابق بیماران برنامه‌ریزی شده است. هدف از این آمارگیری، برآورد کل تعداد بیمارانی است که به عنوان قسمتی از رژیم درمانی، به آنها دیازپام (والیوم) داده می‌شود. تعداد بیماران تحت درمان در هر یک از

مراکـز روانپزشـکی در جدول زیر فهرست شده است. قرار است نمونه‌ای از بیماران با انتخاب نمونـهٔ تصـادفی سـاده‌ای از مراکـز روانپزشـکی گرفـته شود و در هر مرکز روانپزشکی نمونه، زیرنمونه‌ای از بیماران انتخاب شود.

مرکز درمانی	تعداد بیماران	مرکز درمانی	تعداد بیماران
۱	٤۹۱	۱٤	٦۷۲
۲	۸٦٦	۱۵	٤۷۵
۳	۱۸۸	۱٦	٤۳۹
٤	۹۹٤	۱۷	۳۹۲
۵	۲۰۹	۱۸	۵۸٤
٦	۹٦۱	۱۹	۸۸۲
۷	۸۳٤	۲۰	٤۲٤
۸	۹۸۲۰	۲۱	۷۷۵
۹	۳٤۸	۲۲	۲٦۲
۱۰	۲٤٦	۲۳	۹٦۸
۱۱	۳۹۹	۲٤	۵۸٦
۱۲	۱۷۵	۲۵	۸۰۹
۱۳	۱٦٦		

یـک آمارگیـری از ۱۰٪ همـهٔ بیماران در مراکز درمانی ۱، ۳، ۸، ۱۲ و ۱۵ که سال گذشته اجرا شد داده‌های زیر را به دست داده است :

مرکز درمانی	تعداد بیماران	تعداد دریافت‌کنندگان دیازپام
۱	٤٦	۱٤
۳	۱۳	۵
۸	۹٤۲	۳٤۰
۱۲	۱۵	۱
۱۵	٤۲	۲۰

الف. براسـاس داده‌های مربوط به این پنج مرکز درمانی، ضریب همبستگی درون رده‌ای را با توجه به تعداد بیمارانی که دیازپام دریافت می‌کنند محاسبه کنید.

ب. براساس ضریب همبستگی درون رده‌ای بالا، متوسط بهینهٔ اندازهٔ خوشه، $\overline{n}$، برای آمارگیری پیشنهادی از ۲۵ مرکز درمانی که در بالا فهرست شد، اگر مؤلفهٔ هزینهٔ مربوط به خوشه‌ها هزار برابر مؤلفهٔ هزینهٔ واحدهای فهرست‌برداری باشد، چقدر است؟

۱۳.۱۰ براساس اندازهٔ متوسط خوشه‌ای محاسبه شدهٔ بالا، اگر بخواهیم ۹۵٪ مطمئن باشیم که تعداد بیمارانی که دیازپام دریافت می‌کنند در محدودهٔ ۱۰٪ مقدار واقعی برآورد می‌شوند چند مرکز درمانی باید نمونه‌گیری شوند؟

۱٤.۱۰ براساس اندازهٔ خوشه‌ای $\overline{n}$ و m، تعداد خوشه‌های نمونه که در تمرینهای ۱۲.۱۰ و ۱۳.۱۰ به دست آمدند خطای معیار برآورد کل تعداد بیمارانی که دیازپام دریافت می‌کنند چقدر خواهد بود؟

۱۵.۱۰ قرار است یک آمارگیری خانوار به منظور برآورد کردن برخی متغیرها در مورد وضعیت بهداشتی و بهره‌مندی از امکانات بهداشتی اجرا شود. آزمایشگاه تحقیقات آمارگیری که برای اجرای این بررسی طرف قرارداد است، به فهرستهای خانوارها در دفتر سرشماری امریکا دسترسی دارد و می‌تواند خوشه‌هایی با ۱۸ خانوار را تعریف کند که می‌توان نمونه‌ای از خانوارها را از آنها انتخاب کرد. از بررسی دیگری که در مورد فهرستهای مشابه اجرا شده بود ضریبهای همبستگی درون رده‌ای برآورد شده بودند (جدول زیر را ببینید). در بررسی مزبور برآورد شده بود که مؤلفهٔ هزینهٔ مربوط به خوشه‌ها تقریباً یک چهارم هزینهٔ مربوط به واحدهای فهرست‌برداری است. براساس این اطلاعات، یکی از سه نوع متفاوت خوشه‌ها را انتخاب کنید و اندازهٔ مناسب برای خوشهٔ نمونه را تعیین کنید.

	همبستگی درون رده‌ای (δ_x) برای اندازه‌های گوناگون خوشه		
متغیر	$N_i = ٦$	$N_i = ٩$	$N_i = ١٨$
تعداد روز ـ تخت در دو هفتهٔ آخر	۰/۰۲۲	۰/۰۳۸	۰/۰۱۱
تعداد مرخص‌شدگان از بیمارستان در ۱۲ ماه گذشته	۰/۰۵۷	۰/۰٦۹	۰/۰۷۷

۱٦.۱۰ فهرستی از بیمارستانهای موجود در یک ناحیهٔ جغرافیایی روستایی در جدول زیر برحسب بخش نشان داده شده است. یک آمارگیری نمونه‌ای با استفاده از یک طرح نمونه برنامه‌ریزی شده است که در آن بخشها به عنوان خوشه‌ها و بیمارستانها به عنوان واحدهای فهرست‌برداری تلقی می‌شوند، و یک بیمارستان قرار است در هر بخش انتخاب شود. اگر فرض کنیم که کل هزینه‌ها به ازای هر بیمارستان متناسب با تعداد پذیرشهاست، چند بخش باید انتخاب شوند تا ۹۵٪ مطمئن باشیم که کل هزینه‌های روزانه را در بیمارستانهای آن منطقه در محدودهٔ ۲۰٪ مقدار واقعی برآورد می‌کنیم؟

۱۷.۱۰ فرض کنید یک نمونهٔ تصادفی ساده متشکل از پنج خوشه از جامعهٔ نشان داده شده در تمرین ۱٦.۱۰ انتخاب شده و خوشه‌های انتخاب شده، شماره‌های ۵، ۸، ۲۳، ۳۰ و ۳٦ هستند. از هر خوشهٔ نمونه یک نمونهٔ مرحلهٔ دوم متشکل از ۱ واحد فهرست‌برداری بگیرید و از روی این نمونه، میانگین پذیرشها را به ازای هر تخت بیمارستانی برآورد کنید. (در صورت دسترسی، از نرم‌افزار آماری استفاده کنید.)

بخش	بیمارستان	تعداد تخت	متوسط تعداد پذیرشها در روز برای سال ۱۹۸۹
۱	۱	۷۲	٤/۸
۲	۱	۸۷	۸/٤
	۲	۱۰٤	۹/٤
	۳	۳٤	۲/۰
۳	۱	۹۹	۵/۱
٤	۱	٤۸	٤/٤
۵	۱	۹۹	٦/۲
٦	۱	۱۳۱	۹/۱
	۲	۱۸۲	۱۵/۹
۷	۱	٤۲	۲/٤
۸	۱	۳۸	۲/۸
۹	۱	۳٤	۲/۳
۱۰	۱	٤۲	٤/۹
۱۱	۱	۳۹	٤/۰
	۲	۵۹	٤/۱
۱۲	۱	۷٦	۵/۲

بخش	بیمارستان	تعداد تخت	متوسط تعداد پذیرشها در روز برای سال ۱۹۸۹
۱۳	۱	۲۵	۳/۱
	۲	۸۰	۵/۳
۱٤	۱	۵۰	٤/۹
۱۵	۱	۸۸	۷/۱
۱٦	۱	۵۰	٤/٤
۱۷	۱	٦۳	۵/۱
۱۸	۱	٤۵	۳/۹
۱۹	۱	۷۵	۸/۵
۲۰	۱	۱۷	۳/۸
	۲	۱٤۰	۱۱/۹
۲۱	۱	٤٤	٤/۹
۲۲	۱	۱۷۱	۱۲/۰
	۲	۸۵	٤/٦
۲۳	۱	٤۸	۳/۸
	۲	۱۸	۲/۹
۲٤	۱	۵٤	٤/۹
۲۵	۱	٦۸	۳/۸
	۲	٦۸	۵/۵
۲٦	۱	٤٤	۳/۵
۲۷	۱	۳۲	۱/۰
	۲	۹۰	٦/۱
۲۸	۱	۳۵	۲/۹
۲۹	۱	۷۲	۵/۲
۳۰	۱	۱۰٤	٦/٦
۳۱	۱	۸٦	٦/٤
	۲	۹۱	٦/٤
	۳	۵۳	٤/۵
۳۲	۱	۱۰۸	٦/٤
۳۳	۱	۵۰	٤/۹
۳٤	۱	٤۵	۳/۸
۳۵	۱	٦۵	٤/۳
۳٦	۱	٤۸	٤/۹
۳۷	۱	٦۱	۵/۷

۱۸.۱۰ از روی نمونهٔ انتخاب شده در تمرین ۱۷.۱۰، کل تعداد پذیرشها را در ۳۷ بخش ناحیه برای سال ۱۹۸۹ برآورد کنید. برآورد خطای معیار این برآورد چقدر است؟ (در صورت دسترسی، از نرم‌افزار آماری مناسب استفاده کنید.)

۱۹.۱۰ در شهر متوسطی که دارای ۳۰ پمپ بنزین بود یک نمونهٔ تصادفی ساده متشکل از پنج پمپ بنزین گرفته شد. در هر یک از این پنج پمپ بنزین نمونه، یک واحد سیار با تجهیزات جمع‌آوری نمونه‌های ادرار و آب دهان مستقر شد و نمونه‌ای متشکل از یکی از هر ۱۰ اتومبیلی که وارد پمپ بنزین می‌شدند انتخاب شد. از رانندهٔ هر اتومبیل نمونه درخواست می‌شد تا در آمارگیری شرکت کند. آمارگیری شامل مصاحبهٔ کوتاه و ارائهٔ نمونه‌ای از ادرار و آب دهان بود. هر شرکت کننده به عنوان تشویق یک باک بنزین مجانی دریافت می‌کرد و به او اطمینان داده می‌شد که اطلاعات به دست آمده کاملاً محرمانه خواهد بود. هدف اصلی از این آمارگیری به دست آوردن برآوردهایی از تعداد رانندگانی بود که ضمن رانندگی مواد مخدر مصرف کرده بودند. داده‌های زیر به دست آمد:

پمپ بنزین	تعداد افراد مورد آزمایش	تعداد افرادی که علامتهای مصرف مواد مخدر را نشان داده بودند
۱	۱۵	۳
۲	٦	۱
۳	۱۱	۱
٤	٦	۳
٥	۸	۱

الف. براساس این داده‌ها، تعداد رانندگانی را که در سطح شهر علامتهایی از مصرف مواد مخدر نشان می‌دهند برآورد کنید

ب. آیا برآورد مورد استفادهٔ بالا نااریب است؟ نشان دهید که نااریب هست یا نیست.

پ. اگر نااریب نیست به چه اطلاعات دیگری نیاز است تا برآوردی نااریب ایجاد شود؟

کتابشناسی

Virtually every textbook on sampling theory or sample survey methodology contains considerable material on two-stage cluster sampling. Recent review or expository articles relevant to this topic appearing in the Encyclopedia of Biostatistics *are listed below:*

1. Shimizu, 1. M. Multistage sampling. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.
2. Brock, D. B., Beckett, L. A., and Bienias, A. L., Sample surveys. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K, 1998.
3. Freeman, D. H., Optimum allocation in cluster sampling. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K, 1998.

Since two-stage cluster sampling is one of the most common designs used in sample surveys, there is an immense number of published studies presenting substantive findings obtained from such studies. A few of these are mentioned below:

4. Goldberg, J., Levy, P. S., Gelfand, H. M., Mullner, R., Iverson, N., Lemeshow, S., and Rothrock, J., Factors affecting trauma center utilization in Illinois. *Medical Care,* 19: 547, 1981.
5. Brenniman, G. R., Kojola, W. H., Levy, P. S., Carnow, B. W., and Namekata, T., High barium levels in public drinking water and its association with elevated blood pressure. *Archives of Environmental Health,* 36: 28-32, 1981.
6. Barr, D., Hershow, R., Levy, P. S., Furner, F., and Handler, A., Assessing prenatal hepatitis B screening in Illinois using an inexpensive study design adaptable to other jurisdictions. *American Journal of Public Health,* January 1999.

The text by Hansen, Hurwitz, and Madow develops the use of ultimate cluster variance estimates.

7. Hansen, M., Hurwitz, W. and Madow W., *Sample Survey Methods and Theory,* Vol. 1, Wiley, New York, 1953.

فصل ۱۱

نمونه‌گیری خوشه‌ای که در آن خوشه‌ها با احتمال نابرابر نمونه‌گیری می‌شوند: نمونه‌گیری با احتمال متناسب با اندازه

روش‌شناسی نمونه‌گیری خوشه‌ای که در هر دو فصل ۹ و ۱۰ شرح و بسط داده شد منحصراً به آن دسته از طرحهای نمونه‌گیری محدود می‌شد که در آن، خوشه‌ها با احتمال برابر نمونه‌گیری می‌شدند. به عبارت دیگر، در بحث ما دربارهٔ نمونه‌گیری خوشه‌ای تا اینجا، همهٔ خوشه‌های موجود در جامعه، مستقل از تعداد واحدهای شمارش آنها یا در واقع مستقل از هر مشخصهٔ دیگری که ممکن است داشته باشند، احتمال یکسان برای انتخاب شدن در نمونه داشته‌اند. در این فصل نشان می‌دهیم که این‌گونه طرحها در مواردی خاص می‌توانند انتخابهای نمونه‌ایی را نتیجه دهند که قابل اجرا نیستند و یا می‌توانند به برآوردهایی خطی بینجامند (مانند میانگینها، مجموعها، نسبتها) که خطاهای معیار آنها بسیار زیاد باشد. وضعیتهایی خاص که مخصوصاً در برابر این قبیل مشکلات از همه آسیب‌پذیرترند آنهایی هستند که بین واحدهای نمونه‌گیری اولیهٔ آنها از لحاظ تعداد واحدهای شمارش تغییرپذیری قابل ملاحظه‌ای وجود دارد. متأسفانه این وضعیتها از جمله مواردی هستند که در عمل بیش از همه روی می‌دهند.

بـرای اجتـناب از ایـن مشـکلات، در این فصل، روشی را برای نمونه‌گیری خوشه‌ای شرح و بسط مـی‌دهیـم کـه در آن همهٔ خوشه‌های جامعه دارای احتمال یکسان برای انتخاب شدن در نمونه نیستند. تمرکـز مـا بـه خصـوص بـر رده‌ای از طـرحهای نمونـه‌گیری خوشه‌ای است که *نمونه‌گیری با احتمال متناسـب بـا اندازه* نامیده می‌شوند (و عموماً به صورت *نمونه‌گیری PPS* خلاصه می‌شوند). در این رده از طـرحها، احـتمال انتخاب شدن یک خوشه در نمونه متناسب با معیاری از اندازهٔ آن است که معمولاً تعـداد واحدهای شمارش موجود در آن است. به راهی دیگر می‌توان از متغیر دیگری که به سطح متغیر مورد برآورد وابسته است استفاده کرد.

در شـرح و بسـط مفـاهیم مـربوط بـه نمونه‌گیری خوشه‌ای با احتمال نابرابر، در برخی موارد برای روشـن شـدن مطلب از مثالهای تشریحی مربوط به نمونه‌گیری خوشه‌ای یک‌مرحله‌ای استفاده خواهیم کـرد، هر چند که می‌دانیم استفاده از آن بسیار کمتر از نمونه‌گیری خوشه‌ای دو و یا حتی چندمرحله‌ای است.

۱.۱۱ انگیزه برای نمونه‌گیری نکردن با احتمال برابر از خوشه‌ها

جامعـهٔ متشـکل از سـه بیمارسـتان را که در جدول ۱.۱۱ نشان داده شده است در نظر می‌گیریم. فرض کنید قرار است یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از دو تا از این بیمارستانها بگیریم تا کل تعـداد جراحیهای غیرضروری بیماران سرپایی را در بیمارستان برآورد کنیم. آمارگیری مستلزم استخراج اطلاعـات از هـر سابقهٔ نمونه‌ای است و اندازه‌های نمونه برای هر یک از سه نمونهٔ ممکن در زیر نشان داده شده‌اند.

بیمارستانها در نمونه	تعداد سوابق بیماران سرپایی نمونه‌گیری شده			
	بیمارستان ۱	بیمارستان ۲	بیمارستان ۳	کل اندازهٔ نمونه
۱ و ۲	۳۰۰۰	٤۰۰۰	—	۷۰۰۰
۱ و ۳	۳۰۰۰	—	۱۰۰۰۰	۱۳۰۰۰
۲ و ۳	—	٤۰۰۰	۱۰۰۰۰	۱٤۰۰۰

جـدول بـالا بـه وضـوح نشان می‌دهد که این طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده به کل انـدازهٔ نمونـه‌ای بسـیار غیرمنـتظره مـنجر می‌شود (این اندازه می‌تواند حداقل ۷۰۰۰ و حداکثر ۱٤۰۰۰ باشـد) و بـار اسـتخراج اطلاعـــات در میان بیمارستـــانهای نمونه‌گیری شـده بسیار نابرابر خواهد بود (در صـورتی کـه بیمارسـتان ۱ بـرای نمونـه انتخاب شود فقط ۳۰۰۰ سابقه باید از آن انتخاب شود در حـالی که از بیمارستان ۳ باید ۱۰۰۰۰ سابقه انتخاب شود). نابرابری در میان بیمارستانها از لحاظ اندازهٔ

نمونـه و غیـر قـابل پیش‌بینی بودن از نظر اندازهٔ کل نمونه می‌تواند قابل اجرا بودن این آمارگیری را به شدت به مخاطره اندازد.

عـلاوه بـر مشـکلات مربوط به قابل اجرا بودن که در بالا بیان شد، در زیر (با استفاده از فرمولهای مـربوط بـه خطـای معیار در فصل ۹) مشاهده می‌کنیم که یک برآورد خطی از قبیل برآورد کل جامعه، تغییرپذیـری قـابل توجهـی خواهـد داشت. توزیع y'_{clu} برآورد مجموع جراحیهای غیرضروری بیماران سرپایی در هر سه نمونهٔ ممکن در پایین نشان داده شده است.

جدول ۱.۱۱ تعداد جراحیهای انجام شده برای بیماران سرپایی در سال ۱۹۹۷ در سه بیمارستان

بیمارستان	جراحیهای بیماران سرپایی	تعداد جراحیهای غیرضروری بیماران سرپایی
۱	۳۰۰۰	۳۵
۲	٤۰۰۰	۳۸
۳	۱۰۰۰۰	۱۰۰
مجموع	۱۷۰۰۰	۱۷۳

بیمارستانها در نمونه	مجموع نمونه y	برآورد مجموع جامعه y'_{clu}
۲ , ۱	۷۳	۱۰۹/۵
۳ , ۱	۱۳۵	۲۰۲/۵
۳ , ۲	۱۳۸	۲۰۷

مقـدار مـورد انـتظار و خطـای معیار y'_{clu} برآورد مجموع همراه با ضریب تغییرات آن در زیر ارائه شده است:

$$E(y'_{clu}) = ۱۷۳$$
$$SE(y'_{clu}) = ۵۵/۰۴$$
$$V(y'_{clu}) = ۰/۳۲$$

توجـه کنید که این برآورد، وقتی در نظر بگیریم که دو سوم خوشه‌های موجود در جامعه در نمونه قرار گرفـته‌انـد ضـریب تغییـرات زیـادی دارد. عـلت آن اسـت کـه متغیر موردنظر، یعنی تعداد جراحیهای غیرضـروری بیمـاران سـرپایی، عمدتـاً بـه تعـداد جـراحیهای بیمـاران سـرپایی وابسته است و در بین بیمارسـتـــانها از لحـاظ کل تعداد جراحیـــهای بیماران سرپایی تغییرپذیری قابل توجـــهی وجود دارد.

در واقع، بیمارستان ۳ تعداد بیشتری جراحیهای سرپایی داشته و از این رو جراحیهای غیرضروری بیماران سرپایی نیز در این بیمارستان بیش از مجموع بیمارستانهای ۱ و ۲ است. به این ترتیب، نمونه‌ای که بیمارستان شمارهٔ ۳ در آن نباشد تعداد جراحیهای غیرضروری را به صورتی قابل ملاحظه کم برآورد خواهد کرد در حالی که دو نمونه‌ای که شامل بیمارستان شمارهٔ ۳ هستند این مجموع کل را بیش ـ برآورد خواهند کرد. در نمونه‌گیری خوشه‌ای ساده، به طوری که در فصل ۹ و ۱۰ توضیح داده شد، هر بیمارستان صرفنظر از هر اطلاعاتی معلوم دربارهٔ خوشه، شانس یکسانی برای ورود به آمارگیری دارد. همچنین، شیوه‌های برآورد مورد استفاده در نمونه‌گیری خوشه‌ای ساده هیچ استفاده‌ای از اطلاعات مربوط به مشخصه‌های خوشه به عمل نمی‌آورد.

ما نشان خواهیم داد که امکان کاهش خطای نمونه‌گیری یک برآورد مجموع، حاصل از یک نمونهٔ خوشه‌ای، با استفاده از اطلاعات معلوم دربارهٔ خوشه‌ای که ممکن است با متغیر پاسخ همبستگی داشته باشد وجود دارد. این اطلاعات را می‌توان هم در شیوهٔ نمونه‌گیری و هم در فرایند برآورد کردن مورد استفاده قرار داد.

فرض کنیم به جای نمونه‌گیری از خوشه‌ها با احتمال برابر طوری نمونه بگیریم که احتمال قرار گرفتن هر بیمارستان در نمونه، متناسب با تعداد جراحیهای بیماران سرپایی باشد. به طور دقیقتر، شیوه‌ای را برای نمونه‌گیری توصیف می‌کنیم که در آن نمونه‌گیری بدون جایگذاری است و بیمارستانها یک به یک از یک «آوندی» فرضی قرعه‌کشی می‌شوند. فرض کنید که احتمال انتخاب شدن یک بیمارستان در هر نوبت قرعه‌کشی متناسب با تعداد جراحیهای بیماران سرپایی است و هر بار که در قرعه‌کشی انتخاب شد از آوند خارج می‌شود. اگر X_i را برابر با تعداد جراحیهای بیماران سرپایی در بیمارستان i ام فرض کنیم می‌بینیم که P_i، احتمال انتخاب شدن بیمارستان i در اولین قرعه‌کشی از فرمول زیر به دست می‌آید.

$$P_i = \frac{X_i}{X}$$

که در آن

X_i = تعداد جراحیهای بیماران سرپایی در بیمارستان i

X = کل تعداد جراحیهای بیماران سرپایی در هر سه بیمارستان

از مطالب بالا و از نظریهٔ ترکیبیاتی نتیجه می‌گیریم که اگر دو بیمارستان بدون جایگذاری نمونه‌گیری شوند، در آن صورت π_{ij}، احتمال اینکه بیمارستانهای i و j در نمونه انتخاب شوند از فرمول زیر به دست می‌آید.

$$\pi_{ij} = \frac{X_1 X_2}{X}\left(\frac{1}{X - X_1} + \frac{1}{X - X_2}\right) \qquad (1.11)$$

پس برای داده‌های جدول ۱.۱۱ احتمالهای زیر را برای انتخاب شدن داریم:

بیمارستانها در نمونه (j, i)	احتمال ظاهر شدن بیمارستانهای i و j با هم در نمونه [از روی رابطهٔ (۱.۱۱)] π_{ij}
(۲ , ۱)	۰/۱۰۴۷۲
(۳ , ۱)	۰/۳۷۸۱۵
(۳ , ۲)	۰/۵۱۷۱۳

مثلاً از رابطهٔ (۱.۱۱)

$$\pi_{12} = \frac{3000 \times 4000}{17000} \times \left[\frac{1}{17000 - 3000} + \frac{1}{17000 - 4000}\right] = 0/1047$$

به محض این که π_{ij} به دست آمد می‌توانیم احتمال π_i را که بیمارستان i در نمونه قرار بگیرد به دست آوریم:

$$\pi_i = \sum_j \pi_{ij}$$

برای هر یک از سه بیمارستان موجود در جامعه، π_i، احتمال اینکه در نمونه ظاهر شود در زیر نشان داده می‌شود:

بیمارستان	π_i، احتمال قرار گرفتن در نمونه
۱	۰/۴۸۲۸۷
۲	۰/۶۲۱۸۵
۳	۰/۸۹۵۲۸

حالا، براساس برنامهٔ نمونه‌گیری بالا، برآوردگر y'_{hte} را با استفاده از فرمول زیر بیان می‌کنیم

$$y'_{hte} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

نماد «hte » از حروف آغازین Horvitz-Thompson estimator گرفته شده که رده‌ای از برآوردگرهاست که در بخش بعدی به صورتی کلیتر مورد بحث قرار خواهد گرفت.

این برآوردگر برخلاف برآوردگر x'_{clu}، از اطلاعات مربوط به خوشه، هم در نمونه‌گیری و هم در ساختن برآوردگر استفاده می‌کند. شانس انتخاب شدن هر بیمارستان در نمونه برای آن بیمارستانهایی که تعداد بیشتری جراحی سرپایی داشته‌اند بیشتر است زیرا فرض بر این است که به متغیر مورد نظر ـ یعنی تعداد جراحیهای غیرضروری بیماران سرپایی ـ وابسته است. در این شیوهٔ برآورد، به هر بیمارستان برحسب عکس احتمال انتخاب شدن خود در نمونه وزن داده می‌شود تا برآوردگر حاصل نااریب باشد.

توزیع y'_{hte} برای همهٔ نمونه‌های ممکن در زیر نشان داده شده است:

بیمارستانها در نمونه	y'_{hte}	احتمال انتخاب شدن نمونه، π_{ij}
(۲ , ۱)	۱۳۳/۵۹	۰/۱۰۴۷۲
(۳ , ۱)	۱۸۴/۱۸	۰/۳۷۸۱۵
(۳ , ۲)	۱۷۲/۸۰	۰/۵۱۷۱۳

مقدار مورد انتظار y'_{hte} همراه با خطای معیار و ضریب تغییرات آن در زیر نشان داده شده است:

$$E(y'_{hte}) = \sum_{\text{همهٔ نمونه‌ها}} y'_{hte}\pi_{ij} = ۱۳۳/۵۹ \times ۰/۱۰۴۷۲ + ۱۸۴/۱۸ \times ۰/۳۷۸۱۵ + ۱۷۲/۸۰ \times ۰/۵۱۷۱ = ۱۷۳ = Y$$

$$SE(y'_{hte}) = \left(\sum_{\text{همهٔ نمونه‌ها}} (y'_{hte} - ۱۷۳)^۲ \pi_{ij} \right)^{۱/۲} = ۱۴/۴۹$$

$$V(y'_{hte}) = \frac{۱۴/۴۹}{۱۷۳} = ۰/۰۸۴$$

باید توجه داشت که شیوهٔ نمونه‌گیری مبتنی بر انتخاب خوشه‌ها با احتمال نابرابر در ترکیب با استفاده از برآوردگر y'_{hte} در این مثال، ضریب تغییرات ۸/٤٪ را به دست می‌دهد که به مراتب کمتر از ضریب تغییرات ۳۲٪ است که از نمونه‌گیری خوشه‌ها با احتمال برابر به دست می‌آمد. هر دو شیوه، برآوردگرهایی نااریب از کل جامعه تولید می‌کنند.

در مثال بالا نشان دادیم که به وسیلهٔ نمونه‌گیری از خوشه‌ها با احتمال نابرابر، این امکان وجود دارد که برآوردگری به دست آید که خطای معیار آن به صورتی قابل ملاحظه کمتر از خطای معیار به دست آمده از نمونه‌گیری خوشه‌ها با احتمال برابر باشد. در بخشهای بعدی این فصل نشان می‌دهیم که چگونه می‌توان از این قبیل روشها در نمونه‌گیری خوشه‌ای دومرحله‌ای استفاده کرد تا برآوردهایی به

دست آید که معتبر و قابل اعتماد و مبتنی بر نمونه‌هایی باشند که تعداد واحدهای شمارش آنها در داخل خوشه‌ها تقریباً یکسان‌اند. همان‌طور که قبلاً بیان شد، به دلایلی که مربوط به بودجه و امکان‌پذیر بودن کار است، غالباً بسیار مهم است که کل اندازهٔ نمونه، قابل پیش‌بینی باشد و نمونه به شکلی هموار میان خوشه‌ها توزیع شده باشد. در بخش بعد برخی روشهای برآورد را شرح و بسط می‌دهیم که در نمونه‌گیری خوشه‌ها با احتمال نابرابر به کار می‌روند.

۲.۱۱ دو ردهٔ کلی از برآوردگرهای معتبر برای طرحهای نمونه‌ای که در آن واحدها با احتمال نابرابر انتخاب می‌شوند

نوشتگان مربوط به نمونه‌گیری با احتمال نابرابر بدواً پیرامون دو ردهٔ کلی از برآوردگرها تکامل یافته‌اند: یعنی *برآوردگر هورویتز ـ تامپسون* که در بخش قبل به آن اشاره شد و برآوردگری قبل از آن که به *برآوردگر هنسن ـ هورویتز* موسوم است. از این دو مورد با جزئیاتی بیشتر در این بخش بحث می‌کنیم.

۱.۲.۱۱ برآوردگر هورویتز ـ تامپسون

برآوردگر هورویتز ـ تامپسون در اصل به عنوان برآوردگری از یک مجموع در سال ۱۹۵۲ پیشنهاد شد که می‌تواند برای هر طرح نمونه‌گیری، با جایگذاری یا بدون جایگذاری، ساخته شود [۸]. شکل آن برای طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای به صورت زیر است

$$y'_{hte} = \sum_{i=1}^{v} \frac{Y_i}{\pi_i}$$

که در آن

Y_i = مجموع برای i امین خوشهٔ نمونه

π_i = احتمال انتخاب شدن i امین خوشه در نمونه

v = تعداد خوشه‌های متمایز نمونه‌گیری شده

واضح است که هرگاه نمونه‌گیری بدون جایگذاری باشد $v = m$، که m کل تعداد خوشه‌های نمونه‌گیری شده است. می‌توان نشان داد که در هر طرح نمونه‌گیری خوشه ای یک‌مرحله‌ای، y'_{hte} برآوردگری نااریب از Y با خطای معیاری است که از فرمول زیر به دست می‌آید:

$$SE(y'_{hte}) = \sqrt{\sum_{i=1}^{M} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{M} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) Y_i Y_j} \qquad (۳.۱۱)$$

کـه در آن M ، تعداد خوشه‌ها در جامعه، و π_{ij} احتمال این است که خوشه‌های i و j هر دو در نمونه قرار بگیرند.

در مورد مثال تشریحی نشان داده شده در بخش قبل، داریم

$$\sum_{i=1}^{M} \frac{1-\pi_i}{\pi_i} Y_i^2 = ۳۳۵۹/۷۱$$

$$\sum_{i=1}^{M} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) Y_i Y_j = -۳۱۴۹/۷۹$$

و

$$SE(y'_{hte}) = \sqrt{۳۳۵۹/۷۱ + (-۳۱۴۹/۷۹)} = ۱۴/۴۹$$

توجه کنید که این با نتایجی که قبلاً با استفاده از فرمولهای تعاریف به دست آمد توافق دارد.

برآوردگـر زیـر، یعـنی $\hat{Var}(y'_{hte})$ ، یـک برآوردگـر نـااریب از واریـانس $Var(y'_{hte})$ بـرای طـرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای است:

$$\hat{Var}(y'_{hte}) = \sum_{i=1}^{v} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{v} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{Y_i Y_j}{\pi_{ij}} \qquad (٤.١١)$$

ایـن برآوردگـر مشکلاتی دارد به این علت که غالباً بی‌ثبات است و می‌تواند منفی باشد. برآوردگرهای دیگری که کمتر مشکل دارند نیز ابداع شده و در جاهای دیگری مورد بحث قرار گرفته‌اند [۲] تا [٤].

اهمیـت برآوردگـر هورویـتز ـ تامپسون در آن است که برآوردگری نااریب است که تا وقتی بتوان π_i، احـتمال در نمونـه قـرار گرفـتن هـر واحـد را ارزیابی کرد می‌تواند در مورد هر طرح نمونه‌گیری محاسـبه شود. این کار حتی برای پیچیده‌ترین طرحهای نمونه‌گیری نیز غالباً امکان‌پذیر است. تعمیم آن بـه هـرگونه طـرح نمونه‌گیری، آن را در شرح و بسط و یکپارچه کردن نظریهٔ نمونه‌گیری، حائز اهمیت فـوق‌العـاده‌ای ساخته است. عیب آن در بسیاری از وضعیتهای عملی آن است که واریانس آن نه تنها به احـتمالهای شـمول π_i، بـلکه بـه احـتمال تـوأم π_{ij}، یعنی احتمال قرار گرفتن هر جفت از واحدها در نمونه، بستگی دارد. در عمل، به خصوص هرگاه نمونه‌گیری بدون جایگذاری باشد، ارزیابی مقادیر π_{ij} غالباً مشکل و پرزحمت است و برآورد کردن خطاهای نمونه‌گیری را مشکل‌ساز می‌کند. در بخش بعد، برآوردگر دیگری را مورد بحث قرار می‌دهیم که استفاده از آن در عمل غالباً آسانتر است.

۲.۲.۱۱ برآوردگر هنسن ـ هورویتز

این برآوردگر که در زیر برای نمونه‌گیری خوشه‌ای یک‌مرحله‌ای نشان داده می‌شود در سال ۱۹۴۳ توسط هنسن و هورویتز [۷] به عنوان برآوردگری برای مجموع، Y، پیشنهاد شد که هرگاه نمونه‌گیری *با جایگذاری* باشد ناااریب است:

$$y'_{hh} = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{\pi'_i} \qquad (۵.۱۱)$$

که در آن π'_i، احتمال انتخاب واحد i ام در هر قرعه‌کشی نمونه است. خطای معیار این برآوردگر از فرمول زیر به دست می‌آید

$$SE(y'_{hh}) = \sqrt{\frac{1}{m}\sum_{i=1}^{M}\left(\frac{Y_i}{\pi'_i} - Y\right)^2 \pi'_i} \qquad (۶.۱۱)$$

و برآوردگری از خطای معیار از فرمول زیر به دست می‌آید

$$\hat{SE}(y'_{hh}) = \sqrt{\frac{\sum_{i=1}^{m}\left(\frac{Y_i}{\pi'_i} - y'_{hh}\right)^2}{m(m-1)}} \qquad (۷.۱۱)$$

باید توجه داشت که عبارت زیر رادیکال در معادلهٔ (۷.۱۱) برآوردگری ناااریب از واریانس برآوردگر برای هرگونه طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای است.

مثال تشریحی: استفاده از این برآوردگر را با مثال زیر که در جدول ۲.۱۱ آمده است نشان می‌دهیم. ناحیه‌ای دارای چهار خانهٔ سالمندان است و قرار است نمونه‌ای از دو تا از این خانه‌ها با جایگذاری گرفته شود تا Y کل تعداد زنان ۹۰ سال به بالا که در سال ۱۹۹۷ در خانه‌های سالمندان پذیرش

جدول ۲.۱۱ تعداد زنان ۹۰ سال به بالا که در ناحیه‌ای، در سال ۱۹۹۷ در خانه‌های سالمندان پذیرش شده‌اند

خانهٔ سالمندان	تعداد تخت (X_i)	تعداد، Y_i، زنان ۹۰ سال به بالا پذیرش شده در سال ۱۹۹۷	π'_i، احتمال انتخاب خانهٔ سالمندان در نمونه در هر استخراج
۱	۳۵	۳	۰/۱۲۷۲۷
۲	۷۵	۷	۰/۲۷۲۷۳
۳	۱۴۰	۲۴	۰/۵۰۹۰۹
۴	۲۵	۱	۰/۰۹۰۹۱
مجموع	۲۷۵	۳۵	۱/۰۰۰۰۰

شده‌اند برآورد شود. برای نشان دادن برآوردگر هنسن ـ هورویتز، فرض می‌کنیم نمونه‌گیری باجایگذاری است و در هر استخراج خانهٔ سالمندان، احتمال این که خانه‌ای خاص انتخاب شود برابر است با تعداد تختهای آن، X_i، تقسیم بر کل تعداد تختها، X، در هر چهار خانهٔ سالمندان موجود در ناحیه. برای هر خانهٔ سالمندان که انتخاب شود، تعداد اشخاص ۹۰ سال به بالا که در سال ۱۹۹۷ پذیرش شده‌اند از روی سوابق پذیرش تعیین خواهد شد.

در این وضعیت، M ، تعداد خانه‌های سالمندان موجود در ناحیه برابر با ٤ است؛ m ، تعداد واحدهای نمونه‌گیری شده برابر است با ۲ و تعداد نمونه‌های ممکن با جایگذاری برابر است با $٤^{۲}=۱٦$. توزیع y'_{hh}، مقادیر برآورد شدهٔ هنسن ـ هورویتز برای مجموع در جدول ۳.۱۱ نشان داده شده است.

از فرمولهای مربوط به تعاریف می‌بینیم که برآوردگر هنسن ـ هورویتز برآوردگری ناااریب از Y است:

$$E[y'_{hh}] = \sum_r y'_{hh}(r)P(r) = ۲۳/۵۷۱۴۳ \times ۰/۰۱۶۱۹۸ + \cdots + ۲۹/۰۷۱۴۳ \times ۰/۰۰۸۲۶۴ = ۳۵ = Y$$

جدول ۳.۱۱ توزیع برآوردگر هنسن ـ هورویتز، y'_{hh}، در تمام نمونه‌های ممکن متشکل از دو خانهٔ سالمندان که با جایگذاری استخراج شده‌اند

نمونهٔ r	خانه‌های سالمندان در نمونه $(i\ ,\ j)$	$P(r)$ ، احتمال انتخاب شدن نمونهٔ r $(\pi'_i \times \pi'_j)$	$y'_{hh}(r)$
۱	(۱ , ۱)	۰/۰۱٦۲۰	۲۳/۵۷۱٤۳
۲	(۲ , ۱)	۰/۰۳٤۷۱	۲٤/٦۱۹۰۵
۳	(۳ , ۱)	۰/۰٦٤۷۹	۳۵/۳۵۷۱٤
٤	(٤ , ۱)	۰/۰۱۱۵۷	۱۷/۲۸۵۷۱
۵	(۱ , ۲)	۰/۰۳٤۷۱	۲٤/٦۱۹۰۵
٦	(۲ , ۲)	۰/۰۷٤۳۸	۲۵/٦٦٦٦۷
۷	(۳ , ۲)	۰/۱۳۸۸٤	۳٦/٤۰٤۷٦
۸	(٤ , ۲)	۰/۰۲٤۷۹	۱۸/۳۳۳۳۳
۹	(۱ , ۳)	۰/۰٦٤۷۹	۳۵/۳۵۷۱٤
۱۰	(۲ , ۳)	۰/۱۳۸۸٤	۳٦/٤۰٤۷٦
۱۱	(۳ , ۳)	۰/۲۵۹۱۷	٤۷/۱٤۲۸٦
۱۲	(٤ , ۳)	۰/۰٤٦۲۸	۲۹/۰۷۱٤۳
۱۳	(۱ , ٤)	۰/۰۱۱۵۷	۱۷/۲۸۵۷۱
۱٤	(۲ , ٤)	۰/۰۲٤۷۹	۱۸/۳۳۳۳۳
۱۵	(۳ , ٤)	۰/۰٤٦۲۸	۲۹/۰۷۱٤۳
۱٦	(٤ , ٤)	۰/۰۰۸۲٦	۱۱/۰۰۰۰۰

و نیز از روی داده‌های جدول ۳.۱۱ می‌توان دید که خطای معیار آن (همان‌طور که از فرمولهای مربوط به تعریف به دست آمد) به صورت عددی برابر است با آنچه که از روی عبارت (٦.۱۱) محاسبه می‌شد. مشخصاً، داریم:

$$SE(y'_{hh}) = ۹/۱۶$$

در این مثال می‌توان از y'_{hte}، برآوردگر هنسن ـ تامپسون نیز استفاده کرد، زیرا برآوردگری نااریب از مجموع جامعه در نمونه‌گیری با جایگذاری و همچنین در نمونه‌گیری بدون جایگذاری است. ساختن این برآوردگر و توزیع نمونه‌گیری آن در جدولهای ٤.۱۱ و ۵.۱۱ نشان داده شده است.

باز هم با به کار بردن فرمولهای مربوط به تعاریف برای داده‌های جدول ۵.۱۱، می‌توان دید که $E(y'_{hte}) = ۳۵ = X$ و $SE(y'_{hte}) = ۱۰/۸۲$، که در این مورد خاص، اندکی بیشتر از خطای معیار برآوردگر هنسن ـ هورویتز است. ولی، خطاهای معیار هر دوی این برآوردگرها به مراتب کمتر از خطای معیار برآوردگر تورمی معمولی است ($x'_{clu} = (M/m)x$). خطای معیار آن برآوردگر ۲۵/٦٤۱۸ است.

□

۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه

در بخشهای قبلی، مشاهده کردیم که استفاده از نمونه‌گیری با احتمال نابرابر در دو مثال، منجر به برآوردهایی از مجموع شد که خطاهای معیار آنها به مراتب کمتر از خطای معیاری بود که از برآورد بر مبنای نمونه‌گیری از خوشه‌ها با احتمال برابر به دست می‌آمد. در هر دوی این مثالها، نمونه‌گیری از خوشه‌ها بر مبنای مشخصه‌ای معلوم از خوشه‌ها انجام می‌پذیرفت که فرض می‌شد با متغیر موردنظر

جدول ٤.۱۱ تعداد زنان ۹۰ سال به بالا که در سال ۱۹۹۷ در خانه‌های سالمندان یک ناحیه پذیرش شده‌اند

خانهٔ سالمندان	تعداد تختها (X_i)	Y_i، تعداد زنان ۹۰ سال به بالا، پذیرش شده در سال ۱۹۹۷	π_i، احتمال انتخاب خانهٔ سالمندان در نمونه
۱	۳۵	۳	۰/۲۳۸۳٤۷
۲	۷۵	۷	۰/٤۷۱۰۷٤
۳	۱٤۰	۲٤	۰/۷۵۹۰۰۸
٤	۲۵	۱	۰/۱۷۳۵۵٤
مجموع	۲۷۵	۳۵	

وابسته است (مثلاً فرض می‌شد که تعداد تختها در خانهٔ سالمندان با تعـــداد زنان ۹۰ سال به بالا که در دورهٔ زمانی مشخصی پذیرش شده بودند مرتبط باشد). به طور کلی، این راهبردی است که به صورتی گسترده در طراحی آمارگیریهای نمونه‌ای مورد استفاده قرار می‌گیرد و ما در این بخش با جزئیاتی بیشتر از آن بحث خواهیم کرد.

جدول ۵.۱۱ توزیع y'_{hte}، برآوردگر هورویتز ـ تامپسون در همهٔ نمونه‌های ممکن از دو خانهٔ سالمندان که با جایگذاری انتخاب شده‌اند

نمونهٔ r	خانه‌های سالمندان در نمونه $(i\,,\,j)$	$P(r)$، احتمال انتخاب شدن نمونهٔ r $(\pi'_i \times \pi'_j)$	$y'_{hte}(r)$
۱	(۱ , ۱)	۰/۰۱٦۲۰	۱۲/۵۸٦٦۹
۲	(۲ , ۱)	۰/۰۳٤۷۱	۲۷/٤٤٦۳۳
۳	(۳ , ۱)	۰/۰٦٤۷۹	٤٤/۲۰٦۸۹
٤	(٤ , ۱)	۰/۰۱۱۵۷	۱۸/۳٤۸۵۹
۵	(۱ , ۲)	۰/۰۳٤۷۱	۲۷/٤٤٦۳۳
٦	(۲ , ۲)	۰/۰۷٤۳۸	۱٤/۸۵۹٦۵
۷	(۳ , ۲)	۰/۱۳۸۸٤	٤٦/٤۷۹۸٦
۸	(٤ , ۲)	۰/۰۲٤۷۹	۲۰/٦۲۱۵۵
۹	(۱ , ۳)	۰/۰٦٤۷۹	٤٤/۲۰٦۸۹
۱۰	(۲ , ۳)	۰/۱۳۸۸٤	٤٦/٤۷۹۸٦
۱۱	(۳ , ۳)	۰/۲۵۹۱۷	۳۱/٦۲۰۲۱
۱۲	(٤ , ۳)	۰/۰٤٦۲۸	۳۷/۳۸۲۱۱
۱۳	(۱ , ٤)	۰/۰۱۱۵۷	۱۸/۳٤۸۵۹
۱٤	(۲ , ٤)	۰/۰۲٤۷۹	۲۰/٦۲۱۵۵
۱۵	(۳ , ٤)	۰/۰٤٦۲۸	۳۷/۳۸۲۱۱
۱٦	(٤ , ٤)	۰/۰۰۸۲٦	۵/۷٦۱۹۰

ما *نمونه‌گیری با احتمال متناسب با اندازه* را (که با نماد *نمونه‌گیری PPS* خلاصه می‌شود) به عنوان یک رده از نمونه‌گیری با احتمال نابرابر تعریف می‌کنیم که در آن احتمال انتخاب شدن یک واحد در نمونه متناسب با سطح متغیری خاص در آن واحد است. مثالهایی از نمونه‌گیری با احتمال متناسب با اندازه در زیر ارائه می‌شود:

- شهری دارای ۳۵ مغازهٔ خواربارفروشی است و قرار است یک آمارگیری نمونه‌ای به منظور برآورد کردن کل مقدار غذای گربه که طی هفتهٔ گذشته به فروش رفته است اجرا شود. طرح

مزبور مستلزم آن است که نمونه‌ای متشکل از ۱۰ مغازه با احتمال متناسب با کل درآمد ناخالص در سال گذشته انتخاب شود.

- یک سازمان بزرگ حفظ بهداشت (HMO) در دو سال گذشته ادعانامه‌هایی برای ۳۲۹ نفر از بیمه‌شدگان سازمان مراقبتهای پزشکی تنظیم کرده است. هدف از این آمارگیری برآورد کردن کل مبلغی است که سازمان مراقبتهای پزشکی به HMO اضافه پرداخت کرده است. ادعانامه‌ها برای بیمه‌شدگان روی رایانه تنظیم شده‌اند. قرار است نمونه‌ای متشکل از ۲۵ بیمه شده گرفته شود به طوری که احتمال انتخاب یک بیمه شده درنمونه متناسب باشد با کل تعداد ادعانامه‌هایی که از جانب بیمه شده تنظیم شده است.
- قرار است یک نمونهٔ خوشه‌ای دومرحله‌ای گرفته شود که در مرحلهٔ اول آن، بلوکهای شهری با احتمال انتخاب شدن هر بلوک در نمونه متناسب با تعداد معلوم خانوارها در بلوک انتخاب خواهند شد. در مرحلهٔ دوم، قرار است از هر بلوک شهری که در مرحلهٔ اول انتخاب شده است نمونه ای متشکل از ۱۰ خانوار گرفته شود.

در بحثی که متعاقباً دربارهٔ نمونه‌گیری PPS ارائه خواهد شد، توجه خود را منحصراً بر کاربرد آن در نمونه‌گیری خوشه‌ای دومرحله‌ای متمرکز می‌کنیم. همان‌طور که در بالا اشاره شد، هر گاه خوشه‌ها از لحاظ تعداد واحدهای شمارش موجود در آنها تفاوت بسیار زیادی داشته باشند نمونه‌گیری خوشه‌ها با احتمال نابرابر غالباً به برآوردهایی از مشخصه‌های جامعه، به خصوص مجموعهای جامعه‌ای، منجر خواهد شد که خطاهای معیار آنها کمتر از برآوردهایی است که از نمونه‌گیری خوشه‌ها با احتمال برابر به دست می‌آیند.

از میان راههای گوناگون نمونه‌گیری از خوشه‌ها با احتمال نابرابر، معقول به نظر می‌رسد که نمونه با احتمال متناسب با سطح نوعی مشخصهٔ معلوم خوشه که با متغیر موردنظر آمارگیری نمونه‌ای در ارتباط است انتخاب شود. یک مثال آرمانی شده و رهگشا در وضعیت زیر دیده می‌شود. فرض کنید می‌خواهیم Y ، مجموع یک متغیر را در جامعه‌ای برآورد کنیم و می‌دانیم که با X ، مجموع متغیر معلوم دیگری در جامعه متناسب است. به عبارت دیگر، $Y = cX$ ، که c نامعلوم است. فرض کنید که این رابطه در داخل هر خوشه i نیز مصداق دارد یعنی $Y_i = cX_i$. اگر نمونه‌ای از یک خوشه را با احتمال متناسب با مجموع معلوم، X_i، در آن خوشه انتخاب کنیم، در آن صورت برآورد هنسن ـ هورویتز برای Y از فرمول زیر به دست می‌آید

$$y'_{hh} = \frac{Y_i}{\left(\frac{X_i}{X}\right)} = Y_i\left(\frac{X}{X_i}\right) = \left(\frac{Y_i}{X_i}\right)X$$

ولی

$$\frac{Y_i}{X_i} = \frac{Y}{X} = c$$

پس

$$y'_{hh} = \left(\frac{Y}{X}\right)X = Y$$

بنابراین، هر خوشه‌ای که در نمونه انتخاب شود برآوردگری از Y به دست می‌آید که بدون خطاست.

واضح است که مثال ارائه شده در بالا رابطهٔ قطعی کاملی را بین Y و X فرض می‌گیرد که اگر هم در عمل اتفاق بیفتد بسیار نادر است. ولی اگر همبستگی محکمی بین Y و X در داخل هر خوشه وجود داشته باشد در آن صورت، برآوردگری مبتنی بر نمونه‌گیری با احتمال متناسب با اندازه نسبتاً از انتخاب خوشه‌هایی خاص در نمونه تأثیر نمی‌گیرد و از این رو در مقایسه با شیوهٔ برآوردی که از نمونه‌گیری با احتمال متناسب با اندازه استفاده نمی‌کند خطای نمونه‌گیری کمتری خواهد داشت. باید متذکر شد دلایلی که باعث می‌شوند نمونه‌گیری با احتمال متناسب با اندازه، برآوردهایی تولید کند که خطاهای نمونه‌گیری نسبتاً کمی دارند بسیار شبیه به دلایلی‌اند که برآوردگرهای نسبتی را بسیار قابل اعتماد می‌سازند.

۱.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری: استفاده از برآوردگر هنسن ـ هورویتز

همان طور که در بالا اشاره شد، نمونه‌گیری با احتمال متناسب با اندازه یک ردهٔ کلی از طرحهای نمونه‌ای مبتنی بر نمونه‌گیری با احتمال نابرابر است که در آن، احتمال انتخاب یک واحد در نمونه متناسب است با نوعی متغیر معلوم X . برنامهٔ نمونه‌گیری می‌تواند با یا بدون جایگذاری باشد و می‌تواند به هر شکلی از نمونه‌گیری احتمالاتی باشد (مانند نمونه‌گیری تصادفی ساده، نمونه‌گیری سیستماتیک و غیره). در این زیربخش، صورتی خاص از نمونه‌گیری با احتمال متناسب با اندازه را برای نمونه‌گیری خوشه‌ای دومرحله‌ای شرح و بسط می‌دهیم که دارای خواص زیر است:

۱. در مرحلهٔ اول، نمونه‌ای متشکل از m خوشه با جایگذاری و با احتمال متناسب با سطح نوعی متغیر معلوم x در خوشه که فرض می‌شود با متغیر موردنظر، y، در ارتباط است گرفته می‌شود.
۲. از هر یک از m خوشه‌ای که در مرحلهٔ اول انتخاب شده است، یک نمونهٔ تصادفی ساده متشکل از $\overline{n}$ واحد شمارش از میان N_i واحد شمارش موجود در خوشه گرفته می‌شود.
۳. برآورد مجموع جامعه‌ای، y'_{ppswr} ، از فرمول زیر به دست می‌آید

$$y'_{ppswr} = \frac{۱}{m}\sum_{i=۱}^{m}\frac{N_i}{\bar{n}\pi'_i}y_i \qquad (۸.۱۱)$$

که در آن،

$$\pi'_i = \frac{X_i}{X}$$

$y_i = \sum_{j=۱}^{\bar{n}} y_{ij}$ = مجموع نمونه‌ای i امین خوشهٔ نمونه

$\bar{n}$ = تعداد واحدهای شمارش نمونه‌گیری شده به ازای خوشه

برآوردگری که در برابری (۸.۱۱) نشان داده شده است، تعمیم برآوردگر هنسن ـ هورویتز است که برای نمونه‌گیری خوشه‌ای دومرحله‌ای مناسب است. اگر معیار متغیر اندازه، X_i، عبارت از تعداد، N_i واحد شمارش موجود در خوشه باشد، آن‌گاه، برآوردگر برابر است با

$$y'_{ppswr} = \frac{N}{n}\sum_{i=۱}^{m} y_i \qquad (۹.۱۱)$$

همانند عبارتی که در برابری (۷.۱۱) برای خطای معیار برآوردگر هنسن ـ هورویتز در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ارائه شد، از فرمول زیر می‌توان در نمونه‌گیری دومرحله‌ای با احتمال متناسب با اندازه به عنوان برآوردگر خطای معیار y'_{ppswr} استفاده کرد:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{\sum_{i=۱}^{m}\left(\frac{N_i X y_i}{\bar{n}X_i} - y'_{ppswr}\right)^۲}{m(m-۱)}} \qquad (۱۰.۱۱)$$

مثال تشریحی: جامعهٔ متشکل از ده بیمارستان را در نظر می‌گیریم که در جدول ۹.۱۰ نشان داده شده است (داده‌ها در زیر تکرار می‌شوند).

فرض کنید می‌خواهیم کل تعداد بیماران مرده ترخیص شده را از روی یک نمونهٔ خوشه‌ای دومرحله‌ای برآورد کنیم که در آن یک نمونهٔ تصادفی از دو بیمارستان در مرحلهٔ اول با جایگذاری و با احتمال متناسب با اندازه، N_i پذیرش (که از داده‌های اداری معلوم‌اند) انتخاب می‌شود و در مرحلهٔ دوم، یک نمونهٔ تصادفی ساده متشکل از ۵۰ پذیرش بیمار از هر بیمارستان نمونه گرفته می‌شود.

فرض کنید بیمارستانهای ٦ و ۹ انتخاب شده باشند و ۵۰ سابقهٔ پذیرش نمونه‌گیری شده از بیمارستان شمارهٔ ٦، تعداد ۱۰ نفر را دارای شرایط مهلک نشان داده باشد که ۰ نفر از آنها مرده ترخیص شده‌اند؛ و ۵۰ سابقهٔ نمونه‌گیری شده از بیمارستان شمارهٔ ۹ تعداد ۱۰ نفر را دارای شرایط مهلک نشان داده باشد که ۲ نفر از آنها مرده ترخیص شده‌اند.

جدول ۹.۱۰ کل پذیرشهای دارای شرایط مهلک
و کل پذیرشهایی که از ده بیمارستان مرده ترخیص شده‌اند، ۱۹۸۷

بیمارستان	کل پذیرشها N_i	احتمال انتخاب شدن بیمارستان در هر قرعه‌کشی $\pi'_i = N_i / N$	کل افراد دارای شرایط مهلک	کل ترخیص‌شدگان مرده در میان افراد دارای شرایط مهلک
۱	٤۲۸۸	۰/۰۸٥٦٦٤	٥۰۱	٤۲
۲	٥۰۳٦	۰/۱۰۰٦۰۷	۷۸٥	۷۸
۳	۱۱۷۸	۰/۰۲۳٥۳٤	۲۱۳	۱۷
٤	٦۳۸	۰/۰۱۲۷٤٦	۱۷۳	۹
٥	۲۷۰۱۰	۰/٥۳۹٥۹٦	۳٤۰٤	۳۳۸
٦	۱۱۲۲	۰/۰۲۲٤۱٥	۲۱۷	۱۷
۷	۲۱۳٤	۰/۰٤۲٦۳۲	٤۲٤	۳۷
۸	۱۸۲٤	۰/۰۳٦٤۳۹	۲٤٦	۱۸
۹	٤٦۷۲	۰/۰۹۳۳۳٥	۷۷۸	٦۸
۱۰	۲۱٥٤	۰/۰٤۳۰۳۲	۳٤٦	۲۷

از برابریهای (۹.۱۱) و (۱۰.۱۱) اطلاعات زیر را داریم:

$$m = ۲ \qquad N = ٥۰۰٥٦ \qquad \bar{n} = ٥۰$$
$$y_۱ = ۰ \qquad N_۱ = ۱۱۲۲$$
$$y_۲ = ۲ \qquad N_۲ = ۴٦۷۲$$
$$y'_{ppswr} = \frac{٥۰۰٥٦}{۲ \times ٥۰} \times (۰ + ۲) = ۱۰۰۱/۱۲$$
$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{(۰ - ۱۰۰۱/۱۲)^۲ + (۲۰۰۲/۲۴ - ۱۰۰۱/۱۲)^۲}{۲(۲-۱)}} = ۱۰۰۱/۱۲$$

□

مثال تشریحی: حالا به دومین مثال بر مبنای داده‌های جدول ۹.۱۰ می‌پردازیم. فرض کنید یک نمونه با احتمال متناسب با اندازه گرفته‌ایم که در آن پنج بیمارستان با جایگذاری و با احتمال متناسب با N_i، تعداد پذیرشها انتخاب شده‌اند. از هر یک از بیمارستانهای انتخاب شده یک نمونهٔ تصادفی ساده متشکل از ۱۰ پذیرش را انتخاب می‌کنیم. از روی این نمونه می‌خواهیم کل تعداد اشخاص پذیرفته شده در بیمارستان و با بیماری مهلک و نسبت اشخاص ترخیص شدهٔ مرده از میان این اشخاص دچار بیماری مهلک را برآورد کنیم. مانند مثال قبل از تعمیم برآوردگر هنسن ـ هورویتز استفاده خواهیم کرد.

از پنج بار قرعه‌کشی، بیمارستانهای ۲، ۵، ۵، ۵ و ۹ انتخاب و یافته‌های مربوط در جدول ٦.۱۱ ارائه شده‌اند.

جدول ٦.۱۱ نتایج حاصل از نمونه با احتمال متناسب با اندازه

بیمارستان	احتمال انتخاب در هر نوبت قرعه‌کشی (π_i')	تعداد پذیرشهای نمونه‌گیری شده ($\overline{n}$)	تعداد بیماران دارای شرایط مهلک (x_i)	تعداد بیماران ترخیص شده مرده (y_i)
۲	۵۰۳۶ / ۵۰۰۵۶	۱۰	۱	۱
۵	۲۷۰۱۰ / ۵۰۰۵۶	۱۰	۱	۰
۵	۲۷۰۱۰ / ۵۰۰۵۶	۱۰	۲	۰
۵	۲۷۰۱۰ / ۵۰۰۵۶	۱۰	۱	۰
۹	۴۶۷۲ / ۵۰۰۵۶	۱۰	۱	۱

می‌بینیم که $m = ۵$ و $\overline{n} = ۱۰$ ، و از برابری (۹.۱۱):

$$x'_{ppswr} = \frac{۵۰۰۵۶}{۵۰} \times ۶ = ۶۰۰۶/۷۲$$

و

$$y'_{ppswr} = \frac{۵۰۰۵۶}{۵۰} \times ۲ = ۲۰۰۲/۲۴$$

در برآورد کردن خطاهای معیار این مجموعهای برآورد شده متوجه می‌شویم که هرگاه معیار متغیر اندازه، N_i، تعداد واحدهای شمارش باشد فرمول مربوط به خطای معیار برآورد شده که در برابری (۱۰.۱۱) نشان داده شد به شکل زیر تبدیل می‌شود:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{\sum_{i=۱}^{m}\left(\frac{N y_i}{\overline{n}} - y'_{ppswr}\right)^۲}{m(m-۱)}} \qquad (۱۱.۱۱)$$

از برابری (۱۱.۱۱)، خطاهای معیار زیر را محاسبه می‌کنیم:

$$\hat{SE}(x'_{ppswr}) = ۱۰۰۱/۱۲$$

و

$$\hat{SE}(y'_{ppswr}) = ۱۲۲۶/۱۷$$

بالاخره، نسبت ترخیص‌شدگان مرده در میان پذیرفته‌شدگان دچار بیماریهای مهلک با برآورد نسبتی زیر برآورد می‌شود

$$r_{ppswr} = \frac{y'_{ppswr}}{x'_{ppswr}} = \frac{۲۰۰۲/۲۴}{۶۰۰۶/۷۲} = ۰/۳۳$$

برآوردی از خطای معیار این برآوردگر نسبتی را می‌توان با روشهای خطی کردن به دست آورد که ذیلاً در بحث تحلیل نرم‌افزاری این طرح نشان داده خواهد شد.

اکنون استفاده از SUDAAN و STATA را در تهیهٔ برآوردهایی برای این طرح خاص با احتمال متناسب با اندازه نشان خواهیم داد.

پروندهٔ اطلاعاتی مورد استفاده برای SUDAAN دارای ۵۰ سابقه است که هر سابقه اطلاعاتی را از یک سابقهٔ پذیرش نمونه‌گیری شده در مرحلهٔ دوم پس از انتخاب بیمارستان در مرحلهٔ اول شامل است. به این ترتیب، پروندهٔ اطلاعاتی از ۵۰ سابقهٔ پذیرش ـ یعنی ۱۰ سابقه برای هر یک از پنج بیمارستان قرعه‌کشی شده ـ تشکیل شده است. پروندهٔ اطلاعاتی SUDAAN یک پروندهٔ SAS PC به نام *hospslct.ssd.* است. داده‌های موجود در این پرونده در جدول ۷.۱۱ نشان داده شده‌اند و متغیرهای موجود در پرونده در زیر توصیف شده‌اند:

drawing متغیری است که انتخاب ویژهٔ یک بیمارستان را در مرحلهٔ اول نشان می‌دهد. چون ۵ قرعه‌کشی جداگانه و مستقل انجام شده است مقدار این متغیر بین ۱ تا ۵ است.

hospno نشان‌دهندهٔ بیمارستانی خاص است که سابقهٔ پذیرش از آن نمونه‌گیری شده است (و می‌تواند دارای مقادیری از ۱ تا ۱۰ باشد، زیرا ۱۰ بیمارستان در جامعه بوده‌اند).

admiss نشان‌دهندهٔ تعداد پذیرشهایی است که در طی آن سال در بیمارستانهایی که سوابق پذیرش از آنها قرعه‌کشی شده روی داده‌اند.

Lifethrt نشان می‌دهد که آیا سابقهٔ خاص معرف پذیرش بیماری با شرایط مهلک است یا نه ـ اگر جواب مثبت باشد مقدار آن ۱ و اگر منفی باشد ۰ خواهد بود.

dxdead نشان می‌دهد که بیمار پذیرش شده مرده ترخیص شده است یا نه ـ اگر جواب مثبت باشد ۱ و اگر منفی باشد ۰ خواهد بود.

wstar وزن نمونه‌گیری است ـ که در این مورد برابر است با N/n یا $\frac{۵۰۰۵۶}{۵۰} = ۱۰۰۱/۱۲$.

جدول ۷.۱۱ پروندهٔ داده‌ها برای استفاده در مثال تشریحی

drawing	hospno	admiss	lifethrt	dxdead	wstar
1	2	5036	1	1	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
2	5	27010	1	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
3	5	27010	1	0	1001.12
3	5	27010	1	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
4	5	27010	1	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
5	9	4672	1	1	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12

فرمانهای SUDAAN برای این برآورد ذیلاً نشان داده می‌شود:

```
proc descript data = hospslct   filetype = SAS   design = wr  means totals;
nest  drawing / psulev =1;
weight  wstar;
var  lifethrt  dxdead;
proc  ratio  data = hospslct  filetype = SAS  design = wr;
nest  drawing / psulev =1;
weight  wstar;
numer dxdead;
denom lifethrt;
```

فرمان اول، مدول خاص SUDAAN را که قرار است مورد استفاده قرار گیرد، پروندهٔ اطلاعاتی و نوع آن، طرح خاص SUDAAN (که در این مورد WR یا نمونه‌گیری با جایگذاری است)، و محاسبهٔ میانگینها و مجموعها هر دو را نشان می‌دهد. فرمان دوم، خوشه‌بندی را توصیف می‌کند و نشان می‌دهد که واحد نمونه‌گیری اولیه در متغیر *drawing* شناسایی شده است. باید دقیقاً توجه داشت که متغیر *hospno* که نشانهٔ شناسایی بیمارستان است واحد نمونه‌گیری اولیه نیست. در نمونه‌گیری با احتمال متناسب با اندازه (PPS) و با جایگذاری، یک خوشهٔ معین (در این مورد یعنی بیمارستان) می‌تواند بیش از یک بار در نمونه انتخاب شود و هر نوبت قرعه‌کشی یک خوشه یک واحد نمونه‌گیری اولیه به شمار می‌رود. فرمان سوم، وزن نمونه‌گیری را نشان می‌دهد و فرمان چهارم نشان‌دهندهٔ متغیری است که مجموعها و میانگینهای آن باید برآورد شوند. بقیهٔ فرمانها برای برآورد کردن نسبتی به کار می‌روند و گویا هستند. خروجی متناظر SUDAAN برای این برآورد به صورت زیر است:

by: Variable, One.

Variable		One 1
LIFETHRT	Sample Size	50
	Weighted Size	50056.00
	Total	6006.72
	SE Total	1001.12
	Mean	0.12
	SE Mean	0.02
DXDEAD	Sample Size	50
	Weighted Size	50056.00
	Total	2002.24
	SE Total	1226.12
	Mean	0.04
	SE Mean	0.02

by: Variable, One.

Variable		One 1
DXDEAD/ LIFETHRT	Sample Size	50.0000
	Weighted Size	50056.0000
	Weighted X-sum	6006.7200
	Weighted Y-sum	2002.2400
	Ratio Est.	0.3333
	SE Ratio	0.2324

تحلیلی مشابه را می‌توان با استفاده از STATA به وسیلهٔ فرمانهای زیر اجرا کرد:

```
use " a:\hospslct.dta",clear
. svyset pweight wstar
. svyset psu drawing
. svytotal lifethrt dxdead
.svyratio dxdead lifethrt
```

که خروجی زیر را نتیجه می‌دهد:

Survey total estimation

pweight: wstar — Number of obs = 50
Strata: < one > — Number of strata = 1
PSU: drawing — Number of PSUs = 5
Population size = 50056

Total	Estimate	Std. Err.	[%95 Conf.	Interval]	Deff
lifethrt	6006.72	1001.12	3227.165	8786.275	.1856061
dxdead	2002.24	1226.117	-1402.005	5406.485	.765625

.svyratio dxdead lifethrt

Survey ratio estimation

pweight: wstar — Number of obs = 50
Strata: < one > — Number of strata = 1
PSU: drawing — Number of PSUs = 5
Population size = 50056

Ratio	Estimate	Std. Err.	[%95 Conf.	Interval]	Deff
Dxdead/lifethrt	.3333333	.2324056	-.3119279	.9785946	1.429167

□

۲.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه وقتی معیار متغیر اندازه، تعداد واحدهای شمارش نباشد.

گاهی در وضعیتی قرار می‌گیریم که استفاده از N_i، تعداد واحدهای شمارش موجود در داخل خوشه به عنوان معیار متغیر اندازه در نمونه‌گیری با احتمال متناسب با اندازه، نه امکانپذیر است و نه مطلوب. اغلب اوقات، تعداد واحدهای شمارش پیش از نمونه‌گیری معلوم نیست. ولی گاهی اوقات، تعداد واحدهای شمارش موجود در داخل خوشه ممکن است ارتباط خاصی با متغیر پاسخ نداشته باشد و به

این لحاظ استفاده از آن به عنوان معیار برای متغیر اندازه فایدهٔ چندانی نخواهد داشت. مثلاً اگر بیمارستانها خوشه‌ها باشند و برآورد کردن نوعی مشخصهٔ بیماران دچار سرطان پوست موردنظر باشد، تعداد بیماران دچار سرطان پوست که در یک بیمارستان خاص پذیرش شده‌اند احتمال دارد بیشتر با اندازه و کیفیت بخشهای تخصصی بیماریهای پوستی و غده‌شناسی بیمارستان ارتباط داشته باشد تا به اندازهٔ کلی بیمارستان که با مجموع تعداد پذیرشها در آن بیمارستان مشخص می‌شود.

در وضعیتهایی که معیار متغیر اندازه که در ایجاد نمونهٔ PPS به کار رفته است تعداد واحدهای شمارش نیست، فرمولهای مناسب برای برآورد کردن مجموع جامعه و برآورد کردن خطای معیار این برآوردگر مجموع به ترتیب، برابریهای (۸.۱۱) و (۱۰.۱۱) هستند. در هر دو وضعیت، مجموع برآورد شده، مجموعی موزون از مشاهدات نمونه‌ای تک است که به صورت $y'_{ppswr} = \sum_{i=1}^{m} \sum_{j=1}^{\bar{n}} w_i y_{ij}$ است. در وضعیتی که قبلاً بحث شد و در آن معیار متغیر اندازه متناسب با N_i، تعداد واحدهای شمارش موجود در خوشه بود، وزن، w_i، برابر است با N/n و بنابراین، مستقل از خوشه‌ای که از آن نمونه‌گیری شده است، برای هر مشاهدهٔ نمونه یکسان است. در وضعیتی که در آن معیار متغیر اندازه، تعداد واحدهای شمارش موجود در خوشه نیست، وزن، w_i، برابر است با $(N_i X)/(nX_i)$ و بنابراین برای هر خوشه یکسان نیست.

در مثال تشریحی که در بالا مورد بحث قرار گرفت، اگر معیار متغیر اندازهٔ مورد استفاده در نمونه‌گیری خوشه‌ها با احتمال متناسب با اندازه، X_i، تعداد پذیرشهای با شرایط مهلک در بیمارستان باشد، آن‌گاه، همان داده‌های نمونه که در جدول ۷.۱۱ نشان داده شدند برآوردهای زیر را به دست می‌دهند (فرمانهای SUDAAN و خروجی آن در زیر نشان داده شده‌اند):

```
proc  descript data = hspslct2 filetype = sas means totals;
nest drawing/psulev = 1;
weight w2star;
var lifethrt dxdead;
proc ratio data = hspslct2  filetype = sas;
nest drawing/psulev = 1;
weight w2star;
numer dxdead;
denom lifethrt;
```

در این فرمانها، متغیر *w2star* برابر است با $(N_i X)/(nX_i)$.

خروجی مربوط به برآورد مجموعها به صورت زیر است:

LIFETHRT	Sample Size	50
	Weighted Size	51344.99
	Total	6259.18
	SE Total	1277.32
	Mean	0.12
	SE Mean	0.02
DXDEAD	Sample Size	50
	Weighted Size	51344.99
	Total	1760.47
	SE Total	1079.04
	Mean	0.03
	SE Mean	0.02

خروجی مربوط به برآورد نسبت به شکل زیر خواهد بود:

Variable		One 1
DXDEAD/LIFETHRT	Sample Size	50
	Weighted Size	51344.99
	Weighted X-Sum	6259.18
	Weighted Y-Sum	1760.47
	Ratio Est.	0.28
	SE Ratio	0.21

۳.۳.۱۱ چگونگی انتخاب یک نمونه با احتمال متناسب با اندازه و با جایگذاری

از نظر عملیاتی، انتخاب یک نمونهٔ دومرحله‌ای با احتمال متناسب با اندازه و با جایگذاری بسیار آسان است. این روش را باز هم با استفاده از جامعهٔ بیمارستانهای جدول ۹.۱۰ نشان می‌دهیم. باز فرض می‌کنیم که N_i، معیار متغیر اندازه‌ای است که قرار است در انتخاب نمونه مورد استفاده قرار گیرد. ابتدا معیار متغیر اندازه (در این مورد N_i) را انباشته می‌کنیم و اعداد تصادفی را با هر خوشه به صورتی که در جدول ۸.۱۱ نشان داده می‌شود همراه می‌کنیم. تعداد اعداد تصادفی همراه با هر خوشه، به شکلی که در جدول نشان داده شده برابر با N_i است.

اگر قرار باشد m خوشه برای نمونه انتخاب و $\overline{n}$ واحد نمونه‌گیری در هر قرعه‌کشی از خوشه نمونه‌گیری شود، شیوهٔ کار مستلزم (۱) انتخاب یک عدد تصادفی بین ۱ و N ؛ (۲) شناسایی خوشهٔ متناظر با عدد تصادفی؛ (۳) انتخاب یک نمونهٔ تصادفی ساده بدون جایگذاری با $\overline{n}$ واحد شمارش در داخل خوشه؛ و (٤) تکرار این روش تا انتخاب m خوشه و $n = m\overline{n}$ واحد شمارش خواهد بود.

برای نشان دادن این روش، فرض کنید $m = ۵$ و $\overline{n} = ۶$. ابتدا یک عدد تصادفی بین ۰۰۰۰۱ و ۵۰۰۵٦ انتخاب می‌کنیم. فرض کنید این عدد ۳٦۲۰۷ است. این عدد با بیمارستان ۵ متناظر

است. سپس ٦ عدد تصادفی را بین ۰۰۰۱ و ۲۷۰۱۰ بدون جایگذاری انتخاب می‌کنیم و پذیرشهای متناظر با این اعداد را برای نمونه‌های آن خوشهٔ انتخاب شدهٔ خاص در نظر می‌گیریم. بعد یک عدد تصادفی دیگر را از میان اعداد ۰۰۰۰۱ و ٥۰۰٥٦ انتخاب می‌کنیم تا بیمارستان دوم انتخاب شود. فرض کنید این عدد تصادفی ٤۲۷٥۱ است. این عدد با بیمارستان ۸ متناظر است که ۱۸۲٤ پذیرش داشته است. پس ٦ عدد تصادفی را بین ۰۰۰۱ و ۱۸۲٤ بدون جایگذاری انتخاب می‌کنیم تا ٦ پذیرش از این بیمارستان انتخاب شود. این فرایند را ادامه می‌دهیم تا شش پذیرش از هر پنج بیمارستان انتخاب شود. از جدول ۸.۱۱ درمی‌یابیم که در این تمرین سه بار بیمارستان ٥ و یک بار بیمارستانهای ۲ و ۸ انتخاب شده‌اند. پذیرشهای ویژه‌ای که در مرحلهٔ دوم نمونه‌گیری شده‌اند در ستون ششم جدول ۸.۱۱ نشان داده شده‌اند.

٤.۳.۱۱ نمونهٔ مورد نیاز برای نمونهٔ دومرحله‌ای که در آن خوشه‌ها با احتمال متناسب با اندازه و با جایگذاری انتخاب می‌شوند چقدر باید بزرگ باشد؟

در نمونه‌گیری خوشه‌ای دومرحله‌ای، خطای معیار یک پارامتر برآورد شده نه تنها از n ، کل تعداد واحدهای شمارش نمونه‌گیری شده بلکه با استفاده از دو عاملی که n را تشکیل می‌دهند یعنی m ، تعداد خوشه‌های نمونه‌گیری شده و $\overline{n}$، متوسط تعداد واحدهای شمارش نمونه‌گیری شده به ازای هرخوشه تعیین می‌شود. به این ترتیب، مثلاً برآورد مجموع یک جامعه، که از یک نمونهٔ دومرحله‌ای متشکل از ۸۰ واحد شمارش در ۱۰ خوشهٔ نمونه که هر یک ۸ واحد شمارش نمونه دارند به دست آمده است، احتمال دارد خطای معیاری داشته باشد که با برآورد حاصل از یک نمونهٔ متشکل از ۲۰ خوشه که هر یک ٤ واحد شمارش نمونه‌ای دارند فرق داشته باشد. ولی غالباً $\overline{n}$، تعداد واحدهای شمارش که قرار است نمونه‌گیری شوند پیشاپیش بر مبنای هزینه و ملاحظات مربوط به همبستگی میان ـ رده‌ای تعیین می‌شود (مانند مواردی که درفصل ۱۰ بحث شد) به طوری که مسئلهٔ تعیین اندازهٔ نمونه، به مسئلهٔ تعیین m، تعداد خوشه‌هایی که باید نمونه‌گیری شوند، به فرض آنکه $\overline{n}$ واحد شمارش به ازای هر خوشهٔ نمونه انتخاب شود تبدیل خواهد شد. ما درفرمول‌بندیهای خود برای مسئلهٔ اندازهٔ مورد نیاز نمونه در نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری، فرضهای زیر را در نظر می‌گیریم:

۱. $\overline{n}$، تعداد واحدهای شمارش که باید در داخل هر خوشه نمونه‌گیری شوند برای هرخوشهٔ نمونه یکسان و از قبل تعیین شده است.

جدول ۸.۱۱ شیوهٔ نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری

بیمارستان	مجموع پذیرشها N_i	پذیرشهای انباشته	اعداد تصادفی	اعداد تصادفی انتخاب شده برای مرحلهٔ اول	اعداد تصادفی انتخاب شده برای مرحلهٔ دوم
۱	٤٢٨٨	٤٢٨٨	٠٠٠٠١–٠٤٢٨٨		
۲	٥٠٣٦	٩٣٢٤	٠٤٢٨٩–٠٩٣٢٤	٠٨٥٨٩	٠٤٨٠، ٢٣٦٨، ٤١٣٠، ٢١٦٧، ٣٤٧٥، ٣٥٥٣
۳	١١٧٨	١٠٥٠٢	٠٩٣٢٥–١٠٥٠٢		
٤	٦٣٨	١١١٤٠	١٠٥٠٣–١١١٤٠		
٥	٢٧٠١٠	٣٨١٥٠	١١١٤١–٣٨١٥٠	٣٦٢٠٧	٠٩٤٢٩، ١٠٣٦٥، ٠٧١١٩، ٠٢٣٦٨، ٠١٠١١، ٠٧٠٥٦
				١٨٦٠٢	١٦٦٣١، ١٦٨١٥، ٢٠٢٠٦، ٠٥٣٠٠، ٢٢١٦٤، ٢٤٣٦٩
				١٨٧٣٨	١٩٦٨٧، ١١٠٥٢، ١٩٧٤٦، ١٤٣٤٩، ٠٠٦٩٧، ١٩١٢٤
٦	١١٢٢	٣٩٢٧٢	٣٨١٥١–٣٩٢٧٢		
۷	٢١٣٤	٤١٤٠٦	٣٩٢٧٣–٤١٤٠٦		
۸	١٨٢٤	٤٣٢٣٠	٤١٤٠٧–٤٣٢٣٠	٤٢٧٥١	١٦٢٤، ١٦٠١، ١٢٤٨، ٠١١٩،١٨٠٥، ٠٩٠٣
۹	٤٦٧٢	٤٧٩٠٢	٤٣٢٣١–٤٧٩٠٢		
۱۰	٢١٥٤	٥٠٠٥٦	٤٧٩٠٣–٥٠٠٥٦		

۲. برآوردی «مقدماتی» از نمونهٔ خوشه‌ای با احتمال متناسب با اندازه از سطح متغیر موردنظر (که در بحث ما مجموع جامعه‌ای آن متغیر است) وخطای معیار آن موجود است. (بنابراین، ضریب تغییرات برآورد را می‌دانیم.)
۳. می‌خواهیم با اطمینان $(1-\alpha)\times 100$ درصد تعداد خوشه‌های نمونه‌ای مورد نیاز برای برآورد کل سطح متغیری خاص را در جامعه حول $\varepsilon\times 100$ درصد مقدار واقعی آن بر آورد کنیم.

فرض می‌کنیم که معیار متغیر اندازه برای نمونه‌گیری با احتمال متناسب با اندازه، تعداد واحدهای شمارش در داخل هر خوشه است که نتیجه می‌دهد که برآورد مقدماتی خطای معیار از فرمول برابری (۱۱.۱۱) به دست می‌آید. توجه کنید که این برآورد به صورت زیر است:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{G(y_1, y_2, \ldots, y_m)}{m}}$$

که در آن:

$$G(y_1, y_2, \ldots, y_m) = \frac{\sum_{i=1}^{m}\left(\frac{Ny_i}{n} - y'_{ppswr}\right)^2}{m-1}$$

از این فرمول، داریم

$$m = \frac{z^2_{1-\alpha/2} G(y_1, y_2, \ldots, y_m)}{(y'_{ppswr})^2 \varepsilon^2} \qquad (۱۲.۱۱)$$

و برآورد m، تعداد خوشه‌های مورد نیاز برای تأمین ویژگیهای فوق‌الذکر تقریباً برابر است با آنچه در برابری (۱۲.۱۱) نشان داده شده است.

مثال تشریحی: باز هم جامعهٔ متشکل از ۱۰ بیمارستان جدول ۹.۱۰ را در نظر می‌گیریم. فرض کنید می‌خواهیم یک نمونه با احتمال متناسب با اندازه و با جایگذاری بگیریم که مرحلهٔ اول آن از m قرعه‌کشی مستقل از یک بیمارستان تشکیل می‌شود که در هر مرحله از قرعه‌کشی، یک نمونهٔ تصادفی ساده بدون جایگذاری با انتخاب ۱۰ واحد شمارش گرفته می‌شود. باز هم فرض کنید که یک آمارگیری مقدماتی از ۵ بیمارستان به عمل آمده است (با استفاده از همان طرح نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری که قرار است در آمارگیری اصلی به کار گرفته شود). بالاخره فرض می‌کنیم که داده‌های حاصل از بررسی مقدماتی همان داده‌هایی هستند که در بحث قبلی ما در مورد آن آمارگیری توصیف شده‌اند. حالا می‌خواهیم از این داده‌ها برای محاسبهٔ m، تعداد خوشه‌هایی استفاده

کنیم که باید نمونه‌گیری شوند تا ۹۵ درصد مطمئن باشیم که y'_{ppswr}، برآورد کل Y، تعداد پذیرش‌شدگان با شرایط مهلک، حول ۳۰ درصد مقدار واقعی آن باشد.

داده‌های حاصل از آن آمارگیری مقدماتی که برای برآورد کردن m، تعداد خوشه‌های نمونهٔ مورد نیاز ضروری است ذیلاً نشان داده شده‌اند:

$$y'_{ppswr} = ۶۰۰۶/۷۲$$

$$\hat{SE}(y'_{ppswr}) = ۱۰۰۱/۱۲$$

$m^* = ۵$ (تعداد خوشه‌های نمونه‌گیری شده در آمارگیری مقدماتی)

$$G(y_۱,\ldots,y_m) = (۱۰۰۱/۱۲\times\sqrt{۵})^۲ = (۲۲۳۸/۵۷۲)^۲ = ۵۰۱۱۲۰۶/۲۷۲$$

$$\overline{n} = ۱۰$$

$$\varepsilon = ۰/۳$$

$$z_{۱-(\alpha/۲)} = ۱/۹۶$$

با وارد کردن مقادیر بالا در برابری (۱۲.۱۱) داریم

$$m = ۵/۹ \approx ۶$$

بنابراین، شش قرعه‌کشی از بیمارستانها (در مجموع، ۶۰ نمونهٔ پذیرش) لازم است تا ویژگیهای تعیین شده برای برآورد کل تعداد پذیرش‌شدگان دارای شرایط مهلک تأمین شود.

□

۵.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازهٔ تلفنی: روش شماره‌گیری ارقام تصادفی میتوفسکی ـ واکسبرگ[1]

نمونه‌گیری تلفنی به قدری در روش‌شناسی آمارگیری اهمیت پیدا کرده است که یک فصل کامل (فصل ۱۵) را به آن اختصاص داده‌ایم تا به جزئیات این موضوع بپردازیم. این روش در فصل حاضر به عنوان مثالی از نمونه‌گیری با احتمال متناسب با اندازه مطرح می‌شود و نشان می‌دهیم که یکی از فنون آن که کاربرد وسیعی دارد و به نام *روش شماره‌گیری ارقام تصادفی میتوفسکی ـ واکسبرگ* معروف است در واقع صورتی از نمونه‌گیری با احتمال متناسب با اندازه است.

کاربرد مصاحبهٔ تلفنی در آمارگیریهای نمونه‌ای در سالهای اخیر بسیار افزایش یافته است، بدواً به این دلیل که هزینه‌های میدانی ناشی از مراجعات حضوری به خانوارها غالباً بازدارنده است. اگر بپذیریم که همهٔ خانوارها تلفن ندارند و اگر کل خانوارهای دارای تلفن به عنوان جامعهٔ هدف قابل قبول باشد، در آن صورت مصاحبهٔ تلفنی غالباً به عنوان جایگزینی برای مراجعهٔ حضوری به خانوار مورد استفاده قرار می‌گیرد.

[1] Mitofsky-Waksberg

در آمارگیریهای تلفنی می‌توان خانوارها را به چندین طریق برای گنجاندن در نمونه انتخاب کرد. مثلاً از کتاب راهنمای تلفن منطقهٔ هدف می‌توان به عنوان چارچوب نمونه‌گیری استفاده کرد. ولی کتابهای راهنمای تلفن، شماره‌های فهرست نشده را و شماره‌هایی را که بعد از تاریخ انتشار آخرین نسخهٔ کتاب راهنما به خانوارها واگذار شده‌اند ندارند. حذف این خانوارها ممکن است منشأ اریبی زیاد باشد زیرا خانوارهایی که شماره‌هایشان فهرست نشده است یا شماره‌های تازه‌ای گرفته‌اند می‌توانند درصد قابل ملاحظه‌ای از خانوارهای دارای تلفن را تشکیل دهند.

چارچوب نمونه‌گیری دیگری برای آمارگیریهای تلفنی، فهرستی از تمام شماره‌های چهار رقمی است که داخل مراکز تلفن موجودند. یک شمـــارهٔ تلفن از ده رقم تشکیل شده است. سه رقـــم اول *کد ناحیه* را نشان می‌دهند، سه رقم بعدی مرکز *تلفن مربوط* را معین می‌کند و چهار رقم آخر مربوط به *شمارهٔ تلفنی* خاص است. مثلاً برای اینکه شمارهٔ تلفن رئیس دانشکدهٔ بهداشت عمومی دانشگاه ایلی‌نوی در شیکاگو را بگیرید، اول باید کد ۳۱۲ (کد *ناحیهٔ* شیکاگو) را شماره‌گیری کنید، بعد باید شـمارهٔ ۹۹٦ (شمارهٔ مرکز تلفن ویژهٔ تمام دانشگاه ایلـی‌نوی در شیـکاگـو) را بگیریـد و پس از آن شـمارهٔ ٦٦۲۳ (خط اختصاصی رئیس دانشکده) را بگیرید. می‌توانید با شماره‌گیری ده رقم ۳۱۲۹۹٦٦٦۲۳ این موضوع را تحقیق کنید. رئیس فعلی این دانشکده فردی بسیار اجتماعی است و از این‌گونه مکالمات تلفنی استقبال می‌کند. استفاده از این ارقام به عنوان چارچوب نمونه‌گیری به نام *شماره‌گیری ارقام تصادفی* مشهور است و از بسیاری از اریبیهایی که غالباً با استفاده از دفترچه راهنمای تلفن همراه است پرهیز می‌کند. ادامهٔ بحث ما در مورد شماره‌گیری ارقام تصادفی صرفاً براساس روشهایی است که در اصل توسط میتوفسکی شرح و بسط داده شده [۹] و بعدها توسط واکسبرگ پیراسته شده و گسترش یافته است [۱۰].

فهرستی از کدهای ناحیه‌ای و مراکز تلفنی منطقهٔ هدف آمارگیری را می‌توان از دفتر شرکت تلفن محلی تهیه کرد. به این ترتیب، در محدودهٔ یک کد محلی (مثلاً ۳۱۲) و مرکز تلفن (مثلاً ۹۹٦) فقط چهار رقم آخر شماره تلفنها باید به طور تصادفی انتخاب شوند. ولی معلوم می‌شود که نسبت بزرگی (تقریباً ۸۰ درصد) از همهٔ شماره تلفنهایی که با یک کد محلی و مرکز تلفن خاص مشخص شده‌اند یا اصلاً استفاده نمی‌شوند یا به مشاغل، مؤسسات، یا سایر امور غیر از خانوارها اختصاص داده شده‌اند. به عبارت دیگر، فرایند گرفتن شماره‌های تصادفی چهاررقمی با یک کد ناحیه‌ای و مرکز تلفن معین بی‌حاصل خواهد بود زیرا باید تقریباً پنج شماره تلفن را برای به دست آوردن یک خانوار بگیریم. میتوفسکی و واکسبرگ شیوهٔ دیگری را پیشنهاد کرده‌اند که عملکرد شماره‌های متناظر با خانوارها را افزایش می‌دهد. در این شیوه، شمارهٔ تلفنها در خوشه‌هایی که (به جای شش رقم اول) با ۸ رقم اول

مشخص شده‌اند گروه‌بندی می‌شوند. مثلاً اگر جامعهٔ هدف شامل خانوارهایی است که کد ناحیه‌ای آنها ۳۱۲ و مراکز تلفنی آنها ۹۹٦ و ۸۳۵ است، شماره‌ها در ۲۰۰ خوشه براساس ۸ رقم اول به شرح زیر گروه‌بندی می‌شوند:

(۱۰۰ خوشه) ۹۹-۸۳۵-۳۱۲ تا ۰۰-۸۳۵-۳۱۲

(۱۰۰ خوشه) ۹۹-۹۹٦-۳۱۲ تا ۰۰-۹۹٦-۳۱۲

خوشه‌ها به شرح زیر (مثلاً ۱ تا ۲۰۰، مانند بالا) شناسایی و نمونه‌گیری می‌شوند. یک خوشهٔ تصادفی انتخاب می‌شود (مثلاً ٤۷-۹۹٦-۳۱۲) و یک عدد تصادفی بین ۰۰ و ۹۹ (مثلاً ٦۰) گرفته می‌شود. بعد شمارهٔ تلفن (مثلاً ٤۷٦۰-۹۹٦-۳۱۲) شماره‌گیری می‌شود. اگر این شماره تلفن، مربوط به یک خانوار باشد خوشه حفظ می‌شود و شماره‌های دورقمی دیگری گرفته می‌شوند تا به مجموع $\bar{n}$ از ساکنان (شامل اولین خانوار شماره‌گیری شده) برسیم. اگر شمارهٔ اولیه مربوط به یک خانوار نباشد خوشه را کنار می‌گذاریم. این شیوه ادامه می‌یابد تا کل m خوشه نمونه‌گیری شود و در نتیجه مجموع $m\bar{n}$ مصاحبه به دست آید.

شیوهٔ میتوفسکی ـ واکسبرگ بهبود قابل توجهی در نسبت خانوارهای شماره‌گیری شده به بار آورده است که علت آن، گرایش شماره تلفنهای اماکن مسکونی به خوشه‌بندی شدن در «ردیفها»یی از شماره‌های دنباله‌ای (پی‌درپی) بود. طرح نمونه‌ای که در بالا شرح داده شد طرح نمونه‌گیری با احتمال متناسب با اندازه است زیرا اولین مکالمهٔ تلفنی انجام شده درخوشه تعیین می‌کند که خوشه در نمونه قرار می‌گیرد یا نه و احتمال اینکه این تلفن اولیه به مکالمه با یک خانوار بینجامد بستگی دارد به نسبت ۱۰۰ شمارهٔ موجود در خوشه که به خانوارها مربوط می‌شوند.

٤.١١ توضیح بیشتر دربارهٔ نمونه‌گیری با احتمال متناسب با اندازه

در مبحث نمونه‌گیری با احتمال متناسب با اندازه بر یک برنامهٔ نمونه‌گیری و شیوهٔ برآورد کردن خاص متمرکز شدیم. برنامهٔ نمونه‌گیری مستلزم انتخاب خوشه‌هایی با استفاده از نمونه‌گیری تصادفی ساده با جایگذاری و سپس انتخاب یک نمونهٔ تصادفی ساده (بدون جایگذاری) از واحدهای شمارش در هر خوشهٔ نمونه در هر نوبت انتخاب آن خوشه در نمونه بود. شیوهٔ برآورد کردن مستلزم استفاده از برآوردگر هنسن ـ هورویتز بود. از این نوع نمونه‌گیری PPS در آمارگیریها زیاد استفاده می‌شود زیرا برآوردهای خطاهای معیار را می‌توان چون خوشه‌ها مستقل از یکدیگر انتخاب می‌شوند نسبتاً به آسانی به دست آورد. برآوردگر هنسن ـ هورویتز در ترکیب با این نوع نمونه‌گیری مورد استفاده قرار می‌گیرد زیرا محاسبهٔ آن به مراتب از برآوردگر هورویتز ـ تامپسون آسانتر است.

اما، انواع دیگری از نمونه‌گیری با احتمال متناسب با اندازه وجود دارند که گاهی در عمل مورد استفاده قرار می‌گیرند. مثلاً خوشه‌ها را می‌توان با نمونه‌گیری تصادفی ساده بدون جایگذاری یا با شکل دیگری از نمونه‌گیری بدون جایگذاری انتخاب کرد. چون برآوردگر هنسن ـ هورویتز تنها در صورتی می‌تواند مورد استفاده قرار گیرد که نمونه‌گیری با جایگذاری باشد، در نمونه‌گیری PPS که خوشه‌ها بدون جایگذاری انتخاب می‌شوند غالباً از برآوردگر هورویتز ـ تامپسون استفاده می‌شود. خطاهای معیار برآوردهای حاصل را نمی‌توان به آسانی از این قبیل طرحها به دست آورد، هر چند روشهای گوناگونی برای این کار پیشنهاد شده‌اند [۱].

در طرحهای نمونه‌گیری با احتمال متناسب با اندازه، اگر تعداد واحدهای شمارش که در داخل هر خوشهٔ نمونهٔ مرحلهٔ اول انتخاب می‌شوند یکسان باشد آن‌گاه، احتمال انتخاب شدن هر واحد شمارش در نمونه یکسان است و هیچ فرقی نمی‌کند که واحد شمارش به کدام خوشه تعلق داشته باشد. به عبارت دیگر، نمونه *خود ـ وزن* است به این مفهوم که هر واحد شمارش در نمونه معرف همان تعداد واحد شمارش در جامعه است. این حالت معمولاً در نمونه‌گیری خوشه‌ای دو مرحله‌ای ساده صدق نمی‌کند مگر اینکه خوشه‌ها دارای تعداد واحدهای شمارش یکسان N_i، باشند.

۵.۱۱ خلاصه

در این فصل، از طریق مثالهایی نشان دادیم که نمونه‌گیری خوشه‌ای دو مرحله‌ای ساده می‌تواند برآوردهایی به بار آورد که خطاهای نمونه‌گیری آنها بسیار زیاد است، به خصوص اگر خوشه‌ها از لحاظ N_i، تعداد واحدهای شمارش درون خوشه‌ها پراکندگی قابل ملاحظه‌ای داشته باشند.این موضوع به خصوص در مورد برآورد کردن مجموعهای جامعه‌ای صدق می‌کند. همچنین نشان دادیم که استفاده از یک طرح نمونه که در آن خوشه‌ها با احتمال متناسب با نوعی مشخصهٔ خوشه نمونه‌گیری می‌شوند (مانند تعداد واحدهای شمارش موجود در خوشه) می‌تواند به برآوردهایی منجر شود که خطاهای نمونه‌گیری آنها به مراتب کمتر از خطاهای نمونه‌گیری حاصل از خوشه‌هایی با احتمال برابر است. دو ردهٔ کلی از برآوردگرها یعنی برآوردگر هورویتز ـ تامپسون و برآوردگر هنسن ـ هورویتز مورد بحث قرار گرفتند و می‌توانند در ترکیب با نمونه‌گیری با احتمال نابرابر به کار روند. بحث خود را در مورد نمونه‌گیری با احتمال متناسب با اندازه بر وضعیتی متمرکز کردیم که در آن خوشه‌ها به وسیلهٔ نمونه‌گیری تصادفی ساده با جایگذاری انتخاب می‌شوند و احتمال انتخاب شدن متناسب با معیاری از اندازهٔ خوشه است. در هر نوبت از انتخاب خوشه، یک نمونهٔ تصادفی ساده از واحدهای شمارش بدون جایگذاری انتخاب می‌شود و برآورد کردن با استفاده از برآوردگر هنسن ـ هورویتز انجام می‌گیرد. با استفاده از مثالهای تشریحی نشان دادیم که نمونه‌ها چگونه با این شیوه انتخاب می‌شوند،

پارامترها چگونه برآورد می‌شوند، و خطاهای معیار چگونه به دست می‌آیند. فرمولهای مربوط به تعیین اندازهٔ نمونه را برای این طرح شرح و بسط دادیم. بالاخره، در مورد کاربرد نمونه‌گیری با احتمال متناسب با اندازه در آمارگیریهای تلفنی شماره‌گیری با ارقام تصادفی نیز بحث کردیم.

تمرین

بار دیگر جامعهٔ متشکل از ۲۵ مرکز بهداشت روانی محلی را از تمرینهای ۱۲.۱۰ تا ۱۵.۱۰ در نظر می‌گیریم (که جدول مربوط به آن در زیر تکرار می‌شود).

مرکز بهداشت	تعداد بیماران	مرکز بهداشت	تعداد بیماران
۱	٤٩١	١٤	٦٧٢
۲	٨٦٦	١٥	٤٧٥
۳	١٨٨	١٦	٤٣٩
٤	٩٩٤	١٧	٣٩٢
٥	٢٠٩	١٨	٥٨٤
٦	٩٦١	١٩	٨٨٢
۷	٨٣٤	٢٠	٤٢٤
۸	٩٨٢٠	٢١	٧٧٥
۹	٣٤٨	٢٢	٢٦٢
۱۰	٢٤٦	٢٣	٩٦٨
۱۱	٣٩٩	٢٤	٥٨٦
۱۲	١٧٥	٢٥	٨٠٩
۱۳	١٦٦		

۱.۱۱ فرض کنید می‌خواهید برای برآورد کردن کل تعداد بیمارانی که در حال حاضر پروزاک (Prozac) دریافت می‌کنند یک نمونهٔ دو مرحله‌ای با احتمال متناسب با اندازه و با جایگذاری از مراکز بهداشتی بگیرید. باز هم فرض کنید که معیار متغیر اندازه برای نمونه‌گیری با احتمال متناسب با اندازه، تعداد بیماران است و اعداد تصادفی انتخاب شده ۱۱۰۵۲ و ١٢٦١٤ هستند.

الف. دو مرکز بهداشتی نمونه‌گیری شده کدامها هستند؟

ب. فرض کنید یک نمونهٔ تصادفی ساده متشکل از ۲۰ بیمار، در هر نوبت، از مرکز بهداشتی انتخاب شده در آن نوبت، گرفته می‌شود. باز هم فرض کنید که نمونهٔ متناظر با عدد تصادفی ۱۱۰۵۲ شامل ۱۲ بیمار است که پروزاک دریافت می‌کنند و نمونهٔ متناظر با عدد تصادفی ١٢٦١٤ شامل ٦ بیمار تحت درمان با پروزاک است. از برآوردگر

هنسن ـ هورویتز برای برآورد کردن کل تعداد افراد تحت درمان با پروزاک استفاده کنید. برآورد خطای معیار این برآوردگر چقدر است؟

۲.۱۱ مطلوب است برآورد تعداد کره‌ایهایی که در یک شهر معین زندگی می‌کنند. این شهر دارای شش مرکز تلفن است و بررسی آخرین کتاب راهنمای تلفن، فراوانیهای زیر را برای دو نام فامیلی کره‌ای که از همه متداولترند ـ یعنی کیم و پارک ـ نشان داده است.

مرکز تلفن	جمعیت ناحیهٔ جغرافیایی متناظر با مرکز تلفن	تعداد کیم‌ها و پارک‌ها در راهنمای تلفن
۸۳۲	۵۲۳۱	۲۱
۸۵٦	۳۰۱۲	۸
۹۳۵	۲۱۲۳	۷
۹۳٦	۱۲۵٦	۳۵
۹۳۷	۲۵٦۹	۱۷
۹۸۳	۸۳۲۱	۱۰

تصور می‌شود نام فامیلی تقریباً ٦۰ درصد از همهٔ کره‌ایهای ساکن در آن منطقه «کیم» یا «پارک» است و یک خانوار متوسط کره‌ای از چهار نفر تشکیل شده است.

الف. اگر قرار می‌شد از مرکز تلفن به عنوان خوشه استفاده کنید از چه روشی برای نمونه‌گیری خوشه‌ها استفاده می‌کردید؟

ب. از این روش برای انتخاب نمونه‌ای از سه مرکز تلفن استفاده کنید. جزئیات چگونگی انتخاب این نمونه را نشان دهید.

۳.۱۱ از جامعهٔ بخشهایی که در تمرین ۱٦.۱۰، نشان داده شد، یک نمونه با احتمال متناسب با اندازه و با جایگذاری، متشکل از ۵ بخش گرفته شده است و برای هر انتخاب بخش در مرحلهٔ اول، نمونه‌ای از یک بیمارستان در مرحلهٔ دوم گرفته شده است. نمونه‌گیری به شرح زیر اجرا می‌شود:

الف. شیوهٔ انتخاب نمونهٔ بخشها با احتمال متناسب با اندازه همان است که در بخش ۳.۳.۱۱ توصیف شده است.

ب. معیار متغیر اندازه برای نمونه‌گیری از یک بخش، کل تعداد بیمارستانهای موجود در آن بخش است.

پ. اعداد تصادفی که برای انتخاب نمونه انتخاب شده‌اند به شرح زیرند:

دنبالهٔ انتخاب	عدد تصادفی برای انتخاب بخشها در مرحلهٔ اول	عدد تصادفی برای انتخاب بیمارستان در مرحلهٔ دوم
۱	۳۸	۱
۲	۲٤	۱
۳	۲٤	۱
٤	٤۲	۳
٥	۰۸	۱

با استفاده از اطلاعات بالا، بخشها و بیمارستانهای نمونه را تعیین کند.

٤.۱۱ از نمونه‌ای که در تمرین قبل انتخاب شد، کل تعداد پذیرشهای بیمارستانی را در ۳۷ ناحیهٔ بخش طی سال ۱۹۸۹ تعیین کنید. همچنین، خطای معیار این برآوردگر مجموع را تعیین کنید.

٥.۱۱ (برای کسانی که به نرم‌افزارهایی از قبیل STATA یا SUDAAN دسترسی دارند.) کل تعداد پذیرشها را به ازای هر تخت از روی نمونهٔ انتخاب شده در تمرین ۳.۱۱ برآورد کنید. خطای معیار این برآوردگر را نیز تعیین کنید.

کتابشناسی

The text by Brewer referenced below gives a very comprehensive treatment of sampling designs and estimators based on unequal probability sampling.

1. Brewer, K. R. W., and Hanif, M., *Sampling with Unequal Probabilities,* Springer - Verlag, New York, 1983.

The following recent texts give excellent discussions of the Horvitz-Thompson and Hansen-Hurwitz estimators and of the various types of PPS Sampling

2. Hedayat, A. S., and Sinha, B. K., *Design and Inference in Finite Population Sampling,* Wiley, New York, 1991.
3. Lehtonen, R., and Pahkinen, E. J., *Practical Methods for Design and Analysis of Complex Surveys,* Rev. Ed., Wiley, Chichester, U.K., 1997.
4. Thompson, S. K., *Sampling,* Wiley, New York, 1992.

The following expository review in the Enclyclopedia of Statistical Sciences *gives an excellent overview of PPS sampling and contains a long list of useful references on the topic.*

5. Skinner, C. J., Probability proportional to size (PPS) sampling. In *The Encyclopedia of Statistical Sciences,* Johnson, N., and Kotz, S., Eds., Wiley, New York, 1983.

The following entry in the Encyclopedia of Biostatistics *provides a very detailed discussion of practical aspects of PPS sampling including its strengths and weaknesses.*

6. Czaja, R., Sampling with probability proportionate to size. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.

The two original papers on the Hansen-Hurwitz and Horvitz-Thompson estimators are referenced below.

7. Hansen, M. H., and Hurwitz, W. N., On the theory of sampling from finite population. *Annals of Mathematical Statistics,* 14: 333-362, 1943.
8. Horvitz, D. G., and Thompson, D. J., A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47: 663-685, 1952.

The following two articles are the seminal papers on the Mitofsky-Waksberg method of telephone sampling.

9. Mitofsky, W., *Sampling of Telephone Households* (Unpublished CBS Memorandum), 1970.
10. Waksberg, J., Sampling methods for random digit dialing. *Journal of the American Statistical Association,* 73: *40-46,* 1978.

فصل ۱۲

برآورد کردن واریانس در آمارگیریهای نمونه‌ای پیچیده

در فصلهای قبل، روشهای برآورد واریانسهای مشخصه‌های برآورد شدهٔ بیشتر طرحهای نمونه‌گیری گوناگون را که مورد بحث قرار گرفته بودند معرفی کردیم. در بسیاری از این طرحها، واریانس نظری برآوردگری ویژه، تابعی خطی از پارامترهای جامعه است. مثلاً در نمونه‌گیری تصادفی ساده، واریانس یک مجموع برآورد شده، x'، از فرمول زیر به دست می‌آید

$$Var(x') = \frac{N^2}{n}\sigma_x^2\left(\frac{N-n}{N-1}\right)$$

واضح است که چون عبارت بالا تابعی خطی از واریانس جامعه‌ای σ_x^2 است با جایگزین کردن برآورد نااریب $\hat{\sigma}_x^2$ از σ_x^2 در عبارت مربوط به $Var(x')$ که در فصل ۲ ارائه شد، می‌توان برآوردی نااریب به دست آورد.

ولی در بسیاری از طرحهای نمونه‌ای که عملاً به کار برده می‌شوند، فرایند برآورد کردن ممکن است مستلزم طبقه‌بندی، چندین مرحله نمونه‌گیری خوشه‌ای، برآورد نسبتی یا رگرسیونی، پس طبقه‌بندی برای مجموعهای معلوم، و سایر شیوه‌ها باشد و به این نتیجه بینجامد که واریانس برآورد حاصل، یک تابع خطی یا حتی تابعی معلوم از پارامترهای جامعه نباشد. این، در مورد بسیاری

از آمارگیـریهای نمونه‌ای که به وسیلهٔ ادارهٔ سرشماری امریکا[1]، مرکز ملی آمارهای بهداشتی[2]، دفتر آمار کار[3]، و سایر سازمانها اجرا می‌شود مصداق دارد.

بـرای برآورد کردن واریانسهای برآوردهای حاصل از آمارگیریهای نمونه‌ای که مستلزم نمونه‌گیریها و شـیوه‌هـای پیچیـده بـرای بـرآورد کـردن هسـتند ضروری است یکی از دو ردهٔ کلی از روشهایی که مخصوصاً برای این منظور بسط داده شده‌اند، یعنی روشهای *خطی‌سازی* و *تکرار*، به کار برده شود.

فـن *خطـی‌سـازی* بـه وسـیلهٔ کـی‌فیتس [۱] ، وودراف [۲] و دیگران برای آمارگیریهای نمونه‌ای و براســـاس تقـریب سـری تیـلور[4] بسـط داده شده است. در اواخر دههٔ ۱۹۷۰، پژوهشـــگران مؤسسـهٔ مثـلث تحقیـق[5] شـروع بـه ابـداع مطلـبی کـردند کـه سـرانجام تـبدیل به یک نرم‌افزار راحت براساس خطـی‌سازی شد [۳]. این نرم‌افزار، شکل نخستین SUDAAN بود که پس از تکامل بسیار به نرم‌افزاری بـدل شـد کـه کاربـرد وسیعی دارد و در سراسر این کتاب مورد اشاره و بحث بوده است. قابل استفاده بـودن و جاذبـهٔ ایـن نـرم‌افـزار و سـایر نـرم‌افـزارها بـاعث شـده اسـت کـه خطـی‌سازی شاید اکنون پراستفاده‌ترین روش برآورد واریانس برای آمارگیریهای پیچیده باشد.

تکـرار نیـز یـک ردهٔ کـلی از روشـهایی اسـت که با آن برآوردی از واریانس یک پارامتر جامعه‌ای بـرآورد شـده بـه دسـت مـی‌آیـد، به این ترتیب که پارامتر جامعه‌ای برآورد شده به صورت مجموع یا میانگیـنی از چـندین آماره بیان می‌شود که هر یک بر زیرمجموعه‌ای از مشاهدات نمونه‌ای متکی است. واریـانس پارامـتر جامعـه‌ای بـرآورد شده را سپس می‌توان با به دست آوردن برآوردی از واریانس این آمـاره‌هـای «جزء نمونه‌ای» برآورد کرد. مثال ساده‌ای از این فن در فصل ٤، با توجه به برآورد واریانس یـک مشخصهٔ جامعه‌ای برآورد شده، تحت نمونه‌گیری سیستماتیک مکرر نشان داده شد. گـونه‌هایی از این رده از فنون طی چندین سال مورد استفاده قرار گرفته است. هنسن و همکاران [۵] در کتاب درسی خـود کـه در اوایـل دهـهٔ ۱۹۵۰ تألیف کردند به این فنون تحت عنوان *روشـهای گروه تصـادفی* اشاره کـرده‌اند. این روشها در دههٔ ۱۹٦۰ توسـط مـک‌کارتی [٦] و [۷] دقیقتر شده و در سطح گسترده‌ای در مرکـز مـلی آمارهـای بهداشـتی و سـایر سـازمانهای فـدرال مورد استفاده قرار گرفتند. مطالعاتی توسط فرانکل [۸] و بین [۹] انجام گرفته و نشریاتی متعدد به وسیلهٔ لِمی‌شو و همکاران [۱۰] تا [۱٦] به رشتهٔ تحریر درآمده‌اند که سعی در ارزشیابی این روشها داشته‌اند.

[1] U.S. Bureau of the Census
[2] National Center for Health Statistics
[3] Bureau of Labor Statistics
[4] Taylor series approximation
[5] Research Triangle Institute

هـدف از ایـن فصل، معرفی منطق نهفته در پس الگوریتمهای مورد استفاده در بسته‌های نرم‌افزاری گونـاگونی اسـت که در حال حاضر برای برآورد واریانسهای برآوردهای حاصل از آمارگیریهای دارای طـرحهای پیچیـده در دسـترس‌انـد. در بحـثی کـه بـه دنـبال خواهـد آمـد، توجـه خود را بر روشهای خطـی‌سـازی و تکرار متمرکز می کنیم که فعلاً پراستفاده‌ترین روشهای برآورد کردن واریانس به شمار می‌روند.

۱.۱۲ خطی‌سازی

در آمارگیریهای نمونه‌ای پیچیده، برآورد کردن واریانسهای مشخصه‌های جامعه‌ای برآورد شده می‌تواند مشـکل باشـد چه به علت راهی که نمونه با آن راه گرفته شده است و چه به دلیل راهی که برآوردهای مشخصـه‌هـای جامعـه‌ای بـا آن سـاخته شده‌اند. مثلاً، یک مجموع برآورد شده براساس چندین مرحله نمونـه‌گیـری خوشـه‌ای، تعدیـل نسـبتی، و پـس طـبقه‌بندی غالباً ممکن است به راحتی یک برآوردگر مناسـب واریـانس را بـه دسـت ندهـد. در چـنین وضـعیتهایی می‌توان از فن خطی‌سازی برای ساختن تقریـبی بـرای شـکل تـابعی مشخصـهٔ جامعـه‌ای برآورد شده استفاده کرد که تابعی خطی از مشاهدات اصلی است و لذا برای ساختن برآوردگر واریانس مناسب است.

بحـث روش‌شناسـی خطـی‌سـازی که در پی خواهد آمد اساساً گزیده و فشردهٔ مطالبی است که در پیوسـت فـنی کـتابچهٔ راهـنمای SUDAAN ارائـه شـده اسـت کـه همراه با نسخهٔ رایانهٔ شخصی این محصول نرم‌افزاری است [۳].

بـرای سـهولت کـار، فـرض مـی‌کنیم جامعه‌ای داریم که از مشاهدات x_i و y_i دربارهٔ i امین واحد شمارش تشکیل شده است و باز فرض می‌کنیم می‌خواهیم، θ، نوعی مشخصهٔ جامعه‌ای را برآورد کنیم و یـک برآوردگر، $\hat{\theta}$، هم وجود دارد که تابع $f(x,y)$ از متغیرهای x و y است که هر دو تابعی خطی از مشاهدات نمونه‌ای هستند. ولی تابع $f(x,y)$ تابع خطی مشاهدات نمونه‌ای نیست.

بـرای مـثـال، $\hat{\theta}$ ممـکن است $\frac{\bar{x}}{\bar{y}}$، نسبت میانگینهای نمونه‌ای، $\bar{x}$ و $\bar{y}$، باشد. واضح است که $\bar{x}$ و $\bar{y}$ تابعهـای خطـی مشاهدات نمونه‌ای هستند در حالی که $f(\bar{x},\bar{y})$ تابع غیرخطی دو میانگین نمونه‌ای است و از ا ین رو است که در مشاهدات نمونه‌ای نیز غیرخطی است.

محاسـبهٔ واریـانس $\hat{\theta}$ بـا روش خطـی‌سـازی، شـیوه‌ای دومـرحله‌ای اسـت. مرحـلهٔ اول یا مرحلهٔ خطـی‌سـازی مسـتلزم تقریب کردن $f(x,y)$ به وسیلهٔ سری تیلور مرتبهٔ اول است. این مرحله به تقریبی مـی‌انجامد که نسبت به آماره‌های نمونه‌ای $\bar{x}$ و $\bar{y}$، خطی است و به همین علت در مشاهدات نمونه‌ای نیـز خطـی خواهـد بود. همین که این تقریب سری تیلور مرتبهٔ اول برای $\hat{\theta}$ به دست آمد، مرحـلهٔ دوم

متضمن استفاده از روشهای مبتنی بر طرح، از قبیل روشهای توصیف شده در فصلهای ۳ تا ۱۱، برای برآورد واریانس آن خواهد بود. در زیر توضیح می‌دهیم که چگونه این کار به صورت مکانیکی انجام می‌پذیرد.

z_i را که *مقدار خطی شدهٔ* $f(\bar{x},\bar{y})$ برای مشاهدهٔ i نامیده می‌شود به صورت زیر تعریف می‌کنیم:

$$z_i = (\partial f_{\bar{X}})x_i + (\partial f_{\bar{Y}})y_i$$

که در آن

$$\partial f_{\bar{X}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{x}}, \quad \text{محاسبه شده در نقطهٔ } (\bar{X},\bar{Y})$$

و

$$\partial f_{\bar{Y}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{y}}, \quad \text{محاسبه شده در نقطهٔ } (\bar{X},\bar{Y})$$

(که در آنها $\bar{X}$ و $\bar{Y}$ پارامترهای جامعه‌ای نامعلوم‌اند که $\bar{x}$ و $\bar{y}$ برآوردگرهای نااریب آنها هستند).

مثلاً اگر $\hat{\theta} = \bar{x}/\bar{y}$ باشد، آن‌گاه

$$\partial f_{\bar{X}} = \frac{1}{\bar{Y}}, \qquad \partial f_{\bar{Y}} = -\frac{\bar{X}}{\bar{Y}^2}, \qquad z_i = \frac{1}{\bar{Y}}x_i - \frac{\bar{X}}{\bar{Y}^2}y_i$$

بسط سری تیلور مرتبهٔ اول برای $\hat{\theta}$ به صورت زیر است

$$\begin{aligned}\hat{\theta} = f(\bar{x},\bar{y}) &\approx (\partial f_{\bar{X}})(\bar{x}-\bar{X}) + (\partial f_{\bar{Y}})(\bar{y}-\bar{Y}) \\ &= (\partial f_{\bar{X}})\bar{x} + (\partial f_{\bar{Y}})\bar{y} - ((\partial f_{\bar{X}})\bar{X} + (\partial f_{\bar{Y}})\bar{Y})\end{aligned}$$

و

$$Var(\hat{\theta}) \cong Var((\partial f_{\bar{X}})\bar{x} + (\partial f_{\bar{Y}})\bar{y})$$

زیرا عبارت $((\partial f_{\bar{X}})\bar{X} + (\partial f_{\bar{Y}})\bar{Y})$ یک مقدار ثابت است. ولی

$$(\partial f_{\bar{X}})\bar{x} + (\partial f_{\bar{Y}})\bar{y} = (\partial f_{\bar{X}})\sum_{i=1}^{n}\frac{x_i}{n} + (\partial f_{\bar{Y}})\sum_{i=1}^{n}\frac{y_i}{n} = \sum_{i=1}^{n}\frac{z_i}{n} = \bar{z}$$

به این ترتیب، مشکل یافتن واریانس تقریبی $\hat{\theta}$ به مسئلهٔ تعیین واریانس یک ترکیب خطی از مشاهدات نمونه‌ای تبدیل می‌شود و می‌تواند با تعیین طرح نمونه‌ای و استفاده از فرمول مناسب به دست آید. مثلاً اگر طرح، نمونه‌گیری تصادفی ساده باشد، در آن صورت، واریانس $\hat{\theta}$ به صورت زیر برآورد می‌شود

$$\hat{Var}(\hat{\theta}) = \frac{\sum_{i=1}^{n}(\tilde{z}_i - \bar{\tilde{z}})^2}{n(n-1)}\left(\frac{N-n}{N}\right) \qquad (۱.۱۲)$$

که در آن

$$\tilde{z}_i = (\partial f_{\bar{x}})x_i + (\partial f_{\bar{y}})y_i$$

$$\partial f_{\bar{x}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{x}}, \quad \text{محاسبه شده در نقطهٔ } (\bar{x},\bar{y})$$

و

$$\partial f_{\bar{y}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{y}}, \quad \text{محاسبه شده در نقطهٔ } (\bar{x},\bar{y})$$

(زیرا $\bar{X}$ و $\bar{Y}$ پارامترهای مجهول‌اند).

مثال تشریحی: یک مثال بسیار کوچک ساده را انتخاب می‌کنیم تا نشان دهیم خطی‌سازی دقیقاً چگونه عمل می‌کند. فرض کنید هزینه‌های بیمهٔ کمکهای پزشکی را برای یک بیمار حسابرسی می‌کنیم و یک نمونهٔ تصادفی ساده متشکل از ۱۰ درخواست از مجموع ٦٥ درخواست حق بیمه را که از جانب بیمار تنظیم شده است انتخاب کرده‌ایم. داده‌های حاصل از این حسابرسی در جدول ۱.۱۲ نشان داده شده‌اند. می‌خواهیم نسبت مجموع دلارهای اضافی را که از طرف این بیمار پرداخت شده است برآورد کنیم.

جدول ۱.۱۲ پرداختها و اضافه پرداختهای بیمهٔ کمکهای پزشکی مربوط به ۱۰ درخواست حق بیمه که برای بیمار صادر شده است

درخواست حق بیمه	پرداخت (y) به دلار	اضافه پرداخت (x) به دلار	مشاهدهٔ خطی شده $\tilde{z}_i$
۱	۲۱۰	۲۱۰	۰/٤٦٧٥
۲	۷۸	۰	−۰/۱۰۹٤
۳	۳٤۳	۱۲۳	−۰/۰۳٤٦
٤	۲۹۸	۱٥۷	۰/۱٥۱۸
٥	۳٤۹	۰	−۰/٤۸۹٤
٦	۲۱۰	۲۱۰	۰/٤٦٧٥
۷	٥۳٦	۰	−۰/۷٥۱٦
۸	۲۸۹	۱۳٥	۰/۰۸٤٦
۹	۹۸	۰	−۰/۱۳۷٤
۱۰	۳٤٥	۲۳۰	۰/۳٥۰۸

از روی این داده‌ها داریم $\bar{x} = ۱۰۶/۵$ و $\bar{y} = ۲۷۵/۶$. چون از برآورد نسبتی استفاده می‌کنیم، متغیر خطی شدهٔ

$$\tilde{z}_i = \frac{۱}{۲۷۵/۶} x_i - \frac{۱۰۶/۵}{(۲۷۵/۶)^۲} y_i$$

به صورتی که در بالا توصیف شد به دست خواهد آمد. مثلاً

$$\tilde{z}_i = \frac{۱}{۲۷۵/۶} \times ۲۱۰ - \frac{۱۰۶/۵}{(۲۷۵/۶)^۲} \times ۲۱۰ = ۰/۴۶۷۵$$

$\tilde{z}_i$ ها در ستون چهارم جدول ۱.۱۲ نشان داده شده‌اند. با $N = ۶۵$ و $n = ۱۰$، خطای معیار برآورد نسبتی، r، را از برابری (۱.۱۲) محاسبه می‌کنیم که برابر است با ۰/۱۱۵۸.

□

از SUDAAN و STATA، هر دو، برای برآورد کردن واریانسهای آماره‌های برآورد شده از خطی‌سازی استفاده می‌کنند. برای استفاده از هر یک از این نرم‌افزارها لازم است متغیرهای $N(=۶۵)$ و $w(=\frac{۶۵}{۱۰} = ۶/۵)$ را به هر یک از سابقه‌ها در مجموعه داده‌های جدول ۱.۱۲ اضافه کنیم. به متغیر $\tilde{z}_i$ نیاز نخواهیم داشت، زیرا توسط برنامه محاسبه خواهد شد. مجموعهٔ داده‌های حاصل برای استفاده به وسیلهٔ SUDAAN یا STATA در زیر نشان داده می‌شود:

claim	payment	ovpaymnt	*N*	*w*
1	210	210	65	6.5
2	78	0	65	6.5
3	343	123	65	6.5
4	298	157	65	6.5
5	349	0	65	6.5
6	210	210	65	6.5
7	536	0	65	6.5
8	289	135	65	6.5
9	98	0	65	6.5
10	345	230	65	6.5

فرمانهای زیر برای برآورد کردن نسبت x/y و خطای معیار آن توسط STATA به کار می‌روند:

```
use"a: \exmp12_2.dta", clear
. svyset fpc N
. svyset pweight w
. svyratio ovpaymnt/payment
```

این فرمانها، خروجی زیر را تولید می‌کنند:

Survey ratio estimation

pweight: W — Number of obs = 10
Strata: < one > — Number of strata = 1
PSU: <observations> — Number of PSUs = 10
FPC: N — Population size = 65

Ratio	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
ovpaymnt/payment	.3864296	.1158187	.1244294	.6484298	1

تصحیح جامعهٔ متناهی (FPC)، با این فرض است که نمونه‌گیری تصادفی سادهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه (PSU) در داخل هر طبقه، بدون نمونه‌گیری فرعی در داخل واحدهای نمونه‌گیری اولیه، انجام می‌گیرد.

فرمانهای زیر برای برآورد کردن با SUDAAN به کار می‌روند:

```
1  PROC RATIO DATA = EXMP12_2  FILETYPE = SAS  DESIGN = WOR;
2  NEST_ONE_;
3  WEIGHT  W;
4  TOTCNT  N;
5  NUMER  OVPAYMNT;
6  DENOM  PAYMENT;
7  SETENV  COLWIDTH = 15;
8  SETENV  DECWIDTH = 3;
```

این فرمانها، خروجی زیر را تولید می‌کنند:

Number of observations read : 10 Weighted count : 65
Number of observations skipped : 0
(WEIGHT variable nonpositive)
Denominator degrees of freedom : 9

by: Variable, One.

Variable		One 1
OVPAYMNT/PAYMENT	Sample Size	10.000
	Weighted Size	65.000
	Weighted X-sum	17914.000
	Weighted Y-sum	6922.500
	Ratio Est.	0.386
	SE Ratio	0.116

۲.۱۲ روشهای تکرار

روشهای تکرار (که گاهی *روشهای باز نمونه‌گیری* خوانده می‌شوند) واریانس نمونه‌گیری یک آماره را با محاسبهٔ آن آماره برای زیرمجموعه‌ای از نمونه و بررسی تغییرپذیری آن روی مجموعه‌ها برآورد می‌کنند. دو رهیافت کلی برای تکرار که طی سه دههٔ گذشته بسط داده شده‌اند، رهیافت جک‌نایف[1] و رهیافت *تکرار مکرر متعادل*[2] (که به نام رهیافت *نیمه نمونه‌ای متعادل*[3] نیز خوانده می‌شود) هستند. در این بحث از فنون تکرار، تمرکز ما منحصراً بر روشهای تکرار مکرر متعادل (BRR) است. زیرا از لحاظ تاریخی بیشتر از روشهای جک‌نایف در کارهای آمارگیری نمونه‌ای مورد استفاده بوده‌اند.

۱.۲.۱۲ روش تکرار مکرر متعادل

روش تکرار مکرر متعادل بیشترین کاربرد را در رده‌ای از طرحهای نمونه‌ای دارد که در سطح گسترده‌ای مورد استفاده قرار می‌گیرند و در آنها واحدهای نمونه‌گیری اولیه (PSU ها) در L طبقه گروه‌بندی می‌شوند، از هر طبقه، نمونه‌هایی متشکل از دو واحد نمونه‌گیری اولیه انتخاب می‌شود، و از هر واحد نمونه‌گیری اولیه، برآوردهایی مستقل از مشخصه‌های طبقه به دست می‌آید.

حال، روش تکرار مکرر متعادل را با نشان دادن چگونگی برآورد کردن واریانس یک برآورد نسبتی با استفاده از این شیوه به صورتی دقیقتر نشان می‌دهیم. فرض کنید x'_{h1}، x'_{h2}، y'_{h1} و y'_{h2} برآوردهایی از هر یک از دو واحد نمونه‌گیری اولیه برای پارامترهای طبقه‌ای X_h و Y_h در هر یک از L طبقه هستند. (برای این بحث فرض می‌کنیم که طبقه‌ها برای تعیین برآوردهای جامعه‌ای دارای وزن برابرند.) پس، نتیجه می‌گیریم که برآورد نسبتی r برای نسبت جامعه‌ای R بر مبنای همهٔ $2L$ برآورد از فرمول زیر به دست می‌آید

$$r = \frac{x'}{y'}$$

که در آن

$$x' = \sum_{h=1}^{L} \sum_{i=1}^{2} (x'_{hi}/2)$$

و

$$y' = \sum_{h=1}^{L} \sum_{i=1}^{2} (y'_{hi}/2)$$

[1] Jackknifing Approach
[2] Balanced Repeated Replication Approach (BRR)
[3] Balanced Half-sample Approach

برآورد نیمه نمونه‌ای نسبت R را به صورت زیر تعریف می‌کنیم:

$$r_{(k)} = \frac{\sum_{h=1}^{L}[\delta_{kh}x'_{h1} + (1-\delta_{kh})x'_{h2}]}{\sum_{h=1}^{L}[\delta_{kh}y'_{h1} + (1-\delta_{kh})y'_{h2}]}$$

که در آن $(\delta_{k1},\ldots,\delta_{kL})$ یک بردار L ـ بعدی است که عناصر آن برابر با یک یا صفرند. به عبارت دیگر، هر $r_{(k)}$ برآوردی از R است که با گرفتن یکی از دو برآورد از هر طبقه تشکیل شده است. واضح است که 2^L بردار ممکن (δ_k) به صورتی که در بالا نشان داده شده است وجود دارند و از این رو، در کل 2^L برآورد نیمه نمونه‌ای ممکن $r_{(k)}$ موجودند. برآوردی مناسب از واریانس r که از روی 2^L نیمه نمونهٔ ممکن محاسبه شده باشد چنین خواهد بود

$$\hat{Var}(r) = \left(\frac{1}{2^L}\right)\sum_{k=1}^{2^L}(r_{(k)} - \bar{r})^2$$

معمولاً 2^L به قدری بزرگ است که محاسبهٔ همهٔ $r_{(k)}$ های ممکن عملی نیست. زیرمجموعه‌ای از K برآورد نیمه نمونه‌ای را (که در آن $K \le 2^L$) می‌توان با به دست آوردن δ_{ij} از روی ماتریس متعامد از نوعی که پلاکت و برمن توصیف کرده‌اند [۱۷] ساخت. جزئیات این شیوه توسط مک‌کارتی [۶] شرح و بسط داده شده است که نشان می‌دهد این امکان وجود دارد که زیرمجموعهٔ کوچکی از نیمه نمونه‌ها با روش بسیار دقیقی انتخاب شود و با وجود این برآوردهای رضایتبخشی از واریانس به دست آید. در واقع، او نشان داده است که استفاده از زیرمجموعه‌ای کوچک برای برآوردهای خطی، از قبیل میانگینها و مجموعها، دقیقاً همان برآوردی از واریانس را نتیجه می‌دهد که از همهٔ 2^L نیمه نمونه به دست خواهد آمد. مجموعهٔ برآوردهای نیمه نمونه‌ای حاصل را *متعادل* می‌نامند زیرا نشان داده شده است که سهم آنها در واریانس بین طبقه‌ای خنثی می‌شود. فرض کنید مجموعهٔ متعادلی شامل K برآورد نیمه نمونه‌ای از این نوع را در نظر می‌گیریم و برآورد نهایی $\bar{r}$ خود را به صورت زیر تعریف می‌کنیم:

$$\bar{r} = \frac{1}{K}\sum_{k=1}^{K} r_{(k)} \qquad (۲.۱۲)$$

برآوردگر واریانس این برآورد نسبتی $\bar{r}$ از فرمول زیر به دست می‌آید

$$\hat{Var}(\bar{r}) = \left(\frac{1}{K}\right)\sum_{k=1}^{K}(r_{(k)} - \bar{r})^2 \qquad (۳.۱۲)$$

منطق و خواص آماری این برآوردگر واریانسی نیمه نمونه‌ای متعادل توسط مک‌کارتی [٦] و بین [٩] و نیز لِمی‌شو و لِه‌وی [١٠] به تفصیل شرح داده شده‌اند. ما نیز برای این برآوردگر واریانسی، منطقی رهگشا در پیوست فنی این فصل ارائه می‌کنیم و حالا استفاده از این روش را با مثال ساده‌ای نمایش می‌دهیم.

مثال تشریحی: شهری به سه ناحیهٔ خدمات فوریتهای پزشکی[1] تقسیم شده که هر ناحیه شامل پنج ایستگاه آمبولانس است. می‌خواهیم نسبت کسانی را که در کل شهر دچار ایست قلبی شده و پس از دیده شدن توسط پیراپزشکان یکی از ۱۵ ایستگاه آمبولانس توصیف شده در بالا زنده وارد بیمارستان شده‌اند برآورد کنیم. یک آمارگیری نمونه‌ای اجرا شده است که طی آن دو ایستگاه آمبولانس به طور تصادفی از هر یک از این سه ناحیهٔ خدمات اورژانسی نمونه‌گیری شده‌اند. سوابق موجود در هر یک از این شش ایستگاه آمبولانس نمونه‌گیری شده مورد بررسی قرار گرفته و برآوردهایی از تعداد ایستهای قلبی و تعداد و نسبت بیمارانی که دچار ایست قلبی شده و زنده به بیمارستان رسیده‌اند برای ناحیهٔ خدمات اورژانسی تهیه شده است. نتایج این آمارگیری نمونه‌ای در زیر آمده‌اند:

ناحیهٔ خدمات فوریتهای پزشکی (ESA)	ایستگاه آمبولانس	برآورد تعداد ایستهای قلبی برای ESA y'_{hi}	برآورد تعداد بیماران دچار ایست قلبی که زنده به بیمارستان رسیده‌اند برای ESA x'_{hi}
۱	۱	۱۲۰	۲۵
	۲	۷۸	۲٤
۲	۱	۱۸۵	۳۰
	۲	۲۲۸	٤۹
۳	۱	٦۷۰	۸۰
	۲	۵۳۰	۷۰

باید تأکید شود که x'_{hi} و y'_{hi} که در بالا نشان داده شده‌اند برآوردهایی از تعداد ایستهای قلبی و تعداد بیمارانی که زنده به بیمارستان رسیده‌اند براساس داده‌های حاصل از ایستگاه ویژهٔ آمبولانس هستند.

[1] Emergency Service Area (ESA)

واضح است که طرح این آمارگیری نمونه‌ای از طبقه‌بندی بر حسب ناحیهٔ خدمات فوریتهای پزشکی و برآورد کردن برای هر ناحیهٔ خدمات اورژانسی از روی دو ایستگاه آمبولانس که به طور تصادفی انتخاب شده‌اند استفاده می‌کند.

تکرارهای ۸ = $۲^۳$ نیمه نمونهٔ زیر را می‌توان برای سه طبقه تهیه کرد:

تکرار(k)	$\delta_{k۱}$	$۱-\delta_{k۱}$	$\delta_{k۲}$	$۱-\delta_{k۲}$	$\delta_{k۳}$	$۱-\delta_{k۳}$
۱	۱	۰	۱	۰	۱	۰
۲	۱	۰	۱	۰	۰	۱
۳	۱	۰	۰	۱	۱	۰
٤	۱	۰	۰	۱	۰	۱
٥	۰	۱	۱	۰	۱	۰
٦	۰	۱	۱	۰	۰	۱
۷	۰	۱	۰	۱	۱	۰
۸	۰	۱	۰	۱	۰	۱

مثلاً تکرار ۱، مقدار $r_{(k)}$ را به صورت زیر ارائه می‌دهد :

$$r_{(۱)}=\frac{\sum_{h=۱}^{۳}\left[\delta_{ih}x'_{h۱}+(۱-\delta_{ih})x'_{h۲}\right]}{\sum_{h=۱}^{۳}\left[\delta_{ih}y'_{h۱}+(۱-\delta_{ih})y'_{h۲}\right]}=\frac{۱۳۵}{۹۷۵}=۰/۱۳۸۵$$

هشت تکرار، مقادیر زیر از $r_{(k)}$ را به دست می‌دهند:

k	$x'_{(k)}$	$y'_{(k)}$	$r_{(k)}$
۱	۱۳۵	۹۷۵	۰/۱۳۸۵
۲	۱۲۵	۸۳۵	۰/۱٤۹۷
۳	۱٥٤	۱۰۱۸	۰/۱٥۱۳
٤	۱٤٤	۸۷۸	۰/۱٦٤۰
٥	۱۳٤	۹۳۳	۰/۱٤۳٦
٦	۱۲٤	۷۹۳	۰/۱٥٦٤
۷	۱٥۳	۹۷٦	۰/۱٥٦۸
۸	۱٤۳	۸۳٦	۰/۱۷۱۱

از داده‌هایی که در بالا نشان داده شده‌اند پی می‌بریم که برآورد نسبت $\bar{r}$ و واریانس برآورد شدهٔ آن، $\hat{Var}(\bar{r})$، براساس هشت تکرار به صورت زیر به دست می‌آید

$$\bar{r} = \frac{1}{8}\sum_{k=1}^{8} r_{(k)} = ۰/۱۵۳۹$$

و

$$\hat{Var}(\bar{r}) = ۰/۰۰۰۰۹۸$$

پلاکت و برمن [۱۷]، ماتریسهای $m \times m$ را با ابعاد $m = \{4,8,12,\ldots,100\}$ به استثنای $m = 92$ ارائه داده‌اند. بعد ماتریس مورد استفاده در یک مسئلهٔ خاص به تعداد طبقه‌ها بستگی دارد. مک‌کارتی [۶] از ماتریسهایی استفاده کرده است که در آن m مضربی از ۴ و $L \le m \le L+3$، که در آن در همهٔ موارد $L > 2$، $m << 2^L$ است. همین که ماتریس $m \times m$ انتخاب شد همهٔ سطرها بجز تنها اولین ستونهای L مورد استفاده قرار می‌گیرند.

یک مجموعهٔ *متعادل* از چهار تکرار را می‌توان از مجموعهٔ هشت تکرار ممکن با اختیار کردن تکرارهای ۱، ۴، ۶ و ۷ به شرح زیر به دست آورد:

تکرار k	δ_{k1}	$1-\delta_{k1}$	δ_{k2}	$1-\delta_{k2}$	δ_{k3}	$1-\delta_{k3}$
۱	۱	۰	۱	۰	۱	۰
۴	۱	۰	۰	۱	۰	۱
۶	۰	۱	۱	۰	۰	۱
۷	۰	۱	۰	۱	۱	۰

توجه کنید که در این مجموعه، هر برآورد طبقه‌ای x'_{hk} یا y'_{hk} به دفعات برابر ظاهر می‌شود (دو بار در این مثال کوچک)، و با هر برآورد دیگر به تعداد دفعات یکسان (در این مثال یک بار) پدیدار می‌شود. حال برآورد $\bar{r}$ و واریانس آن $\hat{Var}(\bar{r})$ را برای این مجموعهٔ چهار تکرار بررسی می‌کنیم:

$$\bar{r} = \frac{(r_{(1)} + r_{(4)} + r_{(6)} + r_{(7)})}{4} = ۰/۱۵۳۹$$

و

$$\hat{Var}(\bar{r}) = ۰/۰۰۰۰۸۸$$

□

چون هر برآورد طبقه‌ای نیمه نمونه‌ای با تعداد دفعاتی یکسان در مجموعهٔ K نیمه نمونهٔ متعادل ظاهر می‌شود، برآورد $\bar{r}$ با برآورد حاصل از مجموعهٔ کامل ۲^L نیمه نمونه برابر خواهد بود. همچنین، مک‌کارتی [٦] نشان داده است که پایداری برآوردگر واریانسی حاصل از مجموعه‌ای از تکـرارهای نیمه نمونه‌ای متعادل تقریباً همان پایداری برآوردگر واریانسی است که از مجموعهٔ کامل ۲^L تکـرار نیمه نمونه‌ای ساخته می‌شود.

باید توجه داشت که غالباً آمارگیریهایی طراحی می‌شوند که در آنها $۲L$ طبقه وجـــود دارند و از هـر طبقه یک واحد نمونه‌گیری اولیه (PSU) انتخاب می‌شود. برای به کار بردن فن تکرار مکرر متعادل (BRR) واحدهای نمونه‌گیری اولیه، جفت جفت طبقه‌بندی می‌شوند تا L «شبه طبقه» تشکیل شود و فرایند برآورد کردن به همان صورتی که قبلاً شرح داده شد اجرا می‌شود. معمولاً در جفت کردن واحدها دقت می‌شود تا واحدهای نمونه‌گیری اولیه از نظر اطلاعات جمعیت‌شناختی معلوم، حتی‌الامکان شبیه باشند.

استفاده از SUDAAN برای به دست آوردن برآوردهای تکرار مکرر متعادل(BRR) : آخرین نسخهٔ SUDAAN در زمان نگارش متن حاضر (Release 7.5) می‌تواند برآوردهایی از خطاهای معیار را با استفاده از روش تکرار مکرر متعادل (BRR) و نیز برآوردهایی از راه خطی‌سازی به دست آورد. در مثال بالا که در آن از مجموعه‌ای متعادل با چهار تکرار از هشت تکرار ممکن برای برآورد کردن نسبت بیماران دچار ایست قلبی که سالم به بیمارستان رسیده بودند استفاده شده بود، مجموعه داده‌های مورد نیاز برای تحلیل توسط SUDAAN، در زیر نشان داده شده است:

ESA	AMBSTAT	CARDARRS	ALIVE	WT	REPWT1	REPWT4	REPWT6	REPWT7
1	1	120	25	2.5	5	5	0	0
1	2	78	24	2.5	0	0	5	5
2	1	185	30	2.5	5	0	5	0
2	2	228	49	2.5	0	5	0	5
3	1	670	80	2.5	5	0	0	5
3	2	530	70	2.5	0	5	5	0

توجه کنید که در این پرونده شش سابقه، یعنی یک سابقه برای هر نقطهٔ نمونه، و ۹ متغیر وجود دارد. دو متغیر اول، یعنی *ESA* و *AMBSTAT* نشانهٔ ناحیهٔ خدمات اورژانسی و ایستگاه آمبولانس است. دو متغیر بعدی، *CARDARRS* و *ALIVE* به ترتیب نشانهٔ تعداد ایستهای قلبی که در هر ایستگاه آمبولانس نمـونه رسیـدگی شده و تعداد این قبیل بیماران است که زنده به بیمارستان رسیده‌اند. متغیر *WT* وزن نمونه، N/n، است که به هر ایستگاه آمبولانس اختصاص یافته است. برای هر سابقهٔ نمونه در این مثال، $N = ۱۵$، $n = ۶$ و $WT = ۲/۵$ است. متغیر *REPWT1* نشانهٔ وزنی است که در نیمه نمونهٔ

متعادل ۱ به هر مشاهدهٔ نمونه‌ای مورد استفاده در نیمه نمونه داده شده است. چون نیمه نمونهٔ متعادل ۱ فقط شامل سه نقطهٔ نمونه‌ای است، برای سه نقطهٔ نمونه‌ای که در به دست آوردن آن برآورد نیمه نمونه‌ای به کار رفته است $REPWT1 = \frac{۱۵}{۳} = ۵$ و برای سه نقطهٔ نمونه‌ای که در نیمه نمونه قرار نداشته‌اند برابر با صفر خواهد بود. سه متغیر دیگر، یعنی *REPWT4* ، *REPWT6* و *REPWT7* نیز به همین قیاس تعریف می‌شوند.

فرمانهای زیر برای به دست آوردن برآوردهای تکرار مکرر متعادل (BRR) با استفاده از SUDAAN به کار می‌روند:

```
PROC RATIO DATA = AMBLNCE2   FILETYPE = SAS  DESIGN = BRR;
WEIGHT  WT;
REPWGT  REPWT1  REPWT4  REPWT6  REPWT7;
NUMER  ALIVE;
DENOM  CARDARRS;
SETENV  COLWIDTH = 14;
SETENV  DECWIDTH = 5;
```

فرمان اول، نام و نوع پروندهٔ داده‌ها را معرفی می‌کند و حاکی از آن است که برآورد نسبتی قرار است به روش تکرار مکرر متعادل اجرا شود. فرمان دوم، متغیر شامل وزنهای نمونه‌گیری را نشان می‌دهد و فرمان سوم، متغیرهای شامل وزنهای نمونه‌گیری را برای آن نیمه نمونهٔ بخصوص در هر نیمه نمونهٔ متعادل نشان می‌دهد. باقیماندهٔ فرمانها بیانگر متغیرهایی هستند که شامل صورت و مخرج کسر نسبت و قالب خروجی است. خروجی حاصل از این فرمانها در زیر نشان داده شده است:

Variance Estimation Method: BRR
by: Variable, One.

Variable		One - -
ALIVE/CARDARRS	Sample Size	6.00000
	Weighted Size	15.00000
	Weighted X-sum	4527.50000
	Weighted Y-sum	695.00000
	Ratio Est.	0.15351
	SE Ratio	0.00943

برآورد نسبت و خطای معیار آن که از SUDAAN به دست می‌آیند با آنچه قبلاً نشان داده شد تطبیق می‌کنند (جز در مورد تفاوتهایی در سومین رقم دهدهی که ناشی از خطای گرد کردن است).

حالا به مقایسهٔ برآوردهای بالا که با استفاده از SUDAAN به دست آمده است با برآوردهای حاصل از همین داده‌ها با استفاده از روش خطی‌سازی می‌پردازیم. طرح، یک نمونهٔ تصادفی طبقه‌بندی شده است که نواحی خدمات اورژانسی، طبقه‌ها را تشکیل می‌دهند و نمونه‌ای متشکل از ۲ ایستگاه آمبولانس از جامعهٔ متشکل از ۵ ایستگاه آمبولانس در هر ناحیهٔ خدمات اورژانسی گرفته شده است. پروندهٔ داده‌ها برای به دست آوردن برآوردهای خطی‌سازی در زیر نشان داده شده است.

ESA	AMBSTAT	CARDARRS	ALIVE	WT	NAMBSTAT
1	1	120	25	2.5	5
1	2	78	24	2.5	5
2	1	185	30	2.5	5
2	2	228	49	2.5	5
3	1	670	80	2.5	5
3	2	530	70	2.5	5

توجه کنید که متغیر *NAMBSTAT* به هر سابقه اضافه شده است. این متغیر از آن رو لازم است که نمونه‌گیری در داخل نواحی خدمات اورژانسی بدون‌جایگذاری است و نشانهٔ آن است که اندازهٔ جامعه در هر طبقه برابر با پنج است. فرمانهای مورد استفاده برای تهیهٔ برآوردها ذیلاً نشان داده می‌شوند:

```
PROC RATIO DATA = AMBLNCE2   FILETYPE = SAS  DESIGN = STRWOR;
NEST  ESA;
TOTCNT  NAMBSTAT;
WEIGHT  WT;
NUMER  ALIVE;
DENOM  CARDARRS;
SETENV  COLWIDTH = 14;
SETENV  DECWIDTH = 5;
```

حکم مربوط به طرح *design = strwor* نشان می‌دهد که طرح، یک نمونهٔ تصادفی طبقه‌بندی شدهٔ یک‌مرحله ای است که واحدهای شمارش بدون جایگذاری انتخاب شده‌اند.

خروجی حاصل از این مجموعهٔ فرمانها در پایین نشان داده شده است:

Variance Estimation Method: Taylor Series (STRWOR)
by: Variable, One.

Variable		One - -
ALIVE/CARDARRS	Sample Size	6.00000
	Weighted Size	15.00000
	Weighted X-sum	4527.50000
	Weighted Y-sum	695.00000
	Ratio Est.	0.15351
	SE Ratio	0.00760

توجه کنید که خطای معیار نسبت برآورد شده که از خطی‌سازی به دست می‌آید کوچکتر از خطای معیاری است که از روش تکرار مکرر متعادل به دست می‌آید و این دور از انتظار نیست، زیرا روش خطی‌سازی در نظر می‌گیرد که نمونه‌گیری بدون جایگذاری است (و بنابراین از تصحیح جامعهٔ متناهی استفاده می‌کند)، در حالی که در روش تکرار مکرر متعادل چنین چیزی لحاظ نمی‌شود. اگر از روش خطی‌سازی استفاده کرده ولی طرح نمونه‌گیری طبقه‌بندی شده با جایگذاری را به کار برده بودیم (حکم طرح *design = strwr* می‌شد) برآورد خطای معیار نسبت برآورد شده برابر با ۰/۰۰۹۸۱ به دست می‌آمد که به خطای معیار برابر با ۰/۰۰۹۴۳ که از روش تکرار مکرر متعادل به دست می‌آمد بسیار نزدیک است.

۲.۲.۱۲ برآورد کردن به روش جک‌نایف

جک‌نایف روش دیگری است که برای برآورد کردن خطاهای معیار برآوردهای حاصل از آمارگیریهای نمونه‌ای پیچیده به کار می‌رود و مانند روش تکرار مکرر متعادل، مستلزم محاسبهٔ برآورد موردنظر (مانند مجموع، میانگین، نسبت) برای چند نیمه نمونه و سپس محاسبهٔ واریانس این برآوردها روی مجموعهٔ نیمه نمونه‌هاست. برآورد کردن به روش جک‌نایف تاریخچه‌ای طولانی داشته است که به نگارش مقاله‌ای توسط کنویی در دههٔ ۱۹۵۰ برمی‌گردد [۲۰] و صورتهای بسیار متعددی از آن در دست است. در اینجا روش‌شناسی ویژهٔ جک‌نایف را برای برآورد نسبتی که در حال حاضر در SUDAAN مورد استفاده قرار می‌گیرد شرح خواهیم داد و سپس روش‌شناسی برآورد کردن با SUDAAN را مجدداً با استفاده از نمونهٔ ایستگاههای آمبولانس نشان خواهیم داد.

فرض می‌کنیم که L طبقه و n_h نمونه از واحدهای نمونه‌گیری اولیه در داخل طبقهٔ h در اختیار داریم ($h = ۱,\ldots,L$). می‌خواهیم خطای معیار برآورد نسبتی، r، را برآورد کنیم که در آن

$$r = \frac{\sum_{h=۱}^{L}\sum_{i=۱}^{n_h} x'_{hi}}{\sum_{h=۱}^{L}\sum_{i=۱}^{n_h} y'_{hi}}$$

و x'_{hi} و y'_{hi} برآوردهایی بر مبنای داده‌های حاصل از آن واحد نمونه‌گیری اولیهٔ خاص‌اند.

برای هر واحد نمونه‌گیری اولیه در مجموعهٔ داده‌ها، برآورد $r_{(hi)}$ را برای نسبت جامعه‌ای مبتنی بر همهٔ مشاهدات *به استثنای* آنهایی که در واحد نمونه‌گیری اولیهٔ hi قرار دارند به دست می‌آوریم. $r_{(hi)}$ برای یک برآورد نسبتی از فرمول زیر به دست می‌آید

$$r_{(hi)} = \frac{\sum_{h'\neq h}\sum_{i=1}^{n_h} x'_{h'i} + \sum_{i'\neq i} \frac{n_h}{n_h - 1} x'_{hi'}}{\sum_{h'\neq h}\sum_{i=1}^{n_h} y'_{h'i} + \sum_{i'\neq i} \frac{n_h}{n_h - 1} y'_{hi'}}$$

توجه کنید که برآورد، $r_{(hi)}$، از برآوردهای طبقه‌ای حاصل از همهٔ واحدهای نمونه‌گیری اولیه بجز واحد نمونه‌گیری اولیهٔ hi تشکیل شده است و برآورد طبقه‌ای حاصل از آن واحد نمونه‌گیری اولیه بر برآوردهای حاصل از سایر واحدهای نمونه‌گیری اولیه در داخل آن طبقه متکی است که به طرزی مناسب با عامل $(n_h/(n_h - 1))$ تعدیل شده است تا نبود داده‌های مربوط به آن واحد نمونه‌گیری اولیه را بازتاب دهد.

همین که $r_{(hi)}$ ها محاسبه شدند، برآورد خطای معیار r از فرمول زیر به دست می‌آید

$$\hat{SE}(r) = \sqrt{\sum_{h=1}^{L}\sum_{i=1}^{n_h} \frac{n_h - 1}{n_h}(r_{(hi)} - r)^2} \qquad (٤.١٢)$$

مثال تشریحی: با در نظر گرفتن مجدد مثال تشریحی قبل، داده‌های زیر را در اختیار داریم.

ESA(h)	AMBSTAT(i)	x'_{hi}	y'_{hi}	$r_{(hi)}$	r	$(r_{(hi)} - r)^2 / 2$
1	1	692.5	4422.5	.156586	.153506	4.74×10^{-6}
1	2	697.5	4632.5	.150567	.153506	4.32×10^{-6}
2	1	742.5	4635.0	.160194	.153506	2.24×10^{-5}
2	2	647.5	4420.0	.146493	.153506	2.46×10^{-5}
3	1	670.0	4177.5	.160383	.153506	2.36×10^{-5}
3	2	720.0	4877.5	.147617	.153506	1.73×10^{-5}
Total						9.7×10^{-5}

x'_{hi}، y'_{hi} ، و $r_{(hi)}$ همان طور که در مورد x'_{11} ، y'_{11} و $r_{(11)}$ نشان داده شد به دست می‌آیند:

$$x'_{11} = ۲/۵ \times (۲ \times ۲۴ + ۳۰ + ۴۹ + ۸۰ + ۷۰) = ۶۹۲/۵$$
$$y'_{11} = ۲/۵ \times (۲ \times ۷۸ + ۱۸۵ + ۲۲۸ + ۶۷۰ + ۵۳۰) = ۴۴۲۲/۵$$
$$r_{(11)} = \frac{۶۹۲/۵}{۴۴۲۲/۵} = ۰/۱۵۶۵۸۶$$

همان‌طور که قبلاً نشان داده شد، $r = ۰/۱۵۳۵۰۶$ ، و از برابری (٤.١٢) داریم

$$\hat{SE}(r) = \sqrt{۹/۷ \times ۱۰^{-۵}} = ۰/۰۰۹۸۴۹$$

□

این برآورد کردن را می‌توان با استفاده از SUDAAN با فرمانهای زیر به دست آورد:

```
PROC RATIO DATA = AMBLNCE2   FILETYPE = SAS  DESIGN = JACKKNIFE;
NEST  ESA;
WEIGHT  WT;
NUMER  ALIVE;
DENOM  CARDARRS;
SETENV  COLWIDTH = 14;
SETENV  DECWIDTH = 7;
```

که خروجی زیر را تولید می‌کند:

Variance Estimation Method: Jackknife
by: Variable, One.

Variable		One - -
ALIVE/CARDARRS	Sample Size	6.0000000
	Weighted Size	15.0000000
	Weighted X-sum	4527.5000000
	Weighted Y-sum	695.0000000
	Ratio Est.	0.1535064
	SE Ratio	0.0098492

توجه کنید که برآورد خطای معیار نسبت، با آنچه که از استفادهٔ مستقیم از برابری (٤.١٢) به دست آمده است مطابقت دارد و مقدار ٠/٠٠٩٨٤٢ آن بسیار شبیه به مقداری است که قبلاً از روش تکرار مکرر متعادل به دست آمده بود (برآورد خطای معیار r برابر است با ٠/٠٠٩٤٣). هر دوی این مقدارها به مقدار برآورد شدهٔ ٠/٠٠٩٤٣ که با استفاده از روش خطی‌سازی با فرض گرفتن طرح نمونه‌گیری با جایگذاری به دست می‌آمد شبیه‌اند ولی از مقدار به دست آمده از روش خطی‌سازی تحت فرض نمونه‌گیری بدون جایگذاری (که برآورد خطای معیار r برابر است با ٠/٠٠٧٦٠) بیشترند.

فرض پایه‌ای برای استفاده از یکی از دو شیوهٔ تکرار مکرر متعادل یا جک‌نایف برای برآورد کردن واریانس در نمونه‌گیری یک‌مرحله ای، آن است که یا نمونه‌گیری با جایگذاری است یا کسر نمونه‌گیری کوچک (یعنی کمتر از ١٠ درصد) است. برای نمونه‌گیری چندمرحله‌ای باید نمونه‌گیری در مرحلهٔ اول نیز با جایگذاری باشد (یا کسر نمونه‌گیری کوچک باشد). نمونه‌گیری بدون جایگذاری در مراحل بعدی مجاز است.

٣.٢.١٢ برآورد کردن تغییرپذیری مصاحبه‌گر با استفاده از طرحهای نمونه‌گیری تکرار شده (نمونه‌های نافذ بر هم)

از روشهای تکرار می‌توان برای برآورد آن بخشی از تغییرپذیری استفاده کرد که ناشی از تفاوتهایی در فرایند اندازه‌گیری در میان یکایک افرادی است که مسئولیت جمع‌آوری داده‌ها را به عهده دارند. گاهی

برای اندازه‌گیری این تغییرپذیری «مصاحبه‌گر»، از روشی موسوم به *نمونه‌های نافذ بر هم* استفاده می‌شود. این روش را شاید بتوان به بهترین وجه دربارهٔ وضعیتی توضیح داد که در آن نمونهٔ تصادفی ساده‌ای متشکل از n عنصر از جامعه‌ای با N عنصر انتخاب می‌شود. اگر قرار باشد داده‌ها توسط M مصاحبه‌گر جمع‌آوری شوند در آن صورت، به جای گرفتن یک نمونهٔ تصادفی ساده از n عنصر، M نمونهٔ تصادفی سادهٔ مستقل انتخاب می‌کنیم که هر یک شامل $\overline{n} = n/M$ عنصر باشد و هر مجموعهٔ متشکل از $\overline{n}$ عنصر طی فرایندی تصادفی به یک مصاحبه‌گر خاص اختصاص داده می‌شود به طوری که احتمال تخصیص هر یک از M نمونه به هر مصاحبه‌گر یکسان است.

تحت این طرح، هر مصاحبه‌گر برآورد جداگانه‌ای، $\bar{x}_i$، از سطح میانگین $\overline{X}$ برای مشخصهٔ x در جامعه تهیه خواهد کرد و واریانس، σ_x^2 مربوط به توزیع x می‌تواند به وسیلهٔ s_{bx}^2 برآورد شود با این فرض که در فرایند اندازه‌گیری هیچ تفاوتی میان مصاحبه‌گران وجود نداشته است:

$$s_{bx}^2 = \frac{\overline{n}\sum_{i=1}^{M}(\bar{x}_i - \bar{x})^2}{M-1}$$

که در آن $\bar{x}$ میانگین $\bar{x}_i$ است.

برآورد دیگری از σ_x^2 که چه اثرهای مصاحبه‌گر وجود داشته یا نداشته باشند همچنان معتبر است از فرمول زیر به دست می‌آید

$$s_{wx}^2 = \frac{\sum_{i=1}^{M} s_{ix}^2}{M}$$

که در آن، برای $i = 1,\ldots,M$

$$s_{ix}^2 = \frac{\sum_{j=1}^{\overline{n}}(x_{ij} - \bar{x}_i)^2}{\overline{n}-1}$$

تحت فرض صفر مبنی بر اینکه هیچ اثر مصاحبه‌گر وجود ندارد، نسبت s_{bx}^2/s_{wx}^2 با $(M-1)$ و $(n-M)$ درجه آزادی از توزیع F پیروی خواهد کرد و این آمارهٔ آزمون می‌تواند برای آزمون آن فرض به کار رود.

تحت این طرح نمونه‌ای تکرار شده، هر مصاحبه‌گر را می‌توان به صورت خوشه‌ای متشکل از $\overline{n}$ مشاهده نیز در نظر گرفت. با این تفسیر، همبستگی میان رده‌ای را می‌توان به عنوان نشانگری از نسبت واریانس کل ناشی از تغییرپذیری در میان مصاحبه‌گران از لحاظ فرایند اندازه‌گیری نیز به کار برد. δ_x، برآوردگر ضریب همبستگی میان رده‌ای که توسط کالتون [۱۹] و دیگران پیشنهاد شده است به صورت زیر است:

$$\delta_x = \frac{\left(\frac{s_{bx}^2}{s_{wx}^2} - 1\right)}{\left(\frac{s_{bx}^2}{s_{wx}^2} - 1 + \bar{n}\right)} \qquad (۵.۱۲)$$

مثال تشریحی: فرض کنید نمونه‌ای متشکل از ۲۰ کاربر یک بستهٔ نرم‌افزاری حسابداری از روی فهرست کسانی که در شرکت تولید کنندهٔ این نرم‌افزار ثبت‌نام کرده‌اند گرفته شده است. این ۲۰ کاربر به طور تصادفی به ٤ مصاحبه‌گر، یعنی هر مصاحبه‌گر ۵ کاربر، اختصاص داده شده بودند که از روی فهرست اختصاصی خود با هر کاربر مصاحبهٔ تلفنی انجام می‌دادند. متغیر اصلی موردنظر، میزان رضایت از نرم‌افزار مزبور در یک مقیاس پنج درجه‌ای بود که «۱» نشانهٔ ناراضی و «۵» نشانهٔ بسیار راضی بود. داده‌های مربوط به هر یک از این چهار مصاحبه در پایین نشان داده شده‌اند:

	مصاحبه‌گر			
	۱	۲	۳	٤
	۱	۲	۳	٤
	۱	۲	۳	۳
	۲	۲	٤	۵
	۳	۳	٤	۵
	۲	٤	۵	۵
$\bar{x}_i$	۱/۸۰	۲/٦۰	۳/۸۰	٤/٤۰
s_{ix}^2	۰/۷۰	۰/۸۰	۰/۷۰	۰/۸۰

با $M = ٤$ و $\bar{n} = ۵$، داریم

$$\bar{x} = ۳/۱۵ \qquad s_{bx}^2 = ۶/۸۵ \qquad s_{wx}^2 = ۰/۷۵$$

نسبت F چنین است

$$\frac{s_{bx}^2}{s_{wx}^2} = \frac{۶/۸۵}{۰/۷۵} = ۹/۱۳۳$$

که با ۳ و ۱٦ درجهٔ آزادی، بسیار معنی‌دار است. ضریب همبستگی میان رده‌ای، δ_x، به صورت زیر به دست می‌آید

$$\delta_x = \frac{۹/۱۳۳ - ۱}{۹/۱۳۳ - ۱ + ۵} = ۰/۶۱۹۲$$

به این ترتیب می‌توان برآورد کردکه بیش از ٦۱ درصد از کل واریانس، ناشی از اثرهای مصاحبه‌گر است.

□

۳.۱۲ خلاصه

در این فصل از مطالب مربوط به برآورد کردن واریانسها یا خطاهای معیار آماره‌های حاصل از آمارگیریهای نمونه‌ای پیچیده بحث کردیم. به خصوص دربارهٔ برآوردهای واریانس حاصل از سه روش که در سطح گسترده‌ای مورد استفاده قرار می‌گیرند، یعنی *خطی‌سازی*، *تکرار مکرر متعادل*، و *روش جک‌نایف* به تفصیل بحث کردیم و مثالهایی آوردیم. مطالعات تجربی چنین نتیجه داده‌اند که هیچ روشی همواره بهتر از دیگر روشها عمل نکرده است و هر روش می‌تواند برآوردهایی نسبتاً معقول برای واریانس تهیه کند، به شرط آنکه اندازه‌های نمونه‌ای موردنظر به قدر کافی بزرگ باشند.

تمرین

۱.۱۲ $\bar{x}$ را برآوردی از سطح میانگین متغیر x و $\bar{y}$ را برآوردی از سطح میانگین متغیر y در نظر بگیرید که هر دو از نمونه‌گیری تصادفی ساده به دست آمده‌اند. فرض کنید $\bar{z} = \overline{xy}$. از خطی‌سازی در یافتن عبارتی برای برآورد واریانس توزیع $\bar{z}$ استفاده کنید.

۲.۱۲ در یک ایالت بزرگ ۳٤ بخش وجود دارند که در نواحی کلانشهری قرار ندارند. به منظور برآورد کردن کل تعداد بیمارانی که با سندروم اکتسابی نارسایی سیستم ایمنی (ایدز) طی سال تقویمی ۱۹۹۰ در بیمارستانهای این ۳٤ بخش پذیرش شده‌اند، بخشهای مزبور در شش طبقه گروه‌بندی می‌شوند و یک نمونهٔ تصادفی ساده متشکل از دو بخش از هر طبقه انتخاب می‌شود. در هر یک از بخشهای نمونه، یک بیمارستان به صورت نمونهٔ تصادفی ساده گرفته می‌شود و همهٔ سوابق پزشکی بیمارستانی بیمارانی که با تشخیص بیماری ایدز ترخیص شده‌اند در هر بیمارستان نمونه بررسی، خلاصه‌برداری، و شمارش می‌شوند. در جدول زیر، بیمارستانها برحسب طبقه، بخش، و تعداد تخت فهرست شده‌اند.

طبقه‌بندی برای آمارگیری از پذیرش بیماران دچار ایدز

طبقه	بخش	بیمارستان	تخت
۱	۱	۱	۷۲
۱	۲	۱	۸۷
۱	۲	۲	۱۰٤
۱	۲	۳	۳٤
۱	۳	۱	۱۷
۱	۳	۲	۱٤۰
۱	٤	۱	۱۰٤

طبقه‌بندی برای آمارگیری از پذیرش بیماران دچار ایدز (ادامه)

طبقه	بخش	بیمارستان	تخت
۲	۱	۱	۴۴
۲	۲	۱	۹۹
۲	۳	۱	۸۶
۲	۳	۲	۹۱
۲	۳	۳	۵۳
۲	۴	۱	۴۸
۳	۱	۱	۱۷۱
۳	۱	۲	۸۵
۳	۲	۱	۱۰۸
۳	۳	۱	۹۹
۳	۴	۱	۱۳۱
۳	۴	۲	۱۸۲
۴	۱	۱	۴۸
۴	۱	۲	۱۸
۴	۲	۱	۵۰
۴	۳	۱	۴۲
۴	۴	۱	۳۸
۵	۱	۱	۴۲
۵	۲	۱	۵۴
۵	۳	۱	۴۵
۵	۴	۱	۳۴
۶	۱	۱	۳۹
۶	۱	۲	۵۹
۶	۲	۱	۷۶
۶	۳	۱	۶۸
۶	۳	۲	۶۸
۶	۴	۱	۶۵

الف. نشان دهید که چگونه تعداد پذیرش بیماران دچار ایدز را برای ناحیهٔ ۳۴ بخشی با در نظر گرفتن تعداد تختها برآورد می‌کنید.

ب. بیمارستانهایی که عملاً در نمونه قرار گرفته‌اند همراه با تعداد پذیرش بیماران دچار ایدز که در هر بیمارستان نمونه شمارش شده‌اند در جدول زیر نشان داده شده‌اند. از روی

ایـن داده‌ها، کل تعداد پذیرش بیماران مبتلا به ایدز را در سراسر ناحیهٔ ٣٤ بخشی برآورد کنید.

بیمارستانها در نمونه

طبقه	بخش	بیمارستان	تخت	کل مبتلایان به ایدز
١	١	١	٧٢	٢٠
١	٢	١	٨٧	٤٩
٢	٢	١	٩٩	٣٨
٢	٤	١	٤٨	٢٣
٣	٣	١	٩٩	٣٨
٣	٤	١	١٣١	٧٨
٤	٣	١	٤٢	٧
٤	٤	١	٣٨	٢٨
٥	١	١	٤٢	٢٦
٥	٤	١	٣٤	٩
٦	١	١	٣٩	١٨
٦	٢	١	٧٦	٢٠

پ. از مجموعه‌ای از نیمه نمونه‌های متعادل برای برآورد کردن خطای معیار برآورد کل تعـداد پـذیرش مبتلایان به ایدز برای ناحیهٔ ٣٤ بخشی استفاده کنید. از ماتریس پلاکت ـ برمن که در پـایین ارائـه می‌شـود می‌تـوان برای سـاختن برآوردهـای نیمه نمـونه‌ای استـفاده کرد (مک‌کارتی [٦، ص. ١٧]) :

	طبقه					
تکرار	١	٢	٣	٤	٥	٦
١	+	−	−	+	−	+
٢	+	+	−	−	+	−
٣	+	+	+	−	−	+
٤	−	+	+	+	−	−
٥	+	−	+	+	+	−
٦	−	+	−	+	+	+
٧	−	−	+	−	+	+
٨	−	−	−	−	−	−

ت. از روش جک‌نایف برای به دست آوردن خطای معیار برآورد کل تعداد پذیرش مبتلایان بـه ایدز در ناحیهٔ ۳۴ بخشی استفاده کنید. این نتیجه را با برآورد حاصل از روش تکرار مکرر متعادل در قسمت (پ) مقایسه کنید.

۳.۱۲ یک آمارگیری نمونه‌ای از فروشگاههای عرضهٔ ویدیو به عمل آمد که هدف از آن برآورد کردن نسـبت ایـن قـبیل فروشـگاهها بـود کـه براسـاس بـرنامه‌هـای پیـش پرداخت حق عضویت، انگیـزه‌هـایی برای مشتریان فراهم می‌کنند. یک نمونهٔ تصادفی ساده متشکل از ۴۸ فروشگاه از ایـن نـوع گرفـته شد و به هر یک از ۶ مصاحبه‌گر ۸ فروشگاه به طور تصادفی، برای مصاحبهٔ تلفنی، اختصاص داده شد. داده‌های زیر به دست آمدند:

مصاحبه‌گر	تعداد فروشگاههای دارای برنامهٔ عضویت از پیش پرداخت شده
۱	۷
۲	۱
۳	۰
۴	۸
۵	۷
۶	۰

آیا نشانه‌ای از اثر مصاحبه‌گر دیده می‌شود؟ اگر پاسخ «بلی» است، بزرگی این اثر چقدر است؟

۴.۱۲ از روش جـک‌نایف در برآورد کردن خطای معیار نسبت اضافه پرداختی برای داده‌های جـدول ۱.۱۲ استفاده کنید. نتیـجهٔ حاصل چگونه با خطای معیـار حاصـل از روش خطی‌سـازی مقایسه می‌شود؟

پیوست فنی

$X'_{h۱}$ و $X'_{h۲}$ را برآوردهای نااریب مستقلی از دو واحد نمونه‌گیری اولیه برای پارامتر X_h (مانند میانگین، مجموع) در طبقهٔ h ($h = ۱,\ldots,L$) در نظر می‌گیریم و فرض می‌کنیم که $X'_{(۱)},\ldots,X'_{(K)}$ مجموعه‌ای متعادل از برآوردهای نیمه نمونه‌ای باشد. فرض می‌کنیم که L به K قابل قسمت است و برآورد $\hat{X}$ را به صورت زیر تعریف می‌کنیم:

$$\begin{aligned}\hat{X} &= \frac{\sum_{k=۱}^{K} X'_{(k)}}{K} \\ &= \left(\frac{۱}{K}\right)\sum_{h=۱}^{L}\frac{K}{۲}(X'_{h۱}+X'_{h۲}) \\ &= \sum_{h=۱}^{L}\frac{(X'_{h۱}+X'_{h۲})}{۲}\end{aligned}$$

واریانس $\hat{Var}(\hat{X})$ را به صورت زیر تعریف می‌کنیم:

$$\hat{Var}(\hat{X}) = \frac{\sum_{k=۱}^{K}(X_{(k)}-\hat{X})^{۲}}{K}$$

نشان خواهیم داد که $\hat{Var}(\hat{X})$ برآوردی نااریب از $Var(\hat{X})$ است.

به پنج نتیجهٔ زیر توجه کنید

۱.

$$\begin{aligned}E(X'_{h۱}-X'_{h۲})^{۲} &= Var(X'_{h۱}-X'_{h۲})+E^{۲}(X'_{h۱}-X'_{h۲}) \\ &= Var(X'_{h۱}-X'_{h۲}) \\ &= Var(X'_{h۱})+Var(X'_{h۲})\end{aligned}$$

۲.

$$E\left[\sum_{h=۱}^{L}(X'_{h۱}-X'_{h۲})^{۲}\right] = \sum_{h=۱}^{L}\left[Var(X'_{h۱})+Var(X'_{h۲})\right]$$

(این نتیجه پیامد ۱ است.)

۳.

$$\begin{aligned}E(\hat{X}) &= \left(\frac{۱}{K}\right)\sum_{k=۱}^{K}\sum_{h=۱}^{L}E\left[\delta_{kh}X'_{h۱}+(۱-\delta_{kh})X'_{h۲}\right] \\ &= \left(\frac{۱}{K}\right)\sum_{k=۱}^{K}\sum_{h=۱}^{L}X_h \\ &= \left(\frac{۱}{K}\right)K\times X \\ &= X\end{aligned}$$

.۴

$$Var(\hat{X}) = \left(\frac{۱}{K^۲}\right) Var\left(\sum_{k=۱}^{K}\sum_{h=۱}^{L}\left[\delta_{kh}X'_{h۱} + (۱-\delta_{kh})X'_{h۲}\right]\right)$$
$$= \left(\frac{۱}{K^۲}\right)\sum_{k=۱}^{K}\sum_{h=۱}^{L} Var\left[\delta_{kh}X'_{h۱} + (۱-\delta_{kh})X'_{h۲}\right]$$

(این عبارت آخر به این علت درست است که در یک مجموعهٔ متعادل از نیمه نمونه‌ها، مجموع جملات حاصلضرب متقاطع صفر می‌شود.)

$$= \left(\frac{۱}{K^۲}\right)\sum_{h=۱}^{L} Var\left[\left(\frac{K}{۲}\right)X'_{h۱} + \left(\frac{K}{۲}\right)X'_{h۲}\right]$$

(زیرا در یک مجموعه از نیمه نمونه‌های متعادل، هر یک از دو برآورد طبقه‌ای به تعداد دفعات $\left[\frac{K}{۲} =\right]$ ظاهر می‌شود.)

$$= \left(\frac{۱}{۴}\right)\sum_{h=۱}^{L}(Var(X'_{h۱}) + Var(X'_{h۲}))$$

.۵

$$E\left[\sum_{k=۱}^{K}(X'_{(k)} - \hat{X})^۲\right] = \sum_{k=۱}^{K} E(X'_{(k)} - \hat{X})^۲$$
$$= \sum_{k=۱}^{K}\left[Var(X'_{(k)} - \hat{X}) + \left[E(X'_{(k)} - \hat{X})\right]^۲\right]$$
$$= \sum_{k=۱}^{K}\left[Var(X'_{(k)} - \hat{X}) + ۰\right]$$

(از روی نتیجهٔ ۳ و این واقعیت که $E(X'_{(k)}) = X$)

$$= \sum_{k=۱}^{K} Var\left[\sum_{h=۱}^{L}\left(\delta_{kh}X'_{h۱} + (۱-\delta_{kh})X'_{h۲} - \frac{X'_{h۱} + X'_{h۲}}{۲}\right)\right]$$
$$= \sum_{k=۱}^{K} Var\left[\sum_{h=۱}^{L}\left(\left(\delta_{kh} - \frac{۱}{۲}\right)X'_{h۱} + \left(\delta_{kh} - \frac{۱}{۲}\right)X'_{h۲}\right)\right]$$
$$= \sum_{k=۱}^{K}\sum_{h=۱}^{L}\left(\delta_{kh} - \frac{۱}{۲}\right)^۲ Var(X'_{h۱} - X'_{h۲})$$
$$= \sum_{k=۱}^{K}\frac{۱}{۴}\sum_{h=۱}^{L} Var(X'_{h۱} - X'_{h۲})$$
$$= \frac{K}{۴}\sum_{h=۱}^{L}\left[Var(X'_{h۱}) + Var(X'_{h۲})\right]$$
$$= K \times Var(\hat{X})$$

بنابراین،

$$E\left[Var(\hat{X})\right] = E\left[\left(\frac{1}{K}\right)\sum_{k=1}^{K}(X'_{(k)} - \hat{X})^2\right]$$
$$= \left(\frac{1}{K}\right) \times K \times Var(\hat{X})$$
$$= Var(\hat{X})$$

برای این مورد خاص نشان دادیم که $\hat{Var}(\hat{X})$ برآوردگری نااریب از $Var(\hat{X})$ است. با این کـه بـرآوردهای نـاخطی از قبیل برآوردگرهای نسبتی و حاصلضربی نااریب نیستند، اریبی آنها غالباً ناچیز است و استدلالی که در بالا عرضه شد غالباً برای این برآوردهای ناخطی به صورت تقریـبی مصـداق پیـدا مـی‌کـند که موجب استفاده از برآوردگر واریانسی نیمه نمونه‌ای متعادل برای این برآوردگرهای ناخطی می‌شود.

کتابشناسی

The following publications develop linearization methodology.

1. Keyfitz, N., Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52: 503, 1957.
2. Woodruff, R. S., A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* 66: 411, 1971.

The following publications are relevant to software for implementation of the linearization method.

3. Research Triangle Institute, *SUDAAN: Professional Software for SUrvey DAta ANalysis,* Research Triangle Institute, Research Triangle Park, N.C., 1989.
4. SAS Institute, Inc., *SAS Users Guide: Statistics,* SAS Institute Inc., Cary, N.C., 1982.

The following publications are relevant to the discussion of replication methods in this chapter.

5. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods and Theory*, Vol. 1, Wiley, New York, 1953.
6. McCarthy, P. J., *Replication: An Approach to the Analysis of Data from Complex Surveys*, Vital and Health Statistics Series 2, No. 14, DHEW PUb No. (HSM) 73 – 1260, Health Services and Mental Health Administration, U.S. Government Printing Office, Washington, D.C., 1966.
7. McCarthy, P. J., Pseudo-replication: Half samples. *Review of the International Statistical Institute,* 37: 239, 1969.
8. Frankel, M. R., *Inference from Sample Surveys: An Empirical Investigation,* Institute for Social Research, Ann Arbor, MI., 1971.
9. Bean, J. A., *Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution,* Vital and Health

Statistics Series 2, No. 65, DHEW Pub. No. (HRA) 75-1339, Health Resources Administration, U.S. Government Printing Office, Washington, D.C., 1975.

10. Lemeshow, S., and Levy, P. S., Estimating the variances of ratio estimates in complex sample surveys with two primary sampling units per stratum-A comparison of balanced replication and jackknife techniques. *Journal of Statistical Computation and Simulation,* 8: 191, 1979.
11. Lemeshow, S., Half-sample techniques. In *Encyclopedia of Statistical Sciences*, Vol. 3, Wiley, New York, 1983.
12. Lemeshow, S., and Stoddard, A. M., A comparison of alternative estimation strategies for estimating the slope of a linear regression in sample surveys. *Communications in Statistics B: Simulation and Computation*, 13: 153, 1984.
13. Lemeshow, S., and Epp, R., Properties of the balanced half-sample and jackknife variance estimation techniques in the linear case. *Communications in Statistics: Theory and Methods*, A6(13): 1259, 1977.
14. Stanek, E., and Lemeshow, S., The behavior of balanced half-sample variance estimates for linear and combined ratio estimates when strata are paired to form pseudo-strata. *Estadistica,* 32: 71, 1978.
15. Lemeshow, S., The use of unique statistical weights for estimating variances with the balanced half-sample technique. *Journal of Statistical Planning and Inference,* 3: 315, 1979.
16. Lemeshow, S., Hosmer, D. W., and Hislop, D., The effect of non-normality on estimating the variance of the combined ratio estimate. *Communications in Statistics: Simulation and Computation,* B9(4): 371, 1980.
17. Plackett, R. L., and Burman, J. P., The design of optimal multifactorial experiments. *Biometrika,* 33: 305, 1946.

The following references contain discussions of the replicated sample designs or interpenetrating samples.

18. Deming, W. E., *Sample Design in Business Research,* Wiley, New York, 1960.
19. Kalton, G., *Introduction to Survey Sampling,* Sage Publications, Newbury Park, New York, 1983.

The following is considered the original article on jackknife estimation.

20. Quenouille, M. H., Note on bias in estimation. *Biometrika,* 43, 353-360, 1956.

The following recent review articles are relevant to topics covered in this chapter.

21. Frankel, M. R., Resampling procedures in estimation of sampling error. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.
22. Carlson, B. L., Software for statistical analysis of sample survey data. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.
23. Shah, B. V., Linearization methods of variance estimation. In *Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.
24. Ghosh, S., Interpenetrating samples. In *Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.

The following articles show consequences of not using appropriate methods of analysis when analyzing sample survey data.

25. Brogan, D., Pitfalls of using standard statistical software packages for sample

survey data. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.

26. Lemeshow, S., Letenneur, L., Lafont, S., Orgogozo, J. M., and Commenges, D., An Illustration of Analysis Taking into Account Complex Survey Considerations: The Association Between Wine Consumption and Dementia in the PAQUID study. *The American Journal of Epidemiology,* 148 No 3., 298-306, 1998.

In addition to the above literature, Dr Alan Zaslafsky of Harvard University maintains a home page on the Worldwide Web which features information on software for analysis of data from complex sample surveys. The address of this site is: http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html/

قسمت ۳

موضوعهای منتخب در روش‌شناسی آمارگیری نمونه‌ای

فصل ۱۳

بی‌پاسخی و داده‌های گمشده در آمارگیریهای نمونه‌ای

همین که نمونه انتخاب شد، کار میدانی آغاز می‌شود و تلاشی به عمل می‌آید تا داده‌های مطلوب از همهٔ واحدهای شمارش که در نمونه انتخاب شده‌اند جمع‌آوری شود. متأسفانه، به ندرت امکان دارد که در تهیهٔ داده‌های کامل از همهٔ واحدهای نمونه‌گیری شده توفیق حاصل شود. در آمارگیری نمونه‌ای برای بعضی از واحدها ممکن است اصلاً هیچ اطلاعاتی به دست نیاید و برای واحدهای دیگر ممکن است برای بعضی از اقلام سؤالها، ولی نه همهٔ اقلام تعیین شده، اطلاعاتی به دست آید. بی‌پاسخی از نوع قبلی را *بی‌پاسخی واحد* و از نوع بعدی را *بی‌پاسخی سؤال* می‌نامند.

بی‌پاسخی واحد و بی‌پاسخی سؤال هر دو تهدید عمده‌ای برای درستی برآوردهای حاصل از آمارگیریهای نمونه‌ای محسوب می‌شوند و اجتناب از هر دو نوع بی‌پاسخی در نمونه‌گیری از جامعه‌ها بسیار مشکل است. در بسیاری از آمارگیریها ممکن است با تلاش بسیار زیاد و صرف منابع هنگفت، نرخ پاسخگویی حتی به ۵۰٪ از همهٔ واحدهایی برسد که در اصل در نمونه انتخاب شده‌اند. این موضوع حتی هنگامی که در برنامهٔ آمارگیری مراجعهٔ مجدد به خانوارهایی که پاسخگوی آنها در زمان اولین مراجعه در منزل نبوده‌اند پیش‌بینی شده باشد؛ برای پاسخ تلفنی به آمارگیریهای تلفنی؛ یا برای دومین یا سومین ارسال پستی در مورد آمارگیریهای پستی نیز می‌تواند صحت داشته باشد.

افـزایش اسـتفاده از آمارگیریهای نمونه‌ای طی سالها برای تأمین اطلاعات به منظور تصمیم‌گیری، و سـختی فزایـندهٔ بـه دست آوردن نرخهای پاسخگویی بالا در آمارگیریهای نمونه‌ای، منجر به بذل توجه بسیار به این مسئله شده و به ابداع انواع گوناگونی از فنون برای رفتار با بی‌پاسخی و مقادیر گمشده در آمارگیـریهای نمونـه‌ای انجامیده است. در این فصل، دربارهٔ اثر بی‌پاسخی بر درستی برآوردهای حاصل از آمارگیـریهای نمونـه‌ای بحث می‌کنیم و سپس به برخی روشها می‌پردازیم که برای کاهش بی‌پاسخی واحد مورد استفاده قرار گرفته‌اند و بعضی از روشهایی را بررسی می‌کنیم که برای رسیدگی به داده‌های گمشده در وضعیتهای بی‌پاسخی سؤال به کار رفته‌اند.

۱.۱۳ اثر بی‌پاسخی بر درستی برآوردها

مـنظور از اجـرای بیشـتر آمارگیـریها آن اسـت کـه پارامترهای جامعه‌ای از قبیل میانگینها، مجموعها، و نسـبتها بـا بیشترین میزان درستی و قابلیت اعتماد ممکن برآورد شوند. هر یک از شیوه‌های نمونه‌گیری کـه در ایـن کـتاب شرح داده شده است می‌تواند برآوردهایی نااریب (یا دست کم سازگار) از این قبیل پارامـترها تهیـه کنـند، بـه شـرطی که نرخ پاسخگویی به هر سؤال خاص ۱۰۰٪ باشد. واضح است که چـنین چیزی به ندرت اتفاق می‌افتد و بنابراین برآوردهای حاصل، دیگر نااریب نخواهند بود. در واقع، با افزایش نرخ بی‌پاسخی، مقدار اریبی نیز افزایش خواهد یافت.

برای بررسی رسمیتر این ایده، تعریفهای زیر را ارائه می‌کنیم:

N = کل تعداد واحدهای شمارش در جامعه

N_1 = کل تعداد واحدهای پاسخگوی بالقوه در جامعه

N_2 = کل تعداد واحدهای بی‌پاسخ بالقوه در جامعه (یعنی $N_2 = N - N_1$)

$\overline{X}_1$ = میانگین سطح مشخصهٔ x در میان N_1 واحد شمارش بالقوهٔ پاسخگو

$\overline{X}_2$ = میانگین سطح مشخصهٔ x در میان N_2 واحد شمارش بالقوه بی‌پاسخ

$\overline{X} = \dfrac{N_1\overline{X}_1 + N_2\overline{X}_2}{N}$ = میانگین سطح x در میان کل جامعهٔ متشکل از N واحد شمارش

اگـر یـک نمونـهٔ تصـادفی سـاده متشکل از n واحد شمارش انتخاب کنیم و اگر هیچ تلاشی برای گرفـتن داده‌هـا از بـی‌پاسـخهای بـالقوه بـه عمل نیاوریم، میانگین سطح مشخصهٔ x را در واقع به جای مجمـوع N واحد شمارش در کل جامعه، برای زیرگروه N_1 واحد شمارش پاسخ دهنده برآورد می‌کنیم. از بحث دربارهٔ برآوردهایی برای زیرگروههای حاصل ازنمونه‌گیری تصادفی ساده (فصل ۳) دانستیم که اگـر نمونـهٔ متشکل از n واحد شمارش (EU)، n_1 واحد شمارش پاسخگو داشته باشد، و اگر $\bar{x}$ معرف

میانگین سطح χ در میان این n_1 واحد شمارش پاسخگو باشد، آن‌گاه مقدار میانگین χ از فرمول زیربه دست می‌آید

$$E(\bar{x}) = \bar{X}_1$$

و اریبی $\bar{x}$ از فرمول زیر محاسبه می‌شود

$$B(\bar{x}) = \bar{X}_1 - \bar{X} = \left(\frac{N_2}{N}\right)(\bar{X}_1 - \bar{X}_2) \qquad (۱.۱۳)$$

از بررسی رابطهٔ (۱.۱۳) متوجه می‌شویم که اریبی ناشی از بی‌پاسخی مستقل از تعداد n_1 واحدی است که با موفقیت نمونه‌گیری شده‌اند. واضح است که با افزایش اندازهٔ نمونه نمی‌توان این اریبی را کاهش داد و برای کاهش آن باید اقدامات دیگری به عمل آید. یکی از این اقدامات کاهش نسبت پاسخگویان بالقوه N_2 / N است که در یکی از بخشهای بعد مورد بحث قرار خواهد گرفت. به این ترتیب، اثر بی‌پاسخی به نسبت بی‌پاسخها و تفاوت بین میانگین بی‌پاسخهای بالقوه و پاسخگویان بستگی دارد. متأسفانه، پارامترهای N_2، $\bar{X}_1$، و $\bar{X}_2$ به ندرت معلوم‌اند. حال، این ایده‌ها را با یک مثال نشان می‌دهیم.

مثال تشریحی: فرض کنید قرار است یک آمارگیری نمونه‌ای متشکل از ۱۰۰ خانوار که از یک نمونه‌گیری تصادفی ساده به دست آمده‌اند در یک منطقهٔ روستایی شامل ۲۰۰۰ خانوار به منظور برآورد نسبت همهٔ خانوارهای فاقد توالت سیفون‌دار اجرا شود. باز فرض کنید که ۲۰٪ (۴۰۰) از این ۲۰۰۰ خانوار از همکاری با این آمارگیری خودداری کنند یا اگر در نمونه انتخاب شوند قابل دسترسی نباشند (که البته، این موضوع، پیش از آمارگیری معلوم نخواهد شد). به این ترتیب، ۲۰۰۰ خانوار موجود در جامعه از ۴۰۰ خانوار بی‌پاسخ بالقوه و ۱۶۰۰ خانوار پاسخگوی بالقوه تشکیل شده است. بالاخره فرض کنید که ۱۰۰ خانوار (یا ۲۵٪) از ۴۰۰ خانوار بی‌پاسخ بالقوه دارای توالت سیفون‌دار نیستند، در حالی که از ۱۶۰۰ خانوار پاسخگوی بالقوه ۱۶۰ خانوار (یا ۱۰٪) فاقد توالت سیفون‌دار هستند. به این ترتیب، در کل جامعهٔ ۲۰۰۰ خانواری، ۲۶۰ خانوار (یا ۱۳٪) توالت سیفون‌دار ندارند.

اگر در شیوهٔ آمارگیری، هیچ تلاشی برای به دست آوردن داده‌ها از خانوارهای بی‌پاسخ بالقوه به عمل نیاید، توزیع نسبت برآورد شده برای خانوارهای فاقد توالت سیفون‌دار که می‌تواند از آمارگیری به دست آید حول ۰/۱۰ متمرکز خواهد شد (یعنی نسبت خانوارهای فاقد توالت سیفون‌دار در میان ۱۶۰۰ خانوار پاسخگوی بالقوه)، در حالی که مقدار هدف ۰/۱۳ است. به عبارت دیگر، حذف بی‌پاسخهای بالقوه منجر به برآوردی اریب خواهد شد. در این مثال داریم

$$N_1 = ۱۶۰۰ \qquad \bar{X}_1 = ۰/۱۰ \qquad N_2 = ۴۰۰ \qquad \bar{X}_2 = ۰/۲۵ \qquad N = ۲۰۰۰$$

از رابطهٔ (۱.۱۳) داریم

$$B(\bar{x}) = \left(\frac{۴۰۰}{۲۰۰۰}\right)(۰/۱۰ - ۰/۲۵) = -۰/۰۳$$

□

۲.۱۳ روشهایی برای افزایش نرخ پاسخگویی در آمارگیریهای نمونه‌ای

از رابطهٔ (۱.۱۳) می‌توان دید که یکی از راههایی که از آن طریق می‌توان اریبی ناشی از بی‌پاسخی را کاهش داد استفاده از روشی‌شناسیی است که منجر به کاهش تعداد واحدهای شمارش بی‌پاسخ بالقوه می‌شود. در این بخش، برخی روشهای کاهش تعداد خانوارهای بی‌پاسخ بالقوه را برای آمارگیریهای خانوار، فهرست‌وار ذکر می‌کنیم.

۱.۲.۱۳ افزایش تعداد خانوارهایی که تماس با آنها موفقیت آمیز است

در آمارگیریهای خانوار که از مصاحبه‌های مستقیم استفاده می‌شود، حاصل نشدن تماس هنگامی روی می‌دهد که کسی در منزل نباشد. چون در بسیاری از خانوارها، والدین هر دو در طول روز کار می‌کنند و بچه‌ها در مدرسه هستند، احتمال ندارد که تلاش برای تماس گرفتن با اعضای خانوار در طول روز موفقیت‌آمیز باشد. در طرح آمارگیری باید تمهیداتی برای مراجعهٔ مجدد به این خانوارها در اوایل شب اندیشید یا اگر کسب اطلاعات در زمانی در حد معقول کوتاه میسر باشد در اوایل شب با این قبیل خانوارها تماس تلفنی گرفت.

در آمارگیریهای پستی از خانوارها، عدم تماس وقتی روی می‌دهد که آن خانواده دیگر در آدرسی که نامش در آنجا فهرست شده است زندگی نمی‌کند و پرسشنامه قابل ارسال نیست. اگر واحد فهرست‌برداری به جای خانوادهٔ معینی که در آنجا زندگی می‌کند خود آدرس باشد، ممکن است مراجعه به آدرس برای به دست آوردن نام سکنهٔ فعلی آن ضروری باشد. چون یکی از هر پنج خانوادهٔ آمریکایی هر ساله تغییر مکان می‌دهند، این نوع مشکلات در آمارگیریهای پستی، بالقوه زیاد است به خصوص هنگامی که برای به دست آوردن نامها از چارچوبهای جاری استفاده نشود. روشی دیگر برای کاهش این مشکل آن است که روی پاکت مثلاً نوشته شود «آقا و خانم اسمیت یا ساکنان فعلی.»

۲.۲.۱۳ افزایش نرخ تکمیل شدن پرسشنامه‌های پستی

در پرسشنامه‌های پستی غالباً می‌توان نرخ پاسخ‌گیری را با بسته‌بندی کردن خوش‌ظاهر پرسشنامه افزایش داد. مطالبی که برای خانوار ارسال می‌شود باید شامل نامه‌ای باشد که عبارات آن به دقت تنظیم شده باشد و هدف از آمارگیری را شرح دهد، سازمان مسئول اجرای آمارگیری را معرفی کند، و اظهار

دارد کـه اطلاعاتی که از پاسخگو دریافت می‌شود کاملاً محرمانه خواهند ماند و فقط به صورت انبوهه بـرای مقاصـد آمـاری به کار خواهند رفت. اگر اطلاعاتی که پاسخگو ارائه می‌دهد ذاتاً ناراحت‌کننده یا زیان‌آور است عبارت مربوط به محرمانه بودن از اهمیتی خاص برخوردار می‌شود.

آنچـه کـه مربوط به آمارگیری برای پاسخگو ارسال می‌شود باید از کیفیت عالی برخوردار باشد و با پست درجهٔ یک ارسال شود. پاکت برگشتی نیز باید برای پست درجهٔ یک آماده شده باشد. سازمانهایی کـه اقـدام بـه اجـرای آمارگیـریها مـی‌کنـند تقریباً در همه جای دنیا به این نتیجه رسیده‌اند که احتمال پاسخگویی اشخاص به پرسشنامه‌های پستی با ظاهری جذاب و حرفه‌ای به مراتب بیشتر از پاسخگویی بـه پرسشـنامه‌هایی است که ظاهری غیرحرفه‌ای دارند. به علاوه، پرسشنامه‌های بیش از حد طولانی که بـه بیـش از ۳۰ دقیقـه زمان برای پرکردن نیاز دارند بیشتر از پرسشنامه‌های کوتاهتر در معرض مخاطرهٔ کنار گذاشته شدن هستند. مولنر و همکاران [۱۲] یافته‌های یک تجربهٔ کنترل شده را گزارش داده‌اند که بـه منظور نشان دادن اثرهای یک نامهٔ همراه با پرسشنامه‌ای با شکل خاص بر نرخ پاسخگویی، طراحی شده است و طی یک آمارگیری برای مدیران بیمارستان پست شده است.

در بخـش ۳.۱۳ آمارگیـریهایی پسـتی بررسـی شـده‌اند که در طرح خود تمهیداتی را برای مصاحبهٔ حضوری با زیرنمونه‌ای از بی‌پاسخها منظور کرده‌اند.

۳.۲.۱۳ کاهش تعداد خودداریها در مصاحبه‌های رودررو یا تلفنی

خودداری از تکمیل پرسشنامه‌ای که با پست ارسال شده است برای یک پاسخگوی انتخابی بسیار آسان است، زیرا پاسخگو هیچ تماس مستقیمی با سازمان مجری آمارگیری ندارد. خودداری از پاسخ دادن در مصـاحبهٔ تلفـنی تـا انـدازه‌ای مشـکلتر اسـت زیـرا تمـاس صوتی برقرار شده است و از همه مشکلتر خـودداری از جـواب دادن در مصاحبهٔ رودررو است زیرا بین مصاحبه‌گر و پاسخگو تماس به صورت چشم در چشم است.

اگر اقدامات تبلیغاتی مؤثری پیش از آمارگیری اجرا شود می‌تواند در کاهش نرخ بی‌پاسخی، چه در آمارگیـریهای تلفـنی و چه در آمارگیریهای رودررو (و نیز در آمارگیریهای پستی)، مؤثر باشد. ولی این کـار، بـه خصـوص در مـناطق کلانشـهری بـزرگ، به آسانی میسر نیست. مثلاً رسانه‌های خبری رادیو تـلویزیونی و مطـبوعات احـتمالاً راغب نیستند که بهترین زمان و فضای خود را به اعلام آمارگیری در شـرف انجـام از فـراوانی استفاده از شعبه‌های کتابخانه‌ها اختصاص دهند در حالی که خبرهای داغتری (مثلاً پلیس دختر شهردار را در حال فروش کوکایین گرفته است) برای پخش کردن روی داده است.

اگـــر مصاحبـه‌گـر دارای اوراق و اعتبـارنامه‌های مناسب باشد، احتـمال کاهش نرخ بی‌پاسخـی به خصـوص در مصـاحبه‌هـای حضـوری وجـود خواهـد داشـت. ایـن موضـوع، به خصوص در نواحی کلانشهری بزرگ که ترس از وقوع جرایم وجود دارد درست است.

۴.۲.۱۳ استفاده از تأییدیه

اگـر آمارگیـری توسـط یـک سازمان یا نمایندگی رسمی که قلمرو وظایف آن شامل موضوع آمارگیری اسـت تأیید شده باشد ممکن است نرخ پاسخگویی در آمارگیریهای خانوار افزایش یابد. مثلاً آمارگیری بهداشت خانوار ممکن است از تأییدیه انجمن پزشکی محلی بهره‌مند شود. این تأییدیه در آمارگیریهای پسـتی مـی‌توانـد بـه صورت نامه‌ای باشد که همراه با مطالب مربوط به آمارگیری برای پاسخگو ارسال مـی‌شـود. نامـهٔ ضـمیمه بـاید دارای آرم نمایـندگی تأیید کننده بوده توسط یکی از مقامات عالیرتبهٔ آن نمایندگی امضا شده باشد تا حداکثر تأثیر را بر جای بگذارد.

تأیید توسط یک نمایندگی مناسب، به خصوص در آمارگیری از مؤسسات، حائز اهمیت است. مثلاً در آمارگیـری از بیمارستانها، اگر طرح دارای تأییدیه محکمی از انجمن بیمارستانی ایالتی باشد، احتمال موفقیـت آن افـزایش خواهـد یـافت. گـاهی اوقات، اگر سازمان موردنظر به عنوان همکار یا بخشی از کمیتهٔ نظارتی در بررسی مشارکت داده شود، تأییدیه آسانتر به دست خواهد آمد. به این ترتیب می‌توان از مزیت استفاده از تخصص و تجربهٔ آن سازمان نیز به عنوان بخشی از منابع بررسی بهره‌مند شد.

پاداشـی را که در ازای مشارکت در آمارگیری به پاسخگو می‌دهند *انگیزه* می‌نامند. انگیزه‌های نقدی کـه در سـطحی گسـترده در آمارگیـریهای پسـتی و در سـطحی محدودتر در مصاحبه‌های حضوری یا آمارگیـریهای تلفـنی مورد استفاده قرار می‌گیرند در افزایش میزان پاسخگویی مؤثر بوده‌اند. مثلاً اگر در نامه، یک سکهٔ براق ۲۵ سنتی یا یک اسکناس نو تانخوردهٔ یک دلاری قرار داده شود ممکن است میزان پاسـخگویی افـزایش یـابد. از انگیـزه‌های غیرنقدی نیز می‌توان استفاده کرد ولی با انگیزه‌های غیرپولی ممکن است مشکل اریبی بروز کند زیرا این انگیزه‌ها بیشتر احتمال دارد زیرگروههای خاصی از جامعه را جـذب کنـند تـا پول که جاذبهٔ عمومی دارد. مرکز ملی آمارهای بهداشتی در یک بررسی کنترل شده نشـان داده اسـت کـه وعـدهٔ پـرداخت یـک پاداش ده دلاری (به ارزش دلار سال ۱۹۷۱) به فرد نمونهٔ طـراحی شـده در صورت شرکت در یک آزمایش جسمی نسبتاً طولانی همراه با مصاحبه باعث افزایش قـابل توجهـی در میـزان پاسـخگویی شـده است [۱۳]. در کتابی که اردوش دربارهٔ آمارگیریهای پستی نوشته مبحث بسیار خوبی دربارهٔ کاربرد انگیزه‌ها ارائه شده است [۱۰].

۳.۱۳ آمارگیریهای پستی همراه با مصاحبه با بی‌پاسخها

آمارگیریهای پستی به طور کلی کم‌هزینه‌تر از آمارگیریهای خانواری هستند که با مصاحبهٔ حضوری اجرا می‌شوند. ولی غالباً به دست آوردن نرخ پاسخهای کافی برای تأمین مشخصه‌های مربوط به معتبر بودن و قابلیت اطمینان برآوردها از آمارگیریهای پستی مشکل است. اگر نرخ پاسخگویی آغازین به پرسشنامهٔ پستی کم باشد، برآوردهای حاصل به شدت اریب خواهند بود. برای غلبه بر این مشکل می‌توان از یک شیوهٔ نمونه‌گیری دومرحله‌ای استفاده کرد که مرحلهٔ اول آن آمارگیری پستی و مرحلهٔ دوم آن آمارگیری تلفنی و یا مصاحبهٔ حضوری از زیرمجموعه‌ای از کسانی است که به پرسشنامهٔ پستی پاسخ نداده‌اند. این شیوه غالباً می‌تواند برآوردهایی با قابلیت اطمینان زیاد به دست دهد و با هزینه‌ای معقول اجرا شود. این نوع طرح نمونه‌ای به تفصیل در مثال بعد شرح داده شده است.

مثال تشریحی: فرض کنید در جامعه‌ای متشکل از ۳۰۰ پزشک، پرسشنامه‌ای به یک نمونهٔ تصادفی ساده از ۱۰۰ پزشک ارسال شده که در آن از پزشک سؤال شده است که آیا بیمارانی را که نمی‌توانند هزینهٔ خدمات پزشکی را به طور مستقیم یا غیرمستقیم بپردازند می‌پذیرند یا نه. فرض می‌کنیم از ۱۰۰ پرسشنامهٔ ارسال شده ۳۰ پرسشنامه برگشته است و باز فرض می‌کنیم که از ۷۰ پزشکی که به پرسشنامهٔ پستی پاسخ نداده‌اند یک نمونهٔ تصادفی ساده متشکل از ۲۰ پزشک انتخاب شده‌اند و تلاشهایی متمرکز از طریق تلفن و مراجعهٔ حضوری به عمل آمده است تا از این ۲۰ پزشک نیز پاسخهایی گرفته شود و در نتیجهٔ این تلاشهای متمرکز، با ۱۵ نفر از این ۲۰ پزشک با موفقیت مصاحبه شده است. بالاخره فرض می‌کنیم داده‌هایی که در جدول ۱.۱۳ نشان داده شده‌اند از این پاسخگوها به دست آمده‌اند.

برای تعمیم و نشان دادن این وضعیت، نمادهای زیر را معرفی می‌کنیم:

N = تعداد واحدهای شمارش (EU ها) در جامعه.

n = تعداد واحدهای شمارش که در آغاز به صورت پستی نمونه‌گیری شده‌اند.

n_1 = تعداد واحدهای شمارش که به ارسال پستی آغازین پاسخ داده‌اند.

$n_2 = n - n_1$ = تعداد واحدهای شمارش که به ارسال پستی آغازین پاسخ نداده‌اند.

n_2^* = تعدادی از n_2 واحد شمارش بی‌پاسخ که برای تلاش متمرکز (تلفنی یا مصاحبهٔ حضوری) انتخاب شده‌اند.

n_2' = تعدادی از این n_2^* واحد شمارش که برای آنها پاسخهایی با موفقیت به دست آمده است.

جدول ۱.۱۳ داده‌های آمارگیری از پزشکان

	تعداد	تعدادی که پاسخ مثبت داده‌اند
پزشکانی که پرسشنامهٔ پستی را عودت داده‌اند	۳۰	۲۰
پزشکانی که به مصاحبهٔ تلفنی یا حضوری پاسخ داده‌اند	۱۵	۳

$\bar{x}_{۱} = \dfrac{\sum_{i=۱}^{n_{۱}} x_i}{n_{۱}}$ = میانگین سطح x در میان $n_{۱}$ واحد شمارش که در ارسال پستی آغازین با موفقیت با آنها تماس برقرار شده است.

$\bar{x}_{۲} = \dfrac{\sum_{i=۱}^{n'_{۲}} x_i}{n'_{۲}}$ = میانگین سطح x در میان $n'_{۲}$ واحد شمارش که از طریق تلاشهای متمرکز با موفقیت با آنها تماس برقرار شده است.

از برآوردگری که با نماد $\bar{x}_{dub}$ نشان خواهیم داد (چون بر مبنای یک نمونهٔ مضاعف است) می‌توان برای برآورد $\bar{X}$، میانگین جامعهٔ نامعلوم استفاده کرد. این برآوردگر از فرمول زیر به دست می‌آید

$$\bar{x}_{dub} = \frac{n_{۱}\bar{x}_{۱} + n_{۲}\bar{x}_{۲}}{n} \qquad (۲.۱۳)$$

برای این مثال داریم

$$N = ۳۰۰ \qquad n = ۱۰۰ \qquad n_{۱} = ۳۰ \qquad n_{۲} = ۷۰ \qquad n^{*}_{۲} = ۲۰ \qquad n'_{۲} = ۱۵$$

$$\bar{x}_{۱} = \frac{۲۰}{۳۰} = ۰/۶۷ \qquad \bar{x}_{۲} = \frac{۳}{۱۵} = ۰/۲۰$$

و به این ترتیب

$$\bar{x}_{dub} = \frac{۳۰(۰/۶۷) + ۷۰(۰/۲۰)}{۱۰۰} = ۰/۳۴$$

اگر تعداد $n'_{۲}$ واحد شمارش بی‌پاسخ آغازین که با آنها از طریق تلاش متمرکز با موفقیت تماس برقرار می‌شود از نظر عددی نزدیک به $n^{*}_{۲}$ تعداد انتخاب شده برای تلاش متمرکز باشد، در آن صورت $\bar{x}_{dub}$ تقریباً برآوردگری ناریب از میانگین جامعه‌ای نامعلوم $\bar{X}$ خواهد بود.

□

نتایج این مثال را می‌توان به هر آمارگیری مناسبی تعمیم داد.

۱.۳.۱۳ تعیین کسر بهینهٔ بی‌پاسخهای آغازین برای زیرنمونه در تلاشهای متمرکز

فـرض کنید تصمیم بر این است که از شیوهٔ نمونه‌گیری دومرحله‌ای شامل یک پرسشنامهٔ پستی آغازین و پـس از آن از یـک تلاش متمرکز برای به دست آوردن پاسخها از زیرنمونه‌ای از واحدهای شمارشی کـه بـه پرسشـنامهٔ پستی آغازین پاسخ نداده‌اند استفاده شود. تصمیم مهمی که باید گرفته شود آن است کـه نمونهٔ n_2^* که قرار است از n_2 واحدهای شمارش که به پرسشنامهٔ پستی جواب نداده‌اند گرفته شود چقـدر بـاید بـزرگ باشـد. بـرای گرفـتن این تصمیم، راهبردی را پیشنهاد می‌کنیم که با در نظر گرفتن هزینه‌های میدانی و نرخ بی‌پاسخی مورد انتظار به تخصیص بهینه برای زیرنمونه منجر می‌شود.

ابـتدا هزیـنه‌هـای واحـد، مـلازم بـا ایـن طرح نمونه‌گیری را مورد بحث قرار می‌دهیم. فرض کنید *مؤلفه‌های هزینهٔ* زیر تعریف شده‌اند:

C_0 = هزینه به ازای پست کردن پرسشنامه‌های آغازین.

C_1 = هزینه به ازای هر پرسشنامهٔ برگشتی برای پردازش پرسشنامه‌هایی که با پست عودت داده شده‌اند.

C_2 = هزینه به ازای هر پرسشنامه برای به دست آوردن داده‌ها از آن دسته از واحدهای شمارش بی‌پاسخ آغازین که برای تلاش متمرکز تعیین شده‌اند و پردازش داده‌ها به محض دریافت آنها.

مؤلفـهٔ هزینهٔ C_0 شامل هزینهٔ موادی است که در پرسشنامه به کار رفته‌اند (مانند تمبر، پاکت و غیره) به اضـافهٔ هزیـنهٔ نیروی کار لازم در آماده‌سازی این مواد برای ارسال پستی. مؤلفهٔ هزینهٔ C_1 شامل نیروی کـار درگیـر در بازبیـنی و کدگـذاری داده‌هـا و سایر کارها برای آماده‌سازی داده‌ها به اضافهٔ هزینه‌های پـردازش داده‌هـا از قبیل وقت رایانه و برنامه‌ریزی رایانه‌ای است. مؤلفهٔ هزینهٔ C_2 ترکیبی از هزینه‌های میدانی و آماده‌سازی و پردازش داده‌هاست.

برآوردی از *کل هزینه‌های میدانی و فرابری* آمارگیری از فرمول زیر به دست می‌آید

$$C = C_0 n + C_1 n_1 + C_2 n_2' \qquad (۳.۱۳)$$

سـرانجام این که اگر پیش‌بینی شود که نسبت P_1 از کسانی که در آغاز نمونه‌گیری شده‌اند به ارسال پسـتی آغازین پاسخ خواهند داد، در آن صورت *تعداد بهینهٔ* n_2^* *که باید نمونه‌گیری شوند* از فرمول زیر به دست می‌آید

$$n_2^* = n_2 \times \left(\frac{C_0 + C_1 P_1}{C_2 P_1} \right)^{1/2} \qquad (٤.۱۳)$$

مثال تشریحی: در مورد مثال قبل فرض می‌کنیم که مؤلفه‌های هزینه به صورت زیرند:

$C_0 =$ ۱/۵۰ دلار به ازای هر پرسشنامه (بابت هزینه‌های پستی آغازین)

$C_1 =$ ۱۵/۰۰ دلار به ازای هر پرسشنامه (بابت هزینه‌های آماده‌سازی واحد و پردازش)

$C_2 =$ ۴۵/۰۰ دلار بـه ازای هـر واحد شمارش (هزینهٔ واحد برای مصاحبه و پردازش داده‌ها در مورد واحدهای شمارش بی‌پاسخ)

پس P_1، نرخ پاسخ مشاهده شده چنین است

$$P_1 = \frac{۳۰}{۱۰۰} = ۰/۳۰$$

و تعداد بهینه برای زیرنمونه عبارت است از

$$n_2^* = ۷۰ \times \left[\frac{۱/۵۰ + (۱۵)(۰/۳۰)}{۴۵(۰/۳۰)}\right]^{1/2} = ۷۰ \times \frac{۲}{۳} \approx ۴۷$$

به این ترتیب باید دو سوم یا ۴۷ واحد شمارش از میان ۷۰ بی‌پاسخ آغازین در زیرنمونه انتخاب شوند. □

۲.۳.۱۳ تعیین اندازهٔ نمونهٔ مورد نیاز برای آمارگیری پستی دومرحله‌ای

فـرض کنید از اول با آمارگیری پزشکان شروع می‌کنیم و می‌خواهیم بدانیم در آغاز چند پرسشنامه باید ارسـال شـوند. بـا فـرض ایـن کـه ۳۰ درصـد از پزشـکان به پرسشنامهٔ پستی پاسخ دهند و مؤلفه‌های هزیـنه‌ای مـورد اسـتفاده همـان مؤلفـه‌هـای مفـروض در بخش قبل باشند، می‌دانیم که باید دو سوم از بی‌پاسخهای آغازین را برای تلاش متمرکز به صورت زیرنمونه، تهیه کنیم (یعنی $\frac{n_2^*}{n_2} = \frac{۲}{۳}$).

بـرای تعیین تعداد پزشکان مورد نیاز برای نمونه گرفتن در ارسال پستی آغازین، ابتدا تعداد n' را که در صـورت پاسـخ ۱۰۰٪ بـه پرسشـنامهٔ پسـتی مـورد نیـاز اسـت تعیین می‌کنیم. چون در این مورد از نمونـه‌گیری تصادفی ساده استفاده می‌شود برای به دست آوردن n' عبارت تابلوی ۵.۳ (برای نسبت) را بـه کـار مـی‌بـریم. فـرض کـنید حدس می‌زنیم که ۸۰٪ پزشکان، بیمارانی را که توانایی پرداخت بهای خدمـات آنهـا را ندارنـد می‌پذیرند و می‌خواهیم واقعاً مطمئن باشیم که این درصد را در محدودهٔ ۳۰٪ مقدار واقعی آن برآورد می‌کنیم. پس از تابلوی ۵.۳ با $\varepsilon = ۰/۳۰$ ، $p = ۰/۸۰$ ، و $N = ۳۰۰$ ، داریم

$$n' = \frac{۹ \times ۳۰۰ \times ۰/۸ \times ۰/۲}{۹ \times ۰/۸ \times ۰/۲ + ۲۹۹ \times ۰/۳^2 \times ۰/۸^2} = ۲۳ \text{ پزشک}$$

بـه ایـن تـرتیب اگـر نـرخ پاسـخگویی ۱۰۰٪ باشد به نمونه‌ای متشکل از فقط ۲۳ پزشک نیاز خواهیم داشت تا واقعاً مطمئن باشیم که نسبت برآورد شده در محدودهٔ ۳۰ درصدی نسبت واقعی قرار دارد.

ولـی نـرخ پاسـخگویی بـه پرسشنامهٔ پستی فقط ۳۰٪ پیش‌بینی می‌شود و به همین علت، برآورد در معـرض تغییرپذیـری بیشتری قرار دارد. پس به نمونه‌ای با اندازهٔ بزرگتر نیاز داریم. در واقع، n ، تعداد مـورد نیـاز پرسشـنامه‌ها می‌تواند با ضرب کردن n' (از تابلوی ۴.۳) در عاملی که بی‌پاسخی را در نظر گرفته باشد به دست آید. این عامل در فرمول زیر برای n نشان داده شده است:

$$n = n'\left\{1 + (1 - P_1)\left[\left(\frac{C_2 P_1}{C_0 + C_1 P_1}\right)^{1/4} - 1\right]\right\} \qquad (۵.۱۳)$$

پس، در مورد مثال قبل، از روی نتایجی که در بالا و قبل از آن نشان داده شد، داریم

$$\left(\frac{C_2 P_1}{C_0 + C_1 P_1}\right)^{1/4} = \frac{3}{2} \qquad P_1 = 0/30 \qquad n' = 23$$

و بنابراین

$$n = 23\left[1 + 0/7\left(\frac{3}{2} - 1\right)\right] = 31/05 \approx 31$$

پس باید یک نمونهٔ تصادفی ساده متشکل از ۳۱ پزشک برای ارسال پستی آغازین انتخاب کنیم تا واقعاً مطمئن باشیم که مشخصه‌های تعیین شده برای برآورد تأمین خواهد شد.

۴.۱۳ سایر کاربردهای روش‌شناسی نمونه‌گیری مضاعف

روش‌شناسـی نمونـه‌گیری دومرحله‌ای که در بخش قبل به عنوان ابزاری برای پایین آوردن اریبی ناشی از بی‌پاسخی پیشنهاد شد غالباً *نمونه‌گیری مضاعف* نامیده می‌شود و قابل تعمیم به وضعیتهایی است که در آن اریـبیها ناشـی از سـازوکارهایی غیـر از بـی‌پاسـخی هسـتند. در واقـع، روش‌شناسی نمونه‌گیری مضـاعف مسـتلزم جمع‌آوری داده‌های اولیه (مثلاً پرسشنامهٔ پستی) و به دنبال آن، گروه‌بندی عناصر در دو طـبقه اسـت (مانند پاسخگویان به پرسشنامهٔ پستی و بی‌پاسخها). سپس نمونه‌ای از افراد در یکی از طـبقه‌هـا انـتخاب مـی‌شـود (مانـند بی‌پاسخهای پرسشنامهٔ پستی) و برآوردگر نهایی، ترکیبی موزون از بـرآوردهای هـر یک از دو طبقه است [مثلاً رابطهٔ (۲.۱۳)]. در پایین، کاربردی از این شیوهٔ نمونه‌گیری مضاعف را نشان می‌دهیم که با بی‌پاسخی ارتباطی ندارد.

مثال تشریحی: این مثال از تحلیلی گرفته شده است که ریچ و همکاران [۱۴] دربارهٔ داده‌های حاصل از عکسبردای از قفسهٔ سینه که از دومین آمارگیری ملی بهداشت و آزمایش تغذیه جمع‌آوری شده بود، اجرا کرده‌اند. این یک آمارگیری مقطعی بود که از سال ۱۹۷۶ تا ۱۹۸۰ توسط مرکز ملی آمارهای بهداشتی با استفاده از نمونه‌ای متشکل از تقریباً ۲۸۰۰۰ نفر با سنین ۶ ماه تا ۷۴ سال که در امریکا سکونت داشتند اجرا می‌شد. این آمارگیری مشتمل بر گروهی از مصاحبه‌ها، معاینات پزشکی، و تستهای آزمایشگاهی بود. از قریب به ۱۰۰۰۰ آزمودنی ۲۵ سال به بالا عکس قفسهٔ سینه گرفته شد. هر عکس توسط دو پرتوشناس در جستجوی تعداد زیادی از علایم حاکی از بیماری احتمالی ریه بررسی شد. یکی از اقلام مندرج در فهرست نابهنجاریها «گشاد شدن شریانهای ریوی» بود. گشاد شدن شریان ریوی می‌تواند نشانهٔ فشار خون ریوی باشد، بیماری سختی که معمولاً آخرین مرحلهٔ بیماری انسداد مزمن ریه و سایر نابهنجاریهای ریوی است.

از میان ۱۰۱۵۳ عکس قفسهٔ سینهٔ پاسخگویان آمارگیری با ۲۵ تا ۷۴ سال سن که توسط دو پرتوشناس بررسی شد، ۳۲۶ مورد توسط یک یا هر دو پرتوشناس دچار گشادی شریان ریوی تشخیص داده شدند. ولی فقط در ۳۰ مورد (۹/۲٪) *هر دو پرتوشناس* در مورد گشادی شریان توافق داشتند. چون میزان مورد توافق هر دو پرتوشناس تا این اندازه کم بود تصمیم گرفته شد که پرتوشناس سومی تعیین شود تا اندازه‌گیری دقیقتری نه تنها از ۳۲۶ عکس قفسهٔ سینه که در آنها گشادی شریان ریوی توسط یکی یا هر دو پرتوشناس اولیه تشخیص داده شده بود تهیه کند بلکه نمونه‌ای متشکل از ۲۸۸ عکس قفسهٔ سینه را که از میان ۹۸۲۷ (۳۲۶–۱۰۱۵۳) عکسی انتخاب شده‌اند که *فاقد* گشادی شریان ریوی توسط دو پرتوشناس اولیه تشخیص داده شده بودند از نو دقیقتر بخواند.

اگر این واقعیت را در نظر نگیریم که دومین آمارگیری ملی بهداشت و آزمایش تغذیه یک آمارگیری پیچیده است و فرض کنیم که هر دو نمونهٔ اولیه و ثانویه به وسیلهٔ نمونه‌گیری تصادفی ساده به دست آمده‌اند، برآورد زیر را از نرخ شیوع برای افراد دچار گشادی سرخرگهای ریوی خواهیم داشت:

$$p_{dub} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

که در آن

N_1 = تعداد عکسهای قفسهٔ سینه که توسط یک یا هر دو پرتوشناس اولیه مثبت تشخیص داده شده‌اند = ۳۲۶

N_2 = تعداد عکسهای قفسهٔ سینه که توسط هر دو پرتوشناس اولیه منفی تشخیص داده شده‌اند = ۹۸۲۷

$$p_۱ = \frac{x_۱}{N_۱}$$

$$p_۲ = \frac{x_۲}{n_۲}$$

$x_۱ =$ تعداد تأیید شدهٔ مثبت توسط پرتوشناس سوم از میان $N_۱$ تشخیص اولیهٔ مثبت توسط یک یا هر دو پرتوشناس.

$n_۲ =$ تعداد نمونه‌گیری شده از میان $N_۲$ تشخیص منفی توسط هر دو پرتوشناس = ۲۸۸

$x_۲ =$ تعداد تشخیص مثبت توسط پرتوشناس سوم از میان $n_۲$ تشخیص منفی اولیه توسط هر دو پرتوشناس اول و دوم که برای تأیید پرتوشناس سوم نمونه‌گیری شده‌اند.

بر مبنای نتایج پرتوشناس سوم، ارقام زیر به دست آمد

$$p_۱ = ۰/۶۷۵ \qquad p_۲ = ۰/۲۵۷$$

به این ترتیب، برآورد ما از نرخ شیوع گشاد شدن شریان ریوی از طریق این نمونهٔ مضاعف چنین خواهد بود

$$p_{dub} = \frac{۳۲۶(۰/۶۷۵) + ۹۸۲۷(۰/۲۵۷)}{۱۰۱۵۳} = ۰/۲۷۰۴ \quad \text{یا} \quad ٪۲۷/۰۴$$

توجه کنید که اگر ناهماهنگی میان پرتوشناسان اولیه در نظر گرفته نمی‌شد، نرخ شیوع برآورد شده برابر $\frac{۳۲۶}{۱۰۱۵۳} = ٪۳/۲$ می‌شد که به احتمال زیاد، کم‌برآوردی ناخالص از نرخ شیوع می‌بود. □

۵.۱۳ بی‌پاسخی سؤال : روشهای جانهی

پژوهشگرانی که به تحلیل داده‌های حاصل از آمارگیریهای نمونه‌ای می‌پردازند با مشکلاتی مواجه‌اند که یکی از مشکلترین آنها رسیدگی به بی‌پاسخی است. همان طور که در بالا تعریف شد، بی‌پاسخی سؤال به عناصر داده‌های گمشده برمی‌گردد و می‌تواند شامل آن مقادیری از عناصر داده‌ای نیز باشد که آشکارا اشتباه بوده و نمی‌توانند در فرایند تحلیل به کار روند. بی‌پاسخی سؤال هنگامی که داده‌های جمع‌آوری شده در آمارگیری در معرض تحلیل آماری قرار می‌گیرند مشکلات بسیاری به وجود می‌آورد. بسیاری از شیوه‌های آماری از قبیل تحلیل رگرسیونی چندگانه و تحلیل عاملی را، در صورتی که هرگونه مقادیر گمشده وجود داشته باشند، نمی‌توان به فوریت مورد استفاده قرار داد و کاربرد و

تفسیر روشهایی که مقادیر گمشده را لحاظ می‌کنند اگر تحلیلگر خبرگی قابل ملاحظه نداشته باشد غالباً بسیار دشوار است.

در این بخش، ابتدا به بحث دربارهٔ سازوکارهایی می‌پردازیم که باعث بروز مقادیر گمشده می‌شوند و سپس راهبردهایی را ارائه می‌دهیم که می‌توان از آنها در مقابله با این قبیل مشکلات پیروی کرد. بسیاری از این بحث براساس روش‌شناسی توسعه یافته از آغاز دههٔ ۱۹۸۰ استوار است. برای بحث دربارهٔ مسایل مربوط به مقادیر گمشده با جزئیات بیشتر، خواننده را به کتابهای تألیف روبین [۱] و لیتل و روبین [۲] و بررسی مقالاتی جدید توسط لیتل [٦] و بارنارد، روبین و شنکر [۷] ارجاع می‌دهیم.

۱.۵.۱۳ سازوکارهایی که مقادیر گمشده از آنها ناشی می‌شود

درک سازوکارهای تصادفی که باعث بروز مقادیر گمشده می‌شوند به این دلیل اهمیت دارد که هم بر نوع راهبرد تحلیلی که به کار گرفته خواهد شد اثر می‌گذارد و هم بر اعتبار استنباطهای ناشی از تحلیل تأثیر دارد. نمادگذاری زیر توسط لیتل [٦] برای توصیف الگوهای داده‌های گمشده و طبقه‌بندی سازوکارهای زیربنایی این الگوها به کار رفته است.

$Y = (y_{ij})$ را ماتریس فرضی n در k در نظر بگیرید که شامل مقادیر مجموعه‌ای متشکل از k متغیر برای n واحد است که نمونه‌گیری شده‌اند. y_{ij} ممکن است گزارش شده یا گمشده باشد (که در وضعیت اخیر معلوم نخواهند بود). $\mathbf{M} = (m_{ij})$ را ماتریس متناظر n در k برای ۱ ها و صفرها در نظر بگیرید که ۱ نشانهٔ آن است که مقدار y_{ij} موردنظر گمشده است. سازوکار مربوط به رویداد مقادیر گمشده با توزیع ماتریس تصادفی $\mathbf{M}$ از مقادیر گمشده با توجه به ماتریس مقادیر واقعی، Y، مشخص می‌شود. به ویژه، سه مورد از متداولترین سازوکارها را در زیر تعریف می‌کنیم.

گمشدهٔ کاملاً تصادفی. اگر توزیع $\mathbf{M}$ به شرط Y، مستقل از Y باشد، در آن صورت می‌گوییم که سازوکار بی‌پاسخی سؤال به طور *کاملاً تصادفی گمشده* است $(MCAR)$[1].

گمشدهٔ تصادفی. اگر توزیع $\mathbf{M}$ به شرط Y، وابسته به مقادیر مشاهده شدهٔ Y باشد ولی به مقادیر گمشدهٔ Y بستگی نداشته باشد، در آن صورت می‌گوییم که سازوکار بی‌پاسخی سؤال به طور *تصادفی گمشده* است (MAR)[2].

مقادیر گمشدهٔ غیر قابل چشم‌پوشی. اگر توزیع $\mathbf{M}$ به شرط Y، وابسته به مقادیر گمشدهٔ Y باشد، در آن صورت می‌گوییم که سازوکار بی‌پاسخی سؤال *غیر قابل چشم‌پوشی* است.

[1] Missing Completely At Random
[2] Missing At Random

مـثال تشـریحی. یک نمونهٔ تصادفی ساده از پرونده‌های ۳۰۰۰ نفر از اعضای یک سازمان بزرگ حفظ بهداشـت گرفـته شـده اسـت. هـدف از ایـن آمارگیری آن است که از روی داده‌هایی که در اصل طی ارزیـابی آغـازین هـر عضـو جمـع‌آوری شـده اسـت، اطلاعـات مـربوط به مشخصات اعضا را دربارهٔ متغیرهای زیر برآورد کند:

۱. شمارهٔ شناسنامهٔ عضو (ID)
۲. سن (AGE)
۳. جنس (SEX)
٤. تعداد همبسترها طی ۱۲ ماه گذشته (NSEXPART)
۵. بود یا نبود پادگن سطحی هپاتیت B (HBSAG)
٦. پـادگن ویـژهٔ پروستات که در سطحی غیر متعارف بالاست. سطح بالای پادگن ویژهٔ پروستات احتمالاً حاکی از وجود نوعی بدخیمی در غدهٔ پروستات است (PSA).

بـرای نشـان دادن بـرخی از مطالـبی که در این بخش مورد بحث قرار خواهند گرفت یک مجموعه داده‌هـای مصـنوعی بـه نـام *mcardemo.dta* تولیـد کـرده‌ایم (که برای STATA تنظیم شده است). با اسـتفاده از تولیـد کنـنده‌هـای اعـداد تصـادفی، ۳۰۰۰ سـابقه بـا داده‌هـای کامل دربارهٔ هر ٦ متغیر بالا سـاخته‌ایـم. سـپس، بـرای متغیرهای *NSEXPART* ، *HBSAG* و *PSA* ، مطابق با وضعیتی که در پایین شـرح داده مـی‌شـود مقادیـر گمشده را تولید کرده‌ایم. (فرض بر این بوده است که برای متغیرهای سن *AGE* و جنس *SEX* مقادیر گمشده نخواهیم داشت.)

HBSAG . بـرای بـه وجـود آوردن وضـعیتی کـه در آن فقط یک نمونهٔ تصادفی از اعضا از نظر پادگن سـطحی هپاتیت B مورد آزمایش قرار می‌گیرند، مقادیر گمشده را برای آن متغیر از روی یک نمونهٔ تصـادفی سـاده متشـکل از تقریـباً ۵۰٪ اعضـای نمونـه، با استفاده از اعداد تصادفی دارای توزیعی یکنواخت بین ۰ و ۱، تولید کردیم.

PSA . در آغـاز مقادیـر PSA را برای همهٔ زنان معادل ۰ گرفتیم (چون زنها غدهٔ پروستات ندارند) و با اسـتفاده از اعـداد تصادفی به پیروی از یک توزیع لوژستیک با در نظر گرفتن سن، AGE، به عنوان یـک مـتغیر مسـتقل، مقادیـری را تعییـن کـردیم کـه در مورد مردان به طور یکنوا با بالا رفتن سن افزایش پیدا می‌کنند (زیرا وقوع سرطان پروستات با بالا رفتن سن به شدت افزایش پیدا می‌کند). با فـرض ایـن که نه زنان و نه مردان کمتر از ۵۰سال به منظور سنجش این متغیر *PSA* مورد آزمایش قرار نخواهند گرفت، مقادیر گمشده را برای همهٔ زنان و مردان کمتر از ۵۰ سال تهیه کردیم.

NSEXPART . مقادیر اولیه برای این متغیر با استفاده از تولید کننده‌های اعداد تصـادفی طوری ساختـه

شـده بودنـد که برای مردان اندکی بیشتر از زنان باشند، همراه با بالا رفتن سن کاهش پیدا کنند، در مـورد کسـانی که پادگن سطحی هپاتیت B در آنها مثبت است اندکی بیشتر باشند. با این که فرض بـر این بود که همهٔ اعضا در مورد تعداد همبسترها طی ۱۲ ماه گذشته مورد سؤال قرار بگیرند، این فـرض را هـم در نظـر گرفـتیم کـه بسیاری از افرادی که با تعداد زیادی همبستر شده بودند به این سؤال جواب نخواهند داد. برای شبیه‌سازی این مورد، مقادیر گمشده را برای این متغیر با استفاده از اعداد تصادفی دارای توزیع لوژستیکی با پارامترهایی تهیه کردیم که اگر کسی تعداد زیادی همبستر داشت احتمال داشتن مقدار گمشده در مورد این متغیر *NSEXPART* برای او بیشتر باشد.

از شـرح و بسـط بـالا مـتوجه مـی‌شویم که سازوکار «گمشدگی» با توجه به متغیر*HBSAG* از نوع MCAR، گمشدهٔ کاملاً تصادفی است، زیرا احتمال شرطی برای گمشده بودن این متغیر در یک عنصر نمونـه نـه به مقدار واقعی متغیر مزبور و نه به مقدار سایر متغیرهای موجود در مجموعهٔ داده‌ها بستگی دارد. بـرعکس، احـتمال شـرطی برای وجود مقدار گمشده در مورد متغیر *PSA* در یک عنصر نمونه به مقـدار واقعـی آن مـتغیر بسـتگی ندارد، ولی به مقدار متغیرهای جنس *SEX* و سن *AGE* بستگی دارد. بـنابراین، سـازوکار بـرای مقادیـر گمشـدهٔ مربوط به متغیر *PSA* از نوع گمشدهٔ تصادفی، MAR، است (ولـی نـه MCAR، گمشـدهٔ کـاملاً تصـادفی). سـرانجام، سـازوکار مقادیـر گمشـده در مـورد مـتغیر *NSEXPART* نـه از نـوع گمشـدهٔ کاملاً تصادفی، MCAR، است و نه از نوع MAR، گمشدهٔ تصادفی. زیـرا احـتمال شـرطی بـرای وجـود مقدار گمشده در مورد متغیر مزبور در یک عنصر نمونه بستگی به مقـدار واقعـی این متغیر دارد (زیرا تعداد همبسترها احتمالاً سؤالی بسیار حساس است). به این ترتیب، گمشدگی در مورد این متغیر *غیر قابل چشم‌پوشی* است.

میـانگین مقادیر این متغیرها همراه با میانگین مقادیر مزبور اگر مقادیر گمشده وجود نداشته باشد در پایین نشان داده می‌شود:

متغیر	مشاهدات آمارگیری در مورد متغیر	نرخ شیوع یا میانگین حاصل از داده‌های آمارگیری	نرخ شیوع یا میانگین واقعی
نرخ شیوع پادگن سطحی هپاتیت B (HBsAG)	۱٤۸۹	٪٤/۸	٪۵/۱
تعداد همبسترها طی سال گذشته	۲۳٤۳	۱/۲۹	۱/٦۰
نرخ شیوع افزایش پادگن ویژهٔ پروستات (PSA) برای مردان بالاتر از ۵۰ سال	۵۸۰	۳۸/۳	۳۸/۳
نرخ شیوع افزایش پادگن ویژهٔ پروستات (PSA) برای همهٔ مردان	۵۸۰	۳۸/۳	۲۷/۱

از مطالب بالا متوجه می‌شویم که نرخ شیوع HBsAG ی مشاهده شده در آمارگیری نزدیک به مقداری است که در صورت اندازه‌گیری همهٔ افراد به دست می‌آمد. این همان است که انتظار می‌رفت، زیرا HBsAG گمشدهٔ کاملاً تصادفی (MCAR) است. از سوی دیگر، میانگین تعداد همبسترها میانگین واقعی را تا نزدیک به ۲۰٪ کم برآورد می‌کند که این نیز دور از انتظار نیست، زیرا سازوکار تولید مقادیر گمشده، آزمودنیهای دارای تعداد زیادی همبستر را بیشتر در معرض احتمال داشتن مقادیر گمشده قرار می‌دهد. سرانجام، نرخ شیوع بالارفتن پادگن ویژهٔ پروستات PSA (از طریق ساختن) برابر است با نرخ شیوع واقعی آن در مردان بالای ۵۰ سال، ولی اگر این گروه سنی نمایندهٔ همهٔ مردان تلقی شود نرخ شیوع بیش برآورد می‌شود.
□

بسیاری از روشهای کار با داده‌ها تلویحاً حاکی از آنند که داده‌ها یا گمشدهٔ کاملاً تصادفی،MCAR، هستند یا گمشدهٔ تصادفی MAR، و می‌توانند به برآوردهایی اریب برای متغیرهایی منجر شوند که نه گمشدهٔ تصادفی هستند و نه گمشدهٔ کاملاً تصادفی. مثلاً اگر تحلیل فقط به آن عناصری محدود بود که هیچ‌یک از متغیرهای آن دارای مقادیر گمشده نبود، برآورد نرخ شیوع پادگن سطحی هپاتیتB (HBsAG) نااریب می‌شد (چون گمشده کاملاً تصادفی،MCAR است)، ولی برآورد نرخ شیوع نابهنجاری پادگن ویژهٔ پروستات PSA اریب می‌شد، زیرا مجموعه داده‌های به دست آمدهٔ همهٔ مردان کمتر از ۵۰ سال را که کمتر احتمال داشتن مقادیر مربوط به نابهنجاری پروستات را دارند حذف می‌کند. برآورد کردن میانگین تعداد همبسترهای سال گذشته نیز در معرض اریبی خواهد بود زیرا کسانی که تعداد زیادی همبستر داشته‌اند بیشتر احتمال دارد که به این سؤال جواب ندهند.

خواننده می‌تواند برای بحث کاملتر در مورد اثری که سازوکار مقادیر گمشده بر اعتبار برآوردها دارد به آثار لیتل و روبین [۲] و لیتل [٦] رجوع کند.

۲.۵.۱۳ برخی روشها برای تحلیل داده‌ها با وجود مقادیر گمشده

لیتل و روبین [۲] چهار ردهٔ کلی زیر را برای شیوه‌های اجرای تحلیل آماری با وجود بی‌پاسخیهای سؤال، فهرست کرده‌اند.

- *روشهای مربوط به مورد کامل*. این روشها همهٔ واحدهایی را که دارای مقادیر گمشده دربارهٔ هر یک از متغیرهای مورد استفاده در آن تحلیل ویژه‌اند کنار می‌گذارد. شاید این متداولترین روش برای حل مشکل داده‌های ناقص باشد. این روش، حذف همهٔ واحدهای تحلیل دارای اقلام گمشده را که در آن تحلیل خاص مورد استفاده قرا رمی‌گیرند به دنبال دارد. هر چند که این کاربست در حدی وسیع به کار می‌رود و کاربرد فوری روشهای آماری مستلزم داده‌های کامل را میسر می‌سازد،

دلایـلی چـند وجـود دارنـد کـه چرا در کل، شیوهٔ خوبی تلقی نمی‌شود. اول از همه، با حذف همهٔ واحدهـای دارای داده‌هـای گمشـده، انـدازهٔ نمونهٔ موجود به صورتی قابل ملاحظه کوچک شده و افـتی ملازم در دقت نتیجه می‌شود. دوم، اگر افرادی که از تحلیل حذف می‌شوند با آنهایی که باقی می‌مانند تفاوت زیادی داشته باشند، ممکن است برآوردهای حاصل به شدت اریب شوند. سوم، در بعضی طرحهای نمونه‌گیری پیچیده، به افراد وزنهایی آماری،W ، داده می‌شود که از جمله می‌توانند بازتـاب احـتمالهای انتخاب شدن آنان باشند. حذف واحدها براساس مقادیر گمشده، احتمال بسیار دارد که اعتبار این طرحهای وزن‌دار کردن را خنثی کند.

- *روشـهای مبتـنی بـر جـانهی.* این روشـــها معمولاً بر مبنـای اطلاعـاتی معین دربارهٔ متغیـرهایی که تصـور مـی‌رود بـا مقادیـر گمشـــده ارتـباط داشـتـه باشنـد همهٔ مقادیر گمشـده را *جانهی* یـا «پر مـی‌کنند». همین که مقادیر گمشده جانهی شدند، روشهای تحلیلی که مستلزم داده‌های کامل دربارهٔ همـهٔ متغیرها هستند (مانند رگرسیون چندگانه، همبستگی) برای تحلیل داده‌ها به کار برده می‌شوند. روشـهای مبتـنی بـر جـانهی بـه خوبـی مورد استقبال قرا رگرفته و طی سالها در سطحی وسیع در آمارگیـریهای سراسـری عمـده و همچـنین آمارگیـریهای کوچکـتر مورد استفاده قرار گرفته‌اند. در مبحث بعد، توجه را بر این دسته از روشها متمرکز خواهیم کرد.
- *روشـهای تجدید وزن‌دهی.* این روشها سعی می‌کنند مقادیر گمشده را با تعدیل وزنهای نمونه‌گیری برای جبران بی‌پاسخی تطبیق دهند. در حالی که این روش غالباً دربارهٔ بی‌پاسخی واحد به کار برده مـی‌شـود بـرای بـی‌پاسخی سؤال نیز می‌تواند با مدل‌سازی احتمال پاسخگویی یک واحد نمونه به سـؤالی خـاص بـه عنوان تابعی از یک مجموعه از متغیرها به کار رود. این کار عموماً با استفاده از یـک مـدل رگرسـیونی لوژسـتیک یا پروبیت انجام می‌گیرد که دارای متغیر وابستهٔ برابر با «۱» برای واحدهـای دارای مقادیـر گـزارش شده دربارهٔ متغیر موردنظر و «۰» برای واحدهای دارای مقادیر گمشـده دربـارهٔ آن سؤال است. متغیرهای مستقل مورد استفاده در این رگرسیون آنهایی هستند که تصـور مـی‌رود بـا گمشدگی اقلام موردنظر ارتباط داشته باشند و نتیجهٔ این رگرسیون آن است که بـرآوردی از احـتمال پاسـخگویی هـر واحد نمونه به سؤال موردنظر تهیه می‌شود. تحلیل به دست آمده فقط از آن دسته از واحدهایی استفاده خواهد کرد که دارای مقادیر گزارش شده (غیر گمشده) دربـارهٔ سـؤالات موردنظرند و هر واحد را با وارون احتمال برآورد شدهٔ پاسخگویی به سؤال برای آن واحـد (عـلاوه بـر هـر گونه وزن نمونه‌گیری دیگری که معمولاً مورد استفاده قرار گرفته باشد) وزن‌دار خواهـد کـرد. مـنطق این کار آن است که هر واحدی که نمونه‌گیری شده و به سؤال پاسخ

داده است نمایندهٔ W واحد در جامعه است که W = وارون احتمال انتخاب واحد در نمونه و پاسخگویی به سؤال است.

- *روشهای مبتنی بر مدل.* این شیوه‌ها بیشتر در طی ۱۰–۱۵ سال اخیر ابداع شده‌اند و داده‌های دارای مقادیر گمشده را با مدل‌سازی تابع درستنمایی داده‌های ناقص و استفاده از شیوه‌های درستنمایی ماکزیمم تحلیل می‌کنند. روشهایی برای مقادیر گمشدهٔ MAR (گمشدهٔ تصادفی) و مقادیر گمشدهٔ غیر قابل چشم‌پوشی ابداع شده‌اند. بسیاری از این روشها در کتاب تألیف لیتل و روبین [۲] و در بررسی جدیدتری توسط لیتل [٦] مورد بحث قرار گرفته‌اند. درک و استفادهٔ هوشمندانه از این روشها مستلزم تجربهٔ آماری در سطحی نسبتاً بالاست و به همین دلیل در کتاب حاضر از آنها بحث نخواهیم کرد.

۳.۵.۱۳ برخی روشهای جانهی

در زیر چند روش جانهی را که معمولاً بیش از همه به کار می‌روند شرح می‌دهیم:

جانشین کردن میانگین. یکی از متداولترین روشهای جانهی آن است که $\bar{y}$، میانگین مقدار همهٔ افرادی را که مقادیر مربوط به متغیر ویژهٔ y آنها موجود است به هر فردی که مقدار گمشده برای آن سؤال اطلاعاتی دارد نسبت می‌دهند. این روش، برآوردی از میانگین جامعه برای متغیر مزبور را نتیجه می‌دهد که برابر است با برآوردی که اگر محاسبهٔ میانگین فقط برای آن دسته از افرادی که پاسخ دادهٔ مزبور برای آنان موجود است انجام می‌شد به دست می‌آمد. مزیت این روش آن است که مقدار گمشده را با یک مقدار «مورد انتظار» جایگزین می‌کند که دارای پایداری نسبتاً زیادی است. عیب این روش آن است که واریانسهای محاسبه شده با استفاده از داده‌های جانهی شده با این روش، کم نشان داده می‌شوند و همبستگیهای موجود میان متغیرها نیز گمراه کننده خواهند بود. تظریف این روش مستلزم آن است که افراد در زیرگروههایی رده‌بندی شوند و به جای میانگینهای کل، از میانگینهای مربوط به زیرگروهها برای جانهی استفاده شود. این اصلاح، استفاده از اطلاعات مربوط به تفاوتهای موجود میان زیرحوزه‌ها را با توجه به میانگین سطوح متغیرهایی که قرار است جانهی شوند میسر می‌سازد.

روش بی‌درنگ. این روش که در ادارهٔ سرشماری امریکا کاربرد وسیعی دارد تا حدود زیادی از کم برآورد کردن واریانسهای ذاتی در روشهای جانهی که مقادیر گمشده را با میانگین کل میانگینهای زیرحوزه‌ها پر می‌کنند جلوگیری می‌کند. در روش بی‌درنگ، پروندهٔ داده‌های اولیه طوری جور می‌شود که ترتیب سوابق فردی با ساختار طرح نمونه‌ای تطابق دارد (مثلاً افراد یک خوشه با هم در

پـرونده جای داده می‌شوند). خانه‌ها براساس مقادیر جمعیت‌شناختی منتخب یا سایر متغیرها تعیین می‌شوند. برای هر خانه، یک سری اطلاعات ثبتی تهیه می‌شود که شامل سوابق فردی است که همهٔ مـتغیرهای مربوط به او ثبت شده است. با یک مرور سریع در پروندهٔ داده‌ها، خانهٔ مربوط به هر نفر شناسـایی مـی‌شـود، و اگر متغیر موردنظر برای آن شخص گمشده بود، مقدار ثبتی مربوط به همان خانـه بـا مقدار گمشده جایگزین می‌شود. از سوی دیگر، اگر سابقهٔ فرد کامل باشد، سابقهٔ کامل آن فرد جایگزین سابقه‌ای می‌شود که قبلاً به صورت ثبتی به کار گرفته می‌شد. در برخی آمارگیریها که داده‌هـای گمشدهٔ زیادی دارند، آمارهای ثبتی برخی از خانه‌های معین ممکن است کمتر تغییر کنند. بـه همیـن دلیل، همان مقادیر به طور مکرر جانهی خواهند شد. برای اجتناب از این مسئله می‌توان مـوارد چـندگانه‌ای در آمارهـای ثبـتی در نظر گرفت که چرخشی باشند. این روش را در مثال زیر نشان می‌دهیم.

مثال تشریحی: داده‌های زیر از نمونه‌ای متشکل از ۲۰ زن ۷۵ سال به بالا که از میان افراد ساکن در سه شهرک بازنشستگان مخصوص افراد سالخورده انتخاب شده‌اند در آخرین سالروز تولد جمع‌آوری شده اسـت. متغیرهایی که قرار است مقادیر گمشدهٔ آنها جانهی شوند عبارت‌اند از *تحصیـلات* (۱ = ابتدایی، ۲ = دبیرسـتان، ۳ = کمـی تحصـیلات دانشـگاهی، ٤ = فارغ‌التحصیل دانشگاه)، و *نمرهٔ امتحان خلاصه وضعیت ذهنی* (*MMSE*)[1] که یک آزمون غربالگری برای اختلال‌شناختی است.

شهرک بازنشستگان	ساختمان	آزمودنی	تحصیلات	MMSE
۱	۱	۱	۲	۱۷
۱	۱	۲	–	۱۸
۱	۱	۳	۲	۲۰
۱	۲	۱	٤	–
۱	۲	۲	۳	۲۷
۱	۳	۱	۳	۲۰
۱	۳	۲	۲	۱۸
۲	۱	۱	۱	۱۱
۲	۱	۲	۱	–
۲	۱	۳	۲	۱۳
۲	۱	٤	۲	۱۵
۲	۲	۱	–	–
۲	۲	۲	۲	۱٦
۳	۱	۱	۳	۲٤
۳	۱	۲	۳	۲٦
۳	۲	۱	–	۱۵
۳	۲	۲	۲	۱۷
۳	۳	۱	٤	۲٦
۳	۳	۲	۳	–
۳	۳	۳	۳	۲۱

[1] Mini Mental State Examination □

اینک می‌خواهیم براساس شهرک بازنشستگان، خانه‌ها را تعیین کنیم. بر مبنای این رسته‌بندی، آمارهای ثبتی آغازین چنین‌اند:

شهرک بازنشستگان	تحصیلات	MMSE
۱	۲	۱۷
۲	۱	۱۱
۳	۳	۲٤

اولین مقدار گمشده، آموزش مربوط به آزمودنی دوم در شهرک ۱ است. از روی آمارهای ثبتی بالا، مقداری که باید جانهی شود «۲» است. سابقهٔ بعدی کامل است، بنابراین مقدار مربوط به متغیر تحصیلات «۲» و MMSE «۲۰» برای این سابقه به جای مقدار آغازین این آمار ثبتی قرار داده می‌شود. سابقهٔ بعدی هم در شهرک بازنشستگان اول است و مقدار MMSE در آن گمشده است. از روی آمارهای ثبتی برای آن خانه، «۲۰» جانهی می‌شود. سابقهٔ بعدی که دارای مقدار گمشده است دومین سابقه از شهرک مسکونی شمارهٔ ۲ است. در این نقطه، آمارهای ثبتی مربوط به آن خانه مقدار «۱۱» را برای MMSE نشان می‌دهد و همین مقدار نیز جانهی خواهد شد. بقیهٔ مقادیر گمشده به راهی مشابه جانهی می‌شوند و از این شیوه، مجموعه داده‌هایی حاصل می‌شود که در پایین نشان داده شده‌اند (مقادیر جانهی شده پررنگترند):

شهرک بازنشستگان	ساختمان	آزمودنی	تحصیلات	MMSE
۱	۱	۱	۲	۱۷
۱	۱	۲	**۲**	۱۸
۱	۱	۳	۲	۲۰
۱	۲	۱	٤	**۲۰**
۱	۲	۲	۳	۲۷
۱	۳	۱	۳	۲۰
۱	۳	۲	۲	۱۸
۲	۱	۱	۱	۱۱
۲	۱	۲	۱	**۱۱**
۲	۱	۳	۲	۱۳
۲	۱	٤	۲	۱۵
۲	۲	۱	**۲**	**۱۵**
۲	۲	۲	۲	۱٦
۳	۱	۱	۳	۲٤
۳	۱	۲	۳	۲٦
۳	۲	۱	**۳**	۱۵
۳	۲	۲	۲	۱۷
۳	۳	۱	٤	۲٦
۳	۳	۲	۳	**۲٦**
۳	۳	۳	۳	۲۱

□

روش بی‌درنگ صورتهای زیادی دارد. در یکی از این صورتها، آمارهای ثبتی در هر زمان شامل مقادیری از بیش از یک سابقهٔ کامل‌اند و این مقادیر در فرایند جانهی چرخش دارند. این صورت هنگامی مفید واقع می‌شود که داده‌های گمشده بسیار زیادند چون آمارهای ثبتی که در روش بی‌درنگ اصلاح نشده ساخته می‌شوند در هر زمان شامل مقادیری از تنها یک سابقهٔ کامل‌اند، و هنگامی که بسیاری از سوابق کامل نباشند، آمارهای ثبتی گردشی گهگاه خواهند داشت و در نتیجه همان مقادیر به طور مکرر جانهی خواهند شد. صورتی دیگر که به عنوان روش *بادرنگ* موسوم است مانند شیوهٔ بی‌درنگ، خانه‌هایی را شناسایی می‌کند ولی آمارهای ثبتی خانه‌ها از سوابقی تشکیل شده‌اند که عموماً از آمارگیری دیگری غیر از آمارگیری جاری جمع‌آوری شده‌اند. غالباً، این آمارگیری دیگر ممکن است یک آمارگیری قبلی از همین جامعه باشد. برای هر یک از خانه‌های تعیین شده باید حداقل یک مورد از روی آمارگیری قبلی موجود باشد. هرگاه برای خانه‌ای چندین سابقه وجود داشته باشند، انتخاب یکی از آنها برای جانهی در آمارگیری جاری به طور تصادفی صورت می‌گیرد. عیب روش بادرنگ آن است که از داده‌های آمارگیری جاری برای جانهی استفاده نمی‌کند و به همین دلیل جاذبهٔ آن کمتر از روش بی‌درنگ به نظر می‌رسد.

جانهی از روی افرادی که به طور تصادفی انتخاب شده‌اند. این روش مشابه روش بی‌درنگ است جز این که مقادیر مربوط به فردی که به تصادف انتخاب شده است با مقادیر گمشده در سابقهٔ ناقص جایگزین می‌شوند. حتی اگر انتخاب تصادفی در داخل خانه‌هایی، شبیه به آنچه در روش بی‌درنگ تعیین شده است، روی دهد باز هم این شیوه فاقد ویژگی جذاب روش بی‌درنگ، یعنی سود جستن از نظم ذاتی پرونده در فرایند جانهی است. این ویژگی، روش بی‌درنگ را قادر می‌سازد تا از مزیت مهم منطقه‌ای و جغرافیایی سود جوید که ممکن است در ساختمان خانه‌های ثبتی مورد استفاده قرار نگرفته باشند. همچنین، با توجه به سهولت محاسبات، استفاده از روش بی‌درنگ راحت‌تر است.

رگرسیون ذهنی. همان‌طور که از نام آن پیداست، این فرایندی به طور شهودی استنتاجی است که صرفاً بر قضاوت تحلیلگر براساس شواهد موجود متکی است. مثلاً اگر جنسیت یک پاسخگوی نمونه، مقدار گمشده باشد، داده‌های ثبت شده بر روی ابزار آمارگیری دربارهٔ سن در زمان نخستین قاعدگی، تحلیلگر را هدایت می‌کند تا واژهٔ «زن» را جانهی کند. به صورتی مشابه، یک تحلیلگر آشنا با اندازه‌های بدن انسان می‌تواند برآوردی معقول از قد را با توجه به وجود سایر داده‌ها از قبیل سن و جنس تعیین کند.

رگرسیون عینی. سرانجام دربارهٔ روشی بحث می‌کنیم که می‌توان آن را «رگرسیون عینی» نامید. در این روش، معادلات رگرسیونی از مجموعه‌ای از داده‌ها تولید می‌شوند که شامل سوابق کامل‌اند و متغیری که باید جانهی شود به عنوان متغیر وابسته عمل می‌کند. معادلهٔ حاصل ممکن است به صورت زیر باشد

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

که در آن، y متغیری است که باید برای فردی معین جانهی شود و x_1، ... ، x_k متغیرهای کمکی معلوم برای آن شخص‌اند. مقادیری که قرار است به این شیوه جانهی شوند ممکن است نسبت به آنهایی که از روشهای توصیف شدهٔ قبل به دست آمده‌اند برتر باشند، زیرا از اطلاعات مربوط که تصور می‌شد با متغیر مورد جانهی در ارتباط باشند استفاده شده است. این روش همچنین «هماهنگی» مقدار جانهی شده با بقیهٔ سابقهٔ آزمودنی را میسر می‌سازد. همچنین می‌تواند با استفاده از معادلهٔ رگرسیونی برای برآورد مقدار، و سپس به وسیلهٔ تولید یک متغیر تصادفی دارای توزیع نرمال که همان میانگین و واریانس خطای موجود در داده‌ها را داشته باشد ثبت خطای تصادفی را میسر سازد. به این ترتیب ساختار خطای ذاتی در داده‌ها حفظ می‌شود.

۶.۱۳ جانهی چندگانه

روشهایی که در بالا مورد بحث قرار گرفتند به جای هر مقدار گمشده یک مقدار جانهی شدهٔ تک را قرار می‌دهند و تحلیل حاصل، با مقدار جانهی شده به همان طریقی رفتار می‌کند که با مقادیری که واقعاً اندازه‌گیری شده‌اند رفتار می‌شود. در تحلیل، عدم حتمیت دربارهٔ این مقدار گمشده در نظر گرفته نمی‌شود. تا حدودی به دلیل همین مسئله، فنی به نام *جانهی چندگانه* توسط روبین [۱] و سایر پژوهشگران طی دههٔ گذشته ابداع شده است. این روش، جای هر مقدار گمشده را با دو یا چند مقدار جانهی شده پر می‌کند و هر مجموعه از داده‌های به دست آمده را با استفاده از روشهای مربوط به داده‌های کامل تحلیل می‌کند. با ترکیب نتایج این تحلیلها، استنباطهایی به دست می‌آید که می‌توانند عدم حتمیت دربارهٔ مقادیر گمشده را در نظر بگیرند. برتری این روش آن است که کاملاً بر نظریه آماری زمینه‌سازی شده است و کاربرد عملی آن قابل ملاحظه است زیرا می‌تواند از روشهای محاسباتی روز بهره جوید. با این که روش مزبور یک روش جانهی است، به علت تفاوت زیادی که با روشهای بحث شدهٔ قبلی دارد، آن را به صورت یک موضوع «مستقل» ارائه می‌دهیم. مثال تشریحی سادهٔ زیر را در نظر بگیرید.

مثال تشریحی: بیایید همان مثالی را که در بخش ۳.۵.۱۳ نشان دادیم در نظر بگیریم.

شهرک بازنشستگان	ساختمان	آزمودنی	تحصیلات	MMSE
۱	۱	۱	۲	۱۷
۱	۱	۲	–	۱۸
۱	۱	۳	۲	۲۰
۱	۲	۱	٤	–
۱	۲	۲	۳	۲۷
۱	۳	۱	۳	۲۰
۱	۳	۲	۲	۱۸
۲	۱	۱	۱	۱۱
۲	۱	۲	۱	–
۲	۱	۳	۲	۱۳
۲	۱	٤	۲	۱۵
۲	۲	۱	–	–
۲	۲	۲	۲	۱٦
۳	۱	۱	۳	۲٤
۳	۱	۲	۳	۲٦
۳	۲	۱	–	۱۵
۳	۲	۲	۲	۱۷
۳	۳	۱	٤	۲٦
۳	۳	۲	۳	–
۳	۳	۳	۳	۲۱

فرض کنید می‌خواهیم چهار مقدار گمشده برای MMSE را جانهی کنیم و می‌توانیم مقدار گمشده را با مقدار آن متغیر برای هر آزمودنی در همان واحد نمونه‌گیری اولیه (ساختمان) جایگزین کنیم. مقادیر گمشده برای آزمودنی ۱ در واحد نمونه‌گیری اولیهٔ ۲ از طبقهٔ ۱ (که به صورت آزمودنی A نشان داده می‌شود) می‌تواند با مقدار مربوط به آزمودنی دیگر جایگزین شود زیرا در این واحد نمونه‌گیری اولیه فقط دو آزمودنی نمونه‌گیری شده‌اند. مقدار گمشده برای آزمودنی ۲ در واحد نمونه‌گیری اولیهٔ ۱ از طبقهٔ ۲ (که به صورت آزمودنی B نشان داده می‌شود) می‌تواند با مقادیر مربوط به سه آزمودنی دیگر در همان واحد نمونه‌گیری اولیه که مقدار متغیر موردنظر برای آنها گمشده نیست جایگزین شوند. مقدار گمشده برای آزمودنی ۱ در واحد نمونه‌گیری اولیهٔ ۲ از طبقهٔ ۲ (که به صورت آزمودنی C نشان داده می‌شود) می‌تواند تنها با یک مقدار دیگر جایگزین شود و مقدار گمشده برای آزمودنی ۲ در واحد نمونه‌گیری اولیهٔ ۳ از طبقهٔ ۳ (که به صورت آزمودنی D نشان داده می‌شود) می‌تواند با مقادیر مربوط به دو آزمودنی دیگر که داده‌های آنها در مورد متغیر موردنظر در واحد نمونه‌گیری مزبور کامل است جایگزین شود. به این ترتیب، تعداد ترکیبهای آزمودنیهای مختلف که مقادیر آنها می‌توانند در جانهی به کار روند به صورت ٦ = ۲ × ۱ × ۳ × ۱ به دست می‌آید. این ٦ گزینه به شکل زیرند:

ترکیب	آزمودنیها			
	A	B	C	D
١	٢٧	١١	١٦	٢٦
٢	٢٧	١١	١٦	٢١
٣	٢٧	١٣	١٦	٢٦
٤	٢٧	١٣	١٦	٢١
٥	٢٧	١٥	١٦	٢٦
٦	٢٧	١٥	١٦	٢١

فرض کنید یک نمونهٔ تصادفی ساده متشکل از سه تا از این ترکیبها را انتخاب می‌کنیم (مثلاً ٢، ٣ و ٥). مجموعـه داده‌های کامل دربارهٔ متغیر MMSE برای هر یک از این ترکیبها که مقدار گمشدهٔ آنها جانهی شده است در زیر ارائه می‌شود:

آزمودنی	ترکیب		
	١	٢	٣
١	١٧	١٧	١٧
٢	١٨	١٨	١٨
٣	٢٠	٢٠	٢٠
٤	٢٧	٢٧	٢٧
٥	٢٧	٢٧	٢٧
٦	٢٠	٢٠	٢٠
٧	١٨	١٨	١٨
٨	١١	١١	١١
٩	١١	١٣	١٥
١٠	١٣	١٣	١٣
١١	١٥	١٥	١٥
١٢	١٦	١٦	١٦
١٣	١٦	١٦	١٦
١٤	٢٤	٢٤	٢٤
١٥	٢٦	٢٦	٢٦
١٦	١٥	١٥	١٥
١٧	١٧	١٧	١٧
١٨	٢٦	٢٦	٢٦
١٩	٢١	٢٦	٢٦
٢٠	٢١	٢١	٢١
میانگین $(\bar{x})$	١٨/٩٥	١٩/٣٠	١٩/٤٠
خطای معیار $\hat{SE}(\bar{x})$	١/١٣	١/١٢	١/١٢

توجه کنید که هر یک از مجموعه داده‌هایی که به شرح بالا تولید می‌شود برآورد متفاوتی از $\bar{x}$، میانگین نمرهٔ MMSE و برآوردی متفاوت برای $\hat{SE}(\bar{x})$، خطای معیار میانگین به دست می‌دهد. اگر می‌خواستیم استنباطی از میانگین جامعه‌ای نامعلوم، $\bar{X}$، با در نظر گرفتن عدم حتمیت مقادیر جانهی شده تهیه کنیم، می‌توانستیم از روی میانگینها و واریانسهای محاسبه شده، یک مؤلفهٔ «درونی» به دست آوریم که معرف برآورد واریانس میانگین برآورد شدهٔ شرطی روی مقادیر جانهی شده باشد و نیز یک مؤلفهٔ «میانی» به دست آوریم که معرف واریانس برآورد میانگینها در تحققهای متفاوت فرایند جانهی باشد. سپس آنها را ترکیب می‌کردیم تا برآوردی از واریانس کل به دست آوریم. محاسبات مشخص برای این منظور به شرح زیرند.

مؤلفهٔ واریانس درونی

$$s_w^2 = \frac{(۱/۱۳^2 + ۱/۱۲^2 + ۱/۱۲^2)}{۳} = ۱/۲۶$$

مؤلفهٔ واریانس میانی

$$s_b^2 = \frac{\left(۱ + \frac{۱}{k}\right)\sum_{i=۱}^{k}(\bar{x}_i - \bar{x})^2}{(k-۱)}$$

که در آن k تعداد تحققها (در این مورد $k=۳$)، $\bar{x}_i$ برآورد میانگین برای i امین تحقق جانهی، و $\bar{x}$ میانگین این $\bar{x}_i$ هاست. برای این داده‌ها، داریم

$$\bar{x} = \frac{(۱۸/۹۵ + ۱۹/۳۰ + ۱۹/۴۰)}{۳} = ۱۹/۲۲$$

$$s_b^2 = \left(\frac{۴}{۳}\right) \times ۰/۰۵۵۸ = ۰/۰۷۴$$

و واریانس کل و خطای معیار میانگین برآورد شده به صورت زیر به دست می‌آید

$$Var(\bar{x}) = ۰/۰۷۴ + ۱/۲۶ = ۱/۳۳۴$$

$$SE(\bar{x}) = \sqrt{۱/۳۳۴} = ۱/۱۵$$

توجه کنید که در این مثال، خطای معیار برآورد شده از فرایند جانهی چندگانه فقط اندکی بیشتر از خطای معیاری است که با استفاده از تنها هر یک از سه ترکیب بالا برآورد می‌شود (۱/۱۵ در برابر ۱/۱۳ یا ۱/۱۲). این امر نشان می‌دهد که در میان میانگینهای تخصیص‌پذیر به مقادیر ویژه‌ای که نسبت به تغییرپذیری در میانگینهای تخصیص‌پذیر به واریانس نمونه‌گیری توزیع نمره‌های MMSE جانهی شده‌اند تغییرپذیری نسبتاً ناچیزی وجود دارد.

□

بحث بالا دربارهٔ جانهی چندگانه صرفاً برای معرفی موضوع است و شرحی ساده از نحوهٔ استفاده از آن را ارائه می‌دهد. بحث جامعتر مستلزم آشنایی با روشهای استنباط بیزی و سایر مفاهیم نظری

است که خارج از دیدگاه این کتاب است. خوانندهٔ علاقمند به پی‌گیری این زمینه، می‌تواند به متونی که روبین [۱] و لیتل و روبین [۲] تألیف کرده‌اند رجوع کند.

۷.۱۳ خلاصه

در این فصل دربارهٔ مشکلات ناشی از داده‌های ناقص در آمارگیریهای نمونه‌ای بحث کردیم. به خصوص، اثر بی‌پاسخی بر برآوردهای مشاهده شده در آمارگیریهای نمونه‌ای خانوار را شرح دادیم؛ دلایل بروز بی‌پاسخی را مورد بحث قرار دادیم، و روشهای متعددی را برای افزایش نرخ پاسخگویی در این قبیل آمارگیریها پیشنهاد کردیم. در یک طرح نمونه‌ای که در سطحی وسیع به کار می‌رود، پرسشنامه‌هایی برای خانوارهای نمونه با پست ارسال می‌شود و سپس داده‌های مربوط به زیرنمونه‌ای از بی‌پاسخها، با تلاشی متمرکز جمع‌آوری می‌شود. این طرح نمونه‌ای، اگر به طرز مناسبی اجرا شود، به برآوردهایی منتهی می‌شود که با بی‌پاسخی به صورتی جدی اریب نشده‌اند. سرانجام، دربارهٔ مشکل اقلام گمشده در یک پرونده از داده‌ها بحث کردیم و چندین روش جانهی، از جمله جانهی چندگانه را برای تبدیل یک مجموعه داده‌های ناقص به یک مجموعهٔ کامل شرح دادیم.

تمرین

۱.۱۳ در محله‌ای با ۲۰۰ خانوار، مطلوب است اجرای یک آمارگیری پستی به منظور برآوردکردن X (تعداد افراد ۱۸ تا ٦٤ سال سن در محله)، R_1 (نسبت افراد شاغل در میان همهٔ افراد ۱۸ تا ٦٤ سال)، R_2 (متوسط تعداد روزهای کاری تلف شده به ازای هر نفر در سال برای شاغلان ۱۸ تا ٦٤ سال). از روی مجموعه‌ای از ۱۰ خانوار در یک محلهٔ مشابه در همان حوالی، داده‌های زیر به دست آمده‌اند. نمونهٔ موردنظر چقدر باید بزرگ باشد تا واقعاً مطمئن شویم که X را در محدودهٔ ۲۰٪، R_1 را در محدودهٔ ۳۰٪ و R_2 را تا ۲۵٪ مقدار آن برآورد می‌کنیم؟

خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته
الف	٤	۳	۱
ب	۲	۱	۱
پ	۳	۲	۲
ت	۱	۰	۰
ث	۱	۱	۲
ج	۰	۰	۰
چ	۲	۲	۳
ح	٥	٤	۲
خ	۰	۰	۰
د	۲	۱	۲

جدول ۲.۱۳ داده‌های به دست آمده از آمارگیری پستی

خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته	خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته
۱	۰	۰	۰	۲٤	۲	۱	۱
۲	۲	۱	۱	۲۵	*		
۳	*			۲٦	*		
٤	*			۲۷	*		
۵	۱	۱	۰	۲۸	*		
٦	*			۲۹	۲	۰	۰
۷	٤	۱	۰	۳۰	۳	۱	۱
۸	۱	۰	۰	۳۱	٤	۳	۲
۹	٤	۰	۰	۳۲	۳	۳	٦
۱۰	۳	۳	۵	۳۳	۳	۱	۱
۱۱	*			۳٤	۵	۱	۱
۱۲	۰	۰	۰	۳۵	۳	۲	۳
۱۳	۲	۱	۱	۳٦	*		
۱٤	۳	۰	۰	۳۷	٤	۰	۰
۱۵	٤	۳	٦	۳۸	۲	۲	۳
۱٦	*			۳۹	٤	۱	۱
۱۷	۲	۰	۰	٤۰	۵	۱	۱
۱۸	۵	۳	۲	٤۱	۵	۱	۰
۱۹	۲	۲	۱	٤۲	*		
۲۰	۱	۰	۰	٤۳	*		
۲۱	*			٤٤	۳	۳	۵
۲۲	۲	۲	۱	٤۵	۳	۰	۰
۲۳	۰	۰	۰	٤٦	٤	۰	۰

٤٧	٥	١	٠	٧٤	٣	٢	١
٤٨	٢	٢	٠	٧٥	*		
٤٩	٥	١	٠	٧٦	*		
٥٠	٥	٢	٣	٧٧	٠	٠	٠
٥١	٤	٠	٠	٧٨	٢	٠	٠
٥٢	٠	٠	٠	٧٩	*		
٥٣	٠	٠	٠	٨٠	١	١	١
٥٤	٠	٠	٠	٨١	٣	١	١
٥٥	٣	٢	١	٨٢	٤	٢	١
٥٦	٤	١	١	٨٣	٥	٠	٠
٥٧	١	١	١	٨٤	٣	٠	٠
٥٨	٠	٠	٠	٨٥	٤	٣	١
٥٩	٠	٠	٠	٨٦	٣	٣	١
٦٠	*			٨٧	٥	١	٠
٦١	*			٨٨	*		
٦٢	*			٨٩	*		
٦٣	٣	٣	٤	٩٠	١	١	٢
٦٤	٣	٣	١	٩١	٢	٢	٤
٦٥	٢	١	٢	٩٢	١	٠	٠
٦٦	٢	٢	١	٩٣	*		
٦٧	٥	٠	٠	٩٤	٢	١	١
٦٨	*			٩٥	*		
٦٩	٢	٠	٠	٩٦	٤	٢	١
٧٠	٤	٠	٠	٩٧	٢	٢	١
٧١	٢	١	١	٩٨	٤	٢	٠
٧٢	١	٠	٠	٩٩	٠	٠	٠
٧٣	*			١٠٠	*		

جدول ۲.۱۳ داده‌های به دست آمده از آمارگیری پستی (ادامه)

خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته	خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته
۱۰۱	۱	۱	۲	۱۲٤	*		
۱۰۲	۵	۰	۰	۱۲۵	*		
۱۰۳	۱	۰	۰	۱۲٦	۰	۰	۰
۱۰٤	۲	۱	۱	۱۲۷	۳	۱	۰
۱۰۵	*			۱۲۸	۵	٤	۵
۱۰٦	٤	۱	۱	۱۲۹	*		
۱۰۷	۱	۱	۱	۱۳۰	۳	۱	۱
۱۰۸	*			۱۳۱	۳	۱	۱
۱۰۹	۵	۱	۲	۱۳۲	۰	۰	۰
۱۱۰	۳	۲	٤	۱۳۳	٤	۱	۱
۱۱۱	٤	۲	۱	۱۳٤	٤	۳	۵
۱۱۲	۰	۰	۰	۱۳۵	۳	۱	۲
۱۱۳	*			۱۳٦	٤	۳	۱
۱۱٤	۳	۲	۳	۱۳۷	۱	۱	۰
۱۱۵	*			۱۳۸	۵	٤	٦
۱۱٦	٤	۳	۳	۱۳۹	۲	۲	۳
۱۱۷	٤	۳	۰	۱٤۰	٤	۱	۰
۱۱۸	۳	۲	۳	۱٤۱	۲	۱	۲
۱۱۹	۰	۰	۰	۱٤۲	٤	۱	۲
۱۲۰	۰	۰	۰	۱٤۳	۳	۱	۰
۱۲۱	٤	۳	۳	۱٤٤	*		
۱۲۲	۱	۱	۰	۱٤۵	*		
۱۲۳	۰	۰	۰	۱٤٦	۳	۳	۲

۱۴۷	۰	۰	۰	۱۷۴	۲	۱	۱
۱۴۸	۳	۰	۰	۱۷۵	*		
۱۴۹	۱	۱	۰	۱۷۶	*		
۱۵۰	۵	۴	۴	۱۷۷	*		
۱۵۱	۰	۰	۰	۱۷۸	۴	۲	۲
۱۵۲	۵	۴	۲	۱۷۹	۲	۱	۱
۱۵۳	۱	۰	۰	۱۸۰	۱	۰	۰
۱۵۴	۵	۵	۴	۱۸۱	۴	۲	۴
۱۵۵	*			۱۸۲	*		
۱۵۶	۲	۱	۰	۱۸۳	۳	۱	۱
۱۵۷	*		۰	۱۸۴	۳	۲	۰
۱۵۸	۱	۰	۰	۱۸۵	۲	۱	۱
۱۵۹	۴	۲	۲	۱۸۶	۴	۳	۱
۱۶۰	۴	۰	۰	۱۸۷	۴	۱	۱
۱۶۱	۵	۲	۳	۱۸۸	*		
۱۶۲	۳	۱	۱	۱۸۹	۳	۱	۱
۱۶۳	۳	۱	۱	۱۹۰	*		
۱۶۴	۲	۱	۰	۱۹۱	*		
۱۶۵	*			۱۹۲	۳	۲	۰
۱۶۶	۲	۰	۰	۱۹۳	*		
۱۶۷	۴	۲	۰	۱۹۴	*		
۱۶۸	۰	۰	۰	۱۹۵	۲	۲	۲
۱۶۹	۳	۲	۳	۱۹۶	۳	۲	۳
۱۷۰	*			۱۹۷	*		
۱۷۱	۰	۰	۰	۱۹۸	۴	۳	۳
۱۷۲	۳	۱	۱	۱۹۹	۴	۲	۳
۱۷۳	*			۲۰۰	۴	۳	۳

* اطلاعات گمشدهٔ مربوط به این خانوار (یعنی بی‌پاسخها). نگاه کنید به جدول ۳.۱۳.

جدول ۳.۱۳ مقادیر واقعی برای داده‌های گمشده در جدول ۲.۱۳

خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته	خانوار	افراد ۱۸ تا ٦٤ سال	شاغلان ۱۸ تا ٦٤ سال	روزهای کاری تلف شده در ماه گذشته
۳	۳	۲	۲	۱۰۰	۱	۰	۰
٤	۳	۲	۲	۱۰۵	۲	۲	۱
٦	۲	۱	۰	۱۰۸	۱	۱	۰
۱۱	۱	۱	۱	۱۱۳	۳	۳	۲
۱٦	۲	۲	۱	۱۱۵	۰	۰	۰
۲۱	۳	۲	۰	۱۲٤	۲	۲	۱
۲۵	۱	۱	۰	۱۲۵	۱	۱	۰
۲٦	۲	۲	۱	۱۲۹	۲	۲	۲
۲۷	۱	۱	۰	۱٤٤	۱	۱	۱
۲۸	۲	۲	۲	۱٤۵	۲	۲	۲
۳٦	۱	۱	۱	۱۵۵	۱	۱	۱
٤۲	۱	۱	۰	۱۵۷	۲	۲	۱
٤۳	۲	۲	۱	۱٦۵	۱	۱	۰
٦۰	۳	۳	۳	۱۷۰	۲	۲	۲
٦۱	۳	۲	۲	۱۷۳	۲	۲	۲
٦۲	۲	۲	۱	۱۷۵	۲	۲	۱
٦۸	۱	۱	۰	۱۷٦	۳	۳	۲
۷۳	۱	۱	۰	۱۷۷	۲	۱	۰
۷۵	۲	۲	۲	۱۸۲	۲	۲	۰
۷٦	۱	۱	۱	۱۸۸	۲	۲	۲
۷۹	۲	۲	۲	۱۹۰	۳	۳	۱
۸۸	۳	۳	۱	۱۹۱	۲	۱	۱
۸۹	۱	۱	۰	۱۹۳	۳	۳	۱
۹۳	۰	۰	۰	۱۹٤	۳	۲	۱
۹۵	۳	۲	۰	۱۹۷	۳	۲	۱

۲.۱۳ برای داده‌های تمرین ۱.۱۳، فرض کنید ۲۰٪ جامعه به پرسشنامه پاسخ نمی‌دهند. و باز فرض کنید هزینهٔ ارسال پستی آغازین برای هر پرسشنامه ۱/۵۰ دلار، هزینهٔ پردازش پرسشنامه‌های پستی برای هر پرسشنامه ۱۰/۰۰ دلار، و هزینهٔ مصاحبه با بی‌پاسخها هر یک ٦۰/۰۰ دلار (هزینهٔ انجام مصاحبه و پردازش داده‌ها) است. تعیین کنید نسبت بی‌پاسخهایی را که باید نمونه‌گیری شوند و کل تعداد پرسشنامه‌های آغازین را که باید پست شوند.

۳.۱۳ یک نمونهٔ تصادفی ساده از تعداد مورد نیاز خانوارها (به صورتی که در تمرین ۲.۱۳ محاسبه شده است) انتخاب کنید و از روی داده‌هایی که در جدولهای ۲.۱۳ و ۳.۱۳ ارائه شده‌اند با استفاده از نمونه‌گیری مناسب از بی‌پاسخها، X ، R_1، و R_2 را برآورد کنید.

٤.۱۳ داده‌هایی را که در تمرین مربوط به نشان دادن روش بی‌درنگ ارائه شدند در نظر بگیرید. مقادیر گمشده برای MMSE را با استفاده از روش بی‌درنگ به صورتی که در آن مثال به کار رفته است با اصلاحات زیر جانهی کنید. فرض کنید افرادی که دارای مقادیر گمشده‌اند به طور کلی از آنهایی که دارای مقادیر اندازه‌گیری شده از این متغیر بوده‌اند MMSE پایینتری دارند. برای در نظر گرفتن این مطلب، از مقدار جانهی شده‌ای معادل ۸۰٪ مقدار به دست آمده از روش بی‌درنگ استفاده کنید. برآورد میانگین به دست آمده و خطای معیار آن را چگونه با آنچه که از روش بی‌درنگ، بدون این اصلاحات، به دست می‌آمد مقایسه می‌کنید؟

۵.۱۳ مثال تشریحی ارائه شده در بخش ٦.۱۳ را با اصلاحات مورد بحث در تمرین ٤.۱۳ تکرار کنید. برآورد میانگین و خطای معیار آن را که براساس جانهی چندگانه محاسبه می‌شود چگونه با آنچه که در ابتدا به دست می‌آید مقایسه می‌کنید؟

٦.۱۳ از روش بی‌درنگ برای جانهی مقادیر گمشده در داده‌های جدول ۲.۱۳ استفاده کنید. برآورد میانگین به دست آمده و خطای معیار آن را چگونه با آنچه که در صورت پاسخ ۱۰۰٪ به دست می‌آمد مقایسه می‌کنید؟

کتابشناسی

The following books written in the 1980s deal exclusively with issues relating to non-response and statistical analysis with missing data.

1. Rubin, D. B., *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.

2. Little, R. J. A., and Rubin, D. B., *Statistical Analysis With Missing Data*, Wiley, New York, 1987.
3. Kalton, G., *Compensating for Missing Survey Data*, Research Report Series, Institute for Social Research, University of Michigan, Ann Arbor, 1983.

Chapter 4 *of the recent text by Lehtonen and Pahkinen discusses issues relating to the handling of missing data. It gives an especially comprehensive treatment of reweighting methods for unit nonresponse.*

4. Lehtonen, R., and Pahkinen, E. J., *Practical Methods for Design and Analysis of Complex Surveys,* rev. ed., Wiley, Chichester, U.K., 1994.

The following recent reviews and expository articles give overviews of methods used in the handling of missing data due to unit and/or item nonresponse. These reviews are especially useful for their inclusion of references to very recent work on these topics.

5. Dillman, D. A., Call-backs and mail-backs in sample surveys. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester U.K., 1998.
6. Little, R. J., Biostatistical analysis with missing data. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester U.K., 1998.
7. Barnard, J., Rubin, D. B., and Schenker, N., Multiple imputation. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester U.K., 1998.
8. Kviz, F., Nonresponse in sample surveys. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester U.K., 1998.
9. Sudman, S., Response effects in sample surveys. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester U.K., 1998.

The two following classic books give excellent discussions of methods of improving response rates, primarily in mail and telephone surveys.

10. Erdos, P., Professional Mail Surveys, McGraw-Hill, New York, 1970.
11. Dillman, D. A., *Mail and Telephone Surveys. The Total Design Method*, Wiley, New York, 1978.

The following publications are reports of substantive studies which have used methods discussed in this chapter to deal with issues of nonresponse and missing values.

12. Mullner, R., Levy, P. S., Byre, C. S., and Matthews, D., Effects of characteristics of the survey instrument on response rates to a mail survey of community hospitals. *Public Health Reports,* 97: 465, 1982A.
13. National Center for Health Statistics, A Study of the Effect of Renumeration upon Response in the Health and Nutrition Examination Survey, PHS Publication No. 1000, Series 2, No. 67, U.S. Government Printing office, Washington, D.C., 1973.
14. Rich, S., Chomka, E., Hasara, L., Hart, K., Drizd, T., Joo, E., and Levy, P. S., The prevalence of pulmonary hypertension in the U.S. *Chest* 236, 1989.
15. Barr, D., Hershow, R., Levy, P. S., Furner, S., and Handler, A., Assessing prenatal heptatis B screening in Illinois using an inexpensive study design adaptable to other jurisdictions. *American Journal of Public Health*, Jan. 1999.

In January, 1998, *a WINDOWS-based PC program entitled SOLAS became available. This software product can imput missing values using several methods including hot deck and multiple imputation. The users manual includes discussion of each of these methods.*

16. Statistical Solutions Ltd., *SOLAS for Missing Data 1.0,* Statistical Solutions Ltd., Cork, Ireland, 1998.

فصل ۱۴

موضوعهای منتخب در طرح نمونه‌ای و روش‌شناسی برآورد کردن

در این فصل به بحث در مورد برخی فنون نمونه‌گیری و برآورد کردن می‌پردازیم که برای مقاصد خاص ابداع شده‌اند. به صورتی مشخصتر، ابتدا به بحث دربارهٔ برخی روشهای نمونه‌گیری می‌پردازیم که برای استفاده در آمارگیریهای بهداشت که در کشورهای در حال توسعه اجرا می‌شوند ابداع شده‌اند و سپس فنونی را بررسی می‌کنیم که در کشورهای توسعه یافته برای تأمین نیازهایی خاص بسط یافته‌اند.

۱.۱۴ آمارگیریهای برنامهٔ گسترش یافتهٔ ایمن‌سازی سازمان بهداشت جهانی: تعدیلی از نمونه‌گیری با احتمال متناسب با اندازه برای استفاده در کشورهای در حال توسعه

سازمان بهداشت جهانی (WHO)[1]، از طریق برنامهٔ گسترش یافتهٔ ایمن‌سازی (EPI)[2] خود، دسترسی همهٔ کودکان جهان به ایمن‌سازی را هدف قرار داده است. این سازمان به عنوان بخشی از این برنامه، روش‌شناسیی را ابداع کرده است که با استفاده از آمارگیریهای نمونه‌ای نسبتاً سریع و کم‌خرج بتواند

[1] World Health Organization
[2] Expanded Program on Immunization

سطوح ایمن‌سازی کودکان را برآورد کند و در صورت پایین بودن این سطوح، مناطقی را به منظور اجرای برنامه‌های ویژهٔ مصون‌سازی هدف قرار دهد. در این روش، تعدیلی از نمونه‌گیری با احتمال متناسب با اندازه (PPS)[1] به کار می‌رود که در اصل در آمریکا [۱] بسط داده شد و بعدها برای استفاده در برنامهٔ ریشه‌کنی آبله در افریقای غربی [۲] مورد تعدیل قرار گرفت.

با این که آمارگیری برنامهٔ گسترش یافتهٔ ایمن‌سازی می‌تواند برای اهداف دیگری نیز به کار رود، هدف عمده از اجرای آن برآورد کردن پوشش ایمن‌سازی (یعنی نسبت کودکانی که همهٔ ایمن‌سازیهای مورد نیاز را دریافت می‌کنند) در یک ناحیهٔ مشخص هدف است (که می‌تواند یک روستا، شهر، و غیره باشد). طی سالها، کاربستی در این آمارگیریها تکامل یافت که طی آن ۳۰ خوشهٔ نمونه با ۷ کودک از هر خوشه انتخاب و اندازهٔ نمونه از ۲۱۰ کودک تشکیل می‌شد. در واقع، این آمارگیریها به آمارگیریهای «۷ × ۳۰» موسوم شدند.

با این که تاکنون صدها آمارگیری «۷ × ۳۰» اجرا شده است، منطق انتخاب اندازهٔ نمونه، نسبتاً غیرمعمول است. با فرض این که ۵۰٪ جامعهٔ هدف تحت پوشش ایمن‌سازی قرار می‌گیرد، برای ۹۵٪ اطمینان از برآورد کردن نسبت مزبور در محدودهٔ نقاط ۱۰ درصدی مقدار واقعی آن به یک نمونهٔ تصادفی ساده با اندازهٔ ۹٦ کودک نیاز است.ولی نمونه‌گیری تصادفی ساده در محیطی که آمارگیریهای برنامهٔ گسترش یافتهٔ ایمن‌سازی اجرا می‌شوند مقرون به صرفه نیست و به همین دلیل نوعی فن نمونه‌گیری خوشه‌ای باید به صورت معمول به کارگرفته شود. اگر اثر طرح ۲ فرض شود، اندازهٔ نمونهٔ مورد نیاز دو برابر خواهد شد و به ۱۹۲ نفر (یعنی دو برابر ۹٦ نفر) برای نمونه نیاز است. کسانی که روش‌شناسی برنامهٔ گسترش یافتهٔ ایمن‌سازی را ابداع کرده بودند تصمیم داشتند از ۳۰ خوشه استفاده کنند. بنابراین، نمونه‌ای متشکل از ۳۰ خوشه با ۷ نفر از هر خوشه موجب به دست آمدن نمونه‌ای می‌شد که اندازهٔ آن اندکی بزرگتر از نمونهٔ مطلوب بود. این موضوع با جزئیات بیشتر توسط لِمی‌شو و استرا [۳] شرح داده شده است.

خوشه‌ها در آمارگیریهای برنامهٔ گسترش یافتهٔ ایمن‌سازی به جامعهٔ ویژهٔ مورد آمارگیری بستگی دارند و معمولاً از روستاها، شهرها، یا نواحی خدمات بهداشتی تشکیل شده‌اند که اطلاعات جمعیتی مربوط به آنها موجود است. در مرحلهٔ اول، یک نمونهٔ با احتمال متناسب با اندازه (به صورتی که در فصل ۱۱ شرح داده شد) انتخاب می‌شود. ولی در مرحلهٔ دوم به دلایل بودجه‌ای و لوژستیکی، فرایند انتخاب هفت نفر با نمونه‌گیری خوشه‌ای مرسوم تفاوتی بارز دارد. در روش‌شناسی مربوط به برنامهٔ گسترش یافتهٔ ایمن‌سازی، به جای انتخاب تصادفی هفت آزمودنی از میان همهٔ آزمودنیهای موجود،

[1] Probability Proportional to Size

یک خانوار به تصادف برای شروع انتخاب می‌شود، همهٔ اطلاعات مربوط به آزمودنیهای واجد شرایط موجود در آن خانوار جمع‌آوری می‌شود، و سپس به خانواری مراجعه می‌شود که درب ورودی منزل او از لحاظ فیزیکی به خانوار قبلی از همه نزدیکتر است. این فرایند مراجعه به نزدیکترین خانوار بعدی و جمع‌آوری اطلاعات دربارهٔ همهٔ افراد واجد شرایط در آن خانوارها ادامه می‌یابد تا هفت آزمودنی مورد نیاز بررسی شوند.

شیوه‌هایی خاص که برای انتخاب تصادفی اولین خانوار در مرحلهٔ دوم نمونه‌گیری به کار می‌روند به ماهیت خوشه بستگی دارند. مثلاً در یک ناحیهٔ روستایی، ممکن است به مصاحبه‌گر آموزش داد که به مکانی که در داخل خوشه دارای مرکزیت است (از قبیل کلیسا یا بازار) مراجعه کند، به طور تصادفی یک جهت را برای طی طریق انتخاب کند (یعنی شمال، جنوب، شرق، یا غرب) و تعداد K خانوار موجود در آن جهت را از نقطهٔ شروع تا مرز شهر بشمارد. سپس یک عدد تصادفی را بین ۱ و K انتخاب کند تا اولین خانوار تعیین شود. طبق قاعدهٔ عملکرد، داده‌های مربوط به همهٔ کودکان واجد شرایط در خانواری که هفتمین کودک در آن است جمع‌آوری می‌شوند، حتی اگر این کار موجب شود که بیش از ۷ کودک در یک خوشه گنجانده شود.

دریافته‌اند که روش انتخاب کودکان به شرح بالا، در داخل خوشه‌های نمونه برای آمارگیریهای برنامهٔ گسترش یافتهٔ ایمن‌سازی از لحاظ اجرای میدانی، نسبتاً آسان است. هزینه‌های یک انتخاب واقعاً تصادفی در مرحلهٔ دوم با توجه به منابعی که نوعاً به این آمارگیریهای EPI اختصاص داده می‌شود بازدارنده خواهد بود. در تلاشی برای ارزشیابی اثر این انتخاب غیرنمونه‌ای افراد در داخل خوشه‌ها بر برآوردهای مربوط به پوشش ایمن‌سازی، یک مدل شبیه‌سازی رایانه‌ای ابداع شده است [٤]. این مدل، با استفاده از جامعه‌هایی که با مشخصه‌های ویژه به طور مصنوعی ایجاد شده‌اند، نتایج حاصل از روش انتخاب برنامهٔ گسترش یافتهٔ ایمن‌سازی را با نتایج به دست آمده از روشهای سنتی‌تر مقایسه کرده است. معلوم شده است که با وجود ضعف عمل روش برنامهٔ گسترش یافتهٔ ایمن‌سازی در داخل خوشه‌های ویژه‌ای که در آنها پنهانکاریهایی دربارهٔ افراد ایمن‌سازی نشده وجود داشته است، نتایج حاصل از همهٔ خوشه‌ها در مجموع، گرایش به دقیق بودن تا حدود ۱۰ درصد سطوح واقعی جامعه داشته است. این در محدودهٔ اهدافی است که به وسیلهٔ برنامهٔ گسترش یافتهٔ ایمن‌سازی تعیین شده و می‌تواند به این واقعیت نسبت داده شود که تعداد خوشه‌های انتخاب شده زیاد است و اریبیهایی که در داخل خوشه‌ها روی می‌دهند در مجموعهٔ ۳۰ تایی گرایش به سربه‌سر شدن دارند. به این ترتیب، به نظر می‌رسد این روش، هنگام استفاده در سراسر مناطق هدف، کاملاً سودمند باشد ولی در صورت

استفاده در خوشه‌هایی خاص یا برای زیرگروهها می‌تواند برآوردهایی فراهم کند که به کلی غیر قابل قبول باشند.

۲.۱٤ نمونه‌گیری تضمین کیفیت

طرح آمارگیری برنامهٔ گسترش یافتهٔ ایمن‌سازی که در بخش ۱.۱٤ شرح داده شد دارای مزایای عمدهٔ سادگی و کم‌هزینه بودن است که برای برآورد کردن پارامترهایی از قبیل پوشش مایه‌کوبی در کشورها یا نواحی موردنظر، گزینه‌ای جذاب تلقی می‌شود. ولی روش‌شناسی برنامهٔ گسترش یافتهٔ ایمن‌سازی، به این علت که یک فن آمارگیری خوشه‌ای است، نمی‌تواند اطلاعات مربوط به واحدهای جامعه‌ای کوچک یا نواحی بهداشت را فراهم آورد. مدیران برنامهٔ بهداشت برای تمرکز بر فعالیتهای نظارتی به این اطلاعات نیاز دارند. زیرا برخی واحدهای بهداشتی ممکن است در اجرای وظایف خاص خود نسبت به دیگر واحدها با شایستگی کمتری عمل کنند. مدیران می‌توانند با شناسایی واحدهای بهداشتی که به اهداف اظهار شدهٔ خود دست نیافته‌اند (مثلاً پوشش ایمن‌سازی در سطح عالی) تلاشهای مستقیم خود را در آن مناطقی هدایت کنند که به بیشترین بهبود نیاز دارند. شیوهٔ نمونه‌گیری تضمین کیفیت (QAS)[1] به عنوان جایگزین نمونه‌گیری برنامهٔ گسترش یافتهٔ ایمن‌سازی (EPI) برای استفاده به صورت ابزار نظارتی پیشنهاد شده و مورد توجه قابل ملاحظه‌ای قرار گرفته است.

روشهای نمونه‌گیری تضمین کیفیت در اصل برای نمونه‌گیری و بازرسی محصول تولید شده در مواردی به کار برده می‌شدند که ضرورت ایجاب می‌کرد که هزینه‌های نیروی کار و سایر هزینه‌های نمونه‌گیری در سطوح حداقل حفظ شوند. یک نوع نمونه‌گیری تضمین کیفیت، یعنی نمونه‌گیری تضمین کیفیت دسته‌ای[2]، اساساً نمونه‌گیری طبقه‌بندی شده است، ولی اندازه‌های نمونه کوچکتر از آنند که آنچه را تأمین کنند که معمولاً بازه‌های اطمینان باریک در حد قابل قبول برای برآوردها در داخل یک طبقهٔ معین تلقی می‌شوند (و معمولاً به «دسته» یا «بسته» موسوم‌اند). در عوض، براساس این احتمال که تعداد اقلام معیوب یک نمونهٔ مربوط به دسته یا بستهٔ خاص موردنظر، از عدد تعیین شده‌ای کمتر یا برابر با آن باشد، در مورد کیفیت آن دسته تصمیم‌گیری می‌شود. نتایج حاصل از نمونه‌های انتخاب شده از همهٔ بسته‌های دوبه‌دو ناسازگار و فراگیر را می‌توان ترکیب کرد تا برآورد کل دقیقی از متوسط کیفیت کل محصول به دست آید.

متوسط کیفیت محصول غالباً به طور مداوم توسط تولید کننده، مورد نظارت قرار می‌گیرد تا (۱) جایی را که می‌توان در فرایند تولید بهبودی ایجاد کرد شناسایی نمود و (۲) طرح نمونه‌ای را وقتی

[1] Quality Assurance Sampling
[2] Lot Quality Assurance Sampling (LQAS)

کیفیت متوسط محصول تغییر می‌کند تعدیل کرد. به طور کلی، یک دسته واحدی است که «از نظر عملیاتی سودمند» است. مثلاً در یک کاربرد صنعتی، اگر چند ماشین یک قطعهٔ بخصوص را تولید می‌کنند می‌توان دسته‌هایی را که یک ماشین تولید می‌کند انتخاب کرد، به خصوص اگر بیشتر احتمال داشته باشد که هرگونه تغییر در قطعات تولیدی ناشی از طرز کار دستگاه باشد تا نهادهٔ عملگر. بازهٔ نمونه‌گیری تولید کننده باید به اندازهٔ کافی کوتاه باشد تا بتوان هرگونه انحراف در اندازه‌گیریها را، پیش از آن که از حدود تحمل خارج شوند، شناسایی کرد. برای این نوع کاربرد، ارزش دارد که دنبالهٔ اندازه‌گیریها برای شناسایی سریع گرایش به انحراف تحت نظارت قرار گیرند.

برای کاربردهای مربوط به بهداشت عمومی، یک مدیر ملی می‌تواند دسته‌ها را به صورت دریافت‌کنندگان خدمات از یک واحد عملیاتی تک، از قبیل یک تیم ویژهٔ ایمن‌سازی، در یک دورهٔ زمانی مشخص تعریف کند. فاصلهٔ زمانی بین نمونه‌گیریها می‌تواند به بازه‌های بین فصلهای «وقوع زیاد» بیماریهای قابل پیشگیری وابسته باشد، ولی احتمالاً به همین اندازه نیز به مدت زمان و مقدار هزینهٔ مربوط به نمونه‌گیری ارتباط خواهد داشت تا به هرگونه ملاحظات دیگر.

در کاربردهای مربوط به بهداشت عمومی، اگر چنین قضاوت شود که جامعه به طرز مناسبی پوشش داده شده است («یعنی دسته پذیرفته شود»)، در حالی که واقعاً چنین نبوده است، خطای فاحشی روی خواهد داد. برای کنترل این امکان، شیوه‌ای به صورت یک آزمون آماری یک‌طرفه ایجاد شده است. فرض کنید d تعداد افراد مایه‌کوبی نشده در نمونهٔ متشکل از n آزمودنی باشد، و P نسبت واقعی افراد مایه‌کوبی نشده در جامعه‌ای با اندازهٔ N باشد. فرض بر این است که N به نسبت n بسیار بزرگ است. (اگر اتفاقاً N به نسبت n بزرگ نباشد، خواننده می‌تواند به متونی از قبیل متن درسی تألیف براونلی [۲۹، بخش ۱۵.۳] مراجعه کند که چگونگی استفاده از توزیع فوق هندسی را برای ارزشیابی شیوهٔ نمونه‌گیری تضمین کیفیت دسته‌ای نشان می‌دهد.)

فرض صفر چنین است

$H_0 : P \geq P_0$ (یعنی، ۰٫۵۰ ≥ نسبت کودکان مایه‌کوبی نشده)*

در مقابل

$H_a : P < P_0$ (یعنی، ۰٫۵۰ > نسبت کودکان مایه‌کوبی نشده)

جدول چهار خانه‌ای زیر، پیامدهای شیوهٔ آزمون را توصیف می‌کند:

* سطح ۵۰٪ در اینجا برای مثال انتخاب شده است. در واقع هر سطحی را می‌توان انتخاب کرد.

پیامدهای آزمون فرض در شیوهٔ نمونه‌گیری تضمین کیفیت دسته‌ای

جامعهٔ واقعی

تصمیم	به طرز مناسبی مایه‌کوبی نشده	به طرز مناسبی مایه‌کوبی شده
عدم موفقیت در رد H_0: پوشش نامناسب	آزمون، فقدان پوشش مناسب را تشخیص می‌دهد یا نسبت به آن حساس است $1-\alpha$ حساسیت	*مخاطرهٔ تأمین‌کننده* β نرخ مثبت غلط ← رد دسته
رد H_0 : پوشش مناسب	*مخاطرهٔ مصرف‌کننده* α نرخ منفی غلط	آزمون، پوشش مناسب را تشخیص می‌دهد $1-\beta$ ویژگی ← پذیرش دسته

توجه کنید که در این جدول، چون آزمون به صورت یک‌طرفه تهیه شده و چون فرض بر این است که جامعه به طرز مناسبی پوشش داده نشده است، خطای نوع I، پذیرش دسته هنگامی که معیوب است (منفی غلط)، یعنی احتمالی که می‌توانیم تحت کنترل درآوریم، جدیترین خطاست، مگر این که H_0 را رد کنیم. یعنی، با استفاده از مثال ایمن‌سازی، اگر جامعه‌ای(دسته‌ای) از کودکان دارای نسبت ایمن‌سازی شدهٔ قابل قبولی تصور شود درحالی که در حقیقت چنین نیست، تعداد بیشتر کودکان آسیب‌پذیر موجود در جامعه، خطر انتقال بیماری را هنگام بروز آن در دسته افزایش خواهد داد. از این رو است که «هزینهٔ» اعلام این را که جامعه به طرز مناسبی مایه‌کوبی شده است، وقتی حقیقتاً چنین نیست، سنگین در نظر می‌گیریم. از سوی دیگر، خطای نوع II ـ یعنی رد دستهٔ قابل قبول ـ کمتر جدی داوری می‌شود زیرا نتیجهٔ یک تصمیم مثبت غلط آن است که منابع برنامه روی جامعه‌ای متمرکز شود که هم‌اکنون نیز به طرزی مناسب ایمن‌سازی شده است.

مشکل اساسی در نمونه‌گیری تضمین کیفیت دسته‌ای بیش از آن که صرفاً تعیین اندازهٔ نمونه باشد انتخاب توازنی مناسب بین اندازهٔ نمونه و ناحیهٔ بحرانی است. محاسبهٔ احتمال خطای نوع II، β، در همهٔ موارد به مقدار واقعی P ، وقتی فرض می‌شود با P_0 تفاوت دارد، بستگی خواهد داشت. روش محاسبهٔ احتمالها و تعیین اندازه‌های نمونه با استفاده از توزیع دوجمله‌ای انجام خواهد شد و به تفصیل در جاهای دیگر شرح داده شده است [۳۰] و [۳۱].

روش نمونه‌گیری تضمین کیفیت که تا اینجا شرح داده شد به «نمونه‌گیری تکی» موسوم است، زیرا فقط یک نمونه برای تصمیم‌گیری دربارهٔ وضع دسته انتخاب می‌شود. تعدیلی از شیوهٔ نمونه‌گیری تضمین کیفیت دسته‌ای شامل راهبرد «نمونه‌گیری دوگانه» است و می‌تواند تحت شرایط میدانی معین مفید واقع شود.

یک نمونه‌گیری دوگانه یا برنامهٔ نمونه‌گیری دومرحله‌ای را می‌توان برای پایین آوردن هزینه‌های نمونه‌گیری یک آمارگیری مورد استفاده قرار داد. در وضعیت حاضر، دو مقدار بحرانی، d_1^* و d_2^*، تعیین شده‌اند که $d_1^* \le d_2^*$، و دو اندازهٔ نمونه‌ای، n_1 و n_2، مشخص شده‌اند. در مرحلهٔ اول، n_1 نفر مورد مطالعه قرار می‌گیرند. اگر تعداد مشاهده شدهٔ افراد مایه‌کوبی نشده کمتر از d_1^* یا برابر با آن باشد نتیجه می‌گیریم که نسبت واقعی افراد مایه‌کوبی نشده در جامعه به صورتی معنی‌دار کمتر از P_0 است. اگر تعداد مشاهده شدهٔ افراد مایه‌کوبی نشده در مرحلهٔ اول بیشتر از d_1^* ولی کمتر از d_2^* یا برابر با آن باشد، در آن صورت به مرحلهٔ دوم می‌رویم و نمونه‌گیری را ادامه می‌دهیم تا یا d_2^*+1 نفر مایه‌کوبی نشده را مشاهده کنیم که حاکی از سطح پایین پوشش مایه‌کوبی است، یا مجموعاً n_1+n_2 نفر نمونه‌گیری شوند بدون این که از d_2^* بیشتر شوند و این حاکی از سطح بالای پوشش مایه‌کوبی است.

به جای این تصور که نمونه‌گیری دوگانه یک برنامهٔ نمونه‌گیری تکی است که یک مرحله نمونه‌گیری اضافی به دنبال دارد باید دانست که کل اندازهٔ نمونه و مقدار بحرانی هر دو مرحله، با برنامهٔ نمونه‌گیری تکی که قبلاً توصیف شد مطابقت دارد در حالی که مرحلهٔ اول معرف یک نمونهٔ «تحویل یافتهٔ» اولیه است. منشأ صرفه‌جویی بیشتر با نمونه‌گیری دوگانه نیز در همین امر نهفته است. هرگاه نتایج حاصل در مرحلهٔ اول کرانگین باشند، برنامه‌ریز بهداشت می‌تواند بر این اساس نتیجه‌گیری کند که آزمودنیهای نمونه‌گیری شده کمتر از آزمودنیهای نمونه‌گیری شده با استفاده از برنامهٔ نمونه‌گیری یک‌مرحله‌ای است. در غیر این صورت، برنامه‌ریز بهداشت با علم به این که ماکسیمم تعداد نمونه‌گیری شده برابر با اندازهٔ نمونهٔ حاصل از برنامهٔ یک‌مرحله‌ای خواهد بود به نمونه‌گیری ادامه خواهد داد. به عبارت دیگر، استفاده از برنامهٔ نمونه‌گیری دوگانه تضمین می‌کند که تعداد افراد نمونه‌گیری شده کمتر از تعداد حاصل از برنامهٔ نمونه‌گیری یک‌مرحله‌ای بوده یا برابر با آن باشد.

نمونه‌گیری دوگانه در داخل متن طرح نمونه‌گیری تضمین کیفیت دسته‌ای، در عین حالی که می‌تواند تحت شرایطی خاص، صرفه‌جویی در آمارگیری بهداشت را افزایش دهد، ممکن است در شرایطی دیگر اصلاً امکانپذیر نباشد. مثلاً اگر تحلیل نتایج آزمون پزشکی در محل امکانپذیر نباشد، آن‌گاه ممکن است برنامهٔ نمونه‌گیری دوگانه مستلزم آن باشد که اگر نتایج آماری حاصل از نمونهٔ اول کرانگین یا قانع کننده نباشند یک تیم پزشکی برای نمونه‌گیری بیشتر به محلهٔ موردنظر بازگردند. هزینهٔ

زمانی و مراجعه به محل ممکن است با صرفهٔ اقتصادی بالقوهٔ حاصل از نمونه‌گیری دوگانه سربه‌سر شود. ولی اگر هزینهٔ اجرای آزمایش پزشکی به ازای هر آزمودنی زیاد باشد ممکن است برنامه‌ریز بهداشت احساس کند که سود بالقوهٔ حاصل از ناگزیر بودن احتمالی اجرای فقط مرحلهٔ اول نمونه‌گیری، موجب ترجیح نمونه‌گیری دوگانه شود.

نمونه‌گیری تضمین کیفیت دوگانهٔ دسته‌ای، یک فن نمونه‌گیری طبقه‌بندی شده است. این نمونه‌گیری به جای ساختن برآوردهای بازه‌های اطمینان پارامترهای نامعلوم، مسئله را به یک سری آزمونهای فرض تبدیل می‌کند. ولی اطلاعاتی که فراهم می‌کند از نمونه‌گیری تصادفی طبقه‌بندی شدهٔ متداول بیشتر نیست زیرا با استفاده از فن اخیر، اگر اندازه‌های نمونه‌ای به قدر کافی برای تأمین بازه‌های اطمینان سودمند بزرگ باشند، می‌توان بازه‌های اطمینان را برای هر طبقه (دسته) تعیین و دربارهٔ مقادیر پوشش داده شده با هر یک از این قبیل بازه‌ها تصمیم‌گیری کرد.

اگرچه اندازه‌های نمونه‌ای برای هر طبقه در یک طرح آمارگیری تضمین کیفیت دسته‌ای، نوعاً کوچکتر از آنند که بازه‌های اطمینان سودمندی برای برآوردهای مربوط به هر طبقه فراهم کنند، ولی یک طرح آمارگیری تضمین کیفیت دسته‌ای که به صورتی مناسب طراحی شده باشد می‌تواند ابزاری برای آزمون پیوستهٔ طبقه‌ها و رده‌بندی آنها به عنوان «قابل قبول» یا «غیر قابل قبول» از لحاظ یک برآمد ویژه فراهم کند. چون نمونه‌های مربوط به نمونه‌گیری تضمین کیفیت دسته‌ای نسبتاً کوچک‌اند احتمال اجرای این نمونه‌گیری بسیار بیشتر است. شاید بتوان نمونه‌ها را همزمان با انجام سایر وظایفی که کارکنان را به میدان کار می‌کشاند انتخاب کرد. چون قواعد ساده‌ای برای اجرا در نظر گرفته شده‌اند مأمور آمارگیری یا رده‌بندی به حداقل آموزش نیاز دارد. نتایج مربوط به طبقه‌ها را می‌توان ترکیب کرد تا برآوردهایی مناسب برای گروههایی از طبقات از قبیل بخشها، ناحیه‌ها، یا کل کشور فراهم شود.

هر چند که بازه‌های اطمینان همواره بیشتر از یک تصمیم دوحالتی ساده اطلاعات فراهم می‌کنند، ولی اندازه‌های نمونه‌ای مورد نیاز در به دست آوردن برآوردهایی با هر سطحی از دقت مفید، برای طبقه‌های نسبتاً کوچک ممکن است عملی نباشد. در این گونه موارد، یک طرح نمونه‌گیری تضمین کیفیت مناسب می‌تواند جایگزینی باشد که ارزش بررسی دارد.

۳.۱۴ اندازه‌های نمونه برای بررسیهای طولی

یک طرح با اهمیت که در مطالعات مربوط به همه‌گیری شناختی مورد استفاده واقع می‌شود *بررسی همگروهی* است. در این نوع بررسی، افراد را شناسایی و در زمان شناسایی اندازه‌گیری می‌کنند و گاهی سالهای متمادی آنها را برای تعیین متغیرهای برآمد (مثلاً وقوع یک بیماری خاص

موردنظر یا مرگ و میر ناشی از آن) پیگیری می‌نمایند. هدفهای عمدهٔ این قبیل مطالعات عبارت از بررسی روابط بین این متغیرهای برآمد و مشخصه‌های افراد حاضر در زمان شناسایی آغازین است. مثالهای مربوط به این قبیل مطالعات شامل بررسی بیماری رگهای قلبی توسط فرامینگام [۳۲]، بررسی فعالیتهای جسمانی توسط رالف پافنبارگر [۳۳]، و بررسی طولی سالخوردگی [۳٤] است.

این قبیل مطالعات غالباً به اندازهٔ همگروهی بزرگ نیاز دارند که براساس فرضهایی تعیین می‌شود که باید از نظر روابط بین وقوع متغیرهای برآمد و مشخصه‌های موجود در زمان تعیین همگروهها مورد آزمون قرار بگیرند. روشهای مربوط به تعیین این قبیل اندازه‌های همگروهی مورد نیاز، در نوشتگان مربوط به همه‌گیری شناختی شرح داده شده‌اند (مانند اثر روتمن و بویس [۳۵]). ولی به محض این که اندازهٔ همگروهی مورد نیاز تعیین شد، شناسایی تعداد کافی از افراد برای تأمین ملاکهای منظور شدن در گروه ضرورت پیدا می‌کند.

غالباً شناسایی اعضای گروه به وسیلهٔ یک آمارگیری انجام می‌پذیرد که در آن نمونه‌ای از افراد انتخاب می‌شوند و افراد نمونه به منظور تأمین ملاکهای مربوط به گنجانیده شدن در گروه غربالگری می‌شوند. سپس کسانی که ملاکهای موردنظر دربارهٔ آنها مصداق دارد برای بررسیهای بعدی در گروه جای داده می‌شوند. اندازهٔ نمونه‌ای که باید برای این غربالگری انتخاب شود در زیر برای دو طرح نمونه‌ای بیان می‌شود: نمونه‌گیری تصادفی ساده از افراد، و نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده.

۱.۳.۱٤ نمونه‌گیری تصادفی ساده

برای کسب اطمینان $(1-\alpha)100$ درصد برای به دست آوردن حداقل n^* آزمودنی دارای ملاکی خاص، تعداد تقریبی n آزمودنی مورد نیاز برای نمونه‌گیری از فرمول زیر به دست می‌آید

$$n = \left[A + \sqrt{A^2 + \frac{n^*}{P}} \right]^2 \qquad (۱.۱٤)$$

که در آن

$$A = \frac{|z_\alpha|}{2} \times \sqrt{\frac{1-P}{P}}$$

$z_\alpha =$ صدک توزیع نرمال $\alpha \times 100$

$P =$ نسبت افراد در جامعه که ملاکهای منظور شدن در بررسی همگروهی را تأمین می‌کنند (یعنی ورود افراد واجد شرایط برای منظور شدن در گروه)

اندازهٔ نمونه‌ای که از رابطهٔ (۱.۱٤) به دست می‌آید بر مبنای اندازهٔ جامعه است که به مراتب بزرگتر از اندازهٔ نمونه است و مبتنی بر تقریب توزیع نرمال برای توزیع دوجمله‌ای است.

مثال تشریحی: فرض کنید برای یک بررسی مقدماتی پیگیر بلندمدت دربارهٔ تحرک شغلی به ۲۰۰ کارمند اجرایی عالیرتبهٔ بیمارستانها نیاز است. این افراد برای این که در بررسی گنجانیده شوند باید بین ۳۰ تا ۳۹ سال سن داشته باشند. روش وارد کردن این افراد در بررسی به شرح زیر است. از روی پرونده‌ای که شامل اسامی کارمندان اجرایی عالیرتبهٔ تقریباً ٦۰۰۰ بیمارستان محلی در سراسر امریکاست یک نمونهٔ تصادفی ساده گرفته می‌شود. به هر یک از این کارمندان نمونه تلفن کرده، در مورد سن از او سؤال می‌شود. اگر سن او بین ۳۰ تا ۳۹ سال بود، آن‌گاه کوششی به عمل می‌آید تا مشارکت وی در این بررسی جلب شود. حدس بر این است که تقریباً ۳۰٪ از همهٔ کارمندان اجرایی عالیرتبه در دامنهٔ سنی هدف قرار دارند. اگر بخواهیم ۹۵٪ مطمئن باشیم که حداقل ۲۰۰ کارمند اجرایی عالیرتبه را به این ترتیب به دست می‌آوریم (با فرض مشارکت ۸۵٪ از میان همهٔ افراد نمونهٔ واجد شرایط)، چند کارمند اجرایی باید نمونه‌گیری شوند؟

با استفاده از رابطهٔ (۱.۱٤) با $P = ۰/۳۰$، $n^* = ۲۰۰$ و $z_\alpha = -۱/۶۴۵$،

داریم

$$A = \frac{۱/۶۴۵}{۲} \times \sqrt{\frac{۱-۰/۳۰}{۰/۳۰}} = ۱/۲۵۶$$

و

$$n = \left[۱/۲۵۶ + \sqrt{(۱/۲۵۶)^۲ + \left(\frac{۲۰۰}{۰/۳۰} \right)} \right]^۲ \approx ۷۳۵$$

و با در نظر گرفتن ۸۵٪ موفقیت در جلب همکاری، n نهایی مورد نیاز به این صورت خواهد بود

$$n = \frac{۷۳۵}{۰/۸۵} \approx ۸۶۵$$

به این ترتیب باید ۸٦۵ کارمند اجرایی عالیرتبه نمونه‌گیری شوند تا ۹۵٪ مطمئن باشیم که حداقل ۲۰۰ نفر را در گروه سنی هدف انتخاب خواهیم کرد.

□

فرمولبندیی که در بالا شرح داده شد برای نمونه‌گیری تصادفی ساده است. ولی یک طرح نمونه‌گیری که احتمال استفاده از آن برای شناسایی افراد در عمل بیشتر است نمونه‌گیری خوشه‌ای

است که در آن ممکن است خانوار یک خوشه تلقی شود و افراد موجود در هر خانوار نمونه، از نظر داشتن شرایط برای گنجانیده شدن در گروه غربالگری شوند. تعیین ویژگی مشابه برای اندازهٔ نمونه آن است که از خوشه‌ها به تعداد کافی، m تا، نمونه‌گیری شود به طوری که تعداد مورد نیاز n^* از افراد با $(1-\alpha)\times 100$٪ اطمینان برای گنجاینده شدن در گروه شناسایی شوند. با این ویژگی و تحت طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده می‌توان رابطهٔ (۱.۱٤) را با تعدیلهای زیر به کار برد.

۲.۳.۱٤ نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده

$$m = \left[A + \sqrt{A^2 + \frac{n^*}{\overline{N}}} \right]^2 \qquad (۲.۱٤)$$

که در آن

$$A = | z_\alpha | \times \frac{V_N}{2}$$

$$V_N = \frac{\sigma_N}{\overline{N}}$$

$$\sigma_N = \sqrt{\sum_{i=1}^{M} (N_i - \overline{N})^2 / M}$$

M = تعداد خوشه‌ها در جامعه

N_i = تعداد افراد واجد شرایط در خوشهٔ i ، $i = 1,\ldots,M$

$\overline{N} = \sum_{i=1}^{M} N_i / M$ (میانگین تعداد افراد واجد شرایط به ازای خوشه)

باید توجه داشت که $\overline{N}$، میانگین تعداد افراد واجد شرایط به ازای خوشه، مشابه P ، یعنی نسبت افراد واجد شرایط جامعه در تعیین ویژگیها برای اندازهٔ نمونه تحت نمونه‌گیری تصادفی ساده است.

مثال تشریحی: فرض کنید می‌خواهیم نمونه‌ای از خانوارها در یک شهر بزرگ بگیریم و یکایک اعضای خانوار را به منظور شناسایی ۵۰۰ زن مجرد ۲۵ تا ۳۹ سال سن که شاغل تمام وقت‌اند و حداقل یک فرزند مدرسه‌رو دارند غربالگری کنیم. باز فرض کنید که در شهر تقریباً ۳۰۰۰۰ زن با این شرایط وجود دارند و شهر دارای تقریباً ۳۰۰۰۰۰ خانوار است و $\sigma_N \approx \sqrt{\overline{N}}$. چند خانوار باید نمونه‌گیری شوند تا با اطمینان ۹۵٪ حداقل ۵۰۰ زن با ملاکهای بالا شناسایی شوند؟

(میانگین تعداد زنان مجرد به ازای خانوار)

$$\overline{N} = \frac{۳۰۰۰۰}{۳۰۰۰۰۰} = ۰/۱$$

$$\sigma_N = \sqrt{۰/۱} = ۰/۳۱۶$$

$$V_N = \frac{۰/۳۱۶}{۰/۱} = ۳/۱۶$$

$$A = ۱/۶۴۵ \times \frac{۳/۱۶}{۲} = ۲/۶۰$$

$$m = \left[۲/۶۰ + \sqrt{(۲/۶۰)^۲ + \left(\frac{۵۰۰}{۰/۱}\right)} \right]^۲ \approx ۵۳۸۲$$

به این ترتیب باید ۵۳۸۲ خانوار نمونه‌گیری شوند تا ۹۵٪ مطمئن باشیم که حداقل ۵۰۰ زن واجد شرایط شناسایی می‌شوند (با فرض مشارکت ۱۰۰٪ در بین زنان واجد شرایط).

□

۳.۳.۱٤ نمونه‌گیری خوشه‌ای با بیش از یک حوزه

غالباً وضعیت طوری است که می‌خواهیم n^* فرد را در هر یک از چندین حوزهٔ معین برای بررسی آینده شناسایی کنیم. برای مثال، ممکن است بخواهیم ۱۵۰۰ نفر را در هر یک از چهار رستهٔ نژادی ـ جنسی شناسایی کنیم (مرد سفیدپوست، مرد سیاهپوست، زن سفیدپوست، زن سیاهپوست). اگر طرح نمونه‌ای تعیین شده برای شناسایی این افراد، نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده باشد، در آن صورت، تعداد خوشه‌های مورد نیاز که باید نمونه‌گیری شوند تا حداقل n^* نفر در هر یک از این حوزه‌ها به دست آید با فرمول زیر محاسبه می‌شود

$$m = \max\left(m_۱, \ldots, m_H\right) \qquad (۳.۱٤)$$

که در آن

$$m_h = \left[A_h + \sqrt{A_h^۲ + \frac{n^*}{\overline{N}_h}} \right]^۲, \quad h = ۱, \ldots, H \qquad (٤.۱٤)$$

A_h و $\overline{N}_h$ همتاهای ویژهٔ حوزه‌ای پارامترهایی هستند که برای رابطهٔ (۲.۱٤) تعریف شدند و H = تعداد حوزه‌های موردنظر.

می‌توان بدون از دست دادن کلیت، فرض کرد که $m_۱ \le m_۲ \le \cdots \le m_H$. پس، با نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده، این احتمال برای همهٔ حوزه‌ها بجز حوزهٔ H وجود دارد که بیش از تعداد

مورد نیاز برای تأمین ویژگیهای موردنظر، دارای آزمودنیهای واجد شرایط باشند. مطلب مذکور، به دلیل این واقعیت است که در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای معمولی، همین که یک خوشه در نمونه انتخاب شد دیگر نمونه‌گیری فرعی از افراد انجام نمی‌شود. لِه‌وی و همکاران[٥]، [٦] نوعی نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده را ابداع کرده‌اند که در آن خوشه‌های نمونه به طور تصادفی در یکی از گروههای H قرار داده می‌شوند. آنها که در گروه ١ قرار دارند می‌توانند افرادی را از همهٔ حوزه‌های H بگیرند. آنهایی که در گروه ٢ قرار دارند می‌توانند افرادی را از حوزهٔ ٢ تا H بگیرند ولی از حوزهٔ ١ نمی‌توانند بگیرند. همین طور تا خوشه‌های گروه h که فقط می‌توانند افرادی را از حوزهٔ H بگیرند. این تعدیل نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده، به *نمونه‌گیری خوشه‌ای یک‌مرحله‌ای تلسکوپی* موسوم است و به این نتیجه منتهی می‌شود که ویژگیهای موردنظر برای هر حوزه دقیقاً تأمین می‌شود و از مشکل گرفتن تعداد افرادی بیش از حد لازم در بیشتر حوزه‌ها اجتناب می‌شود. این روش در یک آمارگیری بزرگ از اشخاص سالخورده در شانگهای، چین، با موفقیت به کار رفته است [٥]، [٦].

٤.١٤ برآورد کردن شیوع بیماریها از روی بررسیهای غربالگری

غالباً مایل‌اند شیوع یک بیماری یا شرطی در یک جامعه را براساس آزمون غربالگریی برآورد کنند که *حساسیت* و *ویژگی* آن چندان کامل نباشد. این دو اصطلاح به معنای زیرند:

حساسیت، احتمال آن است که شخصی که دچار یک بیماری یا شرط خاص است آزمون غربالگری مثبت داشته باشد.

ویژگی، احتمال آن است که شخصی که دچار یک بیماری یا شرط خاص نیست دارای آزمون غربالگری مثبت نباشد.

آزمون غربالگری معمولاً آزمونی است که در مقایسه با آزمون تشخیصی دقیقتری که در عین اینکه قابل دسترس است برای استفاده در وضعیت آمارگیری، شدنی نیست، برای استفادهٔ میدانی کم‌هزینه و شدنی باشد. مثلاً آزمایش وضعیت ذهنی (MMSE)[1] شامل چند سؤال است که برای غربالگری افراد از لحاظ اختلالات شناختی که می‌تواند نشانهٔ زوال عقل (اختلال مشاعر) باشد یا نباشد [٦] مورد استفاده قرار می‌گیرد در حالی که برای تشخیص این اختلال مشاعر، با هر میزان معقولی از درستی، به یک سری آزمایشات پیچیدهٔ روانی و عصب‌شناختی نیاز خواهد بود.

[1] Mini Mental State Examination

اهداف عمدهٔ برنامه‌های غربالگری معمولاً شناسایی افراد دچار شرط یا بیماری موردنظر برای مداخلهٔ بعدی است، و مسایل آماری این برنامه‌ها مستلزم ارزیابی احتمال این است که فردی که به عنوان مثبت غربالگری شده است واقعاً دچار آن بیماری است و به همین قیاس احتمال این است که فردی که به عنوان منفی غربالگری شده واقعاً دچار آن بیماری نیست (که به ترتیب *مقدار پیشگوی آزمون مثبت* و *مقدار پیشگوی آزمون منفی* نامیده می‌شوند). مباحث مربوط به این مطالب در نوشتگان مربوط به همه‌گیری شناختی فراوان‌اند [۷] و ما در عوض، بحث خود را بر مسئلهٔ برآورد کردن شیوع بیماری از طریق آزمون غربالگری متمرکز می‌کنیم.

روش‌شناسی پایه‌ای، مستلزم انتخاب نمونه‌ای n فردی از جامعه‌ای متشکل از N نفر و اجرای آزمون غربالگری دربارهٔ هر یک از افراد نمونه است. اگر k از n نفر افراد نمونه به عنوان مثبت غربالگری شوند، لِه‌وی و کاس [۸] نشان داده‌اند که در آن صورت $\hat{\pi}$، برآورد درستنمایی ماکسیمم π، شیوع نامعلوم برای افراد دچار آن بیماری از فرمول زیر به دست می‌آید.

$$\hat{\pi} = \frac{\hat{p} + S_p - ۱}{S_e + S_p - ۱} \qquad (۵.۱۴)$$

که در آن، $\hat{p} = \frac{k}{n}$ = نسبت دارندگان آزمون مثبت، S_e = حساسیت آزمون و S_p = ویژگی آزمون. از رابطهٔ (۵.۱۴) می‌توان دید که در وضعیتهایی که مجموع ویژگی و حساسیت کمتر از یک است یا هنگامی که مجموع شیوع مربوط به آزمون و ویژگی آن کمتر از یک است برآورد شیوع $\hat{\pi}$ منفی خواهد بود. گاستویرث [۹] و لیو و لِه‌وی [۱۰] تعدیلهایی از معادلهٔ (۵.۱۴) را پیشنهاد کرده‌اند که امکان برآوردهای منفی را برای شیوع از میان می‌برند. در عمل، یک آزمون غربالگری که حساسیت و یا ویژگی آن به اندازه‌ای کم است که احتمال برآوردهای منفی برای شیوع را به وجود می‌آورد احتمالاً مطلوبیت چندانی به عنوان یک غربال برای بیماری یا شرطی خاص نخواهد داشت.

خطای معیار برآورد شیوع، $\hat{\pi}$، از فرمول زیر به دست می‌آید

$$SE(\hat{\pi}) = \frac{SE(\hat{p})}{(S_e + S_p - ۱)} \qquad (۶.۱۴)$$

که در آن خطای معیار $\hat{p}$، $SE(\hat{p})$، به طرح نمونه‌ای خاص و شیوه‌های برآورد کردن مورد استفاده بستگی خواهد داشت.

مثال تشریحی: فرض کنید یک نمونهٔ تصادفی ساده متشکل از ۱۵۰ عضو از میان اعضای اتحادیه‌ای گرفته شده است که نمایندهٔ معلمان شاغل در هفت ناحیهٔ آموزشی در بخش بزرگی از ایالت ایلی‌نوی

است. این هفت ناحیهٔ آموزشی ٢٥٦٠ آموزگار را در استخدام خود دارند. همهٔ معلمهای نمونه‌گیری شده از لحاظ بیماری آب سیاه تحت آزمون غربالگریی قرار گرفتند که معلوم بود دارای ٩٦٪ حساسیت و ٨٩٪ ویژگی است. از میان ١٥٠ معلم مورد آزمون، ٢٣ نفر مثبت بودند. از روی این داده‌ها مطلوب است برآورد نسبت معلمان دچار آب سیاه در اتحادیه و بازه‌های اطمینان ٩٥٪. برای این نسبت از دانسته‌های خود دربارهٔ نمونه‌گیری تصادفی ساده (فصل ٣)، داریم

$$\hat{p} = \frac{٢٣}{١٥٠} = ٠/١٥٣$$

و

$$\begin{aligned}\hat{SE}(\hat{p}) &= \left(\frac{N-n}{N}\right)^{1/2}\left(\frac{\hat{p}(١-\hat{p})}{n-١}\right)^{1/2} \\ &= \left(\frac{٢٥٦٠-١٥٠}{٢٥٦٠}\right)^{1/2}\left(\frac{٠/١٥٣\times٠/٨٤٧}{١٤٩}\right)^{1/2} \\ &= ٠/٠٢٩\end{aligned}$$

و از رابطه‌های (٥.١٤) و (٦.١٤)، داریم

$$\hat{\pi} = \frac{٠/١٥٣+٠/٨٩-١}{٠/٩٦+٠/٨٩-١} = \frac{٠/٠٤٣}{٠/٨٥} = ٠/٠٥١$$

و

$$\hat{SE}(\hat{\pi}) = \left(\frac{٠/٠٢٩}{٠/٨٥}\right) = ٠/٠٣٤$$

سرانجام، بازه‌های اطمینان ٩٥ درصدی برای π، شیوع نامعلوم آب سیاه در میان همهٔ معلمان در هفت ناحیهٔ آموزشی به صورت زیر است

$$٠/٠٥١-١/٩٦\times٠/٠٣٤ \le \pi \le ٠/٠٥١+١/٩٦\times٠/٠٣٤$$

یا

$$٠ \le \pi \le ٠/١١٨٠$$

توجه داشته باشید که حد پایین به صورت ٠/٠١٦- محاسبه شده و برابر با صفر گرفته شده است، زیرا نسبت منفی مفهومی ندارد.

□

روشی که در بالا ارائه شد فرض را بر این می‌گیرد که حساسیت و ویژگی آزمون غربالگری معلوم‌اند. اگر معلوم نباشند می‌توان از یک طرح نمونه‌گیری دوگانه برای برآورد شیوع نامعلوم، π،

استفاده کرد. در این طرح، زیرنمونه‌ای از آن دسته از افرادی که مثبت غربالگری شده‌اند و زیرنمونهٔ دیگری از آن دسته که منفی غربالگری شده‌اند انتخاب می‌شود. افراد هر دو نمونه مورد آزمایش تشخیصی دقیقتری برای آن بیماری یا شرط خاص قرار می‌گیرند که به عنوان «استاندارد طلایی» برای اثبات یا رد نتایج آزمون غربالگری اولیه مورد استفاده قرار خواهد گرفت. در این نوع طرح، برآورد شیوع، $\hat{\pi}$، از فرمول زیر به دست می‌آید

$$\hat{\pi} = \frac{\frac{k_1}{m_1} \times m + \frac{k_2}{m_2} \times (n-m)}{n} \qquad (۷.۱۴)$$

که در آن

n = تعداد غربالگری شده از جامعهٔ متشکل از N عنصر

m = تعداد مثبت برای آزمون غربالگری

n-m = تعداد منفی برای آزمون غربالگری

m_1 = تعداد نمونه‌گیری شده از گروه متشکل از m فردی که آزمون غربالگری آنها ابتدا مثبت بوده است

k_1 = تعدادی که مثبت بودن آنها از میان m_1 آزمودنی که ابتدا مثبت بوده و مجدداً نمونه‌گیری شده‌اند تأیید شده است

m_2 = تعداد نمونه‌گیری شده از گروه متشکل از n-m فردی که آزمون غربالگری آنها ابتدا منفی بوده است

k_2 = تعداد آزمودنیهایی که از میان m_2 آزمودنی که ابتدا منفی بوده و مجدداً نمونه‌گیری شده‌اند مثبت تشخیص داده شده‌اند

برآوردی از واریانس این برآورد (که با استفاده از واریانس شرطی و نادیده گرفتن تصحیح جامعهٔ متناهی به دست آمده است) با فرمول زیر محاسبه می‌شود

$$\hat{Var}(\hat{\pi}) = \left(\frac{k_1}{m_1} - \frac{k_2}{m_2}\right)^2 \times \frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n} + \left(\frac{m}{n}\right)^2 \frac{\frac{k_1}{m_1}\left(1-\frac{k_1}{m_1}\right)}{m_1} + \left(1-\frac{m}{n}\right)^2 \frac{\frac{k_2}{m_2}\left(1-\frac{k_2}{m_2}\right)}{m_2} \qquad (۸.۱۴)$$

مثال تشریحی: یک نمونهٔ تصادفی ساده متشکل از ۱۲۰ نفر از میان ۴۲۳ نفری که در باجه‌های اخذ عوارض در سیستم بزرگراههای یک ایالت بزرگ کار می‌کنند گرفته شده است. هر کارمند نمونه تحت

یک آزمون سریع غربالگری قرار می‌گیرد که نقص شنوایی را ارزیابی می‌کند. از میان ۱۲۰ کارمندی که به این ترتیب غربالگری شده‌اند ٦٣ نفر علایمی از کاهش شنوایی نشان داده‌اند. یک نمونهٔ تصادفی سادهٔ ۳۰ نفری از میان این ٦٣ نفری که مثبت غربالگری شده‌اند انتخاب شده و مورد آزمایش دقیقتری برای سنجش کاهش شنوایی قرار گرفته‌اند. از این ۳۰ نفر در ۲۳ نفر کاهش شنوایی تأیید شده است. به همین ترتیب، نمونه‌ای ۳۰ نفری از میان ۵۷ نفری که در ابتدا منفی غربالگری شده‌اند گرفته شده و ١٤ نفر از آنها نیز هنگامی که مورد آزمایش دقیقتر قرار گرفتند نقص شنوایی داشتند. از معادلهٔ (٧.١٤) داریم

$$n = ۱۲۰ \qquad m = ۶۳ \qquad n - m = ۵۷$$
$$m_۱ = m_۲ = ۳۰ \qquad k_۱ = ۲۳ \qquad k_۲ = ۱۴$$
$$\hat{\pi} = \frac{\frac{۲۳}{۳۰} \times ۶۳ + \frac{۱۴}{۳۰} \times ۵۷}{۱۲۰}$$

از رابطهٔ (٨.١٤)، خطای معیار $\hat{\pi}$ به صورت زیر برآورد می‌شود

$$\hat{SE}(\hat{\pi}) = \left\{ \left(\frac{۲۳}{۳۰} - \frac{۱۴}{۳۰} \right)^۲ \times \frac{۰/۵۲۵ \times ۰/۴۷۵}{۱۲۰} + (۰/۵۲۵)^۲ \times \frac{۰/۷۶۷ \times ۰/۲۳۳}{۳۰} + (۰/۴۷۵)^۲ \times \frac{۰/۴۶۷ \times ۰/۵۳۳}{۳۰} \right\}^{۱/۲} = ۰/۰۶۱$$

به این ترتیب، بازه‌های اطمینان ۹۵ درصدی برای نسبت π، مبتلایان به بیماری یا شرطی خاص که به منظور آن غربالگری شده‌اند به صورت زیر است

$$۰/۶۲ - ۱/۹۶ \times ۰/۰۶۱ \le \pi \le ۰/۶۲ + ۱/۹۶ \times ۰/۰۶۱$$

یا

$$۰/۵۰ \le \pi \le ۰/۷۴$$

□

مثال بالا اهمیت نمونه‌گیری دوگانه را هنگامی که حساسیت و ویژگی آزمون غربالگری نامعلوم‌اند نشان می‌دهد. اگر نسبت مثبت غربالگری آغازین به کار رفته بود، برآورد به دست آمده، یعنی $\frac{۶۳}{۱۲۰} = ۰/۵۲۵$، به صورتی قابل ملاحظه کمتر از برآوردی می‌شد که در رابطهٔ (٧.١٤) ارائه شده است.

در بررسیهای غربالگری غالباً نه تنها برآورد شیوع بیماری یا شرطی خاص، بلکه برآورد حساسیت و ویژگی آزمون غربالگری خاصی نیز موردنظر است. این کار در صورتی قابل اجراست که نمونه‌ها هم از آنهایی که در ابتدا منفی غربالگری شده‌اند و هم از آنها که ابتدا مثبت غربالگری شده‌اند انتخاب شوند. برآوردهای مناسب، $\hat{S}_e$ و $\hat{S}_p$، برای حساسیت و ویژگی به صورت زیر به دست می‌آیند

$$\hat{S}_e = \frac{m\frac{k_1}{m_1}}{m\left(\frac{k_1}{m_1}-\frac{k_2}{m_2}\right)+n\frac{k_2}{m_2}} \qquad (۹.۱۴)$$

و

$$\hat{S}_p = \frac{(n-m)\left(1-\frac{k_2}{m_2}\right)}{m\left(\frac{k_2}{m_2}-\frac{k_1}{m_1}\right)+n\left(1-\frac{k_2}{m_2}\right)} \qquad (۱۰.۱۴)$$

۵.۱۴ برآورد کردن پیشامدهای نادر: نمونه‌گیری شبکه‌ای

نشان دادیم که تعدیل برنامهٔ نمونه‌گیری تصادفی سادهٔ پایه، بدون تغییر در شیوهٔ برآورد کردن (مثلاً نمونه‌گیری تصادفی طبقه‌بندی شده با تخصیص متناسب) گاهی می‌تواند به برآوردهایی منجر شود که خطای نمونه‌گیری آنها کمتر از برآوردهایی است که از طریق نمونه‌گیری تصادفی ساده به دست آمده‌اند. همچنین بحث کردیم که چگونه یک برنامهٔ نمونه‌گیری تصادفی ساده که از شیوهٔ برآورد کردن تعدیل شده استفاده می کند (مثلاً برآورد نسبتی) می‌تواند به برآوردهایی با خطای نمونه‌گیری کمتر منتهی شود. به راهی مشابه، در این بخش، نشان خواهیم داد که چگونه تعدیل قاعدهٔ شمارش می‌تواند به برآوردهایی منجر شود که خطاهای نمونه‌گیری آنها کمتر از آن است که از طریق قواعد شمارش مرسوم به دست می‌آمد.

قاعدهٔ شمارش، الگوریتمی است که واحدهای شمارش (یا واحدهای فهرست‌برداری) را به واحدهای اولیه ربط می‌دهد. مثلاً در یک آمارگیری نمونه‌ای از خانوارها که برای برآورد کردن کل نوزادانی اجرا می‌شود که در یک جامعه طی دورهٔ زمانی مشخصی متولد می‌شوند، قاعدهٔ واضح شمارش آن است که گزارش کردن تولد به وسیلهٔ خانوار والدین نوزاد میسر باشد. چنین قاعده‌ای برای شمارش، هر عنصر (مثلاً تولد) را به فقط و فقط یک واحد شمارش (مثلاً خانوار) ربط می‌دهد. ولی یک قاعدهٔ شمارش دیگر این امکان را فراهم می‌سازد که گزارش تولد نوزادان به وسیلهٔ خانوار

پدربزرگ و مادربزرگ نوزاد نیز مانند خانوار والدین میسر باشد. در آن صورت این قاعدهٔ شمارش اجازه می‌دهد که یک عنصر به بیش از یک واحد فهرست‌برداری ربط داده شود.

قاعدهٔ شمارشی که ارتباط یک عنصر را به تنها یک واحد شمارش میسر می‌سازد *قاعدهٔ شمارش مرسوم* نامیده می‌شود. قاعدهٔ شمارشی که ارتباط یک عنصر را به بیش از یک واحد شمارش میسر می‌سازد *قاعدهٔ شمارش چندبارگی* نامیده می‌شود. طرحهای نمونه‌ای که از قواعد شمارش چندبارگی استفاده می‌کنند *نمونه‌های شبکه‌ای* نامیده می‌شوند. نمونه‌های شبکه‌ای طی دو دههٔ گذشته مورد توجه قابل ملاحظه‌ای قرار گرفته‌اند، به خصوص در علوم رفتاری و بهداشتی، و به خصوص در وضعیتهایی که متضمن پیشامدهای نادر یا مشخصه‌های کمیاب‌اند. این طرح را با مثال زیر بررسی می‌کنیم.

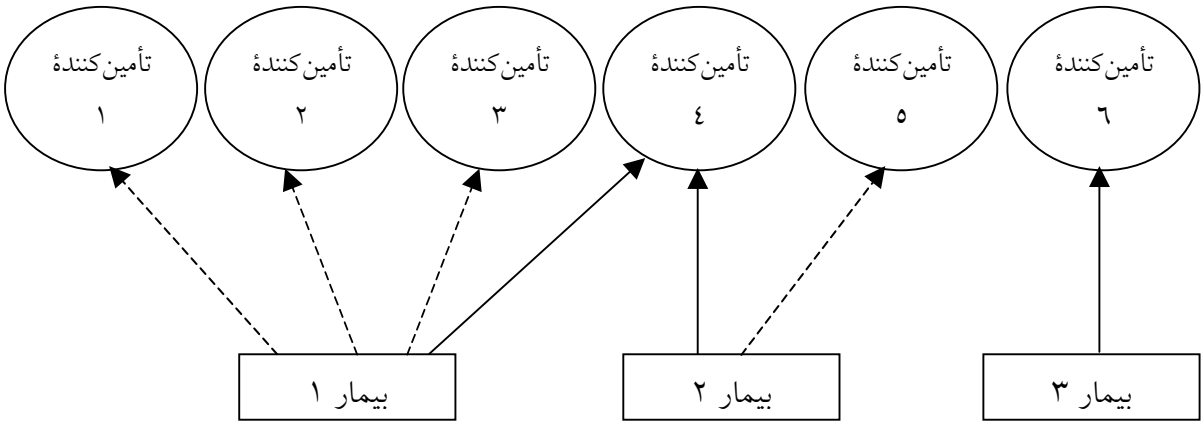
مثال تشریحی: بخشی را در نظر می‌گیریم که دارای ۱۰۰ تأمین کنندهٔ شناسایی شدهٔ مراقبتهای اولیهٔ بهداشتی است. فرض کنید می‌خواهیم یک آمارگیری نمونه‌ای با ده تأمین کنندهٔ مراقبتهای بهداشتی اجرا کنیم تا کل تعداد اشخاصی را که طی یک دورهٔ زمانی خاص تحت درمان سرطان پوست بوده‌اند برآورد کنیم. اگر هر بیمار فقط و فقط توسط یک تأمین کنندهٔ مراقبتهای بهداشتی تحت درمان قرار گرفته بود، فرایند مزبور فرایند ساده‌ای بود. ولی، بیماری که سرطان پوست دارد توسط چند تأمین کنندهٔ مراقبتهای بهداشتی درمان می‌شود (مانند متخصص داخلی، جراح، پرتوشناس، تن‌درمانگر). به این ترتیب، در طراحی آمارگیری نمونه‌ای باید قاعده‌ای برای شمارش تعیین کنیم که ارتباط عناصر را (یعنی اشخاصی را که تحت درمان سرطان پوست بوده‌اند) با بیش از یک واحد شمارش (یعنی تأمین کنندهٔ مراقبتهای بهداشتی) میسر سازد.

فرض می‌کنیم که در طول سال سه نفر تحت درمان سرطان پوست قرار گرفته‌اند. نفر ۱ با چهار پزشک (پزشکان ۱، ۲، ۳ و ٤)، نفر ۲ با دو پزشک (٤ و ۵) و نفر ۳ با یک پزشک (پزشک ٦) درمان می‌شده است. به این ترتیب شبکه‌ای داریم که تأمین کنندگان مراقبتهای بهداشتی را با بیماران مبتلا به سرطان پوست ربط می‌دهد. این شبکه در شکل ۱.۱٤ نشان داده شده است. خطهای توپر این شکل، *اولین* تأمین کننده را (از نظر ترتیب زمانی) نشان می‌دهند که بیمار را تحت درمان قرار داده است.

اگر از یک قاعدهٔ شمارش مرسوم استفاده کنیم ـ مثل ربط دادن بیماران به اولین تأمین کننده‌ای که آنان را تحت درمان گرفته است ـ در آن صورت تأمین کنندهٔ ٤، بیماران ۱ و ۲ را گزارش خواهد داد و تأمین کنندهٔ ٦ بیمار ۳ را گزارش خواهد کرد. ۹۸ تأمین کنندهٔ دیگر هیچ بیماری را گزارش نخواهند کرد. اگر X_i را تعداد بیماران گزارش شده توسط i امین تأمین کننده براساس این قاعدهٔ شمارش در نظر بگیریم، در آن صورت $X_۴ = ۲$، $X_۶ = ۱$ و بقیه $X_i = ۰$ اند. برای یک نمونهٔ تصادفی ساده متشکل

از $n = ۱۰$ تأمین کنندهٔ مراقبتهای بهداشتی، می‌بینیم که برآوردگر $x' = \left(\frac{۱۰۰}{۱۰}\right)x$ با شمارش همهٔ نمونه‌ها دارای توزیع زیر است

x'	فراوانی نسبی
۰	۰/۸۰۹۱
۱۰	۰/۰۹۰۹
۲۰	۰/۰۹۰۹
۳۰	۰/۰۰۹۱



شکل ۱.۱٤ نمونهٔ شبکه‌ای برای تأمین کنندگان مراقبتهای بهداشتی و بیماران مبتلا به سرطان پوست

میانگین توزیع x' یعنی $E(x')$، برابر است با ۳ (مجموع واقعی جامعه) و خطای معیار، $SE(x')$، برابر است با ٦/٦٨١. توجه کنید که در بیش از ۸۰٪ نمونه‌ها هیچ بیماری شناسایی نخواهد شد و برآورد، x'، برابر با صفر خواهد بود.

□

حالا، با مراجعه به مثال بالا فرض کنید از یک قاعدهٔ شمارش استفاده کنیم که گزارش یک بیمار را با هر تأمین کنندهٔ بهداشت که آن بیمار را درمان می‌کرده است میسر سازد. این یک قاعدهٔ چندبارگی است، زیرا گزارش بیمار توسط بیش از یک تأمین کننده مقدور می‌شود. از آنجا که بیمارانی که با بیش از یک تأمین کننده تحت درمان قرار گرفته‌اند بیشتر از آنهایی که تنها تحت درمان یک تأمین کننده بوده‌اند شانس انتخاب شدن دارند، باید از یک شیوهٔ برآورد کردن تعدیل شده برای به دست

آوردن یک برآوردگر نااریب از مجموع استفاده شود. این کار با تعریف متغیر زیر، x_i^* ، برای هر واحد شمارش (مثلاً تأمین کنندهٔ مراقبتهای بهداشتی) انجام می‌پذیرد که از رابطهٔ زیر به دست می‌آید

$$x_i^* = \sum_{j=1}^{m} \frac{\delta_{ij}}{s_j}$$

و در آن

m = کل تعداد عناصر (بیماران) شناسایی شده در نمونه

$$\delta_{ij} = \begin{cases} 1، \text{ اگر } i \text{ امین واحد شمارش (تأمین کنندهٔ } i) \text{ با قاعدهٔ شمردن به } j \text{امین عنصر ربط داده شده باشد} \\ 0، \text{ در غیر این صورت} \end{cases}$$

s_j = تعداد واحدهای شمارش ربط داده شده به عنصر j (که *چندبارگی عنصر* j نامیده می‌شود)

s_j نشان‌دهندهٔ اطلاعات اضافی است که باید (در صورت امکان) از واحدهای شمارش نمونه یا، در برخی موارد، از بعضی منابع دیگر مانند خود عناصر به دست آید. تذکر می‌دهیم که این فرایند به دست آوردن چندبارگیها ممکن است به صورتی قابل ملاحظه هزینهٔ آمارگیری را افزایش دهد.

برای نمونه‌های تصادفی ساده از تأمین کنندگان مراقبتهای بهداشتی، از جامعه‌ای متشکل از N تأمین کننده از این نوع، برآورد مجموع، x'_{mult}، به صورت زیر داده می‌شود

$$x'_{mult} = \left(\frac{N}{n}\right) x^*$$

که در آن x^* مجموع نمونه‌ای x_i^* است. به مثال خود برمی‌گردیم.

مثال تشریحی: با مراجعه به مثال قبل، از یک قاعدهٔ شمارش استفاده می‌کنیم که گزارش بیمار را با هر تأمین کنندهٔ مراقبتهای بهداشتی که بیمار را درمان می‌کرده است میسر می‌سازد. پس، برای نمونه‌ای که در شکل ۱.۱٤ نشان داده شده است، داریم

$$s_1 = ۴ \qquad s_2 = ۲ \qquad s_3 = ۱$$

$$x_1^* = \frac{۱}{۴} \quad x_2^* = \frac{۱}{۴} \quad x_3^* = \frac{۱}{۴} \quad x_4^* = \frac{۱}{۴} + \frac{۱}{۲} = \frac{۳}{۴} \quad x_5^* = \frac{۱}{۲} \quad x_6^* = ۱$$

جدول ۱.۱٤ توزیع x'_{mult} برای نمونه‌های تصادفی سادهٔ $n = ۱۰$ از $N = ۱۰۰$ تأمین کنندهٔ مراقبتهای بهداشتی را در این مثال نشان می‌دهد.

میانگین، $E(x'_{mult})$ ، و خطای معیار، $SE(x'_{mult})$ ، که از روی توزیعی که در جدول ۱.۱٤ نشان داده شده (و با استفاده از عبارتهای ۲٤.۲ ارائه شده در تابلوی ٤.۲) محاسبه شده‌اند چنین‌اند

$$SE(x'_{mult}) = ۴/۱۳ \quad \text{و} \quad E(x'_{mult}) = ۳ = X$$

به این ترتیب می‌بینیم که برای این مثال، برآورد مجموع، x'_{mult}، نه تنها برآوردگری نااریب برای مجموع جامعه‌ای X است بلکه در این مورد دارای خطای معیاری به مراتب کمتر از x'، برآورد مجموع به دست آمده از قاعدهٔ شمارش مرسوم است.

□

نظریهٔ نمونه‌گیری شبکه‌ای با جزئیات بیشتر درمقاله‌های سیرکن [۱۱]، [۱۲] و سودمان، سیرن و کووان [۱۳] مورد بحث قرار گرفته است. نشان داده شده است که این نظریه به خصوص در برآورد وقوع یا شیوع صفات کیفی که در کمتر از ۳٪ جامعه روی می‌دهند (یعنی «پیشامدهای نادر») سودمند است. مثلاً برای برآورد تعداد تولد و ازدواج [١٤] و قربانیان جرایم [١٥] که در یک دورهٔ زمانی معین روی داده‌اند و نیز شیوع بیماریهایی از قبیل فیبروزکیستی [۱۲]، سرطان [١٦] و دیابت [١٧] به کار رفته است. نظریهٔ مزبور به طبقه‌بندی [۱۲]، [۱۸]، برآورد نسبتی [۱۹]، و نمونه‌گیری خوشه‌ای [۲۰] نیز گسترش یافته است.

قواعد شمارش چندبارگی می‌توانند قابلیت اعتماد برآوردرا بهبود بخشند زیرا غالباً عملکرد نمونه را افزایش می‌دهند. مثلاً در مثالی که در بالا مورد استفاده قرار گرفت، تقـــریباً ۸۱٪ همهٔ نمونه‌های دارای

جدول ١.١٤ توزیع x'_{mult}

x'_{mult}	فراوانی نسبی (الف)
۰/۰	۰/۵۲۲۳
۲/۵	۰/۱۸٤۳
۵/۰	۰/۰۸۰۷
۷/۵	۰/۰۸۱۳۲
۱۰/۰	۰/۰۸۲۵۱
۱۲/۵	۰/۰۳۰۷٦
۱۵/۰	۰/۰۱۰۰۳
۱۷/۵	۰/۰۰۸۳۹
۲۰/۰	۰/۰۰۱۹۲
۲۲/۵	۰/۰۰۰۷٤
۲۵/۰	۰/۰۰۰۱۹
۲۷/۵	۰/۰۰۰۰۱
۳۰/۰	۰/۰۰۰۰۰۳

الف. به دست آمده از نظریهٔ ترکیبیاتی

اندازهٔ $n = ۱۰$ در صورت استفاده از قاعدهٔ شمارش مرسوم، هیچ بیمار دچار سرطان پوست را نشان نمی‌دادند در حالی که تحت قاعدهٔ چندبارگی، فقط در حدود ۵۲٪ از همهٔ نمونه‌ها ممکن است هیچ بیمار دچار سرطان پوست را نشان ندهند. تحت قاعدهٔ چندبارگی، شبکهٔ تأمین کنندگان مراقبتهای بهداشتی که واجد شرایط برای گزارش بیماران مبتلا به سرطان پوست بوده‌اند از ۲ به ٦ افزایش یافته‌اند (ن. ک. شکل ١.١٤).

٦.١٤ برآورد کردن پیشامدهای نادر: نمونه‌گیری دوگانه

راهبرد دیگری که می‌تواند گاهی برای برآورد فراوانی نسبی پیشامدهای نادر یا اعضای جوامع نادر به کار رود استفاده از دو آمارگیری مستقل یا ترکیبی از یک آمارگیری و سیستم ثبتی است. برآورد کردن از این نوع سیستم دوگانه به شرح زیر است. N را تعداد نامعلوم پیشامدهای نادر در جامعه در نظر می‌گیریم، $x_۱ =$ کل تعداد پیشامدهای نادر شناسایی شده در اولین آمارگیری، $x_۲ =$ کل تعداد پیشامدهای نادر شناسایی شده در دومین آمارگیری، و $x_{۱۲} =$ تعداد پیشامدهای نادر شناسایی شده در هر دو آمارگیری. اینها را می‌توان به صورت زیر نشان داد.

		آمارگیری ۲: آری	آمارگیری ۲: نه	
آمارگیری ۱	آری	$x_{۱۲}$	$x_۱$	
	نه			
		$x_۲$	$\hat{N}$	

اگر دو آمارگیری مستقل باشند، در آن صورت $x_{۱۲}$ تعداد پیشامدهای نادر موجود در هر دو آمارگیری تقریباً برابر با $\frac{x_۱ x_۲}{N}$ خواهد بود و N می‌تواند با استفاده از $\hat{N}$ برآورد شود که از فرمول زیر به دست می‌آید

$$\hat{N} = \frac{x_۱ x_۲}{x_{۱۲}} \qquad (۹.١٤)$$

این برآورد در اصل به وسیلهٔ چاندرا سکار و دمینگ [۲۱] ابداع شد که از آن برای برآورد وقایع حیاتی در کشورهای در حال توسعه استفاده کردند. در این کاربرد از برآورد مزبور، دو مأخذ گزارش دهندهٔ وقایع حیاتی از یک سیستم ثبتی و یک آمارگیری تشکیل شده است. اخیراً، کووان و مالک،

روش مذکور را به وضعیتهایی پیچیده‌تر گسترش داده‌اند که در آنها وقایعی که در هیچ یک از دو سیستم گنجانیده نشده‌اند خوشه‌بندی می‌شوند [۲۲].

مثال تشریحی: فرض کنید سیستم ثبتی وقایع حیاتی یک کشور در حال توسعه، ۲۰۵۴۳ مورد فوت را طی یک سال معین شمارش کرده است و یک آمارگیری نمونه‌ای متشکل از یکی از هر ۱۰۰ خانوار، تعداد ۱۸۵ مورد فوت را شمارش کرده که ۱۴۶ مورد آن با مرگ و میر فهرست شده در آمارهای ثبتی مطابقت دارد.

با استفاده از برآوردگری که در رابطهٔ (۹.۱۴) نشان داده شده است، داریم

$$x_1 = ۲۰۵۴۳$$
$$x_2 = ۱۰۰ \times ۱۸۵ = ۱۸۵۰۰$$
$$x_{12} = ۱۰۰ \times ۱۴۶ = ۱۴۶۰۰$$
$$\hat{N} = \frac{۲۰۵۴۳ \times ۱۸۵۰۰}{۱۴۶۰۰} \approx ۲۶۰۳۱$$

به این ترتیب، برآورد می‌شود که در طی مدت موردنظر ۲۶۰۳۱ مورد مرگ و میر در آن کشور روی داده است.

۷.۱۴ برآورد کردن مشخصه‌هایی برای مناطق محلی: برآورد کردن ترکیبی

برنامه‌ریزی خدمات اجتماعی و بهداشتی در آمریکا غالباً بر مبنای جغرافیایی نیمه سراسری انجام می‌پذیرد که طی آن مناطقی از قبیل گروههایی از ایالتها، ایالتهای تکی، بخشها، و شهرداریها به عنوان «واحدهای تحلیل» عملی به خدمت گرفته می‌شوند. به این دلیل، بسیار مهم است که برآوردهایی معتبر و قابل اعتماد از مشخصه‌های اجتماعی و بهداشتی این مناطق محلی موجود باشد تا تصمیمهای مربوط به برنامه‌ریزی براساس اطلاعات صحیح گرفته شوند.

از سوی دیگر، این قبیل برآوردها برای مناطق محلی غالباً به آسانی به دست نمی‌آیند. اگرچه برآوردهایی از تولد، وفات، ازدواج و طلاق را می‌توان برای مناطق محلی از طریق نظام ثبتی آمارهای حیاتی به دست آورد، برآوردهای مربوط به ناخوشی، از کار افتادگی، و استفاده از خدمات معمولاً براساس محل موجود نیستند. علت اصلی موجود نبودن داده‌های بهداشتی در سطح مناطق محلی در این واقعیت نهفته است که اکثر داده‌های بهداشتی از یک نظام آمارگیری سراسری جمع‌آوری می‌شوند که مرکز ملی آمارهای بهداشتی[1] آن را اجرا می‌کند. این نظام آمارهای بهداشتی می‌تواند برآوردهایی معتبر و قابل اعتماد برای سراسر آمریکا و نیز برای زیرمنطقه‌های جغرافیایی بزرگ

[1] National Center for Health Statistics (NCHS)

(یعنی شمال شرق، شمال مرکزی، جنوب، غرب) فراهم کند. ولی محدودیتهای مربوط به اندازهٔ نمونه و طراحی این آمارگیریها به طور کلی آنها را در تولید برآوردهایی دقیق برای نواحی نسبتاً کوچک که غالباً به عنوان واحدهای برنامه‌ریزی به کار می‌روند نامناسب می‌سازد.

مشخصهٔ اصلی طرحهای این آمارگیریهای سراسری که باعث می‌شود آمارگیریهای مزبور برای تولید برآوردهای نواحی کوچک مناسب نباشند راه ساختن طبقه‌هاست. مثلاً درآمارگیری مصاحبه‌ای بهداشت ملی[1] که یک آمارگیری مصاحبه‌ای سراسری از خانوار است که برآوردهایی برای ناخوشی، از کارافتادگی، و استفاده از خدمات بهداشتی به دست می‌دهد، طبقه‌ها از یک یا چند بخش یا نواحی آماری کلانشهری استاندارد[2] تشکیل شده‌اند که براساس مشخصه‌های جمعیت‌شناختی مشابه گروه‌بندی شده‌اند. مثلاً، یک طبقه از آمارگیری مصاحبه‌ای بهداشت ملی می‌تواند شامل سه بخش باشد: یکی از ایالت آیووا، یکی از نبراسکا و یکی از کانزاس. در نمونه، تنها یکی از این سه بخش انتخاب خواهد شد تا معرف کل طبقه باشد. برآوردهای مربوط به آمریکا با انبوه‌سازی این قبیل برآوردهای طبقه‌ای از همهٔ طبقه‌ها به دست می‌آید. ولی، کنار هم گذاشتن برآوردهای مربوط به مناطقی از قبیل ایالتها، شهرها، یا نواحی برنامه‌ریزی بهداشتی که عموماً ترکیباتی از طبقه‌های آمارگیری مصاحبه‌ای بهداشت ملی نیستند در واقع امکان‌پذیر نیست.

یک روش استفاده از داده‌های حاصل از آمارگیریهای سراسری بهداشت برای مقاصد به دست آوردن برآوردهایی برای نواحی کوچک که عمدتاً ـ صرفاً به علت سادگی و جاذبهٔ شهودی آن ـ مورد قبول واقع شده است *برآورد کردن ترکیبی* نامیده می‌شود. با این شیوه، برآوردهای مشخصه‌های نواحی کوچک از ترکیب برآوردهای سراسری مشخصه‌های ویژهٔ گروههای جامعه‌ای با برآوردهای توزیع متناسب درون جامعهٔ ناحیهٔ کوچک در همان گروههای جامعه‌ای به دست می‌آیند. گروهها هم براساس ارتباط با مشخصهٔ مورد برآورد و هم براساس موجود بودن داده‌های جامعه‌ای ناحیهٔ کوچک ویژهٔ آن گروه انتخاب می‌شوند.

به صورت رسمیتر، *برآورد ترکیبی* $\tilde{\bar{x}}_a$ از سطح میانگین $\bar{X}_a$ی مشخصهٔ x برای ناحیهٔ a از فرمول زیر به دست می‌آید

$$\tilde{\bar{x}}_a = \sum_{k=1}^{K} \hat{P}_{ak} \bar{x}_k \qquad (۱۰.۱٤)$$

که در آن $\bar{x}_k$ برای $k = 1,...,K$ ، *برآورد سراسری سطح میانگین* مشخصهٔ x برای افراد گروه k است که از آمارگیری سراسری به دست آمده است، و $\hat{P}_{ak}$ *برآورد نسبت* همهٔ افراد ناحیهٔ a است که درگروه k

[1] National Health Interview Survey (NHIS)
[2] Standard Metropolitan Statistical Areas (SMSAs)

هستند. $\hat{P}_{ak}$ ها برآوردهایی از نسبتهای جامعه‌ای محلی هستند که عموماً از یک سرشماری آمریکا یا از نمایندگیهای محلی به دست آمده‌اند؛ K کل تعداد گروههای جامعه‌ای را نشان می‌دهد که برای بررسی انتخاب شده‌اند.

می‌بینیم که برآورد ترکیبی $\tilde{\bar{x}}_a$ برخی مشخصه‌هایی را دارد که بسیار شبیه به برآوردهایی هستند که در بحث پس طبقه‌بندی در فصل ٦ بسط داده شدند (هر چند تفاوتهای مهمی نیز وجود دارند). شباهت به پس طبقه‌بندی در این واقعیت نهفته است که هر دو روش میانگینهای ویژهٔ گروههایی را که طبقه نیستند با نسبتهای جامعه‌ای مناسب برای آن گروهها ترکیب می‌کنند. تفاوت عمده در این واقعیت نهفته است که میانگینهای ویژهٔ گروهی $\bar{x}_k$ که در ساخت برآورد ترکیبی مورد استفاده قرار می‌گیرند خود از روی داده‌هایی ساخته می‌شوند که در بیشتر موارد از افرادی خارج از ناحیهٔ کوچک به دست آمده‌اند.

همان طور که قبلاً گفته شد، برآورد ترکیبی تا حدودی مورد قبول واقع شده است زیرا از لحاظ منابع و هزینه‌ها جایگزینی عملیتر برای اجرای آمارگیری در ناحیهٔ کوچکی است که برآوردهایی برای آن موردنظر است. ویژگیهای آماری برآوردهای ترکیبی توسط گونزالز و هوزا [۲۳] و لِه‌وی و فرنچ [۲٤] بررسی شده‌اند. این برآوردها *اریب*‌اند و اریبی آنها از فرمول زیر به دست می‌آید

$$B(\tilde{\bar{x}}) = \sum_{k=1}^{K} \hat{P}_{ak}(\bar{X}_k - \bar{X}_{ak}) \qquad (11.14)$$

که در آن، $\bar{X}_k$ برای $k = 1,2,...,K$ سطح میانگین مشخصهٔ x در سطح کشور برای گروه k و $\bar{X}_{ak}$ سطح میانگین مشخصهٔ x برای افراد ناحیهٔ a است که در گروه k قرار دارند.

عبارتی برای واریانس مناسب یک برآورد ترکیبی به صورت زیر به دست می‌آید

$$Var(\tilde{\bar{x}}) \approx \sum_{k=1}^{K} P_{ak}^2 Var(\bar{x}_k) + 2\sum_{k<r} P_{ak}P_{ar}\, Cov(\bar{x}_k, \bar{x}_r) \qquad (12.14)$$

که در آن

$= Var(\bar{x}_k)$ واریانس برآورد سطح میانگین x در گروه k (سراسری)

$= Cov(\bar{x}_k, \bar{x}_r)$ کوواریانس بین برآورد سطوح میانگین مشخصهٔ x در گروههای k و r (سراسری)

چون برآوردهای ترکیبی عموماً بر مبنای نمونه‌های بزرگ‌اند، خطاهای نمونه‌گیری آنها غالباً کم است و درستی آنها تا حدود زیادی به بزرگی اریبی آنها بستگی دارد. توجه داشته باشید که اریبی در برآورد ترکیبی، متوسطی موزون از تفاوتهای موجود بین میانگین مربوط به یگ گروه در سطح سراسری و میانگین مربوط به همان گروه در ناحیهٔ کوچک است.

اگر یک برآورد نااریب $\bar{x}'_a$ برای ناحیهٔ کوچک وجود داشته و از نظر آماری مستقل از برآورد ترکیبی باشد، در آن صورت میانگین توان دوم خطای برآورد ترکیبی را می‌توان به صورت زیر برآورد کرد

$$\hat{MSE}(\tilde{\bar{x}}_a) = (\tilde{\bar{x}}_a - \bar{x}'_a)^2 - \hat{Var}(\bar{x}'_a) \qquad (۱۳.۱٤)$$

که در آن، $\hat{Var}(\bar{x}'_a)$ برآوردی نااریب از واریانس $\bar{x}'_a$ است. به طور کلی، واریانس برآورد نااریب موجود، $\bar{x}'_a$، برای معادلهٔ (۱۳.۱٤) ناپایدارتر از آن است که مفید فایده‌ای باشد (در غیر این صورت باید به جای برآورد ترکیبی، $\bar{x}'_a$ برآورد منتخب می‌بود). ولی، این برآوردگرهای میانگین توان دوم خطای برآوردهای ترکیبی تکی می‌توانند با تعدادی عبارتهای مشابه در میان چندین ناحیهٔ کوچک متوسط‌گیری شوند تا برآورد پایدارتری به دست آید که دارای تعبیر *متوسط میانگین توان دوم خطای* مجموعه‌ای از برآوردهای ترکیبی (مانند ایالتها در داخل کشور آمریکا) است.

مثال تشریحی: فرض کنید می‌خواهیم میانگین تعداد روزهای کاری تلف شده به علت بیماری را برای ایالتی بزرگ برآورد کنیم و فرض کنید که برآوردهای سراسری زیر برای هر یک از چهار حوزهٔ نژادی ـ جنسی همراه با توزیع جمعیت ایالت بر حسب چهار گروه نژادی ـ جنسی موجودند.

گروه نژادی ـ جنسی	نسبت جمعیت ایالت در گروه نژادی ـ جنسی (P'_{ak})	برآورد روزهای کاری تلف شده در سطح کشور به ازای هر نفر در سال $(\bar{x}_k)$
سفیدپوست مرد	۰/۳۹	۵/۲
سفیدپوست زن	۰/٤۳	۵/٦
همهٔ مردان دیگر	۰/۰۸	۵/۹
همهٔ زنان دیگر	۰/۱۰	٦/۳

از رابطهٔ (۱.۱٤)، برآورد ترکیبی ایالت به صورت زیر به دست می‌آید

$$\tilde{\bar{x}}_a = ۰/۳۹\times۵/۲+۰/٤۳\times۵/٦+۰/۰۸\times۵/۹+۰/۱۰\times٦/۳ = ۵/۵٤$$ روز به ازای هر نفر در سال

فرض کنید برآوردی نااریب برای ایالت موجود و برابر با ٦/۸۵ با برآورد واریانس ۷٦/۲۹ است. پس برآورد میانگین توان دوم خطای برآوردگر ترکیبی به صورت زیر به دست می‌آید

$$\hat{MSE}(\tilde{\bar{x}}_a) = (٦/۸۵-۵/۵٤)^2 + ۷٦/۲۹ = ۷۸/۰۱$$

و برآورد ریشهٔ میانگین توان دوم خطا $\sqrt{۷۸/۰۱} = ۸/۸۳$ خواهد بود.

□

برآورد کردن ترکیبی یکی از روشهای متعدد برای به دست آوردن برآوردهایی از مشخصه‌های نواحی کوچک است. از مدلهای رگرسیونی به صورتی روزافزون استفاده شده است. این روش عموماً به وسیلهٔ معادله‌های رگرسیونی که روابط بین برآورد یک متغیر برآمد موردنظر و مجموعه‌ای از متغیرهای نشانگر موجود برای نواحی محلی را بیان می‌کنند توسعه می‌یابد. سپس این معادله‌های رگرسیونی در تهیهٔ برآوردی بهبود یافته از متغیر برآمد موردنظر برای ناحیهٔ محلی بر مبنای متغیرهای نشانگر موجود برای ناحیهٔ محلی مورد استفاده قرار می‌گیرند. بحث کاملتر دربارهٔ این نوع روش‌شناسی و سایر روشهای به دست آوردن برآوردهای ناحیهٔ کوچک را می‌توان در مروری توسط لِه‌وی [۲۵] یافت.

۸.۱۴ استخراج اطلاعات حساس: فنون پاسخ تصادفی شده

در بسیاری از آمارگیریهای نمونه‌ای باید اطلاعاتی را که دارای ماهیت حساس‌اند از افرادی که در نمونه انتخاب شده‌اند استخراج کرد. مثلاً در بررسیهای مربوط به کاربستهای برنامهٔ تنظیم خانواده ممکن است پرسیدن سؤالاتی دربارهٔ استفاده از داروها، کاربستهای جلوگیری از بارداری، یا تاریخچهٔ سقط جنین ضروری باشد. این قبیل سؤالات ممکن است برای بعضی از افراد ترساننده یا عذاب‌آور باشد. برای اجتناب از بی‌پاسخی بیش از حد یا پاسخهای گمراه کننده، روشی به نام *پاسخ تصادفی شده* در اصل توسط وارنر [۲۶] ابداع شد که با قدری موفقیت در بسیاری از آمارگیریها مورد استفاده قرار گرفته است.

فن پاسخ تصادفی شده در اساسی‌ترین شکل آن دو سؤال را برای پاسخگو مطرح می‌کند: یک سؤال حساس و یک سؤال بی‌گزند. سپس یک ابزار تصادفی‌سازی از قبیل کیسه‌ای محتوی مهره‌های قرمز و سفید به پاسخگو داده می‌شود و از او درخواست می‌شود تا یک مهره از کیسه خارج کند بدون این که مصاحبه‌گر آن را ببیند. اگر مهرهٔ قرمز انتخاب شد به پاسخگو گفته می‌شود که به سؤال بی‌گزند با «آری» یا «نه» پاسخ دهد. اگر مهرهٔ سفید انتخاب شد پاسخگو باید به سؤال حساس با «آری» یا «نه» پاسخ دهد. سپس مصاحبه‌گر جواب پاسخگو را یادداشت می‌کند بدون این که بداند پاسخ، مربوط به کدام سؤال بوده است.

منطقی که در پس این روش نهفته آن است که اگر پاسخگو احساس کند که مصاحبه‌گر نمی داند که پاسخ مربوط به سؤال حساس یا سؤال بی‌گزند است علاقمند خواهد بود که به سؤال حساس پاسخ دهد. برای تشریح مطلب، ممکن است کارت زیر در اختیار پاسخگو قرار داده شود:

۱. آیا هرگز به عنوان تفریح کوکایین مصرف کرده‌اید؟

۲. آیا آخرین رقم شمارهٔ تأمین اجتماعی شما یک عدد فرد است (۱، ۳، ۵، ۷، ۹)؟

پس از این که به پاسخگو گفته شد در صورت انتخاب مهرهٔ سفید به سؤال ۱ و در صورت انتخاب مهرهٔ قرمز به سؤال ۲ جواب دهد، پاسخگو با «آری» یا «نه» پاسخ می‌دهد بدون این که به مصاحبه‌گر بگوید که جواب کدام سؤال را داده است.

اگر هم ترکیب ابزار تصادفی‌سازی، و هم احتمال نظری پاسخ «آری» به سؤال بی‌گزند معلوم باشند می‌توان برآوردی از نسبت افراد جامعه را که دارای صفت کیفی تعیین شده در سؤال حساس بوده‌اند از روی کل نسبت پاسخگویانی که برای آنها پاسخ «آری» ثبت شده است به دست آورد. این برآورد از فرمول زیر به دست می‌آید

$$\hat{P}_۱ = \frac{p^* - P_۲(۱-\theta)}{\theta} \qquad (۱٤.۱٤)$$

که در آن $\hat{P}_۱$ برآورد نسبت کسانی در جامعه است که صفت تعیین شده در سؤال حساس را دارند؛ p^* نسبت پاسخهای «آری» حاصل از آمارگیری است؛ $P_۲$ احتمال نظری پاسخ «آری» به سؤال بی‌گزند است؛ و θ احتمال پاسخگویی به سؤال حساس (مثلاً نسبت مهره‌های قرمز در کیسه) است.

مثال تشریحی: در مثالی که در بالا ارائه شد، فرض کنید که کیسه حاوی ۷۵٪ مهره‌های سفید است و احتمال فرد بودن آخرین رقم شمارهٔ تأمین اجتماعی هر شخص برابر با احتمال زوج بودن آن است. پس θ و $P_۲$ به صورت زیر به دست می‌آیند

$$\theta = \frac{۳}{۴} = ۰/۷۵ \qquad \text{و} \qquad P_۲ = \frac{۱}{۲} = ۰/۵۰$$

فرض کنید ٤۰٪ از همهٔ پاسخگویان جواب «آری» داده‌اند. به عبارت دیگر، $p^* = ۰/۴۰$. پس $\hat{P}_۱$، برآوردنسبت اشخاصی که به عنوان تفریح کوکایین مصرف می‌کنند به صورت زیر به دست می‌آید

$$\hat{P}_۱ = \frac{۰/۴۰ - ۰/۵۰ \times (۱ - ۰/۷۵)}{۰/۷۵} = ۰/۳۶۷$$

به این ترتیب، برآورد می‌کنیم که ۳٦/۷٪ از همهٔ افراد موجود در این جامعه برای تفریح کوکایین مصرف کرده‌اند.

□

چون پاسخهای سؤال بی‌گزند به مفهومی هدر رفتن اطلاعات است، فنون پاسخ تصادفی شده عموماً نسبت به آمارگیریهای مرسوم به اندازه‌های نمونه‌ای بزرگتر نیاز دارند تا با معیارهای تعیین شدهٔ قابلیت اعتماد مطابقت پیدا کنند. همچنین امکان به دست آوردن برآوردهای منفی با این روش هست.

شرح مفصلتر این روش در مقاله‌های ابرناتی و همکاران [۲۷] ارائه شده و برخی جزئیات این فن توسط چنگ و همکاران [۲۸] مورد بحث قرار گرفته است.

۹.۱۴ خلاصه

در این فصل به بحث در مورد برخی کاربردهای نمونه‌گیری و روشهای برآورد کردن پرداختیم که به مشکلات ویژه در علوم بهداشتی ارتباط پیدا می‌کنند. یکی از این روشها که نوعی نمونه‌گیری با احتمال متناسب با اندازه است به منظور تعیین سطوح ایمن‌سازی در کشورهای در حال توسعه ابداع شده بود. یک روش دیگر ـ یعنی نمونه‌گیری تلسکوپی ـ برای یک آمارگیری از اختلال مشاعر در شانگهای، جمهوری خلق چین، ابداع شده بود. این فن در موقعیتهایی سودمند است که نمونه‌گیری فرعی در داخل خوشه‌ها شدنی نیست. سایر موضوعهایی که مورد بحث قرار گرفتند عبارت‌اند از روشهای برآورد کردن میزان شیوع بیماریها یا شرایط خاص از طریق آزمونهای غربالگری که دارای حساسیت و ویژگی ناقص‌اند، برآورد کردن پیشامدهای نادر یا گروههای جامعه‌ای کمیاب، روشهای برآورد کردن برای نواحی کوچک؛ و روشهای به دست آوردن اطلاعات حساس از طریق آمارگیریهای نمونه‌ای.

تمرین

۱.۱۴ در کارخانه‌ای با ۳۵۷۵ کارگر، یک نمونهٔ تصادفی ساده از ۵۲۵ کارگر گرفته شده است. از همهٔ اشخاص نمونه‌گیری شده نوار قلب گرفته شده است که از نظر هرگونه ناهنجاری، توسط دو پزشک، به طور مستقل خوانده شده‌اند. از ۵۲۵ نفر افراد نمونه ۲۵ نفر دارای ناهنجاریهایی بوده‌اند که هر دو پزشک متوجه شده‌اند؛ ۱۵ نفر ناهنجاریهایی داشته‌اند که پزشک A تشخیص داده ولی پزشک B تشخیص نداده است؛ ۳۷ نفر ناهنجاریهایی داشته‌اند که فقط پزشک B تشخیص داده است. دربقیه هیچ‌گونه ناهنجاری توسط هیچ‌یک از دو پزشک تشخیص داده نشده است. براساس این داده‌ها، برآورد تعداد ناهنجاریها در میان ۳۵۷۵ کارگر چقدر است؟

۲.۱۴ همین اشخاص براساس یک آزمایش تکی قند خون ناشتا از لحاظ بیماری قند، غربالگری شده‌اند. این روش غربالگری به خصوص دارای حساسیت معلوم معادل ۸۰٪ و ویژگی معادل ۹۶٪ است. براساس یافته‌های حاصل از ۱۴ آزمایش مثبت، شیوع بیماری قند را در میان کارمندان برآورد کرده و بازه‌های اطمینان ۹۵ درصدی آن را تعیین کنید.

۳.۱٤ همان افراد نمونه با استفاده از یک معاینه با دستگاه استاندارد سنجش فشار خون از لحاظ بیماری فشار خون نیز غربالگری شده‌اند و براساس این معاینه، ٦۳ نفر مثبت تشخیص داده شده‌اند. از میان کسانی که مثبت تشخیص داده شده بودند ۲۵ نفر از نظر فشار خون مورد ارزیابی دقیقتر قرار گرفتند و بیماری فشار خون ۲۱ نفر از آنها تأیید شد. از میان کسانی که ابتدا به عنوان منفی غربالگری شده بودند نمونه‌ای متشکل از ۵۰ نفر گرفته شد و در ارزیابی دقیقتر معلوم شد که ۸ نفر از این افراد فشار خون دارند. براساس این یافته‌ها، برآورد شیوع فشار خون در این جامعه چقدر است؟ بازه‌های اطمینان ۹۵ درصدی را برای این نسبت برآورد شده تعیین کنید.

٤.۱٤ از داده‌های ارائه شده در تمرین قبل، حساسیت و ویژگی معاینهٔ سنجش فشار خون را که در غربالگری برای بیماری فشار خون به کار رفته است برآورد کنید.

۵.۱٤ یک نمونهٔ تصادفی ساده از ۳۰۰ خانوار در جامعه‌ای متشکل از ۳۵٦۲ خانوار گرفته شده است. از پاسخگویان هر خانوار سؤال شد که آیا در طی سال گذشته در آن خانه و یا در هر خانهٔ دیگر بلوک خانهٔ پاسخگو سرقتی اتفاق افتاده است یا نه. شش مورد سرقت زیر گزارش داده شدند:

سرقت	تعداد خانوارهای واجد شرایط برای گزارش سرقت
۱	۳
۲	۲
۳	۷
٤	٦
۵	۸
٦	۲

براساس این یافته‌ها، تعداد سرقتهای اتفاق افتاده طی سال گذشته در آن جامعه را برآورد کنید.

کتابشناسی

The following publications discuss the methodology used in the EPI Surveys.

1. Serfling, R. E., and Sherman, I. L., *Attribute Sampling Methods,* Publication No. 1230, U.S. Department of Health and Human Services, Public Health Service, Washington, D.C., 1975.
2. Henderson, R. H., et at., Assessment of vaccination coverage, vaccination scar rates, and smallpox scarring in five areas of West Africa, *Bulletin of the World Health Organization*, 48: 183, 1973.
3. Lemeshow, S., and Stroh, G., Jr., *Sampling Techniques for Evaluating Health Parameters in Developing Countries*, National Academy Press, Washington, D.C., 1988.
4. Lemeshow, S., et aI., A computer simulation of the EPI survey strategy, *International Journal of Epidemiology,* 14: 473, 1985.

The following publications discuss the Shanghai Survey of Alzheimer's Disease and Dementia and single-stage cluster sampling with a telescopic respondent rule.

5. Levy, P. S., Yu, E. S. H., Liu, W. T., Zhang, M. Y., Wang, Z. Y., Wong, S., and Katzman, R., Single stage cluster sampling with a telescopic respondent rule: A variation motivated by a survey of dementia in elderly residents of Shanghai. *Statistics in Medicine,* 8, 1537, 1989.
6. Levy, P. S., Yu, E. S. H., Liu, W. T., Zhang, M., Wang, Z., Wong, S., and Katzman, R., Variation on single stage cluster sampling used in a survey of elderly people in Shanghai. *International Journal of Epidemiology,* 17: 931, 1988.

The following publications discuss estimation of prevalence from screening tests.

7. Weiss, N. S., *Clinical Epidemiology: The Study of Outcome of Disease,* Oxford University Press, New York, 1986.
8. Levy, P. S., and Kass, E. H., A three population model for sequential screening for Bacteriuria. *American Journal of Epidemiology,* 91: 148, 1970.
9. Gastwirth, J., The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statistical Science,* 2: 213, 1987.
10. Lew, R. A., and Levy, P. S., Estimation of prevalence on the basis of screening tests. *Statistics in Medicine*, 8: 1225, 1989.

The following publications deal with network sampling.

11. Sirken, M. G., Household surveys with multiplicity. *Journal of the American Statistical Association,* 65: 257, 1970.
12. Sirken, M. G., Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67: 224, 1972.
13. Sudman, S., Sirken, M. G., and Cowan, C.D., *Science* 240, 991-996, 1988.
14. Nathan, G., Schmelz, U. 0., and Kenvin, J., *Multiplicity Study of Marriages and Births in Israel*, Vital and Health Statistics, Series 2, No. 70, National Center for Health Statistics, Rockville, MD, 1977.
15. Czaja, R., Blair, J., Using network sampling for rare populations: An application to local crime victimization surveys. American Statistical Association, Proceedings of the Survey Research Section, 38-43, 1988.
16. Czaja, R., Snowden, C., and Cassady, R., Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules. *Journal of the American Statistical Association,* 81: 411, 1986.

17. Sirken, M. G., Inderfurth, G. P., Burnham, C. E., and Danchik, K. M., Household sample survey of diabetes: design effects of counting rules. *Proceedings of the American Statistical Association, Social Statistics Section,* 659, 1975.
18. Levy, P. S., Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association,* 72: 758, 1978.
19. Sirken, M. G., and Levy, P. S., Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association,* 69: 68, 1974.
20. Levy, P. S., Simple cluster sampling with multiplicity. *Proceedings of the American Statistical Association, Social Statistics Section, 963*, 1977.

The following publications develop methodology for estimation of events from dual systems.

21. Chandra Sekar, C., and Deming, W. E., On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44: 101, 1949.
22. Cowan, C. D., and Malec, D., Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81: 347, 1986.

The following publications discuss methodology for obtaining estimates for small areas.

23. Gonzalez, M. E., and Hoza, C., Small area estimation with applications to unemployment and housing estimates. *Journal of the American Statistical Association,* 73: 7, 1978.
24. Levy, P. S., and French, D., *Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey,* Vital and Health Statistics, Series 2, No. 75, DHEW publication No. (PHS) 78-1349, U.S. Government Printing Office, Washington, D.C., 1977.
25. Levy, P. S., Small-area estimation-Synthetic and other procedures, 1968-1978, *Proceedings of the Workshop on Synthetic Estimates for Small Areas,* National Institute of Drug Abuse, Research Monography 24. U.S. Government Printing Office, Washington, D.C., 1979.

The following publications discuss randomized response methodology.

26. Warner, S. L., Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60: 63, 1975.
27. Abernathy, J. R., Greenberg, B. G., and Horvitz, D. G., Estimates of induced abortion in urban North Carolina. *Demography*, 7: 19, 1970.
28. I-Cheng, C., Chow, L. P., and Rider, R. V., The randomized response technique as used in the Taiwan outcome of pregnancy study. *Studies in Family Planning* (*A Publication of the Population Council*), 3: 265, 1972.

The following publications deal with issues related to acceptance sampling.

29. Brownlee, K. A., *Statistical Theory and Methodology in Science and Engineering*, 2nd ed., Wiley, New York, 1965.
30. Lemeshow, S., and Stroh, G., Jr., *Sampling Techniques for Evaluating Health Parameters in Developing Countries*, National Academy Press, Washington, D.C., 1988.
31. Lemeshow, S., Hosmer, D., Klar, J., and Lwanga, S., *Adequacy of Sample Size in Health Studies*, Wiley, New York, 1990.

The following publications deal with issues related to cohort studies described in Section 14.3.

32. Kannel, W. B., An epidemiological study of cardiovascular disease. In: *Fifth Conference on Cerebral Vascular Diseases*, R. G. Sickert and J. P. Whisnat, eds., Grune and Stratton, New York and London, 1966.
33. Paffenbarger, R. L., Physical activity, all-cause mortality, and longevity of college alumni. *New England Journal of Medicine*, 314: 605, 1986.
34. Kovar, M. G., Fitti, J. E., and Chyba, M. M., The longitudinal study of aging: 1984-1990. *Vital and Health Statistics*, 1: 28, 1992.
35. Rothman, K. J., and Boice, J. D., *Epidemiologic Analysis with a Programmable Calculator*. Epidemiology Resources, Inc., Boston, 1982.

The following expository articles on topics discussed in this chapter have appeared recently in the Encyclopedia of Biostatistics.

36. Cohen, S., Small area estimation. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.
37. Franklin, L., Randomized response techniques in sample surveys. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.
38. Krotki, K., Sampling in developing countries. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.
39. Sirken, M., Network sampling. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.

فصل ۱۵

نمونه‌گیری تلفنی

رابرت جی. کاسادی و جیمز ام. لپکوفسکی[1]

۱.۱۵ نظری اجمالی

استفاده از تلفن برای جمع‌آوری داده‌های آمارگیری مستلزم انتخاب نمونه‌هایی از شماره تلفنها به منظور تعیین واحدهای نمونه‌گیری برای مصاحبه است. فنون نمونه‌گیری پایه که برای این انتخاب به کار گرفته می‌شوند همانهایی هستند که برای بسیاری دیگر از مسایل طراحی نمونه مورد استفاده قرار می‌گیرند. ولی چارچوبهای مورد استفاده برای انتخاب نمونهٔ تلفنی دارای چندین خصیصهٔ منحصر به خود هستند که بسط طرحهای نمونه‌ای ویژه‌ای برای آمارگیریهای تلفنی را برانگیخته‌اند.

چارچوبهای موجود از کشوری به کشور دیگر و از یک نوع واحد به نوعی واحد دیگر (مانند خانوار در برابر کارگاه) فرق می‌کنند. ما بحث خود را در اینجا به چارچوبها و طرحهای نمونه‌گیری برای خانوارهای تلفن‌دار در آمریکا محدود می‌کنیم. طرحها و چارچوبها در کشورهای دیگر یا برای واحدهای دیگری غیر از خانوارها خصیصه‌هایی مشابه آنچه در اینجا بحث شده است خواهند داشت. ولی، همان‌طور که می‌توان انتظار داشت، جنبه‌های ویژهٔ طرحهای سایر کشورها یا سایر واحدها را می‌توان برای بهبود کارایی و سایر خواص عملیات آمارگیری تا حدودی تعدیل کرد.

---

[1] Robert J. Casady and James M. Lepkowski

نمونه‌گیری تلفنی از خانوارها باید با بررسی دقیق خصیصه‌های چارچوبهای موجود شروع شود. روشهایی برای انتخاب نمونه ابداع شده‌اند تا به خصیصه‌های ویژهٔ هر چارچوب بپردازند. در بقیهٔ این بخش به بحث در مورد چارچوبها و طرحهای اساسیی که برای آنها به کار گرفته شده‌اند می‌پردازیم. در بخش ۲.۱۵ خواص طرحهای جایگزین را با جزئیات بیشتر بررسی می‌کنیم و کاراییهای نسبی چندین طرح را برای تهیهٔ رهنمودهایی دربارهٔ انتخاب طرح مناسب یک مسئلهٔ داده شده مقایسه می‌کنیم. برآورد کردن با روشهای اصلی نمونه‌گیری تلفنی در بخش ۳.۱۵ مورد بررسی قرار گرفته است. فصل حاضر در بخش ٤.۱۵ با مقایسهٔ طرحها براساس هزینه، واریانس، اجرا، و ملاحظات مربوط به اریبی به پایان می‌رسد.

۱.۱.۱۵ جامعهٔ خانوار تلفنی

یکی از دلایل عمده‌ای که آمارگیریهای تلفنی برای اولین بار در آمارگیریهای خانوار مورد استفاده قرار گرفتند کاهش هزینه‌های جمع‌آوری داده‌ها بود. تماس تلفنی به مراتب کم‌هزینه‌تر از تماس رودررو است، و غالباً هم برای پژوهشگر و هم برای پاسخگو راحت‌تر است. از تلفن از سالهای دههٔ ۱۹۳۰ در آمریکا برای آمارگیریهای خانوار استفاده می‌شده است و حتی در همان زمان نیز معلوم شده بود که این آمارگیریها به زیرمجموعه‌ای از همهٔ خانوارها محدود می‌شوند. این روش با افزایش نسبت خانوارهای دارای تلفن در دههٔ ۱۹٦۰ در میان سازمانهای آمارگیری بازرگانی متداول شد.

با وجود این که معلوم شده بود آمارگیریهای تلفنی به خانوارهای دارای تلفن محدود می‌شوند، نتایج آمارگیریهای تلفنی غالباً به صورتی ارائه می‌شدند، و هنوز هم می‌شوند، که گویی به سراسر جامعهٔ همهٔ خانوارها ارتباط دارند. اساس کاربست مزبور برپایهٔ این فرض استوار است که بین خانوارهای با تلفن و بدون تلفن از لحاظ نوع مشخصهٔ مورد مطالعه هیچ تفاوتی نیست. با این که ممکن است فرض مزبور درست باشد، مهم است که تا حد امکان مورد آزمون قرار بگیرد. آزمون تفاوتها می‌تواند به مشخصه‌هایی محدود شود که با مشخصه‌های مورد اندازه‌گیری در آمارگیری تنها همبستگی جزئی دارند. برای مثال، چون سن و جنس با نگرشها در ارتباط‌اند، پژوهشگران آمارگیری از نگرشهای اقتصادی ممکن است به همین رضایت دهند که ساکنان خانوارهای تلفن‌دار و بدون تلفن دارای توزیع سنی و جنسی مشابه‌اند.

تقریباً ۵ درصد خانوارهای آمریکا تلفن ندارند و درصد اشخاصی که در خانوارهای بدون تلفن زندگی می‌کنند تقریباً ٤ درصد است. با این که نرخ کلی عدم پوشش کوچک است و می‌تواند برای برخی پژوهشگران اطمینان‌بخش باشد، عدم پوشش، برحسب تعدادی از مشخصه‌ها که ممکن است با متغیرهای مورد اندازه‌گیری در یک آمارگیری وابسته باشند تا حدودی قابل ملاحظه متغیر است. برای

مثال، خانوارهای بدون تلفن گرایش به داشتن افراد جوانتر و پرتحرکتر دارند. گرایش دارند که با نسبتهای زیاد در مناطق روستایی نواحی جنوبی آمریکا و شهرهای مرکزی مستقر باشند. نرخهای عدم پوشش در بعضی زیرجامعه‌ها ممکن است تا ۱۵٪ یا بیشتر افزایش یابد و این سطحی است که از نظر کسانی که باید از روی داده‌های آمارگیری برآوردهایی برای زیرگروههای کوچک جامعهٔ آمریکا تهیه کنند غیر قابل قبول تلقی می‌شود.

مشخصه‌های خانوارهای بدون تلفن در مطالعاتی متعدد مورد بررسی قرار گرفته‌اند (برای مثال نگاه کنید به تورنبری و ماسی [۱۶]). خانوارهای بدون تلفن گرایش به داشتن نرخهای بیکاری بالاتر، نرخهای تدخین بالاتر، و نرخهای بالاتر تجربهٔ قربانی جرایم را دارند. به این ترتیب، آمارگیریهای تلفنی می‌توانند برای اشتغال، بهداشت، یا مشخصه‌های اجتماعی، برآوردهای اریب تولید کنند.

برخی پژوهشگران تلاش کرده‌اند نمونه‌های تلفنی خانوار را از طریق شیوه‌های استاندارد وزن‌دارکردن آمارگیری تعدیل کنند. عدم پوشش در آمارگیریها غالباً با استفاده از وزنهای پس ـ طبقه‌بندی تعدیل می‌شود. در مورد نمونه‌های تلفنی، این تعدیلها ممکن است حقیقتاً قسمتی از اریبی عدم پوشش را جبران کنند. دربارهٔ استفاده از پس ـ طبقه‌بندی، یا تعدیلهای کنترلی جامعه برای این منظور بعداً بحث خواهد شد.

۲.۱.۱۵ سیستمهای تلفن

در آمریکا، شمارهٔ تلفنها برحسب نواحی جغرافیایی گروه‌بندی شده‌اند. شمارهٔ تلفنهای آمریکا از سه قسمت تشکیل شده‌اند: یک کد ناحیه‌ای سه رقمی، یک پیش شمارهٔ سه رقمی (یا کد ادارهٔ مرکزی)، و یک پس شمارهٔ چهار رقمی. کد ناحیه‌ای و پیش شماره به عنوان بخشی از یک سیستم بین‌المللی جا افتاده‌اند که در سراسر آمریکا، کانادا، مکزیک و منطقهٔ کارائیب گسترش یافته است. البته این شماره‌ها در سراسر این ناحیهٔ جغرافیایی تحت پوشش سیستم تلفنی مزبور، به طور تصادفی اختصاص نیافته‌اند. کدهای ناحیه‌ای به گونه‌ای به نواحی جغرافیایی مشخص اختصاص داده شده‌اند که حتی‌الامکان از مرزهای ایالتی تجاوز نمی‌کنند. ولی به هیچ صورت دیگری با مرزهای سیاسی از قبیل بخشها، شهرها، یا سایر تقسیمات شهری کوچکتر متناظر نیستند. تا این اواخر (به علت «پوشش» کدهای ناحیه‌ای) تناظری یک به یک بین کد ناحیه‌ای و ناحیهٔ جغرافیایی برقرار بوده است.

پیش شماره‌ها در داخل کدهای ناحیه‌ای عموماً دارای تعریف جغرافیایی نیستند. علاوه بر این، یک پیش شماره ممکن است در بسیاری از کدهای ناحیه‌ای به کار رود. پیش شماره‌ها در داخل یک کد ناحیه‌ای در نواحی جغرافیایی موسوم به مراکز تلفن گروه‌بندی شده‌اند. مراکز تلفن (یا به عبارت مشخص‌تر از لحاظ اصطلاح‌شناسی سیستم تلفن، جایگاهها) نواحی جغرافیایی هستند که

برای مقاصد نگهداری و تأمین خدمات تلفن تعریف شده‌اند. مثلاً مرکز تلفن دیترویت در داخل کد ناحیه‌ای ۳۱۳ یک ناحیهٔ جغرافیایی است که به طور تقریبی شامل شهر دیترویت و چند ناحیهٔ اطراف آن است. مرکز تلفن دیترویت به بیش از ۱۲۵ پیش شمارهٔ مختلف اختصاص داده شده است.

خانوارها و مشاغل متقاضی خدمات تلفن در داخل ناحیهٔ جغرافیایی که به عنوان مرکز تلفن دیترویت تعریف شده است باید شماره تلفنی دریافت کنند که پیش شمارهٔ آن از این مرکز تلفن خدمات‌رسانی شود. هیچ مرکز تلفن دیگری در داخل همین کد ناحیه‌ای نمی‌تواند از پیش شماره‌های تعیین شده برای مرکز تلفن دیترویت استفاده کند.

در داخل محدودهٔ مرکز تلفن دیترویت ازلحاظ پیش شماره‌ها کمی تفکیک جغرافیایی بیشتر وجود دارد. به یک خانوار که در هر نقطه‌ای در داخل محدودهٔ این مرکز تلفن واقع شده باشد شماره تلفنی با هر پیش شماره که در مرکز موجود باشد اختصاص داده می‌شود. برخی مراکز تلفن که پیش شماره‌های زیادی دارند به تلفنخانه‌هایی تقسیم می‌شوند که هر یک مسئولیت یک بخش فرعی در داخل ناحیهٔ تحت پوشش مرکز تلفن را به عهده دارد. به این ترتیب، در این مراکز تلفن، گروه‌بندی جغرافیایی اندکی، با توجه به پیش شماره‌ها در داخل مرکز تلفن وجود دارد. ولی این طرز تفکیک جغرافیایی پیش شماره‌ها در داخل یک مرکز تلفن نه یک قاعده، بلکه استثناست.

اکثریت مراکز تلفنی آمریکا دارای تعداد کمی از شماره تلفنهای تخصیص یافته هستند. در نتیجه، به بیشتر مراکز تلفن تنها یک پیش شماره اختصاص یافته است. تا این اواخر، مراکز تلفن مربوط به نواحیی بودند که محدودهٔ آنها توسط کمیسیونهای خدمات عمومی تعیین می‌شد و شرکتها می‌توانستند حقوق انحصاری برای تأمین خدمات تلفن به دست آورند. نیازهای خدماتی چنان‌اند که مساحت تحت پوشش یک مرکز تلفن محدود است. با وجود این، تراکم جمعیت در هر مرکز تلفن می‌تواند بسیار متفاوت باشد. به این ترتیب، برخی مراکز تلفن مشتریان بسیار کمی دارند و شماره تلفن برای ارائه به همهٔ مشتریان با یک پیش شمارهٔ خاص به قدر کافی موجود است. برخی دیگر از مراکز تلفن مشتریان بسیار زیادی دارند و پیش شماره‌های متعددی را در اختیار گرفته‌اند.

پس شماره‌ها در مجموعه‌های ۱۰۰۰۰ تایی گروه‌بندی شده‌اند. این شماره‌ها نوعاً توسط کارکنان سرویس تلفن براساس الگوهای واگذاری موجود، در اختیار مشتریان قرار می‌گیرند. به نظر نمی‌رسد هیچ سیستم استاندارد تأمین درخواست مشتریان تازه برای شماره تلفنهای خانگی با یک پیش شمارهٔ خاص وجود داشته باشد.

بررسیهای تجربی، الگوهایی را برای تخصیص پس شماره به پیش شماره‌ها حداقل در مواردی که سراسر سیستم مورد بررسی بوده است ابداع نموده‌اند. در مراکز تلفن دارای پیش شماره‌های متعدد و

مشتریان زیاد به نظر می‌رسد که پیش شماره‌ها و پس شماره‌ها به طور اتفاقی ارائه شده‌اند. در داخل انبوه شماره‌ها، هیچ خوشه‌بندی آشکاری از شماره تلفنهای تخصیص یافته وجود ندارد. ولی در مراکز تلفنی که دارای تنها یک پیش شماره و تعداد مشتریان محدودند، پس شماره‌ها برای کاهش هزینه‌های تخصیص، از قدیم به صورت خوشه‌هایی خاص ارائه شده‌اند. تجهیزات تلفنی الکترومکانیکی قدیمیتر این امکان را برای شرکتهای کوچکتر فراهم می‌ساختند که همهٔ شماره‌ها را در یک «انبوههٔ ۱۰۰۰ تایی» از شماره‌های متوالی اختصاص دهند به طوری که همهٔ پس شماره‌های چهار رقمی با یک رقم شروع می‌شدند. شرکتی که جمعیت کمی را تحت پوشش می‌گرفت فقط باید یک انبوههٔ ۱۰۰۰ تایی یا تعداد محدودی از آنها را خریداری می‌کرد تا تعداد مناسبی از شماره تلفنها را برای مشتریان خود در اختیار داشته باشد. به این ترتیب، شماره تلفنهای موجود در مراکز تلفنی که دارای چند پیش شمارهٔ تکی هستند، در سطح انبوههٔ ۱۰۰۰ تایی خوشه‌بندی شده‌اند. بررسیهای تجربی نشان داده‌اند که این «خوشه‌بندی» شماره‌های واگذار شده با کاهش اندازهٔ انبوهه افزایش می‌یابد. مثلاً، خوشه‌بندی انبوهه‌هایی که با دو رقم اول پس شماره یا «در انبوههٔ ۱۰۰ تایی» تعریف می‌شوند، فشرده‌تر است. بسیاری از روشهای نمونه‌گیری تلفنی که متعاقباً مورد بحث قرار می‌گیرند از این تخصیص خوشه‌بندی شدهٔ پیش شماره‌ها برای افزایش کارایی در شناسایی شماره تلفنهای واگذار شده به واحدهای مسکونی سود جسته‌اند.

۳.۱.۱۵ چارچوبهای نمونه‌گیری

چهار نوع مشکل چارچوب در نمونه‌گیریهای تلفنی وجود دارند. موارد فهرست شده در چارچوب که عناصری از جامعه محسوب نمی‌شوند به عنوان *خالیها* مورد اشاره قرار می‌گیرند، در حالی که عناصری از جامعه که فهرست‌بندی متناظری برای آنها وجود ندارد عناصر *پوشش داده نشده* نامیده می‌شوند. موارد فهرست شده در چارچوب که به عناصر چندگانه‌ای در جامعه اشاره دارند *خوشه* نامیده می‌شوند و عناصری از جامعه که دو یا چند بار در چارچوب، فهرست شده‌اند به نام *فهرست‌برداریهای تکراری* خوانده می‌شوند.

هر یک از این نارساییها می‌تواند به اریبی در برآوردهای آمارگیری یا ناکارآمدی عملیات آمارگیری منجر شود. آمارشناسان نمونه‌گیری، شیوه‌هایی را برای انتخاب کردن ابداع کرده‌اند که اریبی ناشی از این نارساییها را کاهش می‌دهند یا از میان می‌برند. این آمارشناسان در یافتن شیوه‌هایی برای انتخاب کردن به گونه‌ای که این نارساییها کاهش یابد نیز مؤثر بوده‌اند.

برای نمونه‌گیری تلفنی از سه چارچوب اصلی استفاده می‌شود: شمارهٔ تلفنها، کتابچه‌های راهنمای تلفن، و فهرستهای بازرگانی. چارچوب شمارهٔ تلفنها را می‌توان از طریق دو منبع اولیه ایجاد کرد که

هر یک فهرستی از کدهای ناحیه‌ای و ترکیبهایی از پیش شماره‌ها را با پس شماره‌هایی که به صورت تصادفی تهیه شده‌اند ترکیب می‌کند. ترکیبهای کد ناحیه‌ای و پیش شماره‌ها را می‌توان برای بررسیهای محلی از بررسی کتابچه‌های راهنمای تلفنهای محلی به دست آورد که معمولاً شامل فهرستهایی روزآمد از پیش شماره‌های مربوط به مراکز تلفنی موجود در کتابچهٔ راهنما هستند. برای آمارگیریهایی که نواحی جغرافیایی وسیعتری را پوشش می‌دهند می‌توان ترکیبهای کد ناحیه‌ای و پیش شماره‌ها را در آمریکا از یک مؤسسهٔ بازرگانی به نام شرکت تحقیقاتی بل کور[1] (BCR) تهیه کرد.

چارچوب BCR هر ماه بهنگام می‌شود و شامل همهٔ ترکیبهای کد ناحیه‌ای و پیش شماره برای آمریکا، کانادا، مکزیک و منطقهٔ کارائیب است. چارچوب BCR عملاً پوشش کاملی از خانوارهای دارای تلفن را در اختیار قرار می‌دهد ولی این عیب را دارد که تعداد قابل توجهی از موارد فهرست‌برداری شدهٔ آن خالی است. در واقع، کمتر از ۲۵٪ شماره‌های ارائه شده به واحدهای مسکونی اختصاص دارند. به همین علت، استفاده از طرحی مبتنی بر تهیهٔ یک پس شمارهٔ تصادفی همـــراه با کـد ناحیه‌ای و پیش شماره‌ای که از روی چارچوب BCR انتخاب شده باشد از نظر عملیاتی ناکارآمد است. روشهای دیگری براساس چارچوب BCR (به شرح زیر) بسط داده شده‌اند تا از خوشه‌بندی ذاتی شماره تلفنهای خانگی در انبوهه‌های ۱۰۰۰ تایی برای کاهش نسبت خالیهای حاصل از این چارچوب استفاده شود.

چارچوب BCR عیب دیگری دارد که در دو چارچوب دیگر نیز مشترک است. خانوارهایی که از بیش از یک شمارهٔ تلفن برای مقاصد مسکونی استفاده می‌کنند چندین بار در چارچوب وارد می‌شوند. روشهای نمونه‌گیری احتمالاتی مستلزم آن است که تعداد شماره تلفنهای موجود در خانوار به دست آید و برای تهیهٔ وزنی جبران کننده در برآورد کردن مورد استفاده قرار گیرد.

چارچوب دوم، یا کتابچهٔ راهنمای تلفن، در سطح گسترده‌ای به عنوان چارچوب برای بررسیهای محلی مورد استفاده قرار گرفته است. کتابچه‌های راهنما فهرستی از شماره تلفنها را در اختیار قرار می‌دهند که تهیهٔ آن برای یک محله هزینهٔ چندانی ندارد. روشهای نمونه‌گیری ساده از روی فهرست (مانند نمونه‌گیری سیستماتیک) را می‌توان برای انتخاب سریع ولی نه الزاماً سادهٔ نمونه‌ها مورد استفاده قرار داد. جفت و جور کردن چارچوبهای کتابچه‌ای برای نواحی جغرافیایی وسیعتر مشکل است، زیرا هر سال بیش از ۵۰۰۰ کتابچهٔ راهنمای تلفن در سراسر آمریکا منتشر می‌شود. متداول شدن آنها به عنوان چارچوب نمونه‌گیری به دلیل هزینه و راحتی کار و نسبت کم موارد فهرست‌برداری خالی در

[1] Bell Core Research, Inc., (BCR)

مقایسه با چارچوب شمارهٔ تلفن است: تقریباً ۱۰ تا ۱۵ درصد از موارد فهرست شده در راهنمای تلفنهای خانگی آمریکا، دیگر خانگی نیستند.

از سوی دیگر، چارچوبهای کتابچه‌ای دچار عدم پوشش جمعیت خانوارهای دارای تلفن در سطحی قابل ملاحظه به علت شماره‌های فهرست نشده و تغییرات در وضعیت تلفنی خانوارها هستند. درصد خانوارهای تلفن‌داری که در کتابچه‌های راهنمای تلفن آمریکا دیده نمی‌شوند از ۳۵ درصد بیشتر است و از نسبتهای کم (۱۰ درصد) در حومهٔ شهرها و نواحی روستایی تا نسبتهای بیشتر از ۶۰ درصد در برخی مکانهای شهری متغیر است. طراحان آمارگیریها نمی‌توانند این نسبتهای زیاد فهرست‌نشده‌ها یا موارد فهرست شدهٔ قدیمی را بپذیرند و استفاده از روشهای شماره‌گیری ارقام تصادفی یا سایر روشهایی را که پوشش بیشتری در اختیار قرار می‌دهند برمی‌گزینند.

علاوه بر این، سطوح مداخل تکراری در کتابچه‌های تلفن بیشتر از چارچوب BCR است، زیرا مشترکین می‌توانند مداخل اضافی خریداری کنند. مثلاً، زن و شوهری که با دو نام خانوادگی مختلف در یک آدرس به سر می‌برند می‌توانند با پرداخت دستمزدی ناچیز تحت نام خانوادگی هر دو نفر در کتابچهٔ راهنمای تلفن درج شوند. تکرار موجود در فهرست باعث افزایش احتمال انتخاب شدن آن خانوار تلفن‌دار می‌شود که از نظر نمونه‌گیری احتمالاتی باید از طریق تعیین وزنی برای خانوار جبران شود.

چارچوب نوع سوم توسط شرکتهای تجاری در پرونده‌های الکترونیکی براساس کتابچه‌های راهنمای اطلاعات تلفن جمع‌آوری شده در سراسر کشور، تهیه می‌شود. درایه‌های مندرج در کتابچهٔ راهنما (نام، آدرس، و شمارهٔ تلفن) یا وارد رایانه می‌شوند یا در صورت وجود چارچوب الکترونیکی به آن افزوده می‌شوند. فهرستهای مبتنی بر راهنماهای تجاری با سایر فهرستها تکمیل می‌شوند که بزرگترین آنها فهرستهای ثبتی اتومبیلهاست که از تقریباً ۳۰ ایالتی که این قبیل داده‌ها را برای عموم منتشر می‌کنند به دست می‌آیند. پروندهٔ ترکیبی پردازش می‌شود تا برای مقاصد پستی به هر درایه یک کد پستی تخصیص داده شود. چندین شرکت از وجود شماره تلفنها در این قبیل پرونده‌ها برای ایجاد راهنماهای تلفن در سطح کشور و انتخاب و فروش نمونه‌هایی از شماره تلفنها از روی آنها بهره برده‌اند. چارچوبهای تجاری درگیر نسبت کمی از اقلام خالی (تقریباً ۱۰ تا ۱۵ درصد، یعنی برابر با کتابچه‌های راهنما) و عدم توانایی در پوشش شماره‌های فهرست نشده و نیز وجود اقلام تکراری‌اند.

۲.۱۵ طرحهای نمونه‌ای تلفنی

آمارشناسان نمونه‌گیری روشهایی را برای نمونه‌گیری ابداع کرده‌اند که سعی در بهره‌گیری از توانمندیهای هر یک از این سه چارچوب دارند. روشهای مزبور غالباً به صورت یکی از این سه نوع،

رده‌بندی می‌شوند: روشهای نمونه‌گیری با چارچوب فهرست ساده که برای کتابچه‌های راهنمای تلفن مناسب‌اند، روشهای شماره‌گیری ارقام تصادفی مبتنی بر چارچوب شماره تلفن، و روشهایی با کمک فهرست مبتنی بر کتابچه‌های راهنمای تلفن یا فهرستهای تجاری که نمونه‌هایی تولید می‌کنند که شامل شماره تلفنهای فهرست نشده نیز هستند. در اینجا دربارهٔ روشهای نمونه‌گیری مورد استفاده برای کتابچه‌های راهنما بحث نمی‌کنیم. این روشها اصولاً کاربرد روشهایی هستند که قبلاً در فصلهای پیشین کتاب حاضر مورد بحث قرار گرفته‌اند. در عوض به بررسی شماره‌گیری ارقام تصادفی و روشهای با کمک فهرست می‌پردازیم، زیرا کاربردهایی تازه از روشهایی محسوب می‌شوند که در جاهای دیگر این کتاب توصیف شده‌اند.

۱.۲.۱۵ طرحهای نمونه‌ای با استفاده از چارچوب BCR

همان‌گونه که قبلاً توصیف شد، ساده‌ترین و مستقیمترین رهیافت برای استفاده از چارچوب BCR، انتخاب تصادفی شمارهٔ تلفنها از روی چارچوب، تلفن کردن به شماره‌های انتخاب شده، و اجرای مصاحبه در صورت رسیدن به یک خانوار است. شماره‌ها انتخاب و گرفته می‌شوند تا به اندازهٔ نمونهٔ مطلوب خانوارهای موجود در آن حوزه، مثلاً n، برسیم. به طوری که پیش از این اشاره شد تنها در حدود ۲۵٪ شماره تلفنهای نمونه به خانوارها تخصیص یافته‌اند. بنابراین، شماره تلفنهای مورد نیاز، مثلاً n'، بسیار بیشتر از n خواهد بود. تعداد مورد انتظار شماره تلفنهای مورد نیاز n/p است که در آن p نسبت شماره تلفنهایی است که به واحدهای مسکونی خانوار اختصاص یافته‌اند. پس برای جبران موارد فهرست‌برداری شده که واجد شرایط نیستند باید نمونهٔ شمارهٔ تلفنهای انتخاب شده از چارچوب BCR تقریباً چهار برابر نمونهٔ مطلوب متشکل از n خانوار تلفن‌دار باشد.

به طور کلی، تعیین وضعیت یک شماره تلفن مستلزم صرف هزینه است به خصوص در مورد شماره تلفنهایی که در اختیار خانوارها نیستند. غالباً باید برای تعیین وضعیت، یک شماره تلفن چندین بار شماره‌گیری شود. چون برای هر نوع نتیجهٔ حاصل از شماره‌گیری باید شیوه‌ها مشخص شوند، استفاده از BCR (یا هر فهرستی که دارای نسبت زیادی از اقلام خالی است) هزینه‌های اداری و عملیاتی آمارگیری تلفنی را به شدت افزایش خواهد داد. (ن. ک. لپکوفسکی [۱۰]).

خلاصه کردن هزینه‌های عملیاتی استفاده از هر چارچوب، برای مقایسهٔ طرحهای گوناگون، مفید خواهد بود. هزینه‌های استفاده از BCR را می‌توان در یک مدل هزینه‌ای ساده به صورت زیر خلاصه کرد. فرض کنید C_0 هزینهٔ تعیین وضعیت شماره‌ای باشد که به خانواری در محدودهٔ مورد نظر (خانواری که تلفن دارد) واگذار نشده باشد؛ C_1 نشان‌دهندهٔ هزینهٔ تعیین وضعیت شماره‌ای در اختیار

خانوار در محدودهٔ مورد نظر باشد؛ و $C_۲$ نشان‌دهندهٔ هزینهٔ اجرای مصاحبه برای آمارگیری باشد. پس کل هزینهٔ آمارگیری براساس چارچوب BCR از این فرمول به دست می‌آید

$$C = n(C_۱ + C_۲) + (n' - n)C_۰$$

با منظور کردن نسبت شماره تلفنهایی که در محدوده قرار دارند، کل هزینهٔ مورد انتظار برای یک آمارگیری با شماره‌گیری ارقام به صورت تصادفی ساده از فرمول زیر به دست می‌آید

$$E(C) = n\big((C_۱ + C_۲) + C_۰(۱ - p)/p\big)$$

با در نظر گرفتن p در همسایگی ۰/۲۵، مؤلفهٔ هزینهٔ مورد انتظار ناشی از تلفن زدنهای بی‌حاصل [یعنی $nC_۰(۱-p)/p$] ، نسبت قابل ملاحظه‌ای از کل هزینهٔ مورد انتظار C را تشکیل خواهد داد. انگیزهٔ همهٔ طرحهای تلفنی که در بخشهای زیر توصیف شده‌اند اشتیاق به کاهش نسبت هزینه‌های ناشی از تلفن زدنهای بی‌حاصل بوده است.

۱.۱.۲.۱۵ *طرح میتوفسکی ـ واکسبرگ*

طرح رقم تصادفی دومرحله‌ای که میتوفسکی [۱۱] پیشنهاد کرد و توسط واکسبرگ [۱۸] شرح و بسط کاملتری داده شد در چنان سطح گسترده‌ای در آمارگیریهای تلفنی به کار برده شده است که تقریباً با آمارگیریهای تلفنی با شماره‌گیری ارقام تصادفی مترادف شده است. روش مذکور از این واقعیت بهره می‌برد که شمارهٔ تلفنهای واگذار شده به خانوارهای ساکن، گرایش به خوشه‌بندی در انبوهه‌هایی از شماره تلفنهای متوالی دارند. در حالی که در مجموع فقط ۲۵ درصد از شماره‌های موجود در چارچوب BCR در اختیار خانوارها هستند، در میان انبوهه‌های ۱۰۰ تایی از شماره‌های متوالی که حداقل یک شماره تلفن آن مربوط به یک خانوار باشد ۶۰ درصد شماره تلفنها خانگی هستند. واضح است که اگر بتوان انبوهه‌های صدتایی با یک یا چند شماره تلفن خانگی را شناسایی کرد و نیز اگر نمونه‌گیری به همین انبوهه‌ها محدود باشد، در آن صورت، نسبت تلفن زدنهای بی‌حاصل به طور قابل ملاحظه‌ای کاهش خواهد یافت و هزینه‌های عملیاتی در مجموع پایین خواهد آمد.

فن میتوفسکی ـ واکسبرگ با گروه‌بندی شماره‌های موجود در چارچوب BCR در مجموعه‌های ۱۰۰ تایی از شماره‌های متوالی با استفاده از کد ناحیه‌ای، پیش شمارهٔ سه رقمی و دو رقم اول پس شماره برای تعیین *انبوهه* (یا *انبوههٔ* ۱۰۰ تایی) آغاز می‌شود. در مرحلهٔ اول، انبوهه‌های ۱۰۰ تایی به صورت تصادفی با جایگذاری انتخاب می‌شوند. یک شماره تلفن در داخل انبوهه به طور تصادفی انتخاب و شماره‌گیری می‌شود. اگر معلوم شود که شمارهٔ انتخاب شده خانگی است انبوههٔ مزبور برای نمونه‌گیری مرحلهٔ دوم حفظ می‌شود. این فرایند ادامه می‌یابد تا نمونهٔ تعیین شدهٔ متشکل از m انبوههٔ

۱۰۰ تایی به دست آید. شمارهٔ تلفنها در داخل هر انبوههٔ ۱۰۰ تایی حفظ شده، به صورت تصادفی بدون جایگذاری انتخاب می‌شوند تا مجموع k شماره تلفن خانگی (شامل اولین شماره‌ای که برای حفظ انبوهه به کار رفته است) شناسایی شوند. وضعیت خانگی بودن یک شماره تلفن فقط با گرفتن آن شماره تعیین می‌شود.

به این ترتیب، در فن میتوفسکی ـ واکسبرگ از طرح دو مرحله‌ای استفاده می‌شود که در آن، انبوههای ۱۰۰ تایی با احتمال متناسب با تعداد شماره تلفنهای خانگی در مرحلهٔ اول و نمونه‌ای با اندازهٔ ثابت در مرحلهٔ دوم از خانوارهای ساکن این اماکن انتخاب می‌شوند. به این ترتیب، نمونهٔ $n = mk$ خانوار ساکن با احتمال برابر (ولی نامعلوم) انتخاب می‌شود. کارایی فن میتوفسکی ـ واکسبرگ از این واقعیت ناشی می‌شود که شماره تلفنهای واجد شرایط در نسبت تقریباً کوچکی از انبوههای ۱۰۰ تایی متمرکز شده‌اند.

برای تعیین هزینهٔ استفاده از این طرح دیگر نمونه‌ای برای چارچوب BCR، فرض می‌کنیم t نسبت انبوههای ۱۰۰ تایی بدون شماره تلفنهای واجد شرایط باشد. در آن صورت، کل تعداد مورد انتظار شمارهٔ تلفنها در سراسر مراحل اول و دوم به صورت $n(۱-t(k-۱)/k)/p$ است و کل هزینهٔ عملیاتی مورد انتظار به صورت زیر خواهد بود:

$$E(C) = n\big((C_۱ + C_۲) + C_۰(۱ - p - t(k-۱)/k)/p\big)$$

واضح است که هم تعداد تلفن زدنهای مورد انتظار و هم هزینهٔ مورد انتظار با افزایش k کاهش می‌یابد. t در سطح کل کشور در همسایگی ۰/۶۵ است، بنابراین حتی مقادیر نه چندان زیاد k می‌تواند به صرفه‌جوییهای قابل ملاحظه در هزینه‌ها بینجامد.

با این که فن میتوفسکی ـ واکسبرگ روشی ظریف برای بهبود کارایی آمارگیری تلفنی است، در عمل مشکلاتی وجود دارند. واضحترین آنها این است که برخی از انبوههای صدتایی ممکن است کمتر از مقدار مورد نیاز k خانوار واجد شرایط داشته باشند که در این حالت به اجبار باید به همهٔ شماره‌های موجود در انبوهه تلفن شود. در این صورت وزن دادن برای جبران آن ضروری خواهد بود. مشکل دوم آن است که همیشه این امکان وجود ندارد که وضعیت واجد شرایط بودن یک شمارهٔ انتخابی با دقت تعیین شود. در مرحلهٔ اول، چنین حالتی می‌تواند به حذف یا گنجاندن نادرست انبوههایی ۱۰۰ تایی منجر شود. در مرحلهٔ دوم ممکن است در پایان دورهٔ آمارگیری هنوز تکلیف بعضی شماره‌ها روشن نشده باشد به طوری که کمتر از k خانوار واجد شرایط در انبوهه شناسایی شده باشند. بالاخره این که مشکل پیچیده‌تر، موضوع همبستگی درون انبوهه‌ای است که با جزئیات بیشتر در بخش ۳.۱۵، هنگام بررسی برآورد کردن مورد بحث قرار خواهد گرفت.

۲.۱.۲.۱۵ *طرح پوتوف*

طرح پیشنهادی پوتوف [۱۲] شبیه طرح میتوفسکی ـ واکسبرگ است با این تفاوت که واجد شرایط بودنش به ردهٔ وسیعتری از شماره تلفنها گسترش یافته است. پوتوف پیشنهاد کرده است که به جای تعیین واجد شرایط بودن یک انبوههٔ ۱۰۰ تایی بر مبنای این که آیا شمارهٔ تلفن به واحد مسکونی اختصاص دارد یا نه، بهتر است واجد شرایط بودن وضعیت آن شماره تلفن، پس از شماره‌گیری تعریف شود. او شماره تلفنهایی را که تحت برآمدهای گوناگون واجد شرایط بوده‌اند *شماره‌های خوش‌یمن* نامیده است. شماره های خوش‌یمن نوعاً نه تنها شماره تلفنهای خانگی خانوارها را در برمی‌گیرند بلکه شماره‌هایی را هم شامل می‌شوند که بدون جواب زنگ می‌خورند و وضعیت خانگی بودن آنها پس از شماره‌گیری نیز نامعلوم است. این تعریف گسترده‌تر، از مقدار غربالگری مورد نیاز برای مرحلهٔ اول و مقدار جایگذاری مورد نیاز در مرحلهٔ دوم می‌کاهد. پوتوف [۱۲]، [۱۳] همچنین پیشنهاد کرده است که در مرحلهٔ اول در هر انبوههٔ ۱۰۰ تایی، $c \geq ۲$ شماره تلفن انتخاب شود. در این صورت، نمونه‌گیری در مرحلهٔ دوم به تعداد شماره‌های خوش‌یمن مشاهده شده در مرحلهٔ اول بستگی خواهد داشت. از جزئیات روش پوتوف در این کتاب به تفصیل بحث نخواهد شد.

طرح نمونه‌گیری پوتوف نمونه‌ای از شماره‌های واجد شرایط با احتمال برابر به دست می‌دهد. جایگذاری فقط برای تعداد کمی از پیش شماره‌های ناحیه‌ای مورد نیاز است و ابهامهای مربوط به وضعیت شماره تلفنهای شماره‌گیری شده در مرحلهٔ اول را کاهش می‌دهد. همچنین، وقتی c ، تعداد شماره تلفنهای انتخاب شده به ازای هر انبوههٔ ۱۰۰ تایی در مرحلهٔ اول زیاد می‌شود، شانسهای به دست آوردن انبوهه‌ای که در مرحلهٔ دوم خالی شود کاهش پیدا می‌کند. اجرای طرح پوتوف برای تعیین اندازهٔ نمونه‌ای مناسب، مستلزم شناختی دربارهٔ نسبت شماره‌های خوش‌یمنی است که واقعاً واجد شرایط‌اند. ساختار اجرایی این روش، پیچیده‌تر و نیازهای آموزشی آن برای کارکنان عملیاتی به نسبت طرح میتوفسکی ـ واکسبرگ بیشتر است.

۲.۲.۱۵ طرحهای نمونه‌ای که از شماره تلفنهای خانگی منتشر شده استفاده می‌کنند

همان طور که در بالا توصیف شد ۸۵ تا ۹۰ درصد شماره تلفنهای موجود در فهرستهای تجاری به خانوارهای ساکن در اماکن مسکونی مربوط می‌شوند. انتخاب سیستماتیک سر راست شماره تلفنها (انتخاب تصادفی فقط در صورتی امکان‌پذیر است که فهرست به ترتیب شماره‌گذاری شده باشد) از روی چنین فهرستی به مراتب کارآمدتر از طرحهایی خواهد بود که برای نمونه‌گیری از روی فهرست BCR مورد استفاده قرار می‌گیرند. متأسفانه فهرست نوعاً مبتنی بر راهنمای تلفن، به علت منسوخ بودن شماره‌های فهرست شده و فهرست نشده فقط شامل تقریباً ۷۰ درصد خانوارهای دارای تلفن خانگی

است. مقایسهٔ شماره تلفنهای منتشر شده و منتشر نشدهٔ خانوارهای تلفن‌دار نشان می‌دهد که اگر خانوارهایی که شمارهٔ تلفنهای آنها منتشر نشده است از چارچوب نمونه‌گیری حذف شوند اریبی قابل ملاحظه‌ای نتیجه خواهد شد (برانر و برانر [۲]). طرحهایی که در این بخش مورد بحث قرار گرفتند می‌کوشند تا از کارایی ذاتی نمونه‌گیری بر مبنای راهنمای تلفن بهره‌مند شوند و در عین حال پوشش طرح را گسترش دهند تا کل جمعیت دارای تلفن خانگی را شامل شوند.

۱.۲.۲.۱۵ شماره‌گیری رقم به اضافهٔ یک

شماره‌گیری رقم به اضافهٔ یک، شیوه‌ای با استفاده از راهنمای تلفن است که در آن نمونه‌ای از شماره تلفنها از روی دفترچهٔ راهنما انتخاب و به پس شمارهٔ تلفن عدد صحیح یک اضافه می‌شود. مثلاً به پس شمارهٔ هر شماره تلفنی که از راهنمای تلفن انتخاب شود در این شماره‌گیری به پس شماره «یک» اضافه می‌شود. به این ترتیب اگر شمارهٔ انتخابی ۷۳٤-۹۳٦-۰۰۲۱ باشد جایگزین آن، شمارهٔ ۷۳٤-۹۳٦-۰۰۲۲ خواهد بود. نمونهٔ شماره تلفنها که به این ترتیب به دست می‌آید به طور کلی شامل شماره‌های فهرست شده و فهرست نشده، هر دو، خواهد بود. علاوه بر این، در مقایسه با طرح شماره‌گیری رقم تصادفی ساده نسبت بیشتری از شماره‌های نتیجه‌بخش به دست خواهد داد.

متأسفانه این شیوه دارای چند مشکل نظری و عملیاتی است. بررسیهای تجربی مزبور نشان داده‌اند که انتخاب خانوارهای دارای تلفن خانگی در روش شماره‌گیری به اضافهٔ یک دارای اریبی قابل توجهی است. به علاوه احتمالهای انتخاب شدن شماره‌های موجود در جامعهٔ هدف، نابرابر و نامعلوم‌اند. در واقع، احتمال انتخاب شدن برخی شماره‌های فهرست نشده ممکن است صفر باشد مگر این که شماره‌های فهرست نشده به صورت یکنواخت با شماره‌های فهرست شده آمیخته باشند. تعمیمهای طرح مذکور که در آن d رقم آخر (دو یا چند رقم آخر) با d رقم دیگر که به طور تصادفی به دست آمده‌اند جایگزین می‌شوند پیشنهاد شده است.

طرحی که در ارتباط نزدیک با طرح بالاست براساس بازه‌های نیمه باز شماره تلفنها توسط فرانکل و فرانکل [۵] پیشنهاد شده است. در راهنماهایی که به ترتیب شماره تنظیم شده‌اند، خوشه، شامل یک شماره تلفن فهرست شده همراه با همهٔ شماره‌های دیگر تا شمارهٔ فهرست شدهٔ بعدی ولی بدون منظور نمودن آن تعریف شده است. نمونه‌ای از خوشه‌ها از روی راهنمای تلفن، صرفاً با انتخاب یک نمونهٔ تصادفی ساده از شمارهٔ تلفنهای موجود در راهنما، گرفته می‌شود. سپس همهٔ شماره‌های موجود در داخل خوشهٔ انتخابی شماره‌گیری می‌شوند. این روش به احتمالهای معلوم غیرصفر، برای انتخاب شدن همهٔ خانوارهای تلفنی تحقق می‌بخشد. ولی تغییرپذیری بالقوه زیاد اندازهٔ خوشه می‌تواند مشکلات عملیاتی دشوار را مطرح سازد. اندازهٔ نمونه می‌تواند نسبت به یک سطح هدف به شدت تغییر کند. به

علاوه، این روش در معرض مشکلات برآورد کردن قرار دارد چرا که اندازهٔ نمونه و خوشه هر دو متغیرهای تصادفی‌اند.

۲.۲.۲.۱۵ *نمونه‌گیری دومرحله‌ای*

یک طرح نمونه‌گیری دومرحله‌ای با استفاده از فهرست راهنمای تلفن توسط سودمن [۱۵] پیشنهاد شده است. در این شیوه که در اصل توسط استاک [۱٤] پیشنهاد شده بود از انبوهه‌های ۱۰۰۰ تایی شماره تلفنها (که با شش رقم اول شماره تلفن ۱۰ رقمی مشخص می‌شوند) به عنوان واحد نمونه‌گیری مرحلهٔ اول استفاده می‌شود. انتخاب انبوهه‌های ۱۰۰۰ تایی شبیه انتخاب مرحلهٔ اول در روش میتوفسکی ـ واکسبرگ است با این تفاوت که برای انتخاب نمونهٔ مرحلهٔ اول از راهنمای شماره تلفنهای فهرست شده استفاده می‌شود. به این ترتیب، احتمال انتخاب شدن در مرحلهٔ اول متناسب با تعداد شماره تلفنهای فهرست شده در انبوههٔ ۱۰۰۰ تایی است. در مرحلهٔ دوم، انتخاب شماره تلفنها ادامه می‌یابد تا تعداد ثابتی از شماره تلفنهای فهرست شده که از پیش تعیین شده است انتخاب شود. سعی بر آن است که خانوارهای دارای شماره تلفنهای فهرست شده و فهرست نشده هر دو مورد مصاحبه قرار بگیرند. باید توجه داشت که احتمال انتخاب شدن شماره تلفنهای فهرست نشده در انبوهه‌های ۱۰۰۰ تایی که فاقد هر نوع شمارهٔ فهرست شده‌اند صفر است ولی نسبت خانوارهای تلفن‌دار در این قبیل انبوهه‌های ۱۰۰۰ تایی کوچک است و این در بیشتر موارد مشکلی جدی محسوب نمی‌شود. نگران کننده‌تر، این واقعیت است که تعیین وضعیت فهرست شدن، به خود ـ گزارش پاسخگو بستگی دارد که می‌تواند خطا باشد. یک «راهنمای شماره تلفنهای به ترتیب معکوس» که در بسیاری از نواحی کلانشهری موجود است می تواند این منشأ خطا را از میان بردارد.

شیوهٔ سودمن برخلاف روش میتوفسکی ـ واکسبرگ، خوشه‌هایی با اندازهٔ نابرابر برای خانوارهای تلفن‌دار تولید خواهد کرد (هر چند خوشه‌هایی که برای خانوارهای تلفن‌دار فهرست شده تولید می‌کند دارای اندازهٔ برابرند). معمولاً تغییرپذیری اندازهٔ خوشه چندان زیاد نیست. همچنین، وجود خوشه‌های خالی محتمل است ولی با انبوهه‌های ۱۰۰۰ تایی به جای ۱۰۰ تایی در روش میتوفسکی ـ واکسبرگ این امر چندان جای نگرانی ندارد.

۳.۲.۱۵ طرحهایی که از چارچوب BCR و شماره تلفنهای منتشر شده استفاده می‌کنند

باید توجه داشت که طرحهای مورد بحث در بخش ۱.۲.۱۵ تنها به چارچوب BCR نیاز دارند در حالی که طرحهای مورد بحث در بخش ۲.۲.۱۵ تنها به فهرستی منتشر شده از شماره تلفنهای خانگی نیازمندند. طرحهایی که در بخش حاضر مورد بحث قرار می‌گیرند به هر دو نیاز دارند. ایدهٔ اصلی

زیربنایی این طرحها آن است که خواص پوششی مطلوب چارچوب BCR، با کارایی نسبتاً زیاد نمونه‌گیری از چارچوب متشکل از شماره تلفنهای فهرست شده، پیوند داده شود.

۱.۳.۲.۱۵ طرحهایی با چارچوب دوگان

یک نمونهٔ شماره‌گیری رقم تصادفی متشکل از n_B خانوار تلفنی از روی چارچوب BCR انتخاب می‌شود و همزمان نمونه‌ای متشکل از n_D خانوار تلفنی از چارچوب فهرست راهنمای تلفن انتخاب می‌شود. با فرض این که n'_B و n'_D به ترتیب تعداد شماره تلفنهای مورد نیاز از هر چارچوب برای به دست آوردن اندازهٔ نمونهٔ مطلوب باشند، هزینهٔ طرح با چارچوب دوگان از فرمول زیر به دست می‌آید

$$C = (n_B + n_D)(C_1 + C_2) + C_0(n'_B + n'_D - n_B - n_D)$$

هزینهٔ مورد انتظار برای آمارگیری با چارچوب دوگان به صورت زیر است

$$E(C) = n(C_1 + C_2 + C_0(\lambda(1-p_B)/p_B + (1-\lambda)(1-p_D)/p_D))$$

که در آن $n = n_B + n_D$ کل اندازهٔ نمونه، $\lambda = n_B/n$ نسبت کل نمونهٔ تعیین شده برای چارچوب BCR، p_B نسبت شمارهٔ تلفنها در چارچوب BCR که به خانوارهای دارای تلفن خانگی واگذار شده‌اند، و p_D نسبت شمارهٔ تلفنها در چارچوب راهنمای تلفن که به خانوارهای دارای تلفن خانگی واگذار شده‌اند. چون p_B در همسایگی ۰/۲۵ و p_D معمولاً در حدود ۰/۸۵ است، هزینهٔ مورد انتظار (برای کل اندازهٔ نمونه‌ای ثابت n) با کاهش λ کم خواهد شد.

برای ترکیب داده‌های حاصل از دو چارچوب به منظور برآورد کردن، راههای ممکن متعددی وجود دارند. به طور کلی، برآوردگرهای چارچوب دوگان پیچیده‌تر از برآوردگرهای مربوط به طرحهای مورد بحث پیشین‌اند. گروز و لپکوفسکی [٦] بحث مفصلی دربارهٔ مسئلهٔ برآورد کردن از چارچوب دوگان و مسئلهٔ انتساب نمونه به دو چارچوب برای تحقق کمترین هزینه به ازای یک واریانس توصیف شده ارائه کرده‌اند.

برای اجرای روش چارچوب دوگان، وضعیت راهنمای تلفنی (یعنی فهرست شده یا فهرست نشده) هر خانواری که تلفن خانگی دارد باید روی نمونهٔ BCR معلوم باشد. برای اجتناب از به کار گرفتن گزارشهای بالقوه غیر قابل اعتماد پاسخگویان دربارهٔ وضعیت در فهرست بودن آنها، می‌توان شماره تلفنهای انتخاب شده از روی چارچوب BCR را با فهرست راهنمای تلفن در هنگام انتخاب نمونه جور کرد. اگر چارچوب راهنمای تلفن، آدرس شماره‌های فهرست شده را داشته باشد این امکان

وجود خواهد داشت که پیشاپیش نامه‌هایی به منظور بهبود نرخ پاسخگویی ارسال شود. به طور کلی، طرح چارچوب دوگان مستلزم عملیات اجرایی نسبتاً پیچیده‌تری است. نیاز به جور کردن نمونهٔ BCR با چارچوب راهنمای تلفن برای تعیین شماره‌های واجد شرایط انتخاب از هر دو چارچوب، ارسال نامه‌های پیشاپیش، و استفاده از برآوردگر پیچیده‌تر باعث افزایش هزینه‌ها می‌شوند. مزایای بیشتر بودن نرخ پاسخگویی به مراتب بیشتر از جبران کردن هزینه‌های ارسال نامه‌های پیشاپیش است.

۲.۳.۲.۱۵ *طبقه‌بندی مبتنی بر راهنمای تلفن*

از فهرست راهنمای تلفن می‌توان به منظور طبقه‌بندی چارچوب BCR برای بهبود کارایی نمونه‌گیری از روی چارچوب BCR استفاده کرد. در یک کاربرد نوعی، از فهرست راهنمای تلفن برای شناسایی انبوهه‌های ۱۰۰ تایی در چارچوب BCR که دارای یک یا چند شماره تلفن فهرست شده در راهنمای تلفن هستند استفاده می‌شود. سپس چارچوب BCR به دو طبقه افراز می‌شود: یک طبقه شامل همهٔ شماره تلفنها در انبوهه‌های ۱۰۰ تایی است که یک یا چند شماره تلفن فهرست شده دارند و طبقهٔ دیگر شامل همهٔ شماره تلفنهای دیگر است. به طبقهٔ اول به عنوان طبقهٔ با چگالی بالا اشاره می‌شود در حالی که طبقهٔ دوم را طبقهٔ باقیمانده می‌نامند. سپس نمونه‌های شماره‌گیری با رقم تصادفی ساده از هر طبقه انتخاب می‌شود به طوری که از طبقهٔ با چگالی بالا، نمونه‌ای به مراتب بزرگتر انتخاب می‌شود زیرا تقریباً همهٔ شماره تلفنهای خانگی در این طبقه قرار دارند.

در طرح طبقه‌بندی شده سعی بر این است که از همان جنبهٔ خوشه‌بندی چارچوب شماره تلفنها در طرح میتوفسکی ـ واکسبرگ استفاده شود. شماره تلفنهای خانوارهای دارای تلفن خانگی، اعم از فهرست شده و فهرست نشده، گرایش به خوشه‌بندی در انبوهه‌های ۱۰۰ تایی دارند. اگر انبوهه‌های محتوی این قبیل شماره تلفنها را بتوان به میزان بیشتری شناسایی و نمونه‌گیری کرد، آنگاه کارایی نمونه‌گیری می‌تواند افزایش بسیار زیادی پیدا کند. کاسادی و لپکوفسکی [۳] پی برده‌اند که در سطح کل کشور، نسبت شماره تلفنهای موجود در چارچوب BCR که به طبقهٔ با چگالی بالا اختصاص می‌یابند تقریباً ۰/۳۸ خواهد بود و بیش از ۹۷٪ شماره‌های واگذار شده به خانوارهای دارای تلفن خانگی را شامل خواهد شد. آنها نشان داده‌اند که نسبت شماره تلفنهای تخصیص یافته به خانوارها در طبقهٔ با چگالی بالا تقریباً ۰/۵۵ است در حالی که نسبت شماره تلفنهای تخصیص یافته به خانوارها در طبقهٔ باقیمانده فقط حدود ۰/۰۲ است.

با فرض این که یک نمونهٔ شماره‌گیری ارقام تصادفی متشکل از n_1 خانوار تلفن‌دار از طبقهٔ با چگالی بالا و یک نمونهٔ متشکل از n_2 خانوار تلفن‌دار از طبقهٔ باقیمانده انتخاب شده باشد، هزینهٔ طرح طبقه‌بندی از فرمول زیر به دست می‌آید

$$C = (n_۱ + n_۲)(C_۱ + C_۲) + C_۰(n'_۱ + n'_۲ - n_۱ - n_۲)$$

که در آن $n'_۱$ و $n'_۲$ تعداد شماره تلفنهای مربوط به هر طبقه برای دستیابی به اندازه‌های نمونه‌ای مطلوب است. هزینهٔ مورد انتظار نمونهٔ طبقه‌بندی شده چنین است

$$E(C) = n(C_۱ + C_۲ + C_۰(\gamma(۱-p_۱)/p_۱ + (۱-\gamma)(۱-p_۲)/p_۲))$$

که در آن n، کل اندازهٔ نمونه، $\gamma = n_۱/n$، نسبت کل نمونهٔ تخصیص یافته به طبقهٔ با چگالی بالا، $p_۱$، نسبت شماره تلفنهای واگذار شده به خانوارهای دارای تلفن خانگی در طبقهٔ با چگالی بالا، و $p_۲$ نسبت شماره تلفنهای واگذار شده به خانوارهای دارای تلفن خانگی در طبقهٔ باقیمانده است. چون $p_۱$ تقریباً ۰/۵۵ و $p_۲$ معمولاً در حدود ۰/۰۲ است، هزینهٔ مورد انتظار (برای n، کل اندازهٔ ثابت نمونه) با افزایش γ کاهش خواهد یافت. انتساب نمونه به طبقات برای به حداقل رساندن هزینهٔ مربوط به یک واریانس ثابت (یا به حداقل رساندن واریانس برای یک هزینهٔ ثابت) توسط کاسادی و لپکوفسکی [۳] به تفصیل مورد بحث قرار گرفته است.

احتمالهای انتخاب شدن در داخل طبقات، معلوم، مثبت، و برابرند، بنابراین، برآورد کردن مجموعهای جامعه‌ای سرراست است. برآورد کردن میانگین جامعه‌ای در سطح طبقه نیز سرراست است، ولی برآورد کردن میانگین کل جامعه مستلزم آن است که کل جامعهٔ دارای تلفن خانگی، برای هر طبقه برآورد شود و یک برآوردگر نسبتی برای برآورد میانگین جامعه‌ای مورد استفاده قرار گیرد. بحث مفصلتر دربارهٔ برآورد میانگینها و واریانسها در بخش ۳.۱۵ ارائه خواهد شد.

تحت مدل هزینه‌ای نسبتاً ساده‌ای که در بالا ارائه شد، طرح مزبور به طرزی مناسب با طرح میتوفسکی ـ واکسبرگ مطابقت می‌کند. در کاربست، ثابت شده است که طبقه‌بندی مبتنی بر راهنمای تلفن در نمونه‌گیری شماره‌گیری رقم تصادفی ساده در داخل طبقات از جهت اجرا و مدیریت دارای برتری است. این طرح با دو هزینه همراه است که در مدل ساده وجود ندارند: هزینهٔ خود فهرست تجاری و هزینهٔ طبقه‌بندی چارچوب BCR. هزینهٔ فهرست تجاری برحسب فروشنده فرق می‌کند ولی برای عملیات آمارگیری بزرگ مداوم، مؤلفهٔ هزینه‌ای نسبتاً ناچیزی محسوب می‌شود. این هزینه‌ها هر دو ثابت‌اند و می‌توانند در چندین بررسی تقسیط شوند تا تأثیر آنها بر هر مطالعهٔ تکی تا حد بسیار زیادی کاهش یابد.

۳.۳.۲.۱۵ برش دادن براساس راهنمای تلفن

یک مورد خاص در روش طبقه‌بندی شده که در آخرین بخش مورد بحث قرار گرفت هیچ نمونه‌ای را به طبقهٔ باقیمانده منتسب نمی‌کند؛ یعنی این‌که چارچوب BCR با برداشتن طبقهٔ باقیمانده بریده می‌شود. نرخ دستیابی به هدف که به شدت افزایش یافته است همراه با سایر مزایای طرح طبقه‌بندی مبتنی بر راهنمای تلفن، جذابیت بسیار زیادی به این رهیافت بخشیده‌اند. از معایب آشکار آن این است که وقتی چارچوب بریده شده باشد همهٔ جامعهٔ هدف پوشش داده نمی‌شود. در مثالی که در بالا ارائه شد، تقریباً ۲٪ خانوارهای تلفن‌دار پوشش داده نخواهند شد. ولی، شواهد تجربی (بریک و همکاران [۱]، کونور و هیرینگا [٤]) نشان می‌دهند که جامعهٔ خارج از حوزه از نظر بسیاری از متغیرها بسیار شبیه به جامعهٔ هدف است و اریبی ناشی از برش دادن بسیار اندک خواهد بود. همان طور که قبلاً اشاره شد، تقریباً ۵٪ جامعهٔ خانوارهای آمریکا جزو جامعهٔ تلفن‌دار نیستند و هرگونه اریبی اضافی ناشی از برش دادن چارچوب BCR احتمالاً ناچیز خواهد بود.

۳.۱۵ برآورد کردن

در محاسبهٔ برآوردها از روی نمونه‌ها باید جنبه‌های احتمالاتی این طرحها نیز در نظر گرفته شوند. در اینجا اصول اساسی این قبیل برآورد کردن به طور مختصر برای میانگینها (و به طور ضمنی برای نسبتها) و واریانسهای نمونه‌گیری آنها توصیف شده‌اند. به علاوه، در بعضی موارد، پس طبقه‌بندی، یا تعدیل کنترل جامعه، در داده‌های آمارگیری تلفنی به کار رفته است تا در تعدیل نمونهٔ خانوارهای تلفن‌دار با توزیع همهٔ خانوارها تلاش شود.

۱.۳.۱۵ برآورد میانگینها

برای طرح شماره‌گیری رقم تصادفی ساده، فرض کنید $\bar{y}_{RDD}$ میانگین سادهٔ n مشاهده از y، متغیر خانوار باشد. به همین ترتیب فرض کنید $\bar{y}_{MW}$ میانگین سادهٔ mk مشاهدهٔ تحت طرح میتوفسکی ـ واکسبرگ باشد. هم $\bar{y}_{RDD}$ و هم $\bar{y}_{MW}$ برای میانگین جامعه‌ای $\bar{Y}$ از لحاظ طرح نااریب‌اند. به علاوه،

$$\text{var}(\bar{y}_{MW}) \cong \frac{\sigma^2}{mk}(1+\rho(k-1)) \qquad \text{و} \qquad \text{var}(\bar{y}_{RDD}) = \frac{\sigma^2}{n}$$

که در آن σ^2 واریانس جامعه‌ای و ρ همبستگی درون انبوههٔ ۱۰۰ تایی برای متغیر y است.

برآورد کردن میانگین جامعه برای طرحهای طبقه‌بندی شدهٔ مبتنی بر راهنمای تلفن قدری پیچیده‌تر است. نمونه‌گیری در داخل طبقه، شماره‌گیری رقم تصادفی است. پس $\bar{y}_h$ (میانگین سادهٔ n_h مشاهده از طبقهٔ h^n) برای میانگین طبقه‌ای $\bar{Y}_h$ نااریب است. نتیجه آن که

$$y'_t = \sum_{h=1}^{H} N_h \left(\frac{n_h}{n'_h} \right) \bar{y}_h$$

در انبوهش جامعه‌ای مقادیر y برای خانوارهای تلفن‌دار تقریباً نااریب است و

$$N'_t = \sum_{h=1}^{H} N_h \left(\frac{n_h}{n'_h} \right)$$

برای کل تعداد خانوارهای تلفن‌دار، مثلاً N_t، تقریباً نااریب است. به این ترتیب، برآوردگر نسبتی $\bar{y}_{strat} = \left(y'_t / N'_t \right)$، برای میانگین جامعه‌ای تقریباً نااریب است و

$$\text{var}(\bar{y}_{strat}) \cong \sum_{h=1}^{H} \frac{z_h^2 \sigma_h^2 (1 + (1 - p_h)\lambda_h)}{n_h}$$

که در آن p_h نسبت شماره تلفنهای واگذار شده به خانوارهای دارای تلفن خانگی در طبقهٔ h ام است، z_h نسبت جمعیت خانوارهای تلفن‌دار منظور شده در طبقهٔ h ام است، و $\lambda_h = \left(\bar{Y}_h - \bar{Y} \right)^2 / \sigma_h^2$.

چندین نکتهٔ آماری دیگر هست که باید هنگام استفاده از طرحهای تلفنی به یاد داشت:

۱. به طور کلی، برای برآورد کردن میانگینهای زیرردهای به برآوردگرهای نسبتی نیاز است که عبارتهای واریانسی نسبتاً سادهٔ بالا در مورد آنها کاربرد ندارند.

۲. این طرحها نمونه‌هایی از خانوارها را ارائه می‌دهند، نه اشخاص را. اگر اشخاص در درون خانوارها انتخاب شوند، آن‌گاه وزن‌دار کردن اضافی و برآوردگرهای پیچیده‌تری مورد نیازند.

۳. برای داشتن برآوردگرهای نااریب باید وزنهای مربوط به خانوارهای دارای چند تلفن تعدیل شود تا احتمال بیشتر برای انتخاب شدن آنها به حساب آید.

٤. برآوردگرهای بالا برای دستیابی به اندازهٔ نمونه‌ای ثابت بر استفاده از شماره‌گیری رقم تصادفی مبتنی هستند. این مستلزم آن است که وضعیت همهٔ شماره تلفنهای انتخاب شده تعیین شده باشد، که این نیز به نوبهٔ خود مستلزم دفترداری دقیق و نظارت از نزدیک است. چون برای هر انبوههٔ ۱۰۰ تایی که نگهداری می‌شود اندازه‌های نمونه ای ثابت مورد نیاز است، روش میتوفسکی ـ واکسبرگ پیچیده‌تر می‌شود و به همین دلیل نیاز به کنترل شدیدتر نیز اهمیتی بیشتر دارد.

۲.۳.۱۵ برآورد کردن واریانس نمونه‌گیری

به منظور برآورد $Var(\bar{y}_{rdd})$ ، فرض کنید y_i مقدار متغیر y برای i امین خانوار انتخابی باشد. برآوردگر نااریب برای $Var(\bar{y}_{rdd})$ از فرمول زیر به دست می‌آید

$$\hat{Var}(\bar{y}_{rdd}) = \frac{\hat{\sigma}^2}{n}$$

که در آن،

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{rdd})^2}{n-1}$$

برای نمونه‌گیری میتوفسکی ـ واکسبرگ فرض می‌کنیم y_{ij} مقدار متغیر y برای j امین خانوار انتخاب شده در i امین انبوهۀ ۱۰۰ تایی نگهداری شده باشد. برآوردگر نااریب برای $Var(\bar{y}_{mw})$ از فرمول زیر به دست می‌آید

$$\hat{Var}(\bar{y}_{mw}) = \frac{1}{m}\frac{\sum_{i=1}^{m}(\bar{y}_i - \bar{y}_{mw})^2}{m-1}$$

که در آن،

$$\bar{y}_i = \frac{\sum_{j=1}^{k} y_{ij}}{k}$$

برای طرح طبقـــه‌بندی شده فرض می‌کنیم y_{hi} مقدار متغیر y برای i امین خانوار انتخاب شده در h امین طبقه باشد. کاربرد فن خطی‌سازی در مورد برآوردگر نسبتی $\bar{y}_{strat}$، برآوردگر واریانسی زیر را به دست می‌دهد

$$\hat{Var}(\bar{y}_{strat}) = \sum_{h=1}^{H}\frac{\hat{z}_h^2\hat{\sigma}_h^2(1+(1-\hat{p}_h)\hat{\lambda}_h)}{n_h}$$

که در آن،

$$\hat{p}_h = \frac{n_h}{n'_h} \qquad \hat{z}_h = \frac{N_h\hat{p}_h}{N'_t} \qquad \hat{\sigma}_h^2 = \frac{\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2}{n_h - 1} \qquad \hat{\lambda}_h = (\bar{y}_h - \bar{y}_{strat})^2 / \hat{\sigma}_h^2$$

اگرچه نتایج به تفصیل ارائه نشده‌اند، فن خطی‌سازی می‌تواند در به دست آوردن برآوردگرهایی برای واریانس برآوردگرهای نسبتی مورد نیاز مربوط به میانگینهای زیررده نیز مورد استفاده قرار بگیرد.

۳.۳.۱۵ پس طبقه‌بندی

در نظریهٔ نمونه‌گیری سنتی، پس طبقه‌بندی هنگامی مطرح می‌شود که متغیرهایی که باید برای ایجاد طبقات مورد استفاده قرار گیرند در زمان انتخاب کردن در دسترس نباشند. یعنی ممکن است علاقه‌مند به افراز جامعه به G پس طبقه، با استفاده از متغیرهای جمع‌آوری شده در طی آمارگیری باشند. مانند نمونه‌گیری طبقه‌بندی شده با انتساب متناسب، بهبود در دقت با تعدیل مناسب برای برآورد کردن واریانس امکان‌پذیر است. پس طبقه‌بندی مستلزم آن است که پس طبقه‌ها برای هر عنصر نمونه معلوم باشد و وزنهای پس طبقه، مثلاً W_g، برای هر پس طبقه موجود باشند. وزنهای پس طبقه‌ای باید از یک منبع خارجی از قبیل یک سرشماری، پیش‌بینیهای سرشماری، یا سوابق اداری به دست آمده باشند. مثلاً، پس طبقه‌های مبتنی بر سن و جنسیت در صورتی می‌توانند برای پاسخگویان ایجاد شوند که فراوانیها یا نسبتهای W_g جامعه‌ای مناسب را برای گروههای سنی و جنسی بتوان در جامعه پیدا کرد.

در نمونه‌گیری تلفنی غالباً پس طبقه‌بندی نه تنها با جمعیت ساکن در خانوارهای تلفن‌دار بلکه با جمعیت ساکن در همهٔ خانوارها تعدیل می‌شود. این حالت از پس طبقه‌بندی به منظور به دست آوردن برآوردهایی به کار برده می‌شود که به مفهومی خاص با توزیع جمعیت، نه تنها در خانوارهای تلفن‌دار، بلکه در همهٔ خانوارها تعدیل شده باشند.

گامهای پس طبقه‌بندی را می‌توان به صورت زیر خلاصه کرد:

۱. تفکیک نمونه در G پس طبقه براساس مشاهدهٔ برخی مشخصه‌ها.
۲. به دست آوردن وزنهای نمونه‌ای W_g برای جامعه، نوعاً از روی یک منبع خارجی از قبیل یک آمارگیری بزرگتر، یک سرشماری، پیش‌بینیهای سرشماری، یا سوابق اداری.
۳. محاسبهٔ میانگینهای $\bar{y}_g$ برای مشخصهٔ موردنظر، جداگانه برای هر پس طبقه، و محاسبهٔ میانگین کل $\bar{y}_{ps} = \sum_{g=1}^{G} W_g \bar{y}_g$.
٤. برای برآوردهای واریانسی، استفاده از

$$\hat{Var}(\bar{y}_{ps}) \cong \frac{1}{n}\left[\sum_{g=1}^{G} W_g s_g^2 + \sum_{g=1}^{G} W_g (1-W_g)\frac{s_g^2}{N_g}\right]$$

یا

$$\hat{Var}(\bar{y}_{ps}) \cong \frac{1}{n}\sum_{g=1}^{G} W_g s_g^2\left[1+\frac{1-W_g}{N_g}\right]$$

که در آن، s_g^2 برآورد عنصر در داخل پس طبقه، و N_g اندازهٔ جامعه برای پس طبقهٔ g است. شکل برآورد s_g^2 به طرح نمونه‌ای بستگی خواهد داشت.

تقریباً در همهٔ وضعیتهای عملی، $\bar{y}_{ps}$، برآورد پس طبقه‌بندی، واریانسهایی کوچکتر از واریانس برآورد میانگین بدون پس طبقه‌بندی خواهد داشت.

از پس طبقه‌بندی غالباً به عنوان *تعدیل کنترل جامعه* نیز نام برده می‌شود. وزنهای پس طبقه‌بندی شده برای هر عنصر بسط داده می‌شوند. برآوردهای موزون که با استفاده از وزنهای پس طبقه‌بندی شده محاسبه می‌شوند نسبت به توزیع بیرونی که به صورت W_g نشان داده می‌شود «تعدیل» می‌شوند. در مورد طرح شماره‌گیری رقم تصادفی، اثرهای این تعدیل را در صورتی می‌توان با وضوح بیشتر مشاهده کرد که برآورد میانگین پس طبقه‌بندی شده را مجدداً به شکل زیر بیان کنیم. فرض می‌کنیم r معرف تعداد پاسخگویان در نمونه و r_g معرف تعداد پاسخگویان در پس طبقهٔ g ام باشد. علاوه بر این، فرض می‌کنیم y_{gi} نشانگر مقدار مشخصهٔ y برای i امین پاسخگو در g امین پس طبقه است. در این صورت می‌توان میانگین پس طبقه را برحسب عنصر وزنها، W_{gi}، به صورت زیر نوشت:

$$\bar{y}_{ps} = \sum_{g=1}^{G} W_g \bar{y}_g = \frac{\sum_{g=1}^{G} N_g \bar{y}_g}{N} = \frac{\sum_{g=1}^{G} \left(\frac{r}{N}\right)\left(\frac{N_g}{r_g}\right)\sum_{i=1}^{r_g} y_{gi}}{\sum_{g=1}^{G}\left(\frac{r}{N}\right)N_g}$$

$$= \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} y_{gi}}{\sum_{g=1}^{G}\frac{N_g/N}{1/r}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} y_{gi}}{\sum_{g=1}^{G}\sum_{i=1}^{r_g}\left(\frac{1}{r_g}\right)\frac{N_g/N}{1/r}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} y_{gi}}{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi}}$$

یعنی وزن، W_{gi}، حاصل بخش نسبت در جمعیت در g امین پس طبقه و نسبت در نمونه در g امین پس طبقه است: $W_{gi} = (N_g/N)/(r_g/r)$. بنابراین، پس طبقه‌بندی یک نمونه از پاسخگویان خانوارهای تلفنی برای توزیعی براساس همهٔ خانوارها، تعدیلی همزمان برای بی‌پاسخی و بی‌پوششی خانوارهای بدون تلفن فراهم می‌کند.

پس طبقه‌بندی نمونه‌های تلفنی جنبه‌هایی دارد که توجه به آنها حائـز اهمیت است. اول، نوعـاً، W_g ها داده‌های سرشماری یا سایر داده‌های مربوط به همهٔ خانوارها و نه صرفاً خانوارهای دارای تلفن هستند. دوم، در حالی که تعدیل پس طبقه‌بندی می‌تواند به عنوان تعدیلی هم برای بی‌پاسخی و هم برای بی‌پوششی تلقی شود، در عمل غالباً پس از شکلی از جبران بی‌پاسخی از طریق

وزن دادن، به کار برده می‌شود. سوم، ممکن است به دست آوردن وزنهای جامعه‌ای، W_g، برای رده‌بندی متقاطع کامل برای جامعه امکان‌پذیر نباشد، ولی توزیعهای حاشیه‌ای ممکن است موجود باشند. شیوه‌های تعدیل نسبتی با دادن وزن (ن. ک. کالتون و کاسپرزیک [۷]) می‌تواند برای تولید یک توزیع کامل از رده‌بندی متقاطع براساس توزیعهای حاشیه‌ای مورد استفاده قرار گیرد. مثلاً وزنهای جامعه‌ای ممکن است برای سن و تحصیلات موجود باشند، ولی برای رده‌بندی متقاطع آنها موجود نباشند. برآورد نسبتی با دادن وزن می‌تواند برای تولید رده‌بندی متقاطع براساس یک مدل «اثرهای اصلی» برای سن و تحصیلات مورد استفاده واقع شود. پس از آن، وزنهای رده‌بندی متقاطع با دادن وزن برای جامعه در مورد توزیع پاسخگو به کار خواهند رفت تا وزنهایی در سطح عنصر به صورتی که در بالا اشاره شد تولید شوند.

۴.۱۵ مقایسهٔ طرحها

۱.۴.۱۵ مبادلهٔ هزینه ـ واریانس

تابع هزینه در بخش ۲.۱.۲.۱۵ و واریانس نمونه‌گیری یک برآورد میانگین در بخش ۱.۳.۱۵ اندازهٔ نمونه‌ای k در درون انبوههٔ ۱۰۰ تایی را تعیین می‌کنند که هزینهٔ مورد انتظار یک واریانس نمونه‌گیری ثابت را برای طرح میتوفسکی ـ واکسبرگ به حداقل می‌رساند (یا واریانس را برای یک هزینهٔ ثابت به حداقل می‌رساند). یک عبارت صریح برای مقدار بهینهٔ k را می‌توان در مقالهٔ واکسبرگ [۱۸] یافت. به همین ترتیب، تابع هزینه در بخش ۲.۳.۲.۱۵ و واریانس برآورد میانگین در بخش ۱.۳.۱۵ انتسابی از نمونه را به طبقه‌ها تعیین می‌کنند که هزینهٔ مورد انتظار یک واریانس ثابت را برای طبقه‌بندی مبتنی بر راهنمای تلفن به حداقل می‌رساند (یا واریانس را برای یک هزینهٔ ثابت به حداقل می‌رساند). عبارتهای صریح برای انتساب نمونه را می‌توان در مقالهٔ کاسادی و لپکوفسکی [۳] یافت.

با استفاده از مقادیر مورد قبول عوامل هزینه‌ای و پارامترهای جامعه‌ای برای مدلهای هزینه‌ای ساده و عبارتهای مربوط به واریانس که در بالا ذکر شدند، کاسادی و لپکوفسکی چنین نتیجه‌گیری کرده‌اند که طرح میتوفسکی ـ واکسبرگ و طبقه‌بندی مبتنی بر راهنمای تلفن هر دو نسبت به طرح شماره‌گیری رقم تصادفی ساده بهبود قابل ملاحظه‌ای داشته‌اند. آنها همچنین به این نتیجه رسیده‌اند که براساس مدل هزینه‌ای ساده به تنهایی، دو رهیافت مزبور از نظر کارایی تفاوتی اندک با یکدیگر دارند. ولی چنانچه امکان اریبی بیشتر را بتوان تحمل کرد، در آن صورت، طرح برشی از همه کارآمدتر است.

۲.۴.۱۵ ملاحظات اجرایی

ویژگیهای سیستمهای تلفن که بر اجرای طرحهای توصیف شده در بخش اخیر تأثیر می‌گذارند بسیار فراوان‌اند. در این زیربخش، چندین مورد را که مهمترند مورد بحث قرار می‌دهیم.

شناسایی وضعیت خانگی بودن هر شماره تلفن که از شماره‌گیری رقم تصادفی یا نمونه‌هایی به یاری فهرست تولید شده است همیشه فرایندی آسان نیست. شماره‌هایی که جواب می‌دهند باید از نظر کاربری خانگی بررسی شوند و آنهایی که مشترکاً برای مقاصد خانگی و تجاری مورد استفاده قرار می‌گیرند باید به طرز مناسبی رده‌بندی شوند (معمولاً هر کاربری خانگی کافی است که شمارهٔ تلفن به عنوان خانگی رده‌بندی شود). بعضی شماره‌ها فوراً به عنوان غیرخانگی شناسایی می‌شوند زیرا دایر نیستند و وضعیت ثبتی آنها نیز به وضوح این وضعیت را نشان می‌دهد. بسیاری از شماره‌هایی که دایر نیستند از نظر ثبتی نیز به جایی ارتباط پیدا نمی‌کنند که وضعیت آنها را نشان دهـد، ولی به یـک «ماشین زنگ‌زن» متصل‌اند. به این ترتیب، مصاحبه‌گرانی که شمارهٔ تلفنها را برای تعیین وضعیت خانگی بودن آنها غربالگری می کنند نمی‌توانند بین شمارهٔ تلفنهای خانگی که زنگ می خورند ولی کسی در منزل نیست و شمارهٔ تلفنهایی که فعلاً دایر نیستند تفاوت قایل شوند.

این مشکل اخیر در پیاده‌سازی برخی طرحها دارای اهمیتی خاص است. سامان دادن به شماره‌هایی که بدون جواب زنگ می‌خورند در طرحهای شماره‌گیری رقم تصادفی دو مرحله‌ای که مستلزم جایگذاری شماره تلفنهای غیر خانگی است، به خصوص در ایامی که جمع‌آوری داده‌های آمارگیری با محدودیت زمانی همراه است، بسیار مشکل است. بسیاری از سازمانهای آمارگیری با شماره تلفنهایی که در اوقات گوناگون روز و روزهای مختلف هفته بدون جواب زنگ می‌خورند به عنوان غیر خانگی رفتار می‌کنند. اگر رده‌بندی به عنوان غیر خانگی در طی دورهٔ بررسی دیر انجام شده باشد شمارهٔ جایگزین مدت زمان نسبتاً کوتاهی خواهد داشت تا به آن تلفن شود. غالباً شماره‌های جایگزین همان تنوع زمانی را در طی روز و در روزهای هفته که شماره‌های اصلی در اختیار داشته‌اند برای تلفن کردن در اختیار ندارند. سازمانهای آمارگیری شیوه‌های نمونه‌گیریی را که نمونهٔ ثابتی از شماره تلفنها ارائه می‌دهند به شیوه‌ای که ممکن است در اواخر دورهٔ آمارگیری شماره تلفنهای جدیدی تولید کند ترجیح می‌دهند.

در پایان دورهٔ بررسی باید شماره تلفنهایی که بدون جواب زنگ خورده‌اند و چندین بار به آنها تلفن شده است به عنوان خانگی یا غیر خانگی رده‌بندی شوند تا بررسی بسته شود. اگر به شماره‌ای در ساعتها و روزهای گوناگون تلفن شده است می‌توان آن را به طور دلخواه به عنوان غیر خانگی رده‌بندی کرد. به این ترتیب در برابر نرخ پاسخ به آمارگیری شمارش نخواهد شد، زیرا به عنوان غیرنمونه‌ای

رده‌بندی شده است. از سوی دیگر، شماره تلفنهایی که بدون جواب زنگ خورده‌اند ولی به دفعات کافی به آنها تلفن نشده است نوعاً به عنوان خانگی و بدون پاسخ رده‌بندی می‌شوند و به محاسبه‌ای محافظه‌کارانه از نرخ پاسخ می‌انجامند.

برای غلبه بر این مشکلات و کاهش هزینه‌های غربالگری شماره تلفنها از لحاظ خانگی بودن، سیستمهای غربالگری خودکار ابداع شده‌اند تا حداقل، شماره تلفنهایی را که به برنامه‌های ضبط شده متصل‌اند شناسایی کرده، معلوم کنند که دایرند یا نه. برنامهٔ ضبط شده برای «خارج از سرویس» نوعاً با یک «فاصلهٔ سه گام کامل موسیقی» شروع می‌شود که با صدای زنگ تلفن همراه نیست. نوعی سخت‌افزار و نرم‌افزار انحصاری تهیه شده است که تلفنها را شماره‌گیری و طنین فاصلهٔ سه گام را کشف می‌کند. شماره تلفنهایی که دارای برنامهٔ ضبط شدهٔ فاصلهٔ سه گام هستند از نمونه‌گیری بعدی حذف می‌شوند. شماره‌هایی که برنامهٔ سه گام را ندارند غالباً «صدای قطع زنگ» را تجربه می‌کنند که طی آن در حالی که سخت‌افزار ارتباط را قطع می‌کند صدای زنگ تلفن برای لحظه‌ای شنیده می‌شود.

آمارگیریهایی که در سطح ایالت یا کل کشور اجرا می‌شوند دارای مرزهایی جغرافیایی برای جامعهٔ خود هستند که با مرزهای کد ناحیه‌ای مطابقت دارند. شماره تلفنهای نمونه‌ای که در داخل کدهای ناحیه‌ای نمونه تولید می‌شوند به مکانهای مسکونی داخل ناحیهٔ جغرافیایی هدف اختصاص خواهند داشت. بسیاری از آمارگیریها جامعه‌های با تعریف جغرافیایی را هدف قرار داده‌اند که مرزهای آنها با کدهای ناحیه‌ای و مرزهای مراکز تلفن جور نمی‌شوند. در این قبیل موارد می‌توان جامعه را دوباره تعریف و آن را به جامعهٔ ساکن در محدودهٔ مراکز تلفن تعیین شده محدود کرد، یا می‌توان نمونه‌ای از یک مجموعه از مراکز تلفن انتخاب کرد که سراسر ناحیهٔ جغرافیایی را پوشش دهد ولی نواحی خارج از هدف را نیز شامل شود. در آن صورت باید شماره تلفنها نه تنها برای تعیین وضعیت خانگی بودن بلکه برای مکان‌یابی اماکن مسکونی براساس خود ـ گزارش پاسخگو نیز غربالگری شوند. رده‌بندی شماره تلفنهایی که بدون جواب زنگ می‌خورند در این قبیل آمارگیریهای غربالگری، به مراتب مسئله‌سازتر است.

تشخیص دوباره‌کاریها در هر یک از چارچوبها نیز نوعاً مستلزم خود ـ گزارش پاسخگو است. از خانوارهای پاسخگو سؤال می‌شود که آیا بیش از یک شماره تلفن به آن خانوار اختصاص دارد یا نه و اگر پاسخ مثبت بود شمارهٔ این قبیل تلفنها پرسیده می‌شود. این شماره تلفن خود ـ گزارشی که از طریق آن می‌توان به خانوار دسترسی پیدا کرد بعداً در تولید وزن برای برآورد کردن استفاده می‌شود. بسیاری از سازمانهای آمارگیری، ارتباطهای اشتباه و شماره‌گیریهای غلط تلفنچی را نیز بازبینی می‌کنند.

شماره تلفنهایی که اشتباهاً شماره‌گیری شده‌اند مانند ارتباطهای اشتباه حذف می‌شوند تا از پیچیدگی بیشتر کارها در فرایند وزن دادن برای مدخلهای تکراری خانوار پرهیز شود.

آمارگیریهای علوم اجتماعی و بهداشتی نیز غالباً یک فرد واجد شرایط را در خانوار برای مصاحبهٔ بیشتر انتخاب می‌کنند. مثلاً در یک آمارگیری مربوط به رضایت از وضعیت ازدواج، یک فرد بزرگسال انتخاب خواهد شد تا از تأثیر گفت‌وگوی بزرگسالان دربارهٔ محتوای آمارگیری بین مصاحبه‌ها بر پاسخها، جلوگیری شود. انتخاب پاسخگو باید در مراحل اولیهٔ مصاحبه انجام پذیرد. شیوهٔ انتخاب عینی پاسخگو که کیش [۸] توصیف کرده است در آمارگیریهای تلفنی برای این منظور کاربرد وسیعی داشته است ولی این شیوه به نتیجه‌ای نامطلوب ـ یعنی نرخ افزایش یافتهٔ بی‌پاسخی ـ می‌انجامد. خانوارها تمایل ندارند در آمارگیریی شرکت کنند که اولین سؤالهایی که در آن مطرح شده است برای به دست آوردن فهرستی از افراد واجد شرایط ساکن در خانوار است. روشهای دیگر، شامل شیوهٔ توصیف شده توسط ترولدال و کارتر [۱۷]، و روش نزدیکترین تاریخ تولد است (برای توصیف آن نگاه کنید به لاوراکاس [۹]). نشان داده شده است که این شیوهٔ آخر دارای اریبی در انتخاب کردن است ولی همچنان مورد استفاده قرار می‌گیرد زیرا کاربرد آن آسان است و نگرانیهای مربوط به افزایش نرخهای بی‌پاسخی را که با شیوهٔ انتخاب کردن درون خانواری کیش همراه است ندارد.

سرانجام این که دستگاههای پاسخگوی تلفن و تلفنهای همراه مشکلات روزافزونی را برای عملیات نمونه‌گیری تلفنی مطرح ساخته‌اند. ماشینهای پاسخگو تا حدود زیادی تشخیص سریع واحدهای مسکونی را میسر ساخته‌اند. می‌توان پیامهایی گذاشت که از خانوار درخواست شود با یک شماره تلفن مجانی تماس بگیرد و تلفن کردن به خانوارهای دارای ماشین پاسخگو را می‌توان در اوقات گوناگون روز و روزهای مختلف هفته زمانبندی کرد تا برای دستیابی به خانوار در زمانی که کسی به تلفن جواب می‌دهد تلاش شود. تلفنهای همراه مشکلی دیگر را مطرح می‌کنند. آیا این شماره تلفنها خانگی هستند یا تجاری؟ به علاوه ممکن است مشترک مورد نظر با جواب دادن به تلفن هزینه‌ای متحمل شود. شمارهٔ تلفنهای همراه اکثراً در پیش شماره‌های جداگانه‌ای قرار داده نشده‌اند بلکه مانند سایر شماره تلفنها با همان پیش شماره‌ها مخلوط شده‌اند. این توزیع، شناسایی را مشکل می‌کند. با وجود این به تلفنهای همراه ممکن است سریعتر از تلفن خانگی جواب داده شود. به علاوه، پرسشگری که خوب آموزش دیده باشد می‌تواند ترتیبی اتخاذ کند تا به خانواری با شمارهٔ دیگر تلفن بزند و به این ترتیب هزینهٔ مشترک تلفن را کاهش دهد.

۳.٤.۱۵ انتخاب از میان طرحهای گوناگون

همان طور که قبلاً اشاره شد، گزینش از میان طرحهای گوناگون تا اندازهٔ زیادی به در نظر گرفتن ویژگیهای مربوط به هزینه و خطای هر طرح مبتنی است. نوعاً سه عامل عمدهٔ هزینه‌ای در نظر گرفته می‌شود: هزینهٔ تولید نمونهٔ شماره تلفنها، هزینهٔ غربالگری نمونه، و «راحتی» کار با شیوهٔ نمونه‌گیری موردنظر در پیاده‌سازی (هزینه‌ای که غالباً کمی کردن آن مشکل است). از جهت خطا، دو نگرانی عمده وجود دارند: پوشش جامعهٔ خانوارهای تلفن‌دار و واریانس نمونه‌گیری.

اگر سه رقیب اصلی را با توجه به این مشخصه‌ها مورد آزمایش قرار دهیم می‌توانیم درک کنیم که چرا سازمانها امروزه از میان طرحهای مختلف، گزینشهای خاصی را به عمل می‌آورند. مثلاً، تولید شماره تلفنها در طرح نمونه‌ای شماره‌گیری رقم تصادفی دو مرحله‌ای میتوفسکی ـ واکسبرگ کم‌خرج است. غـربالگری در مرحلهٔ دوم کارایی قرار دارد زیرا بیش از ٦۰٪ شماره تلفنها خانگی‌اند. طرح میتوفسکی ـ واکسبرگ در اجرا مشکلات چندی را مطرح می‌کند، از جمله جایگذاری شماره تلفنهای غیر خانگی و خوشه‌های خالی شده که می‌توانند برای برخی عملیات آمارگیری، ناراحتیهای قابل توجهی باشند و روشهای دیگری که از این مشکلات دوری کنند جاذبهٔ زیادی پیدا می‌کنند. از لحاظ خطا، طرح میتوفسکی ـ واکسبرگ واقعاً پوشش کامل جامعهٔ خانوارهای تلفن‌دار را تأمین می‌کند. واریانسهای نمونه‌گیری بیشتر از طرحهای نمونه‌ای عنصری است و علت آن، انتخاب نمونهٔ خوشه‌ای و افزایشهای واریانس ناشی از همگنی درونی انبوههٔ ۱۰۰ تایی در میان عناصر نمونه است. یعنی این که اثرهای طرح برای نمونه‌های میتوفسکی ـ واکسبرگ بیشتر از یک است.

طرح طبقه‌بندی شده دارای مجموعهٔ مشخصه‌هایی تا حدودی متفاوت است. هزینه‌های تولید نمونه می‌توانند زیاد باشند. نمونهٔ طبقه‌ای فهرست شده را می‌توان با هزینهٔ نسبتاً کمی به ازای تعداد نمونه از یک شرکت نمونه‌گیری تجاری خرید ولی نمونهٔ طبقه‌ای فهرست نشده مستلزم طبقه‌بندی بیشتـر شماره تلفنها و نمونـه‌های شمـاره‌گیری رقم تصـادفی دو مرحله‌ای است که از هر طبقهٔ فهرست نشده انتخاب شده باشد. هزینه‌های غربالگری نیز بیشتر از طرح میتوفسکی ـ واکسبرگ است زیرا تقریباً ۵۰٪ شماره تلفنهای نمونه در طبقهٔ فهرست شده خانگی هستند و در طبقهٔ فهرست نشده درصد بسیار کمی خانگی‌اند. با توجه به این که روشهای نمونه‌گیری مختلف در طبقه‌ها مورد استفاده قرار می‌گیرند، راحتی انتخاب نمونه برای طرح طبقه‌بندی شده کمتر از طرح میتوفسکی ـ واکسبرگ است. ولی طرح طبقه‌بندی شده نیاز به جایگزین کردن شماره تلفنها و خوشه‌های خالی شده را از میان می‌برد. طرح طبقه‌بندی شده در واقع کل جامعهٔ خانوارهای تلفن‌دار را پوشش می‌دهد. واریانسهای نمونه‌گیری برای

طرح طبقه‌بندی شده کمتر از طرح میتوفسکی ـ واکسبرگ خواهد بود زیرا نمونه‌گیری از عناصر، دارای واریانسهای کمتری است و می‌توان بهبودهایی را در دقت ناشی از طبقه‌بندی انتظار داشت.

طرح برشی به نسبت طرحهای میتوفسکی ـ واکسبرگ و طبقه‌بندی شده دارای عیب عدم پوشش خانوارهای تلفن‌دار در انبوهه‌های ۱۰۰ تایی فاقد شماره تلفنهای فهرست شده است. سطح عدم پوشش پایین است و بررسیهای تجربی نشان داده‌اند که تفاوت بین جوامع پوشش داده شده و پوشش داده نشده برای بسیاری از مشخصه‌ها اندک است. نمونه‌های انتخاب شده با استفاده از طرح برشی هنگامی که از شرکتهای نمونه‌گیری تجاری به دست آمده باشند گرانبها نیستند. هزینه‌های غربالگری طرح برشی نسبت به طرحهای میتوفسکی ـ واکسبرگ و طبقه‌بندی شده در حد متوسط است زیرا تقریباً ۵۵٪ شماره تلفنهای تولید شده خانگی خواهند بود. طرح برشی در میان سه طرحی که در اینجا مورد بررسی قرار گرفته‌اند از همه راحت‌تر است زیرا به هیچ نوع جایگذاری شماره تلفنها نیاز ندارد، نمونه فقط از طبقهٔ فهرست شده انتخاب می‌شود، و هیچ نمونه‌گیری دو مرحله‌ای از روی طبقهٔ فهرست نشده لازم نیست. واریانسهای نمونه‌گیری برآوردها برای طرح برشی باید از همه کمتر باشد زیرا این یک نمونهٔ عنصری طبقه‌بندی شده بدون نمونه‌گیری خوشه‌ای است.

به این ترتیب، کسی که در حوزهٔ نمونه‌گیری کار می‌کند با موضوع گزینش از میان طرحهایی که پوشش کامل را تأمین می کنند ولی با تعدادی نابسامانیها در انتخاب همراه‌اند و یک طرح که پوشش کمتری را تأمین می‌کند ولی با راحتیهایی در انتخاب همراه است مواجه می‌شود. در کاربست فعلی با توجه به شواهد تجربی مربوط به اندازهٔ اریبی ناشی از عدم پوشش خانوارهای تلفن‌دار در انبوهه‌های ۱۰۰ تایی فاقد شماره تلفنهای فهرست شده، طرح برشی اخیر را ترجیح می‌دهند. یعنی، دست‌اندرکاران، روشهای نمونه‌گیری برشی را برای آمارگیریهای تلفنی براساس یک روش کلاسیک ولی غیررسمی مبادلهٔ هزینه ـ خطا انتخاب می‌کنند.

۵.۱۵ خلاصه

در این فصل، روشهای مناسبی را برای نمونه‌گیری از خانوارهای تلفن‌دار با استفاده از چارچوب شمارهٔ تلفن و چارچوب مبتنی بر راهنمای تلفن شرح و بسط دادیم. شیوه‌های شماره‌گیری رقم تصادفی دو مرحله‌ای به صورتی که میتوفسکی ابداع کرده و واکسبرگ شرح و بسط داده است در سطحی گسترده مورد استفاده قرار می‌گیرند. سایر رهیافتهای جاری به نسبت شیوهٔ مزبور دارای نارساییهایی از نظر پوشش و هزینه هستند. این نارساییها از طریق طرحهای نمونه‌ای تلفنی که از اطلاعات شماره تلفنهای فهرست شده برای بهبود کارایی هزینهٔ شماره‌گیری رقم تصادفی استفاده می‌کنند مورد

رسیدگی قرار گرفته‌اند. چارچوب شماره تلفن به یک طبقه که اطلاعات شماره تلفنهای فهرست شده برای آن در سطح انبوهه‌های ۱۰۰ تایی موجود است و یک طبقه که برای آن چنین اطلاعاتی موجود نیست تقسیم می‌شود. کاراییهای طرحهای نمونه‌گیری گوناگون برای این طرح طبقه‌بندی شده با شماره‌گیری رقم تصادفی ساده و فن میتوفسکی ـ واکسبرگ مقایسه شده است. افزایش نه چندان زیاد کارایی این طرحهای جایگزین را نسبت به فن میتوفسکی ـ واکسبرگ می‌توان تقریباً برای همهٔ این قبیل طرحها نشان داد. همچنین دربارهٔ استفاده از پس طبقه‌بندی برای استنباطهایی از روی جامعهٔ خانوارهای تلفن‌دار برای یک جامعهٔ وسیعتر و دربارهٔ مسایل عملیاتی که در اجرای روشهای نمونه‌گیری تلفنی مطرح می‌شوند به طور خلاصه بحث کردیم.

تمرین

۱.۱۵ قرار است یک آمارگیری تلفنی با $n = ۱۶۰۰$ مصاحبهٔ تکمیل شدهٔ مطلوب در یک منطقهٔ کلانشهری بزرگ اجرا شود. نرخ مورد انتظار برای تلفنهای خانگی دایر ۳۵٪ است. یک بزرگسال با ۲۱ سال سن یا مسنتر قرار است از هر خانوار انتخاب شود (واجد شرایط بودن خانوار مورد انتظار ۹۵٪ است) و نرخ پاسخگویی مورد انتظار ۷۰٪ است. طرح نمونه‌ای شماره‌گیری رقم تصادفی دو مرحله‌ای میتوفسکی ـ واکسبرگ پیشنهاد می‌شود.

الف. اگر اندازهٔ خوشه را در $b = ۸$ شماره تلفن خانگی، ثابت فرض کنیم، چند شماره تلفن اولیه باید تولید کرد و چند خوشه باید به دست آورد؟

ب. اگر نرخ مورد انتظار دایر بودن تلفن ۶۵٪ باشد چند شماره تلفن باید در مرحلهٔ دوم تولید شود؟

پ. چند مشکل در طی مصاحبه پیش‌بینی می‌شود. با یک جملهٔ تنها توضیح دهید که آیا این شماره تلفنها به عنوان «نمونه یا غیرنمونه» رده‌بندی می‌شوند یا نه و نحوهٔ رفتار با هر یک را چگونه توصیه می‌کنید:

۱. ارتباط غلط

۲. تماس با خانوار برقرار شده و قرار مصاحبه برای زمان دیگری گذاشته شده است. وقتی در زمان مقرر دوباره شماره‌گیری می‌شود، شماره تلفن خارج از سرویس است.

۳. تلفن شماره‌گیری شده مورد استفاده یک شرکت حسابداری کوچک در یک خانهٔ مسکونی است ولی برای مکالمات شخصی نیز مورد استفاده قرار می‌گیرد.

٤. در پایان بررسی، تعداد کمی از شماره تلفنها باقی مانده‌اند که پس از یک تا ۲۰ بار شماره‌گیری هنوز بدون جواب زنگ می‌خورند.

۲.۱۵ یک نمونهٔ تلفنی در یک مرکز تلفن انتخاب شده است. یکصد (۱۰۰) شماره تلفن متوالی در یک انبوههٔ ۱۰۰ تایی که در مرحلهٔ اول طرح میتوفسکی ـ واکسبرگ انتخاب شده‌اند به ترتیب تصادفی قرار گرفته‌اند. فرض کنید که در یکی از انبوههای انتخاب شده به چهل شماره تلفن اول فهرست تصادفی شده تلفن شده و نتایج زیر برای هر یک به دست آمده است:

(۱) ۹۳٦-۰۰۷۱	مصاحبه	(۲۱) ۹۳٦-۰۰۲٦	مصاحبه
(۲) ۹۳٦-۰۰۵۳	کسب و کار	(۲۲) ۹۳٦-۰۰٤۹	کسب و کار
(۳) ۹۳٦-۰۰٤٦	مشغول (٦ بار تلفن)	(۲۳) ۹۳٦-۰۰۸٤	قطع شده
(٤) ۹۳٦-۰۰۸٤	واگذار نشده	(۲٤) ۹۳٦-۰۰۰۳	زنگ بدون جواب
(۵) ۹۳٦-۰۰۸۹	واگذار نشده	(۲۵) ۹۳٦-۰۰۹٤	قطع شده
(٦) ۹۳٦-۰۰۷۹	کسب و کار	(۲٦) ۹۳٦-۰۰۱۹	کسب و کار
(۷) ۹۳٦-۰۰۵۰	امتناع	(۲۷) ۹۳٦-۰۰۳۱	مصاحبه
(۸) ۹۳٦-۰۰۲٤	مصاحبه	(۲۸) ۹۳٦-۰۰۵٦	امتناع
(۹) ۹۳٦-۰۰۲۱	ارتباط اشتباه	(۲۹) ۹۳٦-۰۰۱٦	زنگ بدون جواب
(۱۰) ۹۳٦-۰۰۰۱	قطع شده	(۳۰) ۹۳٦-۰۰۹۳	واگذار نشده
(۱۱) ۹۳٦-۰۰۳۸	مصاحبه	(۳۱) ۹۳٦-۰۰۸۷	زنگ بدون جواب
(۱۲) ۹۳٦-۰۰۹۰	تلفن پولی	(۳۲) ۹۳٦-۰۰٤۹	کسب و کار
(۱۳) ۹۳٦-۰۰٤۸	گوشی را می‌گذارند (۵ بار)	(۳۳) ۹۳٦-۰۰۱۰	نوفه
(۱٤) ۹۳٦-۰۰٤۵	زنگ بدون جواب (۱۵)	(۳٤) ۹۳٦-۰۰۹۳	واگذار نشده
(۱۵) ۹۳٦-۰۰۱٤	زنگ بدون جواب (۱۱)	(۳۵) ۹۳٦-۰۰۹۵	واگذار نشده
(۱٦) ۹۳٦-۰۰۰۲	پیام‌ده بیمارستان	(۳٦) ۹۳٦-۰۰۲۲	کسب و کار
(۱۷) ۹۳٦-۰۰۱۷	امتناع	(۳۷) ۹۳٦-۰۰۱۱	مصاحبه
(۱۸) ۹۳٦-۰۰۳۰	ماشین پاسخگو	(۳۸) ۹۳٦-۰۰۷۵	امتناع
(۱۹) ۹۳٦-۰۰۹٦	واگذار نشده	(۳۹) ۹۳٦-۰۰۳۷	واگذار نشده
(۲۰) ۹۳٦-۰۰٤٤	به شمارهٔ ۱۱۰۱-۷٤۷ تغییر یافته است	(٤۰) ۹۳٦-۰۰۱۲	واگذار نشده

برای هر یک از انواع نتایج زیر، بیان کنید که اگر در پایان بررسی این نتایج را به دست می‌آوردید آنها را به عنوان خانگی رده‌بندی می کردید یا غیر خانگی:

الف. قطع شده
ب. خارج از سرویس
پ. زنگ بدون پاسخ برای ۱۱ بار تلفن یا بیشتر
ت. ارتباط اشتباه (با ۲ بار شماره‌گیری)
ث. ماشین پاسخگو
ج. شماره تلفن عوض شده
چ. گوشی را می‌گذارند یا زنگ بدون جواب برای تلفنهای متعدد

۳.۱۵ ویژگیهای طرح زیر را برای یک طرح نمونهٔ تلفنی طبقه‌بندی شده در نظر بگیرید

طبقه	شرح	W_h	C_h
۱	انبوههٔ ۱۰۰ تایی فهرست شده	۰/۳۵	۰/۷۵ دلار
۲	انبوههٔ ۱۰۰ تایی فهرست نشده		
	پیش‌شمارهٔ فهرست شده		
	انبوههٔ ۱۰۰۰ تایی فهرست نشده	۰/۲٤	۷/۳۰
۳	باقیمانده	۰/٤۱	۱/٤۵

الف. یک انتساب بهینه برای یک طرح سه طبقه‌ای تهیه کنید که هزینه‌های غربالگری آن از ۱۲۵۰ دلار بیشتر نباشد. فرض کنید واریانسهای مربوط به عنصر در سراسر طبقه‌ها تقریباً یکسان‌اند. به عناصر موجود در هر طبقه، وزنهای مورد نیاز جبران کنندهٔ احتمالهای نابرابر برای انتخاب شدن را اختصاص دهید.

ب. فرض کنید تصمیم بر این است که طبقهٔ ۲ به علت هزینه‌های غربالگری شماره تلفنهای موجود در آن حذف شود (یعنی استفاده از طرح چارچوب برشی). حالا از کدام انتساب بهینه استفاده خواهد شد؟ وزنهای مورد نیاز عناصر برای جبران احتمالهای نابرابر برای انتخاب شدن کدامها خواهند بود؟

کتابشناسی

1. Brick, J. M., Waksberg, J., Kulp, D., and Starer, A., Bias in list assisted telephone samples. *Public Opinion Quarterly,* 59(2): 218-235, 1995.
2. Brunner, J. A., and Brunner, G. A., Are voluntarily unlisted telephone subscribers really different? *Journal of Marketing Research*, 8: 121-124, 1971.

3. Casady, R. J., and Lepkowski, J. M., Stratified telephone sampling designs. *Survey Methodology*, 19(1): 103-113, 1993.
4. Connor, J., and Heeringa, S., Evaluation of two cost-efficient RDD designs. Paper Presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL, 1992.
5. Frankel, M. R., and Frankel, L., Some recent developments in sample survey design. *Journal of Marketing Research*, 14: 280-293, 1977.
6. Groves, R. M., and Lepkowski, J. M., Dual frame mixed mode survey designs. *Journal of Official Statistics*, 1(3): 263-286, 1985.
7. Kalton, G., and Kasprzyk, D., The treatment of missing survey data. *Survey Methodology*, 12: 1-16, 1988.
8. Kish, L., *Survey Sampling*, Wiley, New York, 1965.
9. Lavrakas, P. J., *Telephone Survey Methods: Sampling, Selection, and Supervision*. Sage Publications, Newbury Park, Calif., 1987.
10. Lepkowski, J. M., Telephone sampling methods in the United States. In *Telephone Survey Methodology*, Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., II, and Waksberg, J., Eds., Wiley, New York, 1988.
11. Mitofsky, W., Sampling of Telephone Households. Unpublished CBS News memorandum, 1970.
12. Potthoff, R. F., Some generalizations of the Mitofsky-Waksberg techniques for random digit dialing. *Journal of the American Statistical Association*, 82(298): 409-418, 1987.
13. Potthoff, R. F., Generalizations of the Mitofsky-Waksberg technique for random digit dialing: Some added topics. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 615-620, 1987.
14. Stock, J. S., How to improve samples based on telephone listings. *Journal of Marketing Research*, 2(3): 50-51, 1962.
15. Sudman, S., The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10(2): 204-207, 1973.
16. Thornberry, O. T., and Massey, J. T., Trends in U.S. telephone coverage across time and subgroups. In *Telephone Survey Methodology*, Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nichols, W. L., II, and Waksberg, J., Eds., Wiley, New York, 1988.
17. Troldahl, V. C., and Carter, R. E., Jr., Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1(2): 71-76, 1964.
18. Waksberg, J., Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 19(1): 103-113, 1978.

The following review was written for the Encyclopedia of Biostatistics and contains, in a different format, much of the material that is in this chapter.

19. Casady, R. J., and Lepkowski, J. M., Telephone sampling. In *The Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.

فصل ۱۶

راهبردهایی برای تحلیل مبتنی بر طرح داده‌های آمارگیری نمونه‌ای

محققان، آمارگیریهای نمونه‌ای را نوعاً برای تهیهٔ برآوردهای معتبر و قابل اعتماد از پارامترهای جامعه‌ای از قبیل میانگینها، مجموعها، و نسبتها طراحی می‌کنند، ولی بیش از پیش به ارزشیابی روابط درونی میان متغیرها نیز علاقه‌مندند. بسیاری از اوقات از نرم‌افزارهای آماری که در سطح وسیعی در دسترس قرار دارند برای تحلیل این قبیل داده‌ها استفاده شده است، به خصوص هنگامی که در تحلیل، مدلسازی چندمتغیره موردنظر بوده است. متأسفانه، این‌گونه نرم‌افزارها به طور کلی بسیاری از ویژگیهای طرح آمارگیری خاصی را که مورد تحلیل قرار دارد در نظر نمی‌گیرند. در این کتاب، بر استفاده از SUDAAN و STATA در تحلیل داده‌های آمارگیری تمرکز شده است، زیرا این نرم‌افزارها برخلاف بسیاری از نرم‌افزارهای آماری دیگری که فعلاً در دسترس‌اند از این توانایی برخوردارند که داده‌های آمارگیری را طوری تحلیل کنند که ویژگیهای مربوط به طرح نیز در نظر گرفته شوند. همان طور که قبلاً اشاره کردیم، استفادهٔ ما از این دو محصول بازتاب آشنایی خود ما با آنهاست و نباید به عنوان تأیید آنها در مقابل سایر نرم‌افزارهایی که با تواناییهای مشابه فعلاً در دسترس قرار دارند یا آنهایی که ممکن است در آینده در دسترس باشند تعبیر شود. راهبردهایی که در این فصل و در سراسر این کتاب به اجمال بیان شده‌اند بر نظریهٔ نمونه‌گیری متکی‌اند و باید مستقل از و قابل تبدیل به هر

محصول نرم‌افزاری باشند که مدعی اجرای تحلیلی است که *بر طرح* داده‌های آمارگیری نمونه‌ایی مبتنی است که به صورت زیر تعریف شده است.

تحلیل مبتنی بر طرح داده‌های آمارگیری نمونه‌ای، تحلیلی است که ویژگیهای ملازم با طرح نمونه‌ای را در نظر می‌گیرد. غالباً این ویژگیها از جمله شامل طبقه‌بندی، خوشه‌بندی، کسرهای نمونه‌گیری به اندازهٔ کافی بزرگ برای بهره‌گیری از کاربرد تصحیح جامعهٔ متناهی، و وزنهای نمونه‌گیری هستند. تحلیلی که مبتنی بر طرح نباشد، یعنی ویژگیهای طرح را، از قبیل آنچه در بالا فهرست شد، در نظر نگیرد به عنوان تحلیل *مبتنی بر مدل* نامیده خواهد شد. کاربرد ما از این اصطلاح اخیر وسیعتر از آن است که در نوشتگان به کار می‌رود زیرا که تحت تحلیل مبتنی بر مدل، تحلیلی را می‌گنجانیم که تحت هیچ مدل صریحاً بیان شده‌ای انجام نشده است، بلکه به نظر می‌رسد نمونه‌گیری تصادفی مستقل از یک جامعهٔ نامتناهی را فرض می‌گیرد (فرضهای «کلاسیک»).

در بخشهای قبل نشان دادیم که هرگاه ویژگیهای طرح آمارگیری در تحلیل نادیده گرفته شود، این امکان وجود دارد که نتایج نادرست باشند. در این فصل نشان خواهیم داد که چگونه ویژگیهای مربوط به طرح شناسایی می‌شوند، و سپس به صورتی مناسب در تحلیل داده‌های آمارگیری نمونه‌ای منظور می‌شوند. همچنین نشان خواهیم داد که چگونه غفلت از منظور کردن این ویژگیها می‌تواند به برآوردهای نادرست و استنباطهای نامعتبر منجر شود.

۱.۱۶ مراحل مورد نیاز برای اجرای تحلیل مبتنی بر طرح

گامهای زیر برای اجرای تحلیل مبتنی بر طرح مورد نیازند.

۱. شناسایی عناصر زیر در طرح نمونه‌ای:
   طبقه‌بندی
   خوشه‌بندی متغیرهای مورد استفاده
   اندازه‌های جامعهٔ مورد نیاز برای تعیین تصحیح جامعهٔ متناهی
۲. تعیین وزن برای هر آزمودنی نمونه براساس اطلاعات بالا.
۳. تعیین یک وزن نمونه‌گیری نهایی برای هر سابقهٔ نمونه که هرگونه تعدیل را که برای بی‌پاسخی و پس‌طبقه‌بندی مورد نیاز است در نظر بگیرد.
٤. حصول اطمینان از اینکه همهٔ داده‌های مربوط به طبقه‌بندی، خوشه‌بندی، و اندازهٔ جامعه که برای یک تحلیل مناسب مبتنی بر طرح مورد نیازند، برای هر سابقهٔ نمونه شناسایی شده‌اند.

۵. تعییـن شیوه و مجموعهٔ فرمانها به منظور اجرای تحلیل مورد نیاز برای بستهٔ نرم‌افزاری خاصی که قرار است به کار رود.
٦. اجرای تحلیل و تفسیر دقیق یافته‌ها

مثال تشریحی: ایالتی را در نظر می‌گیریم که شامل ۵ ناحیه و دارای ۵۷ آسایشگاه سالمندان به صورتی است که در جدول ۱.۱٦ نشان داده شده‌اند. فرض کنید یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از دو آسایشگاه سالمندان از هر ناحیه می‌گیریم و از هر آسایشگاه سالمندان نمونه‌ای متشکل از پنج پذیرش را انتخاب می‌کنیم تا کل تعداد بیمارانی که در طی سال ۱۹۹۷ در آسایشگاههای سالمندان این پنج ناحیه پذیرش شده‌اند و منبع تأمین هزینهٔ آنها بیمهٔ کمکهای پزشکی بوده است تعیین شود. داده‌های حاصل در جدول ۲.۱٦ ارائه شده و متغیرهایی که در این جدول آمده‌اند به صورت زیر توصیف شده‌اند:

REGION ناحیه را نشان می‌دهد
NURSHOME آسایشگاه سالمندان نمونه را نشان می‌دهد
PATIENT بیمار نمونه را نشان می‌دهد
MEDICAID در صورتی که بیمار به هزینهٔ بیمهٔ کمکهای پزشکی بستری شده باشد برابر با «۱» و در غیر اینصورت برابر با «۰» است
RGNHOMES کل تعداد آسایشگاههای سالمندان در ناحیه است
NHADMISS کل تعداد پذیرشها در آسایشگاه سالمندان طی سال ۱۹۹۷ است

جدول ۱.۱٦ تعداد آسایشگاههای سالمندان در پنج ناحیهٔ یک ایالت

ناحیه	تعداد آسایشگاههای سالمندان در ناحیه	کل پذیرشها در آسایشگاههای سالمندان ناحیه طی سال ۱۹۹۷
۱	۱۲	۱۲٤٥
۲	۲۰	۳۸۸۷
۳	۱۱	٥۸۳
٤	۸	٤۰۰
٥	٦	۲۳۷٦

گامهای شش‌گانهٔ فهرست شدهٔ بالا را مرور خواهیم کرد تا برآوردی مبتنی بر طرح برای کل تعداد بیماران پذیرش شده در آسایشگاههای سالمندان طی سال ۱۹۹۷ به دست آوریم که منبع تأمین هزینهٔ آنها بیمهٔ کمکهای پزشکی بوده است.

جدول ۲.۱٦ داده‌های حاصل از نمونهٔ آسایشگاههای سالمندان

REGION	NURSHOME	PATIENT	MEDICAID	RGNHOMES	NHADMISS
1	1	1	1	12	123
1	1	2	1	12	123
1	1	3	1	12	123
1	1	4	0	12	123
1	1	5	1	12	123
1	2	1	0	12	89
1	2	2	0	12	89
1	2	3	1	12	89
1	2	4	0	12	89
1	2	5	0	12	89
2	1	1	1	20	231
2	1	2	0	20	231
2	1	3	1	20	231
2	1	4	0	20	231
2	1	5	1	20	231
2	2	1	0	20	187
2	2	2	0	20	187
2	2	3	0	20	187
2	2	4	1	20	187
2	2	5	0	20	187
3	1	1	1	11	43
3	1	2	1	11	43
3	1	3	1	11	43
3	1	4	0	11	43
3	1	5	1	11	43
3	2	1	1	11	49
3	2	2	1	11	49
3	2	3	1	11	49
3	2	4	1	11	49
3	2	5	0	11	49
4	1	1	0	8	56
4	1	2	1	8	56
4	1	3	1	8	56
4	1	4	0	8	56
4	1	5	0	8	56
4	2	1	0	8	38
4	2	2	0	8	38
4	2	3	0	8	38
4	2	4	0	8	38
4	2	5	1	8	38
5	1	1	1	6	359
5	1	2	0	6	359
5	1	3	1	6	359
5	1	4	1	6	359
5	1	5	0	6	359
5	2	1	0	6	460
5	2	2	1	6	460
5	2	3	0	6	460
5	2	4	1	6	460
5	2	5	0	6	460

۱. این یک نمونهٔ خوشه‌ای طبقه‌بندی شدهٔ دو مرحله‌ای است که در مرحلهٔ اول آن، آسایشگاههای سالمندان با نمونه‌گیری تصادفی سادهٔ بدون جایگذاری در داخل هر طبقه انتخاب می‌شوند و بیماران با نمونه‌گیری تصادفی ساده در داخل هر آسایشگاه انتخاب شدهٔ سالمندان برگزیده می‌شوند. متغیر طبقه‌بندی، REGION است، متغیر خوشه‌بندی NURSHOME است، متغیر RGNHOMES تعداد آسایشگاههای سالمندان در داخل هر ناحیه است و متغیری است که برای محاسبهٔ تصحیح جامعهٔ متناهی برای مرحلهٔ اول نمونه‌گیری به کار می‌رود، متغیر NHADMISS کل تعداد بیماران پذیرش شده طی سال ۱۹۹۷ در هر یک از آسایشگاههای سالمندان نمونه است و به عنوان متغیر مورد استفاده برای محاسبهٔ تصحیح جامعهٔ متناهی در مرحلهٔ دوم نمونه‌گیری به کار می‌رود.

۲. وزن نمونه‌گیری برای هر سابقهٔ نمونه به صورت زیر به دست می‌آید:

فرض می‌کنیم M_h = تعداد آسایشگاههای سالمندان در ناحیهٔ h باشد.

پس، $f_{۱h}$، احتمال انتخاب شدن هر آسایشگاه سالمندان در نمونه، از فرمول زیر به دست می‌آید

$$f_{۱h} = \frac{۲}{M_h}$$

قرار می‌دهیم، N_{hi} = تعداد پذیرشها در آسایشگاه سالمندان i در داخل طبقهٔ h برای سال ۱۹۹۷.

پس کسر نمونه‌گیری مرحلهٔ دوم، $f_{۲hi}$، از فرمول زیر به دست می‌آید

$$f_{۲hi} = \frac{۵}{N_{hi}}$$

به این ترتیب، احتمال کلی گنجانیده شدن یک بیمار در نمونه از فرمول زیر به دست می‌آید

$$f_{hi} = \frac{۲}{M_h} \times \frac{۵}{N_{hi}}$$

و وزن نمونه‌گیری، w_{hi}، از فرمول زیر محاسبه می‌شود

$$w_{hi} = \frac{M_h N_{hi}}{۱۰} \qquad (۱.۱۶)$$

۳. در این مثال تشریحی، فرض بر این است که پاسخگویی کامل است و هیچ برنامه‌ای برای پس‌طبقه‌بندی وجود ندارد.

٤. بررسی جدول ۲.۱٦ نشان می‌دهد که متغیرهای مربوط به طبقه‌بندی، خوشه‌بندی، و تصحیح جامعهٔ متناهی برای هر سابقهٔ نمونه موجود است. اگرچه در جدول ۲.۱٦ صریحاً نشان داده نشده است ولی برای هر سابقهٔ نمونه یک متغیر، *WEIGHT* ایجاد کرده‌ایم که از معادلهٔ (۱.۱٦) به دست آمده است.

٥. برای این مثال تشریحی از SUDAAN استفاده خواهیم کرد زیرا STATA برای طرحهای نمونه‌ای خوشه‌ای از تصحیح جامعهٔ متناهی استفاده نمی‌کند.
فرمانهای مناسب برای SUDAAN در زیر نشان داده شده‌اند:

```
PROC DESCRIPT DATA = "A:CH16ILLI" FILETYPE = SAS DESIGN = WOR
DEFF TOTALS;
NEST REGION NURSHOME;
TOTCNT RGNHOMES NHADMISS;
VAR MEDICAID;
WEIGHT WEIGHT;
SETENV COLWIDTH = 15;
SETENV DECWIDTH = 4;
```

اولین سطر فرمان، شیوهٔ مناسب SUDAAN را برای برآورد کردن مجموع نشان می‌دهد، مجموعهٔ داده‌های مناسب را مکان‌یابی می‌کند، نشان می‌دهد که پروندهٔ اطلاعاتی مربوط به SAS است، طرح نمونه‌ای مناسب را تعیین می‌کند (هنگامی که تصحیح جامعهٔ متناهی مورد نیاز باشد طرح بدون جایگذاری WOR لازم است).

سطر دوم فرمان یا گزارهٔ "nest " به معنای تودرتو، متغیرهای طبقه‌بندی و خوشه‌بندی را نشان می‌دهد (اگر توضیح دیگری داده نشده باشد، اولین متغیر نشان داده شده در گزارهٔ تودرتو به صورت متغیر طبقه‌بندی تعبیر می‌شود).

سطر سوم یا گزارهٔ "totcnt " متغیرهای جامعه را که برای محاسبهٔ تصحیح جامعهٔ متناهی در هر مرحله از نمونه‌گیری مورد نیازند نشان می‌دهد.

سطر چهارم فرمان، متغیری (متغیرهایی) را مشخص می‌کند که قرار است برآورد (برآوردهایی) برای آن (آنها) تهیه شود.

سطر پنجم، متغیر وزن نمونه‌گیری را نشان می‌دهد.
دو سطر آخر، مشخصات خروجی را ارائه می‌دهند.

٦. خروجی حاصل در زیر نشان داده شده است:

Number of observations read: 50 Weighted count: 8791
Deonominator degrees of freedom: 5

Variance Estimation Method: Taylor Series (WOR)
by: Variable, One.

Variable		One - -
MEDICAID	Sample Size	50.0000
	Weighted Size	8791.0000
	Total	4180.2000
	SE Total	1112.7529
	Mean	0.4755
	SE Mean	0.1052
	DEFF Mean # 4	2.1748
	DEFF Total # 4	3.1479

یافته‌هایی که در بالا نشان داده شده‌اند، برآورد می‌کنند که در طی سال ۱۹۹۷ تعداد ٤۱۸۰ بیمار در آسایشگاههای سالمندان این ایالت پذیرش شده‌اند که هزینه‌های آنها توسط بیمهٔ کمکهای پزشکی پرداخت می‌شده است. خطای معیار این برآورد مجموع ۱۱۱۳ و اثر طرح آن برای مجموع ۳/۱۵ است که دلالت بر آن دارد که برای این متغیر مقدار بسیار زیادی خوشه‌بندی انجام شده است.

□

۲.۱۶ مسایل تحلیل آمارگیریهای نمونه‌ای «نوعی»

همهٔ آمارگیریهای نمونه‌ای مشتمل بر برخی ویژگیهای طرح‌اند ولی الزاماً همهٔ آنهایی را که در بالا فهرست شدند شامل نیستند، و بعضی دارای طرحهایی به مراتب پیچیده‌ترند. در اینجا سودمند است که یک آمارگیری نمونه‌ای «نوعی» توصیف شود. نوع طرحی که اینجا به تصویر خواهیم کشید از نوع اکثر وضعیتهای آمارگیریی است که در زندگی واقعی روی می‌دهند. البته، وضعیتهایی پیچیده‌تر هم وجود دارند به خصوص در برنامه‌های آمارگیری بزرگی که به صورت جاری در سطح ملتها نگهداری می‌شوند. ولی، غالب اوقات حتی این آمارگیریها هم می‌توانند برای هدفهای تحلیل ساده شوند تا با وضعیت نوعیی که در اینجا شرح خواهیم داد شباهت پیدا کنند و نتایج آنها غالباً به آنچه نزدیک خواهند بود که اگر پارامترهای طرح آمارگیری کل توصیف می‌شدند صحیح می‌بودند.

حالا به بررسی یک آمارگیری می‌پردازیم که طرح آن بسیار شبیه به طرح آمارگیری توصیف شده در بخش ۱.۱٦ است، ولی با اندازهٔ نمونه‌ای بزرگتر، با برخی پیچیدگیهایی که قبلاً بررسی نشده‌اند، و با دستور کار تحلیل به مراتب پیچیده‌تر. به صورتی مشخصتر، طرح این آمارگیری به گونه‌ای است که همهٔ عناصر جامعه در یکی از L طبقهٔ دوبه‌دو ناسازگار و فراگیر گروه‌بندی شده‌اند. در داخل طبقهٔ

h ام، جامعه به M_h خوشه [به نام واحدهای نمونه‌گیری اولیه (PSUها)] تقسیم و از میان آنها m_h خوشه برای گنجاندن در نمونه انتخاب شده‌اند. در داخل هر خوشهٔ نمونه، مرحلهٔ دوم نمونه‌گیری اجرا می‌شود و n_{hi} واحد شمارش انتخاب می‌شوند. به هر واحدی که از یک واحد نمونه‌گیری اولیه انتخاب می‌شود یک *وزن نمونه‌گیری* به صورتی که قبلاً گفته شد داده می‌شود. این وزن را می‌توان به صورت تعداد واحدهایی در جامعه در نظر گرفت که این واحد معرف آن است و در این مثال با طبقه و واحد نمونه‌گیری اولیه‌ای که واحد مزبور در آن قرار دارد تعیین می‌شود. وزن نمونه‌گیری مجدداً به صورت معکوس احتمال گنجانیده شدن واحد در نمونه محاسبه می‌شود. هنگامی که اندازه‌های جامعه معلوم باشد این وزنها را می‌توان به صورت *تصحیح جامعهٔ متناهی* در تحلیل وارد کرد.

مثال تشریحی: برای نشان دادن این مفاهیم، تحلیل داده‌های جمع‌آوری شده به عنوان بخشی از بررسی PAQUID[1] را در نظر می‌گیریم که یک نمونهٔ خوشه‌ای طبقه‌بندی شده از شهروندان سالخوردهٔ فرانسوی است [۱] و [۲]. ساکنانی که در تاریخ ۳۱ دسامبر ۱۹۸۷ دارای ٦٥ سال سن یا بیشتر بوده و در آسایشگاهها به سر نمی‌بردند و در دو ناحیه یا «استان» در جنوب غربی فرانسه: ژیروند (با ۱۰۰۰۰ کیلومتر مربع وسعت و ۱۱۲۷٥٤٦ نفر جمعیت) و دوردونْی (با ۹۰٦۰ کیلومتر مربع مساحت و ۳۷۷۳٥٦ نفر جمعیت) زندگی می‌کردند در این بررسی گنجانیده شده بودند. در مجموع ۳۷۷۷ نفر از ساکنان محلی به منظور شناسایی عوامل پایه‌ای و طول عمر که ممکن بود با اختلال شناخت، زوال عقل، و بیماری آلزایمر در ارتباط باشند انتخاب شدند. داده‌های پایه‌ای در سالهای ۱۹۸۹–۱۹۸۸ جمع‌آوری شدند و آزمودنیها از آن پس تاکنون به صورت ادواری تحت بررسی بوده‌اند.

استانهای ژیروند و دوردونْی به ترتیب از ٥٤۳ و ٥٥٥ ناحیهٔ کوچکتر تشکیل شده‌اند که *بخش* نامیده می‌شوند. برای انتخاب آزمودنیهای بررسی، بخشهای داخل هر استان براساس اندازهٔ جمعیت در یکی از چهار گروه زیر جای داده شدند:

گروه ۱ : بخشهای دارای ٥۰۰۰۰ نفر جمعیت و بیشتر
گروه ۲ : بخشهای دارای ۱۰۰۰۰ تا ٤۹۹۹۹ نفر جمعیت
گروه ۳ : بخشهای دارای ۲۰۰۰ تا ۹۹۹۹ نفر جمعیت
گروه ٤ : بخشهای دارای کمتر از ۲۰۰۰ نفر جمعیت

در داخل هر گروه، بخشهایی با احتمال متناسب با اندازهٔ بخش انتخاب شدند. در استان دوردونْی هیچ بخشی با بیش از ٥۰۰۰۰ نفر جمعیت وجود نداشت. در مجموع، ۳۷ بخش از ژیروند و ۳۸ بخش از دوردونْی انتخاب شدند.

[1] Personnes Agees Quid

در داخـل هـر بخش انتخاب شده، آزمودنیها به طور تصادفی از روی فهرستهای انتخاباتی هر بخش انـتخاب و برحسـب سـن و جـنس طـبقه‌بـندی شـدند. از آزمودنیهای انتخاب شده درخواست شد تا موافقـت کتـبی خـود را بـرای مشارکت در بررسی ارائه کنند. نتیجه آن شد که کل جامعهٔ مورد بررسی شامل ۳۷۷۷ آزمودنی: ۲۷۹۲ نفر از ژیروند و ۹۸۵ نفر از دوردونْی به دست آمد.

بـا همـهٔ آزمودنیهـای ایـن بررسـی از طریق تلفن یا (در صورت امکان) پست برای کسب اجازه به مـنظور مشـارکت دادن آنهـا در بررسـی تماس گرفته شد. با کسانی که موافقت خود را برای مشارکت اعـلام کـرده بودنـد (٦٨/٥٪) در منزلشـان توسـط یـک روانشـناس آموزش دیده برای به دست آوردن داده‌هـای پایـه‌ای مصاحبه شد. یک پرسشنامهٔ دارای ساختار به هر شرکت کننده اختصاص داده شد. به آزمودنیهـا در فاصـله‌هـای زمـانی مـنظم مـراجعه مـی‌شـد تـا با توجه به اختلال شناخت و زوال عقل غربالگری شوند.

واضـح است که این یک نمونهٔ تصادفی ساده نیست، بلکه، در عوض، یک آمارگیری نمونه‌ای است کـه بسـیاری از ویژگیهایی را که قبلاً به اجمال مورد بررسی قرار گرفته‌اند با خود دارد. در بحثی که در پی خواهد آمد نمادهای زیر به کار گرفته خواهند شد:

- h به یکی از ۷ طبقه اشاره دارد (که برحسب استان و گروه تعریف شده است)
- M_h تعداد خوشه‌ها (یا بخشها)یی است که طبقهٔ h ام را تشکیل می‌دهند، $h = ۱,...,۷$
- m_h تعداد خوشه‌های انتخاب شده از طبقهٔ h ام است، $h = ۱,...,۷$
- i به تعداد بخش در داخل طبقه اشاره می‌کند
- N_{hi} تعـداد افـراد ٦٥ سـال بـه بالاست که در i امین بخش نمونه‌گیری شده در داخل طبقهٔ h (براساس داده‌های سرشماری) زندگی می‌کنند
- N_h تعـداد افـراد ٦٥ سـال بـه بالاسـت که در طبقهٔ h (براساس داده‌های سرشماری) زندگی می‌کنند
- n_{hi} تعداد افرادی است که عملاً در بخش i ، داخل طبقهٔ h بررسی شده‌اند.

جدول ۳.۱٦ خلاصه‌ای از فرایند انتخاب را نشان می‌دهد.

می‌توان چنین تصور کرد که ۷ طبقه داریم و نمونه‌ها در داخل هر طبقه انتخاب می‌شوند. در ژیروند فقـط سـه بخش با بیش از ۵۰۰۰۰ نفر جمعیت وجود داشت (طبقهٔ ۱) و هر یک از این بخشها انتخاب شـدند. در دوردونْی فقـط ۲ بخش با جمعیتی بین ۱۰۰۰۰ و ٤۹۹۹۹ نفر وجود داشت (طبقهٔ ۳) و هر دو انـتخاب شـدند. در مـورد ایـن دو طـبقه، چـون $m_h = M_h$، احـتمال انـتخاب یـک واحـد خـاص (n_{hi}/N_{hi}) اسـت، بـنابراین، وزن نمونـه‌گیـری برای این دو طبقه $N_{hi}/n_{hi} = ۱/(n_{hi}/N_{hi})$ خواهد بود.

بـرای پـنج طـبقهٔ دیگـر، احـتمال تقریـبی انـتخاب یک واحد خاص برابر است با حاصلضرب احتمال تقریـبی انـتخاب بخـش مـورد نظـر بـا اسـتفاده از احـتمال متناسب با اندازهٔ (PPS) طرح نمونه‌گیری، $m_h \times (N_{hi}/N_h)$ (اثـبات این معادلـــه را در پایان همین فصل ببینید)، ضربدر احتمال انتخاب فرد در بخش انتخاب شده، n_{hi}/N_{hi}. یعنی احتمال انتخاب یک نفر در این پنج طبقه به صورت زیر است

$$\left(m_h \times \frac{N_{hi}}{N_h}\right) \times \left(\frac{n_{hi}}{N_{hi}}\right) = \frac{m_h n_{hi}}{N_h}$$

نتیجه می‌شود که وزن نمونه‌گیری در این طبقه‌ها $(N_h/(m_h n_{hi}))$ است.

جدول ۳.۱٦ انتخاب آزمودنیهای نمونه از استانهای ژیروند و دوردونْی

استان	گروه	طبقه	M_h	m_h	$\sum_{i=1}^{m_h} N_{hi}$	$\sum_{i=1}^{m_h} n_{hi}$
ژیروند	۱	۱	۳	۳	٤۹۷۸٦	۱۲۳۱
ژیروند	۲	۲	۱٤	٤	۳۵۲۵۳	٤۷۳
دوردونْی	۲	۳	۲	۲	۱۱۷٦٤	۷۱
ژیروند	۳	٤	۷۳	۱۰	۳۷۰۲٤	۳٦۱
دوردونْی	۳	۵	۲٦	۱۳	۱۷٦۸۷	۳۱۱
ژیروند	٤	٦	٤۵۳	۲۰	٤٤۹۸۰	۷۲۷
دوردونْی	٤	۷	۵۲۷	۲۳	٤٦۲۵۰	٦۰۳
مجموع			۱۰۹۸	۷۵	۲٤۲۷٤٤	۳۷۷۷

اگـرچه ایـن وزنها، همان طور که موردنظر است، به این مفهوم صحیح‌اند که تعداد افراد موجود در جامعـه‌ای را بـه تصـویر مـی‌کشـند که هر آزمودنی موردنظر در بررسی، معرف آن است، ولی مجموع وزنهـای داخـل زیرگـروههای سـنی ـ جنسی مشخص شده، ممکن است درست به مجموع اندازه‌های جامعـه‌ای معـلوم در داخل این زیرگروهها نرسد. برای اصلاح این وضعیت، یک تعدیل پس طبقه‌بندی لازم اسـت تـا وزن اولیه را به گونه‌ای اصلاح کند که مجموع وزنها برای همهٔ افراد نمونه‌گیری شده در هر یک از زیرگروههای سنی ـ جنسی با اندازه‌های جامعه‌ای معلوم در داخل هر یک از این زیرگروهها دقیقاً یکسان شود. زیرگروهها عبارت‌اند از:

مردان در سنین ٦۵ تا ۷٤ سال
زنان در سنین ٦۵ تا ۷٤ سال
مردان در سنین ۷۵ تا ۸٤ سال
زنان در سنین ۷۵ تا ۸٤ سال
مردان در سنین ۸۵ سال و بیشتر
زنان در سنین ۸۵ سال و بیشتر

مـثلاً در طـبقهٔ ۱ ژیـروند ۱۰۱۰۵ مرد ٦٥ تا ۷٤ سال در فهرستهای انتخاباتی وجود داشتند. این رقم نشان دهندهٔ ۲۰/۳٪ همهٔ آزمودنیهای موجود در فهرست انتخاباتی آن طبقه بود. ولی در نمونه، ۲۸٤ مرد بیـن سـنین ٦٥ تـا ۷٤ سال بودند که معرف ۲۳/۰۷٪ نمونهٔ مربوط به آن طبقه‌اند. بنابراین، می‌توانیم هر وزن را با ضرب کردن آن در نسبت $\frac{۰/۲۰۳}{۰/۲۳۰۷} = ۰/۸۸$ تعدیل کنیم. اگر این عاملهای تعدیل را c_{hijk} بنامیم کـه در آن h نشـانگر طـبقه، i نشانگر واحد نمونه‌گیری اولیه، j نشانگر زیرگروه، و k نشانگر فرد باشد، در آن صورت می‌توان وزنهای آماری نهایی، w_{hijk}، را به صورت زیر بیان کرد

$$w_{hijk} = \left(\frac{N_h}{m_h n_{hi}} \right) \times c_{hijk}$$

□

بحـث بـالا نشـان مـی‌دهـد کـه بررسـی PAQUID را مـی‌توان به عنوان یک طرح طبقه‌بندی شدهٔ چـندمرحله‌ای توصـیف کـرد کـه در مرحلهٔ اول نمونه‌گیری، بخشها (یا واحدهای نمونه‌گیری اولیه) از طـبقه‌هـا انـتخاب مـی‌شـوند و در مرحـلهٔ دوم، آزمودنیها از روی فهرستهای انتخاباتی در داخل بخش منـتخب بـه طـور تصـادفی انـتخاب می‌شوند. همان طور که در فصلهای قبلی بیان شد، تا همین اواخر تحـلیل داده‌هـای آمارگیـری از این نوع نسبتاً مشکل بود زیرا بیشتر نرم‌افزارهای آماری فرض را بر این مـی گیـرند کـه داده‌هـا بـا نمونه‌گیری تصادفی ساده انتخاب شده‌اند. این فرض می‌تواند هنگام تحلیل داده‌هـای آمارگیـری نمونـه‌ای کـه از طـرح نمونـه‌گیـری تصـادفی ساده به دست نیامده‌اند باعث بروز خطاهـایی در بـرآورد کـردن و اسـتنباط شود. این قبیل برنامه‌ها ممکن است برآوردهای نقطه‌ای اریب تولید کنند و ممکن است خطاهای معیار این برآوردها را بسیار کم برآورد کنند. همان طور که بارها در فصـلهای پیشین این کتاب ذکر شد، برنامه‌هایی همچون SUDAAN و STATA امکان تحلیل داده‌های آمارگیری نمونه‌ای را به شکلی مناسب فراهم می‌سازند.

بـرای تحـلیل این داده‌ها از قابلیتهای آمارگیری نمونه‌ای STATA استفاده می‌کنیم. برای این منظور، بـه سـادگی مشـخص مـی‌کـنیم که هفت طبقه وجود دارند و بخشها واحدهای نمونه‌گیری اولیه را در داخـل آنها تشکیل می‌دهند. همچنین وزنهای آماری، w_{hijk}، را تهیه می‌کنیم که با پیروی از فرایندی که در بالا توصیف شد محاسبه می‌شوند.

پارامـترهای طرح آمارگیری برای STATA با استفاده از فرمانهای *svyset* مشخص می‌شوند. در این مـثال، مـتغیرهای شامل اطلاعات مربوط به طبقه، وزن نمونه‌گیری، و و احد نمونه‌گیری اولیه به ترتیب STRATUM، WEIGHT و PSU نـامگذاری شـده‌انـد. این مشخصات به صورت بخشی از مجموعه داده‌ها ذخیره شده‌اند و هر بار که یک تحلیل آمارگیری اجرا شود فراخوانده می‌شوند:

```
. svyset strata STRATUM
. svyset pweight WEIGHT
. svyset psu PSU
```

هدف از این تحلیل آن است که وقوع زوال عقل در میان آن دسته از افرادی که در زمان آزمایش آغازین آمارگیری زوال عقل *نداشته‌اند* در طی سه سال برآورد شود. به این دلیل فقط آن دسته از آزمودنیها را که در زمان مشاهدهٔ آغازین دچار زوال عقل نبوده‌اند $(n = ۲۲۷۳)$ در نظر می‌گیریم. به عنوان اولین تحلیل، محاسبهٔ نسبت این آزمودنیها را که نمرهٔ آزمایش کوچک وضعیت ذهنی MMSE[1] آنان برابر با ۲۱ یا کمتر بوده است در نظر می‌گیریم. آزمایش کوچک وضعیت ذهنی آزمایشی کوتاه است که اختلال شناختی را غربالگری می‌کند و نمرهٔ آن از صفر تا ۳۰ است که نمره‌های پایینتر نشانهٔ احتمال بیشتر وقوع زوال عقل است. با در نظر گرفتن متغیر mmse01 به عنوان نشانگر نمرهٔ ۲۱ یا کمتر برای آزمایش کوچک وضعیت ذهنی، می‌بینیم که نادیده گرفتن طرح آمارگیری به برآورد زیر برای نسبت و بازهٔ اطمینان ملازم با آن منتهی می‌شود:

Variable	obs	Mean	Std. Err.	[95% Conf.	Interval]
Mmse01	2242	.0816236	.0057836	.0702818	.0929653

برای منظور کردن پارامترهای طرح آمارگیری، فرمان زیر را برای STATA صادر می‌کنیم:

```
. svymean mmse01
```

و نتایج زیر را به دست می‌آوریم:

```
Survey mean estimation
pweight:    finalwt          Number of obs    =      2242
Strata:     newstrat         Number of strata =         7
PSU:        parish           Number of PSUs   =        75
                             Population size  =  14863.41
```

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
mmse01	. 082957	.0079147	.0671634	.0987506	1.845315

توجه داشته باشید که در حالی که برآوردهای نقطه‌ای کاملاً قابل مقایسه‌اند (تقریباً ۸٪ دارای نمرهٔ ۲۱ یا کمتر برای آزمایش کوچک وضعیت ذهنی)، خطای معیار نسبت برآورد شده براساس طرح آمارگیری، ۰/۰۰۷۹، در مقایسه با خطای معیار برآورد شدهٔ ۰/۰۰۵۸ حاصل از نادیده گرفتن پارامترهای

[1] Mini – Mental State Examination

طـرح آمارگیـری اسـت. *اثـر طـرح* (کـه در خروجی STATA به صورت" Deff " نشان داده می‌شود) نسـبت واریـانس پارامـتر بـرآورد شده با در نظر گرفتن پارامترهای طرح آمارگیری به واریانس پارامتر برآورد شده با فرض نمونه‌گیری تصادفی ساده است. در این حالت

$$\text{Deff} = \frac{۰/۰۰۷۹^{۲}}{۰/۰۰۵۸^{۲}} = ۱/۸۵$$

با اینکه برآوردهای نقطه‌ای ممکن است در دو طرح تحلیل قابل مقایسه باشند، خطاهای معیار میانگینها در تحـلیل مبتـنی بـر طـرح بـه مـراتب بیشترند. این همان چیزی است که در مورد برآوردهای نسبت نمونـه‌ای انـتظار می‌رفت. واضح است که نادیده گرفتن پارامترهای طرح آمارگیری به کم برآورد کردن واریانس نمونه‌ای تا میزان ۸۵٪ و به دست آوردن بازه‌های اطمینان بیش از حد باریک، منتهی می‌شود.

نوعاً، اثرهای طرح برای برآورد میانگینها، مجموعها، و نسبتها به مراتب بیشتر از یک است. اثر طرح مـی‌توانـد بـه عـنوان مقیـاس تورم در واریانس تصور شود که ناشی از همگنی درون خوشه‌هاست. اثر طـرح را مـی‌تـوان به صورت $۱+\delta_y(\overline{n}-۱)$ بیان کرد که در آن، δ_y، *ضریب همبستگی درون خوشه‌ای* (یعـنی *ضـریب همبسـتگی درون رده‌ای*)، و $\overline{n}$ متوسط تعداد واحدهای نمونه‌گیری شده به ازای خوشه اسـت. همـان طـور کـه قـبلاً بیان شد، دامنهٔ ضریب همبستگی درون خوشه‌ای می‌تواند از مقادیر منفی کوچـک (هـنگامی کـه داده‌هـای داخل خوشه‌ها به شدت ناهمگن باشند) تا یک (هنگامی که داده‌های داخل خوشه‌ها بسیار همگن باشند) کشیده شود. فقط در موارد بسیار نادر که داده‌ها در داخل خوشه‌ها بسیار ناهمگن باشند اثر طرح کمتر از یک خواهد بود.

یـک هـدف مهم بررسی مزبور این بود که بررسی شود که آیا بین مصرف شراب و وقوع زوال عقل رابطـه‌ای وجـود دارد یـا نه، و اگر وجود دارد برآورد قدرت این رابطه چقدر است. به طور مشخصتر، بـرآمد مـورد توجـه در ایـن بررسی این بود که آیا آزمودنی مزبور در مدت زمان پیگیری ۳ ساله دچار زوال عقـل بـوده اسـت یا نه و عامل مخاطرهٔ اصلی موردنظر مصرف شراب با گروه‌بندی در رسته‌های زیر بوده است یا نه :

شراب =
- ۰ بدون مصرف شراب
- ۱ مصرف شراب تا $\frac{۱}{۴}$ لیتر در روز
- ۲ مصرف شراب بیش از $\frac{۱}{۴}$ لیتر در روز

در این تحلیل، متغیر وابسته، داشتن یا نداشتن زوال عقل است که بر تعداد زیادی از آزمایشـهای عصبی ـ روانی متکی است که به مراتب از آزمایش کوچک وضعیت ذهنی پیچیده‌ترند. جدول‌بندی

متقاطع وضعیت زوال عقل برای دورهٔ سه سالهٔ پیگیری (که براساس این تعداد کثیر از آزمایشها تعیین می‌شود) در مقابل مصرف شراب بدون توجه به طرح آمارگیری در جدول ٤.١٦ نشان داده شده است.

جدول ٥.١٦ نشان می‌دهد که همان نسبتهای بخت و بازه‌های اطمینان را می‌توان با استفاده از مدل رگرسیونی لوژستیک تحت فرض نمونه‌گیری تصادفی ساده به دست آورد.

جدول ٤.١٦ پیوند مصرف شراب با وقوع زوال عقل (تحلیل مبتنی بر مدل)

مصرف شراب	وقوع زوال عقل				نسبتهای بخت (OR)*	بازه‌های اطمینان ٩٥ درصد
	آری		نه			
	تعداد	درصد	تعداد	درصد		
هیچ	٤٨	٤/٩	٩٢٣	٩٥/١	١/٠٠٠	
روزی $\frac{١}{٤}$ لیتر و کمتر	٤٧	٥/١	٨٧٥	٩٤/٩	١/٠٣٣	٠/٦٨٤–١/٥٦١
بیشتر از $\frac{١}{٤}$ لیتر در روز	٤	١/١	٣٧٦	٩٨/٩	٠/٢٠٥	٠/٠٧٣–٠/٥٧١
مجموع	٩٩		٢١٧٤			

*در مقایسه با هیچ به عنوان گروه مرجع.

با استفاده از ویژگیهای طرح آمارگیری، یک رگرسیون لوژستیک اجرا می‌شود که DEMENTIA (زوال عقل) متغیر وابسته و دو متغیر نشانگر برای مصرف شراب (WINE_1 و WINE_2) متغیرهای مستقل آن هستند. این کار در نرم‌افزار STATA با فرمان زیر اجرا می‌شود:

```
. svylogit DEMENTIA WINE_1 WINE_2, or
```

از این فرمان خروجی زیر به دست می‌آید:

```
Survey logistic regression

pweight:  WEIGHT                        Number of obs    =      2273
Strata:   STRATUM                       Number of strata =         7
PSU:      PSU                           Number of PSUs   =        75
                                        Population size  =  148779.94
                                        F(  2,     67)   =      5.99
                                        Prob > F         =    0.0041

------------------------------------------------------------------------------
   DEMENTIA | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     WINE_1 |   1.040389   .2807133     0.147   0.884     .6072504    1.7824760
     WINE_2 |    .1691458  .0885213    -3.395   0.001     .0595280     .4806197
------------------------------------------------------------------------------
```

جدول ٦.١٦ این تحلیل را در چارچوبی مشابه چارچوب ارائه شده در جدول ٥.١٦ ارائه می‌دهد.

جدول ٥.١٦ تحلیل رگرسیونی لوژستیک از مصرف شراب و وقوع زوال عقل با فرض نمونه‌گیری تصادفی ساده (تحلیل مبتنی بر مدل)

مصرف شراب	نسبت بخت	خطای معیار نسبت بخت	بازه‌های اطمینان ۹۵ درصد
هیچ	۱/۰۰۰	—	—
روزی $\frac{۱}{۴}$ لیتر و کمتر	۱/۰۳۳	۰/۲۱۷	[۰/٦۸٤ , ۱/٥٦]
بیشتر از $\frac{۱}{۴}$ لیتر در روز	۰/۲۰٥	۰/۱۰۷	[۰/۰۷۳ , ۰/٥۷۱]

نسبتهای بخت در هر دو تحلیل شبیه‌اند ولی اثر حافظ بودن شراب در تحلیلی که طرح نمونه‌گیری را به صورتی صحیح گزارش می‌کرد ۱۸٪ بیشتر به نظر می‌رسد. جالب توجه است که این مصرف متوسط یا زیاد شراب، ولی نه کم آن است که به نظر می‌رسد اثر حفاظت در برابر زوال عقل داشته باشد. به علاوه، اثرهای طرح برای ضریبهای رگرسیونی به شدت اثرهای آن برای میانگینها یا نسبتها نیستند. در برخی موارد، اثرهای طرح عملاً کمتر از ۱ بود. جالب است که اثرهای طرح برای ضریبهای رگرسیونی را می‌توان با $۱+(\overline{n}-۱)\delta_x\delta_y$ تقریب زد. چون این عبارت به حاصلضرب ضریبهای همبستگی درون خوشه‌ای برای هر دو متغیر وابسته و مستقل بستگی دارد، که هر دو طبق تعریف، کوچکتر یا برابر با ۱ هستند، بزرگی این اثر کوچکتر یا برابر است با آنچه که در مورد میانگینها، مجموعها، یا نسبتها دیده شد. به علاوه، توجه به این نکته مهم است که δ_x و δ_y الزاماً در یک راستا نیستند. امکان دارد که داده‌های داخل خوشه‌ها نسبت به یک متغیر، ناهمگن و نسبت به متغیری دیگر همگن باشند. در این حالت، حاصلضرب ضریبهای همبستگی درون خوشه‌ای، منفی و اثر طرح حاصل از آن کمتر از ۱ خواهد بود.

جدول ٦.١٦ تحلیل رگرسیونی لوژستیک از مصرف شراب و وقوع زوال عقل با در نظر گرفتن پارامترهای آمارگیری نمونه‌ای (تحلیل مبتنی بر طرح)

مصرف شراب	نسبت بخت	خطای معیار نسبت بخت	بازه‌های اطمینان ۹۵ درصد
هیچ	۱/۰۰۰	—	—
روزی $\frac{۱}{۴}$ لیتر و کمتر	۰/۰٤۰	۰/۲۸۱	[۰/٦۰۷ , ۱/۷۸۲]
بیشتر از $\frac{۱}{۴}$ لیتر در روز	۰/۱٦۹	۰/۰۸۹	[۰/۰٦۰ , ۰/٤۸۱]

اخیراً، سازمانهای آمارگیری از قبیل مرکز ملی آمارهای بهداشتی[1] و ادارهٔ سرشماری امریکا[2] داده‌هایی را از آمارگیری بهداشت ملی و بررسی تغذیه (NHANES III)[3] و آمارگیری مصاحبه‌ای بهداشت ملی (NHIS)[4] از طریق سی‌دی‌رام CD-ROM برای استفادهٔ پژوهشگران علاقمند توزیع کرده‌اند. در مقابل غنای این قبیل پایگاههای اطلاعاتی و فقدان تاریخی نسبی دسترسی به نرم‌افزارهای مناسبی که بتوانند با کفایت طرح آمارگیری را در بررسیها گزارش کنند، پژوهشگران پیچیدگیهای آمارگیری را به سادگی نادیده گرفته‌اند و داده‌ها را به صورتی تحلیل کرده‌اند که گویی نتیجهٔ یک نمونهٔ تصادفی ساده‌اند. دسترسی به برنامه‌های جدید همچون STATA و SUDAAN قابلیتهای تحلیلی تازه‌ای که برای اجرای تحلیلهای مناسب ضروری‌اند در اختیار تحلیلگران داده‌ها قرار داده است.

۳.۱٦ خلاصه

در این فصل، یک شیوهٔ شش مرحله‌ای برای اجرای تحلیل مبتنی بر طرح داده‌های حاصل از آمارگیریهای نمونه‌ای ارائه شد تا بتوان آن را با هر نرم‌افزاری که قابلیت اجرای چنین تحلیلی را داشته باشد مورد استفاده قرار داد. این شیوه را با استفاده از یک طرح نمونه‌ای خوشه‌ای طبقه‌بندی شدهٔ دومرحله‌ای نسبتاً ساده نشان دادیم. سپس با استفاده از یک مثال که طرح نمونه‌ای پیچیده‌تری دارد نشان دادیم که چگونه غفلت در استفاده از تحلیل مبتنی بر طرح می تواند به نتایج گمراه کننده بینجامد.

پیوست فنی

اثبات این که احتمال انتخاب یک واحد نمونه‌گیری اولیهٔ (PSU) ویژه، i، تحت نمونه‌گیری با احتمال متناسب با اندازه از فرمول زیر به دست می‌آید

$$P\{i \in \text{ نمونه }\} = \frac{N_i}{N} m$$

در زیر آمده است. در این فرمول

N_i = تعداد واحدهای شمارش در واحد نمونه‌گیری اولیهٔ (PSU) i

N = کل تعداد واحدهای شمارش در جامعه

m = کل تعداد واحدهای نمونه‌گیری اولیه (PSU) در نمونه

[1] National Center for Health Statistics
[2] U.S. Bureau of the Census
[3] National Health and Nutrition Examination Survey
[4] National Health Interview Survey

اثبات. فرض کنید نمونه‌گیری واحدهای نمونه‌گیری اولیه (PSU ها) با جایگذاری است:

$$P\{PSU\ i \notin \text{نمونه}\} = \left(1 - \frac{N_i}{N}\right)^m \approx 1 - m\frac{N_i}{N}$$

(بسط قضیهٔ دوجمله‌ای تا جملهٔ اول)

پس

$$P\{PSU\ i \in \text{نمونه}\} = 1 - P\{PSU\ i \notin \text{نمونه}\} = \frac{N_i}{N}m$$

کتابشناسی

The following articles describe the design and basic findings of the PAQUID Study:

1. Lemeshow, S., Letenneur, L., Dartigues, J. F., Lafont, S., Orgogozo, J. M., and Commenges, D., An illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID study. *American Journal of Epidemiology,* 148(3): 298-306, 1998.
2. Barberger-Gateau, P., Dartigues, J. F., Commenges, D., Gagnon, M., Letenneur, L., Canet, C., Miquel, J. L., Nejjari, C., Tessier, J. F., Berr, C., Dealberto, M. J., Decamps, A., Alperovitch, A., and Salamon, R., Paquid: An interdisciplinary epidemiologic study of cerebral and functional aging. *Annals of Gerontology*, 383-392, 1992.
3. Letenneur, L., Commenges, D., Dartigues, J. F., and Barberger-Gateau, P. Incidence of dementia and Alzheimer's disease in elderly community residents of South-Western France. *International Journal of Epidemiology*, 23: 1256-1261, 1994.

The following article furnishes another example of the problems that result from ignoring design features in the analysis of data from sample surveys.

4. Brogan D., Software Packages for Analysis of Sample Survey Data, Misuse of Standard Packages. In *Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.

Also many of the references cited in Chapter 12 *deal with general issues in analysis of data from complex sample surveys.*