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.
۸ن۹ل/HB۸۴۹/۴۹ ۳۰۴/۶۰۱۵۱۹۵۲
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**

© ۱۳۸۱ پژوهشکدهٔ آمار
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

همهٔ حقوق این اثر برای پژوهشکدهٔ آمار محفوظ است. هیچ بخشی از این کتاب را نمی‌توان بدون اجازهٔ کتبی از ناشرش تکثیر یا به هر شکلی و با هر وسیله‌ای ذخیره کرد. استفاده یا تقلید از طرح جلد، ممنوع است.

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

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 $\frac{\bar{X}}{\bar{Y}}$ 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

$$R = \frac{\bar{X}}{\bar{Y}}$$

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

$$r=\frac{\bar{x}}{\bar{y}} \qquad \text{و} \qquad r=\frac{x}{y}$$

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 , ٦	۲۷۵۰۰۰	۸۲۰۰۰۰	۰/۳۳۵
۱ , ۷	۲۷۰۰۰۰	۹۸۰۰۰۰	۰/۲۷٦
۱ , ۸	۲۵۰۰۰۰	۷۵۰۰۰۰	۰/۳۳۳
۲ , ۳	۱۲۵۰۰۰	۵۰۰۰۰۰	۰/۲۵۰
۲ , ٤	۲۵۰۰۰۰	۸۰۰۰۰۰	۰/۳۱۳
۲ , ۵	۲۰۰۰۰۰	٦۵۰۰۰۰	۰/۳۰۸
۲ , ٦	۲۲۵۰۰۰	۷۲۰۰۰۰	۰/۳۱۳
۲ , ۷	۲۲۰۰۰۰	۸۸۰۰۰۰	۰/۲۵۰
۲ , ۸	۲۰۰۰۰۰	٦۵۰۰۰۰	۰/۳۰۸
۳ , ٤	۲۷۵۰۰۰	۹۰۰۰۰۰	۰/۳۰٦
۳ , ۵	۲۲۵۰۰۰	۷۵۰۰۰۰	۰/۳۰۰
۳ , ٦	۲۵۰۰۰۰	۸۲۰۰۰۰	۰/۳۰۵
۳ , ۷	۲٤۵۰۰۰	۹۸۰۰۰۰	۰/۲۵۰
۳ , ۸	۲۲۵۰۰۰	۷۵۰۰۰۰	۰/۳۰۰
٤ , ۵	۳۵۰۰۰۰	۱۰۵۰۰۰۰	۰/۳۳۳
٤ , ٦	۳۷۵۰۰۰	۱۱۲۰۰۰۰	۰/۳۳۵
٤ , ۷	۳۷۰۰۰۰	۱۲۸۰۰۰۰	۰/۲۸۹
٤ , ۸	۳۵۰۰۰۰	۱۰۵۰۰۰۰	۰/۳۳۳
۵ , ٦	۳۲۵۰۰۰	۹۷۰۰۰۰	۰/۳۳۵
۵ , ۷	۳۲۰۰۰۰	۱۱۳۰۰۰۰	۰/۲۸۳
۵ , ۸	۳۰۰۰۰۰	۹۰۰۰۰۰	۰/۳۳۳
٦ , ۷	۳٤۵۰۰۰	۱۲۰۰۰۰۰	۰/۲۸۸
٦ , 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.

$$SE(r)\approx\left(\frac{R}{\sqrt{n}}\right)\times\left(V_x^{۲}+V_y^{۲}-۲\rho_{xy}V_xV_y\right)^{۱/۲}\times\sqrt{\frac{N-n}{N-۱}} \qquad (۱.۷)$$

که در آن ρ_{xy}، ضریب همبستگی بین x و y است و به صورت زیر تعریف می‌شود

$$\rho_{xy}=\frac{\sum_{i=۱}^{N}(X_i-\bar{X})(Y_i-\bar{Y})/N}{\sigma_x\sigma_y} \qquad (۲.۷)$$

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

$$V(\bar{y})=\left(\frac{۱}{\bar{Y}}\right)\times\left(\frac{\sigma_y}{\sqrt{n}}\right)\times\sqrt{\frac{N-n}{N-۱}}=\left(\frac{۱}{۴۳۷۵۰۰}\right)\times\left(\frac{۱۵۲۷۰۴/۸}{\sqrt{۲}}\right)\times\sqrt{\frac{۸-۲}{۸-۱}}=۰/۲۲۸۵$$

محاسبه با معادلهٔ (۱.۷) مقدار

$$SE(r)\approx\left(\frac{۰/۳۰۵۷}{\sqrt{۲}}\right)\times\left[(۰/۳۶۹۵)^{۲}+(۰/۳۴۹۰)^{۲}-۲(۰/۹۲۷۲)(۰/۳۶۹۵)(۰/۳۴۹۰)\right]^{۱/۲}\times\sqrt{\frac{۸-۲}{۸-۱}}=۰/۰۲۷۷$$

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 $V(\bar{y})=۰/۰۴۹۹\leq ۰/۰۵$ 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

□

خطای معیار برآورد یک نسبت را می‌توان 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:

$$\frac{s_y}{\sqrt{n} \times \bar{y}} \times \sqrt{\frac{N-n}{N}} \leq 0.05$$

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 $100(1-\alpha)$ درصدی را می‌توان به صورت زیر ساخت

$$r - z_{1-(\alpha/2)} \hat{SE}(r) \leq R \leq r + z_{1-(\alpha/2)} \hat{SE}(r)$$

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، ٥، 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 ۰/۰۰٦۳ 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

$x_{(۲)} =$ ۲۲۵۶۰ کل تعداد نوزادان در زیرحوزهٔ ۲ بیمارستانهای نمونه که وضعیت پادگن سطحی هپاتیت B مادرانشان در پروندهٔ پزشکی ثبت نشده است

$y_{(۱)} = ۴۷۸۵۰$ کل تعداد نوزادان در نمونهٔ زیرحوزهٔ ۱ بیمارستانها

$y_{(۲)} = ۳۸۹۲۹$ کل تعداد نوزادان در نمونهٔ زیرحوزهٔ ۲ بیمارستانها

$N_۱ = ۶۶$ تعداد جامعهٔ معلوم بیمارستانهای زیرحوزهٔ ۱

$N_۲ = ۳۵$ 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.

□

۴.۷ 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 فصل ۳ شرح داده شده است برآورد کنیم. $\binom{۶}{۲}$=۱۵ 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	٦٦٥٧	۲۲٦۹
۲	٤٠٥٧	۱۳۲٤
۳	۳٦٤۲	۹٥۲
٤	٥۳۲٠	۱٥٥۸
٥	٤٤۸٠	۱۳٥۲
٦	٥۸۸٠	۱۷۹٦
مجموع	۳٠٠۳٦	۹۲٥۱

* 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	۱٠۷۷۹	۲ , ٦	۹۳٦٠
۱ , ۳	۹٦٦۳	۳ , ٤	۷٥۳٠
۱ , ٤	۱۱٤۸۱	۳ , ٥	٦۹۱۲
۱ , ٥	۱٠۸٦۳	۳ , ٦	۸۲٤٤
۱ , ٦	۱۲۱۹٥	٤ , ٥	۸۷۳٠
۲ , ۳	٦۸۲۸	٤ , ٦	۱٠٠٦۲
۲ , ٤	۸٦٤٦	٥ , ٦	۹٤٤٤
۲ , ٥	۸٠۲۸		

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)

$$E(x'') = \frac{۱۰۰۷۳+۹۳۹۴+\cdots+۹۱۲۷}{۱۵} = ۹۲۱۱/۰۷$$

$$SE(x'') = \left[\frac{(۱۰۰۷۳-۹۲۱۱/۰۷)^۲+(۹۳۹۴-۹۲۱۱/۰۷)^۲+\cdots+(۹۱۲۷-۹۲۱۱/۰۷)^۲}{۱۵}\right]^{۱/۲} = ۴۶۶/۳۸$$

$$MSE(x'') = Var(x'') + B^۲(x'') = ۲۱۷۵۱۲/۳۳ + (۹۲۱۱/۰۷-۹۲۵۱)^۲ = ۲۱۹۱۰۶/۷۳$$

جدول ٦.۷ 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 نااریب 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

$$x'' = \left(\frac{x'}{y'}\right) \times Y = ۸۷۸/۹۲۳۰ \times ۵۰ = ۴۳۹۴۶/۱۵ \quad \text{کامیون ـ مایل}$$

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:

$$SE(x'') = \left(\frac{YR}{\sqrt{n}}\right) \times \left(V_x^۲ + V_y^۲ - ۲\rho_{xy} V_x V_y\right)^{۱/۲} \times \sqrt{\frac{N-n}{N-۱}} \qquad (۶.۷)$$

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 ٤ 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:

$$\hat{SE}(x'') = \left(\frac{۵۰\times ۸۷۸/۵۸۴۸}{\sqrt{۳}}\right)\times\Big[۰/۰۲۴۵ + ۰/۰۰۳۸۸۳ - ۲\times ۰/۹۳۳۷ \times\sqrt{۰/۰۲۱۶}\times\sqrt{۰/۰۰۳۸۸۳}\Big]^{۱/۲}\times\sqrt{\frac{۸-۳}{۸-۱}} = ۱۹۶۳/۰۴$$

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:

$$x'' - ۱/۹۶\times\hat{SE}(x'') \le X \le x'' + ۱/۹۶\times\hat{SE}(x'')$$
$$۴۳۹۴۶/۱۵ - ۱/۹۶\times ۱۹۶۳/۰۴ \le X \le ۴۳۹۴۶/۱۵ + ۱/۹۶\times ۱۹۶۳/۰۴$$
$$۴۰۰۹۸ \le X \le ۴۷۷۹۳/۷۱$$

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 $(\hat{V}_y^{۲})^{۱/۲}$ 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 (۶.۷) 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:

$$\frac{Var(x'')}{Var(x')} = \frac{V_x^۲ + V_y^۲ - ۲\rho_{xy}V_xV_y}{V_x^۲}$$

چون

$$\begin{aligned} Var(x'') &= Y^۲ \times Var(r) \\ &= Y^۲ \times \left(\frac{R^۲}{n}\right) \times \left(V_x^۲ + V_y^۲ - ۲\rho_{xy}V_xV_y\right) \times \left(\frac{N-n}{N-۱}\right) \end{aligned}$$

و

$$\begin{aligned} Var(x') &= \left(\frac{N^۲}{n}\right) \times \sigma_x^۲ \times \left(\frac{N-n}{N-۱}\right) \\ &= \left(\frac{N^۲}{n}\right) \times \left(\bar{X}^۲ V_x^۲\right) \times \left(\frac{N-n}{N-۱}\right) = \left(\frac{X^۲}{n}\right) \times V_x^۲ \times \left(\frac{N-n}{N-۱}\right) \end{aligned}$$

نتیجه می‌گیریم که واریانس x'' وقتی کمتر از واریانس x' خواهد بود که

$$V_x^۲ + V_y^۲ - ۲\rho_{xy}V_xV_y < V_x^۲$$

یا به طور هم‌ارز هرگاه

$$\frac{V_y}{V_x} < ۲\rho_{xy} \qquad (۸.۷)$$

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.

۶.۷ 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

$$\hat{SE}(x'') = \left[\frac{۵۰\times ۸۷۸/۵۸۴۸}{\sqrt{۳}}\right](۰/۰۲۴۵-۰/۰۰۴۴۳۵)^{۱/۲}\sqrt{\frac{۸-۳}{۸-۱}} = ۳۰۳۶/۳۳$$

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.

$$b = \frac{\sum_{i=۱}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=۱}^{n}(y_i-\bar{y})^{۲}} \qquad (۱٦.۷)$$

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	۱۱۵	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:

$$x' = ۲\times(۱۰۳+۱۱۵) = ۴۳۶$$
$$y' = ۲\times(۶۰+۷۰) = ۲۶۰$$
$$\bar{x} = \frac{۱۰۳+۱۱۵}{۲} = ۱۰۹$$
$$\bar{y} = \frac{۶۰+۷۰}{۲} = ۶۵$$
$$b = \frac{(۱۰۳-۱۰۹)(۶۰-۶۵)+(۱۱۵-۱۰۹)(۷۰-۶۵)}{(۶۰-۶۵)^۲+(۷۰-۶۵)^۲} = ۱/۲۰$$
$$Y = ۱۸۱$$

و

$$x''' = x' + b(Y - y') = ۴۳۶ + ۱/۲۰(۱۸۱-۲۶۰) = ۳۴۱/۲$$

برآورد نسبتی x'' به صورت زیر به دست می‌آید

$$x'' = \left(\frac{x'}{y'}\right)\times Y = \left(\frac{۴۳۶}{۲۶۰}\right)\times ۱۸۱ = ۳۰۳/۵$$

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	(اریبی)^{۲}	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 $h=۱,...,L$ ، 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	٤۸/٤۰
۸	٤	۵۷٤	٦۰/۰۰
۹	۲	٤۹۸	٤۸/٤۰
۱۰	۵	۳۷۲	۸۸/۸۰
۱۱	۳	۳۷۸	۲٦/۸۰
۱۲	٦	۳۷۲	۳۹/٦۰
۱۳	٤	۳٦۰	۵۸/۸۰
۱٤	٤	٤۵۰	۵٤/۲۰
۱۵	۲	۵٤۰	٤٤/۲۰
۱٦	۵	٤۵۰	۷۵/٤۰
۱۷	۳	٤۱٤	٤۵/۲۰
۱۸	٤	٤۹۸	۷۲/۰۰
۱۹	۲	۵۱۰	۲۱/۲۰
۲۰	٤	٤۳۸	۵۵/٤۰
۲۱	۲	۳۹٦	۵۱/۹۰
۲۲	۵	۳٤۸	٤٦/٦۰
۲۳	۳	٤٦۲	۷۹/٦۰
۲٤	٤	٤۱٤	۳۳/٦۰
۲۵	۷	۳۹۰	۷۵/٦۰
۲٦	۳	٤٦۲	٦۹/٦۰
۲۷	۳	٤۱٤	۵۷/٤۰
۲۸	٦	۵۷۰	۱۲٦/۰۰
۲۹	۲	٤٦۲	۳۹/۱۰
۳۰	۲	٤۱٤	٤۳/۲۰
۳۱	٦	٤۱٤	۳٦/٤۰
۳۲	٤	٤۰۲	٤۰/۲۰
۳۳	۲	۳۷۸	٤۱/٤۰

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؟

۴.۷ 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 ۳۵۲۴ 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
۴	۲	۶
۵	۳	۳
۶	۷	۰
۷	۱	۰
۸	۱	۲
۹	۰	۰
۱۰	۴	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.

۶.۷ 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 ســادهٔ 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 اولین اهدای خون	تعداد اشخاص	وضعیت از لحاظ HIV	
		مثبت	منفی
۳	۱۸	۲	۱٦
٤	۱۰	۲	۸
٥	۸	۱	۷
٦	۷	۰	۷
۷	٦	۲	٤
۸	٥	۰	٥
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: ٤۳۷۵

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 ۱٤۰۰۰۰ 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.