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 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

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

واضح است که چون عبارت بالا تابعی خطی از واریانس جامعه‌ای $\sigma_x^{۲}$ است با جایگزین کردن برآورد نااریب $\hat{\sigma}_x^{۲}$ از $\sigma_x^{۲}$ 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 آمارگیریهای نمونه‌ای که به وسیلهٔ ادارهٔ سرشماری امریکا[1]، مرکز ملی آمارهای بهداشتی[2]، دفتر آمار کار[3]، و سایر سازمانها اجرا می‌شود مصداق دارد.

برای برآورد کردن واریانسهای برآوردهای حاصل از آمارگیریهای نمونه‌ای که مستلزم نمونه‌گیریها و شیوه‌های پیچیده برای برآورد کردن هستند 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.

فن *خطی‌سازی* به وسیلهٔ کی‌فیتس [۱] ، وودراف [۲] و دیگران برای آمارگیریهای نمونه‌ای و براساس تقریب سری تیلور[4] بسط داده شده است. در اواخر دههٔ ۱۹۷۰، پژوهشگران مؤسسهٔ مثلث تحقیق[5] شروع به ابداع مطلبی کردند که سرانجام تبدیل به یک نرم‌افزار راحت براساس خطی‌سازی شد [۳]. این نرم‌افزار، شکل نخستین SUDAAN بود که پس از تکامل بسیار به نرم‌افزاری بدل شد که کاربرد وسیعی دارد و در سراسر این کتاب مورد اشاره و بحث بوده است. قابل استفاده بودن و جاذبهٔ این نرم‌افزار و سایر نرم‌افزارها باعث شده است که خطی‌سازی شاید اکنون پراستفاده‌ترین روش برآورد واریانس برای آمارگیریهای پیچیده باشد.

تکرار نیز یک ردهٔ کلی از روشهایی است که با آن برآوردی از واریانس یک پارامتر جامعه‌ای برآورد شده به دست می‌آید، به این ترتیب که پارامتر جامعه‌ای برآورد شده به صورت مجموع یا میانگینی از چندین آماره بیان می‌شود که هر یک بر زیرمجموعه‌ای از مشاهدات نمونه‌ای متکی است. واریانس پارامتر جامعه‌ای برآورد شده را سپس می‌توان با به دست آوردن برآوردی از واریانس این آماره‌های «جزء نمونه‌ای» برآورد کرد. مثال ساده‌ای از این فن در فصل ٤، با توجه به برآورد واریانس یک مشخصهٔ جامعه‌ای برآورد شده، تحت نمونه‌گیری سیستماتیک مکرر نشان داده شد. گونه‌هایی از این رده از فنون طی چندین سال مورد استفاده قرار گرفته است. هنسن و همکاران [۵] در کتاب درسی خود که در اوایل دههٔ ۱۹۵۰ تألیف کردند به این فنون تحت عنوان *روشهای گروه تصادفی* اشاره کرده‌اند. این روشها در دههٔ ۱۹٦۰ توسط مک‌کارتی [٦] و [۷] دقیقتر شده و در سطح گسترده‌ای در مرکز ملی آمارهای بهداشتی و سایر سازمانهای فدرال 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، $\hat{\theta}$ ممـــکن 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

متضـمن اسـتـفاده از روشـــهای مبتـنی بر طرح، از قبیل روشـهای توصیف شـده در فصلهای ۳ تا ۱۱، بـرای بـرآورد واریـانس 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

$$\tilde{z}_i = (\partial f_{\bar{x}})x_i + (\partial f_{\bar{y}})y_i$$

$$\partial f_{\bar{x}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{x}}, \quad (\bar{x},\bar{y}) \text{ محاسبه شده در نقطهٔ}$$

و

$$\partial f_{\bar{y}} = \frac{\partial f(\bar{x},\bar{y})}{\partial \bar{y}}, \quad (\bar{x},\bar{y}) \text{ محاسبه شده در نقطهٔ}$$

(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	۱۵۷	۰/۱۵۱۸
۵	۳٤۹	۰	−۰/٤۸۹٤
٦	۲۱۰	۲۱۰	۰/٤٦٧۵
۷	۵۳٦	۰	−۰/۷۵۱٦
۸	۲۸۹	۱۳۵	۰/۰۸٤٦
۹	۹۸	۰	−۰/۱۳۷٤
۱۰	۳٤۵	۲۳۰	۰/۳۵۰۸

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 $x'_{h۱}$، $x'_{h۲}$، $y'_{h۱}$ و $y'_{h۲}$ 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 $۲L$ برآورد از فرمول زیر به دست می‌آید

$$r = \frac{x'}{y'}$$

که در آن

$$x' = \sum_{h=۱}^{L} \sum_{i=۱}^{۲} \left(x'_{hi} / ۲ \right)$$

و

$$y' = \sum_{h=۱}^{L} \sum_{i=۱}^{۲} \left(y'_{hi} / ۲ \right)$$

[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:

$$r_{(k)} = \frac{\sum_{h=1}^{L}\left[\delta_{kh}x'_{h1} + (1-\delta_{kh})x'_{h2}\right]}{\sum_{h=1}^{L}\left[\delta_{kh}y'_{h1} + (1-\delta_{kh})y'_{h2}\right]}$$

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 [۶] و بین [۹] و نیز لِمی‌شو و لِه‌وی [۱۰] 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	۲۴
۲	۱	۱۸۵	۳۰
	۲	۲۲۸	۴۹
۳	۱	۶۷۰	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
۵	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	۱۵٤	۱۰۱۸	۰/۱۵۱۳
٤	۱٤٤	۸۷۸	۰/۱٦٤۰
۵	۱۳٤	۹۳۳	۰/۱٤۳٦
٦	۱۲٤	۷۹۳	۰/۱۵٦٤
۷	۱۵۳	۹۷٦	۰/۱۵٦۸
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

$$\bar{r} = \frac{۱}{۸}\sum_{k=۱}^{۸} r_{(k)} = ۰/۱۵۳۹$$

و

$$\hat{Var}(\bar{r}) = ۰/۰۰۰۰۹۸$$

پلاکت و برمن [۱۷]، ماتریسهای $m \times m$ را با ابعاد $m = \{۴,۸,۱۲,\ldots,۱۰۰\}$ به استثنای $m = ۹۲$ ارائه داده‌اند. بعد ماتریس مورد استفاده در یک مسئلهٔ خاص به تعداد طبقه‌ها بستگی دارد. مک‌کارتی [٦] از ماتریسهایی استفاده کرده است که در آن m مضربی از ٤ و $L \le m \le L + ۳$، که در آن در همهٔ موارد $L > ۲$، $m << ۲^L$ 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، ٤، ٦ و ۷ به شرح زیر به دست آورد:

تکرار k	$\delta_{k۱}$	$۱-\delta_{k۱}$	$\delta_{k۲}$	$۱-\delta_{k۲}$	$\delta_{k۳}$	$۱-\delta_{k۳}$
۱	۱	۰	۱	۰	۱	۰
٤	۱	۰	۰	۱	۰	۱
٦	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:

$$\bar{r} = \frac{(r_{(۱)} + r_{(۴)} + r_{(۶)} + r_{(۷)})}{۴} = ۰/۱۵۳۹$$

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

$$r_{(hi)} = \frac{\sum_{h'\neq h}\sum_{i=۱}^{n_h} x'_{h'i} + \sum_{i'\neq i}\frac{n_h}{n_h - ۱} x'_{hi'}}{\sum_{h'\neq h}\sum_{i=۱}^{n_h} y'_{h'i} + \sum_{i'\neq i}\frac{n_h}{n_h - ۱} y'_{hi'}}$$

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 $(n_h/(n_h - ۱))$ 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

$$\hat{SE}(r) = \sqrt{\sum_{h=۱}^{L}\sum_{i=۱}^{n_h}\frac{n_h - ۱}{n_h}(r_{(hi)} - r)^۲} \qquad (٤.۱۲)$$

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 $x'_{۱۱}$ ، $y'_{۱۱}$ و $r_{(۱۱)}$ نشان داده شد به دست می‌آیند:

$$x'_{۱۱} = ۲/۵\times(۲\times۲۴+۳۰+۴۹+۸۰+۷۰) = ۶۹۲/۵$$
$$y'_{۱۱} = ۲/۵\times(۲\times۷۸+۱۸۵+۲۲۸+۶۷۰+۵۳۰) = ۴۴۲۲/۵$$
$$r_{(۱۱)} = \frac{۶۹۲/۵}{۴۴۲۲/۵} = ۰/۱۵۶۵۸۶$$

همان‌طور که قبلاً نشان داده شد، $r = ۰/۱۵۳۵۰۶$ ، و از برابری (٤.۱۲) داریم

$$\hat{SE}(r) = \sqrt{۹/۷\times۱۰^{-۵}} = ۰/۰۰۹۸۴۹$$

□

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 (۴.۱۲) به دست آمده است مطابقت دارد و مقدار ۰/۰۰۹۸۴۲ آن بسیار شبیه به مقداری است که قبلاً از روش تکرار مکرر متعادل به دست آمده بود (برآورد خطای معیار r برابر است با ۰/۰۰۹۴۳). هر دوی این مقدارها به مقدار برآورد شدهٔ ۰/۰۰۹۴۳ 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 ۰/۰۰۷۶۰) 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 ۱۰ 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.

۳.۲.۱۲ 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

بـرای انـدازه‌گیـری ایـن تغییرپذیـری «مصـاحبه‌گر»، از روشی موسوم به *نمونه‌های نافذ بر هم* استفاده مـی‌شـود. این روش را شاید بتوان به بهترین وجه دربارهٔ وضعیتی توضیح داد که در آن نمونهٔ تصادفی سـاده‌ای متشـکل از n عنصـر از جامعه‌ای با N عنصر انتخاب می‌شود. اگر قرار باشد داده‌ها توسـط M مصاحبـه‌گر جمع‌آوری شوند در آن صورت، به جای گرفتن یک نمونهٔ تصـادفی سـاده از n عنصر، M نمونـهٔ تصـادفی سـادهٔ مستقل انتخاب می‌کنیم که هر یک شامل $\bar{n} = n/M$ عنصر باشد و هر مجموعهٔ متشکل از $\bar{n}$ 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 $\bar{X}$ برای مشخصهٔ x در جامعه تهیه خواهد کرد و واریانس، $\sigma_x^۲$ مربوط به توزیع x می‌تواند به وسیلهٔ $s_{bx}^۲$ برآورد شود با این فرض که در فرایند اندازه‌گیری هیچ تفاوتی میان مصاحبه‌گران وجود نداشته است:

$$s_{bx}^۲ = \frac{\bar{n}\sum_{i=۱}^{M}(\bar{x}_i - \bar{x})^۲}{M-۱}$$

که در آن $\bar{x}$ میانگین $\bar{x}_i$ است.

برآورد دیگری از $\sigma_x^۲$ که چه اثرهای مصاحبه‌گر وجود داشته یا نداشته باشند همچنان معتبر است از فرمول زیر به دست می‌آید

$$s_{wx}^۲ = \frac{\sum_{i=۱}^{M} s_{ix}^۲}{M}$$

که در آن، برای $i = ۱,\ldots,M$

$$s_{ix}^۲ = \frac{\sum_{j=۱}^{\bar{n}}(x_{ij} - \bar{x}_i)^۲}{\bar{n}-۱}$$

تحت فرض صفر مبنی بر اینکه هیچ اثر مصاحبه‌گر وجود ندارد، نسبت $s_{bx}^۲/s_{wx}^۲$ با $(M-۱)$ 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 $\bar{n}$ 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 اندازه‌گیـری نیز به کار برد. δ_x، برآوردگر ضریب همبستگی میان رده‌ای که توسط کالتون [۱۹] و دیگران پیشنهاد شده است به صورت زیر است:

$$\delta_x = \frac{\left(\frac{s_{bx}^{۲}}{s_{wx}^{۲}} - ۱\right)}{\left(\frac{s_{bx}^{۲}}{s_{wx}^{۲}} - ۱ + \bar{n}\right)} \qquad (۵.۱۲)$$

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
$s_{ix}^{۲}$	۰/۷۰	۰/۸۰	۰/۷۰	۰/۸۰

با $M = ٤$ و $\bar{n} = ۵$، داریم

$$\bar{x} = ۳/۱۵ \qquad s_{bx}^{۲} = ۶/۸۵ \qquad s_{wx}^{۲} = ۰/۷۵$$

نسبت F چنین است

$$\frac{s_{bx}^{۲}}{s_{wx}^{۲}} = \frac{۶/۸۵}{۰/۷۵} = ۹/۱۳۳$$

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 ۳۴ 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 ۳۴ 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	۱۰۴
۱	۲	۳	۳۴
۱	۳	۱	۱۷
۱	۳	۲	۱۴۰
۱	۴	۱	۱۰۴

طبقه‌بندی برای آمارگیری از پذیرش بیماران دچار ایدز (ادامه)

طبقه	بخش	بیمارستان	تخت
۲	۱	۱	٤٤
۲	۲	۱	۹۹
۲	۳	۱	۸٦
۲	۳	۲	۹۱
۲	۳	۳	۵۳
۲	٤	۱	٤۸
۳	۱	۱	۱۷۱
۳	۱	۲	۸۵
۳	۲	۱	۱۰۸
۳	۳	۱	۹۹
۳	٤	۱	۱۳۱
۳	٤	۲	۱۸۲
٤	۱	۱	٤۸
٤	۱	۲	۱۸
٤	۲	۱	۵۰
٤	۳	۱	٤۲
٤	٤	۱	۳۸
۵	۱	۱	٤۲
۵	۲	۱	۵٤
۵	۳	۱	٤۵
۵	٤	۱	۳٤
٦	۱	۱	۳۹
٦	۱	۲	۵۹
٦	۲	۱	۷٦
٦	۳	۱	٦۸
٦	۳	۲	٦۸
٦	٤	۱	٦۵

الف. نشان دهید که چگونه تعداد پذیرش بیماران دچار ایدز را برای ناحیهٔ ۳٤ 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 تعـــداد پـذیرش مبتلایان به ایدز برای ناحیهٔ ٣٤ بخشی استفاده کنید. از ماتریس پلاکت ـ برمن که در پـاییـن 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 به 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 نسبت این قبیل فروشگاهها بود که براساس برنامه‌های پیش پرداخت حق عضویت، انگیزه‌هایی برای مشتریان فراهم می‌کنند. یک نمونهٔ تصادفی ساده متشکل از ۴۸ فروشگاه از این نوع گرفته 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 جک‌نایف در برآورد کردن خطای معیار نسبت اضافه پرداختی برای داده‌های جدول ۱.۱۲ استفاده کنید. نتیجهٔ حاصل چگونه با خطای معیار حاصل از روش خطی‌سازی مقایسه می‌شود؟

پیوست فنی

X'_{h1} و X'_{h2} را برآوردهای نااریب مستقلی از دو واحد نمونه‌گیری اولیه برای پارامتر X_h (مانند میانگین، مجموع) در طبقهٔ h ($h=1,\ldots,L$) در نظر می‌گیریم و فرض می‌کنیم که $X'_{(1)},\ldots,X'_{(K)}$ 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:

$$\begin{aligned}\hat{X} &= \frac{\sum_{k=1}^{K} X'_{(k)}}{K}\\ &= \left(\frac{1}{K}\right)\sum_{h=1}^{L}\frac{K}{2}(X'_{h1}+X'_{h2})\\ &= \sum_{h=1}^{L}\frac{(X'_{h1}+X'_{h2})}{2}\end{aligned}$$

واریانس $\hat{Var}(\hat{X})$ را به صورت زیر تعریف می‌کنیم:

$$\hat{Var}(\hat{X}) = \frac{\sum_{k=1}^{K}(X_{(k)}-\hat{X})^2}{K}$$

نشان خواهیم داد که $\hat{Var}(\hat{X})$ برآوردی نااریب از $Var(\hat{X})$ است.

به پنج نتیجهٔ زیر توجه کنید

۱.

$$\begin{aligned}E(X'_{h1}-X'_{h2})^2 &= Var(X'_{h1}-X'_{h2})+E^2(X'_{h1}-X'_{h2})\\ &= Var(X'_{h1}-X'_{h2})\\ &= Var(X'_{h1})+Var(X'_{h2})\end{aligned}$$

۲.

$$E\left[\sum_{h=1}^{L}(X'_{h1}-X'_{h2})^2\right] = \sum_{h=1}^{L}\left[Var(X'_{h1})+Var(X'_{h2})\right]$$

(این نتیجه پیامد ۱ است.)

۳.

$$\begin{aligned}E(\hat{X}) &= \left(\frac{1}{K}\right)\sum_{k=1}^{K}\sum_{h=1}^{L}E\left[\delta_{kh}X'_{h1}+(1-\delta_{kh})X'_{h2}\right]\\ &= \left(\frac{1}{K}\right)\sum_{k=1}^{K}\sum_{h=1}^{L}X_h\\ &= \left(\frac{1}{K}\right)K\times X\\ &= X\end{aligned}$$

.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.)

$$= \left(\frac{۱}{۴}\right)\sum_{h=۱}^{L}\left(Var(X'_{h۱}) + Var(X'_{h۲})\right)$$

.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،

$$E\left[Var(\hat{X})\right] = E\left[\left(\frac{۱}{K}\right)\sum_{k=۱}^{K}(X'_{(k)} - \hat{X})^{۲}\right]$$
$$= \left(\frac{۱}{K}\right) \times K \times Var(\hat{X})$$
$$= Var(\hat{X})$$

برای این مورد خاص نشان دادیم که $\hat{Var}(\hat{X})$ برآوردگری نااریب از $Var(\hat{X})$ است. با این که برآوردهای ناخطی 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 تقریبی مصداق پیدا می‌کند 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/