

نمونه‌گیری

روشها و کاربردها

پل اس. لهوی

استنلی لمی شو

مترجم

گیتی مختاری امیرمجدی



پژوهشکده آمار

لوی، پل	Levy, Paul S.
نمونه‌گیری: روشها و کاربردها / پل اس. لهوی، استنلی لمی شو؛ مترجم گیتی مختاری امیرمجدی	
-- تهران: مرکز آمار ایران، پژوهشکده آمار، ۱۳۸۱.	
۵۸۵ ص. : جدول، نمودار.	
فهرست‌نویسی بر اساس اطلاعات فیفا.	ISBN 964-365-150-9 ریال : ۳۰۰۰۰
عنوان اصلی: Sampling of Populations: Methods and Applications.	
کتاب‌نامه.	
نمایه	
۱. جمعیت - روشهای آماری. ۲. آمارگیری نمونه‌ای. الف. لمشو، استنلی Lemeshow, Stanley	
ب. مختاری امیرمجدی، گیتی، ۱۳۳۳- ، مترجم. ج. مرکز آمار ایران، پژوهشکده آمار. د. عنوان.	
ان ۹/۴۹/۴۹ HB8۴۹/۴۹/۴۹	۳۰۴/۶۰۱۵۱۹۵۲
۱۳۸۱	
کتابخانه ملی ایران	۴۰۶۴۷-۸۱ م

- 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 : مؤسسه انتشارات ستایش

© ۱۳۸۱ پژوهشکده آمار

شماره 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

همه حقوق این اثر برای پژوهشکده آمار محفوظ است. هیچ بخشی از این کتاب را نمی‌توان بدون اجازه 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 تیترو میترا، و قلم لاتین Times New Roman.

چاپ و صحافی شده در ایران.

چاپ یکم

شمارگان: ۶۰۰

پیشنهاد برای نحوه نقل مطلب، جدول یا نمودار از این کتاب، به صورت زیر است:

مختاری امیرمجدی، گیتی (۱۳۸۱). نمونه‌گیری: روشها و کاربردها. لهوی، پل اس.؛ لمی شو، استنلی. ترجمه از انگلیسی به فارسی. تهران: پژوهشکده آمار.

شابک ۹۶۴ - ۳۶۵ - ۱۵۰ - ۹

ISBN 964-365-150-9

بها: سی و پنج هزار ریال

فصل ۱۶

راهنمایی برای تحلیل مبتنی بر طرح داده‌های آمارگیری نمونه‌ای

محققان، آمارگیریه‌ای نمونه‌ای را نوعاً برای تهیه برآوردهای معتبر و قابل اعتماد از پارامترهای جامعه‌ای از قبیل میانگینها، مجموعهها، و نسبتها طراحی می‌کنند، ولی بیش از پیش به ارزشیابی روابط درونی میان متغیرها نیز علاقه‌مندند. بسیاری از اوقات از نرم‌افزارهای آماری که در سطح وسیعی در دسترس قرار دارند برای تحلیل این قبیل داده‌ها استفاده شده است، به خصوص هنگامی که در تحلیل، مدلسازی چندمتغیره موردنظر بوده است. متأسفانه، این‌گونه نرم‌افزارها به طور کلی بسیاری از ویژگیهای طرح آمارگیری خاصی را که مورد تحلیل قرار دارد در نظر نمی‌گیرند. در این کتاب، بر استفاده از SUDAAN و STATA در تحلیل داده‌های آمارگیری تمرکز شده است، زیرا این نرم‌افزارها برخلاف بسیاری از نرم‌افزارهای آماری دیگری که فعلاً در دسترس‌اند از این توانایی برخوردارند که داده‌های آمارگیری را طوری تحلیل کنند که ویژگیهای مربوط به طرح نیز در نظر گرفته شوند. همان‌طور که قبلاً اشاره کردیم، استفاده ما از این دو محصول بازتاب آشنایی خود ما با آنهاست و نباید به عنوان تأیید آنها در مقابل سایر نرم‌افزارهایی که با تواناییهای مشابه فعلاً در دسترس قرار دارند یا آنهایی که ممکن است در آینده در دسترس باشند تعبیر شود. راهنمایی که در این فصل و در سراسر این کتاب به اجمال بیان شده‌اند بر نظریه نمونه‌گیری متکی‌اند و باید مستقل از و قابل تبدیل به هر

محصول نرم‌افزاری باشند که مدعی اجرای تحلیلی است که بر طرح داده‌های آمارگیری نمونه‌ای مبتنی است که به صورت زیر تعریف شده است.

تحلیل مبتنی بر طرح داده‌های آمارگیری نمونه‌ای، تحلیلی است که ویژگیهای ملازم با طرح نمونه‌ای را در نظر می‌گیرد. غالباً این ویژگیها از جمله شامل طبقه‌بندی، خوشه‌بندی، کسرهای نمونه‌گیری به اندازه کافی بزرگ برای بهره‌گیری از کاربرد تصحیح جامعه متناهی، و وزنهای نمونه‌گیری هستند. تحلیلی که مبتنی بر طرح نباشد، یعنی ویژگیهای طرح را، از قبیل آنچه در بالا فهرست شد، در نظر نگیرد به عنوان تحلیل مبتنی بر مدل نامیده خواهد شد. کاربرد ما از این اصطلاح اخیر وسیعتر از آن است که در نوشتگان به کار می‌رود زیرا که تحت تحلیل مبتنی بر مدل، تحلیلی را می‌گنجانیم که تحت هیچ مدل صریحاً بیان شده‌ای انجام نشده است، بلکه به نظر می‌رسد نمونه‌گیری تصادفی مستقل از یک جامعه نامتناهی را فرض می‌گیرد (فرضهای «کلاسیک»).

در بخشهای قبل نشان دادیم که هرگاه ویژگیهای طرح آمارگیری در تحلیل نادیده گرفته شود، این امکان وجود دارد که نتایج نادرست باشند. در این فصل نشان خواهیم داد که چگونه ویژگیهای مربوط به طرح شناسایی می‌شوند، و سپس به صورتی مناسب در تحلیل داده‌های آمارگیری نمونه‌ای منظور می‌شوند. همچنین نشان خواهیم داد که چگونه غفلت از منظور کردن این ویژگیها می‌تواند به برآوردهای نادرست و استنباطهای نامعتبر منجر شود.

۱.۱۶ مراحل مورد نیاز برای اجرای تحلیل مبتنی بر طرح

گامهای زیر برای اجرای تحلیل مبتنی بر طرح مورد نیازند.

۱. شناسایی عناصر زیر در طرح نمونه‌ای:
 - طبقه‌بندی
 - خوشه‌بندی متغیرهای مورد استفاده
 - اندازه‌های جامعه مورد نیاز برای تعیین تصحیح جامعه متناهی
۲. تعیین وزن برای هر آزمودنی نمونه براساس اطلاعات بالا.
۳. تعیین یک وزن نمونه‌گیری نهایی برای هر سابقه نمونه که هرگونه تعدیل را که برای بی‌پاسخی و پس طبقه‌بندی مورد نیاز است در نظر بگیرد.
۴. حصول اطمینان از اینکه همه داده‌های مربوط به طبقه‌بندی، خوشه‌بندی، و اندازه جامعه که برای یک تحلیل مناسب مبتنی بر طرح مورد نیازند، برای هر سابقه نمونه شناسایی شده‌اند.

۵. تعیین شیوه و مجموعه فرمانها به منظور اجرای تحلیل مورد نیاز برای بسته نرم‌افزاری خاصی که قرار است به کار رود.

۶. اجرای تحلیل و تفسیر دقیق یافته‌ها

مثال تشریحی: ایالتی را در نظر می‌گیریم که شامل ۵ ناحیه و دارای ۵۷ آسایشگاه سالمندان به صورتی است که در جدول ۱.۱۶ نشان داده شده‌اند. فرض کنید یک نمونه تصادفی طبقه‌بندی شده متشکل از دو آسایشگاه سالمندان از هر ناحیه می‌گیریم و از هر آسایشگاه سالمندان نمونه‌ای متشکل از پنج پذیرش را انتخاب می‌کنیم تا کل تعداد بیمارانی که در طی سال ۱۹۹۷ در آسایشگاههای سالمندان این پنج ناحیه پذیرش شده‌اند و منبع تأمین هزینه آنها بیمه کمکهای پزشکی بوده است تعیین شود. داده‌های حاصل در جدول ۲.۱۶ ارائه شده و متغیرهایی که در این جدول آمده‌اند به صورت زیر توصیف شده‌اند:

REGION	ناحیه را نشان می‌دهد
NURSHOME	آسایشگاه سالمندان نمونه را نشان می‌دهد
PATIENT	بیمار نمونه را نشان می‌دهد
MEDICAID	در صورتی که بیمار به هزینه بیمه کمکهای پزشکی بستری شده باشد برابر با «۱» و در غیر اینصورت برابر با «۰» است
RGNHOMES	کل تعداد آسایشگاههای سالمندان در ناحیه است
NHADMISS	کل تعداد پذیرشها در آسایشگاه سالمندان طی سال ۱۹۹۷ است

جدول ۱.۱۶ تعداد آسایشگاههای سالمندان در پنج ناحیه یک ایالت

ناحیه	تعداد آسایشگاههای سالمندان در ناحیه	کل پذیرشها در آسایشگاههای سالمندان ناحیه طی سال ۱۹۹۷
۱	۱۲	۱۲۴۵
۲	۲۰	۳۸۸۷
۳	۱۱	۵۸۳
۴	۸	۴۰۰
۵	۶	۲۳۷۶

گامهای شش‌گانه فهرست شده بالا را مرور خواهیم کرد تا برآوردی مبتنی بر طرح برای کل تعداد بیماران پذیرش شده در آسایشگاههای سالمندان طی سال ۱۹۹۷ به دست آوریم که منبع تأمین هزینه آنها بیمه کمکهای پزشکی بوده است.

جدول ۲.۱۶ داده‌های حاصل از نمونه‌آسایشگاههای سالمندان

REGION	NURSHOME	PATIENT	MEDICAID	RGNHOMES	NHADMISS
1	1	1	1	12	123
1	1	2	1	12	123
1	1	3	1	12	123
1	1	4	0	12	123
1	1	5	1	12	123
1	2	1	0	12	89
1	2	2	0	12	89
1	2	3	1	12	89
1	2	4	0	12	89
1	2	5	0	12	89
2	1	1	1	20	231
2	1	2	0	20	231
2	1	3	1	20	231
2	1	4	0	20	231
2	1	5	1	20	231
2	2	1	0	20	187
2	2	2	0	20	187
2	2	3	0	20	187
2	2	4	1	20	187
2	2	5	0	20	187
3	1	1	1	11	43
3	1	2	1	11	43
3	1	3	1	11	43
3	1	4	0	11	43
3	1	5	1	11	43
3	2	1	1	11	49
3	2	2	1	11	49
3	2	3	1	11	49
3	2	4	1	11	49
3	2	5	0	11	49
4	1	1	0	8	56
4	1	2	1	8	56
4	1	3	1	8	56
4	1	4	0	8	56
4	1	5	0	8	56
4	2	1	0	8	38
4	2	2	0	8	38
4	2	3	0	8	38
4	2	4	0	8	38
4	2	5	1	8	38
5	1	1	1	6	359
5	1	2	0	6	359
5	1	3	1	6	359
5	1	4	1	6	359
5	1	5	0	6	359
5	2	1	0	6	460
5	2	2	1	6	460
5	2	3	0	6	460
5	2	4	1	6	460
5	2	5	0	6	460

۱. این یک نمونه خوشه‌ای طبقه‌بندی شده دو مرحله‌ای است که در مرحله اول آن، آسایشگاه‌های سالمندان با نمونه‌گیری تصادفی ساده بدون جایگذاری در داخل هر طبقه انتخاب می‌شوند و بیماران با نمونه‌گیری تصادفی ساده در داخل هر آسایشگاه انتخاب شده سالمندان برگزیده می‌شوند. متغیر طبقه‌بندی، REGION است، متغیر خوشه‌بندی NURSHOME است، متغیر RGNHOMES تعداد آسایشگاه‌های سالمندان در داخل هر ناحیه است و متغیری است که برای محاسبه تصحیح جامعه متناهی برای مرحله اول نمونه‌گیری به کار می‌رود، متغیر NHADMISS کل تعداد بیماران پذیرش شده طی سال ۱۹۹۷ در هر یک از آسایشگاه‌های سالمندان نمونه است و به عنوان متغیر مورد استفاده برای محاسبه تصحیح جامعه متناهی در مرحله دوم نمونه‌گیری به کار می‌رود.

۲. وزن نمونه‌گیری برای هر سابقه نمونه به صورت زیر به دست می‌آید:
فرض می‌کنیم $M_h =$ تعداد آسایشگاه‌های سالمندان در ناحیه h باشد.
پس، f_{1h} ، احتمال انتخاب شدن هر آسایشگاه سالمندان در نمونه، از فرمول زیر به دست می‌آید

$$f_{1h} = \frac{2}{M_h}$$

قرار می‌دهیم، $N_{hi} =$ تعداد پذیرشها در آسایشگاه سالمندان i در داخل طبقه h برای سال ۱۹۹۷.

پس کسر نمونه‌گیری مرحله دوم، f_{2hi} ، از فرمول زیر به دست می‌آید

$$f_{2hi} = \frac{5}{N_{hi}}$$

به این ترتیب، احتمال کلی گنجانیده شدن یک بیمار در نمونه از فرمول زیر به دست می‌آید

$$f_{hi} = \frac{2}{M_h} \times \frac{5}{N_{hi}}$$

و وزن نمونه‌گیری، w_{hi} ، از فرمول زیر محاسبه می‌شود

$$w_{hi} = \frac{M_h N_{hi}}{10} \quad (1.16)$$

۳. در این مثال تشریحی، فرض بر این است که پاسخگویی کامل است و هیچ برنامه‌ای برای پس طبقه‌بندی وجود ندارد.

۴. بررسی جدول ۲.۱۶ نشان می‌دهد که متغیرهای مربوط به طبقه‌بندی، خوشه‌بندی، و تصحیح جامعه‌متناهی برای هر سابقه نمونه موجود است. اگرچه در جدول ۲.۱۶ صریحاً نشان داده نشده است ولی برای هر سابقه نمونه یک متغیر، *WEIGHT* ایجاد کرده‌ایم که از معادله (۱.۱۶) به دست آمده است.
۵. برای این مثال تشریحی از SUDAAN استفاده خواهیم کرد زیرا STATA برای طرح‌های نمونه‌ای خوشه‌ای از تصحیح جامعه‌متناهی استفاده نمی‌کند. فرمانهای مناسب برای SUDAAN در زیر نشان داده شده‌اند:

```
PROC DESCRIPT DATA = "A:CH16ILLI" FILETYPE = SAS DESIGN = WOR
DEFF TOTALS;
NEST REGION NURSHOME;
TOTCNT RGNHOMES NHADMISS;
VAR MEDICAID;
WEIGHT WEIGHT;
SETENV COLWIDTH = 15;
SETENV DECWIDTH = 4;
```

اولین سطر فرمان، شیوه مناسب SUDAAN را برای برآورد کردن مجموع نشان می‌دهد، مجموعه داده‌های مناسب را مکان‌یابی می‌کند، نشان می‌دهد که پرونده اطلاعاتی مربوط به SAS است، طرح نمونه‌ای مناسب را تعیین می‌کند (هنگامی که تصحیح جامعه‌متناهی مورد نیاز باشد طرح بدون جایگذاری WOR لازم است).

سطر دوم فرمان یا گزاره "nest" به معنای تودرتو، متغیرهای طبقه‌بندی و خوشه‌بندی را نشان می‌دهد (اگر توضیح دیگری داده نشده باشد، اولین متغیر نشان داده شده در گزاره تودرتو به صورت متغیر طبقه‌بندی تعبیر می‌شود).

سطر سوم یا گزاره "totcnt" متغیرهای جامعه را که برای محاسبه تصحیح جامعه‌متناهی در هر مرحله از نمونه‌گیری مورد نیازند نشان می‌دهد.

سطر چهارم فرمان، متغیری (متغیرهایی) را مشخص می‌کند که قرار است برآورد (برآوردهایی) برای آن (آنها) تهیه شود.

سطر پنجم، متغیر وزن نمونه‌گیری را نشان می‌دهد. دو سطر آخر، مشخصات خروجی را ارائه می‌دهند.

۶. خروجی حاصل در زیر نشان داده شده است:

Number of observations read: 50 Weighted count: 8791
 Deominator degrees of freedom: 5

Variance Estimation Method: Taylor Series (WOR)
 by: Variable, One.

Variable		One --
MEDICAID	Sample Size	50.0000
	Weighted Size	8791.0000
	Total	4180.2000
	SE Total	1112.7529
	Mean	0.4755
	SE Mean	0.1052
	DEFF Mean # 4	2.1748
	DEFF Total # 4	3.1479

یافته‌هایی که در بالا نشان داده شده‌اند، برآورد می‌کنند که در طی سال ۱۹۹۷ تعداد ۴۱۸۰ بیمار در آسایشگاههای سالمندان این ایالت پذیرش شده‌اند که هزینه‌های آنها توسط بیمه کمکهای پزشکی پرداخت می‌شده است. خطای معیار این برآورد مجموع ۱۱۱۳ و اثر طرح آن برای مجموع ۳/۱۵ است که دلالت بر آن دارد که برای این متغیر مقدار بسیار زیادی خوشه‌بندی انجام شده است.

□

۲.۱۶ مسایل تحلیل آمارگیریهای نمونه‌ای «نوعی»

همه آمارگیریهای نمونه‌ای مشتمل بر برخی ویژگیهای طرح‌اند ولی الزاماً همه آنها را که در بالا فهرست شدند شامل نیستند، و بعضی دارای طرحهایی به مراتب پیچیده‌ترند. در اینجا سودمند است که یک آمارگیری نمونه‌ای «نوعی» توصیف شود. نوع طرحی که اینجا به تصویر خواهیم کشید از نوع اکثر وضعیتهای آمارگیری است که در زندگی واقعی روی می‌دهند. البته، وضعیتهایی پیچیده‌تر هم وجود دارند به خصوص در برنامه‌های آمارگیری بزرگی که به صورت جاری در سطح ملتها نگهداری می‌شوند. ولی، غالب اوقات حتی این آمارگیریها هم می‌توانند برای هدفهای تحلیل ساده شوند تا با وضعیت نوعی که در اینجا شرح خواهیم داد شباهت پیدا کنند و نتایج آنها غالباً به آنچه نزدیک خواهند بود که اگر پارامترهای طرح آمارگیری کل توصیف می‌شدند صحیح می‌بودند.

حالا به بررسی یک آمارگیری می‌پردازیم که طرح آن بسیار شبیه به طرح آمارگیری توصیف شده در بخش ۱.۱۶ است، ولی با اندازه نمونه‌ای بزرگتر، با برخی پیچیدگیهایی که قبلاً بررسی نشده‌اند، و با دستور کار تحلیل به مراتب پیچیده‌تر. به صورتی مشخصتر، طرح این آمارگیری به گونه‌ای است که همه عناصر جامعه در یکی از L طبقه دوه‌دو ناسازگار و فراگیر گروه‌بندی شده‌اند. در داخل طبقه

h ام، جامعه به M_h خوشه [به نام واحدهای نمونه‌گیری اولیه (PSUها)] تقسیم و از میان آنها m_h خوشه برای گنجاندن در نمونه انتخاب شده‌اند. در داخل هر خوشه نمونه، مرحله دوم نمونه‌گیری اجرا می‌شود و n_{hi} واحد شمارش انتخاب می‌شوند. به هر واحدی که از یک واحد نمونه‌گیری اولیه انتخاب می‌شود یک وزن نمونه‌گیری به صورتی که قبلاً گفته شد داده می‌شود. این وزن را می‌توان به صورت تعداد واحدهایی در جامعه در نظر گرفت که این واحد معرف آن است و در این مثال با طبقه و واحد نمونه‌گیری اولیه‌ای که واحد مزبور در آن قرار دارد تعیین می‌شود. وزن نمونه‌گیری مجدداً به صورت معکوس احتمال گنجانیده شدن واحد در نمونه محاسبه می‌شود. هنگامی که اندازه‌های جامعه معلوم باشد این وزن‌ها را می‌توان به صورت تصحیح جامعه متناهی در تحلیل وارد کرد.

مثال تشریحی: برای نشان دادن این مفاهیم، تحلیل داده‌های جمع‌آوری شده به عنوان بخشی از بررسی PAQUID^۱ را در نظر می‌گیریم که یک نمونه خوشه‌ای طبقه‌بندی شده از شهروندان سالخورده فرانسوی است [۱] و [۲]. ساکنانی که در تاریخ ۳۱ دسامبر ۱۹۸۷ دارای ۶۵ سال سن یا بیشتر بوده و در آسایشگاه‌ها به سر نمی‌بردند و در دو ناحیه یا «استان» در جنوب غربی فرانسه: ژیروند (با ۱۰۰۰۰ کیلومتر مربع وسعت و ۱۱۲۷۵۴۶ نفر جمعیت) و دوردون^۲ (با ۹۰۶۰ کیلومتر مربع مساحت و ۳۷۷۳۵۶ نفر جمعیت) زندگی می‌کردند در این بررسی گنجانیده شده بودند. در مجموع ۳۷۷۷ نفر از ساکنان محلی به منظور شناسایی عوامل پایه‌ای و طول عمر که ممکن بود با اختلال شناخت، زوال عقل، و بیماری آلزایمر در ارتباط باشند انتخاب شدند. داده‌های پایه‌ای در سالهای ۱۹۸۸-۱۹۸۹ جمع‌آوری شدند و آزمودنیها از آن پس تاکنون به صورت ادواری تحت بررسی بوده‌اند. استانهای ژیروند و دوردون^۲ به ترتیب از ۵۴۳ و ۵۵۵ ناحیه کوچکتر تشکیل شده‌اند که بخش نامیده می‌شوند. برای انتخاب آزمودنیهای بررسی، بخشهای داخل هر استان براساس اندازه جمعیت در یکی از چهار گروه زیر جای داده شدند:

گروه ۱: بخشهای دارای ۵۰۰۰۰ نفر جمعیت و بیشتر

گروه ۲: بخشهای دارای ۱۰۰۰۰ تا ۴۹۹۹۹ نفر جمعیت

گروه ۳: بخشهای دارای ۲۰۰۰ تا ۹۹۹۹ نفر جمعیت

گروه ۴: بخشهای دارای کمتر از ۲۰۰۰ نفر جمعیت

در داخل هر گروه، بخشهایی با احتمال متناسب با اندازه بخش انتخاب شدند. در استان دوردون^۲ هیچ بخشی با بیش از ۵۰۰۰۰ نفر جمعیت وجود نداشت. در مجموع، ۳۷ بخش از ژیروند و ۳۸ بخش از دوردون^۲ انتخاب شدند.

^۱ Personnes Ages Quid

در داخل هر بخش انتخاب شده، آزمودنیها به طور تصادفی از روی فهرستهای انتخاباتی هر بخش انتخاب و برحسب سن و جنس طبقه‌بندی شدند. از آزمودنیهای انتخاب شده درخواست شد تا موافقت کتبی خود را برای مشارکت در بررسی ارائه کنند. نتیجه آن شد که کل جامعه مورد بررسی شامل ۳۷۷۷ آزمودنی: ۲۷۹۲ نفر از ژیروند و ۹۸۵ نفر از دوردون^۱ی به دست آمد.

با همهٔ آزمودنیهای این بررسی از طریق تلفن یا (در صورت امکان) پست برای کسب اجازه به منظور مشارکت دادن آنها در بررسی تماس گرفته شد. با کسانی که موافقت خود را برای مشارکت اعلام کرده بودند (۶۸/۵٪) در منزلشان توسط یک روانشناس آموزش دیده برای به دست آوردن داده‌های پایه‌ای مصاحبه شد. یک پرسشنامه دارای ساختار به هر شرکت کننده اختصاص داده شد. به آزمودنیها در فاصله‌های زمانی منظم مراجعه می‌شد تا با توجه به اختلال شناخت و زوال عقل غربالگری شوند.

واضح است که این یک نمونهٔ تصادفی ساده نیست، بلکه، در عوض، یک آمارگیری نمونه‌ای است که بسیاری از ویژگیهایی را که قبلاً به اجمال مورد بررسی قرار گرفته‌اند با خود دارد. در بحثی که در پی خواهد آمد نمادهای زیر به کار گرفته خواهند شد:

- h به یکی از ۷ طبقه اشاره دارد (که برحسب استان و گروه تعریف شده است)
- M_h تعداد خوشه‌ها (یا بخشها)یی است که طبقهٔ h ام را تشکیل می‌دهند، $h = 1, \dots, 7$
- m_h تعداد خوشه‌های انتخاب شده از طبقهٔ h ام است، $h = 1, \dots, 7$
- i به تعداد بخش در داخل طبقه اشاره می‌کند
- N_{hi} تعداد افراد ۶۵ سال به بالاست که در i امین بخش نمونه‌گیری شده در داخل طبقهٔ h (براساس داده‌های سرشماری) زندگی می‌کنند
- N_h تعداد افراد ۶۵ سال به بالاست که در طبقهٔ h (براساس داده‌های سرشماری) زندگی می‌کنند
- n_{hi} تعداد افرادی است که عملاً در بخش i ، داخل طبقهٔ h بررسی شده‌اند.

جدول ۳.۱۶ خلاصه‌ای از فرایند انتخاب را نشان می‌دهد.

می‌توان چنین تصور کرد که ۷ طبقه داریم و نمونه‌ها در داخل هر طبقه انتخاب می‌شوند. در ژیروند فقط سه بخش با بیش از ۵۰۰۰۰ نفر جمعیت وجود داشت (طبقهٔ ۱) و هر یک از این بخشها انتخاب شدند. در دوردون^۱ فقط ۲ بخش با جمعیتی بین ۱۰۰۰۰ و ۴۹۹۹۹ نفر وجود داشت (طبقهٔ ۳) و هر دو انتخاب شدند. در مورد این دو طبقه، چون $m_h = M_h$ ، احتمال انتخاب یک واحد خاص (n_{hi}/N_{hi}) است، بنابراین، وزن نمونه‌گیری برای این دو طبقه $1/(n_{hi}/N_{hi}) = N_{hi}/n_{hi}$ خواهد بود.

برای پنج طبقه دیگر، احتمال تقریبی انتخاب یک واحد خاص برابر است با حاصلضرب احتمال تقریبی انتخاب بخش مورد نظر با استفاده از احتمال متناسب با اندازه (PPS) طرح نمونه‌گیری، $m_h \times (N_{hi}/N_h)$ (اثبات این معادله را در پایان همین فصل ببینید)، ضربدر احتمال انتخاب فرد در بخش انتخاب شده، n_{hi}/N_{hi} . یعنی احتمال انتخاب یک نفر در این پنج طبقه به صورت زیر است

$$\left(m_h \times \frac{N_{hi}}{N_h}\right) \times \left(\frac{n_{hi}}{N_{hi}}\right) = \frac{m_h n_{hi}}{N_h}$$

نتیجه می‌شود که وزن نمونه‌گیری در این طبقه‌ها $(N_h/(m_h n_{hi}))$ است.

جدول ۳.۱۶ انتخاب آزمودنی‌های نمونه از استانهای ژیروند و دوردون

استان	گروه	طبقه	M_h	m_h	$\sum_{i=1}^{m_h} N_{hi}$	$\sum_{i=1}^{m_h} n_{hi}$
ژیروند	۱	۱	۳	۳	۴۹۷۸۶	۱۲۳۱
ژیروند	۲	۲	۱۴	۴	۳۵۲۵۳	۴۷۳
دوردون	۲	۳	۲	۲	۱۱۷۶۴	۷۱
ژیروند	۳	۴	۷۳	۱۰	۳۷۰۲۴	۳۶۱
دوردون	۳	۵	۲۶	۱۳	۱۷۶۸۷	۳۱۱
ژیروند	۴	۶	۴۵۳	۲۰	۴۴۹۸۰	۷۲۷
دوردون	۴	۷	۵۲۷	۲۳	۴۶۲۵۰	۶۰۳
مجموع			۱۰۹۸	۷۵	۲۴۲۷۴۴	۳۷۷۷

اگرچه این وزنها، همان طور که موردنظر است، به این مفهوم صحیح‌اند که تعداد افراد موجود در جامعه‌ای را به تصویر می‌کشند که هر آزمودنی موردنظر در بررسی، معرف آن است، ولی مجموع وزنها داخل زیرگروه‌های سنی - جنسی مشخص شده، ممکن است درست به مجموع اندازه‌های جامعه‌ای معلوم در داخل این زیرگروه‌ها نرسد. برای اصلاح این وضعیت، یک تعدیل پس طبقه‌بندی لازم است تا وزن اولیه را به گونه‌ای اصلاح کند که مجموع وزنها برای همه افراد نمونه‌گیری شده در هر یک از زیرگروه‌های سنی - جنسی با اندازه‌های جامعه‌ای معلوم در داخل هر یک از این زیرگروه‌ها دقیقاً یکسان شود. زیرگروه‌ها عبارت‌اند از:

مردان در سنین ۶۵ تا ۷۴ سال	زنان در سنین ۶۵ تا ۷۴ سال
مردان در سنین ۷۵ تا ۸۴ سال	زنان در سنین ۷۵ تا ۸۴ سال
مردان در سنین ۸۵ سال و بیشتر	زنان در سنین ۸۵ سال و بیشتر

مثلاً در طبقه ۱ ژیروند ۱۰۱۰۵ مرد ۶۵ تا ۷۴ سال در فهرستهای انتخاباتی وجود داشتند. این رقم نشان دهنده ۲۰/۳٪ همه آزمودنیهای موجود در فهرست انتخاباتی آن طبقه بود. ولی در نمونه، ۲۸۴ مرد بین سنین ۶۵ تا ۷۴ سال بودند که معرف ۲۳/۰۷٪ نمونه مربوط به آن طبقه‌اند. بنابراین، می‌توانیم هر وزن را با ضرب کردن آن در نسبت $\frac{0/203}{0/2307} = 0/88$ تعدیل کنیم. اگر این عاملهای تعدیل را c_{hijk} بنامیم که در آن h نشانگر طبقه، i نشانگر واحد نمونه‌گیری اولیه، j نشانگر زیرگروه، و k نشانگر فرد باشد، در آن صورت می‌توان وزنهای آماری نهایی، w_{hijk} ، را به صورت زیر بیان کرد

$$w_{hijk} = \left(\frac{N_h}{m_h n_{hi}} \right) \times c_{hijk}$$

□

بحث بالا نشان می‌دهد که بررسی PAQUID را می‌توان به عنوان یک طرح طبقه‌بندی شده چندمرحله‌ای توصیف کرد که در مرحله اول نمونه‌گیری، بخشها (یا واحدهای نمونه‌گیری اولیه) از طبقه‌ها انتخاب می‌شوند و در مرحله دوم، آزمودنیها از روی فهرستهای انتخاباتی در داخل بخش منتخب به طور تصادفی انتخاب می‌شوند. همان طور که در فصلهای قبلی بیان شد، تا همین اواخر تحلیل داده‌های آمارگیری از این نوع نسبتاً مشکل بود زیرا بیشتر نرم‌افزارهای آماری فرض را بر این می‌گیرند که داده‌ها با نمونه‌گیری تصادفی ساده انتخاب شده‌اند. این فرض می‌تواند هنگام تحلیل داده‌های آمارگیری نمونه‌ای که از طرح نمونه‌گیری تصادفی ساده به دست نیامده‌اند باعث بروز خطاهایی در برآورد کردن و استنباط شود. این قبیل برنامه‌ها ممکن است برآوردهای نقطه‌ای اریب تولید کنند و ممکن است خطاهای معیار این برآوردها را بسیار کم برآورد کنند. همان طور که بارها در فصلهای پیشین این کتاب ذکر شد، برنامه‌هایی همچون SUDAAN و STATA امکان تحلیل داده‌های آمارگیری نمونه‌ای را به شکلی مناسب فراهم می‌سازند.

برای تحلیل این داده‌ها از قابلیت‌های آمارگیری نمونه‌ای STATA استفاده می‌کنیم. برای این منظور، به سادگی مشخص می‌کنیم که هفت طبقه وجود دارند و بخشها واحدهای نمونه‌گیری اولیه را در داخل آنها تشکیل می‌دهند. همچنین وزنهای آماری، w_{hijk} ، را تهیه می‌کنیم که با پیروی از فرایندی که در بالا توصیف شد محاسبه می‌شوند.

پارامترهای طرح آمارگیری برای STATA با استفاده از فرمانهای `svyset` مشخص می‌شوند. در این مثال، متغیرهای شامل اطلاعات مربوط به طبقه، وزن نمونه‌گیری، و واحد نمونه‌گیری اولیه به ترتیب STRATUM، WEIGHT و PSU نامگذاری شده‌اند. این مشخصات به صورت بخشی از مجموعه داده‌ها ذخیره شده‌اند و هر بار که یک تحلیل آمارگیری اجرا شود فراخوانده می‌شوند:

```
. svyset strata STRATUM
. svyset pweight WEIGHT
. svyset psu PSU
```

هدف از این تحلیل آن است که وقوع زوال عقل در میان آن دسته از افرادی که در زمان آزمایش آغازین آمارگیری زوال عقل نداشته‌اند در طی سه سال برآورد شود. به این دلیل فقط آن دسته از آزمودنیها را که در زمان مشاهده آغازین دچار زوال عقل نبوده‌اند ($n = 2273$) در نظر می‌گیریم. به عنوان اولین تحلیل، محاسبه نسبت این آزمودنیها را که نمره آزمایش کوچک وضعیت ذهنی MMSE¹ آنان برابر با ۲۱ یا کمتر بوده است در نظر می‌گیریم. آزمایش کوچک وضعیت ذهنی آزمایشی کوتاه است که اختلال شناختی را غربالگری می‌کند و نمره آن از صفر تا ۳۰ است که نمره‌های پایینتر نشانه احتمال بیشتر وقوع زوال عقل است. با در نظر گرفتن متغیر mmse01 به عنوان نشانگر نمره ۲۱ یا کمتر برای آزمایش کوچک وضعیت ذهنی، می‌بینیم که نادیده گرفتن طرح آمارگیری به برآورد زیر برای نسبت و بازه اطمینان ملازم با آن منتهی می‌شود:

Variable	obs	Mean	Std. Err.	[95% Conf. Interval]
Mmse01	2242	.0816236	.0057836	.0702818 .0929653

برای منظور کردن پارامترهای طرح آمارگیری، فرمان زیر را برای STATA صادر می‌کنیم:

```
. svymeans mmse01
```

و نتایج زیر را به دست می‌آوریم:

Survey mean estimation

```
pweight: finalwt      Number of obs =      2242
Strata:   newstrat     Number of strata =       7
PSU:     parish       Number of PSUs =      75
                          Population size = 14863.41
```

Mean	Estimate	Std. Err.	[95% Conf. Interval]	Deff
mmse01	.082957	.0079147	.0671634 .0987506	1.845315

توجه داشته باشید که در حالی که برآوردهای نقطه‌ای کاملاً قابل مقایسه‌اند (تقریباً ۸٪ دارای نمره ۲۱ یا کمتر برای آزمایش کوچک وضعیت ذهنی)، خطای معیار نسبت برآورد شده براساس طرح آمارگیری، ۰/۰۰۷۹، در مقایسه با خطای معیار برآورد شده ۰/۰۰۵۸ حاصل از نادیده گرفتن پارامترهای

¹ Mini – Mental State Examination

طرح آمارگیری است. اثر طرح (که در خروجی STATA به صورت "Deff" نشان داده می‌شود) نسبت واریانس پارامتر برآورد شده با در نظر گرفتن پارامترهای طرح آمارگیری به واریانس پارامتر برآورد شده با فرض نمونه‌گیری تصادفی ساده است. در این حالت

$$\text{Deff} = \frac{0.0079^2}{0.0058^2} = 1.85$$

با اینکه برآوردهای نقطه‌ای ممکن است در دو طرح تحلیل قابل مقایسه باشند، خطاهای معیار میانگینها در تحلیل مبتنی بر طرح به مراتب بیشترند. این همان چیزی است که در مورد برآوردهای نسبت نمونه‌ای انتظار می‌رفت. واضح است که نادیده گرفتن پارامترهای طرح آمارگیری به کم برآورد کردن واریانس نمونه‌ای تا میزان ۸۵٪ و به دست آوردن بازه‌های اطمینان بیش از حد باریک، منتهی می‌شود.

نوعاً، اثرهای طرح برای برآورد میانگینها، مجموعها، و نسبتها به مراتب بیشتر از یک است. اثر طرح می‌تواند به عنوان مقیاس تورم در واریانس تصور شود که ناشی از همگنی درون خوشه‌هاست. اثر طرح را می‌توان به صورت $(\bar{n} - 1) + \delta_y$ بیان کرد که در آن، δ_y ضریب همبستگی درون خوشه‌ای (یعنی ضریب همبستگی درون رده‌ای)، و \bar{n} متوسط تعداد واحدهای نمونه‌گیری شده به ازای خوشه است. همان طور که قبلاً بیان شد، دامنه ضریب همبستگی درون خوشه‌ای می‌تواند از مقادیر منفی کوچک (هنگامی که داده‌های داخل خوشه‌ها به شدت ناهمگن باشند) تا یک (هنگامی که داده‌های داخل خوشه‌ها بسیار همگن باشند) کشیده شود. فقط در موارد بسیار نادر که داده‌ها در داخل خوشه‌ها بسیار ناهمگن باشند اثر طرح کمتر از یک خواهد بود.

یک هدف مهم بررسی مزبور این بود که بررسی شود که آیا بین مصرف شراب و وقوع زوال عقل رابطه‌ای وجود دارد یا نه، و اگر وجود دارد برآورد قدرت این رابطه چقدر است. به طور مشخصتر، برآمد مورد توجه در این بررسی این بود که آیا آزمودنی مزبور در مدت زمان پیگیری ۳ ساله دچار زوال عقل بوده است یا نه و عامل مخاطره اصلی موردنظر مصرف شراب با گروه‌بندی در رسته‌های زیر بوده است یا نه :

$$\left. \begin{array}{l} 0 \text{ بدون مصرف شراب} \\ 1 \text{ مصرف شراب تا } \frac{1}{4} \text{ لیتر در روز} \\ 2 \text{ مصرف شراب بیش از } \frac{1}{4} \text{ لیتر در روز} \end{array} \right\} = \text{شراب}$$

در این تحلیل، متغیر وابسته، داشتن یا نداشتن زوال عقل است که بر تعداد زیادی از آزمایشهای عصبی - روانی متکی است که به مراتب از آزمایش کوچک وضعیت ذهنی پیچیده‌ترند. جدول‌بندی

مقاطع وضعیت زوال عقل برای دوره سه ساله پیگیری (که براساس این تعداد کثیر از آزمایشها تعیین می‌شود) در مقابل مصرف شراب بدون توجه به طرح آمارگیری در جدول ۴.۱۶ نشان داده شده است. جدول ۵.۱۶ نشان می‌دهد که همان نسبتهای بخت و بازه‌های اطمینان را می‌توان با استفاده از مدل رگرسیونی لوژستیک تحت فرض نمونه‌گیری تصادفی ساده به دست آورد.

جدول ۴.۱۶ پیوند مصرف شراب با وقوع زوال عقل (تحلیل مبتنی بر مدل)

بازه‌های اطمینان ۹۵ درصد	نسبتهای بخت (OR)*	وقوع زوال عقل				مصرف شراب
		نه درصد	تعداد	آری درصد	تعداد	
	۱/۰۰۰	۹۵/۱	۹۲۳	۴/۹	۴۸	هیچ
۰/۶۸۴-۱/۵۶۱	۱/۰۳۳	۹۴/۹	۸۷۵	۵/۱	۴۷	روزی ۱/۴ لیتر و کمتر
۰/۰۷۳-۰/۵۷۱	۰/۲۰۵	۹۸/۹	۳۷۶	۱/۱	۴	بیشتر از ۱/۴ لیتر در روز
			۲۱۷۴		۹۹	مجموع

* در مقایسه با هیچ به عنوان گروه مرجع.

با استفاده از ویژگیهای طرح آمارگیری، یک رگرسیون لوژستیک اجرا می‌شود که DEMENTIA (زوال عقل) متغیر وابسته و دو متغیر نشانگر برای مصرف شراب (WINE_1 و WINE_2) متغیرهای مستقل آن هستند. این کار در نرم‌افزار STATA با فرمان زیر اجرا می‌شود:

. svylogit DEMENTIA WINE_1 WINE_2, or

از این فرمان خروجی زیر به دست می‌آید:

Survey logistic regression						
pweight:	WEIGHT	Number of obs =	2273			
Strata:	STRATUM	Number of strata =	7			
PSU:	PSU	Number of PSUs =	75			
		Population size =	148779.94			
		F (2, 67) =	5.99			
		Prob > F =	0.0041			
DEMENTIA	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
WINE_1	1.040389	.2807133	0.147	0.884	.6072504	1.7824760
WINE_2	.1691458	.0885213	-3.395	0.001	.0595280	.4806197

جدول ۶.۱۶ این تحلیل را در چارچوبی مشابه چارچوب ارائه شده در جدول ۵.۱۶ ارائه می‌دهد.

جدول ۵.۱۶ تحلیل رگرسیونی لوژستیک از مصرف شراب و وقوع زوال عقل
با فرض نمونه‌گیری تصادفی ساده (تحلیل مبتنی بر مدل)

بازه‌های اطمینان	خطای معیار	نسبت بخت	مصرف شراب
۹۵ درصد	نسبت بخت		
—	—	۱/۰۰۰	هیچ
[۰/۶۸۴, ۱/۵۶]	۰/۲۱۷	۱/۰۳۳	روزی $\frac{1}{4}$ لیتر و کمتر
[۰/۰۷۳, ۰/۵۷۱]	۰/۱۰۷	۰/۲۰۵	بیشتر از $\frac{1}{4}$ لیتر در روز

نسبتهای بخت در هر دو تحلیل شبیه‌اند ولی اثر حافظ بودن شراب در تحلیلی که طرح نمونه‌گیری را به صورتی صحیح گزارش می‌کرد ۱۸٪ بیشتر به نظر می‌رسد. جالب توجه است که این مصرف متوسط یا زیاد شراب، ولی نه کم آن است که به نظر می‌رسد اثر حفاظت در برابر زوال عقل داشته باشد. به علاوه، اثرهای طرح برای ضریبهای رگرسیونی به شدت اثرهای آن برای میانگینها یا نسبتها نیستند. در برخی موارد، اثرهای طرح عملاً کمتر از ۱ بود. جالب است که اثرهای طرح برای ضریبهای رگرسیونی را می‌توان با $1 + (\bar{n} - 1)\delta_x\delta_y$ تقریب زد. چون این عبارت به حاصلضرب ضریبهای همبستگی درون خوشه‌ای برای هر دو متغیر وابسته و مستقل بستگی دارد، که هر دو طبق تعریف، کوچکتر یا برابر با ۱ هستند، بزرگی این اثر کوچکتر یا برابر است با آنچه که در مورد میانگینها، مجموعها، یا نسبتها دیده شد. به علاوه، توجه به این نکته مهم است که δ_x و δ_y الزاماً در یک راستا نیستند. امکان دارد که داده‌های داخل خوشه‌ها نسبت به یک متغیر، ناهمگن و نسبت به متغیری دیگر همگن باشند. در این حالت، حاصلضرب ضریبهای همبستگی درون خوشه‌ای، منفی و اثر طرح حاصل از آن کمتر از ۱ خواهد بود.

جدول ۶.۱۶ تحلیل رگرسیونی لوژستیک از مصرف شراب و وقوع زوال عقل
با در نظر گرفتن پارامترهای آمارگیری نمونه‌ای (تحلیل مبتنی بر طرح)

بازه‌های اطمینان	خطای معیار	نسبت بخت	مصرف شراب
۹۵ درصد	نسبت بخت		
—	—	۱/۰۰۰	هیچ
[۰/۶۰۷, ۱/۷۸۲]	۰/۲۸۱	۰/۰۴۰	روزی $\frac{1}{4}$ لیتر و کمتر
[۰/۰۶۰, ۰/۴۸۱]	۰/۰۸۹	۰/۱۶۹	بیشتر از $\frac{1}{4}$ لیتر در روز

اخیراً، سازمانهای آمارگیری از قبیل مرکز ملی آمارهای بهداشتی^۱ و اداره سرشماری امریکا^۲ داده‌هایی را از آمارگیری بهداشت ملی و بررسی تغذیه (NHANES III)^۳ و آمارگیری مصاحبه‌ای بهداشت ملی (NHIS)^۴ از طریق سی‌دی‌رام CD-ROM برای استفاده پژوهشگران علاقمند توزیع کرده‌اند. در مقابل غنای این قبیل پایگاههای اطلاعاتی و فقدان تاریخی نسبی دسترسی به نرم‌افزارهای مناسبی که بتوانند با کفایت طرح آمارگیری را در بررسیها گزارش کنند، پژوهشگران پیچیدگیهای آمارگیری را به سادگی نادیده گرفته‌اند و داده‌ها را به صورتی تحلیل کرده‌اند که گویی نتیجه یک نمونه تصادفی ساده‌اند. دسترسی به برنامه‌های جدید همچون STATA و SUDAAN قابلیت‌های تحلیلی تازه‌ای که برای اجرای تحلیلهای مناسب ضروری‌اند در اختیار تحلیلگران داده‌ها قرار داده است.

۳.۱۶ خلاصه

در این فصل، یک شیوه شش مرحله‌ای برای اجرای تحلیل مبتنی بر طرح داده‌های حاصل از آمارگیریهای نمونه‌ای ارائه شد تا بتوان آن را با هر نرم‌افزاری که قابلیت اجرای چنین تحلیلی را داشته باشد مورد استفاده قرار داد. این شیوه را با استفاده از یک طرح نمونه‌ای خوشه‌ای طبقه‌بندی شده دومرحله‌ای نسبتاً ساده نشان دادیم. سپس با استفاده از یک مثال که طرح نمونه‌ای پیچیده‌تری دارد نشان دادیم که چگونه غفلت در استفاده از تحلیل مبتنی بر طرح می‌تواند به نتایج گمراه کننده بینجامد.

پیوست فنی

اثبات این که احتمال انتخاب یک واحد نمونه‌گیری اولیه (PSU) ویژه، i ، تحت نمونه‌گیری با احتمال متناسب با اندازه از فرمول زیر به دست می‌آید

$$P\{i \in \text{نمونه}\} = \frac{N_i}{N} m$$

در زیر آمده است. در این فرمول

$$N_i = \text{تعداد واحدهای شمارش در واحد نمونه‌گیری اولیه (PSU) } i$$

$$N = \text{کل تعداد واحدهای شمارش در جامعه}$$

$$m = \text{کل تعداد واحدهای نمونه‌گیری اولیه (PSU) در نمونه}$$

¹ National Center for Health Statistics

² U.S. Bureau of the Census

³ National Health and Nutrition Examination Survey

⁴ National Health Interview Survey

اثبات. فرض کنید نمونه‌گیری واحدهای نمونه‌گیری اولیه (PSU ها) با جایگذاری است:

$$P\{\text{نمونه } PSU i \notin\} = \left(1 - \frac{N_i}{N}\right)^m \approx 1 - m \frac{N_i}{N}$$

(بسط قضیه دو جمله‌ای تا جمله اول)

پس

$$P\{\text{نمونه } PSU i \in\} = 1 - P\{\text{نمونه } PSU i \notin\} = \frac{N_i}{N} m$$

کتابشناسی

The following articles describe the design and basic findings of the PAQUID Study:

1. Lemeshow, S., Letenneur, L., Dartigues, J. F., Lafont, S., Orgogozo, J. M., and Commenges, D., An illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID study. *American Journal of Epidemiology*, 148(3): 298-306, 1998.
2. Barberger-Gateau, P., Dartigues, J. F., Commenges, D., Gagnon, M., Letenneur, L., Canet, C., Miquel, J. L., Nejjari, C., Tessier, J. F., Berr, C., Dealberto, M. J., Decamps, A., Alperovitch, A., and Salamon, R., Paquid: An interdisciplinary epidemiologic study of cerebral and functional aging. *Annals of Gerontology*, 383-392, 1992.
3. Letenneur, L., Commenges, D., Dartigues, J. F., and Barberger-Gateau, P. Incidence of dementia and Alzheimer's disease in elderly community residents of South-Western France. *International Journal of Epidemiology*, 23: 1256-1261, 1994.

The following article furnishes another example of the problems that result from ignoring design features in the analysis of data from sample surveys.

4. Brogan D., Software Packages for Analysis of Sample Survey Data, Misuse of Standard Packages. In *Encyclopedia of Biostatistics*, Armitage, P. A., and Colton, T., Eds., Wiley, Chichester, U.K., 1998.

Also many of the references cited in Chapter 12 deal with general issues in analysis of data from complex sample surveys.