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 ۵

طبقه‌بندی و نمونه‌گیری تصادفی طبقه‌بندی شده

در دو نوع نمونه‌گیری که تا اینجا مورد بحث قرار گرفتند یعنی در نمونه‌گیری تصادفی ساده و نمونه‌گیری سیستماتیک، لازم است که نمونه‌ای از جامعه به صورت یک کل گرفته شود و در هیچ یک از آنها لازم نیست که زیرحوزه‌ها یا زیرگروههای جامعه قبل از گرفتن نمونه شناسایی شوند. ولی گاهی اوقات می‌توان چارچوب نمونه‌گیری را به گروهها یا *طبقه‌هایی* افراز و نمونه‌گیری را جداگانه در داخل هر طبقه اجرا کرد. این طرح نمونه‌گیری را *نمونه‌گیری طبقه‌بندی شده* می‌نامند. اگر برای انتخاب نمونه در داخل هر طبقه از نمونه‌گیری تصادفی ساده استفاده شود طرح نمونه را *نمونه‌گیری تصادفی طبقه‌بندی شده* می‌نامند.

در این فصل برخی «مبانی» طبقه‌بندی و نمونه‌گیری تصادفی طبقه‌بندی شده را معرفی می‌کنیم. اینها شامل بحث دربارهٔ روشهای گرفتن نمونه، تعریف اصطلاحها و نمادگذاریهای مشخصه‌های جامعه و نمونه که در بحث طرحهای نمونه با طبقه‌بندی به کار رفته‌اند، و برآوردهای بنیادی مشخصه‌های جامعه خواهند بود. در فصل بعد با جزئیات بیشتری به روشهای استفاده از طرحهای طبقه‌بندی از راههایی که احتمال تولید قابل اعتمادترین برآوردها بیشتر است می پردازیم:

۱.۵ نمونهٔ تصادفی طبقه‌بندی شده چیست؟

همان طور که در بالا اشاره شد، *نمونهٔ تصادفی طبقه‌بندی شده* یک برنامهٔ نمونه‌گیری است که در آن جامعه به L طبقهٔ دو به دو ناسازگار و جامع طبقه‌بندی می‌شود، و یک نمونهٔ تصادفی ساده متشکل از n_h عنصر از داخل هر طبقهٔ h گرفته می‌شود. نمونه‌گیری در داخل هر طبقه به طور مستقل اجرا می‌شود. در واقع می‌توانیم طرح نمونه‌گیری تصادفی ساده را به صورت L نمونهٔ تصادفی سادهٔ مجزا تصور کنیم.

از نظر عملیاتی، نمونهٔ تصادفی طبقه‌بندی شده به همان طریق نمونهٔ تصادفی ساده گرفته می‌شود ولی نمونه‌گیری در داخل هر طبقه به طور جداگانه و مستقل انجام می‌گیرد. اگر $N_1, N_2, \ldots, N_L$ معرف تعداد واحدهای نمونه‌گیری در داخل هر طبقه و $n_1, n_2, \ldots, n_L$ معرف تعداد واحدهای نمونه‌گیری انتخاب شده به طور تصادفی در داخل هر طبقه باشند، آنگاه، تعداد کل نمونه‌های تصادفی طبقه‌بندی شدهٔ ممکن برابر خواهد بود با:

$$\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \cdots \times \binom{N_L}{n_L}$$

که از $\binom{N}{n}$، کل تعداد نمونه‌های تصادفی سادهٔ ممکن، کوچکتر بوده و یا با آن برابر است.

برای مثال، اگر سه طبقه داشته باشیم و $N_1 = ۳$، $N_2 = ۵$ و $N_3 = ٦$ باشد تعداد کل نمونه‌های ممکن متشکل از $n_1 = ۱$ عنصر از طبقهٔ ۱، $n_2 = ۲$ عنصر از طبقهٔ ۲ و $n_3 = ٤$ عنصر از طبقهٔ ۳ چنین است:

$$\binom{۳}{۱} \times \binom{۵}{۲} \times \binom{۶}{۴} = ۴۵۰$$

تعداد کل نمونه‌های تصادفی ساده متشکل از ۷ عنصر از ١٤ عنصر جامعه عبارت است از

$$\binom{۱۴}{۷} = ۳۴۳۲$$

احتمال انتخاب شدن یک عنصر در نمونه بستگی دارد به طبقهٔ خاصی که آن عنصر در آن گروه‌بندی شده است و می‌توان نشان داد که برابر است با $\frac{n_h}{N_h}$ (در صورتی که آن عنصر در طبقهٔ h باشد). در مثالی که هم اکنون شرح دادیم، احتمال انتخاب شدن یک عنصر برابر است با $\frac{۱}{۳}$ برای عناصر طبقهٔ ۱، $\frac{۲}{۵}$ برای عناصر طبقهٔ ۲ و $\frac{۴}{۶}$ برای عناصر طبقهٔ ۳.

۲.۵ چگونگی گرفتن نمونهٔ تصادفی طبقه‌بندی شده

همان طور که در بالا شرح داده شد، نمونهٔ تصادفی طبقه‌بندی شده صرفاً با گرفتن نمونه‌های تصادفی ساده متشکل از n_h عنصر ($h = ۱,\ldots,L$) به طور مستقل در داخل هر طبقه تهیه می‌شود. انتخاب عملی نمونه، با استفاده از هر یک از روشهایی که قبلاً برای گرفتن نمونهٔ تصادفی ساده شرح داده شد اجرا می‌شود.

مثلاً، اگر کسی بخواهد از گزارشهای ذخیره شده در پروندهٔ اطلاعاتی رایانه نمونه‌ای بگیرد می‌تواند از یک الگوریتم برای گرفتن نمونهٔ تصادفی طبقه‌بندی شده که در کتاب راهنمای SAS [۷] توصیف شده است استفاده کند. همچنین مدول تکمیلی SYSTAT، SAMPLE نیز می‌تواند برای گرفتن نمونهٔ تصادفی طبقه‌بندی شده مورد استفاده قرار گیرد.

۳.۵ چرا نمونه‌گیری طبقه‌بندی شده؟

نمونه‌گیری طبقه‌بندی شده به این دلیل در انواع خاصی از آمارگیریها به کار می‌رود که آسانی درک نمونه‌گیری تصادفی ساده را با افزایش قابل توجهی در قابلیت اعتماد بالقوه ترکیب می‌کند. هر گاه بخواهیم برآوردهای جداگانه‌ای از پارامترهای جامعه برای هر یک از زیرحوزه‌ها در داخل کل جامعه به دست آوریم و علاوه بر آن بخواهیم مطمئن باشیم که نمونهٔ ما نمایندهٔ جامعه است استفاده از این فن راحت است. مثلاً، فرض کنید می‌خواهیم تعداد کل تختها را در بیمارستانهای یک ایالت مشخص برآورد کنیم. می‌دانیم که اکثر بیمارستانها، اندازه‌های کوچک یا متوسط دارند و تنها چندتایی بیمارستانهای بزرگ موجودند. می‌دانیم که این بیمارستانهای بزرگ بخش قابل توجهی از کل تعداد تختها را دارند.

حالا فرض کنید تصمیم بر این است که یک نمونهٔ تصادفی ساده از بیمارستانهای این ایالت انتخاب کنیم، تعداد تختها را در هر یک از نمونه‌هایی که به این ترتیب انتخاب شده‌اند تعیین، و با استفاده از روشهای فصل ۳، کل تعداد تختها را در تمام بیمارستانهای سراسر ایالت برآورد کنیم. مشکل این شیوه آن است که شانس زیادی وجود دارد که نمونهٔ ما تعداد بسیار زیاد یا بسیار کمی از بیمارستانهای بسیار بزرگ را در خود جای دهد. در نتیجه ممکن است نمونه به صورتی مناسب معرف جامعه نباشد.

راه حل ما برای این مشکل آن است که واحدهای نمونه‌گیری (بیمارستانها) را قبل از نمونه‌گیری بر اساس اندازه (یعنی کوچک، متوسط، بزرگ) در سه گروه طبقه‌بندی کنیم و سپس با استفاده از فنون نمونه‌گیری تصادفی ساده، از هر یک از گروههای سه‌گانه تعداد معینی از بیمارستانها را انتخاب نماییم.

سپس برآورد تعداد کل تختها را می‌توان از ترکیب نتایج سه طبقه به دست آورد. این، اساس نمونه‌گیری طبقه‌بندی شده است.

برای نشان دادن ایده‌ها و مزایای طبقه‌بندی به یک مثال نگاهی می‌اندازیم.

مثال تشریحی: فرض کنید جاده‌ای با ۲٤ مایل طول از مناطقی عبور می‌کند که می‌توانند به صورت شهری و روستایی طبقه‌بندی شوند و این جاده به هشت قطعه تقسیم شده است که هر یک دارای طولی برابر ۳ مایل است. نمونه‌ای شامل سه قطعه گرفته شده و در هر قطعهٔ نمونه‌گیری شده تجهیزات ویژه‌ای نصب گردیده است تا مجموع مایل‌های (مسافتهای) طی شده توسط وسایل نقلیه، اتومبیل و کامیون، در آن بخش طی یک سال خاص شمارش شود. به علاوه، گزارشی از تمام حوادثی که در هر قطعهٔ نمونه اتفاق می‌افتد نیز ثبت شود.

مقدار مسافت طی شده توسط کامیونها و تعداد حوادثی که در آن کامیون دخالت داشته است طی یک دورهٔ معین برای هر یک از قطعه‌های هشتگانهٔ جامعه در جدول ۱.۵ ارائه شده است.

فرض کنید یک نمونهٔ تصادفی ساده متشکل از سه قطعه را به منظور برآورد کل تعداد کامیون ـ مایل طی شده در جاده گرفته‌ایم. برای یک جامعهٔ متشکل از هشت قطعه، ٥٦ نمونهٔ ممکن سه قطعه‌ای وجود دارند. توزیع نمونه‌گیری تعداد کامیون ـ مایل طی شده در جاده در جدول ۲.۵ ارائه شده است.

جدول ۱.۵ کامیون ـ مایل و تعداد حوادثی که در آن کامیون مداخله داشته است بر حسب قطعه‌های جاده

قطعه	نوع	تعداد کامیون ـ مایل (×۱۰۰)	تعداد حوادث با مداخلهٔ کامیون
۱	شهری	٦۳۲۷	۸
۲	روستایی	۲٥٥٥	٥
۳	شهری	۸٦۹۱	۹
٤	شهری	۷۸۳٤	۹
٥	روستایی	۱٥۸٦	٥
٦	روستای	۲۰۳٤	۱
۷	روستایی	۲۰۱٥	۹
۸	روستایی	۳۰۱۲	٤

جدول ۲.۵ توزیع نمونه‌گیری x' برای ۵۶ نمونهٔ ممکن سه قطعه‌ای

قطعه‌ها در نمونه	x'	قطعه‌ها در نمونه	x'
(۱,۲,۳)	٤٦٨٦١/٣٣	(۲,٤,۷)	٣٣٠٧٧/٣٣
(۱,۲,٤)	٤٤٥٧٦	(۲,٤,۸)	٣٥٧٣٦
(۱,۲,٥)	٢٧٩١٤/٦٧	(۲,٥,٦)	١٦٤٦٦/٦٧
(۱,۲,٦)	٢٩١٠٩/٣٣	(۲,٥,۷)	١٦٤١٦
(۱,۲,۷)	٢٩٠٥٨/٦٧	(۲,٥,۸)	١٩٠٧٤/٦٧
(۱,۲,۸)	٣١٧١٧/٣٣	(۲,٦,۷)	١٧٦١٠/٦٧
(۱,۳,٤)	٦٠٩٣٨/٦٧	(۲,٦,۸)	٢٠٢٦٩/٣٣
(۱,۳,٥)	٤٤٢٧٧/٣٣	(۲,۷,۸)	٢٠٢١٨/٦٧
(۱,۳,٦)	٤٥٤٧٢	(۳,٤,٥)	٤٨٢٩٦
(۱,۳,۷)	٤٥٤٢١/٣٣	(۳,٤,٦)	٤٩٤٩٠/٦٧
(۱,۳,۸)	٤٨٠٨٠	(۳,٤,۷)	٤٩٤٤٠
(۱,٤,٥)	٤١٩٩٢	(۳,٤,۸)	٥٢٠٩٨/٦٧
(۱,٤,٦)	٤٣١٨٦/٦٧	(۳,٥,٦)	٣٢٨٢٩/٣٣
(۱,٤,۷)	٤٣١٣٦	(۳,٥,۷)	٣٢٧٧٨/٦٧
(۱,٤,۸)	٤٥٧٩٤/٦٧	(۳,٥,۸)	٣٥٤٣٧/٣٣
(۱,٥,٦)	٢٦٥٢٥/٣٣	(۳,٦,۷)	٣٣٩٧٣/٣٣
(۱,٥,۷)	٢٦٤٧٤/٦٧	(۳,٦,۸)	٣٦٦٣٢
(۱,٥,۸)	٢٩١٣٣/٣٣	(۳,۷,۸)	٣٦٥٨١/٣٣
(۱,٦,۷)	٢٧٦٦٩/٣٣	(٤,٥,٦)	٣٠٥٤٤
(۱,٦,۸)	٣٠٣٢٨	(٤,٥,۷)	٣٠٤٩٣/٣٣
(۱,۷,۸)	٣٠٢٧٧/٣٣	(٤,٥,۸)	٣٣١٥٢
(۲,۳,٤)	٥٠٨٨٠	(٤,٦,۷)	٣١٦٨٨
(۲,۳,٥)	٣٤٢١٨/٦٧	(٤,٦,۸)	٣٤٣٤٦/٦٧
(۲,۳,٦)	٣٥٤١٣/٣٣	(٤,۷,۸)	٣٤٢٩٦
(۲,۳,۷)	٣٥٣٦٢/٦٧	(٥,٦,۷)	١٥٠٢٦/٦٧
(۲,۳,۸)	٣٨٠٢١/٣٣	(٥,٦,۸)	١٧٦٨٥/٣٣
(۲,٤,٥)	٣١٩٣٣/٣٣	(٥,۷,۸)	١٧٦٣٤/٦٧
(۲,٤,٦)	٣٣١٢٨	(٦,۷,۸)	١٨٨٢٩/٣٣

جدول ۳.۵ دو طبقه برای داده‌های جدول ۱.۵

طبقهٔ ۱ (قطعات شهری)		طبقهٔ ۲ (قطعات روستایی)	
قطعه	۱۰۰۰× تعداد کامیون ـ مایل	قطعه	۱۰۰۰× تعداد کامیون ـ مایل
۱	٦۳۲۷	۲	۲۵۵۵
۳	۸٦۹۱	۵	۱۵۸٦
٤	۷۸۳٤	٦	۲۰۳٤
		۷	۲۰۱۵
		۸	۳۰۱۲

برآوردهای کل تعداد کامیون ـ مایل که از این طریق به دست می‌آید از ۱۵۰۲٦/٦۷ تا ٦۰۹۳۸/٦۷ تغییر می‌کنند و میانگین توزیع نمونه‌گیری x' برابر است با ۳٤۰۵٤ یعنی مجموع جامعه، و خطای معیار x' برابر است با ۱۰۵۳٦/۹.

حالا فرض کنید که به جای گرفتن یک نمونهٔ تصادفی ساده متشکل از سه قطعه از جامعه‌ای با هشت قطعه، اول قطعه‌ها را به دو طبقه تقسیم کنیم، یکی شامل قطعات شهری، دیگری شامل قطعات روستایی، به صورتی که در جدول ۳.۵ نشان داده شده است.

حالا می‌توانیم یک نمونه متشکل از یک قطعه از طبقهٔ ۱ و دو قطعه از طبقهٔ ۲ انتخاب کنیم و کل تعداد کامیون ـ مایل را به وسیلهٔ برآورد x'_{str} به صورت زیر برآورد کنیم:

$$x'_{str} = x'_{۱} + x'_{۲}$$

که در آن

$x'_{۱}$ = تعداد کامیون ـ مایل برآورد شده در سه قطعهٔ تشکیل دهندهٔ طبقهٔ ۱

$x'_{۲}$ = تعداد کامیون ـ مایل برآورد شده در پنج قطعهٔ تشکیل دهندهٔ طبقهٔ ۲

از نظر مفهومی، هر طبقه را یک زیرجامعه در نظر می‌گیریم، در داخل هر زیرجامعه به طور مستقل نمونه می‌گیریم، و برآوردی برای کل جامعه با انبوه سازی برآورد تک تک طبقه‌ها در همهٔ زیرجامعه‌ها یا طبقات به دست می‌آوریم. بنابراین، اگر یک نمونهٔ تصادفی ساده از یک قطعه از طبقهٔ ۱ و دو قطعه از طبقهٔ ۲ بگیریم در مجموع $\binom{۳}{۱} \times \binom{۵}{۲} = ۳۰$ نمونهٔ ممکن به دست می‌آوریم. پس از آن اگر از شیوهٔ برآورد که در بالا با معادلهٔ مربوط به x'_{str} ارائه شد استفاده کنیم توزیع نمونه‌گیری x'_{str} را به دست می‌آوریم که در جدول ٤.۵ نشان داده شده است.

$E(x'_{str})$، میانگین توزیع نمونه‌گیری x'_{str} برابر است با ۳٤۰۵٤ (تابلوی ۳.۲ را ببینید) که همان مجموع واقعی جامعه است. $SE(x'_{str})$، خطای معیار کل برآورد شدهٔ x'_{str} برابر است با ۳۲۹۷/٦ که

جدول ٤.٥ توزیع نمونه‌گیری x'_{str} برای ۳۰ نمونهٔ ممکن از سه قطعه

طبقهٔ ۱	طبقهٔ ۲	$x'_۱(=۳\bar{x}_۱)$	$x'_۲(=۵\bar{x}_۲)$	$x'_{str}=(x'_۱+x'_۲)$
۱	(۲,۵)	۱۸۹۸۱	۱۰۳۵۲/۵	۲۹۳۳۳/۵
۱	(۲,٦)	۱۸۹۸۱	۱۱٤۷۲/۵	۳۰٤۵۳/۵
۱	(۲,۷)	۱۸۹۸۱	۱۱٤۲۵	۳۰٤۰٦
۱	(۲,۸)	۱۸۹۸۱	۱۳۹۱۷/۵	۳۲۸۹۸/۵
۱	(۵,٦)	۱۸۹۸۱	۹۰۵۰	۲۸۰۳۱
۱	(۵,۷)	۱۸۹۸۱	۹۰۰۲/۵	۲۷۹۸۳/۵
۱	(۵,۸)	۱۸۹۸۱	۱۱٤۹۵	۳۰٤۷٦
۱	(٦,۷)	۱۸۹۸۱	۱۰۱۲۲/۵	۲۹۱۰۳/۵
۱	(٦,۸)	۱۸۹۸۱	۱۲٦۱۵	۳۱۵۹٦
۱	(۷,۸)	۱۸۹۸۱	۱۲۵٦۷/۵	۳۱۵٤۸/۵
۳	(۲,۵)	۲٦۰۷۳	۱۰۳۵۲/۵	۳٦٤۲۵/۵
۳	(۲,٦)	۲٦۰۷۳	۱۱٤۷۲/۵	۳۷۵٤۵/۵
۳	(۲,۷)	۲٦۰۷۳	۱۱٤۲۵	۳۷٤۹۸
۳	(۲,۸)	۲٦۰۷۳	۱۳۹۱۷/۵	۳۹۹۹۰/۵
۳	(۵,٦)	۲٦۰۷۳	۹۰۵۰	۳۵۱۲۳
۳	(۵,۷)	۲٦۰۷۳	۹۰۰۲/۵	۳۵۰۷۵/۵
۳	(۵,۸)	۲٦۰۷۳	۱۱٤۹۵	۳۷۵٦۸
۳	(٦,۷)	۲٦۰۷۳	۱۰۱۲۲/۵	۳٦۱۹۵/۵
۳	(٦,۸)	۲٦۰۷۳	۱۲٦۱۵	۳۸٦۸۸
۳	(۷,۸)	۲٦۰۷۳	۱۲۵٦۷/۵	۳۸٦٤۰/۵
٤	(۲,۵)	۲۳۵۰۲	۱۰۳۵۲/۵	۳۳۸۵٤/۵
٤	(۲,٦)	۲۳۵۰۲	۱۱٤۷۲/۵	۳٤۹۷٤/۵
٤	(۲,۷)	۲۳۵۰۲	۱۱٤۲۵	۳٤۹۲۷
٤	(۲,۸)	۲۳۵۰۲	۱۳۹۱۷/۵	۳۷٤۱۹/۵
٤	(۵,٦)	۲۳۵۰۲	۹۰۵۰	۳۲۵۵۲
٤	(۵,۷)	۲۳۵۰۲	۹۰۰۲/۵	۳۲۵۰٤/۵
٤	(۵,۸)	۲۳۵۰۲	۱۱٤۹۵	۳٤۹۹۷
٤	(٦,۷)	۲۳۵۰۲	۱۰۱۲۲/۵	۳۳٦۲٤/۵
٤	(٦,۸)	۲۳۵۰۲	۱۲٦۱۵	۳٦۱۱۷
٤	(۷,۸)	۲۳۵۰۲	۱۲۵٦۷/۵	۳٦۰٦۹/۵

خیلی کمتر از خطای معیار x'، یعنی برآورد معادل کل جامعهٔ تحت نمونه‌گیری تصادفی ساده است. این نتایج در جدول ۵.۵ خلاصه شده‌اند.

به صورت شهودی می‌توان دید که چرا برنامهٔ نمونه‌گیری مبتنی بر طبقه‌بندی برآوردی را به دست می‌دهد که خطای معیار آن کمتر از برآورد نظیر بر مبنای نمونه‌گیری تصادفی ساده است. برای x'، برآورد مجموع، با استفاده از نمونه‌گیری تصادفی ساده ۵۶ مقدار ممکن به دست می‌آید در حالی که از طبقه‌بندی ۳۰ مقدار برای x'_{str} به دست می‌آید. همچنین، امتحان کردن دامنهٔ تغییرات دو توزیع نشان می‌دهد که طبقه‌بندی، آن دسته از نمونه‌ها را که منجر به برآوردهای بسیار زیاد یا بسیار کم برای مجموع می‌شد حذف کرده است. سه قطعهٔ شهری دارای مقادیر زیاد و پنج قطعهٔ روستایی دارای مقادیر کم مشخصه‌های مورد اندازه‌گیری (کامیون – مایل) بودند. طبقه‌بندی اطمینان می‌دهد که حداقل یک قطعهٔ شهری و یک قطعهٔ روستایی در نمونه انتخاب می‌شوند و به این ترتیب امکان برآوردهای بسیار زیاد و بسیار کم مجموع را از میان می‌برد.

جدول ۵.۵ مقایسهٔ نتایج نمونه‌گیری تصادفی ساده و طبقه‌بندی

	طرح طبقه‌بندی	
	نمونه‌گیری تصادفی ساده	طبقه‌بندی
تعداد عناصر در نمونه	۳	*۳
تعداد نمونه‌های ممکن	۵۶	۳۰
میانگین توزیع برآورد مجموع	۳۴۰۵۴†	۳۴۰۵۴†
خطای معیار برآورد مجموع	۱۰۵۳۶/۹	۳۲۹۷/۶
دامنهٔ تغییرات توزیعهای برآورد مجموعها	۴۵۹۱۲	۱۲۰۰۷

* یک عنصر از طبقهٔ ۱ و دو عنصر از طبقهٔ ۲.

† این مقدار همان مجموع جامعه است.

□

طبقه‌بندی نسبت به نمونه‌گیری تصادفی ساده، سه امتیاز عمده دارد.

۱. با شرایط معین مفروض، دقت آن نسبت به نمونه‌گیری تصادفی ساده ممکن است افزایش پیدا کند (یعنی ممکن است خطاهای معیار به دست آمده از این شیوهٔ برآورد، کمتر باشند).
۲. امکان به دست آوردن برآوردهایی با دقت مشخص برای هر یک از طبقه‌ها وجود دارد.
۳. ممکن است جمع‌آوری اطلاعات برای نمونهٔ طبقه‌بندی شده نیز به دلایل سیاسی یا اداری به همان آسانی جمع‌آوری اطلاعات برای یک نمونهٔ تصادفی ساده باشد. در این صورت، با

گرفتن نمونهٔ طبقه‌بندی شده چیزی را از دست نمی‌دهیم زیرا، خطاهای معیار حاصل از این طرح به ندرت ممکن است از خطاهای معیار نمونه‌گیری تصادفی ساده بیشتر شود.

راهبرد مورد استفاده برای ساختن طبقات مستلزم دو گام است. ابتدا پارامتر مورد نظر جامعه برای برآورد کردن را تعیین می‌کنیم. سپس جامعه را نسبت به متغیر دیگری که تصور می‌شـــود با متغیر مورد نظر مربوط باشد طبقه‌بندی می‌کنیم. اگر فرض ما راجع به این پیوند درست باشد این گام دوم تضمین می‌کند که طبقه‌ها نسبت به متغیر تحت بررسی همگن‌اند.

برای مثال، اگر بخواهیم تعداد تختها را در تمام بیمارستانهای ایالت برآورد کنیم ممکن است بخواهیم بیمارستانها را از لحاظ زیربنا (که اطلاعی است که به خاطر مالیات به راحتی قابل دسترسی است) با این استدلال طبقه‌بندی کنیم که بیمارستانهایی که فضای بیشتری دارند تختهای بیشتری خواهند داشت. از سوی دیگر، اگر به متوسط هزینهٔ روزانهٔ هر تخت بیمارستانی به ازای هر بیمار علاقه‌مندیم می‌توانیم طبقه‌بندی بیمارستانها را بر مبنای محل جغرافیایی آنان با این استدلال انجام دهیم که بیمارستانهای مناطقی از ایالت که دارای اقتصاد پر رونقی‌اند می‌توانند هزینه‌های سنگینتری از بیماران خود مطالبه کنند تا بیمارستانهایی که در مناطق با اقتصاد کساد واقع شده‌اند.

در بیشتر موقعیتهای عملی، طبقه‌بندی جامعه نسبت به متغیر تحت بررسی صرفاً به دلایل مربوط به هزینه و عملی بودن، کاری مشکل است. در بسیاری از موارد جامعه به راحت‌ترین طریق با استفاده از معیارهای اداری (مثل نواحی رأی‌گیری)، معیارهای جغرافیایی (مثل شمال، جنوب، شرق یا غرب) یا سایر معیارهای طبیعی (مثل جنس یا سن) طبقه‌بندی می‌شود. طبقه‌بندی بر اساس راحتی غیرمنطقی نیست زیرا برآورد یک تک پارامتر در آمارگیری مدرن متداول نیست. در عوض، اطلاعات زیادی در مورد هر واحد نمونه‌گیری جمع‌آوری می‌شود و پارامترهای زیادی مورد توجه‌اند. واضح است که آن چه برای یک متغیر می‌تواند راهبرد طبقه‌بندی بهینه‌ای برای تهیه طبقات نسبتاً همگن باشد ممکن است نسبت به یک متغیر دیگر، طبقات بسیار ناهمگن فراهم نماید. اهمیت دارد که آماردان قبل از تصمیم‌گیری در مورد ملاک مناسب برای طبقه‌بندی، دامنهٔ داده‌هایی را که قرار است جمع‌آوری شوند مورد توجه قرار دهد. این قبیل مطالب گاهی اوقات در گزارشهای آمارگیریهای بهداشتی که از طبقه‌بندی استفاده کرده‌اند مورد بحث قرار می‌گیرند.

عیب عمدهٔ نمونه‌گیری طبقه‌بندی شده آن است که نیازمند شناسایی یکایک واحدهای شمارش بر حسب طبقه پیش از اجرای نمونه‌گیری است. اگر چنین اطلاعاتی به سهولت فراهم نباشد این روش به ندرت ممکن است عملی باشد.

ممکن است بخواهید طبقه‌بندی را با راهبرد نمونه‌گیری خوشه‌ای (فصلهای ۸ تا ۱۱) مقایسه کنید که در آن ممکن است صرفه‌جوییهایی عمده در وقت و هزینه امکان‌پذیر شود. مع‌هذا، به دلیل مزایایی که در بالا شرح داده شد، طبقه بندی فن بسیار توانایی است که در سطح وسیعی مورد استفاده قرار می‌گیرد.

۴.۵ پارامترهای جامعه برای طبقات

پارامترهای جامعه برای طبقات را می‌توان با همان نمادهایی تعریف کرد که برای تعریف پارامترهای جامعه به طور کلی مورد استفاده قرار دادیم. در این بخش نمادهایی را معرفی می‌کنیم که در سراسر بحث از برنامه‌های نمونه‌گیری بر مبنای طبقه‌بندی به کار خواهند رفت.

جامعه‌ای را در نظر می‌گیریم که دارای N واحد اولیه است که در L طبقهٔ دو به دو ناسازگار و جامع به گونه‌ای گروه‌بندی شده‌اند که طبقهٔ ۱ دارای N_1 واحد اولیه، طبقهٔ ۲ دارای N_2 واحد اولیه، ... ، و طبقهٔ L دارای N_L واحد اولیه است. به عبارت دیگر، $N=\sum_{h=1}^{L} N_h$ *اندازهٔ جامعه* است.

فرض کنید یک متغیر یا مشخصهٔ x را در جامعه در نظر گرفته‌ایم. از نماد $X_{h,i}$ برای نشان دادن *مقدار مشخصهٔ* x برای iامین واحد اولیه در داخل طبقهٔ h استفاده می‌کنیم. به عبارت دیگر، واحدهای اولیه داخل هر طبقهٔ ویژهٔ h از ۱ تا N_h شماره‌گذاری می‌شوند.

مثال تشریحی: فرض کنید می‌خواهیم متوسط هزینهٔ روزانهٔ دارو به ازای هر بیمار را در یک بیمارستان برآورد کنیم. تصمیم می‌گیریم بیمارستان را بر حسب خدمات (درمانی، جراحی، زنان و زایمان، و همهٔ خدمات دیگر با هم) طبقه‌بندی کنیم و واحدهای اولیه را به صورت بیماران در یک روز معین تعریف می‌کنیم.

فرض کنید که در یک روز تعیین شده ۲۵۰ بیمار در بیمارستان است که ۱۰۰ نفر برای درمان، ۷۵ نفر برای جراحی، ۵۰ نفر زنان و زایمان و ۲۵ نفر برای سایر خدمات مراجعه کرده‌اند. پس با استفاده از نمادهایی که در بالا معرفی شدند داریم:

$$N=۲۵۰ \qquad N_۱=۱۰۰ \qquad N_۲=۷۵ \qquad N_۳=۵۰ \qquad N_۴=۲۵$$

اگر $X_{h,i}$ نشان‌دهندهٔ مقدار متغیر x برای iامین واحد اولیه در داخل طبقهٔ h باشد، آنگاه مثلاً برای طبقهٔ ۲ داریم:

$X_{۲,۱}$ = مقدار متغیر x برای عنصر ۱ در داخل طبقهٔ ۲

$X_{۲,۲}$ = مقدار متغیر x برای عنصر ۲ در داخل طبقهٔ ۲

و به همین ترتیب تا می‌رسیم به

$X_{۲,۷۵}$ = مقدار متغیر x برای عنصر ۷۵ در داخل طبقهٔ ۲

عناصر داخل سایر طبقه‌ها نیزبه همین ترتیب تعریف می‌شوند.

□

پارامترهای جامعه برای طبقات را به روشی تعریف می‌کنیم که پارامترهای جامعهٔ کل را تعریف کردیم.

مجموع یا مقدار تجمعی متغیر x *در داخل یک طبقهٔ* h را با نماد X_{h+} نشان می‌دهیم که از فرمول زیر به دست می‌آید:

$$X_{h+} = \sum_{i=۱}^{N_h} X_{h,i}$$

مجموع کل جامعه با جمع کردن مجموع طبقه‌ها به دست می‌آید، یا

$$X = \sum_{h=۱}^{L} \sum_{i=۱}^{N_h} X_{h,i} = \sum_{h=۱}^{L} X_{h+}$$

سطح *میانگین* یک مشخصهٔ x برای طبقهٔ h با نماد $\overline{X}_h$ نشان داده می‌شود و از فرمول زیر به دست می‌آید:

$$\overline{X}_h = \frac{\sum_{i=۱}^{N_h} X_{h,i}}{N_h} = \frac{X_{h+}}{N_h}$$

میانگین $\overline{X}$ متغیر x برای تمام جامعه از فرمول زیر

$$\overline{X} = \frac{X}{N}$$

یا معادل جبری آن،

$$\overline{X} = \frac{\sum_{h=۱}^{L} N_h \overline{X}_h}{N} = \sum_{h=۱}^{L} W_h \overline{X}_h$$

به دست می‌آید، که در آن،

$$W_h = \frac{N_h}{N}$$

به عبارت دیگر، میانگین $\overline{X}$ برای تمام جامعه عبارت است از *متوسط موزون* میانگینهای تکی طبقه‌ها، $\overline{X}_h$، با وزنهای ($W_h = \frac{N_h}{N}$) متناسب با تعداد عناصر در هر طبقه.

واریانس $\sigma^۲_{hx}$ توزیع متغیر x در داخل یک طبقهٔ خاص h به صورت متوسط توان دوم انحرافها حول میانگین طبقه تعریف و با فرمول زیر نشان داده می‌شود:

$$\sigma^۲_{hx} = \frac{\sum_{i=۱}^{N_h} (X_{h,i} - \overline{X}_h)^۲}{N_h}$$

ضرایب تغییرات و واریانسهای نسبی برای هر طبقه به همان صورتی تعریف می‌شوند که برای جامعه‌ای که به طبقات گروه‌بندی نشده است تعریف شد. *ضریب تغییرات* برای توزیع در داخل یک طبقهٔ خاص از فرمول زیر به دست می‌آید:

$$V_{hx} = \frac{\sigma_{hx}}{\overline{X}_h}$$

و واریانس نسبی صرفاً عبارت است از توان دوم ضریب تغییرات.

بـرای راحتی کار در ارجاعهای بعدی، پارامترهای جامعه برای نمونه‌گیری طبقه‌بندی شده در تابلوی ۱.۵ خلاصه شده‌اند. حالا ببینیم این فرمولها در عمل چگونه به کار می‌روند.

مثال تشریحی: جامعه‌ای متشکل از ١٤ خانواده را در نظر بگیرید که در سه بلوک شهری زندگی می‌کنند. اگر خانواده‌ها را واحدهای اولیه، بلوکها را به عنوان طبقه‌ها، و اندازهٔ خانواده‌ها را مشخصهٔ x در نظر بگیریم، وضعیتی خواهیم داشت که در جدول ٦.۵ نشان داده شده است.

تابلوی ۱.۵ پارامترهای طبقه‌ها و جامعه برای نمونه‌گیری طبقه‌بندی شده

	تمام جامعه	در داخل طبقه
مجموع		
(۱.۵)	$X = \sum_{h=1}^{L}\sum_{i=1}^{N_h} X_{h,i} = \sum_{h=1}^{L} X_{h+}$	$X_{h+} = \sum_{i=1}^{N_h} X_{h,i}$
میانگین		
(۲.۵)	$\overline{X} = \frac{\sum_{h=1}^{L} N_h \overline{X}_h}{N} = \sum_{h=1}^{L} W_h \overline{X}_h = \frac{X}{N}$	$\overline{X}_h = \frac{\sum_{i=1}^{N_h} X_{h,i}}{N_h} = \frac{X_{h+}}{N_h}$
نسبت		
(۳.۵)	$P_y = \frac{\sum_{h=1}^{L} N_h P_{hy}}{N} = \sum_{h=1}^{L} W_h P_{hy}$	$P_{hy} = \frac{\sum_{i=1}^{N_h} Y_{h,i}}{N_h}$
واریانس		
(٤.۵)		$\sigma_{hx}^2 = \frac{\sum_{i=1}^{N_h} (X_{h,i} - \overline{X}_h)^2}{N_h}$
واریانس نسبی		
(۵.۵)		$V_{hx}^2 = \frac{\sigma_{hx}^2}{\overline{X}_h^2}$

در این تعریفها L تعداد طبقه‌ها، N_h تعداد عناصر در طبقهٔ h، N تعداد کل عناصر، $X_{h,i}$ مقدار x برای iامین عنصر در طبقهٔ h، $y_{h,i}$ نشانهٔ بود یا نبود یک صفت کیفی دوحالتی y برای iامین عنصر در طبقهٔ h، و $W_h = \frac{N_h}{N}$ نسبتی از کل جامعه که متعلق به طبقهٔ h است.

جدول ۶.۵ طبقه‌ها برای جامعهٔ متشکل از ۱۴ خانواده

بلوک	خانواده	اندازهٔ خانواده
۱	۱	۴
	۲	۳
	۳	۴
۲	۱	۴
	۲	۶
	۳	۴
	۴	۸
	۵	۸
۳	۱	۲
	۲	۳
	۳	۲
	۴	۲
	۵	۲
	۶	۳

بر حسب نمادهایی که در بالا معرفی شدند، مقادیر x برای هر طبقه به شرح زیرند:

طبقهٔ ۱ ($N_۱ = ۳$)	طبقهٔ ۲ ($N_۲ = ۵$)	طبقهٔ ۳ ($N_۳ = ۶$)
$X_{۱,۱} = ۴$	$X_{۲,۱} = ۴$	$X_{۳,۱} = ۲$
$X_{۱,۲} = ۳$	$X_{۲,۲} = ۶$	$X_{۳,۲} = ۳$
$X_{۱,۳} = ۴$	$X_{۲,۳} = ۴$	$X_{۳,۳} = ۲$
	$X_{۲,۴} = ۷$	$X_{۳,۴} = ۲$
	$X_{۲,۵} = ۸$	$X_{۳,۵} = ۲$
		$X_{۳,۶} = ۳$

مجموع متغیر x در داخل هر طبقهٔ h با استفاده از معادلهٔ (۱.۵) به شرح زیر است:

$$X_{۱+} = ۴ + ۳ + ۴ = ۱۱$$
$$X_{۲+} = ۴ + ۶ + ۴ + ۷ + ۸ = ۲۹$$
$$X_{۳+} = ۲ + ۳ + ۲ + ۲ + ۲ + ۳ = ۱۴$$

مجموع برای تمام جامعه دوباره با استفاده از معادلهٔ (۱.۵) عبارت است از:

$$X = ۱۱ + ۲۹ + ۱۴ = ۵۴$$

میانگینهای جامعه‌ای برای طبقات با استفاده از معادلهٔ (۲.۵) به صورت زیرند:

$$\overline{X}_{۱} = \frac{۱۱}{۳} = ۳/۶۷ \qquad \overline{X}_{۲} = \frac{۲۹}{۵} = ۵/۸ \qquad \overline{X}_{۳} = \frac{۱۴}{۶} = ۲/۳۳$$

میانگین جامعهٔ کل از معادلهٔ (۲.۵) به صورت زیر به دست می‌آید:

$$\overline{X} = \frac{۳}{۱۴} \times ۳/۶۷ + \frac{۵}{۱۴} \times ۵/۸ + \frac{۶}{۱۴} \times ۲/۳۳ = ۳/۸۵۷$$

واریانسهای طبقات با استفاده از معادلهٔ (٤.۵) عبارت‌اند از:

$$\sigma^{۲}_{۱x} = \frac{(۴ - ۳/۶۷)^{۲} + (۳ - ۳/۶۷)^{۲} + (۴ - ۳/۶۷)^{۲}}{۳} = ۰/۲۲۲$$

$$\sigma^{۲}_{۲x} = \frac{(۴ - ۵/۸)^{۲} + (۶ - ۵/۸)^{۲} + \cdots + (۸ - ۵/۸)^{۲}}{۵} = ۲/۵۶$$

$$\sigma^{۲}_{۳x} = \frac{(۲ - ۲/۳۳)^{۲} + (۳ - ۲/۳۳)^{۲} + \cdots + (۳ - ۲/۳۳)^{۲}}{۶} = ۰/۲۲۲$$

واریانسهای نسبی برای طبقات از معادلهٔ (۵.۵) به شرح زیر به دست می‌آیند:

$$V^{۲}_{۱x} = \frac{۰/۲۲۲}{(۳/۶۷)^{۲}} = ۰/۰۱۶۵$$
$$V^{۲}_{۲x} = \frac{۲/۵۶}{(۵/۸)^{۲}} = ۰/۰۷۶۱$$
$$V^{۲}_{۳x} = \frac{۰/۲۲۲}{(۲/۳۳)^{۲}} = ۰/۰۴۰۹$$

□

۵.۵ آماره‌های نمونه برای طبقات

فرض کنید که در داخل یک طبقهٔ خاص h به نحوی یک نمونهٔ n_h عنصری از N_h عنصر موجود در طبقه انتخاب و هر عنصر نمونه را نسبت به نوعی متغیر x اندازه‌گیری کرده‌ایم. برای راحتی کار، عناصر

نمونه را از ۱ تا n_h شماره‌گذاری می‌کنیم و قرار می‌گذاریم که $x_{h,۱}$ مقدار متغیر x برای عنصر شـــمارهٔ «۱» نمونه و $x_{h,۲}$ مقدار متغیر x برای عنصر شمارهٔ «۲» نمونه و $x_{h.n_h}$ مقدار x برای عنصر شمارهٔ « n_h » نمونه باشد. اگر نمونه‌ای متشکل از n_h عنصر در داخل هر طبقهٔ h گرفته شود، در آن صورت کل اندازهٔ نمونهٔ n از راه زیر به دست می‌آید:

$$n = \sum_{h=۱}^{L} n_h$$

که در آن L، تعداد طبقاتی است که عناصر جامعه در آنها گروه‌بندی شده‌اند.

مثلاً، در مثال قبل اگر یک عنصر از طبقهٔ ۱، دو عنصر از طبقهٔ ۲ و چهار عنصر از طبقهٔ ۳، انتخاب کنیم خواهیم داشت

$$n_۱ = ۱ \quad , \quad n_۲ = ۲ \quad , \quad n_۳ = ۴$$

و

$$n = ۱ + ۲ + ۴ = ۷$$

اگر مقادیر چهار عنصر انتخاب شده از طبقهٔ ۳ عبارت باشند از $X_{۳,۱}$، $X_{۳,۳}$، $X_{۳,۴}$ و $X_{۳,۶}$ در آن صورت خواهیم داشت:

$$x_{۳,۱} = X_{۳,۱} = ۲ \qquad x_{۳,۳} = X_{۳,۴} = ۲$$
$$x_{۳,۲} = X_{۳,۳} = ۲ \qquad x_{۳,۴} = X_{۳,۶} = ۳$$

مجموع نمونه‌ای و میانگین *نمونه‌ای برای یک طبقهٔ خاص* از فرمول زیر به دست می‌آیند:

$$x_{h+} = \sum_{i=۱}^{n_h} x_{h,i} \qquad \text{و} \qquad \bar{x}_h = \frac{\sum_{i=۱}^{n_h} x_{h,i}}{n_h} = \frac{x_{h+}}{n_h}$$

۶.۵ برآورد پارامترهای جامعه از روی نمونه‌گیری تصادفی طبقه‌بندی شده

اگر انتخاب عناصر نمونه به طور مستقل در داخل هر طبقه با استفاده از نمونه‌گیری تصادفی ساده انجام پذیرفته باشد، در آن صورت این طرح را *طرح نمونه‌گیری تصادفی طبقه‌بندی شده* می‌نامند و برآوردهای مناسب پارامترهای جامعه همراه با خطاهای معیار آنها در تابلوی ۲.۵ نشان داده شده‌اند.

در ایـن تعـاریف، L تعـداد طبقات، N_h تعداد عناصر در طبقهٔ h، N کل تعداد عناصر، $x_{h,i}$ مقدار x بـرای iامیـن عنصر در طبقهٔ h و $y_{h,i}$ معرف بود یا نبود یک صفت کیفی دو حالتی y برای iامین عنصر در طبقهٔ h است.

همچنین $S_{hx}^{۲}$، برآورد واریانس توزیع x برای طبقهٔ h است که از فـرمول زیر به دست می‌آید:

$$S_{hx}^{۲} = \frac{\sum_{i=۱}^{n_h}(x_{h,i} - \bar{x}_h)^{۲}}{n_h - ۱} \qquad (۹.۵)$$

تابلوی ۲.۵ برآوردهای پارامترهای جامعه و خطاهای معیار این برآوردها در نمونه‌گیری طبقه‌بندی شده

پارامتر	خطای معیار برآورد برای تمام جامعه	برآورد برای تمام جامعه	برآورد برای طبقه
مجموع			
(۶.۵)	$\hat{SE}(x'_{\text{str}}) = \sqrt{\sum_{h=۱}^{L} \frac{N_h^{۲} S_{hx}^{۲}}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$	$x'_{\text{str}} = \sum_{h=۱}^{L} x'_h$	$x'_h = N_h \bar{x}_h$
میانگین			
(۷.۵)	$\hat{SE}(\bar{x}_{\text{str}}) = \sqrt{\sum_{h=۱}^{L}\left(\frac{N_h}{N}\right)^{۲} \frac{S_{hx}^{۲}}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$	$\bar{x}_{\text{str}} = \frac{\sum_{h=۱}^{L} N_h \bar{x}_h}{N}$	$\bar{x}_h = \frac{\sum_{i=۱}^{n_h} x_{h,i}}{n_h}$
نسبت			
(۸.۵)	$\hat{SE}(p_{y,\text{str}}) = \sqrt{\sum_{h=۱}^{L}\left(\frac{N_h}{N}\right)^{۲} \frac{p_{hy}(۱-p_{hy})}{n_h - ۱}\left(\frac{N_h - n_h}{N_h}\right)}$	$p_{y,\text{str}} = \frac{\sum_{h=۱}^{L} y_{hi}}{N}$	$p_{hy} = \frac{\sum_{i=۱}^{n_h} y_{h,i}}{n_h}$

مثال تشریحی: با استفاده از STATA یکصد و پنجاه و هشت (۱۵۸) بیمارستان در ناحیه‌ای بر حسب سطح خدمات زایمان در سه طبقه گروه‌بندی شده‌اند که با متغیر *oblevel* نشان داده می‌شود. از این ۱۵۸ بیمارستان ناحیه، ٤۲ بیمارستان در طبقهٔ ۱ (تخصصی)، ۹۹ بیمارستان در طبقهٔ ۲ (متوسط) و ۱۷ بیمارستان در طبقـــهٔ ۳ (درجهٔ سه) قرار دارند. یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۱۵ بیمارستان از این جامعه گرفته شده است: ٤ بیمارستان در طبقهٔ ۱؛ ۵ بیمارستان در طبقهٔ ۲؛ و ٦ بیمارستان در طبقهٔ ۳. این نمونهٔ خاص متشکل از ۱۵ بیمارستان همراه با متغیرهای مربوط برای برآورد، در پروندهٔ اطلاعاتی STATA به صورت *hospsamp.dta* ثبت شده‌اند. این پرونده در صفحهٔ بعد نشان داده شده است.

Hospno	oblevel	weighta	tothosp	births
15	1	10.50	42	480
80	1	10.50	42	426
86	1	10.50	42	342
136	1	10.50	42	174
7	2	19.80	99	2022
26	2	19.80	99	576
62	2	19.80	99	1999
90	2	19.80	99	482
101	2	19.80	99	836
28	3	2.83	17	3108
34	3	2.83	17	4674
39	3	2.83	17	2539
102	3	2.83	17	1610
119	3	2.83	17	4618
149	3	2.83	17	1781

معانی متغیرها به شرح زیرند:

hospno = کد شناسایی برای بیمارستان خاص.

oblevel = سطح خدمات زایمان برای هر بیمارستان.

weighta = وزن نمونه‌گیری $\frac{N_h}{n_h}$ برای بیمارستان خاص.

tothosp = تعداد کل جامعهٔ بیمارستانها در طبقه‌ای که بیمارستان نمونه قرار دارد.

births = تعداد رویدادهای تولد در بیمارستان در طی سال قبل.

از فرمانهای زیر می‌توان برای به دست آوردن برآوردهای کل تعداد تولدها و خطاهای معیار برآوردها استفاده کرد.

```
. use "a:\ hospsamp.dta", clear
. svyset pweight weighta
. svyset strata oblevel
. svyset fpc tothosp
. svytotal births
. svytotal births, by(oblevel)
```

فرمان اول نشان می‌دهد که داده‌ها در فایل اطلاعاتی STATA به نام *hospsamp.dta* قرار دارند.

فرمان دوم نشان می‌دهد که وزن نمونه‌گیری برای بیمارستان در متغیر *weighta* قرار دارد.

فرمان سوم نشان می‌دهد که متغیر نشانگر طبقه در متغیر *oblevel* قرار دارد.

فرمان چهارم نشان می‌دهد که N_h، تعداد بیمارستانهای طبقه در متغیر *tothosp* قرار دارد. در اینجا نیز این فرمان برای گزارشهای متعلق به یک طبقه یکسان است ولی برای گزارشهای مربوط به بیمارستانها در طبقه‌های مختلف یکسان نیست. تصحیح جامعهٔ متناهی از روی این متغیر محاسبه می‌شود.

فرمان پنجم نشان می‌دهد که کل تعداد تولدها باید از روی داده‌های نمونه برآورد شود.

فرمان ششم نشان می‌دهد که کل تعداد تولدها برای هر طبقه باید از این داده‌ها برآورد شود.

خروجی STATA که از فرمانهای بالا تولید شده به شرح زیر است:

```
Survey total estimation

pweight:  weighta                    Number of obs     =     15
Strata:   oblevel                    Number of strata  =      3
PSU:      <observations>             Number of PSUs    =     15
FPC:      tothosp                    Population size   =    158

Total      | Estimate     Std. Err.     [95% Conf.  Interval]        Deff
-----------+-------------------------------------------------------------
births     | 183983       34014.35      109872.1    258093.9         .7035476
```

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

```
Survey total estimation

pweight:  weighta                    Number of obs     =     15
Strata:   oblevel                    Number of strata  =      3
PSU:      <observations>             Number of PSUs    =     15
FPC:      tothosp                    Population size   =    158

Total  Subpop  | Estimate     Std. Err.     [95% Conf.  Interval]        Deff
---------------+-------------------------------------------------------------
births         |
oblevel = = 1  | 14931        2669.857      9113.882    20748.12         .1564799
oblevel = = 2  | 117117       33067.68      45068.7     189165.3         1.089406
oblevel = = 3  | 51935        7508.403      35575.6     68294.4          .0330073
```

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

مثال تشریحی با استفاده از SUDAAN: دوباره مثالی را که در آن ۱۵۸ بیمارستان در یک ناحیه بر حسب سطح خدمات زایمان در سه طبقه گروه‌بندی شده‌اند در نظر می‌گیریم، که در آن یک نمونهٔ تصادفی طبقه‌بندی شده گرفتیم و ٤ بیمارستان از طبقهٔ ۱، ۵ بیمارستان از طبقهٔ ۲ و ٦ بیمارستان از طبقهٔ ۳ انتخاب کردیم. این نمونهٔ خاص متشکل از ۱۵ بیمارستان در بحث استفاده از STATA برای تحلیل این داده‌ها در بالا نشان داده شد. حالا فرض می‌کنیم که پروندهٔ اطلاعاتی *hospsamp.dta* یک

پـروندهٔ اطـلاعـاتی SAS PC اسـت (*hospsamp.ssd*). در زیر نشـان خواهیـم داد که چگونه نرم‌افـزار آمـاری SUDAAN مـی‌توانـد بـرآوردهـا و خطاهـای معیـار را بـرای ایـن طـرح نمونه‌گیری تصادفی طبقه‌بندی شده تولید کند.

دوبـاره بـا فـرض ایـن کـه *hospsamp* بـه پـروندهٔ اطلاعـاتی SAS ارسال شده است، می‌توانیم از فـرمانهای زیـر در بـه دسـت آوردن بـرآوردهای کل تولدها برای هر طبقه و برای ۱۵۸ بیمارستانی که جامعه را تشکیل می‌دهند، استفاده کنیم.

```
PROC DESCRIPT DATA = HOSPSAMP FILETYPE = SAS  DESIGN = WOR TOTALS;
NEST OBLEVEL;
WEIGHT WEIGHTA;
TOTCNT TOTHOSP;
VAR BIRTHS;
SUBGROUP OBLEVEL;
LEVELS 3;
SETENV COLWIDTH=20;
SETENV DESWIDTH=3;
```

سـطر اول اطلاعـاتی دربـارهٔ پـروندهٔ اطلاعـاتی ارائه می‌دهد و حاکی از آن است که نمونه‌گیری بدون جایگذاری است و قرار است برآورد مجموعها را به دست آوریم.

سطر دوم نشان می‌دهد که گزارشهای نمونه در داخل متغیر طبقه *oblevel* جای داده شده‌اند.

سـطر سـوم متغیری را که قرار است به عنوان وزن نمونه‌گیری به کار رود مشخص می‌سازد که در این مورد *weighta* نامیده می‌شود و وارون کسر نمونه‌گیری خاص طبقه است.

سطر چهارم نشان می‌دهد که متغیر *tothosp* شامل کل تعداد بیمارستانهای جامعـهٔ آمـاری اسـت، (۴۲ بیمارستان برای طبقهٔ ۱، ۹۹ بیمارستان برای طبقهٔ ۲ و ۱۷ بیمارستان برای طبقهٔ ۳).

سطر پنجم نشان می‌دهد که کل تعداد تولدها قرار است برآورد شود.

سطرهای ششم و هفتم نشان می‌دهند که کل تعداد تولدها برای هر طبقه قرار است برآورد شود.

سطرهای ۸ و ۹ فورمت خروجی را تعیین می‌کنند.

خروجی زیر از فرمانهای فوق‌الذکر SUDAAN تولید شده است.

بـرآوردهای حاصل از SUDAAN و STATA با یکدیگر مطابقت دارند و خواننده می‌تواند تحقیق کند که با برآوردهای حاصل از فرمولهای تابلوی ۲.۵ نیز مطابقت دارند.

□

۷.۵ خلاصه

در ایـن فصـل مـروری بـر طبقه‌بندی داشتیم. مفهوم طبقه‌بندی را معرفی و تعریف کردیم که منظور از نمونـهٔ تصـادفی طـبقه‌بندی شده چیست. مواردی را به بحث گذاشتیم که ممکن بود طبقه‌بندی در آنها

مناسب باشد و مزایای طبقه‌بندی را نسبت به نمونه‌گیری تصادفی ساده و نیز معایب عمدهٔ آن را توضیح دادیم. عبارتهایی را که برای مشخص کردن پارامترهای جامعه به کار می‌روند و آماره‌های تحت نمونه‌گیری تصادفی ساده را ارائه کردیم. سرانجام نشان دادیم که چگونه می‌توان برآوردهایی از طریق نمونه‌گیری تصادفی طبقه‌بندی شده، همراه با خطاهای معیار آنها، با استفاده از دو نرم‌افزار SUDAAN و STATA به دست آورد.

```
1  PROC DESCRIPT DATA = HOSPSAMP FILETYPE = SAS  DESIGN = WOR
   TOTALS;
2  NEST OBLEVEL;
3  WEIGHT WEIGHTA;
4  TOTCNT TOTHOSP;
5  VAR BIRTHS;
6  SUBGROUP OBLEVEL;
7  LEVELS 3;
8  SETENV COLWIDTH=20;
9  SETENV DECWIDTH=3;
```

Number of observations read : 15 Weighted count: 158
Number of observations skipped : 0
(WEIGHT variable nonpositive)
denominator degrees of freedom : 12

by: Variable, OBLEVEL.

Variable		OBLEVEL Total	1
BIRTHS	Sample Size	15.000	4.000
	Weighted Size	158.000	42.000
	Total	183982.904	14931.000
	SE Total	34014.329	2669.857
	Mean	1164.449	355.500
	SE Mean	215.281	63.568

by: Variable. OBLEVEL.

Variable		OBLEVEL 2	3
BIRTHS	Sample Size	5.000	6.000
	Weighted Size	99.000	17.000
	Total	117116.928	51934.977
	SE Total	33067.664	7508.399
	Mean	1183.000	3055.000
	SE Mean	334.017	441.671

تمرین

۱.۵ در یک نمونهٔ تصادفی ساده در مورد۱۰۰۰ دونده از میان ۱۰۰۰۰ دونده‌ای که در مسابقهٔ ماراتون ۱۹۹۵ شیکاگو شرکت کرده بودند، نتیجهٔ آزمایش خون ۳۵ نفر برای مصرف استروییید و سایر داروهایی که فعالیت را زیاد می‌کنند مثبت بود. وقتی نتایج بر حسب زمان تکمیل مسابقه دسته‌بندی شدند جدول زیر به دست آمد.

زمان تکمیل مسابقه (ساعت)	تعداد در نمونه	تعداد نتیجهٔ مثبت برای مصرف دارو	درصد مثبت
کمتر از ۲/۵	۱۰۰	۲۵	۲۵/۰۰
۲/۵– ٤	۵۰۰	۷	۱/٤۰
بیشتر از ٤	٤۰۰	۳	۰/۷۵
جمع	۱۰۰۰	۳۵	۳/۵۰

الف. خطای معیار نسبتی که برای مصرف داروی فعالیت‌زا مثبت برآورد شده چقدر است؟

ب. با بررسی (بدون محاسبه) نرخهای ارائه شده در بالا فکر می‌کنید طبقه‌بندی می‌توانست منجر به برآورد اساساً بهتری شود؟ چرا آری یا چرا نه؟

۲.۵ فرض کنید که در تمرین ۱.۵ از ۱۰۰۰۰ نفری که مسابقهٔ ماراتون را به پایان رساندند ۲۰۰۰ نفر در زمانی کمتر از ۲/۵ ساعت، ٦۰۰۰ نفر بین ۲/۵ تا ٤ ساعت و ۲۰۰۰ نفر در زمانی بیشتر از ٤ ساعت مسابقه را تکمیل کرده باشند. در مورد نتایج نمونه‌گیری به صورتی که در تمرین ۱.۵ ارائه شد اظهار نظر کنید.

۳.۵ اگر یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۳۳۳ نفر در هر یک از گروههای سه‌گانه گرفته می‌شد، برآورد درصد مثبت که از این طریق به دست می‌آمد احتمالاً چقدر می‌شد؟

٤.۵ تعداد ۲۰۰۰ دوندهٔ دیگر نیز وارد بازی شدند ولی آن را به پایان نرساندند. از میان این عده به تصادف برای ٦۰۰ نفر پرسشنامه‌هایی از طریق پست ارسال شد که ۵۰۰ نفر آن را پر کردند. یکی از اقلام این پرسشنامه از پاسخگو می‌خواست که متوسط تعداد مایلهایی را که در هفته طی مدت ۸ هفته قبل از مسابقهٔ ماراتون پیموده است برآورد کند. میانگین، $\bar{x}$، تعداد مایلهای طی شده ۳۲/٤ با انحراف معیار S_x معادل ۷/۳ بود. یک نمونهٔ تصادفی ساده متشکل از ۵۰۰ نفر از ۱۰۰۰۰ دونده‌ای که مسابقهٔ ماراتون را تکمیل کرده بودند ٤۰۰ پاسخگو داشت که

متوسط تعداد مایلهای هفتگی آنها ٤٦/٨ با انحراف معیاری معادل ٦/٢ مایل بود. متوسط مایلهای طی شده در هفته طی مدت مورد نظر برای همهٔ کسانی که وارد مسابقهٔ ماراتون شدند چقدر است؟

۵.۵ داده‌های زیر مربوط به ٦ سازمان حفظ بهداشت برای سال ۱۹۸۸ در یک شهر با اندازهٔ متوسط است:

تعداد کارکنان تأمین کنندهٔ مراقبت از بیماران و تعداد رویارویی با بیمار طی سال ۱۹۸۸ توسط ٦ سازمان

HMO: سازمان حفظ بهداشت	تعداد پزشکان تأمین کنندهٔ مراقبت ازبیماران	تعداد رویارویی با بیمار
۱	۱۰	۲۲۰۰۰
۲	٦	۱٤۰۰۰
۳	٤	۱۰۲۰۰
٤	۳۰	۷۰۰۰۰
۵	۷	۱۵۰۰۰
٦	۳	۵۰۰۰

چون این طرح مستلزم استفادهٔ نوبتی از هزینه و زمان است پیشنهاد می‌شود که یک نمونهٔ دوتایی از این سازمانهای حفظ بهداشت برای برآورد کل تعداد رویارویی با بیماران در این شش سازمان طی سال ۱۹۸۸ گرفته شود.

الف. همهٔ نمونه‌های تصادفی سادهٔ دوتایی را شمارش و کل تعداد رویارویی با بیمار در سال ۱۹۸۸ را برای ٦ سازمان برآورد کنید.

ب. خطای معیار برآورد مجموع از روی طرح نمونهٔ تعیین شده در بالا چقدر است؟

پ. سازمانهای شمارهٔ ۱، ۲، ۳، ۵ و ٦ را در یک طبقه و سازمان شمارهٔ ٤ را (به تنهایی) در یک طبقهٔ دیگر گروه‌بندی کنید. از روی این دو طبقه، تمام نمونه‌های تصادفی طبقه‌بندی شدهٔ دوتایی را شمارش کنید. خطای معیار برآورد مجموع حاصل از این برنامهٔ نمونه‌گیری چقدر است؟

ت. در مورد مطلوبیت استفاده از طبقه‌بندی در این وضعیت در مقابل نمونه‌گیری تصادفی ساده اظهار نظر کنید.

۶.۵ فرض کنید که در وضعیت ارائه شده در تمرین ۵.۵، برآورد تعداد رویارویی در سال ۱۹۸۸ به ازای هر پزشک مورد نظر است. قسمتهای (الف) تا (ت) تمرین ۵.۵ را برای این وضعیت برآورد تکرار کنید. آیا در این وضعیت نمونه‌گیری تصادفی طبقه‌بندی شده بهتر از نمونه‌گیری تصادفی ساده است؟ چرا بهتر است و چرا نیست؟

۷.۵ در درمانگاه بزرگی واقع در یک بیمارستان مرکز شهر، ۵۶ بیماری که دارای عفونت بدون علامتهای ظاهری ایدز هستند با یک داروی آزمایشی تحت درمان قرار گرفته‌اند که تصور می‌شود توانایی بازدهی برخی از کارکردهای سیستم دفاعی بدن را که با عفونت HIV همراه است دارند. از این بیماران شمارش گلبولی CD4 برای ۱۲ نفر در ویزیت اول کمتر از ۲۵۰ بود، برای ۲۰ نفر بین ۲۵۰ تا ۴۰۰ بود و برای ۲۴ نفر بیشتر از ۴۰۰ بود. (هر چه این شمارش کمتر باشد وضعیت بیمار وخیمتر است). می‌خواهیم یک نمونهٔ تصادفی طبقه‌بندی شده متشکل از ۳۰ بیمار یعنی ۱۰ نفر از هر یک از گروههای سه‌گانهٔ بالا بگیریم با این هدف که وقوع پیشامدهای ناشی از ایدز را در یک دورهٔ ۱۲ ماهه در میان این بیماران برآورد کنیم.

الف. چند نمونهٔ متشکل از ۳۰ بیمار به شرحی که در بالا گذشت امکان پذیر است؟

ب. با فرض این که بیمار می‌تواند بیش از یک پیشامد حاکی از ایدز طی این مدت داشته باشد به صورت جبری نشان دهید که چگونه وقوع پیشامدهای حاکی از ایدز را از روی نمونه برای دورهٔ ۱۲ ماهه برآورد می‌کنید.

پ. چند نمونهٔ ۳۰ تایی تحت نمونه‌گیری تصادفی ساده بدون طبقه‌بندی امکان پذیر است؟

۸.۵ در زیر نتایج آمارگیری نمونه‌ای که در تمرین ۷.۵ توصیف شد ارائه شده است:

	تعداد افراد		
تعداد پیشامدها	$CD4 < ۲۵۰$	$۲۵۰ \le CD4 < ۴۰۰$	$CD4 \ge ۴۰۰$
۰	۵	۷	۹
۱	۱	۲	۱
۲	۲	۱	۰
۳	۲	۰	۰

از روی این داده‌ها وقوع پیشامدهای حاکی از ایدز را در جامعهٔ هدف برآورد کنید. خطای معیار برآورد نرخ وقوع چقدر است؟ (باید فرمول خطای معیار را از «اصول اولیه» به دست آورید).

۹.۵ از روی داده‌های تمرین ۸.۵ نسبت بیماران دارای یک یا چند پیشامد حاکی از ایدز را برآورد کنید. خطای معیار این نسبت برآورد شده چقدر است؟

۱۰.۵ یک نمونهٔ تصادفی ساده (بدون طبقه‌بندی) متشکل از ۳۰ نفر از این ۵۶ بیمار، داده‌های زیر را به دست داده است:

	تعداد افراد		
تعداد پیشامدها	CD 4 < ۲۵۰	۲۵۰ ≤ CD 4 < ٤۰۰	CD 4 ≥ ٤۰۰
۰	۳	۸	۱۲
۱	۱	۲	۰
۲	۰	۱	۱
۳	۲	۰	۰

وقوع پیشامدهای حاکی از ایدز را از روی این داده‌ها برآورد کنید و خطای معیار این برآورد را به دست آورید. این نتایج را با نتایج حاصل از طرح نمونه‌گیری طبقه‌بندی شده مقایسه کنید. کدام طرح، برآورد قابل اعتمادتری را تولید می‌کند؟ چرا؟

کتابشناسی

The following articles present sample surveys in which stratification is used.

1. Hemphill, F. M., A sample survey of home injuries, *Public Health Reports,* 67: 1026, 1952.
2. Horvitz, D. G., Sampling and field procedures of the Pittsburgh morbidity survey, *Public Health Reports,* 67: 1003, 1952.
3. Goldberg, J., Levy, P. S., Mullner, R., Gelfand, H., Iverson, N., Lemeshow, S., and Rothrock, J., Factors affecting trauma center utilization in Illinois, *Medical Care* 19: 547, 1981.
4. Stasny, E. A., Toomey, B. G., and First, R. J., Estimating the rate of rural homelessness: A study of nonurban Ohio, *Survey Methodology,* 20: 87, 1994.
5. Barner, B. M., and Levy, P. S., State-wide shoulder belt usage by type of roadway and posted speed limit: A three year comparison, *38th Annual Proceedings Association for the Advancement of Automotive Medicine,* Association for the Advancement of Automotive Medicine, Des Plaines Ill., 1994.

The following software programs have the capability of taking stratified samples.

6. Frankel, M. R., and Spencer, B. D., *SAMPLE. A Supplementary Module for SYSTAT,* SYSTAT, Inc., Evanston, Ill., 1990.
7. SAS Institute Inc., *SAS Language and Procedures,* Usage 2, Version 6, 1st ed., Cary, N.C., SAS Institute Inc., 1991, pp. 649.

In addition, there is material on stratification in virtually every text on sampling theory and survey methodology, including those listed in the bibliography sections of earlier chapters.