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 ۱۱

نمونه‌گیری خوشه‌ای که در آن خوشه‌ها با احتمال نابرابر نمونه‌گیری می‌شوند: نمونه‌گیری با احتمال متناسب با اندازه

روش‌شناسی نمونه‌گیری خوشه‌ای که در هر دو فصل ۹ و ۱۰ شرح و بسط داده شد منحصراً به آن دسته از طرحهای نمونه‌گیری محدود می‌شد که در آن، خوشه‌ها با احتمال برابر نمونه‌گیری می‌شدند. به عبارت دیگر، در بحث ما دربارهٔ نمونه‌گیری خوشه‌ای تا اینجا، همهٔ خوشه‌های موجود در جامعه، مستقل از تعداد واحدهای شمارش آنها یا در واقع مستقل از هر مشخصهٔ دیگری که ممکن است داشته باشند، احتمال یکسان برای انتخاب شدن در نمونه داشته‌اند. در این فصل نشان می‌دهیم که این‌گونه طرحها در مواردی خاص می‌توانند انتخابهای نمونه‌ایی را نتیجه دهند که قابل اجرا نیستند و یا می‌توانند به برآوردهایی خطی بینجامند (مانند میانگینها، مجموعها، نسبتها) که خطاهای معیار آنها بسیار زیاد باشد. وضعیتهایی خاص که مخصوصاً در برابر این قبیل مشکلات از همه آسیب‌پذیرترند آنهایی هستند که بین واحدهای نمونه‌گیری اولیهٔ آنها از لحاظ تعداد واحدهای شمارش تغییرپذیری قابل ملاحظه‌ای وجود دارد. متأسفانه این وضعیتها از جمله مواردی هستند که در عمل بیش از همه روی می‌دهند.

برای اجتناب از این مشکلات، در این فصل، روشی را برای نمونه‌گیری خوشه‌ای شرح و بسط می‌دهیم که در آن همهٔ خوشه‌های جامعه دارای احتمال یکسان برای انتخاب شدن در نمونه نیستند. تمرکز ما به خصوص بر رده‌ای از طرحهای نمونه‌گیری خوشه‌ای است که *نمونه‌گیری با احتمال متناسب با اندازه* نامیده می‌شوند (و عموماً به صورت *نمونه‌گیری PPS* خلاصه می‌شوند). در این رده از طرحها، احتمال انتخاب شدن یک خوشه در نمونه متناسب با معیاری از اندازهٔ آن است که معمولاً تعداد واحدهای شمارش موجود در آن است. به راهی دیگر می‌توان از متغیر دیگری که به سطح متغیر مورد برآورد وابسته است استفاده کرد.

در شرح و بسط مفاهیم مربوط به نمونه‌گیری خوشه‌ای با احتمال نابرابر، در برخی موارد برای روشن شدن مطلب از مثالهای تشریحی مربوط به نمونه‌گیری خوشه‌ای یک‌مرحله‌ای استفاده خواهیم کرد، هر چند که می‌دانیم استفاده از آن بسیار کمتر از نمونه‌گیری خوشه‌ای دو و یا حتی چندمرحله‌ای است.

۱.۱۱ انگیزه برای نمونه‌گیری نکردن با احتمال برابر از خوشه‌ها

جامعهٔ متشکل از سه بیمارستان را که در جدول ۱.۱۱ نشان داده شده است در نظر می‌گیریم. فرض کنید قرار است یک نمونهٔ خوشه‌ای یک‌مرحله‌ای ساده متشکل از دو تا از این بیمارستانها بگیریم تا کل تعداد جراحیهای غیرضروری بیماران سرپایی را در بیمارستان برآورد کنیم. آمارگیری مستلزم استخراج اطلاعات از هر سابقهٔ نمونه‌ای است و اندازه‌های نمونه برای هر یک از سه نمونهٔ ممکن در زیر نشان داده شده‌اند.

بیمارستانها در نمونه	تعداد سوابق بیماران سرپایی نمونه‌گیری شده			
	بیمارستان ۱	بیمارستان ۲	بیمارستان ۳	کل اندازهٔ نمونه
۱ و ۲	۳۰۰۰	٤۰۰۰	—	۷۰۰۰
۱ و ۳	۳۰۰۰	—	۱۰۰۰۰	۱۳۰۰۰
۲ و ۳	—	٤۰۰۰	۱۰۰۰۰	۱٤۰۰۰

جدول بالا به وضوح نشان می‌دهد که این طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ساده به کل اندازهٔ نمونه‌ای بسیار غیرمنتظره منجر می‌شود (این اندازه می‌تواند حداقل ۷۰۰۰ و حداکثر ۱٤۰۰۰ باشد) و بار استخراج اطلاعات در میان بیمارستانهای نمونه‌گیری شده بسیار نابرابر خواهد بود (در صورتی که بیمارستان ۱ برای نمونه انتخاب شود فقط ۳۰۰۰ سابقه باید از آن انتخاب شود در حالی که از بیمارستان ۳ باید ۱۰۰۰۰ سابقه انتخاب شود). نابرابری در میان بیمارستانها از لحاظ اندازهٔ

نمونـه و غیـر قـابل پیش‌بینی بودن از نظر اندازهٔ کل نمونه می‌تواند قابل اجرا بودن این آمارگیری را به شدت به مخاطره اندازد.

عـلاوه بـر مشـکلات مربوط به قابل اجرا بودن که در بالا بیان شد، در زیر (با استفاده از فرمولهای مـربوط بـه خطـای معیار در فصل ۹) مشاهده می‌کنیم که یک برآورد خطی از قبیل برآورد کل جامعه، تغییرپذیـری قـابل توجهـی خواهـد داشت. توزیع y'_{clu} برآورد مجموع جراحیهای غیرضروری بیماران سرپایی در هر سه نمونهٔ ممکن در پایین نشان داده شده است.

جدول ۱.۱۱ تعداد جراحیهای انجام شده برای بیماران سرپایی در سال ۱۹۹۷ در سه بیمارستان

بیمارستان	جراحیهای بیماران سرپایی	تعداد جراحیهای غیرضروری بیماران سرپایی
۱	۳۰۰۰	۳۵
۲	٤۰۰۰	۳۸
۳	۱۰۰۰۰	۱۰۰
مجموع	۱۷۰۰۰	۱۷۳

بیمارستانها در نمونه	مجموع نمونه y	برآورد مجموع جامعه y'_{clu}
۲ , ۱	۷۳	۱۰۹/۵
۳ , ۱	۱۳۵	۲۰۲/۵
۳ , ۲	۱۳۸	۲۰۷

مقـدار مـورد انـتظار و خطـای معیار y'_{clu} برآورد مجموع همراه با ضریب تغییرات آن در زیر ارائه شده است:

$$E(y'_{clu}) = ۱۷۳$$
$$SE(y'_{clu}) = ۵۵/۰۴$$
$$V(y'_{clu}) = ۰/۳۲$$

توجـه کنید که این برآورد، وقتی در نظر بگیریم که دو سوم خوشه‌های موجود در جامعه در نمونه قرار گرفـته‌انـد ضـریب تغییـرات زیـادی دارد. عـلت آن اسـت کـه متغیر موردنظر، یعنی تعداد جراحیهای غیرضـروری بیمـاران سـرپایی، عمدتـاً بـه تعـداد جـراحیهای بیمـاران سـرپایی وابسته است و در بین بیمارسـتـانها از لحـاظ کل تعداد جراحیـهای بیماران سرپایی تغییرپذیری قابل توجـهی وجود دارد.

در واقع، بیمارستان ۳ تعداد بیشتری جراحیهای سرپایی داشته و از این رو جراحیهای غیرضروری بیماران سرپایی نیز در این بیمارستان بیش از مجموع بیمارستانهای ۱ و ۲ است. به این ترتیب، نمونه‌ای که بیمارستان شمارهٔ ۳ در آن نباشد تعداد جراحیهای غیرضروری را به صورتی قابل ملاحظه کم برآورد خواهد کرد در حالی که دو نمونه‌ای که شامل بیمارستان شمارهٔ ۳ هستند این مجموع کل را بیش ـ برآورد خواهند کرد. در نمونه‌گیری خوشه‌ای ساده، به طوری که در فصل ۹ و ۱۰ توضیح داده شد، هر بیمارستان صرفنظر از هر اطلاعاتی معلوم دربارهٔ خوشه، شانس یکسانی برای ورود به آمارگیری دارد. همچنین، شیوه‌های برآورد مورد استفاده در نمونه‌گیری خوشه‌ای ساده هیچ استفاده‌ای از اطلاعات مربوط به مشخصه‌های خوشه به عمل نمی‌آورد.

ما نشان خواهیم داد که امکان کاهش خطای نمونه‌گیری یک برآورد مجموع، حاصل از یک نمونهٔ خوشه‌ای، با استفاده از اطلاعات معلوم دربارهٔ خوشه‌ای که ممکن است با متغیر پاسخ همبستگی داشته باشد وجود دارد. این اطلاعات را می‌توان هم در شیوهٔ نمونه‌گیری و هم در فرایند برآورد کردن مورد استفاده قرار داد.

فرض کنیم به جای نمونه‌گیری از خوشه‌ها با احتمال برابر طوری نمونه بگیریم که احتمال قرار گرفتن هر بیمارستان در نمونه، متناسب با تعداد جراحیهای بیماران سرپایی باشد. به طور دقیقتر، شیوه‌ای را برای نمونه‌گیری توصیف می‌کنیم که در آن نمونه‌گیری بدون جایگذاری است و بیمارستانها یک به یک از یک «آوندی» فرضی قرعه‌کشی می‌شوند. فرض کنید که احتمال انتخاب شدن یک بیمارستان در هر نوبت قرعه‌کشی متناسب با تعداد جراحیهای بیماران سرپایی است و هر بار که در قرعه‌کشی انتخاب شد از آوند خارج می‌شود. اگر X_i را برابر با تعداد جراحیهای بیماران سرپایی در بیمارستان i ام فرض کنیم می‌بینیم که P_i، احتمال انتخاب شدن بیمارستان i در اولین قرعه‌کشی از فرمول زیر به دست می‌آید.

$$P_i = \frac{X_i}{X}$$

که در آن

X_i = تعداد جراحیهای بیماران سرپایی در بیمارستان i

X = کل تعداد جراحیهای بیماران سرپایی در هر سه بیمارستان

از مطالب بالا و از نظریهٔ ترکیبیاتی نتیجه می‌گیریم که اگر دو بیمارستان بدون جایگذاری نمونه‌گیری شوند، در آن صورت π_{ij}، احتمال اینکه بیمارستانهای i و j در نمونه انتخاب شوند از فرمول زیر به دست می‌آید.

$$\pi_{ij} = \frac{X_۱ X_۲}{X}\left(\frac{۱}{X - X_۱} + \frac{۱}{X - X_۲}\right) \qquad (۱.۱۱)$$

پس برای داده‌های جدول ۱.۱۱ احتمالهای زیر را برای انتخاب شدن داریم:

بیمارستانها در نمونه $(j، i)$	احتمال ظاهر شدن بیمارستانهای i و j با هم در نمونه [از روی رابطهٔ (۱.۱۱)] π_{ij}
(۲ , ۱)	۰/۱۰٤۷۲
(۳ , ۱)	۰/۳۷۸۱۵
(۳ , ۲)	۰/۵۱۷۱۳

مثلاً از رابطهٔ (۱.۱۱)

$$\pi_{۱۲} = \frac{۳۰۰۰ \times ۴۰۰۰}{۱۷۰۰۰} \times \left[\frac{۱}{۱۷۰۰۰ - ۳۰۰۰} + \frac{۱}{۱۷۰۰۰ - ۴۰۰۰}\right] = ۰/۱۰۴۷$$

به محض این که π_{ij} به دست آمد می‌توانیم احتمال π_i را که بیمارستان i در نمونه قرار بگیرد به دست آوریم:

$$\pi_i = \sum_j \pi_{ij}$$

برای هر یک از سه بیمارستان موجود در جامعه، π_i، احتمال اینکه در نمونه ظاهر شود در زیر نشان داده می‌شود:

بیمارستان	π_i، احتمال قرار گرفتن در نمونه
۱	۰/٤۸۲۸۷
۲	۰/٦۲۱۸۵
۳	۰/۸۹۵۲۸

حالا، براساس برنامهٔ نمونه‌گیری بالا، برآوردگر y'_{hte} را با استفاده از فرمول زیر بیان می‌کنیم

$$y'_{hte} = \sum_{i=۱}^{n} \frac{y_i}{\pi_i}$$

نماد «hte » از حروف آغازین Horvitz-Thompson estimator گرفته شده که رده‌ای از برآوردگرهاست که در بخش بعدی به صورتی کلیتر مورد بحث قرار خواهد گرفت.

این برآوردگر برخلاف برآوردگر x'_{clu}، از اطلاعات مربوط به خوشه، هم در نمونه‌گیری و هم در ساختن برآوردگر استفاده می‌کند. شانس انتخاب شدن هر بیمارستان در نمونه برای آن بیمارستانهایی که تعداد بیشتری جراحی سرپایی داشته‌اند بیشتر است زیرا فرض بر این است که به متغیر مورد نظر ـ یعنی تعداد جراحیهای غیرضروری بیماران سرپایی ـ وابسته است. در این شیوهٔ برآورد، به هر بیمارستان برحسب عکس احتمال انتخاب شدن خود در نمونه وزن داده می‌شود تا برآوردگر حاصل نااریب باشد.

توزیع y'_{hte} برای همهٔ نمونه‌های ممکن در زیر نشان داده شده است:

بیمارستانها در نمونه	y'_{hte}	احتمال انتخاب شدن نمونه، π_{ij}
(۲ , ۱)	۱۳۳/۵۹	۰/۱۰٤۷۲
(۳ , ۱)	۱۸٤/۱۸	۰/۳۷۸۱۵
(۳ , ۲)	۱۷۲/۸۰	۰/۵۱۷۱۳

مقدار مورد انتظار y'_{hte} همراه با خطای معیار و ضریب تغییرات آن در زیر نشان داده شده است:

$$E(y'_{hte}) = \sum_{\text{همهٔ نمونه‌ها}} y'_{hte}\pi_{ij} = ۱۳۳/۵۹ \times ۰/۱۰۴۷۲ + ۱۸۴/۱۸ \times ۰/۳۷۸۱۵ + ۱۷۲/۸۰ \times ۰/۵۱۷۱ = ۱۷۳ = Y$$

$$SE(y'_{hte}) = \left(\sum_{\text{همهٔ نمونه‌ها}} (y'_{hte} - ۱۷۳)^{۲} \pi_{ij} \right)^{۱/۲} = ۱۴/۴۹$$

$$V(y'_{hte}) = \frac{۱۴/۴۹}{۱۷۳} = ۰/۰۸۴$$

باید توجه داشت که شیوهٔ نمونه‌گیری مبتنی بر انتخاب خوشه‌ها با احتمال نابرابر در ترکیب با استفاده از برآوردگر y'_{hte} در این مثال، ضریب تغییرات ۸/٤٪ را به دست می‌دهد که به مراتب کمتر از ضریب تغییرات ۳۲٪ است که از نمونه‌گیری خوشه‌ها با احتمال برابر به دست می‌آمد. هر دو شیوه، برآوردگرهایی نااریب از کل جامعه تولید می‌کنند.

در مثال بالا نشان دادیم که به وسیلهٔ نمونه‌گیری از خوشه‌ها با احتمال نابرابر، این امکان وجود دارد که برآوردگری به دست آید که خطای معیار آن به صورتی قابل ملاحظه کمتر از خطای معیار به دست آمده از نمونه‌گیری خوشه‌ها با احتمال برابر باشد. در بخشهای بعدی این فصل نشان می‌دهیم که چگونه می‌توان از این قبیل روشها در نمونه‌گیری خوشه‌ای دومرحله‌ای استفاده کرد تا برآوردهایی به

دست آید که معتبر و قابل اعتماد و مبتنی بر نمونه‌هایی باشند که تعداد واحدهای شمارش آنها در داخل خوشه‌ها تقریباً یکسان‌اند. همان‌طور که قبلاً بیان شد، به دلایلی که مربوط به بودجه و امکان‌پذیر بودن کار است، غالباً بسیار مهم است که کل اندازهٔ نمونه، قابل پیش‌بینی باشد و نمونه به شکلی هموار میان خوشه‌ها توزیع شده باشد. در بخش بعد برخی روشهای برآورد را شرح و بسط می‌دهیم که در نمونه‌گیری خوشه‌ها با احتمال نابرابر به کار می‌روند.

۲.۱۱ دو ردهٔ کلی از برآوردگرهای معتبر برای طرحهای نمونه‌ای که در آن واحدها با احتمال نابرابر انتخاب می‌شوند

نوشتگان مربوط به نمونه‌گیری با احتمال نابرابر بدواً پیرامون دو ردهٔ کلی از برآوردگرها تکامل یافته‌اند: یعنی *برآوردگر هورویتز ـ تامپسون* که در بخش قبل به آن اشاره شد و برآوردگری قبل از آن که به *برآوردگر هنسن ـ هورویتز* موسوم است. از این دو مورد با جزئیاتی بیشتر در این بخش بحث می‌کنیم.

۱.۲.۱۱ برآوردگر هورویتز ـ تامپسون

برآوردگر هورویتز ـ تامپسون در اصل به عنوان برآوردگری از یک مجموع در سال ۱۹۵۲ پیشنهاد شد که می‌تواند برای هر طرح نمونه‌گیری، با جایگذاری یا بدون جایگذاری، ساخته شود [۸]. شکل آن برای طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای به صورت زیر است

$$y'_{hte} = \sum_{i=1}^{v} \frac{Y_i}{\pi_i}$$

که در آن

Y_i = مجموع برای i امین خوشهٔ نمونه

π_i = احتمال انتخاب شدن i امین خوشه در نمونه

v = تعداد خوشه‌های متمایز نمونه‌گیری شده

واضح است که هرگاه نمونه‌گیری بدون جایگذاری باشد $v = m$، که m کل تعداد خوشه‌های نمونه‌گیری شده است. می‌توان نشان داد که در هر طرح نمونه‌گیری خوشه ای یک‌مرحله‌ای، y'_{hte} برآوردگری نااریب از Y با خطای معیاری است که از فرمول زیر به دست می‌آید:

$$SE(y'_{hte}) = \sqrt{\sum_{i=1}^{M} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{M} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) Y_i Y_j} \qquad (۳.۱۱)$$

کـه در آن M ، تعداد خوشه‌ها در جامعه، و π_{ij} احتمال این است که خوشه‌های i و j هر دو در نمونه قرار بگیرند.

در مورد مثال تشریحی نشان داده شده در بخش قبل، داریم

$$\sum_{i=۱}^{M} \frac{۱-\pi_i}{\pi_i} Y_i^۲ = ۳۳۵۹/۷۱$$

$$\sum_{i=۱}^{M} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) Y_i Y_j = -۳۱۴۹/۷۹$$

و

$$SE(y'_{hte}) = \sqrt{۳۳۵۹/۷۱ + (-۳۱۴۹/۷۹)} = ۱۴/۴۹$$

توجه کنید که این با نتایجی که قبلاً با استفاده از فرمولهای تعاریف به دست آمد توافق دارد.

برآوردگـر زیـر، یعـنی $\hat{Var}(y'_{hte})$ ، یـک برآوردگـر نـااریب از واریـانس $Var(y'_{hte})$ بـرای طـرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای است:

$$\hat{Var}(y'_{hte}) = \sum_{i=۱}^{v} \frac{۱-\pi_i}{\pi_i} Y_i^۲ + \sum_{i=۱}^{v} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{Y_i Y_j}{\pi_{ij}} \qquad (٤.۱۱)$$

ایـن برآوردگـر مشکلاتی دارد به این علت که غالباً بی‌ثبات است و می‌تواند منفی باشد. برآوردگرهای دیگری که کمتر مشکل دارند نیز ابداع شده و در جاهای دیگری مورد بحث قرار گرفته‌اند [۲] تا [٤].

اهمیـت برآوردگـر هورویـتز ـ تامپسون در آن است که برآوردگری نااریب است که تا وقتی بتوان π_i، احـتمال در نمونـه قـرار گرفـتن هـر واحـد را ارزیابی کرد می‌تواند در مورد هر طرح نمونه‌گیری محاسـبه شود. این کار حتی برای پیچیده‌ترین طرحهای نمونه‌گیری نیز غالباً امکان‌پذیر است. تعمیم آن بـه هـرگونه طـرح نمونه‌گیری، آن را در شرح و بسط و یکپارچه کردن نظریهٔ نمونه‌گیری، حائز اهمیت فـوق‌العـاده‌ای ساخته است. عیب آن در بسیاری از وضعیتهای عملی آن است که واریانس آن نه تنها به احـتمالهای شـمول π_i، بـلکه بـه احـتمال تـوأم π_{ij}، یعنی احتمال قرار گرفتن هر جفت از واحدها در نمونه، بستگی دارد. در عمل، به خصوص هرگاه نمونه‌گیری بدون جایگذاری باشد، ارزیابی مقادیر π_{ij} غالباً مشکل و پرزحمت است و برآورد کردن خطاهای نمونه‌گیری را مشکل‌ساز می‌کند. در بخش بعد، برآوردگر دیگری را مورد بحث قرار می‌دهیم که استفاده از آن در عمل غالباً آسانتر است.

۲.۲.۱۱ برآوردگر هنسن ـ هورویتز

این برآوردگر که در زیر برای نمونه‌گیری خوشه‌ای یک‌مرحله‌ای نشان داده می‌شود در سال ۱۹۴۳ توسط هنسن و هورویتز [۷] به عنوان برآوردگری برای مجموع، Y، پیشنهاد شد که هرگاه نمونه‌گیری *با جایگذاری* باشد نااریب است:

$$y'_{hh} = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{\pi'_i} \qquad (۵.۱۱)$$

که در آن π'_i، احتمال انتخاب واحد i ام در هر قرعه‌کشی نمونه است. خطای معیار این برآوردگر از فرمول زیر به دست می‌آید

$$SE(y'_{hh}) = \sqrt{\frac{1}{m}\sum_{i=1}^{M}\left(\frac{Y_i}{\pi'_i} - Y\right)^2 \pi'_i} \qquad (۶.۱۱)$$

و برآوردگری از خطای معیار از فرمول زیر به دست می‌آید

$$\hat{SE}(y'_{hh}) = \sqrt{\frac{\sum_{i=1}^{m}\left(\frac{Y_i}{\pi'_i} - y'_{hh}\right)^2}{m(m-1)}} \qquad (۷.۱۱)$$

باید توجه داشت که عبارت زیر رادیکال در معادلهٔ (۷.۱۱) برآوردگری نااریب از واریانس برآوردگر برای هرگونه طرح نمونه‌گیری خوشه‌ای یک‌مرحله‌ای است.

مثال تشریحی: استفاده از این برآوردگر را با مثال زیر که در جدول ۲.۱۱ آمده است نشان می‌دهیم. ناحیه‌ای دارای چهار خانهٔ سالمندان است و قرار است نمونه‌ای از دو تا از این خانه‌ها با جایگذاری گرفته شود تا Y کل تعداد زنان ۹۰ سال به بالا که در سال ۱۹۹۷ در خانه‌های سالمندان پذیرش

جدول ۲.۱۱ تعداد زنان ۹۰ سال به بالا که در ناحیه‌ای، در سال ۱۹۹۷ در خانه‌های سالمندان پذیرش شده‌اند

خانهٔ سالمندان	تعداد تخت (X_i)	تعداد، Y_i، زنان ۹۰ سال به بالا پذیرش شده در سال ۱۹۹۷	π'_i، احتمال انتخاب خانهٔ سالمندان در نمونه در هر استخراج
۱	۳۵	۳	۰/۱۲۷۲۷
۲	۷۵	۷	۰/۲۷۲۷۳
۳	۱۴۰	۲۴	۰/۵۰۹۰۹
۴	۲۵	۱	۰/۰۹۰۹۱
مجموع	۲۷۵	۳۵	۱/۰۰۰۰۰

شده‌اند برآورد شود. برای نشان دادن برآوردگر هنسن ـ هورویتز، فرض می‌کنیم نمونه‌گیری باجایگذاری است و در هر استخراج خانهٔ سالمندان، احتمال این که خانه‌ای خاص انتخاب شود برابر است با تعداد تختهای آن، X_i، تقسیم بر کل تعداد تختها، X، در هر چهار خانهٔ سالمندان موجود در ناحیه. برای هر خانهٔ سالمندان که انتخاب شود، تعداد اشخاص ۹۰ سال به بالا که در سال ۱۹۹۷ پذیرش شده‌اند از روی سوابق پذیرش تعیین خواهد شد.

در این وضعیت، M ، تعداد خانه‌های سالمندان موجود در ناحیه برابر با ٤ است؛ m ، تعداد واحدهای نمونه‌گیری شده برابر است با ۲ و تعداد نمونه‌های ممکن با جایگذاری برابر است با ۱٦=$٤^{۲}$. توزیع y'_{hh}، مقادیر برآورد شدهٔ هنسن ـ هورویتز برای مجموع در جدول ۳.۱۱ نشان داده شده است.

از فرمولهای مربوط به تعاریف می‌بینیم که برآوردگر هنسن ـ هورویتز برآوردگری نااریب از Y است:

$$E[y'_{hh}] = \sum_r y'_{hh}(r)P(r) = ۲۳/۵۷۱۴۳ \times ۰/۰۱۶۱۹۸ + \cdots + ۲۹/۰۷۱۴۳ \times ۰/۰۰۸۲۶۴ = ۳۵ = Y$$

جدول ۳.۱۱ توزیع برآوردگر هنسن ـ هورویتز، y'_{hh}، در تمام نمونه‌های ممکن متشکل از دو خانهٔ سالمندان که با جایگذاری استخراج شده‌اند

نمونهٔ r	خانه‌های سالمندان در نمونه $(i\,,j)$	$P(r)$ ، احتمال انتخاب شدن نمونهٔ r $(\pi'_i \times \pi'_j)$	$y'_{hh}(r)$
۱	(۱ , ۱)	۰/۰۱٦۲۰	۲۳/۵۷۱٤۳
۲	(۲ , ۱)	۰/۰۳٤۷۱	۲٤/٦۱۹۰۵
۳	(۳ , ۱)	۰/۰٦٤۷۹	۳۵/۳۵۷۱٤
٤	(٤ , ۱)	۰/۰۱۱۵۷	۱۷/۲۸۵۷۱
۵	(۱ , ۲)	۰/۰۳٤۷۱	۲٤/٦۱۹۰۵
٦	(۲ , ۲)	۰/۰۷٤۳۸	۲۵/٦٦٦٦۷
۷	(۳ , ۲)	۰/۱۳۸۸٤	۳٦/٤۰٤۷٦
۸	(٤ , ۲)	۰/۰۲٤۷۹	۱۸/۳۳۳۳۳
۹	(۱ , ۳)	۰/۰٦٤۷۹	۳۵/۳۵۷۱٤
۱۰	(۲ , ۳)	۰/۱۳۸۸٤	۳٦/٤۰٤۷٦
۱۱	(۳ , ۳)	۰/۲۵۹۱۷	٤۷/۱٤۲۸٦
۱۲	(٤ , ۳)	۰/۰٤٦۲۸	۲۹/۰۷۱٤۳
۱۳	(۱ , ٤)	۰/۰۱۱۵۷	۱۷/۲۸۵۷۱
۱٤	(۲ , ٤)	۰/۰۲٤۷۹	۱۸/۳۳۳۳۳
۱۵	(۳ , ٤)	۰/۰٤٦۲۸	۲۹/۰۷۱٤۳
۱٦	(٤ , ٤)	۰/۰۰۸۲٦	۱۱/۰۰۰۰۰

و نیز از روی داده‌های جدول ۳.۱۱ می‌توان دید که خطای معیار آن (همان‌طور که از فرمولهای مربوط به تعریف به دست آمد) به صورت عددی برابر است با آنچه که از روی عبارت (٦.۱۱) محاسبه می‌شد. مشخصاً، داریم:

$$SE(y'_{hh}) = ۹/۱۶$$

در این مثال می‌توان از y'_{hte}، برآوردگر هنسن ـ تامپسون نیز استفاده کرد، زیرا برآوردگری نااریب از مجموع جامعه در نمونه‌گیری با جایگذاری و همچنین در نمونه‌گیری بدون جایگذاری است. ساختن این برآوردگر و توزیع نمونه‌گیری آن در جدولهای ٤.۱۱ و ۵.۱۱ نشان داده شده است.

باز هم با به کار بردن فرمولهای مربوط به تعاریف برای داده‌های جدول ۵.۱۱، می‌توان دید که $E(y'_{hte}) = ۳۵ = X$ و $SE(y'_{hte}) = ۱۰/۸۲$، که در این مورد خاص، اندکی بیشتر از خطای معیار برآوردگر هنسن ـ هورویتز است. ولی، خطاهای معیار هر دوی این برآوردگرها به مراتب کمتر از خطای معیار برآوردگر تورمی معمولی است ($x'_{clu} = (M/m)x$). خطای معیار آن برآوردگر ۲۵/٦٤۱۸ است.

□

۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه

در بخشهای قبلی، مشاهده کردیم که استفاده از نمونه‌گیری با احتمال نابرابر در دو مثال، منجر به برآوردهایی از مجموع شد که خطاهای معیار آنها به مراتب کمتر از خطای معیاری بود که از برآورد بر مبنای نمونه‌گیری از خوشه‌ها با احتمال برابر به دست می‌آمد. در هر دوی این مثالها، نمونه‌گیری از خوشه‌ها بر مبنای مشخصه‌ای معلوم از خوشه‌ها انجام می‌پذیرفت که فرض می‌شد با متغیر موردنظر

جدول ٤.۱۱ تعداد زنان ۹۰ سال به بالا که در سال ۱۹۹۷ در خانه‌های سالمندان یک ناحیه پذیرش شده‌اند

خانهٔ سالمندان	تعداد تختها (X_i)	Y_i، تعداد زنان ۹۰ سال به بالا، پذیرش شده در سال ۱۹۹۷	π_i، احتمال انتخاب خانهٔ سالمندان در نمونه
۱	۳۵	۳	۰/۲۳۸۳٤۷
۲	۷۵	۷	۰/٤۷۱۰۷٤
۳	۱٤۰	۲٤	۰/۷۵۹۰۰۸
٤	۲۵	۱	۰/۱۷۳۵۵٤
مجموع	۲۷۵	۳۵	

وابسته است (مثلاً فرض می‌شد که تعداد تختها در خانهٔ سالمندان با تعـــداد زنان ۹۰ سال به بالا که در دورهٔ زمانی مشخصی پذیرش شده بودند مرتبط باشد). به طور کلی، این راهبردی است که به صورتی گسترده در طراحی آمارگیریهای نمونه‌ای مورد استفاده قرار می‌گیرد و ما در این بخش با جزئیاتی بیشتر از آن بحث خواهیم کرد.

جدول ۵.۱۱ توزیع y'_{hte}، برآوردگر هورویتز ـ تامپسون در همهٔ نمونه‌های ممکن از دو خانهٔ سالمندان که با جایگذاری انتخاب شده‌اند

نمونهٔ r	خانه‌های سالمندان در نمونه $(i\,,j)$	$P(r)$، احتمال انتخاب شدن نمونهٔ r $(\pi'_i \times \pi'_j)$	$y'_{hte}(r)$
۱	(۱ , ۱)	۰/۰۱۶۲۰	۱۲/۵۸۶۶۹
۲	(۲ , ۱)	۰/۰۳۴۷۱	۲۷/۴۴۶۳۳
۳	(۳ , ۱)	۰/۰۶۴۷۹	۴۴/۲۰۶۸۹
۴	(۴ , ۱)	۰/۰۱۱۵۷	۱۸/۳۴۸۵۹
۵	(۱ , ۲)	۰/۰۳۴۷۱	۲۷/۴۴۶۳۳
۶	(۲ , ۲)	۰/۰۷۴۳۸	۱۴/۸۵۹۶۵
۷	(۳ , ۲)	۰/۱۳۸۸۴	۴۶/۴۷۹۸۶
۸	(۴ , ۲)	۰/۰۲۴۷۹	۲۰/۶۲۱۵۵
۹	(۱ , ۳)	۰/۰۶۴۷۹	۴۴/۲۰۶۸۹
۱۰	(۲ , ۳)	۰/۱۳۸۸۴	۴۶/۴۷۹۸۶
۱۱	(۳ , ۳)	۰/۲۵۹۱۷	۳۱/۶۲۰۲۱
۱۲	(۴ , ۳)	۰/۰۴۶۲۸	۳۷/۳۸۲۱۱
۱۳	(۱ , ۴)	۰/۰۱۱۵۷	۱۸/۳۴۸۵۹
۱۴	(۲ , ۴)	۰/۰۲۴۷۹	۲۰/۶۲۱۵۵
۱۵	(۳ , ۴)	۰/۰۴۶۲۸	۳۷/۳۸۲۱۱
۱۶	(۴ , ۴)	۰/۰۰۸۲۶	۵/۷۶۱۹۰

ما *نمونه‌گیری با احتمال متناسب با اندازه* را (که با نماد *نمونه‌گیری PPS* خلاصه می‌شود) به عنوان یک رده از نمونه‌گیری با احتمال نابرابر تعریف می‌کنیم که در آن احتمال انتخاب شدن یک واحد در نمونه متناسب با سطح متغیری خاص در آن واحد است. مثالهایی از نمونه‌گیری با احتمال متناسب با اندازه در زیر ارائه می‌شود:

- شهری دارای ۳۵ مغازهٔ خواربارفروشی است و قرار است یک آمارگیری نمونه‌ای به منظور برآورد کردن کل مقدار غذای گربه که طی هفتهٔ گذشته به فروش رفته است اجرا شود. طرح

مزبور مستلزم آن است که نمونه‌ای متشکل از ۱۰ مغازه با احتمال متناسب با کل درآمد ناخالص در سال گذشته انتخاب شود.

- یک سازمان بزرگ حفظ بهداشت (HMO) در دو سال گذشته ادعانامه‌هایی برای ۳۲۹ نفر از بیمه‌شدگان سازمان مراقبتهای پزشکی تنظیم کرده است. هدف از این آمارگیری برآورد کردن کل مبلغی است که سازمان مراقبتهای پزشکی به HMO اضافه پرداخت کرده است. ادعانامه‌ها برای بیمه‌شدگان روی رایانه تنظیم شده‌اند. قرار است نمونه‌ای متشکل از ۲۵ بیمه شده گرفته شود به طوری که احتمال انتخاب یک بیمه شده درنمونه متناسب باشد با کل تعداد ادعانامه‌هایی که از جانب بیمه شده تنظیم شده است.
- قرار است یک نمونهٔ خوشه‌ای دومرحله‌ای گرفته شود که در مرحلهٔ اول آن، بلوکهای شهری با احتمال انتخاب شدن هر بلوک در نمونه متناسب با تعداد معلوم خانوارها در بلوک انتخاب خواهند شد. در مرحلهٔ دوم، قرار است از هر بلوک شهری که در مرحلهٔ اول انتخاب شده است نمونه ای متشکل از ۱۰ خانوار گرفته شود.

در بحثی که متعاقباً دربارهٔ نمونه‌گیری PPS ارائه خواهد شد، توجه خود را منحصراً بر کاربرد آن در نمونه‌گیری خوشه‌ای دومرحله‌ای متمرکز می‌کنیم. همان‌طور که در بالا اشاره شد، هر گاه خوشه‌ها از لحاظ تعداد واحدهای شمارش موجود در آنها تفاوت بسیار زیادی داشته باشند نمونه‌گیری خوشه‌ها با احتمال نابرابر غالباً به برآوردهایی از مشخصه‌های جامعه، به خصوص مجموعهای جامعه‌ای، منجر خواهد شد که خطاهای معیار آنها کمتر از برآوردهایی است که از نمونه‌گیری خوشه‌ها با احتمال برابر به دست می‌آیند.

از میان راههای گوناگون نمونه‌گیری از خوشه‌ها با احتمال نابرابر، معقول به نظر می‌رسد که نمونه با احتمال متناسب با سطح نوعی مشخصهٔ معلوم خوشه که با متغیر موردنظر آمارگیری نمونه‌ای در ارتباط است انتخاب شود. یک مثال آرمانی شده و رهگشا در وضعیت زیر دیده می‌شود. فرض کنید می‌خواهیم Y ، مجموع یک متغیر را در جامعه‌ای برآورد کنیم و می‌دانیم که با X، مجموع متغیر معلوم دیگری در جامعه متناسب است. به عبارت دیگر، $Y = cX$ ، که c نامعلوم است. فرض کنید که این رابطه در داخل هر خوشه i نیز مصداق دارد یعنی $Y_i = cX_i$. اگر نمونه‌ای از یک خوشه را با احتمال متناسب با مجموع معلوم، X_i، در آن خوشه انتخاب کنیم، در آن صورت برآورد هنسن ـ هورویتز برای Y از فرمول زیر به دست می‌آید

$$y'_{hh} = \frac{Y_i}{\left(\frac{X_i}{X}\right)} = Y_i\left(\frac{X}{X_i}\right) = \left(\frac{Y_i}{X_i}\right)X$$

ولی

$$\frac{Y_i}{X_i} = \frac{Y}{X} = c$$

پس

$$y'_{hh} = \left(\frac{Y}{X}\right)X = Y$$

بنابراین، هر خوشه‌ای که در نمونه انتخاب شود برآوردگری از Y به دست می‌آید که بدون خطاست.

واضح است که مثال ارائه شده در بالا رابطهٔ قطعی کاملی را بین Y و X فرض می‌گیرد که اگر هم در عمل اتفاق بیفتد بسیار نادر است. ولی اگر همبستگی محکمی بین Y و X در داخل هر خوشه وجود داشته باشد در آن صورت، برآوردگری مبتنی بر نمونه‌گیری با احتمال متناسب با اندازه نسبتاً از انتخاب خوشه‌هایی خاص در نمونه تأثیر نمی‌گیرد و از این رو در مقایسه با شیوهٔ برآوردی که از نمونه‌گیری با احتمال متناسب با اندازه استفاده نمی‌کند خطای نمونه‌گیری کمتری خواهد داشت. باید متذکر شد دلایلی که باعث می‌شوند نمونه‌گیری با احتمال متناسب با اندازه، برآوردهایی تولید کند که خطاهای نمونه‌گیری نسبتاً کمی دارند بسیار شبیه به دلایلی‌اند که برآوردگرهای نسبتی را بسیار قابل اعتماد می‌سازند.

۱.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری: استفاده از برآوردگر هنسن ـ هورویتز

همان طور که در بالا اشاره شد، نمونه‌گیری با احتمال متناسب با اندازه یک ردهٔ کلی از طرحهای نمونه‌ای مبتنی بر نمونه‌گیری با احتمال نابرابر است که در آن، احتمال انتخاب یک واحد در نمونه متناسب است با نوعی متغیر معلوم X . برنامهٔ نمونه‌گیری می‌تواند با یا بدون جایگذاری باشد و می‌تواند به هر شکلی از نمونه‌گیری احتمالاتی باشد (مانند نمونه‌گیری تصادفی ساده، نمونه‌گیری سیستماتیک و غیره). در این زیربخش، صورتی خاص از نمونه‌گیری با احتمال متناسب با اندازه را برای نمونه‌گیری خوشه‌ای دومرحله‌ای شرح و بسط می‌دهیم که دارای خواص زیر است:

۱. در مرحلهٔ اول، نمونه‌ای متشکل از m خوشه با جایگذاری و با احتمال متناسب با سطح نوعی متغیر معلوم x در خوشه که فرض می‌شود با متغیر موردنظر، y، در ارتباط است گرفته می‌شود.
۲. از هر یک از m خوشه‌ای که در مرحلهٔ اول انتخاب شده است، یک نمونهٔ تصادفی ساده متشکل از $\bar{n}$ واحد شمارش از میان N_i واحد شمارش موجود در خوشه گرفته می‌شود.
۳. برآورد مجموع جامعه‌ای، y'_{ppswr} ، از فرمول زیر به دست می‌آید

$$y'_{ppswr} = \frac{1}{m}\sum_{i=1}^{m}\frac{N_i}{\bar{n}\pi'_i}y_i \qquad (۸.۱۱)$$

که در آن،

$$\pi'_i = \frac{X_i}{X}$$

$y_i = \sum_{j=1}^{\bar{n}} y_{ij}$ = مجموع نمونه‌ای i امین خوشهٔ نمونه

$\bar{n}$ = تعداد واحدهای شمارش نمونه‌گیری شده به ازای خوشه

برآوردگری که در برابری (۸.۱۱) نشان داده شده است، تعمیم برآوردگر هنسن ـ هورویتز است که برای نمونه‌گیری خوشه‌ای دومرحله‌ای مناسب است. اگر معیار متغیر اندازه، X_i، عبارت از تعداد، N_i واحد شمارش موجود در خوشه باشد، آن‌گاه، برآوردگر برابر است با

$$y'_{ppswr} = \frac{N}{n}\sum_{i=1}^{m} y_i \qquad (۹.۱۱)$$

همانند عبارتی که در برابری (۷.۱۱) برای خطای معیار برآوردگر هنسن ـ هورویتز در نمونه‌گیری خوشه‌ای یک‌مرحله‌ای ارائه شد، از فرمول زیر می‌توان در نمونه‌گیری دومرحله‌ای با احتمال متناسب با اندازه به عنوان برآوردگر خطای معیار y'_{ppswr} استفاده کرد:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{\sum_{i=1}^{m}\left(\frac{N_i X y_i}{\bar{n}X_i} - y'_{ppswr}\right)^2}{m(m-1)}} \qquad (۱۰.۱۱)$$

مثال تشریحی: جامعهٔ متشکل از ده بیمارستان را در نظر می‌گیریم که در جدول ۹.۱۰ نشان داده شده است (داده‌ها در زیر تکرار می‌شوند).

فرض کنید می‌خواهیم کل تعداد بیماران مرده ترخیص شده را از روی یک نمونهٔ خوشه‌ای دومرحله‌ای برآورد کنیم که در آن یک نمونهٔ تصادفی از دو بیمارستان در مرحلهٔ اول با جایگذاری و با احتمال متناسب با اندازه، N_i پذیرش (که از داده‌های اداری معلوم‌اند) انتخاب می‌شود و در مرحلهٔ دوم، یک نمونهٔ تصادفی ساده متشکل از ۵۰ پذیرش بیمار از هر بیمارستان نمونه گرفته می‌شود.

فرض کنید بیمارستانهای ٦ و ۹ انتخاب شده باشند و ۵۰ سابقهٔ پذیرش نمونه‌گیری شده از بیمارستان شمارهٔ ٦، تعداد ۱۰ نفر را دارای شرایط مهلک نشان داده باشد که ۰ نفر از آنها مرده ترخیص شده‌اند؛ و ۵۰ سابقهٔ نمونه‌گیری شده از بیمارستان شمارهٔ ۹ تعداد ۱۰ نفر را دارای شرایط مهلک نشان داده باشد که ۲ نفر از آنها مرده ترخیص شده‌اند.

جدول ۹.۱۰ کل پذیرشهای دارای شرایط مهلک
و کل پذیرشهایی که از ده بیمارستان مرده ترخیص شده‌اند، ۱۹۸۷

بیمارستان	کل پذیرشها N_i	احتمال انتخاب شدن بیمارستان در هر قرعه‌کشی $\pi'_i = N_i / N$	کل افراد دارای شرایط مهلک	کل ترخیص‌شدگان مرده در میان افراد دارای شرایط مهلک
۱	٤۲۸۸	۰/۰۸٥٦٦٤	٥۰۱	٤۲
۲	٥۰۳٦	۰/۱۰۰٦۰۷	۷۸٥	۷۸
۳	۱۱۷۸	۰/۰۲۳٥۳٤	۲۱۳	۱۷
٤	٦۳۸	۰/۰۱۲۷٤٦	۱۷۳	۹
٥	۲۷۰۱۰	۰/٥۳۹٥۹٦	۳٤۰٤	۳۳۸
٦	۱۱۲۲	۰/۰۲۲٤۱٥	۲۱۷	۱۷
۷	۲۱۳٤	۰/۰٤۲٦۳۲	٤۲٤	۳۷
۸	۱۸۲٤	۰/۰۳٦٤۳۹	۲٤٦	۱۸
۹	٤٦۷۲	۰/۰۹۳۳۳٥	۷۷۸	٦۸
۱۰	۲۱٥٤	۰/۰٤۳۰۳۲	۳٤٦	۲۷

از برابریهای (۹.۱۱) و (۱۰.۱۱) اطلاعات زیر را داریم:

$$m = ۲ \qquad N = ٥۰۰٥٦ \qquad \bar{n} = ٥۰$$
$$y_۱ = ۰ \qquad N_۱ = ۱۱۲۲$$
$$y_۲ = ۲ \qquad N_۲ = ۴۶۷۲$$
$$y'_{ppswr} = \frac{٥۰۰٥٦}{۲ \times ٥۰} \times (۰ + ۲) = ۱۰۰۱/۱۲$$
$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{(۰ - ۱۰۰۱/۱۲)^۲ + (۲۰۰۲/۲۴ - ۱۰۰۱/۱۲)^۲}{۲(۲-۱)}} = ۱۰۰۱/۱۲$$

□

مثال تشریحی: حالا به دومین مثال بر مبنای داده‌های جدول ۹.۱۰ می‌پردازیم. فرض کنید یک نمونه با احتمال متناسب با اندازه گرفته‌ایم که در آن پنج بیمارستان با جایگذاری و با احتمال متناسب با N_i، تعداد پذیرشها انتخاب شده‌اند. از هر یک از بیمارستانهای انتخاب شده یک نمونهٔ تصادفی ساده متشکل از ۱۰ پذیرش را انتخاب می‌کنیم. از روی این نمونه می‌خواهیم کل تعداد اشخاص پذیرفته شده در بیمارستان و با بیماری مهلک و نسبت اشخاص ترخیص شدهٔ مرده از میان این اشخاص دچار بیماری مهلک را برآورد کنیم. مانند مثال قبل از تعمیم برآوردگر هنسن ـ هورویتز استفاده خواهیم کرد.

از پنج بار قرعه‌کشی، بیمارستانهای ۲، ۵، ۵، ۵ و ۹ انتخاب و یافته‌های مربوط در جدول ٦.۱۱ ارائه شده‌اند.

جدول ٦.۱۱ نتایج حاصل از نمونه با احتمال متناسب با اندازه

بیمارستان	احتمال انتخاب در هر نوبت قرعه‌کشی (π'_i)	تعداد پذیرشهای نمونه‌گیری شده ($\bar{n}$)	تعداد بیماران دارای شرایط مهلک (x_i)	تعداد بیماران ترخیص شده مرده (y_i)
۲	$\frac{۵۰۳۶}{۵۰۰۵۶}$	۱۰	۱	۱
۵	$\frac{۲۷۰۱۰}{۵۰۰۵۶}$	۱۰	۱	۰
۵	$\frac{۲۷۰۱۰}{۵۰۰۵۶}$	۱۰	۲	۰
۵	$\frac{۲۷۰۱۰}{۵۰۰۵۶}$	۱۰	۱	۰
۹	$\frac{۴۶۷۲}{۵۰۰۵۶}$	۱۰	۱	۱

می‌بینیم که $m = ۵$ و $\bar{n} = ۱۰$ ، و از برابری (۹.۱۱):

$$x'_{ppswr} = \frac{۵۰۰۵۶}{۵۰} \times ۶ = ۶۰۰۶/۷۲$$

و

$$y'_{ppswr} = \frac{۵۰۰۵۶}{۵۰} \times ۲ = ۲۰۰۲/۲۴$$

در برآورد کردن خطاهای معیار این مجموعهای برآورد شده متوجه می‌شویم که هرگاه معیار متغیر اندازه، N_i، تعداد واحدهای شمارش باشد فرمول مربوط به خطای معیار برآورد شده که در برابری (۱۰.۱۱) نشان داده شد به شکل زیر تبدیل می‌شود:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{\sum_{i=۱}^{m}\left(\frac{N\,y_i}{\bar{n}} - y'_{ppswr}\right)^۲}{m\,(m-۱)}} \qquad (۱۱.۱۱)$$

از برابری (۱۱.۱۱)، خطاهای معیار زیر را محاسبه می‌کنیم:

$$\hat{SE}(x'_{ppswr}) = ۱۰۰۱/۱۲$$

و

$$\hat{SE}(y'_{ppswr}) = ۱۲۲۶/۱۷$$

بالاخره، نسبت ترخیص‌شدگان مرده در میان پذیرفته‌شدگان دچار بیماریهای مهلک با برآورد نسبتی زیر برآورد می‌شود

$$r_{ppswr} = \frac{y'_{ppswr}}{x'_{ppswr}} = \frac{۲۰۰۲/۲۴}{۶۰۰۶/۷۲} = ۰/۳۳$$

برآوردی از خطای معیار این برآوردگر نسبتی را می‌توان با روشهای خطی کردن به دست آورد که ذیلاً در بحث تحلیل نرم‌افزاری این طرح نشان داده خواهد شد.

اکنون استفاده از SUDAAN و STATA را در تهیهٔ برآوردهایی برای این طرح خاص با احتمال متناسب با اندازه نشان خواهیم داد.

پروندهٔ اطلاعاتی مورد استفاده برای SUDAAN دارای ۵۰ سابقه است که هر سابقه اطلاعاتی را از یک سابقهٔ پذیرش نمونه‌گیری شده در مرحلهٔ دوم پس از انتخاب بیمارستان در مرحلهٔ اول شامل است. به این ترتیب، پروندهٔ اطلاعاتی از ۵۰ سابقهٔ پذیرش ـ یعنی ۱۰ سابقه برای هر یک از پنج بیمارستان قرعه‌کشی شده ـ تشکیل شده است. پروندهٔ اطلاعاتی SUDAAN یک پروندهٔ SAS PC به نام *hospslct.ssd.* است. داده‌های موجود در این پرونده در جدول ۷.۱۱ نشان داده شده‌اند و متغیرهای موجود در پرونده در زیر توصیف شده‌اند:

drawing متغیری است که انتخاب ویژهٔ یک بیمارستان را در مرحلهٔ اول نشان می‌دهد. چون ۵ قرعه‌کشی جداگانه و مستقل انجام شده است مقدار این متغیر بین ۱ تا ۵ است.

hospno نشان‌دهندهٔ بیمارستانی خاص است که سابقهٔ پذیرش از آن نمونه‌گیری شده است (و می‌تواند دارای مقادیری از ۱ تا ۱۰ باشد، زیرا ۱۰ بیمارستان در جامعه بوده‌اند).

admiss نشان‌دهندهٔ تعداد پذیرشهایی است که در طی آن سال در بیمارستانهایی که سوابق پذیرش از آنها قرعه‌کشی شده روی داده‌اند.

Lifethrt نشان می‌دهد که آیا سابقهٔ خاص معرف پذیرش بیماری با شرایط مهلک است یا نه ـ اگر جواب مثبت باشد مقدار آن ۱ و اگر منفی باشد ۰ خواهد بود.

dxdead نشان می‌دهد که بیمار پذیرش شده مرده ترخیص شده است یا نه ـ اگر جواب مثبت باشد ۱ و اگر منفی باشد ۰ خواهد بود.

wstar وزن نمونه‌گیری است ـ که در این مورد برابر است با N/n یا $\frac{۵۰۰۵۶}{۵۰} = ۱۰۰۱/۱۲$.

جدول ۷.۱۱ پروندهٔ داده‌ها برای استفاده در مثال تشریحی

drawing	hospno	admiss	lifethrt	dxdead	wstar
1	2	5036	1	1	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
1	2	5036	0	0	1001.12
2	5	27010	1	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
2	5	27010	0	0	1001.12
3	5	27010	1	0	1001.12
3	5	27010	1	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
3	5	27010	0	0	1001.12
4	5	27010	1	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
4	5	27010	0	0	1001.12
5	9	4672	1	1	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12
5	9	4672	0	0	1001.12

فرمانهای SUDAAN برای این برآورد ذیلاً نشان داده می‌شود:

```
proc descript data = hospslct   filetype = SAS   design = wr  means totals;
nest  drawing / psulev =1;
weight  wstar;
var  lifethrt  dxdead;
proc  ratio  data = hospslct  filetype = SAS  design = wr;
nest  drawing / psulev =1;
weight  wstar;
numer dxdead;
denom lifethrt;
```

فرمان اول، مدول خاص SUDAAN را که قرار است مورد استفاده قرار گیرد، پروندهٔ اطلاعاتی و نوع آن، طرح خاص SUDAAN (که در این مورد WR یا نمونه‌گیری با جایگذاری است)، و محاسبهٔ میانگینها و مجموعها هر دو را نشان می‌دهد. فرمان دوم، خوشه‌بندی را توصیف می‌کند و نشان می‌دهد که واحد نمونه‌گیری اولیه در متغیر *drawing* شناسایی شده است. باید دقیقاً توجه داشت که متغیر *hospno* که نشانهٔ شناسایی بیمارستان است واحد نمونه‌گیری اولیه نیست. در نمونه‌گیری با احتمال متناسب با اندازه (PPS) و با جایگذاری، یک خوشهٔ معین (در این مورد یعنی بیمارستان) می‌تواند بیش از یک بار در نمونه انتخاب شود و هر نوبت قرعه‌کشی یک خوشه یک واحد نمونه‌گیری اولیه به شمار می‌رود. فرمان سوم، وزن نمونه‌گیری را نشان می‌دهد و فرمان چهارم نشان‌دهندهٔ متغیری است که مجموعها و میانگینهای آن باید برآورد شوند. بقیهٔ فرمانها برای برآورد کردن نسبتی به کار می‌روند و گویا هستند. خروجی متناظر SUDAAN برای این برآورد به صورت زیر است:

by: Variable, One.		
Variable		One 1
LIFETHRT	Sample Size Weighted Size Total SE Total Mean SE Mean	50 50056.00 6006.72 1001.12 0.12 0.02
DXDEAD	Sample Size Weighted Size Total SE Total Mean SE Mean	50 50056.00 2002.24 1226.12 0.04 0.02
by: Variable, One.		
Variable		One 1
DXDEAD/ LIFETHRT	Sample Size Weighted Size Weighted X-sum Weighted Y-sum Ratio Est. SE Ratio	50.0000 50056.0000 6006.7200 2002.2400 0.3333 0.2324

تحلیلی مشابه را می‌توان با استفاده از STATA به وسیلهٔ فرمانهای زیر اجرا کرد:

```
use “ a:\hospslct.dta”,clear
. svyset pweight wstar
. svyset psu drawing
. svytotal lifethrt dxdead
.svyratio dxdead lifethrt
```

که خروجی زیر را نتیجه می‌دهد:

```
Survey total estimation

pweight:  wstar                          Number of obs    =      50
Strata:   < one >                        Number of strata =       1
PSU:      drawing                        Number of PSUs   =       5
                                         Population size  =   50056
```

Total	Estimate	Std. Err.	[%95 Conf.	Interval]	Deff
lifethrt	6006.72	1001.12	3227.165	8786.275	.1856061
dxdead	2002.24	1226.117	-1402.005	5406.485	.765625

```
.svyratio dxdead lifethrt

Survey ratio estimation

pweight:  wstar                          Number of obs    =      50
Strata:   < one >                        Number of strata =       1
PSU:      drawing                        Number of PSUs   =       5
                                         Population size  =   50056
```

Ratio	Estimate	Std. Err.	[%95 Conf.	Interval]	Deff
Dxdead/lifethrt	.3333333	.2324056	-.3119279	.9785946	1.429167

□

۲.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازه وقتی معیار متغیر اندازه، تعداد واحدهای شمارش نباشد.

گاهی در وضعیتی قرار می‌گیریم که استفاده از N_i، تعداد واحدهای شمارش موجود در داخل خوشه به عنوان معیار متغیر اندازه در نمونه‌گیری با احتمال متناسب با اندازه، نه امکانپذیر است و نه مطلوب. اغلب اوقات، تعداد واحدهای شمارش پیش از نمونه‌گیری معلوم نیست. ولی گاهی اوقات، تعداد واحدهای شمارش موجود در داخل خوشه ممکن است ارتباط خاصی با متغیر پاسخ نداشته باشد و به

این لحاظ استفاده از آن به عنوان معیار برای متغیر اندازه فایدهٔ چندانی نخواهد داشت. مثلاً اگر بیمارستانها خوشه‌ها باشند و برآورد کردن نوعی مشخصهٔ بیماران دچار سرطان پوست موردنظر باشد، تعداد بیماران دچار سرطان پوست که در یک بیمارستان خاص پذیرش شده‌اند احتمال دارد بیشتر با اندازه و کیفیت بخشهای تخصصی بیماریهای پوستی و غده‌شناسی بیمارستان ارتباط داشته باشد تا به اندازهٔ کلی بیمارستان که با مجموع تعداد پذیرشها در آن بیمارستان مشخص می‌شود.

در وضعیتهایی که معیار متغیر اندازه که در ایجاد نمونهٔ PPS به کار رفته است تعداد واحدهای شمارش نیست، فرمولهای مناسب برای برآورد کردن مجموع جامعه و برآورد کردن خطای معیار این برآوردگر مجموع به ترتیب، برابریهای (۸.۱۱) و (۱۰.۱۱) هستند. در هر دو وضعیت، مجموع برآورد شده، مجموعی موزون از مشاهدات نمونه‌ای تک است که به صورت $y'_{ppswr} = \sum_{i=1}^{m}\sum_{j=1}^{\bar{n}} w_i y_{ij}$ است. در وضعیتی که قبلاً بحث شد و در آن معیار متغیر اندازه متناسب با N_i، تعداد واحدهای شمارش موجود در خوشه بود، وزن، w_i، برابر است با N/n و بنابراین، مستقل از خوشه‌ای که از آن نمونه‌گیری شده است، برای هر مشاهدهٔ نمونه یکسان است. در وضعیتی که در آن معیار متغیر اندازه، تعداد واحدهای شمارش موجود در خوشه نیست، وزن، w_i، برابر است با $(N_i X)/(nX_i)$ و بنابراین برای هر خوشه یکسان نیست.

در مثال تشریحی که در بالا مورد بحث قرار گرفت، اگر معیار متغیر اندازهٔ مورد استفاده در نمونه‌گیری خوشه‌ها با احتمال متناسب با اندازه، X_i، تعداد پذیرشهای با شرایط مهلک در بیمارستان باشد، آن‌گاه، همان داده‌های نمونه که در جدول ۷.۱۱ نشان داده شدند برآوردهای زیر را به دست می‌دهند (فرمانهای SUDAAN و خروجی آن در زیر نشان داده شده‌اند):

```
proc  descript data = hspslct2 filetype = sas means totals;
nest drawing/psulev = 1;
weight w2star;
var lifethrt dxdead;
proc ratio data = hspslct2  filetype = sas;
nest drawing/psulev = 1;
weight w2star;
numer dxdead;
denom lifethrt;
```

در این فرمانها، متغیر *w2star* برابر است با $(N_i X)/(nX_i)$.

خروجی مربوط به برآورد مجموعها به صورت زیر است:

LIFETHRT	Sample Size Weighted Size Total SE Total Mean SE Mean	50 51344.99 6259.18 1277.32 0.12 0.02
DXDEAD	Sample Size Weighted Size Total SE Total Mean SE Mean	50 51344.99 1760.47 1079.04 0.03 0.02

خروجی مربوط به برآورد نسبت به شکل زیر خواهد بود:

Variable		One 1
DXDEAD/LIFETHRT	Sample Size Weighted Size Weighted X-Sum Weighted Y-Sum Ratio Est. SE Ratio	50 51344.99 6259.18 1760.47 0.28 0.21

۳.۳.۱۱ چگونگی انتخاب یک نمونه با احتمال متناسب با اندازه و با جایگذاری

از نظر عملیاتی، انتخاب یک نمونهٔ دومرحله‌ای با احتمال متناسب با اندازه و با جایگذاری بسیار آسان است. این روش را باز هم با استفاده از جامعهٔ بیمارستانهای جدول ۹.۱۰ نشان می‌دهیم. باز فرض می‌کنیم که N_i، معیار متغیر اندازه‌ای است که قرار است در انتخاب نمونه مورد استفاده قرار گیرد. ابتدا معیار متغیر اندازه (در این مورد N_i) را انباشته می‌کنیم و اعداد تصادفی را با هر خوشه به صورتی که در جدول ۸.۱۱ نشان داده می‌شود همراه می‌کنیم. تعداد اعداد تصادفی همراه با هر خوشه، به شکلی که در جدول نشان داده شده برابر با N_i است.

اگر قرار باشد m خوشه برای نمونه انتخاب و $\bar{n}$ واحد نمونه‌گیری در هر قرعه‌کشی از خوشه نمونه‌گیری شود، شیوهٔ کار مستلزم (۱) انتخاب یک عدد تصادفی بین ۱ و N ؛ (۲) شناسایی خوشهٔ متناظر با عدد تصادفی؛ (۳) انتخاب یک نمونهٔ تصادفی ساده بدون جایگذاری با $\bar{n}$ واحد شمارش در داخل خوشه؛ و (٤) تکرار این روش تا انتخاب m خوشه و $n = m\bar{n}$ واحد شمارش خواهد بود.

برای نشان دادن این روش، فرض کنید $m = ۵$ و $\bar{n} = ۶$. ابتدا یک عدد تصادفی بین ۰۰۰۰۱ و ۵۰۰۵٦ انتخاب می‌کنیم. فرض کنید این عدد ۳٦۲۰۷ است. این عدد با بیمارستان ۵ متناظر

است. سپس ٦ عدد تصادفی را بین ۰۰۰۱ و ۲۷۰۱۰ بدون جایگذاری انتخاب می‌کنیم و پذیرشهای متناظر با این اعداد را برای نمونه‌های آن خوشهٔ انتخاب شدهٔ خاص در نظر می‌گیریم. بعد یک عدد تصادفی دیگر را از میان اعداد ۰۰۰۰۱ و ۵۰۰۵٦ انتخاب می‌کنیم تا بیمارستان دوم انتخاب شود. فرض کنید این عدد تصادفی ٤۲۷۵۱ است. این عدد با بیمارستان ۸ متناظر است که ۱۸۲٤ پذیرش داشته است. پس ٦ عدد تصادفی را بین ۰۰۰۱ و ۱۸۲٤ بدون جایگذاری انتخاب می‌کنیم تا ٦ پذیرش از این بیمارستان انتخاب شود. این فرایند را ادامه می‌دهیم تا شش پذیرش از هر پنج بیمارستان انتخاب شود. از جدول ۸.۱۱ درمی‌یابیم که در این تمرین سه بار بیمارستان ۵ و یک بار بیمارستانهای ۲ و ۸ انتخاب شده‌اند. پذیرشهای ویژه‌ای که در مرحلهٔ دوم نمونه‌گیری شده‌اند در ستون ششم جدول ۸.۱۱ نشان داده شده‌اند.

٤.۳.۱۱ نمونهٔ مورد نیاز برای نمونهٔ دومرحله‌ای که در آن خوشه‌ها با احتمال متناسب با اندازه و با جایگذاری انتخاب می‌شوند چقدر باید بزرگ باشد؟

در نمونه‌گیری خوشه‌ای دومرحله‌ای، خطای معیار یک پارامتر برآورد شده نه تنها از n ، کل تعداد واحدهای شمارش نمونه‌گیری شده بلکه با استفاده از دو عاملی که n را تشکیل می‌دهند یعنی m ، تعداد خوشه‌های نمونه‌گیری شده و $\bar{n}$، متوسط تعداد واحدهای شمارش نمونه‌گیری شده به ازای هرخوشه تعیین می‌شود. به این ترتیب، مثلاً برآورد مجموع یک جامعه، که از یک نمونهٔ دومرحله‌ای متشکل از ۸۰ واحد شمارش در ۱۰ خوشهٔ نمونه که هر یک ۸ واحد شمارش نمونه دارند به دست آمده است، احتمال دارد خطای معیاری داشته باشد که با برآورد حاصل از یک نمونهٔ متشکل از ۲۰ خوشه که هر یک ٤ واحد شمارش نمونه‌ای دارند فرق داشته باشد. ولی غالباً $\bar{n}$، تعداد واحدهای شمارش که قرار است نمونه‌گیری شوند پیشاپیش بر مبنای هزینه و ملاحظات مربوط به همبستگی میان ـ رده‌ای تعیین می‌شود (مانند مواردی که درفصل ۱۰ بحث شد) به طوری که مسئلهٔ تعیین اندازهٔ نمونه، به مسئلهٔ تعیین m، تعداد خوشه‌هایی که باید نمونه‌گیری شوند، به فرض آنکه $\bar{n}$ واحد شمارش به ازای هر خوشهٔ نمونه انتخاب شود تبدیل خواهد شد. ما درفرمول‌بندیهای خود برای مسئلهٔ اندازهٔ مورد نیاز نمونه در نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری، فرضهای زیر را در نظر می‌گیریم:

۱. $\bar{n}$، تعداد واحدهای شمارش که باید در داخل هر خوشه نمونه‌گیری شوند برای هرخوشهٔ نمونه یکسان و از قبل تعیین شده است.

جدول ۸.۱۱ شیوهٔ نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری

بیمارستان	مجموع پذیرشها N_i	پذیرشهای انباشته	اعداد تصادفی	اعداد تصادفی انتخاب شده برای مرحلهٔ اول	اعداد تصادفی انتخاب شده برای مرحلهٔ دوم
۱	٤۲۸۸	٤۲۸۸	۰۰۰۰۱–۰٤۲۸۸		
۲	۵۰۳٦	۹۳۲٤	۰٤۲۸۹–۰۹۳۲٤	۰۸۵۸۹	۰٤۸۰، ۲۳٦۸، ٤۱۳۰، ۲۱٦۷، ۳٤۷۵، ۳۵۵۳
۳	۱۱۷۸	۱۰۵۰۲	۰۹۳۲۵–۱۰۵۰۲		
٤	٦۳۸	۱۱۱٤۰	۱۰۵۰۳–۱۱۱٤۰		
۵	۲۷۰۱۰	۳۸۱۵۰	۱۱۱٤۱–۳۸۱۵۰	۳٦۲۰۷	۰۹٤۲۹، ۱۰۳٦۵، ۰۷۱۱۹، ۰۲۳٦۸، ۰۱۰۱۱، ۰۷۰۵٦
				۱۸٦۰۲	۱٦٦۳۱، ۱٦۸۱۵، ۲۰۲۰٦، ۰۵۳۰۰، ۲۲۱٦٤، ۲٤۳٦۹
				۱۸۷۳۸	۱۹٦۸۷، ۱۱۰۵۲، ۱۹۷٤٦، ۱٤۳٤۹، ۰۰٦۹۷، ۱۹۱۲٤
٦	۱۱۲۲	۳۹۲۷۲	۳۸۱۵۱–۳۹۲۷۲		
۷	۲۱۳٤	٤۱٤۰٦	۳۹۲۷۳–٤۱٤۰٦		
۸	۱۸۲٤	٤۳۲۳۰	٤۱٤۰۷–٤۳۲۳۰	٤۲۷۵۱	۱٦۲٤، ۱٦۰۱، ۱۲٤۸، ۱۸۰۵،۰۱۱۹، ۰۹۰۳
۹	٤٦۷۲	٤۷۹۰۲	٤۳۲۳۱–٤۷۹۰۲		
۱۰	۲۱۵٤	۵۰۰۵٦	٤۷۹۰۳–۵۰۰۵٦		

۲. برآوردی «مقدماتی» از نمونهٔ خوشه‌ای با احتمال متناسب با اندازه از سطح متغیر موردنظر (که در بحث ما مجموع جامعه‌ای آن متغیر است) وخطای معیار آن موجود است. (بنابراین، ضریب تغییرات برآورد را می‌دانیم.)

۳. می‌خواهیم با اطمینان $(1-\alpha)\times 100$ درصد تعداد خوشه‌های نمونه‌ای مورد نیاز برای برآورد کل سطح متغیری خاص را در جامعه حول $\varepsilon\times 100$ درصد مقدار واقعی آن بر آورد کنیم.

فرض می‌کنیم که معیار متغیر اندازه برای نمونه‌گیری با احتمال متناسب با اندازه، تعداد واحدهای شمارش در داخل هر خوشه است که نتیجه می‌دهد که برآورد مقدماتی خطای معیار از فرمول برابری (۱۱.۱۱) به دست می‌آید. توجه کنید که این برآورد به صورت زیر است:

$$\hat{SE}(y'_{ppswr}) = \sqrt{\frac{G(y_1, y_2, \ldots, y_m)}{m}}$$

که در آن:

$$G(y_1, y_2, \ldots, y_m) = \frac{\sum_{i=1}^{m}\left(\frac{Ny_i}{\bar{n}} - y'_{ppswr}\right)^2}{m-1}$$

از این فرمول، داریم

$$m = \frac{z^2_{1-\alpha/2}\, G(y_1, y_2, \ldots, y_m)}{(y'_{ppswr})^2 \varepsilon^2} \qquad (۱۲.۱۱)$$

و برآورد m، تعداد خوشه‌های مورد نیاز برای تأمین ویژگیهای فوق‌الذکر تقریباً برابر است با آنچه در برابری (۱۲.۱۱) نشان داده شده است.

مثال تشریحی: باز هم جامعهٔ متشکل از ۱۰ بیمارستان جدول ۹.۱۰ را در نظر می‌گیریم. فرض کنید می‌خواهیم یک نمونه با احتمال متناسب با اندازه و با جایگذاری بگیریم که مرحلهٔ اول آن از m قرعه‌کشی مستقل از یک بیمارستان تشکیل می‌شود که در هر مرحله از قرعه‌کشی، یک نمونهٔ تصادفی ساده بدون جایگذاری با انتخاب ۱۰ واحد شمارش گرفته می‌شود. باز هم فرض کنید که یک آمارگیری مقدماتی از ۵ بیمارستان به عمل آمده است (با استفاده از همان طرح نمونه‌گیری با احتمال متناسب با اندازه و با جایگذاری که قرار است در آمارگیری اصلی به کار گرفته شود). بالاخره فرض می‌کنیم که داده‌های حاصل از بررسی مقدماتی همان داده‌هایی هستند که در بحث قبلی ما در مورد آن آمارگیری توصیف شده‌اند. حالا می‌خواهیم از این داده‌ها برای محاسبهٔ m، تعداد خوشه‌هایی استفاده

کنیم که باید نمونه‌گیری شوند تا ۹۵ درصد مطمئن باشیم که y'_{ppswr}، برآورد کل Y، تعداد پذیرش‌شدگان با شرایط مهلک، حول ۳۰ درصد مقدار واقعی آن باشد.

داده‌های حاصل از آن آمارگیری مقدماتی که برای برآورد کردن m، تعداد خوشه‌های نمونهٔ مورد نیاز ضروری است ذیلاً نشان داده شده‌اند:

$$y'_{ppswr} = ۶۰۰۶/۷۲$$

$$\hat{SE}(y'_{ppswr}) = ۱۰۰۱/۱۲$$

$$m^* = ۵$$ (تعداد خوشه‌های نمونه‌گیری شده در آمارگیری مقدماتی)

$$G(y_۱, \ldots, y_m) = (۱۰۰۱/۱۲ \times \sqrt{۵})^۲ = (۲۲۳۸/۵۷۲)^۲ = ۵۰۱۱۲۰۶/۲۷۲$$

$$\bar{n} = ۱۰$$

$$\varepsilon = ۰/۳$$

$$z_{۱-(\alpha/۲)} = ۱/۹۶$$

با وارد کردن مقادیر بالا در برابری (۱۲.۱۱) داریم

$$m = ۵/۹ \approx ۶$$

بنابراین، شش قرعه‌کشی از بیمارستانها (در مجموع، ٦۰ نمونهٔ پذیرش) لازم است تا ویژگیهای تعیین شده برای برآورد کل تعداد پذیرش‌شدگان دارای شرایط مهلک تأمین شود.

□

۵.۳.۱۱ نمونه‌گیری با احتمال متناسب با اندازهٔ تلفنی: روش شماره‌گیری ارقام تصادفی میتوفسکی ـ واکسبرگ[1]

نمونه‌گیری تلفنی به قدری در روش‌شناسی آمارگیری اهمیت پیدا کرده است که یک فصل کامل (فصل ۱۵) را به آن اختصاص داده‌ایم تا به جزئیات این موضوع بپردازیم. این روش در فصل حاضر به عنوان مثالی از نمونه‌گیری با احتمال متناسب با اندازه مطرح می‌شود و نشان می‌دهیم که یکی از فنون آن که کاربرد وسیعی دارد و به نام *روش شماره‌گیری ارقام تصادفی میتوفسکی ـ واکسبرگ* معروف است در واقع صورتی از نمونه‌گیری با احتمال متناسب با اندازه است.

کاربرد مصاحبهٔ تلفنی در آمارگیریهای نمونه‌ای در سالهای اخیر بسیار افزایش یافته است، بدواً به این دلیل که هزینه‌های میدانی ناشی از مراجعات حضوری به خانوارها غالباً بازدارنده است. اگر بپذیریم که همهٔ خانوارها تلفن ندارند و اگر کل خانوارهای دارای تلفن به عنوان جامعهٔ هدف قابل قبول باشد، در آن صورت مصاحبهٔ تلفنی غالباً به عنوان جایگزینی برای مراجعهٔ حضوری به خانوار مورد استفاده قرار می‌گیرد.

[1] Mitofsky-Waksberg

در آمارگیریهای تلفنی می‌توان خانوارها را به چندین طریق برای گنجاندن در نمونه انتخاب کرد. مثلاً از کتاب راهنمای تلفن منطقهٔ هدف می‌توان به عنوان چارچوب نمونه‌گیری استفاده کرد. ولی کتابهای راهنمای تلفن، شماره‌های فهرست نشده را و شماره‌هایی را که بعد از تاریخ انتشار آخرین نسخهٔ کتاب راهنما به خانوارها واگذار شده‌اند ندارند. حذف این خانوارها ممکن است منشأ اریبی زیاد باشد زیرا خانوارهایی که شماره‌هایشان فهرست نشده است یا شماره‌های تازه‌ای گرفته‌اند می‌توانند درصد قابل ملاحظه‌ای از خانوارهای دارای تلفن را تشکیل دهند.

چارچوب نمونه‌گیری دیگری برای آمارگیریهای تلفنی، فهرستی از تمام شماره‌های چهار رقمی است که داخل مراکز تلفن موجودند. یک شمـــارهٔ تلفن از ده رقم تشکیل شده است. سه رقـــم اول *کد ناحیه* را نشان می‌دهند، سه رقم بعدی مرکز *تلفن مربوط* را معین می‌کند و چهار رقم آخر مربوط به *شمارهٔ تلفنی* خاص است. مثلاً برای اینکه شمارهٔ تلفن رئیس دانشکدهٔ بهداشت عمومی دانشگاه ایلی‌نوی در شیکاگو را بگیرید، اول باید کد ۳۱۲ (کد *ناحیهٔ* شیکاگو) را شماره‌گیری کنید، بعد باید شـمارهٔ ۹۹٦ (شمارهٔ مرکز تلفن ویژهٔ تمام دانشگاه ایلـی‌نوی در شیـکاگـو) را بگیریـد و پس از آن شـمارهٔ ٦٦٢٣ (خط اختصاصی رئیس دانشکده) را بگیرید. می‌توانید با شماره‌گیری ده رقم ۳۱۲۹۹٦٦٦۲۳ این موضوع را تحقیق کنید. رئیس فعلی این دانشکده فردی بسیار اجتماعی است و از این‌گونه مکالمات تلفنی استقبال می‌کند. استفاده از این ارقام به عنوان چارچوب نمونه‌گیری به نام *شماره‌گیری ارقام تصادفی* مشهور است و از بسیاری از اریبیهایی که غالباً با استفاده از دفترچه راهنمای تلفن همراه است پرهیز می‌کند. ادامهٔ بحث ما در مورد شماره‌گیری ارقام تصادفی صرفاً براساس روشهایی است که در اصل توسط میتوفسکی شرح و بسط داده شده [۹] و بعدها توسط واکسبرگ پیراسته شده و گسترش یافته است [۱۰].

فهرستی از کدهای ناحیه‌ای و مراکز تلفنی منطقهٔ هدف آمارگیری را می‌توان از دفتر شرکت تلفن محلی تهیه کرد. به این ترتیب، در محدودهٔ یک کد محلی (مثلاً ۳۱۲) و مرکز تلفن (مثلاً ۹۹٦) فقط چهار رقم آخر شماره تلفنها باید به طور تصادفی انتخاب شوند. ولی معلوم می‌شود که نسبت بزرگی (تقریباً ۸۰ درصد) از همهٔ شماره تلفنهایی که با یک کد محلی و مرکز تلفن خاص مشخص شده‌اند یا اصلاً استفاده نمی‌شوند یا به مشاغل، مؤسسات، یا سایر امور غیر از خانوارها اختصاص داده شده‌اند. به عبارت دیگر، فرایند گرفتن شماره‌های تصادفی چهاررقمی با یک کد ناحیه‌ای و مرکز تلفن معین بی‌حاصل خواهد بود زیرا باید تقریباً پنج شماره تلفن را برای به دست آوردن یک خانوار بگیریم. میتوفسکی و واکسبرگ شیوهٔ دیگری را پیشنهاد کرده‌اند که عملکرد شماره‌های متناظر با خانوارها را افزایش می‌دهد. در این شیوه، شمارهٔ تلفنها در خوشه‌هایی که (به جای شش رقم اول) با ۸ رقم اول

مشخص شده‌اند گروه‌بندی می‌شوند. مثلاً اگر جامعهٔ هدف شامل خانوارهایی است که کد ناحیه‌ای آنها ۳۱۲ و مراکز تلفنی آنها ۹۹۶ و ۸۳۵ است، شماره‌ها در ۲۰۰ خوشه براساس ۸ رقم اول به شرح زیر گروه‌بندی می‌شوند:

(۱۰۰ خوشه) ۹۹-۸۳۵-۳۱۲ تا ۰۰-۸۳۵-۳۱۲

(۱۰۰ خوشه) ۹۹-۹۹۶-۳۱۲ تا ۰۰-۹۹۶-۳۱۲

خوشه‌ها به شرح زیر (مثلاً ۱ تا ۲۰۰، مانند بالا) شناسایی و نمونه‌گیری می‌شوند. یک خوشهٔ تصادفی انتخاب می‌شود (مثلاً ۴۷-۹۹۶-۳۱۲) و یک عدد تصادفی بین ۰۰ و ۹۹ (مثلاً ۶۰) گرفته می‌شود. بعد شمارهٔ تلفن (مثلاً ۴۷۶۰-۹۹۶-۳۱۲) شماره‌گیری می‌شود. اگر این شماره تلفن، مربوط به یک خانوار باشد خوشه حفظ می‌شود و شماره‌های دورقمی دیگری گرفته می‌شوند تا به مجموع $\bar{n}$ از ساکنان (شامل اولین خانوار شماره‌گیری شده) برسیم. اگر شمارهٔ اولیه مربوط به یک خانوار نباشد خوشه را کنار می‌گذاریم. این شیوه ادامه می‌یابد تا کل m خوشه نمونه‌گیری شود و در نتیجه مجموع $m\bar{n}$ مصاحبه به دست آید.

شیوهٔ میتوفسکی ـ واکسبرگ بهبود قابل توجهی در نسبت خانوارهای شماره‌گیری شده به بار آورده است که علت آن، گرایش شماره تلفنهای اماکن مسکونی به خوشه‌بندی شدن در «ردیفها»یی از شماره‌های دنباله‌ای (پی‌درپی) بود. طرح نمونه‌ای که در بالا شرح داده شد طرح نمونه‌گیری با احتمال متناسب با اندازه است زیرا اولین مکالمهٔ تلفنی انجام شده درخوشه تعیین می‌کند که خوشه در نمونه قرار می‌گیرد یا نه و احتمال اینکه این تلفن اولیه به مکالمه با یک خانوار بینجامد بستگی دارد به نسبت ۱۰۰ شمارهٔ موجود در خوشه که به خانوارها مربوط می‌شوند.

۴.۱۱ توضیح بیشتر دربارهٔ نمونه‌گیری با احتمال متناسب با اندازه

در مبحث نمونه‌گیری با احتمال متناسب با اندازه بر یک برنامهٔ نمونه‌گیری و شیوهٔ برآورد کردن خاص متمرکز شدیم. برنامهٔ نمونه‌گیری مستلزم انتخاب خوشه‌هایی با استفاده از نمونه‌گیری تصادفی ساده با جایگذاری و سپس انتخاب یک نمونهٔ تصادفی ساده (بدون جایگذاری) از واحدهای شمارش در هر خوشهٔ نمونه در هر نوبت انتخاب آن خوشه در نمونه بود. شیوهٔ برآورد کردن مستلزم استفاده از برآوردگر هنسن ـ هورویتز بود. از این نوع نمونه‌گیری PPS در آمارگیریها زیاد استفاده می‌شود زیرا برآوردهای خطاهای معیار را می‌توان چون خوشه‌ها مستقل از یکدیگر انتخاب می‌شوند نسبتاً به آسانی به دست آورد. برآوردگر هنسن ـ هورویتز در ترکیب با این نوع نمونه‌گیری مورد استفاده قرار می‌گیرد زیرا محاسبهٔ آن به مراتب از برآوردگر هورویتز ـ تامپسون آسانتر است.

اما، انواع دیگری از نمونه‌گیری با احتمال متناسب با اندازه وجود دارند که گاهی در عمل مورد استفاده قرار می‌گیرند. مثلاً خوشه‌ها را می‌توان با نمونه‌گیری تصادفی ساده بدون جایگذاری یا با شکل دیگری از نمونه‌گیری بدون جایگذاری انتخاب کرد. چون برآوردگر هنسن ـ هورویتز تنها در صورتی می‌تواند مورد استفاده قرار گیرد که نمونه‌گیری با جایگذاری باشد، در نمونه‌گیری PPS که خوشه‌ها بدون جایگذاری انتخاب می‌شوند غالباً از برآوردگر هورویتز ـ تامپسون استفاده می‌شود. خطاهای معیار برآوردهای حاصل را نمی‌توان به آسانی از این قبیل طرحها به دست آورد، هر چند روشهای گوناگونی برای این کار پیشنهاد شده‌اند [۱].

در طرحهای نمونه‌گیری با احتمال متناسب با اندازه، اگر تعداد واحدهای شمارش که در داخل هر خوشهٔ نمونهٔ مرحلهٔ اول انتخاب می‌شوند یکسان باشد آن‌گاه، احتمال انتخاب شدن هر واحد شمارش در نمونه یکسان است و هیچ فرقی نمی‌کند که واحد شمارش به کدام خوشه تعلق داشته باشد. به عبارت دیگر، نمونه *خود ـ وزن* است به این مفهوم که هر واحد شمارش در نمونه معرف همان تعداد واحد شمارش در جامعه است. این حالت معمولاً در نمونه‌گیری خوشه‌ای دو مرحله‌ای ساده صدق نمی‌کند مگر اینکه خوشه‌ها دارای تعداد واحدهای شمارش یکسان N_i، باشند.

۵.۱۱ خلاصه

در این فصل، از طریق مثالهایی نشان دادیم که نمونه‌گیری خوشه‌ای دو مرحله‌ای ساده می‌تواند برآوردهایی به بار آورد که خطاهای نمونه‌گیری آنها بسیار زیاد است، به خصوص اگر خوشه‌ها از لحاظ N_i، تعداد واحدهای شمارش درون خوشه‌ها پراکندگی قابل ملاحظه‌ای داشته باشند.این موضوع به خصوص در مورد برآورد کردن مجموعهای جامعه‌ای صدق می‌کند. همچنین نشان دادیم که استفاده از یک طرح نمونه که در آن خوشه‌ها با احتمال متناسب با نوعی مشخصهٔ خوشه نمونه‌گیری می‌شوند (مانند تعداد واحدهای شمارش موجود در خوشه) می‌تواند به برآوردهایی منجر شود که خطاهای نمونه‌گیری آنها به مراتب کمتر از خطاهای نمونه‌گیری حاصل از خوشه‌هایی با احتمال برابر است. دو ردهٔ کلی از برآوردگرها یعنی برآوردگر هورویتز ـ تامپسون و برآوردگر هنسن ـ هورویتز مورد بحث قرار گرفتند و می‌توانند در ترکیب با نمونه‌گیری با احتمال نابرابر به کار روند. بحث خود را در مورد نمونه‌گیری با احتمال متناسب با اندازه بر وضعیتی متمرکز کردیم که در آن خوشه‌ها به وسیلهٔ نمونه‌گیری تصادفی ساده با جایگذاری انتخاب می‌شوند و احتمال انتخاب شدن متناسب با معیاری از اندازهٔ خوشه است. در هر نوبت از انتخاب خوشه، یک نمونهٔ تصادفی ساده از واحدهای شمارش بدون جایگذاری انتخاب می‌شود و برآورد کردن با استفاده از برآوردگر هنسن ـ هورویتز انجام می‌گیرد. با استفاده از مثالهای تشریحی نشان دادیم که نمونه‌ها چگونه با این شیوه انتخاب می‌شوند،

پارامترها چگونه برآورد می‌شوند، و خطاهای معیار چگونه به دست می‌آیند. فرمولهای مربوط به تعیین اندازهٔ نمونه را برای این طرح شرح و بسط دادیم. بالاخره، در مورد کاربرد نمونه‌گیری با احتمال متناسب با اندازه در آمارگیریهای تلفنی شماره‌گیری با ارقام تصادفی نیز بحث کردیم.

تمرین

بار دیگر جامعهٔ متشکل از ۲۵ مرکز بهداشت روانی محلی را از تمرینهای ۱۲.۱۰ تا ۱۵.۱۰ در نظر می‌گیریم (که جدول مربوط به آن در زیر تکرار می‌شود).

مرکز بهداشت	تعداد بیماران	مرکز بهداشت	تعداد بیماران
۱	٤۹۱	١٤	٦۷۲
۲	۸٦٦	۱۵	٤۷۵
۳	۱۸۸	١٦	٤۳۹
٤	۹۹٤	۱۷	۳۹۲
۵	۲۰۹	۱۸	۵۸٤
٦	۹٦۱	۱۹	۸۸۲
۷	۸۳٤	۲۰	٤۲٤
۸	۹۸۲۰	۲۱	۷۷۵
۹	۳٤۸	۲۲	۲٦۲
۱۰	۲٤٦	۲۳	۹٦۸
۱۱	۳۹۹	۲٤	۵۸٦
۱۲	۱۷۵	۲۵	۸۰۹
۱۳	۱٦٦		

۱.۱۱ فرض کنید می‌خواهید برای برآورد کردن کل تعداد بیمارانی که در حال حاضر پروزاک (Prozac) دریافت می‌کنند یک نمونهٔ دو مرحله‌ای با احتمال متناسب با اندازه و با جایگذاری از مراکز بهداشتی بگیرید. باز هم فرض کنید که معیار متغیر اندازه برای نمونه‌گیری با احتمال متناسب با اندازه، تعداد بیماران است و اعداد تصادفی انتخاب شده ۱۱۰۵۲ و ۱۲٦۱٤ هستند.

الف. دو مرکز بهداشتی نمونه‌گیری شده کدامها هستند؟

ب. فرض کنید یک نمونهٔ تصادفی ساده متشکل از ۲۰ بیمار، در هر نوبت، از مرکز بهداشتی انتخاب شده در آن نوبت، گرفته می‌شود. باز هم فرض کنید که نمونهٔ متناظر با عدد تصادفی ۱۱۰۵۲ شامل ۱۲ بیمار است که پروزاک دریافت می‌کنند و نمونهٔ متناظر با عدد تصادفی ۱۲٦۱٤ شامل ٦ بیمار تحت درمان با پروزاک است. از برآوردگر

هنسن ـ هورویتز برای برآورد کردن کل تعداد افراد تحت درمان با پروزاک استفاده کنید. برآورد خطای معیار این برآوردگر چقدر است؟

۲.۱۱ مطلوب است برآورد تعداد کره‌ایهایی که در یک شهر معین زندگی می‌کنند. این شهر دارای شش مرکز تلفن است و بررسی آخرین کتاب راهنمای تلفن، فراوانیهای زیر را برای دو نام فامیلی کره‌ای که از همه متداولترند ـ یعنی کیم و پارک ـ نشان داده است.

مرکز تلفن	جمعیت ناحیهٔ جغرافیایی متناظر با مرکز تلفن	تعداد کیم‌ها و پارک‌ها در راهنمای تلفن
۸۳۲	۵۲۳۱	۲۱
۸۵۶	۳۰۱۲	۸
۹۳۵	۲۱۲۳	۷
۹۳۶	۱۲۵۶	۳۵
۹۳۷	۲۵۶۹	۱۷
۹۸۳	۸۳۲۱	۱۰

تصور می‌شود نام فامیلی تقریباً ۶۰ درصد از همهٔ کره‌ایهای ساکن در آن منطقه «کیم» یا «پارک» است و یک خانوار متوسط کره‌ای از چهار نفر تشکیل شده است.

الف. اگر قرار می‌شد از مرکز تلفن به عنوان خوشه استفاده کنید از چه روشی برای نمونه‌گیری خوشه‌ها استفاده می‌کردید؟

ب. از این روش برای انتخاب نمونه‌ای از سه مرکز تلفن استفاده کنید. جزئیات چگونگی انتخاب این نمونه را نشان دهید.

۳.۱۱ از جامعهٔ بخشهایی که در تمرین ۱۶.۱۰، نشان داده شد، یک نمونه با احتمال متناسب با اندازه و با جایگذاری، متشکل از ۵ بخش گرفته شده است و برای هر انتخاب بخش در مرحلهٔ اول، نمونه‌ای از یک بیمارستان در مرحلهٔ دوم گرفته شده است. نمونه‌گیری به شرح زیر اجرا می‌شود:

الف. شیوهٔ انتخاب نمونهٔ بخشها با احتمال متناسب با اندازه همان است که در بخش ۳.۳.۱۱ توصیف شده است.

ب. معیار متغیر اندازه برای نمونه‌گیری از یک بخش، کل تعداد بیمارستانهای موجود در آن بخش است.

پ. اعداد تصادفی که برای انتخاب نمونه انتخاب شده‌اند به شرح زیرند:

دنبالهٔ انتخاب	عدد تصادفی برای انتخاب بخشها در مرحلهٔ اول	عدد تصادفی برای انتخاب بیمارستان در مرحلهٔ دوم
۱	۳۸	۱
۲	۲٤	۱
۳	۲٤	۱
٤	٤۲	۳
٥	۰۸	۱

با استفاده از اطلاعات بالا، بخشها و بیمارستانهای نمونه را تعیین کند.

٤.۱۱ از نمونه‌ای که در تمرین قبل انتخاب شد، کل تعداد پذیرشهای بیمارستانی را در ۳۷ ناحیهٔ بخش طی سال ۱۹۸۹ تعیین کنید. همچنین، خطای معیار این برآوردگر مجموع را تعیین کنید.

٥.۱۱ (برای کسانی که به نرم‌افزارهایی از قبیل STATA یا SUDAAN دسترسی دارند.) کل تعداد پذیرشها را به ازای هر تخت از روی نمونهٔ انتخاب شده در تمرین ۳.۱۱ برآورد کنید. خطای معیار این برآوردگر را نیز تعیین کنید.

کتابشناسی

The text by Brewer referenced below gives a very comprehensive treatment of sampling designs and estimators based on unequal probability sampling.

1. Brewer, K. R. W., and Hanif, M., *Sampling with Unequal Probabilities,* Springer - Verlag, New York, 1983.

The following recent texts give excellent discussions of the Horvitz-Thompson and Hansen-Hurwitz estimators and of the various types of PPS Sampling

2. Hedayat, A. S., and Sinha, B. K., *Design and Inference in Finite Population Sampling,* Wiley, New York, 1991.
3. Lehtonen, R., and Pahkinen, E. J., *Practical Methods for Design and Analysis of Complex Surveys,* Rev. Ed., Wiley, Chichester, U.K., 1997.
4. Thompson, S. K., *Sampling,* Wiley, New York, 1992.

The following expository review in the Enclyclopedia of Statistical Sciences *gives an excellent overview of PPS sampling and contains a long list of useful references on the topic.*

5. Skinner, C. J., Probability proportional to size (PPS) sampling. In *The Encyclopedia of Statistical Sciences,* Johnson, N., and Kotz, S., Eds., Wiley, New York, 1983.

The following entry in the Encyclopedia of Biostatistics *provides a very detailed discussion of practical aspects of PPS sampling including its strengths and weaknesses.*

6. Czaja, R., Sampling with probability proportionate to size. In *Encyclopedia of Biostatistics,* Armitage, P. A., and Colton, T. D., Eds., Wiley, Chichester, U.K., 1998.

The two original papers on the Hansen-Hurwitz and Horvitz-Thompson estimators are referenced below.

7. Hansen, M. H., and Hurwitz, W. N., On the theory of sampling from finite population. *Annals of Mathematical Statistics,* 14: 333-362, 1943.
8. Horvitz, D. G., and Thompson, D. J., A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47: 663-685, 1952.

The following two articles are the seminal papers on the Mitofsky-Waksberg method of telephone sampling.

9. Mitofsky, W., *Sampling of Telephone Households* (Unpublished CBS Memorandum), 1970.
10. Waksberg, J., Sampling methods for random digit dialing. *Journal of the American Statistical Association,* 73: *40-46,* 1978.