۳

نمونه‌گیری با احتمال متغیر

۰.۳ مقدمه

آنچه در فصل دوم تحت عنوان نمونه‌گیری تصادفی ساده شرح دادیم مبتنی بر این بود که احتمال انتخاب همهٔ نمونه‌های ممکن به‌حجم n از جامعهٔ به‌حجم N، و در نتیجه احتمال انتخاب همهٔ واحدهای جامعه برای شرکت در یک نمونه یکسان باشد. در عمل، این یکسان بودن احتمال انتخاب واحدها، گاهی به‌حصول برآوردهایی منجر می‌شود که با وجود نااریب بودن، گاهی به‌دلیل بزرگ بودن واریانس جامعه، از واقعیت دورند. برای روشن شدن مطلب به‌ذکر مثالی می‌پردازیم. فرض کنید بخواهیم جمعیت شهرنشین استانی را برآورد کنیم. اگر استان شامل ده شهر باشد که جمعیت آنها برحسب ۱۰۰۰ نفر به‌ترتیب ۱۰۰، ۱۱۰، ۱۵۰، ۱۳۰، ۲۱۰، ۷۰، ۵۰۰، ۵۴۰، ۷۰۰، و ۸۰ است، کل جمعیت شهرنشین استان برابر با ۲۵۹۰ است. اگر بخواهند با نمونه‌ای تصادفی به‌حجم ۴، جمعیت کل را برآورد کنند ممکن است نمونهٔ حاصل ۱۰۰، ۸۰، ۱۳۰، ۷۰ باشد که چون میانگین نمونه برابر با ۹۵ است جمعیت کل برابر با ۹۵۰ برآورد می‌شود. می‌بینید که این برآورد، با اینکه نااریب است به‌دلیل بزرگی واریانس از واقعیت، یعنی از ۲۵۹۰ بسیار دور است. علت این اختلاف بارز این است که در روش نمونه‌گیری تصادفی، شانس انتخاب شهری با جمعیت ۵۰۰ برای شرکت در نمونه، با شانس انتخاب شهری با جمعیت مثلاً ۷۰ برای شرکت در نمونه یکی است که ممکن است به انتخاب نمونه‌ای مثل نمونهٔ بالا منجر شود که برآوردی غیرمنطقی را به‌دست دهد. عقل سلیم می‌گوید که در انتخاب واحدهای نمونه، به شهری که جمعیت بیشتری

دارد باید شانس انتخاب بیشتری داده شود. دادن احتمال به واحدها برای شرکت در نمونه، درواقع دادن وزنهایی است که عمدتاً واریانس جامعه را تقلیل می‌دهد. همین ایده، نمونه‌گیری جدیدی را به‌نام نمونه‌گیری با احتمال متغیر القا می‌کند.

۱.۳ تعریف نمونه‌گیری با احتمال متغیر

یک جامعه با N واحد را در نظر می‌گیریم و فرض می‌کنیم Y_i معرف مقدار صفت تحت مطالعه برای iامین واحد جامعه $(i = ۱, ۲, \ldots, N)$ باشد. به‌علاوه فرض می‌کنیم که $p_i(i = ۱, ۲, \ldots, N)$ معرف احتمال انتخاب iامین واحد جامعه برای عضویت در نمونه‌ای مفروض باشد. بدیهی است $\sum_{i=۱}^{N} p_i = ۱$. در این نمونه‌گیری احتمال انتخاب واحدهای جامعه برای شرکت در نمونه ثابت نیست و معمولاً از واحدی به واحد دیگر تغییر می‌کند. اگر در حالت خاص، p_iها متناسب با اندازهٔ Y_i ها باشند نمونه‌گیری را تصادفی با احتمال متناسب با اندازه می‌نامند و آن را نمونه‌گیری PPS* می‌گویند.

نمونه‌گیری تصادفی با احتمال متغیر را می‌توان به دو روش با جایگذاری و بدون جایگذاری انجام داد. در روش با جایگذاری احتمال انتخاب واحد Y_i برای شرکت در نمونه به‌همان مقداری است که از قبل به‌صورت p_i مشخص شده است، زیرا در هر انتخاب با کل جامعه سروکار داریم، اما در نمونه‌گیری بدون جایگذاری با وضعی کمی پیچیده روبه‌رو هستیم. بدیهی است که نحوهٔ انتخاب واحدها با روش انتخاب نمونهٔ تصادفی ساده متفاوت است. قبل از ورود به بحث برآورد پارامترهای جامعه با استفاده از نمونهٔ تصادفی با احتمال متغیر، شیوه‌هایی را برای انتخاب نمونهٔ با احتمال متغیر که خاص روش با جایگذاری است شرح می‌دهیم.

۲.۳ انتخاب نمونهٔ با احتمال متغیر و با جایگذاری به‌شیوهٔ مجموع تراکمی

در نمونه‌گیری تصادفی ساده برای انتخاب واحدهای نمونه، از جدول اعداد تصادفی استفاده می‌کنیم، زیرا احتمال انتخاب همهٔ واحدها برای شرکت در نمونه یکسان است ولی در نمونه‌گیری با احتمال متغیر، چون احتمال تخصیص‌یافته به هر واحد در حالت کلی با احتمالهای منسوب به واحدهای دیگر برای عضویت در نمونه فرق می‌کند استفاده مستقیم از جدول اعداد تصادفی میسر نیست. برای انتخاب نمونه اعداد صحیح $X_۱$، $X_۲$، $\ldots$ ، X_N را که متناسب با احتمالهای انتخاب واحدهای $Y_۱$، $Y_۲$، $\ldots$ ، Y_N جامعه هستند در نظر می‌گیریم. این اعداد صحیح وقتی p_iها اعدادی گویا باشند همیشه وجود دارند. مثلاً اگر $N = ۵$ و احتمالهای متناظر با ۵ واحد جامعه به‌ترتیب $\frac{۱}{۶}$، $\frac{۱}{۳}$، $\frac{۱}{۵}$، $\frac{۱}{۴}$، و $\frac{۱}{۲۰}$ باشند، وقتی احتمالها را هم مخرج کنیم هم‌ارز با $\frac{۱۰}{۶۰}$، $\frac{۲۰}{۶۰}$، $\frac{۱۲}{۶۰}$، $\frac{۱۵}{۶۰}$ و $\frac{۳}{۶۰}$

---

* PPS نمادی برای اصطلاح Probability Proportional to Size

خواهند بود. لذا این احتمالها با اعداد صحیح ۱۰، ۲۰، ۱۲، ۱۵، و ۳ متناسب‌اند و برای این مثال

$$X_۱ = ۱۰,\ X_۲ = ۲۰,\ X_۳ = ۱۲,\ X_۴ = ۱۵,\ X_۵ = ۳$$

اعداد صحیح متناظر با $Y_۱$، $Y_۲$، ... ، $Y_۵$ هستند. ما حالتی را که احتمالها گنگ‌اند نادیده می‌گیریم. اصطلاحاً دنبالهٔ $X_۱$، $X_۲$، ... ، X_N را دنبالهٔ صفتهای کمکی متناظر با $Y_۱$، $Y_۲$، ... ، Y_N می‌نامیم، لذا

$$\begin{array}{llllll} \text{صفت اصلی}: & Y_۱ & Y_۲ \cdots & Y_i & \cdots & Y_N \\ \text{صفت کمکی}: & X_۱ & X_۲ \cdots & X_i & \cdots & X_N \end{array}$$

حال برای انتخاب نمونهٔ با احتمال متغیر، به اولین واحد جامعهٔ صفت اصلی، عدد $T_۱ = X_۱$ و به دومین واحد جامعه عدد $T_۲ = X_۱ + X_۲$، و در حالت کلی به iامین واحد جامعه $(i = ۱, \ldots, N)$ عدد $T_i = X_۱ + X_۲ + \cdots + X_i$ را نسبت می‌دهیم. بدیهی است عدد منسوب به Y_N برابر $T_N = \sum_{i=۱}^{N} X_i$ است. در این صورت جدول زیر را به‌دست می‌آوریم.

جدول ۱.۳ نمایش احتمالهای انتخاب متناسب با اندازهٔ واحدها

صفت اصلی	صفت کمکی	T_i	p_i
$Y_۱$	$X_۱$	$X_۱$	$X_۱/T_N$
$Y_۲$	$X_۲$	$X_۱ + X_۲$	$X_۲/T_N$
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
Y_i	X_i	$X_۱ + X_۲ + \cdots + X_i$	X_i/T_N
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
Y_N	X_N	$X_۱ + X_۲ + \cdots \quad + X_N$	X_N/T_N

با رجوع به جدول اعداد تصادفی، عددی از ۱ تا T_N انتخاب می‌کنیم. این عدد را R می‌نامیم. عدد R را با T_iها در ستون سوم، مقایسه می‌کنیم. عدد R یا با یکی از T_iها برابر است و یا بین دو T_iی متوالی، یعنی مثلاً بین $T_{i-۱}$ و T_i، قرار دارد. پس $T_{i-۱} < R \leq T_i$. در این صورت واحد iام با مقدار Y_i را به‌عنوان واحد نمونه انتخاب می‌کنیم. باید ثابت کنیم که انتخاب واحد iام متناسب با X_i صورت گرفته است. تعداد حالتهای مساعد برای اینکه R در رابطهٔ $T_{i-۱} < R \leq T_i$ قرار گیرد برابر است با $T_i - T_{i-۱}$، یعنی

$$T_i - T_{i-۱} = (X_۱ + X_۲ + \cdots + X_i) - (X_۱ + X_۲ + \cdots + X_{i-۱}) = X_i$$

و چون R را از بین اعداد ۱ تا T_N انتخاب کرده‌ایم، لذا تعداد حالتهای ممکن انتخاب R برابر با T_N است. پس

$$P(X_{i-۱} < R \leq X_i) = P(\text{انتخاب } Y_i \text{ برای شرکت در نمونه})$$

$$= \frac{X_i}{T_N} = \frac{X_i}{X_۱ + \cdots + X_N} = p_i$$

یعنی واحد iام که با شیوهٔ بالا انتخاب می‌کنیم و واحدی از نمونه است با احتمال p_i برگزیده می‌شود. پس از انتخاب واحد iام، آن را به‌جامعه برمی‌گردانیم و این فرایند را n بار تکرار می‌کنیم تا از N واحد جامعه، نمونه‌ای به‌حجم n انتخاب شود. بدیهی است که امکان دارد برخی از واحدهای نمونهٔ منتخب، تکراری باشند.

مثال ۱.۳ صفتی که تحت بررسی است میزان محصول گندم ۱۰ روستاست. می‌خواهیم به‌کمک نمونه‌ای به حجم ۴، میزان محصول ۱۰ روستا را برآورد کنیم. مساحت زمین زیرکشت هر یک از ده روستا را می‌دانیم. در این مثال، چهار واحد نمونه را به روش تصادفی با احتمال متغیر، که احتمالها متناسب با صفت کمکی مساحتهای زیر کشت‌اند، مشخص می‌کنیم.

میزان محصول ۱۰ روستا را به‌ترتیب Y_1، Y_2، . . . ، Y_{10} می‌نامیم. این مقادیر در دست نیستند. مساحت زمین زیرکشت متناظر با Y_iها را می‌دانیم و آنها را با X_1، X_2، . . . ، X_{10} نشان می‌دهیم. مقادیر X_i در جدول زیر آمده‌اند. با توجه به این جدول، به‌شیوهٔ مجموع تراکمی ۴ واحد نمونه را تعیین می‌کنیم. در ستون سوم جدول، مقادیر T_i را مطابق شرحی که در بیان شیوهٔ انتخاب نمونه ذکر شد، مشخص کرده‌ایم. حال به جدول اعداد تصادفی رجوع می‌کنیم و از اعداد ۱ تا ۱۸۵۰ چهار عدد انتخاب می‌کنیم. فرض کنید اعداد ۷۲، ۵۵۱، ۱۷۳۰، و ۹۸۰ به‌دست آیند. به‌ترتیب داریم

$$۷۲ < T_1$$
$$T_3 < ۵۵۱ < T_4$$
$$۱۷۳۰ = T_8$$
$$T_4 < ۹۸۰ < T_5$$

شمارهٔ واحد	محصول	مساحت زمین زیرکشت	T_i
۱	Y_1	۱۰۰	۱۰۰
۲	Y_2	۹۰	۱۹۰
۳	Y_3	۳۰۰	۴۹۰
۴	Y_4	۱۱۰	۶۰۰
۵	Y_5	۴۰۰	۱۰۰۰
۶	Y_6	۱۵۰	۱۱۵۰
۷	Y_7	۵۰۰	۱۶۵۰
۸	Y_8	۸۰	۱۷۳۰
۹	Y_9	۷۰	۱۸۰۰
۱۰	Y_{10}	۵۰	۱۸۵۰

لذا با توجه به استدلالی که در شیوهٔ انتخاب واحدهای نمونه ارائه شد. نمونهٔ تصادفی مورد نظر

عبارت است از واحدهای ۱، ۴، ۸، و ۵ جامعه با مقادیر

$$Y_۱,\ Y_۴,\ Y_۸,\ Y_۵$$

پس باید به روستاهای شمارهٔ ۱، ۴، ۸، و ۵ مراجعه و میزان محصول گندم آنها را ثبت کنیم. ممکن بود چهار عدد تصادفی منتخب، مثلاً اعداد ۱۵۰، ۱۷۱، ۱۵۰۱، و ۱۸۰۲ باشند. در این صورت

$$T_۱ < ۱۵۰ < T_۲\ ,\quad T_۱ < ۱۷۱ < T_۲$$
$$T_۶ < ۱۵۰۱ < T_۷,\quad T_۹ < ۱۸۰۲ < T_{۱۰}$$

و لذا، نمونهٔ منتخب به‌صورت زیر است: واحدهای ۲، ۲، ۷، و ۱۰ با مقادیر

$$Y_۲,\ Y_۲,\ Y_۷,\ Y_{۱۰}$$

می‌بینید که در این حالت واحد $Y_۲$ یا روستای شمارهٔ ۲، دوبار در نمونه شرکت می‌کند. ▲

شیوهٔ مجموع تراکمی، مستلزم تشکیل ستون T_iها برای همهٔ واحدهای مقادیر X_i هاست که مسلماً وقتی جامعه بزرگ است انجام آن حتی با نرم‌افزارهای کامپیوتری هم پرهزینه و وقتگیر است. روش زیر را که لاهیری برای انتخاب نمونه با احتمال متغیر پیشنهاد کرده است و روشی با هزینهٔ کمتر است، عرضه می‌کنیم.

۳.۳ انتخاب نمونهٔ با احتمال متغیر به‌شیوهٔ لاهیری

به ستون اندازه‌های X_i مراجعه می‌کنیم و بزرگترین مقدار X_i را مشخص و آن را M می‌نامیم. اگر N حجم جامعه باشد، یک زوج عدد تصادفی $(i,\ j)$ را با شرایط $۱ \le i \le N$ و $۱ \le j \le M$ انتخاب می‌کنیم. وقتی i انتخاب شد X_i را در نظر می‌گیریم، در این صورت اگر داشته باشیم $j \le X_i$، واحد iام را به‌عنوان واحدی از نمونهٔ مطلوب اختیار می‌کنیم. اگر داشته باشیم $j > X_i$ زوج انتخاب‌شدهٔ $(i,\ j)$ را نادیده می‌گیریم و زوج تصادفی دیگری را انتخاب و فرایند قبلی را تکرار می‌کنیم. این فرایند را تا انتخاب n واحد نمونه ادامه می‌دهیم. قبل از اثبات اینکه استفاده از این شیوه، به هر واحد نمونه، احتمال انتخابی متناسب با X_i را تخصیص می‌دهد، به‌ذکر مثالی می‌پردازیم.

مثال ۲.۳ به مثال ۱.۳ برمی‌گردیم و این بار به‌روش لاهیری نمونه‌ای به‌حجم ۴ با احتمال متناسب با X_iها به‌دست می‌آوریم. چون بزرگترین X_i برابر با ۵۰۰ است، پس $M = ۵۰۰$ و از طرفی $N = ۱۰$. زوج $(i,\ j)$ از اعداد تصادفی را به‌قسمی انتخاب می‌کنیم که $۱ \le i \le ۱۰$ و $۱ \le j \le ۵۰۰$، یعنی عدد اول را با استفاده از یک ستون جدول اعداد تصادفی و عدد دوم را با استفاده از سه ستون انتخاب می‌کنیم. فرض می‌کنیم زوج $(۵, ۲۰۱)$ به‌دست آید. به‌ازای $i = ۵$

با مراجعه به جدول داده‌ها، مشاهده می‌شود که $X_5 = 400$ و چون $j = 201 < X_5$، لذا واحد $i = 5$ با مقدار Y_5 اولین واحد نمونه است. مجدداً زوجی تصادفی با شرایط مورد نظر انتخاب می‌کنیم. فرض می‌کنیم زوج (۶,۴۰۳) نتیجه شود، چون

$$j = 403 > X_6 = 150$$

این زوج تصادفی را نادیده می‌گیریم و زوج جدیدی از اعداد تصادفی را برمی‌گزینیم. ممکن است به دفعاتی متعدد موفق به‌انتخاب واحد جدیدی برای نمونه نشویم. اگر به‌عنوان مثال زوجهای تصادفی بعدی به‌ترتیب (۲, ۸۳)، (۸, ۳۰۲)، (۵, ۹۷)، و (۳, ۲۵) باشند به‌سهولت می‌توانیم تحقیق کنیم که با زوج اول، Y_2 به‌عنوان دومین واحد نمونه انتخاب می‌شود. زوج بعدی یعنی (۸, ۳۰۲) را باید نادیده گرفت و با زوج (۵, ۹۷)، واحد ۵ را به‌عنوان سومین واحد نمونه برمی‌گزینیم. از زوج آخر، یعنی (۳, ۲۵)، واحد سه، چهارمین واحد نمونه خواهد بود. لذا مقادیر واحدهای نمونهٔ مورد نظر به‌روش لاهیری به‌صورت

$$Y_5,\ Y_2,\ Y_5,\ Y_3$$

است، که در آن واحد پنجم، تکراری است. ▲

با توجه به مثال بالا می‌توان حالتی را تصور کرد که انتخابهای متوالی زوجهای $(i,\ j)$ هرگز به انتخاب واحدی منجر نشود. در ادامهٔ مطلب خواهیم دید که احتمال وقوع چنین حالتی صفر است. اینک که با ذکر مثال، شیوهٔ لاهیری برای انتخاب واحدهای نمونه مشخص شد ثابت می‌کنیم که با این شیوه، احتمال انتخاب واحد iام متناسب با X_i است.

برهان شیوهٔ لاهیری. قبلاً متذکر می‌شویم که هر زوج $(i,\ j)$ که به انتخاب Y_i منجر نشود به زوج ناموثر موسوم است و اگر زوجی به انتخاب Y_i منجر شود آن را زوج مؤثر می‌نامند. حال زوج $(Y_i,\ X_i)$ را در نظر می‌گیریم و احتمال انتخاب Y_i را در بار اولِ انتخابِ زوج $(i,\ j)$ با $P_1(Y_i)$ نشان می‌دهیم. به‌همین ترتیب احتمال انتخاب Y_i را در بار دوم انتخاب زوج $(i,\ j)$ با $P_2(Y_i)$ نمایش می‌دهیم و الی‌آخر. احتمال $P_1(Y_i)$ را می‌توان به‌صورت زیر نوشت

$P_1(Y_i) = P(j \le X_i$ ،$1 \le j \le M$ ابتدا i از اعداد ۱ تا N انتخاب شود و سپس برای)

چون انتخاب i و j مستقل از هم هستند، پس داریم

$$P_1(Y_i) = P(\text{انتخاب } i \text{ از اعداد ۱ تا } N) \cdot P(j \le X_i | 1 \le j \le M) \quad (۱.۳)$$

اما

$$P(\text{انتخاب } i \text{ از اعداد ۱ تا } N) = \frac{1}{N} \quad (۲.۳)$$

از طرفی تعداد حالتهای ممکن برای انتخاب j برابر با M است و حالتهای مساعد برای وقوع پیشامد $j \le X_i$ برابر با X_i است، پس

$$\begin{cases} P(j \le X_i) = \frac{X_i}{M} \\ P(j > X_i) = ۱ - \frac{X_i}{M} = \frac{M - X_i}{M} \end{cases} \qquad (۳.۳)$$

اگر رابطه‌های (۲.۳) و (۳.۳) را در (۱.۳) منظور کنیم، نتیجه می‌شود

$$P_۱(Y_i) = \frac{۱}{N} \cdot \frac{X_i}{M} \qquad (۴.۳)$$

درواقع، این احتمال، احتمال مؤثر بودن انتخاب زوج $(i,\ j)$ است.

برای اینکه زوج $(i,\ j)$ در اولین انتخاب مؤثر نباشد باید یکی از پیشامدهای زیر رخ دهد

$i = ۱$، آن‌گاه $j > X_۱$

$i = ۲$، آن‌گاه $j > X_۲$

...

...

$i = N$ آن‌گاه $j > X_N$

لذا،

$$P((i,\ j) \text{ مؤثر نبودن زوج }) = P(i=۱, j>X_۱) + P(i=۲, j>X_۲) + \cdots + P(i=N,\ j>X_N) \qquad (۵.۳)$$

اما

$$P(i = ۱,\ j > X_۱) = P(i = ۱) \cdot P(j > X_۱) = \frac{۱}{N} \cdot \frac{M - X_۱}{M}$$

$$P(i = ۲,\ j > X_۲) = P(i = ۲) \cdot P(j > X_۲) = \frac{۱}{N} \cdot \frac{M - X_۲}{M}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$P(i = N,\ j > X_N) = P(i = N) \cdot P(j > X_N) = \frac{۱}{N} \cdot \frac{M - X_N}{M}$$

اگر این رابطه‌ها را در (۵.۳) قرار دهیم، نتیجه می‌شود که

$$P((i,\ j)\ \text{مؤثر نبودن زوج}) = \sum_{i=1}^{N} \frac{1}{N} \cdot \frac{M - X_i}{M} = \frac{1}{NM} \sum_{i=1}^{N} (M - X_i)$$
$$= \frac{1}{NM}(MN - N\overline{X}) = 1 - \frac{\overline{X}_N}{M} \qquad (۶.۳)$$

که در آن، $\overline{X}_N$ میانگین X_iهاست. حال $P_2(Y_i)$ را محاسبه می‌کنیم

$$P_2(Y_i) = P((i,\ j)\ \text{انتخاب}\ Y_i\ \text{در دومین انتخاب})$$
$$= P((i,\ j)\ \text{مؤثر بودن انتخاب در بار دوم} \cap (i,\ j)\ \text{مؤثر نبودن انتخاب در بار اول})$$
$$= P(\text{مؤثر بودن انتخاب در بار دوم}) \cdot P(\text{مؤثر نبودن انتخاب در بار اول}) =$$

که با توجه به رابطه‌های (۴.۳) و (۶.۳)، داریم

$$P_2(Y_i) = \left(1 - \frac{\overline{X}_N}{M}\right) \frac{1}{N}(X_i/M) \qquad (۷.۳)$$

به‌همین ترتیب

$$P_3(Y_i) = P\left((i,j)\ \begin{matrix}\text{مؤثر نبودن انتخاب}\\ \text{در بار اول}\end{matrix} \cap (i,j)\ \begin{matrix}\text{مؤثر نبودن انتخاب}\\ \text{در بار دوم}\end{matrix} \cap (i,j)\ \begin{matrix}\text{مؤثر بودن انتخاب}\\ \text{در بار سوم}\end{matrix}\right)$$
$$= \left(1 - \frac{\overline{X}_N}{M}\right)\left(1 - \frac{\overline{X}_N}{M}\right) \frac{1}{N}(X_i/M)$$

ممکن است متوالیاً، به‌دفعات بسیار زیاد، انتخاب $(i,\ j)$ مؤثر نبوده و سپس انتخاب بعدی مؤثر باشد. حتی می‌توان تصور کرد که تعداد مؤثر نبودنها به بینهایت هم بگراید. حال اگر هدف تعیین احتمال انتخاب واحد i با مقدار Y_i به‌عنوان واحدی از نمونه باشد باید در یکی از پیشامدهای متوالی که احتمال متناظر با آنها را در بالا مشخص کردیم Y_i انتخاب شود. پس

$$P(\text{انتخاب}\ Y_i\ \text{در نمونه}) = \frac{1}{N} \cdot \frac{X_i}{M} + \left(1 - \frac{\overline{X}_N}{M}\right) \frac{1}{N}\ \frac{X_i}{M}$$
$$+ \left(1 - \frac{\overline{X}_N}{M}\right)^2 \frac{1}{N}\ \frac{X_i}{M} + \cdots$$

طرف دوم این برابری، یک سری هندسی است که قدر نسبت آن $(\text{۱} - \frac{\overline{X}_N}{M})$ است. چون M بزرگترین X_i است، $\frac{\overline{X}_N}{M}$ کوچکتر از واحد است و لذا $(\text{۱} - \frac{\overline{X}_N}{M} < \text{۱})$ و مثبت است مگر اینکه همه X_iها با هم برابر باشند که نمونه‌گیری به نمونه‌گیری تصادفی ساده تبدیل می‌شود. با این توضیح، سری طرف دوم رابطهٔ بالا یک سری همگراست و وقتی تعداد انتخابهای ناموٌثر متوالی به بینهایت بگراید، داریم

$$\lim P(\text{انتخاب } Y_i \text{ در نمونه}) = \frac{\frac{\text{۱}}{N}\frac{X_i}{M}}{\text{۱} - \left(\text{۱} - \frac{\overline{X}_N}{M}\right)} = \frac{X_i}{N\overline{X}_N} = \frac{X_i}{T_N}$$

یعنی احتمال انتخاب Y_i به‌روش لاهیری برای شرکت در نمونه متناسب با X_i است. در این روش برای انتخاب واحدها نیازی به داشتن مجموع تراکمی X_iها نداریم و انتخابها با محاسباتی کمتر صورت می‌گیرند.

تبصره. از رابطهٔ (۶.۳) نتیجه می‌شود که انتخابهای متوالی $(i,\ j)$ نمی‌توانند همیشه ناموٌثر باشند زیرا احتمال ناموٌثر بودن $(i,\ j)$ برابر با $\text{۱} - \frac{\overline{X}_N}{M}$ است که همیشه از ۱ کمتر است و لذا پیشامدی حتمی نخوهد بود.

۴.۳ روش خرد کردن برای اصلاح شیوهٔ لاهیری

دیدیم که احتمال موٌثر نبودن انتخاب زوج $(i,\ j)$ برابر $\left(\text{۱} - \frac{\overline{X}_N}{M}\right)$ است. هرچه این مقدار کوچکتر باشد یعنی هرچه کسر $\frac{\overline{X}_N}{M}$ به یک نزدیکتر باشد احتمال بالا کوچکتر بوده و در نتیجه تعداد انتخابهای ناموٌثر که موجب صرف هزینه و وقت می‌شود کمتر است. مقدار $\frac{\overline{X}_N}{M}$ وقتی به یک نزدیک است که M به میانگین X_iها نزدیک باشد. اگر مقادیر X_iها پراکندگی زیاد نداشته باشند، یعنی M به $\overline{X}_N$ نزدیک باشد این منظور حاصل می‌شود. وقتی M، یعنی بزرگترین X_i با سایر X_iها و در نتیجه با $\overline{X}_N$ فاصلهٔ زیاد داشته باشد می‌توان با روش خرد کردن که ذیلاً شرح می‌دهیم این فاصله را کم و تعداد انتخابهای ناموٌثر را نیز کم کرد. این روش همان‌طور که توضیح دادیم برای کوچکتر کردن $\text{۱} - \frac{\overline{X}_N}{M}$ ابداع شده است. در این روش آن واحدی را که مقدار X آن متناظر با M است به دو یا چند واحد تفکیک می‌کنیم. با این عمل تعداد واحدهای جامعه یک یا چند واحد اضافه می‌شود ولی مقدار M، یعنی ماکسیمم X_iها هم کوچک می‌شود. البته با این عمل $\overline{X}_N$ هم کوچک می‌شود ولی تفکیک را به‌صورتی انجام می‌دهیم که $\frac{\overline{X}_N}{M}$ نهایتاً بزرگ و احتمال $\text{۱} - \frac{\overline{X}_N}{M}$ کوچک و در نتیجه تعداد انتخابهای ناموٌثر کم شود. برای توضیح مطلب به مثال زیر توجه کنید.

مثال ۳.۳ در جدول زیر میزان محصول گندم ۸ ده را با مساحت زمین زیرکشت آنها درج کرده‌ایم. هدف، انتخاب نمونه‌ای از این جامعه با احتمال متناسب با مساحت زیرکشت است.

شمارهٔ واحد	میزان محصول Y_i	مساحت زمین زیرکشت X_i
۱	$Y_۱$	۴۵
۲	$Y_۲$	۵۰
۳	$Y_۳$	۷۵
۴	$Y_۴$	۳۰
۵	$Y_۵$	۱۲۰
۶	$Y_۶$	۶۵
۷	$Y_۷$	۳۵
۸	$Y_۸$	۶۰

با روش لاهیری احتمال نامؤثر بودن انتخاب زوج $(i,\ j)$، با توجه به مقادیر $N = ۸$ و $\overline{X}_N = ۶۰$، و $M = ۱۲۰$، برابر است با $۱ - \frac{\overline{X}_N}{M} = ۰{,}۵$. یعنی باید انتظار داشته باشیم که نصف زوجهای $(i,\ j)$ در انتخاب هر نمونه، نامؤثر باشند که مسلماً با اتلاف وقت و هزینه همراه‌اند. حال با استفاده از روش خرد کردن، روستایی که مساحت زیرکشت آن ۱۲۰ است به دو روستا با مساحتهای زیرکشت ۷۵ و ۴۵ تقسیم می‌کنیم. در نتیجه حجم جامعه $N = ۹$ و $M = ۷۵$ و $\overline{X}_N = \frac{۱۶۰}{۳}$ خواهد شد، لذا

$$۱ - \frac{\overline{X}_N}{M} = ۱ - \frac{۱۶۰/۳}{۷۵} = \frac{۱۳}{۴۵} \simeq ۰{,}۲۹$$

ملاحظه می‌کنید که با این روش احتمال نامؤثر بودن زوج $(i,\ j)$ از ۰٫۵ به ۰٫۲۹ کاهش می‌یابد.▲

۵.۳ برآورد میانگین در نمونه‌گیری با احتمال متغیر و با جایگذاری

همان‌طور که در ابتدای فصل متذکر شدیم، حجم جامعه را N و مقدار صفت اصلی واحد iام جامعه را Y_i و احتمال انتخاب این واحد در نمونهٔ به‌حجم n و با جایگذاری را با p_i نشان می‌دهیم که $\sum_{i=۱}^{N} p_i = ۱$. فرض می‌کنیم با یکی از شیوه‌های قبل نمونه‌ای به‌حجم n انتخاب کرده‌ایم. حال می‌خواهیم برآوردکننده‌ای برای $\overline{Y}_N$، میانگین جامعه، به‌دست آوریم.

قضیهٔ ۱.۳ در نمونه‌گیری تصادفی با جایگذاری و با احتمال متغیر و به‌حجم n، آمارهٔ $\frac{۱}{n}\sum_{i=۱}^{n} \frac{Y_i}{np_i}$ برآوردکنندهٔ ناااریب میانگین جامعه است.

برهان. فرض کنیم نمونه‌ای که با احتمال متغیر و با جایگذاری انتخاب شده است دارای مقادیر و احتمالهای متناظر زیر باشد:

$$\begin{matrix} Y_۱ & Y_۲ & \cdots & Y_i & \cdots & Y_n \\ p_۱ & p_۲ & \cdots & p_i & \cdots & p_n \end{matrix}$$

بدیهی است $\sum_{i=۱}^{n} p_i < ۱$. اگر میانگین جامعه $\overline{Y}_N$ باشد می‌خواهیم از روی نمونه برآوردکننده‌ای

برای $\overline{Y}_N$ بیابیم. روی مقادیر واحدهای جامعه تبدیل متغیر

$$Z_i = \frac{Y_i}{Np_i} \qquad i = ۱,\ ۲, \cdots, N$$

را اعمال می‌کنیم تا جامعه Z_iها به‌وجود آید. میانگین این جامعه به‌صورت زیر به‌دست می‌آید

$$\overline{Z}_N = \sum_{i=۱}^{N} Z_i p_i = \sum_{i=۱}^{N} \frac{Y_i}{Np_i} \cdot p_i = \frac{۱}{N} \sum_{i=۱}^{N} Y_i = \overline{Y}_N$$

یعنی میانگین جامعهٔ Z_iها برابر با میانگین جامعهٔ Y_iهاست، پس به‌جای تعیین برآوردکننده‌ای برای $\overline{Y}_N$ باید برآوردکننده‌ای برای $\overline{Z}_N$ بیابیم. ولی با توجه به آنچه در آمار مقدماتی دیده‌اید، اگر از جامعهٔ متغیر تصادفی Z که تابع جرم احتمال آن $P(Z = Z_i) = p_i$، $i = ۱, \cdots, N$ است نمونه‌ای با جایگذاری، که استقلال واحدها را تأمین می‌کند، به‌تصادف انتخاب کنیم، و نمونه به صورت $Z_۱$، $Z_۲$، … ، Z_n باشد، آن‌گاه $\overline{Z}_n = \frac{۱}{n}\sum_{i=۱}^{n} Z_i$ برآوردکنندهٔ نااریب $\overline{Z}_N$ جامعه است. پس

$$\overline{Z}_n = \hat{\overline{Z}}_N = \hat{\overline{Y}}_N$$

اما با توجه به $Z_i = \frac{Y_i}{Np_i}$، برآوردکنندهٔ نااریب $\overline{Y}_N$، یعنی $\overline{Z}_n$ به‌صورت زیر است

$$\hat{\overline{Y}}_N = \overline{Z}_n = \frac{۱}{n} \sum_{i=۱}^{n} \frac{Y_i}{Np_i}$$

و قضیه ثابت می‌شود. این برآوردکننده منسوب به هنسن و هورویتس است. □

مثال ۴.۳ از جامعه‌ای با ۵۰ واحد، نمونه‌ای به‌حجم ۶ و با احتمال متغیر و با جایگذاری انتخاب کرده‌ایم. مقادیر واحدها و احتمالهای متناظر با آنها در جدول زیر آمده‌اند

Y_i	۱۰	۱۲	۲۰	۸	۱۴	۱۶
p_i	$\frac{۱}{۵۰}$	$\frac{۱}{۴۵}$	$\frac{۱}{۶۰}$	$\frac{۱}{۴۰}$	$\frac{۱}{۴۵}$	$\frac{۱}{۷۰}$

برآورد نااریب میانگین جامعه را بیابید.

ابتدا تبدیل متغیر $Z_i = \frac{Y_i}{Np_i}$ را برای جدول داده‌ها اجرا می‌کنیم تا Z_iهای متناظر به‌دست آیند

$$Z_۱ = \frac{Y_۱}{Np_۱} = \frac{۱۰}{۵۰\,(\frac{۱}{۵۰})} = ۱۰ \quad , \quad Z_۲ = \frac{۱۲}{۵۰\,(\frac{۱}{۴۵})} = ۱۰٫۸$$

$$Z_۳ = \frac{۲۰}{۵۰\,(\frac{۱}{۶۰})} = ۲۴ \quad , \quad Z_۴ = \frac{۸}{۵۰\,(\frac{۱}{۴۰})} = ۶٫۴$$

$$Z_۵ = \frac{۱۴}{۵۰\,(\frac{۱}{۴۵})} = ۲۵٫۲ \quad , \quad Z_۶ = \frac{۱۶}{۵۰\,(\frac{۱}{۷۰})} = ۲۲٫۴$$

پس نمونهٔ Z_iها به‌صورت زیر است

$$۱۰,\quad ۱۰٫۸,\quad ۲۴,\quad ۶٫۴,\quad ۲۵٫۲,\quad ۲۲٫۴$$

بنابراین

$$\hat{\overline{Y}}_N = \overline{Z}_n = \frac{۱}{۶}[۱۰ + ۱۰٫۸ + \cdots + ۲۲٫۴] \simeq ۱۶٫۴۶$$

که برآورد نااریب میانگین جامعهٔ به‌حجم ۵۰ی Y_iهاست. ▲

۶.۳ واریانس برآوردکنندهٔ میانگین جامعه در نمونه‌گیری با احتمال متغیر و با جایگذاری

مسلماً نمونه‌های متعددی به‌حجم n، با روشهای مذکور می‌توان از جامعهٔ مورد نظر انتخاب کرد. لذا پس از تبدیل Y_iهای نمونه به Z_iها، $\overline{Z}_n$های متعددی خواهیم داشت. هرچه $\overline{Z}_n$ها پراکندگی کمتری داشته باشند، برآمد تصادفی یکی از آنها به‌عنوان برآورد نااریب $\overline{Y}_N$ با دقتی بیشتر همراه است. برای داشتن ایده‌ای از پراکندگی $\overline{Z}_n$ها، واریانس $\overline{Z}_n$ تصادفی را محاسبه می‌کنیم. می‌دانیم

$$\mathrm{V}(\overline{Z}_n) = E(\overline{Z}_n^۲) - [E(\overline{Z}_n)]^۲$$

اما $\overline{Z}_n = \frac{۱}{n}\sum_{i=۱}^{n} Z_i$ و $E(\overline{Z}_n) = \overline{Z}_N$، پس

$$\begin{aligned}
\mathrm{V}(\overline{Z}_n) &= E\left[\left(\frac{۱}{n}\sum_{i=۱}^{n} Z_i\right)^۲\right] - \overline{Z}_N^۲ \\
&= \frac{۱}{n^۲}E\left[\left(\sum_{i=۱}^{n} Z_i\right)^۲\right] - \overline{Z}_N^۲ \\
&= \frac{۱}{n^۲}E\left[\sum_{i=۱}^{n} Z_i^۲ + \sum_{(i\neq j)=۱}^{n} Z_iZ_j\right] - \overline{Z}_N^۲ \\
&= \frac{۱}{n^۲}\left[\sum_{i=۱}^{n} E(Z_i^۲) + \sum_{(i\neq j)=۱}^{n} E(Z_iZ_j)\right] - \overline{Z}_N^۲ \qquad (۸.۳)
\end{aligned}$$

اما Z_iها متناظر با احتمالهای p_iها هستند، لذا

$$E(Z_i^۲) = \sum_{i=۱}^{N} p_i Z_i^۲ \qquad (۹.۳)$$

از طرفی Z_i مستقل از Z_j است، پس

$$E(Z_i Z_j) = E(Z_i)E(Z_j) = \overline{Z}_N \cdot \overline{Z}_N = \overline{Z}_N^2 \tag{۱۰.۳}$$

اگر رابطه‌های (۹.۳) و (۱۰.۳) را در (۸.۳) منظور کنیم، نتیجه می‌شود

$$\begin{aligned}
V(\overline{Z}_n) &= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{i=1}^{N} p_i Z_i^2 + \sum_{(i\neq j)=1}^{n} \overline{Z}_N^2\right] - \overline{Z}_N^2 \\
&= \frac{1}{n^2}\left[n\sum_{i=1}^{N} p_i Z_i^2 + n(n-1)\overline{Z}_N^2\right] - \overline{Z}_N^2 \\
&= \frac{1}{n}\sum_{i=1}^{N} p_i Z_i^2 + \frac{n-1}{n}\overline{Z}_N^2 - \overline{Z}_N^2
\end{aligned}$$

و سرانجام

$$V(\overline{Z}_n) = \frac{1}{n}\left(\sum_{i=1}^{N} p_i Z_i^2 - \overline{Z}_N^2\right)$$

عبارت داخل پرانتز برابر با واریانس جامعهٔ Z_i هاست. اگر واریانس این جامعه را با σ_Z^2 نشان دهیم، داریم

$$V(\overline{Z}_n) = \frac{\sigma_Z^2}{n}$$

یعنی با توجه به اینکه $\overline{Z}_n$ برآوردکنندهٔ $\overline{Y}_N$ است

$$V(\hat{\overline{Y}}_N) = \frac{\sigma_Z^2}{n}$$

از طرفی با توجه به تبدیل $Z_i = \frac{Y_i}{Np_i}$

$$\begin{aligned}
\sigma_Z^2 &= \sum_{i=1}^{N} p_i Z_i^2 - \overline{Z}_N^2 = \sum_{i=1}^{N} p_i \cdot \left(\frac{Y_i}{Np_i}\right)^2 - \overline{Y}_N^2 \\
&= \sum_{i=1}^{N} \frac{Y_i^2}{N^2 p_i} - \overline{Y}_N^2
\end{aligned}$$

پس

$$V(\hat{\overline{Y}}_N) = V(\overline{Z}_n) = \frac{1}{n}\left[\sum_{i=1}^{N} \frac{Y_i^2}{N^2 p_i} - \overline{Y}_N^2\right] \tag{۱۱.۳}$$

تبصرهٔ ۱. اگر قرار دهیم $p_i = \frac{1}{N}$، نمونه‌گیری با احتمال متغیر به نمونه‌گیری تصادفی سادهٔ با جایگذاری تبدیل می‌شود. رابطهٔ $Z_i = \frac{Y_i}{Np_i}$ به $Z_i = Y_i$ تبدیل می‌شود، و

$$V(\hat{\overline{Y}}_N) = V(\overline{Z}_n) = V(\overline{Y}_n) = \frac{\sigma_Y^2}{n}$$

تبصرهٔ ۲. اگر احتمال انتخاب Y_i متناسب با مقدار Y_i باشد، آن‌گاه

$$p_i = \frac{Y_i}{\sum_{i=1}^{N} Y_i}$$

و در نتیجه،

$$Y_i = p_i \sum_{i=1}^{N} Y_i$$

که اگر آن را در رابطهٔ $Z_i = \frac{Y_i}{Np_i}$ منظور کنیم، نتیجه می‌شود

$$Z_i = \frac{p_i \sum_{i=1}^{N} Y_i}{Np_i} = \overline{Y}_N$$

یعنی Z_i به‌ازای هر i همیشه برابر با $\overline{Y}_N$ است و لاجرم $\sigma_Z^2 = 0$ و این نمونه‌گیری کاراتر از نمونه‌گیری تصادفی ساده با جایگذاری است.

۷.۳ برآورد واریانس برآوردکنندهٔ میانگین در نمونه‌گیری با احتمال متغیر و با جایگذاری

در رابطه $V(\hat{\overline{Y}}_N) = \frac{\sigma_Z^2}{n}$، چون σ_Z^2 مجهول است نمی‌توانیم واریانس برآوردکنندهٔ میانگین جامعه را به‌دست آوریم، لذا باید از روی نمونه برآوردی برای این واریانس بیابیم. می‌توان نشان داد که s_Z^2، تغییرات نمونه، برآوردکننده‌ای ناااریب برای σ_Z^2 جامعه است (نظیر مورد نمونه‌گیری تصادفی ساده با جایگذاری). می‌دانیم

$$s_Z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2$$

پس

$$E(s_Z^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} Z_i^2 - n\overline{Z}_n^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(Z_i^2) - nE(\overline{Z}_n^2)\right] \qquad (۱۲.۳)$$

از طرفی

$$\mathrm{V}(\overline{Z}_n) = E(\overline{Z}_n^{۲}) - \overline{Z}_N^{۲}$$

پس

$$E(\overline{Z}_n^{۲}) = \frac{\sigma_Z^{۲}}{n} + \overline{Z}_N^{۲} \qquad (۱۳.۳)$$

ضمناً می‌دانیم که

$$E(Z_i^{۲}) = \sum_{i=۱}^{N} p_i Z_i^{۲} \qquad (۱۴.۳)$$

اگر (۱۳.۳) و (۱۴.۳) را در (۱۲.۳) قرار دهیم، نتیجه می‌شود

$$\begin{aligned} E(s_Z^{۲}) &= \frac{۱}{n-۱}\left[\sum_{i=۱}^{n}\sum_{i=۱}^{N} p_i Z^{۲} - n\left(\frac{\sigma_Z^{۲}}{n} + \overline{Z}_N^{۲}\right)\right] \\ &= \frac{۱}{n-۱}\left[n\sum_{i=۱}^{N} p_i Z_i^{۲} - n\overline{Z}_N^{۲} - \sigma_Z^{۲}\right] \\ &= \frac{۱}{n-۱}\left[n\left(\sum_{i=۱}^{N} p_i Z_i^{۲} - \overline{Z}_N^{۲}\right) - \sigma_Z^{۲}\right] \end{aligned}$$

اما عبارت داخل پرانتز برابر با $\sigma_Z^{۲}$ است، پس

$$E(s_Z^{۲}) = \frac{۱}{n-۱}(n\sigma_Z^{۲} - \sigma_Z^{۲}) = \sigma_Z^{۲}$$

یعنی $s_Z^{۲}$ برآوردکنندهٔ ناااریب $\sigma_Z^{۲}$ است. لذا با توجه به برابری

$$\mathrm{V}(\hat{\overline{Y}}_N) = \mathrm{V}(\overline{Z}_n) = \frac{\sigma_Z^{۲}}{n}$$

برآوردکنندهٔ ناااریب این واریانس به‌صورت زیر است

$$\hat{\mathrm{V}}(\hat{\overline{Y}}_N) = \frac{s_Z^{۲}}{n}$$

از برآوردکنندهٔ هنسن-هورویتس، برآوردکنندهٔ ناااریب مجموع واحدهای جامعه به‌صورت زیر به‌دست می‌آید

$$\hat{T} = N\hat{\overline{Y}}_N = \frac{۱}{n}\sum_{i=۱}^{n}\frac{Y_i}{p_i}$$

که با توجه به (۱۱.۳)، داریم

$$\mathrm{V}(\hat{T}) = \mathrm{V}(N\hat{\overline{Y}}_N) = \frac{۱}{n}\sum_{i=۱}^{N} p_i \left(\frac{Y_i}{p_i} - T\right)^۲$$

برآوردکنندهٔ نااریب این واریانس به‌وضوح عبارت است از

$$\hat{\mathrm{V}}(\hat{T}) = \frac{۱}{n(n-۱)}\sum_{i=۱}^{n} \left(\frac{Y_i}{p_i} - \hat{T}\right)^۲$$

یک بازهٔ اطمینان $(\alpha - ۱)$ درصد برای مجموع کل واحدهای جامعه، مبتنی بر نرمال بودن توزیع بزرگ نمونه‌ای برای $\hat{T}$، به‌صورت زیر است

$$\hat{T} \pm z\sqrt{\hat{\mathrm{V}}(\hat{T})}$$

که z نقطهٔ $\alpha/۲$ بالایی توزیع نرمال استاندارد است. برای حجمهای نمونه‌ای کمتر از ۵۰، باید از توزیع t با $n - ۱$ درجه آزادی استفاده کرد.

مثال ۵.۳ با توجه به داده‌های مثال ۴.۳ برآوردی نااریب برای واریانس برآوردکنندهٔ میانگین جامعه بیابید.

در مثال ۴.۳، پس از استفاده از تبدیل $Z_i = \frac{Y_i}{Np_i}$، به‌جای نمونهٔ به‌حجم ۶ از جامعهٔ Y_iها، نمونهٔ به‌حجم ۶ از جامعهٔ Z_iها را به‌صورت زیر به‌دست آوردیم

$$۱۰,\quad ۱۰٫۸,\quad ۲۴,\quad ۶٫۴,\quad ۲۵٫۲,\quad ۲۲٫۴$$

مقدار $\hat{\overline{Y}}_N$ حاصل از این نمونه برابر ۱۶٫۴۶ است. این مقدار رخدادی از متغیر تصادفی $\hat{\overline{Y}}_N$ است. حال می‌خواهیم واریانس این متغیر تصادفی را برآورد کنیم. ابتدا $s_Z^۲$ را به‌دست می‌آوریم

$$\begin{aligned} s_Z^۲ &= \frac{۱}{n-۱}\sum_{i=۱}^{n}(Z_i - \overline{Z}_n)^۲ = \frac{۱}{n-۱}\left(\sum_{i=۱}^{n} Z_i^۲ - n\overline{Z}_n^۲\right) \\ &= \frac{۱}{۵}[۱۰^۲ + \cdots + ۲۲٫۴^۲ - ۶(۱۶٫۴۶)^۲] = \frac{۱}{۵}(۱۹۷۰٫۴۰ - ۱۶۲۵٫۵۹) \\ &\simeq ۶۸٫۹۶ \end{aligned}$$

پس

$$\hat{\mathrm{V}}(\hat{\overline{Y}}_N) = \frac{۱}{۶}(۶۸٫۹۶) \simeq ۱۱٫۴۹$$

تبصره. برای کاربردها می‌توان $\hat{\mathrm{V}}(\hat{\overline{Y}}_N)$ را به‌صورت زیر نوشت

$$\begin{aligned}\hat{\mathrm{V}}(\hat{\overline{Y}}_N) &= \frac{s_Z^2}{n} = \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z}_n)^2\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Y_i}{Np_i}-\hat{\overline{Y}}_N\right)^2\\ &= \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}\left(\frac{Y_i}{p_i}-N\hat{\overline{Y}}_N\right)^2 \qquad (15.3)\end{aligned}$$

۸.۳ مقایسهٔ نمونه‌گیری تصادفی با احتمال متغیر و نمونه‌گیری تصادفی ساده (حالت با جایگذاری) در برآورد میانگین جامعه

از جامعهٔ به‌حجم N نمونه‌ای تصادفی به‌حجم n، با احتمال متغیر و با جایگذاری به‌دست می‌آوریم. با این داده‌ها میانگین جامعه را برآورد می‌کنیم. حال تصور می‌کنیم این داده‌ها حاصل یک نمونه‌گیری تصادفی ساده با جایگذاری است. یک بار دیگر با این تصور، میانگین جامعه را برآورد می‌نماییم. اگر این دو برآوردکننده را به‌ترتیب با $\overline{Z}_n$ و $\overline{Y}_{\mathrm{Ran}}$ نشان دهیم، برای مقایسهٔ دقت این دو برآوردکننده باید واریانس آنها را با هم مقایسه کنیم. اگر مقادیر واحدهای نمونه Y_1، Y_2، ... ، Y_n باشند، مطابق (۱۱.۳) داریم

$$\mathrm{V}(\overline{Z}_n) = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{N^2p_i}-\overline{Y}_N^2\right]$$

و

$$\begin{aligned}\mathrm{V}(\hat{Y}_{\mathrm{Ran}}) &= \frac{\sigma^2}{n} = \frac{1}{nN}\sum_{i=1}^{N}(Y_i-\overline{Y}_N)^2 = \frac{1}{nN}\left[\sum_{i=1}^{N}Y_i^2-N\overline{Y}_N^2\right]\\ &= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{N}-\overline{Y}_N^2\right]\end{aligned}$$

برای مقایسه دقت دو روش تفاضل دو واریانس را حساب می‌کنیم

$$\begin{aligned}d = \mathrm{V}(\overline{Z}_n)-\mathrm{V}(\overline{Y}_{\mathrm{Ran}}) &= \left[\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i^2}{N^2p_i}-\frac{1}{n}\overline{Y}_N^2\right]-\left[\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i^2}{N}-\frac{1}{n}\overline{Y}_N^2\right]\\ &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{Y_i^2}{Np_i}-Y_i^2\right)\\ &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{Np_i}-1\right)Y_i^2\end{aligned}$$

حال اگر به‌جای p_i مقدار آن $\frac{X_i}{\sum X_i}$ یا $\frac{X_i}{N\overline{X}_N}$ را قرار دهیم نتیجه می‌شود که

$$d = \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{\sum X_i}{NX_i} - 1\right)Y_i^2 = \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{\overline{X}_N}{X_i} - 1\right)Y_i^2$$

پس

$$d = \frac{1}{nN}\left(\overline{X}_N\sum_{i=1}^{N}\frac{Y_i^2}{X_i} - \sum_{i=1}^{N}Y_i^2\right) \qquad (۱۶.۳)$$

اگر داشته باشیم

$$\overline{X}_N\sum_{i=1}^{N}\frac{Y_i^2}{X_i} < \sum_{i=1}^{n}Y_i^2$$

در نتیجه، $d > 0$ و نمونه‌گیری با احتمال متغیر و با جایگذاری دقیقتر از نمونه‌گیری تصادفی ساده با جایگذاری است. نابرابری بالا را می‌توان به‌صورت زیر نوشت

$$\sum_{i=1}^{N}\frac{Y_i^2}{X_i} < \sum_{i=1}^{N}\frac{Y_i^2}{\overline{X}_N}$$

این نابرابری برحسب پارامترهای جامعهٔ اصلی و جامعهٔ کمکی بیان شده است، لذا کاربرد عملی در تعیین اینکه کدام‌یک از دو نمونه‌گیری در جامعه‌ای مفروض مناسبتر است ندارد. نابرابری اخیر را گاهی به‌صورت زیر بیان می‌کنند

$$\sum_{i=1}^{N}(X_i - \overline{X}_N)\frac{Y_i^2}{X_i} > 0 \qquad (۱۷.۳)$$

راج نشان داده است که این رابطه وقتی X_iها و $\frac{Y_i^2}{X_i}$ ها به‌صورت مثبت وابسته‌اند برقرار است و در چنین حالتی نمونه‌گیری با احتمال متغیر و با جایگذاری بهتر از نمونه‌گیری تصادفی سادهٔ با جایگذاری است.

تبصره. اگر رابطهٔ بین صفت اصلی و صفت کمکی به‌صورت خطی $Y = a + bX$، یعنی همبستگی بین Y و X کامل باشد با استفاده از نابرابری نتیجه می‌شود که نمونه‌گیری با احتمال متغیر وقتی دقت کمتری دارد که

$$\frac{\overline{X} - \tilde{X}}{\tilde{X}\sigma_X^2} > \frac{b^2}{a^2}$$

که در آن $\tilde{X} = \frac{N}{\sum(\frac{1}{X_i})}$.

۹.۳ نمونه‌گیری با احتمال متغیر و بدون جایگذاری

جامعه‌ای N واحدی با مقادیر $Y_۱$، $Y_۲$، . . . ، Y_i، . . . ، Y_N را در نظر می‌گیریم. فرض می‌کنیم $p_۱$، $p_۲$، . . . ، p_i، . . . ، p_N احتمالهای متناظر با واحدهای جامعه باشند. می‌خواهیم با روش بدون جایگذاری نمونه‌ای به‌حجم n از این جامعه انتخاب کنیم به‌قسمی که احتمال انتخاب واحد iام جامعه، $i = ۱, \ldots , N$ در هر استخراج برابر p_i، N , . . . ,۱ $= i$ باشد. واحد اول را نظیر حالت با جایگذاری به یکی از دو روشی که متذکر شدیم به‌دست می‌آوریم. اگر مثلاً واحد iام در انتخاب اول برای عضویت در نمونه انتخاب شود همان‌طور که دیدیم

$$P(\text{در بار اول } Y_i \text{ استخراج}) = p_i , \qquad i = ۱, \ldots , N$$

حال این واحد را به جامعه برنمی‌گردانیم. جامعه‌ای که فعلاً در اختیار داریم دارای مقادیر

$$Y_۱, \ Y_۲, \ldots , Y_{i-۱}, \ Y_{i+۱}, \ldots , Y_N$$

است که حجم آن ۱ $- N$ است. احتمالهای متناظر با این واحدها در جامعهٔ اصلی با دنبالهٔ زیر مشخص می‌شود

$$p_۱, \ p_۲, \ldots , p_{i-۱}, \ p_{i+۱}, \ldots , p_N$$

در مورد این جامعهٔ جدید و احتمالهای متناظر با واحدهای آن، نمی‌توانیم به دو روش مذکور، واحد دوم نمونه را انتخاب کنیم زیرا مجموع این احتمالها برابر با یک نیست. برای تهیهٔ واحدهای نمونه در حالت بدون جایگذاری پیشنهادهایی متعدد که کم و بیش با پیچیدگی همراه‌اند وجود دارند. این روشها را می‌توان عمدتاً در مراجع [۸] و [۱۳] یافت. از جملهٔ این روشها، روش رائو و همکاران، روش دوربین[۱]، روش سانتر[۲]، و روش هدایت-لین است. ما برای تعیین واحدهای دوم، سوم، . . . ، و nام نمونه بدون جایگذاری به اثبات و ذکر جزئیات مطالب نمی‌پردازیم، و تنها روشی را که به برآورد مرتب موسوم است به‌صورتی مشروح مطرح می‌کنیم. در پایان نیز به‌اختصار شیوهٔ نمونه‌گیری رائو-هارتلی-کوکران را بررسی می‌نماییم.

۱.۹.۳ برآورد مرتب

پس از انتخاب اولین واحد نمونه که فرض می‌کنیم واحد iام جامعه است، جامعهٔ باقی‌مانده را که به‌حجم ۱ $- N$ است با مقادیر و احتمالهای متناظری که در زیر آورده‌ایم در نظر می‌گیریم

$$\begin{array}{ccccccc} Y_۱ & Y_۲ & \ldots & Y_{i-۱}, & Y_{i+۱}, & \ldots & Y_N \\ \dfrac{p_۱}{۱-p_i} & \dfrac{p_۲}{۱-p_i} & \ldots & \dfrac{p_{i-۱}}{۱-p_i}, & \dfrac{p_{i+۱}}{۱-p_i}, & \ldots, & \dfrac{p_N}{۱-p_i} \end{array} \qquad (۱۸.۳)$$

1. Durbin 2. Sunter

بدین طریق، احتمالهای متناظر با واحدها در جامعهٔ اصلی را با تقسیم بر $1 - p_i$ به احتمالهایی متناسب که مجموع آنها یک است تبدیل کرده‌ایم. واضح است که

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{p_j}{1-p_i} = \frac{1}{1-p_i}\sum_{\substack{j=1 \\ j\neq i}}^{N} p_j = \frac{1-p_i}{1-p_i} = 1$$

حال در این جامعهٔ جدید به‌حجم $N - 1$، با توجه به احتمالهای متناظر با واحدها به یکی از دو روش مجموع تراکمی یا لاهیری، واحدی را که اولین انتخاب از این جامعه است به‌عنوان دومین واحد نمونهٔ بدون جایگذاری از جامعهٔ اصلی برمی‌گزینیم. با همین فرایند جامعهٔ باقی‌مانده را که به‌حجم $N - 2$ است در نظر می‌گیریم، و اگر Y_i و Y_j دو واحد منتخب قبلی باشند، این جامعه و احتمالهای متناظر با واحدهای آن را به‌صورت زیر می‌نویسیم

$$\begin{matrix} Y_1 & \ldots & Y_{j-1} & Y_{j+1} & \ldots & Y_{i-1} & Y_{i+1} & \ldots & Y_N \\ \frac{p_1}{1-p_i-p_j} & \ldots & \frac{p_{j-1}}{1-p_i-p_j} & \frac{p_{j+1}}{1-p_i-p_j} & \ldots & \frac{p_{i-1}}{1-p_i-p_j} & \frac{p_{i+1}}{1-p_i-p_j} & \ldots & \frac{p_N}{1-p_i-p_j} \end{matrix}$$

باز هم احتمالها متناسب‌اند و مجموع آنها یک است. به یکی از دو روش نمونه‌گیری با جایگذاری، یک واحد از این جامعه را به‌عنوان سومین واحد نمونهٔ جامعهٔ اصلی انتخاب می‌کنیم. این فرایند را تا انتخاب n واحد نمونه ادامه می‌دهیم. بدیهی است انجام این فرایند به‌کمک کامپیوتر چندان مشکل نخواهد بود.

حال نمونهٔ حاصل از این فرایند را در نظر می‌گیریم، اولین واحد نمونه، واحد شمارهٔ i با مقدار Y_i از جامعهٔ اصلی است و احتمال متناظر با آن p_i است. اگر قرار دهیم

$$Z_1 = \frac{Y_i}{Np_i} \tag{۱۹.۳}$$

داریم

$$E(Z_1) = \sum_{i=1}^{N} \frac{Y_i}{Np_i} \cdot p_i = \overline{Y}_N$$

دومین واحد نمونه، واحد شمارهٔ j با مقدار Y_j از جامعه است، و احتمال متناظر با آن p_j است. اگر قرار دهیم

$$Z_2 = \frac{1}{N}\left[Y_i + Y_j \frac{1-p_i}{p_j}\right] \tag{۲۰.۳}$$

آن‌گاه برای iی ثابت، با توجه به (۱۸.۳)

$$E\left[Y_j\frac{1-p_i}{p_j}\middle|Y_i\right] = \sum_{\substack{j=1\\ j\neq i}}^{N} Y_j\frac{1-p_i}{p_j}\cdot\frac{p_j}{1-p_i}$$
$$= \sum_{\substack{j=1\\ j\neq i}}^{N} Y_j = N\overline{Y}_N - Y_i$$

لذا

$$E(Z_2) = \frac{1}{N}\left[E(Y_i) + E\left(Y_j\frac{1-p_i}{p_j}\right)\right]$$

چون i ثابت است

$$= \frac{1}{N}[Y_i + N\overline{Y}_N - Y_i] = \overline{Y}_N$$

اگر روابطی مشابه با (۲۰.۳) برای واحدهای سوم و چهارم و ... و nام نمونه بنویسیم، میانگین هر Z_i، در دنبالهٔ Z_1، Z_2، ... ، Z_n برابر با $\overline{Y}_N$ است. لذا میانگین

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

نیز برابر با $\overline{Y}_N$ است. یعنی $\overline{Z}_n$ برآوردکنندهٔ نااریب $\overline{Y}_N$ است.

در مورد روابط مشابه با (۲۰.۳)، می‌توان مطلب را به‌صورت کلی زیر مطرح کرد:
اگر Y_1، Y_2، ... ، Y_n مقادیر واحدهای نمونه باشند که با توضیح بالا، بدون جایگذاری انتخاب شده‌اند، و اگر p_1، ... ، p_n احتمالهای متناظر با آنها فرض شوند، متغیر Z_i را که تعمیمی از (۲۰.۳) است به‌صورت زیر می‌نویسیم

$$Z_i = \frac{1}{N}\left[Y_1 + Y_2 + \cdots + Y_{i-1} + Y_i\frac{1-(p_1+p_2+\cdots+p_{i-1})}{p_i}\right] \quad (۲۱.۳)$$
$$i = 1,\ 2, \ldots, n$$

در این صورت $\overline{Z}_n$ برای دنبالهٔ Z_1، Z_2، ... ، Z_n برآوردکنندهٔ نااریب $\overline{Y}_N$ است. می‌توان ثابت کرد که برای $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$

$$\hat{V}(\overline{Z}_n) = \hat{V}(\hat{\overline{Y}}_N) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2 \quad (۲۲.۳)$$

که امکان می‌دهد واریانس برآوردکنندهٔ $\overline{Y}_N$ را برآورد کنیم.

مثال ۶.۳ جدول زیر درآمد روزانهٔ ۱۰ خانوار و تعداد افراد هر خانوار را نشان می‌دهد. Y_i درآمد روزانه و X_i تعداد افراد خانوادهٔ iام است که اولی صفت اصلی و دومی صفت کمکی است.

شمارهٔ خانوار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Y_i برحسب ۱۰۰۰	۶	۸	۷	۶٫۵	۵	۴	۹	۷	۸	۷٫۵
X_i	۳	۴	۴	۳	۲	۳	۵	۳	۴	۳
T_i	۳	۷	۱۱	۱۴	۱۶	۱۹	۲۴	۲۷	۳۱	۳۴

می‌خواهیم با نمونه‌ای به‌حجم ۴ بدون جایگذاری، برآوردی برای میانگین درآمد روزانه به‌دست آوریم. ابتدا با توجه به شیوهٔ مجموع تراکمی، با مراجعه به جدول اعداد تصادفی، عددی از اعداد ۱ تا ۳۴ اختیار می‌کنیم. فرض کنید ۲۲ به‌دست آید. در این صورت $Y_{۷} = ۹$ اولین واحد نمونه است. احتمالهای متناظر با Y_iها در جامعهٔ اصلی به‌صورت زیرند

$$\frac{۳}{۳۴},\ \frac{۴}{۳۴},\ \frac{۴}{۳۴},\ \frac{۳}{۳۴},\ \frac{۲}{۳۴},\ \frac{۳}{۳۴},\ \frac{۵}{۳۴},\ \frac{۳}{۳۴},\ \frac{۴}{۳۴},\ \frac{۳}{۳۴}$$

چون $Y_{۷}$ را به‌عنوان اولین واحد نمونه انتخاب کرده‌ایم و نمونه‌گیری بدون جایگذاری است، جامعه‌ای جدید به‌حجم ۹ تشکیل می‌شود که احتمهالهای متناظر با واحدهای آن، طبق آنچه گفتیم از $\frac{p_i}{1-p_i}$ به‌دست می‌آیند. چون $p_{۷} = \frac{۵}{۳۴}$، داریم

واحدهای جامعه	$Y_{۱}$	$Y_{۲}$	$Y_{۳}$	$Y_{۴}$	$Y_{۵}$	$Y_{۶}$	$Y_{۸}$	$Y_{۹}$	$Y_{۱۰}$
احتمالهای متناظر	$\frac{۳}{۲۹}$	$\frac{۴}{۲۹}$	$\frac{۴}{۲۹}$	$\frac{۳}{۲۹}$	$\frac{۲}{۲۹}$	$\frac{۳}{۲۹}$	$\frac{۳}{۲۹}$	$\frac{۴}{۲۹}$	$\frac{۳}{۳۹}$

بنابراین، جدول زیر را داریم

شمارهٔ خانوار	۱	۲	۳	۴	۵	۶	۸	۹	۱۰
Y_i برحسب ۱۰۰۰	۶	۸	۷	۶٫۵	۵	۴	۷	۸	۷٫۵
X_i	۳	۴	۴	۳	۲	۳	۳	۴	۳
T_i	۳	۷	۱۱	۱۴	۱۶	۱۹	۲۲	۲۶	۲۹

ملاحظه می‌کنید سه سطر اول، همان سه سطر اول جدول قبل هستند که $Y_{۷}$ آنها کنار گذاشته شده است.

به‌شیوهٔ مجموع تراکمی، با مراجعه به جدول اعداد تصادفی، عددی از اعداد ۱ تا ۲۹ اختیار می‌کنیم. فرض کنید ۹ به‌دست آید، لذا $Y_{۳} = ۷$ دومین واحد نمونه است. از حذف $Y_{۳}$ در جدول بالا سطر T_i به‌صورت زیر درمی‌آید

$$\begin{array}{llllllllll} Y_i: & Y_{۱} & Y_{۲} & Y_{۴} & Y_{۵} & Y_{۶} & Y_{۸} & Y_{۹} & Y_{۱۰} \\ T_i: & ۳ & ۷ & ۱۰ & ۱۲ & ۱۵ & ۱۸ & ۲۲ & ۲۵ \end{array}$$

از اعداد ۱ تا ۲۵ عددی به تصادف برمی‌گزینیم. فرض کنید ۲۴ به‌دست آید. پس $Y_{۱۰} = ۷٫۵$ سومین واحد نمونه است. پس از حذف این واحد، داریم

$$
\begin{array}{llllllll}
Y_i : & Y_۱ & Y_۲ & Y_۴ & Y_۵ & Y_۶ & Y_۸ & Y_۹ \\
T_i : & ۳ & ۷ & ۱۰ & ۱۲ & ۱۵ & ۱۸ & ۲۲
\end{array}
$$

از اعداد ۱ تا ۲۲، عددی به تصادف اختیار می‌کنیم. فرض کنید ۱۱ به‌دست آید، پس $Y_۵ = ۵$ چهارمین واحد نمونه است. بنابراین نمونهٔ منتخب به‌صورت زیر است

i شمارهٔ واحد منتخب :	۷	۳	۱۰	۵
Y_i مقادیر :	۹	۷	۷٫۵	۵
احتمالهای متناظر :	$\frac{۵}{۳۴}$	$\frac{۴}{۳۴}$	$\frac{۳}{۳۴}$	$\frac{۲}{۳۴}$

Z_iها را مطابق فرمول (۲۱.۳) محاسبه می‌کنیم

$$Z_۱ = \frac{Y_۱}{Np_۱} = \frac{۹}{۱۰\left(\frac{۵}{۳۴}\right)} = \frac{۳۰۶}{۵۰}$$

$$Z_۲ = \frac{۱}{N}\left(Y_۱ + Y_۲ \cdot \frac{۱-p_۱}{p_۲}\right) = \frac{۱}{۱۰}\left[۹+۷\left(\frac{\frac{۲۹}{۳۴}}{\frac{۴}{۳۴}}\right)\right] = \frac{۱}{۱۰}\left[۹+\frac{۲۰۳}{۴}\right] = \frac{۱۲۱}{۲۰}$$

$$Z_۳ = \frac{۱}{N}\left(Y_۱ + Y_۲ + Y_۳ \cdot \frac{۱-p_۱-p_۲}{p_۳}\right) = \frac{۱}{۱۰}\left[۹+۷+۷٫۵\left(\frac{\frac{۲۵}{۳۴}}{\frac{۳}{۳۴}}\right)\right] = \frac{۱۵۷}{۲۰}$$

$$Z_۴ = \frac{۱}{N}\left(Y_۱ + Y_۲ + Y_۳ + Y_۴ \cdot \frac{۱-p_۱-p_۲-p_۳}{p_۴}\right) = \frac{۱}{۱۰}\left[۲۳٫۵+۵\left(\frac{\frac{۲۲}{۳۴}}{\frac{۲}{۳۴}}\right)\right] = \frac{۱۵۷}{۲۰}$$

$$\hat{\overline{Y}}_N = \overline{Z}_n = \frac{۱}{۴}\left(\frac{۳۰۶}{۵۰} + \frac{۱۲۱}{۲۰} + \frac{۱۵۷}{۲۰} + \frac{۱۵۷}{۲۰}\right) \simeq ۶٫۹۶$$

از رابطهٔ (۲۲.۳) می‌توان برآورد واریانس $\hat{\overline{Y}}_N$ را که همان برآورد واریانس $\hat{Z}_n$ است محاسبه کرد.▲

۲.۹.۳ طرح نمونه‌گیری رائو-هارتلی-کوکران و برآوردکنندهٔ وابسته به آن

در این بخش رهبرد نمونه‌گیری معروف و راحت دیگری را که رائو و همکاران او پیشنهاد کرده‌اند و مبتنی بر صفت کمکی است متذکر می‌شویم. موضوع را با گروهبندی تصادفی جامعه تحت بررسی آغاز می‌کنیم.

جامعه P به‌حجم N و مجموعهٔ اعداد صحیح $N_۱$، $N_۲$، . . . ، N_n را در نظر می‌گیریم، به‌قسمی که برای $۱ \leq i \leq n$ ،$N_i \geq ۱$ و $N = \sum_{i=۱}^{n} N_i$. در این صورت می‌گوییم $\{G_۱, G_۲, \ldots, G_n\}$ یک گروهبندی تصادفی $(N, n; N_۱, N_۲, \ldots, N_n)$ از جامعهٔ P است اگر

$$P = \bigcup_{i=۱}^{n} G_i$$

و برای هر $۱ \leq i \leq n$، تعداد عناصر موجود در G_i عبارت باشد از $|G_i| = N_i$ و به‌علاوه برای هر i در تشکیل G_i، هر N_iتایی جامعهٔ P شانس برابر برای انتخاب شدن داشته باشد. یک راه

تشکیل چنین گروهبندی تصادفی این است که فرض کنیم G_1 یک نمونهٔ تصادفی از P است که به‌روش تصادفی ساده به‌حجم N_1 انتخاب شده است، G_2 نمونه‌ای تصادفی از $P - G_1$ است که از جامعهٔ به‌حجم $N - N_1$ انتخاب شده است و نظایر آن. در مرحلهٔ $(n-1)$ام، نمونه G_{n-1} به‌حجم N_{n-1} از جامعهٔ به‌حجم $N - \sum_{i=1}^{n-2} N_i$ به‌روش تصادفی ساده انتخاب می‌شود، و بالاخره در مرحله nام، G_n از باقیماندهٔ واحدهای جامعه تشکیل می‌شود.

حال طرح نمونه‌گیری رائو-هارتلی-کوکران (RHC) را از جامعهٔ P به‌حجم N با احتمالهای متناظر p_1، p_2، ... ، p_N برای ورود به‌نمونه شرح می‌دهیم.

روش، شامل دو مرحله است. در زیر N_1، N_2، ... ، N_n را به‌عنوان مقادیر دلخواه ولی ثابت در نظر می‌گیریم که $N_i \geq 1$، $1 \leq i \leq n$ و $\sum_{i=1}^{n} N_i = N$.

مرحلهٔ ۱. ابتدا یک گروهبندی تصادفی $(N, n; N_1, N_2, \ldots, N_n)$ جامعهٔ P را که مثلاً به‌صورت $\{G_1, G_2, \ldots, G_n\}$ است تشکیل می‌دهیم.

مرحلهٔ ۲. با استفاده از روش PPS نمونه‌گیری مستقل در داخل هر گروه، یک واحد از هر گروه استخراج می‌کنیم. بنابر روش PPS، با فرض داشتن گروهبندی تصادفی، شانس انتخاب iامین واحد در نمونه، اگر $i \in G_K$، برابر با $\frac{p_i}{P_K}$ است که در آن $P_K = \sum_{j \in G_K} p_j$ با شرط $1 \leq K \leq n$، $1 \leq i \leq N$.

برآوردکنندهٔ مجموع جامعه که با طرح RHC پیشنهاد می‌شود به‌صورت زیر است

$$\hat{T}_{\mathrm{RHC}} = \sum_{K=1}^{n} Y_{i_K} P_K / p_{i_K}$$

که در آن i_1، i_2، ... ، i_K، ... ، i_n نمونهٔ منتخبی است که واحد i_K از G_K، $1 \leq K \leq n$ انتخاب شده است. در زمینهٔ طرح RHC قضیهٔ زیر را بدون اثبات ذکر می‌کنیم.

قضیهٔ ۲.۳ تحت طرح نمونه‌گیری RHC، برآوردکنندهٔ $\hat{T}_{\mathrm{RHC}}$ برای مجموع واحدهای جامعه نااریب است و واریانس آن به‌صورت زیر است

$$\mathrm{V}(\hat{T}_{\mathrm{RHC}}) = \left\{ \left(\sum_{K=1}^{n} N_K^2 - N \right) \Big/ (N^2 - N) \right\} \left\{ \sum_{i=1}^{N} \left(\frac{Y_i}{p_i} - T \right)^2 p_i \right\}$$

و به‌علاوه، برآورد نااریب این واریانس به‌صورت زیر است

$$\hat{\mathrm{V}}(\hat{T}_{\mathrm{RHC}}) = \left\{ \left(\sum_{K=1}^{n} N_K^2 - N \right) \Big/ \left(N^2 - \sum_{K=1}^{n} N_K^2 \right) \right\} \left\{ \sum_{K=1}^{n} \left(\frac{Y_{i_K}}{p_{i_K}} - \hat{T}_{\mathrm{RHC}} \right)^2 P_K \right\}$$

تبصره. تا اینجا فرض کردیم N_iها متفاوت، دلخواه ولی تثبیت شده‌اند و $N_K \geq ۱$، $۱ \leq K \leq n$، $\sum_{K=۱}^{n} N_K = N$. با بررسی فرمول $\mathrm{V}(\hat{T}_{\mathrm{RHC}})$ می‌توان نتیجه گرفت که انتخاب $N_۱$، $N_۲$، ...، N_n که این واریانس را مینیمم می‌کند متناظر است با

$$N_۱ = N_۲ = \cdots = N_n = \frac{N}{n} \qquad \text{اگر } N \text{ بر } n \text{ تقسیم‌پذیر باشد}$$

و

$$\begin{cases} N_۱ = N_۲ = \cdots = N_u = m + ۱ \\ N_{u+۱} = N_{u+۲} = \cdots = N_n = m \end{cases} \qquad \text{اگر } N = mn + u\text{، } ۱ \leq u \leq n - ۱$$

به‌عبارت دیگر برای اینکه طرح نمونه‌گیری RHC را به مفیدترین صورت درآوریم باید گروهبندی تصادفی را به‌نحوی انجام دهیم که حجمهای گروهی تا حد ممکن به‌هم نزدیک باشند. چنگ[1] و لی[2] (۱۹۸۳، ۱۹۸۷) و داس‌گوپتا و سینها (۱۹۸۳) بررسیهایی در زمینهٔ ویژگیهایی نظیر مینیماکس بودن نمونه‌گیری RHC با انتخاب اپتیمم حجمهای گروهی انجام داده‌اند.

مثال ۷.۳ جامعه‌ای متشکل از ۱۵ واحد است. در زیر مقادیر صفتهای کمکی متناظر با آنها را آورده‌ایم

شمارهٔ واحدها i:	۱	۲	۳	۴	۵	۶	۷	۸
صفت کمکی X_i:	۲٫۳	۳	۴٫۱	۲٫۶	۵٫۳	۶	۲٫۸	۵٫۲

شمارهٔ واحدها i:	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵
صفت کمکی X_i:	۱۱٫۳	۷٫۲	۸	۳٫۵	۴٫۲	۳٫۸	۵٫۲

می‌خواهیم مجموع واحدهای جامعه را برمبنای نمونه‌ای به‌حجم ۳ با استفاده از شیوهٔ RHC برآورد کنیم. قرار می‌دهیم $N_۱ = N_۲ = N_۳ = ۵$ که متناظر با انتخاب اپتیمم N_i هاست.

مرحلهٔ ۱. به جدول اعداد تصادفی رجوع می‌کنیم. فرض می‌کنیم گروهبندی تصادفی زیر حاصل شود

$$G_۱ = \{۷,\ ۳,\ ۱۲,\ ۸,\ ۱\} \qquad G_۲ = \{۹,\ ۱۴,\ ۶,\ ۲,\ ۱۰\}$$
$$G_۳ = \{۴,\ ۵,\ ۱۱,\ ۱۳,\ ۱۵\}$$

مرحلهٔ ۲. با توجه به داده‌ها $\sum X_i = ۷۴٫۵$. لذا از $p_i = \frac{X_i}{\sum X_i}$ همهٔ مقادیر p_i، $۱ \leq i \leq ۱۵$ مشخص می‌شوند. حال با واحدهای $G_۱$، $G_۲$، و $G_۳$، مقادیر p_i متناظر را همراه می‌کنیم و با استفاده از روش PPS از هرکدام یک واحد انتخاب می‌کنیم. می‌توان فرض کرد سه واحد منتخب عبارت‌اند از

$$i_۱ = ۷ \qquad i_۲ = ۲ \qquad i_۳ = ۱۳$$

1. cheng 2. Li

حال به واحد شمارهٔ ۷، ۲، و ۱۳ جامعه مراجعه کرده مشخصهٔ مورد نظر را مشاهده می‌کنیم. اگر اندازهٔ مشخصه‌ها به‌صورت زیر باشند

$$Y_{۷} = ۳۰ \qquad Y_{۲} = ۳۲ \qquad Y_{۱۳} = ۴۵$$

آن‌گاه محاسبات زیر را انجام می‌دهیم

$$P_{۱} = \sum_{i\in G_{۱}} p_i = \frac{X_{۷}+X_{۳}+X_{۱۲}+X_{۸}+X_{۱}}{\sum_i X_i} = \frac{۲٫۸+۴٫۱+۳٫۵+۵٫۲+۲٫۳}{۷۴٫۵}$$
$$= ۰٫۲۴۰۳$$

$$P_{۲} = \sum_{i\in G_{۲}} p_i = \frac{X_{۹}+X_{۱۴}+X_{۶}+X_{۲}+X_{۱۰}}{\sum_i X_i} = \frac{۱۱٫۳+۳٫۸+۶+۳+۷٫۲}{۷۴٫۵}$$
$$= ۰٫۴۲۰۱$$

$$P_{۳} = \sum_{i\in G_{۳}} p_i = ۰٫۳۳۹۶$$

$$p_{i_{۱}} = p_{۷}, \qquad \frac{p_{۷}}{P_{۱}} = \frac{\frac{۲٫۸}{۷۴٫۵}}{\frac{۱۷٫۹}{۷۴٫۵}} = \frac{۲٫۸}{۱۷٫۹} = ۰٫۱۵۶۴$$

$$p_{i_{۲}} = p_{۲}, \qquad \frac{p_{۲}}{P_{۲}} = \frac{۳}{۳۱٫۳} = ۰٫۰۹۵۸$$

$$p_{i_{۳}} = p_{۱۳}, \qquad \frac{p_{۱۳}}{p_{۳}} = \frac{۴٫۲}{۲۵٫۳} = ۰٫۱۶۶۰$$

لذا

$$\hat{T}_{\mathrm{RHC}} = \sum_{K=۱}^{n} Y_{i_K} P_K / p_{i_K}$$
$$= ۳۰\left(\frac{۱}{۰٫۱۵۶۴}\right) + ۳۲\left(\frac{۱}{۰٫۰۹۵۸}\right) + ۴۵\left(\frac{۱}{۰٫۱۶۶۰}\right)$$
$$\# ۱۹۱٫۸۱۵۹ + ۳۳۴٫۰۲۹۲ + ۲۷۱٫۰۸۴۳$$
$$= ۷۹۶٫۹۲۹۴$$

از رابطهٔ $\hat{\mathrm{V}}(\hat{T}_{\mathrm{HRC}})$ برآورد واریانس $\hat{T}_{\mathrm{RHC}}$ برابر با ۲۰۸٫۱۴ به‌دست می‌آید. ▲

تمرینها

۱ . مشخصهٔ مورد بررسی، میزان پنبهٔ تولیدشده در ۱۰ روستای گرگان است. مساحت زمین زیر کشت پنبه در این ۱۰ روستا به‌ترتیب ۱۴، ۱۲، ۱۰، ۱۶، ۱۸، ۲۰، ۱۰، ۱۴، ۱۷، ۲۰ هکتار است.

ای بررسی میزان پنبه، ابتدا باید نمونه‌ای به‌حجم ۴ با احتمال متغیر و متناسب با مساحتهای متناظر انتخاب کنیم. این نمونه را یک بار با روش جایگذاری و بار دیگر با روش بدون جایگذاری مشخص کنید.

۲. از جامعه‌ای که به‌حجم ۵۰ است، نمونه‌ای به‌حجم ۸، با احتمال متغیر و به روش تصادفی با جایگذاری انتخاب کرده، مشخصهٔ واحدهای منتخب را همراه با احتمال متناظر با آنها در جدول زیر ثبت کرده‌ایم

Y_i	۱۰۰	۱۰۴	۱۱۵	۱۱۲	۱۲۰	۱۲۵	۱۲۲	۱۵۰
p_i	۳/۱۰۰	۲/۱۰۰	۱/۵۰	۱/۵۰	۱/۴۰	۱/۶۰	۱/۴۵	۲/۹۰

الف) برآورد نااریبی برای میانگین جامعه بیابید.

ب) برآورد نااریبی برای واریانس برآوردکنندهٔ میانگین جامعه به‌دست آورید.

۳. جامعه‌ای مرکب از ۱۲ واحد است می‌خواهیم نمونه‌ای به‌حجم ۴، با احتمال متغیر، و به روش تصادفی با جایگذاری از این جامعه انتخاب کنیم. مقدار و احتمالهای متناظر با واحدها به‌صورت زیرند:

Y_i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}
p_i	۱/۱۲	۱/۱۸	۱/۹	۱/۱۲	۱/۶	۱/۱۸	۱/۹	۱/۱۲	۱/۱۵	۱/۱۸	۱/۱۲	۲/۴۵

نمونه را یک بار با روش لاهیری و یک‌بار با روش مجموع تراکمی به‌دست آورید.

۴. فرض کنید از جامعهٔ بالا نمونه‌ای به‌حجم ۴، بدون جایگذاری و با احتمال متغیر انتخاب کرده‌ایم. مشخصهٔ ۴ واحد منتخب را اندازه‌گیری کرده و با احتمالهای متناظر در جدول زیر آورده‌ایم

Y_i	۱۰	۱۱	۸	۱۳
p_i	۲/۴۵	۱/۹	۱/۶	۱/۱۲

الف) میانگین جامعه را برآورد کنید.

ب) برآورد واریانس برآوردکنندهٔ میانگین جامعه را به‌دست آورید.

۵. از جامعه مذکور در تمرین ۳، نمونه‌ای به‌حجم ۴، بدون جایگذاری، انتخاب کنید و میانگین جامعه را برآورد نمایید.

۶. جامعه‌ای متشکل از ۱۲ واحد است. در زیر مقادیر صفتهای کمکی متناظر با آنها را آورده‌ایم

شمارهٔ واحد i:	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
صفت کمکی X_i:	۲	۵	۷	۳	۱	۱٫۲	۱٫۸	۲٫۳	۳٫۷	۴	۳٫۱	۱٫۹

مجموع واحدهای جامعه را برمبنای نمونه‌ای به‌حجم ۴ به‌شیوهٔ RHC برآورد کنید.

تمرینهای چهارگزینه‌ای

۱. در جامعه‌ای به‌حجم ۱۰۰ اگر بزرگترین واحد جامعهٔ کمکی ۶۰۰ باشد و بخواهیم با روش

لاهیری نمونه‌ای با جایگذاری و با احتمال متغیر انتخاب کنیم، احتمال مؤثر بودن انتخاب زوج $(۵, j)$ برابر است با

الف) $\frac{X_۵}{۶۰۰}$ ب) $\frac{X_۵}{۱۰۰}$ ج) $\frac{X_۵}{۶۰۰۰۰}$ د) $\frac{X_j}{۵}$

۲ . در یک نمونه‌گیری تصادفی با احتمال متغیر، میانگین صفت کمکی ۴۸ است. اگر بزرگترین واحد صفت کمکی ۶۰ باشد و نمونه‌گیری با جایگذاری و با روش لاهیری انجام شود، احتمال ناموثر بودن هر انتخاب برابر است با

الف) $\frac{۴}{۵}$ ب) $\frac{۱}{۴۸}$ ج) $\frac{۱}{۶۰}$ د) $\frac{۱}{۵}$

۳ . در یک نمونه‌گیری تصادفی با احتمال متغیر و با جایگذاری نمونه‌ای به‌حجم ۵ گرفته‌ایم. اگر حجم جامعه ۵۰ و مجموع $\frac{Y_i}{p_i}$های نمونه ۵۰۰ باشد برآورد نااریب میانگین جامعه برابر است با

الف) ۲ ب) ۱۰ ج) $\frac{۱}{۲}$ د) ۱۰۰

۴ . در جامعه‌ای به‌حجم ۱۰ نمونه‌ای به‌حجم ۲ با احتمال متغیر و بدون جایگذاری گرفته‌ایم. اندازهٔ این ۲ واحد به‌ترتیب ۱۲ و ۱۶ و احتمالهای متناظر با آنها به‌ترتیب $\frac{۱}{۸}$ و $\frac{۱}{۱۰}$ است. در این صورت برآورد میانگین جامعه برابر است با

الف) ۱۵ ب) ۱۵٫۵ ج) ۱۴ د) ۱۲٫۴

۵. اگر نمونه‌ای به‌روش تصادفی با احتمال متغیر و با جایگذاری از جامعه‌ای به‌حجم ۵۰ بگیریم و اگر در جامعه $\sum \frac{Y_i^۲}{X_i} = ۲$ و $\sum Y_i^۲ = ۱۰۰$ و $\sum X_i = ۴۰۰۰$، آن‌گاه دقت نمونه‌گیری مزبور از نمونه‌گیری تصادفی سادهٔ با جایگذاری

الف) بیشتر است ب) کمتر است ج) دقتها یکسان‌اند د) نمی‌توان قضاوت کرد.

۴

نمونه‌گیری تصادفی با طبقه‌بندی

۰.۴ مقدمه

(در نمونه‌گیری تصادفی ساده، واریانس برآورد میانگین جامعه علاوه‌بر اینکه به حجم نمونه بستگی دارد به تغییرپذیری مشخصهٔ تحت بررسی هم وابسته است.) اگر جامعه خیلی ناهمگن باشد و به‌دلیل محدودیتهای اقتصادی نتوانیم حجم نمونه را بزرگ اختیار کنیم تقریباً غیرممکن است که با روش نمونه‌گیری تصادفی ساده، برآوردی به حد کافی دقیق برای پارامتر مورد نظر جامعه بیابیم. مثلاً در بررسیهای مربوط به مشخصه‌های کارکنان شرکتها در شهری بزرگ، تعداد کارکنان در برخی از شرکتها بیش از ۱۰۰۰ نفر و در بعضی محدود به ده نفر یا کمتر است. در این صورت انتخاب نمونه تصادفی ساده از جامعهٔ شرکتها، با توجه به ناهمگنی تعداد کارکنان، از نمونه‌ای به نمونهٔ دیگر با نوسانات شدید همراه است. اما اگر ممکن باشد که شرکتها را از نظر تعداد کارکنان به چند طبقه با حجم خیلی کم، کم، متوسط، زیاد، و خیلی زیاد تقسیم کنیم و از هر طبقه برای بررسی مشخصهٔ مورد نظر به انتخاب نمونه بپردازیم از نوسانات شدید در نمونه‌های مختلف جلوگیری خواهد شد. این رهیافت، نمونه‌گیری دیگری را تحت عنوان نمونه‌گیری با طبقه‌بندی مطرح می‌کند. این روش، تکنیکی بسیار متداول است که به‌دلایل زیاد انجام می‌شود. عمده‌ترین این دلایل به شرح زیرند:

۱. اگر برای بعضی از زیرجامعه‌های یک جامعه، داده‌ها و اطلاعاتی با دقت معلوم بخواهند لازم است که هر زیرجامعه را یک طبقه به حساب آورند.

۲. تشکیلاتی که در یک کشور، مسئول انجام نمونه‌گیری برای ارائهٔ نتایج به سازمانهای ذی‌ربط

است، در هر یک از نواحی مختلف کشور واحدهایی دارد. کارکنان هر واحد دربارهٔ ویژگیهای مورد نظر در ناحیهٔ خود، اطلاعاتی دقیقتر از سایرین دارند و لذا اگر نمونه‌گیری در هر ناحیه، به‌عنوان یک طبقه، به‌صورتی مستقل از نواحی دیگر صورت گیرد با دقتی بیشتر همراه بوده، به‌علاوه از لحاظ میزان هزینه و نحوهٔ سازماندهی کار نمونه‌گیری، تسهیلاتی بیشتر فراهم می‌شوند. همگنی تقریبی بعضی از صفات تحت نمونه‌گیری در یک ناحیه نیز، به گونه‌ای که خواهیم دید به بالا بردن کارایی نمونه‌گیری با طبقه‌بندی کمک عمده‌ای خواهد کرد. مشکلات نمونه‌گیری در بخشهای مختلف یک جامعهٔ بزرگ، به‌صورتی بارز، متفاوت‌اند، که لاجرم نمی‌توان رفتار و سیاستی یکسان در همهٔ جامعه داشت و طبیعتاً تقسیم جامعه به طبقه‌ها امری منطقی است.

۳. با طبقه‌بندی کردن جامعه می‌توان دقت برآورد مجموع جامعه را کنترل کرد. زیرا می‌توان یک جامعهٔ ناهمگن را به زیرجامعه‌های تقریباً همگن تقسیم کرد. منظور از زیرجامعه یا طبقهٔ همگن، طبقه‌ای است که اندازه‌ها از واحدی به واحد دیگر تغییر عمده‌ای ندارند و می‌توان در چنین طبقه‌ای با نمونه‌ای به حجم اندک، برآوردی دقیق از صفت تحت بررسی به‌دست آورد. برآوردهایی که جداگانه از طبقه‌های مختلف فراهم می‌شوند سرانجام ترکیب‌شده و برآوردی دقیق برای صفت مورد نظر درکل جامعه به‌دست می‌آید. چون انتخاب نمونه‌ها از طبقه‌های مختلف مستقل از هم انجام می‌شود واریانس هر برآوردکننده در طبقه‌ها با هم جمع می‌شوند تا واریانس برآوردکننده در کل جامعه با احتساب ضرایبی که بعداً از آن گفتگو می‌کنیم به‌دست آید. لذا اصل طبقه‌بندی مبتنی بر آن چنان افرازی از جامعه است که واریانس برآوردکننده در هر طبقه تا حد ممکن کوچک باشد. برای این هدف، لاجرم باید در هر طبقه واحدهایی همگن قرار گیرند.

در این مقدمه لازم است تذکر دهیم که در بعضی از جامعه‌ها تقسیم جامعه به طبقات را باید با توجه به ویژگیهای قسمتهای مختلف جامعه انجام داد ولی در اکثر جامعه‌ها طبقه‌ها به‌صورت طبیعی و با ساختار جامعه از قبل مشخص شده‌اند. مثالهایی نوعی از جامعه‌هایی که به‌طور طبیعی طبقه‌بندی شده‌اند به‌شرح زیرند:

الف) جامعهٔ دانش‌آموزان مدارس خودبه‌خود به‌وسیلهٔ کلاس‌بندی مدارس و جنس دانش‌آموز به‌صورت طبقه‌ها درمی‌آید. یک طبقه در این جامعه، مثلاً «کلاسِ چهارم دختران» است.

ب) جامعهٔ افراد مالیات‌دهنده، به‌وسیلهٔ شهر، جنس، و دامنهٔ درآمد گزارش‌شده طبقه‌بندی می‌شود. تعیین دامنهٔ درآمدها دلخواه است. یک طبقه در این جامعه، مثلاً «مردان شهر تهران با درآمد سالیانه‌ای در دامنهٔ [دومیلیون-یک‌میلیون] تومان» است.

ج) برای جامعهٔ خانوارها در یک کشور، هر استان کشور یک طبقه است. بدیهی است مرز هر استان باید دقیقاً تعریف شود.

د) جامعهٔ مؤسسات خرده‌فروشی به‌وسیلهٔ نوع جنسی که عرضه می‌شود (ادویه، گوشت، ...) و مکان مؤسسه، خودبه‌خود طبقه‌بندی می‌شود.

بدیهی است بعضی از ویژگیهایی‌که طبقه‌ها را به وجود می‌آورند جنبهٔ کیفی و برخی جنبهٔ

کمی دارند. مثلاً جنس دانش‌آموز، یا نام خرده‌فروشی جنبهٔ کیفی و میزان درآمد جنبهٔ کمّی دارد. معمولاً اگر یک جامعه به‌صورت طبیعی طبقه‌بندی نشده باشد و بخواهیم خود ما به‌کار طبقه‌بندی بپردازیم باید ابتدا نمونه‌ای نسبتاً بزرگ به‌روش تصادفی ساده از جامعه بگیریم و سپس با در نظر گرفتن ناهمگنی و پراکندگی این نمونه، طبقاتی را با کرانهای مناسب انتخاب کنیم، به‌طوری که اگر واحدهای این نمونه را در طبقه‌های ساخته‌شده توزیع کنیم در هر طبقه همگنی موجود باشد. بعداً دربارهٔ تعیین کرانهای طبقات گفتگو خواهیم کرد.

۱.۴ تعریف نمونهٔ تصادفی با طبقه‌بندی

در نمونه‌گیری با طبقه‌بندی، همان‌طور که توضیح دادیم، ابتدا جامعهٔ به حجم N را به L زیرجامعه به حجمهای N_1، N_2، ...، N_h، ...، N_L تقسیم می‌کنیم، به‌قسمی که این زیرجامعه‌ها متداخل نباشند و هر واحد جامعه به یک و تنها به یک زیرجامعه متعلق باشد. درواقع زیرجامعه‌ها افرازی از جامعه را به‌وجود می‌آورند. پس

$$\boxed{N = \sum_{h=1}^{L} N_h} \quad \text{جامعه}$$

هر زیرجامعه را یک طبقه می‌نامیم. این طبقه‌ها همان‌گونه که در بالا اشاره کردیم به‌گونه‌ای تعیین می‌شوند که واحدهای با حداکثر همگنی در هر طبقه قرار بگیرند. برای به‌دست آوردن برآوردهای دقیق باید مقادیر $N_h (h = 1, 2, \ldots, L)$ را بدانیم. وقتی طبقه‌ها و حجم آنها مشخص شد از هر طبقه نمونه‌ای انتخاب می‌شود. این نمونه را می‌توان با هر روشی انتخاب کرد. بعداً که با نمونه‌گیریهای سیستماتیک، خوشه‌ای، براورد رگرسیونی، برآورد نسبتی آشنا شدید، می‌توانید در هر طبقه نمونه را با یکی از این روشها هم به‌دست آورید. انتخاب نمونه در هر طبقه مستقل از سایر طبقه‌هاست. حجمهای نمونه‌ها در طبقه‌ها را به‌ترتیب با n_1، n_2، ...، n_h، ...، n_L نشان می‌دهیم به‌طوری که

$$n = \sum_{h=1}^{L} n_h$$

معرف حجم نمونه از کل جامعه است. اگر نمونه‌ای که از هر طبقه انتخاب می‌شود، به‌روش نمونه‌گیری تصادفی سادهٔ بدون جایگذاری باشد، نمونه‌گیری را نمونه‌گیری تصادفی با طبقه‌بندی می‌نامند و نمونهٔ متشکل از کل نمونه‌های طبقات را نمونهٔ تصادفی با طبقه‌بندی می‌گویند. در این فصل، این نوع نمونه‌گیری را بررسی می‌کنیم.

نظریهٔ نمونه‌گیری با طبقه‌بندی، از ویژگیهای برآوردهای نمونهٔ با طبقه‌بندی و از انتخاب بهترین حجم نمونه در هر طبقه برای تأمین دقت ماکسیمم بحث می‌کند. ابتدا فرض می‌کنیم طبقه‌ها پیشاپیش ساخته شده‌اند، و مطالب مربوط به برآوردها را مطرح می‌کنیم، سپس در گام بعد از چگونگی ساختن طبقه‌ها گفتگو می‌نماییم.

۲.۴ نمادها و برخی از تعریفها

تعداد طبقه‌ها را با L نشان می‌دهیم. زیرنویس h، معرف شمارهٔ طبقه است که از ۱ تا L تغییر می‌کند. زیرنویس i برای تعیین شمارهٔ واحدها در داخل هر طبقه به‌کار می‌رود. N تعداد کل افراد جامعه است. نمادهای زیر برای طبقهٔ h، L ,... ,۲ ,۱ $=h$ تعریف می‌شوند

میانگین جامعه	$\overline{Y}_N$
تعداد کل واحدها در طبقهٔ hام	N_h
تعداد واحدهای نمونهٔ از طبقهٔ hام	n_h
مقدار صفت واحد iام در طبقهٔ hام	$Y_{hi},\quad i=1,2,\ldots,n_h$
نسبت تعداد واحدهای طبقهٔ hام به تعداد واحدهای کل جامعه یا وزن طبقهٔ hام	$W_h=\dfrac{N_h}{N}$
کسر نمونه‌گیری برای طبقهٔ hام	$f_h=\dfrac{n_h}{N_h}$
میانگین طبقهٔ hام	$\overline{\mathcal{Y}}_h=\dfrac{1}{N_h}\sum_{i=1}^{N_h}Y_{hi}$
میانگین نمونهٔ طبقه hام	$\overline{Y}_h=\dfrac{1}{n_h}\sum_{i=1}^{n_h}Y_{hi}$
تغییرات طبقهٔ hام	$S_h^2=\dfrac{1}{N_h-1}\sum_{i=1}^{N_h}(Y_{hi}-\overline{\mathcal{Y}}_h)^2$
تغییرات نمونه‌ای طبقهٔ hام	$s_h^2=\dfrac{1}{n_{h-1}}\sum_{i=1}^{n_h}(Y_{hi}-\overline{Y}_h)^2$

واریانس ← واریانس نمونه

گاهی اوقات S_h^2 و s_h^2 را به‌ترتیب واریانس طبقهٔ h و واریانس نمونهٔ این طبقه می‌نامند. توجه کنید که واحدهای هر طبقه و نمونهٔ آن را یکسان نشان داده‌ایم. خواننده از محتوای مطلب درهر جا متوجه خواهد شد که واحد مورد بحث به طبقه تعلق دارد یا به نمونهٔ طبقه. اگر از L طبقه که به حجمهای N_1، N_2، ... ، N_L، هستند، L نمونه به‌ترتیب به حجمهای n_1، n_2، ... ، n_L به‌روش تصادفی ساده اختیار کنیم، دنباله‌ای با $\sum_{h=1}^{L}n_h=n$ عضو خواهیم داشت که همان نمونهٔ تصادفی با طبقه‌بندی است. میانگین موزون میانگینهای نمونه‌ای طبقه‌ها را میانگین نمونهٔ با طبقه‌بندی می‌نامیم و آن را با $\overline{Y}_{st}$ نمایش می‌دهیم

$$\boxed{\overline{Y}_{st}}=\frac{1}{N}\sum_{h=1}^{L}N_h\overline{Y}_h=\sum_{h=1}^{L}W_h\overline{Y}_h \qquad (۱.۴)$$

نکته: (توجه کنید که این نمونه، با نمونهٔ تصادفی به حجم n از کل جامعه تفاوت دارد.)

قضیهٔ ۱.۴ اگر در طبقهٔ h، $h = ۱, \ldots, L$، $\overline{Y}_h$ برآوردکنندهٔ نااریب میانگین طبقهٔ hام باشد، آن‌گاه $\overline{Y}_{st}$ برآوردکنندهٔ نااریب میانگین کل جامعه، یعنی $\overline{Y}_N$ است.

برهان. با توجه به رابطهٔ (۱.۴)، داریم

صفحه ۱

$$\boxed{E(\overline{Y}_{st})} = E\left[\sum_{h=۱}^{L} W_h \overline{Y}_h\right] = \sum_{h=۱}^{L} W_h E(\overline{Y}_h)$$

اما چون بنابر فرض قضیه، $\overline{Y}_h$ برآوردکنندهٔ نااریب $\overline{\mathcal{Y}}_h$ است، پس

$$E(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h \overline{\mathcal{Y}}_h$$

از طرفی $\overline{\mathcal{Y}}_h = \frac{۱}{N_h} \sum_{i=۱}^{N_h} Y_{hi}$، پس

$$E(\overline{Y}_{st}) = \sum_{h=۱}^{L} \left[\frac{W_h}{N_h} \sum_{i=۱}^{N_h} Y_{hi}\right]$$

اما $W_h = \frac{N_h}{N}$ و لذا

$$E(\overline{Y}_{st}) = \frac{۱}{N} \sum_{h=۱}^{L} \sum_{i=۱}^{N_h} Y_{hi}$$

ضریب $\frac{۱}{N}$ در عبارت طرف دوم برابر مجموع همهٔ واحدهای جامعه است، پس

$$E(\overline{Y}_{st}) = \overline{Y}_N \qquad (۲.۴)$$

لذا $\hat{\overline{Y}}_N = \overline{Y}_{st}$، برآوردکننده‌ای نااریب است. لازم است تذکر دهیم که برای هر نوع نمونه‌گیری در داخل طبقه‌ها، وقتی $\overline{Y}_h$ برآوردکنندهٔ نااریب میانگین طبقه، h، $h = ۱, ۲, \ldots, L$ باشد، حکم قضیهٔ بالا صحیح است. ◻

فرع. اگر T_N مجموع واحدهای جامعه باشد، آن‌گاه $N\overline{Y}_{st}$ برآوردکنندهٔ نااریب T_N است، زیرا از رابطهٔ $E(\overline{Y}_{st}) = \overline{Y}_N$ نتیجه می‌شود که $E(N\overline{Y}_{st}) = N\overline{Y}_N$ و چون $T_N = N\overline{Y}_N$، پس $E(N\overline{Y}_{st}) = T_N$، یعنی $N\overline{Y}_{st}$ برآوردکنندهٔ نااریب T_N است. نکته

قضیهٔ ۲.۴ اگر نمونه‌های طبقه‌ها مستقل از هم باشند

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲ \mathrm{V}(\overline{Y}_h) \qquad (۳.۴)$$

برهان. $\overline{Y}_{st}$ تابعی خطی از $\overline{Y}_h$ها با ضریبهای ثابت W_h، $h = ۱, \ldots, L$ است، لذا

$$\mathrm{V}(\overline{Y}_{st}) = \mathrm{V}(W_۱\overline{Y}_۱ + W_۲\overline{Y}_۲ + \cdots + W_L\overline{Y}_L)$$

چون نمونه‌ها در طبقه‌ها مستقل از یکدیگرند، $\overline{Y}_h$ها از هم مستقل‌اند و بنابراین

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} \mathrm{V}(W_h\overline{Y}_h) = \sum_{h=۱}^{L} W_h^۲\mathrm{V}(\overline{Y}_h)$$

□

رابطهٔ بالا نشان می‌دهد که $\mathrm{V}(\overline{Y}_{st})$ به $\overline{Y}_h$ها و وزنها بستگی دارد. اگر واحدهای هر طبقه همگن باشند واریانس طبقه‌ها کوچک هستند. به‌عبارت دیگر اگر بتوان جامعهٔ با پراکندگی زیاد را به طبقاتی تقسیم کنیم که واحدهای هر طبقه نزدیک به‌هم باشند $\overline{Y}_N$ تقریباً با خطای اندک برآورد می‌شود. درواقع استفاده از وزنهای طبقاتی $\frac{N_h}{N}$ است که موجب می‌شود $\overline{Y}_{st}$ تقریباً با خطای اندک $\overline{Y}_N$ را برآورد کند. رابطهٔ (۳.۴) به‌نوع نمونه‌گیری طبقات بستگی ندارد.

قضیهٔ ۳.۴ در نمونه‌گیری تصادفی با طبقه‌بندی، واریانس $\overline{Y}_{st}$ به‌صورت زیر است

$$\begin{aligned}\mathrm{V}(\overline{Y}_{st}) &= \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{S_h^۲}{n_h} = \sum_{h=۱}^{L} W_h^۲(۱ - f_h)\frac{S_h^۲}{n_h} \\ &= \sum_{h=۱}^{L}\frac{W_h^۲S_h^۲}{n_h} - \sum_{h=۱}^{L}\frac{W_hS_h^۲}{N} \qquad (۴.۴)\end{aligned}$$

برهان. چون در طبقهٔ h، $h = ۱, \ldots, L$، نمونه‌گیری به‌روش تصادفی سادهٔ بدون جایگذاری انجام می‌شود، بنابراین

$$\mathrm{V}(\overline{Y}_h) = \left(\frac{۱}{n_h} - \frac{۱}{N_h}\right) S_h^۲ = \frac{N_h - n_h}{N_h}\cdot\frac{S_h^۲}{n_h}$$

اینک با توجه به (۳.۴)

$$\begin{aligned}\mathrm{V}(\overline{Y}_{st}) &= \sum_{h=۱}^{L} W_h^۲\mathrm{V}(\overline{Y}_h) = \sum_{h=۱}^{L} W_h^۲\frac{N_h - n_h}{N_h}\frac{S_h^۲}{n_h} \\ &= \sum_{h=۱}^{L}\frac{N_h^۲}{N^۲}\cdot\frac{۱}{N_h}(N_h - n_h)\frac{S_h^۲}{n_h} \\ &= \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{S_h^۲}{n_h}\end{aligned}$$

که درستی برابری اول را در حکم قضیه نشان می‌دهد. از طرفی

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲ \frac{N_h - n_h}{N_h} \cdot \frac{S_h^۲}{n_h} = \sum_{h=۱}^{L} W_h^۲ \left(۱ - \frac{n_h}{N_h}\right) \frac{S_h^۲}{n_h}$$

چون $\frac{n_h}{N_h} = f_h$، پس

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲ (۱ - f_h) \frac{S_h^۲}{n_h}$$

که درستی برابری دوم را در حکم قضیه نشان می‌دهد. (درستی آخرین برابری را در حکم قضیه تحقیق کنید.)

فرع ۱ . اگر در همهٔ طبقه‌ها f_hها کوچک و قابل اغماض باشند، آن‌گاه

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲ \frac{S_h^۲}{n_h} = \frac{۱}{N^۲} \sum_{h=۱}^{L} N_h^۲ \frac{S_h^۲}{n_h} \qquad (۵.۴)$$

فرع ۲ . اگر T_N مجموع واحدهای جامعه باشد، آن‌گاه همان‌طور که دیدیم $N\overline{Y}_{st}$ برآوردکننده نااریب T_N است و

$$\mathrm{V}(\hat{T}_N) = \mathrm{V}(N\overline{Y}_{st}) = N^۲ \mathrm{V}(\overline{Y}_{st}) = N^۲ \cdot \frac{۱}{N^۲} \sum_{h=۱}^{L} N_h (N_h - n_h) \frac{S_h^۲}{n_h}$$

$$= \sum_{h=۱}^{L} N_h (N_h - n_h) \frac{S_h^۲}{n_h} \qquad (۶.۴)$$

□

۳.۴ نمونه‌گیری با طبقه‌بندی و با تخصیص متناسب

اگر در نمونه‌گیری با طبقه‌بندی حجم نمونه‌های طبقه‌ها با حجم طبقه‌ها متناسب باشند، نمونه‌گیری را با تخصیص متناسب می‌نامند. درواقع در این نوع نمونه‌گیری

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \qquad h = ۱, \ldots, L$$

یا

$$\frac{n_h}{N_h} = \frac{n}{N} \qquad h = ۱, \ldots, L$$

که نتیجه می‌دهد به‌ازای L , … ,۲ ,$۱ = h$، $f_h = f$. به‌عبارت دیگر در این نوع نمونه‌گیری با طبقه‌بندی، کسر نمونه‌گیری در همهٔ طبقه‌ها یکسان و برابر با $\frac{n}{N}$ است.

قضیهٔ ۴.۴ در نمونه‌گیری با طبقه‌بندی و با تخصیص متناسب

$$\mathrm{V}(\overline{Y}_{st}) = \frac{۱ - f}{n} \sum_{n=۱}^{L} W_h S_h^۲ \qquad (۷.۴)$$

برهان. با توجه به (۴.۴)، رابطهٔ کلی

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲ (۱ - f_h) \frac{S_h^۲}{n_h}$$

را در نظر می‌گیریم. چون نمونه‌گیری با تخصیص متناسب است، همهٔ f_hها برابر f هستند، پس

$$\mathrm{V}(\overline{Y}_{st}) = \sum_{h=۱}^{L} \frac{N_h}{N} \cdot W_h (۱ - f) \frac{S_h^۲}{n_h}$$

چون، $\frac{N_h}{N} = \frac{n_h}{n}$، داریم

$$\mathrm{V}(\overline{Y}_{st}) = (۱ - f) \sum_{h=۱}^{L} W_h \cdot \frac{n_h}{n} \cdot \frac{S_h^۲}{n_h}$$
$$= \frac{۱ - f}{n} \sum_{h=۱}^{L} W_h S_h^۲$$

□

فرع. وقتی نمونه‌گیری با تخصیص متناسب بوده و مقدار واریانس در همهٔ طبقه‌ها یکسان باشد، آن‌گاه اگر این واریانس مشترک را با $S_w^۲$ نشان دهیم همهٔ $S_h^۲$ها برابر با $S_w^۲$ هستند. پس برابری (۷.۴) به‌صورت زیر درمی‌آید

$$\mathrm{V}(\overline{Y}_{st}) = \frac{۱ - f}{n} \sum_{h=۱}^{L} W_h S_h^۲ = \frac{۱ - f}{n} S_w^۲ \sum_{h=۱}^{L} W_h = (۱ - f) \frac{S_w^۲}{n} \qquad (۸.۴)$$

که در آن، f، کسر نمونه‌گیری در کل جامعه است.

مثال ۱.۴ جدول زیر معرف تعداد ساکنین ۶۴ شهر بزرگ در سال ۱۹۳۰ برحسب هزار نفر است. در این جدول، جمعیت ۶۴ شهر در سال ۱۹۳۰ و در دو طبقه ثبت شده است. طبقهٔ اول شامل ۱۶ شهر با جمعیت بیشتر و طبقهٔ دوم شامل ۴۸ شهر با جمعیت کمتر است.

جمعیت ۶۴ شهر (برحسب ۱۰۰۰) در سال ۱۹۳۰*

Y_{hi}ها

h:	۱	۲		
	۹۰۰	۳۶۴	۲۰۹	۱۱۳
	۸۲۲	۳۱۷	۱۸۳	۱۱۵
	۷۸۱	۳۲۸	۱۶۳	۱۲۳
	۸۰۵	۳۰۲	۲۵۳	۱۵۴
	۶۷۰	۲۸۸	۲۳۲	۱۴۰
	۱۲۳۸	۲۹۱	۲۶۰	۱۱۹
	۵۷۳	۲۵۳	۲۰۱	۱۳۰
	۶۳۴	۲۹۱	۱۴۷	۱۲۷
	۵۷۸	۳۰۸	۳۹۲	۱۰۰
	۴۸۷	۲۷۲	۱۶۴	۱۰۷
	۴۴۲	۲۸۴	۱۴۳	۱۱۴
	۴۵۱	۲۵۵	۱۶۹	۱۱۱
	۴۵۹	۲۷۰	۱۳۹	۱۶۳
	۴۶۴	۲۱۴	۱۷۰	۱۱۶
	۴۰۰	۱۹۵	۱۵۰	۱۲۲
	۳۶۶	۲۶۰	۱۴۳	۱۳۴

* داده‌ها از تکنیکهای نمونه‌گیری تألیف کوکران اقتباس شده‌اند.

قرار است کل جمعیت ۶۴ شهر در سال ۱۹۳۰ را از روی نمونه‌ای به حجم ۲۴ برآورد کنند. انحراف معیار برآوردکنندهٔ کل جمعیت را (۱) به‌وسیلهٔ نمونهٔ تصادفی ساده، (۲) به‌وسیلهٔ نمونهٔ تصادفی با طبقه‌بندی و با تخصیص متناسب، (۳) به‌وسیلهٔ نمونهٔ تصادفی با طبقه‌بندی و استخراج ۱۲ واحد از هر طبقه بیابید.

این جامعه به بسیاری از جامعه‌های صنعتی شباهت دارد که بعضی واحدها، مثلاً شهرهای بزرگ، سهمی عمده از کل را دارند و در تغییرپذیری نقشی مهمتر از بقیه واحدها بازی می‌کنند. در جدول زیر، جدول داده‌ها، مجموعهای طبقاتی، و مجموعهای مربعات ارائه شده‌اند. در این مثال فقط از داده‌های سال ۱۹۳۰ استفاده کرده‌ایم.

مجموع و مجموع مربعات

طبقه	$\sum Y_{hi}$	$\sum(Y_{hi}^{۲})$
۱	۱۰۰۷۰	۷۱۴۵۴۵۰
۲	۹۴۹۸	۲۱۴۱۷۲۰

برای کل جامعه در سال ۱۹۳۰، به‌دست می‌آوریم

$$T_N = ۱۹۵۶۸ \qquad S^۲ = ۵۲۴۴۸$$

برای ۳ موردی که در صورت مثال مطرح شده‌اند برآورد کل جامه را با نمادهای $\hat{T}_{ran}$، $\hat{T}_{prop}$ و $\hat{T}_{equal}$ نشان می‌دهیم. طبق فرض ۱۶ $= N_۱$، ۴۸ $= N_۲$.

۱. برای نمونه‌گیری تصادفی ساده

$$V(\hat{T}_{ran}) = N^۲\left(۱ - \frac{n}{N}\right)\frac{S^۲}{n} = (۶۴)^۲\left(۱ - \frac{۲۴}{۶۴}\right)\frac{۵۲۴۴۸}{۲۴} = ۵۵۹۴۴۵۳$$

و لذا $\sigma(\hat{T}_{ran}) = ۲۳۶۵$.

۲. برای دو طبقه به‌ترتیب

$$S_۱^۲ = ۵۳۸۴۳ \qquad S_۲^۲ = ۵۵۸۱$$

توجه کنید واریانس مربوط به طبقهٔ ۱ که شامل شهرهای پرجمعیت است تقریباً ۱۰ برابر واریانس طبقهٔ ۲ است.

در تخصیص متناسب، داریم ۶ $= n_۱$ و ۱۸ $= n_۲$. بنابر (۷.۴)

$$V(\overline{Y}_{st}) = \frac{۱ - f}{n}\sum_{h=۱}^{L} W_h S_h^۲ = \frac{N - n}{nN^۲}\sum_{h=۱}^{L} N_h S_h^۲$$

از ضرب طرفین در $N^۲$، داریم

$$V(N\overline{Y}_{st}) = V(\hat{T}_{prop}) = \frac{N - n}{n}\sum_{h=۱}^{L} N_h S_h^۲$$
$$= \frac{۶۴ - ۲۴}{۲۴}[(۱۶)(۵۳۸۴۳) + (۴۸)(۵۵۸۱)] = ۱۸۸۲۲۹۳$$

و لذا

$$\sigma(\hat{T}_{prop}) = ۱۳۷۲$$

۳. برای ۱۲ $= n_۱ = n_۲$، فرمول کلی

$$V(\overline{Y}_{st}) = \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{S_h^۲}{n_h}$$

را به‌کار می‌بریم. از ضرب طرفین در N^2، داریم

$$V(N\bar{Y}_{st}) = V(\hat{T}_{equal}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

پس

$$V(\hat{T}_{equal}) = \frac{16(16-12)(53843)}{12} + \frac{(28)(28-12)(5581)}{12}$$
$$= 1090827$$

و $\sigma(\hat{T}_{equal}) = 1044$.

در این مثال، نمونه‌گیری با حجمهای نمونه‌ای برابر در دو طبقه دقیقتر از تخصیص متناسب است، ولی هر دو از نمونه‌گیری تصادفی ساده بهترند. ▲

تبصره. احتمال اینکه واحدی متعلق به طبقهٔ h به‌عنوان عضوی از نمونهٔ با طبقه‌بندی انتخاب شود برابر با $\frac{n_h}{N_h}$ است. احتمال اینکه دو واحد متعلق به‌طبقهٔ h به‌عنوان اعضای نمونهٔ با طبقه‌بندی انتخاب شوند $n_h(n_h-1)/N_h(N_h-1)$ است. اگر دو واحد به دو طبقهٔ مختلف مثلاً طبقه‌های h و K متعلق باشند احتمال اینکه به‌عنوان اعضای نمونه با طبقه‌بندی انتخاب شوند برابر با $n_h \cdot n_K/N_h N_K$ است. تمام این احتمالها با احتمال انتخاب هر واحد در نمونه‌گیری تصادفی ساده متفاوت‌اند.

۴.۴ برآورد واریانس برآوردکنندهٔ میانگین در نمونه‌گیری با طبقه‌بندی

در نمونه‌گیری تصادفی سادهٔ بدون جایگذاری دیدیم که s^2ی نمونه برآوردکنندهٔ ناریب S^2ی جامعه است. لذا s_h^2، یعنی واریانس نمونه‌ای در طبقهٔ hام برآوردکنندهٔ ناریب S_h^2، یعنی واریانس طبقه hام است. با توجه به این مطلب قضیهٔ زیر را بیان می‌کنیم.

قضیهٔ ۵.۲ در نمونه‌گیری تصادفی سادهٔ با طبقه‌بندی، برآوردکنندهٔ واریانس برای برآوردکنندهٔ میانگین جامعه برابر است با

$$\hat{V}(\bar{Y}_{st}) = \hat{\sigma}^2(\bar{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} \qquad (9.4)$$

برهان. بنابر (۲.۴) داریم

$$V(\bar{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

با توجه به برابری

$$E(s_h^۲) = S_h^۲ \qquad h = ۱,\ ۲, \ldots, L$$

رابطهٔ بالا به‌صورت زیر درمی‌آید

$$\mathrm{V}(\overline{Y}_{st}) = \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{E(s_h^۲)}{n_h}$$

چون ، برای مقدار ثابت a، $E(aX) = aE(X)$ که در آن X، متغیری تصادفی است، پس رابطهٔ بالا را می‌توان چنین نوشت

$$\mathrm{V}(\overline{Y}_{st}) = E\left[\frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{s_h^۲}{n_h}\right]$$

این برابری، معرف آن است که عبارت داخل کروشه برآوردکنندهٔ ناااریب $\mathrm{V}(\overline{Y}_{st})$ است، یعنی

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h(N_h - n_h)\frac{s_h^۲}{n_h}$$

و یا

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \sum_{h=۱}^{L} W_h^۲(۱ - f_h)\frac{s_h^۲}{n_h} \qquad (۱۰.۴)$$

در عمل، با داشتن تنها یک نمونه در هر طبقه، $s_h^۲$ها به‌صورت مقادیری عددی هستند و لذا برآوردی نااریب، برای واریانس $\overline{Y}_{st}$ به‌دست می‌آید. صورت دیگر (۱۰.۴)، با توجه به آخرین برابری (۴.۴) به‌صورت زیر است

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \sum_{h=۱}^{L}\frac{W_h^۲ s_h^۲}{n_h} - \sum_{h=۱}^{L}\frac{W_h s_h^۲}{N} \qquad (۱۱.۴)$$

برای محاسبهٔ $\hat{\mathrm{V}}(\overline{Y}_{st})$ باید از هر طبقه حداقل دو واحد انتخاب شود، زیرا در غیر این صورت محاسبهٔ $s_h^۲$ به‌دلیل صفر شدن $n_h - ۱$ در رابطهٔ

$$s_h^۲ = \frac{۱}{n_h - ۱}\sum_{i=۱}^{n_h}(Y_{hi} - \overline{Y}_h)^۲$$

میسر نیست. برای حالت نمونه‌گیری با طبقه‌بندی و با تخصیص متناسب، از (۷.۴) داریم

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \frac{۱ - f}{n}\sum_{h=۱}^{L} W_h s_h^۲$$

که برآوردکننده‌ای نااریب است. □

۵.۴ حدود اطمینان میانگین و مجموع واحدهای جامعه در نمونه‌گیری با طبقه‌بندی

برای یک نمونهٔ تصادفی سادهٔ با طبقه‌بندی به حجم n از جامعه‌ای به حجم N، تعداد نمونه‌های ممکن $\prod_{h=1}^{L}\binom{N_h}{n_h}$ است. مثلاً اگر ۴ طبقه به حجمهای ۱۰، ۸، ۹، ۱۲ داشته باشیم و از این طبقه‌ها به‌ترتیب ۴، ۲، ۳، ۵ واحد به‌تصادف انتخاب کنیم تعداد نمونه‌های ممکن برابر است با

$$\binom{۱۰}{۴}\binom{۸}{۲}\binom{۹}{۳}\binom{۱۲}{۵} = (۱۲۰)(۲۸)(۸۴)(۷۹۲) = ۳۹۱۱۸۴۶۴$$

بنابراین در حالت کلی حداکثر به‌همین تعداد هم $\overline{Y}_{st}$ خواهیم داشت. اگر تعداد واحدهایی که $\overline{Y}_{st}$ را تولید می‌کنند یعنی n بزرگ باشد، مثلاً بیش از ۲۰، فرض نرمال بودن توزیع $\overline{Y}_{st}$ را به‌شرط متناهی بودن واریانسهای طبقات، و با توجه به قضیهٔ حد مرکزی می‌توان پذیرفت. با توجه به این مقدمه، برای تهیهٔ بازهٔ اطمینان، ابتدا توزیع $\overline{Y}_{st}$ را نرمال می‌گیریم. لذا متغیر تصادفی $Z = \frac{\overline{Y}_{st}-\overline{Y}_N}{\sigma(\overline{Y}_{st})}$ دارای توزیع نرمال استاندارد است. اگر z مقدار متغیر نرمال استاندارد متناظر با احتمال تجمعی $\frac{\alpha}{۲} - ۱$ باشد، بازهٔ اطمینان $(\alpha - ۱)۱۰۰$ درصد $\overline{Y}_N$ به‌صورت زیر است

$$[\overline{Y}_{st} - z\sigma(\overline{Y}_{st}), \quad \overline{Y}_{st} + z\sigma(\overline{Y}_{st})] \qquad (۱۲.۴)$$

این بازه، بازه‌ای تصادفی است زیرا $\overline{Y}_{st}$ تصادفی است. برای یک نمونهٔ مشخص، این بازه ثابت است. معمولاً $\sigma(\overline{Y}_{st})$ مجهول است و باید به‌جای آن، $\hat{\sigma}(\overline{Y}_{st})$ را از رابطهٔ (۹.۴) منظور کرد. لذا برآورد بازهٔ اطمینان به‌صورت زیر است

$$[\overline{Y}_{st} - z\hat{\sigma}(\overline{Y}_{st}), \quad \overline{Y}_{st} + z\hat{\sigma}(\overline{Y}_{st})] \qquad (۱۳.۴)$$

اگر جمله‌های بازه‌های (۱۲.۴) (۱۳.۴) را در N ضرب کنیم، بازهٔ اطمینان و برآورد بازهٔ اطمینان برای مجموع واحدهای جامعه، یعنی T_N به‌دست می‌آید. برآورد بازهٔ اطمینان برای T_N با ضریب اطمینان $\alpha - ۱$ به‌صورت زیر است

$$[N\overline{Y}_{st} - zN\hat{\sigma}(\overline{Y}_{st}), \quad N\overline{Y}_{st} + zN\hat{\sigma}(\overline{Y}_{st})]$$

همان‌طور که اشاره شد می‌پذیریم که $\overline{Y}_{st}$ دارای توزیع نرمال است و z را از جدول توزیع نرمال استاندارد به‌دست می‌آوریم. اگر تعداد واحدهایی که از طبقه‌ها انتخاب می‌کنیم کم باشد باید مقدار z را به‌جای توزیع نرمال از توزیع t به‌دست آوریم. در این بازهٔ اطمینان $\overline{Y}_{st}$ و $\hat{\sigma}(\overline{Y}_{st})$ تصادفی هستند. توزیع $\hat{\sigma}(\overline{Y}_{st})$ در حالت کلی بسیار پیچیده است، ولی یک روش تقریبی برای تخصیص

با توجه به برابری

$$E(s_h^2) = S_h^2 \qquad h = 1, 2, \ldots, L$$

رابطهٔ بالا به‌صورت زیر درمی‌آید

$$\mathrm{V}(\overline{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{E(s_h^2)}{n_h}$$

چون ، برای مقدار ثابت a، $E(aX) = aE(X)$ که در آن X، متغیری تصادفی است، پس رابطهٔ بالا را می‌توان چنین نوشت

$$\mathrm{V}(\overline{Y}_{st}) = E\left[\frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}\right]$$

این برابری، معرف آن است که عبارت داخل کروشه برآوردکنندهٔ نااریب $\mathrm{V}(\overline{Y}_{st})$ است، یعنی

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}$$

و یا

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \sum_{h=1}^{L} W_h^2(1 - f_h)\frac{s_h^2}{n_h} \qquad (۱۰.۴)$$

در عمل، با داشتن تنها یک نمونه در هر طبقه، s_h^2ها به‌صورت مقادیری عددی هستند و لذا برآوردی نااریب، برای واریانس $\overline{Y}_{st}$ به‌دست می‌آید. صورت دیگر (۱۰.۴)، با توجه به آخرین برابری (۴.۴) به‌صورت زیر است

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L}\frac{W_h s_h^2}{N} \qquad (۱۱.۴)$$

برای محاسبهٔ $\hat{\mathrm{V}}(\overline{Y}_{st})$ باید از هر طبقه حداقل دو واحد انتخاب شود، زیرا در غیر این صورت محاسبهٔ s_h^2 به‌دلیل صفر شدن $n_h - 1$ در رابطهٔ

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(Y_{hi} - \overline{Y}_h)^2$$

میسر نیست. برای حالت نمونه‌گیری با طبقه‌بندی و با تخصیص متناسب، از (۷.۴) داریم

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \frac{1 - f}{n}\sum_{h=1}^{L} W_h s_h^2$$

که برآوردکننده‌ای نااریب است. □

۵.۴ حدود اطمینان میانگین و مجموع واحدهای جامعه در نمونه‌گیری با طبقه‌بندی

برای یک نمونهٔ تصادفی سادهٔ با طبقه‌بندی به حجم n از جامعه‌ای به حجم N، تعداد نمونه‌های ممکن $\prod_{h=1}^{L}\binom{N_h}{n_h}$ است. مثلاً اگر ۴ طبقه به حجمهای ۱۰، ۸، ۹، ۱۲ داشته باشیم و از این طبقه‌ها به‌ترتیب ۴، ۲، ۳، ۵ واحد به‌تصادف انتخاب کنیم تعداد نمونه‌های ممکن برابر است با

$$\binom{۱۰}{۴}\binom{۸}{۲}\binom{۹}{۳}\binom{۱۲}{۵} = (۱۲۰)(۲۸)(۸۴)(۷۹۲) = ۳۹۱۱۸۴۶۴$$

بنابراین در حالت کلی حداکثر به‌همین تعداد هم $\overline{Y}_{st}$ خواهیم داشت. اگر تعداد واحدهایی که $\overline{Y}_{st}$ را تولید می‌کنند یعنی n بزرگ باشد، مثلاً بیش از ۲۰، فرض نرمال بودن توزیع $\overline{Y}_{st}$ را به‌شرط متناهی بودن واریانسهای طبقات، و با توجه به قضیهٔ حد مرکزی می‌توان پذیرفت. با توجه به این مقدمه، برای تهیهٔ بازهٔ اطمینان، ابتدا توزیع $\overline{Y}_{st}$ را نرمال می‌گیریم. لذا متغیر تصادفی $Z = \frac{\overline{Y}_{st}-\overline{Y}_N}{\sigma(\overline{Y}_{st})}$ دارای توزیع نرمال استاندارد است. اگر z مقدار متغیر نرمال استاندارد متناظر با احتمال تجمعی $۱-\frac{\alpha}{۲}$ باشد، بازهٔ اطمینان $۱۰۰(۱-\alpha)$ درصد $\overline{Y}_N$ به‌صورت زیر است

$$[\overline{Y}_{st} - z\sigma(\overline{Y}_{st}), \quad \overline{Y}_{st} + z\sigma(\overline{Y}_{st})] \tag{۱۲.۴}$$

این بازه، بازه‌ای تصادفی است زیرا $\overline{Y}_{st}$ تصادفی است. برای یک نمونهٔ مشخص، این بازه ثابت است. معمولاً $\sigma(\overline{Y}_{st})$ مجهول است و باید به‌جای آن، $\hat{\sigma}(\overline{Y}_{st})$ را از رابطهٔ (۹.۴) منظور کرد. لذا برآورد بازهٔ اطمینان به‌صورت زیر است

$$[\overline{Y}_{st} - z\hat{\sigma}(\overline{Y}_{st}), \quad \overline{Y}_{st} + z\hat{\sigma}(\overline{Y}_{st})] \tag{۱۳.۴}$$

اگر جمله‌های بازه‌های (۱۲.۴) (۱۳.۴) را در N ضرب کنیم، بازهٔ اطمینان و برآورد بازهٔ اطمینان برای مجموع واحدهای جامعه، یعنی T_N به‌دست می‌آید. برآورد بازهٔ اطمینان برای T_N با ضریب اطمینان $۱-\alpha$ به‌صورت زیر است

$$[N\overline{Y}_{st} - zN\hat{\sigma}(\overline{Y}_{st}), \quad N\overline{Y}_{st} + zN\hat{\sigma}(\overline{Y}_{st})]$$

همان‌طور که اشاره شد می‌پذیریم که $\overline{Y}_{st}$ دارای توزیع نرمال است و z را از جدول توزیع نرمال استاندارد به‌دست می‌آوریم. اگر تعداد واحدهایی که از طبقه‌ها انتخاب می‌کنیم کم باشد باید مقدار z را به‌جای توزیع نرمال از توزیع t به‌دست آوریم. در این بازهٔ اطمینان $\overline{Y}_{st}$ و $\hat{\sigma}(\overline{Y}_{st})$ تصادفی هستند. توزیع $\hat{\sigma}(\overline{Y}_{st})$ در حالت کلی بسیار پیچیده است، ولی یک روش تقریبی برای تخصیص

درجهٔ آزادی، وقتی از توزیع t استفاده می‌کنیم، تخصیص ساترتوایت[1] است. می‌توانیم $\hat{\sigma}^2(\overline{Y}_{st})$ را چنین بنویسیم

$$\hat{\sigma}^2(\overline{Y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} g_h s_h^2 \qquad (۱۴.۴)$$

که در آن

$$g_h = \frac{N_h(N_h - n_h)}{n_h}$$

درجهٔ آزادی مؤثر را با n_e نشان می‌دهیم که به‌صورت زیر محاسبه می‌شود

$$n_e = \frac{\left(\sum_{h=1}^{L} g_h s_h^2\right)^2}{\sum_{h=1}^{L} \frac{g_h^2 s_h^4}{n_h - 1}} \qquad (۱۵.۴)$$

مقدار n_e همیشه بین کوچکترین مقدار از مقادیر $n_h - 1$، $h = 1, 2, \ldots, L$ و مجموع این مقادیر واقع می‌شود. اگر توزیع جامعه دارای چاولگی مثبت باشد، فرمول (۱۵.۴) درجهٔ آزادی را بیشتر برآورد می‌کند.

۶.۴ تخصیص اپتیمم

در نمونه‌گیری تصادفی با طبقه‌بندی، معمولاً اگر هزینهٔ انتخاب هر واحد نمونه در همهٔ طبقه‌ها یکسان باشد، نمونه‌گیر از طبقه‌های با حجم بیشتر نمونه‌ای با حجم بزرگتر انتخاب می‌کند، ولی اگر هزینهٔ انتخاب هر واحد متفاوت باشد برای انتخاب حجم نمونه فقط حجم طبقه نمی‌تواند ملاک انتخاب حجم نمونه باشد. در این صورت خط‌مشی نمونه‌گیر باید بر این دو نکته متکی باشد که n_h را طوری انتخاب کند که برای بودجه‌ای معین، $V(\overline{Y}_{st})$ مینیمم شود و یا برای مقدار از پیش تعیین‌شدهٔ $V(\overline{Y}_{st})$، هزینهٔ نمونه‌گیری در کل مینیمم باشد.

اگر در طبقه hام هزینهٔ هر واحد نمونه‌گیری C_h فرض شود، وقتی حجم نمونه در این طبقه n_h باشد، هزینهٔ نمونه‌گیری از این طبقه برابر با $C_h n_h$ است و لذا هزینهٔ نمونه‌گیری از L طبقه برابر با $\sum_{h=1}^{L} C_h n_h$ خواهد شد. اگر هزینهٔ رفت و آمد بین طبقات و هزینه‌های اداری و غیره را C_0 فرض کنیم، هزینهٔ کل نمونه‌گیری، C، با معادلهٔ زیر بیان می‌شود

$$C = C_0 + \sum_{h=1}^{L} C_h n_h \qquad (۱۶.۴)$$

1. Satterthwaite

اگر در حالتی خاص هزینه‌های اداری وجود نداشته باشند و $C_\circ$ فقط همان هزینهٔ رفت و آمد بین طبقه‌ها باشد نشان داده‌اند که می‌توان $C_\circ$ را برابر با $\sum t_h\sqrt{n_h}$ گرفت، که در آن t_h هزینهٔ رفت و آمد به‌ازای هر واحد است. ما در این کتاب تابع هزینه را به صورت تابع هزینه (۱۶.۴) در نظر گرفته‌ایم که برحسب حجم نمونه، معادله‌ای خطی است.

قضیهٔ ۶.۴ در نمونه‌گیری تصادفی با طبقه‌بندی که تابع هزینه به‌صورت (۱۶.۴) است، برآورد واریانس $\overline{Y}_{st}$ برای مقدار مشخص C و یا هزینهٔ C برای واریانس مشخص V، وقتی مینیمم می‌شود که n_h متناسب با $W_hS_h/\sqrt{C_h}$ باشد.

برهان. داریم

$$C = C_\circ + \sum_{h=1}^{L} C_h n_h$$

از طرفی اگر $\mathrm{V}(\overline{Y}_{st})$ را به قصد خلاصه‌نویسی فقط با V نشان دهیم، بنابر (۲.۴)، داریم

$$\mathrm{V} = \mathrm{V}(\overline{Y}_{st}) = \sum_{h=1}^{L}\frac{W_h^2S_h^2}{n_h} - \sum_{h=1}^{L}\frac{W_hS_h^2}{N} \tag{I}$$

آنچه باید مورد بحث قرار دهیم عبارت‌اند از:

(۱) مقادیر n_h را طوری انتخاب کنیم که وقتی C مشخص است مقدار V مینیمم باشد.

(۲) مقادیر n_h را طوری انتخاب کنیم که وقتی V مشخص است مقدار C مینیمم باشد.

در واقع مسأله عبارت از انتخاب n_hها برای مینیمم کردن V به‌ازای Cی ثابت یا مینیمم کردن C به‌ازای Vی ثابت است. از رابطهٔ I داریم

$$\mathrm{V} + \sum_{h=1}^{L}\frac{W_hS_h^2}{N} = \sum_{h=1}^{L}\frac{W_h^2S_h^2}{n_h} \tag{II}$$

جملهٔ دوم طرف اول (II) ثابت است زیرا فقط به پارامترهای جامعه بستگی دارد.

از طرفی، تابع هزینه را به‌صورت زیر می‌نویسم

$$C - C_\circ = \sum_{h=1}^{L} C_h n_h \tag{III}$$

اگر قرار دهیم $\mathrm{V}' = \mathrm{V} + \sum_{h=1}^{L}\frac{W_hS_h^2}{N}$ و $C' = C - C_\circ$، از ضرب طرفین دو برابری II و III داریم

$$C'\mathrm{V}' = \left(\sum_{h=1}^{L}\frac{W_h^2S_h^2}{n_h}\right)\left(\sum_{h=1}^{L}C_hn_h\right) \tag{IV}$$

وقتی V ثابت باشد V' هم ثابت است و باید مقادیر n_h را طوری تعیین کنیم که C و در نتیجه C' مینیمم شود. وقتی C ثابت باشد C' هم ثابت است و باید V و در نتیجه V' مینیمم شود. پس هر دو قسمت مسأله هم‌ارز این هستند که در رابطهٔ (IV) وقتی V' ثابت است C' را مینیمم کنیم و وقتی C' ثابت است V' را مینیمم کنیم. لذا در هر دو حالت باید مقادیر n_h را به‌قسمی بیابیم که طرف دوم (IV) مینیمم شود. برای انجام این کار یادآور می‌شویم که از اتحاد جبری

$$\left(\sum a_i^2\right)\left(\sum b_i^2\right)-\left(\sum a_ib_i\right)^2=\sum_{\substack{i \\ i<j}}\sum_{j}(a_ib_j-a_jb_i)^2$$

که در آن a_i و b_i مثبت‌اند و حدود $\sum$ برای تمام جملات یکی است، به‌سهولت نابرابری زیر که به نابرابری کوشی-شوارتس موسوم است نتیجه می‌شود

$$\left(\sum a_i^2\right)\left(\sum b_i^2\right)\geq\left(\sum a_ib_i\right)^2$$

برابری وقتی برقرار می‌شود که به‌ازای هر مقدار i داشته باشیم $\frac{b_i}{a_i}=C^{te}$، زیرا باید طرف دوم اتحاد که مجموع چندین مربع است برابر با صفر باشد، و ناچار هر یک از جملات باید صفر باشد که به تناسب

$$\frac{b_j}{a_j}=\frac{b_i}{a_i}\,,\qquad j\neq i$$

منجر می‌شود که همان شرط $\frac{b_i}{a_i}=C^{te}$ را به‌دست می‌دهد. حال اگر در نابرابری کوشی-شوارتس قرار دهیم $a_i=\frac{W_hS_h}{\sqrt{n_h}}$ و $b_i=\sqrt{C_hn_h}$، داریم

$$\left(\sum\frac{W_h^2S_h^2}{n_h}\right)\left(\sum C_hn_h\right)\geq\left(\sum W_hS_h\sqrt{C_h}\right)^2$$

اما طرف اول این نابرابری با $C'V'$ برابر است و وقتی مینیمم می‌شود که با طرف دوم برابر باشد و آن هم وقتی رخ می‌دهد که داشته باشیم

$$\frac{b_h}{a_h}=\frac{\sqrt{C_hn_h}}{\frac{W_hS_h}{\sqrt{n_h}}}=\frac{n_h\sqrt{C_h}}{W_hS_h}=\mathrm{c}^{\mathrm{te}}=K$$

یعنی

$$n_h=\frac{W_hS_h}{\sqrt{C_h}}\cdot K$$

و در نتیجه

$$n=\sum_{h=1}^{L}n_h=\sum_{h=1}^{L}\frac{W_hS_h}{\sqrt{C_h}}\cdot K$$

از تقسیم نظیر به نظیر طرفین در برابری اخیر بر یکدیگر داریم

$$\frac{n_h}{n} = \frac{W_h S_h/\sqrt{C_h}}{\sum_{h=1}^{L}(W_h S_h/\sqrt{C_h})}$$

یا

$$n_h = \frac{W_h S_h/\sqrt{C_h}}{\sum_{h=1}^{L}(W_h S_h/\sqrt{C_h})} \cdot n = \frac{N_h S_h/\sqrt{C_h}}{\sum_{h=1}^{L}(N_h S_h/\sqrt{C_h})} \cdot n \qquad (۱۷.۲)$$

یعنی $n_h \propto N_h S_h/\sqrt{C_h}$ یا $n_h \propto W_h S_h/\sqrt{C_h}$. □

از این قضیه بی‌درنگ می‌توان نتایج زیر را توجیه کرد:

۱. چون n_h نسبت مستقیم با N_h دارد هرچه حجم طبقه بیشتر باشد حجم نمونه باید بزرگتر باشد.

۲. چون n_h با S_h نسبت مستقیم دارد، هرچه تغییرات طبقه، یعنی ناهمگنی واحدها بیشتر باشد باید حجم نمونه بزرگتر باشد.

۳. چون n_h با $\sqrt{C_h}$ و در نتیجه با C_h نسبت معکوس دارد هرچه C_h کوچکتر باشد، یعنی نمونه‌گیری در طبقه‌ای ارزانتر تمام شود، باید حجم نمونهٔ آن بیشتر باشد.

از معادلهٔ (۱۷.۲) مقادیر $n_h(h = ۱, \ldots, L)$ برحسب n به‌دست می‌آیند ولی n مجهول است. لذا n را برای دو مورد تحت بررسی، یعنی معلوم بودن C یا V، به‌دست می‌آوریم.

الف. اگر هزینهٔ کل نمونه‌گیری، C، از قبل تثبیت شده باشد، معادلهٔ (۱۷.۲) مقادیر $n_h(h = ۱, \ldots, L)$ را به‌گونه‌ای می‌دهد که V مینیمم باشد. اگر مقدار n_h را از معادلهٔ (۱۷.۲) در معادلهٔ تابع هزینه قرار دهیم، داریم

$$C - C_۰ = \sum_{h=۱}^{L} C_h \cdot \frac{N_h S_h/\sqrt{C_h}}{\sum_{h=۱}^{L}(N_h S_h/\sqrt{C_h})} \cdot n$$

از این معادله، مقدار n به‌صورت زیر نتیجه می‌شود

$$n = \frac{(C - C_۰) \sum_{h=۱}^{L}(N_h S_h/\sqrt{C_h})}{\sum_{h=۱}^{L}(N_h S_h \sqrt{C_h})} \qquad (۱۸.۲)$$

وقتی C، هزینهٔ کل نمونه‌گیری، مشخص باشد و مقادیر N_h، S_h، و C_h به‌ازای $h = ۱, \ldots, L$ معلوم باشند، ابتدا فرمول (۱۸.۲) را به‌کار می‌بریم و مقدار n را می‌یابیم. در گام بعد به‌کمک فرمول (۱۷.۲) مقادیر n_hها را معین می‌کنیم، و به نمونه‌گیری می‌پردازیم.

ب. اگر V از قبل تثبیت شده باشد، مقادیر n_hها را از فرمول (۱۷.۴) در رابطهٔ (۴.۴) قرار می‌دهیم

$$\mathrm{V}=\sum_{h=1}^{L}\left[W_h^2S_h^2/n_h-\frac{W_hS_h^2}{N}\right]$$

$$\mathrm{V}=\sum_{h=1}^{L}W_hS_h^2\left[W_h\Big/\frac{W_hS_h/\sqrt{C_h}}{\sum W_hS_h/\sqrt{C_h}}n\right]-\sum_{h=1}^{L}\frac{W_h}{N}S_h^2$$

$$\mathrm{V}=\frac{1}{n}\sum_{h=1}^{L}\left[(W_hS_h\sqrt{C_h})\left(\sum_{h=1}^{L}W_hS_h/\sqrt{C_h}\right)\right]-\sum_{h=1}^{L}\frac{W_hS_h^2}{N}$$

وقتی V معلوم است از معادلهٔ بالا n به‌صورت زیر به‌دست می‌آید

$$n=\frac{\left[\sum_{h=1}^{L}W_hS_h\sqrt{C_h}\right]\left[\sum_{h=1}^{L}W_hS_h/\sqrt{C_h}\right]}{\mathrm{V}+\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2} \qquad (۱۹.۴)$$

لذا ابتدا از (۱۹.۴) با داشتن W_h و S_h و C_h مقدار n را به‌دست می‌آوریم و سپس از (۱۷.۴) مقادیر n_h را مشخص می‌کنیم.

تبصرهٔ ۱. یک حالت خاص مهم، حالتی است که هزینهٔ نمونه‌گیری برای هر واحد در تمام طبقات یکسان باشد، یعنی $C_h=c$. در این صورت $C=C_0+cn$ درمی‌آید و تخصیص اپتیمم برای هزینهٔ ثابت به تخصیص اپتیمم برای حجم نمونه‌ای ثابت تبدیل می‌شود، زیرا $n=\frac{C-C_0}{c}$. سپس مقادیر n_h از (۱۷.۴) که به‌صورت زیر خلاصه می‌شود به‌دست می‌آیند

$$n_h=\frac{W_hS_h/\sqrt{c}}{\sum_{h=1}^{L}(W_hS_h/\sqrt{c})}\cdot n=\frac{W_hS_h}{\sum_{h=1}^{L}W_hS_h}n=\frac{N_hS_h}{\sum_{h=1}^{L}N_hS_h}\cdot n \qquad (۲۰.۴)$$

تخصیص (۲۰.۴) را گاهی تخصیص نیمن[1] یا انتساب نیمن می‌نامند.

تبصرهٔ ۲. در تبصرهٔ ۱ دیدیم که برای $C_h=c$ و cی ثابت واریانس $\overline{Y}_{st}$ به‌ازای مقادیر n_h حاصل از (۲۰.۴) مینیمم می‌شود. مقدار این مینیمم را می‌توانیم به‌دست آوریم. می‌دانیم

$$\mathrm{V}=\mathrm{V}(\overline{Y}_{st})=\sum_{h=1}^{L}\frac{W_h^2S_h^2}{n_h}-\sum_{h=1}^{L}\frac{W_hS_h^2}{N}$$

1. Neyman

به‌جای n_h مقدار آن را از (۲۰.۴) قرار می‌دهیم تا V ی اپتیمم به‌دست آید

$$V_{opt} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{nW_h S_h / \sum_{h=1}^{L} W_h S_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N}$$

$$= \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \qquad (21.4)$$

جملهٔ دوم سمت راست، معرف fpc است.

تبصرهٔ ۳. برای استفاده از فرمولهای (۱۷.۴) تا (۲۱.۴) باید مقادیر N_h یا مقادیر W_h و S_h^2 نیز معلوم باشند. در کاربردها معمولاً مقادیر N_h و در نتیجه W_h معلوم‌اند. اما باید با نمونه‌گیری تصادفی مقدماتی از هر طبقه، مثل مورد نمونه‌گیری تصادفی، ابتدا برآوردهایی نااریب برای S_h^2 به‌دست آوریم و در فرمولهای مذکور به‌جای مقادیر S_h^2 برآوردهای نااریب s_h^2 را قرار دهیم. در این صورت n و n_h و V_{opt} به‌ترتیب به $\hat{n}$ و $\hat{n}_h$ و $\hat{V}_{opt}$ تبدیل می‌شوند، و حجمهای نمونه‌ای و واریانس اپتیمم با تقریب به‌دست می‌آیند.

مثال ۲.۴ مؤسسه‌ای تحقیقاتی برای تعیین متوسط مدت زمانی که بیماران مبتلا به دیابت در بیمارستانهای شهری بستری می‌شوند تعداد بیمارانی را که در یک سال در سه بیمارستان شهر بستری بوده‌اند در نظر می‌گیرند. این تعداد به‌ترتیب ۳۰۰، ۱۲۰، و ۱۸۰ است. از روی نمونه‌گیری مقدماتی در هر بیمارستان واریانس تعداد روزهای بستری بودن بیماران تقریباً برابر $\frac{۲۹۹}{۳}$، $\frac{۲۳۸}{۱۵}$، و $\frac{۱۷۹}{۵}$ به‌دست آمده است. هزینهٔ کسب اطلاعات دربارهٔ هر بیمار در سه بیمارستان به‌ترتیب ۴، ۹، ۱۶ است.

الف) مؤسسه ابتدا تصمیم می‌گیرد که کلاً n فرد از سه بیمارستان را به‌عنوان نمونه انتخاب کند. با اطلاعات بالا، در هر بیمارستان چه سهمی از n را باید در نظر گرفت تا تخصیص، اپتیمم باشد؟

ب) اگر مؤسسه بودجه‌ای برابر ۸۸۰ واحد پول که شامل هزینه‌های اداری نیست برای انجام تحقیق تخصیص دهد، با این تخصیص دقیقاً حجم نمونه‌ای که باید از هر بیمارستان انتخاب کرد چقدر است؟

ج) قبل از اجرای نمونه‌گیری، اهمیت مطلب ایجاب می‌کند که برآورد متوسط روزهای بستری بودن را با دقتی خاص معین کنند، به شکلی که واریانس برآورد این متوسط برابر ۴٫۰ باشد. در این صورت از هر بیمارستان به چه حجمی باید نمونه گرفت؟

د) اگر از سه بیمارستان به‌ترتیب ۹۸، ۱۱، ۱۸ بیمار را به‌عنوان نمونه انتخاب کنیم و میانگین مدت بستری بودن بیماران این سه نمونه به‌ترتیب ۳۰، ۲۶، و ۲۴ روز باشد برآورد نااریبی برای متوسط مدت بستری بودن جامعهٔ بیماران دیابتی به‌دست آورید. اگر تغییرات این سه نمونه به‌ترتیب ۸۰، ۱۰، و ۳۰ باشند برآورد واریانس متوسط مدت بستری بودن در هر بیمارستان و برآورد واریانس برآوردکنندهٔ متوسط مدت بستری بودن در بیمارستانها را به‌دست آورید.

ه) یک بازهٔ اطمینان ۹۵ درصدی برای میانگین واقعی مدت بستری بودن بیابید.

الف) بیماران دیابتی به ۳ طبقه تقسیم شده‌اند به‌طوری که

$$N_۱ = ۳۰۰ \quad N_۲ = ۱۲۰ \quad N_۳ = ۱۸۰ \quad N = ۶۰۰$$

واریانسهای طبقات به‌ترتیب عبارت‌اند از $\sigma_۱^۲ = \frac{۲۹۹}{۳}$، $\sigma_۲^۲ = \frac{۲۳۸}{۱۵}$، $\sigma_۳^۲ = \frac{۱۷۹}{۵}$، و

$$S_h^۲ = \frac{N_h}{N_h - ۱}\sigma_h^۲ \quad h = ۱,\ ۲,\ ۳$$

پس

$$S_۱^۲ = \frac{۳۰۰}{۲۹۹} \cdot \frac{۲۹۹}{۳} = ۱۰۰$$
$$S_۲^۲ = \frac{۱۲۰}{۱۱۹} \cdot \frac{۲۳۸}{۱۵} = ۱۶$$
$$S_۳^۲ = \frac{۱۸۰}{۱۷۹} \cdot \frac{۱۷۹}{۵} = ۳۶$$

لذا $S_۱ = ۱۰$، $S_۲ = ۴$، و $S_۳ = ۶$.

حال با توجه به تخصیص اپتیمم

$$n_h = \frac{N_h S_h / \sqrt{C_h}}{\sum (N_h S_h / \sqrt{C_h})} \cdot n$$

و داریم $C_۱ = ۴$، $C_۲ = ۹$، $C_۳ = ۱۶$. محاسبات را به‌صورت زیر انجام می‌دهیم

$$\begin{cases} N_۱ S_۱ / \sqrt{C_۱} & = \dfrac{(۳۰۰)(۱۰)}{۲} = ۱۵۰۰ \\ N_۲ S_۲ / \sqrt{C_۲} & = \dfrac{(۱۲۰)(۴)}{۳} = ۱۶۰ \\ N_۳ S_۳ / \sqrt{C_۳} & = \dfrac{(۱۸۰)(۶)}{۴} = ۲۷۰ \end{cases}$$

$$\sum_{h=۱}^{۳} N_h S_h / \sqrt{C_h} = ۱۹۳۰$$

بنابر فرمول بالا، داریم

$$n_۱ = \frac{۱۵۰۰}{۱۹۳۰} n \simeq ۰/۷۸ n$$
$$n_۲ = \frac{۱۶۰}{۱۹۳۰} n \simeq ۰/۰۸۲ n$$
$$n_۳ = \frac{۲۷۰}{۱۹۳۰} n \simeq ۰/۱۳۹ n$$

ب) براساس فرض مسأله $C - C_{\circ} = ۸۸۰$. می‌دانیم برای Cی ثابت، بنابر (۱۸.۲)

$$n = (C - C_{\circ})\frac{\sum(N_h S_h/\sqrt{C_h})}{\sum N_h S_h \sqrt{C_h}}$$

داریم

$$\begin{cases} N_۱ S_۱ \sqrt{C_۱} = ۳۰۰(۱۰)(۲) = ۶۰۰۰ \\ N_۲ S_۲ \sqrt{C_۲} = ۱۲۰(۴)(۳) = ۱۴۴۰ \\ N_۳ S_۳ \sqrt{C_۳} = ۱۸۰(۶)(۴) = ۴۳۲۰ \end{cases}$$

و

$$\sum_{h=۱}^{۳} N_h S_h \sqrt{C_h} = ۱۱۷۶۰$$

پس

$$n = \frac{(۸۸۰)(۱۹۳۰)}{۱۱۷۶۰} \simeq ۱۴۵$$

و در نتیجه با توجه به نتایج قسمت الف، داریم

$$\begin{aligned} n_۱ &\simeq ۰٫۷۸n = ۰٫۷۸(۱۴۵) \simeq ۱۱۳ \\ n_۲ &\simeq ۰٫۰۸۲n = ۰٫۰۸۲(۱۴۵) \simeq ۱۲ \\ n_۳ &\simeq ۰٫۱۳۹n = ۰٫۱۳۹(۱۴۵) \simeq ۲۰ \end{aligned}$$

ج) از رابطهٔ (۱۹.۲) برای واریانس از پیش تعیین‌شدهٔ V، با توجه به‌رابطهٔ $W_h = \frac{N_h}{N}$، داریم

$$n = \frac{(\sum N_h S_h \sqrt{C_h})(\sum N_h S_h/\sqrt{C_h})}{VN^۲ + \sum N_h S_h^۲}$$

از طرفی

$$\begin{cases} N_۱ S_۱^۲ = ۳۰۰ \times ۱۰^۲ = ۳۰۰۰۰ \\ N_۲ S_۲^۲ = ۱۲۰ \times ۴^۲ = ۱۹۲۰ \\ N_۳ S_۳^۲ = ۱۸۰ \times ۶^۲ = ۶۴۸۰ \end{cases}$$

و

$$\sum_{h=۱}^{۳} N_h S_h^۲ = ۳۸۴۰۰$$

پس

$$n = \frac{(۱۹۳۰)(۱۱۷۶۰)}{۰٫۴(۶۰۰)^۲ + ۳۸۴۰۰} \simeq ۱۲۵$$

و از آنجا

$$\begin{cases} n_۱ = ۱۲۵ \times ۰٫۷۸ \simeq ۹۸ \\ n_۲ = ۱۲۵ \times ۰٫۰۸۲ \simeq ۱۱ \\ n_۳ = ۱۲۵ \times ۰٫۱۳۹ \simeq ۱۸ \end{cases}$$

مجموع n_iها از n بیشتر می‌شود زیرا در محاسبهٔ تقریبی آنها برای کم نشدن دقت نمونه‌گیری تقریب اضافی را منظور کرده‌ایم.

د) در نمونه‌گیری تصادفی از سه‌طبقه به‌ترتیب داریم

$$n_۱ = ۹۸, \quad \overline{Y}_۱ = ۳۰, \quad N_۱ = ۳۰۰, \quad s_۱^۲ = ۸۰, \quad W_۱ = \frac{۳۰۰}{۶۰۰} = \frac{۱}{۲}$$

$$n_۲ = ۱۱, \quad \overline{Y}_۲ = ۲۶, \quad N_۲ = ۱۲۰, \quad s_۲^۲ = ۱۰, \quad W_۲ = \frac{۱۲۰}{۶۰۰} = \frac{۱}{۵}$$

$$n_۳ = ۱۸, \quad \overline{Y}_۳ = ۲۴, \quad N_۳ = ۱۸۰, \quad s_۳^۲ = ۳۰, \quad W_۳ = \frac{۱۸۰}{۶۰۰} = \frac{۳}{۱۰}$$

لذا

$$\hat{\overline{Y}}_N = \sum_{h=۱}^{۳} W_h \hat{\overline{Y}}_h = \frac{۱}{۲}(۳۰) + \frac{۱}{۵}(۲۶) + \frac{۳}{۱۰}(۲۴) = ۲۷٫۴$$

برای تعیین برآورد واریانس متوسط مدت بستری بودن در هر بیمارستان، باید از فرمول برآورد واریانس میانگین نمونه‌گیری تصادفی استفاده کنیم. به‌شرح زیر

$$\hat{\mathrm{V}}(\overline{Y}_۱) = \left(\frac{۱}{n_۱} - \frac{۱}{N_۱}\right) s_۲^۲ = \left(\frac{۱}{۹۸} - \frac{۱}{۳۰۰}\right) ۸۰ \simeq ۰٫۵۵۰$$

$$\hat{\mathrm{V}}(\overline{Y})_۲ = \left(\frac{۱}{n_۲} - \frac{۱}{N^۲}\right) s_۲^۲ = \left(\frac{۱}{۱۱} - \frac{۱}{۱۲۰}\right) ۱۰ \simeq ۰٫۸۲۶$$

$$\hat{\mathrm{V}}(\overline{Y}_۳) = \left(\frac{۱}{n_۳} - \frac{۱}{N_۳}\right) s_۳^۲ = \left(\frac{۱}{۱۸} - \frac{۱}{۱۸۰}\right) ۳۰ = ۱٫۵$$

و بالاخره

$$\hat{\mathrm{V}}(\overline{Y}_{st}) = \frac{۱}{N^۲}\sum_{h=۱}^{۳} N_h(N_h - n_h)\frac{s_h^۲}{n_h}$$
$$= \frac{۱}{(۶۰۰)^۲}\left[۳۰۰(۳۰۰-۹۸)\frac{۸۰}{۹۸}+۱۲۰(۱۲۰-۱۱)\frac{۱۰}{۱۱}+۱۸۰(۱۸۰-۱۸)\frac{۳۰}{۱۸}\right]$$
$$\simeq ۰٫۳۰۵$$

▲

۷.۴ مقایسهٔ دقت برآوردکننده‌ها در نمونه‌گیری تصادفی با طبقه‌بندی و نمونه‌گیری تصادفی ساده

به‌صورت شهودی استنباط ما این است که نمونه‌گیری تصادفی با طبقه‌بندی همیشه برآوردکننده‌ای از میانگین را ارائه می‌دهد که واریانسش از واریانس برآوردکنندهٔ میانگین نمونه‌گیری تصادفی ساده کوچکتر است، اما این احساس شهودی همیشه درست نیست. مواردی وجود دارند که دقت نمونه‌گیری تصادفی ساده بیش از نمونه‌گیری تصادفی با طبقه‌بندی است. اگر مقادیر n_h خیلی از تخصیص اپتیمم دور باشند ممکن است واریانس برآوردکنندهٔ میانگین بزرگ باشد. درواقع حتی وقتی طبقه‌بندی برای حجم تثبیت‌شدهٔ نمونه با تخصیص اپتیمم انجام شود امکان دارد واریانس برآوردکننده بزرگ باشد. در این بخش مقایسه‌ای بین نمونه‌گیری تصادفی ساده و نمونه‌گیری تصادفی با طبقه‌بندی در حالتهای تخصیص اپتیمم و متناسب انجام می‌دهیم. این مقایسه نشان خواهد داد در چه مواقعی انجام طبقه‌بندی مناسب است. قبل از آغاز مقایسه یادآور می‌شویم که واریانس میانگین نمونهٔ تصادفی را با نماد V_{ran} و واریانس $\overline{Y}_{st}$ در حالت تخصیص اپتیمم را با نماد V_{opt} و واریانس در حالت تخصیص متناسب را با نماد V_{prop} نشان می‌دهیم.

قضیهٔ ۷.۴ اگر $\frac{۱}{N_h}$ قابل اغماض باشد. آن‌گاه

$$\mathrm{V}_{opt} \le \mathrm{V}_{prop} \le \mathrm{V}_{ran} \qquad (۲۲.۴)$$

که در آن منظور از تخصیص اپتیمم تخصیص نیمن، برای n ثابت است، یعنی با $n_h \propto N_h S_h$.

برهان. قبلاً دیدیم که به‌ترتیب مطابق با (۵.۲)، (۷.۴)، و (۲۱.۴)

$$\mathrm{V}_{ran} = (۱-f)\frac{S^۲}{n}$$
$$\mathrm{V}_{prop} = \frac{۱-f}{n}\sum W_h S_h^۲ = \frac{\sum W_h S_h^۲}{n} - \frac{\sum W_h S_h^۲}{N}$$
$$\mathrm{V}_{opt} = \frac{(\sum W_h S_h)^۲}{n} - \frac{\sum W_h S_h^۲}{N}$$

از طرفی تغییرات جامعه به‌صورت زیر است

$$S^{۲} = \frac{۱}{N-۱}\sum_{h=۱}^{L}\sum_{i=۱}^{N_h}(Y_{hi} - \overline{Y}_N)^{۲}$$
$$= \frac{۱}{N-۱}\sum_{h=۱}^{L}\sum_{i=۱}^{N_h}[(Y_{hi} - \overline{\mathcal{Y}}_h) + (\overline{\mathcal{Y}}_h - \overline{Y}_N)]^{۲}$$

پس

$$(N-۱)S^{۲} = \sum_{h=۱}^{L}\sum_{i=۱}^{N_h}(Y_{hi} - \overline{\mathcal{Y}}_h)^{۲} + \sum_{h=۱}^{L}\sum_{i=۱}^{N_h}(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲}$$
$$+ ۲\sum_{h=۱}^{L}\sum_{i=۱}^{N_h}(Y_{hi} - \overline{\mathcal{Y}}_h)(\overline{\mathcal{Y}}_h - \overline{Y}_N)$$
$$= \sum_{h=۱}^{L}\sum_{i=۱}^{N_h}(Y_{hi} - \overline{\mathcal{Y}}_h)^{۲} + \sum_{h=۱}^{L} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲}$$

سومین جمله در عبارت وسط برابر با ۰ است (استدلال به‌عهدهٔ دانشجوست). پس

$$(N-۱)S^{۲} = \sum_{h=۱}^{L}(N_h - ۱)S_h^{۲} + \sum_{h=۱}^{L} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲} \qquad (۲۳.۴)$$

زیرا برای طبقهٔ hام

$$S_h^{۲} = \frac{۱}{N_h - ۱}\sum_{i=۱}^{N_h}(Y_{hi} - \overline{\mathcal{Y}}_h)^{۲}$$

اگر $\frac{۱}{N_h}$ قابل اغماض باشد مسلماً $\frac{۱}{N}$ نیز قابل اغماض است. پس از رابطهٔ (۲۳.۴) داریم

$$S^{۲} = \sum_{h=۱}^{L}\frac{N_h - ۱}{N-۱}S_h^{۲} + \sum_{h=۱}^{L}\frac{N_h}{N-۱}(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲}$$
$$= \sum_{h=۱}^{L}\frac{\frac{N_h}{N} - \frac{۱}{N}}{۱ - \frac{۱}{N}}S_h^{۲} + \sum_{h=۱}^{L}\frac{\frac{N_h}{N}}{۱ - \frac{۱}{N}}(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲}$$
$$\simeq \sum_{h=۱}^{L}\frac{N_h}{N}S_h^{۲} + \sum_{h=۱}^{L}\frac{N_h}{N}(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲}$$

پس

$$S^{۲} \simeq \sum_{h=۱}^{L} W_h S_h^{۲} + \sum_{h=۱}^{L} W_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^{۲} \qquad (۲۴.۴)$$

اگر این مقدار را به‌جای $S^۲$ در V_{ran} قرار دهیم، نتیجه می‌شود

$$V_{ran} = (۱-f)\frac{S^۲}{n} \simeq \frac{۱-f}{n}\sum_{h=۱}^{L} W_h S_h^۲ + \frac{۱-f}{n}\sum_{h=۱}^{L} W_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^۲$$

که اولین جملهٔ طرف دوم برابر V_{prop} است. پس

$$V_{ran} = V_{prop} + \frac{۱-f}{n}\sum_{h=۱}^{L} W_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^۲ \qquad (۲۵.۴)$$

از این برابری نتیجه می‌شود که با توجه به نامنفی بودن جمله دوم طرف دوم

$$V_{prop} \leq V_{ran} \qquad (۲۶.۴)$$

برابری وقتی برقرار می‌شود که $\overline{\mathcal{Y}}_h$ها با $\overline{Y}_N$ برابر باشند. از طرفی با توجه به روابط ابتدای برهان

$$V_{prop} - V_{opt} = \frac{۱}{n}\left[\sum_{h=۱}^{L} W_h S_h^۲ - \left(\sum_{h=۱}^{L} W_h S_h\right)^۲\right]$$

اگر قرار دهیم $\overline{S} = \sum_{h=۱}^{L} W_h S_h$

$$\begin{aligned} V_{prop} - V_{opt} &= \frac{۱}{n}\left[\sum_{h=۱}^{L} W_h S_h^۲ - \overline{S}^۲\right] \\ &= \frac{۱}{n}\left[\sum_{h=۱}^{L} W_h (S_h - \overline{S})^۲\right] \qquad (۲۷.۴) \end{aligned}$$

زیرا

$$\begin{aligned} &\sum_{h=۱}^{L} W_h \overline{S}^۲ - ۲\sum_{h=۱}^{L} W_h S_h \overline{S} = \overline{S}^۲ \sum_{h=۱}^{L} W_h - ۲\overline{S}\sum_{h=۱}^{L} W_h S_h \\ &= \overline{S}^۲ - ۲\overline{S}\cdot\overline{S} = -\overline{S}^۲ \end{aligned}$$

پس $V_{prop} - V_{opt} \geq ۰$، زیرا در (۲۷.۴) طرف آخر عبارتی نامنفی است. لذا

$$V_{opt} \leq V_{prop} \qquad (۲۸.۴)$$

برابری وقتی برقرار است که S_hها با $\overline{S}$ برابر باشند. از ترکیب (۲۶.۴) و (۲۸.۴) نتیجه می‌شود که

$$V_{opt} \leq V_{prop} \leq V_{ran}$$

که برهان را تکمیل می‌کند. □

نتیجه. از روابط (۲۵.۴) و (۲۷.۴) رابطهٔ زیر نتیجه می‌شود

$$V_{ran} = V_{opt} + \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \overline{S})^2 + \frac{1-f}{n}\sum_{h=1}^{L} W_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2 \qquad (۲۹.۴)$$

این رابطه نشان می‌دهد که اگر دو نمونه‌گیری تصادفی ساده و تصادفی با طبقه‌بندی در حالت تخصیص اپتیمم را در نظر بگیریم، دو مؤلفه در تقلیل واریانس دخالت دارند. مؤلفهٔ اول (جملهٔ آخر طرف دوم رابطه) وقتی کوچک است که میانگینهای طبقه‌ها خیلی از هم فاصله نداشته باشند، و مؤلفهٔ دوم (جملهٔ میانی طرف دوم) وقتی کوچک است که تفاوت بین انحراف معیارهای طبقات کم باشد. این مؤلفه، معرف تفاوت بین واریانس تخصیص اپتیمم و واریانس تخصیص متناسب است. اگر N_h قابل اغماض نباشد و اگر به‌جای S^2 از (۲۳.۴) استفاده کنیم، پس از محاسبات لازم، نتیجه می‌شود

$$V_{ran} = V_{prop} + \frac{1-f}{n(N-1)}\left[\sum_{h=1}^{N} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2 - \frac{1}{N}\sum_{h=1}^{L}(N-N_h)S_h^2\right]$$

با توجه به عبارت داخل کروشه، اگر داشته باشیم

$$\sum_{h=1}^{L} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2 < \frac{1}{N}\sum_{h=1}^{L}(N-N_h)S_h^2 \qquad (۳۰.۴)$$

آن‌گاه مقدار عبارت داخل کروشه منفی است و $V_{ran} \le V_{prop}$. یعنی با شرط (۳۰.۴) نمونه‌گیری تصادفی ساده کاراتر از نمونه‌گیری با طبقه‌بندی و با تخصیص متناسب است.

اگر همهٔ S_h^2ها، $h = 1, \ldots, L$، باهم برابر باشند و مقدار مشترک آنها را با S_W^2 نشان دهیم به‌قسمی که تخصیص متناسب به مفهوم نیمن، اپتیمم باشد، در این صورت نابرابری (۳۰.۴) به‌صورت زیر درمی‌آید

$$\sum_{h=1}^{L} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2 < \frac{1}{N}\sum_{h=1}^{L}(N-N_h)S_W^2$$

و یا

$$\sum_{h=1}^{L} N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2 < \frac{S_W^2}{N}(NL-N) = (L-1)S_W^2$$

یا

$$\sum_{h=1}^{L} \frac{N_h(\overline{\mathcal{Y}}_h - \overline{Y}_N)^2}{L-1} < S_W^2 \qquad (۳۱.۴)$$

اگر تحلیل واریانس را به‌خاطر داشته باشید از این رابطه نتیجه می‌گیرید که باید میانگین مربعات در

بین طبقات، کوچکتر از میانگین مربعات در درون طبقات باشد، یا به‌عبارت دیگر نسبت F باید کوچکتر از ۱ باشد. در چنین حالتی نمونه‌گیری تصادفی کاراتر از نمونه‌گیری تصادفی با طبقه‌بندی در حالت تخصیص متناسب است.

۸.۴ حالت خاص تخصیص نیمن

وقتی n_hها را برای مقداری تثبیت‌شده از n محاسبه می‌کنیم ممکن است یک یا چند n_h از N_hهای مربوط بزرگتر شوند. اگر تخصیص مورد نظر، تخصیص نیمن باشد و مثلاً داشته باشیم $n_1 > N_1$، آن‌گاه خط‌مشی معمول این است: اگر جامعه بیش از دو طبقه داشته باشد

$$\hat{n}_1 = N_1 \,, \qquad \hat{n}_h = (n - N_1)\frac{W_h S_h}{\sum_{h=2}^{L} W_h S_h} \qquad h \geq 2 \qquad (۳۲.۴)$$

درواقع چون $n_1 > N_1$ نامفهوم است حداکثر $\hat{n}_1 = N_1$ است، سپس از رابطهٔ $n_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \cdot n$ مربوط به تخصیص نیمن استفاده می‌کنیم. بدین‌طریق که ابتدا N_1 را از حجم کل نمونه کم می‌کنیم و $n - N_1$ واحد باقیماندهٔ نمونه را از بین $h - 1$ طبقه دیگر با استفاده از فرمول تخصیص نیمن انتخاب می‌کنیم که (۳۲.۴) را نتیجه می‌دهد. اگر $n_1 > N_1$ و $n_2 > N_2$ به‌دست آمده باشند، خط‌مشی معمول چنین است

$$\hat{n}_1 = N_1 \,, \qquad \hat{n}_2 = N_2 \,, \qquad \hat{n}_h = (n - N_1 - N_2)\frac{W_h S_h}{\sum_{h=3}^{L} W_h S_h} \qquad h \geq 3 \qquad (۳۳.۴)$$

استدلال مربوط به درستی رابطهٔ (۳۳.۴) شبیه استدلال حالت قبل است. به‌طورکلی هر تعداد از n_hهایی را که بزرگتر از N_h مربوط باشند با N_h برآورد می‌کنیم و سپس مجموع این N_hها را از کل n می‌کاهیم و $n - \sum N_h$ واحد باقیمانده را از طبقات باقیمانده با استفاده از فرمول تخصیص نیمن به‌دست می‌آوریم. نکته‌ای که در این مورد حائز اهمیت است این است که فرمول

$$\mathrm{V}_{min}(\overline{Y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

که در تخصیص نیمن برای محاسبهٔ واریانس مینیمم به‌کار می‌رود در این حالت برقرار نیست. اگر $\sum'$ معرف مجموعیابی روی طبقه‌هایی باشد که برای آنها $\hat{n}_h < N_h$ است، فرمول واریانس مینیمم به‌صورت زیر حاصل می‌شود

$$\mathrm{V}_{min}(\overline{Y}_{st}) = \frac{(\sum' W_h S_h)^2}{n'} - \frac{\sum' W_h S_h^2}{N}$$

که در آن n' مجموع حجمهای نمونه‌های طبقه‌های با شرط $\hat{n}_h < N_h$ است.

۹.۴ برآورد حجم نمونهٔ کل وقتی وزن نمونه‌ها معلوم است

برای تعیین مقدار n، تحت تخصیص اپتیمم فرمولهایی ارائه دادیم. در این بخش برای هر تخصیصی، فرمولهایی، معرفی می‌شود. فرض می‌کنیم که $\overline{Y}_{st}$ دارای واریانس معلوم V بوده، s_h برآورد S_h و مقادیر $w_h = \frac{n_h}{n}$ معلوم باشند. می‌دانیم بنابر (۱۱.۴)

$$\mathrm{V} \simeq \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2$$

اگر به‌جای n_h مقدار آن را در این رابطه قرار دهیم، نتیجه می‌شود

$$\mathrm{V} \simeq \frac{1}{n}\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2$$

از این برابری، رابطهٔ کلی زیر را برای $\hat{n}$ به‌دست می‌آوریم

$$\hat{n} = \frac{\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h}}{\mathrm{V} + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2} \qquad (۳۴.۴)$$

اگر N بزرگ باشد، آن‌گاه

$$n_0 = \frac{1}{\mathrm{V}}\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h} \qquad (۳۵.۴)$$

و اگر N بزرگ نباشد، از ترکیب (۳۴.۴) و (۳۵.۴) داریم

$$\hat{n} = \frac{n_0}{1 + \frac{1}{N\mathrm{V}}\sum_{h=1}^{L} W_h s_h^2} \qquad (۳۶.۴)$$

در حالتهای خاص، فرمولها به‌صورتهایی درمی‌آیند که برای محاسبه راحت‌ترند. چند مورد را در زیر می‌آوریم:

الف) تخصیص را نیمن می‌گیریم (n تثبیت شده است): وقتی $n_h \simeq n\frac{W_h s_h}{\sum W_h s_h}$، یعنی $\frac{n_h}{n} = w_h \simeq \frac{W_h s_h}{\sum W_h s_h}$ آن‌گاه صورت کسر (۳۴.۴) به‌صورت زیر درمی‌آید

$$\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{W_h s_h / \sum_{h=1}^{L} W_h s_h} = \left(\sum_{h=1}^{L} W_h s_h\right)^2$$

و در نتیجه (۳۴.۴) چنین نوشته می‌شود

$$\hat{n} = \frac{(\sum_{h=1}^{L} W_h s_h)^2}{V + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2} \qquad (۳۷.۴)$$

ب) تخصیص متناسب. در تخصیص متناسب $w_h = \frac{n_h}{n} = \frac{N_h}{N} = W_h$. بنابراین (۳۵.۴) را می‌توان به‌صورت زیر نوشت

$$n_0 = \frac{1}{V}\sum_{h=1}^{L} W_h s_h^2 \qquad (۳۸.۴)$$

و (۳۶.۴) به‌صورت زیر درمی‌آید

$$\hat{n} = \frac{n_0}{1 + \frac{n_0}{N}} \qquad (۳۹.۴)$$

۱۰.۴ تعیین حجم نمونه وقتی واریانس برآوردکنندهٔ مجموع مقادیر واحدها از قبل مشخص شده است

اگر مایل باشیم n را به‌قسمی بیابیم که $V(\hat{T}_N)$ برابر مقدار معین V باشد، در فرمولهای قبل به‌جای V مقدار $\frac{V}{N^2}$ را قرار می‌دهیم تا مقدار n مورد نظر به‌دست آید، زیرا در فرمولهای بالا V برابر با واریانس $\overline{Y}_{st}$ است در حالی که در این بخش V معرف $V(\hat{T}_N)$ یا $V(N\overline{Y}_{st})$ است که برابر با $N^2V(\overline{Y}_{st})$ است. از قرار دادن $\frac{V}{N^2}$ به‌جای V در رابطهٔ (۳۴.۴)، برای حالت کلی داریم

$$\hat{n} = \frac{\sum_{h=1}^{L} W_h^2 s_h^2/w_h}{\frac{V}{N^2} + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2} = \frac{\sum_{h=1}^{L} \frac{N_h^2}{N^2} s_h^2/w_h}{\frac{V}{N^2} + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N} s_h^2}$$

و یا به‌صورت خلاصه

$$\hat{n} = \frac{\sum_{h=1}^{L} N_h^2 s_h^2/w_h}{V + \sum_{h=1}^{L} N_h s_h^2} \qquad (۴۰.۴)$$

برای تخصیص نیمن (n تثبیت‌شده) از (۳۷.۴) داریم

$$\hat{n} = \frac{(\sum_{h=1}^{L} W_h s_h)^2}{\frac{V}{N^2} + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2} = \frac{(\sum_{h=1}^{L} \frac{N_h}{N} s_h)^2}{\frac{V}{N^2} + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N} s_h^2}$$

و یا به‌صورت خلاصه

$$\hat{n} = \frac{(\sum_{h=۱}^{L} N_h s_h)^۲}{V + \sum_{h=۱}^{L} N_h s_h^۲} \qquad (۴۱.۴)$$

و برای تخصیص متناسب از (۳۸.۴) و (۳۹.۴) داریم

$$n_۰ = \frac{۱}{V/N^۲} \sum_{h=۱}^{L} W_h s_h^۲ = \frac{N^۲}{V} \sum_{h=۱}^{L} \frac{N_h}{N} s_h^۲$$

و یا

$$n_۰ = \frac{N}{V} \sum_{h=۱}^{L} N_h s_h^۲ \qquad (۴۲.۴)$$

و

$$\hat{n} = \frac{n_۰}{۱ + \frac{n_۰}{N}}$$

مثال ۳.۴* برای برآورد تعداد کسانی که از آنها در سال اول دانشکده‌های یک دانشگاه ثبت نام شده است، ۱۹۶ مسؤول ثبت نام دانشگاه را در نظر گرفته‌اند. این مسؤولین را ابتدا در ۷ طبقه منظور و سپس یک طبقه را به‌علت کوچکی تعداد در ۶ طبقهٔ دیگر ادغام کرده‌اند. مقادیر N_hها در جدول زیر آمده‌اند. قرار است از هر طبقه تعدادی را به‌عنوان نمونه انتخاب کنند و از روی تعداد ثبت نام آنها، تعداد کل ثبت نامها را برآورد کنند. مقدار S_h را از روی ثبت نامهای دو سال قبل به‌صورت مقادیری از s_h که در جدول آمده‌اند برآورد کرده‌اند. می‌خواهند n، حجم کل نمونه را به‌قسمی بیابند که ضریب تغییرات در کل ثبت نام ۵٪ باشد. کل تعداد ثبت نامهای در سال پیش دانشگاه ۵۶۴۷۲ بوده است.

با توجه به میزان ثبت نام دو سال پیش و تعریف ضریب تغییرات، خطای معیار مطلوب $\sigma = (۰٫۰۵)(۵۶۴۷۲) = ۲۸۲۴$، پس واریانس مطلوب

$$V = (۲۸۲۴)^۲ = ۷۹۷۴۹۷۶$$

ممکن است این انتقاد وارد باشد که ضریب تغییرات دو سال قبل برای ثبت نامهای فعلی صادق نبوده و ضریب تغییرات فعلی بیشتر است. اما فرض می‌کنیم که این ضریب ثابت مانده باشد. در جدول زیر مقادیر N_h، s_h و $N_h s_h$ را قبل از تعیین n آورده‌ایم. فرمول مربوط برای تعیین n، فرمول (۴۱.۴) است که برای تخصیص اپتیمم در برآورد مجموع به‌کار می‌رود. با ۱۹۶ واحد در این

* اقتباس از اثر کوکران: تکنیکهای نمونه‌گیری (۱۹۷۷) وایلی.

جامعه، نمی‌توان از fpc صرف‌نظر کرد. اما ابتدا fpc را نادیده می‌گیریم و $n_\circ$ را حساب می‌کنیم

$$n_\circ = \frac{(\sum N_h s_h)^۲}{V} = \frac{(۲۶۸۴۱)^۲}{۷۹۷۴۹۷۶} = ۹۰٫۳۴$$

و برای محاسبهٔ n بدون نادیده گرفتن fpc، داریم

$$\hat{n} = \frac{n_\circ}{۱ + \frac{۱}{V}\sum N_h s_h^۲} = \frac{۹۰٫۳۴}{۱ + \frac{۴۶۴۰۳۸۷}{۷۹۷۴۹۷۶}} \simeq ۵۷$$

با توجه به رابطهٔ $\hat{n}_h = n \cdot \frac{N_h s_h}{\sum N_h s_h}$ می‌توان مقادیر حجم نمونه‌ای هر طبقه را مشخص کرد. این مقادیر هم در جدول زیر آمده‌اند

طبقه	N_h	s_h	$N_h s_h$	$N_h s_h^۲$	$\hat{n}_h$
۱	۱۳	۳۲۵	۴۲۲۵	۱۳۷۳۱۲۵	۹
۲	۱۸	۱۹۰	۳۴۲۰	۶۴۹۸۰۰	۷
۳	۲۶	۱۸۹	۴۹۱۴	۹۲۸۷۴۶	۱۱
۴	۴۲	۸۲	۳۴۴۴	۲۸۲۴۰۸	۷
۵	۷۳	۸۶	۶۲۷۸	۵۳۹۹۰۸	۱۳
۶	۲۴	۱۹۰	۴۵۶۰	۸۶۶۴۰۰	۱۰
مجموع	۱۹۶		۲۶۸۴۱	۴۶۴۰۳۸۷	۵۷

▲

۱۱.۴ نمونه‌گیری با طبقه‌بندی برای برآورد نسبتها

فرض می‌کنیم هر واحد جامعه به یکی از دو ردهٔ مجزا از هم C و C' متعلق باشد. می‌خواهیم در این جامعه نسبت واحدهایی را بیابیم که در ردهٔ معین C می‌افتند. اگر جامعه در L طبقه سازمان یافته باشد، آن‌گاه فرض می‌کنیم A_h تعداد واحدهایی از طبقه hام است که در ردهٔ C می‌افتند و a_h تعداد واحدهایی از نمونهٔ تصادفی این طبقه باشد که در ردهٔ C هستند. پس

$$P_h = \frac{A_h}{N_h} \quad , \quad p_h = \frac{a_h}{n_h} \quad h = ۱,\ ۲, \ldots, L$$

که P_h، نسبت واحدهایی از طبقهٔ hام جامعه و p_h نسبت واحدهایی از نمونهٔ طبقهٔ hام است که در ردهٔ C می‌افتند. قبلاً در نمونه‌گیری تصادفی ساده دیدیم که p_h برآوردکنندهٔ ناریب P_h است. اگر p_{st} را به‌صورت

$$p_{st} = \sum_{h=۱}^{L} W_h p_h \qquad (۴۳.۴)$$

تعریف کنیم. واضح است که

$$E(p_{st}) = \sum_{h=1}^{L} W_h E(p_h) = \sum_{h=1}^{L} W_h P_h = \sum_{h=1}^{L} W_h \frac{A_h}{N_h}$$
$$= \sum_{h=1}^{L} \frac{N_h}{N} \cdot \frac{A_h}{N_h} = \frac{1}{N} \sum_{h=1}^{L} A_h$$

اما $\sum_{h=1}^{L} A_h$ برابر مجموع واحدهایی از جامعه است که در ردهٔ C هستند. اگر قرار دهیم

$$P = \frac{\sum_{h=1}^{L} A_h}{N}$$

نتیجه می‌شود $E(p_{st}) = P$ یعنی p_{st} در (۴۳.۴) برآوردکنندهٔ نااریب نسبت واحدهایی از جامعه است که در ردهٔ C هستند.

قضیهٔ ۸.۴ در نمونه‌گیری تصادفی با طبقه‌بندی

$$V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \cdot \frac{P_h Q_h}{n_h} \qquad (۴۴.۴)$$

برهان. این، حالتی خاص از قضیهٔ کلی واریانس برآورد میانگین در نمونه‌گیری با طبقه‌بندی است. قبلاً دیدیم که

$$V(\overline{Y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

اگر Y_{hi} را در حالت خاص، وقتی در ردهٔ C است برابر ۱ و در غیر این‌صورت برابر ۰ بگیریم، بنابر آنچه در نمونه‌گیری تصادفی برای تعیین نسبتها دیدیم

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \qquad (۴۵.۴)$$

که در آن $Q_h = 1 - P_h$. اگر (۴۵.۴) را در رابطهٔ قبلی قرار دهیم، نتیجه می‌شود

$$V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{N_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$
$$= \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \cdot \frac{P_h Q_h}{n_h}$$

□

تبصره. در کاربردها، حتی وقتی که fpc قابل اغماض نباشد $\frac{N_h}{N_h-1}$ را برابر با ۱ می‌گیرند و فرمول واریانس را به‌صورت تقریبی زیر به‌کار می‌برند

$$\begin{aligned} V(p_{st}) &\simeq \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h-n_h)\frac{P_hQ_h}{n_h} \\ &= \sum_{h=1}^{L}\frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{P_hQ_h}{n_h} \\ &= \sum_{h=1}^{L} W_h^2(1-f_h)\frac{P_hQ_h}{n_h} \qquad (۴۶.۴) \end{aligned}$$

فرع ۱ . اگر بتوانیم fpc را نادیده بگیریم، خواهیم داشت

$$V(p_{st}) \simeq \sum_{h=1}^{L} W_h^2\frac{P_hQ_h}{n_h} \qquad (۴۷.۴)$$

فرع ۲ . اگر تخصیص، متناسب باشد در (۲۲.۲) برابری $\frac{n_h}{N_h} = \frac{n}{N}$ را منظور می‌کنیم. چون $\frac{n_h}{n} = \frac{N_h}{N} = \frac{N_h-n_h}{N-n}$، پس

$$\begin{aligned} V(p_{st}) &= \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2}{N_h-1}(N-n)\frac{n_h}{n}\frac{P_hQ_h}{n_h} \\ &= \frac{N-n}{N}\cdot\frac{1}{nN}\sum_{h=1}^{L}\frac{N_h^2}{N_h-1}P_hQ_h \qquad (۴۸.۴) \\ &\simeq \frac{1-f}{n}\sum_{h=1}^{L}W_hP_hQ_h \qquad (۴۹.۴) \end{aligned}$$

در برابری آخر $N_h - 1$ را به‌تقریب برابر با N_h فرض کرده‌ایم.

فرع ۳ . چون P_h در عمل مجهول است، لذا نمی‌توان $V(p_{st})$ را به‌دست آورد. برای تعیین $\hat{V}(p_{st})$، همان‌طورکه در نمونه‌گیری تصادفی ساده برای نسبتها دیدیم از برآورد $\frac{P_hQ_h}{n_h}$ که تقریباً برابر با $\frac{p_hq_h}{n_h-1}$ است، استفاده می‌کنیم، بنابراین

$$\hat{V}(p_{st}) \simeq \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2(N_h-n_h)}{N_h-1}\frac{p_pq_h}{n_h-1} \qquad (۵۰.۴)$$

فرع ۴ . بهترین انتخاب مقادیر n_hها برای مینیمم کردن $V(p_{st})$ از قضیهٔ کلی (۶.۴) نتیجه می‌شود:

الف) اگر حجم نمونه ثابت باشد برای اینکه واریانس مینیمم شود باید $n_h \propto N_h S_h$ باشد
ولی از (۴۵.۴) داریم

$$S_h = \sqrt{\frac{N_h}{N_h - ١} P_h Q_h}$$

پس

$$n_h \propto N_h \sqrt{N_h/(N_h - ١)} \cdot \sqrt{P_h Q_h} \simeq N_h \sqrt{P_h Q_h}$$

که در آن $N_h - ١ \simeq N_h$ فرض شده است. پس از رابطهٔ

$$n_h = n \frac{N_h S_h}{\sum_{h=١}^{L} N_h S_h}$$

داریم

$$n_h \simeq n \cdot \frac{N_h \sqrt{P_h Q_h}}{\sum_{h=١}^{L} N_h \sqrt{P_h Q_h}} \qquad (۵۱.۴)$$

ب) اگر هزینهٔ $C = C_\circ + \sum_{h=١}^{L} C_h n_h$ ثابت باشد، وقتی واریانس p_{st} مینیمم است که داشته باشیم

$$n_h = n \frac{N_h S_h / \sqrt{C_h}}{\sum N_h S_h / \sqrt{C_h}}$$

که اگر به‌جای S_h مقدار تقریبی آن $\sqrt{P_h Q_h}$ را قرار دهیم

$$n_h \simeq n \frac{N_h \sqrt{P_h Q_h / C_h}}{\sum_{h=١}^{L} N_h \sqrt{P_h Q_h / C_h}} \qquad (۵۲.۴)$$

مقدار n از رابطهٔ (۱۸.۴) با قرار دادن $\sqrt{P_h Q_h}$ به‌جای S_h به‌دست می‌آید

$$n = \frac{(C - C_\circ) \sum (N_h \sqrt{P_h Q_h / C_h})}{\sum (N_h \sqrt{P_h Q_h C_h})} \qquad (۵۳.۴)$$

مثال ۴.۴ جامعهٔ زیر با دو طبقه داده شده است. نمونه‌هایی با حجم ۲ $= n_١$ و ۲ $= n_٢$ از هر طبقه انتخاب کنید و نشان دهید که $E(p_{st}) = P$.

طبقهٔ I	طبقهٔ II
۱	۱
۰	۰
۱	۰

می‌توانیم $۹ = \binom{۳}{۲}\binom{۳}{۲}$ نمونهٔ ممکن به حجم $n = n_۱ + n_۲ = ۴$ از این جامعه استخراج کنیم. این ۹ نمونهٔ ممکن در زیر فهرست شده‌اند

I	II	$p_۱$	$p_۲$	p_{st}
	۱, ۰	۰٫۵	۰٫۵	۰٫۵
۱, ۰	۱, ۰		۰٫۵	۰٫۵
	۰, ۰		۰	۰٫۲۵
	۱, ۰	۱	۰٫۵	۰٫۷۵
۱, ۱	۱, ۰		۰٫۵	۰٫۷۵
	۰, ۰		۰	۰٫۵
	۱, ۰	۰٫۵	۰٫۵	۰٫۵
۰, ۱	۱, ۰		۰٫۵	۰٫۵
	۰, ۰		۰	۰٫۲۵

تشخیص چگونگی تشکیل جدول به‌عهدهٔ خواننده است.

$$E(p_{st}) = \frac{۱}{۹}(۰٫۵ + ۰٫۵ + \cdots + ۰٫۲۵) = \frac{۴٫۵}{۹} = ۰٫۵$$

از طرفی با توجه به جامعهٔ اصلی

$$P = \frac{۳}{۶} = ۰٫۵$$

پس

$$E(p_{st}) = P$$

∆

مثال ۵.۴ یک بررسی مقدماتی در سه شهر کوچک انجام داده‌اند و نسبت خانواده‌هایی را که دو فرزند یا بیشتر دارند برآورد کرده‌اند. از داده‌های حاصل از بررسی مقدماتی استفاده کنید و تعیین کنید که برای برآورد p_{st} با تخصیص نیمن، وقتی حجم کل نمونه n است، حجم نمونهٔ هر شهر چقدر باید باشد؟

شهر	خانوداه	p_h	p_hq_h	$\sqrt{p_hq_h}$	$N\cdot\sqrt{p_hq_h}$
A	۲۰۰۰	۰٫۱۰	۰٫۰۹	۰٫۳	۶۰۰
B	۳۰۰۰	۰٫۱۵	۰٫۱۲۷۵	۰٫۳۵	۱۰۵۰
C	۵۰۰۰	۰٫۲۰	۰٫۱۶	۰٫۴	۲۰۰۰
مجموع	۱۰۰۰۰				۳۶۵۰

از فرمول نیمن داریم

$$\hat{n}_h \simeq n\cdot\frac{N_h\sqrt{p_hq_h}}{\sum N_h\sqrt{p_hq_h}}$$

با توجه به محاسبات جدول بالا که، سه ستون اول آن داده‌های حاصل از بررسی است، داریم

$$\hat{n}_۱=\frac{۶۰۰}{۳۶۵۰}\,n=\frac{۱۲}{۷۳}n$$

$$\hat{n}_۲=\frac{۱۰۵۰}{۳۶۵۰}\,n=\frac{۲۱}{۷۳}$$

$$\hat{n}_۳=\frac{۲۰۰۰}{۳۶۵۰}\,n=\frac{۴۰}{۷۳}\,n$$

▲

۱۲.۴ اثر انحراف از تخصیص اپتیمم

در این بخش کاهش دقت نتیجهٔ نمونه‌گیری را به‌دلیل عدم توفیق در دستیابی به تخصیص اپتیمم مورد بحث قرار می‌دهیم.

فرض کنید که قصد داریم برای مقداری مفروض از n، از تخصیص نیمن استفاده کنیم. حجم نمونه‌ای n'_h در طبقهٔ hام باید به‌صورت زیر باشد

$$n'_h=\frac{n(W_hS_h)}{\sum W_hS_h} \qquad (۵۴.۴)$$

می‌دانیم که واریانس مینیمم حاصل از این تخصیص چنین است

$$\mathrm{V}_{min}(\overline{Y}_{st})=\frac{(\sum W_hS_h)^۲}{n}-\frac{\sum W_hS_h^۲}{N} \qquad (۵۵.۴)$$

در عمل، چون مقادیر S_hها معلوم نیستند فقط این تخصیص را با تقریب می‌توان انجام داد. اگر $\hat{n}_h$ حجم نمونه‌ایی مورد استفاده در طبقه hام باشد مقدار واریانسی که حاصل می‌شود به‌صورت

زیر است

$$\mathrm{V}(\overline{Y}_{st}) = \sum \frac{W_h^{۲} S_h^{۲}}{\hat{n}_h} - \frac{\sum W_h S_h^{۲}}{N}$$

افزایش واریانس ناشی از تخصیص ناکامل عبارت است از

$$\mathrm{V}(\overline{Y}_{st}) - \mathrm{V}_{min}(\overline{Y}_{st}) = \sum \frac{W_h^{۲} S_h^{۲}}{\hat{n}_h} - \frac{۱}{n}\left(\sum W_h S_h\right)^{۲} \qquad (۵۶.۴)$$

از رابطهٔ (۵۴.۴) داریم

$$W_h S_h = \frac{n'_h}{n} \sum W_h S_h$$

اگر این مقدار را در (۵۶.۴) قرار دهیم، نتیجهٔ زیر حاصل می‌شود

$$\begin{aligned}\mathrm{V}(\overline{Y}_{st}) - \mathrm{V}_{min}(\overline{Y}_{st}) &= \frac{(\sum W_h S_h)^{۲}}{n^{۲}}\left[\sum \frac{n_h'^{۲}}{\hat{n}_h} - n\right] \\ &= \frac{(\sum W_h S_h)^{۲}}{n^{۲}} \sum \frac{(\hat{n}_h - n'_h)^{۲}}{\hat{n}_h} \qquad (۵۷.۴)\end{aligned}$$

اگر در رابطهٔ (۵۵.۴)، $N \longrightarrow \infty$، آن‌گاه این رابطه به‌صورت زیر درمی‌آید

$$\mathrm{V}_{min}(\overline{Y}_{st}) = \frac{(\sum W_h S_h)^{۲}}{n} \qquad (۵۸.۴)$$

از تقسیم روابط (۵۷.۴) و (۵۸.۴) نتیجه می‌شود

$$\frac{\mathrm{V}(\overline{Y}_{st}) - \mathrm{V}_{min}(\overline{Y}_{st})}{\mathrm{V}_{min}(\overline{Y}_{st})} = \frac{۱}{n} \sum \frac{(\hat{n}_h - n'_h)^{۲}}{\hat{n}_h} \qquad (۵۹.۴)$$

این مقدار، افزایش نسبی واریانس را، وقتی تخصیص به‌صورت ناکامل انجام می‌شود نشان می‌دهد. در این رابطه $\hat{n}_h$ مقدار حجم نمونه در طبقهٔ hام است که به‌جای مقدار n'_h که مربوط به تخصیص اپتیمم است مورد استفاده واقع شده است. اگر fpc قابل چشم‌پوشی نباشد، در رابطهٔ (۵۹.۴) نماد (=) به‌نماد «$\geq$» تبدیل می‌شود.

اگر $g_h = |\hat{n}_h - n'_h|/\hat{n}_h$، (۵۹.۴) به‌صورت زیر درمی‌آید

$$\frac{\mathrm{V} - \mathrm{V}_{min}}{\mathrm{V}_{min}} = \sum_{h=۱}^{L} \frac{\hat{n}_h}{n} g_h^{۲}$$

که درواقع میانگین موزون $g_h^{۲}$ است. بنابراین، حد بالایی محافظه‌کارانهٔ $\frac{\mathrm{V}-\mathrm{V}_{min}}{\mathrm{V}_{min}}$ برابر $g^{۲}$ است که در آن g، بزرگترین تفاوت نسبی در هر طبقه است.

مثلاً اگر $g = ۰٫۲$ یا ۲۰٪ باشد، نمو نسبی واریانس نمی‌تواند از $(۰٫۲)^۲$ تجاوز کند. اگر $g = ۰٫۳$ یا ۳۰٪ باشد، نمو نسبی واریانس حداکثر ۹٪ است.

۱۳.۴ مسألهٔ تخصیص، برای بیش از یک صفت

گاهی اوقات در یک نمونه‌گیری، چند صفت از واحدهای نمونه به‌طور همزمان مشاهده و اندازه‌گیری می‌شوند. چون بهترین تخصیص مقادیر n_h برای یک صفت الزاماً بهترین تخصیص برای صفت دیگر نیست (زیرا مقادیر S_h صفتها با هم متفاوت‌اند) بنابراین معلوم نیست بین مقادیر مختلف n_h که متناظر با صفتهای مختلف برای هر طبقه به‌دست می‌آیند کدام n_h را باید برگزید، و نمونه‌گیری از آن طبقه را با آن حجم نمونه‌ای انجام داد. در این مورد اولین گامی که باید برداشت این است که بین صفتهای مختلف، صفاتی را که از نظر بررسی نمونه‌ای در درجهٔ اول اهمیت‌اند انتخاب کنیم و بقیه را کنار بگذاریم. این کار موجب تقلیل صفات برای تعیین مقادیر n_h می‌شود. اگر داده‌های مقدماتی خوبی موجود باشند، آن‌گاه می‌توانیم تخصیص اپتیمم را برای هر یک از صفات بیابیم و ببینیم که در هر طبقه میزان ناهماهنگی حجمهای حاصل از صفات چقدر است. اگر بین صفات، میزان همبستگی زیاد باشد اختلاف این حجمها نسبتاً کم خواهد بود. در هر حال بین این حجمها در هر طبقه باید با مصالحه یکی را انتخاب کرد. مثال زیر نحوهٔ عمل را مشخص می‌کند.

مثال ۶.۴ داده‌های این مثال را جسن[1] در یک بررسی ارائه داده است. ایالت آیووا از نظر جغرافیایی به ۵ منطقه تقسیم شد و سازمانهای عمدهٔ کشاورزی در هر منطقه مشخص شد. این ۵ منطقه به‌عنوان ۵ طبقه در یک بررسی از صنعت شیر به‌کار رفت. سه قلم از مهمترین اقلام تحت بررسی به‌صورت (۱) تعداد گاوهایی که در روز دوشیده شده‌اند، (۲) تعداد گالنهای شیر در روز، و (۳) کل درآمد روزانه از محصولات شیر مشخص شدند. از یک بررسی در چند سال قبل مقادیر s_h طبقات به‌صورتی بوده‌اند که در جدول ۱ نشان داده‌ایم. در جدول ۲، تخصیصهای اپتیمم نیمن بر پایهٔ این s_hها برای هر قلم، در نمونه‌ای به حجم ۱۰۰۰ سازمان داده شده است.

جدول ۱. وزنها و مقادیر s_h طبقات

طبقه	$W_h = \dfrac{N_h}{N}$	s_h برای گاوهای شیرده	s_h برای گالنهای شیر	s_h برای درآمد روزانهٔ محصولات شیر
۱	۰٫۱۹۷	۴٫۶	۱۱٫۷	۳۳۲
۲	۰٫۱۹۱	۳٫۴	۹٫۸	۳۵۷
۳	۰٫۲۱۹	۳٫۳	۷٫۰	۲۴۶
۴	۰٫۱۸۴	۲٫۸	۶٫۵	۱۷۳
۵	۰٫۲۰۸	۳٫۷	۹٫۸	۲۷۹

1. Jessen

جدول ۲. مقادیر n_h اپتیمم متناظر با سه صفت و مقادیر n_h تخصیص متناسب

	تخصیص				
		اپتیمم برای			متوسط
طبقه	متناسب	گاوها	گالنها	درآمدها	n_h
۱	۱۹۷	۲۵۴	۲۵۸	۲۳۶	۲۵۰
۲	۱۹۱	۱۸۲	۲۰۹	۲۴۶	۲۱۲
۳	۲۱۹	۲۰۳	۱۷۱	۱۹۴	۱۸۹
۴	۱۸۴	۱۴۵	۱۳۴	۱۱۵	۱۳۱
۵	۲۰۸	۲۱۶	۲۲۸	۲۰۹	۲۱۸

به‌طوری که از جدول ۲ دیده می‌شود تخصیصهای اپتیمم در ستونهای ۲ و ۳ و ۴ تفاوتهای نسبتاً کمی دارند و تفاوت تمام آنها با تخصیص متناسب در یک جهت است. مثلاً در طبقهٔ اول، با تخصیص متناسب، باید از ۱۹۷ سازمان نمونه‌گیری کرد در حالی که حجم نمونه برای طبقهٔ اول با تخصیص اپتیمم از ۲۳۶ تا ۲۵۸ است. در ستون آخر متوسط تقریبی سه حجم نمونه‌ای اپتیمم را از طبقهٔ اول برابر با ۲۵۰ نوشته‌ایم که مصالحه‌ای بین سه حجم نمونه‌ای حاصل از تخصیص اپتیمم برای سه صفت است. در جدول ۳، واریانسهای نمونه‌ای مورد انتظار $\bar{y}_{st}$ را برای تخصیصهای اپتیمم سه‌گانه و برای حالت مصالحه و تخصیص متناسب آورده‌ایم. فرمولهایی که برای محاسبه به‌کار برده‌ایم به‌صورت زیرند

$$V_{opt} = \frac{(\sum W_h s_h)^2}{n}$$

$$^{*}V_{com} = \sum \frac{(W_h S_h)^2}{n_h}$$

$$V_{prop} = \frac{\sum W_h s_h^2}{n}$$

اگر ممکن می‌بود که سه‌بار، هر بار برای یک صفت با تخصیص اپتیمم، عمل نمونه‌گیری را انجام داد، دقت حاصل نزدیک دقت تخصیص حالت مصالحه می‌شد. آنچه جالب است این است که دقت تخصیص متناسب کمی کمتر از دقت تخصیص حالت مصالحه است.

جدول ۳. واریانسهای مورد انتظار برآورد میانگین

نوع تخصیص	گاوها	گالنها	درآمدها
اپتیمم	۰٫۰۱۲۷	۰٫۰۸۰۰	۷۶٫۹
مصالحه‌ای	۰٫۰۱۲۸	۰٫۰۸۰۲	۷۷٫۶
متناسب	۰٫۰۱۳۱	۰٫۰۸۳۷	۸۰٫۹

* com مخفف اصطلاح compromise به‌معنای مصالحه است.

۱۴.۴ ساختن طبقات

در این بخش، سؤالهایی را به‌صورت زیر مطرح می‌کنیم. بهترین مشخصه برای ساختن طبقات چیست؟ چگونه می‌توان کرانهای طبقات را تعیین کرد؟

واضح است که بهترین مشخصه متغیر Y، توزیع فراوانی آن است. بعد از آن توزیع فراوانی متغیری کمکی است که همبستگی بسیار قوی با متغیر Y دارد. دالنیوس[1]، با فرض معلوم بودن تعداد طبقات، برای تعیین بهترین کرانهای طبقات تحت تخصیص متناسب و تخصیص نیمن، معادلاتی را ارائه داده است. روشهای تقریبی سریعتر دیگری نیز به‌وسیلهٔ پژوهشگران دیگر عرضه شده‌اند. ما در اینجا تخصیص نیمن را به‌دلیل کاربرد فراوان آن، و به‌دلیل اینکه عموماً بهتر از تخصیص متناسب عمل می‌کند در نظر می‌گیریم. بدواً می‌پذیریم که طبقات با استفاده از مقادیر متغیر Y ساخته می‌شوند. ابتدا راهی نظری و سپس طریقی کاربردی را بررسی می‌کنیم.

فرض کنید $Y_۰$ کوچکترین و Y_L بزرگترین مقدار Y در جامعه باشند. مسألهٔ مورد نظر، یافتن کرانهای طبقاتی یعنی مقادیر $Y_۱$، $Y_۲$، $\ldots$ ، $Y_{L-۱}$ هستند به‌قسمی که L طبقه به‌وجود آید و ضمناً واریانس برآوردکنندهٔ میانگین جامعه، یعنی

$$\mathrm{V}(\overline{Y}_{st}) = \frac{۱}{n}\left(\sum_{h=۱}^{L} W_h S_h\right)^۲ - \frac{۱}{N}\sum_{h=۱}^{L} W_h S_h^۲$$

مینیمم شود. اگر $\frac{۱}{N}$ قابل اغماض باشد می‌توان از جملهٔ دوم طرف دوم صرف‌نظر کرد، لذا کافی است که $Y_۱$، $Y_۲$، $\ldots$ ، $Y_{L-۱}$ را که مشخص‌کنندهٔ کران طبقه‌ها هستند به‌قسمی بیابیم که $\sum W_h S_h$ مینیمم شود. در این مجموع، چون Y_h در بازه‌های $[Y_{h-۱},\ Y_h]$ و $[Y_h,\ Y_{h+۱}]$ ظاهر می‌شود لذا Y_h در جمله‌های $W_h S_h$ و $W_{h+۱}S_{h+۱}$ ظاهر خواهد شد. این دوبازه به‌ترتیب طبقه‌های hام و $h+۱$ام را مشخص می‌کنند. با این توضیح داریم

$$\frac{\partial}{\partial Y_h}\left(\sum W_h S_h\right) = \frac{\partial}{\partial Y_h}(W_h S_h) + \frac{\partial}{\partial Y_h}(W_{h+۱}S_{h+۱}) \qquad (۶۰.۴)$$

سعی می‌کنیم مقادیر دو جملهٔ عبارت سمت راست (۶۰.۴) را حساب کنیم. برای این منظور فرض می‌کنیم $f_Y(y)$ تابع چگالی یا تابع فراوانی نسبی Yها باشد. اگر t معرف Yهایی باشد که در طبقه hام قرار دارند آن‌گاه وزن طبقه hام، همان احتمال مربوط به این طبقه است، یعنی

$$W_h = \int_{Y_{h-۱}}^{Y_h} f(t)dt = F(Y_h) - F(Y_{h-۱})$$

که $F(t)$ تابع توزیع متناظر با $f(t)$ است، لذا

$$\frac{\partial W_h}{\partial Y_h} = f(Y_h) \qquad (۶۱.۴)$$

1. Dalenius

میانگین Yهای بازهٔ (Y_{h-1}, Y_h)، یعنی میانگین Yهای واقع در طبقهٔ hام، با توجه به اینکه تابع چگالی در این بازه $f(t)/W_h$ است عبارت است از

$$\mu_h = \int_{Y_{h-1}}^{Y_h} t\,\frac{f(t)}{W_h}dt$$

چون t معرف Yهای طبقه hام، و $\frac{f(t)}{W_h}$ معرف تابع چگالی متناظر با این طبقه است، پس S_h^2 که می‌توان آن را تقریباً متناظر با واریانس t دانست به‌صورت زیر است

$$\begin{aligned} S_h^2 &= E(t-E(t))^2 = E(t^2) - [E(t)]^2 \\ &= \int_{Y_{h-1}}^{Y_h} t^2\frac{f(t)}{W_h}dt - \left[\int_{Y_{h-1}}^{Y_h} t\frac{f(t)}{W_h}dt\right]^2 \end{aligned}$$

بنابراین

$$W_hS_h^2 = \int_{Y_{h-1}}^{Y_h} t^2 f(t)dt - \frac{\left[\int_{Y_{h-1}}^{Y_h} tf(t)dt\right]^2}{W_h}$$

با یادآوری دستور مشتقگیری ازانتگرالها نسبت به پارامتری که در حدود انتگرال و عبارت زیر انتگرال وجود دارد، یعنی با یادآوری فرمول

$$\frac{\partial}{\partial\theta}\int_{\alpha(\theta)}^{\beta(\theta)} f(x,\theta)dx = f[\beta(\theta),\theta]\cdot\beta'(\theta) - f[\alpha(\theta),\theta]\alpha'(\theta) + \int_{\alpha(\theta)}^{\beta(\alpha)}\frac{\partial}{\partial\theta}f(x,\theta)dx$$

از طرفین رابطهٔ بالا به‌نسبت به Y_h مشتق می‌گیریم

$$\begin{aligned} &S_h^2\frac{\partial W_h}{\partial Y_h} + 2W_hS_h\cdot\frac{\partial S_h}{\partial Y_h} \\ &= Y_h^2 f(Y_h) - \frac{2\left[\int_{Y_{h=1}}^{Y_h} tf(t)dt\right]Y_hf(Y_h)\cdot W_h - \frac{\partial W}{\partial Y_h}\left[\int_{Y_{h-1}}^{Y_h} tf(t)dt\right]^2}{W_h^2} \\ &= Y_h^2f(Y_h) - 2Y_hf(Y_h)\frac{\int_{Y_{h-1}}^{Y_h} tf(t)dt}{W_h} + \left[\int_{Y_{h-1}}^{Y_h} tf(t)dt/W_h\right]^2\cdot f(Y_h) \end{aligned}$$

و یا با توجه به مقدار μ_h

$$S_h^2\frac{\partial W_h}{\partial Y_h} + 2W_hS_h\frac{\partial S_h}{\partial Y_h} = Y_h^2f(Y_h) - 2Y_hF(Y_h)\mu_h + f(Y_h)\cdot\mu_h^2 \quad (۶۲.۴)$$

با توجه به (۶۱.۴)، به دو طرف رابطهٔ (۶۲.۴) مقدار ثابت $S_h^2 f(Y_h) = S_h^2 \frac{\partial W_h}{\partial Y_h}$ را نظیر به نظیر اضافه می‌کنیم. نتیجه می‌شود

$$2S_h^2 \frac{\partial W}{\partial Y_h} + 2W_h S_h \frac{\partial S_h}{\partial Y_h} = Y_h^2 f(Y_h) - 2Y_h \mu_h f(Y_h) + \mu_h^2 f(Y_h) + S_h^2 f(Y_h)$$

اگر طرفین را بر $2S_h$ تقسیم کنیم پس از خلاصه کردن، داریم

$$S_h \frac{\partial W}{\partial Y_h} + W_h \frac{\partial S_h}{\partial Y_h} = \frac{1}{2S_h} f(Y_h)[Y_h^2 - 2Y_h\mu_h + \mu_h^2 + S_h^2]$$

یا

$$\frac{\partial (W_h S_h)}{\partial Y_h} = \frac{1}{2} f(Y_h) \frac{(Y_h - \mu_h)^2 + S_h^2}{S_h} \qquad (۶۳.۴)$$

به‌طریقی مشابه می‌توان نتیجه گرفت که

$$\frac{\partial (W_{h+1} S_{h+1})}{\partial Y_h} = -\frac{1}{2} f(Y_h) \frac{(Y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}} \qquad (۶۴.۴)$$

حال اگر بخواهیم که $V(\overline{Y}_{st})$ مینیمم شود باید با توجه به (۶۰.۴) داشته باشیم

$$\frac{\partial}{\partial Y_h}(W_h S_h) + \frac{\partial}{\partial Y_h}(W_{h+1} S_{h+1}) = 0$$

یا

$$\frac{\partial}{\partial Y_h}(W_h S_h) = -\frac{\partial}{\partial Y_h}(W_{h+1} S_{h+1})$$

بنابر رابطهٔ (۶۳.۴) و(۶۴.۴) نتیجه می‌شود که

$$\frac{(Y_h - \mu_h)^2 + S_h^2}{S_h} = \frac{(Y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}} \qquad (h = 1, 2, \dots, L-1)$$

بدین طریق دستگاهی با $L - 1$ معادله به‌دست می‌آید که در صورت معلوم بودن برآوردهایی مقدماتی از S_h، S_{h+1}، و μ_h، مقادیر Y_h را می‌توان یافت. متأسفانه این معادلات به‌دلیل وابستگی S_h و μ_h به Y_h کاربرد عملی ندارند. دالنیوس و هاجز روشی تقریبی و سریع برای مینیمم کردن $\sum W_h S_h$ به‌شرح زیر ارائه کرده‌اند. قرار می‌دهیم

$$Z_h = Z(Y_h) = \int_{Y_0}^{Y_h} \sqrt{f(t)}\,dt \qquad (۶۵.۴)$$

اگر تعداد طبقات زیاد بوده و عرض آنها کم باشد $f(y)$ را می‌توان تقریباً در هر طبقه ثابت گرفت، یعنی توزیع Y را در هر طبقه توزیع یکنواخت فرض کرد. بنابراین

$$W_h = \int_{Y_{h-1}}^{Y_h} f(t)dt \simeq f_h(Y_h - Y_{h-1}) \qquad (66.4)$$

از طرفی می‌دانیم که اگر متغیر Y روی فاصله $(a,\ b)$ دارای توزیع یکنواخت باشد آن‌گاه

$$\mathrm{V}(Y) = \frac{(b-a)^2}{12}$$

چون در اینجا فرض کرده‌ایم که Y روی فاصله $(Y_{h-1},\ Y_h)$ توزیع یکنواخت دارد. لذا به‌طور تقریبی

$$S_h^2 \simeq \frac{(Y_h - Y_{h-1})^2}{12}$$

و لذا

$$S_h \simeq \frac{(Y_h - Y_{h-1})}{\sqrt{12}} \qquad (67.4)$$

با توجه به (۶۵.۴) می‌توان نوشت

$$Z_h = \int_{Y_0}^{Y_h} \sqrt{f(t)}dt \quad , \quad Z_{h-1} = \int_{Y_0}^{Y_{h-1}} \sqrt{f(t)}dt$$

پس

$$Z_h - Z_{h-1} = \int_{Y_{h-1}}^{Y_h} \sqrt{f(t)}dt$$

چون توزیع در فاصلهٔ $(Y_{h-1},\ Y_h)$ یکنواخت است

$$Z_h - Z_{h-1} \simeq \sqrt{f_h}(Y_h - Y_{h-1}) \qquad (68.4)$$

در این رابطه و در (۶۶.۴)، f_h مقدار ثابت $f(Y)$ در طبقهٔ hام است. از (۶۸.۴) داریم

$$\sum_{h=1}^{L}(Z_h - Z_{h-1})^2 \simeq \sum_{h=1}^{L} f_h(Y_h - Y_{h-1})^2$$
$$= \sum_{h=1}^{L} f_h(Y_h - Y_{h-1})(Y_h - Y_{h-1})$$

با توجه به (۶۶.۴)

$$\sum_{h=۱}^{L}(Z_h - Z_{h-۱})^۲ \simeq \sum_{h=۱}^{L} W_h(Y_h - Y_{h-۱})$$

اگر به‌جای $Y_h - Y_{h-۱}$ مقدار آن را از (۶۷.۴) قرار دهیم نتیجه می‌شود که

$$\sum_{h=۱}^{L}(Z_h - Z_{h-۱})^۲ \simeq \sqrt{۱۲}\sum_{h=۱}^{L} W_h S_h = \sum_{h=۱}^{L} f_h(Y_h - Y_{h-۱})^۲ \qquad (۶۹.۴)$$

برای اینکه $\sum W_h S_h$ مینیمم شود باید مقدار

$$\sum_{h=۱}^{L}(Z_h - Z_{h-۱})^۲$$

مینیمم شود. مجموع h فاصلهٔ طبقاتی مقداری ثابت است، زیرا

$$\sum_{h=۱}^{L}(Z_h - Z_{h-۱}) = Z_L - Z_۰$$

اینک h فاصله به‌صورت $Z_h - Z_{h-۱}$ داریم که طول هر کدام مثبت و مجموع آنها مقداری ثابت است، بنابراین $\sum_{h=۱}^{L}(Z_h - Z_{h-۱})^۲$ وقتی مینیمم است که $Z_h - Z_{h-۱}$، یعنی مقدار

$$\sqrt{f_h}(Y_h - Y_{h-۱}), \qquad h = ۱,\ ۲,\ \ldots, L$$

به‌ازای هر مقدار h، ثابت باشد. لذا اگر فاصله‌های طبقه‌ها را به‌صورتی اختیار کنیم که این مقادیر در هر طبقه تقریباً همانند باشند، کرانهای طبقه‌ها به‌گونه‌ای خواهند بود که واریانس برآوردکنندهٔ میانگین جامعه مینیمم باشد.

با توضیحات بالا دالنیوس و هاجز قاعدهٔ عملی زیر را ارائه می‌دهند:

اگر $f(Y)$ معلوم باشد، مقادیر تجمعی $\sqrt{f(Y)}$ را محاسبه می‌کنیم. سپس Y_h ها را به‌قسمی می‌یابیم که بر دامنهٔ مقادیر $\sqrt{f(Y)}$ فاصله‌های برابر ایجاد کنند. باید توجه داشت که برای تعیین کرانها ضروری است که f یعنی تابع توزیع فراوانی جامعه معلوم باشد. مثال زیر نحوهٔ استفاده از این قاعده را نشان می‌دهد.

مثال ۷.۴ داده‌های جدول ۱، توزیع فراوانی درصد وامهای بانکی را در ۱۳۴۳۸ شعبهٔ بانک کشور آمریکا نشان می‌دهد که به وامهای صنعتی تخصیص یافته‌اند. توزیع چاوله است و مُد آن در دم چپ قرار دارد. در ستون مربوط به مقادیر تجمعی $\sqrt{f}$، به‌عنوان مثال $۵۸٫۹ = \sqrt{۳۴۳۶}$،۰ و $۱۰۹٫۱ = \sqrt{۲۵۱۶} + \sqrt{۳۴۶۴}$،۰ و نظایر آن. فرض کنید مایلیم ۵ طبقه داشته باشیم. چون

مجموع مقادیر تجمعی $\sqrt{f}$ برابر با ۳۹۵٫۵ است، نقاط تقسیم باید در ۷۷٫۹، ۱۵۵٫۸، ۲۳۳٫۷، و ۳۱۱٫۶ باشند. نزدیکترین نقاط موجود در جدول ۲ آمده‌اند.

جدول ۱. محاسبهٔ کرانهای طبقات به‌وسیلهٔ قاعدهٔ $\sqrt{f}$ تجمعی*

مقدار تجمعی $\sqrt{f(t)}$	$f(Y)$	٪ (وامهای صنعتی / کل وامها)	مقدار تجمعی $\sqrt{f(Y)}$	$f(Y)$	٪ (وامهای صنعتی / کل وامها)
۳۴۰٫۳	۱۲۶	۵۰-۵۵	۵۸٫۹	۳۴۶۴	۰-۵
۳۵۰٫۶	۱۰۷	۵۵-۶۰	۱۰۹٫۱	۲۵۱۶	۵-۱۰
۳۵۹٫۷	۸۲	۶۰-۶۵	۱۵۵٫۵	۲۱۵۷	۱۰-۱۵
۳۶۶٫۸	۵۰	۶۵-۷۰	۱۹۵٫۳	۱۵۸۱	۱۵-۲۰
۳۷۳٫۰	۳۹	۷۰-۷۵	۲۲۹٫۱	۱۱۴۲	۲۰-۲۵
۳۷۸٫۰	۲۵	۷۵-۸۰	۲۵۶٫۴	۷۴۹	۲۵-۳۰
۳۸۲٫۰	۱۶	۸۰-۸۵	۲۷۹٫۰	۵۱۲	۳۰-۳۵
۳۸۶٫۴	۱۹	۸۵-۹۰	۲۹۸٫۴	۳۷۶	۳۵-۴۰
۳۸۷٫۸	۲	۹۰-۹۵	۳۱۴٫۷	۲۶۵	۴۰-۴۵
۳۸۹٫۵	۳	۹۵-۱۰۰	۳۲۹٫۱	۲۰۷	۴۵-۵۰

* این مثال از مرجع [۳] اقتباس شده است.

جدول ۲. نزدیکترین نقاط موجود به‌نقاط تقسیم

	طبقه				
	۱	۲	۳	۴	۵
کرانها	٪۰-۵	٪۵-۱۵	٪۱۵-۲۵	٪۲۵-۴۵	٪۴۵-۱۰۰
بازهٔ مربوط به $\sqrt{f}$ تجمعی	۵۸٫۹	۹۶٫۶	۷۳٫۶	۸۵٫۶	۷۴٫۸

▲

توجه. رابطهٔ (۶۹.۴) یک پیامد فوری و جالب دارد: اگر W_hS_h ثابت باشد، تخصیص نیمن، حجم نمونه‌ای ثابت $n_h = \frac{n}{L}$ را برای همهٔ طبقه‌ها به‌دست می‌دهد. در مقایسه با روشهای دیگر، تخصیص $n_h = \frac{n}{L}$ در این حالت روشی رضایتبخش است. برای مطالعهٔ بیشتر در زمینهٔ ساختن طبقه‌ها، خواننده را به مرجع [۳] ارجاع می‌دهیم.

۱۵.۴ طبقه‌بندی بعد از انتخاب نمونه

گاهی اوقات وقتی مایلیم طبقه‌بندی را برحسب متغیری کلیدی انجام دهیم با این مشکل روبه‌رو هستیم که نمی‌توانیم واحدها را تا بعد از انتخاب در طبقهٔ صحیح آنها قرار دهیم. مثلاً اگر بخواهیم نتیجهٔ نظرخواهی خاصی را برحسب جنس نظردهنده طبقه‌بندی کنیم و نظرخواهی را با مصاحبهٔ تلفنی انجام دهیم قبل از تماس با فرد نمی‌توانیم پاسخگو را در طبقهٔ زنان یا مردان منظور کنیم.

همین‌طور اگر حسابرسی بخواهد حسابهایی را که با نمونه‌گیری برای حسابرسی انتخاب می‌کند در دو طبقه عمده فروش و خرده فروش قرار دهد تا بعد از انتخاب حساب و بررسی آن نمی‌تواند این طبقه‌بندی را انجام دهد.

فرض کنید نمونهٔ تصادفی ساده‌ای به حجم n برای نظرخواهی انتخاب شود. پاسخگویان نمونه را پس از انتخاب می‌توان به دو قسمت $n_۱$ مرد و $n_۲$ زن تقسیم کرد. اگر بخواهیم $\overline{Y}_N$ جامعه را به‌وسیله $\overline{Y}_{st}$ برآورد کنیم باید $\frac{N_۱}{N}$ و $\frac{N_۲}{N}$ را بدانیم. توجه کنید که در این حالت $n_۱$ و $n_۲$ دو متغیر تصادفی‌اند و از قبل نمی‌دانیم که در n تماس تلفنی با پاسخگویان چند پاسخگوی زن و چند پاسخگوی مرد خواهیم داشت. البته مجموع این دو متغیر تصادفی، مساوی مقدار ثابت n است. این نمونه‌گیری گرچه به‌ظاهر نمونه‌گیری با طبقه‌بندی است ولی درواقع به‌دلیل تصادفی بودن مقادیر $n_۱$ و $n_۲$ با تعریف نمونه‌گیری با طبقه‌بندی مطابقت نمی‌کند. می‌توان نشان داد که اگر $\frac{N_h}{N}$ معلوم، و برای هر طبقه $n_h \geq ۲۰$، آن‌گاه طبقه‌بندی پس از انتخاب نمونه که گاهی آن را طبقه‌بندی پسین می‌گویند تقریباً همان دقت نمونه‌گیری تصادفی با طبقه‌بندی در تخصیص متناسب را دارد [۱۵].

طبقه‌بندی پس از انتخاب نمونه، غالباً وقتی نمونهٔ تصادفی ساده دقیقاً برحسب گروهبندیهای جامعه‌ای حالت تعادلی ندارد، انتخابی مناسب است. فرض کنید مثلاً نمونهٔ تصادفی ساده‌ای به حجم $n = ۱۰۰$ از جامعه‌ای انتخاب کرده‌ایم که باید به‌طور برابر بین زن و مرد تقسیم شود. صفت مورد نظر، وزن پاسخگویان به کیلوگرم است و هدف نمونه‌گیری برآورد کردن متوسط وزن در جامعه است. نمونهٔ منتخب اطلاعات زیر را فراهم کرده است:

مرد	زن
$n_۱ = ۲۰$	$n_۲ = ۸۰$
$\overline{y}_۱ =$ ۹۰ کیلوگرم	$\overline{y}_۲ =$ ۵۵ کیلوگرم

$$\overline{Y} = \frac{۲۰}{۱۰۰}(۹۰) + \frac{۸۰}{۱۰۰}(۵۵) = ۶۲$$

با توجه به اینکه تعداد مردان پاسخگو نسبت به تعداد زنان پاسخگو کم است به‌نظر می‌رسد که برآورد $\overline{Y} = ۶۲$ برآوردی کمتر از واقعیت است. می‌توانیم این برآورد را به‌صورت زیر تصحیح کنیم

$$\overline{Y}_{st} = \frac{N_۱}{N}\overline{y}_۱ + \frac{N_۲}{N}\overline{y}_۲ = \frac{۱}{۲}(۹۰) + \frac{۱}{۲}(۵۵) = ۷۲٫۵$$

این برآورد به‌نظر بیشتر واقعی می‌آید زیرا در این محاسبه به زنان و مردان وزنهای برابر نسبت داده شده است. توجه دارید که گرچه $N_۱$، $N_۲$، و N نامعلوم‌اند ولی از قبل فرض شده است که نسبت دو جنس برابرند. مسلماً این $\overline{Y}_{st}$ واریانسی دارد که با واریانس مربوط به $\overline{Y}_{st}$ در نمونه‌گیری تصادفی با طبقه‌بندی تفاوت دارد، زیرا نمونه‌گیری با طبقه‌بندی، قبل از انتخاب نمونه طرّاحی نشده است. اما، واریانس تقریبی را می‌توان به‌صورت زیر به‌دست آورد.

مثل سابق $\frac{N_h}{N} = W_h$، $h = ۱, \ldots, L$، ولی n_hها تصادفی‌اند و داریم

$$\begin{cases} E(n_h) = nW_h \qquad h = ۱,\ ۲, \ldots, L \\ \mathrm{V}(n_h) = nW_h(۱ - W_h) \end{cases}$$

اگر n_h ثابت می‌بود

$$\begin{aligned} \hat{\mathrm{V}}(\overline{Y}_{st}) &= \sum_{h=۱}^{L} W_h^۲ \frac{N_h - n_h}{N_h} \frac{s_h^۲}{n_h} \\ &= \sum_{h=۱}^{L} W_h^۲ \frac{s_h^۲}{n_h} - \frac{۱}{N} \sum_{h=۱}^{L} W_h s_h^۲ \end{aligned} \qquad (۷۰.۴)$$

اما n_h تصادفی است. در این وضعیت یک تقریب کلی از $\hat{\mathrm{V}}(\overline{Y}_{st})$ بدین‌طریق به‌دست می‌آید که به‌جای $\frac{۱}{n_h}$ مقدار امید آن را قرار دهیم. متأسفانه تعیین امید عکس یک متغیر تصادفی در حالت کلی ساده نیست، اما می‌توان نشان داد که تقریبی خوب به‌صورت زیر است

$$E\left(\frac{۱}{n_h}\right) \simeq \frac{۱}{nW_h} + \frac{۱ - W_h}{n^۲ W_h^۲} \qquad (۷۱.۴)$$

اگر (۷۱.۴) را در (۷۰.۴) قرار دهیم، نتیجه می‌شود

$$\begin{aligned} \hat{\mathrm{V}}_P(\overline{Y}_{st}) &= \frac{۱}{n} \sum_{h=۱}^{L} W_h s_h^۲ + \frac{۱}{n^۲}(۱ - W_h)s_h^۲ - \frac{۱}{N} \sum_{h=۱}^{L} W_h s_h^۲ \\ &= \left(\frac{۱}{n} - \frac{۱}{N}\right) \sum_{h=۱}^{L} W_h s_h^۲ + \frac{۱}{n^۲} \sum_{h=۱}^{L} (۱ - W_h) s_h^۲ \end{aligned}$$

و سرانجام

$$\hat{\mathrm{V}}_P(\overline{Y}_{st}) = \frac{N - n}{Nn} \sum_{h=۱}^{L} W_h s_h^۲ + \frac{۱}{n^۲} \sum_{h=۱}^{L} (۱ - W_h) s_h^۲ \qquad (۷۲.۴)$$

که در آن، زیرنویس P نمادی برای طبقه‌بندی پسین* است. در (۷۲.۴) جملهٔ اول، واریانس میانگین نمونهٔ طبقه‌بندی تحت تخصیص متناسب است. جملهٔ دوم همیشه نامنفی است و مقدار افزایش واریانسی راکه می‌توان به‌جای طبقه‌بندی پیشین از طبقه‌بندی پسین انتظار داشت نشان می‌دهد. توجه کنید که در مخرج کسر جملهٔ $n^۲$ قرار دارد و معنای آن است که این جمله معمولاً خیلی کوچک است.
به‌طور خلاصه، تقریب (۷۱.۴) تنها وقتی خوب است که n بزرگ و n_h مثبت باشد. همین‌طور، مقدار افزایش واریانس، به‌شرط بزرگ بودن n، کوچک خواهد بود. پس، طبقه‌بندی پسین تنها وقتی

* حرف P نمادی برای Poststrafication، به‌معنای طبقه‌بندی پسین، است.

نتایج خوبی می‌دهد که n بزرگ و تمام n_hها نیز نسبتاً بزرگ باشند. یک پیامد کاربردی از این نتیجه آن است که طبقه‌بندی پسین را نمی‌توان در طبقات زیاد انجام داد.

مثال ۸.۴ مدیر شرکتی می‌داند که ۴۰٪ حسابهای قابل وصولش از عمده‌فروشیها و ۶۰٪ بقیه از خرده‌فروشیها به‌دست می‌آیند. اما، مشخص کردن حسابهای فردی بدون بررسی هر پرونده و مطالعهٔ آن، مشکل است. حسابرسی می‌خواهد که نمونه‌ای به حجم $n = ۱۰۰$ از بین حسابها، برای برآورد متوسط مقدار حسابهای قابل وصول شرکت تهیه کند. نمونه‌ای تصادفی استخراج می‌کند که بعد از بررسی، ۷۰٪ واحدهای نمونه مربوط به عمده‌فروشیها و ۳۰٪ مربوط به خرده‌فروشیهاست. پس از انجام طبقه‌بندی روی این نمونه، نتایج نمونه‌ای زیر حاصل شده‌اند

خرده‌فروشی	عمده‌فروشی
$n_۲ = ۳۰$	$n_۱ = ۷۰$
$\bar{y}_۲ = ۲۸۰$	$\bar{y}_۱ = ۵۲۰$
$s_۲ = ۹۰$	$s_۱ = ۲۱۰$

پارامتر μ، متوسط مقدار حسابهای قابل وصول شرکت را برآورد کنید.

چون نسبت مشاهده‌شده حسابهای عمده‌فروشی یعنی ۷ر۰، از نسبت واقعی این حسابها در جامعه، یعنی ۴ر۰ خیلی فاصله دارد، به‌نظر می‌رسد که طبقه‌بندی بعد از نمونه‌گیری تصادفی طبقه‌بندی نامناسب است. می‌دانیم

$$\overline{Y}_{st} = \left(\frac{N_۱}{N}\right)\bar{y}_۱ + \left(\frac{N_۲}{N}\right)\bar{y}_۲ = (۰{,}۴)(۵۲۰) + (۰{,}۶)(۲۸۰) = ۳۷۶$$

با استفاده از (۷۲.۴) و با اغماض از ضریب تصحیح جامعهٔ متناهی، داریم

$$\begin{aligned}\hat{\mathrm{V}}_P(\overline{Y}_{st}) &= \frac{۱}{n}\sum_{i=۱}^{۲} W_h s_h^۲ + \frac{۱}{n^۲}\sum_{h=۱}^{۲}(۱ - W_h)s_h^۲ \\ &= \frac{۱}{۱۰۰}[۰{,}۴(۲۱۰)^۲ + ۰{,}۶(۹۰)^۲] + \frac{۱}{(۱۰۰)^۲}[۰{,}۶(۲۱۰)^۲ + ۰{,}۴(۹۰)^۲] \\ &= ۲۲۵ + ۲{,}۹۷ = ۲۲۷{,}۹۷\end{aligned}$$

در عبارت $\hat{\mathrm{V}}_P(\overline{Y}_{st})$، اولین جمله، مقداری است که اگر نمونه را از قبل طبقه‌بندی می‌کردیم (و این نتایج نمونه‌ای حاصل می‌شد) به‌دست می‌آمد. جملهٔ دوم بهایی است که برای عدم طبقه‌بندی از قبل می‌پردازیم. ▲

۱۶.۴ نمونه‌گیری مضاعف برای طبقه‌بندی

تا اینجا فرض بر این بود که مقادیر $W_h = \frac{N_h}{N}$، L ،$h = ۱, ۲, \ldots$، قبل از شروع نمونه‌گیری، مقادیر ثابت معلومی هستند. ولی همیشه این‌طور نیست. مثلاً ممکن است بخواهیم جامعهٔ رأی‌دهندگان

را برحسب جنس، سطح درآمد، یا سطح تحصیلات طبقه‌بندی کنیم ولی از روی فهرست ثبت اسامی رأی‌دهندگان، اطلاعات لازم برای طبقه‌بندی این صفات موجود نیست.

اندیشهٔ اصلی نمونه‌گیری مضاعف (نمونه‌گیری دو مرحله‌ای) برای طبقه‌بندی، نسبتاً ساده است، ولی برآورد واریانس میانگین در این حالت پیچیده است. فرض کنید اطلاعات مقدماتی نظیر جنس رأی‌دهندگان برای طبقه‌بندی به‌آسانی به‌دست آیند اما اطلاعات مفصلتر، مانند نظر سیاسی آنها یا علت مخالفت آنها با موضوع رأی‌گیری به‌آسانی در اختیار نبوده و نیاز به مصاحبه داشته باشد. بدیهی است انجام مصاحبه با همهٔ رأی‌دهندگان کاری وقتگیر، پرهزینه و اصولاً نشدنی است. در این صورت نمونه‌ای بزرگ برای تشکیل طبقات و نمونه‌ای خیلی کوچکتر برای گردآوری داده‌ها اختیار می‌کنند. مثلاً برای مشخص کردن جنس رأی‌دهندگان نمونه‌ای بزرگ انتخاب می‌کنیم (مرحلهٔ اول) و برای تکمیل اطلاعات به تصادف با عده‌ای مصاحبه می‌نماییم (مرحلهٔ دوم).

فرض کنید نمونهٔ مرحلهٔ اول به حجم n' برای تعیین اینکه واحدها در کدام طبقات می‌افتند مورد استفاده قرار گیرد. گیریم

$$w'_h = \frac{n'_h}{n'} \qquad h = 1, \ldots, L$$

نسبت اولین بخشی از نمونه باشد که در طبقه hام می‌افتد. واضح است که w'_h برآوردکنندهٔ نااریب W_h است (چرا؟). بدیهی است نمونهٔ اول به‌تصادف انتخاب شده است.

در مرحلهٔ دوم، از n'_h واحد متعلق به طبقه hام، به‌صورتی تصادفی، n_h واحد انتخاب می‌کنیم. اندازه‌های تحت بررسی (اطلاعات مفصلتر) را از این n_h واحد به‌دست می‌آوریم و میانگین اندازه‌ها را $\overline{Y}_h$ و میزان تغییر آنها را s_h می‌گیریم. این عمل را برای همهٔ L طبقه انجام می‌دهیم. اگر $\overline{Y}_N$ میانگین جامعه‌ای این صفت باشد، داریم

$$\hat{\overline{Y}}_N = \overline{Y}'_{st} \simeq \sum_{h=1}^{L} w'_h \overline{Y}_h$$

اگر کسرهای نمونه‌گیری مرحلهٔ دوم در هر طبقه، یعنی $\frac{n_h}{N_h}$ کوچک باشند و N بزرگ فرض شود، واریانس تقریبی $\overline{Y}_{st}$ به‌صورت زیر برآورد می‌شود

$$\hat{\mathrm{V}}(\overline{Y}'_{st}) = \frac{n'}{n'-1} \sum_{h=1}^{L} \left(w'^{2}_h - \frac{w'_h}{n'} \right) \frac{s^2_h}{n_h} + \frac{w'_h(\overline{Y}_h - \overline{Y}'_{st})^2}{n'} \Bigg]$$

(برای اثبات به [۳] رجوع شود)

اگر n' آن‌قدر بزرگ باشد که w'/n'_h قابل اغماض باشد این برآورد واریانس به‌صورت زیر خلاصه می‌شود

$$\hat{\mathrm{V}}(\overline{Y}'_{st}) = \sum_{h=1}^{L} \left[\frac{w'^{2}_h s^2_h}{n_h} + \frac{w'_h(\overline{Y}_h - \overline{Y}'_{st})^2}{n'} \right] \qquad (۷۳.۴)$$

برآورد واریانس $\overline{Y}_{st}$ را در حالت کلی به‌خاطر بیاورید. این برآورد واریانس را به‌صورت زیر می‌توان نوشت

$$\hat{V}(\overline{Y}_{st}) = \frac{۱}{N^۲}\sum_{h=۱}^{L} N_h^۲\left(\frac{N_h - n_h}{N_h}\right)\left(\frac{s_h^۲}{n_h}\right) \qquad (۷۴.۴)$$

اگر (۷۴.۴) را با (۷۳.۴) مقایسه کنیم می‌بینیم که صرف‌نظر از تصحیح جامعهٔ متناهی، قسمت اول سمت راست برابری (۷۳.۴) با سمت راست (۷۴.۴) یکی است، که در آن به‌جای $\frac{N_h}{N}$ یعنی W_h مقدار w'_h گذاشته شده است و قبلاً هم گفتیم که w'_h برآوردکنندهٔ ناااریب W_h است. جملهٔ دوم (۷۳.۴) مؤلفهٔ اضافی واریانس است که به‌دلیل عدم اطلاع از وزنهای دقیق طبقات به‌وجود آمده است. مطلب را با مثالی تشریح می‌کنیم.

مثال ۹.۴ می‌خواهند از روی فهرست ثبت نامها و حجم دانشکده‌های کشوری، متوسط تعداد ثبت نام را در سال معینی برآورد کنند. دانشکده‌های خصوصی، ثبت نامهایی کمتر از دانشکده‌های دولتی دارند. لذا دو طبقهٔ خصوصی و دولتی در نظر گرفته شده‌اند. تعداد دانشکده‌های خصوصی و دولتی را نمی‌دانند. ولی تشخیص آن با انتخاب هر واحد نمونه سریعاً مشخص می‌شود. از نمونه‌ای به حجم $n' = ۱۴۱$ دانشکده نتیجه شده است که

$$\text{خصوصی} : n'_۱ = ۸۴ \qquad \text{دولتی} : n'_۲ = ۵۷$$

زیرنمونه‌هایی به حجم ۱۱ دانشکدهٔ خصوصی و ۱۲ دانشکدهٔ دولتی به‌تصادف از نمونهٔ بالا انتخاب کرده، تعداد ثبت نامهای آنها را در سال مورد نظر تعیین نموده‌اند. داده‌ها در جدول زیر آمده‌اند.

دولتی $n_۲ = ۱۲$		خصوصی $n_۱ = ۱۱$	
دانشکده	تعداد ثبت نام	دانشکده	تعداد ثبت نام
۴۵۲	۷۳۳۲	۱۲۲	۱۶۱۸
۱۳۱	۲۳۵۶	۸۸	۱۱۴۰
۹۹۶	۲۱۸۷۹	۶۵	۱۰۰۰
۵۰	۹۳۶	۵۵	۱۲۲۵
۱۰۶	۱۲۹۳	۷۹	۷۹۱
۳۲۶	۵۸۹۴	۷۹	۱۶۰۰
۵۰۶	۸۵۰۰	۴۰	۷۴۶
۳۷۱	۶۴۹۱	۷۵	۱۷۰۱
۱۰۷	۷۸۱	۳۲	۷۰۱
۲۹۸	۷۲۵۵	۴۲۸	۶۹۱۸
۱۲۸	۲۱۳۶	۱۱۰	۱۰۵۰
۲۸۰	۵۳۸۰		

برآورد میانگین تعداد ثبت نامها و برآورد واریانس برآوردکنندهٔ میانگین را محاسبه کنید.
بنابر آنچه در بالا گفتیم

$$\overline{Y}'_{st} = w'_1\overline{Y}_1 + w'_2\overline{Y}_2$$

در این مثال که داده‌ها واقعی هستند W_1 و W_2 را درکل جامعه نداریم. w'_1 و w'_2 را از روی نمونه اول بدست می‌آوریم

$$w'_1 = \frac{۸۴}{۱۴۱} \quad . \quad w'_2 = \frac{۵۷}{۱۴۱}$$

این دو مقدار برآورد نااریب W_1 و W_2 هستند. با توجه به داده‌های جدول با محاسبهٔ $\overline{Y}_1$، $\overline{Y}_2$ و s_1، s_2، به‌ترتیب داریم

$$\overline{Y}'_{st} = \left(\frac{۸۴}{۱۴۱}\right)(۱۶۸۱) + \left(\frac{۵۷}{۱۴۱}\right)(۵۸۵۳)$$
$$= (۰{,}۶۰)(۱۶۸۱) + (۰{,}۴۰)(۵۸۵۳) = ۳۳۴۹{,}۸$$

و

$$\hat{V}(\overline{Y}'_{st}) = \frac{1}{n_1}(w'_1 s_1)^2 + \frac{1}{n_2}(w'_2 s_2)^2 + \frac{1}{n'}[w'_1(\overline{Y}_1 - Y'_{st})^2 + w'_2(\overline{Y}_2 - Y'_{st})^2]$$

برای نوشتن این رابطه از (۷۳.۴) استفاده کرده‌ایم. پس

$$\hat{V}(\overline{Y}'_{st}) = \frac{۱}{۱۱}[(۰{,}۶۰)(۱۷۷۳)]^2 + \frac{۱}{۱۲}[(۰{,}۴۰)(۵۷۶۳)]^2$$
$$+ \frac{۱}{۱۴۱}[(۰{,}۶۰)(۱۶۸۱ - ۳۳۴۹{,}۸)^2 + (۰{,}۴۰)(۵۸۵۳ - ۳۳۴۹{,}۸)^2]$$
$$= ۵۴۵۷۰۸{,}۰۵ + ۲۹۶۲۶{,}۵۲ = ۵۷۵۳۳۴{,}۵۷$$

دراین محاسبه با توجه به داده‌های جدول از $s_1 = ۱۷۷۳$ و $s_2 = ۵۷۶۳$ استفاده شده است. جملهٔ دوم در محاسبهٔ برآورد واریانس، یعنی ۲۹۶۲۶٫۵۲ نسبتاً بزرگ به‌نظر می‌رسد. توجه دارید که این مقدار، افزایش واریانس ناشی از به‌کار بردن w'_1 و w'_2 به‌جای W_1 و W_2 است که نامعلوم‌اند. البته با وجود ظاهر بزرگ این افزایش، باکمی دقت ملاحظه می‌شود که این مقدار فقط ۵٪ کل واریانس است.▲

تمرینها

۱ . میزان محصول سیب درختان باغی برحسب ده کیلوگرم به‌شرح جدول زیر است. محصولها برحسب ردیف درختان در جدول آمده‌اند. سه ردیف آخر، درختان جوانتر هستند. جامعه را به دو

طبقه درختان جوان و درختان دیگر تقسیم می‌کنیم. از طبقهٔ اول ۲ درخت با محصولهای ۵ و ۴ و از طبقهٔ دوم ۶ درخت با محصولهای ۶، ۴، ۷، ۶، ۷، ۹ به‌تصادف انتخاب می‌کنیم. برآورد میانگین محصول درختان باغ را به‌دست آورید. برآورد واریانس این برآوردکننده را معین کنید. اگر از کل جدول استفاده کنیم واریانس دقیق این میانگین چقدر است؟ در سطح معنادار بودن ۵ درصد، از روی برآورد میانگین محصول درختان باغ، بازهٔ اطمینانی برای میانگین محصول درختان باغ به‌دست آورید. اگر دو نمونهٔ بالا یک نمونهٔ تصادفی (ساده) از جامعهٔ درختان به‌حساب آید، برآورد و میانگین محصول درختان باغ و برآورد واریانس این برآوردکننده چقدر است؟

۷	۸	۵	۶	۶	۱۰	۷	۶
۵	۴	۴	۷	۶	۶	۳	۸
۴	۷	۸	۱۰	۸	۴	۶	۶
۶	۴	۶	۴	۸	۷	۹	۸
۶	۳	۹	۹	۷	۸	۱۱	۹
۵	۳	۴	۳	۵	۲	۴	۳
۵	۳	۳	۴	۵	۴	۳	۴
۴	۵	۴	۳	۳	۴	۳	۵
۴۲	۳۷	۴۳	۴۶	۴۸	۴۵	۴۶	۴۹

۲ . جمعیت سه شهر کوچک به‌ترتیب $N_1 = ۴۰۰۰۰$، $N_۲ = ۲۰۰۰۰$، و $N_۳ = ۳۰۰۰۰$ نفر است. می‌خواهیم برای بررسی مشخصه‌ای در این ۳ شهر نمونه‌ای تصادفی با طبقه‌بندی شامل ۴۰۰ واحد انتخاب کنیم. از روی سرشماریهای گذشته داریم $S_1 = ۲۰$، $S_۲ = ۱۲$، $S_۳ = ۱۴$. می‌دانیم هزینهٔ انتخاب هر واحد ۳۶ است. حجم نمونه‌ای که باید از هر شهر انتخاب کنیم در دو حالت زیر به‌دست آورید

الف) وقتی تخصیص، متناسب با حجم است. ب) وقتی تخصیص اپتیمم است.

۳ . در یک بررسی جمعیت شناختی روستاهای یک استان از روش نمونه‌گیری تصادفی با طبقه‌بندی استفاده می‌شود. هر شهرستان یک طبقه منظور می‌شود. اگر هزینهٔ جمع‌آوری اطلاعات برای هر واحد ۱۰ باشد و هزینه‌های اداری و غیره روی هم ۱۰۰۰۰۰ باشند، اندازهٔ اپتیمم n حجم نمونه را طوری بیابید که واریانس میانگین نمونه دارای کوچکترین مقدار شود. هزینهٔ کل پیش‌بینی شده برای این بررسی ۸۰۰۰۰۰ است. اطلاعات موجود در جدول زیر آمده است. با این اطلاعات حجم نمونهٔ هر طبقه را تعیین کنید.

طبقه	۱	۲	۳	۴	۵	۶	۷	۸	۹
تعداد روستا	۱۱۰	۸۲	۶۶	۵۲	۲۳	۶۸	۱۱۰	۹۰	۱۷۰
جمعیت متوسط	۴۰۰	۸۰۰	۹۰۰	۱۱۰۰	۱۹۰۰	۶۰۰	۴۵۰	۳۸۰	۳۰۰
S_h	۵۰۰	۹۰۰	۹۲۰	۱۱۶۰	۱۹۵۰	۵۰۰	۷۸۰	۵۰۰	۵۰۰

۴ . یک مؤسسه برای تعیین متوسط درآمد روزانهٔ روستاییان سه روستا قصد دارد نمونه‌گیری

انجام دهد. تعداد روستاییان ۳ روستا به‌ترتیب ۱۵۰، ۶۰، و ۹۰ نفرند. از روی نمونه‌گیری مقدماتی واریانسهای درآمدهای ۳ روستا به‌ترتیب $\frac{۱۲۹}{۶}$، $\frac{۵۹}{۱۵}$، و $\frac{۸۹}{۱۰}$ به‌دست آمده است. هزینهٔ تهیهٔ یک واحد نمونه در سه روستا به‌ترتیب ۴، ۹، ۱۶ است.

الف) مؤسسه ابتدا تصمیم می‌گیرد n فرد، جمعاً از ۳ روستا، انتخاب کند. با اطلاعات بالا، از هر روستا چه سهمی از n را باید برای نمونه‌گیری در نظر گرفت تا واریانس برآوردکنندهٔ میانگین مینیمم باشد؟

ب) اگر بودجه‌ای برابر ۴۴۰ که شامل هزینه‌های اداری نیست برای انجام تحقیق تخصیص یابد، با این بودجه دقیقاً حجم نمونه‌ای که باید از هر روستا انتخاب شود چقدر است؟

ج) قبل از اجرای نمونه‌گیری، ممکن است تصمیم بگیرند که برآورد متوسط درآمد با دقتی خاص محاسبه شود. بدین معنا که واریانس برآوردکننده برابر $\frac{۸۵}{۱۰۰}$ باشد. در این صورت از هر روستا چه تعداد واحد نمونه باید برگزید؟

د) اگر از ۳ روستا به‌ترتیب ۱۶، ۶، ۹ روستایی به‌تصادف انتخاب شوند و میانگین درآمد روزانهٔ این ۳ نمونه به‌ترتیب ۱۰۰، ۱۲۰، ۱۱۰، باشد، برآوردهای نااریبی برای متوسط درآمد روزانهٔ جامعهٔ هر روستا بیابید. اگر واریانس این ۳ نمونه به‌ترتیب ۱۵، ۵، و $\frac{۱۰}{۲}$ باشد برآورد واریانس برآوردکنندهٔ متوسط درآمد روزانهٔ هر یک از ۳ روستا و برآورد واریانس برآوردکنندهٔ متوسط درآمد روزانه جامعهٔ ۳ روستا را تعیین کنید.

۵. برای تعیین برآورد متوسط مدت زمانی که مبتلایان به یک نوع بیماری اعصاب در بیمارستانهای شهری بستری می‌شوند، تعداد بیمارانی را که در یک سال در ۳ بیمارستان موجود در شهر بستری بوده‌اند در نظر می‌گیرند. تعداد بیماران به‌ترتیب ۶۰۰، ۲۲۰، و ۳۶۰ نفر بوده‌اند. از روی یک بررسی مقدماتی واریانس تعداد روزهای بستری بودن بیماران تقریباً برابر با $\frac{۵۱۱}{۶}$، $\frac{۲۲۹}{۱۵}$، و $\frac{۲۵۹}{۱۰}$ به‌دست آمده است. هزینهٔ کسب اطلاع دربارهٔ یک بیمار در ۳ بیمارستان به‌ترتیب ۲، ۲، و ۹ است.

الف) اگر بودجه‌ای معادل ۱۱۸۲ که هزینه‌های اداری را شامل نیست برای تحقیق تخصیص دهند حجم نمونهٔ هر بیمارستان چقدر باید باشد تا واریانس برآوردکنندهٔ مورد نظر مینیمم شود؟

ب) اگر از ۳ بیمارستان به‌ترتیب ۱۵۰، ۲۲، و ۵۲ بیمار به‌عنوان نمونه انتخاب کنند و متوسط مدت بستری بودن آنها به‌ترتیب ۲۰، ۲۵، و ۲۲ روز باشد برآوردی نااریب برای متوسط مدت بستری بودن در هر ۳ بیمارستان به‌صورت جداگانه و همچنین برآوردی نااریب برای متوسط مدت بستری بودن جامعهٔ بیماران این نوع بیماری به‌دست آورید. اگر تغییرات این سه نمونه به‌ترتیب ۹۰، ۱۵، و ۳۵ باشد برآورد واریانس برآوردکنندهٔ متوسط مدت بستری بودن در هر بیمارستان و برآورد واریانس برآوردکنندهٔ متوسط مدت بستری بودن در بیمارستانها را محاسبه کنید.

ج) از سوابق موجود در بیمارستانها دریافته‌اند که از نمونه‌های بالا به‌ترتیب ۳۰، ۵، و ۱۰ نفر از نظر ارثی گرفتار این بیماری شده‌اند. برآوردی نااریب برای نسبت افراد بیماری را که در کل جامعهٔ بیماران مذکور به‌دلیل وراثت بیمار شده‌اند بیابید و تعداد این افراد را برای کل جامعهٔ در یک سال برآورد کنید. برآورد واریانس این برآوردکننده را به‌دست آورید.

د) اگر بخواهند قبل از نمونه‌گیری قسمت الف، متوسط مدت بستری بودن را با دقتی خاص

بیابند به‌قسمی که واریانس برآوردکنندهٔ این متوسط برابر ۲ر۰ باشد در این صورت حجم نمونه‌ای که از هر بیمارستان باید بگیرند چقدر است؟

۶ . در یک نمونه‌گیری تصادفی با طبقه‌بندی، تعداد طبقات ۴ و وزن طبقات به‌ترتیب عبارت‌اند از $\frac{۱}{۳}$، $\frac{۱}{۴}$، $\frac{۱}{۶}$، و $\frac{۱}{۴}$ و $N = ۲۴۰$. اگر تعداد واحدهای نمونه در ۴ طبقه به‌ترتیب ۲۰، ۱۵، ۱۰، و ۱۵ و تغییرات نمونه‌ای در طبقات به‌ترتیب ۱۲، ۱۰، ۸، ۱۲ و میانگینهای نمونه‌ای در طبقات به‌ترتیب، ۱۴، ۱۲، ۸، ۱۵ باشند برآورد میانگین جامعه و برآورد واریانس این برآوردکننده را بیابید. در صورتی که توزیع $\overline{Y}_{st}$ را نرمال بگیریم بازهٔ اطمینانی برای $\overline{Y}_N$، میانگین جامعه، با ضریب اطمینان ۹۵ر۰ بیابید. اگر این توزیع نرمال نباشد با استفاده از روش ساترتوایت بازهٔ اطمینانی برای $\overline{Y}_N$ با همان ضریب اطمینان بیابید.

۷ . از دفاتر ثبت سابقهٔ ۴ بیمارستان تعداد کل افرادی که گروه خونی آنها معین شده است به‌ترتیب عبارت‌اند از ۱۲۰۰، ۱۵۴۰، ۷۶۰، و ۱۳۰۰. از این ۴ دفتر به‌تصادف ۵۰، ۶۰، ۳۰، ۶۰ نفر را انتخاب می‌کنند. از همان دفاتر دریافته‌اند که ۲۰، ۲۲، ۱۰، ۱۸ نفر دارای گروه خونی A هستند.

الف) برآوردی ناریب برای نسبت افرادی که در کل جامعهٔ مراجعه کنندگان دارای گروه خونی A هستند بیابید و تعداد این افراد را برآورد کنید.

ب) برآورد واریانس برآوردکنندهٔ این نسبت را تعیین کنید.

ج) اگر در این مسأله تعداد افراد نمونه معلوم نباشد و قصد داشته باشیم جمعاً ۷۰ نفر انتخاب کنیم و از سوابق قبلی این بیمارستانها بدانیم نسبت افرادی که دارای گروه خونی A هستند به‌ترتیب ۴ر۰، ۳۴ر۰، ۳۲ر۰، ۳۵ر۰ است. معین کنید که برای نمونه‌گیری مورد نظر از هر بیمارستان، با تخصیص نیمن، تقریباً چند نفر باید انتخاب کنیم؟

۸ . در یک نمونه‌گیری از دو طبقه مایلیم به‌جای $n_۱$ و $n_۲$ در انتساب نیمن داشته باشیم $n_۱ = n_۲$. اگر $\mathrm{V}(\overline{Y}_{st})$ معرف واریانس در حالت $n_۱ = n_۲$، و $\mathrm{V}_{opt}(\overline{Y}_{st})$، واریانس مربوط به تخصیص نیمن باشد نشان دهید وقتی N بزرگ باشد

$$\frac{\mathrm{V}(\overline{Y}_{st}) - \mathrm{V}_{opt}(\overline{Y}_{st})}{\mathrm{V}_{opt}(\overline{Y}_{st})} = \left(\frac{r-۱}{r+۱}\right)^۲$$

که در آن $r = \frac{n_۱}{n_۲}$.

۹ . فرض کنید دو گروه از مصرف‌کنندگان یک قلم کالا را در نظر گرفته و نظر آنها را دربارهٔ کیفیت کالا خواسته‌ایم. اگر ۱ معرف نظر موافق و ۰ معرف نظر ناموافق باشد، داریم

گروه ۱: ۱, ۱, ۰, ۱, ۰, ۱, ۱, ۱

گروه ۲: ۱, ۰, ۰, ۰, ۱, ۰, ۱, ۰

الف) P و P_hها را حساب کنید.

ب) دو نمونه به حجم $n_h = ۳$ انتخاب کنید و $p_۱$ و $p_۲$ را به‌دست آورید.

ج) با استفاده از نتایج قسمت (ب) مقدار P را برآورد کنید.

۱۰ . برای جامعه‌ای فرضی از دانشجویان ۴ دانشکده، اطلاعاتی به‌صورت جدول زیر داریم. تعداد دانشجویان هر دانشکده در ستون دوم جدول، حجم نمونهٔ هر دانشکده در ستون سوم و نسبت دانشجویانی که از هر دانشکده حداقل یک‌بار در سال به پزشک مراجعه کرده‌اند در ستون آخر ثبت شده‌اند. نسبت دانشجویانی را برآورد کنید که در طول سال گذشته حداقل یک‌بار به پزشک مراجعه کرده‌اند. همچنین $\hat{\mathrm{V}}(p_{st})$ را حساب کنید

دانشکده	N_h	n_h	p_h
۱	۲۰۰۰	۱۰۰	۰٫۲
۲	۱۶۰۰	۸۰	۰٫۳
۳	۱۲۰۰	۶۰	۰٫۴
۴	۱۲۰۰	۶۰	۰٫۳

۱۱ . به مثال ۱.۴ رجوع کنید و آنچه را که دربارهٔ جمعیت ۶۴ شهر در ۱۹۳۰ خواسته بودیم دربارهٔ جمعیت ۶۴ شهر در ۱۹۲۰ نیز بیابید.

۱۲ . پژوهشگری می‌خواهد برآوردی از متوسط فروش سالیانهٔ ۵۶ شرکت را به‌دست آورد. برای این کار نمونه‌ای به حجم ۱۵ شرکت انتخاب می‌کند. داده‌های فراوانی مربوط به این شرکتها به‌صورت طبقه‌بندی با فاصله‌های ۵۰۰۰۰ دلار موجودند که در جدول زیر ارائه شده‌اند. برای $L = ۳$ بهترین تخصیص کرانهای این طبقات را تعیین کنید.

فراوانی	درآمد (برحسب ۱۰۰۰ دلار)	فراوانی	درآمد (برحسب ۱۰۰۰ دلار)
۵	۳۰۰-۳۵۰	۱۱	۱۰۰-۱۵۰
۸	۳۵۰-۴۰۰	۱۴	۱۵۰-۲۰۰
۳	۴۰۰-۴۵۰	۹	۲۰۰-۲۵۰
۲	۴۵۰-۵۰۰	۴	۲۵۰-۳۰۰

۱۳ . جامعه‌ای به ۳ طبقه افراز شده است. حجم طبقات به‌ترتیب ۲۰۰، ۳۰۰، و ۴۰۰ است. مقادیر S_h طبقه‌ها به‌ترتیب ۲ و ۴ و ۵ هستند. بودجهٔ کل نمونه‌گیری، صرف‌نظر، از هزینه‌های مشترک اداری و غیره ۳۶۰۰ و هزینهٔ نمونه‌گیری هر واحد در هر طبقه برابر ۵۰ است.

الف) اگر تخصیص نیمن مورد نظر باشد، حجم نمونه در هر طبقه چقدر است؟

ب) اگر از ۳ طبقه به‌ترتیب ۸، ۲۴، و ۴۰ واحد به‌تصادف بدون جایگذاری انتخاب کنیم، به‌شرط آنکه میانگین این نمونه‌ها به‌ترتیب ۸، ۱۰، و ۱۲ باشند، برآورد میانگین جامعه چقدر است؟ برآورد واریانس این برآوردکننده چقدر است؟

ج) با ضریب اطمینان ۹۵ درصد، به‌شرط آنکه توزیع تقریبی $\overline{Y}_{st}$ نرمال باشد، بازهٔ اطمینانی برای میانگین جامعه بیابید.

۱۴ . نمونه‌گیری تصادفی با طبقه‌بندی همیشه برآوردکننده‌ای با واریانسی کوچکتر از واریانس برآوردکنندهٔ متناظرش در نمونه‌گیری تصادفی ساده تولید نمی‌کند. مسأله زیر این مطلب را نشان می‌دهد.
یک توزیع‌کننده مواد غذایی در شهری بزرگ می‌خواهد بداند که آیا تقاضاها آن‌قدر هستند که فرآورده‌ای جدید را به بخش توزیعش اضافه کند یا نه. برای یاری به اخذ تصمیم، طرحی می‌کند که این فرآوردهٔ جدید را به نمونه‌ای تصادفی از فروشگاههایی که مشتری او هستند بدهد و میانگین فروش ماهیانه این فرآورده را برآورد کند. چهار زنجیره از فروشگاهها مشتری او هستند. هر زنجیره را یک طبقه می‌گیرد. در زنجیرهٔ ۱، جمعاً ۲۴ فروشگاه، در زنجیرهٔ ۲، به تعداد ۳۶ فروشگاه، در زنجیرهٔ ۳، به‌تعداد ۳۰ فروشگاه، و در زنجیرهٔ ۴، به‌تعداد ۳۰ فروشگاه وجود دارند. فروشگاه توان مادی و اداری تهیهٔ داده‌های ماهیانهٔ ۲۰ فروشگاه را دارد. توزیع‌کننده از تخصیص متناسب برای انجام یک نمونه‌گیری با طبقه‌بندی استفاده می‌کند. از چهار طبقه نمونه‌هایی تصادفی اختیار می‌کند. نتایج فروش ماهیانه در فروشگاههای منتخب به شرح جدول زیر است

طبقهٔ ۱	طبقهٔ ۲	طبقهٔ ۳	طبقهٔ ۴
۹۴	۹۱	۱۰۸	۹۲
۹۰	۹۹	۹۶	۱۱۰
۱۰۲	۹۳	۱۰۰	۹۴
۱۱۰	۱۰۵	۹۳	۹۱
	۱۱۱	۹۳	۱۱۳
	۱۰۱		

میانگین فروش ماهیانه این فرآورده را در کل ۴ زنجیره برآورد کرده و حدود خطای آن را مشخص کنید. اگر نمونه‌ها را یک نمونهٔ تصادفی ساده از جامعهٔ فروشگاهها بگیریم واریانس برآوردکنندهٔ حاصل از این نمونه را با واریانس $\overline{Y}_{st}$ قسمت بالا مقایسه کنید. چرا کارایی نمونه‌گیری تصادفی ساده بیش از نمونه‌گیری تصادفی با طبقه‌بندی است؟

۱۵ . متغیر تصادفی Y دارای توزیع‌یکنواخت روی بازهٔ $(a,\ a+c)$ است. دامنهٔ تغییرات متغیر را به L قسمت برابر تقسیم می‌کنیم تا L طبقه ایجاد شود. از هر طبقه نمونه‌ای تصادفی به حجم $\frac{n}{L}$ می‌گیریم. بار دیگر از کل جامعه نمونه‌ای تصادفی به حجم n تهیه می‌کنیم. اگر واریانس برآوردکنندهٔ میانگین جامعه را در نمونه‌گیری تصادفی ساده با $V_۱$ و در نمونه‌گیری با طبقه‌بندی با $V_۲$ نمایش دهیم ثابت کنید که $V_۲ = \frac{V_۱}{L^۲}$.

تمرینهای چهارگزینه‌ای

۱ . متغیر تصادفی Y دارای توزیع یکنواخت روی بازهٔ (۵ ,۲) است. دامنهٔ تغییرات متغیر را به ۳ قسمت برابر تقسیم می‌کنیم تا سه‌طبقه به‌وجود آیند. از هر طبقه نمونه‌ای تصادفی به حجم ۳۰ می‌گیریم. بار دیگر از کل جامعه نمونه‌ای تصادفی به حجم ۹۰ می‌گیریم. اگر واریانس برآوردکنندهٔ

میانگین جامعه را در نمونه‌گیری از کل جامعه با V_1 و در نمونه‌گیری از طبقات با V_2 نشان دهیم، آن‌گاه

الف) $V_1 = 3V_2$ ب) $V_1 = \frac{1}{3}V_2$ ج) $V_2 = \frac{V_1}{4}$ د) $V_1 = V_2$

۲. در یک نمونه‌گیری تصادفی با طبقه‌بندی از تخصیص نیمن استفاده می‌شود. جامعه از ۴ طبقه تشکیل شده است. در این طبقات مقادیر W_hS_h با اعداد ۲، ۳، ۲، ۲ متناسب‌اند. اگر حجم نمونهٔ طبقه سوم ۶۰ باشد حجم کل نمونه برابر است با

الف) ۱۶۵ ب) ۱۵۰ ج) ۱۳۵ د) ۱۲۰

۳. در یک نمونه‌گیری با طبقه‌بندی تابع هزینه به‌صورت $C = \sum_{i=1}^{2} C_i n_i$ است. n_i حجم نمونه در طبقهٔ i و C_i هزینه تهیهٔ یک واحد نمونه در طبقهٔ i است. اگر W_i و S_i به‌ترتیب وزن و انحراف معیار تقریبی طبقه i باشد، داریم

طبقه iام	W_i	S_i	C_i
۱	۰٫۴	۱۰	۴۰۰
۲	۰٫۶	۲۰	۹۰۰

اگر n حجم کل نمونه در طبقات فرض شود، برای $V(\overline{Y}_{st})$ تثبیت شده، وقتی C مینیمم است که $\frac{n_1}{n}$ برابر باشد با

الف) $\frac{1}{4}$ ب) $\frac{2}{5}$ ج) $\frac{1}{3}$ د) $\frac{2}{7}$

۴. جامعه‌ای از دو طبقه تشکیل شده است. حجم طبقات به‌ترتیب ۱۰ و ۱۵ است. می‌خواهیم نمونه‌ای به حجم ۱۰، با روش متناسب با حجم، از دو طبقه اختیار کنیم. تعداد نمونه‌های ممکن برابر است با

الف) $\binom{25}{10}$ ب) ۲۱۰ ج) $\binom{15}{6}$ د) $\binom{15}{6}$۲۱۰

۵. وقتی از نمونه‌گیری با طبقه‌بندی برای برآورد میانگین جامعه استفاده می‌کنیم که

الف) پراکندگی در درون هر طبقه زیاد باشد ب) پراکندگی در درون هر طبقه کم باشد

ج) چارچوب نمونه‌گیری را نداشته باشیم د) حجم نمونه بزرگ باشد

۶. در یک نمونه‌گیری تصادفی با طبقه‌بندی که تنها دارای دو طبقه به وزنهای $\frac{1}{3}$ و $\frac{2}{3}$ است، s^2ی نمونه‌های دو طبقه به‌ترتیب ۴ و ۷ است. اگر حجم جامعه ۹۹ و حجم کل نمونه ۹ باشد و نمونه‌گیری با تخصیص متناسب صورت گیرد، برآورد واریانس برآوردکنندهٔ میانگین جامعه برابر است با تقریباً

الف) ۰٫۶ ب) ۰٫۵۴ ج) ۶ د) ۰٫۵

۷. در یک نمونه‌گیری با طبقه‌بندی و با L طبقه، واریانس برآوردکنندهٔ مجموع واحدهای جامعه ۹ برابر مجموع $N_hS_h^2$ هاست. در این صورت کسر نمونه‌گیری در کل جامعه برابر است با

الف) $\frac{1}{9}$ ب) $\frac{1}{10}$ ج) $\frac{1}{L}$ د) $\frac{9}{L}$

۸. در یک نمونه‌گیری با دو طبقه، کسرهای نمونه‌گیری طبقات کوچک و قابل اغماض‌اند. اگر حجم نمونهٔ طبقهٔ اول ۱۶ و وزن این طبقه $\frac{1}{2}$ باشد و اگر مقادیر S^2ی طبقات برابر با ۴ باشند و بخواهیم واریانس $\overline{Y}_{st}$ برابر با $\frac{1}{18}$ شود حجم طبقه دوم را باید برابر با مقدار زیر اختیار کنیم

الف) ۳۲ ب) ۶۴ ج) ۸۰ د) ۹۶

۹ . در یک نمونه‌گیری با طبقه‌بندی، اطلاعات زیر دردست است

طبقهٔ iام	W_i	$S_i^{۲}$	C_h
۱	$\frac{۱}{۳}$	۴	۴
۲		۹	۹

اگر بودجه نمونه‌گیری صرف‌نظر از مخارج اداری برابر ۲۲۰ باشد برای اینکه $\mathrm{V}(\overline{Y}_{st})$ مینیمم شود حجم کل نمونه را باید برابر مقدار زیر انتخاب کرد

الف) ۶۰ ب) ۴۸۴ ج) ۳۰ د) ۲۱

۱۰ . در یک نمونه‌گیری با طبقه‌بندی، ۴ طبقه داریم. در این طبقه‌ها $N_hS_h/\sqrt{C_h}$ متناسب با اعداد ۲، ۳، ۳ و ۴ هستند. اگر تخصیص اپتیمم بوده و حجم کل نمونه ۱۸۰ باشد حجم نمونهٔ طبقهٔ ۴ برابر است با

الف) ۸۴ ب) ۳۰ ج) ۴۵ د) ۶۰