فصلنامه گزیده‌مطالب آماری
سال ۱۷، شماره ۲، تابستان ۱۳۸۵، صص ۱۴۸ - ۱۳۱

محاسبه واریانس برای داده‌های حاصل از نمونه‌گیری چند مرحله‌ای و براورد نسبتی

آرمان بیداربخت‌نیا

مرکز آمار ایران

چکیده. در این مقاله روش‌هایی برای براورد واریانس مورد بررسی قرار می‌گیرد که اثر روش‌های پیچیده نمونه‌گیری و همچنین استفاده از براوردگرهای نسبتی و پساطبقه‌بندی را منعکس می‌کنند. نخست براورد واریانس در طرح‌های مختلف نمونه‌گیری بررسی می‌شود و سپس با استفاده از روش خطی‌سازی سری تیلور، براورد واریانس برای براوردگر نسبتی محاسبه می‌شود. در نهایت با استفاده از یک مثال، براورد واریانس یک براوردگر پیچیده به‌گونه‌ای بیان می‌شود که اثر روش نمونه‌گیری، روش براورد و همچنین وزن‌دهی را نیز منعکس می‌کند.

۱- مقدمه

ممکن است نمونه‌ای با استفاده از روش‌های نمونه‌گیری چند مرحله‌ای انتخاب شده و سپس عناصر آن به‌منظور براورد مقادیر کل جامعه وزن‌دهی شده باشند. به این ترتیب براورد نسبتی پس‌طبقه‌بندی شده تورمی ایجاد می‌کند که نه تنها اندازه بلکه ترکیب آن مثل سن– جنس– نژاد و مناطق سکونت را نیز منعکس می‌کند. سپس این داده‌های وزن‌دهی شده برای براورد میانگین، نسبت و دیگر پارامترهای مورد علاقه به کار برده می‌شوند. هدف اصلی مقاله این است که نشان دهد چگونه براوردگرهای واریانس، مراحل واقعی روش‌های نمونه‌گیری و براورد نسبتی را منعکس می‌کنند.

واژگان کلیدی: براوردگر نسبتی؛ براورد واریانس؛ روش خطی‌سازی.

بعضی از واریانس‌ها را می‌توان با ترکیب دو روش شناخته‌شده خطی‌سازی براوردهای نسبتی (وودراف، ۱۹۷۱) و براوردگر واریانس تعمیم‌یافته برای نمونه‌گیری چند مرحله‌ای (کندال و استورات، ۱۹۶۸) تقریب زد. با این ترکیب، می‌توان ویژگی‌های این دو روش را در براوردهای واریانس انعکاس داد.

روش‌های نخست، براورد نسبتی را با استفاده از بسط سری‌های تیلور خطی نموده و سهم متغیر از بسط خطی برای براوردهای نسبتی را حفظ می‌کند. سپس واریانس نسبت با واریانس سهم متغیر از نسبت‌های خطی شده، تقریب زده می‌شود.

روش‌های دوم، حالت تعمیم‌یافته واریانس برای داده‌های جمع‌بندی‌شده را برای هر طرحی که در انتخاب نمونه به کار رفته باشد، شامل می‌شوند.

بعد از این روش‌ها دو دسته علائم مجموع می‌تواند وجود داشته باشد که دسته اول از براوردهای نسبتی و دسته دوم از طرح‌های نمونه‌گیری ناشی می‌شود. واریانس تعمیم‌یافته می‌تواند با تعویض علائم مجموع، انتقال آن‌ها به جلو برای نمونه‌گیری و جمع بستن آن‌ها برای براورد نسبتی به دست آید. به این ترتیب فقط علائم مجموع ناشی از روش‌های نمونه‌گیری باقی می‌ماند. سپس می‌توانیم فرمول واریانسی که قبلاً تهیه شده است را برای این نتیجه نهایی به کار ببریم.

این روش‌ها می‌تواند برای براوردهای واریانس داده‌هایی به کار رود که توسط مراکز آمارهای بهداشتی و دیگر مراکز دولتی گردآوری شده باشد که معمولاً از براوردها و طرح‌های نمونه‌گیری پیچیده بهره می‌گیرند و ممکن است که هنوز به فکر هر دو ویژگی براوردهای واریانس نباشند. هیدیروگلو و رائو (۱۹۸۳) و شاه (۱۹۸۱) این روش‌ها را به‌ترتیب برای تحلیل داده‌های آمارگیری بهداشتی کانادا و برنامه خطاهای استاندارد برای داده‌های آمارگیری به کار برده‌اند؛ فرمول‌های واریانس در مورد نخست برای نمونه‌گیری با جایگذاری با احتمال‌های برابر به کار رفته است، در حالی‌که برای مورد اخیر نمونه‌گیری دو مرحله‌ای بدون جایگذاری با احتمال‌های برابر استفاده شده است. در عمل برای نمونه‌گیری چند مرحله‌ای به ندرت در همه مراحل از یک روش نمونه‌گیری استفاده می‌شود. برای مثال، ممکن است نمونه‌ها در مرحله اول با جایگذاری و با احتمال‌های متناسب با اندازه جامعه (pps) و در مرحله دوم بدون جایگذاری و با احتمال‌های برابر

انتخاب شود.

بخش ۲ به معرفی برخی نمادهایی که در بخش‌های بعد به کار می‌رود، پرداخته است. بخش ۳ براوردهای واریانس تعمیم‌یافته برای سرجمع‌ها، با استفاده از نمونه‌گیری چند مرحله‌ای را ارائه می‌کند. در بخش ۴، براوردهای نسبتی پس‌طبقه‌بندی شده، خطی شده‌اند و فقط سهم متغیر حفظ شده است. سپس واریانس نسبت‌ها را با آن سهم متغیر نسبت‌های خطی شده تقریب می‌زنیم. نهایتاً یک مثال و چند دستورالعمل در بخش ۵ آمده است.

۲- علائم

فرض کنید که جامعه به L طبقه مستقل طبقه‌بندی شده باشد که با $s=1,\ldots,L$ نمادگذاری می‌شود و اعضای طبقه sام در N_{1s} واحد نمونه‌گیری مقدماتی (PSU) گروه‌بندی شده باشند که با $i=1,\ldots,N_{1s}$ نمادگذاری می‌شود و iامین PSU شامل N_{2si} عضو است که با $j=1,\ldots,N_{2si}$ نشان داده می‌شود. علائم متناظر برای نمونه با حرف کوچک n و با زیرنویس‌های مشابه، همان‌گونه که در جدول (۱) مشاهده می‌کنیم، نشان داده می‌شود. از آن‌جا که واریانس برای سرجمع‌های این طبقات جمع‌پذیر است، برای نشان دادن واریانس حاصل از یک طبقه، از زیرنویس s برای طبقات صرف نظر می‌کنیم.

جدول (۲) واریانس‌های سرجمع‌ها را به‌همراه نوع سرجمع و طرح نمونه‌گیری نشان می‌دهد.

برخی از فرمول‌ها برای این واریانس‌ها در بخش ۳ مورد بحث قرار می‌گیرد. واریانس نسبت‌ها در جدول (۱) می‌تواند خطی شده و به این ترتیب با سرجمع‌ها در دسته‌های مشابه قرار گیرد.

جدول ۱ـ علائم برای داده‌های نمونه‌ای، در حالی‌که خوشه‌بندی در دو مرحله انجام شده و در هر طبقه نمونه‌گیری سه مرحله‌ای اجرا شده باشد

	جامعه	نمونه
واحدهای مرحله اول	N_1	n_1
واحدهای مرحله دوم	N_{2i}	n_{2i}
واحدهای مرحله سوم	N_{2ij}	n_{2ij}
شاخص مرحله اول	$i=1,\ldots,N_1$	$i=1,\ldots,n_1$
شاخص مرحله دوم	$j=1,\ldots,N_{2i}$	$j=1,\ldots,n_{2i}$
شاخص مرحله سوم	$k=1,\ldots,N_{2ij}$	$k=1,\ldots,n_{2ij}$
شاخص سلول‌ها	$h=1,\ldots q$	$h=1,\ldots q$
مقادیر کل:	$N=\sum_i^{N_1}\sum_j^{N_{2i}} N_{2ij}$	$n=\sum_i^{n_1}\sum_j^{n_{2i}} n_{2i}$
سرجمع‌ها برای سلول	$Y_{h2}=\sum_i^{N_1}\sum_j^{N_{2i}}\sum_k^{N_{2ij}} y_{hijk}$	$y_{h2}=\sum_i^{n_1}\sum_j^{n_{2i}}\sum_k^{n_{2ij}} y_{hijk}$
	$X_{h2}=\sum_i^{N_1}\sum_j^{N_{2i}}\sum_k^{N_{2ij}} x_{hijk}$	$x_{h2}=\sum_i^{n_1}\sum_j^{n_{2i}}\sum_k^{n_{2ij}} x_{hijk}$
نسبت در سلول:	Y_{h2}/N	y_{h2}/n
نسبت:	$R_{h2}=X_{h2}/Y_{h2}$	$r_{h2}=x_{h2}/y_{h2}$

x_{hijk} و y_{hijk} متغیرهایی برای صفات x و y به‌ترتیب هستند.

۳- واریانس

نمونه‌گیری می‌تواند با احتمال برابر یا نابرابر و یا با احتمال‌های متناسب با اندازه (pps)، با جایگذاری و یا بدون جایگذاری و با طرح‌های متقارن یا نامتقارن انجام شود. در هر مرحله، ممکن است که هر ترکیبی از این گزینه‌ها را در نظر داشته باشیم.

جدول ۲ـ واریانس‌ها با انواع طرح و سرجمع

انواع طرح		سرجمع		
		نمونه‌گیری یک مرحله‌ای[1]	نمونه‌گیری دو مرحله‌ای[2]	نمونه‌گیری سه مرحله‌ای[6]
احتمال‌های نابرابر	WR[3]	$\underset{1}{\mathrm{var}}(y_{h1})$	$\mathrm{var}(y_{h2})$	$\mathrm{var}(y_{h3})$
	Wo[4]	$\underset{1}{\mathrm{var}}(y_{h1})$	$\mathrm{var}(y_{h2})$	$\mathrm{var}(y_{h3})$
احتمال‌های برابر	WR[3]	$\underset{1}{\mathrm{var}}(y_{h1})$	$\mathrm{var}(y_{h2})$	$\mathrm{var}(y_{h3})$
	Wo[4]	$\underset{1}{\mathrm{var}}(y_{h1})$	$\mathrm{var}(y_{h2})$	$\mathrm{var}(y_{h3})$
ترکیب‌ها[5]			$\mathrm{var}(y_{h2})$	$\mathrm{var}(y_{h3})$

[1] $y_{h1} = \sum_i^{n_1} y_{hi}$ ؛ [2] $y_{h2} = \sum_i^{n_1}\sum_j^{n_{2i}} y_{hij}$ ؛ [3] باجایگذاری؛ [4] بدون جایگذاری؛

[5] ترکیبی از نمونه‌گیری احتمالاتی با احتمال‌های برابر و نابرابر؛ [6] $y_{h3} = \sum_i^{n_1}\sum_j^{n_{2i}}\sum_k^{n_{3ij}} y_{hijk}$.

یک فرمول واریانس تعمیم‌یافته برای هر براورد $\hat{\theta}_h$ در hامین سلول، بر اساس احتمال‌های انتخاب کاملاً دلخواه ارائه می‌کنیم. به این ترتیب واریانس کل عبارت از مجموع واریانس‌ها برای تمام طبقات است.

علامت E برای عملگر مقدار مورد انتظار، var برای واریانس و vâr برای براورد نااریب var به کار رفته است. می‌توان نوشت:

$$\mathrm{var}(\hat{\theta}_h) = \underset{1}{\mathrm{var}}(\underset{>1}{E}(\hat{\theta}_h)) + \underset{1}{E}(\underset{>1}{\mathrm{var}}(\hat{\theta}_h)) \qquad (1)$$

که « ۱< » علامتی برای ارائه تمام مراحل نمونه‌گیری پس از مرحله نخست است.

اگر $\hat{\theta}_h = y_{h1}$، تعریف شده در جدول (۱) باشد، براوردگر نااریب آن می‌تواند به‌شکل زیر نوشته شود:

$$\hat{\text{var}}(\hat{\theta}_h) = \underset{1}{\hat{\text{var}}}(\hat{\theta}_h) + \sum_i^{n_1} \pi_i^{(1)} \underset{>1}{\hat{\text{var}}}(y_{hi}) \qquad (2)$$

که $\pi_i^{(1)}$ عبارت است از احتمال این که واحد iام در n_1 واحد نمونه‌گیری مقدماتی (PSU) باشد.

رابطه (۱) ممکن است به سه مولفه تقسیم شود:

$$\text{var}(\hat{\theta}_h) = \underset{1}{\text{var}}\ \underset{2}{E}\ \underset{>3}{E}(\hat{\theta}_h) + \underset{1}{E}\ \underset{2}{\text{var}}\ \underset{>3}{E}(\hat{\theta}_h) + \underset{1}{E}\ \underset{2}{E}\ \underset{>3}{\text{var}}(\hat{\theta}_h). \qquad (3)$$

اگر $y_{hi} = \sum_{j=1}^{n_{2i}} y_{hij}$، در رابطه (۲) جایگزین شود، براورد ناریب (۳) را می‌توان به‌شکل زیر نوشت:

$$\hat{\text{var}}(\hat{\theta}_h) = \underset{1}{\hat{\text{var}}}(\hat{\theta}_h) + \sum_i^{n_1} \pi_i^{(1)} \underset{2}{\hat{\text{var}}}(y_{hi}) + \sum_i^{n_1} \pi_i^{(1)} \sum_j^{n_{2i}} \pi_{ij}^{(2)} \underset{>2}{\hat{\text{var}}}(y_{hij}) \qquad (4)$$

که $\pi_{ij}^{(2)}$ عبارت است از احتمال این که واحد j ام از مرحله دوم در iامین واحد انتخاب شده در مرحله اول باشد. این رابطه برای مراحل بعدی نمونه‌گیری نیز قابل تعمیم است.

فرمول‌های بالا را می‌توان به این ترتیب خلاصه نمود: یک براوردگر ناریب برای واریانس در نمونه‌گیری چند مرحله‌ای، وقتی که مرحله اول نمونه‌گیری بدون جایگذاری باشد، از مجموع دو مولفه حاصل شده است. مولفه اول واریانس را در حالتی که فقط نمونه‌گیری مرحله اول اجرا شده باشد براورد می‌کند. مولفه دوم عبارت از مجموع موزون براوردها درون واحدهای انتخاب شده در مرحله اول، برای واریانس حاصل از مراحل دیگر نمونه‌گیری است (واحدهای مرحله اول ثابت فرض می‌شوند)؛ وزن‌ها عبارت از احتمال‌های انتخاب این واحدهای مرحله اول می‌باشد. (دوربین، ۱۹۵۳).

اگر نمونه‌گیری در مرحله اول به صورت جایگذاری انجام شده باشد، با توجه به این که $\pi_i^{(1)} \to 0$، فقط جمله اول در (۴) باقی می‌ماند. در این مورد، برای یک نمونه‌گیری چند مرحله‌ای با هر تعداد مرحله نمونه‌گیری، وقتی که در مرحله اول برای هر انتخاب احتمال‌های نامساوی به کار برده می‌شود، در حالی که مراحل دیگر دلخواه بوده و در واحدهای متمایز انتخاب شده در مرحله اول استخراج به صورت مستقل انجام می‌شود،

براورد واریانس‌ها ساده است.

واریانس‌های مربوط به چند وضعیت متفاوت برای نمونه‌گیری را ملاحظه کنید.

الف) $\operatorname{var}(y_{h1})$ برای حالت بدون جایگذاری با احتمال نابرابر در یک مرحله:

فرض کنید ${}_r p_i$ احتمال انتخاب شدن iامین فرد در rامین انتخاب باشد و

$$\sum_{i}^{N_1} {}_r p_i = 1 \quad , \quad \pi_i = \sum_{r}^{n_1} {}_r p_i \quad , \quad \pi_{ii'} = \sum_{r \neq s}^{n_1} \sum {}_r p_i \; {}_s p_j$$

کندال و استوارت (۱۹۶۸ صفحه ۱۷۲) نشان داده‌اند که

$$\underset{1}{\operatorname{var}}(y_{h1}) = \sum_{i=1}^{N_1} \pi_i (1-\pi_i)\, y_{hi}^2 + \sum_{i \neq i'}^{N_1} \sum (\pi_{ii'} - \pi_i \pi_{i'})\, y_{hi}\, y_{hi'} \qquad (5)$$

با استفاده از:

$$E\Big(\sum_{i \neq i'}^{n} \sum g(y_i\, y_{i'})\Big) = \sum_{i \neq i'}^{N} \sum \pi_{ii'}\, g(y_i\, y_i) \quad \text{و} \quad E\Big(\sum_{i}^{n} g(y_i)\Big) = \sum_{i}^{N} \pi_i\, g(y_i)$$

برای هر تابع g از مشاهدات، براورد ناریب (۵) به صورت زیر داده شده است:

$$\underset{1}{\widehat{\operatorname{var}}}(y_{h1}) = \frac{1}{2} \sum_{i \neq i'}^{n_1} \sum \frac{(\pi_i \pi_{i'} - \pi_{ii'})}{\pi_{ii'}} (y_{hi} - y_{hi'})^2 , \qquad (6)$$

برای نمونه‌گیری یک مرحله‌ای، در فرمول عمومی (۲)، $\widehat{\operatorname{var}}(y_{h1}) = \underset{1}{\widehat{\operatorname{var}}}(y_{h1})$

ب) $\operatorname{var}(y_{h2})$ برای حالت بدون جایگذاری با احتمال نابرابر در نمونه‌گیری دو مرحله‌ای:

$$\widehat{\operatorname{var}}(y_{h2}) = \underset{1}{\widehat{\operatorname{var}}}(y_{h2}) + \sum_{i}^{n_1} \pi_i^{(1)} \underset{2}{\widehat{\operatorname{var}}}(y_{hi+}) ,$$

که جمله اول با استفاده از (۶) معلوم است و جمله دوم عبارت از مجموع موزون واریانس‌ها برای مراحل دوم در واحدهای انتخاب شده در مرحله اول است.

پ) $\hat{\text{var}}_1(y_{h1})$ برای حالت بدون جایگذاری با احتمال برابر در نمونه‌گیری یک مرحله‌ای:

داریم $F_1 = \frac{n_1}{N_1} = \pi_i$ و $\pi_{ii'} = \frac{n_1}{N_1}\,\frac{(n_1-1)}{(N_1-1)}$. با استفاده از این عبارت‌ها، می‌توان (۶) را به صورت زیر نوشت:

$$\hat{\text{var}}_1(y_{h1}) = (1-F_1)\frac{n_1}{n_1-1}\sum_{i}^{n_1}(y_{hi}-\bar{y}_h)^2 \tag{۷}$$

که $\bar{y}$ میانگین y_i ها است.

ت) $\hat{\text{var}}_1(y_{h1})$ برای حالت با جایگذاری با احتمال نابرابر:

حال باید حالت $i = i'$ را برای $\pi_{ii'}$ در نظر بگیریم، اما جمله π_i و $\pi_{i'}$ در مجموع دوگانه هنوز هم باید پسوندهای متفاوتی داشته باشد. رابطه (۶) برای این نمونه‌گیری نیز برقرار است.

ث) $\hat{\text{var}}_1(y_{h1})$ برای حالت با جایگذاری با احتمال برابر:

در این مورد تئوری ساده‌تر می‌شود، یعنی $\pi_i = np_i$ و $\pi_{ii'} = n(n-1)\,p_i p_{i'}$ که p_i عبارت است از احتمال این که هر کدام از انتخاب‌های نمونه‌گیری با جایگذاری شامل عضو iام باشد و تحت این تعاریف می‌توان (۶) را به‌شکل زیر نوشت:

$$\text{var}(y_{hl}) = \frac{n_1}{2}\sum_{i}^{n_1}\sum_{j} p_i\, p_j (y_i - y_j)^2\,, \tag{۸}$$

از آن‌جا که شرایطی مشابه با (۶) برقرار است، با قرار دادن $i = i'$، براوردگر ناریب (۸) را می‌توان به‌شکل زیر نوشت:

$$\hat{\text{var}}_1(y_{h_1}) = \frac{n_1}{(n_1-1)}\sum_{i}^{n_1}(y_{hi}-\bar{y}_h)^2\,, \tag{۹}$$

که تفاوت آن با (۷) فقط در $(1-F_1)$، یعنی فاکتور مربوط به بدون جایگذاری بودن طرح

است. بخشی از هدف طرح نمونه (یعنی انتخاب $\pi_{ii'}$ و سپس π_i) کاهش واریانس براوردگر تا حد ممکن است. ما می‌توانیم مجموعه‌هایی از $\pi_{ii'}$ بیابیم که در تولید واریانس کوچک برای تمام براوردهایی که ممکن است به کار ببریم، مؤثر باشند. برور (۱۹۶۳) مقادیر $\pi_{i'}$ و $\pi_{ii'}$ را در حالتی که دو واحد نمونه داشته باشیم ($n = 2$) معرفی کرد که ویژگی‌های مطلوب واریانس کوچک در (ح) و $(\pi_i\pi_{i'} - \pi_{ii'}) > 0$ که در (۶) نشان داده شد را دارا است.

ج) $\mathrm{var}\,(y_{h3})$ برای حالت بدون جایگذاری با احتمال برابر:

$$y_{h3} = \sum_{i}^{n_1} \sum_{j}^{n_{2i}} \sum_{k}^{n_{3ij}} y_{hijk} \qquad (10)$$

فرض کنید که نمونه‌گیری با احتمال برابر و بدون جایگذاری در هر یک از سه مرحله نمونه‌گیری در طبقه انجام می‌شود. همان‌گونه که در قسمت (پ) انجام شد، چنین احتمال‌هایی را در فرمول عمومی (۴) جایگزین می‌کنیم. می‌توان نشان داد که

$$(11)\quad \mathrm{var}(y_{h3}) = n_1\sigma^2_{y_h}(1-F_1) + F_1\sum_{i}^{N_1} n_{2i}\sigma^2_{y_{hi}}(1-F_{2i}) + F_1\sum_{i}^{N_1} F_{2i}\sum_{j}^{N_{2i}} n_{3ij}\sigma^2_{hij}(1-F_{3ij})$$

که $F_1 = n_1 / N_1$، $F_{2i} = n_{2i} / N_{2i}$ و $F_{3i} = n_{3ij} / N_{3ij}$ کسرهای نمونه‌گیری هستند.

$$\sigma^2_{y_{hi}} = \frac{1}{N_1 - 1}\sum_{i}^{N_1}(y_{hi} - \frac{1}{N_1}\sum_{i}^{N_1} y_{hi})^2$$

$$\sigma^2_{y_{hi}} = \frac{1}{N_{2i} - 1}\sum_{j}^{N_{2i}}(y^*_{hij} - \frac{1}{N_{2i}}\sum_{j}^{N_{2i}} y^*_{hij})^2$$

$$\sigma^2_{hij} = \frac{1}{N_{3ij} - 1}\sum_{k}^{N_{3ij}}(y_{hijk} - \frac{1}{N_{3ij}}\sum_{k}^{N_{3ij}} y_{hijk})^2$$

$$y_{hi} = \frac{n_{ni}}{N_{2i}}\sum_{j}^{N_{21}} y^*_{hij}\ ,\ y^*_{hij} = \frac{n_{3ij}}{N_{3ij}}\sum_{k}^{N_{3ij}} y_{hijk}$$

برای داده‌های متقارن، یعنی $n_1 n_{2i} n_{3ij} = n_1 n_2 n_3 = n$، براورد ناریب (۱۱) به

صورت زیر داده شده است:

$$(۱۲)\quad \hat{\mathrm{var}}(y_{h_{۲}})=n^{۲}\left(\frac{s_{i}^{۲}}{n_{۱}}(۱-F_{۱})+\frac{n_{۱}}{N_{۱}}\frac{s_{۲}^{۲}}{n_{۱}n_{۲}}(۱-F_{۲})+\frac{n_{۱}}{N_{۱}}\frac{n_{۲}}{N_{۲}}\frac{s_{۳}^{۲}}{n_{۱}n_{۲}n_{۳}}(۱-F_{۳})\right)$$

در (۱۲,۳) همه عبارت‌ها بعد از عبارت اول در کسرهای نمونه‌گیری مرحله قبل، $(n_{۱}/N_{۱})(n_{۲}/N_{۲})$.....، ضرب شده و $s^{۲}$ نیز جایگزین $\sigma^{۲}$ شده است. توجه کنید که اگر $n_{۱}/N_{۱}$ قابل چشم‌پوشی باشد، تمام عبارت‌های دیگر پس از عبارت نخست نیز قابل چشم پوشی هستند.

چ) $\mathrm{var}(y_{h_{۲}})$ برای حالت بدون جایگذاری با احتمال برابر در نمونه‌گیری دو مرحله‌ای:

نتایج دو مرحله‌ای، با قرار دادن $N_{۱}=n_{۱}=۱$ در (۱۱) پس از تغییرات مناسب در زیرنویس‌ها حاصل می‌شود.

ح) $\mathrm{var}(y_{h۳})$ برای حالت با جایگذاری با احتمال متناسب با اندازه جامعه (pps) برای دو مرحله نخست و بدون جایگذاری با احتمال برابر برای مرحله سوم:

استفاده از نمونه‌گیری با احتمال‌های برابر در نمونه‌گیری چند مرحله‌ای به‌دلیل بزرگ شدن واریانس، به ندرت صورت می‌گیرد. وقتی که واحدها از نظر اندازه به‌طور قابل ملاحظه‌ای متفاوت باشند، نمونه‌گیری با احتمال برابر منجر به بزرگ شدن واریانس‌ها می‌شود. در مقابل، در حالت متقارن وقتی که تمام واحدها در هر مرحله‌ای اندازه یکسان داشته باشند، این مشکل پیش نمی‌آید. بنا بر این ناچار هستیم که برای کاهش واریانس نمونه‌گیری، طرح نمونه‌گیری دیگری را جستجو کنیم.

با تنوع دادن به احتمال‌ها در هر مرحله، ممکن است به هدف بالا دست یابیم. اگر احتمال کلی انتخاب یک عضو منفرد در یک نمونه‌گیری چند مرحله‌ای $\frac{n}{N}$ باشد، آن نمونه‌گیری را به‌دلیل این که اعضا نمونه به‌طور مساوی وزن‌دهی شده‌اند، خود وزن گویند. بنا بر این، واریانس نمونه می‌تواند برای برخی براوردها کاهش یابد. یک راه ساده برای دست‌یابی به طرح نمونه‌گیری pps خود وزن عبارت است از انتخاب $n_{۱}$ واحد نمونه‌گیری مقدماتی (PSU) با احتمال‌های $p_{i}^{(۱)}$ در هر انتخاب، $n_{۲i}$ واحد مرحله دوم از

هر یک از n_1 واحد نمونه‌گیری مقدماتی (PSU) با احتمال‌های $p_{ij}^{(2)}$ و n_{2ij} واحد مرحله سوم با احتمال‌های $p_{ijk}^{(3)}$ در هر انتخاب، که $P_i^{(1)} = \frac{N_{2i+}}{N}, P_{ij}^{(2)} = \frac{N_{2ij}}{N_{2i+}}$ و $P_{ijk}^{(3)} = \frac{1}{N_{2ij}}$ و زیرنویس « + » به مفهوم جمع‌بستن روی زیرنویس مورد نظر است.

لذا طرح نمونه‌گیری pps یک شرط لازم برای کاهش واریانس نمونه فراهم می‌کند، اگر $n_1 n_{2i} n_{2ij} = n_1 n_2 n_3 = n$ و $(n_1\ p_i^{(1)})(n_{2i}\ p_{ij}^{(2)})(n_{2ij}\ p_{ijk}^{(3)}) = \frac{n}{N}$ که احتمال انتخاب کلی برای هر واحد مقدماتی است.

برای مقدار کل y_{h3} در (۱۰) از نمونه‌گیری pps با جایگذاری برای دو مرحله اول و نمونه‌گیری با احتمال برابر بدون جایگذاری برای مرحله سوم، می‌توان نشان داد که

$$\begin{aligned} \mathrm{var}(y_{h3}) = {} & \frac{n_1}{N}\sum_{i}^{N_1} N_{2i+}(T_i - \bar{T})^2 + \frac{n_1}{N}\sum_{i}^{N_1} n_{2i}\sum_{j}^{N_{2i}} N_{2ij}(T_{ij} - \bar{T}_i)^2 \\ & + \frac{n_1}{N}\sum_{i=1}^{N_1} n_{2i}\sum_{j=1}^{N_{2i}} n_{2ij}\ N_{2ij}\ \sigma_{hij}^2\ (1 - F_{2ij}) \end{aligned} \qquad (13)$$

که $m_{ij} = \frac{1}{n_{2ij}}\sum_{k}^{n_{2ij}} y_{hijk}$; $\mu_{ij} = E(m_{ij})$; $T_i = \sum_{j}^{N_{2i}} T_{ij}$; $T_{ij} = n_{2ij}\mu_{ij}$ و $\bar{T} = n_1 \sum_{i}^{N_1} (\frac{N_{2i+}}{N})\bar{T}_i$ که $\bar{T}_i = \sum_{j}^{N_{2i}} \frac{(N_{2ij})}{N_{2i+}} T_{ij}$، $N_{2i+} = \sum_{j=1}^{N_{2i}} N_{2ij}$، $\sigma_{hij}^2 = \frac{1}{N_{2ij} - 1}\sum_{k=1}^{N_{2ij}} (y_{hijk} - \mu_{ij})^2$.

ممکن است یک براورد ناریب از (۱۳) یافت شود که نتیجه آن با بحث سابق در مورد نمونه‌گیری باجایگذاری در یک مرحله، سازگار باشد.

خ) $\mathrm{var}(y_{h2})$ برای pps با جایگذاری در مرحله اول و حالت بدون جایگذاری با احتمال برابر در مرحله دوم:

نتیجه برای حالت دو مرحله‌ای، با قرار دادن $n_1 = N_1 = 1$ و ایجاد تغییرات مناسب در

علائم، از (۱۳) به دست می‌آید.

به‌طور مشابه، می‌توانیم واریانس نسبت و کواریانس نسبت‌ها را نیز به دست آوریم.

۴- خطی‌سازی

براوردهای نسبتی می‌تواند با یک بسط سری‌های تیلور تحت جمع، خطی شود. سپس واریانس سهم متغیر از این بسط، شبیه به واریانس نسبت اصلی است. هر نسبت از متغیرهای $u_1,\ldots,u_k$ را با تابع $f(u_1,u_2,\ldots,u_k)$ نشان می‌دهیم. داریم:

$$\mathrm{var}(f(u_1,\ldots,u_k)) \approx \mathrm{var}(\sum_i^k u_i \frac{\partial F(u_1,\ldots,u_k)}{\partial u_i}) \quad (۱۴)$$

که $(i=1,\ldots,k)\quad E(u_i)=U_i$ و علامت « $\approx$ » به مفهوم این است که هر دو طرف علامت به‌طور تقریبی مساوی هستند.

مثال ۱) فرض کنید x و y متغیرهای تصادفی با مقادیر مورد انتظار X و Y باشند. نسبت $\frac{x}{y}$ را در نظر بگیرید. واریانس نسبت با واریانس سهم متغیر از یک بسط خطی از آن، تقریب زده شده است:

$$\mathrm{var}(\frac{x}{y}) \approx \mathrm{var}(\frac{x-Ry}{Y}) \quad (۱۵)$$

که $R=\frac{X}{Y}$.

مثال ۲) نسبت $X'=(\frac{x}{y})\hat{Y}$ اغلب برای اهداف براورد استفاده شده است که x و y متغیر هستند در حالی که $\hat{Y}$ یک عدد معلوم است.

$$\mathrm{var}(\frac{x}{y}\hat{Y}) \approx \mathrm{var}(\frac{\hat{Y}}{Y}(x-Ry)) \quad (۱۶)$$

با استفاده از این دو روش که در بخش ۳ و ۴ ارائه شد، می‌توانیم یک واریانس برای

براورد نسبتی پیچیده به دست آوریم. ما این واریانس را با استفاده از یک مثال واقعی در بخش ۵ ارائه خواهیم کرد.

۵- یک مثال و خلاصه بحث

براورد x برای مشخصه X جامعه در آمارگیری جاری جمعیتی به‌شکل زیر است: (گزارش فنی ۴۰، اداره سرشماری، صفحه ۱۵۵)

$$X' = \sum_{a}^{A} \frac{x_{aSR} + \sum_{c=1}^{C} \frac{x_{acNS}}{z_c} Z_c}{y_{aSR} + \sum_{c=1}^{C} \frac{y_{acNS}}{z_c} Z_c} \hat{Y}_a \qquad (۱۷)$$

که زیرنویس $c = 1,\ldots,C$، (C برابر با ۴۸ سلول از رسته رنگ- محل سکونت) برای طبقات غیر خود نماینده (NS) از تعدیل نسبتی اول و $a = 1,\ldots,A$ (A برابر با ۶۰ سلول از رسته‌های سن- جنسیت- نژاد) از تعدیل نسبتی دوم حاصل می‌شود. ما رابطه (۱۷) را به‌شکل زیر بیان می‌کنیم.

$$X' = \sum_{a}^{A} \frac{x'_a}{y'_a} \hat{Y}_a \qquad (۱۸)$$

که x'_a و y'_a در (۱۷) تعریف شده‌اند، عبارت‌های x_{aSR}، y_{aSR}، x_{acNS}، y_{acNS}، z_c، Z_c و $\hat{Y}_a$ در (۱۷) به‌شکل زیر تعریف می‌شوند:

x_{aSR} = مقدار کل موزون نمونه، از آخرین واحدهای نمونه‌گیری (USU) در واحدهای نمونه‌گیری مقدماتی (PSUs) خود نماینده (SR)، برای جامعه با ویژگی مطلوب x در aامین رسته سن- جنسیت- نژاد. وزن‌ها عبارت از عکس احتمال انتخاب USUها هستند. در عمل، وزن‌ها که با w_{sij} نشان داده می‌شوند، شامل وزن‌دهی خاص و فاکتورهای تعدیل برای بی‌پاسخی نیز هستند.

y_{aSR} = همان‌طور که برای x_{aSR} گفته شد، اما برای مقدار کل جامعه.

x_{acNS} = همان‌طور که برای x_{aSR} گفته شد، اما برای cامین رسته نژاد– محل سکونت در مورد جامعه غیر خود نماینده (NS)

y_{acNS} = همان‌طور که برای x_{ac} گفته شد، اما برای مقدار کل جامعه.

z_c = مقدار کل براورد شده در سرشماری سال ۱۹۸۰ برای جامعه NSR در cامین رسته ادغام شده نژاد– محل سکونت، بر اساس جمعیت سرشماری ۱۹۸۰ برای PSUهای نمونه NSR که روی تمام طبقات NS وزن‌دهی و جمع شده است.

z_{csij} = ۱ اگر (sij)امین فرد در PSU مربوط به نمونه NSR متعلق به cامین رسته رنگ– محل سکونت باشد و در دیگر موارد z_{csij} = ۰ .

Z_C = جمعیت سرشماری ۱۹۸۰ در طبقات NS در cامین رسته ادغام شده نژاد– محل سکونت.

Y_a = مقدار کل مستقل جامعه برای ایالات متحده در aامین رسته سن– جنسیت– نژاد برای ماه جاری CPS.

برای واریانس X'، نخست X' را تحت علامت مجموع که روی a جمع می‌بندد، خطی می‌کنیم و سپس فقط سهم متغیر از بسط خطی را در نظر می‌گیریم. در مرحله دوم علامت‌های مجموع را تعویض کرده، آن‌ها را برای نمونه‌گیری به جلو منتقل می‌کنیم و برای براوردهای نسبت جمع می‌بندیم. در مرحله سوم واریانس سرجمع حاصل می‌تواند با استفاده از فرمولی که در بخش ۳ ارائه شد، به دست آید.

این مراحل به‌شکل زیر نشان داده می‌شود:

مرحله اول– (۱۸) را خطی کرده و سهم متغیر را از آن استخراج نمایید. واریانس براورد نسبت با آن سهم خطی شده تقریب زده می‌شود.

$$\mathrm{var}(X') \approx \mathrm{var}(\sum_{a}^{A} \gamma_a (x'_a - \frac{\tilde{X}_a}{\tilde{Y}_a} y'_a)) \qquad (۱۹)$$

از (۱۷) و (۱۹) می‌توان نوشت:

$$\mathrm{var}(X') \approx \mathrm{var}(\sum_{a}^{A} \gamma_a ((x_{aSR} + \sum_{c}^{C} \frac{x_{acNS}}{z_c} Z_c) - \frac{\tilde{X}_a}{\tilde{Y}_a} (y_{aSR} + \sum_{c}^{C} \frac{y_{acNS}}{z_c} Z_c)))$$

که $\gamma_a = \frac{\hat{Y}_a}{\tilde{Y}_a}$، $\tilde{X}_a = E(x'_a)$ و $\tilde{Y}_a = E(y'_a)$.

عبارات مربوط به طبقات SR و طبقات NS را به‌طور مجزا جمع‌آوری می‌کنیم و براورد نسبتی را با جمع زدن روی c برای دومین مرتبه خطی می‌کنیم، سپس به دست می‌آید:

$$\operatorname{var}(X') \approx \operatorname{var}(\sum_a^A \gamma_a (x_{aSR} - \frac{\tilde{X}_a}{\tilde{Y}_a} y_{aSR}) + \sum_a^A \gamma_a \sum_c^C \frac{Z_c}{\tilde{Z}_c}(x_{acNS} - \frac{\tilde{X}_a}{\tilde{Y}_a} y_{acNS} + \frac{\tilde{X}_a}{\tilde{Y}_a}\frac{\tilde{Y}_{ac}}{\tilde{Z}_c} z_c - \frac{\tilde{X}_{ac}}{\tilde{Z}_c} z_c)) \tag{۲۰}$$

که

$\tilde{Z}_c = E(z_c)$، $\tilde{Y}_{ac} = E(y_{acNS})$، $\tilde{X}_{ac} E(x_{acNS})$، $x_{aSR} = \sum_s^{L_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} x_{asij}$،

$y_{aSR} = \sum_S^{L_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} y_{asij}$، $x_{acNS} = \sum_S^{n'_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} x_{acsij}$،

$y_{acNS} = \sum_S^{n'_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} y_{acsij}$، $z_c = \sum_S^{n'_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} z_{csij}$

L_1 تعداد SR-PSUها، n_{2s} تعداد واحدهای نمونه‌گیری مرحله دوم (SSUs) در sامین SR-PSU و n_{3si} تعداد واحدهای نمونه‌گیری مرحله سوم (TSUs) در iامین SSU می‌باشد. n'_1 تعداد NS-PSUهای نمونه‌گیری شده است.

مرحله دوم- با استفاده از تعاریف بالا از x_{aSR}، y_{aSR}، x_{acNS}، y_{acNS} و z_c با مجموع‌های سه‌گانه، ما علامت‌های جمع ناشی از نمونه‌گیری را به جلوی علامت‌های براورد نسبتی برای سن– جنسیت– رنگ و سکونت– نژاد، منتقل می‌کنیم. برای سن– جنسیت– رنگ و سکونت– نژاد، روی a و c جمع می‌بندیم، فقط علائم مجموع مربوط به نمونه‌گیری باقی می‌ماند، (۲۰) را می‌توان به صورت زیر نوشت:

$$(۲۱)\quad \mathrm{var}(X') \cong$$

$$\mathrm{var}\Big(\sum_{S}^{L_1}\sum_{i}^{n_{S2}}\sum_{j}^{n_{s2i}}(B_{+sij}+C_{+sij})+\sum_{S}^{n'_1}\sum_{i}^{n_{S2}}\sum_{j}^{n_{s2i}}(B'_{++sij}-C'_{++sij}+D'_{++sij}-D_{++sij})\Big)$$

که می‌تواند به صورت زیر بیان شود:

$$(۲۲)\qquad \mathrm{var}(X') = \mathrm{var}\Big(\sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} t_{sij} + \sum_{S}^{n'_1}\sum_{i}^{n_{S2}}\sum_{j}^{n_{S2I}} t'_{sij}\Big)$$

که t_{sij} و t'_{sij} در آن‌ها به‌خوبی تعریف شده‌اند. شش عبارت در (۲۱) به صورت زیر تعریف می‌شوند:

(۲۱-۱):

$$\sum_{a}^{A}\gamma_a x_{aSR} = \sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{2si}}\sum_{a}^{A} B_{asij} = \sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{2si}} B_{+sij}\,,\ B_{asij}=\gamma_a\, w_{sij}x_{asSRij};$$

(۲۱-۲):

$$\sum_{a}^{A}\frac{\widetilde{X}_a}{\widetilde{Y}_a}\gamma_a y_{aSR} = \sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{as2i}}\sum_{a}^{A} C_{asij} = \sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} C_{sij}\,,\ C_{asij}=\frac{\widetilde{X}_a}{\widetilde{Y}_a}\gamma_a w_{sij}y_{asSRij};$$

برای عبارت دوم،

(۲۱-۳):

$$\sum_{a}^{A}\sum_{c}^{C}\gamma_a x_{ac} = \sum_{S}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}}\sum_{a}^{A}\sum_{c}^{C} B'_{acsij} = \sum_{s}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} B'_{++tsij}\,,\ B'_{acsij}=\gamma_a\, w_{sij}\, x_{acNSsij};$$

(۲۱-۴):

$$\sum_{a}^{A}\sum_{c}^{C}\gamma_a\frac{\widetilde{X}_a}{\widetilde{Y}_{ac}}y_{ac} = \sum_{S}^{n_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}}\sum_{a}^{A}\sum_{c}^{C} C'_{acsij} = \sum_{s}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} C'_{++sij}\,,\ C'_{acsij}=\gamma_a\frac{\widetilde{X}_a}{\widetilde{Y}_{ac}}w_{sij}y_{acNSsij};$$

(۲۱-۵):

$$\sum_{a}^{A}\sum_{c}^{C}\gamma_a \frac{\tilde{X}_a}{\tilde{Y}_a}\frac{\tilde{Y}_{ac}}{\tilde{Z}_c} Z_c = \sum_{S}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}}\sum_{a}^{A}\sum_{c}^{C} D'_{acsij} = \sum_{s}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} D'_{++sij} ,$$

$$D'_{csij} = \gamma_a \frac{\tilde{X}_a}{\tilde{Y}_a}\frac{\tilde{Y}_{ac}}{\tilde{Z}_c} w_{sij} z_{acsij};$$

(۲۱-۶):

$$\sum_{a}^{A}\sum_{c}^{C}\gamma_a \frac{\tilde{Y}_{ac}}{\tilde{Z}_c} z_c = \sum_{S}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}}\sum_{a}^{A}\sum_{c}^{C} D_{acsij} = \sum_{s}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} D_{++sij}$$

$$D_{acsij} = \gamma_a \frac{\tilde{Y}_{ac}}{\tilde{Z}_C} w_{sij} z_{csij};$$

مرحله سوم– واریانس X' – عبارت‌های اول و دوم به ترتیب حاصل از SR-PSUها و برای NS-PSUها هستند. می‌توانیم واریانس‌های آن‌ها را به‌طور جداگانه محاسبه کرده و با یکدیگر جمع کنیم.

از آن‌جا که برای طبقه‌های غیر خود– نماینده، نمونه‌ها برای دو مرحله اول با طرح PPS با جایگذاری و برای مرحله آخر با احتمال‌های برابر بدون جایگذاری انتخاب شده است، وقتی که داده‌های طبقات خود– نماینده، همان‌گونه که در (۷) نشان داده شد، فقط از دو مرحله حاصل شده باشند، می‌توانیم (۶) را برای براورد واریانس طبقات غیر خود– نماینده به کار ببریم. واریانس X' عبارت است از:

$$\operatorname{var}\left(\sum_{S}^{L_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} t_{sij}\right) + \operatorname{var}\left(\sum_{s}^{n'_1}\sum_{i}^{n_{s2}}\sum_{j}^{n_{s2i}} t'_{sij}\right) = \sigma^2_{sSR} + \sigma^2_{sNS}$$

عبارت اول مجموع L_1 واحد نمونه‌گیری مقدماتی(PSUs) خود– نماینده است، در حالی که عبارت دوم مجموع n'_1 واحد نمونه‌گیری مقدماتی(PSUs) غیر خود– نماینده است.

ممکن است فردی بخواهد واریانس نسبت X'/Y' را داشته باشد، که X' همان‌طور که نشان داده شد خطی شده است، Y' براورد نسبتی دیگری است که با

استفاده از (۱۷) حاصل می‌شود و $\mathrm{var}(X'/Y')$ نیز به طریقی مشابه به دست می‌آید.

ممکن است که به‌طور مکرر از روش دلتا برای $\mathrm{var}(X')$ استفاده کنیم. با استفاده از این مثال، کاربرد نخست می‌تواند برای رسته‌های سن– جنسیت– نژاد به کار رود، کاربرد دوم برای سلول‌های سکونت– رنگ و در نهایت کاربرد سوم برای طرح سه مرحله‌ای مورد استفاده قرار گیرند. در اینجا فرض‌های دیگری، بجز آن‌هایی که از تقریب سری‌های تیلور حاصل می‌شود، فرض دیگری وجود ندارد.

روش‌هایی که در این مقاله ارائه شد، از این جهت که می‌توانیم روش‌های نمونه‌گیری و براورد را روی واریانس نمونه منعکس کنیم، ممکن است یکی از بهترین روش‌های براورد واریانس باشد. عملکرد این روش با استفاده از یک روش تجربی قابل تشخیص است.

این روش ممکن است که برای متغیرهای گسسته نیز به‌شکل متغیرهای پیوسته به کار گرفته شود.

مرجع‌ها

[1] Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society, B, 15, 262.

[2] Hidiroglou, M. A. and Rao, J.N.K. (1983). Chi-Square Tests for the Analysis of Three Way Contingency Tables from the Canada Health Survey. Statistics Canada, Ottawa, Canada.

[3] Shah, B.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Unpublished document, RTI.

[4] Kendall M. G. and Stuart A. S. (1968). The Advanced theory of Statistics, Vol. 3. Hafner Publishing Company, New York.

[5] Woodruff, Ralph S.(1971). Simple Method for Approximating Variance of a Complicated Estimate. Journal of the American Statistical Association, Volume 66, June, pp411-414.