گزیده‌مطالب آماری
سال ۱۹، شماره‌ی ۲، پاییز و زمستان ۱۳۸۷، صص ۲۴۳-۲۲۵

برخی از کاربردهای رایج نمونه‌گیری سیستماتیک

لی-چون زنگ[†]

مرکز آمار نروژ

مترجم: شیرین گلچی

پژوهشکده‌ی آمار

چکیده. نمونه‌گیری سیستماتیک فنی است که در آمارگیری نمونه‌ای به صورت گسترده مورد استفاده قرار می‌گیرد و اجرای آن چه زمانی که واحدها با احتمال برابر انتخاب می‌شوند و چه در حالت انتخاب با احتمالات متناسب با اندازه‌های کمکی آسان است. این فن در صورتی که بتوان به اثرهای مطلوب طبقه‌بندی در فهرست کردن واحدها دست یافت، می‌تواند از کارایی زیادی برخوردار باشد. عیب اصلی آن این است که روش نااریبی برای براورد واریانس نمونه‌گیری وجود ندارد و این‌که زمانی که مرتب کردن جامعه بر اساس اطلاعات نادرست صورت گرفته باشد، نمونه‌گیری سیستماتیک ممکن است ضعیف عمل کند. در این مقاله جنبه‌ای از نمونه‌گیری سیستماتیک را که در گذشته به آن توجه زیادی نشده است، مورد بررسی قرار می‌دهیم. نشان داده شده است که در برخی موقعیت‌های متعارف، زمانی که تحت یک مدل مفروض برای جامعه، نمونه‌گیری سیستماتیک به طور متوسط دارای کارایی یکسانی با دیگر روش‌های نمونه‌گیری تصادفی متناظر است، واریانس نمونه‌گیری در نمونه‌گیری سیستماتیک به مراتب بیش‌تر نوسان می‌کند. استفاده از نمونه‌گیری سیستماتیک با مخاطره‌ای در ارتباط است که به طور کلی با بزرگ شدن کسر نمونه‌گیری افزایش می‌یابد. این می‌تواند برای نمونه‌های بزرگ از جامعه‌های کوچک در حالت نمونه‌گیری یک‌مرحله‌ای، یا زیرنمونه‌های بزرگ از زیرجامعه‌های کوچک در حالت نمونه‌گیری چندمرحله‌ای، بسیار مخرب باشد.

[†] Zhang, L.C. (2008). On Some Common Practices of Systematic Sampling. *Journal of Official Statistics*, **24**, 557-569.

واژگان کلیدی: تصمیم آماری؛ مخاطره‌ی بیزی مرتبه دوم؛ طرح استوار؛ آمارگیری پانلی.
دریافت: ۱۳۸۸/۲/۱ پذیرش: ۱۳۸۸/۳/۴

۱- مقدمه

داده‌های گم‌شده مشکل عمومی در تحقیقات مبتنی بر آمارگیری است. نادیده گرفتن هر گونه داده‌های گم‌شده با استفاده از یک تحلیل موردی کامل می‌تواند نتایجی اریب را ایجاد کند. اریبی هنگامی اتفاق می‌افتد که شرکت‌کنندگان دارای داده‌های کامل، به‌طور سیستماتیک با شرکت‌کنندگان دارای داده‌های گم‌شده تفاوت داشته باشند. به‌خصوص مطالعات طولی مستعد چنین اریبی‌هایی هستند زیرا داده‌های گم‌شده به‌مرور زمان درنتیجه‌ی بی‌پاسخی و انصراف شرکت‌کنندگان انباشته می‌شود. یکی از روش‌های جبران داده‌های گم‌شده جانهی است. طی بیست سال گذشته انبوه نوشتگان درباره‌ی نظریه و روش‌شناسی جانهی رشد قابل ملاحظه‌ای داشته است و نرم افزارهایی نیز در راستای آن تکامل یافته است. ولی در زمینه‌ی جانهی داده‌های طولی کار نسبتاً کمی صورت گرفته است.

برای جانهی چند رهیافت نظری وجود دارد. راگوناتان [۱۵] این قبیل رهیافت‌ها را بازنگری و سه رده را شناسایی کرده است: معادلات براورد موزون، جانهی چندگانه و فرمول‌های مبتنی بر درستنمایی. ابراهیم و دیگران [۸] رهیافت تماماً بیزی را به‌عنوان رده‌ی چهارم درنظر گرفتند.

معادلات براوردگر موزون (WEE)، آمارهای ثبتی دارای داده‌های کامل را وزن‌دار می‌کنند تا موارد مشابه دارای داده‌های گم‌شده را جبران کنند. در این اواخر نوشتگان بر بهبود براورد واریانس ([۱۹] و [۲۰]) تمرکز داشته‌اند زیرا WEE، هنگامی‌که تعدیل نشده باشد، واریانس واقعی داده‌ها را کم‌براورد می‌کند. اجرای WEE در حال حاضر به جای شیوه‌های استاندارد در بسته‌های نرم‌افزار آماری اصلی، بر الگوریتم‌های مدل-ویژه و کاربر-تعریف‌شده تکیه دارد. جانهی چندگانه (MI) از شبیه‌سازی بیز برای پرکردن داده‌های گم‌شده استفاده می‌کند که از تلفیق نتایج به دست آمده از چندین مجموعه‌ی جانهی‌شده‌ی مکرر به دست می‌آیند. برای مشاهده‌ی پوشش فراگیر جانهی چندگانه به رابین [۲۲] مراجعه کنید. مدل‌های تماماً بیزی (FB) روش‌های MI را با شبیه‌سازی توأم توزیع متغیرهای دارای داده‌های گم‌شده و نیز پارامترهای نامعلوم در یک معادله‌ی رگرسیون گسترش می‌دهند. در FB مدل‌های تحلیل و جانهی کاملاً و به‌طور هم‌زمان

مشخص هستند. همچنین فنون ماکسیمم درستنمایی نیز بر مدل‌های کاملاً مشخص تکیه دارند اما از این جهت که براوردهای پارامترها با استفاده از تقریب‌های مبتنی بر درستنمایی به جای شبیه‌سازی بیزی ایجاد شده‌اند، با FB فرق دارند.

رهیافت‌های ماکسیمم درستنمایی برای جانهی اغلب در بسته‌های نرم‌افزاری اصلی، غیر قابل کنترل‌اند. اجرای روش‌ها بر فرض‌های محکم در مورد الگوهای گم‌شدگی که مکرراً در بررسی داده‌های پیچیده زیر پا گذاشته می‌شوند، تکیه دارد. در حالی که شیوه‌های MI در تعدادی بسته‌های نرم‌افزاری مانند SAS [۲۳]، Stata [۲۷]، S-Plus [۹] و R [۱۸] وجود دارند، این روش‌ها عموماً بر این فرض تکیه دارند که داده‌ها، نرمال چندمتغیره‌اند یا می‌توانند با توزیع نرمال چندمتغیره تقریب زده شوند [۲۴]. بیش‌تر کارهای اخیر درباره‌ی معادلات رگرسیونی زنجیره‌ای، منجر به افزوده شدن بسته‌هایی شده است که داده‌های رسته‌ای را نیز شامل می‌شوند: MICE در S-PLUS [۲۹]، Ice در Stata [۲۱] و IVEware برای SAS [۱۶]. ولی تألیف‌کنندگان هنوز مشکلاتی در اضافه کردن داده‌های طولی در روش‌های جانهی با استفاده از این برنامه‌ها دارند. فنون FB در جانهی طولی بسیار مناسب هستند، زیرا می‌توانند ساختار سلسله مراتبی را در فرایند مدل‌سازی ادغام کنند و مانند رگرسیون زنجیره‌ای، این قابلیت را دارند که به‌طور سیستماتیک به داده‌های رسته‌ای نیز بپردازند. بسته‌های نرم‌افزار WinBUGS [۲۶] و MLwiN [۱۷] هر دو از چارچوب FB استفاده می‌کنند. کاولس [۵] و وودورت [۳۰] هر دو شرح مختصر و مفیدی برای WinBUGS تهیه کرده‌اند، درحالی‌که کارپنتر و کنوارد [۲] وکانگدون [۳] مثال‌های مقدماتی جانهی FB با داده‌های گم‌شده را مطرح کرده‌اند. پتیت [۱۳] و کیو و همکاران [۱۴] تحلیل‌های دقیقی درزمینه‌ی داده‌های رسته‌ای گم‌شده ارائه کرده‌اند.

هدف این مقاله نشان دادن توانایی WinBUGS برای جبران داده‌های طولی گم‌شده با تمرکز خاص بر داده‌های گم‌شده‌ی متغیر کمکی است. ما این کار را با نگاه بر تحلیل طولی میزان وقوع دیابت در زنان استرالیایی انجام می‌دهیم. در بخش ۲ مثالی تشویق‌کننده از مطالعه‌ی طولی استرالیا درباره‌ی سلامت زنان را معرفی می‌کنیم. در بخش ۳ یک مدل کاملاً بیزی برای میزان وقوع دیابت بدون و با داده‌های گم‌شده‌ی متغیر کمکی را تعیین می‌کنیم. در بخش ۴ اجرای آن را در WinBUGS ارائه کرده و

نتایج را در بخش ۵ توصیف می‌کنیم. در بخش ۶ یک ماکروی کلی SAS (که WinBUGS نامیده می‌شود) برای تحلیل مدل‌های طولی با داده‌های گم‌شده‌ی متغیر کمکی ارائه می‌دهیم. با یک بحث و تعدادی توصیه در بخش ۷ نتیجه‌گیری می‌کنیم.

۲- جامعه‌های همگن

ابتدا نمونه‌گیری سیستماتیک با احتمالات برابر را از یک ترتیب ثابت جامعه که ممکن است با متغیر مورد نظر ناهمبسته تلقی شود، در نظر بگیرید. فرض کنید اندازه‌ی نمونه n و فاصله‌ی نمونه‌گیری k باشد. برای سادگی فرض می‌کنیم k به صورت طبیعی یک مقدار صحیح است که در $N=nk$ صدق می‌کند. mامین نمونه‌ی سیستماتیک را با s_m نمایش می‌دهیم، یعنی $s_m=\{m,m+k,m+۲k,\ldots,m+(n-۱)k\}$. فرض کنید $\bar{y}_m$ میانگین نمونه‌ی متناظر باشد که براوردگر ناااریبی از میانگین جامعه است که با $\bar{Y}=\frac{۱}{N}\sum_{i\in U}y_i$ نشان داده می‌شود. واریانس نمونه‌گیری $\bar{y}_m$ به صورت $\mathrm{V}_{sys}=k^{-۱}\sum_{m=۱}^{k}\left(\bar{y}_m-\bar{Y}\right)^۲$ به دست می‌آید که ممکن است از واریانس میانگین نمونه‌ی تصادفی ساده که با $\mathrm{V}_{srs}=\left(n^{-۱}-N^{-۱}\right)\sigma^۲$ نشان داده می‌شود و در آن $\sigma^۲=\frac{۱}{N-۱}\sum_{i\in U}\left(y_i-\bar{Y}\right)^۲$، بیش‌ترباشد یا نباشد.

همان‌طور که قبلاً اشاره شد دو نتیجه وجود دارد که نشان می‌دهد SRS و SYS دارای مخاطره‌های بیزی یکسان هستند، یعنی $\mathrm{E}_\theta\left(\mathrm{V}_{srs}\right)=\mathrm{E}_\theta\left(\mathrm{V}_{sys}\right)$. ما مخاطره‌ی بیزی مرتبه‌ی دوم آن‌ها را تحت مدل همگن

$$\mathrm{E}(y_i)=\mu \quad , \quad \mathrm{E}\left((y_i-\mu)^r\right)=\mu_r<\infty \qquad (۳)$$

برای $i\in U$، که در آن برای سادگی، E را جایگزین E_θ کرده‌ایم و به ازای هر $i\neq j$، y_i از y_j مستقل است، بررسی خواهیم کرد. داریم $\mathrm{E}\left(\mathrm{V}_{srs}\right)=\mathrm{E}\left(\mathrm{V}_{sys}\right)=\left(\frac{۱}{n}-\frac{۱}{N}\right)\mu_۲=\left(۱-\frac{۱}{k}\right)\mu_{۲,n}$ که در آن $\mu_{r,n}$ نمایانگر گشتاور مرکزی rام $\bar{y}_m$ است. این، حالت خاصی از قضیه‌ی کلی‌تر ۸-۵ کاکران [۲] است که در آن واریانس مدل y_i می‌تواند روی واحدها تغییر کند.

علاوه بر این، بنابر نتیجه‌ی واریانسِ واریانسِ تجربی (به عنوان مثال [۱۲])، بـه دسـت می‌آوریم

$$\mathrm{V}\left(\mathrm{V}_{srs}\right)=\left(\frac{۱}{n}-\frac{۱}{N}\right)^{۲}\mathrm{V}\left(\sigma^{۲}\right)=\left(۱-\frac{۱}{k}\right)^{۲}\mu_{۲,n}^{۲}\left(\frac{۲}{N-۱}+\frac{\gamma_{۲}}{N}\right)$$

که در آن $\gamma_{۲}=\frac{\mu_{۴}}{\mu_{۲}^{۲}-۳}$ ضریب کشیدگی y_i است. به همین ترتیب داریم

$$\mathrm{V}\left(\mathrm{V}_{sys}\right)=\left(\frac{k-۱}{k}\right)^{۲}\mathrm{V}\left(\frac{۱}{k-۱}\sum_{m=۱}^{k}\left(\overline{y}_{m}-\overline{Y}\right)^{۲}\right)=\left(۱-\frac{۱}{k}\right)^{۲}\mu_{۲,n}^{۲}\left(\frac{۲}{k-۱}+\frac{\gamma_{۲,n}}{k}\right)$$

که در آن $\gamma_{۲,n}=\frac{\gamma_{۲}}{n}$ ضریب کشیدگی $\overline{y}_{m}$ است. نتیجه می‌شود که ضریب تغییراتِ (CV) V_{srs} و V_{sys} به ترتیب عبارت‌اند از

$$\mathrm{CV}\left(\mathrm{V}_{srs}\right)=\left(۱-\frac{۱}{k}\right)\left(\frac{۲}{N-۱}+\frac{\gamma_{۲}}{N}\right)^{\frac{۱}{۲}}$$

(۴)

و

$$\mathrm{CV}\left(\mathrm{V}_{sys}\right)=\left(۱-\frac{۱}{k}\right)\left(\frac{۲}{k-۱}+\frac{\gamma_{۲}}{k}\right)^{\frac{۱}{۲}}$$

در صـورتی کـه جامعـه بـه انـدازه‌ی کـافی بـزرگ باشـد داریـم $\mathrm{CV}\left(\mathrm{V}_{sys}\right)\approx\left(۱-\frac{۱}{k}\right)\sqrt{\frac{۲}{k-۱}}$ که همراه با کسر نمونه‌گیری افزایش می‌یابد. بـرای مثـال، کسر نمونه‌گیری کل در آمارگیری نیروی کار نـروژ در حـدود $\frac{۱}{۱۴۰}$ اسـت، بـه نحـوی کـه $\mathrm{CV}\left(\mathrm{V}_{sys}\right)\approx$ ٪۱۲. در مقایسه، $\mathrm{CV}\left(\mathrm{V}_{srs}\right)=\mathrm{O}\left(\frac{۱}{\sqrt{N}}\right)$ و مخـاطره‌ی بیـزی مرتبـه‌ی دوم نمونه‌گیری تصادفی ساده ناچیز است. انتخاب نمونه‌های سیستماتیک از یک فهرست ظاهراً تصادفی، اما در حقیقت ثابت جامعه یـک اقـدام مبتنـی بـر شـانس و بـدون انتظـار حصول بهبودی در مقایسه با نمونه‌گیری تصادفی ساده است. به سادگی، کنتـرل کمتـری روی واریانس واقعی نمونه‌گیری وجود دارد که ممکن است به شکل قابـل ملاحظـه‌ای از امیـد ریاضـی خـود انحـراف داشـته باشـد. ایـن امـر بـرای نمونـه‌گیـری سیـستماتیک طبقه‌بندی‌شده در مقایسه با نمونه‌گیری تصادفی ساده‌ی طبقه‌بندی‌شده نیز صـادق اسـت. در نمونه‌گیری دومرحله‌ای که در آن نمونه‌گیری سیستماتیک برای زیرنمونه‌گیری واحـدها داخل یک واحد نمونـه‌گیـری اولیـه (PSU) مـورد اسـتفاده قـرار مـی‌گیـرد، آن‌چـه روی

مخاطره‌ی بیزی مرتبه‌ی دوم تأثیرگذار است کسرهای نمونه‌گیری داخل PSUها است. بنابراین نمونه‌گیری سیستماتیک در حالت نمونه‌گیری چندمرحله‌ای می‌تواند مخاطره‌ی بیزی مرتبه‌ی دوم زیادی داشته باشد، حتی اگر کسر نمونه‌گیری کلی کوچک باشد.

۳- جامعه‌های رگرسیونی نسبتی

حال وضعیت نمونه‌گیری سیستماتیک πps را در نظر بگیرید. در این حالت قاعده‌ی « هر kامین » برای مجموع تجمعی یک متغیر کمکی به کار می‌رود که با x_i برای $i \in U$ نشان داده می‌شود. هر فهرست ثابت U می‌تواند مورد استفاده قرار بگیرد. برای سادگی فرض می‌کنیم x_iها مقادیر صحیح‌اند. فرض کنید $X = \sum_{i \in U} x_i$، آن‌گاه طول فاصله توسط $k = \frac{X}{n}$ داده می‌شود که فرض می‌کنیم k نیز به صورت طبیعی یک عدد صحیح است. اگر از زاویه‌ی دیگری به قضیه نگاه کنیم، نمونه‌گیری سیستماتیک با احتمالات برابر همان نمونه‌گیری سیستماتیک πps است که در آن $x_i \equiv ۱$. واحد i ممکن است در x_i نمونه‌ی سیستماتیک مختلف ظاهر شود. فرض می‌کنیم احتمال شمول به گونه‌ای است که به ازای همه‌ی $i \in U$، $\pi_i = \frac{n x_i}{X} < ۱$. بر اساس هر نمونه‌ی سیستماتیک πps که با s_m برای $m = ۱, \ldots, k$، نشان داده می‌شود، براوردگر Y عبارت است از

$$\hat{Y}_m = \sum_{i \in s_m} \frac{y_i}{\pi_i} = \frac{X}{n} \sum_{i \in s_m} b_i = X \bar{b}_m$$

که $b_i = \frac{y_i}{x_i}$ و $\bar{b}_m = \sum_{i \in s_m} \frac{b_i}{n}$.

داریم $\mathrm{E}_{sys}\left(\bar{b}_m\right) = \frac{Y}{X}$ و $E_{sys}\left(\bar{y}_m\right) = Y$ و

$$\mathrm{V}_{sys}\left(\hat{Y}_m\right) = X^{۲}\, \mathrm{V}_{sys}\left(\bar{b}_m\right) = X^{۲} \left\{ \frac{۱}{k} \sum_{m=۱}^{k} \left(\bar{b}_m - \frac{Y}{X} \right)^{۲} \right\}$$

حال $\bar{b}_m$ بهترین براوردگر خطی ناریب (BLUE) برای β تحت مدل زیر است

$$y_i = x_i \beta + x_i \varepsilon_i \qquad (۵)$$

که در آن $\mathrm{E}(\varepsilon_i)=0$ و $\mathrm{V}(\varepsilon_i)=\mu_2$ و برای $i\neq j\in U$، $\mathrm{cov}(\varepsilon_i,\varepsilon_j)=0$، یعنی یک مدل رگرسیونی نسبتی با واریانس باقیمانده‌ی متناسب با x_i^2. داریم

$$\mathrm{V}_{sys}(\bar{b}_m)=\frac{1}{k}\sum_{m=1}^{k}(\bar{b}_m-\beta)^2-\left(\frac{Y}{X}-\beta\right)^2\doteq\frac{1}{k}\sum_{m=1}^{k}(\bar{b}_m-\beta)^2=\frac{1}{k}\sum_{m=1}^{k}\bar{\varepsilon}_m^2\overset{def}{=}Z$$

که در آن $\bar{\varepsilon}_m=\sum_{i\in s_m}\frac{\varepsilon_i}{n}$، به شرط آن‌که جامعه به اندازه‌ی کافی بزرگ باشد و لذا جمله‌ی $\left(\frac{Y}{X}-\beta\right)^2$ در مقایسه با Z در مرتبه‌ی پایین‌تری باشد. داریم $\mathrm{E}(Z)=\mu_{2,n}$. به علاوه

$$Z^2=\frac{1}{k^2}\left(\sum_{m=1}^{k}\bar{\varepsilon}_m^4+\sum_{m,p;p\neq m}\bar{\varepsilon}_m^2\bar{\varepsilon}_p^2\right)$$

که در آن $\bar{\varepsilon}_m$ و $\bar{\varepsilon}_p$ لزوماً از هم مستقل نیستند زیرا ممکن است واحدهایی در هر دوی s_m و s_p ظاهر شوند. هر چند داریم

$$\mathrm{E}(\bar{\varepsilon}_m^2\bar{\varepsilon}_p^2)=n^{-4}\sum_{\substack{(i_1,i_2)\in s_m\\(j_1,j_2)\in s_p}}\mathrm{E}(\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_{j_1}\varepsilon_{j_2})$$

که در آن $\mathrm{E}(\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_{j_1}\varepsilon_{j_2})$ صفر نیست و در واقع مثبت است تنها اگر به صورت $\mathrm{E}(\varepsilon_i^4)$ یا $\mathrm{E}(\varepsilon_i^2\varepsilon_j^2)$ باشد. فرض کنید s_{mp} نمایانگر فصل مشترک s_m و s_p باشد. داریم

$$\sum_{\substack{(i_1,i_2)\in s_m\\(j_1,j_2)\in s_p}}\mathrm{E}(\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_{j_1}\varepsilon_{j_2})=\mathrm{E}\left(2\sum_{i\neq j\in s_{mp}}\varepsilon_i^2\varepsilon_j^2+\sum_{\substack{i\in s_m\\j\in s_p}}\varepsilon_i^2\varepsilon_j^2\right)$$

جمله‌ی اول با اختیار کردن $(i_1,i_2)=(j_1,j_2)$ و $(i_1,i_2)=(j_2,j_1)$، $i_1\neq i_2$، و جمله‌ی دوم با اختیار کردن $i_1=i_2$ و $j_1=j_2$، به دست می‌آید. فرض کنید s_m^p نمایانگر واحدهایی از s_m باشد که s_p شامل آن‌ها نمی‌شود و s_p^m نمایانگر واحدهایی از s_p باشد که s_m آن‌ها را شامل نمی‌شود. داریم

$$\sum_{i\in s_m,j\in s_p}\varepsilon_i^2\varepsilon_j^2=\sum_{i,j\in s_{mp}}\varepsilon_i^2\varepsilon_j^2+\sum_{i\in s_{mp},j\in s_p^m}\varepsilon_i^2\varepsilon_j^2+\sum_{i\in s_{mp},j\in s_m^p}\varepsilon_i^2\varepsilon_j^2+\sum_{i\in s_p^m,j\in s_m^p}\varepsilon_i^2\varepsilon_j^2$$

و

$$\sum_{i,j\in s_{mp}} \varepsilon_i^{۲}\varepsilon_j^{۲} = \sum_{i\in s_{mp}} \varepsilon_i^{۴} + \sum_{i\neq j\in s_{mp}} \varepsilon_i^{۲}\varepsilon_j^{۲}$$

لذا به دست می‌آوریم

$$n^{۴}\,\mathrm{E}\left(\overline{\varepsilon}_m^{۲}\overline{\varepsilon}_p^{۲}\right) = ۲\sum_{i\neq j\in s_{mp}} \mu_{۲}^{۲} + \left(\sum_{i\in s_{mp}} \mu_{۴} - \sum_{i\in s_{mp}} \mu_{۲}^{۲}\right)$$
$$+ \left(\sum_{i\in s_{mp}} \mu_{۲}^{۲} + \sum_{i\neq j\in s_{mp}} \mu_{۲}^{۲} + \sum_{\substack{i\in s_{mp}\\ j\in s_p^m}} \mu_{۲}^{۲} + \sum_{\substack{i\in s_{mp}\\ j\in s_m^p}} \mu_{۲}^{۲} + \sum_{\substack{i\in s_p^m\\ j\in s_m^p}} \mu_{۲}^{۲}\right)$$
$$= ۲c_{mp}\left(c_{mp}-۱\right)\mu_{۲}^{۲} + c_{mp}\left(\gamma_{۲}+۲\right)\mu_{۲}^{۲} + n^{۲}\mu_{۲}^{۲} = \mu_{۲}^{۲}\left(n^{۲} + c_{mp}\gamma_{۲} + ۲c_{mp}^{۲}\right)$$

که در آن c_{mp} تعداد واحدهای مشترک s_m و s_p است و دو جمله‌ی دربردارنده‌ی c_{mp} تنها در صورتی که s_{mp} تهی نباشد وجود خواهند داشت. چهارمین گشتاور مرکزی $\overline{\varepsilon}_m$ را با $\mu_{۴,n}$ نمایش می‌دهیم. حال داریم (۶)

$$\mathrm{V}\left(Z\right) = \frac{۱}{k}\left(\mu_{۴,n} - \mu_{۲,n}^{۲}\right) + \frac{\mu_{۲,n}^{۲}}{k^{۲}n^{۲}}\left(\gamma_{۲}\sum_{m\neq p} c_{mp} + ۲\sum_{m\neq p} c_{mp}^{۲}\right) = \mu_{۲,n}^{۲}\lambda_n\left(\gamma_{۲}\right)$$

که در آن

$$\lambda_n\left(\gamma_{۲}\right) = ۲\left(\frac{۱}{k} + \frac{\sum_{m\neq p} c_{mp}^{۲}}{X^{۲}}\right) + \gamma_{۲}\frac{\sum_{i=۱}^{N} x_i^{۲}}{X^{۲}}$$

زیرا $\mu_{۴,n} = \mu_{۲,n}^{۲}\left(\gamma_{۲,n}+۳\right)$، $X = nk$ و $\sum_{m\neq p} c_{mp} = \sum_{i=۱}^{N} x_i\left(x_i-۱\right)$. جمله‌ی $\sum_{m\neq p} c_{mp}^{۲}$ به طور کلی حل‌نشدنی به نظر می‌رسد، اما با ترتیب معلوم واحدها می‌توان آن را محاسبه کرد. نتیجه می‌شود که $\mathrm{CV}\left(\mathrm{V}_{sys}\left(\hat{Y}_m\right)\right) \doteq \mathrm{CV}\left(Z\right) = \sqrt{\lambda_n\left(\gamma_{۲}\right)}$.

در ضمن انواع گوناگونی از دیگر روش‌های نمونه‌گیری تصادفی πps وجود دارد. به

راحتی می‌توان نشان داد که در حالت نمونه‌گیری پواسون (PS)، ضریب تغییرات $V_{ps}\left(\hat{Y}\right)$ از مرتبه‌ی $o\left(\frac{1}{\sqrt{N}}\right)$ تحت مدل (۵) است. نتایج به طور کلی برای هر طرح نمونه‌گیری با اندازه‌ی ثابت πps صادق است به شرط آن‌که واریانس نمونه‌گیری آن از طریق یک جمله‌ی تصحیح جامعه‌ی متناهی با واریانس نمونه‌گیری PS مرتبط باشد.

۴- یک مثال عددی

برای توضیح نتایج (۴) و (۶) از لحاظ عددی، فرض کنید نمونه‌ای را شامل ۱۰ واحد از جامعه‌ای به اندازه‌ی ۱۰۰ که با $U=\{1,2,\ldots,100\}$ نشان داده می‌شود، در نظر گرفته‌ایم. متغیرهای کمکی به سادگی به صورت $x_i=i$ داده شده‌اند. متغیرهای آمارگیری y_i باید تحت مدل زیر شبیه‌سازی شوند

$$y_i = x_i + x_i^a \varepsilon_i \qquad (۷)$$

که در آن $\varepsilon_i \overset{iid}{\sim} N\left(0,\sigma^2\right)$ و $0 \le a \le 1$.

لذا واریانس شرطی y_i به شرط x_i برابر است با $x_i^{2a}\sigma^2$. در حالت $a=0$، $y_i - x_i$ از مدل همگن (۳) پیروی می‌کند. در حالت $a=1$ مدل (۵) را با $\beta=1$ داریم که استفاده از آن می‌تواند انگیزه‌ی نمونه‌گیری πps را فراهم کند.

ابتدا نمونه‌گیری πps را در نظر بگیرید. فرض کنید $a=0, 0/25, 0/5, 1$ و $\sigma=0/01, 0/1$. توجه کنید که در این‌جا تا زمانی که مقادیر منفی y بتوانند با احتمال‌های غیر قابل اغماض تولید شوند، که در این صورت پایه و اساس نمونه‌گیری πps زیر سؤال می‌رود، σ نمی‌تواند خیلی بزرگ شود. برای هر زوج (a,σ) یک جامعه‌ی $\theta=\left(y_1,\ldots,y_{100}\right)^T$ تولید می‌کنیم که برای آن سه واریانس نمونه‌گیری محاسبه می‌شود. اولین واریانس برای نمونه‌گیری سیستماتیک πps به دست می‌آید. دومین واریانس برای SPS است که یک روش نمونه‌گیری πps تصادفی تقریبی است و این واریانس با مونت کارلوی ساده محاسبه می‌شود. نهایتاً واریانس مجانبی نظری نمونه‌گیری را برای نمونه‌گیری سیستماتیک πps با جایگشت تصادفی θ پیش از انتخاب یک نمونه‌ی سیستماتیک محاسبه می‌کنیم که عبارت است از [۴]

$$V_{asy}=\sum_{i=1}^{100}\pi_i\left(1-\frac{n-1}{n}\pi_i\right)\left(\frac{y_i}{\pi_i}-\frac{Y}{n}\right)^2$$

و می‌تواند برای محک زدن کارایی دو واریانس دیگر مورد استفاده قرار بگیرد.

جدول ۱- نتایج شبیه‌سازی برای نمونه‌گیری πps بر حسب درصد

طرح		کارایی نسبی		CV واریانس نمونه‌گیری		
		سیستماتیک	*SPS*	سیستماتیک	*SPS*	نظری
$a=1$	$\sigma=0/01$	۹۸	۱۰۰	۴۱	۱۶	۱۶
	$\sigma=0/1$	۱۰۱	۱۰۰	۴۱	۱۶	۱۶
$a=0/5$	$\sigma=0/01$	۱۰۱	۱۰۰	۳۴	۱۵	۱۵
	$\sigma=0/1$	۹۹	۹۹	۳۴	۱۵	۱۵
$a=0/25$	$\sigma=0/01$	۱۰۰	۹۹	۲۸	۱۸	۱۸
	$\sigma=0/1$	۹۹	۹۹	۲۹	۱۸	۱۸
$a=0$	$\sigma=0/01$	۱۰۰	۹۹	۳۷	۳۶	۳۵
	$\sigma=0/1$	۱۰۱	۹۹	۴۰	۳۸	۳۷

کارایی نسبی: نسبت بین متوسط واریانس نمونه‌گیری πps سیستماتیک (یا نمونه‌گیری SPS) و متوسط نظری واریانس V_{asy}.

CV واریانس نمونه‌گیری: ضریب تغییرات واریانس‌های نمونه‌گیری.

شبیه‌سازی‌ها برای ۱۰۰۰ θی تولید شده به صورت مستقل تکرار می‌شوند. نتایج در جدول ۱ ارائه شده‌اند که در آن کارایی نسبی (RE) برابر است با نسبت $\frac{E(V_\delta)}{E(V_{asy})}$. موارد زیر را ملاحظه می‌کنیم:

۱- مشاهده می‌شود که نمونه‌گیری سیستماتیک πps و SPS هر دو کارایی نسبی حول ۱۰۰ درصد به دست می‌دهند، به گونه‌ای که این دو روش نمونه‌گیری از لحاظ اصل مخاطره‌ی بیزی معادل‌اند.

۲- تحت مدل (۵) یعنی $a=1$، CV نمونه‌گیری سیستماتیک πps می‌تواند از (۶) مشتق شود. از نرمال بودن ε_iها داریم $\gamma_2=0$ و از Uی داده شده داریم

$\sum_{m\neq p} c_{mp}^{۲} = ۲۰۷۹۵۰۰$، به طوری که $\sqrt{\lambda_n(\gamma_۲)} = ۰/۴۱۰$ که به $\sigma^۲$ بستگی ندارد. CV نظری توسط شبیه‌سازی‌ها تأیید می‌شود. مشاهده‌می‌شود که نمونه‌گیری سیستماتیک πps دارای مخاطره‌ی بیزی مرتبه‌ی دوم به مراتب بیش‌تری نسبت به نمونه‌گیری تصادفی πps است.

۳- برای aی داده شده مخاطره‌ی بیزی مرتبه‌ی دوم روی σ تغییرات زیادی ندارد. برای $۰/۲۵ < a < ۱$ مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری تصادفی πps تقریباً ثابت است و به مقدار قابل توجهی از مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری سیستماتیک πps کمتر است. مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری تصادفی πps با نزدیک شدن a به صفر، به سرعت افزایش می‌یابد، اما کوچک‌تر از مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری سیستماتیک πps باقی می‌ماند. به طور خلاصه نمونه‌گیری تصادفی πps با توجه به اصل تصمیم بیز استوار تحت مدل (۵) ترجیح داده می‌شود و این انتخاب نسبت به انحراف از فرض $a = ۰$، که صورتی از تشخیص نادرست $\pi(\theta)$ است، استوار است.

حال نمونه‌گیری سیستماتیک با احتمالات برابر را در نظر بگیرید. یک نتیجه‌ی کلی وجود دارد که اظهار می‌دارد زمانی که واریانس داخل نمونه‌ای از واریانس جامعه بیش‌تر است، نمونه‌گیری سیستماتیک از SRS کاراتر است و این به خاطر تجزیه‌ی واریانس جامعه به صورت زیر است

$$\sum_{i\in U}(y_i - \bar{Y})^۲ = \sum_{m=۱}^{k}\sum_{i\in s_m}(y_i - \bar{y}_m)^۲ + \sum_{m=۱}^{k} n(\bar{y}_m - \bar{Y})^۲.$$

یعنی تجزیه‌ی واریانس جامعه به تغییرات داخل k نمونه‌ی سیستماتیک و تغییرات بین این نمونه‌ها. از آن‌جا که V_{sys} با مؤلفه‌ی دوم متناسب است به ازای یک θی داده شده زمانی مینیمم می‌شود که مؤلفه‌ی اول ماکسیمم شود. بر اساس ترتیب متناظر واحدها، نمونه‌گیری سیستماتیک می‌تواند به صورت بالقوه به حصول کارایی نسبت به نمونه‌گیری تصادفی ساده منجر شود. برای مثال حالت غایی را تحت مدل (۷) با $\sigma = ۰$ در نظر بگیرید یعنی $y_i = x_i$. ترتیب بهینه برای یک نمونه‌ی سیستماتیک شامل ۱۰ واحد، ایجاد

تناوب بین ترتیب کاهشی و افزایشی یکی از هر ۱۰ واحد از جامعه است ([۱۰] مثال ۳-۴-۲)، که توسط $U_{opt}=(۱,\ldots,۱۰,۲۰,\ldots,۱۱,۲۱,\ldots,۳۰,۴۰,\ldots,۳۱,\ldots,۱۰۰,\ldots,۹۱)$ نشان داده می‌شود و در این حالت $\mathrm{V}_{sys}(\overline{y}_s)=۰$.

البته در عمل هرگز y_iها دقیقاً معلوم نیستند. ولی، حال که $\overline{x}_m=\frac{X}{N}$ ثابتی از نمونه‌گیری است، ترتیب U_{opt} تحت مدل (۷) با $a=۰$ بهینه باقی می‌ماند. براوردگر مبتنی بر یک نمونه‌ی سیستماتیک انتخاب شده از U_{opt} با احتمالات برابر عبارت است از

$$\hat{Y}_m = X + N\overline{\varepsilon}_m$$

که در آن $X=\sum_{i=۱}^{N}x_i$ و $\overline{\varepsilon}_m=\frac{۱}{n}\sum_{i\in s_m}(y_i-x_i)$. این براوردگر همان براوردگر تفاضل ([۱۰]، فصل ۳-۶) بر اساس یک نمونه‌ی تصادفی ساده است. به عبارت دیگر کارایی نمونه‌گیری سیستماتیک بر اساس U_{opt} می‌تواند با استفاده‌ی توأم نمونه‌گیری تصادفی ساده و براوردگر تفاضل نیز حاصل شود. البته مخاطره‌ی بیزی مرتبه‌ی دوم برای استراتژی دوم تنها از مرتبه‌ی $O\left(\frac{۱}{\sqrt{N}}\right)$ است. این وضعیت در جدول (۲) نشان داده شده است که در آن RE برابر $\frac{\mathrm{E}(\mathrm{V}_{sys})}{\mathrm{E}(\mathrm{V}_{srs})}$ است و CV_δ، CV واریانس نمونه‌ای واقعی تولید شده توسط $\delta=SYS,SRS$ است. موارد زیر را ملاحظه می‌کنیم:

۱- همان‌طور که انتظار می‌رفت استفاده‌ی توأم از نمونه‌گیری تصادفی ساده و براوردگر تفاضل به اندازه‌ی نمونه‌گیری سیستماتیک بهینه تحت مدل (۷) با $a=۰$ کارامد است. کارایی نمونه‌گیری سیستماتیک با انحراف از فرض $a=۰$، یعنی حرکت a از ۰ به سمت ۱، به آهستگی کم می‌شود.

۲- مخاطره‌ی بیزی مرتبه دوم نمونه‌گیری سیستماتیک به‌مراتب بیش‌تر از این مخاطره برای نمونه‌گیری تصادفی ساده تحت فرض $a=۰$ است. برای هر دو روش تغییرات برای $۰\le a\le ۱$ خیلی کم است. به طور خلاصه استفاده‌ی مرکب نمونه‌گیری تصادفی ساده و براوردگر تفاضل بر اساس اصل تصمیم بیز استوار تحت مدل (۷) با $a=۰$ بر نمونه‌گیری سیستماتیک بهینه ترجیح داده می‌شود و این انتخاب نسبت به انحراف از این فرض استوار است.

جدول ۲- نتایج شبیه‌سازی برای نمونه‌گیری با احتمالات برابر بر حسب درصد: نمونه‌گیری سیستماتیک بر اساس U_{opt} در مقابل استفاده‌ی توأم از نمونه‌گیری تصادفی ساده و براوردگر تفاضل

	$a=0$			$a=0/5$			$a=1$		
	RE	CV_{sys}	CV_{srs}	RE	CV_{sys}	CV_{srs}	RE	CV_{sys}	CV_{srs}
$\sigma=0/01$	۱۰۴	۴۷	۱۶	۱۰۷	۴۸	۱۸	۱۱۲	۴۸	۱۹
$\sigma=0/1$	۹۹	۴۹	۱۶	۱۰۷	۴۷	۱۷	۱۱۲	۴۷	۲۰

RE: نسبت بین متوسط واریانس نمونه‌گیری سیستماتیک و نمونه‌گیری تصادفی ساده.
CV: ضریب تغییرات واریانس‌های نمونه‌گیری.

۵- نمونه‌گیری سیستماتیک برای چندین موقعیت

یک نمونه‌ی سیستماتیک، پس از این که انتخاب شد، ممکن است برای چندین موقعیت متعاقب مورد استفاده قرار بگیرد. چنین نمونه‌ای ممکن است تشکیل یک گروه در یک طرح پانلی چرخشی بدهد، مانند طرح LFS (آمارگیری نیروی کار) در اغلب کشورها. همچنین می‌تواند هسته‌ی اصلی یک آمارگیری پانلی باشد، با واحدهای تکمیلی که گاه‌گاه برای به حساب آوردن بازسازی طبیعی جامعه به نمونه اضافه می‌شوند. برای ساده کردن بحث در این‌جا فرض می‌کنیم که یک نمونه‌ی سیستماتیک در موقعیت اول انتخاب شده و پیش از کنار گذاشته شدن برای تمامی موقعیت‌های بعدی مورد استفاده قرار گرفته است، و جامعه‌ی U در طول این دوره یکسان باقی مانده است.

در این صورت نتایج (۴) و (۶) مستقیماً برای کل دوره‌ی فعال پانل صدق می‌کنند. به صورت واضح‌تر، فرض کنید $\boldsymbol{y}_i=(y_{i1},\ldots,y_{it},\ldots,y_{iT})^T$ متغیر مورد بررسی متناظر با $i\in U$ باشد. نتایج (۴) و (۶) مستقیماً برای هر تابعی از $\boldsymbol{y}_i$ صدق می‌کنند. برای مثال، فرض کنید $\boldsymbol{y}_i$ از ۴ معیار وضعیت اشتغال در هر یک از ۴ فصل سال تقویمی در آمارگیری نیروی کار تشکیل شده باشد. اشتغال متوسط سالیانه از میانگین y_{i1} تا y_{i4} حاصل می‌شود. با انتخاب یک نمونه‌ی سیستماتیک در موقعیت اول، خطر نوسان واریانس براوردگر نرخ اشتغال متوسط سالیانه و همچنین خطر نوسان واریانس در هر فصل را می‌پذیریم. علاوه بر این یکی از کاربردهای مهم داده‌های پانلی، براورد تغییرات در جامعه

است. فرض کنید $z_i = (z_{i۲},\ldots,z_{iT})^T$، که در آن برای $t = ۲,\ldots,T$، $z_{it} = y_{it} - y_{i,t-۱}$ تغییرات دوره به دوره است. باز هم نتایج (۴) و (۶) مستقیماً برای هر z_{it} صدق می‌کنند، به نحوی که براورد تغییرات ممکن است به خاطر نمونه‌گیری سیستماتیک مخاطره‌ی بیزی مرتبه‌ی دوم زیادی داشته باشد.

ملاحظات بالا خودهمبستگی قوی محتمل داخل y_i را که اغلب در جامعه‌های طبیعی دیده می‌شود، در نظر نمی‌گیرند. به علاوه؟، یک آزمون شرطی نیز مورد نیاز است. ساده‌ترین حالت را که در آن $T = ۲$ و y_{it} یک متغیر رسته‌ای مانند وضعیت اشتغال است، در نظر بگیرید. به عنوان یک مدل ساده از وابستگی بین $y_{i۱}$ و $y_{i۲}$، احتمالات تغییر وضعیت مارکوف p_{ab} را برای $y_{i۲} = b$ به شرط $y_{i۱} = a$، مستقل برای $i \neq i \in U$ در نظر می‌گیریم. به شرط $y_{i۱} = a$ این به یک مدل جامعه‌ی همگن (۳) منتهی می‌شود. میانگین نمونه‌ی سیستماتیک $z_i = y_{i۲} - y_{i۱}$ به صورت زیر داده می‌شود

$$\overline{z}_m = \sum_{a;y_{i۱}=a} \frac{n_a}{n} \overline{z}_{m,a}$$

که در آن n_a تعداد واحدهای با $y_{i۱} = a$ است و $\overline{z}_{m,a}$ میانگین تغییرات میان آن‌ها است. صورت بسته‌ی واریانس شرطی $\mathrm{V}_{sys}\left(\overline{z}_m \mid \{y_{i۱}; i \in U\}\right)$ به طور کلی پیچیده است.

در عوض هر ترتیبی را که در آن واحدها با توجه به مقدار $y_{i۱}$ تقسیم‌بندی شده‌اند در نظر بگیرید. فرض کنید $\frac{N_a}{k}$ به طور طبیعی و به ازای کلیه‌ی aها یک مقدار صحیح است که در آن N_a تعداد واحدهای با $y_{i۱} = a$ در جامعه است. در این صورت n_a و $\overline{y}_{m,t=۱}$ هر دو ثابت‌های نمونه‌گیری هستند، به نحوی که واریانس $\overline{z}_m$ به سادگی، واریانس $\overline{y}_{m,t=۲}$ است. حال نتیجه‌ی (۴) می‌تواند برای $\overline{y}_{m,a,t=۲}$ صدق کند، یعنی داخل هر بخش از $y_{i۱}$ تحت مدل تغییر وضعیت مارکوف به نحوی که مخاطره‌های بیزی مرتبه‌ی دوم $\overline{y}_{m,a,t=۲}$ به شرط $\{y_{i۱}; i \in U\}$ مستقیماً به سمت $\overline{z}_m$ می‌رود. در نظر گرفتن این حالت خاص نشان می‌دهد که مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری سیستماتیک می‌تواند برای براوردگرهای تغییرات در جامعه‌های خودهمبسته و همچنین زمانی که واریانس به صورت شرطی ارزیابی می‌شود زیاد باشد. این موضوع را به وسیله‌ی یک مطالعه‌ی شبیه‌سازی در بخش بعد آزمون می‌کنیم.

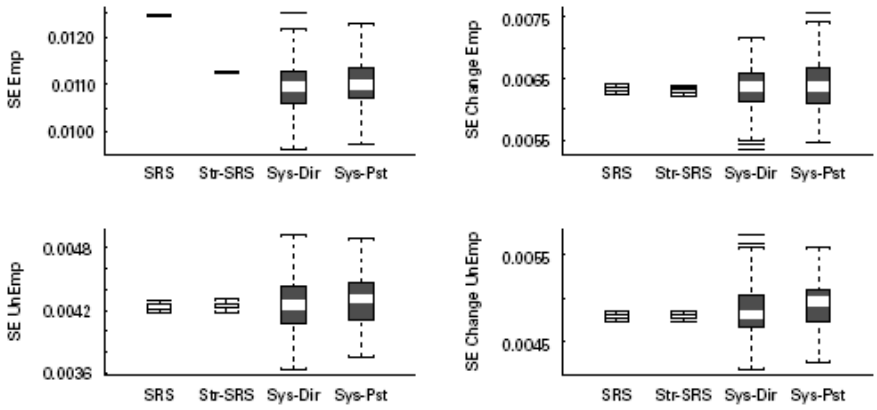
۶- شبیه‌سازی: پویایی بازار کار

در این بخش پویایی بازار کار را با استفاده از سرشماری سال ۲۰۰۱ نروژ و آمارگیری نیروی کار نروژ شبیه‌سازی می‌کنیم. از سرشماری سال ۲۰۰۱ وضعیت اشتغال را به دست می‌آوریم که به سه دسته‌ی « شاغل »، « بیکار » و « غیر فعال » طبقه‌بندی شده است و متغیر مورد نظر جامعه در $t=۱$ در نظر گرفته می‌شود. سپس از آمارگیری نیروی کار در فصل آخر سال ۲۰۰۴ و فصل اول سال ۲۰۰۵ یک ماتریس تغییر وضعیت ۳×۳ برای وضعیت اشتغال بین دو فصل به دست می‌آوریم. با استفاده از این احتمالات تغییر وضعیت مارکوف قادریم یک وضعیت اشتغال در جامعه را در $t=۲$ شبیه‌سازی کنیم. جامعه در داخل هر یک از ۱۹ شهرستان نروژ به تفکیک شهر، سن، جنس و شماره‌ی شناسایی فردی (PIN) طبقه‌بندی شده است که PIN می‌تواند با وضعیت اشتغال مورد نظر ناهمبسته در نظر گرفته شود.

چهار راهبرد متفاوت را در نظر می‌گیریم: ۱- نمونه‌گیری سیستماتیک با احتمالات برابر در $t=۱$ و براورد بر اساس وزن‌دهی مستقیم که با $Sys-Dir$ نشان داده می‌شود، ۲- نمونه‌گیری تصادفی ساده در $t=۱$ و براورد بر اساس وزن‌دهی مستقیم که با SRS نشان داده می‌شود، ۳- نمونه‌گیری تصادفی طبقه‌بندی‌شده با تخصیص متناسب نسبت به جنس و سن (کلاً ۲۲ گروه) که به براورد طبقه‌بندی‌شده منتهی و با $Str-SRS$ نمایش داده می‌شود، و ۴- نمونه‌گیری سیستماتیک با احتمالات برابر و براورد پس‌طبقه‌بندی با ۲۲ گروه سنی-جنسیتی به‌عنوان پس‌طبقه، که با $Sys-Pst$ نمایش داده می‌شود.

شبیه‌سازی‌ها برای هر یک از ۱۹ شهرستان نروژ به صورت جداگانه انجام می‌شود تا بازتابی از طرح طبقه‌بندی شده‌ی آمارگیری نیروی کار نروژ باشد. یک نمونه‌ی انتخاب شده در $t=۱$، در $t=۲$ نیز مورد استفاده قرار می‌گیرد و اندازه‌ی نمونه‌های داخل شهرستان از آمارگیری نیروی کار نروژ گرفته شده‌اند. نتایج برای همه‌ی شهرستان‌ها خیلی شبیه است. در این‌جا ما وضعیت را تنها برای Østfold در شکل ۱ نشان می‌دهیم. در این حالت نمونه‌گیری سیستماتیک می‌تواند به عنوان یک طبقه‌بندی تلویحی نسبت به شهر، سن و جنسیت در نظر گرفته شود. اثرات طبقه‌بندی فقط برای نرخ اشتغال در $t=۲$ قابل توجه‌اند (Emp)، که در حدود ۲۰ درصد کاهش واریانس را در مقایسه با SRS

نشان می‌دهد. هرچند، بیش‌تر اثر طبقه‌بندی می‌تواند با طبقه‌بندی تنها بر اساس جنسیت و سن به دست آید. توجه کنید که طبقه‌بندی بر اساس شهر علاوه بر جنسیت و سن، به خاطر تعداد زیاد طبقات غیرکاربردی است. در همه‌ی حالت‌های دیگر نمی‌توان انتظار بهبود کارایی را با استفاده از نمونه‌گیری سیستماتیک داشت.



شکل ۱- نمودار جعبه‌ای انحراف معیار (SE) نرخ اشتغال در $t = ۲$ (Emp)، تغییر نرخ اشتغال (Change Emp)، نرخ بیکاری در $t = ۲$ (UnEmp) و تغییر نرخ بیکاری (Change UnEmp) برای شهرستان Østfold: نمونه‌گیری تصادفی ساده با وزن‌دهی مستقیم (SRS)، نمونه‌گیری تصادفی طبقه‌بندی‌شده با تخصیص متناسب (Str-SRS)، نمونه‌گیری سیستماتیک (Sys-Dir) و نمونه‌گیری سیستماتیک با براورد پس‌طبقه‌بندی (Sys-Pst)

مشاهده می‌شود که در حالی که مخاطره‌های بیزی مرتبه‌ی دوم SRS و $Str - SRS$ برای جامعه‌ای به این اندازه (در حدود صد و هفتاد و نه هزار نفر) ناچیز است، این مخاطره‌ها تحت نمونه‌گیری سیستماتیک همچنین زمانی که واریانس‌ها مانند این‌جا به صورت شرطی محاسبه می‌شوند، قابل ملاحظه‌اند. ضریب تغییرات V_{sys} برای Emp ۱۱ درصد، برای تغییر Emp ۱۵/۸ درصد، برای UnEmp ۱۶/۴ درصد و برای تغییر UnEmp ۱۵/۴ درصد است. این مقادیر با مقدار غیر شرطی نظری که توسط (۴) داده می‌شوند و برای Østfold تقریباً برابر است با $\sqrt{2f} = \sqrt{\frac{۲}{۱۳۴/۶}} = ٪۱۲/۲$، قابل مقایسه‌اند. لذا در موقعیت‌های خاص، نوسان واریانس ممکن است اثر مورد انتظار

طبقه‌بندی روی براورد Emp را از بین ببرد. توجه کنید که مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری سیستماتیک نمی‌تواند به وسیله‌ی پس‌طبقه‌بندی کاهش داده شود.

به خصوص برای براورد تغییرات که در این‌جا نگرانی اصلی ما است ضریب تغییرات واریانس نمونه‌گیری سیستماتیک تقریباً برابر حالت براورد سطح است. لذا استفاده از نمونه‌گیری سیستماتیک می‌تواند باعث شود که واریانس نمونه‌گیری واقعی یک براوردگر تغییرات، به شدت در طول زمان تغییر کند. برای مثال، اگر واریانس واقعی ۱۵ درصد بالاتر از امید ریاضی بین فصل اول و دوم و ۱۵ درصد پایین‌تر از امید ریاضی بین فصل دوم و سوم باشد، آن‌گاه دو براورد تغییرات دارای ۳۰ درصد اختلاف در واریانس‌های نمونه‌گیری هستند که از استفاده‌ی تنها از نمونه‌گیری سیستماتیک ناشی شده است. حال که در این‌جا ضریب تغییرات واریانس هر دوی براوردگرهای تغییرات در حدود ۱۵ درصد است، این یک سناریوی غیر معمول نیست. در حالت افراطی‌تری که دو انحراف معیار بالا یا پایین‌تر از واریانس نمونه‌گیری ورد انتظارند، واریانس واقعی یک براوردگر تغییرات تقریباً دو برابر (یعنی $\frac{۱/۳}{۰/۷}$) دیگری است. قطعاً نگه‌داشتن این مقدار به عنوان یک ویژگی طرح نمونه‌گیری نامطلوب است.

۷- خلاصه

در قسمت‌های قبل مفهوم مخاطره‌ی بیزی مرتبه‌ی دوم و اصل تصمیم استوار بیز را معرفی کردیم. تعدادی از وضعیت‌هایی را در نظر گرفتیم که در آن‌ها نمونه‌گیری سیستماتیک به طور معمول جایگزین روش‌های نمونه‌گیری تصادفی دیگری می‌شود که به همان اندازه کارا هستند. نشان داده شده است که این کار می‌تواند مخاطره‌های بیزی مرتبه‌ی دوم بزرگ، یعنی نوسانات واریانس نمونه‌گیری واقعی در هر دوی آمارگیری نمونه‌ای مقطعی و طولی ایجاد کند. این می‌تواند برای نمونه‌های بزرگ گرفته شده از جامعه‌های کوچک، یا زیرنمونه‌های بزرگ از زیرجامعه‌های کوچک، بسیار زیان‌آور باشد. استفاده ازنمونه‌گیری سیستماتیک برای راحتی در چنین موقعیت‌هایی بر پایه‌ی شانس و بدون هیچ‌گونه توقعی برای بهبود کارایی است.

نمونه‌گیری سیستماتیک به فراوانی درموقعیت‌هایی که ما مورد بررسی قرار نداده‌ایم نیز به کار می‌رود. کاکران [۲] مثال‌های متعددی، شامل استفاده‌ی معمول از نمونه‌گیری سیستماتیک یک یا دوبعدی درآمارگیری‌های جنگل‌داری و اراضی ذکر کرده است. چنین موقعیت‌هایی می‌توانند مشابه آنچه در این‌جا انجام شده است مورد مطالعه قرار بگیرند، اما نیازمند مدل‌های جامعه‌ی نسبتاً خاص شامل همبستگی‌هایی در طول زمان ومکان‌اند که خارج از چارچوب این مقاله‌اند. به علاوه، همان‌طور که قبلاً اشاره شد، جمع‌آوری یک نمونه‌ی سیستماتیک ممکن است در چنین موقعیت‌هایی هنوز هم آسان‌تر باشد. سرانجام، تعادل در نمونه‌گیری سیستماتیک می‌تواند در آمارگیری‌های اراضی و بوم‌شناسی نقش برجسته‌ای ایفا کند. برای مثال گستردگی یکنواخت نواحی نمونه‌گیری برای هموارسازی فضایی داده‌ها ممکن است مهم‌تر از تصادفی بودن نمونه تلقی شود.

ما نمونه‌گیری سیستماتیک را از نقطه نظر یک تصمیم آماری مورد مطالعه قرار داده‌ایم که در آن تابع زیان به صورت واریانس نمونه‌گیری براوردگر آمارگیری تعریف می‌شود. توجه کنید که برای استنباط مدل–مبنا که در آن واریانس یک براوردگر تنها تحت مدل جامعه ارزیابی می‌شود مخاطره‌ی بیزی مرتبه‌ی دوم نمونه‌گیری سیستماتیک با مخاطره‌ی بیزی مرتبه‌ی دوم یک روش نمونه‌گیری تصادفی دیگر تفاوتی ندارد به شرط آن‌که نمونه‌گیری در هر دو حالت فاقد اطلاعات مفید باشد. در واقع نمونه‌گیری سیستماتیک گاهی به‌عنوان گام اول در ساختن نمونه‌های متعادل مختلف مفید در نظر گرفته می‌شود [۱۱].

مرجع‌ها

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.). Springer. NY: New York.

[2] Cochran, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley and Sons.

[3] French, S. (1986). *Decision Theory: An Introduction to the Mathematics of Rationality*. Chichester: Ellis Horwood Ltd.

[4] Hartley, H.O. and Rao, J.N.K. (1962). Sampling with Unequal Probabilities and Without Replacement. *Annals of Mathematical Statistics*, **33**, 350-374.

[5] Madow, W.G. (1949). On the Theory of Systematic Sampling, II. *Annals of Mathematical Statistics*, **20**, 333-354.

[6] Madow, W.G. (1953). On the Theory of Systematic Sampling, III. *Annals of Mathematical Statistics*, **24**, 101-106.

[7] Madow, W.G. and Madow, L.H. (1944). On the Theory of Systematic Sampling. *Annals of Mathematical Statistics*, **15**, 1-24.

[8] Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, **14**, 149-162.

[9] Raj, D. (1965). Variance Estimation in Randomized Systematic Sampling with Probability Proportionate to Size. *Journal of the American Statistical Association*, **60**, 278-284.

[10] Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

[11] Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: Wiley.

[12] Wetherill, B.G. (1981). *Intermediate Statistical Methods*: Chapman and Hall: London.

[13] Zigmond, A.S.; Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, **67**, 361–370.

شیرین گلچی
فوق لیسانس آمار
تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان شهید فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.
پیام‌نگار: golchishirin@yahoo.com