گزیده‌مطالب آماری
سال ۱۹، شماره‌ی ۲، پاییز و زمستان ۱۳۸۷، صص ۱۸۴–۱۶۹

نمونه‌ی تصادفی متعادل‌شده و نحوه‌ی انتخاب آن

فاطمه هرندی، * فرهاد مهران

پژوهشکده‌ی آمار

چکیده. هدف نمونه‌گیری، نتیجه‌گیری درباره‌ی کل جامعه بر اساس اطلاعات بخشی از آن است. به‌طور کلی از دو دیدگاه می‌توان مطلوبیت نمونه را ارزیابی کرد. دیدگاه اول به نمایانگر بودن نمونه اشاره می‌کند، به این معنی که تا چه حد ساختار اساسی جامعه را در بر دارد، و دیدگاه دوم به تصادفی بودن نمونه با این مفهوم که آیا واحدهای انتخاب شده در نمونه و همچنین واحدهای انتخاب نشده همگی احتمال انتخاب مثبت و مشخص داشته‌اند، توجه می‌نماید. در این مقاله ابتدا نمونه‌گیری متعادل‌شده (Balanced Sampling) به‌عنوان یک روش نمونه‌گیری احتمالی که انتخاب نمونه‌ی نمایانگر از جامعه‌ی مورد بررسی را امکان‌پذیر می‌سازد، تشریح می‌شود. سپس نحوه‌ی انتخاب نمونه‌ی متعادل‌شده به‌روش مکعبی (Cube Method) بیان می‌شود. سرانجام با انجام یک شبیه‌سازی با استفاده از یک مجموعه داده‌های واقعی، کارایی این روش با روش نمونه‌گیری PPS مقایسه می‌شود. شبیه‌سازی برای سه گروه مختلف از متغیرهای مبنای تعادل اجرا می‌شود. نتایج شبیه‌سازی حاکی از آن است که استفاده از نمونه‌گیری متعادل‌شده موجب بهبود کارایی براوردها می‌شود و با افزودن متغیرهای تعادل مناسب، میزان افزایش کارایی محسوس‌تر می‌شود.

۱– مقدمه

بیش از یک قرن از اولین باری که پیشنهاد استفاده از روش نمونه‌گیری از جوامع محدود به جای سرشماری مطرح شده است می‌گذرد [۸]. امروزه آمارگیری نمونه‌ای در سطحی

واژگان کلیدی: نمونه‌گیری احتمالی؛ نمونه‌ی نمایانگر؛ نمونه‌گیری متعادل‌شده؛ روش مکعبی؛ شبیه‌سازی؛ کارایی.

دریافت: ۱۳۸۷/۱۲/۳، پذیرش: ۱۳۸۸/۳/۳

* نویسنده‌ی عهده‌دار مکاتبات

گسترده توسط سازمان‌های دولتی، محققان بازار، طراحان نظرسنجی، محققان اجتماعی و سایر افراد استفاده می‌شود که ناشی از به وجود آمدن روش‌های نمونه‌گیری کارا برای جمع‌آوری مقرون به صرفه‌ی داده‌ها طی این سال‌ها بوده است.

بررسی تاریخچه آمارگیری‌های نمونه‌ای نشان می‌دهد که در سال‌های اولیه تنها نمایانگر بودن نمونه حائز اهمیت بوده و ارزش هر نمونه‌گیری با این معیار که نمونه‌ی حاصل از آن تا چه حد نمایانگر جامعه مورد بررسی است، سنجیده می‌شده است. این دیدگاه به روش نمونه‌گیری توجه نداشته و عمدتاً مبتنی بر نمونه‌های قصدی بوده است [۷]. با گذشت زمان و آشکار شدن ضعف‌های نمونه‌گیری‌های غیرتصادفی [۹]، آنچه که در درجه‌ی اول اهمیت قرار گرفت، تصادفی بودن روش نمونه‌گیری بود به نحوی که امروزه تصادفی بودن انتخاب نمونه، اصل لازم هر آمارگیری از جامعه‌ی محدود، به ویژه در تولید آمار رسمی، تلقی می‌شود؛ هر چند به تنهایی برای اعتبار نتایج آمارگیری کافی نیست. برای معتبر نمودن نتایج آمارگیری مبتنی بر نمونه‌های تصادفی لازم است از تمامی اطلاعات موجود در مورد جامعه‌ی مورد بررسی قبل و بعد از نمونه‌گیری استفاده شود تا از نمایانگر بودن نمونه‌ی انتخابی اطمینان حاصل شود. در واقع تنها با تلفیق این دو ویژگی در یک نمونه‌گیری، می‌توان به براوردهایی با دقت بالا از پارامترهای مورد نظر در جامعه‌ی هدف دست یافت.

نمونه‌گیری متعادل‌شده به‌عنوان یکی از روش‌های تلفیق این دو ویژگی، بحثی است که از دیرباز و حتی قبل از طرح نظریه‌ی نمونه‌گیری احتمالی توسط نیمن [۹] مطرح بوده است. اما از آن‌جا که روش‌های کارا برای انتخاب نمونه‌ی متعادل‌شده‌ای که در عین حال نمونه‌ای تصادفی هم باشد، وجود نداشت و روش‌های موجود نیز یا محدود به شرایط خاص بوده و یا امکان عملیاتی شدن نداشته‌اند ([۹]؛ [۱۴]؛ [۱۰]؛ [۳]؛ [۲]؛ [۴]؛ [۶]؛ [۱۳])، استفاده از این شیوه‌ی نمونه‌گیری در آمارگیری‌های ملی وارد نشده بود.

در سال ۲۰۰۴، دویل و تیه روشی به نام روش مکعبی را برای انتخاب نمونه‌ی متعادل‌شده پیشنهاد کردند [۵] که امکان انتخاب نمونه‌ی تصادفی با احتمال‌های انتخاب برابر یا نابرابر را که در عین حال بر اساس متغیرهای متعددی نمایانگر نیز باشد، فراهم

می‌آورد. این روش هم‌اکنون توسط مراکز آمار فرانسه و سوئیس برای آمارگیری‌های ملی مورد استفاده قرار می‌گیرد [۱۲].

در بخش دوم این مقاله ضمن مروری کوتاه بر مفاهیم پایه‌ای در نمونه‌گیری احتمالی، مفهوم نمونه‌ی متعادل‌شده و نحوه‌ی انتخاب آن به‌روش مکعبی بیان می‌شود. در بخش سوم مقاله چگونگی شبیه‌سازی انجام شده و نتایج حاصل از آن ارائه می‌شود. بخش چهارم مقاله به بیان نتایج و پیشنهادات حاصل از پژوهش اختصاص یافته است.

۲- نمونه‌ی متعادل‌شده ونحوه‌ی انتخاب آن

در هر نمونه‌گیری وقتی واژه‌ی نمونه (sample) را استفاده می‌کنیم دو مفهوم متفاوت می‌تواند مد نظر باشد که برای احتراز از اشتباه می‌توان با اضافه نمودن پیشوندهای مناسب به واژه‌ی مذکور این دو مفهوم را از یکدیگر تفکیک نمود [۱]. در این صورت منظور از **کل‌نمونه** ($\mathbf{s}$)، مجموعه واحدهایی است که نمونه را تشکیل می‌دهند و منظور از **تک‌نمونه** (s_k)، kامین واحد نمونه است با تعریف زیر،

$$s_k = \begin{cases} 1, & \text{اگر واحد } k\text{ام در نمونه‌ی } \mathbf{s} \text{ نباشد} \\ 0, & \text{اگر واحد } k\text{ام در نمونه‌ی } \mathbf{s} \text{ باشد} \end{cases}$$

بنا بر این هر کل‌نمونه‌ای را می‌توان به‌صورت برداری از تک‌نمونه‌ها نشان داد.

$$\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)'$$

متناظر با هر یک از این دو مفهوم، یک احتمال انتخاب نیز وجود دارد که به‌صورت زیر تعریف می‌شود.

$$\pi_k = \Pr(s_k = 1) > 0 \quad ; \quad k = 1, \ldots, N \; ;$$

π_k احتمال انتخاب تک‌نمونه‌ی kام است به نحوی که در هر نمونه‌گیری احتمالی، مقادیر π_k باید برای تمامی واحدهای جامعه مورد بررسی بزرگ‌تر از صفر باشد و

$$\mathrm{p}(\mathbf{s}) = \Pr(\mathbf{s} \in \mathbf{S})$$

احتمال انتخاب کل‌نمونه‌ی $\mathbf{s}$ است که در آن $\mathbf{S}$ مجموعه‌ی کل‌نمونه‌های ممکن در جامعه‌ی مورد بررسی است (در نمونه‌گیری بدون جایگذاری و با اندازه‌ی نمونه‌ی غیر ثابت برابر است با 2^{N}) و مقادیر $p(\mathbf{s})$ در رابطه‌ی زیر صدق می‌کنند.

$$\sum_{\mathbf{s}\in\mathbf{S}} p(\mathbf{s}) = 1$$
$$\circ \le p(\mathbf{s}) \le 1 \quad ; \quad \mathbf{s}\in\mathbf{S}$$

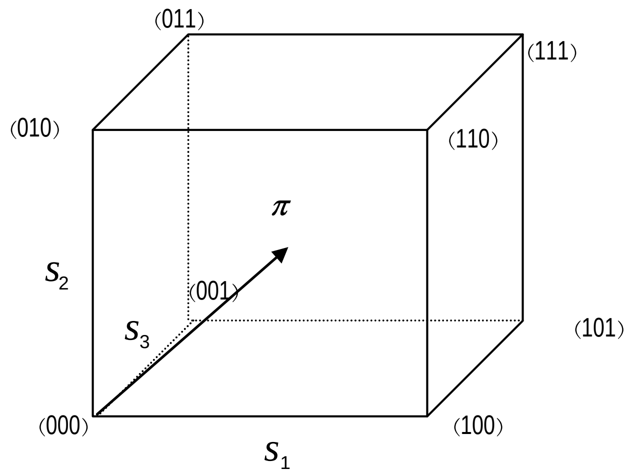
به بیان دیگر بعضی از کل‌نمونه‌ها می‌توانند شانسی برای انتخاب نداشته باشند و این موضوع خللی به احتمالی بودن طرح نمونه‌گیری وارد نمی‌کند.

به‌طور کلی در هر نمونه‌گیری احتمالی، طرح نمونه‌گیری با زوج $(\mathbf{S}, p(\cdot))$ یعنی مجموعه‌ی کل‌نمونه‌های ممکن و احتمال انتخاب هر یک از این کل‌نمونه‌ها تعریف می‌شود. بسته به طرح انتخابی بعضی از کل‌نمونه‌های ممکن می‌توانند شانس انتخاب صفر داشته باشند (برای مثال در نمونه‌گیری سیستماتیک بسیاری از $p(\cdot)$ها صفر هستند زیرا کل‌نمونه‌های شامل واحدهای مجاور شانسی برای انتخاب شدن ندارند). به بیان دیگر احتمال‌های انتخاب کل‌نمونه‌ها از طرح انتخاب شده برای نمونه‌گیری تأثیر می‌پذیرند. هرندی و همکاران [۱] با ذکر یک مثال نشان دادند که با تغییر طرح انتخاب نمونه، علی‌رغم ثابت بودن مقادیر π_k، مقادیر $p(\cdot)$ تغییر می‌کند.

احتمال انتخاب تک‌نمونه‌ی kام را می‌توان از جمع احتمال‌های انتخاب کل‌نمونه‌هایی که واحد kام را در بر دارند نیز به دست آورد، یعنی

$$\pi_k = \sum_{\mathbf{s}\ni \mathbf{k}} p(\mathbf{s}).$$

طبق نظریه‌ی نمونه‌گیری، پس از فهرست کردن تمامی کل‌نمونه‌های ممکن و تعیین احتمال انتخاب هر یک، یکی از آن‌ها به تصادف انتخاب می‌شود. شکل ۱ نمایش هندسی کل‌نمونه‌های ممکن در یک جامعه‌ی فرضی شامل ۳ واحد را نشان می‌دهد.



شکل ۱- نمایش هندسی انتخاب نمونه در جامعه‌ای به اندازه‌ی $N = 3$

در این شکل هر کدام از محورها نماینده‌ی یکی از تک‌نمونه‌ها و رئوس مکعب حاصل، کل‌نمونه‌های ممکن را نشان می‌دهد. π نیز احتمال انتخاب تک‌نمونه‌ها را نشان می‌دهد.

در حالت کلی در هر نمونه‌گیری احتمالی، فضای نمونه‌گیری را می‌توان به‌صورت یک ابَر مکعب N-بعدی در نظر گرفت که هر یک از ابعاد آن، یکی از تک‌نمونه‌ها و هر یک از رئوس آن یکی از کل‌نمونه‌های ممکن را نشان می‌دهد. الگوریتم نمونه‌گیری، رفتن از π به یکی از رئوس به‌صورت تصادفی است که بسته به طرح نمونه‌گیری انتخابی، شانس انتخاب رئوس مختلف می‌تواند برابر، نابرابر و یا حتی در مواردی صفر باشد. بنا بر این احتمال‌های انتخاب کل‌نمونه‌ها، در قالب قیود زیر و بسته به الگوریتم انتخابی برای نمونه‌گیری تعیین می‌شوند.

$$\begin{cases} \sum_{\mathbf{s}\in\mathbf{S}} \mathrm{p}(\mathbf{s}) = 1 \\ 0 \le \mathrm{p}(\mathbf{s}) \le 1 \quad ; \quad \mathbf{s}\in\mathbf{S} \\ \pi_k = \sum_{\mathbf{s}\ni k} \mathrm{p}(\mathbf{s}) \end{cases} \tag{۱}$$

در یک طرح نمونه‌گیری تصادفی متعادل‌شده، تنها به کل‌نمونه‌هایی شانس انتخاب داده می‌شود که علاوه بر مقید بودن به قیود مذکور بتوانند براوردهایی دقیق (برابر با مقدار واقعی معلوم) از متغیرهای کمکی همبسته با متغیرهای مورد بررسی تولید نمایند. به بیان دیگر در نمونه‌گیری متعادل‌شده

$$Q=\left\{\mathbf{s}\in\mathbf{S}\ \middle|\ \hat{t}_{\mathbf{x}\boldsymbol{\pi}}=t_{\mathbf{x}}\right\}$$

مجموعه‌ی کل‌نمونه‌های ممکن با احتمال غیر صفر است. در رابطه‌ی فوق، $\mathbf{x}$ بردار p متغیر کمکی مربوط به واحد kام، $\hat{t}_{x\pi}=\sum_{k\in U}\frac{x_k s_k}{\pi_k}$ براورد هورویتز تامپسون $\mathbf{x}$ و $t_x=\sum_{k\in U}\mathbf{x}_k$ مجموع معلوم متغیر کمکی $\mathbf{x}$ است. بنا بر این

$$\operatorname{var}\left(\hat{t}_{x_\pi}\right)=0.$$

ایده‌ی اصلی در نمونه‌گیری متعادل‌شده این است که با توجه به توانایی طرح نمونه‌گیری متعادل‌شده در ارائه‌ی براورد دقیق متغیرهای کمکی، انتظار می‌رود واریانس نمونه‌گیری سایر براوردهای مورد نظر در آمارگیری نیز بسته به میزان همبستگی آن‌ها با متغیرهای کمکی، کم و بیش کاهش یابد. به سادگی می‌توان نشان داد که برخی طرح‌های متداول نمونه‌گیری، طرح‌هایی متعادل‌شده روی متغیرهای کمکی خاص می‌باشند [۱۲].

برای انتخاب نمونه‌ی متعادل‌شده، طبق نظریه‌ی نمونه‌گیری ابتدا باید تمامی کل‌نمونه‌های ممکن فهرست شوند سپس برای این که ارزش هر کل‌نمونه از نظر میزانی که شرط تعادل را محقق می‌کند مشخص شود به هر کل‌نمونه، هزینه‌ای منتسب شود به این ترتیب که اگر کل‌نمونه‌ی $\mathbf{s}$ متعادل باشد، هزینه‌ی صفر و در غیر این صورت متناسب با میزان دور بودن آن از تحقق شرط تعادل، هزینه‌ی مثبتی به آن منتسب شود.

به این منظور لازم است یک تابع هزینه تعریف شود [۱۲]. در مرحله‌ی بعد باید مقادیر احتمال‌های انتخاب کل‌نمونه‌ها به نحوی تعیین شوند که کل‌نمونه‌های دارای کمترین هزینه، بیش‌ترین شانس انتخاب را داشته باشند و برعکس. پس کافی است امید ریاضی تابع هزینه‌ی زیر نسبت به $p(\cdot)$ها مینیمم شود (با لحاظ قیود رابطه‌ی (۱) و با فرض معلوم بودن مقادیر π_k).

$$\sum_{\mathbf{s}\in\mathbf{S}} \mathrm{p}(\mathbf{s})\mathrm{Cost}(\mathbf{s})$$

در رابطه‌ی فوق $\mathrm{Cost}(\mathbf{s})$ هزینه‌ی انتخاب کل‌نمونه‌ی $\mathbf{s}$ است.

پس از حل معادله‌ی بالا و به دست آوردن مقادیر $\mathrm{p}(\cdot)$ها، با توجه بـه احتمـال‌هـای انتخاب تعیین شده، یکی از کل‌نمونه‌ها به تصادف انتخاب می‌شود. بـرای دیـدن مثـالی از کاربرد روش بالا برای انتخاب نمونه‌ی متعادل‌شده به [۱] رجوع کنید.

استفاده از روش بالا برای انتخاب نمونـه‌ی متعـادل‌شـده کـه تیـه [۱۲] آن را روش کـل شماری نامیده، علی‌رغم سادگی در جوامع واقعی که اغلب بزرگ هستند، عملی نیست زیرا تعداد کل‌نمونه‌های ممکن بسیار زیاد است (برای مثال در نمونه‌گیری بدون جایگـذاری بـا $N = 20$، تعداد کل‌نمونه‌های ممکن ۱۰۴۸۵۷۶ است). این محدودیت خاص نمونه‌گیـری متعادل‌شده نیست زیرا در نمونه‌گیری از جوامع واقعی نیز هیچ‌گاه تمـامی کـل‌نمونـه‌هـای ممکن و احتمال انتخاب آن‌ها را فهرست نمی‌کنیم بلکه در عمل بر اساس مقادیر π_k هـا، واحدهای نمونه را تک‌تک انتخاب می‌کنیم. در نمونه‌گیری متعادل‌شده نیز می‌توان بـدون تعیین مقادیر احتمال انتخاب تک‌تک کل‌نمونه‌ها ($\mathrm{p}(\cdot)$ها)، یکی از آن‌ها را انتخاب کـرد. روش اخیر برای انتخاب نمونه‌ی متعادل‌شده که دویل و تیـه [۵] آن را روش مکعبـی نـام نهاده‌اند بر مبنای نمایش هندسی فضای نمونه‌گیری و مفهوم انتخاب نمونه در ایـن فـضا استوار است.

همان‌گونه که قبلاً اشاره شد، در هر نمونه‌گیری تصادفی بدون جایگذاری، مجموعه‌ی کل‌نمونه‌های ممکن را می‌توان به‌صورت رئوس یک ابَر مکعب N-بعدی در نظر گرفت. بنا بر این در هر طرحی که مقادیر π_k آن مشخص باشد، کافی است روشی تصادفی برای رسیدن از بردار π به یکی از رئوس مکعب N-بعدی تعریف شـود کـه در واقـع همـان الگوریتم نمونه‌گیری است. در روش مکعبی، برای انتخاب نمونـه‌ی متعـادل‌شـده، تحقـق شرط تعادل نیز در الگوریتم نمونه‌گیری لحاظ می‌شود بدون این که به اصل تصادفی بودن انتخاب خدشه‌ای وارد شود. مراحل کار انتخاب نمونه در روش مکعبی به ترتیب زیر است؛

- ابتدا شکل هندسی معادلات تعادل (Balancing Equations، $\hat{\mathrm{t}}_{\mathbf{x}\boldsymbol{\pi}} = \mathrm{t}_{\mathbf{x}}$) در فـضای N-بعـدی تعیـین مـی‌شـود کـه یـک زیرفـضای آفـین

(Affine Subspace) با بعد $N - p$ است و با نماد Q نـشان داده مـی‌شـود. طی مراحل بعدی سعی می‌شود به‌صورت تصادفی یک کل‌نمونه در Q انتخاب شود و اگر هیچ کل‌نمونه‌ای در Q قرار نگیـرد تـلاش مـی‌شـود کـل‌نمونـه‌ای حتی‌الامکان نزدیک به Q به‌صورت تصادفی انتخاب شود.

- در مرحله‌ی دوم، اشتراک زیر فضای Q با مکعب N-بعدی مشخص می‌شود.

$$\mathrm{K} = \left\{ [0,1]^N \cap Q \right\}$$

K یک چند وجهی محدب $(N - p)$-بعدی است. بدیهی اسـت کـه رئـوس K لزوماً منطبق بر رئوس مکعب نیستند. تیه [۱۲] نشان داده است که حـداکثر p (تعداد متغیرهای تعادل) نقطه در هر رأس K، مقادیری به‌جز صفر و یـک دارند.

- مرحله‌ی بعد انتخاب تصادفی یکـی از رئـوس K اسـت[۱] کـه مرحلـه‌ی پـرواز (Flight Phase) نامیده می‌شود. چنان‌چه رأس انتخاب شده یکی از کل‌نمونه‌ها (منطبق بر یکی از رئوس مکعب) باشد، کـار انتخـاب نمونـه خاتمـه مـی‌یابـد و کل‌نمونه‌ی انتخابی دقیقاً متعادل‌شده خواهد بود اما اگـر رأس انتخـابی یکـی از کل‌نمونه‌ها نباشد به این مفهوم است که دسترسی به نمونه‌ی دقیقاً متعادل‌شـده امکان‌پذیر نیست و تنها می‌توان به نمونه‌ی تقریباً متعادل‌شده دسـت یافـت. بـه همین دلیل مرحله دیگری به کار انتخاب نمونه اضافه می‌شود که مرحله‌ی فرود (Landing Phase) نام دارد.

- در مرحله‌ی فرود با روش گرد کردن تـصادفی از رأس انتخـاب شـده‌ی K بـه یکی از رئوس مجاور در مکعب (یعنی یکی از کل‌نمونه‌های ممکـن) مـی‌رسـیم. نحوه‌ی انجام این کار مشابه انتخاب تصادفی یک کل‌نمونه در روش کل شماری است که قبلاً توضیح داده شده است. به این منظور لازم است ابتدا بـا توجـه بـه معادلات تعادل برای هر یک از کل‌نمونه‌های مجـاور رأس انتخـاب‌شـده‌ی K، هزینه محاسبه شود. سپس با مینیمم کردن امیـد ریاضـی تـابع هزینـه، احتمـال انتخاب این کل‌نمونه‌ها تعیین و در نهایت یکی از آن‌ها به تصادف انتخاب شـود. در هر حال کل‌نمونه‌ای که در مرحله‌ی فرود انتخاب می‌شود، تقریباً متعادل‌شـده

خواهد بود و از آن‌جا که هر رأس K از جمله رأس انتخاب شده در مرحله‌ی پرواز، حد اکثر p نقطه‌ی غیر از صفر یا یک دارد، حد اکثر در مورد p واحد جامعه، تکلیف انتخاب یا عدم انتخاب در مرحله‌ی فرود و از طریق گرد کردن تصادفی تعیین می‌شود.

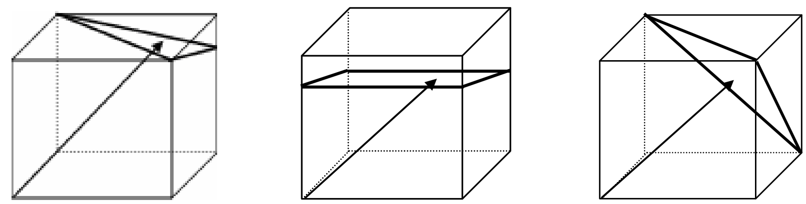
بنا بر این در انتخاب نمونه‌ی متعادل‌شده به‌روش مکعبی، همیشه نمی‌توان کل‌نمونه‌ای دقیقاً متعادل‌شده داشت. به‌طور کلی در یک نمونه‌گیری متعادل‌شده به‌روش مکعبی بسته به مقادیر متغیرهای کمکی و احتمال‌های انتخاب تک‌نمونه‌ها ($\mathbf{x}_k$ ، $t_{\mathbf{x}}$ و π_k) یکی از سه حالت زیر رخ می‌دهد.

الف- تمامی رئوس K منطبق بر رأس‌هایی از مکعب N-بعدی هستند. در این حالت طرح نمونه‌گیری دقیقاً متعادل‌شده خواهد بود به این معنی که در تمامی تکرارهای ممکن نمونه‌گیری، کل‌نمونه‌های انتخابی دقیقاً متعادل‌شده خواهند بود. این حالت در عمل به ندرت اتفاق می‌افتد،

ب- هیچ یک از رئوس K منطبق بر رأسی از مکعب نیستند. در این حالت تنها می‌توان طرح نمونه‌گیری تقریباً متعادل‌شده داشت به این مفهوم که در هیچ یک از تکرارهای ممکن نمونه‌گیری، کل‌نمونه‌ی انتخابی دقیقاً متعادل‌شده نخواهد بود. در واقع در این حالت هیچ یک از کل‌نمونه‌های ممکن در نمونه‌گیری در Q واقع نمی‌شوند و تنها می‌توان کل‌نمونه‌ای نزدیک به Q را به تصادف انتخاب کرد، و

پ- تنها برخی از رئوس K منطبق بر رئوسی از مکعب هستند. در این صورت در تکرار نمونه‌گیری، برخی از کل‌نمونه‌های انتخابی دقیقاً متعادل‌شده (کل‌نمونه‌های منطبق بر رئوس K) و سایر کل‌نمونه‌های انتخابی، تقریباً متعادل‌شده خواهند بود بنا بر این طرح نمونه‌گیری، تنها بعضی مواقع متعادل‌شده خواهد بود.

شکل زیر نمایش هندسی حالت‌های الف، ب و پ را وقتی $N = 3$ و با در نظر گرفتن سه حالت مختلف فرضی برای K و با وجود یک متغیر تعادل $(p = 1)$ نشان می‌دهد.



(پ) (ب) (الف)

شکل ۲- نمایش هندسی سه حالت ممکن چند وجهی دو بعدی K درداخل مکعب سه بعدی $[\circ,1]^3$

به‌طور کلی نیاز به گرد کردن تصادفی تقریباً در تمامی کاربردهای نمونه‌گیری متعادل‌شده کم و بیش وجود دارد هر چند تقریب ناشی از آن معمولاً قابل صرف نظر کردن است مگر این که اندازه‌ی نمونه خیلی کوچک باشد. به هر حال به دلیل گرد کردن به‌صورت تصادفی، خدشه‌ای به اصل انتخاب تصادفی و احتمال‌های انتخاب از پیش تعیین شده (π_k ها) وارد نمی‌شود. البته لازم است به این نکته توجه شود که گرد کردن، مسئله‌ای خاص نمونه‌گیری متعادل‌شده نیست و در روش‌های معمول نمونه‌گیری نیز گاهی نیاز به گرد کردن وجود دارد هر چند از آن‌جایی که معمولاً این کار به‌صورت غیر تصادفی انجام می‌شود، اصل تصادفی بودن انتخاب و یا احتمال‌های انتخاب تعیین شده (π_k ها) خدشه‌دار نیز می‌شوند اما با فرض ناچیز بودن میزان خدشه‌ی وارد شده، از آن صرف‌نظر می‌شود. مثالی در این زمینه یک نمونه‌گیری طبقه‌بندی با انتساب متناسب بین طبقات است که در آن π_k ها برای تمامی واحدها ($k = 1,2,\ldots,N$) و در همه‌ی طبقات یکسان است. هنگام توزیع نمونه بین طبقات، چنان‌چه اندازه‌های نمونه‌ی طبقه‌ها (مقادیر n_h)، اعدادی غیر صحیح باشند، معمولاً با گرد کردن به بالای اعشار بزرگ‌تر از ۴ و گرد کردن به پایین سایر موارد، به عدد صحیح تبدیل می‌شوند. با این کار در واقع بخشی از سهم تعدادی از طبقات کاسته شده و به سهم طبقات دیگری افزوده می‌شود که موجب تغییر احتمال انتخاب واحدها در طبقات مختلف می‌گردد. اما معمولاً با فرض ناچیز بودن میزان تغییر حاصل، از آن صرف‌نظر می‌شود. در این حالت روش صحیح که مانع تغییر

π_kها در طبقات مختلف شود، گرد کردن به‌صورت تصادفی و بـا احتمـال‌هـای متفـاوت است (برای مثال اگر سهم طبقه‌ای $n_h = 22/6$ شود باید به تـصادف و بـا احتمـال‌هـای متفاوت به ۲۲ یا ۲۳ گرد شود) و این مشابه همان شیوه‌ای است که در روش مکعبی برای گرد کردن استفاده می‌شود.

۳- شبیه‌سازی

برای ارزیابی عملکرد نمونه‌گیـری متعـادل‌شـده بـه‌روش مکعبـی، کـارایی آن را بـا روش نمونه‌گیری PPS که خود نیز روشی کارا در نمونه‌گیری احتمالی اسـت، مقایـسه نمـودیم. داده‌های مورد استفاده، فهرست حوزه‌های سرشماری استان قزوین در سال ۱۳۸۵ شـامل ۷۲۱ حوزه بوده است. برای هر حوزه اطلاعات جمعیتی و خـانواری شـامل تعـداد خـانوار، تعداد مرد، تعداد زن، تعداد محصل، تعداد باسواد، تعداد شاغل، تعداد بیکـار، تعـداد مهـاجر، جمعیت کل و جمعیت هر یک از گروه‌های عمده سنی از سرشـماری ۱۳۸۵ وجـود داشـته است. از فهرست مذکور، نمونه‌ای به اندازه‌ی ۱۱۶ حـوزه (مـشابه نـسبت نمونـه‌گیـری از خوشه‌ها در آمارگیری نیروی کار استان قزوین) به دو روش مکعبی و PPS انتخـاب شـده است. در هر دو روش نمونه‌گیری، احتمال انتخاب حوزه‌ها متناسب با تعداد خانوار حـوزه و به‌صورت زیر محاسبه گردید.

$$\pi_k = \frac{116 M_k}{\sum_{k \in U} M_k}$$

که در آن M_k، تعداد خانوار حوزه‌ی kام و $\sum_{k \in U} M_k$ تعداد کل خانوارهای استان قزوین است. نمونه PPS به‌روش سیستماتیک و با استفاده از فایل مرتب‌شده‌ی حوزه‌ها بر اساس تعداد خانوار حوزه، انتخاب شد. برای نمونه‌گیری متعـادل‌شـده، ۳ گـروه مختلـف از متغیرهای کمکی به شرح زیر مورد استفاده قرار گرفت.

گروه ۱ متغیرهای کمکی: واحد[۲]، کل جمعیت حوزه، تعداد خانوار حوزه؛

گروه ۲ متغیرهای کمکی: واحد،کل جمعیت حوزه، تعداد خانوار حوزه، تعداد بیکار حوزه؛

گروه ۳ متغیرهای کمکی: واحد،کل جمعیت حوزه، تعداد خانوار حوزه، تعداد بیکار حوزه، تعداد مهاجر حوزه و تعداد افراد ۶۵ ساله و بیش‌تر حوزه.

متناظر با هر گروه از متغیرهای کمکی یک نمونه‌ی متعادل‌شده انتخاب شد که به‌ترتیب با BS۱، BS۲ و BS۳ نام‌گذاری شده‌اند. به این ترتیب در مجموع چهار نوع نمونه‌گیری وجود دارد. براوردها در همه‌ی موارد به‌روش هورویتز–تامپسون محاسبه شده‌اند. هر روش نمونه‌گیری را با ۱۰۰۰ تکرار مستقل انجام دادیم و بر اساس آن برای هر متغیر، مقادیر تجربی متوسط مقدار مورد براورد ($MEAN$)، متوسط اریبی، انحراف استاندارد و میانگین توان دوم خطا (MSE)، را محاسبه کردیم. برای این که بتوان عملکرد روش‌ها برای متغیرهای مختلف را مقایسه نمود، شاخص میانگین توان دوم خطای نسبی با تعریف زیر نیز محاسبه شد.

$$RMSE = \frac{\sqrt{MSE}}{MEAN}$$

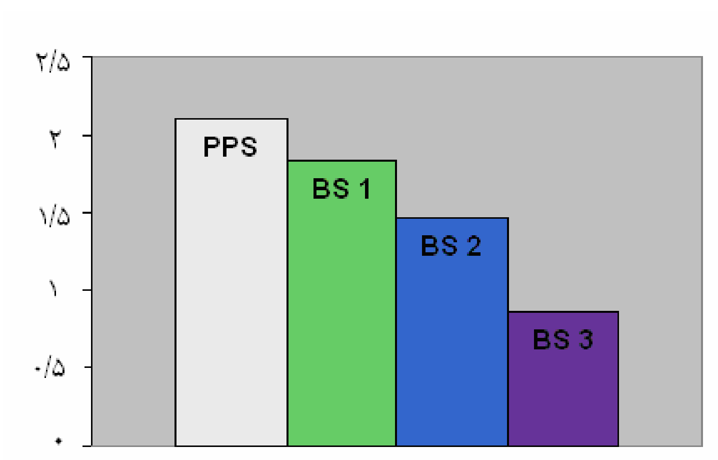
از متوسط مقادیر شاخص $RMSE$ روی تمامی متغیرهای مورد بررسی، به‌عنوان شاخص کلی ارزیابی روش استفاده شد. کار شبیه‌سازی با نرم‌افزار R انجام شده است. برای انتخاب نمونه‌ی متعادل‌شده نیز از برنامه‌ی نوشته‌شده توسط تیه و ماتی [۱۱] تحت عنوان Sample cube که در R وجود دارد، استفاده شده است.

جدول ۱، مقادیر میانگین توان دوم خطای نسبی ($RMSE$) هر یک از متغیرها به ازای هر یک از ۴ شیوه‌ی نمونه‌گیری را به‌صورت درصد نشان می‌دهد. سطر آخر جدول مذکور، متوسط مقادیر شاخص مزبور (روی کلیه‌ی متغیرها) برای هر یک از روش‌های نمونه‌گیری را نشان می‌دهد.

شکل ۳، مقادیر سطر آخر جدول مذکور را به تصویر کشیده است. این مقادیر به‌عنوان شاخص کلی در مقایسه‌ی بین ۴ روش نمونه‌گیری مورد استفاده قرار گرفته‌اند.

جدول ۱- میانگین توان دوم خطای نسبی ($RMSE$) براورد متغیرها به‌تفکیک طرح نمونه‌گیری (درصد)

شرح	*PPS*	*BS* **1**	*BS* **2**	*BS* **3**
جمعیت	۰/۵۲	۰/۱۷	۰/۱۶	۰/۱۶
تعداد خانوار	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰
تعداد بیکار	۶/۶۳	۵/۹۱	۰/۸۴	۰/۹۴
تعداد مهاجر	۴/۴۱	۵/۸۳	۵/۷۶	۰/۹۶
جمعیت ۶۵ ساله	۷/۲۴	۴/۹۰	۴/۶۴	۰/۸۱
تعداد مرد	۰/۶۴	۰/۳۳	۰/۳۴	۰/۳۳
تعداد زن	۰/۶۰	۰/۳۱	۰/۳۰	۰/۲۸
جمعیت ۱۴–۰ ساله	۱/۶۸	۱/۳۲	۱/۳۱	۱/۱۴
جمعیت ۲۴–۱۵ ساله	۱/۱۹	۰/۸۱	۰/۷۸	۰/۷۳
جمعیت ۳۴–۲۵ ساله	۱/۲۸	۱/۲۶	۱/۲۲	۱/۱۴
جمعیت ۶۴–۳۵ ساله	۱/۲۰	۱/۲۶	۱/۳۰	۱/۲۷
تعداد محصل	۲/۲۱	۱/۶۸	۱/۵۸	۱/۱۷
تعداد شاغل	۰/۸۱	۱/۱۳	۱/۰۱	۱/۰۰
تعداد باسواد	۱/۱۲	۰/۸۹	۰/۸۵	۰/۶۱
متوسط	۲/۱۰	۱/۸۴	۱/۴۳	۰/۷۵



شکل ۳- متوسط میانگین توان دوم خطای نسبی براوردها به‌تفکیک طرح نمونه‌گیری

نتایج به دست آمده به‌خوبی نشان می‌دهد که نمونه‌گیری متعادل‌شده در مقایسه با PPS کاراتر است و با افزایش تعداد متغیرهای کمکی مبنای تعادل، کارایی نمونه‌گیری متعادل‌شده نیز افزایش می‌یابد. علاوه براین افزایش کارایی نه تنها برای متغیرهای مبنای تعادل بلکه برای متغیرهای همبسته با آن‌ها نیز مشاهده می‌شود.

۴- نتیجه‌گیری

همان طور که انتظار می‌رفت نتایج شبیه‌سازی تأیید نمود که استفاده از نمونه‌گیری متعادل‌شده موجب بهبود کارایی براوردها می‌شود و با افزودن متغیرهای تعادل مناسب، میزان افزایش کارایی محسوس‌تر می‌شود. از آن‌جا که کارایی نمونه‌گیری متعادل‌شده با روش PPS مقایسه شده که خود روشی کارا در بین روش‌های نمونه‌گیری مورد استفاده در آمارگیری‌های ملی است، افزایش کارایی ناشی از کاربرد نمونه‌گیری متعادل‌شده ارزش بیش‌تری می‌یابد.

از طرفی در آمارگیری‌های ملی معمولاً از چارچوب‌هایی استفاده می‌شود که کم و بیش حاوی اطلاعات کمکی مرتبط با متغیرهای مورد نظر آمارگیری حداقل در سطح واحدهای نمونه‌گیری مرحله‌ی اول (Primary Sampling Units: PSUs) هستند. بنا بر این استفاده از روش نمونه‌گیری متعادل‌شده در این آمارگیری‌ها امکان‌پذیر می‌باشد. نکته‌ی مهم، تصمیم‌گیری در مورد متغیرهای تعادل مناسب است به نحوی که حتی‌الامکان برای هر کدام از متغیرهای اصلی مورد نظر در آمارگیری یک متغیر کمکی همبسته به‌عنوان متغیر تعادل منظور شود و در عین حال تعداد متغیرهای تعادل بی‌رویه افزایش نیابد.

مسئله‌ی دیگری که لازم است در استفاده از نمونه‌گیری متعادل‌شده مورد توجه قرار گیرد، چگونگی کیفیت اطلاعات کمکی است. درواقع افزایش کارایی تنها در صورتی کاملاً محقق می‌شود که اطلاعات کمکی از کیفیت خوبی برخوردار باشد (اطلاعات کمکی موجود در چارچوب نمونه‌گیری غلط یا ناقص نباشد). به هر حال، صحیح یا به‌هنگام نبودن اطلاعات کمکی منجر به اریبی نخواهد شد و چنانچه اطلاعات کمکی مورد استفاده از

کیفیت مطلوب برخوردار نباشد، کارایی طرح در عمل از کارایی مورد انتظـار کمتـر خواهـد بود .

سپاس‌گزاری

از پژوهشکده‌ی آمار برای حمایت مالی طرح پژوهشی مرتبط با این مقاله تشکر می‌کنیم.

توضیحات

[۱] نحوه‌ی انتخاب تصادفی رأسی از K در [۱۲] آمده است و در هرندی و همکاران [۱] با ذکر یک مثال عددی تشریح شده است.

[۲] متغیری با مقدار ثابت ۱ برای تمامی حوزه‌ها. متعادل نمودن روی این متغیر موجب می‌شود که معادله‌ی تعادل: $\hat{N} = N$ که در آن N تعداد حوزه‌های جامعه مورد بررسی و $\hat{N}$ براورد آن است، محقق گردد.

مرجع‌ها

[۱] هرندی، فاطمه؛ مهران،فرهاد؛ فرید روحانی،محمدرضا؛ فلاح‌محـسن‌خـانی،زهره (۱۳۸۶). گـزارش طرح پژوهشی «نمونه‌گیری متعادل‌شده و امکان‌سـنجی اسـتفاده از آن در آمـارگیری‌هـای ملـی». پژوهشکده‌ی آمار، تهران.

[2] Ardilly, P.(1991). Echantillonnage representatif optimum a probabilities inegales. *Annales d'Economie et de Statistique*, **23**, 91-113.

[3] Deville, J. C.; Grosbras, J. M.; Roth, N. (1988). Efficient sampling algorithms and balanced sample. Pages 255- 266 of: *COMPSTAT, Proceedings in Computational Statistics*. Heidelberg: Physica Verlag.

[4] Deville, J. C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilization of auxiliary information. *In: Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Orebro (Sweden).

[5] Deville, J. C.; Tille, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, **91**, 893- 912.

[6] Hedayat, A.S.; Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, **44**, 237- 247.

[7] Jensen, A. (1926), *Bull. Inst. Stat.*, **22**, 359- 381.

[8] Kiaer, A. N. (1895- 1896). *Bull. Int. Statist. Inst*., **9**, Liv. 2, 176- 183.

[9] Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**, 558- 606.

[10] Thionet, P. (1953). *La theorie des sondages*. Paris: INSEE, Imprimerie nationale.

[11] Tille, Y.; Matei, A. (2005). *The R package Sampling*. The Comprehensive R Archive Network, http: //cran. R- project. org/, Manual of the Contributed Packages.

[12] Tille, Y. (2006). *Sampling Algorithms*. Springer, New York.

[13] Valliant, R.; Dorfman, A. H.; Royall, R. M. (2000). *Finite Population Sampling and Inference, A prediction Approach*. Wiley, New York.

[14] Yates, F. (1946). A Review of Recent Statistical Developments in Sampling and Sampling Surveys. *Journal of the Royal Statistical Society*, **109**, 12- 43.

فاطمه هرندی

فوق لیسانس آمار

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان شهید فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.

پیام‌نگار: harandi@srtc.ac.ir

فرهاد مهران

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان شهید فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.

پیام‌نگار: