

نمونه‌ی تصادفی متعادل شده و نحوه‌ی انتخاب آن

فاطمه هرنندی،* 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 متعادل شده (Balanced Sampling) به‌عنوان یک روش نمونه‌گیری احتمالی که انتخاب نمونه‌ی نمایانگر از جامعه‌ی مورد بررسی را امکان‌پذیر می‌سازد، تشریح می‌شود. سپس نحوه‌ی انتخاب نمونه‌ی متعادل شده 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 متعادل شده موجب بهبود کارایی برآوردها 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 متعادل شده؛ 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 برآوردهایی با دقت بالا از پارامترهای مورد نظر در جامعه‌ی هدف دست یافت.

نمونه‌گیری متعادل شده به عنوان 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 متعادل شده‌ای 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 متعادل شده 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 و نحوه‌ی 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 **کل نمونه** (S)، مجموعه واحدهایی است که نمونه را تشکیل می‌دهند و منظور از **تک نمونه** (S_k)، k امین واحد نمونه است با تعریف زیر،

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

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

$$s = (s_1, \dots, s_k, \dots, s_N)'$$

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

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

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

$$p(s) = \Pr(s \in S)$$

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

$$\sum_{s \in S} p(s) = 1$$

$$0 \leq p(s) \leq 1 \quad ; \quad s \in S$$

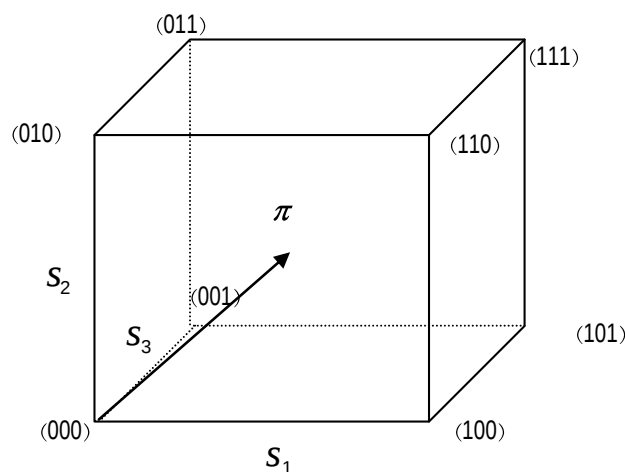
به بیان دیگر بعضی از کل نمونه‌ها می‌توانند شانس برای انتخاب نداشته باشند و این موضوع خللی به احتمالی بودن طرح نمونه‌گیری وارد نمی‌کند.

به‌طور کلی در هر نمونه‌گیری احتمالی، طرح نمونه‌گیری با زوج $(S, p(\cdot))$ یعنی مجموعه‌ی کل نمونه‌های ممکن و احتمال انتخاب هر یک از این کل نمونه‌ها تعریف می‌شود. بسته به طرح انتخابی بعضی از کل نمونه‌های ممکن می‌توانند شانس انتخاب صفر داشته باشند (برای مثال در نمونه‌گیری سیستماتیک بسیاری از $p(\cdot)$ ها صفر هستند زیرا کل نمونه‌های شامل واحدهای مجاور شانس برای انتخاب شدن ندارند). به بیان دیگر احتمال‌های انتخاب کل نمونه‌ها از طرح انتخاب شده برای نمونه‌گیری تأثیر می‌پذیرند. هرنندی 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 تک نمونه‌ی k ام را می‌توان از جمع احتمال‌های انتخاب کل نمونه‌هایی که واحد k ام را در بر دارند نیز به دست آورد، یعنی

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

طبق نظریه‌ی نمونه‌گیری، پس از فهرست کردن تمامی کل نمونه‌های ممکن و تعیین احتمال انتخاب هر یک، یکی از آن‌ها به تصادف انتخاب می‌شود. شکل ۱ نمایش هندسی کل نمونه‌های 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، کل نمونه‌های ممکن را نشان می‌دهد. π نیز احتمال انتخاب تک‌نمونه‌ها را نشان می‌دهد.

در حالت کلی در هر نمونه‌گیری احتمالی، فضای نمونه‌گیری را می‌توان به صورت یک اُبر مکعب N -بعدی در نظر گرفت که هر یک از ابعاد آن، یکی از تک‌نمونه‌ها و هر یک از رئوس آن یکی از کل نمونه‌های ممکن را نشان می‌دهد. الگوریتم نمونه‌گیری، رفتن از π به یکی از رئوس به صورت 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 کل نمونه‌ها، در قالب قیود زیر و بسته به الگوریتم انتخابی برای نمونه‌گیری تعیین می‌شوند.

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

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

$$Q = \{s \in S \mid \hat{t}_{x\pi} = t_x\}$$

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

$$\text{var}(\hat{t}_{x\pi}) = 0.$$

ایده‌ی اصلی در نمونه‌گیری متعادل‌شده این است که با توجه به توانایی طرح نمونه‌گیری متعادل‌شده در ارائه‌ی برآورد دقیق متغیرهای کمکی، انتظار می‌رود واریانس نمونه‌گیری سایر برآوردهای 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 کل نمونه‌های ممکن فهرست شوند سپس برای این که ارزش هر کل نمونه از نظر میزانی که شرط تعادل را محقق می‌کند مشخص شود به هر کل نمونه، هزینه‌ای منتسب شود به این ترتیب که اگر کل نمونه‌ی 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 کل نمونه‌ها به نحوی تعیین شوند که کل نمونه‌های 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).

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

در رابطه‌ی فوق $\text{Cost}(s)$ هزینه‌ی انتخاب کل نمونه‌ی s است.

پس از حل معادله‌ی بالا و به دست آوردن مقادیر $p(\cdot)$ ها، با توجه به احتمال‌های انتخاب تعیین شده، یکی از کل نمونه‌ها به تصادف انتخاب می‌شود. برای دیدن مثالی از کاربرد روش بالا برای انتخاب نمونه‌ی متعادل‌شده به [۱] رجوع کنید.

استفاده از روش بالا برای انتخاب نمونه‌ی متعادل‌شده که تیه [۱۲] آن را روش کل شماری نامیده، علی‌رغم سادگی در جوامع واقعی که اغلب بزرگ هستند، عملی نیست زیرا تعداد کل نمونه‌های ممکن بسیار زیاد است (برای مثال در نمونه‌گیری بدون جایگذاری با $N = 20$ ، تعداد کل نمونه‌های ممکن 1048576 است). این محدودیت خاص نمونه‌گیری متعادل‌شده نیست زیرا در نمونه‌گیری از جوامع واقعی نیز هیچ‌گاه تمامی کل نمونه‌های ممکن و احتمال انتخاب آن‌ها را فهرست نمی‌کنیم بلکه در عمل بر اساس مقادیر π_k ها، واحدهای نمونه را تک‌تک انتخاب می‌کنیم. در نمونه‌گیری متعادل‌شده نیز می‌توان بدون تعیین مقادیر احتمال انتخاب تک‌تک کل نمونه‌ها ($p(\cdot)$ ها)، یکی از آن‌ها را انتخاب کرد. روش اخیر برای انتخاب نمونه‌ی متعادل‌شده که دوپل و تیه [۵] آن را روش مکعبی نام نهاده‌اند بر مبنای نمایش هندسی فضای نمونه‌گیری و مفهوم انتخاب نمونه در این فضا استوار است.

همان‌گونه که قبلاً اشاره شد، در هر نمونه‌گیری تصادفی بدون جایگذاری، مجموعه‌ی کل نمونه‌های ممکن را می‌توان به‌صورت رئوس یک ابر 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 شود که در واقع همان الگوریتم نمونه‌گیری است. در روش مکعبی، برای انتخاب نمونه‌ی متعادل‌شده، تحقق 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، $t_x \pi = t_x$) در

فضای N -بعدی تعیین می‌شود که یک زیرفضای آفین

(Affine Subspace) با بعد $N - p$ است و با نماد Q نشان داده می‌شود. طی مراحل بعدی سعی می‌شود به صورت تصادفی یک کل نمونه در Q انتخاب شود و اگر هیچ کل نمونه‌ای در Q قرار نگیرد تلاش می‌شود کل نمونه‌ای حتی‌الامکان نزدیک به Q به صورت تصادفی انتخاب شود.

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

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

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

• مرحله‌ی بعد انتخاب تصادفی یکی از رئوس K است^۱ که مرحله‌ی پرواز (Flight Phase) نامیده می‌شود. چنانچه رأس انتخاب شده یکی از کل نمونه‌ها (منطبق بر یکی از رئوس مکعب) باشد، کار انتخاب نمونه خاتمه می‌یابد و کل نمونه‌ی انتخابی دقیقاً متعادل شده خواهد بود اما اگر رأس انتخابی یکی از کل نمونه‌ها نباشد به این مفهوم است که دسترسی به نمونه‌ی دقیقاً متعادل شده امکان‌پذیر نیست و تنها می‌توان به نمونه‌ی تقریباً متعادل شده دست یافت. به همین دلیل مرحله دیگری به کار انتخاب نمونه اضافه می‌شود که مرحله‌ی فرود (Landing Phase) نام دارد.

• در مرحله‌ی فرود با روش گرد کردن تصادفی از رأس انتخاب شده‌ی K به یکی از رئوس مجاور در مکعب (یعنی یکی از کل نمونه‌های ممکن) می‌رسیم. نحوه‌ی انجام این کار مشابه انتخاب تصادفی یک کل نمونه در روش کل شماری است که قبلاً توضیح داده شده است. به این منظور لازم است ابتدا با توجه به معادلات تعادل برای هر یک از کل نمونه‌های مجاور رأس انتخاب شده‌ی K ، هزینه محاسبه شود. سپس با مینیمم کردن امید ریاضی تابع هزینه، احتمال انتخاب این کل نمونه‌ها تعیین و در نهایت یکی از آن‌ها به تصادف انتخاب شود. در هر حال کل نمونه‌ای که در مرحله‌ی فرود انتخاب می‌شود، تقریباً متعادل شده

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.

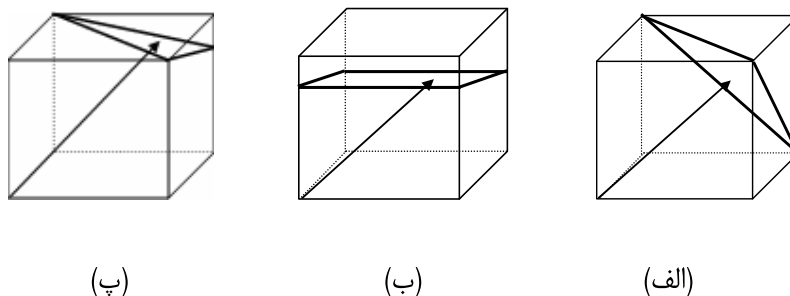
بنا بر این در انتخاب نمونه‌ی متعادل‌شده به‌روش مکعبی، همیشه نمی‌توان کل نمونه‌ای دقیقاً متعادل‌شده داشت. به‌طور کلی در یک نمونه‌گیری متعادل‌شده به‌روش مکعبی بسته به مقادیر متغیرهای کمکی و احتمال‌های انتخاب تک‌نمونه‌ها (π_k و t_x, x_k) یکی از سه حالت زیر رخ می‌دهد.

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، کل نمونه‌های 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، کل نمونه‌ی انتخابی دقیقاً متعادل‌شده نخواهد بود. در واقع در این حالت هیچ یک از کل نمونه‌های ممکن در نمونه‌گیری در Q واقع نمی‌شوند و تنها می‌توان کل نمونه‌ای 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 کل نمونه‌های انتخابی دقیقاً متعادل‌شده (کل نمونه‌های منطبق بر رئوس K) و سایر کل نمونه‌های 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.



شکل ۲- نمایش هندسی سه حالت ممکن چند وجهی دو بعدی K در داخل مکعب سه بعدی $[0,1]^3$

به طور کلی نیاز به گرد کردن تصادفی تقریباً در تمامی کاربردهای نمونه‌گیری متعادل شده 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 به صورت 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 متعادل شده نیست و در روش‌های معمول نمونه‌گیری نیز گاهی نیاز به گرد کردن وجود دارد هر چند از آن جایی که معمولاً این کار به صورت 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 صرف نظر می‌شود. مثالی در این زمینه یک نمونه‌گیری طبقه‌بندی با انتساب متناسب بین طبقات است که در آن π_k ها برای تمامی واحدها ($k = 1, 2, \dots, N$) و در همه‌ی طبقات یکسان است. هنگام توزیع نمونه بین طبقات، چنانچه اندازه‌های نمونه‌ی طبقه‌ها (مقادیر π_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 صرف نظر می‌شود. در این حالت روش صحیح که مانع تغییر

π_k ها در طبقات مختلف شود، گرد کردن به صورت تصادفی و با احتمال‌های متفاوت است (برای مثال اگر سهم طبقه‌ای $n_h = 22/6$ شود باید به تصادف و با احتمال‌های 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 نمونه‌گیری متعادل‌شده به روش مکعبی، کارایی آن را با روش نمونه‌گیری PPS که خود نیز روشی کارا در نمونه‌گیری احتمالی است، مقایسه نمودیم. داده‌های مورد استفاده، فهرست حوزه‌های سرشماری استان قزوین در سال ۱۳۸۵ شامل ۷۲۱ حوزه بوده است. برای هر حوزه اطلاعات جمعیتی و خانواری شامل تعداد خانوار، تعداد مرد، تعداد زن، تعداد محصل، تعداد باسواد، تعداد شاغل، تعداد بیکار، تعداد مهاجر، جمعیت کل و جمعیت هر یک از گروه‌های عمده سنی از سرشماری ۱۳۸۵ وجود داشته است. از فهرست مذکور، نمونه‌ای به اندازه‌ی ۱۱۶ حوزه (مشابه نسبت نمونه‌گیری از خوشه‌ها در آمارگیری نیروی کار استان قزوین) به دو روش مکعبی و PPS انتخاب شده است. در هر دو روش نمونه‌گیری، احتمال انتخاب حوزه‌ها متناسب با تعداد خانوار حوزه و به صورت زیر محاسبه گردید.

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

که در آن M_k ، تعداد خانوار حوزه‌ی k ام و $\sum_{k \in U} M_k$ تعداد کل خانوارهای استان قزوین است. نمونه PPS به روش سیستماتیک و با استفاده از فایل مرتب‌شده‌ی حوزه‌ها بر اساس تعداد خانوار حوزه، انتخاب شد. برای نمونه‌گیری متعادل‌شده، ۳ گروه مختلف 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 بیش تر حوزه. متناظر با هر گروه از متغیرهای کمکی یک نمونه‌ی متعادل شده انتخاب شد که به ترتیب با BS_1 ، BS_2 و BS_3 نام گذاری شده‌اند. به این ترتیب در مجموع چهار نوع نمونه‌گیری وجود دارد. برآوردها در همه‌ی موارد به روش 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 برآورد ($MEAN$)، 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.

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

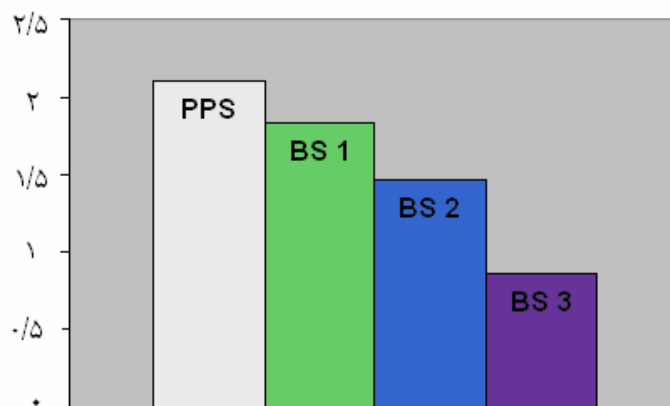
از متوسط مقادیر شاخص $RMSE$ روی تمامی متغیرهای مورد بررسی، به عنوان شاخص کلی ارزیابی روش استفاده شد. کار شبیه‌سازی با نرم‌افزار R انجام شده است. برای انتخاب نمونه‌ی متعادل شده نیز از برنامه‌ی نوشته شده توسط تیه و ماتی [۱۱] تحت عنوان Sample cube که در R وجود دارد، استفاده شده است.

جدول ۱، مقادیر میانگین توان دوم خطای نسبی ($RMSE$) هر یک از متغیرها به ازای هر یک از ۴ شیوه‌ی نمونه‌گیری را به صورت 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 به عنوان شاخص کلی در مقایسه‌ی بین ۴ روش نمونه‌گیری مورد استفاده قرار گرفته‌اند.

جدول ۱- میانگین توان دوم خطای نسبی (RMSE) برآورد متغیرها به تفکیک طرح نمونه‌گیری (درصد)

شرح	PPS	BS1	BS2	BS3
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 پاسواد	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 برآوردها به تفکیک طرح نمونه‌گیری

نتایج به دست آمده به خوبی نشان می‌دهد که نمونه‌گیری متعادل شده در مقایسه با PPS کارا تر است و با افزایش تعداد متغیرهای کمکی مبنای تعادل، کارایی نمونه‌گیری متعادل شده نیز افزایش می‌یابد. علاوه بر این 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 متعادل شده موجب بهبود کارایی برآوردها می‌شود و با افزودن متغیرهای تعادل مناسب، میزان افزایش کارایی محسوس‌تر می‌شود. از آن‌جا که کارایی نمونه‌گیری متعادل شده 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 متعادل شده 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 متعادل شده 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 به عنوان متغیر تعادل منظور شود و در عین حال تعداد متغیرهای تعادل بی‌رویه افزایش نیابد.

مسئله‌ی دیگری که لازم است در استفاده از نمونه‌گیری متعادل شده مورد توجه قرار گیرد، چگونگی کیفیت اطلاعات کمکی است. در واقع 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 انتظار کمتر خواهد 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 هرنندی و همکاران [۸] 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 $\hat{N}$ برآورد آن است، محقق گردد.

مرجع‌ها

- [۱] هرنندی، فاطمه؛ مهران، فرهاد؛ فرید روحانی، محمدرضا؛ فلاح‌محسن‌خانی، زهره (۱۳۸۶). گزارش طرح پژوهشی «نمونه‌گیری متعادل‌شده و امکان‌سنجی استفاده از آن در آمارگیری‌های ملی». پژوهشکده‌ی آمار، تهران.
- [2] Ardilly, P.(1991). Echantillonnage representatif optimum a probabilités inégales. *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 هرنندی

فوق لیسانس آمار

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

پیام‌نگار: harandi@srtc.ac.ir

فرهاد مهران

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

پیام‌نگار:

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.