مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۲، شماره‌ی ۱، بهار و تابستان ۱۳۹۰، صص ۶۳ – ۷۱

مقایسه‌ی کارایی نمونه‌گیری متعادل‌شده و PPS تحت شرایط یکسان و بررسی تأثیر اندازه‌ی نمونه بر آن‌ها

فاطمه هرندی،[†] زهره فلاح محسن‌خانی[†,*] و محمدرضا فرید روحانی[‡]

[†] پژوهشکده‌ی آمار
[‡] دانشگاه شهید بهشتی

چکیده: در این مقاله ابتدا کارایی روش نمونه‌گیری متعادل‌شده در مقایسه با روش نمونه‌گیری با احتمال متناسب با اندازه (PPS)[1] در شرایط یکسان یعنی در حالتی که در هر دو روش تنها از اطلاعات یک متغیر کمکی استفاده شود، بررسی می‌شود. سپس با انجام یک شبیه‌سازی، کارایی مزبور در صورت تغییر کسر نمونه‌گیری[2] بررسی می‌شود. یافته‌ها حاکی از آن است که در شرایط یکسان از نظر اطلاع کمکی مورد استفاده، فارغ از اندازه‌ی کسر نمونه‌گیری، کاربرد نمونه‌گیری متعادل‌شده به بهره‌ای[3] بیش از نمونه‌گیری PPS منجر می‌شود و با افزایش کسر نمونه‌گیری میزان بهره‌ی حاصل نیز افزایش می‌یابد.

واژگان کلیدی: نمونه‌گیری متعادل‌شده؛ نمونه‌گیری PPS؛ روش مکعبی؛ اندازه‌ی نمونه؛ کارایی؛ بهره.

۱- مقدمه

برای دستیابی به راهبرد نمایانگر در نمونه‌گیری از جامعه‌های متناهی یعنی راهبردی که از طریق روش نمونه‌گیری، روش براورد یا هر دوی آن‌ها، امکان دست‌یابی به براوردهایی دقیق از ویژگی‌های مورد نظر جامعه‌ی مورد مطالعه را فراهم آورد، روش‌های مختلفی طراحی شده است. ویژگی مشترک همه‌ی این روش‌ها، استفاده از اطلاعات کمکی مرتبط با ویژگی‌های مورد نظر جامعه در مرحله‌ی طراحی نمونه‌گیری، مرحله‌ی براورد یا در هر دو

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۸۹/۶/۳، پذیرش: ۱۳۹۰/۵/۱۵.

مرحله است [۱]. از جمله روش‌های مطرح در مرحله‌ی طراحی نمونه‌گیری، روش نمونه‌گیری PPS است که در اغلب آمارگیری‌های ملی حد اقل در یکی از مرحله‌های نمونه‌گیری مورد استفاده قرار می‌گیرد. یکی دیگر از روش‌های مورد توجه در این زمینه روش نمونه‌گیری متعادل‌شده[۴] است که هر چند از دیرباز مطرح بوده است [۴] اما تنها در سال‌های اخیر دویل و تیه موفق شده‌اند روشی عملی به‌نام روش مکعبی را برای انتخاب نمونه‌ی متعادل‌شده‌ی احتمالی ارایه دهند [۳] که از قابلیت لازم برای استفاده در آمارگیری‌های ملی نیز برخوردار است و هم اکنون در مراکز آماری چند کشور مورد استفاده قرار می‌گیرد [۸]. هر چند تمام مطالعه‌هایی که تا کنون برای ارزیابی عمل‌کرد نمونه‌گیری متعادل‌شده به‌روش مکعبی انجام شده است دلالت بر کاراتر بودن این روش در مقایسه با روش نمونه‌گیری PPS دارد، که خود روشی کارا در نمونه‌گیری احتمالی است [۱، ۳]، اما نکته‌ی مهمی که در این مطالعه‌ها بررسی نشده، یکسان نبودن شرایط مقایسه (به‌جز روش نمونه‌گیری) است. در واقع در تمام این مطالعه‌ها از اطلاعات بیش از یک متغیر کمکی برای متعادل کردن نمونه‌ها استفاده شده در صورتی که در نمونه‌گیری PPS بنا بر ماهیت آن، تنها از مقدارهای یک متغیر کمکی به‌عنوان معیار اندازه برای گزینش واحدها استفاده شده است. بنا بر این در مورد نتیجه‌های این مطالعه‌ها این ابهام وجود دارد که کاراتر بودن نتیجه‌های حاصل از نمونه‌گیری متعادل‌شده از PPS، ناشی از ماهیت ذاتی این روش نمونه‌گیری یا استفاده‌ی گسترده‌تر آن از اطلاعات کمکی موجود درباره‌ی جامعه‌ی مورد بررسی است. به این ترتیب، آن‌چه که در این مطالعه‌ها مورد بررسی قرار نگرفته است، علت کاراتر بودن نمونه‌گیری متعادل‌شده از روش مکعبی است و همچنین این موضوع که اندازه‌ی نمونه چه نقشی در کارایی مزبور دارد.

در این مقاله سعی شده است به سؤال‌های بالا پاسخ داده شود. بدین منظور، ابتدا در بخش دوم این مقاله به معرفی اجمالی دو روش نمونه‌گیری متعادل‌شده و PPS و نحوه‌ی مقایسه‌ی این دو روش در شرایط یکسان پرداخته می‌شود. در بخش سوم مقاله، شبیه‌سازی طرح‌ریزی‌شده در پژوهش حاضر برای بررسی کارایی نمونه‌گیری متعادل‌شده در مقایسه با PPS در شرایط یکسان و تأثیر اندازه‌ی نمونه بر این کارایی تشریح می‌شود. بخش چهارم مقاله به بیان نتایج حاصل از شبیه‌سازی اختصاص یافته است.

۲- مقایسه‌ی نمونه‌گیری متعادل‌شده و PPS در شرایط یکسان

به‌طور کلی در هر نمونه‌گیری احتمالی، از یک جامعه‌ی متناهی با N واحد، اگر مجموعه‌ی واحدهایی که نمونه را تشکیل می‌دهند (کل‌نمونه) با حرف $\boldsymbol{S}$ و k امین واحد نمونه (تک‌نمونه) را با علامت S_k، با تعریف زیر درنظر بگیریم:

$$S_k = \begin{cases} ۱ & \text{اگر واحد } k \text{ ام در نمونه‌ی } \boldsymbol{S} \text{ باشد} \\ ۰ & \text{اگر واحد } k \text{ ام در نمونه‌ی } \boldsymbol{S} \text{ نباشد.} \end{cases}$$

هر طرح نمونه‌گیری را می‌توان با زوج $(\boldsymbol{S}, p(\cdot))$ یعنی مجموعه‌ی کل‌نمونه‌های ممکن و احتمال انتخاب هر یک از این کل‌نمونه‌ها تعریف کرد به نحوی که:

$$(۱) \qquad \begin{cases} \sum_{\boldsymbol{s}\in\boldsymbol{S}} p(\boldsymbol{s}) = ۱ \\ ۰ \le p(\boldsymbol{s}) \le ۱; \quad \boldsymbol{s} \in \boldsymbol{S} \\ ۰ < \pi_k = \sum_{k\in\boldsymbol{s}} p(\boldsymbol{s}) \le ۱; \quad k = ۱,۲,\ldots,N. \end{cases}$$

π_k احتمال انتخاب تک‌نمونه‌ی k ام است. در هر نمونه‌گیری احتمالی، مقدارهای π_k باید برای تمام واحدهای جامعه‌ی مورد بررسی، بزرگ‌تر از صفر باشد. هر چند احتمال‌های انتخاب کل‌نمونه‌ها ($p(\cdot)$ ها) از طرح انتخاب‌شده برای نمونه‌گیری تأثیر می‌پذیرند و بسته به طرح انتخاب‌شده، 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).

۲-۱- نمونه‌گیری متعادل‌شده به روش مکعبی[۵]

در نمونه‌گیری متعادل‌شده، سعی بر این است که به کل‌نمونه‌هایی که از جنبه‌ی اطلاعات کمکی موجود در زمان طراحی نمونه‌گیری، نمایانگر مناسبی برای جامعه‌ی مورد بررسی نیستند شانسی برای انتخاب داده نشود. بدین منظور تلاش می‌شود تنها به کل نمونه‌هایی شانس انتخاب داده شود که علاوه بر صدق کردن در شرط‌های رابطه‌ی (۱) بتوانند براوردهایی دقیق (برابر با مقدار واقعی معلوم) از متغیرهای کمکی همبسته با متغیرهای

مورد بررسی تولید کنند. طرح نمونه‌گیری $(\boldsymbol{S}, p(\cdot))$، متعادل‌شده بر اساس متغیرهای کمکی x نامیده می‌شود اگر و فقط اگر $\boldsymbol{Q} = \{\boldsymbol{s} \in \boldsymbol{S} \mid \hat{\boldsymbol{t}}_{X\pi} = \boldsymbol{t}_X\}$ مجموعه‌ی کل‌نمونه‌های ممکن با احتمال غیر صفر باشد. به بیان دیگر نمونه‌گیری متعادل‌شده، طرحی است که در آن تنها کل‌نمونه‌هایی شانس انتخاب غیر صفر دارند که معادله‌ی تعادل $\hat{\boldsymbol{t}}_{X\pi} = \boldsymbol{t}_X$ را محقق می‌سازند.

در رابطه‌ی بالا، x بردار p متغیر کمکی مربوط به واحد kام است که مؤلفه‌های آن مبنای تعادل قرار می‌گیرند (متغیرهای مبنای تعادل)، $\hat{\boldsymbol{t}}_{X\pi} = \sum_{k\in U} \frac{\boldsymbol{x}_k \boldsymbol{s}_k}{\pi_k}$ براورد هورویتز تامپسون مجموع x و $\boldsymbol{t}_X = \sum_{k\in U} \boldsymbol{x}_k$ مجموع معلوم x است.

برای انتخاب نمونه‌ی متعادل‌شده به روش مکعبی [۸]، کل‌نمونه‌های ممکن به‌صورت رئوس یک ابرمکعب N بعدی در نظر گرفته می‌شوند. سپس شکل هندسی معادله‌های تعادل (Balancing Equations، $\hat{\boldsymbol{t}}_{X\pi} = \boldsymbol{t}_X$) در فضای N بعدی که یک زیرفضای آفین (Affine Subspace) با بعد $N - p$ است، ترسیم می‌شود. طی مراحل بعدی سعی می‌شود به‌صورت تصادفی یک کل‌نمونه در زیرفضای مزبور انتخاب شود (کل‌نمونه‌ی متعادل‌شده) و اگر هیچ کل‌نمونه‌ای در این زیرفضا قرار نگیرد تلاش می‌شود کل‌نمونه‌ای حتی الامکان نزدیک به آن (کل‌نمونه‌ی تقریباً متعادل‌شده) به‌صورت تصادفی انتخاب شود. برای آگاهی از جزئیات نحوه‌ی انتخاب تصادفی یک کل‌نمونه‌ی متعادل‌شده یا تقریباً متعادل‌شده به [۱] مراجعه کنید.

۲-۲- نمونه‌گیری PPS

در نمونه‌گیری PPS، مقدار π_k متناسب با بزرگی اندازه‌ی متغیر کمکی واحد kام تعیین می‌شود. معیار اندازه، مقادیر متغیر کمکی موجود در چارچوب آمارگیری است که در صورت انتخاب متغیر مناسب، به افزایش کارایی نمونه‌گیری PPS در مقایسه با نمونه‌گیری تصادفی ساده منجر خواهد شد. برای مباحث تفصیلی‌تر در مورد این نوع نمونه‌گیری به [۵، ۶ و ۹] مراجعه کنید.

۳-۲- یکسان‌سازی شرایط

برای این که بتوان مقایسه‌ی درستی بین دو روش نمونه‌گیری متعادل‌شده و PPS انجام داد لازم است تا حد امکان تمام شرایط به‌جز روش انتخاب نمونه یکسان باشد. بدین منظور باید در هر دو روش نمونه‌گیری از اطلاعات کمکی یکسانی استفاده شود و علاوه بر آن احتمال‌های انتخاب تک‌نمونه‌ها (π_k ها) نیز برای دو روش نمونه‌گیری یکسان باشد.

۳- شبیه‌سازی

از آن جا که در انتخاب نمونه‌ی متعادل‌شده به روش مکعبی، همیشه نمی‌توان کل‌نمونه‌ای دقیقاً متعادل‌شده داشت و کیفیت تعادل در این روش به مقدارهای متغیرهای کمکی مورد استفاده بستگی دارد [۱، ۲ و ۸]. برای بررسی کارایی نمونه‌گیری متعادل‌شده به روش مکعبی در مقایسه با PPS در شرایط یکسان، یک شبیه‌سازی با داده‌های تصادفی انجام شده است. در این مرحله علاوه بر تعیین کارایی کلی، تأثیر اندازه‌ی نمونه بر این کارایی نیز مورد بررسی قرار گرفته است.

بدین منظور ابتدا داده‌های مورد نیاز در مطالعه با استفاده از نرم‌افزار R تولید شده است. در گام اول، ۱۰۰۰ داده‌ی تصادفی از یک توزیع گاما با میانگین و انحراف استاندارد ۲ تولید شد. این اعداد به‌عنوان مقدارهای متغیر کمکی مورد استفاده در هر دو شیوه‌ی نمونه‌گیری در نظر گرفته شدند. در گام دوم، چندجمله‌ای‌های درجه‌ی m از متغیر کمکی مزبور با ضریب‌های تصادفی دارای توزیع یکنواخت و جمله‌ی خطای تصادفی نرمال استاندارد تولید شدند. برای درجه‌ی چندجمله‌ای‌ها (m)، مقدارهای ۱، ۲، ۳، ۴، ۵ و ۱۰ در نظر گرفته شدند. هر یک از این چندجمله‌ای‌ها، به‌عنوان مقدارهای یکی از متغیرهای مورد نظر در آمارگیری (y) لحاظ شدند. به این ترتیب داده‌های جامعه‌ی آماری مورد بررسی به‌صورت ماتریسی با ابعاد (۷ × ۱۰۰۰) به دست آمد ($[x, y_1, y_2, y_3, y_4, y_5, y_6]$).

پس از ساخت ماتریس داده‌ها، ابتدا احتمال انتخاب تک‌نمونه‌ها تعیین می‌شود. برای مثال، احتمال انتخاب تک‌نمونه‌ی k ام (π_k)، به‌صورت زیر تعیین و برای هر دو روش نمونه‌گیری، مورد استفاده قرار می‌گیرد.

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

در رابطه‌ی بالا، M_k مقدار متغیر کمکی برای واحد kام، $\sum_{k \in U} M_k$ جمع کل مقادیر متغیر کمکی در جامعه‌ی مورد بررسی و n تعداد نمونه است.

بر اساس هر یک از دو روش نمونه‌گیری، تعداد ۱۰۰۰۰ بار از جامعه‌ی مذکور نمونه‌ی n تایی انتخاب، و بر اساس آن برای هر متغیر، مقدارهای تجربی متوسط مقدار مورد براورد (MEAN) و میانگین توان دوم خطا (MSE) محاسبه شد. برای انتخاب نمونه‌ی متعادل‌شده به روش مکعبی از برنامه‌ی نوشته‌شده توسط تیه و ماتی [۷] با عنوان Sample Cube که در R وجود دارد، استفاده شد. براوردها در همه‌ی حالت‌ها به‌روش هورویتز- تامپسون محاسبه شدند. برای تعیین تأثیر اندازه‌ی نمونه، محاسبات بالا برای کسرهای نمونه‌گیری مختلف شامل ۰/۰۱، ۰/۱، ۰/۱۵ و ۰/۲، جداگانه انجام شد.

برای این که بتوان عمل‌کرد روش‌ها را مقایسه کرد، برای متغیرهای پاسخ مختلف، شاخص براورد میانگین توان دوم خطای نسبی با تعریف زیر برای هر دو روش نمونه‌گیری و برای کسرهای نمونه‌گیری مذکور محاسبه شد.

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

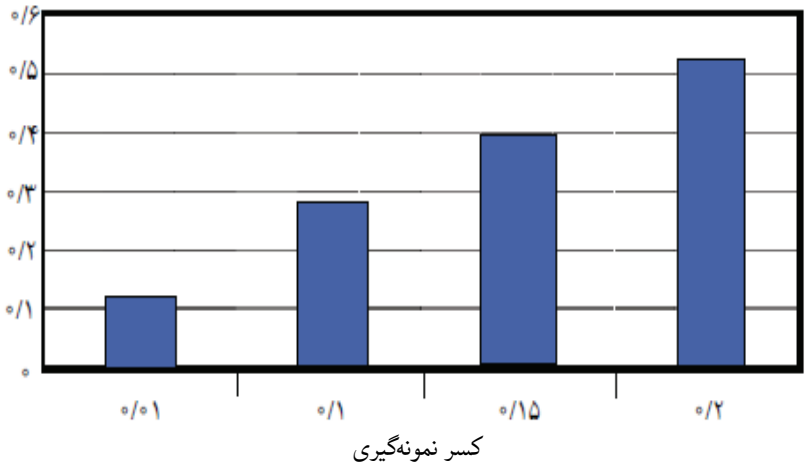
از متوسط مقدارهای شاخص مذکور روی شش متغیر مورد بررسی به‌عنوان شاخص کلی ارزیابی روش استفاده شد. بر اساس شاخص کلی ارزیابی روش نمونه‌گیری، بهره‌ی حاصل از نمونه‌گیری متعادل‌شده به روش مکعبی به‌صورت زیر محاسبه شد:

$$\mathrm{EF\,(CUBE)} = ۱ - \frac{\mathrm{RMSE\,(CUBE)}}{\mathrm{RMSE\,(PPS)}}$$

۴- نتیجه‌گیری

یافته‌های حاصل از شبیه‌سازی نشان می‌دهد که حتی در شرایط یکسان از نظر اطلاعات کمکی مورد استفاده نیز بهره‌ی حاصل از نمونه‌گیری متعادل‌شده بیش از نمونه‌گیری PPS

است. بنا بر این کاراتر بودن نمونه‌گیری متعادل‌شده نسبت به PPS، ماهیت ذاتی این روش نمونه‌گیری است که ناشی از صفر یا کم کردن احتمال انتخاب کل‌نمونه‌های نامناسب است. میزان افزایش کارایی با افزایش کسر نمونه‌گیری بیش‌تر می‌شود. شکل ۱، بهره‌ی حاصل از نمونه‌گیری متعادل‌شده [EF(CUBE)] را به ازای مقدارهای مختلف کسر نمونه‌گیری به‌صورت درصد نشان می‌دهد.



شکل ۱- بهره‌ی نمونه‌گیری متعادل‌شده نسبت به نمونه‌گیری PPS برای مقدارهای مختلف کسر نمونه‌گیری

توضیحات

1. Probability Proportional to size Sampling
2. Sampling Fraction
3. Gain
4. Balanced Sampling
5. CUBE Method

مرجع‌ها

[۱] هرندی، فاطمه؛ مهران، فرهاد؛ فرید روحانی، محمدرضا؛ فلاح محسن‌خانی، زهره (۱۳۸۶). گزارش طرح پژوهشی «*نمونه‌گیری متعادل‌شده و امکان‌سنجی استفاده از آن در آمارگیری‌های ملی*». پژوهشکده‌ی آمار، تهران.

[۲] هرندی، فاطمه؛ مهران، فرهاد (۱۳۸۷). *نمونه‌ی تصادفی متعادل‌شده و نحوه‌ی انتخاب آن*. *گزیده‌ی مطالب آماری*، سال ۱۹ شماره‌ی ۲ ، پاییز و زمستان ۱۳۸۷، صص ۱۸۶-۱۷۱.

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

[4] 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.

[5] Sarndal, C.E.; Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

[6] Thompson, S.K. (2002). *Sampling*, 2nd ed. Wiley, New York.

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

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

[9] Verma, V. (2002). *Sampling Methods*. Training Handbook, Revised, SIAP, Tokyo.

فاطمه هرندی
فوق لیسانس آمار
تهران، خیابان سید جمال‌الدین اسدآبادی، خیابان ۲۵، شماره‌ی ۵، پژوهشکده‌ی آمار.
رایانشانی: harandi@srtc.ac.ir

زهره فلاح محسن‌خانی
فوق لیسانس آمار
تهران، خیابان سید جمال‌الدین اسدآبادی، خیابان ۲۵، شماره‌ی ۵، پژوهشکده‌ی آمار.
رایانشانی: zohrehf@srtc.ac.ir

محمدرضا فریدروحانی
دکتری آمار
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه آمار.
رایانشانی: m-faridrohani@sbu.ac.ir