کاربرد نمونه‌گیری دوفازی در برآوردگر رگرسیونی عام

حمید بیدرام[۱]

چکیده

نمونه‌گیری دو فازی[۲] یکی از روشهای نمونه‌گیری است که دارای کاربردهای متعدد و پتانسیل عملی بالایی می‌باشد. این روش که اولین بار توسط نیمن[۳] (۱۹۳۸) معرفی شد، هنگامی مورد استفادهٔ بهینه قرار می‌گیرد که در جامعه یا چارچوب نمونه‌گیری با متغیرهای کمکی مواجه باشیم. از این رو در این مقاله ضمن معرفی این روش نمونه‌گیری، ارتباط آن با برآوردگر رگرسیونی عام[۴] که بر مبنای متغیرهای کمکی استوار است، بیان و کاربرد این روش نمونه‌گیری را در برآوردگر مذکور مطرح می‌نمائیم.

واژه‌های کلیدی: نمونه‌گیری تصادفی ساده، نمونه‌گیری دوفازی، روش نمونه‌گیری عام.

۱. مقدمه

وقتی چارچوب نمونه‌گیری شامل اطلاعات کمکی کم و یا اصلا فاقد اطلاعات کمکی باشد این اطلاعات ممکن است از دو راه زیر بدست آید.

الف) استفاده از یک طرح خیلی ساده مثل SI (طرح نمونه‌گیری تصادفی ساده بدون جایگذاری) که با برآودگر π (برآوردگر هارویتز- تامپسون)[۲] ، ترکیب شده باشد. در اینجا، دقت قابل قبول در برآوردها ممکن است از طریق یک نمونه با حجم زیاد بدست آیند ولی این حجم زیاد ممکن است با هزینه زیادی همراه باشد.

ب) جمع‌آوری اطلاعات از طریق یک ساختار جدید، چارچوب اطلاع‌رسانی زیاد و سپس ترکیب طرح نمونه‌گیری با برآوردگر رگرسیونی، در اینجا یک نمونهٔ کوچک ممکن است دقت کافی را دارا باشد. به هر حال در اینجا نیز ممکن است هنوز بواسطهٔ هزینه‌های موجود در ایجاد ساختار جدید، موضوع هزینه‌های بررسی بالا باشد.

اما راه سومی که در اینجا معرفی خواهیم کرد یک سازگاری بین دو راه بالا با نمونهٔ انتخاب شده در روش دوفازی برقرار می‌کند. روش به این صورت است که

---

[۱] گروه آمار، دانشگاه اصفهان

[۲] Two – Phase Sampling

[۳] Neyman

[۴] General Regression Estimation

۱ـ در فاز اول، یک نمونه نسبتاً بزرگ به نام s_a از عناصر توسط یک طرح نمونه‌گیری ساده $p_a(\cdot)$ انتخاب می‌کنیم. برای عناصر در s_a از روی متغیرهای کمکی (یا متغیر کمکی) اطلاعات کم هزینه‌ای را جمع‌آوری می‌کنیم.

۲ـ با کمک اطلاعات کمکی انتخاب شده در فاز اول، یک نمونه s از s_a در فاز دوم با استفاده از طرح $p(\cdot \mid s_a)$ انتخاب می‌کنیم؛ این یک زیرنمونه خواهد بود. مطالعه متغیر هدف Y توسط مشاهدات آن در نمونهٔ فاز دوم انجام می‌گیرد. این روش یک روش نمونه‌گیری به نام **نمونه‌گیری دو فازی** می‌باشد. گسترش این نمونه‌گیری به بیش از دو فاز نمونه‌گیری چندفازی نامیده می‌شود. این نمونه‌گیری دارای مزایای زیادی است که از جمله مزیت‌های این روش نمونه‌گیری، ایجاد یک چارچوب با اطلاع‌رسانی زیاد است که تا حد امکان اطلاعات جامعه توسط این چارچوب منعکس شده و لزومی به داشتن اطلاعات از چارچوب اولیه نیست.

حسن دیگر این نمونه‌گیری این است که در هنگام عدم وجود پاسخ در طرح نمونه‌گیری، با استفاده از فاز دوم می‌توان به رفع عدم پاسخ[۵] پرداخت. بدین صورت که چون در انتخاب نمونهٔ فاز دوم فرض بر این است که نمونهٔ فاز اول داده شده است، کلیهٔ عدم پاسخها که در نمونهٔ فاز اول بدست آمده‌اند از چارچوب نمونه‌گیری فاز دوم که از فاز اول بدست می‌آید، کنار گذاشته می‌شود و با این کار در فاز دوم با مشکل عدم پاسخ مواجه نخواهیم شد. از این خصوصیت در اغلب بررسیهای نمونه‌ای که با عدم پاسخ مواجه‌اند استفاده می‌کنند[۳].

۲. نمادها و انتخاب برآوردگرها

نمونه‌ای که از فاز اول بدست می‌آید را با s_a و حجم آن را با n_{s_a} نمایش می‌دهیم، طرح نمونه‌گیری مورد استفاده در فاز اول را با $p_a(\cdot)$ و تابع احتمال آن را با $p_a(s_a)$ نمایش می‌دهیم که $p_a(s_a)$ به معنای احتمال انتخاب نمونه s_a ، می‌باشد. احتمالات انتخاب متناظر با این فاز برای هر $k,l \in U$ (که U مجموعه عناصر جامعه است) عبارتند از

$$\pi_{ak} = \sum_{s_a \ni k} p_a(s_a) \quad (۱)$$

و

$$\pi_{akl} = \sum_{s_a \ni k,l} p_a(s_a) \quad (۲)$$

که π_{ak} احتمال انتخاب عضو kام و π_{akl} احتمال انتخاب توأم اعضای k و l در فاز اول می‌باشد. با نمونهٔ داده شدهٔ s_a ، نمونهٔ فاز دوم s به حجم n_s مطابق با طرح $p(\cdot \mid s_a)$ بدست می‌آید؛ بطوری‌که $p(s \mid s_a)$ احتمال شرطی انتخاب نمونه s است. احتمالات انتخاب تحت طرح مذکور برای هر $k,l \in s_a$ عبارت است از

$$\pi_{k|s_a} = \sum_{s \ni k} p(s \mid s_a) \quad (۳)$$

و

$$\pi_{kl|s_a} = \sum_{s \ni k,l} p_a(s \mid s_a) \quad (۴)$$

که $\pi_{k|s_a}$ احتمال انتخاب عضو kام و $\pi_{kl|s_a}$ احتمال انتخاب توأم اعضای k و l در فاز دوم می‌باشد. به شرطی که نمونهٔ فاز اول یعنی s_a داده شده باشد. قدم بعدی پیدا کردن یک برآوردگر نااریب برای مجموع کل جامعه $t = \sum_U Y_k$ می‌باشد[۱]. برای این کار یک برآوردگر π (برآوردگر هارویتز ـ تامپسون) به کار می‌رود، یعنی

$$\hat{t}_\pi = \sum_s \breve{Y}_k = \sum_s \frac{Y_k}{\pi_k} \quad (۵)$$

که $\pi_k = p(k \in s)$ ، احتمال انتخاب عنصر kام است. واضح است که برآوردگر (۵) نیاز به محاسبه π_k دارد ولی این کار همیشه در نمونه‌گیری دوفازی امکان‌پذیر نیست. زیرا

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

که $p(s)$ احتمال انتخاب نمونهٔ s است و

$$p(s) = \sum_{s_a \supset s} p_a(s_a) p(s \mid s_a)$$

و بنابراین

---

[۵] Nonresponse

$$\pi_k = \sum_{s\ni k}\sum_{s_a\supset s} p_a(s_a)p(s \mid s_a)$$

$$= \sum_{\substack{s_a\ni k s_a\supset s \\ s\ni k}}\sum p_a(s_a)p(s \mid s_a)$$

$$= \sum_{s_a\ni k} p_a(s_a)[\sum_{\substack{s_a\supset s \\ s\ni k}} p(s \mid s_a)] \quad (۶)$$

$$= \sum_{s_a\ni k} p_a(s_a)\pi_{k|s_a}$$

پس برای تعیین π_k در عمل، باید احتمالات $p_s(s_a)$ را برای هر s_a (که معمولا آن را می‌دانیم) بدست آوریم. همچنین باید $\pi_{k|s_a}$ را برای هر s_a (که معمولا نمی‌دانیم زیرا ممکن است $\pi_{k|s_a}$ به نتیجه نمونه از فاز اول وابسته باشد) بدست آوریم.ولی در عمل، همیشه امکان محاسبهٔ $\pi_{k|s_a}$ وجود ندارد و در نتیجه π_k قابل محاسبه نیست. لذا برآوردگر زیر را معرفی می کنیم.

۳. برآوردگر π^*

از آنجایی که برآوردگر π همیشه نمی‌تواند در عمل مورد استفاده قرار گیرد دنبال یک برآوردگر نااریب هستیم که در نمونه‌گیری دوفازی از آن استفاده نمائیم. فرض کنید

$$\sum_{s_a}\breve{Y}_{ak} = \sum_{s_a}\frac{Y_{ak}}{\pi_{ak}}$$

یک برآوردگر π برای $t = \sum_U Y_k$ باشد بطوریکه برای هر $Y_k, k \in s_a$ و π_{ak} معلوم باشد. ولی همانطور که قبلا هم ذکر شد مقادیر Y_k برای نمونهٔ دوم (یا نمونهٔ فاز دوم) یعنی فقط برای $k \in s$ قابل مشاهده و دسترسی می‌باشد، بنابراین برای s_a داده شده برآوردگر $\sum_{s_a}\breve{Y}_{ak}$ را به صورت یک برآوردگر شرطی π نااریب به صورت

$$\sum_s \frac{\breve{Y}_{ak}}{\pi_{k|s_a}} = \sum_s \frac{Y_k}{\pi_{ak}\pi_{k|s_a}} \quad (۷)$$

در نظر می گیریم[۱]. این برآوردگر برای $\pi_{k|s_a}$ که برای هر s_a از فاز اول معلوم و مشخص است یک برآوردگر قابل استفاده است. با معرفی

$$\pi_k^* = \pi_{ak}\cdot\pi_{k|s_a} \quad (۸)$$

متوجه می‌شویم که در حقیقت در (۷) به مقادیر $\sum Y_k$ وزنی برابر $1/\pi_k^*$ تعلق گرفته است، بنابراین داریم

$$\sum_s \frac{Y_k}{\pi_{ak}\cdot\pi_k \mid s_a} = \sum_s \frac{Y_k}{\pi_k^*} = \sum_s \breve{\breve{Y}}_k$$

این برآوردگر نااریب جدید را برآوردگر π^* می‌نامیم و به صورت زیر نمایش می‌دهیم.

$$\hat{t}_{\pi^*} = \sum_s \frac{Y_k}{\pi_k^*} = \sum_s \breve{\breve{Y}}_k \quad (۹)$$

تذکر ۱: واضح است که $\hat{t}_{\pi^*}$ با $\hat{t}_\pi$ داده شده در (۵) یکسان نیست چون $\pi_k \neq \pi_k^*$ و علت آن این است که برای هر $k \in s_a$

$$p(k \in s \mid k \in s_a) = \frac{p(k \in s)}{p(k \in s_a)} = \frac{\pi_k}{\pi_{ak}} \quad (۱۰)$$

که بطور کلی با $\pi_{k|s_a}$ اختلاف دارد. در اینجا $\pi_{k|s_a}$ احتمال این است که نمونهٔ فاز دوم s تحت یک طرح بخصوص شامل عضو k ام باشد وقتی s_a در فاز اول داده شده است. از طرفی π_k/π_{ak} احتمال شرطی انتخاب عضو k ام در فاز دوم است به شرطی که عضو k ام در نمونه s_a وجود داشته باشد. حال با استفاده از (۹) نتیجه می گیریم که بطور کلی π_k و π_k^* یکسان نیستند، زیرا

$$\pi_k = p(k \in s) = p(k \in s_a)p(k \in s \mid k \in s_a)$$

$$= \pi_{ak}(\pi_k/\pi_{ak}) \neq \pi_{ak}\cdot\pi_k \mid s_a = \pi_k^*$$

٤. برآوردگر رگرسیونی عام برای نمونه‌گیری دو فازی

برآوردگر π^* داده شده در (۹) منحصرا روی وزن عناصر بنا شده است. احتمالات انتخاب طرح فاز دوم نمونه‌گیری ممکن است به اطلاعات جمع‌آوری شده در فاز اول بستگی داشته باشد ولی متغیرهای کمکی در فرمول برآوردگر π^* ظاهر نمی‌شوند. در اینجا می‌خواهیم استفادهٔ صحیح و سادهٔ متغیرهای کمکی را به شکلی از برآوردگر رگرسیونی در نمونه‌گیری دوفازی بیان کنیم. حال متغیرهای کمکی ممکن است دو نوع باشند.

الف) مقادیر بدست آمده توسط مشاهدهٔ عناصر در نمونهٔ s_a از فاز اول ، یعنی مقادیری که در چارچوب استفاده شده برای فاز دوم ظاهر می‌شوند (همانطور که می‌دانیم وقتی نمونهٔ اول انتخاب شد از آن به عنوان چارچوب نمونه‌گیری برای نمونهٔ فاز دوم استفاده می‌کنیم).

ب) مقادیر قابل دسترس برای همهٔ N عنصر جامعهٔ U یعنی مقادیر داده شده در چارچوب اولیه.

همانطور که ذکر شد ایدهٔ اصلی در نمونه‌گیری دوفازی این است که بتوانیم از متغیرهای کمکی نوع (الف) استفاده کنیم. در اکثر کاربردها، اطلاعاتی از نوع (ب) در اختیار نیست. قبل از اینکه در اینجا به ارتباط برآوردگر رگرسیونی عام بپردازیم لازم است بر حسب فازهای نمونه‌گیری متغیرهای کمکی را نمادگذاری کنیم.

۱- فرض کنید $\mathbf{X}_k$ یک بردار از J متغیر کمکی قابل دسترسی برای هر $k \in s_a$ باشد.

۲- فرض کنید $\mathbf{X}_{1k}$ یک بردار از J_1 متغیر کمکی قابل دسترسی برای هر $k \in U$ باشد.

فرض کنید که $\mathbf{X}_k$ شامل متغیرهایی که از قبل برای هر $k \in U$ معلوم و همچنین شامل متغیرهایی معلوم فقط برای هر $k \in s_a$ باشد، از طرفی می‌توان نوشت

$$\mathbf{X}_k = (\mathbf{X}'_{1k}, \mathbf{X}'_{2k})'$$

که در آن $\mathbf{X}_{1k}$ بردار J_1 تایی از مقادیر معلوم برای هر $k \in U$ بوده و $\mathbf{X}_{2k}$ برداری از $J_2 = J - J_1$ مقدار ثبت شده با استفاده از K عنصر فقط در نمونهٔ فاز اول s_a باشد. در برآوردگر رگرسیونی که در [۱] مفصلا مورد بحث واقع شده است مقادیر پیش‌بینی مربوط به $\mathbf{X}_k$ را با $\hat{Y}_k$ از s به s_a و مقادیر پیش‌بینی مربوط به $\mathbf{X}_{1k}$ را با $\hat{Y}_{1k}$ از s_a به U نشان می‌دهیم. در جدول شماره یک به طور خلاصه این مقادیر را نشان داده‌ایم.

هدف در این بخش، بهبود برآوردگر π^* با فرض اینکه طرح‌ها در هر دو فاز ثابت باشند، می‌باشد که با بدست آوردن برآوردگر رگرسیونی عام در زیر به هدف خود خواهیم رسید.

برآوردگر رگرسیونی عام را، که سارندال و سونیسون (۱۹۸۷) [۲] معرفی کردند و جزئیات آن در[۱] آمده است با استفاده از نمادها و متغیرهای بیان شده، معرفی می‌کنیم. در اینجا دو منبع برای بدست آوردن متغیرهای کمکی در نظر می‌گیریم؛ فرض کنید $\mathbf{X}_{1k}$ برای هر $k \in U$ معلوم بوده و $\mathbf{X}_{2k}$ مقادیر معلوم برای هر $k \in s_a$ باشد. بنابراین برای یک عنصر $k \in s_a$ اطلاعات کامل را بطور خلاصه در بردار

$$\mathbf{X}_k = (\mathbf{X}'_{1k}, \mathbf{X}'_{2k})' \qquad (11)$$

قرار داده که متغیر مورد مطالعه Y_k برای مقادیر $k \in s$ (فاز دوم) قابل مشاهده است. نهایتا برآوردگر رگرسیونی عام بر طبق روش نمونه‌گیری دوفازی به صورت زیر معرفی می‌شود

$$\hat{t}_r = \sum_{u} \hat{Y}_{1k} + \sum_{s_a} \frac{\hat{Y}_k - \hat{Y}_{1k}}{\pi_{ak}} + \sum_{s} \frac{Y_k - \hat{Y}_k}{\pi^*_k} \qquad (12)$$

$$= \sum_{s} \frac{Y_k}{\pi^*_k} + \left(\sum_{U} \hat{Y}_{1k} - \sum \frac{\hat{Y}_{1k}}{\pi_{ak}}\right) + \left(\sum_{s_a} \frac{\hat{Y}_k}{\pi_{ak}} - \sum_{s} \frac{\hat{Y}_k}{\pi^*_k}\right)$$

که $\hat{Y}_k$ و $\hat{Y}_{1k}$ مقادیر پیش‌بینی شده از برازش یک مدل رگرسیونی مناسب، می‌باشد و π^*_k و π_{ak} به ترتیب در (۹) و (۱) معرفی شده‌اند.

جدول ۱

مجموعه عناصر	بردارهای متغیرهای کمکی
جامعهٔ U	$\mathbf{X}_{1k}$
نمونهٔ فاز اول s_a	$\mathbf{X}_k = (\mathbf{X}'_{1k}, \mathbf{X}'_{2k})'$
نمونهٔ فاز دوم s	$\mathbf{X}_k = (\mathbf{X}'_{1k}, \mathbf{X}'_{2k})'$
متغیرهای مشاهده شده	**مقادیر پیش بینی شده**
–	$\hat{Y}_{1k}$
–	$\hat{Y}_k , \hat{Y}_{1k}$
Y_k	$\hat{Y}_k$

مراجع

[۱] بیدرام، حمید، ۱۳۷۹ ، برآوردگر رگرسیونی عام و کاربرد آن در بررسی هزینه خانوار، پایان نامهٔ کارشناسی ارشد آمار، دانشگاه صنعتی اصفهان.

[2] Sardnal, C.E., Swensson, B. and wretman, J., 1990. *Model Assisted Survey Sampling*, Springer-Verlag, New Youk.

[3] Zieschang, K.D., 1990. *Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey,* Journal of the American Statistical Association, 85, 986- 1001.