

برآورد خطای پرسشگر در آمارگیری‌ها

معصومه موحدی،* حمیدرضا نواب‌پور

دانشگاه علامه طباطبائی

چکیده. اگر آمارگیری به صورت رو در رو توسط پرسشگر اجرا شود، حضور پرسشگر تأثیر به‌سزایی بر هزینه‌های آمارگیری و کیفیت داده‌ها خواهد داشت. خطای پرسشگر به دلیل عدم دقت پرسشگرها به وجود می‌آید و باعث خطا در پاسخ می‌شود. در این مقاله ضمن مرور خطاهای غیرنمونه‌گیری با انجام یک شبیه‌سازی، واریانس میانگین برای داده‌های شبیه‌سازی شده که تحت تأثیر خطای پرسشگر قرار گرفته‌اند برآورد می‌شوند. همچنین واریانس برآوردگر در یک آمارگیری فرضی با به کارگیری سه مجموعه از پرسشگرها محاسبه و مقایسه می‌شود تا تأثیر بارکار پرسشگرها بر دقت برآوردها مورد مقایسه قرار گیرد. همچنین نتیجه‌های به دست آمده با در نظر گرفتن خطای پرسشگر با فرمول‌های کلاسیک (بدون در نظر گرفتن خطای پرسشگر) مقایسه می‌شود. نتیجه‌ی شبیه‌سازی نشان می‌دهد که هر چه بارکار پرسشگر بیش‌تر باشد دقت کاهش می‌یابد.

۱- مقدمه

فعالیت‌های مختلف یک آمارگیری به چهار زمینه‌ی اصلی تقسیم می‌شوند: ساخت چارچوب نمونه‌گیری، طراحی طرح نمونه‌گیری، انتخاب واحدهای نمونه‌ای و جلب همکاری آن‌ها، و گردآوری داده‌ها و انتقال آن‌ها به رایانه. خطاهایی که ممکن است در هر یک از این مرحله‌ها رخ دهند عبارت‌اند از: خطای چارچوب، خطای نمونه‌گیری، خطای بی‌پاسخی و خطای اندازه‌گیری. در این تقسیم‌بندی خطای چارچوب، خطای بی‌پاسخی و خطای اندازه‌گیری، خطاهای غیرنمونه‌گیری نامیده می‌شوند. چنان‌چه در آمارگیری‌ها

واژگان کلیدی: خطای پرسشگر؛ شرایط ایده‌آل؛ بارکار پرسشگر؛ همبستگی درون پرسشگری.

دریافت: ۱۳۸۸/۱/۲۴، پذیرش: ۱۳۸۸/۷/۵

* نویسنده‌ی عهده‌دار مکاتبات

خطاهای غیرنمونه‌گیری رخ ندهند شرایط آمارگیری شرایطی ایده‌آل نامیده می‌شود. وجود شرایط ایده‌آل در آمارگیری‌ها تقریباً غیر ممکن است یعنی آمارگیری‌ها فارغ از خطاهای غیرنمونه‌گیری نیستند. بنا بر این از آن‌جا که شرایط ایده‌آل در آمارگیری‌ها وجود ندارد، نمی‌توان از فرمول‌های برآورد کلاسیک استفاده کرد. در صورت استفاده از این فرمول‌ها به دلیل وجود خطا، آماره‌ها از اعتبار کافی برخوردار نخواهند بود. تقریباً در اکثر آمارگیری‌هایی که از پرسشگر برای گردآوری داده‌ها استفاده می‌شود خطای پرسشگر رخ می‌دهد. خطای پرسشگر جزء خطای اندازه‌گیری می‌باشد. وجود خطای پرسشگر در آمارگیری مانند سایر خطاهای غیرنمونه‌گیری باعث از بین رفتن شرایط ایده‌آل می‌شود، بنا بر این باید برای برآورد پارامترها، برآوردگرهای دیگری یافت [۳].

۲- خطای پرسشگر در آمارگیری‌ها

تمام مطالعه‌ها این نتیجه‌گیری را تأیید می‌کنند که پرسشگرها به نوعی در ایجاد انگیزه برای پاسخگویان نقش دارند، فرایندهایی که بر اساس آن‌ها این نتیجه‌ها حاصل شده است از مطالعه‌ای به دست آمده که از سوی فالولر و مانگیون [۲] گزارش شده است. در آمارگیری‌هایی که برای گردآوری داده‌ها از پرسشگر استفاده می‌شود، برای کاهش تأثیر پرسشگرها بر پاسخ‌ها، باید پرسشگرها به خوبی آموزش دیده باشند و همچنین نباید بارکار زیاد داشته باشند. برای داشتن پرسشگرهایی آموزش دیده با بارکار کم، هزینه‌ی کافی باید در اختیار باشد. عامل هزینه در تمام آمارگیری‌ها عامل مهمی است. باید توجه کرد که در صورت استفاده از چنین آمارگیری‌هایی چنانچه هزینه‌ی کافی در اختیار نباشد، آموزش پرسشگرها و تعداد آن‌ها تحت تأثیر قرار خواهند گرفت و در نتیجه آماره‌های آمارگیری تحت تأثیر خطای پرسشگر قرار می‌گیرند.

خطای اندازه‌گیری یکی از خطاهای غیرنمونه‌گیری است که در اکثر آمارگیری‌ها بروز می‌کند. این خطا تفاوت پاسخ صحیح و پاسخی است که به پایگاه داده‌های نهایی راه پیدا می‌کند. یکی از منابع خطای اندازه‌گیری، خطای پرسشگر می‌باشد. خطای پرسشگر همانند دیگر منابع خطای اندازه‌گیری باعث تغییر در پاسخ واقعی می‌شود. روش‌های

نمونه‌گیری برای مدل‌بندی خطای اندازه‌گیری از جمله خطای پرسشگر با کار هنس، هورویتس و مادو [۵] شروع شده است.

۳- مدل تحلیل واریانس

چند مدل برای برآورد واریانس در حضور خطای پرسشگر پیشنهاد شده است، مدل زیر که از مدل ارائه شده توسط هنس، هورویتس و برشارد [۴] اقتباس شده است مدل تحلیل واریانس می‌باشد. در ذیل این مدل به‌طور مختصر توضیح داده شده است [۱].

در مدل تحلیل واریانس، مدل خطای اندازه‌گیری را می‌توان به‌صورت زیر نوشت.

$$(۱) \quad y_{ij} = \mu_{ij} + b_i + e_{ij}$$

که در آن μ_{ij} مقدار واقعی پاسخ و y_{ij} مقدار مشاهده شده، b_i خطای تصادفی ناشی از پرسشگر i ام و e_{ij} خطای تصادفی برای واحد j ام مربوط به پرسشگر i ام می‌باشد که به‌عنوان تفاوت میان مشاهده‌ی y_{ij} و $\mu_{ij} + b_i$ برای واحد j ام تعریف می‌شود، که با اثر پرسشگر ناهمبسته می‌باشد و لذا انحرافی اضافی است که به پاسخگو هنگام مصاحبه از سوی پرسشگر i ام ارتباط دارد.

$$e_{ij} = y_{ij} - \mu_{ij} - b_i$$

فرض می‌شود که b_i ها اثرهای تصادفی دارای میانگین ۰ و واریانس σ_b^2 هستند. بنا بر این به‌دلیل وجود استقلال بین مؤلفه‌های موجود در سمت راست مدل، واریانس مشاهده‌ی y_{ij} به‌صورت زیر به دست می‌آید.

$$\text{var}(y_{ij}) = \sigma_\mu^2 + \sigma_b^2 + \sigma_e^2$$

که در این جا $\sigma_e^2 = \sum_i \sum_j \sigma_{ij}^2 / n$ و

$$\text{Cov}(y_{ij}, y_{i'j'}) = \begin{cases} 0 & i \neq i' \\ \sigma_b^2 & i = i', j \neq j' \end{cases}$$

برای سادگی فرض می‌شود که برای تمام i ها، $m_i = m$ و همچنین $n = mI$ ، I تعداد پرسشگرها می‌باشد. از آن‌جا که $E(\bar{y}) = \mu$ است، می‌توان نتیجه گرفت که $\text{Bias}(\bar{y}) = 0$.

اگر ضریب همبستگی داخل پرسشگری (ρ_y) به صورت زیر تعریف شود،

$$(۲) \quad \rho_y = \frac{\sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2}$$

$\hat{\sigma}_\mu^2$ ، $\hat{\sigma}_b^2$ و $\hat{\sigma}_e^2$ به ترتیب براورهای واریانس‌های مربوط به خطا، مقدارهای واقعی داده‌ها و خطای پرسشگر می‌باشند.

اگر کسر نمونه‌گیری نادیده گرفته شود، واریانس میانگین به صورت زیر به دست می‌آید.

$$V(\bar{y}) = \frac{1}{n}(\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2)[1 + (m-1)\rho_y]$$

$$= \frac{V(y_{ij})}{n}[1 + (m-1)\rho_y]$$

۴- یک مطالعه‌ی شبیه‌سازی

با توجه به این که به دست آوردن داده‌هایی که دارای خطای پرسشگر باشند و این خطا را به وضوح مشخص کرده باشند در عمل یافت نمی‌شوند، لذا داده‌های این مطالعه توسط نرم‌افزار R از جامعه‌ی نامحدود شبیه‌سازی شده‌اند.

۵- تولید داده‌ها

با استفاده از نرم‌افزار R جامعه‌ای به اندازه‌ی ۱۰۰۰۰۰ واحد برای متغیر پاسخ از توزیع نرمال با میانگین ۲ و انحراف معیار ۳/۲ تولید شد. از این جامعه یک نمونه‌ی تصادفی به اندازه‌ی ۱۰۰ انتخاب شد. در مرحله‌ی اول نمونه‌ی انتخاب شده به صورت تصادفی بین ۲۰ پرسشگر تقسیم شد و واریانس برآوردگر میانگین در حضور خطایی که این ۲۰ پرسشگر ایجاد می‌کنند، برآورد شد. در مرحله‌ی دوم این نمونه بین ۱۰ پرسشگر و در مرحله‌ی سوم بین ۵ پرسشگر به صورت تصادفی تقسیم شد. سپس واریانس برآوردگر مورد نظر در سه مرحله برآورد و مورد بررسی قرار می‌گیرند تا تأثیر بارکارهای متفاوت پرسشگرها مقایسه شوند.

۶- طرح مطالعه

برای به دست آوردن برآوردها در حضور خطای پرسشگر داده‌ها باید به نحوی تولید شوند که دارای خطای پرسشگر باشند. برای رسیدن به چنین داده‌هایی گام‌های زیر اجرا شده است.

- ۱- در اولین گام نمونه‌ای به اندازه‌ی ۱۰۰ ($n = 100$) به روش تصادفی ساده بدون جایگذاری از جامعه‌ی ساخته شده، انتخاب می‌شود. برای این نمونه مقادیرهای واقعی، μ_{ij} در دست است. برای ساختن y_{ij} از مدل‌های معرفی شده در گام‌های بعدی استفاده می‌شود.
- ۲- نمونه‌ی به دست آمده در گام اول به تصادف بین پرسشگرها تقسیم می‌شود. فرض می‌شود تعداد مصاحبه‌های پرسشگرها یکسان هستند.
- ۳- داده‌های به دست آمده در نمونه به دلیل تأثیر پرسشگر با مقدار واقعی در جامعه متفاوت هستند ولی این داده‌ها با مقدار واقعی در جامعه همبستگی دارند. در این گام مدلی در نظر گرفته شده است که نشان‌دهنده‌ی همبستگی میان داده‌های نمونه با مقادیرهای واقعی خود می‌باشد. مدل در نظر گرفته شده برای این منظور به صورت زیر می‌باشد.

$$(۴) \quad y_{ij} = 0/2 \mu_{ij} + e_{ij}^*$$

ضریب ۰/۲ نشان‌دهنده‌ی همبستگی میان واحدهای نمونه‌ای با واحدهای واقعی خود در جامعه است. μ_{ij} نیز همان‌طور که در بالا گفته شد مقدار واقعی پاسخ i ام مربوط به پرسشگر j ام می‌باشد. e_{ij}^* خطای تصادفی می‌باشد، این خطا از توزیع نرمال با میانگین ۰ و انحراف معیار ۲ تولید می‌شود، این خطا نشان‌دهنده‌ی انحرافی است که داده‌های نمونه از مقادیرهای واقعی خود در جامعه دارند.

- ۴- در این گام باید مدلی در نظر گرفت که نشان‌دهنده‌ی تأثیر پرسشگرها بر روی داده‌ها باشد. در گام قبل با استفاده از مدل برآزش داده شده همبستگی بین واحدهای نمونه‌ای و واحدهای مشابه آن‌ها در جامعه ایجاد شد. از آن‌جا که

پرسشگرها بر داده‌های مربوط به خود تأثیر می‌گذارند، باید میان داده‌های مربوط به یک پرسشگر همبستگی وجود داشته باشد. در این گام واحدهای نمونه‌ای به گونه‌ای ساخته می‌شوند که بین واحدهای یک پرسشگر همبستگی ایجاد شود. با استفاده از مدل زیر می‌توان این همبستگی را ایجاد نمود.

$$(۵) \quad y_{ij} = 0/3 y_{i(j-1)} + e_{ij}^{\circ}$$

که در آن ضریب ۰/۳ نشان‌دهنده‌ی تأثیری است که پرسشگر i ام بر داده‌های مربوط به خود می‌گذارد. $y_{i(j-1)}$ مقدار داده‌ی $(j-1)$ ام مربوط به پرسشگر i ام می‌باشد. e_{ij}° خطای تصادفی می‌باشد. این خطا از توزیع نرمال با میانگین ۰ و انحراف معیار ۱/۵ تولید می‌شود، توجه کنید که این خطا ممکن است شامل خطاهای دیگری غیر از خطای نمونه‌گیری مانند خطای پاسخگو نیز باشد.

۵- پس از به دست آوردن y_{ij} ها، کوواریانس مربوط به هر پرسشگر محاسبه می‌شود. با توجه به فرمول زیر کوواریانس مربوط به هر پرسشگر مقدار ثابت و برابر واریانس خطای پرسشگر است.

$$\text{cov}(y_{ij}, y_{i'j'}) = \begin{cases} 0 & i \neq i' \\ \sigma_b^2 & i = i', j \neq j' \end{cases}$$

در ذیل کوواریانس میان داده‌های مربوط به یک پرسشگر محاسبه شده است. همان‌طور که انتظار می‌رفت این مقدار ثابت است و از آن می‌توان به‌عنوان واریانس خطای پرسشگر استفاده نمود.

$$\begin{aligned} \text{cov}(y_{ij}, y_{i'j'}) &= \text{cov}(y_{ij}, 0/3 y_{ij} + e_{ij}^{\circ}) = 0/3 \text{var}(y_{ij}) \\ &= 0/3 \text{var}(0/2 \mu_{ij} + e_{ij}^*) = 1/2 \text{var}(\mu_{ij}) + 0/3 \text{var}(e_{ij}^*) \end{aligned}$$

توجه کنید که فرض می‌شود e_{ij}° مستقل از y_{ij} می‌باشد. انحراف معیار μ_{ij} برابر ۳/۲ و انحراف معیار e_{ij}^* برابر ۲ است، با جایگذاری این دو مقدار در عبارت بالا به دست می‌آید

$$\text{cov}(y_{ij}, y_{ij'}) = 1/2(10/24) + 0/3(4) = 13/488 \approx 13/49$$

$$\Rightarrow \sigma_b^2 \approx 3/49 \quad \Rightarrow \sigma_b \approx 3/67.$$

۶- خطای تصادفی (e_{ij}) دارای توزیع نرمال با میانگین صفر و واریانس دل‌خواه می‌باشند. از آن‌جا که در قسمت قبل واریانس خطای پرسشگر به دست آمد، با استفاده از عبارت زیر می‌توان انحراف معیار خطای تصادفی را به دست آورد.

$$e_{ij} = y_{ij} - \mu_{ij} - b_i \quad \Rightarrow \quad \text{var}(e_{ij}) = \text{var}(y_{ij} - \mu_{ij} - b_i)$$

$$\Rightarrow \quad \sigma_e = \sqrt{\text{var}(y_{ij} - \mu_{ij} - b_i)}$$

b_i ها خطای پرسشگر می‌باشند. برای تولید این خطاها تعداد ۱۰۰ داده از توزیع نرمال با میانگین صفر و انحراف معیار به دست آمده از فرمول بالا، تولید می‌شود.

۷- پس از تولید خطای مدل (۱)، می‌توان خطای پرسشگر را به صورت زیر به دست آورد.

$$b_i = y_{ij} - \mu_{ij} - e_{ij}$$

۸- ضریب همبستگی پرسشگر توسط فرمول زیر برآورد می‌شود.

$$\hat{\rho}_y = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_\mu^2 + \hat{\sigma}_b^2 + \hat{\sigma}_e^2}$$

$\hat{\sigma}_\mu^2$ ، $\hat{\sigma}_b^2$ و $\hat{\sigma}_e^2$ به ترتیب برآورهای واریانس‌های مربوط به خطا، مقدرهای واقعی داده‌ها و خطای پرسشگر می‌باشند.

۹- سپس می‌توان واریانس میانگین نمونه‌ای را توسط فرمول زیر برآورد نمود.

$$\text{var}(\hat{y}) = \frac{1}{n} (\hat{\sigma}_\mu^2 + \hat{\sigma}_b^2 + \hat{\sigma}_e^2) [1 + (m-1)\hat{\rho}_y]$$

گام‌های ذکر شده برای تولید داده‌ها برای ۱۰۰۰ نمونه‌ی مستقل به اندازه‌ی ۱۰۰ و برای تعداد ۵، ۱۰ و ۲۰ پرسشگر با بارکارهای متفاوت انجام شده است. همچنین برآورد آماره‌ها به دو صورت یک‌بار با در نظر گرفتن خطای پرسشگر و بار دیگر بدون در نظر گرفتن خطای پرسشگر مورد محاسبه قرار گرفته است، به این ترتیب می‌توان تأثیر اثر پرسشگر را در آماره‌ها مشاهده کرد.

۷- نتایج مطالعه‌ی شبیه‌سازی

در جدول ۱ برآورد میانگین، برآورد واریانس میانگین و بازه‌ی اطمینان به تفکیک بارکارهای متفاوت و به دو روش (با در نظر گرفتن خطای پرسشگر و بدون در نظر گرفتن خطای پرسشگر) با استفاده از ۱۰۰۰ نمونه‌ی مستقل به اندازه‌ی ۱۰۰ از جامعه‌ی شبیه‌سازی شده، آورده شده است.

جدول ۱- میانگین نمونه‌ای، برآورد واریانس میانگین نمونه‌ای و بازه‌ی اطمینان به تفکیک بارکار پرسشگرها

		۲۰ پرسشگر	۱۰ پرسشگر	۵ پرسشگر
با در نظر گرفتن خطای پرسشگر	بارکار هر پرسشگر	۵	۱۰	۲۰
	میانگین	۱	۱/۰۳	۱/۰۲
	واریانس	۱/۴۳	۲/۳۹	۴/۲۵
	بازه‌ی اطمینان ۹۵٪	(۰/۱۶۸/۳۶)	(۰/۱۷۸/۲۸)	(۰/۱۸/۲)
بدون در نظر گرفتن خطای پرسشگر	بارکار هر پرسشگر	۵	۱۰	۲۰
	میانگین	۱	۱/۰۳	۱/۰۲
	واریانس	۰/۰۶۶۲	۰/۰۶۷	۰/۰۶۶۳
	بازه‌ی اطمینان ۹۵٪	(۰/۱۹۸/۰۶)	(۰/۱۹۹/۰۷)	(۰/۱۹۶/۰۴)

در ابتدا تأثیر خطای پرسشگر بر آماره‌ها مورد مقایسه قرار می‌گیرد. همان‌طور که در جدول ۱ مشاهده می‌شود، هنگامی که واریانس با در نظر گرفتن خطای پرسشگر محاسبه می‌شود بزرگ‌تر می‌شود. این نتیجه کاملاً طبیعی است به این دلیل که وجود خطای پرسشگر باعث بزرگ‌تر شدن واریانس خواهد شد و چنان‌چه این خطا در نظر گرفته نشود همچنان که در قسمت پائین جدول ۱ نشان داده شده است واریانس کوچک‌تر می‌شود. برآورد واریانس اخیر هنگامی حاصل می‌شود که به تعداد پرسشنامه، پرسشگر استخدام شود.

برآورد واریانس با کاهش تعداد پرسشگرها و افزایش بارکار آن‌ها بزرگ‌تر می‌شود. همچنین بازه‌ی اطمینان متناظر نیز پهن‌تر می‌شوند. می‌توان نتیجه گرفت افزایش بارکار

پرسشگرها باعث کاهش دقت نتایج و در نتیجه کاهش دقت آماره‌های نمونه‌گیری می‌شود.

۸- خلاصه

مطالعه‌ی شبیه‌سازی انجام شده نتایج زیر را به دست می‌دهد.

- ۱- هنگامی که در یک آمارگیری از پرسشگرها برای گردآوری اطلاعات استفاده می‌شود، اطلاعات به دست آمده تحت تأثیر پرسشگرها قرار می‌گیرند. چنانچه پرسشگر نتواند ارتباط مناسبی با پاسخگو ایجاد نماید، حضور او باعث می‌شود که پاسخگو نتواند به راحتی به پرسش‌ها پاسخ دهد، به خصوص در مورد سؤال‌های حساس پاسخگو هنجارهای اجتماعی و شرایط محیطی را در نظر گرفته و ممکن است پاسخ درستی ارائه ندهد. وقتی پرسشگرها بر پاسخ‌ها اثر می‌گذارند، واریانس آماره‌ها افزایش می‌یابد. اگر بدون توجه به این خطا آماره‌ها محاسبه شوند بدون شک این آماره‌ها کم‌دقت خواهند بود.
- ۲- اندازه‌ی بارکار پرسشگر می‌تواند اثر پرسشگرها بر واریانس آماره‌ها را مینیمم کند. با بالا رفتن بارکار پرسشگرها واریانس افزایش می‌یابد. وقتی بارکار پرسشگرها بیش‌تر می‌شود دقت آن‌ها در به دست آوردن پاسخ‌های درست کمتر می‌شود و در اکثر مواقع آن‌ها به فکر تمام کردن هر چه سریع‌تر کار می‌باشند و بر درستی پاسخ‌های به دست آمده و ثبت صحیح آن‌ها کمتر توجه می‌کنند. در نتیجه، در صورت وجود اعتبار کافی استفاده از پرسشگر بیش‌تر در یک مطالعه‌ی خاص و سپردن تعداد مصاحبه‌های کمتر به هر یک از آن‌ها، روشی برای کاهش اثر پرسشگرها بر برآوردهای خطاهای استاندارد است.

مرجع‌ها

- [1] Biemer, P. P.; Groves, M. R.; Lybey, E. L.; Mathiowetz, A. N.; Sudman, S. (1991). *Measurement Errors in Surveys*, New York, John Wiley and Sons.
- [2] Fowler, F. J.; Mangione, T. W. (1990). *Standardized Survey Interviewing*

- Minimizing Interviewer Related Error*. Beverlly Hills, CA, Sage Publications.
- [3] Groves, R. M.; Fowler, F. J.; Micck, P.; Couper, J. M.; Lepkowski, E. S.; Tourangean, G. (2004). *Survey Methodology*. Wiley Interscience.
- [4] Hansen, M.; Hurwitz, W.; Bershada, M. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*, **38**, 359-374.
- [5] Hansen, M. H.; Hurwitz, W. N.; Madow, W. G. (1953). *Survey Methods and Theory*, Vol. I: Methods and Applications, Vol. II: Theory, New York, John Wiley and Sons.

معصومه موحدی

فوق لیسانس آمار

تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشگاه علامه طباطبایی، دانشکده اقتصاد، گروه آمار.
پامنگار: msmh_movahedi@yahoo.com

حمیدرضا نوابپور

دانشیار آمار

تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشگاه علامه طباطبایی، دانشکده اقتصاد، گروه آمار.
پامنگار: h.navvabpour@srtc.ac.ir

