مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۶، شماره‌ی ۱، بهار و تابستان ۱۳۹۴، صص ۵۱-۶۴

مقایسه‌ی براورد میانگین جامعه بر اساس روش‌های جانهی نسبتی در آمارگیری‌های نمونه‌ای

آسیه رشیدی‌نژاد

مرکز آمار ایران

چکیده: وجود داده‌های گمشده در بیش‌تر آمارگیری‌ها امری اجتناب‌ناپذیر است. اما در تمامی طرح‌های آماری قبل از اجرای آن، مطمئناً راهکار مناسبی برای برخورد با این مشکل در نظر گرفته می‌شود. از جمله مواردی که برای برخورد با داده‌های گمشده در نظر گرفته می‌شود، جانهی و وزن‌دهی است که هر کدام مزیت‌ها و مشکلاتی دارند، اما استفاده از مجموعه داده‌ای که هیچ‌یک از این دو روش را در تحلیل به کار نگیرد، منجر به اریبی براوردها می‌شود و در نتیجه تحلیل‌های اشتباهی را به همراه خواهد داشت. در این مقاله به بررسی براوردگری که برای میانگین جامعه در حضور بی‌پاسخی پیشنهاد شده [۱۲] و از براوردگر والش به‌دست آمده است [۱۶]، پرداخته می‌شود. همچنین نشان داده می‌شود که براوردگر میانگین به‌دست آمده از روش پیشنهادی بهتر از براوردگر حاصل از روش جانهی نسبتی و براوردگر TSS (Toutenberg, Srivastava and Shalabh) است و میانگین مربع خطای براوردگر حاصل کم‌تر از براوردگر به‌دست آمده بر اساس روش جانهی نسبتی است.

واژگان کلیدی: مکانیسم گمشدگی؛ جانهی؛ داده‌های گمشده؛ براوردگر TSS؛ براوردگر والش؛ روش جانهی نسبتی.

۱- مقدمه

وقتی که یک مجموعه داده شامل مقادیر گمشده است، تحلیل‌گر باید تصمیم بگیرد که چطور با آن برخورد کند. در گذشته رایج‌ترین تصمیم استفاده از تحلیل داده‌های کامل بود و فقط تحلیل بر اساس داده‌های کامل صورت می‌گرفت و واحدهایی با داده‌های گمشده

دریافت: ۱۳۹۳/۱۰/۱۵، پذیرش: ۱۳۹۴/۱۰/۳۰.

برای هر متغیر در تحلیل، وجود نداشت. این روش شامل محدودیت‌هایی است؛ از جمله اندازه‌ی نمونه‌ی کم‌تر، که منجر به کاهش شدید توان و همچنین منجر به اریبی نتایج می‌شود. در مقابل جانهی داده‌های گمشده روشی کاراست. برای مثال [۷]، [۸] و [۱۱] در مواردی از جانهی استفاده کردند. اما باید به این نکته توجه کرد در حالتی که داده‌های گمشده توسط یکی از روش‌های جانهی براورد شود و وقتی که براوردگرهای معمولی برای براورد پارامترهای جامعه استفاده شود استنتاج‌های اریبی را به همراه خواهد داشت. در این مقاله براوردگرهای میانگین جامعه در حالتی که روش‌های جانهی نسبتی استفاده می‌شود، بررسی شده است. بر اساس [۹] و [۱۵] فرض شده است که بعضی واحدها در نمونه از پاسخگویی اجتناب می‌کنند؛ لذا ویژگی مورد مطالعه‌ی آن‌ها موجود نمی‌باشد، در حالی که متغیرهای کمکی تمامی مشاهدات در نمونه موجود است؛ بنا بر این گمشدگی تنها در ویژگی مورد مطالعه رخ می‌دهد. روش جانهی نسبتی عموماً در فرایندهای آمارگیری‌های نمونه‌ای استفاده می‌شود. در این مقاله ابتدا به بررسی مکانیسم‌های گمشدگی بر اساس رده‌بندی [۱۰] پرداخته می‌شود، سپس روش جدید [۱۲] که بر اساس مقادیر کمکی داده‌های گمشده را جانهی می‌کند، بیان شده و در آخر مطالعه‌ی شبیه‌سازی ارایه شده است.

۲- مکانیسم‌های گمشدگی بر اساس رده‌بندی روبین

سه نوع مکانیسم گمشدگی وجود دارد. برای توضیح انواع مکانیسم‌های گمشدگی، متغیر Y را در نظر می‌گیریم که به دو بخش (Y_{obs}, Y_{mis}) تقسیم شده است و همان‌طور که از نام آن مشخص است Y_{obs} نشان دهنده‌ی بخش مشاهده شده و $\mathrm{Y_{mis}}$ نمایانگر بخش گمشده‌ی متغیر Y است.

گمشدگی کاملاً تصادفی (Missing Completely At Random) (MCAR): در این حالت گمشدگی در متغیر Y (یعنی Y_{mis}) به هیچ‌یک از متغیرهای دیگر و حتی خود متغیر Y بستگی ندارد به عبارت دیگر احتمال این‌که یک مشاهده گمشده باشد به متغیرهایی که مشاهده شده و متغیر Y_{mis} که گمشده است، وابسته نیست. بنا بر این در این حالت تجزیه و تحلیل بر اساس داده‌های مشاهده‌شده مناسب است و براوردهای نااریبی از پارامترها نتیجه خواهد داد. به عنوان مثال شرکت‌کنندگان در یک آمارگیری با

انداختن سکه تصمیم می‌گیرند که در آمارگیری شرکت کنند یا خیر. تشخیص این نوع گمشدگی با آزمون امکان‌پذیر است [۲]. کارشناسان روش‌هایی را برای آزمون مکانیسم گمشدگی کاملاً تصادفی ارایه داده‌اند به عنوان مثال مقایسه‌های آزمون t یک متغیره و آزمون گمشدگی کاملاً تصادفی [۶].

گمشدگی تصادفی (Missing At Random) (MAR): در این مورد، گمشدگی در متغیر Y، به بخش گمشده‌ی متغیر، Y_{mis} وابسته نیست و فقط بستگی به متغیرهای مشاهده‌شده دارد؛ به عبارت دیگر احتمال این‌که یک مشاهده گمشده باشد تنها به اندازه‌های مشاهده‌شده وابسته است. به عنوان مثال، مردان نسبت به زنان تمایل کم‌تری برای شرکت در آمارگیری مربوط به بررسی میزان افسردگی دارند. برای تشخیص مکانیسم گمشدگی تصادفی الگوریتمی ارایه شده است [۳].

گمشدگی غیرتصادفی (Missing Not At Random) (MNAR): در این حالت گمشدگی در متغیر Y بستگی به مقادیر گمشده‌ی متغیر Y (یعنی Y_{mis}) دارد، حتی بعد از این‌که روی سایر متغیرها مشروط شده باشد. به عبارت دیگر در این حالت احتمال گمشدگی یک مشاهده وابسته به مقدار گمشده است. به عنوان مثال، افرادی با درامد زیاد تمایلی به گزارش درامد خود ندارند. ذکر این نکته حائز اهمیت است که هیچ تجزیه و تحلیلی از داده‌های دارای گمشدگی غیر تصادفی، بدون انجام تحلیل حساسیت کامل نیست[۴].

۳- جانهی نسبتی

فرض کنید $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ میانگین جامعه متناهی $\Omega = \{1, 2, \dots, i, \dots N\}$ باشد. نمونه‌ی تصادفی بدون جایگذاری، به اندازه‌ی n برای براورد $\bar{Y}$ انتخاب می‌شود و مجموعه‌ی نمونه با S نشان داده می‌شود. همچنین r تعداد واحدهای پاسخگو از n واحد نمونه است و مجموعه‌ی واحدهای پاسخگو را با R_p نشان داده و مجموعه واحدهای بی‌پاسخ را با R_p^c نمایش می‌دهیم. برای هر واحد $i \in R_p^c$، مقدار y_i مشاهده شده است. اما برای هر $i \in R_p^c$ مقدار y_i گمشده و نوع مکانیسم گمشدگی MCAR فرض شده است. در این‌جا جانهی با استفاده از مقادیر کمکی انجام می‌شود. ابتدا متغیر کمکی X را

در نظر می‌گیریم به طوری که x_i مقدار X برای واحد i ام، معلوم و مثبت باشد، به عبارت دیگر داده‌های $X_s = \{x_i; i \in S\}$ معلوم هستند. عبارات زیر را تعریف می‌کنیم:

$$\varepsilon = \frac{\bar{y}_r}{\bar{Y}} - ۱, \qquad \delta = \frac{\bar{x}_r}{\bar{X}} - ۱, \qquad \eta = \frac{\bar{x}_n}{\bar{X}} - ۱.$$

بر اساس مفاهیم نمونه‌گیری دو مرحله‌ای [۹] و [۱] عبارات زیر نتیجه می‌شود:

$$E(\varepsilon) = E(\delta) = E(\eta) = ۰$$

و

$$E\left(\varepsilon^۲\right) = \left(\frac{۱}{r} - \frac{۱}{N}\right) C_y^۲, \qquad E\left(\eta^۲\right) = \left(\frac{۱}{n} - \frac{۱}{N}\right) C_x^۲,$$

$$E\left(\delta^۲\right) = \left(\frac{۱}{r} - \frac{۱}{N}\right) C_x^۲, \qquad E(\varepsilon\delta) = \left(\frac{۱}{r} - \frac{۱}{N}\right) \rho C_y C_x,$$

$$E(\delta\eta) = \left(\frac{۱}{n} - \frac{۱}{N}\right) C_x^۲, \qquad E(\varepsilon\eta) = \left(\frac{۱}{r} - \frac{۱}{N}\right) \rho C_y C_x,$$

که $C_x^۲ = \frac{S_x^۲}{\bar{X}^۲}$، $C_y^۲ = \frac{S_y^۲}{\bar{Y}^۲}$، $\rho = \frac{S_{xy}}{S_x S_y}$. نمادگذاری از مرجع [۵] استفاده شده است. اگر واحد iام نیاز به جانهی داشته باشد مقدار $\hat{b}x_i$ جایگزین می‌شود که $\hat{b} = \sum_{i=۱}^{r} y_i / \sum_{i=۱}^{r} x_i$. داده‌ها بعد از جانهی به صورت زیر است:

$$y_i' = \begin{cases} y_i & if \;\; i \epsilon R_p \\ \hat{b}x_i & if \;\; i \in R_p^c. \end{cases}$$

این روش جانهی، جانهی نسبتی نامیده می‌شود و تحت این روش براوردگر نقطه‌ای میانگین جامعه، $\bar{y}_n = \frac{۱}{n}\sum_{i=۱}^{n} y_i'$، به‌صورت زیر تبدیل می‌شود.

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}, \qquad (۱)$$

که $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$، $\bar{x}_r = \frac{1}{r}\sum_{i=1}^{r} x_i$ و $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$. در بخش بعدی به مطالعه‌ی خواص براوردگر نسبتی و براوردگر t_4، پرداخته می‌شود.

۱-۳- براوردگر نسبتی

قضیه‌ی ۱. اریبی براوردگر نسبتی $\bar{y}_{RAT}$ عبارت است از:

$$Bias\,(\bar{y}_{RAT}) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(C_x^2 - \rho C_x C_y\right). \qquad (۲)$$

اثبات: براوردگر $\bar{y}_{RAT}$ در معادله‌ی (۱) بر حسب η و δ و ε به‌صورت زیر نوشته می‌شود:

$$\bar{y}_{RAT} = \bar{Y}[1 + \varepsilon + \eta - \delta + \delta^2 + \varepsilon\eta - \varepsilon\delta - \delta\eta + O(\varepsilon^2)] \qquad (۳)$$

که عبارت $O(\varepsilon^2)$ تقریب مرتبه‌ی دوم را نشان می‌دهد. با گرفتن امید ریاضی از دو طرف رابطه‌ی (۳) و محاسبه‌ی انحراف آن از میانگین واقعی، رابطه‌ی (۲) به‌دست می‌آید.

قضیه‌ی ۲. میانگین مربع خطای براوردگر نسبتی، $\bar{y}_{RAT}$، به‌صورت زیر به‌دست می‌آید:

$$MSE(\bar{y}_{RAT}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{xy}\} \qquad (۴)$$

که $R = \frac{\bar{Y}}{\bar{X}}$.

اثبات:

$$\begin{aligned} MSE(\bar{y}_{RAT}) &= E[\bar{Y}_{RAT} - \bar{Y}]^2 \\ &= E[\varepsilon + \eta - \delta]^2 \\ &= E[\varepsilon^2 + \eta^2 + \delta^2 + 2\varepsilon\eta - 2\varepsilon\delta - 2\delta\eta]. \end{aligned}$$

با گذاشتن مقدار مورد انتظار برای هر عبارت، رابطه‌ی (۴) ثابت می‌شود.

با مقایسه‌ی روش جانهی نسبتی با روش جانهی میانگینی، $\bar{y}_m = \frac{1}{r}\sum_{i=1}^{r} y_i$، که از میانگین مقادیر مشاهده‌شده استفاده می‌شود و واریانس آن به‌صورت زیر است:

$$V(\bar{y}_m) = \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 \qquad (5)$$

و S_y^2 واریانس مقادیر مشاهده‌شده است، دیده می‌شود که جانهی با استفاده از روش نسبتی بهتر از جانهی با روش میانگینی است اگر شرط $\beta > \frac{R}{2}$ که $\beta = \frac{S_{xy}}{S_x^2}$ برقرار باشد.

۲-۳- براوردگر t_4

براوردگر t_4، [۱۴]، با نمادگذاری تعریف‌شده، به‌صورت زیر نوشته می‌شود:

$$t_4 = \bar{y}_r \left[\frac{r^2\bar{x}_r + (n+r)(n-r)\{(n\bar{x}_n - r\bar{x}_r)/(n-r)\}}{rn\bar{x}_r + n(n-r)\{(n\bar{x}_n - r\bar{x}_r)/(n-r)\}}\right]$$

$$= \bar{y}_r \left[\frac{r^2\bar{x}_r + (n+r)(n\bar{x}_n - r\bar{x}_r)}{rn\bar{x}_r + n(n\bar{x}_n - r\bar{x}_r)}\right]$$

$$= \bar{y}_r \left[1 + \frac{r}{n\bar{x}_n}(\bar{x}_n - \bar{x}_r)\right].$$

بنابر این داریم:

$$t_4 = \bar{y}_r + \left(\frac{r\bar{y}_r}{n\bar{x}_n}\right)(\bar{x}_n - \bar{x}_r)$$

براوردگر t_4 در حقیقت یک نوع براوردگر رگرسیونی است که ضریب رگرسیونی آن توسط $\frac{r\bar{y}_r}{n\bar{x}_n}$ به‌دست می‌آید. میانگین مربع خطای براوردگر t_4 به‌صورت زیر به‌دست می‌آید:

$$MSE\left(t_4\right) = E\left[t_4 - \bar{Y}\right]^2 \approx E\left[\bar{Y}\varepsilon + \frac{r}{n}\bar{Y}(\eta - \delta)(1+\varepsilon)(1 - \eta + \cdots)\right]^2$$

$$\approx \bar{Y}^2 E\left[\varepsilon^2 + \left(\frac{r}{n}\right)^2\left(\eta^2 + \delta^2 - 2\delta\eta\right) + 2\left(\frac{r}{n}\right)(\varepsilon\eta - \varepsilon\delta)\right]$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[S_y^2 + \left(\frac{r}{n}\right)^2 R^2 S_x^2 - 2\left(\frac{r}{n}\right) R S_{xy}\right]$$

توجه شود که براوردگر $\bar{y}_{RAT}$ کاراتر از براوردگر t_4است اگر

$$MSE(\bar{y}_{RAT}) < MSE(t_4) \qquad (6)$$

نامساوی (۶) برقرار می‌شود اگر $\beta > \frac{R}{2}(\frac{r}{n} + 1)$. در بخش ۴ روش جانهی جدید بررسی می‌شود.

۴- روش جدید جانهی

در روش جدید مطرح‌شده [۱۲]، داده‌ها بعد از جانهی به‌صورت زیر است:

$$y_i' = \begin{cases} y_i & i \in R_p \\ \bar{y}_r \left[\dfrac{(n-r)\bar{x}_n + \alpha r(\bar{x}_n - \bar{x}_r)}{\alpha \bar{x}_r + (1-\alpha)\bar{x}_n}\right] \dfrac{x_i}{\sum_{i \in R_p^C} x_i} & i \in R_p^C \end{cases}$$

که α مقدار ثابتی است و نیاز است طوری انتخاب شود که واریانس براوردگر میانگین جامعه مینیمم شود. براوردگر نقطه‌ای، $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i'$، از میانگین جامعه $\bar{Y}$ تحت روش جدید جانهی به‌صورت زیر است:

$$\bar{y}_p = \frac{\bar{y}_r \bar{x}_n}{\alpha \bar{x}_r + (1-\alpha)\bar{x}_n} \qquad (7)$$

اگر در رابطه‌ی (۷) مقدار $\alpha = 0$ باشد پس $\bar{y}_p = \bar{y}_r$ و اگر $\alpha = 1$ باشد آن‌گاه $\bar{y}_p = \bar{y}_{RAT}$. لذا روش مطرح‌شده‌ی جانهی به نوعی سازش بین روش‌های جانهی نسبتی و میانگینی است. براوردگر رابطه‌ی (۷) مشابه براوردگر شناخته‌شده‌ی میانگین جامعه است که رابطه‌ی (۸) آن را نشان می‌دهد [۱۶]. همان‌طور که ملاحظه می‌شود در براوردگر ارایه‌شده در رابطه‌ی (۷) از مقادیر مشاهده‌شده استفاده شده است در حالی‌که در $\bar{y}_w$ از میانگین‌های نمونه‌ی nتایی استفاده می‌شود.

$$\bar{y}_w = \frac{\bar{y}_n \bar{X}}{\alpha \bar{x}_n + (1-\alpha)\bar{X}} \tag{8}$$

قضیه‌ی ۳. اریبی براوردگر مطرح شده $\bar{y}_p$ در رابطه‌ی (۷) به‌صورت زیر است:

$$\begin{aligned} Bias\left(\bar{y}_p\right) = \bar{Y}\Bigg[&\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_y C_x + \alpha^2\left(\frac{1}{r}-\frac{1}{N}\right)C_x^2 \\ &+(1-\alpha)^2\left(\frac{1}{n}-\frac{1}{N}\right)C_x^2 \\ &-\alpha\left\{\left(\frac{1}{r}-\frac{1}{N}\right)\rho C_y C_x + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right\} \\ &+2\alpha(1-\alpha)\left(\frac{1}{n}-\frac{1}{N}\right)C_x^2 \\ &-(1-\alpha)\left\{\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_y C_x + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right\}\Bigg] \end{aligned} \tag{9}$$

اثبات: براوردگر $\bar{y}_p$ بر حسب ε، δ و η به‌صورت زیر نوشته می‌شود:

$$\begin{aligned} \bar{y}_p = \bar{Y}\Big[&1+\varepsilon-\alpha(\delta-\eta)+\varepsilon\eta+\alpha^2\delta^2+\left(1-\alpha\right)^2\eta^2-\alpha(\varepsilon\delta+\delta\eta) \\ &-\left(1-\alpha\right)\left(\varepsilon\eta+\eta^2\right)+2\alpha(1-\alpha)\delta\eta\Big] \end{aligned} \tag{10}$$

با گرفتن امید ریاضی از دو طرف رابطه‌ی (۱۰) و محاسبه‌ی انحراف آن از میانگین واقعی، رابطه‌ی (۹) به‌دست می‌آید.

قضیه‌ی ۴. مقدار مینیمم میانگین مربع خطای براوردگر مطرح‌شده $\bar{y}_p$، که مقدار بهینه‌ی α برابر $\rho\frac{C_y}{C_x}$ است، به‌صورت زیر است:

$$\text{Min } MSE\left(\bar{y}_p\right) = MSE(\bar{y}_{RAT}) - \left(\frac{1}{r}-\frac{1}{n}\right)S_x^2\left(\frac{S_{xy}}{S_x^2}-\frac{\bar{Y}}{\bar{X}}\right)^2 \tag{11}$$

عبارت (۱۱) نشان می‌دهد که روش مطرح‌شده‌ی جانهی بهتر از روش جانهی نسبی است به شرط این که $\frac{S_{xy}}{S_x^2} \neq \frac{\bar{Y}}{\bar{X}}$. هر چه تفاوت مقدار بین $\frac{S_{xy}}{S_x^2}$ و $\frac{\bar{Y}}{\bar{X}}$ بیش‌تر باشد، کارایی روش

مطرح‌شده جانهی بیش‌تر می‌شود. در این روش، یافته‌ها مشابه نتایج بسط داده‌شده در مرجع [۱۳] است. لازم است به این نکته اشاره شود که $MSE(\bar{y}_p) < MSE(t_۴)$ مگر این‌که $r = n$ یا $\frac{S_{xy}}{S_x^۲}=(\frac{r}{n})R$. بنابر این $\bar{y}_p = t_۴$ است، اگر

$$\alpha = \frac{r\bar{x}_n}{[(n+r)\bar{x}_n + r\bar{x}_r]}$$

۱-۴- اطلاعات کمکی چندگانه

فرض کنید $X = (x_{ij})_{n\times p}$، $(i = ۱, \dots, n,\ j = ۱, \dots, p)$ یک ماتریس $n \times p$ از بردارهای کمکی باشد. فرض شده که اطلاعات متغیر کمکی کامل است. اما پاسخ‌ها تنها برای متغیر مورد مطالعه گمشده‌اند. در این وضعیت روش جدید جانهی توسط عبارت زیر داده شده است:

$$y_i' = \begin{cases} y_i & i \in R_p \\ \bar{y}_r \left[\dfrac{n\prod_{j=۱}^{p}\bar{x}_{nj} - r\prod_{j=۱}^{p}\{\alpha_j\bar{x}_{rj}+(۱-\alpha_j)\bar{x}_{nj}\}}{\prod_{j=۱}^{p}\{\alpha_j\bar{x}_{rj}+(۱-\alpha_j)\bar{x}_{nj}\}}\right]\dfrac{\sum_{j=۱}^{p}x_{ij}}{\sum_{i\in R_p^C}\sum_{j=۱}^{p}x_{ij}} & i \in R_p^C \end{cases}$$

تحت این روش جانهی، براوردگر نقطه‌ای میانگین جامعه به‌صورت زیر است:

$$\bar{y}_{mult} = \bar{y}_r \prod_{j=۱}^{p} \frac{\bar{x}_{nj}}{\alpha_j\bar{x}_{rj} + (۱-\alpha_j)\bar{x}_{nj}}$$

که تعمیمی از براوردگر والش برای اطلاعات چندگانه متغیر کمکی است. کم‌ترین واریانس براوردگر $\bar{y}_{mult}$ عبارت است از:

$$v(\bar{y}_{mult}) = \left(\frac{۱}{n} - \frac{۱}{N}\right)S_y^۲ + \left(\frac{۱}{r} - \frac{۱}{n}\right)S_y^۲(۱ - R^۲_{y,x_۱,\dots,x_p})$$

که $R^۲_{y,x_۱,\dots,x_p}$ ضریب همبستگی چندگانه را نشان می‌دهد.

۵- مطالعه‌ی شبیه‌سازی

به‌منظور بررسی کارایی نسبی روش جانهی مطرح‌شده با روش جانهی نسبتی و براوردگر $t_۴$ روابط زیر محاسبه می‌شوند:

$$RE(\bar{y}_p, \bar{y}_{RAT}) = \frac{MSE(\bar{y}_{RAT})}{MSE(\bar{y}_p)} \times ۱۰۰\% = RE_{RAT/p}$$

$$RE\left(\bar{y}_p, t_۴\right) = \frac{MSE(t_۴)}{MSE(\bar{y}_p)} \times ۱۰۰\% = RE_{t_۴/p}$$

جامعه‌ای شامل $N = ۱۰۰۰$ واحد آماری را از توزیع نرمال دو متغیره، $(X,Y) \sim N(۰, ۰, ۱, ۱, ۰/۸)$، شبیه‌سازی شده است. فرض می‌شود که نمونه‌ی اولیه‌ای به اندازه‌ی n از جامعه‌ای به اندازه‌ی N انتخاب شود به طوری که $n = f_s N$ و f_s نسبت نمونه‌گیری متناهی است که بین ۵% تا ۴۰% تغییر می‌کند. توجه کنید که در بیش‌تر آمارگیری‌های واقعی حدود ۲% تا ۵% جامعه در نمونه حضور دارند. همچنین $r = [f_r N]$ که $\lfloor\ \rfloor$ نماد جز صحیح است و f_r کسر نرخ پاسخ را نشان می‌دهد. چون در مطالعاتی مشابه نرخ پاسخ حدود ۲۰% را مورد بررسی قرار گرفت [۹]، در این مطالعه به طور مشابه نرخ پاسخ از۱۰% تا ۳۰% در نظر گرفته می‌شود. لازم به ذکر است که گمشدگی تنها در متغیر Y رخ می‌دهد. کارایی نسبی برای مقادیر متفاوت f_s و f_r در جدول ۱ نشان داده شده است.

جدول ۱- مقایسه‌ی کارایی نسبی براورد میانگین جامعه بر اساس روش مطرح‌شده با دو روش براورد نسبتی و براورد $t_۴$

اندازه‌ی نمونه	r	f_s	f_r	$RE_{RAT/p}$	$RE_{t_۴/p}$
	۵		۰/۱۰	۱۱۲/۷۲	۱۴۷/۱۴
	۷		۰/۱۵	۱۱۲/۱۹	۱۴۱/۸۵
۵۰	۱۰	۰/۰۵	۰/۲۰	۱۱۰/۲۰	۱۳۲/۶۳
	۱۲		۰/۲۵	۱۰۹/۴۳	۱۲۶/۵۰
	۱۵		۰/۳۰	۱۰۸/۳۵	۱۱۷/۳۷

ادامه‌ی جدول ۱- مقایسه‌ی کارایی نسبی براورد میانگین جامعه بر اساس روش مطرح‌شده با دو روش براورد نسبتی و براورد $t_۴$

اندازه‌ی نمونه	r	f_s	f_r	$RE_{RAT/p}$	$RE_{t_۴/p}$
	۱۰		۰/۱۰	۱۱۲/۷۴	۱۴۷/۳۱
	۱۵		۰/۱۵	۱۱۲/۲۳	۱۴۲/۰۵
۱۰۰	۲۰	۰/۱۰	۰/۲۰	۱۱۰/۲۶	۱۳۲/۸۲
	۲۵		۰/۲۵	۱۰۹/۴۷	۱۲۶/۶۸
	۳۰		۰/۳۰	۱۰۸/۴۰	۱۱۷/۵۱
	۱۵		۰/۱۰	۱۱۲/۷۸	۱۴۷/۵۹
	۲۲		۰/۱۵	۱۱۲/۲۸	۱۴۲/۲۶
۱۵۰	۳۰	۰/۱۵	۰/۲۰	۱۱۰/۳۲	۱۳۲/۰۴
	۳۷		۰/۲۵	۱۰۹/۵۱	۱۲۶/۸۴
	۴۵		۰/۳۰	۱۰۸/۴۲	۱۱۷/۶۶
	۲۰		۰/۱۰	۱۱۲/۸۲	۱۴۷/۷۶
	۳۰		۰/۱۵	۱۱۲/۳۵	۱۴۲/۴۴
۲۰۰	۴۰	۰/۲۰	۰/۲۰	۱۱۰/۳۷	۱۳۲/۵۹
	۵۰		۰/۲۵	۱۰۹/۵۴	۱۲۶/۰۲
	۶۰		۰/۳۰	۱۰۸/۴۸	۱۱۷/۸۴
	۲۵		۰/۱۰	۱۱۲/۸۵	۱۴۷/۹۴
	۳۷		۰/۱۵	۱۱۲/۴۰	۱۴۲/۶۳
۲۵۰	۵۰	۰/۲۵	۰/۲۰	۱۱۰/۴۲	۱۳۲/۷۷
	۶۲		۰/۲۵	۱۰۹/۵۷	۱۲۶/۲۱
	۷۵		۰/۳۰	۱۰۸/۵۲	۱۱۸/۰۱
	۳۰		۰/۱۰	۱۱۲/۸۹	۱۴۸/۱۳
	۴۵		۰/۱۵	۱۱۲/۴۳	۱۴۲/۹۱
۳۰۰	۶۰	۰/۳۰	۰/۲۰	۱۱۰/۴۷	۱۳۲/۹۴
	۷۵		۰/۲۵	۱۰۹/۶۲	۱۲۶/۳۹
	۹۰		۰/۳۰	۱۰۸/۵۸	۱۱۸/۱۷
	۳۵		۰/۱۰	۱۱۲/۹۲	۱۴۸/۳۱
	۵۲		۰/۱۵	۱۱۲/۴۷	۱۴۳/۰۸
۳۵۰	۷۰	۰/۳۵	۰/۲۰	۱۱۰/۵۱	۱۳۳/۱۳
	۸۷		۰/۲۵	۱۰۹/۶۶	۱۲۶/۵۷
	۱۰۵		۰/۳۰	۱۰۸/۶۳	۱۱۸/۳۴

ادامه‌ی جدول ۱- مقایسه‌ی کارایی نسبی براورد میانگین جامعه بر اساس روش مطرح‌شده با دو روش براورد نسبتی و براورد $t_۴$

$RE_{t_۴/p}$	$RE_{RAT/p}$	f_r	f_s	r	اندازه‌ی نمونه
۱۴۸/۴۹	۱۱۲/۹۷	۰/۱۰		۴۰	
۱۴۳/۲۵	۱۱۲/۵۱	۰/۱۵		۶۰	
۱۳۳/۳۳	۱۱۰/۵۵	۰/۲۰	۰/۴۰	۸۰	۴۰۰
۱۲۶/۷۴	۱۰۹/۷۲	۰/۲۵		۱۰۰	
۱۱۸/۴۹	۱۰۸/۶۷	۰/۳۰		۱۲۰	

با ملاحظه‌ی کارایی نسبی جدول ۱، براورد میانگین جامعه بر اساس روش جانهی مطرح‌شده نسبت به براورد میانگین جامعه از روش براورد جانهی نسبتی و براورد $t_۴$ در اندازه‌های نمونه‌ی ۵۰ ، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰، ۳۰۰، ۳۵۰ و ۴۰۰ و تمامی حالت‌های نرخ پاسخ، f_r، مشخص می‌شود که براورد میانگین جامعه از روش جانهی مطرح‌شده بهتر از براورد میانگین جامعه با استفاده از روش‌های نسبتی و $t_۴$ است. همچنین هر چه میزان نرخ پاسخ کم‌تر باشد، کارایی نسبی روش مطرح‌شده بهتر از دو روش دیگر است. علاوه بر این هر اندازه تعداد نمونه بیش‌تر باشد کارایی نسبی براورد میانگین جامعه با روش جدید بهتر از کارایی نسبی براورد میانگین جامعه بر اساس دو روش نسبتی و $t_۴$ است. بنا بر این انتخاب براوردگر در آمارگیری‌هایی که برای جبران بی‌پاسخی از جانهی استفاده می‌کنند مهم است و استفاده از براوردگرهای معمولی مناسب نیست.

مرجع‌ها

[1] Arnab, R. and Singh, S. (2006). On the Estimation of Variance from Missing Data, *Metron*, LXIV **2**, 161–171.

[2] Enders, C.K. (2010). *Applied Missing Data Analysis*. Guilford Press, New York and London.

[3] Fleiss, J.L., Levin, B. and Paik, M.C. (2002). *Statistical Methods for Rates and Proportions*. 3rd ed. John Wiley & Sons, New York.

[4] Ibrahim, J.G., Chen, M.H. and Lipsitz, S.R. (2008). Bayesian Methods for Generalized Linear Models with Covariates Missing at Random, *Canadian Journal of Statistics*. **30**, 55–78.

[5] Lee, H., Rancourt, E. and Särndal, C.E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values, *J. Official Statist*. **10**, 231–243.

[6] Little, R.J.A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J. Am. Stat. Assoc.*, **83**, 1198–1202.

[7] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, John Wiley, New York.

[8] Rao, C.R. and Toutenburg, H. (1995). *Linear Models Least Squares and Alternatives*, Springer, New York.

[9] Rao, J.N.K. and Sitter, R.R. (1995). Variance Estimation under Two Phase Sampling with Application to Imputation for Missing Data, *Biometrika*, **82**, 453–460.

[10] Rubin, D.B. (1976). Inference and Missing Data, *Biometrika*, **63**, 581-592.

[11] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Sample Surveys*, John Wiley, New York.

[12] Singh, S. (2009). A New Method of Imputation in Survey Sampling, *Statistics*, **43**, 499–511.

[13] Singh, S. and Deo, B. (2003). Imputation by Power Transformation, *Statist*. Pap. 555–579.

[14] Toutenberg, H., Srivastava, V.K. and Shalabh, (2008). Amputation versus Imputation of Missing Values through Ratio Method in Sample Surveys, *Statist*. **49**, 237–247.

[15] Tracy, D.S. and Osahan, S.S. (1994). Random Non_response on Study Variable versus on Study as Well as Auxiliary Variables, *Statistica*, **54**, 163–168.

[16] Walsh, J.E. (1968). Generalization of Ratio Estimator for Population, *Sankhyä A*, **32**, 99–103.

آسیه رشیدی‌نژاد
فوق لیسانس آمار
تهران، خیابان فاطمی، نبش رهی‌معیری، مرکز آمار ایران.
رایانشانی: a_rashidinezhad@sci.org.ir