گزیده‌مطالب آماری
سال۲۰، شماره‌ی ۲، پاییز و زمستان ۱۳۸۸، صص ۲۱۱-۲۲۷

براوردیابی پارامترهای مجهول در نمونه‌گیری چرخشی

مینا توحیدی* و محمدرضا نمازی راد

دانشگاه شیراز

چکیده: در این مقاله ضمن معرفی آمارگیری چرخشی به‌عنوان یك روش کارا در مقایسه با آمارگیری پانلی و آمارگیری از واحدهای مستقل، با استفاده از مدل‌های خطی تعمیم‌یافته یك روش جدید برای براوردیابی تمامی ترکیبات خطی از میانگین‌های دوره‌های مختلف (از جمله اختلاف میانگین‌ها و مجموع میانگین‌ها در دوره‌های مختلف) ارائه می‌شود. سپس، به کمک یك شبیه‌سازی مونت کارلو، عملکرد تجربی روش معرفی‌شده برای دو مقطع متوالی آمارگیری مورد ارزیابی قرار می‌گیرد. نتایج به دست آمده نشان می‌دهند با استفاده از این روش، براوردیابی تغییرات پارامتر در دو مقطع متوالی آمارگیری چرخشی با دقت بالاتری امکان‌پذیر می‌شوند. همچنین، با استفاده از نتایج به دست آمده می‌توان نتیجه گرفت که با افزایش ضریب تداخل، کارایی براوردگر جدید در دو مقطع متوالی آمارگیری چرخشی نیز افزایش می‌یابد.

۱. مقدمه

با بررسی اجمالی طرح‌های مختلف نمونه‌گیری، مشاهده می‌شود که بسیاری از این نمونه‌گیری‌ها، در مقاطع کاملاً معین (معمولاً به‌صورت ماهانه، فصلی، سالانه و یا دو سال یك‌بار) تکرار می‌شوند. آمارگیری‌های مکرر[۱] علاوه بر این‌که از وضعیت جامعه‌ی مورد مطالعه در زمان جاری اطلاعاتی به‌دست می‌دهند، امکان شناخت عوامل تأثیرگذار بر جامعه‌ی مورد مطالعه در طی زمان را نیز فراهم می‌کنند.

الگوی ترکیب در نمونه‌گیری مکرر ممکن است به‌گونه‌ای باشد که در هر زمان:

واژگان کلیدی: آمارگیری مکرر؛ آمارگیری چرخشی؛ آمارگیری پانلی؛ آمارگیری از واحدهای مستقل؛ شبیه‌سازی مونت کارلو؛ کارایی؛ مدل‌های خطی تعمیم‌یافته.
دریافت: ۱۳۸۸/۲/۱۵، پذیرش: ۱۳۸۸/۱۰/۲۰
* نویسنده‌ی عهده‌دار مکاتبات

- نمونه شامل همان مجموعه‌ی واحدهایی شود که در دوره‌ی قبل، نمونه‌گیری شده‌اند (آمارگیری پانلی ۲).
- نمونه شامل مجموعه‌ی واحدهایی کاملاً متفاوت با واحدهای نمونه‌گیری‌شده در دوره‌ی قبل باشد (آمارگیری از واحدهای مستقل).
- نمونه، مجموعه‌ای از واحدهای نمونه‌گیری‌شده در دوره‌ی قبل (واحدهای جور شده) و مجموعه‌ای از واحدهایی که در دوره‌ی قبل، نمونه‌گیری نشده‌اند (واحدهای جور نشده) را شامل شود (آمارگیری چرخشی).

کِنی و دوربین هدف اصلی از تدوین طرح‌های دوره‌ای را ترسیم روندهای موجود در سطح جامعه می‌دانند [۶]. هِندِرسون در سال ۱۹۱۶ با استفاده از میانگین متحرک به براورد تغییرات در طرح‌های دوره‌ای پرداخت [۴]. فِلِگی در سال ۱۹۶۳ روشی را برای انتخاب واحدهای آماری با احتمال متناسب در دوره‌های مختلف یك نمونه‌گیری چرخشی ارائه کرد [۳].

آمارگیری از نمونه‌های مستقل، ساده‌ترین نوع آمارگیری در طول زمان است. آمارگیری از نمونه‌های مستقل در مقطع‌های زمانی (مثلاً هر فصل یا ماه)، نمونه‌ی جداگانه‌ای را از جامعه‌ی آماری تحت بررسی، انتخاب می‌کند. بنا بر این، پس از چند مقطع زمانی، چند نمونه‌ی «مستقل» از هم وجود خواهند داشت که بر اساس هر یک می‌توان پارامتر مورد نظر را برای آن مقطع زمانی خاص براورد و در نهایت روند را در طی زمان شناسایی کرد. در این روش به‌علت عدم همبستگی میان داده‌های دوره‌های متوالی براورد تفاضل میانگین با خطای بالایی همراه است.

برای مطالعه‌ی تغییرات پارامترهای مختلف اقتصادی و اجتماعی در سطح یك جامعه‌ی آماری، باید نمونه‌های مشابه در مقاطع زمانی مختلف مورد بررسی قرار گیرند. آمارگیری پانلی به مشاهده‌ی واحدهای مشابه در طول زمان می‌پردازد. آمارگیری پانلی دارای مزایایی خاص نسبت به دیگر روش‌های نمونه‌گیری مکرر است. موزر و کالتون [۷] در مطالعات خود دریافتند که وقتی در طرح نمونه‌گیری، براورد تفاضل دو یا چند پارامتر مد نظر قرار می‌گیرد، داده‌های پانلی در مقایسه با داده‌های حاصل از دیگر روش‌های نمونه‌گیری در طی زمان و با اندازه‌ی نمونه‌ی یکسان، بسیار دقیق‌تر عمل می‌کنند. نمونه‌گیری پانلی از معدود روش‌هایی است که اجازه‌ی مطالعه‌ی تغییرات در سطح واحدهای کوچك را به آمارشناس می‌دهد. همچنین، برخی خطاها در نمونه‌گیری‌های

ابتدایی (از قبیل بی‌پاسخی، عدم درج صحیح اطلاعات و ...) در مراحل بعد قابل مرتفع شدن است و به این ترتیب هزینه‌های تصحیح اطلاعات، بسیار کاهش می‌یابند.

از سوی دیگر آمارگیری پانلی دارای اشکالاتی نیز می‌باشد. از آن‌جا که مراجعه به واحدهای یکسان در دراز مدت با مشکلاتی از قبیل بی‌پاسخی و مهاجرت مواجه می‌شود، خطای آمارگیری از دوره‌ای به دوره‌ی دیگر رو به افزایش می‌نهد. از سوی دیگر تعمیم نتایج حاصل از آمارگیری در این روش مشکل است چرا که فقط بخشی از جامعه‌ی آماری، در دوره‌های متوالی در آمارگیری حضور دارند.

طبق تعریف، در نمونه‌گیری چرخشی ممکن است برخی واحدها برای چند دوره‌ی زمانی متوالی در نمونه باقی بمانند و یا برخی واحدها پس از خروج از نمونه، مجدداً در دوره‌های بعدی وارد نمونه شوند. منظور از چرخش نمونه، جایگزین شدن همه یا تعدادی از واحدهای نمونه با واحدهای جدید از یك دوره به دوره‌ی دیگر نمونه‌گیری است.

در نمونه‌گیری‌های چرخشی، نمونه‌های مربوط به دوره‌های مختلف آمارگیری با توجه به طرح، دارای تداخلی از پیش تعیین‌شده هستند. این تداخل، روند محاسبه‌ی براوردگرها را دچار پیچیدگی‌هایی می‌کند. این شیوه‌ی آمارگیری، تلفیقی از دو شیوه‌ی آمارگیری از واحدهای مستقل و آمارگیری پانلی است. از آن‌جا که نمونه‌گیری چرخشی برای بررسی پارامترهای مختلف در سطح جامعه در دوره‌های زمانی طولانی به‌عنوان روشی مناسب در نظر آمارشناسان پذیرفته شده است و استفاده از این روش نمونه‌گیری در مراکز مختلف آماری و از جمله‌ی آن‌ها مرکز آمار ایران در حال گسترش است، انجام یك بررسی جامع در مورد چگونگی براورد پارامترهای کلیدی با واریانس کوچك در طرح‌های نمونه‌گیری چرخشی، امری ضروری به‌نظر می‌رسد. در این مقاله ابتدا با استفاده از یك مدل خطی تعمیم‌یافته، به براوردیابی تمامی ترکیبات خطی از میانگین‌های دوره‌های مختلف می‌پردازیم. سپس، به کمك یك شبیه‌سازی مونت کارلو، عملکرد تجربی روش معرفی‌شده برای دو مقطع متوالی آمارگیری مورد ارزیابی قرار می‌گیرد.

۲- یك مدل عمومی نمونه‌گیری چرخشی

در این بخش یك مدل عمومی برای نمونه‌گیری چرخشی معرفی می‌شود و در بخش‌های بعد این مدل برای براوردیابی استفاده خواهد شد. یك طرح نمونه‌گیری چرخشی را می‌توان با استفاده از یك ماتریس به‌طور کامل تعریف کرد. ماتریسِ در نظر گرفته‌شده برای این

طرح نمونه‌گیری با علامت N نشان داده می‌شود و اعضای این ماتریس به‌صورت n_{ij} تعریف می‌شوند، که $i=1,\dots,r$ و $j=1,\dots,b$. در تعریف این ماتریس، علامت b نشان‌دهنده‌ی تعداد واحدهای آماری شرکت‌کننده در نمونه‌گیری و علامت r معرف تعداد واحدهای آماری مورد مراجعه در هر دوره‌ی آماری است. متغیر t تعیین‌کننده‌ی تعداد دوره‌های نمونه‌گیری است [۴].

برای تشکیل ماتریس N توجه به نکات زیر ضروری است:

۱. با توجه به تعداد واحدهای آماری مورد مراجعه در هر دوره‌ی آماری (r واحد آماری)، r درایه‌ی اول از سطر اول ماتریس N برابر ۱ و مابقی درایه‌های سطر اول برابر صفر تعریف می‌شوند.

$$\begin{cases} n_{1j}=1; & j=1,\dots,r \\ n_{1j}=0; & j=r+1,\dots,b. \end{cases}$$

۲. اگر s میزان انتقال در یک طرح نمونه‌گیری چرخشی از یک دوره به دوره‌ی بعد تعریف شده باشد، برای تشکیل سطر دوم ماتریس N، تمامی درایه‌های سطر اول به میزان s واحد تغییر مکان داده و در یک حرکت دوره‌ای به جلو رانده می‌شوند. به این ترتیب برای هر مقدار $p=1,\dots,b-s$، درایه‌ی $(p+s)$ام از سطر دوم برابر با درایه‌ی p ام از سطر اول خواهد بود.

۳. سطرهای دیگر نیز به‌همین ترتیب با جابجایی درایه‌ها تشکیل می‌شوند.

۴. در نهایت، سطر اول در حرکت دوره‌ای تکرار خواهد شد. در این مرحله ماتریس N بدون در نظر گرفتن سطر تکراری تشکیل می‌شود.

در جدول ۱ مثال ساده‌ای از این ماتریس نشان داده شده است. در این مثال، $b=8$، $r=5$، $s=2$ و $t=4$ در نظر گرفته شده است.

جدول ۱- نمایش ماتریسی یک طرح نمونه‌گیری چرخشی با فرض $b=8$، $r=5$ و $s=2$

۱	۱	۱	۱	۱	۰	۰	۰
۰	۰	۱	۱	۱	۱	۱	۰
۱	۰	۰	۰	۱	۱	۱	۱
۱	۱	۱	۰	۰	۰	۱	۱

در بخش بعد، یك مدل خطی برای داده‌های حاصل از نمونه‌گیری چرخشی ارائه کرده و به برآورد پارامترهای مجهول در مدل ارائه‌شده خواهیم پرداخت.

۳- مدل خطی جمع‌پذیر (Linear Additive Model) با خطاهای وابسته

با توجه به این‌که در نمونه‌گیری چرخشی برخی واحدهای نمونه در دوره‌های متوالی تکرار می‌شوند، به نظر می‌رسد که توزیع مشاهدات به یکدیگر وابسته باشند. به‌همین دلیل بِلهاوْس [۱] برای اولین بار یك مدل خطی با خطاهای وابسته را برای مشاهدات در یك نمونه‌گیری چرخشی توصیه می‌کند. این مدل به‌صورت زیر معرفی می‌شود:

$$Y_{ij} = \mu_i + \varepsilon_{ij}; \quad i = ۱, ۲, \ldots, t, \quad j = ۱, ۲, \ldots, r \tag{۱}$$

که در آن Y_{ij}، نشان‌دهنده‌ی j امین مقدار مشاهده‌شده در دوره‌ی iام، μ_i نشان‌دهنده‌ی میانگین خصیصه‌ی مورد نظر در دوره‌ی iام، و ε_{ij} ها معرف خطاهای تصادفی وابسته با میانگین صفر و واریانس $\sigma^۲$ هستند. همچنین، با این فرض که ضریب همبستگی بین دو مشاهده‌ی تکراری در دوره‌ی iام و $(i+k)$ام، $\rho(k)$ می‌باشد، کوواریانس بین ε_{ij} ها به‌صورت زیر به‌دست می‌آیند:

$$\operatorname{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = -\frac{\sigma^۲}{b-۱}$$
$$j \neq j', \quad j = ۱, \ldots, r, \quad j' = ۱, \ldots, r, \quad i = ۱, \ldots, t \tag{۲}$$

اگر s میزان انتقال در یك طرح نمونه‌گیری چرخشی از یك دوره به دوره‌ی بعد باشد، آنگاه، برای هر مقدار مفروض برای k، اگر $۰ < r - ks$ و $b - r > ks$ باشد، کوواریانس بین خطاهای تصادفی در دوره‌های iام و $(i+k)$ام برابر است با:

$$\begin{cases} \operatorname{cov}\{\varepsilon_{(i+k)j}, \varepsilon_{i(j+ks)}\} = \sigma^۲ \rho(k) & j = ۱, \ldots, (r-ks) \\ \operatorname{cov}\{\varepsilon_{(i+k)j}, \varepsilon_{ij'}\} = ۰ & \text{برای مقادیر دیگر } j \text{ و } j' \end{cases} \tag{۳}$$

اما اگر $۰ < r - ks$ و $b - r > ks$ باشد، آنگاه:

$$(۴)\quad \begin{cases} \operatorname{cov}\{\varepsilon_{(i+k)j}, \varepsilon_{i(j+ks)}\} = \sigma^{۲}\rho(k) & j = ۱,\ldots,(r-ks) \\ \operatorname{cov}\{\varepsilon_{(i+k)(b-ks+j)}, \varepsilon_{ij}\} = \sigma^{۲}\rho(k) & j = ۱,\ldots,[r-(b-ks)] \\ \operatorname{cov}\{\varepsilon_{(i+k)j}, \varepsilon_{ij'}\} = ۰ & \text{برای مقادیر دیگر } j \text{ و } j' \end{cases}$$

و اگر $۰ > r-ks$ و $b-r < ks$ باشد، آن‌گاه:

$$(۵)\quad \begin{cases} \operatorname{cov}\{\varepsilon_{(i+k)(b-ks+j)}, \varepsilon_{ij}\} = \sigma^{۲}\rho(k) & j = ۱,\ldots,[r-(b-ks)] \\ \operatorname{cov}\{\varepsilon_{(i+k)j}, \varepsilon_{ij'}\} = ۰ & \text{برای مقادیر دیگر } j \text{ و } j' \end{cases}$$

لازم به ذکر است، در فرمول‌های (۳) تا (۵)، i مقادیر ۱ تا $(t-k)$ را می‌پذیرد.

بلهاوس برای تخمین μ_i ها، فقط براوردگرهای خطی از میانگین‌های نمونه‌ای دوره‌های مختلف را مورد بررسی قرار داد [۱]. اما در این بخش، به منظور یافتن براوردگرهای بهینه، تمامی توابع خطی از Y_{ij} ها در نظر گرفته شده و با بکارگیری روش کم‌ترین توان‌های دوم خطا، براوردگرهای مطلوب برای هر ترکیب خطی از میانگین دوره‌های مختلف بدست آمده است. در بخش بعد برتری ارزنده‌ی این براوردها در تخمین تفاضل میانگین‌ها نیز نسبت به دیگر براوردها از طریق شبیه‌سازی عددی نشان داده می‌شود.

مدل خطی معرفی‌شده در رابطه‌ی (۱) را می‌توان به شکل ماتریسی زیر نمایش داد:

$$(۶)\qquad \underline{Y} = A\,\underline{\mu} + \underline{\varepsilon}$$

که در آن $\underline{Y}$ یك بردار rt بعدی است که Y_{ij} ها را شامل می‌شود. همچنین، $\underline{\mu}$ نشان‌دهنده‌ی بردار میانگین‌های خصیصه‌ی مورد نظر در جامعه و در دوره‌های مختلف است. در این رابطه، A یك ماتریس $rt \times t$ بعدی (ماتریس طرح مدل) است که به صورت زیر تعریف می‌شود:

$$(۷)\qquad A = \begin{bmatrix} \underline{۱}_r & \underline{۰} & \cdots & \underline{۰} \\ \underline{۰} & \underline{۱}_r & \cdots & \underline{۰} \\ \vdots & & & \\ \underline{۰} & \underline{۰} & \cdots & \underline{۱}_r \end{bmatrix}$$

در ماتریس قبل، $\underline{1}_r$ یك بردار r بعدی با مؤلفه‌های "۱" است.

در رابطه‌ی (۶)، $\underline{\varepsilon}$ بردار خطای تصادفی با میانگین صفر و ماتریس واریانس-کوواریانس $V\,(\underline{\varepsilon})$ است که مؤلفه‌های آن به‌وسیله‌ی رابطه‌های (۲) تا (۵) قابل تعریف هستند.

در زیر به بررسی دو حالت در برآوردیابی بردار $\underline{\mu}$ پرداخته شده است:

آ) اگر برای مقادیر مختلف k، $\rho(k)$ و σ^{2} مشخص باشند، ماتریس $\rho=\frac{1}{\sigma^{2}}V\,(\underline{\varepsilon})$ یك ماتریس با مؤلفه‌های مشخص بوده و برآوردگر تعمیم‌یافته‌ی کم‌ترین توان‌های دوم خطا برای بردار میانگین عبارت است از:

$$\underline{\hat{\mu}}=(A'\rho^{-1}A)^{-1}A'\rho^{-1}\underline{Y} \qquad (۸)$$

و واریانس $\underline{\hat{\mu}}$ برابر است با:

$$V\,(\underline{\hat{\mu}})=(A'\rho^{-1}A)^{-1}\,\sigma^{2}. \qquad (۹)$$

ب) وقتی که ضریب همبستگی بین مشاهدات و واریانس خطا در مدل در دوره‌های مختلف نامشخص باشند، باید از برآورد این کمیت‌ها استفاده نمود. یك برآورد گشتاوری برای $\rho(k)$، با استفاده از ضریب همبستگی نمونه‌ای بین مشاهدات در دو دوره با فاصله‌ی k، به‌صورت زیر تعریف می‌شود:

$$\hat{\rho}(k)=\frac{1}{t}\sum_{i=1}^{t}\hat{\rho}\{i\,,\,(i+k)\} \qquad (۱۰)$$

که در آن $\hat{\rho}\{i\,,\,(i+k)\}$، ضریب همبستگی نمونه‌ای مشاهدات تکراری در دوره‌ی iام و $(i+k)$ام می‌باشد (دوره‌ها به‌صورت دایره‌ای در نظر گرفته می‌شوند). همچنین، اگر S_i^{2} واریانس نمونه‌ای در دوره‌ی iام باشد، یك برآورد مناسب برای σ^{2} عبارت است از:

$$S_p^{2}=\frac{1}{t}\sum_{i=1}^{t}S_i^{2} \qquad (۱۱)$$

در نتیجه یك براوردگر مناسب برای بردار میانگین خصیصه‌ی مورد نظر در دوره‌های مختلف عبارت است از:

$$\hat{\hat{\mu}} = (A'\hat{\rho}^{-۱}A)^{-۱}A'\hat{\rho}^{-۱}\underline{Y} \qquad (۱۲)$$

در رابطه‌ی (۱۲)، $\hat{\rho}$ ، معرف براوردی برای ماتریس ضریب همبستگی ρ است که برای بردار $\underline{\varepsilon}$ تعریف شده بود. ماتریس $\hat{\rho}$ مانند ماتریس ρ است، با این تفاوت که در آن به جای $\rho(k)$ ، از $\hat{\rho}(k)$ استفاده می‌شود.

یادآوری چند نکته در ارتباط با مطالب قبل لازم به نظر می‌رسد:

۱. اگر ماتریس ضرائب همبستگی ρ مشخص باشد، براوردگر تعریف شده‌ی $\underline{\hat{\mu}}$ در رابطه‌ی (۸)، بهترین براوردگر خطی ناااریب برای $\underline{\mu}$ خواهد بود.

۲. در صورتی‌که ماتریس ضرائب همبستگی، نامشخص باشد، براوردگر $\underline{\hat{\hat{\mu}}}$ می‌تواند براوردگر مناسبی برای $\underline{\mu}$ باشد. ولی باید به این نکته توجه داشت که این براوردگر ناااریب نخواهد بود.

۳. با استفاده از رابطه‌های (۲) تا (۵) به‌راحتی نمی‌توان درایه‌های ماتریس ρ را تعیین نمود. لذا به‌منظور ارائه‌ی یك راه ساده برای مشخص‌سازی ماتریس ρ ، ابتدا بردار مشاهدات مربوط به کل جامعه را به‌صورت زیر در نظر می‌گیریم:

$$(۱۳) \quad \underline{Y}^{*'} = \left(\underline{Y}'_{۱}, \underline{Y}'_{۲}, \ldots, \underline{Y}'_{t}\right), \qquad \underline{Y}'_{i} = \left(Y_{i۱}, Y_{i۲}, \ldots, Y_{ib}\right)$$

با درنظر گرفتن بردار تعریف‌شده‌ی کل مشاهدات جامعه، ماتریس واریانس-کوواریانس مربوط به $\underline{Y}^{*}$ به شکل زیر قابل تعریف است:

$$V(\underline{Y}^{*}) = \begin{bmatrix} \Sigma & \rho(۱)\sigma^{۲}I_b & \rho(۲)\sigma^{۲}I_b & \cdots & \rho(t-۱)\sigma^{۲}I_b \\ \rho(۱)\sigma^{۲}I_b & \Sigma & \rho(۱)\sigma^{۲}I_b & \cdots & \rho(t-۲)\sigma^{۲}I \\ \rho(۲)\sigma^{۲}I_b & \rho(۱)\sigma^{۲}I_b & \Sigma & \cdots & \rho(t-۳)\sigma^{۲}I \\ \vdots & & & & \\ \rho(t-۱)\sigma^{۲}I & \rho(t-۲)\sigma^{۲}I & \rho(t-۳)\sigma^{۲}I & \cdots & \Sigma \end{bmatrix}$$

در رابطه‌ی بالا، $\rho(k)$ نشان‌دهنده‌ی ضریب همبستگی مشاهدات تکراری در دو دوره با فاصله‌ی k بوده و I_b ماتریس یکه‌ی $b\times b$ بعدی است. همچنین، Σ ،ماتریس واریانس- کوواریانس مربوط به بردارهای $\underline{Y}_i$ می‌باشد، که به‌صورت زیر تعریف می‌شود:

$$\Sigma=\begin{bmatrix} \sigma^{2} & -\frac{\sigma^{2}}{b-1} & \cdots & -\frac{\sigma^{2}}{b-1} \\ -\frac{\sigma^{2}}{b-1} & \sigma^{2} & \cdots & -\frac{\sigma^{2}}{b-1} \\ \vdots & & & \\ -\frac{\sigma^{2}}{b-1} & -\frac{\sigma^{2}}{b-1} & \cdots & \sigma^{2} \end{bmatrix} \qquad (14)$$

با توجه به روابط بالا، ماتریس واریانس-کوواریانس $\underline{\varepsilon}$ ، به‌صورت زیر محاسبه می‌گردد:

$$V(\underline{\varepsilon})=T\,V(\underline{Y}^{*})\,T' \qquad (15)$$

به‌طوری که ماتریس تبدیل T بر اساس ماتریس طرح نمونه‌گیری (N)، به‌شکل زیر تعریف می‌شود:

$$T=\begin{bmatrix} D^{(1)} & 0 & \cdots & 0 \\ 0 & D^{(2)} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & D^{(t)} \end{bmatrix}_{rt\times tb} \qquad (16)$$

که در آن ماتریس‌های $D^{(k)}$، با بعد $r\times b$، ماتریس‌هایی با مؤلفه‌های $d_{ij}^{(k)}$ هستند به‌طوری که:

$$\begin{cases} d_{jj}^{(k)}=1 & \text{اگر} \quad j=(k-1)s+i\le b;\quad i=1,\dots,r \\ d_{(j-b)(j-b)}^{(k)}=1 & \text{اگر} \quad j=(k-1)s+i> b;\quad i=1,\dots,r \\ d_{ij}^{(k)}=0 & \text{برای مقادیر دیگر } i \text{ و } j \end{cases} \qquad (17)$$

لازم به ذکر است، ماتریس واریانس-کوواریانس $\underline{Y}^{*}$ در رابطه‌ی (۱۳) با این فرض محاسبه شده است که σ^{2} (واریانس جامعه)، در دوره‌های مختلف یکسان باشد. در صورتی‌که اختلافی میان واریانس دوره‌های مختلف وجود داشته باشد، باید این تفاوت در ماتریس $V(\underline{Y}^{*})$ نمایان شود.

در انتها به این نکته توجه کنید که در مدل خطی ارائه‌شده در رابطه‌ی (۶)، و با تعریف ماتریس A به شکل نشان داده‌شده در رابطه‌ی (۷)، تمامی ترکیبات خطی از میانگین‌های دوره‌های مختلف براوردپذیر هستند. به عبارت دیگر، برای هر مقادیر ثابت $c_1,\dots,c_t$، پارامتر $\sum_{i=1}^{t} c_i\mu_i$، براوردپذیر خواهد بود. بنا بر این، با استفاده از این روش می‌توان براوردگرهای خطی نااریب با کم‌ترین واریانس برای پارامترهای میانگین دوره، تغییرات بین دوره‌ها و میانگین چند دوره‌ی متوالی ارائه داد [۲].

به‌طور کلی و با استفاده از مطالب پیش‌گفته، بهترین براوردگر خطی نااریب برای $\sum_{i=1}^{t} c_i\mu_i$ عبارت است از:

$$\sum_{i=1}^{t} c_i\hat{\mu}_i=\underline{C}'\,\underline{\hat{\mu}} \qquad (۱۸)$$

که در آن $\underline{C}'=(c_1,\dots,c_t)$ و $\underline{\hat{\mu}}$ براورد میانگین ارائه‌شده در رابطه‌ی (۸) می‌باشد. واریانس این براوردگر نیز برابر است با:

$$\underline{C}'(A'\rho^{-1}A)^{-1}\underline{C}\,\sigma^{2} \qquad (۱۹)$$

به‌عنوان مثال بهترین براوردگر خطی نااریب برای میانگین دوره‌ی iام (μ_i)، iامین مؤلفه‌ی بردار $\underline{\hat{\mu}}$ در رابطه‌ی (۸) بوده و واریانس این براوردگر برابر است با:

$$\operatorname{var}(\hat{\mu}_i)=v_{ii}\,\sigma^{2} \qquad (۲۰)$$

که v_{ii} مؤلفه‌ی سطر iام و ستون iام از ماتریس $(A'\rho^{-1}A)^{-1}$ می‌باشد.

۴- شبیه‌سازی

در این بخش، از شبیه‌سازی مونت کارلو برای ارزیابی روش معرفی‌شده برای براوردیابی در نمونه‌گیری چرخشی استفاده می‌شود. برای این منظور، فرض می‌کنیم هدف آمارگیری، براورد تغییرات میانگین متغیر X از یك جامعه با جمعیت ۲۰۰۰۰۰ در دو مقطع زمانی متوالی است. برای آن‌که آمارگیری مکرر توجیه‌پذیر باشد، توزیع متغیر X در دو مقطع زمانیِ متوالی متفاوت در نظر گرفته شده و داده‌های این دو مقطع به گونه‌ای تولید می‌شوند که دارای همبستگی باشند. در نهایت و پس از انجام نمونه‌گیری، براورد تغییرات پارامتر به روش معرفی‌شده در بخش (۲) مورد ارزیابی تجربی قرار می‌گیرد.

۴-۱- طرح مطالعه

ابتدا اندازه‌ی نمونه‌ها را برای هر مقطع و با توجه به جامعه‌های آماری مشخص کرده و سپس نمونه‌گیری به روش چرخشی صورت می‌پذیرد. برای آن‌که اثر میزان تداخل نیز در این ارزیابی لحاظ شود، نمونه‌گیری چرخشی، در پنج وضعیت مختلف (تداخل‌های ۱۰، ۲۰، ۳۰، ۴۰ و ۵۰ درصد) صورت می‌گیرد. برای مثال، در آمارگیری چرخشی با تداخل ۳۰ درصد، ۳۰ درصد از مقادیر نمونه‌های مقطع دوم از نمونه‌های مقطع اول به دست می‌آیند و بقیه‌ی واحدها متفاوت (مستقل) از واحدهای مقطع اول انتخاب می‌شوند. لازم به ذکر است، یکسان بودن واحدهای نمونه‌ای در دو مقطع، به معنای برابری مقدار متغیر آن‌ها در دو مقطع نیست، زیرا جامعه‌های آماری در دو مقطع زمانی از یك توزیع و با پارامترهای مختلف تولید می‌شوند.

نمونه‌گیری در دو مقطع زمانی به روش تصادفی ساده و با در نظر گرفتن درصد تداخل، انجام می‌شود. پس از انجام روند نمونه‌گیری، با استفاده از میانگین توان دوم خطا، براورد تفاضل میانگین جامعه در دو دوره‌ی مورد نظر با براورد ارائه‌شده در بخش ۲ مقایسه می‌شود. همچنین، تفاضل میانگین نمونه‌ای در دو مقطع، با استفاده از نتایج شبیه‌سازی قابل بررسی خواهد بود. لازم به ذکر است، از ترکیب پنج حالت نمونه‌گیری که هر یک در چهار وضعیت همبستگی شبیه‌سازی شده‌اند، به ۲۰ حالت مختلف شبیه‌سازی می‌رسیم. نتایج حاصل از این شبیه‌سازی پاسخ‌گویی به پرسش‌های زیر را امکان‌پذیر می‌سازند:

آیا روش ارائه‌شده در این مقاله کارایی لازم را دارد؟

آیا افزایش میزان تداخل می‌تواند بر دقت روش مورد نظر اثرگذار باشد؟
افزایش همبستگی بین داده‌های دو مقطع متوالی چه اثری در کارایی این روش دارد؟

۲-۴- تولید داده‌ها و نتایج حاصل

برای ایجاد داده‌های مورد نیاز در مطالعه‌ی حاضر، از نرم‌افزار SPSS استفاده شده است. در گام اول ۲۰۰۰۰۰ داده‌ی تصادفی از یک توزیع نرمال با میانگین ۱۰۰۰ و انحراف معیار ۱۰۰ تولید شده است. این اعداد در کل مراحل شبیه‌سازی به‌عنوان جامعه‌ی آماری در مقطع زمانی اول در نظر گرفته می‌شوند. در گام دوم و در هر مرحله از شبیه‌سازی، با استفاده از داده‌های به‌دست آمده در گام اول و با در نظر گرفتن همبستگی مورد نیاز (ρ)، داده‌های مربوط به جامعه‌ی آماری در مقطع زمانی دوم با توجه به رابطه‌ی زیر تولید می‌شوند:

$$X_{۲} = \rho \cdot X_{۱} + z \qquad (۲۱)$$

که در آن z یک متغیر تصادفی با توزیع $N(\alpha, (۱-\rho^{۲})\times۱۰۰)$ و مستقل از $X_{۱}$ است. ضریب همبستگی میان دو متغیر $X_{۱}$ و $X_{۲}$ به دست آمده از رابطه‌ی (۲۱) برابر ρ خواهد بود. این مطلب به‌کمک رابطه‌های زیر قابل اثبات است.

$$\mathrm{cov}(X_{۱}, X_{۲}) = \mathrm{cov}(\rho \cdot X_{۱} + z, X_{۱}) = \rho \ \mathrm{var}(X_{۱})$$

$$Var(X_{۲}) = \rho^{۲} \ \mathrm{var}(X_{۱}) + \mathrm{var}(z) = \rho^{۲} \ \mathrm{var}(X_{۱}) + \mathrm{var}(X_{۱})(۱-\rho^{۲}) = \mathrm{var}(X_{۱})$$

$$\rho(X_{۲}, X_{۱}) = \frac{\rho \, \mathrm{var}(X_{۱})}{\sqrt{\mathrm{var}(X_{۱}) \ \mathrm{var}(X_{۲})}} = \frac{\rho \ \mathrm{var}(X_{۱})}{\sqrt{\mathrm{var}(X_{۱}) \ \mathrm{var}(X_{۱})}} = \frac{\rho \ \mathrm{var}(X_{۱})}{\mathrm{var}(X_{۱})} = \rho$$

از آن‌جا که انتظار می‌رود کارایی روش‌های براوردیابی در نمونه‌گیری چرخشی تحت تأثیر میزان همبستگی داده‌های دو مقطع باشد، شبیه‌سازی تحت چهار مقدار همبستگیِ ۰/۲، ۰/۴، ۰/۶ و ۰/۸ تکرار شده است تا بدین وسیله، اثر میزان همبستگی بر کارایی، سنجیده شود. مشخصات جامعه‌های آماری برای چهار وضعیت گفته شده به شرح زیر است:

$$1: X_1 \sim N(1000, 100) \quad X_2 \sim N(700, 100) \quad \rho_{X_1 X_2} = 0/2$$
$$2: X_1 \sim N(1000, 100) \quad X_2 \sim N(900, 100) \quad \rho_{X_1 X_2} = 0/4$$
$$3: X_1 \sim N(1000, 100) \quad X_2 \sim N(1100, 100) \quad \rho_{X_1 X_2} = 0/6$$
$$4: X_1 \sim N(1000, 100) \quad X_2 \sim N(1300, 100) \quad \rho_{X_1 X_2} = 0/8$$

که در آن X_1 و X_2 به‌ترتیب متغیرهای مورد نظر در مقاطع زمانیِ متوالی اول و دوم هستند.

در مقاطع مختلف شبیه‌سازی برای به دست آوردن اندازه‌ی نمونه در مقاطع زمانی اول و دوم، از فرمول نمونه‌گیری تصادفی ساده‌ی بدون جایگذاری با حاشیه‌ی خطای (Margin of Error) ۱۰ و سطح اطمینان ۹۵ درصد استفاده شده است.

$$n_1 = \frac{\left(\frac{z}{d}\right)^2 \operatorname{var}(X_1)}{1+\frac{1}{N}\left(\frac{z}{d}\right)^2 \operatorname{var}(X_1)} = \frac{\left(\frac{1/96}{10}\right)^2 (100)^2}{1+\frac{1}{200000}\left(\frac{1/96}{10}\right)^2 (100)^2} \cong 384$$

لازم به ذکر است، برای هر یک از حالات ۲۰گانه، ۱۰۰۰ نمونه‌ی ۳۸۴ تایی انتخاب می‌شود تا بر اساس آن ارزیابی تجربی صورت گیرد.

پس از اتمام عملیات نمونه‌گیری، به‌منظور براورد تغییرات میانگین در دو مقطع زمانی، از دو روش استفاده می‌شود و سپس نتایج حاصل با یکدیگر مقایسه می‌شوند. در روش اول و بدون در نظر گرفتن تداخل استفاده‌شده در روند نمونه‌گیری، تفاضل میانگین‌های نمونه‌ای را به‌عنوان براوردگر تغییرات میانگین در دو مقطع زمانی در نظر گرفته و میانگین توان دوم خطای آن از رابطه‌ی زیر به‌دست می‌آید:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{1000}(\hat{\theta}_i - \theta)^2}{1000} \qquad (22)$$

که در آن

$$\hat{\theta}_i = \hat{\bar{X}}_{2i} - \hat{\bar{X}}_{1i} \qquad i = 1, 2, \dots, 1000$$

و

$$\theta = \mu_۲ - \mu_۱ = \begin{cases} -۳۰۰ & \text{for case } ۱ \\ -۱۰۰ & \text{for case } ۲ \\ ۱۰۰ & \text{for case } ۳ \\ ۳۰۰ & \text{for case } ۴ \end{cases}.$$

در رابطه‌ی بالا منظور از θ، تفاضل واقعی دو جامعه (وضعیت جامعـه در دو مقطع زمانی) است. همان‌گونه که شرح داده شد، برای انجام نمونه‌گیری، تعداد ۳۸٤ نمونـه از دو جامعه انتخاب می‌شود و این عمل ۱۰۰۰ بار رخ می‌دهد. در هر بار نمونـه‌گیـری، از ۳۸٤ واحدِ نمونه‌گیری‌شده، میانگینی برای هر یك از جامعه‌ها به‌دست می‌آید، که تفاضل ایـن دو مقدار، به‌عنوان براوردی برای θ در نظر گرفته می‌شود. این مقـدار در نمونـه‌گیـری iام بـا علامـت $\hat{\theta}_i$ نـشان داده مـی‌شـود ($i \in \{۱, ۲, \ldots, ۱۰۰۰\}$) و در رابطــه‌ی (۲۰) بــرای محاسبه‌ی میزان خطا مورد استفاده قرار می‌گیرد.

در روش دوم و به‌منظور براورد خطای تغییـرات میـانگین، از رابطـه‌ی (۱۸) اسـتفاده می‌شود. نتایج حاصل در جدول (۲) قابل مشاهده است.

جدول ۲- میانگین توان دوم خطا برای دو روش براوردیابی در هر یک از حالات ۲۰گانه‌ی شبیه‌سازی

روش آمارگیری		تفاضل میانگین‌های نمونه‌ای دو دوره همبستگی بین داده‌های دو مقطع متوالی				براورد حاصل از رابطه‌ی (۱۸) همبستگی بین داده‌های دو مقطع متوالی			
		۰/۲	۰/۴	۰/۶	۰/۸	۰/۲	۰/۴	۰/۶	۰/۸
آمارگیری چرخشی	۱۰% تداخل	۵۱۵	۴۹۱	۴۶۰	۴۵۴	۳۹۵	۳۵۷	۲۴۶	۱۸۲
	۲۰% تداخل	۴۹۳	۴۶۵	۴۱۸	۴۰۳	۳۶۷	۲۷۳	۱۷۵	۱۵۱
	۳۰% تداخل	۴۹۰	۴۳۴	۳۸۲	۳۴۹	۳۲۶	۲۴۷	۱۵۱	۱۰۲
	۴۰% تداخل	۴۹۷	۴۲۶	۳۶۳	۳۲۵	۳۰۸	۲۰۳	۱۳۴	۸۵
	۵۰% تداخل	۴۵۷	۴۰۰	۳۱۳	۲۶۶	۲۷۶	۱۴۶	۹۸	۴۵

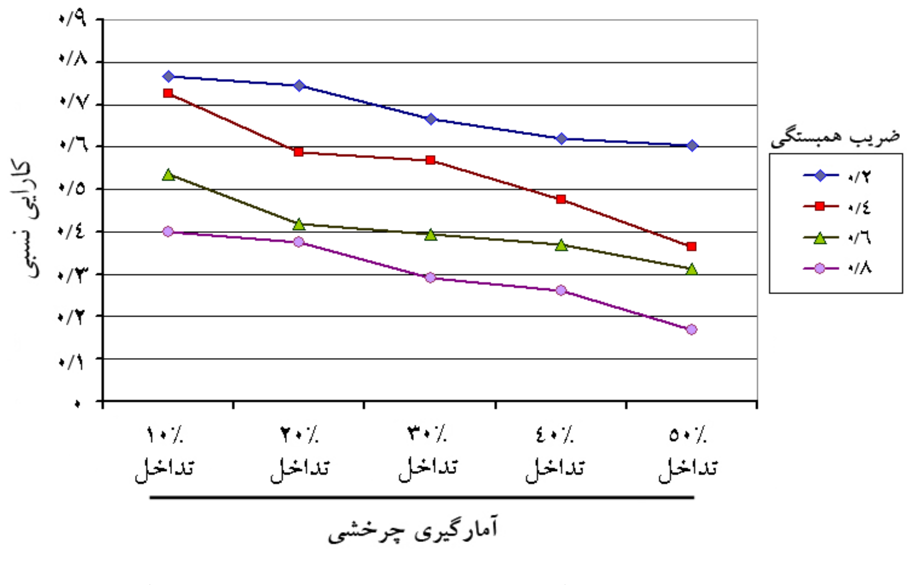
نتایج نشان می‌دهند که در آمارگیری‌های چرخشی، با افزایش درصد تـداخل، همـواره از مقدار میانگین توان دوم خطای براورد کاسته می‌شود.

به‌منظور بررسی نتایج به‌دست آمده، به مقایسه‌ی براورد کارایی نسبی (Relative Efficiency) پرداخته می‌شود. برای این منظور از رابطه‌ی زیر استفاده شده است:

$$RE_{ij} = \frac{\operatorname{var}(\underline{C}'\hat{\underline{\mu}}_{ij})}{\operatorname{var}(\bar{X}_{۲ij} - \bar{X}_{۱ij})} \qquad i = ۱, ۲, ۳, ۴; \quad j = ۱, ۲, ۳, ۴, ۵ \qquad (۲۳)$$

که در این رابطه، $\operatorname{var}(\underline{C}'\hat{\underline{\mu}}_{ij})$ نشان‌دهنده‌ی واریانس براورد تفاضل میانگین در دو مقطع نمونه‌گیری چرخشی معرفی‌شده در رابطه‌ی (۱۸)، برای همبستگی i ام و تداخل j ام می‌باشد. همچنین، $\operatorname{var}(\bar{X}_{۲ij} - \bar{X}_{۱ij})$ نشان‌دهنده‌ی واریانس تفاضل میانگین نمونه‌ای در دو مقطع متوالی برای همبستگی i ام و تداخل j ام است. بنا بر این، مقداری کم‌تر از "یک" بر کارا بودن روش معرفی‌شده در این مقاله نسبت به تفاضل میانگین نمونه‌ای برای براورد تغییرات در دو دوره‌ی متوالی دلالت دارد.

شکل ۱ مقدارهای مربوط به کارایی نسبی برای دو روش به کار رفته با استفاده از رابطه‌ی (۲۱) را نشان می‌دهد.



شکل ۱- مقایسه‌ی کارایی براورد تفاضل میانگین دو دوره با استفاده از رابطه‌ی (۱۸) و تفاضل میانگین نمونه‌ای با در نظر گرفتن ضریب‌های همبستگی مختلف.

۳-۴- بحث و نتیجه‌گیری نهایی

با بررسی یافته‌های حاصل از شبیه‌سازی انجام‌شده، می‌توان به سئوال‌های مطرح‌شده در بخش ۱-۴ پاسخ داد. در شبیه‌سازی انجام‌شده، ابتدا با در نظر گرفتن این‌که آمارگیری به‌روش نمونه‌گیری تصادفی ساده صورت پذیرفته است، از تفاضل میانگین نمونه‌ای برای براورد تفاضل میانگین دو دوره استفاده شده است. در این روش به میزان تداخل در براورد خطا توجه نشده است. در گام بعد با استفاده از روش جدید معرفی شده در این مقاله به براورد خطای تغییرات میانگین پرداخته شده است که در این روش، میزان تداخل در براوردیابی اثر مستقیم دارد. نتایج آورده‌شده در جدول ۲ نشان می‌دهد که با استفاده از روش ارائه‌شده در این مقاله، براورد جدید معرفی‌شده با خطای کم‌تری همراه خواهد بود. همچنین، با دقت در شکل ۱ می‌توان نتیجه گرفت که با افزایش میزان تداخل در نمونه‌گیری از مقاطع مختلف، بر کارایی روش مزبور افزوده می‌شود. نکته‌ی دیگر آن‌که با افزایش هبستگی میان داده‌های مقاطع نمونه‌گیری نیز کارایی این روش افزایش می‌یابد.

در پایان خاطر نشان می‌شود، شبیه‌سازی حاضر فقط مدل ساده‌ای از نمونه‌گیری چرخشی را ارائه کرده است. پیش‌بینی می‌شود روش معرفی‌شده در این مقاله در نمونه‌گیری‌هایی با الگوی چرخش پیچیده‌تر کارایی چشم‌گیرتری از خود به نمایش بگذارد. همچنین، در شبیه‌سازی آورده‌شده فقط به بررسی براورد تفاضل میانگین پرداخته شده است. انتخاب این پارامتر بدان سبب است که کاهش خطا در براورد تغییرات مزیت ارزنده‌ی استفاده از آمارگیری چرخشی نسبت به دیگر روش‌های متداول آمارگیری مکرر است. اما نباید فراموش کرد که روش معرفی‌شده در مقاله‌ی حاضر برای براوردیابی همه‌ی ترکیبات خطی از میانگین‌های دوره‌های مختلف کارایی دارد.

توضیحات

۱. Repeated Survey
۲. Panel Survey
۳. Rotation Survey

مرجع‌ها

[1] Bellhouse, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *Journal of Statistical Planning and Inferences,* **21**, 69–74.

[2] Christensen, R. (1987). *The Theory of Linear Models*. Springer,Verlag.

[3] Fellegi, I.P. (1963). Sampling with and without replacement: Rotating and non–rotating samples. *J. Amer. Stat. Ass.,* **58**, 183–201.

[4] Henderson, R. (1916), Note on graduation by adjusted average. *Transactions of the American Society of Actuaries,* **17**, 43–48.

[5] John, J.A., Wolock, F.M., and David, H.A. (1972). Cyclic designs, national Bureau of Standards, *Applied Mathematics Series,* 62.

[6] Kenny, P.B., and Durbin, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society*, Series A, **145**, 1–41.

[7] Moser, C.A. and Kalton, G. (1979). *Survey Methods in Social Investigation*. Second Edition, Heinemann Educational Books, London.

مینا توحیدی
دکتری آمار
شیراز، چهارراه ادبیات، دانشکده‌ی علوم دانشگاه شیراز، بخش آمار.
رایانشانی: mtowhidi@susc.ac.ir

محمدرضا نمازی راد
دانشجوی دکتری آمار کاربردی
شیراز، خیابان عفیف‌آباد، کوچه‌ی ۶/ ۲۰، پلاک ۴۳ کد پستی: ۱۴۸۳۴– ۷۱۸۳۸.
رایانشانی: mr_namazi@yahoo.com