

برآورد طرح پایه میانگین جامعه متناهی در نمونه‌گیری از مجموعه رتبه‌دار

زهرا شیرکوند[†] و روشنگر علی‌اکبری صبا[‡]

[†] دانشگاه علمی- کاربردی

[‡] پژوهشکده آمار

چکیده. این مقاله روش نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری در جامعه‌های متناهی را معرفی می‌کند و به نحوه‌ی محاسبه احتمال شمول نمونه‌ها تحت این روش می‌پردازد. برآوردگرهای هورویتز-تامپسون و هنس-هورویتز را با استفاده از متغیر کمکی تحت این طرح معرفی کرده و براساس داده‌های هزینه و درآمد خانوارهای استان تهران در سال ۱۳۹۰ نشان می‌دهد که برخی از برآوردگرهای حاصل از روش نمونه‌گیری از مجموعه‌ی رتبه‌دار کاراتر از برآوردگرهای نمونه‌گیری تصادفی ساده است.

واژگان کلیدی: جامعه متناهی، برآوردگر هورویتز-تامپسون، برآوردگر هنس-هورویتز، احتمال شمول، برآوردگر نسبی، نمونه‌گیری از مجموعه‌ی رتبه‌دار.

۱- مقدمه

ارائه‌ی روش‌های جدید نمونه‌گیری برای افزایش کارایی و دقت برآوردهای به دست آمده، یکی از مهم‌ترین اهداف نظریه‌ی نمونه‌گیری است. گاهی واحدهای برخی از جامعه‌های تحت مطالعه را می‌توان به‌آسانی و با صرف کم‌ترین هزینه، رتبه‌بندی کرد. روش نمونه‌گیری از مجموعه‌ی رتبه‌دار (RSS) و تعمیم‌های آن، اغلب روش‌های کاراتری را برای

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۹۹/۲/۳، پذیرش: ۱۳۹۹/۱۲/۱۳.

برآورد پارامترهای جامعه مورد مطالعه، ارائه می‌کنند. در روش‌های نمونه‌گیری از مجموعه‌ی رتبه‌دار برآوردگر ساده میانگین جامعه اغلب دارای دقت بیش‌تری نسبت به برآوردگر حاصل از نمونه‌گیری تصادفی ساده با همان تعداد واحد اندازه‌گیری شده است، بنا بر این می‌توان استنباط‌های درست‌تری درباره‌ی جامعه‌ی مورد مطالعه انجام داد. این افزایش کارایی در نتیجه رتبه‌بندی واحدها حاصل می‌شود که با کمترین هزینه صورت گرفته است. اگر متغیری در اختیار باشد که به‌آسانی مشاهده و رتبه‌بندی شود و همبستگی بالایی با متغیر مورد مطالعه داشته باشد، رتبه‌بندی بر اساس این متغیر انجام می‌شود. شایان ذکر است اگر مقدار ترتیبی صفت مورد مطالعه با مقدار واقعی آن در جامعه برابر باشد، رتبه‌بندی RSS کامل است و در غیر این صورت رتبه‌بندی ناکامل است.

ایده‌ی نمونه‌گیری از مجموعه‌ی رتبه‌دار اولین بار توسط مک‌اینتایر [۹] مطرح شد. استاکس [۱۱] استفاده از متغیر کمکی در نمونه‌گیری از مجموعه‌ی رتبه‌دار را در نظر گرفت. مسئله‌ی استفاده از طرح نمونه‌گیری از مجموعه‌ی رتبه‌دار در جامعه‌های متناهی اولین بار توسط تاکاهاسی و فوتات‌سویا [۱۲] مورد مطالعه قرار گرفت. ال-صالح و سماوی [۴] احتمال شمول در نمونه‌گیری از مجموعه‌ی رتبه‌دار را مورد مطالعه قرار دادند. جعفری جوزانی و جانسون [۸] برآورد طرح‌پایه با استفاده از نمونه‌گیری از مجموعه‌ی رتبه‌دار در جامعه‌های متناهی را مورد بررسی قرار دادند. طرح RSS به‌صورت قابل ملاحظه‌ای بر اساس استنباط مدل‌پایه گسترش یافته است و تنها تعداد اندکی از مقاله‌ها از جمله مقاله‌های [۱۰]، [۱۲]، [۵] و [۸] بر استنباط طرح‌پایه متمرکز هستند. این مقاله برآوردهایی تحت طرح نمونه‌گیری از مجموعه‌ی رتبه‌دار برای میانگین هزینه و درآمد خانوارهای استان تهران در سال ۱۳۹۰ بر اساس روش نمونه‌گیری از مجموعه‌ی رتبه‌دار ارائه و نشان می‌دهد این برآوردها نسبت به برآوردهای متناظر در نمونه‌گیری تصادفی ساده بهتر عمل می‌کنند و دقت بالاتری دارند.

۲- برآورد طرح پایه نمونه‌گیری از مجموعه‌ی رتبه‌دار با استفاده از متغیر کمکی

۲-۱- معرفی روش

جامعه‌ی متناهی شامل N عضو با واحدهای $U = \{U_1, \dots, U_N\}$ شامل زوج‌های مرتب $(y_1, x_1), \dots, (y_N, x_N)$ داریم، که در آن y ها متغیر مورد مطالعه و x ها متغیر کمکی برای رتبه‌بندی هستند. فرض می‌کنیم $x_{i:U}$ نشان‌دهنده‌ی i امین مقدار ترتیبی x در جامعه است و برای یک نمونه تصادفی ساده S به اندازه‌ی n ، نشان‌دهنده‌ی i امین مقدار x در نمونه است. برای نمونه‌گیری از مجموعه‌ی رتبه‌دار، m و k ثابت را به‌گونه‌ای انتخاب می‌کنیم که در رابطه‌ی $n = mk \leq N$ صدق کند. روش نمونه‌گیری به این صورت است که:

(۱) یک نمونه‌ی تصادفی ساده بدون جایگذاری $s_{1,1}$ با اندازه k از جامعه انتخاب می‌کنیم، واحدهای آن را بر اساس رتبه‌بندی قضاوتی (به صورت صعودی) مرتب کرده، کوچکترین واحد رتبه‌بندی شده‌ی متغیر کمکی $(x_{1:s_{1,1}})$ را در نظر می‌گیریم و واحد متناظر با آن در جامعه، $y_{[1:s_{1,1}]}$ ، را به عنوان اولین واحد نمونه انتخاب می‌کنیم. همه‌ی واحدها را به جامعه برمی‌گردانیم.

(۲) یک نمونه‌ی تصادفی بدون جایگذاری دیگر $s_{1,2}$ با اندازه‌ی k از جامعه انتخاب و واحدهای آن را مرتب می‌کنیم. واحد دارای دومین رتبه‌ی متغیر کمکی $x_{2:s_{1,2}}$ را در نظر می‌گیریم و واحد متناظر با آن در جامعه $(y_{[2:s_{1,2}]})$ را به عنوان واحد دوم نمونه انتخاب می‌کنیم. همه‌ی واحدها را به جامعه برمی‌گردانیم.

(۳) این فرایند را ادامه می‌دهیم تا از نمونه‌ی k ام واحد متناظر با k امین (بزرگترین) رتبه، $y_{[k:s_{1,k}]}$ ، به عنوان آخرین واحد در نمونه انتخاب شود. این مرحله‌ها نمونه‌ی مجموعه‌ی رتبه‌دار زیر را نتیجه می‌دهد:

$$rss_1 = \{y_{[1:s_{1,1}]}, \dots, y_{[k:s_{1,k}]}\}$$

با تکرار این فرایند به تعداد m چرخه، نمونه‌ی مجموعه‌ی رتبه‌دار با اندازه نمونه‌ی $n = mk$ به دست می‌آید:

$$RSS = \underbrace{\{y_{[1:s_{1,1}]}, \dots, y_{[k:s_{1,k}]} \}}_{RSS_1}, \dots, \underbrace{\{y_{[1:s_{r,1}]}, \dots, y_{[k:s_{r,k}]} \}}_{RSS_r}, \dots, \underbrace{\{y_{[1:s_{m,1}]}, \dots, y_{[k:s_{m,k}]} \}}_{RSS_m}$$

که در آن $y_{[j:s_{r,j}]}$ نشان‌دهنده‌ی واحد انتخاب شده‌ی j ام از نمونه‌ی $s_{r,j}$ در چرخه‌ی r ام است. به دلیل این‌که بعد از هر مرحله نمونه‌گیری بدون جایگذاری، همه‌ی واحدها به جامعه برگردانده می‌شوند، ممکن است نمونه شامل عضوهای تکراری باشد. این روش را نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری می‌نامیم و آن را با $RSSWR(m, k)$ نمایش می‌دهیم. این روش نمونه‌گیری سطح صفر طرح نمونه‌گیری دیش‌پاند و همکارانش [۶] است. فرض کنید RSS^* و RSS_r^* به ترتیب نشان‌دهنده‌ی نمونه‌های شامل عضوهای غیرتکراری جامعه $[i:U]$ در RSS و RSS_r ($r = 1, 2, \dots, m$) است. با فرض $g: U \rightarrow R$ جمع بر روی این مجموعه‌ها به ترتیب به صورت $\sum_{RSS^*} g(y_{[i:U]})$ و $\sum_{RSS_r^*} g(y_{[i:U]})$ می‌باشد. در زیر تعدادی از ویژگی‌های این روش را بررسی می‌کنیم.

(۱) طرح $RSSWR(m, 1)$ با طرح نمونه‌گیری تصادفی ساده با جایگذاری با m تکرار مستقل $SRSWR(m)$ برابر است.

(۲) در طرح $RSSWR(m, k)$ رتبه‌بندی اغلب بر اساس متغیری کمکی انجام می‌شود، که هم به آسانی مشاهده می‌شود و هم همبستگی بالایی با متغیر مورد مطالعه دارد.

۲-۲- محاسبه احتمال شمول تحت طرح نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری

احتمال شمول یک واحد در جامعه احتمال این است که آن واحد جامعه در نمونه انتخاب شود. در روش SRS همه واحدهای جامعه احتمال شمول یکسانی دارند. به همین دلیل است که SRS هیچ کنترلی بر روی واحدهایی که به عنوان نمونه انتخاب می‌شوند، ندارد درحالی که با استفاده از RSS می‌توان یک نمونه‌ی کنترل شده به دست آورد. احتمال

شمول مرتبه‌ی دوم تحت روش نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری اولین بار توسط جعفری جوزانی و جانسون (۲۰۱۰) [۸] مورد بررسی قرار گرفت. آن‌ها برای این تعریف، از روش نمونه‌گیری شرح داده شده توسط دیش‌پاند و همکارانش (۲۰۰۶) [۶] استفاده کردند.

۲-۲-۱- محاسبه احتمال شمول مرتبه‌ی اول و دوم

احتمال شمول مرتبه اول $\pi_{i:U}^{(m)}$ برابر است با احتمال این‌که واحد $[i:U]$ در rSS^* وجود داشته باشد. احتمال شمول مرتبه دوم $\pi_{ii':U}^{(m)}$ را، احتمال این‌که هر دو واحد $[i:U]$ و $[i':U]$ در rSS^* وجود داشته باشند، تعریف می‌کنیم.

لم ۱. فرض می‌کنیم $\alpha_{ij} = \frac{\binom{i-1}{j-1} \binom{N-i}{k-j}}{\binom{N}{k}}$ که $i = 1, 2, \dots, N$ و $j = 1, 2, \dots, k$ در این صورت احتمال شمول مرتبه اول و دوم تحت روش نمونه‌گیری شرح داده شده به صورت زیر تعریف می‌شود:

$$(۱) \quad \pi_{i:U}^{(m)} = 1 - \prod_{j=1}^k (1 - \alpha_{ij})^m$$

برای $i \neq i'$ داریم:

$$(۲) \quad \pi_{ii':U}^{(m)} = 1 - \prod_{j=1}^k (1 - \alpha_{ij})^m - \prod_{j=1}^k (1 - \alpha_{i'j})^m + \prod_{j=1}^k (1 - \alpha_{ij} - \alpha_{i'j})^m.$$

اثبات:

فرض می‌کنیم $S_{r,j}$ ، زامین نمونه‌ی تصادفی ساده (با اندازه k) در چرخه‌ی r ام است، $A_{i,r,j}$ را به صورت $A_{i,r,j} = \{x_{j:S_{r,j}} = x_{i:U}\}$ تعریف می‌کنیم. این پیشامد زمانی اتفاق می‌افتد که زامین x ترتیبی در $S_{r,j}$ با i امین مقدار x ترتیبی در جامعه برابر باشد. به آسانی می‌توان نشان داد

$$p\{A_{i,r,j}\} = p\{x_{j:s} = x_{i:U}\} = \left(\alpha_{ij} = 0 \quad i < j \text{ (اگر)} \right) \\ \frac{\binom{i-1}{j-1} \binom{N-i}{k-j}}{\binom{N}{k}} = \alpha_{ij}$$

واضح است که پیشامد $A_i = \bigcup_{r=1}^m \bigcup_{j=1}^k A_{i,r,j}$ زمانی رخ می‌دهد که $x_{i:U}$ (به همین ترتیب $y_{[i:U]}$) حداقل یک بار در RSS ظاهر شود. به دلیل استقلال بین $S_{r,j}$ ها، احتمال شمول برای $x_{i:U}$ (به همین ترتیب $y_{[i:U]}$) برابر است با

$$\pi_{i:U}^{(m)} = p(A_i) = p\left\{ \bigcup_{r=1}^m \bigcap_{j=1}^k A_{i,r,j} \right\} = 1 - p\left\{ \bigcap_{r=1}^m \bigcap_{j=1}^k A_{i,r,j}^c \right\} \\ = 1 - \prod_{j=1}^k (1 - \alpha_{ij})^m$$

برای احتمال شمول مرتبه‌ی دوم داریم

$$\pi_{ii':U}^{(m)} = p\{A_i \cap A_{i'}\} = 1 - p\{A_i^c\} - p\{A_{i'}^c\} + p\{A_i^c \cap A_{i'}^c\} \\ \text{و } p\{A_i^c \cap A_{i'}^c\} = \prod_{j=1}^k (1 - \alpha_{ij} - \alpha_{i'j})^m \text{ با استفاده از استدلال‌های یکسان بالا}$$

احتمال شمول مرتبه‌ی دوم به دست می‌آید.

۳- مطالعه کاربردی

در این بخش تلاش می‌شود تا به صورت کاربردی نشان داده شود برآوردهای هورویتز-تامپسون، هورویتز-تامپسون نسبی، هنس-هورویتز و هنس-هورویتز نسبی برای میانگین (یا مجموع) جامعه تحت طرح $RSSWR$ کاراتر از این برآوردها تحت طرح SRS است.

۳-۱- انتخاب نمونه‌ها به روش نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری

افرون (۱۹۷۹) روش نمونه‌گیری خودگردان را بر اساس ایده‌ی بازنمونه‌گیری از داده‌ها برای برآورد شاخص‌های درستی برآوردها مانند آریبی و واریانس ارائه نمود. در این

بخش با استفاده از نمونه‌های خودگردان و داده‌های آمارگیری از هزینه و درآمد خانوارهای شهری سال ۱۳۹۰ استان تهران، برتری استفاده از نمونه‌گیری از مجموعه‌ی رتبه‌دار را نسبت به نمونه‌گیری تصادفی ساده در انتخاب واحدهای نمونه‌ی مرحله‌ی اول نشان می‌دهیم. با توجه به عدم دسترسی به چارچوب آمارگیری، ابتدا با تکرار داده‌های موجود، جامعه‌ای به اندازه‌ی ۳۷۳ واحد نمونه‌گیری اولیه (مطابق با واقعیت) شبیه‌سازی می‌کنیم. با توجه به اهداف اصلی آمارگیری از هزینه و درآمد خانوارها که عمده‌ترین آن‌ها ارائه‌ی برآورد میانگین هزینه‌ی خوراکی، هزینه‌ی غیرخوراکی، هزینه‌ی کل و درآمد سالانه‌ی خانوارهای هر یک از استان‌های کشور با دقت مورد نظر است، در این بخش دقت برآورد هزینه‌ی کل و درآمد سالانه‌ی خانوارها را با استفاده از روش RSSWR در نمونه‌گیری مرحله‌ی اول مورد بررسی قرار می‌دهیم. همان‌گونه که در بخش‌های قبل بیان شد برای استفاده از این روش نمونه‌گیری اطلاعات یک متغیر کمکی با همبستگی بالا برای رتبه‌بندی واحدهای نمونه‌گیری شده مورد نیاز است. بدین منظور از اطلاع درآمد مورد انتظار خانوارها استفاده می‌کنیم. این متغیر با استفاده از اطلاعات خانوارها و امکانات و تسهیلات و ویژگی‌های آن‌ها در چارچوب نمونه‌گیری برای هر یک از خانوارهای جامعه، پیش از نمونه‌گیری برآورد شده است. زاهدیان و همکارانش [۲] ضمن معرفی این متغیر نشان دادند درآمد مورد انتظار خانوارها و درآمد خانوارها، که اطلاع آن در آمارگیری از هزینه و درآمد خانوارها جمع‌آوری می‌شود، همبستگی بالایی دارند.

در این بخش برای برآورد کارایی برآوردها از نمونه‌گیری خودگردان استفاده می‌کنیم. برای انجام مطالعه از داده‌های آمارگیری از هزینه و درآمد خانوارهای شهری ۱۳۹۰ استان تهران استفاده می‌کنیم. این آمارگیری که به صورت سالانه در تمام استان‌های کشور اجرا می‌شود یکی از آمارگیری‌های خانواری مرکز آمار ایران است که اطلاعاتی در خصوص هزینه‌ها و درآمدهای سالانه‌ی خانوارها ارائه می‌کند. اندازه‌ی نمونه‌ی این آمارگیری برای استان تهران در سال ۱۳۹۰ برابر ۱۴۶۰ خانوار در ۲۹۰ واحد نمونه‌گیری اولیه بوده است. روش نمونه‌گیری مورد استفاده در این آمارگیری نمونه‌گیری چندمرحله‌ای طبقه‌بندی شده بوده که در مرحله‌ی اول واحدهای نمونه‌گیری اولیه به روش تصادفی ساده از بین ۳۷۳ واحد جامعه انتخاب شده است (برای مطالعه‌ی جزئیات به طرح نمونه‌گیری آمارگیری از هزینه و درآمد خانوارها، ۱۳۹۰ مراجعه کنید). متغیر درآمد مورد انتظار خانوارها را به دلیل همبستگی بالا با متغیر درآمد ($\rho = 0.83$) به عنوان متغیر رتبه‌بندی انتخاب

می‌کنیم. احتمال شمول واحدهای جامعه را بر اساس روش RSSWR که در بخش ۲-۲-۱ بیان کردیم، به صورت زیر محاسبه می‌کنیم

$$\pi_{i:U}^{(m)} = 1 - \prod_{j=1}^n (1 - \alpha_{ij})^m \quad i = 1, 2, \dots, 373$$

از جامعه در اندازه‌های مختلف $n = 100, 200, 300$ نمونه‌هایی به روش نمونه‌گیری از مجموعه‌ی رتبه‌دار با جایگذاری انتخاب می‌کنیم و کارایی برآوردها را در ۵۰۰۰ بار تکرار نمونه‌گیری می‌سنجیم.

۲-۳-۲- برآوردهای در نظر گرفته شده تحت طرح RSSWR

۳-۲-۱- راوردگر هوریتز-تامپسون

بر اساس طرح RSSWR برآوردها $\hat{t}_{y,RSS,\pi} = \sum_{r_{SS}^*} \frac{y_{[i:U]}}{\pi_{i:U}}$ برای \hat{t}_y (مجموع جامعه) ناریب است (جعفری جوزانی و جانسون، ۲۰۱۰). با توجه به هدف این بررسی که برآورد میانگین متغیرهای هزینه و درآمد خانوارها است، برآورد میانگین در این روش را به‌طور مشابه به صورت زیر تعریف می‌کنیم:

$$(3) \quad \hat{\mu}_{y,RSS,\pi} = \sum_{r_{SS}^*} \frac{y_{[i:U]}}{\pi_{i:U}} / \sum_{r_{SS}^*} \frac{1}{\pi_{i:U}}$$

که در آن $y_{[i:U]}$ مقادیر متغیر هزینه یا درآمد در هر نمونه‌ی شامل عضوهای غیر تکراری است و $\pi_{i:U}$ احتمال شمول واحد i ام است.

۳-۲-۲- برآوردها نسبتی هوریتز-تامپسون

بر اساس طرح RSSWR برآوردها نسبتی هوریتز-تامپسون مجموع جامعه به صورت زیر است [۸]:

$$\hat{t}_{y,RSS,\pi,ra} = t_z \frac{\hat{t}_{y,RSS,\pi}}{\hat{t}_{z,RSS,\pi}}$$

که در آن t_z مجموع مقادیر متغیر کمکی در واحدهای جامعه است. برآوردها میانگین در این روش را به‌طور مشابه به صورت زیر تعریف می‌کنیم:

$$(۴) \quad \hat{\mu}_{y,RSS,\pi,ra} = \mu_z \frac{\hat{\mu}_{y,RSS,\pi}}{\hat{\mu}_{z,RSS,\pi}}$$

که در آن $\hat{\mu}_{y,RSS,\pi}$ و $\hat{\mu}_{z,RSS,\pi}$ به ترتیب برآورد میانگین متغیرهای اصلی و کمکی به روش برآورد هورویتر-تامپسون برای میانگین جامعه هستند و μ_z ، میانگین معلوم مقادیر متغیر کمکی برای واحدهای جامعه است. در این بخش برای محاسبه‌ی برآوردگرهای نسبتی از اطلاع درآمد مورد انتظار خانوارها که مقدار آن برای کل واحدهای جامعه معلوم است، به عنوان اطلاع متغیر کمکی استفاده می‌شود.

۳-۲-۳- برآوردگر هنسن-هورویتر

برآورد مجموع جایگزین برآوردگر هنسن-هورویتر برای t_y (مجموع مقادیر واحدها در جامعه) ناریب است و به صورت زیر تعریف می‌شود [۸]:

$$\hat{t}_{y,RSS,pwr} = \frac{N}{mk} \sum_{r=1}^m \sum_{j=1}^k y_{j:sr,j}$$

با توجه به این که تعداد چرخه‌ها در این مطالعه برابر یک در نظر گرفته شده است، برآوردگر میانگین در این روش را به طور مشابه به صورت زیر تعریف می‌کنیم:

$$(۵) \quad \hat{\mu}_{y,RSS,pwr} = \frac{1}{k} \sum_{j=1}^k y_{j:sr,j}$$

که در آن $y_{j:sr,j}$ مقدار متغیر اصلی برای واحد j ام در نمونه است.

۳-۲-۴- برآوردگر هنسن-هورویتر نسبتی

تحت طرح RSSWR با متغیر مورد مطالعه درآمد و هزینه خانوارها و متغیر کمکی و رتبه‌بندی درآمد مورد انتظار خانوارها برآوردگر هنسن-هورویتر نسبتی تقریباً ناریب برای t_y به صورت زیر است [۸]:

$$\hat{t}_{y,RSS,pwr,ra} = t_z \frac{\hat{t}_{y,RSS,pwr}}{\hat{t}_{z,RSS,pwr}}$$

برآوردگر میانگین در این روش به طور مشابه به صورت زیر محاسبه می‌شود:

$$(۶) \quad \hat{\mu}_{y,RSS,pwr,ra} = \mu_z \frac{\hat{\mu}_{y,RSS,pwr}}{\hat{\mu}_{z,RSS,pwr}}$$

که در آن $\hat{\mu}_{y,RSS,pwr}$ و $\hat{\mu}_{z,RSS,pwr}$ به ترتیب برآورد میانگین متغیرهای مورد مطالعه (درآمد یا هزینه) و متغیر کمکی (درآمد مورد انتظار) به روش هنس-هورویتز برای میانگین جامعه هستند.

۳-۳- کارایی و مقایسه برآوردها

در این بخش کارایی برآوردها را برای میانگین درآمد و هزینه‌ی خانوارها ارزیابی می‌کنیم. برآوردها را به صورت $\hat{\mu} = \frac{1}{5000} \sum_{i=1}^{5000} \hat{\mu}_i$ محاسبه می‌کنیم. برای محاسبه برآوردها از نرم‌افزار SAS استفاده می‌کنیم. اثر طرح برآوردها را معیاری برای کارایی آن‌ها در نظر می‌گیریم. با توجه به نااریب (یا تقریباً نااریب) بودن برآوردها در این بخش فرض می‌شود میانگین توان دوم خطاهای هر برآوردها تقریباً برابر واریانس آن برآوردها است. برای هر برآوردها تعریف شده در این بخش با اندازه نمونه‌های مختلف، اثر طرح را به صورت زیر محاسبه می‌کنیم.

$$deff(\hat{\mu}) = \frac{Var(\hat{\mu})}{\left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n}}$$

که در آن $n = \sum_U \pi_{i:U}$ و مخرج، واریانس برآوردها تحت طرح نمونه‌گیری تصادفی ساده بدون جایگذاری (SRSWOR) است. صورت کسر به روش زیر برآورد می‌شود. با توجه به نااریب یا تقریباً نااریب بودن برآوردها در این بخش فرض می‌شود میانگین توان دوم خطاها برای هر برآوردها تقریباً برابر واریانس آن برآوردها است.

$$Var(\hat{\mu}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\mu}_i - \bar{\mu})^2 \approx M\hat{S}E(\hat{\mu})$$

که در آن $\hat{\mu}_i$ مقدار برآورد مورد نظر در تکرار i ام نمونه‌گیری ($i = 1, 2, \dots, 5000$) و $\bar{\mu}$ میانگین این برآوردها است. جدول‌های ۱ تا ۳ اثر طرح برآوردهای ۳ تا ۶ را با فرض رتبه‌بندی ناکامل، با متغیر درآمد مورد انتظار خانوارها وقتی اندازه‌ی نمونه به ترتیب ۱۰۰، ۲۰۰ و ۳۰۰ است، نشان می‌دهد.

جدول ۱- اثر طرح برآوردگرها تحت رتبه‌بندی با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 100$

متغیر مورد مطالعه	$\rho(x, y)$	$cv(x)/cv(y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۰٫۷۷۵	۰٫۰۲۸	۰٫۰۲۹	۰٫۴۱۲	۰٫۳۸۱
هزینه	۰٫۲۷	۰٫۶۹۲	۰٫۰۹۱	۰٫۰۹۲	۱٫۱۷۱	۱٫۱۷۹

جدول ۲- اثر طرح برآوردگرها تحت طرح رتبه‌بندی با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 200$

متغیر مورد مطالعه	$\rho(x, y)$	$cv(x)/cv(y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۰٫۷۷۵	۰٫۴۴۲	۰٫۲۹۵	۰٫۶۰۰	۰٫۵۹۹
هزینه	۰٫۲۷	۰٫۶۹۲	۰٫۹۱۸	۰٫۹۳۲	۱٫۸۶۹	۱٫۸۶۹

با مشاهده‌ی این جدول‌ها می‌توان نتیجه گرفت:

- (۱) وقتی همبستگی بین متغیر مورد مطالعه و متغیر کمکی بالا است، اثر طرح مقدار کمی دارد. به عبارت دیگر استفاده از نمونه‌گیری از مجموعه‌ی رتبه‌دار و برآوردگرهای تعریف شده کارایی بیشتری دارند. همان‌طور که مشاهده می‌شود، همبستگی بالا بین متغیر درآمد و درآمد مورد انتظار، کارایی بهتری را نتیجه می‌دهد.
- (۲) برای متغیر درآمد که همبستگی بالایی با متغیر رتبه‌بندی دارد، برآوردگر هورویتز-تامپسون نسبتی کارایی بیشتری نسبت به سایر برآوردگرها دارد. در این حالت ناپایداری برآوردگر $\hat{\mu}_{y,RSS,\pi}$ با توجه به این‌که هر دو برآوردگرهای $\hat{\mu}_{y,RSS,\pi}$ و $\hat{\mu}_{z,RSS,\pi}$ دارای اندازه نمونه یکسانی هستند، کاهش می‌یابد.

جدول ۳- اثر طرح برآوردگرها تحت رتبه‌بندی با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 300$

متغیر مورد مطالعه	$\rho(x, y)$	$cv(x)/cv(y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۰٫۷۷۵	۰٫۴۸۸	۰٫۲۸۳	۱٫۲۸۷	۱٫۲۹۵
هزینه	۰٫۲۷	۰٫۶۹۲	۰٫۹۰۷	۰٫۹۳۵	۴٫۰۴۰	۴٫۰۵۵

- (۳) برآوردگر $\hat{\mu}_{y,RSS,pwr,ra}$ نسبتاً خوب عمل می‌کند. اگرچه برآوردگر $\hat{\mu}_{y,RSS,\pi,ra}$ برتری قابل توجهی دارد اما برآوردگر $\hat{\mu}_{y,RSS,pwr,ra}$ تنها نیاز به دانستن x_i ها در نمونه دارد در حالی که برآوردگر $\hat{\mu}_{y,RSS,\pi,ra}$ نیازمند دانستن همه‌ی x_i ها در محاسبه احتمال‌های شمول است.

۴) وقتی اندازه‌ی نمونه افزایش می‌یابد و به اندازه‌ی جامعه نزدیک‌تر می‌شود کارایی نمونه‌گیری از مجموعه‌ی رتبه‌دار در جامعه‌ی متناهی نسبت به نمونه‌گیری تصادفی ساده بدون جایگذاری کاهش می‌یابد. این موضوع در جدول‌های ۱ تا ۳ مشاهده می‌شود. علاوه بر اثر طرح در ادامه برای مقایسه کارایی روش‌ها میانگین توان دوم خطاهای نسبی روش‌ها در اندازه نمونه‌های مختلف به صورت زیر محاسبه می‌شود.

$$RMSE(\hat{\mu}) = \frac{MSE(\hat{\mu})}{\bar{\mu}_U}$$

که در آن $\bar{\mu}_U$ میانگین جامعه است.

جدول ۴- کارایی نسبی برآوردها با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 100$

متغیر مورد مطالعه	$\rho(x, y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۸۱۰۰۹	۶۱۴۷۲	۷۴۵۵۶	۶۸۸۷۹
هزینه	۰٫۲۷	۲۶۷۱۵۴	۲۶۹۱۲۸	۲۹۲۵۴۹	۲۹۴۳۶۶

با توجه به جدول‌های ۴، ۵ و ۶ نیز می‌توان نتیجه‌های مشابه جدول‌های ۱، ۲ و ۳ به دست آورد.

جدول ۵- کارایی نسبی برآوردها با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 200$

متغیر مورد مطالعه	$\rho(x, y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۳۹۹۰۳	۲۶۶۴۳	۳۴۳۹۰	۱۶۶۱۳۲۰
هزینه	۰٫۲۷	۱۱۴۴۲۵	۱۱۶۲۰۱	۱۴۷۹۱۱	$۵۲۰۸۱۸E + ۱۵$

جدول ۶- کارایی نسبی برآوردها با متغیر درآمد مورد انتظار و اندازه‌ی نمونه $n = 300$

متغیر مورد مطالعه	$\rho(x, y)$	$\hat{\mu}_{y,RSS,\pi}$	$\hat{\mu}_{y,RSS,\pi,ra}$	$\hat{\mu}_{y,RSS,pwr}$	$\hat{\mu}_{y,RSS,pwr,ra}$
درآمد	۰٫۸۳	۲۴۲۹۳	۱۴۰۶۳	۲۰۷۵۴	۲۰۸۷۸
هزینه	۰٫۲۷	۶۲۲۶۷	۶۴۲۲۵	۹۰۱۲۵	$۸۷۴۳۶۵E + ۱۱۱$

۴- نتیجه‌گیری

در این مقاله از داده‌های آمارگیری از هزینه و درآمد خانوارهای شهری سال ۱۳۹۰ در استان تهران استفاده کردیم و با به‌کارگیری روش نمونه‌گیری خودگردان یک مطالعه‌ی شبیه‌سازی با ۵۰۰۰ مرتبه تکرار نمونه‌گیری انجام دادیم. برآوردهای طرح‌پایه برای متغیرهای درآمد و هزینه‌ی خانوارها را تحت طرح RSSWR به دست آوردیم. کارایی این

برآوردگرها به همبستگی بین متغیر مورد مطالعه و متغیر رتبه‌بندی بستگی دارد. برآوردگرهای $\hat{\mu}_{y,RSS,\pi,ra}$ و $\hat{\mu}_{y,RSS,pwr,ra}$ کارایی خوبی دارند. مشاهدات نتیجه می‌دهد که وقتی متغیر کمکی همبستگی بالایی با متغیر مورد مطالعه دارد، اثر طرح برآوردگرها (در مقایسه با SRSWOR) کمتر از یک است. نتایج نشان می‌دهد که استفاده از RSSWR در عمل به‌عنوان یک روش بهتر که منجر به بهبود دقت برآوردگرهای میانگین جامعه‌ی متناهی می‌شود، منطقی است. البته باید توجه داشت که هرچه اندازه نمونه بیش‌تر می‌شود، کارایی برآوردگرها تحت طرح RSSWR نسبت به تصادفی ساده کمتر می‌شود. همچنین بهبود دقت برآوردگرها در این روش نمونه‌گیری مستلزم استفاده از متغیر رتبه‌بندی مناسب با ضریب همبستگی بالا با متغیر مورد مطالعه است.

مرجع‌ها

- [۱] دستورالعمل طرح آمارگیری (۱۳۹۰). آمارگیری از هزینه و درآمد خانوار شهری/روستایی. مرکز آمار ایران.
- [۲] زاهدیان، علیرضا؛ علی‌اکبری صبا، روشنگر، دانش‌پرور، نیما و باقری، فرید (۱۳۹۲). تعیین توزیع دو دهک پایین جامعه در مناطق جغرافیایی. پژوهشکده آمار، تهران.
- [۳] علی‌اکبری صبا، روشنگر (۱۳۸۷). روش نمونه‌گیری از مجموعه‌ی رتبه‌دار و برخی کاربردهای آن. رساله‌ی دکتری، دانشگاه آزاد اسلامی، واحد علوم تحقیقات تهران.
- [4] Al-Saleh, M.F. and Samawi, H.M. (2007). A note on inclusion Probability in Ranked Set Sampling and Some of Its Variations, *Test*, **16**, 198-209.
- [5] Barabesi, L. and Pisani, C. (2002). Ranked Set Sampling for Replicated Sampling Designs, *Biometrics*, **58**, 586-592.
- [6] Deshpand, J.V., Frey, J., and Ozturk, O. (2006). Nonparametric Ranked Set Sampling Confidence Intervals for Quantiles of a Finite Population, *Environ, Ecol. Stat.*, **13**, 25-40.
- [7] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics*, **7**, 1-26
- [8] Jafari Jozani, M. and Johnson, B.C. (2010). Design Based Estimation

- for Ranked Set Sampling in Finite Population, Environ. Ecol. Statistic, **18**, 663-685.
- [9] McIntyre, G.A. (1952). A Method for Unbiased Selective Sampling Using Ranked Sets, Aust. J. Agric. Res., **3**, 385-390.
- [10] Patil, G.P., Sinha, A.K. and Tailie, C. (1995). Finite Population corrections for Ranked Set Sampling, J. Ann. Inst. Stat., **47**, 621-638.
- [11] Stokes, S.L. (1977). Ranked Set Sampling with Conomitant Variables, Communications in Statics-Theory and Methods, **12**, 1207-1211.
- [12] Takahasi, K. and Futatsuya, M. (1988). Ranked Set Sampling from Finite Population, Prok. Inst. Stat. Math, **36**, 55-65.

زهرا شیرکوند

فوق لیسانس آمار اقتصادی و اجتماعی
تهران، ورامین، دانشگاه علمی-کاربردی.
رایانشانی: h90.shirkavand@gmail.com

روشنگ علی اکبری صبا

دکترای آمار
تهران، فاطمی، باباطاهر، پژوهشکدهی آمار
رایانشانی: roshanaksaba22@gmail.com