# کاربردی از نمونه گیری غیرمستقیم با استفاده از روش سهم وزن تعمیم یافته

سهیلا جبلی‌فرد ، حمید رضا نواب پور[۱]

چکیده:

برای انتخاب نمونه‌های مورد نیاز در آمارگیری‌های اجتماعی – اقتصادی، لازم است، چارچوب نمونه گیری از پیش مشخص باشد. متأسفانه در برخی موارد ممکن است فهرستی از واحدهای مورد نظر وجود نداشته باشد، اما فهرست دیگری شامل واحدهایی که با واحدهای جامعه‌ی هدف در ارتباط هستند، موجود باشد. در این حالت می‌توان با انتخاب یک نمونه از جامعه‌ای که دارای چارچوب کامل است و با استفاده از تناظر موجود بین این جامعه و جامعه‌ی هدف، پارامترهای جامعه‌ی هدف را برآورد کرد. اگر اتصال‌های بین واحدهای دو جامعه یک به یک نباشند، مسئله‌ی وزن برآورد و یا احتمال گزینش برای واحدهای مورد بررسی در جامعه‌ی هدف نیز مطرح می‌شود. در این مقاله ضمن مروری بر نمونه‌گیری غیرمستقیم، روش اتصال رکوردی و روش سهم وزن تعمیم یافته، کاربرد آنها در دو جامعه‌ای که از طریق روش اتصال رکوردی با یکدیگر مرتبط شده‌اند، بررسی می‌شود.

**واژه‌های کلیدی:** نمونه گیری غیرمستقیم، روش سهم وزن تعمیم یافته، اتصال رکوردها، وزن اتصال.

## ۱ مقدمه

روش‌های آمارگیری نمونه‌ای از جامعه‌هایی که شامل افرادی با ویژگی نادر (ابتلا به $HIV^+$)، جمعیت‌هایی با یک خصیصه‌ی حساس (اعتیاد به مواد مخدر)، و یا واحدهای متحرک و سیار (عشایر و مهاجرین غیرقانونی) هستند، در چند دهه‌ی اخیر بسیار توسعه یافته‌اند. در این گونه آمارگیری‌ها، چون نسبت افراد مورد نظر در جامعه‌ی اصلی کوچک است، لذا تعداد اندکی از آنها در نمونه قرار می‌گیرند. این امر باعث افزایش خطای نمونه گیری می‌شود. از سوی دیگر با افزایش اندازه‌ی نمونه برای به دست آوردن تعداد بیشتری از افراد مورد نظر، هزینه‌های مربوط به گردآوری داده‌ها افزایش می‌یابند. مشکل دیگر در این گونه آمارگیری‌ها، فقدان چارچوب و یا ناقص بودن آن است. چون برای انتخاب واحدهای نمونه‌ای وجود چارچوب لازم است، لذا استفاده از روش‌های متداول نمونه گیری با هزینه و یا خطای زیاد همراه است. یکی از طرح‌هایی که در این گونه موارد استفاده می‌شود، نمونه گیری غیرمستقیم[۲] است. روش‌های مختلفی در این نوع از نمونه گیری معرفی شده‌اند که عبارتند از نمونه گیری گلوله برفی [۳]، نمونه گیری خوشه‌ای سازوار [۹] و نمونه گیری شبکه‌ای [۱۰]. در هر یک از این روش‌ها، مفهوم اتصال بین واحدهای دو جامعه‌ی مرتبط، مورد توجه است. اگر اتصال‌های

[۱] دانشیار گروه آمار– دانشگاه علامه طباطبایی
[۲] Indirect Sampling

بین واحدهای دو جامعه یک به یک نباشند، مسئله‌ی وزن برآورد و یا احتمال گزینش برای واحدهای نمونه‌ای در جامعه‌ی هدف مطرح می‌شود. برای حل این مسئله روش سهم وزن تعمیم یافته[۳]، توسط لاوالی و کارون [۴]، لاوالی و دویل [۶]، لاوالی [۷] و لاوالی و دویل [۵] معرفی شده است.روش سهم وزن تعمیم یافته یک وزن برآورد برای واحدهای مورد بررسی در جامعه‌ی هدف فراهم می‌کند. این وزن برآورد متناظر با یک میانگین موزون از وزن‌های واحدهای نمونه‌ی اولیه است. در این مقاله ابتدا در بخش ۲ نحوه‌ی به دست آوردن برآورد در جامعه‌ی هدف را با استفاده از روش سهم وزن تعمیم یافته با ذکر مثال توضیح داده، سپس در بخش ۳، روش‌های اتصال رکوردی و نحوه‌ی محاسبه وزن‌های اتصال بین واحدهای دو جامعه‌ای که با یکدیگر در اتصالند، تشریح می‌شوند. در بخش ۴ توضیح می‌دهیم که چگونه می‌توان با استفاده از وزن‌های اتصال به دست آمده از طریق روش اتصال رکوردی در دو جامعه‌ای که از این طریق با یکدیگر اتصال برقرار کرده‌اند، و به کار بردن روش سهم وزن تعمیم یافته برآوردگر جامعه‌ی هدف را محاسبه کرد. در بخش آخر با ارائه یک مثال کاربردی برآورد پارامتر جامعه‌ی هدف را با استفاده از روش نمونه گیری غیرمستقیم به دست آورده و با مقدار به دست آمده‌ی آن از طریق یکی از روش‌های نمونه گیری کلاسیک مقایسه خواهیم کرد.

## ۲ برآورد در نمونه گیری غیرمستقیم با استفاده از روش سهم وزن تعمیم یافته

فرض می‌شود جامعه‌ی $U^A$ با جامعه‌ی هدف $U^B$، در ارتباط است، در واقع بین واحدهای جامعه‌ی $U^A$ و واحدهای جامعه‌ی $U^B$ تناظر یا اتصال برقرار است. این اتصال‌ها می‌توانند یک به یک، چند به یک و یا چند به چند باشند. برای مثال جامعه‌ی $U^A$ می‌تواند جامعه‌ی والدین و جامعه‌ی $U^B$ جامعه‌ی فرزندان آنها باشند. نمونه‌ی $s^A$ به اندازه‌ی $m^A$ واحد از جامعه‌ی $U^A$ به اندازه‌ی $M^A$ با استفاده از یکی از طرح‌های نمونه‌گیری انتخاب می‌شود. احتمال انتخاب واحد $j$ برابر با $\pi_j^A$ است، به طوری که برای هر $j \in U^A$: $\pi_j^A > 0$. جامعه‌ی $U^B$ شامل $M^B$ واحد است. این جامعه شامل $N$ خوشه‌ی $U_i^B$، $i = 1, \ldots, N$ و خوشه‌ی $i$ شامل $M_i^B$ واحد است، همچنین $U^B = \bigcup_{i=l}^{N} U_i^B$. فرض می‌شود که بین واحد $j$ از جامعه‌ی $U^A$ و واحد $k$ در خوشه‌ی $i$ از جامعه $U^B$ اتصال برقرار است. این اتصال توسط متغیر نشانگر $lj,ik$ به صورت زیر تعریف می‌شود: اگر یک اتصال بین $j \in U^A$ و $ik \in U^B$ وجود داشته باشد: $l_{j,ik} = 1$ و در غیر این صورت صفر است.

محدودیت زیر برای این‌که روش سهم وزن تعمیم یافته[۴] برآورد نااریب تولید کند، اساسی است [۷]: هرخوشه‌ی $i$ از $U^B$ باید حداقل با یک واحد $j$ از $U^A$ در اتصال باشد. بنابراین تعداد اتصال‌های واحدهای $U^A$ با

[۳] Generalized Weight Share Method
[۴] GWSM

هر خوشه‌ی $i$ از $U^B$ باید:

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} > 0, \quad i = 1, \ldots, N \qquad (1)$$

برای هر واحد $j$ که در $s^A$ گزینش می‌شود، واحدهای $ik$ در $U^B$ که یک اتصال غیرصفر با واحد $j$ دارند، مشخص می‌شوند. برای هر واحد مشخص شده‌ی $ik$ می‌توان به فهرست $M_i^B$ واحد در خوشه $i$ دسترسی پیدا کرد. مجموعه‌ی $n$ خوشه‌ی مشخص شده در $U^B$ توسط واحدهای $j \in s^A$ را با $\Omega^B$ نمایش می‌دهیم. هدف به دست آوردن برآورد مقدار کل برای $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$، در جامعه‌ی $U^B$ برای صفت $y$ است.

## گام‌های روش سهم وزن تعمیم یافته

این روش شامل گام‌های زیر است:

**گام اول:** برای هر واحد $k$ از خوشه‌ی $i \in \Omega^B$ وزن اولیه $w'_{ik}$ به صورت زیر تعیین می‌شود:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$$

که در آن:

$$t_j = \begin{cases} 1 & j \in s^A \\ 0 & j \notin s^A. \end{cases}$$

ابتدا مشخص می‌شود که هر یک از اعضای جامعه‌ی $U^A$ با کدام خوشه از $U^B$ در ارتباط است. سپس هر یک از اعضای خوشه‌ی در اتصال که با اعضای نمونه در اتصال هستند، در وزن اولیه به کار برده می‌شوند.

**گام دوم:** برای هر $k$ از خوشه‌ی $i \in \Omega^B$ تعداد اتصال‌های $L_{ik}^B$ محاسبه می‌شود. کمیت $L_{ik}^B$ تعداد اتصال‌های بین واحدهای $U^A$ و واحد $k$ درخوشه‌ی $i$ام از $U^B$ را نمایش می‌دهد.

$$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$$

**گام سوم:** وزن نهایی $w_i$ از رابطه‌ی زیر به دست می‌آید:

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$$

**گام چهارم:** برای هر $k \in U_i^B$، $w_{ik} = w_i$ در نظر گرفته می‌شود، یعنی به هر عضو $k$ از خوشه‌ی $i$ وزن $w_i$ نسبت داده می‌شود.

$$\begin{aligned} w_{ik} &= \sum_{k=1}^{M_i^B} \frac{w'_{ik}}{L_{ik}^B} \sum_{k=1}^{M_i^B} \frac{1}{L_{ik}^B} \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik} \\ &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{1}{L_{ik}^B} \sum_{k=1}^{M_i^B} l_{j,ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{L_{j,i}}{L_{ik}^B}. \end{aligned}$$

که در آن $L_{j,i} \sum_{k=1}^{M_i^B} l_{j,ik}$. یک برآوردگر برای $Y^B$، برآوردگر هورویتز– تامپسون است که به صورت زیر تعریف می‌شود:

$$Y_H^B T = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} \frac{1}{\pi_{ik}^B} l y_{ik}$$

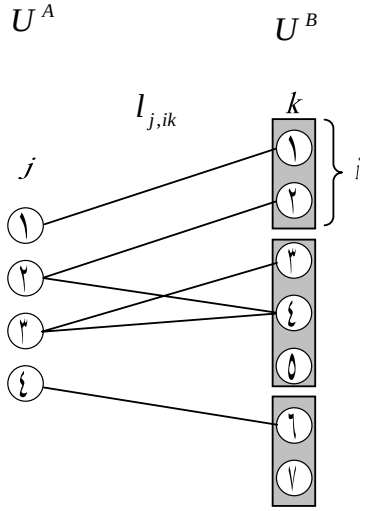
که در آن $\pi_{ik}^B$ احتمال شمول واحد $k$ از خوشه‌ی $i$ در جامعه‌ی $U^B$ و $y_{ik}$ اندازه‌ی صفت مورد نظر فرد $k$ در خوشه‌ی $i$ است. برای محاسبه $Y_{HT}^B$ باید اطلاع دقیقی از $\pi_{ik}^B$ در دست باشد، ولی چون چارچوب جامعه‌ی $U^B$ در دست نیست، لذا محاسبه مقادیر $\pi_{HT}^B$ میسر نمی‌باشد. برای محاسبه اریبی و یا واریانس برآوردگر $Y^B$ لازم است تعریف جدیدی ارائه شود: اگر برای هر $k \in i$، تعریف کنیم $Z_{ik} = \frac{Y_i}{L_i^B}$ که در آن $Y_i = \sum_{i=1}^{M_i^B} y_{ik}$ و $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$، آن‌گاه $\hat{Y}^B$ به صورت زیر به دست می‌آید:

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j$$

که در آن $\hat{Y}^B$ کمیت برآوردگر هورویتز — تامپسون متغیر $Z^j$ است. با توجه به این تعریف جدید $\hat{Y}^B$ می‌تواند هم به فرم تابعی از عناصر $ik$ در جامعه‌ی $U^B$ نوشته شود و هم به صورت تابعی از واحدهای $j$ در جامعه‌ی $U^A$. در این صورت واریانس $\hat{Y}^B$ از رابطه‌ی زیر به دست می‌آید:

$$Var(\hat{Y}^B) = \sum_{j=1}^{M^A} \sum_{t=1}^{M^A} \frac{(\pi_{jt}^A - \pi_j^A \pi_t^A)}{\pi_j^A \pi_t^A} Z_j Z_t$$

که در آن $\pi_{jt}$ احتمال توام گزینش واحدهای $j$ و $t$ است. برآوردگر $\hat{Y}^B$ نسبت به طرح نمونه گیری برای $Y^B$ نااریب است [۵] . برای نشان دادن نحوه‌ی محاسبه‌ی وزن نهایی، مثال ساده‌ای ارائه می‌شود. شکل زیر نشان می‌دهد که واحدهای $j = ۱$ و $j = ۲$ گزینش شده‌اند. برای $j = ۱$ خوشه‌ی $i = ۱$ و برای $j = ۲$ خوشه‌های $i = ۱$ و $i = ۲$ مشخص شده‌اند.




**شکل ۱. مثالی از نحوه‌ی اتصال‌های دو جامعه**

در جدول ۱ وزن اولیه، تعداد اتصال‌ها و در آخر وزن نهایی برای هر واحد گزینش شده از جامعه‌ی $U^B$ در شکل ۱ محاسبه شده‌اند.

**جدول ۱. نحوه‌ی محاسبه وزن‌های هر خوشه**

| خوشه‌ی $i$ | عضو $k$ از خوشه‌ی $i$ ام | $w'_{ik}$ | $L_{ik}^B$ | $w_i$ |
|---|---|---|---|---|
| ۱ | ۱ | $\frac{۱}{\pi_۱^A}$ | ۱ | $\frac{۱}{۲}[\frac{۱}{\pi_۱^A} + \frac{۱}{\pi_۲^A}]$ |
| ۱ | ۲ | $\frac{۱}{\pi_۲^A}$ | ۱ | $\frac{۱}{۲}[\frac{۱}{\pi_۱^A} + \frac{۱}{\pi_۲^A}]$ |
| ۲ | ۱ | ۰ (زیرا $t_۳ = ۰$) | ۱ | $\frac{۱}{۳}[۰ + \frac{۱}{\pi_۲^A} + ۰] = \frac{۱}{۳\pi_۲^A}$ |
| ۲ | ۲ | $\frac{۱}{\pi_۲^A} + ۰ = \frac{۱}{\pi_۲^A}$ | ۲ | $\frac{۱}{۳\pi_۲^A}$ |
| ۲ | ۳ | ۰ (زیرا برای هر $j$، $l_{j۲۳} = ۰$) | ۰ | $\frac{۱}{۳\pi_۲^A}$ |

در نتیجه برآورد مقدار کل $Y^B$ به صورت زیر به دست می‌آید:

$$\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik}$$

$$= \frac{۱}{۲}[\frac{۱}{\pi_۱^A} + \frac{۱}{\pi_۲^A}]y_{۱۱} + \frac{۱}{۲}[\frac{۱}{\pi_۱^A} + \frac{۱}{\pi_۲^A}]y_{۱۲} + \frac{y_{۱۲}}{۳\pi_۲^A} + \frac{y_{۲۲}}{۳\pi_۲^A} + \frac{y_{۲۳}}{۳\pi_۲^A}$$

راه دیگر برای محاسبه $\hat{Y}^B$ از طریق محاسبه $Z_j$ ها است. در این مثال: $z_{۱۱} = z_{۱۲} = \frac{y_{۱۱}+y_{۱۲}}{۲}$

و $z_{۲۱} = z_{۲۲} = z_{۲۳} = \frac{y_{۲۱}+y_{۲۲}+y_{۲۳}}{۳}$ زیرا $L_۱ = ۲$، $L_۲ = ۳$ و $Y_۱ = y_{۱۱} + y_{۱۲}$، $Y_۲ = y_{۲۱} + y_{۲۲} + y_{۲۳}$.

در نتیجه:

$$\begin{aligned}\hat{Y}^B &= \frac{Z_۱}{\pi_۱^A} + \frac{Z_۲}{\pi_۲^A} = \frac{Z_۱ ۱}{\pi_۱^A} + (\frac{Z_{۱۲}}{\pi_۲^A} + \frac{Z_{۲۲}}{\pi_۲^A}) \\ &= \frac{y_{۱۱}+y_{۱۲}}{۲\pi_۱^A} + \frac{y_{۱۱}+y_{۱۲}}{۲\pi_۲^A} + \frac{y_{۲۱}+y_{۲۲}+y_{۲۳}}{۳\pi_۲^A}\end{aligned}$$

اگر رابطه‌ی (۱) برقرار نباشد، یعنی خوشه‌ای در $U^B$ وجود داشته باشد که در اتصال با هیچ واحدی از $U^A$ نباشد، وزن برآورد، مقدار $Y^B$ را کم برآورد می‌کند. یک راه حل برای تعدیل اریبی، ادغام این خوشه با خوشه‌های دیگر است. این ادغام به این صورت انجام می‌شود که چون خوشه‌ی مورد نظر دارای هیچ اتصالی با واحدهای $U^A$ نیست، بنابراین اگر با خوشه‌ی دیگری ادغام شود، در متغیر $Z_{ik} = \frac{Y_i}{L_i^B} B$ مقدار $Y_i$ افزایش ولی مقدار $L_i^B$ ثابت می‌ماند، لذا مقدار واریانس افزایش می‌یابد. بنابراین ادغام این خوشه با خوشه‌ای باید صورت گیرد که این مقدار افزایش در واریانس مینیمم شود. به این ترتیب نااریبی برآوردگر حاصل از این روش تضمین می‌شود [۴].

## ۳ اتصال رکوردی

اغلب سازمان‌ها نیاز به تشخیص چندبارگی‌ها[۵] در پایگاه‌های اطلاعاتی بزرگ دارند. در یک دفتر ثبت جمعیت، برخی موجودیت‌ها (افراد یا شرکت‌ها) ممکن است تحت دو یا چند شماره ثبتی فهرست شوند. برای تشخیص چندبارگی‌ها، نام، نشانی، تاریخ تولد و سایر اطلاعات کمکی مورد نیاز است. چندبارگی‌ها در یک فایل زمانی به وجود می‌آیند که آن فایل توسطیک منبع خارجی به روز شود، در حالی که شناسه‌های واحدها (شماره‌های ثبتی) در دسترس نیستند و یا با خطا ثبت شده‌اند.

اتصال رکوردی[۶] روش شناسی ادغام رکوردهای متناظر از دو یا چند فایل و یافتن چندبارگی‌های درون یک فایل است. اتصال رکوردی ابزار مهمی در تولید داده‌های آماری به خصوص در ارتباط با سرشماری‌ها است. افزایش کیفیت داده‌ها با استفاده از دفترهای ثبت جمعیت به جای سرشماری، با جانهی داده‌های گمشده یا به روز درآوردن آنها، بهبود در پوشش سرشماری‌ها برای جلوگیری از کم گزارش دهی و ایجاد چارچوب از کاربردهای مهم اتصال رکوردی است [۱۱]. در علوم کامپیوتر یک رکورد، ساختار داده‌ای در رابطه با یک موجودیت است که شامل چند فیلد مجزا می‌باشد. هر یک از این فیلدها را می‌توان از طریق نام‌شان فراخوانی کرد. رکوردها در یک منبع اطلاعاتی نمایشگر مشاهده‌ی موجودیت‌های یک جامعه‌ی خاص هستند. یک شناسه‌ی یکتا و یا یک متغیر، یک فیلد مشترک بین دو پایگاه داده است که توسطآن فرایند جورکردن رکوردهای مربوطبه یک موجودیت، با مقایسه دو به دوی رکوردهای دو فایل انجام می‌شود. فیلدهای به کار گرفته شده برای مقایسه می‌توانند نام، نشانی، جنس، سن و ... باشند.

نیوکمب مفاهیم اتصال رکوردی را بر مبنای نسبت بخت‌های فراوانی‌ها و قاعده‌های تصمیم برای تعیین

[۵] Duplication
[۶] Record Linkage

جورها و ناجورها معرفی کرد [۸]. فله گی و سانتر این روش را به صورت یک مدل ریاضی فرمول بندی کردند [۲]. اگر یک شناسه‌ی یکتا یا کلید موجودیت موردنظر در دسترس باشد، اتصال رکوردی به صورت قطعی انجام می‌شود. این نوع از اتصال به جورسازی ریاضی نیز شناخته شده است. در این نوع اتصال فرض می‌شود که متغیرهای شناسا فارغ از خطا هستند و رکوردهای مربوط به یک موجودیت دقیقاً بر اساس همین متغیرها با هم جور می‌شوند. اگر متغیرهای شناسا به دلیل خطا در ثبت و یا فقدان اطلاعات تکمیلی قابل دسترس در طول زمان دچار تغییر شده باشند، در این صورت متغیر شناسای یکتای فارغ از خطا که بتواند توسط تمام داده‌های منابع اطلاعاتی به اشتراک گذاشته شود، وجود ندارد. در نتیجه اتصال رکوردی احتمالاتی به کار گرفته می‌شود. این نوع از اتصال به جورسازی آماری نیز معروف است. در این نوع از جورسازی اتصال بر اساس چندین متغیر انجام می‌شود. فرض می‌شود منبع $A$، $N_A$ رکورد و منبع $B$ ، $N_B$ رکورد دارد. بین هر رکورد از $A$ و رکوردهای $B$ یک تناظر پنهانی وجود دارد. بنابراین $N_A \times N_B$ زوج رکورد وجود دارند که وضعیت‌های جور بودن یا جور نبودن آنها باید مشخص شود. بنابراین $A \times B = \{(a,b), a \in A, b \in B\}$ یک زوج رکورد، جورند[۷]، اگر هر دو رکورد به یک موجودیت اشاره داشته باشند. یک زوج رکورد، ناجورند[۸] ، اگر هر یک از رکوردها به موجودیت‌های متمایزی اشاره داشته باشند. در فضای حاصل ضرب $A \times B$ ، دو مجموعه‌ی جدا از هم $M$ و $U$ به صورت زیر تعریف می‌شوند: مجموعه‌ی $M$ شامل زوج رکوردهای جور $M = \{(a,b) : a = b; a \in A, b \in B\}$ و مجموعه‌ی $U$ شامل زوج رکوردهای ناجور $U = \{(a,b) : a \neq b; a \in A, b \in B\}$ است. فرآیند اتصال رکوردی قصد دارد تعیین کند که هر زوج رکورد به کدام مجموعه تعلق دارد:

— اگر زوج مورد نظر یک اتصال باشد: مجموعه‌ی $L$ (تصمیم $A_1$ )،

— اگر زوج مورد نظر یک نااتصال باشد: مجموعه ی $N$ (تصمیم $A_3$)، و

— اگر زوج مورد نظر امکان اتصال درست را داشته باشد (بلاتکلیف): مجموعه‌ی $C$ (تصمیم $A_2$).

برای این که مشخص شود این دسته از زوج‌ها اتصال جور هستند و یا ناجور، به اطلاعات تشخیصی تکمیلی نیاز است. یک بردار مقایسه یا یک بردار توافق $\gamma$ از فضای مقایسه‌ی $\Gamma$ (فضای همه‌ی بردارهای مقایسه) برای هر زوج رکورد $(\alpha(a), \beta(b))$ سطح توافق بین رکوردهای $a$ و $b$ را معرفی می‌کند: $\alpha(a) \times \beta(b) \to \Gamma\gamma \in \Gamma$. مقایسه‌ی $\gamma$ می‌تواند بر اساس یک الگوی ساده «توافق / عدم توافق» که مقادیر (۱ و ۰) را می‌گیرد، باشد. به طور مثال در یک بردار سه گانه‌ی مقایسه:

— $\gamma_1$: زوج‌هایی که در نام خانوادگی با یکدیگر توافق دارند،

— $\gamma_2$: زوج هایی که در نام با یکدیگر توافق دارند، و

— $\gamma_3$: زوج هایی که در آدرس با یکدیگر توافق دارند.

---

[۷] Matched
[۸] Unmatched

در این صورت بردار $\gamma = (\gamma_1, \gamma_2, \gamma_3)' = (1, 0, 1)'$ بر این که زوج مورد نظر فقط در نام خانوادگی و آدرس با یکدیگر توافق دارند، دلالت دارد.

قاعده‌ی اتصال $F$ نگاشتی از $\Gamma$ به یک مجموعه از تابع تصمیم‌های تصادفی $Dd(\gamma)$ است. یک قاعده‌ی اتصال وضعیت زوج‌ها را بر اساس الگوی توافق تعیین می‌کند: $F : \Gamma \to L, C, N$. قاعده‌ی اتصال اپتیمال، قاعده‌ای است که احتمال قرار گرفتن یک زوج را در مجموعه‌ی زوج‌های بلاتکلیف $(C)$، مینیمم می‌کند. برای یک بردار مقایسه $\gamma$ در $\Gamma$ ، $m(\gamma)$ چنین تعریف می‌شود: احتمال شرطی مشاهده‌ی $\gamma$ به طوری که زوج رکورد، یک جور واقعی باشد $m(\gamma) = P(\gamma|(a,b) \in M)$ یا $m(\gamma) = P(\gamma|M)$. به طور مشابه $u(\gamma)$ چنین تعریف می‌شود: احتمال شرطی مشاهده‌ی به طوری که زوج رکورد یک اتصال ناجور واقعی باشد $u(\gamma) = P(\gamma|(a,b) \in U)$ یا $u(\gamma) = P(\gamma|U)$. اگر این احتمال‌های شرطی معلوم باشند، برای هر الگوی $\gamma$ نسبت درست نمایی و یا نسبت توافق به صورت زیر تعریف می‌شود:

$$R(\gamma) = \frac{P(\gamma \in \Gamma|M)}{P(\gamma \in \Gamma|u)} = \frac{m(\gamma)}{u(\gamma)}$$

نسبت توافق تعیین کننده‌ی توان تشخیص بردار مقایسه است. احتمال‌های موجود در $R(\gamma)$ پارامترهای جورسازی نامیده می‌شوند. نسبت $R(\gamma)$ و یا هر تبدیل صعودی یکنوا از آن مثل لگاریتم به عنوان وزن جورسازی تعریف می‌شود. کمیت $\ln(R(\gamma))$ را شاخص توافق می‌نامند که به صورت زیر نوشته می‌شود.

$$\ln(R(\gamma)) = \sum_{k=1}^{K} \ln\left(\frac{P(\gamma_k|M)}{P(\gamma_k|U)}\right)$$

وزن توافق بر روی یک متغیر جورسازی $k$ عبارت است از $\ln\left(\frac{m_k}{u_k}\right)$ و وزن عدم توافق بر روی این متغیر عبارت است از $\ln\left(\frac{1-m_k}{1-u_k}\right)$. برای مثال وزن کل جورسازی برای یک بردار مقایسه $\gamma = (1, 0, 1)'$ عبارت است از: $\ln\left(\frac{m_1}{u_1}\right) + \ln\left(\frac{1-m_2}{1-u_2}\right) \ln\left(\frac{m_3}{u_3}\right)$. فرض اساسی در این جا وجود استقلال شرطی است یعنی:

$$P(\gamma|M) = P(\gamma_1|M)P(\gamma_2|M)\dots P(\gamma_k|M)$$

و

$$P(\gamma|U) = P(\gamma_1|U)P(\gamma_2|U)\dots P(\gamma_k|U)$$

به‌طور کلی اتصال‌ها یا نااتصال‌ها به تخصیص‌های تحت قاعده‌ی تصمیم و جورها یا ناجورها به وضعیت‌های واقعی زوج‌ها اشاره دارند. روش‌های مختلفی برای تعیین پارامترهای شرطی $P(\gamma|M)$ و $P(\gamma|U)$ ارائه شده است. راه مستقیمی که فله گی و سانتر [۲] معرفی کردند، به صورت زیر است:

احتمال حاشیه‌ای توافق بر روی مشخصه‌ای از مجموعه‌ی جورها:

اگر $X$ نام کوچک $(x_1)$، نام خانوادگی $(x_2)$ و یا سن $(x_3)$ باشد: $P(X$ توافق در $|M) = 0/9$،

اگر $X$ نام خیابان $(x_4)$، پلاک $(x_5)$ و یا جنس $(x_6)$ باشد: $P(X$ توافق در $|M) = 0/8$،

احتمال حاشیه‌ای توافق بر روی مشخصه‌ای از مجموعه‌ی ناجورها:

اگر $X$ نام کوچک $(x_1)$، نام خانوادگی $(x_2)$ و یا سن $(x_3)$ باشد: $P(X$ توافق در $|U) = 0/1$،

اگر $X$ نام خیابان $(x_4)$، پلاک $(x_5)$ و یا جنس $(x_6)$

باشد: $P(X$ توافق در $|M) = ۰/۲$.

احتمال‌های $P(\gamma_k|M)$ و $P(\gamma_k|U)$ احتمال‌های $m_k$ و $u_k$ برای هر متغیر جورسازی $k$ خوانده می‌شوند. این احتمال‌ها اغلب با استفاده از یک اطلاع پیشین ارزیابی می‌شوند. احتمال‌های $m$ اغلب پیشین ۰/۹ می‌گیرند و احتمال‌های $u$ اغلب پیشین ۰/۱ می‌گیرند. راه دیگر برآورد $m$ و $u$ با استفاده از داده‌های جور شده است. برای هر $a \in \alpha$ و $b \in \beta$ و $\gamma(a,b) \in \Gamma$ این برآوردها به صورت زیر قابل محاسبه‌اند:

$$\hat{m}_k = \frac{\sum_{(a,b)\in L}[\gamma_k(a,b) = ۱]}{\sum_{\forall(a,b) I[(a,b)\in L]}}$$

و

$$\hat{u}_k = \frac{\sum_{(a,b)\in N}[\gamma_k(a,b) = ۱]}{\sum_{\forall(a,b) I[(a,b)\in N]}}$$

یعنی برای به دست آوردن $\hat{m}_k$، باید فراوانی توافق‌ها برای مقایسه‌ی $\gamma_k$ را در مجموعه‌ی زوج‌های اتصال یافته پیدا و بر تعداد زوج‌هایی که به عنوان اتصال شناخته شده‌اند، تقسیم کرد. همچنین برای به دست آوردن $\hat{u}_k$ ، باید فراوانی توافق‌ها برای مقایسه‌ی $\gamma_k$ را در مجموعه‌ی زوج هایی که اتصال نیافته‌اند، پیدا و بر تعداد زوج‌هایی که به عنوان نااتصال شناخته شده‌اند، تقسیم کرد [۱۱]. اگر $\gamma(a,b) \in \Gamma$ شامل یک الگوی ساده توافق (یک) و عدم توافق (صفر) در رابطه با سه متغیر که در فرض استقلال شرطی صدق می‌کند، باشد، آن گاه ثابت‌های برداری (احتمال‌های حاشیه‌ای) وجود دارند، به طوری که: $\mathbf{m} = (m_۱, m_۲, \ldots, m_k)'$ و $\mathbf{u} = (u_۱, u_۲, \ldots, u_k)'$ و برای هر: $\gamma \in \Gamma$:

$$P(\gamma|M) = \prod_{i=۱}^{K} m_i^{\gamma_i}(۱ - m_i)^{(۱-\gamma_i)}$$

و $P(\gamma|U) = \prod_{i=۱}^{K} u_i^{\gamma_i}(۱ - u_i)^{(۱-\gamma_i)}$.

اگر $\gamma \in \Gamma$ بیش از سه متغیر را معرفی کند، آن گاه تکنیک‌های حل معادلات نظیر روش گشتاورها و یا روش ماکسیمم سازی مقدار مورد انتظار (الگوریتم $EM$ ) به کار می‌رود. وقتی که مجموعه‌ی داده‌ها کامل نیست و یا داده‌های گمشده دارد، از الگوریتم $EM$ استفاده می‌شود. این روش به سرعت به جواب‌های حدی یکتا با نقاط شروع مختلف همگرا می‌شود.

دو نوع خطای متعارف در اتصال رکوردی وجود دارند: خطای نوع $I$ عبارت است از احتمال این که یک مقایسه‌ی ناجور به اشتباه یک اتصال تلقی شود، و خطای نوع $II$ عبارت است از احتمال این که یک مقایسه‌ی جور به اشتباه یک نااتصال تلقی شود. شرایطی که تحت آنها این خطاها رخ می‌دهند، در جدول ۲ نشان داده شده‌اند.

جدول ۲. انواع خطاها در فرآیند اتصال رکوردی

| | جور | ناجور |
|---|---|---|
| اتصال دارند | جور واقعی | جور غلط |
| اتصال ندارند | ناجور غلط | ناجور واقعی |

## ۴ روش سهم وزن تعمیم یافته در اتصال رکوردی

متغیر نشانگر $l_{j,ik}$ فقط نشان می‌دهد که آیا اتصال بین $j \in U^B$ و $ik \in U^B$ وجود دارد یا نه، ولی اهمیت اتصال‌ها را مشخص نمی‌کند. لذا به جای استفاده از متغیر $l_{j,ik}$، متغیر کمی $\theta_{j,ik}$ که نمایانگر احتمال درستی اتصال زوج $(j,ik)$ است، به کار برده می‌شود [۴]. این متغیر میزان اهمیت اتصال $l_{j,ik}$ را نشان

می‌دهد. تناظر بین واحدهای دو جامعه با ماتریس اتصال $\Theta_{AB} = [\theta_{ji}^{AB}]_{N^A \times N^B}$ که در آن $\theta_{ji}^{AB} \neq 0$ است، مشخص می‌شود:

$$\Theta_{AB} = \begin{pmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & \cdots & \cdots & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \theta_{ji}^{AB} & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \cdots & \theta_{N^A \times N^B}^{AB} \end{pmatrix}$$

اتصال رکوردی احتمالاتی، بین رکوردهای دو جمعیت $U^A$ و $U^B$ با به کارگیری یک فرآیند احتمالاتی، اتصال برقرار می‌کند. روش سهم وزن تعمیم یافته در حالتی که $U^A$ و $U^B$ با به کار بردن اتصال رکوردی و قاعده‌ی تصمیم زیر با یکدیگر اتصال می‌یابند، به عنوان رهیافتی کلاسیک شناخته می‌شود.

$$D(j,k) = \begin{cases} \text{اتصال} & \theta_{j,ik} \geq \theta_{high} \\ \text{بلاتکلیف} & \theta_{low} < \theta_{j,ik} < \theta_{high} \\ \text{نااتصال} & \theta_{j,ik} \leq \theta_{low} \end{cases}$$

با در نظر گرفتن هر یک از رهیافت‌های زیر می‌توان روش سهم وزن تعمیم یافته را با توجه به وزن‌های اتصال که از فرآیند اتصال رکوردی نتیجه می‌شوند، سازگار کرد.

۱ — اولین رهیافت استفاده از همه‌ی اتصال‌های غیرصفر که توسط فرآیند اتصال رکوردی، وزن‌های اتصال هر کدام شان، مشخص شده است، را فراهم می‌آورد. وقتی همه اتصال‌های غیرصفر با روش سهم وزن تعمیم یافته به‌کار گرفته می‌شوند، ممکن است اهمیت بیشتری به اتصال‌هایی که وزن بزرگتری دارند، در مقایسه با آنهایی که وزن اتصال کوچکتری دارند، داده شود. طبق تعریف برای هر زوج $(j, ik)$ از فضای $A \times B$ وزن اتصال $\theta_{j,ik}$ بیانگر احتمال درستی اتصال زوج $(j, ik)$ است. به همین دلیل دیگر از متغیر نشانگر $l_{j,ik}$ که فقط بیان می کند که آیا یک اتصال بین واحد $j$ از $U^A$ و واحد $ik$ از $U^B$ برقرار شده است یا نه، استفاده نمی‌شود و به‌جای آن از وزن اتصال $\theta_{j,ik}$ که در اولین گام‌های فرآیند اتصال رکوردی، محاسبه شده است، استفاده می‌شود. ۲— دومین رهیافت به‌کارگیری همه‌ی اتصال‌های غیر صفر که وزن‌های اتصال بیشتر از یک کران مفروض $\theta_{high}$ دارند، را پیشنهاد می‌کند. چون ممکن است اکثر زوج رکوردهای فضای $A \times B$ از فایل‌های $U^A$ و $U^B$ وزن‌های اتصال غیرصفر داشته باشند و در عمل این وزن‌های اتصال، نسبتا کوچک و ناچیز باشند، به نظر نامحتمل می‌رسد که با این مقدار، زوج رکوردها درست اتصال یافته باشند. لذا در این حالت بهتر است وزن‌های اتصالی درنظر گرفته شوند که بزرگتر از یک کران بالای $\theta_{high}$ باشند. در این رهیافت متغیر نشانگر $l_{j,ik}$ برای تعیین این که اتصال وجود دارد یا نه، به کار گرفته نمی‌شود و به جای آن وزن اتصال $\theta_{j,ik}$ که بزرگتر از کران $\theta_{high}$ است در نظر گرفته می‌شود. وزن‌های کمتر از $\theta_{high}$ صفر فرض می‌شوند. بنابراین تعریف می‌کنیم: $\theta_{j,ik}^T = \theta_{j,ik}$ اگر $\theta_{j,ik} \geq \theta_{high}$ و صفر در غیر این صورت.

## ۵ کاربرد

برای نشان دادن کاربردی از نمونه گیری غیرمستقیم، دو جامعه‌ای که در بخش ۱.۵ تعریف شده‌اند، با استفاده از روش اتصال رکوردی به یکدیگر اتصال یافته و مقدار میانگین متغیر مورد نظر با روش نمونه گیری غیرمستقیم در جامعه‌ی هدف با استفاده از روش سهم وزن تعمیم یافته برآورد می‌شود. سپس این برآورد با مقدار به دست

آمده‌ی آن از طریق نمونه گیری مستقیم مقایسه می‌شود.

## ۱.۵ طرح مطالعه

برای برآورد میانگین نمره‌ی اکتساب شده‌ی داوطلبان پسر گروه تجربی منطقه‌ی ۱۸ آموزش و پرورش تهران که در جلسه‌ی آزمون سراسری سال ۱۳۸۳ حاضر بوده‌اند، فایلی از اطلاعات این داوطلبان شامل نام، نام خانوادگی، جنس، سال تولد، نام پدر، شماره شناسنامه و نتایج مربوط به آزمون مانند نمره‌ها ی آنان، در اختیار است (جامعه‌ی $U^B$ شامل ۲۰۰ رکورد). خوشه‌های این جامعه دبیرستان‌ها(۱۳ دبیرستان) هستند. چون در این جامعه همه‌ی داوطلبان کد دبیرستان خود را مشخص نکرده‌اند، و یا برخی از متغیرها نیاز به ویرایش یا جانهی دارند (برای مثال ممکن است داوطلب کد منطقه‌ی خود را درج نکرده و یا آن را اشتباه درج کرده باشد)، در این صورت برای انجام نمونه گیری، از جامعه‌ی دیگری که در ارتباط با این جامعه است و چارچوبی کامل دارد، استفاده می‌شود. این جامعه شامل اطلاعات تقاضانامه‌ی داوطلب مانند نام، نام خانوادگی، جنس، سال تولد، نام پدر و شماره شناسنامه است. در این جامعه نتایج آزمون داوطلبان وجود ندارند (جامعه‌ی $U^A$ شامل ۱۵۴ رکورد). اطلاعات موجود در این جامعه ویرایش و اصلاح شده‌اند. چون برخی اطلاعات مربوط به یک داوطلب در جامعه‌ی $U^A$ و برخی دیگر در جامعه‌ی $U^B$ وجود دارند، بنابراین به منظور ایجاد فایلی جامع از اطلاعات داوطلبان، لازم است اتصال رکوردی به منظوریافتن زوج رکوردهای جور انجام شود.

برای یافتن رکوردهای متناظر با یک فرد از کلید ساخته شده از دو یا چند متغیر استفاده شد (اتصال احتمالاتی). کلید مرتب سازی بر اساس ۴ متغیر نام، نام خانوادگی، شماره شناسنامه و نام پدر ساخته شد. برای انجام اتصال رکوردی برنامه‌ای به یکی از زبان‌های برنامه نویسی (در کامپیوترهای بزرگ با زبان برنامه نویسی $PL/۱$ و در کامپیوترهای شخصی با زبان برنامه نویسی $SAS$) نوشته شد. در ابتدا دو مجموعه‌ی جورها و ناجورها مشخص نیستند، لذا باید یک مقدار پیشین به هر یک از احتمال‌های حاشیه‌ای $m_i$ و $U_i$، $(i = ۱, ۲, ۳, ۴)$ بر اساس درجه‌ی اهمیت جور شدن رکوردها براساس آن متغیر اختصاص داد. چون احتمال جور شدن دو رکورد براساس نام خانوادگی در رکوردهای جور از اهمیت بیشتری برخوردار است، لذا $m_۱ = ۰/۹۵$ و در نتیجه احتمال جور شدن بر اساس نام خانوادگی در مجموعه‌ی ناجورها $u_۱ = ۰/۰۵$ است. احتمال جور شدن براساس نام در مجموعه‌ی جورها $m_۲ = ۰/۸$ و در مجموعه‌ی ناجورها $u_۲ = ۰/۲$، احتمال جور شدن براساس شماره‌ی شناسنامه در مجموعه‌ی جورها $m_۳ = ۰/۹$ و در مجموعه‌ی ناجورها $u_۳ = ۰/۱$، احتمال جور شدن براساس نام پدر در مجموعه‌ی جورها$m_۴ = ۰/۷$ و در مجموعه‌ی ناجورها $u_۴ = ۰/۳$ در نظر گرفته شدند. با مقایسه هر یک از رکوردهای فایل $U^A$ با تک تک رکوردهای فایل $U^B$، بردار مقایسه‌ی $\gamma = (\gamma_۱, \gamma_۲, \gamma_۳, \gamma_۴)'$ برای هر زوج مقایسه به دست آمد. این بردارها ترکیب‌های مختلف از آرایش صفرها و یک‌ها هستند و حداکثر ۱۶ حالت را شامل می‌شوند. اگر نتیجه‌ی مقایسه

$\gamma_i$ برابر یک باشد، آن گاه $\theta_i = \log\left(\frac{m_i}{u_i}\right)$ و اگر نتیجه برابر صفر باشد، $\theta_i = \log\left(\frac{1-m_i}{1-u_i}\right)$. سرانجام داریم: $\theta_{ji}^{AB} = \sum_{i=1}^{4} \theta_i$ . برای کلیه رکوردهایی که از مقایسه‌ی آنها با یکدیگر، $\gamma = (0,0,0,0)'$ نتیجه شد، وزن $\gamma = (\gamma_1,\gamma_2,\gamma_3,\gamma_4)'$ اتصال صفر اختصاص داده شد. بعد از این که بردار برای تمام زوج مقایسه‌ها محاسبه شد، مقادیر وزن‌های اتصال در اولین بار اجرای برنامه اتصال رکوردی برای هر زوج به دست آمدند. با توجه به این که مقادیر به دست آمده‌ی $\theta_i$ برای هر یک از زوج مقایسه‌ها بر اساس یک اتصال قوی، مقداری مثبت و بر اساس یک اتصال ضعیف مقداری منفی است، چنین تصمیم گرفته شد که کلیه زوج‌هایی که وزن اتصال مثبت دارند، در مجموعه‌ی جورها ($M$)، و کلیه زوج‌هایی که وزن اتصال منفی دارند، در مجموعه‌ی ناجورها ($U$) قرار گیرند. با در نظر گرفتن این مقدار کران برای وزن‌های به دست آمده و رابطه‌های (۲) مقادیر جدید برای $\hat{m}_i$ ها و $\hat{u}_i$ ها به دست آمدند. با توجه به مقادیر برآورد شده، بار دیگر رکوردهای فایل $U^A$ با تک تک رکوردهای فایل $U^B$ مقایسه شدند، مقدار وزن‌های اتصال جدید برای هر زوج مقایسه محاسبه شدند. خروجی برنامه‌ی اتصال رکوردی برای تعیین اتصال‌های موجود بین واحدهای دو جامعه و همچنین وزن‌های مربوط به هر اتصال، به کار برده شد. دو رهیافت یکی استفاده از همه‌ی اتصال‌های غیرصفر با متغیر نشانگر $l$ و دیگری استفاده از وزن‌های اتصال بزرگتر از یک مقدار آستانه‌ی مفروض در نظر گرفته شدند. سرانجام آماره‌های به دست آمده در هر رهیافت برآورد و با یکدیگر مقایسه شدند.

## ۲.۵ نمونه گیری غیرمستقیم

همان گونه که اشاره شد به دلیل نقص چارچوب جامعه‌ی $U^B$ امکان نمونه گیری مستقیم از این جامعه وجود ندارد. بنابراین ابتدا نمونه‌ای از جامعه‌ی $U^A$ گزینش می‌شود و سپس با استفاده از اتصال رکوردهای دو جامعه، نمونه‌ی نهایی از جامعه‌ی $U^B$ به دست می‌آید. چون ممکن است بین واحدهای جامعه‌ی $U^A$ و واحدهای جامعه‌ی $U^B$ یک یا چند اتصال برقرار باشد، برای انجام نمونه گیری غیرمستقیم روش سهم وزن تعمیم یافته به کار گرفته شد. طرح نمونه گیری از جامعه اول، تصادفی ساده بدون جایگذاری بود.

رکوردهای جامعه‌ی $U^A$ به ترتیب شماره گذاری شدند. برای این که نشان داده شود که اندازه‌ی نمونه تاثیری در تفاوت عملکرد دو رهیافت ندارد، دو نمونه‌ی تصادفی ساده بدون جایگذاری، با دو کسر نمونه گیری متفاوت گزینش شدند (یک نمونه ۳۰ تایی و یک نمونه ۵۰ تایی). برای راحتی کار در همان ابتدا قبل از اجرای برنامه‌ی اتصال رکوردی، واحدهای جامعه‌ی $U^B$ برحسب خوشه مرتب شدند. این کار برای این انجام شد که اعضای هر خوشه کنار یکدیگر قرار گیرند. سپس به هریک از واحدهای این جامعه به ترتیب شماره‌ای داده شد. هر یک از خوشه‌های جامعه‌ی $U^B$ بررسی شدند که آیا با واحدهای جامعه‌ی $U^A$ در اتصال هستند (برای بررسی نااریب بودن برآوردگر حاصل) و سپس خوشه‌هایی که با عناصر نمونه در اتصال اند، مشخص شدند. مجموعه $\Omega^B$ در این کاربرد همان ۱۳ تا خوشه شد که این مسئله نااریبی برآوردگر حاصل را تضمین می‌کرد. احتمال شمول در نمونه گیری تصادفی ساده

برای نمونه ۳۰ تایی برابر است با: $\pi_i = \frac{m^A}{M^A} = \frac{۳۰}{۱۵۴}$ و برای نمونه ۵۰ تایی برابر است با: $\pi_i = \frac{m^A}{M^A} = \frac{۵۰}{۱۵۴}$. وزن هر خوشه براساس گام‌های تعریف شده در بخش ۲ محاسبه شدند. در رهیافت دوم در هر یک از فرمول‌های تشریح شده در گام‌های اساسی به جای $l_{j,ik}$ ها، $\theta_{j,ik}$ ها به کار برده شدند. در آخر مقدار $\hat{Y}^B$ به ازای اندازه‌های مختلف نمونه و در هر رهیافت به دست آمد. چون جمع $y_{ik}$ ها در هر خوشه و همچنین جمع اتصال‌های هر خوشه با جامعه‌ی $U^A$ مشخص هستند، با درنظر گرفتن متغیر $Z_{ik} = \frac{Y_i}{L_i^B}$ مقادیر $Z - i = \sum_{i=۱}^{N} \sum_{k=۱}^{M_i^B} l_{j,ik} z_{ik}$ به دست آمد و از این طریق نیز مقدار $\hat{Y}_B$ محاسبه شد. برای برآورد واریانس $\hat{Y}^B$ لازم است تعداد زیادی نمونه از همین جامعه تولید شوند. این کار توسط روش شبیه سازی مونت کارلو صورت گرفت. استفاده از این روش بر اساس دنباله‌ای از اعداد تصادفی، خواص برآوردگرها مطالعه می‌شوند. ابتدا مقدار میانگین و واریانس مقادیر $Z_i$ ها را محاسبه کرده، سپس با استفاده از شبیه سازی مونت کارلو ۵۰۰ بار نمونه‌هایی به اندازه‌ی ۳۰ و ۵۰ از جامعه‌ای با این میانگین و واریانس تولید شد [۴].

$$\bar{Z} = \frac{۱}{M^A} \sum_{j=۱}^{M^A} Z_j$$

$$S^۲ = \frac{۱}{M^A} \sum_{j=۱}^{M^A} (Z_j - \bar{Z})^۲$$

$$\hat{\bar{E}}(Y) = \frac{۱}{۵۰۰} \sum_{t=۱}^{۵۰۰} \bar{Y}_i$$

$$\hat{Var}(Y) = \frac{۱}{۵۰۰} \sum_{t=۱}^{۵۰۰} \bar{Y}_t - E(\bar{Y}))^۲$$

$$\hat{CV}(\hat{Y}) = ۱۰۰ \times \frac{\sqrt{\hat{Var}(\bar{Y})}}{\hat{E}(\bar{Y})}$$

## ۳.۵ برآورد میانگین نمره‌ی کل داوطلبان

در رهیافت اول و برای نمونه‌ی اول میانگین نمره‌ی کل ۴۵۲۱/۸۵ و در نمونه‌ی دوم ۶۱۷۵/۸۳ برآورد شد. در رهیافت دوم برای نمونه‌ی اول برآورد میانگین نمره‌ی کل ۶۲۴۶/۸۸ و در نمونه‌ی دوم ۵۹۴۶/۵۵ به دست آمد. جدول ۳ عملکرد دو رهیافت را باهم مقایسه می‌کند.

جدول ۳. مقایسه‌ی عملکرد دو رهیافت با یکدیگر

| | نمونه‌ی ۳۰ تائی | نمونه‌ی ۵۰ تائی |
|---|---|---|
| رهیافت اول | برآورد انحراف معیار میانگین = ۱۳۰/۵۵<br>برآورد ضریب تغییرات = ۹۵/۲ | برآورد انحراف معیار میانگین = ۳۹/۱۰۳<br>برآورد ضریب تغییرات = ۶۴/۱ |
| رهیافت دوم | برآورد انحراف معیار میانگین = ۴۹/۵۸<br>برآورد ضریب تغییرات = ۹۴/۰ | برآورد انحراف معیار میانگین = ۴۰/۴۰<br>برآورد ضریب تغییرات = ۰/۶۶ |

برای مقایسه‌ی برآورد حاصل از روش سهم وزن تعمیم یافته با برآورد حاصل از یکی از روش‌های نمونه گیری کلاسیک، با توجه به این که جامعه‌ی $U^B$ ماهیت خوشه‌ای دارد، برآورد میانگین نمره‌ی کل را توسط نمونه گیری خوشه‌ای یک مرحله‌ای با اندازه‌های نابرابرخوشه‌ها به دست آوردیم. نتیجه در جدول ۴ آورده شده است.

جدول ۴. مقایسه‌ی عملکرد دو رهیافت با نمونه گیری خوشه‌ای

| | نمونه گیری خوشه‌ای | رهیاف اول | رهیافت دوم |
|---|---|---|---|
| برآورد میانگین | ۴۶/۳۲۲۵ | ۹۴/۴۴۲۷ | ۳۹/۶۲۰۰ |

## ٦ نتیجه گیری

با توجه به نتایج به دست آمده در جدول ۴ ، ملاحظه می‌شود که کارایی برآورد حاصل از روش سهم وزن در هر یک از رهیافت‌ها بیشتر از برآورد حاصل از روش خوشه‌ای یک مرحله‌ای است. همچنین در جدول ۳ ملاحظه می‌شود در رهیافتی که از متغیر نشانگر استفاده می‌کند، چه در نمونه‌ی اول و چه در نمونه‌ی دوم مقدار انحراف معیار برآورد شده بیشتر از مقدار آن در رهیافت استفاده از وزن‌های اتصال بزرگتر از یک آستانه‌ی مفروض است (رهیافت دوم). دلیل این امر آن است که در رهیافت دوم از اتصال‌هایی استفاده می‌شود که دارای وزن اتصال درخور توجه هستند و اتصال‌های با وزن کوچک (اتصال های ضعیف) نادیده گرفته می‌شوند. به دلیل مذکور مقدار ضریب تغییرات در رهیافت استفاده از وزن‌های اتصال بزرگتر از یک آستانه‌ی مفروض کمتر است از رهیافتی که از متغیر نشانگر استفاده می‌شود.

## مراجع


[1] Abowd, J. and Vilhuber, L. (2005), workshop on *Record Linking*, Cornell University.

[2] Fellegi, P. and Sunter, A. (1969), A theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

[3] Goodman, L.A. (1961), Snowball sampling, *Annals of Mathematical Statistics*, 32(1), 148-170.

[4] Lavallee, P. and Caron, P. (2001), Estimation using the generalized weight share method: The case of record linkage, *Survey Methodology*, 27(2), 155- 169.

[5] Lavallee, P. and Deville, J. (2006), Indirect sampling: The foundations of the generalized weight share method, *Survey Methodology*, 32(2), 156-176.

[6] Lavallee, P. and Deville, J. (2002), Theoretical foundations of the generalized weight share method, *ICRASS: International Conference on Recent Advances in Survey Sampling.*

[7] Lavallee, P. (2004), Indirect sampling: A practical solution for populations difficult to reach, *Statistics Canada symposium.*

[8] Newcomb, H.B., Kennedy, S.J., Axford, S.J. and James, A.P. (1959), Automatic linkage of vital records, *Science*, 130, 954-959.

[9] Thompson, S.K. (1990), Adaptive cluster sampling, *Journal of the American Statistical Associatino*, 85(412), 1050-1059.

[10] Thompson, S.K. and Seber, G.A. (1996), *Adaptive Sampling*, Willey, New York.

[11] Winkler, E. (1995), *Matching and Record Linking In Business Survey Methods*, (Cox, Binder, Chinnappa, Colledge and Kott, Editeurs), Willey, New York.