فصلنامه گزیده‌مطالب آماری
سال ۱۷، شماره ۲، تابستان ۱۳۸۵، صص ۴۶-۲۹

ساخت و کاربرد وزن‌های نمونه

زهرا رضایی قهرودی

مرکز آمار ایران

چکیده. در این مقاله، مراحل مختلف وزن‌دهی و کاربرد آن در محاسبه براورد مشخصه‌های مورد نظر در آمارگیری‌های خانوار ارائه شده است. به‌خصوص، تصحیح وزن‌های نمونه به‌منظور جبران عیوب و نقایص مختلف در نمونه انتخابی بیان شده است. در این مقاله توجه ما محدود به براوردهایی است که در بیش‌تر گزارش‌های آماری تهیه می‌شود که با ارائه مثال‌هایی از نمونه‌گیری‌های واقعی که در کشورهای در حال توسعه و کشورهای دیگری که موقعیت‌های آمارگیری مشابه دارند، به بررسی روش‌های مختلف وزن‌دهی می‌پردازیم.

۱- مقدمه

نمونه‌گیری‌های خانوار، عمدتاً بر اساس طرح‌های نمونه‌گیری پیچیده به‌منظور کنترل هزینه است. نتایج نمونه‌گیری دارای ویژگی‌هایی است که ممکن است منجر به اریبی و انحرافات دیگر بین نمونه و جامعه مرجع شود. این ویژگی‌ها شامل انتخاب واحدهای با احتمال‌های نابرابر، پوشش ندادن کل جامعه و بی‌پاسخی است. برای اصلاح این ویژگی‌ها از روش وزن‌دهی استفاده می‌کنیم تا نهایتاً به براوردهای مناسب در مورد مشخصه‌های مورد نظر دست یابیم. به عبارت دیگر، اهداف وزن‌دهی شامل موارد زیر است:

- تصحیح احتمال‌های نابرابر انتخاب؛
- تصحیح بی‌پاسخی واحد؛

واژگان کلیدی: وزن پایه؛ تعدیل بی‌پاسخی؛ پسا طبقه‌بندی؛ براورد در حوزه.

- تصحیح توزیع نمونه موزون برای متغیرهای اصلی مورد نظر (مانند وزن، نژاد، جنسیت و...) به‌منظور تطابق با توزیع جامعه.

روش‌های مورد استفاده برای هر یک از حالت‌های فوق در بخش‌های بعدی با جزئیات بیش‌تر ارائه می‌شود. در بخش ۲ وزن‌های نمونه در غالب طرح نمونه‌گیری چند مرحله‌ای شامل تصحیح وزن‌های نمونه، به‌منظور محاسبه واحدهای تکراری و واحدهایی که واجد شرایط بودن آن‌ها نامشخص است، ارائه شده است. بخش ۳ وزن‌دهی را به‌منظور تصحیح احتمال‌های نابرابر انتخاب مورد بررسی قرار داده و در بخش ۴ و ۵ به‌ترتیب موضوع بی‌پاسخی و روش‌های جبران بی‌پاسخی، عدم پوشش و روش‌های جبران آن به‌همراه مثال‌های عددی بیان شده است. در بخش ۶ افزایش واریانس براوردها که ناشی از وزن‌دهی است، به‌همراه مثال عددی ارائه شده، در بخش ۷ به مسأله پیرایش وزن‌ها پرداخته شده است و نهایتاً در بخش آخر نتیجه‌گیری از مطالب گفته‌شده، بیان می‌شود.

۲- ایجاد وزن‌های نمونه

وزن‌دهی اغلب به‌عنوان اولین مرحله تحلیل داده‌های نمونه‌گیری استفاده می‌شود، که این کار معمولاً با ساخت وزن پایه برای هر واحد نمونه‌گیری به‌منظور تصحیح احتمال‌های نابرابر انتخاب، آغاز می‌شود. در حالت کلی، وزن پایه هر واحد نمونه‌گیری، عکس احتمال انتخاب آن واحد در نمونه است. به این صورت که اگر احتمال وجود یک واحد در نمونه با p_i نمایش داده شود، وزن پایه w_i به‌صورت $w_i = 1/p_i$ است. به‌عنوان مثال یک واحد نمونه که با احتمال ۱/۵۰ انتخاب شود، نماینده ۵۰ واحد در جامعه‌ای است که نمونه از آن گرفته شده است.

در طرح‌های چند مرحله‌ای، وزن‌های پایه معرف احتمال انتخاب در هر مرحله است. به‌عنوان مثال در طرح دو مرحله‌ای که i امین واحد نمونه‌گیری با احتمال p_i در مرحله اول و j امین خانوار درون i امین واحد نمونه‌گیری با احتمال $p_{j(i)}$ در مرحله دوم انتخاب شود،

احتمال انتخاب هر خانوار در نمونه (p_{ij}) از رابطه زیر محاسبه می‌شود.

$$p_{ij} = p_i \times p_{j(i)}$$

بنا بر این عکس احتمال فوق، وزن پایه‌ای است که از تصحیح احتمال‌های نابرابر انتخاب حاصل شده است. به همین ترتیب، اگر $w_{ij,b}$ وزن پایه j امین خانوار، $w_{ij,nr}$ وزن جبران بی‌پاسخی و $w_{ij,nc}$ وزن جبران عدم پوشش در نظر گرفته شود، در این صورت وزن نهایی j امین خانوار از i امین واحد نمونه‌گیری شده از رابطه زیر به دست می‌آید.

$$w_{ij} = w_{ij,b} \times w_{ij,nr} \times w_{ij,nc}$$

شایان ذکر است که اندیس b، nr و nc در وزن‌ها به‌ترتیب برای نشان دادن وزن پایه، وزن مربوط به بی‌پاسخی و وزن مربوط به عدم پوشش است.

۱-۲- تصحیح وزن پایه برای واجدین شرایط نامعلوم

در مرحله جمع‌آوری داده در آمارگیری‌های خانوار، مثال‌ها و شواهدی وجود دارد که مسأله واجد شرایط بودن خانوارها مطرح می‌شود. به‌عنوان مثال، یک پرسشگر ممکن است در زمان جمع‌آوری داده هیچ فردی را در واحد مسکونی نمونه‌گیری شده، مشاهده نکند (حتی بعد از مراجعه مجدد). در چنین مواردی مشخص نیست که آیا واحد مسکونی اشغال شده است یا نه. اگر این واحد مسکونی واقعاً اشغال شده باشد، باید به‌عنوان یک واحد مسکونی بی‌پاسخ طبقه‌بندی شود. در غیر این صورت، این واحد مسکونی واجد شرایط نیست. بعضی مواقع پرسشگر فرض می‌کند که اگر هیچ فردی در واحد مسکونی مشاهده نشود، آن واحد مسکونی اشغال نشده است و در نتیجه واجد شرایط نیست، که این‌ها فرضیات ناصحیحی است که اغلب به اشتباه منجر به نرخ بالای پاسخ می‌شود.

بنا بر این زمانی که واجد شرایط بودن برخی از واحدهای مسکونی نمونه‌گیری شده نامعلوم است، وزنی‌دهی این مسأله را تعدیل می‌کند. یک روش ساده، در نظر گرفتن نسبتی از واحدهای مسکونی در نمونه است که وضعیت واجد شرایط بودن یا نبودن آن‌ها مشخص است. از همین نسبت برای واحدهای مسکونی‌ای که واجد شرایط بودن آن‌ها

نامعلوم است، استفاده می‌شود. به‌عنوان مثال نمونه‌ای از ۳۰۰ واحد مسکونی که وضعیت پاسخ‌گویی آن‌ها در جدول زیر ثبت شده است را در نظر بگیرید.

جدول ۱- وضعیت پاسخ‌گویی واحدهای مسکونی

طبقه‌بندی جواب‌گویی	تعداد واحدهای مسکونی
واحدهای با مصاحبه کامل	۲۱۵
واحدهای واجد شرایط بی‌پاسخ	۲۵
واحدهای غیر واجد شرایط	۱۰
واحدهایی که واجد شرایط بودن آن‌ها نامعلوم است	۵۰

در بین واحدهای مسکونی که وضعیت واجد شرایط بودن آن‌ها مشخص و معلوم است، نسبت واحدهای مسکونی واجد شرایط $\frac{۲۱۵+۲۵}{۲۱۵+۲۵+۱۰}=۰/۹۶$ است. بنا بر این می‌توان فرض کرد که همین نسبت (۰/۹۶) از واحدهای مسکونی که واجد شرایط بودن آن‌ها نامعلوم است، واجد شرایط هستند. به عبارت دیگر، ۹۶ درصد از ۵۰ واحد مسکونی که واجد شرایط بودن آن‌ها نامعلوم است، واجد شرایط هستند. اکنون ضریب زیر برای تعدیل وزن واحدهای مسکونی واجد شرایط (واحدهای مسکونی با مصاحبه کامل و واحدهای مسکونی واجد شرایط که پاسخ نداده‌اند)، محاسبه می‌شود:

$$F_{ue}=\frac{\sum_{c} w_{ij,b}+\sum_{nr} w_{ij,b}+\varepsilon\times\sum_{ue} w_{ij,b}}{\sum_{c} w_{ij,b}+\sum_{nr} w_{ij,b}}=۱+\varepsilon\frac{\sum_{ue} w_{ij,b}}{\sum_{c} w_{ij,b}+\sum_{nr} w_{ij,b}},$$

که در آن ε براورد نسبت واحدهای واجد شرایطی است که واجد شرایط بودن آن‌ها نامعلوم است. (در مثال فوق $\varepsilon=۰/۹۶$). جمع‌بندی روی c، nr و ue در فرمول فوق به‌ترتیب جمع وزن‌های پایه واحدهای مسکونی با مصاحبه کامل، بی‌پاسخ‌های واجد

شرایط و واحدهای مسکونی‌ای است که واجد شرایط بودن آن‌ها نامعلوم است. بنا بر این وزن پایه تعدیل‌یافته واحدهای مسکونی واجد شرایط (واحدهای مسکونی با مصاحبه کامل و واحدهای مسکونی واجد شرایط که پاسخ نداده‌اند) با ضرب کردن وزن پایه در فاکتور F_{ue} به دست می‌آید.

۲-۲- تعدیل وزن واحدهای تکراری

اگر مشخص شود که بعضی از واحدها در چارچوب تکرار شده‌اند، احتمال انتخاب بالاتر چنین واحدهایی می‌تواند به‌وسیله فاکتورهای تعدیل جبران شود. اغلب واحدهای تکراری بعد از انتخاب نمونه، مشخص شده و احتمال انتخاب چنین واحدهایی به‌صورت زیر تعدیل می‌شود:

فرض کنید i امین واحد نمونه‌گیری دارای احتمال انتخاب p_{i1} است و $k-1$ رکورد اضافی که در چارچوب نمونه‌گیری توسط این واحد تکراری، مشخص شده است هر یک دارای احتمال انتخاب معلوم $p_{i2}, \ldots, p_{ik}$ است. بنا بر این احتمال انتخاب تعدیل‌یافته واحد نمونه‌گیری مورد نظر به‌صورت زیر تعریف می‌شود:

$$p_i = 1-(1-p_{i1})(1-p_{i2})\ldots(1-p_{ik})$$

وزن متناظر با این واحد نمونه‌گیری $1/p_i$ است.

اکنون روش‌های ساخت وزن‌های نمونه را به‌منظور تعدیل سه ویژگی احتمال‌های انتخاب نابرابر، بی‌پاسخی و عدم پوشش به‌همراه چند مثال مورد بحث و بررسی قرار می‌دهیم.

۳- تعدیل احتمال‌های انتخاب نابرابر

نمونه‌گیری دو مرحله‌ای شامل انتخاب مناطق به‌عنوان واحدهای انتخاب شده در مرحله اول و خانوارها به‌عنوان واحدهای انتخابی مرحله دوم می‌باشد. در مرحله اول، نمونه‌ای با حجم n از N واحد جامعه انتخاب می‌شود و سپس m خانوار از هر یک از M_i خانوار موجود

در مناطق نمونه‌گیری شده انتخاب می‌شوند. احتمال انتخاب هر خانوار بستگی بـه تعـداد خانوارها در منطقه‌ای که در آن واقع شده است، دارد. فرض کنید M_i معرف تعداد خانوارها در i امین منطقه است. بنا بر این احتمال انتخـاب یـک منطقـه، n/N و احتمـال شـرطی انتخاب یک خانوار در i امین منطقه نمونه‌گیری شده، m/M_i است. احتمـال کـل انتخـاب یک خانوار از رابطه زیر محاسبه می‌شود:

$$p_{ij} = p_i \times p_{j(i)} = \frac{n}{N} \times \frac{m}{M_i} = \frac{nm}{N} \times \frac{1}{M_i}$$

و وزن خانوار نمونه‌گیری شده بر اساس این طرح نمونه‌گیری به‌صورت زیر است:

$$w_i = \frac{1}{p_{ij}} = \frac{N}{nm} \times M_i$$

۴- تعدیل وزن برای بی‌پاسخی

این یک امر نادر است که در آمارگیری‌ها، تمام اطلاعـات مـورد نظـر از همـه واحـدهای نمونه‌گیری به دست آید. به‌عنوان مثال، برخی از خانوارها ممکن اسـت پاسـخگوی هـیچ اطلاعاتی نباشند (بی‌پاسخی واحد) در حالی که خانوارهای دیگـر ممکـن اسـت تنهـا بـه تعدادی از سوالات پاسخ دهند (بی‌پاسخی قلم). بنا بر این این مسأله بسیار حـائز اهمیـت است که موارد بی‌پاسخی را هر چند کم باشد، در نظر بگیریم تا با این کـار امکـان اریـب شدن براوردهای مورد نظر طرح را کاهش دهیم. به‌عنوان مثال فردی که در یک منطقـه شهری زندگی می‌کند و درامد نـسبتاً بـالایی دارد، ممکـن اسـت بـا احتمـال کمتـری در نمونه‌گیری شرکت کند. بنا بر این عدم حضور بخش زیادی از این قشر جامعه می‌تواند در براوردهای میانگین درامد خانوارها، پیشرفت تحصیلی و... تأثیرگذار باشد.

۴-۱- کاهش اریبی بی‌پاسخی در آمارگیری‌های خانوار

اریبی ناشی از بی‌پاسخی برای میانگین نمونه، تابعی از دو فاکتور زیر است:

- نسبتی از جامعه که پاسخ نداده‌اند؛

- اختلاف میانگین مشخصه مورد نظر در جامعه بین گروه‌های پاسخگو و بی‌پاسخ.

کاهش اریبی ناشی از بی‌پاسخی، مستلزم پایین بودن نرخ بی‌پاسخی و اختلاف کم بین افراد و خانوارهای پاسخگو و بی‌پاسخ است. از طریق حفظ رکوردهای ثبت شده مربوط به هر واحد نمونه‌گیری، امکان براورد نرخ بی‌پاسخی برای کل نمونه و یا زیر حوزه‌های مورد نظر وجود دارد. بنا بر این مطالعات خاصی در این زمینه می‌تواند به ارزیابی اختلافات بین واحدهای بی‌پاسخ و پاسخگو بپردازد. (گروز و کاپر ۱۹۹۸)

۲-۴- جبران بی‌پاسخی

روش‌های متعددی برای افزایش نرخ پاسخ‌دهی و کاهش اریبی ناشی از بی‌پاسخی در آمارگیری‌های خانوار وجود دارد. در یک مصاحبه ممکن است مصاحبه‌کننده مجبور باشد چند مرتبه تماس بگیرد و یا به افراد مراجعه کند. در واقع مصاحبه‌کننده باید با تلاش‌های متعدد، مصاحبه خود را با خانوار نمونه‌گیری شده به اتمام برساند. نرخ پاسخ‌دهی بالا از طریق آموزش بهتر مصاحبه‌گران امکان‌پذیر است. با وجود تلاش‌های زیاد در این زمینه، بی‌پاسخی امری غیر قابل اجتناب است. بنا بر این طراحان نمونه‌گیری اغلب تعدیل‌هایی را برای جبران بی‌پاسخی ارائه می‌دهند. سه روش زیر روش‌های تعدیل بی‌پاسخی واحد است.

- تعدیل حجم نمونه با در نظر گرفتن حجم نمونه بالاتر از آن‌چه مورد نیاز است، صورت گیرد تا از این طریق بتوانیم بی‌پاسخی‌های مورد انتظار را نیز به حساب آوریم؛
- جانشین کردن؛ فرایند جایگذاری خانوارهای بی‌پاسخ با خانوارهای دیگری که در نمونه نبوده به‌طوری که خانوارها نسبت به مشخصه خاص، شبیه خانوارهای بی‌پاسخ باشند؛
- تعدیل وزن نمونه به‌منظور جبران بی‌پاسخی.

توصیه می‌شود بی‌پاسخی واحد در آمارگیری خانوار از طریق تعدیل وزن‌های نمونه که در بخش ۳-۴ با یک مثال عددی ارائه خواهد شد، صورت گیرد. مشکلات و مسائل

متعددی در ارتباط با جانشین کردن وجود دارد که می‌توان در کالتون (۱۹۸۳) با این مشکلات آشنا شد.

۳-۴- تعدیل بی‌پاسخی وزن‌های نمونه

روش تعدیل وزن‌های نمونه به‌منظور جبران بی‌پاسخی که در آمارگیری‌های خانوار مورد استفاده قرار می‌گیرد، در چهار مرحله زیر خلاصه شده است.

مرحله اول: به‌کارگیری وزن‌های اولیه (برای احتمال‌های انتخاب نابرابر و تعدیل‌های دیگر که در بخش‌های ۲ و ۳ بحث شده است، در صورت لزوم).

مرحله دوم: تقسیم‌بندی نمونه به زیرگروه‌ها و محاسبه نرخ پاسخ موزون برای هر زیر گروه.

مرحله سوم: استفاده از عکس نرخ پاسخ برای زیرگروه‌ها به‌منظور تعدیل بی‌پاسخی.

مرحله چهارم: محاسبه وزن تعدیل شده بی‌پاسخی برای iامین واحد به‌صورت زیر است:

$$w_i = w_{1i} \times w_{2i}$$

که در آن w_{1i} وزن اولیه، w_{2i} وزن تعدیل بی‌پاسخی و در نتیجه w_i وزن نهایی برای واحد iام پس از تعدیل بی‌پاسخی است.

مثال ۱: با استفاده از روش نمونه‌گیری چند مرحله‌ای طبقه‌بندی شده، نمونه‌ای به حجم ۱۰۰۰ خانوار از دو منطقه (شمال و جنوب) یک کشور انتخاب شده است. خانوارها در شمال کشور با نرخ ۱/۱۰۰ و در جنوب با نرخ ۱/۲۰۰ نمونه‌گیری شده‌اند. نرخ پاسخ در مناطق شهری کمتر از نرخ پاسخ در مناطق روستایی است.

هدف براورد نسبت و تعداد خانوارهایی است که دسترسی به مراقبت‌های اولیه بهداشت دارند. فرض کنید n_h تعداد خانوارهای نمونه‌گیری شده در طبقه hام، r_h تعداد خانوارهای واجد شرایط پاسخگو و t_h تعداد خانوارهای واجد شرایطی است که به اطلاعات پرسشنامه پاسخ داده‌اند و به مراقبت‌های اولیه بهداشت نیز دسترسی دارند. بنا بر این وزن تعدیل شده بی‌پاسخی برای خانوارها در طبقه hام به‌صورت

$$w_h = w_{1h} \times w_{2h}$$

است که در آن $w_{2h} = n_h / r_h$، وزن مربوط به تعدیل بی‌پاسخی در طبقه hام می‌باشد (برای سادگی از اندیس i صرف نظر شده است). اطلاعات در سطح طبقات در جدول زیر خلاصه شده است.

جدول ۲- اطلاعات خانوارهای نمونه‌گیری شده به تفکیک طبقات

طبقات	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
شمال- شهری	۱۰۰	۸۰	۷۰	۱۰۰	۱/۲۵	۱۲۵	۱۰۰۰۰	۸۷۵۰
شمال- روستایی	۳۰۰	۱۲۰	۱۰۰	۱۰۰	۲/۵۰	۲۵۰	۳۰۰۰۰	۲۵۰۰۰
جنوب- شهری	۲۰۰	۱۷۰	۱۵۰	۲۰۰	۱/۱۸	۲۳۶	۴۰۱۲۰	۳۵۴۰۰
جنوب- روستایی	۴۰۰	۳۶۰	۱۸۰	۲۰۰	۱/۱۱	۲۲۲	۷۹۹۲۰	۳۹۹۶۰
کل	۱۰۰۰	۷۳۰	۵۰۰				۱۶۰۰۴۰	۱۰۹۱۱۰

بنا بر این نسبت خانوارهایی که دسترسی به مراقبت‌های اولیه بهداشت دارند به‌صورت زیر براورد می‌شود:

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_r} = \frac{۱۰۹۱۱۰}{۱۶۰۰۴۰} = ۰/۶۸۲ \quad , \quad ٪۶۸/۲$$

و تعداد خانوارهای براورد شده که دسترسی به مراقبت‌های اولیه بهداشت دارند به‌صورت

$$\hat{t} = \sum w_h t_h = ۱۰۹۱۱۰$$

است. شایان ذکر است که براورد غیر وزنی نسبت خانوارهایی که دسترسی به مراقبت‌های اولیه بهداشت دارند، تنها با استفاده از اطلاعات خانوارهایی که به پرسشنامه پاسخ داده‌اند، به‌صورت زیر براورد می‌شود:

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{۵۰۰}{۷۳۰} = ۰/۶۸۵ \quad , \quad ٪۶۸/۵$$

و براورد نسبت خانوارهایی که دسترسی به مراقبت‌های اولیه بهداشت دارند بـا اسـتفاده از وزن‌های اولیه و بدون در نظر گرفتن تعدیل بی‌پاسخی به‌صورت

$$\hat{p}_{۱} = \frac{\sum w_{۱h} t_h}{\sum w_{۱h} r_r} = \frac{۸۳۰۰۰}{۱۲۶۰۰۰} = ۰/۶۵۹ \quad , \quad ٪۶۵/۹$$

است. نتایج فوق نشان می‌دهد که اختلاف قابل قبولی بین نسبت براورد شده بـا در نظـر گرفتن وزن‌های اولیه در مقایسه بـا نـسبت بـراورد شـده بـا در نظـر گـرفتن وزن‌هـای تعدیل‌شده بی‌پاسخی وجود دارد. در صورتی که این اختلاف بین نسبت غیر وزنی و نسبت تعدیل شده بی‌پاسخی جزئی است.

بعد از تصحیح وزن‌های بی‌پاسخی، در بخش بعدی، تعدیل وزن‌ها به‌منظـور جبـران عدم پوشش بیان می‌شود.

۵- تعدیل وزن برای عدم پوشش

عدم پوشش به عیب چارچوب نمونه‌گیری برای پوشش جامعه هدف و برخی از واحـدهای نمونه که احتمال انتخاب در نمونه انتخابی را ندارند، اشاره می‌کند. این مشکل تنها یکی از عیوب متعدد چارچوب‌های نمونه‌گیـری اسـت کـه در انتخـاب نمونـه در آمـارگیری‌هـای کشورهای در حال توسعه با آن مواجه هستیم. برای دستیابی به اطلاعات بیش‌تر در زمینه مشکلات موجود در چارچوب‌های نمونه‌گیری و برخی از راه‌حل‌هـای ممکـن بـه یانـسانه (۲۰۰۳) مراجعه شود. عدم پوشش یکی از نگرانی‌هـای عمـده در آمـارگیری‌هـای خـانوار بالاخص در کشورهای در حال توسعه است. بنا بر این شناسایی، ارزیـابی و کنتـرل عـدم پوشش در آمارگیری‌های خانوار باید به‌عنوان یکی از مسائل اصلی در ادارات ملی آمـار در کشورهای در حال توسعه مورد بررسی قرار گیرد.

در این بخش به شناسایی برخی از منابع عدم پوشش در بررسی‌هـای خـانوار اشـاره می‌شود و از یک روش جبـران عـدم پوشـش، کـه بـه تعـدیل آمـاری وزن‌هـا از طریـق

پساطبقه‌بندی معروف است، اشاره می‌شود.

۱-۵- منابع عدم پوشش در آمارگیری‌های خانوار

اغلب آمارگیری‌های خانوار در کشورهای در حال توسعه بر اساس نمونه‌گیری چند مرحله‌ای طبقه‌بندی صورت می‌گیرد. واحدها در مرحله اول عموماً مناطق جغرافیایی هستند. در مرحله دوم فهرستی از خانوارها یا واحدهای مسکونی از واحدهای انتخاب شده، تهیه می‌شود و در آخرین مرحله نمونه‌ای از اعضای خانه یا ساکنین خانوار از نمونه انتخابی تهیه می‌شود. بنا بر این عدم پوشش ممکن است در هر سه سطح انتخاب واحدهای مرحله اول، خانوار و افراد وجود داشته باشد.

چون واحدهای مرحله اول نمونه‌گیری عمدتاً بر اساس مناطق شمارش شده و اطلاعات سرشماری نفوس و مسکن قبلی است، انتظار می‌رود که کل حوزه جغرافیایی جامعه هدف پوشش داده شود. همچنین عدم پوشش در اولین مرحله انتخاب نمونه عموماً کم است. عدم پوشش در این مرحله در کشورهای در حال توسعه، به سختی عدم پوشش در مراحل بعدی نیست. اما به هر حال عدم پوشش در اولین مرحله انتخاب واحد نمونه‌گیری در بیش‌تر آمارگیری‌ها وجود دارد. به‌عنوان مثال، یک آمارگیری ممکن است برای براورد جمعیت وارد شده در یک کشور یا یک منطقه از کشور طراحی شده باشد، که بعضی از مناطق مربوط به اولین مرحله انتخاب نمونه ممکن است به‌جهت مسائلی چون جنگ و ناآرامی‌های درون کشوری و بلایای طبیعی یا دلایل دیگر غیر قابل دسترس باشند. همچنین مناطق پرت با خانوارهای کم‌جمعیت اغلب از چارچوب نمونه‌گیری در آمارگیری‌های خانوار حذف می‌شوند، زیرا برای پوشش آن‌ها باید هزینه زیادی پرداخت شود و این واحدها نماینده کوچکی از جامعه می‌باشند و تأثیر کمی در نتایج کلی جامعه دارند. بنا بر این در نتایج ارائه شده از چنین آمارگیری‌هایی، باید حذف از چنین مناطقی به وضوح توضیح داده شود. زمانی که بخشی از جامعه را پوشش نداده‌ایم، نباید چنین تلقی شود که نتایج آمارگیری برای کل کشور یا منطقه می‌باشد، بنا بر این مسأله عدم پوشش باید به‌طور کامل در گزارش‌های آماری ثبت شده باشد.

عوامل متعدد دیگری در ایجاد عدم پوشش مؤثر است، که به تعدادی از آن‌ها اشاره

می‌کنیم. یکی از این عوامل تعریف ناصحیح از مفاهیم مربوط به هر آمارگیری است. عوامل دیگری شامل از قلم‌افتادگی غیر عمدی واحدهای مسکونی از فهرست‌های تهیه شده در زمان عملیات میدانی، از قلم‌افتادگی به‌جهت خطاهای اندازه‌گیری، عدم حضور اعضای غایب خانوار و از قلم‌افتادگی به‌جهت عدم درک صحیح از مفاهیم آمارگیری وجود دارد. جزئیات بیش‌تر در مورد منابع عدم پوشش در لپکوسکی (۲۰۰۳) و منابع موجود در آن بیان شده است. روش‌های زیر برای کنترل عدم پوشش در آمارگیری‌های خانوار وجود دارد:

- بهبود دستورالعمل‌های میدانی از طریق استفاده از چارچوب‌های چندگانه و بهبود دستورالعمل‌های فهرست‌برداری؛
- جبران عدم پوشش از طریق تعدیل وزن‌ها.

در مثال زیر از روش تعدیل وزن‌ها (روش دوم) استفاده شده است. روش تعدیل وزن معروف به پساطبقه‌بندی است. در این روش، جبران عدم پوشش از طریق تعدیل توزیع نمونه‌گیری وزنی برای متغیرهای خاص، انجام می‌گیرد به‌طوری که این توزیع با توزیع جامعه مورد نظر مطابقت داشته باشد.

مثال۲: در مثال قبل فرض کنید تعداد خانوارها در شمال کشور ۴۵۰۲۵ و در جنوب کشور ۱۱۵۸۰۰ است. کل نمونه وزنی به‌ترتیب برای مناطق شمالی و جنوبی با اعمال وزن بی‌پاسخی برابر با ۴۰۰۰۰ و ۱۲۰۰۴۰ است. بنا بر این برای محاسبه وزن‌ها به روش زیر عمل می‌کنیم:

مرحله اول: محاسبه فاکتورهای پساطبقه‌بندی

برای منطقه شمال $$w_{2h} = \frac{N_h}{\hat{N}_h} = \frac{۴۵۰۲۵}{۴۰۰۰۰} = ۱/۱۲۶$$

برای منطقه جنوب $$w_{2h} = \frac{N_h}{\hat{N}_h} = \frac{۱۱۵۸۰۰}{۱۲۰۰۴۰} = ۰/۹۶۵$$

که در آن $\hat{N}_h$ کل نمونه وزنی است.

مرحله دوم: محاسبه وزن‌های نهایی تعدیل شده

$$w_f = w_h \times w_{2h}$$

که نتایج در جدول زیر خلاصه شده است:

جدول ۳- اعمال وزن پساطبقه‌بندی با استفاده از اطلاعات جدول قبلی

طبقات	r_h	t_h	w_h	w_f	$w_f r_h$	$w_f t_h$
شمال_ شهری	۸۰	۷۰	۱۲۵	۱۴۰/۷۵	۱۱۲۵۶	۹۸۴۹
شمال_ روستایی	۱۲۰	۱۰۰	۲۵۰	۲۸۱/۴۰	۳۳۷۶۸	۲۸۱۴۰
جنوب_ شهری	۱۷۰	۱۵۰	۲۳۶	۲۲۷/۷۷	۳۸۷۰۹	۳۴۱۵۵
جنوب_ روستایی	۳۶۰	۱۸۰	۲۲۲	۲۱۴/۲۰	۷۷۱۱۲	۳۸۵۵۶
کل	**۷۳۰**	۵۰۰			**۱۶۰۸۴۵**	**۱۱۰۷۰۰**

بنا بر این براورد نسبت خانوارهایی که به مراقبت‌های اولیه بهداشت دسترسی دارند برابر است با:

$$\hat{p}_f = \frac{\sum w_f t_h}{\sum w_f r_h} = \frac{۱۱۰۷۰۰}{۱۶۰۸۴۵} = ۰/۶۸۸ \quad \text{یا} \quad ٪۶۸/۸$$

همان‌طور که نتایج جدول نشان می‌دهد، با وزن‌های تعدیل‌شده از روش پساطبقه‌بندی، تعداد نمونه وزنی برای مناطق شمالی و جنوبی به‌ترتیب $(۱۱۲۵۶ + ۳۳۷۶۸) = ۴۵۰۲۴$ و $(۳۸۷۰۹ + ۷۷۱۱۲) = ۱۱۵۸۲۱$ است که به اعداد کنترلی که در بالا به آن اشاره شده است، بسیار نزدیک است.

۶- افزایش واریانس به‌دلیل وزن‌دهی

با وجود این‌که وزن‌دهی باعث کاهش اریبی براوردها می‌شود، می‌تواند باعث افزایش واریانس براوردها نیز شود. برای روشن شدن مطلب، طرح نمونه‌گیری یک مرحله‌ای

طبقه‌بندی شده، با احتمال برابر در هر طبقه را در نظر می‌گیریم. اگر واریانس طبقات (واریانس درون واحدهای طبقه) در هر طبقه یکسان نباشد، در نظر گرفتن وزن‌های نابرابر در طبقات، دقت براوردهای طرح را بالا می‌برد و در صورتی که واریانس طبقات در هر طبقه یکسان باشد، در نظر گرفتن وزن‌های نابرابر منجر به افزایش واریانس براوردها می‌شود. بنا بر این اثر وزن‌دهی، در افزایش واریانس جامعه براورد شده به‌وسیله عامل زیر اندازه‌گیری می‌شود.

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2}$$

که در آن $n = \sum_h n_h$ حجم نمونه قابل قبول یا واجد شرایط، w_h وزن نهایی و n_h حجم واقعی نمونه در طبقه hام است. اکنون در مثال زیر به بررسی این مطلب می‌پردازیم.

مثال۳: محاسبه فاکتور افزایش واریانس با استفاده از اطلاعات مثال قبل با وزن‌های نهایی w_f و n_h به‌عنوان حجم نمونه طبقه hام.

جدول ۴- اطلاعات موردنیاز برای محاسبه فاکتور افزایش واریانس با استفاده از جدول قبلی

طبقات	r_h	w_f	$w_f r_h$	$w_f^2 r_h$
شمال- شهری	۸۰	۱۴۰/۷۵	۱۱۲۶۰	۱۵۸۴۲۸۲
شمال- روستایی	۱۲۰	۲۸۱/۴۰	۳۳۷۶۸	۹۵۰۲۳۱۵
جنوب- شهری	۱۷۰	۲۲۷/۷۷	۳۸۷۰۹	۸۸۱۴۰۳۹
جنوب- روستایی	۳۶۰	۲۱۴/۲۰	۷۷۱۱۲	۱۶۵۱۷۳۹۰
کل	**۷۳۰**		**۱۶۰۸۴۹**	**۳۶۴۱۸۰۲۶**

با توجه به فرمول فوق، فاکتور افزایش واریانس برابر $L = ۷۳۰ \times \frac{۳۶۴۱۸۰۲۶}{(۱۶۰۸۴۹)^2} = ۱/۰۳$ است. به این معنا که به‌دلیل استفاده از روش وزن‌دهی به‌منظور جبران بی‌پاسخی و عدم پوشش، افزایش واریانس در براوردهای آماری تقریباً ۳ درصد است.

۷- پیرایش وزن‌ها

زمانی که وزن‌دهی برای جبران عیوب و ویژگی‌های مربوط به طرح‌های آمارگیری محاسبه می‌شود، بررسی توزیع وزن‌های تعدیل شده نیز لازم است. وزن‌های خیلی بزرگ، اگر تنها روی بخش کوچکی از اقلام واحدهای نمونه‌گیری تأثیر بگذارند، می‌توانند باعث افزایش قابل توجه‌ای در واریانس براوردهای آماری شوند. بنا بر این لازم است روی وزن‌های خیلی بزرگ پیرایش صورت گیرد، تا از این طریق بتوانیم واریانس براوردهای مربوطه را کنترل کنیم. پیرایش وزن‌ها، اکثراً بعد از مرحله تعدیل وزن بی‌پاسخی صورت می‌گیرد.

مادامی که پیرایش وزن‌ها منجر به کاهش واریانس براوردها شود، می‌تواند منجر به اریبی در براوردها نیز شود. در بعضی موارد، کاهش واریانس به‌دلیل پیرایش وزن‌های خیلی بزرگ، ممکن است باعث افزایش اریبی شود. بنا بر این پیرایش وزن‌ها، باید زمانی که ضرورت دارد، صورت گیرد. به عبارت دیگر زمانی باید از این روش استفاده کرد که اریبی به‌وجود آمده به‌واسطه پیرایش وزن‌ها از اهمیت کمتری نسبت به کاهش واریانس‌ها برخوردار باشد.

در هر طرح طبقه‌بندی، فرایند پیرایش وزن‌ها باید در هر طبقه صورت گیرد. فرایند با تعیین یک کران بالا برای وزن‌های اصلی آغاز می‌شود و سپس تعدیل کل وزن‌ها به‌گونه‌ای صورت می‌گیرد که جمع وزن‌های پیرایش شده مشابه جمع وزن‌های اصلی باشد. w_{hi} را وزن نهایی iامین واحد از طبقه hام و w_{hB} را کران بالا برای وزن‌های مشخص شده در طبقه hام در نظر می‌گیریم. وزن پیرایش شده برای iامین واحد نمونه‌گیری شده در طبقه hام به‌صورت زیر تعریف می‌شود.

$$w_{hi(T)} = \begin{cases} w_{hi} & w_{hi} < w_{hB} \\ w_{hB} & w_{hi} \geq w_{hB} \end{cases}$$

وزن‌های پیرایش شده برای کل نمونه می‌تواند به این صورت که جمع آن‌ها دقیقاً مشابه جمع وزن‌های معمول است تعدیل شود. این کار با محاسبه فاکتور زیر صورت می‌گیرد.

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}}$$

جمع‌بندی در نسبت فوق روی همه طبقات و همه واحدها در نمونه است و وزن پیرایش تعدیل شده به‌صورت زیر تعریف می‌شود.

$$w^*_{h(t)} = F_T \times W_{h(T)}$$

روشن است که، $\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$ است و این همان هدف مطلوب و مورد نظر ما است. اکنون در مثالی به‌شرح روش پیرایش وزن‌ها می‌پردازیم. این مثال تنها برای فهم این مطلب بیان شده است و بر اساس یک وضعیت واقعی نیست.

مثال۴: در جدول زیر دو ستون اول تعداد کل واحدهای نمونه‌گیری شده و وزن‌های نهایی در ۷ طبقه است. بیش‌ترین وزن ۲۵۰ انتخاب شده است و همان‌طور که در ستون سوم جدول نشان داده شده است، وزن‌های اصلی در ۲۵۰ بریده شده است.

جدول ۵- محاسبه وزن مربوط به پیرایش وزن‌ها

n_h	w_h	$w_{h(T)}$	$n_h w_h$	$n_h w_{h(T)}$	$n_h w^*_{h(T)}$
۸۰	۱۴۰/۷۵	۱۴۰/۷۵	۱۱۲۶۰	۱۱۲۶۰	۱۱۸۲۳/۰۰
۱۰۰	۱۵۰/۲۵	۱۵۰/۲۵	۱۵۰۲۵	۱۵۰۲۵	۱۵۷۷۶/۲۵
۱۲۵	۱۷۵/۰۰	۱۷۵/۰۰	۲۱۸۷۵	۲۱۸۷۵	۲۲۹۶۸/۷۵
۱۵۰	۲۰۰/۰۰	۲۰۰/۰۰	۳۰۰۰۰	۳۰۰۰۰	۳۱۵۰۰/۰۰
۱۲۰	۲۵۰/۰۰	۲۵۰/۰۰	۳۰۰۰۰	۳۰۰۰۰	۳۱۵۰۰/۰۰
۱۲۰	۲۷۵/۱۳	۲۵۰/۰۰	۳۳۰۱۵	۳۰۰۰۰	۳۱۵۰۰/۰۰
۱۷۰	۲۸۵/۴۰	۲۵۰/۰۰	۴۸۵۱۸	۴۲۵۰۰	۴۴۶۲۵/۰۰
۸۶۵			**۱۸۹۶۹۳**	**۱۸۰۶۶۰**	**۱۸۹۶۹۳**

فاکتور F_T با استفاده از اطلاعات جدول فوق به‌صورت زیر است:

$$F_T = \frac{\sum_i n_h w_{hi}}{\sum_i n_h w_{hi(T)}} = \frac{۱۸۹۶۹۳}{۱۸۰۶۶۰} = ۱/۰۵$$

پس از این مرحله، وزن‌های پیرایش شده از ضرب هر وزن در فاکتور $F_T = ۱/۰۵$ به دست می‌آید، به‌گونه‌ای که رابطه $\sum n_h w_h \sum n_h w^*_{h(T)} = ۱۸۹۶۹۳$ برقرار باشد.

۸- نتیجه‌گیری

وزن‌دهی به‌عنوان بخش جدایی‌ناپذیر از تحلیل‌های آماری آمارگیری‌های خانوار در کشورهای در حال توسعه از اهمیت ویژه‌ای برخوردار است. در اغلب طرح‌ها و دستورالعمل‌های آمارگیری از روش وزن‌دهی استفاده می‌شود. دیدگاه مورد قبول عامه این بوده است که استفاده از وزن‌ها، تحلیل‌ها را پیچیده‌تر می‌کند. اما پیشرفت‌های زیادی در زمینه تکنولوژی کامپیوتر در دهه گذشته، این موضوع را رد کرده است و وجود سخت‌افزارها و نرم‌افزارهای متعدد و برنامه‌های کامپیوتری مشخص برای تحلیل داده‌ها که در بسیاری از کشورهای در حال توسعه در دسترس می‌باشد، کار محاسبات را بسیار آسان کرده است.

همان‌طور که قبلاً بحث شد، وزن‌دهی باعث کاهش اریبی ناشی از مشکلات عدم پوشش و بی‌پاسخی می‌شود. عدم پوشش و بی‌پاسخی انواع مختلفی از خطاهای ناشی از نقص یک آمارگیری طراحی شده برای تهیه اطلاعات از برخی واحدهای جامعه هدف می‌باشد. برای آمارگیری‌های خانوار در کشورهای در حال توسعه، عدم پوشش مشکل جدی‌تری نسبت به بی‌پاسخی است که در این مقاله به روش‌های تعدیل آماری وزن‌های پایه برای جبران مشکلات غیر قابل اجتناب اشاره شده است.

وجود نرم‌افزارهای آماری که در دسترس همه می‌باشد، مسأله وزن‌دهی را به‌عنوان یک کار معمول در تحلیل‌های آماری مربوط به آمارگیری‌های خانوار حتی در کشورهای در حال توسعه درآورده است. توجه دقیق به مسأله وزن‌دهی به‌جهت وجود خطاهای جدی

۱- مقدمه

از سال ۱۹۵۰ به بعد که رایانه، در تحلیل و ذخیره‌سازی داده‌ها به کار رفت، حجم اطلاعات ذخیره شده در آن پس از حدود ۲۰ سال دو برابر شد و همزمان با پیشرفت فناوری اطلاعات، حجم داده‌ها در پایگاه داده‌ها هر دو سال یک بار، دو برابر شد و همچنان با سرعت بیش‌تری نسبت به گذشته حجم اطلاعات ذخیره شده بیش‌تر و بیش‌تر می‌شود. با وجود شبکه جهانی وب، سیستم‌های یکپارچه اطلاعاتی، سیستم‌های یکپارچه بانکی، تجارت الکترونیکی و... لحظه به لحظه به حجم داده‌ها در پایگاه داده‌ها اضافه شده و باعث به‌وجود آمدن انبارهای (توده‌های) عظیمی از داده‌ها شده است، به‌طوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده‌ها را بیش از پیش نمایان کرده است.

شدت رقابت‌ها در عرصه‌های علمی، اجتماعی، اقتصادی، سیاسی و نظامی نیز اهمیت سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. بنا بر این نیاز به طراحی سیستم‌هایی که قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران با تأکید بر حداقل مداخله انسانی باشند از یک سو و روی آوردن به روش‌های تحلیل متناسب با حجم داده‌های حجیم از سوی دیگر، به‌خوبی احساس می‌شود. در حال حاضر، داده‌کاوی مهم‌ترین فناوری برای بهره‌برداری مؤثر، صحیح و سریع از داده‌های حجیم است و اهمیت آن رو به فزونی است.

با توجه به‌وجود اطلاعات ارزشمند در پایگاه‌های داده‌ای در اواخر دهه ۸۰ میلادی، تلاش برای استخراج و استفاده از اطلاعات پایگاه‌های داده‌ای شروع شد. داده‌کاوی فرایندی است که در آغاز دهه ۹۰ پا به عرصه ظهور گذاشته و با نگرشی نو، به مسئله استخراج اطلاعات از پایگاه داده‌ها می‌پردازد. در سال ۱۹۸۹ و ۱۹۹۱ کارگاه‌های کشف دانش از پایگاه داده‌ها توسط پیاتتسکی و همکارانش و در فاصله سال‌های ۱۹۹۱ تا ۱۹۹۴کارگاه‌های فوق، توسط فایاد و پیاتتسکی و دیگران برگزار شد. به‌طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط « فیاض » در اولین کنفرانس بین‌المللی « کشف دانش و داده‌کاوی » در سال ۱۹۹۵ مطرح شد. از سال ۱۹۹۵ داده‌کاوی به صورت جدی وارد مباحث آمار شد [۸] و در سال ۱۹۹۶، اولین شماره مجله کشف دانش از پایگاه داده‌ها