فصلنامه گزیده‌مطالب آماری
سال ۱۷، شماره ۱، بهار ۱۳۸۵، صص ۱۱۱-۹۹

استفاده از روش چنگک‌زنی تعمیم‌یافته به‌منظور ایجاد سازگاری در جدول‌های حاصل از سرشماری

آرمان بیداربخت‌نیا

مرکز آمار ایران

چکیده. روش‌های تعدیل وزنی، اغلب در آمارگیری‌های نمونه‌ای به‌منظور سازگار کردن توزیع‌های نمونه‌ای با توزیع‌های جامعه به کار می‌رود. این روش‌ها، علاوه بر آمارگیری‌های نمونه‌ای، ممکن است در سرشماری همراه با نمونه‌گیری نیز مورد استفاده قرار گیرند. در سرشماری همراه با نمونه‌گیری، مسئله سازگاری جدول‌ها از اهمیت خاصی برخوردار بوده و مستلزم استفاده از روش‌های براورد پیچیده‌ای است. این مقاله به معرفی دو روش پس‌طبقه‌بندی و چنگک‌زنی و چگونگی استفاده از این روش‌ها به‌منظور ایجاد سازگاری در جدول‌های حاصل از سرشماری عمومی نفوس و مسکن ۱۳۸۵ می‌پردازد.

۱- مقدمه

در آمارگیری‌های خانواری، اغلب برای جبران آثار ناشی از بی‌پاسخی و نقص چارچوب، از روش‌های تعدیل وزنی برای براورد پارامتر(های) جامعه استفاده می‌شود. در این روش‌ها وزن‌های پایه‌ای طرح با استفاده از اطلاعات کمکی موجود به‌گونه‌ای تعدیل می‌شوند که توزیع‌های نمونه‌ای با توزیع‌های جامعه‌ای همگون شود. اطلاعات کمکی از منابع مختلفی فراهم می‌شود. داده‌های ثبتی، پیش‌بینی‌های جمعیتی، نتایج آمارگیری‌های معتبر مثل نیروی کار و هزینه و درامد خانوار و اطلاعات حاصل از مرحله اول در آمارگیری‌های چند

واژگان کلیدی: تعدیل وزنی؛ وزن پایه‌ای طرح؛ براوردگر هورویتز- تامپسون؛ سرشماری همراه با نمونه‌گیری.

مرحله‌ای، از جمله منابعی است که می‌توان اطلاعات کمکی را از آن‌ها استخراج کرد. روش‌های متعددی برای انجام این تعدیل‌ها به کار می‌رود. پس‌طبقه‌بندی، چنگک‌زنی (Raking)، مدل‌بندی رگرسیون لوژستیکی و براورد رگرسیونی تعمیم‌یافته، روش‌هایی هستند که با توجه به اهداف و نوع داده‌های کمکی می‌توان برای انجام تعدیل وزنی به کار برد. روش‌های نام‌برده همگی عضو خانوده‌ای از براوردگرها به‌نام براوردگرهای کالیبره (Calibration Estimators) هستند.

تعدیل‌های وزنی به‌واسطه کارکرد آن‌ها در ایجاد سازگاری بین توزیع‌های نمونه‌ای و توزیع‌های جامعه‌ای، علاوه بر آمارگیری‌های نمونه‌ای، ممکن است در سرشماری‌ها نیز مورد استفاده قرار گیرند.

مرکز آمار ایران تصمیم دارد تا به‌منظور بالا بردن کیفیت نتایج حاصل از سرشماری عمومی نفوس و مسکن، از طریق کاهش خطاهای غیر نمونه‌گیری و همچنین ارائه اطلاعات بیش‌تر در نواحی کوچک جغرافیایی، برای زیرگروه‌های مختلف جمعیتی و در عین حال کاهش بار پاسخگو در سرشماری عمومی نفوس و مسکن ۱۳۸۵، این سرشماری را به‌روش سرشماری همراه با نمونه‌گیری انجام دهد. به این ترتیب برخی از اطلاعات (اقلام عمومی) برای تمام افراد، خانوارها و واحدهای مسکونی در جامعه جمع‌آوری می‌شوند و اطلاعات دیگری تحت عنوان اقلام تفصیلی فقط از بخشی از افراد، خانوارها و واحدهای مسکونی که به‌عنوان نمونه انتخاب شده‌اند، جمع‌آوری می‌گردد. به این ترتیب در سرشماری عمومی نفوس و مسکن ۱۳۸۵ از سه نوع فرم برای جمع‌آوری اطلاعات مربوط به خانوارها استفاده می‌شود. فرم ۲ (پرسشنامه خانوار– اقلام عمومی) که فقط شامل اطلاعات عمومی است. فرم ۳ (پرسشنامه خانوار– اقلام عمومی و تفصیلی) که علاوه بر اقلام عمومی، اقلام تفصیلی را نیز شامل می‌شود و فرم ۴ (پرسشنامه خانوار مؤسسه‌ای)[1]. نحوه انجام سرشماری به این ترتیب است که در هر یک از حوزه‌های سرشماری (محدوده کار یک مامور سرشماری)، برای هر کدام از خانوارهای مؤسسه‌ای، فرم ۴، برای هر کدام از خانوارهای معمولی غیر ساکن، فرم ۳ و برای خانوارهای معمولی ساکن و خانوارهای گروهی، همزمان با تهیه فهرست خانوارهای حوزه، بر اساس یک فاصله نمونه‌گیری از پیش تعیین شده و به‌روش سیستماتیک خطی، خانوارهایی به‌عنوان

خانوارهای نمونه انتخاب شده و برای این خانوارها فرم ۳ و بـرای بقیـه خانوارهـای حـوزه فرم ۲ تکمیل می‌شود.

واضح است که در جدول‌های حاصل از نتایج سرشماری، جدول‌های مربوط به اقـلام عمومی به‌طور مستقیم از فایل متشکل از مجموعه فرم‌های ۲ و ۳ و ۴ استخراج می‌شـود، در حالی‌که جدول‌های مربوط به اقلام تفصیلی و همچنین تقاطع اقلام عمومی با تفصیلی (دو یا چند طرفه) فقط با استفاده از اطلاعات فرم ۳ و با بـه‌کـارگیری روش‌هـای بـراورد، استخراج می‌شود.

آن‌چه در این جدول‌ها بیش از هر عامـل دیگـری اهمیـت دارد، سـازگاری آن‌هـا بـا جدول‌هایی است که فقط از اقلام عمومی و به‌طور مستقیم حاصل شده‌اند. به این ترتیـب در براوردهای انجام شده برای استخراج نتایج سرشماری عمومی نفـوس و مـسکن ۱۳۸۵، از تعدیل‌های وزنی مناسب استفاده می‌شود که اطلاعات کمکی برای انجام این تعدیل‌هـا، شمارش‌های حاصل از فرم‌های ۲ و ۳ برای اقـلام عمـومی در هـر یـک از سـطوح و یـا ترکیب‌هایی از سطوح اقلام عمومی است.

دو روشی که در سرشماری عمومی نفوس و مـسکن ۱۳۸۵ بـرای بـراورد اطلاعـات مربوط به اقلام تفـصیلی مـورد اسـتفاده قـرار مـی‌گیرنـد، عبارتنـد از پـس‌طبقـه‌بنـدی و چنگک‌زنی تعمیم‌یافته که در این مقاله علاوه بر معرفی هر دو روش، بـه بررسـی مفـصل چنگک‌زنی تعمیم‌یافته در قالب یک مثال پرداخته می‌شود، ضمن این‌که مراحل محاسباتی این روش به‌گونه‌ای که برای برنامه‌نویسی نیز آسان باشد، شرح داده شده است.

۲- پس‌طبقه‌بندی

جامعه U با N عضو را در نظر بگیرید که نمونه S با n عضو از آن انتخاب شده اسـت. پاسخ jامین واحد جامعه را با Y_j، $j = ۱,\ldots,N$ و پاسخ اخذ شده از iامین واحد نمونه را با y_i، $i = ۱,\ldots,n$، نشان می‌دهیم. هدف، بـراورد مقـدار کـل جامعـه بـرای متغیـر مـورد بررسی Y است. فرض کنید پس از انجام نمونه‌گیری جامعه را به G طبقـه تقـسیم‌بنـدی کرده و اندازه جمعیت در طبقه gام را با N_g و اندازه نمونه متعلق بـه آن را بـا n_g نـشان می‌دهیم که $g = ۱,\ldots,G$. مثلاً اگر جمعیت فعال کشور را بر اسـاس همـه ترکیـب‌هـای

ممکن حاصل از سطوح متغیرهای جنس (زن و مرد)، وضع سواد (باسواد و بی‌سواد) و وضع فعالیت (شاغل و بیکار) طبقه‌بندی کنیم، خواهیم داشت:

اگر π_i احتمال انتخاب شدن عضو iام در نمونه باشد، وزن پایه برای عضو iام عبارت است از:

$$d_i = \frac{1}{\pi_i}$$

که این عدد در سرشماری همراه با نمونه‌گیری تقریباً برابر با فاصله نمونه‌گیری است. به این ترتیب براورد پس‌طبقه‌بندی برای مقدار کل جامعه، $t_y = \sum_{j=1}^{N} Y_j$، عبارت است از:

$$\hat{t}_{y(pos)} = \sum_{g=1}^{G} \frac{N_g}{\hat{N}_g} \hat{t}_{y(g)} \qquad (1)$$

که در آن $\hat{t}_{y(g)} = \sum_{i=1}^{n_g} d_{gi} y_{gi}$، براوردگر هوروتیز– تامپسون برای مقدار کل متغیر Y در پس‌طبقه gام و $\hat{N}_g = \sum_{i=1}^{n_g} d_{gi}$ براورد تعداد در پس‌طبقه gام است و d_{gi} عبارت است از: وزن پایه عضو iام در پس‌طبقه jام.

با توجه به این‌که با تمام اقلام سرشماری به شکل متغیرهای گسسته برخورد می‌شود، لذا متغیر Y در روابط بالا همواره به صورت یک متغیر دو حالتی است که به شکل زیر تعریف می‌شود.

$$Y_i = \begin{cases} 1 & \text{نمونه } i \text{ام دارای صفت مورد نظر باشد} \\ 0 & \text{نمونه } i \text{ام دارای صفت مورد نظر نباشد} \end{cases}$$

به این ترتیب براورد جمعیت دارای صفت مورد نظر در پس‌طبقه gام به صورت $\hat{t}_{y(g)} = \sum_{i=1}^{n'_g} d_{gi}$ محاسبه می‌شود که n'_g عبارت است از تعداد عناصر نمونه متعلق به پس‌طبقه gام که دارای صفت مورد نظر هستند و d_{gi} وزن پایه برای واحد نمونه iام در پس‌طبقه gام است.

این روش را به‌دلیل این‌که برای تمام ترکیب‌های ممکن از سطوح متغیرهای پس‌طبقه‌بندی سازگاری ایجاد می‌کند، روش پس‌طبقه‌بندی کامل (Complete Poststratification) یا وزن‌دهی خانه‌ای (Cell Weighting) گویند.

در مواردی اندازه نمونه در برخی از پس‌طبقه‌ها صفر است یک روش آن است که برخی پس‌طبقه‌ها در یکدیگر ادغام می‌شوند.

اما روشی که بیش‌تر مورد استفاده قرار می‌گیرد و کمتر با مشکل اندازه نمونه صفر مواجه است، روش « پس‌طبقه‌بندی ناقص » (Incomplete Poststratification) است که پس‌طبقه‌بندی را روی حاشیه‌ها انجام می‌دهد. این روش به‌طور کلی شبیه به‌روش پس‌طبقه‌بندی است، با این تفاوت که در این روش فقط برای حاشیه‌های یک متغیر سازگاری ایجاد می‌شود. مزیت این روش بر روش پس‌طبقه‌بندی کامل، طبق آن‌چه گفته شد، این است که در این‌جا کمتر با مشکل اندازه نمونه صفر مواجه هستیم. برای نمونه، فرض کنید در مثال قبل فقط برای متغیر جنس (مرد و زن)، سازگاری مورد نیاز باشد. به این ترتیب $G = ۲$، به این ترتیب بر خلاف مثال قبل که برای تمام ترکیب‌های ممکن حاصل از سطوح سه متغیر جنس، وضع سواد و وضع فعالیت نیاز به ایجاد سازگاری بود، در این حالت که فقط سازگاری برای متغیر جنس لازم است، احتمال این‌که اندازه نمونه در سطوح مورد نظر برای سازگاری (هر یک از سطوح متغیر جنس) صفر باشد خیلی کمتر است.

در پس‌طبقه‌بندی برای حاشیه‌ها اگر پس‌طبقه‌بندی شامل بیش از یک متغیر باشند مثلاً در مورد ذکر شده، اگر بخواهیم به‌طور هم‌زمان برای حاشیه‌های سه متغیر سازگاری ایجاد شود، یعنی $G = ۲ + ۲ + ۲ = ۶$، دیگر روش پس‌طبقه‌بندی قابل استفاده نیست. برای ایجاد سازگاری با شرایط بالا، در سرشماری عمومی نفوس و مسکن ۱۳۸۵ از روش چنگک‌زنی تعمیم‌یافته استفاده شده است که در ادامه معرفی می‌شود.

شایان ذکر است که در سرشماری عمومی نفوس و مسکن ۱۳۸۵، در مورد جدول‌های خانواری و واحد مسکونی، به‌دلیل نیاز به سازگاری در تمام ترکیب‌های ممکن از سطوح متغیرهای پس‌طبقه‌بندی، برای ایجاد سازگاری در جدول‌ها از روش پس‌طبقه‌بندی کامل استفاده شده است. البته مشکل خانه‌های با اندازه نمونه صفر در این

مورد نیز همچنان باقی است، که با ادغام برخی از سطوح می‌توان آن را برطرف نمود.

۳- چنگک‌زنی

در پس‌طبقه‌بندی برای حاشیه‌ها، اگر بیش از یک متغیر پس‌طبقه‌بندی داشته باشیم، روشی مشابه با پس‌طبقه‌بندی ناقص، ولی به صورت تکراری مورد استفاده قرار می‌گیرد. این روش را چنگک‌زنی یا برازش متناسب تکراری (Iterative Proportional Fitting) گوییم.

فرض کنید بخواهیم برای حاشیه‌های متغیرهای جنس و وضع سواد به‌طور همزمان سازگاری ایجاد شود. این‌کار باید به صورت تکراری و در چند مرحله انجام گیرد. در مرحله اول، براورد پس‌طبقه‌بندی ناقص را روی حاشیه‌های سطری (یکی از متغیرهای جنس و وضع سواد) به دست می‌آوریم تا برای حاشیه‌های آن متغیر، سازگاری ایجاد شود. در این صورت مشاهده می‌شود که برای حاشیه‌های ستونی، سازگاری وجود ندارد. در مرحله دوم، براورد پس‌طبقه‌بندی ناقص را برای حاشیه‌های ستونی به دست می‌آوریم تا برای آن‌ها نیز سازگاری ایجاد شود. با انجام مرحله دوم، سازگاری حاصل از مرحله نخست از بین می‌رود، لذا در مرحله سوم، مجدداً برای حاشیه‌های سطری، براوردهای پس‌طبقه‌بندی ناقص را محاسبه می‌کنیم. این فرایند به‌طور یک در میان روی حاشیه‌های سطری و ستونی تا جایی ادامه می‌یابد که برای تمام حاشیه‌ها، سازگاری ایجاد شود.

در یک آمارگیری به وسعت سرشماری همراه با نمونه‌گیری، به‌طور هم‌زمان به‌دنبال ایجاد سازگاری در تعداد زیادی متغیر با سطوح نسبتاً زیاد هستیم. برای رفع این مشکل در سرشماری عمومی نفوس و مسکن ۱۳۸۵ حالت تعمیم‌یافته‌ای از روش چنگک‌زنی مورد استفاده قرار می‌گیرد.

۴- چنگک‌زنی تعمیم‌یافته

در این روش به‌دنبال وزن‌هایی هستیم که بر اساس یک تابع فاصله مشخص، کمترین فاصله تا وزن‌های پایه‌ای را داشته و در عین حال براوردهایی تولید کنند که با جمعیت کل

برای صفت مورد نظر در حاشیه‌ها برابر باشند.

یک ارائه $X_{n\times L}$ را برای متغیرهای کمکی در نظر بگیرید که در آن n تعداد اعضای نمونه و L تعداد سطوحی است که می‌خواهیم برای آن‌ها سازگاری ایجاد شود. برای عضو iام نمونه، بردار متغیرهای کمکی به صورت $\underset{\sim}{X}_i = (x_{i1},\dots,x_{iL})'$ تعریف می‌شود. وزن پایه برای عضو iام نمونه برابر با d_i و وزن نهایی برابر با $w_i = g_i d_i$ است که g_i را فاکتور تعدیل وزنی گوییم. برای به دست آوردن براوردهای تعدیل شده، به‌دنبال w_i هایی هستیم که علاوه بر مینیمم کردن تابع فاصله

$$\Delta(\underset{\sim}{w},\underset{\sim}{d}) = \sum_{i=1}^{n}\left(w_i \log\left(\frac{w_i}{d_i}\right) - w_i + d_i \right) \qquad (۲)$$

در معادلات سازگاری $\sum_{i=1}^{n} w_i \underset{\sim}{X}_i = \underset{\sim}{t}_x$ نیز صدق کنند. $\underset{\sim}{t}_x$ عبارت است از یک بردار ستونی L بعدی که هر یک از عناصر آن مقدار کل معلوم جامعه برای یکی از سطوح متغیرهای کمکی است. مقادیر w_i مورد نظر پس از حل معادلات لاگرانژ و با استفاده از روش‌های مبتنی بر تکرار، حاصل می‌شود. این تکرارها تا جایی ادامه می‌یابد که سازگاری کامل برقرار شده یا بر اساس معیارهای از پیش تعیین شده به همگرایی مورد نظر برسیم.

روش محاسبه w_i با استفاده از حل معادلات لاگرانژ به‌گونه‌ای که برای برنامه‌نویسی نیز مناسب باشد، در پیوست آمده است. بر پایه این روش، به کمک رابطه‌های (۵) تا (۷) مقادیر w_i مشخص می‌شوند. به این ترتیب عامل تعدیل g_i با استفاده از رابطه زیر به دست می‌آید:

$$g_i = \exp(X_i' \underset{\sim}{\lambda}) \qquad (۳)$$

که $\underset{\sim}{\lambda}$ بردار ضرایب لاگرانژ است. وزن نهایی برای عضو iام نمونه به صورت $w_i = g_i d_i$ حاصل می‌شود.

۵- مثال

یک جامعه ۱۰۰۰۰۰ نفری را در نظر بگیرید. فرض کنید ویژگی‌های جنس، سن و وضع

سواد به‌ترتیب در ۲ سطح، ۵ سطح و ۲ سطح از تمام افراد جامعه مورد پرسش قرار می‌گیرد. فرض کنید برای ۳۰۰۰ نفر از این جامعه، به‌طور تصادفی علاوه بر ویژگی‌های فوق، توانایی صحبت کردن به زبان فارسی[۲] نیز در ۳ سطح پرسش می‌شود. جدول‌هایی که از اطلاعات مربوط به تمام افراد به‌طور مستقیم حاصل می‌شود، جدول جنس– سن و جنس– وضع سواد است و جدول‌های حاصل از نمونه که باید براورد شود، سن– جنس– توانایی صحبت کردن به زبان فارسی و جنس– وضع سواد– توانایی صحبت کردن به زبان فارسی است. شایان ذکر است که ویژگی‌های سن و جنس برای تمام افراد و وضع سواد فقط برای افراد بالای ۶ سال پرسیده شده است. به این ترتیب علاقه‌مند هستیم تا جدول‌های حاصل از براورد با جدول‌های مستقیم از جامعه سازگاری داشته باشد. به این ترتیب سطوحی که می‌خواهیم برای آن‌ها سازگاری ایجاد شود، عبارتند از $۲\times۵=۱۰$ سطح برای جدول سن– جنس و $۲\times۲=۴$ سطح از جدول جنس– وضع سواد، یعنی $L=۱۰+۴=۱۴$.

نتایج مستقیم از جامعه در جدول ۱ و جدول ۲ نشان داده شده است.

با توجه به وزن پایه که عبارت است از $d_i=\frac{۱۰۰}{۳}$ $(i=۱,\ldots,۳۰۰۰)$، براوردگر هوروتیز– تامپسون برای هر یک از جدول‌ها نمونه‌ای در جدول ۳ و جدول ۴ قابل مشاهده است. با مقایسه این دو جدول با جدول ۱ و جدول ۲ مشخص می‌شود، که سرجمع‌های حاصل، با یکدیگر سازگاری ندارند. اما پس از تعدیل وزن‌های پایه برای واحدهای نمونه با استفاده از روش چنگک‌زنی تعمیم‌یافته که نتایج آن در جدول ۵ و جدول ۶ مشاهده می‌شود، خواهیم دید که سازگاری مورد نظر با جدول ۱ و جدول ۲ حاصل شده است. همه محاسبات با برنامه‌نویسی به‌وسیله نرم‌افزار SAS/IML انجام شده است. برنامه مورد نظر پس از ۳ تکرار به همگرایی رسیده است.

جدول۱- جمعیت افراد به‌تفکیک جنس و وضع سواد

جنس	وضع سواد		کل
	بی‌سواد	باسواد	
زن	۷۳۶۷	۲۹۳۰۳	۳۶۶۷۰
مرد	۷۴۲۴	۲۹۳۸۶	۳۶۸۱۰
کل	۱۴۷۹۱	۵۸۶۸۹	۷۳۴۸۰

جدول۲- جمعیت افراد به‌تفکیک جنس و سن

گروه سنی	جنس		کل
	زن	مرد	
۱	۶۲۴۹	۶۲۱۴	۱۲۴۶۳
۲	۲۴۴۲۱	۲۴۴۷۹	۴۸۹۰۰
۳	۱۵۸۸۳	۱۵۸۶۳	۳۱۷۴۶
۴	۳۰۸۶	۳۲۳۸	۶۳۲۴
۵	۲۸۱	۲۸۶	۵۶۷
کل	۴۹۹۲۰	۵۰۰۸۰	۱۰۰۰۰۰

جدول۳- براورد هورویتز– تامپسون تعداد افراد بر حسب سن و توانایی صحبت کردن بـه زبـان فارسـی به‌تفکیک جنس

مرد					زن				
گروه سنی	توانایی صحبت کردن به زبان فارسی			کل	گروه سنی	توانایی صحبت کردن به زبان فارسی			کل
	۱	۲	۳			۱	۲	۳	
۱	۵۱۳۳	۱۴۰۰	۱۶۷	۶۷۰۰	۱	۴۳۰۰	۱۴۳۳	۶۷	۵۸۰۰
۲	۱۷۴۳۳	۵۷۰۰	۳۳۳	۲۳۴۶۷	۲	۱۸۵۳۳	۵۶۶۷	۳۶۶	۲۴۵۶۷
۳	۱۱۸۶۷	۴۵۳۳	۵۳۳	۱۶۹۳۳	۳	۱۱۷۰۰	۳۷۰۰	۱۳۳	۱۵۵۳۳
۴	۲۵۳۳	۹۶۷	۱۰۰	۳۶۰۰	۴	۲۳۳۳	۵۳۳	۶۷	۲۹۳۳
۵	۲۳۳	۶۷	۰	۳۰۰	۵	۱۳۳	۳۳	۰	۱۶۷
کل	۳۷۲۰۰	۱۲۶۶۷	۱۱۳	۵۱۰۰۰	کل	۳۷۰۰۰	۱۱۳۶۷	۶۳۳	۴۹۰۰۰

جدول۴- براورد هورویتز– تامپسون تعداد افراد بر حسب وضع سواد و توانایی صحبت کردن به زبان فارسی به‌تفکیک جنس

مرد					زن				
وضع سواد	توانایی صحبت کردن به زبان فارسی			کل	وضع سواد	توانایی صحبت کردن به زبان فارسی			کل
	۱	۲	۳			۱	۲	۳	
بی‌سواد	۵۸۰۰	۱۹۳۳	۱۶۷	۷۴۶۷	بی‌سواد	۶۱۰۰	۱۳۳۳	۳۳	۷۴۶۷
باسواد	۲۱۲۳۳	۷۸۳۳	۷۳۳	۲۸۶۳۳	باسواد	۲۱۴۶۷	۶۷۳۳	۴۳۳	۲۸۶۳۳
کل	۲۷۰۳۳	۹۷۶۷	۹۰۰	۳۶۱۰۰	کل	۲۷۵۶۷	۸۰۶۷	۴۶۶	۳۶۱۰۰

جدول۵- براورد تعدیل‌شده تعداد افراد بر حسب سن و توانایی صحبت کردن به زبان فارسی به‌تفکیک جنس

مرد					زن				
گروه سنی	توانایی صحبت کردن به زبان فارسی			کل	گروه سنی	توانایی صحبت کردن به زبان فارسی			کل
	۱	۲	۳			۱	۲	۳	
۱	۴۷۶۱	۱۲۹۸	۱۵۵	۶۲۱۴	۱	۴۶۳۳	۱۵۴۴	۷۲	۶۲۴۹
۲	۱۸۱۹۷	۵۹۳۸	۳۴۵	۲۴۴۷۹	۲	۱۸۴۲۰	۵۶۳۶	۳۶۵	۲۴۴۲۱
۳	۱۱۱۰۳	۴۲۵۶	۵۰۴	۱۵۸۶۳	۳	۱۱۹۵۸	۳۷۸۸	۱۳۷	۱۵۸۸۳
۴	۲۲۸۰	۸۷۰	۸۸	۳۲۳۸	۴	۲۴۵۲	۵۶۳	۷۱	۳۰۸۶
۵	۲۲۳	۶۳	۰	۲۸۶	۵	۲۲۵	۵۶	۰	۲۸۱
کل	۳۶۵۶۳	۱۲۴۲۵	۱۰۹۲	۵۰۰۸۰	کل	۳۷۶۸۸	۱۱۵۸۷	۶۴۵	۴۹۹۲۰

جدول۶- براورد تعدیل‌شده تعداد افراد بر حسب وضع سواد و توانایی صحبت کردن به زبان فارسی به‌تفکیک جنس

مرد					زن				
وضع سواد	توانایی صحبت کردن به زبان فارسی			کل	وضع سواد	توانایی صحبت کردن به زبان فارسی			کل
	۱	۲	۳			۱	۲	۳	
بی‌سواد	۵۴۴۰	۱۸۲۷	۱۵۷	۷۴۲۴	بی‌سواد	۶۰۲۰	۱۳۱۵	۳۲	۷۳۶۷
باسواد	۲۰۹۸۱	۷۶۹۶	۷۰۹	۲۹۳۸۶	باسواد	۲۱۹۷۳	۶۸۸۸	۴۴۲	۲۹۳۰۳
کل	۲۴۶۲۱	۹۵۲۳	۸۶۶	۳۶۸۱۰	کل	۲۷۹۹۳	۸۲۰۳	۴۷۴	۳۶۶۷۰

۶- سپاسگزاری

با سپاس فراوان از اعضای محترم کمیته سرشماری توأم با نمونه‌گیری[۳] که کاربردی کردن این روش و استفاده از آن در سرشماری عمومی نفوس و مسکن ۱۳۸۵ بدون راهنمایی‌های ارزشمند و پیشنهادهای سازنده ایشان، میسر نبود.

توضیحات

[۱] برای اطلاع از جزئیات مربوط به فرم‌های مزبور، به راهنماهای مربوط در سرشماری عمومی نفوس و مسکن ۱۳۸۵ مراجعه شود.

[۲] در اجرای آزمایشی سال ۱۳۸۴ سرشماری نفوس و مسکن ۱۳۸۵ « توانایی صحبت کردن به زبان فارسی »، یکی از پرسش‌هایی است که در آزمایش پرسیده شده اما در اجرای اصلی این سرشماری منظور نشده است.

[۳] « کمیته سرشماری توأم با نمونه‌گیری » یکی از کمیته‌های گروه تهیه طرح سرشماری عمومی نفوس و مسکن ۱۳۸۵ است.

مرجع‌ها

[1] Devill, J. C., and Sarndal, C-E. and Sautory. O. (1993), "Generalized Raking Procedures in Survey Sampling," journal of the American Statistical Association, 88, 1013-1020.

[2] Groves, R.M, Dillman, D.A., Eltinge, J.L., Little, R.J.(2002), Survey Nonresponse, New York : Willey.

[3] Kalton, G., Flores-Cervantes, I. (2003), "Weighting Methods," Journal of Official Statistics, Vol. 19, No. 2, 81-97.

[4] Singh, A.C., and Mohl, C.A. (1996), "Understanding calibration estimators in survey sampling," Survey Methodology. 22, 107-115.

پیوست: نحوه محاسبه w_iها در روش چنگک‌زنی تعمیم‌یافته

برای محاسبه w_iها به‌شیوه زیر عمل می‌کنیم.

فرض کنید I متغیر کمکی داریم. می‌خواهیم وزن‌هایی تولید کنیم که براوردهای نمونه‌ای حاصل از آن وزن‌ها برای L سطح حاصل از حاشیه‌های متغیرهای کمکی و یا ترکیب‌هایی از سطوح آن‌ها، با مقدارهای معلوم جامعه سازگار باشد.

ارائه $X_{n\times L}$ را طوری تشکیل می‌دهیم که به‌ازای هر واحد نمونه یک سطر و به‌ازای هر سطح، یک ستون داشته باشد. در هر یک از ستون‌ها، عنصر iام نمونه ($i=1,\dots,n$) بسته به این‌که عنصر مزبور به سطح مربوط تعلق داشته یا نداشته باشد، به‌ترتیب مقدار یک یا صفر می‌گیرد.

$\underset{\sim}{d}$: عبارت است از یک بردار ستونی با n سطر که درایه‌های آن وزن پایه برای اعضای نمونه هستند.

$\underset{\sim}{t}_x$: یک بردار ستونی با L سطر که درایه‌های آن عبارت است از جمعیت حاصل از سرشماری، برای سطح mام، ($m=1,\dots,L$).

$w^{(v)}$: یک بردار ستونی با n سطر که درایه iام آن برابر است با وزن تولید شده برای نمونه iام در تکرار vام.

$\hat{\underset{\sim}{t}}_x^{(v)}$: یک بردار ستونی با L سطر که درایه‌های آن عبارت است از براورد موزون تعداد افراد برای سطح mام در تکرار vام.

$$\hat{\underset{\sim}{t}}_x^{(v)} = X' w^{(v)} \qquad (۴)$$

برای محاسبه وزن‌های $w^{(v)}$ باید ابتدا فاکتور $f^{(v)}$ را با استفاده از رابطه زیر به دست آوریم.

$$f^{(v)} = X(X'\Gamma_{(v-1)}X)^{-1}(\underset{\sim}{t}_x - \hat{\underset{\sim}{t}}_x^{(v-1)}) \qquad (۵)$$

در رابطه‌ی بالا داریم:

$\Gamma_{(v-1)}$: یک ارائه قطری $n \times n$ که عنصر iام روی قطر آن عبارت است از $w_i^{(v-1)}$ که وزن تولید شده برای نمونه iام در تکرار $(v-1)$ام است.

برای حالتی که $v=1$ باشد باید موارد زیر را در نظر داشت.

$$w^{(0)} = \underset{\sim}{d}$$

به این ترتیب داریم:

$$\underset{\sim}{\hat{t}}_x^{(0)} = X'\underset{\sim}{d}$$

$\Gamma_{(0)}$: ارائه قطری است که عنصر iام روی قطر آن عبارت است از وزن پایه برای نمونه iام.

پس از محاسبه $f^{(v)}$ در تکرار vام، بردار $g^{(v)}$ به صورت زیر محاسبه می‌شود.

$$g^{(v)} = g^{(v-1)} \# \exp(f^{(v)}) \qquad (۶)$$

که نماد# نشان‌دهنده این است که درایه‌های متناظر دو بردار در یکدیگر ضرب معمولی شود و $\exp(f^{(v)})$ برداری است که هر عنصر آن، تابع exp عنصر متناظر در $f^{(v)}$ است.

سپس بردار وزن $w^{(v)}$ در تکرار vام با استفاده از رابطه زیر محاسبه می‌شود.

$$w^{(v)} = \underset{\sim}{d} \# g^{(v)} \qquad (۷)$$

پس از انجام هر تکرار یک شاخص محاسبه می‌شود.

$$d = norm_{(v)} - norm_{(v-1)}$$
$$norm_{(v)} = \left\| t_x - \underset{\sim}{\hat{t}}_x^{(v)} \right\|$$

که نماد $\| \ \|$ نشانگر نرم تفاضل هندسی دو بردار است. در صورتی که $d \le \varepsilon$ باشد تکرارها متوقف شده و بردار $w^{(v)}$ در آخرین تکرار، بردار وزن‌های نهایی است که با استفاده از آن می‌توان تمام جدول‌های مربوط را به دست آورد. مقدار ε عدد مثبت و کوچکی است که بر اساس عواملی از قبیل زمان اجرای برنامه و میزان سازگاری مورد نظر تعیین می‌شود.