مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۳، شماره‌ی ۲، پاییز و زمستان ۱۳۹۱، صص ۱۹۹ – ۲۲۴

براورد مجموع در آمارگیری‌های دوچارچوبی

مرجان نورینی* و انور قیطولی

مرکز آمار ایران

چکیده: در آمارگیری‌های نمونه‌ای، چارچوبی که تقریباً تمام واحدهای جامعه‌ی هدف را پوشش دهد و نمونه‌گیری از آن کم‌هزینه باشد، کم‌تر یافت می‌شود. در حالی که ممکن است چارچوب‌های دیگری وجود داشته باشند که هر یک کسری از جامعه‌ی هدف را پوشش داده و در عین حال نمونه‌گیری از آن‌ها نیز کم‌هزینه باشد. در این مقاله نخست مفاهیم، مزایا و معایب چارچوب چندگانه تشریح شده و سپس روش‌های مختلف برای براورد پارامترها مطرح شده است. در نهایت با استفاده از شبیه‌سازی مونت کارلویی، این روش‌ها با یکدیگر مقایسه شده‌اند.

واژگان کلیدی: آمارگیری نمونه‌ای؛ براورد مجموع؛ چارچوب دوگان؛ شبیه‌سازی مونت کارلویی.

۱- مقدمه

یکی از مهم‌ترین مشکلات در اجرای آمارگیری‌های نمونه‌ای، چارچوب‌های فهرستی ناقص و یا تاریخ گذشته‌ای (Out of date) هستند که برای انتخاب نمونه مورد استفاده قرار می‌گیرند. در آمارگیری‌های نمونه‌ای که این نوع چارچوب‌ها مورد استفاده قرار می‌گیرند، براورد پارامترهای جامعه دچار اریبی می‌شود. روزامد کردن یک فهرست، کار پرهزینه و دشواری است که البته به یُمن پیشرفت‌هایی که در مدیریت بانک‌های اطلاعاتی به وجود آمده، این کار تا حدودی امکان‌پذیر شده است. در هر حال، مهم‌ترین و پرهزینه‌ترین مرحله در روزامدسازی یک فهرست، عملیات جمع‌آوری داده‌های مورد نیاز است که باید با استفاده از راهکارهای مناسب، هزینه را کاهش و در عین حال دقت را افزایش داد.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۱/۱۰/۲۶، پذیرش: ۱۳۹۲/۹/۴.

گاهی اوقات برای به دست آوردن فهرست واحدهای نمونه‌گیری (چارچوب)، از چند فهرست که در ارتباط با جامعه‌ی مورد بررسی است، استفاده می‌شود. در این حالت فرض می‌شود که این چارچوب‌های مختلف تقریباً کل جامعه را پوشش می‌دهند. مثلاً می‌توان با استفاده از منابع اداری، فهرست به دست آمده از یک سرشماری را روزامد کرد. در این حالت چارچوب حاصل، بر اساس دو یا چند فهرست است. این روش را تنها وقتی می‌توان به کار برد که فهرست‌های مختلف، اطلاعات مورد نیاز برای تکمیل چارچوب را شامل شوند و اتصال داده‌ها و رکوردها به یکدیگر به خوبی انجام گیرد، به‌طوری که نتایج قابل اطمینانی را به دنبال داشته باشد. در غیر این صورت چارچوب حاصل ناقص بوده و واحدهای تکراری زیادی خواهد داشت. راه دیگر، استفاده از فهرست‌های مختلف به روش چارچوب چندگانه است. به عبارت دیگر، این فهرست‌ها ترکیبی از چارچوب‌های فهرستی و یا ترکیبی از چارچوب فهرستی به همراه چارچوب ناحیه‌ای هستند. در این روش، از ترکیب براوردگرهای حاصل از واحدهای نمونه‌گیری متعلق به بخش غیر متداخل چارچوب‌ها با براوردگر حاصل از واحدهای نمونه‌گیری متعلق به بخش متداخل چارچوب‌ها، یک براوردگر به دست می‌آید.

۲- تاریخچه

شاید بتوان گفت اولین شالوده‌ی آمارگیرهای چندچارچوبی در سال ۱۹۴۹ با آمارگیری از فروشگاه‌های خرده‌فروشی که مجری آن دفتر سرشماری آمریکا بود، گذاشته شد. در این نمونه‌گیری که از دو چارچوب ناحیه‌ای و فهرستی استفاده شده بود، واحدهای آماری مشترک در دو چارچوب، شناسایی و از چارچوب ناحیه‌ای حذف شدند. بدین ترتیب با وجود پیچیدگی طرح، براورد مجموع جامعه از ترکیب براوردهای حاصل از دو چارچوب به دست آمد. از آنجایی‌که گاه شناسایی واحدهای آماری مشترک در دو یا چند چارچوب در حین عملیات میدانی امکان‌پذیر نیست، انجام نمونه‌گیری و تکمیل پرسشنامه برای واحدهای آماری مشترک، اجتناب‌ناپذیر است. به‌منظور جبران ورود بیش از یک بار واحدهای مشترک در نمونه هنگام براورد، استفاده از تعدیل وزنی ضروری است. یکی از این تعدیل‌های وزنی که اغلب مورد استفاده قرار می‌گیرد، انتساب اوزانی به واحدهای مشترک در نمونه است که با امید ریاضی تعداد دفعات انتخابشان رابطه دارد.

هارتلی [۹] نظریه‌ی مقدماتی در مورد چارچوب چندگانه را توسعه داد و اولین کسی بود که انتساب وزن به واحدهای مشترک را مد نظر قرار داد. وی در بررسی‌های خود استفاده از تنها دو چارچوب با نمادهای A و B را در نظر گرفت. روشی که او ارایه نمود، بر اساس انتساب وزن θ به واحدهای مشترک متعلق به چارچوب A و وزن $(1-\theta)$ به واحدهای مشترک متعلق به چارچوب B بود. بدین ترتیب می‌توان دو چارچوب با تعدادی واحد مشترک را به دو طبقه‌ی مجزا تبدیل نمود.

به علاوه، وی فرمول‌هایی را برای براورد مجموع، مقدار بهینه‌ی θ و نیز کسرهای نمونه‌گیری بهینه ارایه کرد. بعد از او نیز مقاله‌های دیگری در این زمینه ارایه شد که از آن جمله می‌توان به [۲]، [۵]، [۶]، [۱۱]، [۱۳]، [۱۵]، [۱۶] و [۱۷] اشاره کرد.

اخیراً استفاده از نمونه‌گیری‌های دوچارچوبی، که یکی از چارچوب‌ها را فهرست شماره‌ی تلفن افراد و دیگری را یک چارچوب ناحیه‌ای تشکیل می‌دهد، در برخی از کشورها مرسوم شده است. چرا که استفاده از چارچوبی شامل فهرست تلفن‌ها و انجام نمونه‌گیری به‌صورت متمرکز و تکمیل پرسشنامه از طریق مصاحبه‌ی تلفنی و نظارت هم‌زمان پرسشنامه توسط کارشناس در حین تکمیل پرسشنامه، هم خطای نمونه‌گیری را کاهش داده و هم هزینه‌ی نمونه‌گیری را به نحو چشمگیری تقلیل می‌دهد. از چنین طرح‌هایی می‌توان به طرح ملی بررسی آماری جنایات در آمریکا که در سال ۱۹۸۵ انجام شد، اشاره کرد. کسدی و دیگران [۳] بررسی‌هایی را در مورد استفاده از چارچوب‌های دوگان انجام دادند که در آن‌ها از ترکیب چارچوب ناحیه‌ای که یک چارچوب کامل است و یک چارچوب تلفنی استفاده کردند. همچنین [۱۴] تأثیر نسبت‌های مختلف عدم پاسخگویی در نتایج را آزمون کرد. این بررسی‌ها توسط [۷] تعمیم و توسعه داده شدند. آن‌ها همچنین در مورد مزایای طرح‌های «چندچارچوبی- چند روشی» (Multiframe-Multimode)، با توجه به ساختار اجرایی (اداری) متفاوتی که برای استفاده از دو روش جمع‌آوری داده‌ها به‌طور هم‌زمان، مورد نیاز است، کاوش‌ها و بررسی‌هایی را انجام دادند. در مجموع بررسی‌های تلفنی دارای نرخ پاسخ پایین‌تری هستند. عقیده بر این است که بعضی از انواع پاسخگویان (افراد مسن و کسانی که در گفتار و شنوایی دارای مشکلاتی هستند) در مصاحبه‌های تلفنی نسبت به مصاحبه‌های حضوری در جواب دادن به سؤالات، مشکلات بیش‌تری دارند. این مسئله میزان پرسش‌ها و ابهامات پیرامون کیفیت داده‌های جمع‌آوری شده تحت دو روش را افزایش می‌دهد.

در ایالت‌های متحده، بررسی‌های اولیه برای استفاده از دو چارچوب در طرح‌های نمونه‌گیری کشاورزی، از اواسط دهه‌ی ۱۹۵۰ مورد توجه قرار گرفت و نمونه‌گیری مبتنی بر این روش در بسیاری از نمونه‌گیری‌های کشاورزی در این کشور در دهه‌ی ۱۹۶۰ صورت گرفت. بدین ترتیب می‌توان مشاهده کرد که استفاده‌ی توأم از دو چارچوب در طرح‌های آماری کشاورزی از سال‌ها پیش در ایالت‌های متحده رایج شده است. لیکن به دلیل نیاز به فن‌آوری پیشرفته برای ایجاد یک چارچوب ناحیه‌ای، تاکنون تعداد اندکی از کشورها از این روش بهره جسته‌اند. سازمان خواربار و کشاورزی ملل متحد، FAO، به‌دلیل مزایایی که نمونه‌گیری‌های دوچارچوبی دارند، استفاده از این روش را به سایر کشورها توصیه کرده است. البته استفاده از طرح دوچارچوبی در طرح‌های آماری کشاورزی مستلزم وجود چارچوب ناحیه‌ای و فهرستی است. برای این منظور، تهیه‌ی یک چارچوب ناحیه‌ای کشاورزی مستلزم صرف هزینه‌ی بسیار، نیروی کارشناسی بسیار ماهر و فن‌آوری پیشرفته در زمینه‌ی سیستم‌های ماهواره‌ای و کامپیوتری است. لذا تهیه‌ی یک چارچوب ناحیه‌ای کشاورزی برای بار اول در هر کشور نیاز به صرف هزینه، زمان و دقت بسیار دارد.

۳- دلایل استفاده از چارچوب چندگانه

چارچوب آماری اساس و مبنای یک طرح آمارگیری نمونه‌ای را تشکیل می‌دهد. در صورت وجود مشکلاتی از قبیل عدم پوشش کامل، وجود واحدهای آمارگیری مشابه و پوشش فراتر از حد لازم در یک چارچوب آماری، خطای برآورد پارامترها افزایش پیدا می‌کند. در مرحله‌ی برآورد که لازم است نتایج را از نمونه به جامعه تعمیم دهیم در صورت وجود مشکل یا نقص در چارچوب، برآوردهای ارایه‌شده اریب خواهند شد، بنابراین باید به چارچوب آماری مورد استفاده توجه کافی معطوف شود.

از روش چارچوب چندگانه در اجرای آمارگیری‌هایی استفاده می‌شود که در آن‌ها چارچوب‌هایی با پوشش ناقص، وجود دارند. این چارچوب‌های ناقص تحت عنوان «چارچوب‌های مکمل» نیز شناخته می‌شوند. دستورالعمل چارچوب چندگانه این امکان را ایجاد می‌کند که چارچوب‌های ناقص با یکدیگر تداخل داشته باشند. هیچ تلاشی برای حذف عناصری از یک چارچوب ناقص که در عین حال به چارچوب ناقص دیگر تعلق دارد صورت نمی‌گیرد و نمونه‌ها از همه‌ی چارچوب‌ها انتخاب می‌شوند.

به‌طور کلی موارد استفاده بیش از یک چارچوب بنا به دلایلی است که در زیر به به آن‌ها اشاره می‌کنیم:

۱. گاهی ممکن است چارچوبی که همه‌ی واحدهای جامعه‌ی مورد مطالعه را پوشش می‌دهد در دسترس نباشد، اما امکان دستیابی به پوشش کامل با تلفیقی از دو یا چند چارچوب فراهم شود. در چنین حالتی به‌منظور دسترسی به پوشش مناسب، از دو چارچوب یا بیش‌تر به‌طور هم‌زمان استفاده می‌شود.

۲. گاهی نیز ممکن است یک چارچوب، پوشش کامل را برای جامعه‌ی مورد مطالعه فراهم کند اما چارچوب ناقص دیگری موجود باشد که هزینه‌ی آمارگیری از آن کم‌تر از هزینه‌ی آمارگیری از چارچوب کامل باشد. در این شرایط به‌دلیل پایین‌تر بودن هزینه‌ی آمارگیری از چارچوب ناقص، می‌توان با هزینه‌ای مشخص و ثابت از دو چارچوب استفاده و اندازه‌ی نمونه را بزرگ‌تر کرده و واریانس براورد را کاهش داد.

۳. گاهی ممکن است صفت مورد بررسی مربوط به جامعه‌ی خاصی باشد که نسبت به کل جامعه در اقلیت است. به‌عنوان مثال می‌توان به جامعه‌ی ناشنوایان، جامعه‌ی افرادی که مبتلا به یک نوع بیماری خاص هستند و یا دسترسی به میزان تولید محصولی خاص که سطح زیر کشت نسبتاً کمی را از کل سطح زیر کشت در محدوده‌ی مورد بررسی به خود اختصاص داده است، اشاره کرد. در اصطلاح به چنین جوامعی، جوامع کمیاب گفته می‌شود. بدیهی است که با استفاده از یک چارچوب ناحیه‌ای که پوشش کاملی را ارایه می‌دهد، احتمال دسترسی به این واحدهای خاص بسیار کم بوده، لذا می‌توان با بهره‌گیری از یک چارچوب دیگر که فهرستی ناقص از واحدهای آمارگیری جامعه‌ی مورد مطالعه است، آمارگیری را انجام داد. بسته به هدف، هزینه و دقت براورد می‌توان چارچوب فهرستی را سرشماری یا نمونه‌گیری کرد. برای مثال درصورت بررسی مربوط به جامعه‌ی ناشنوایان، می‌توان فهرستی از افراد ناشنوا را از مدارس ناشنوایان، مؤسسه‌ی ارایه‌ی خدمات به ناشنوایان و ... تهیه کرد. در این شرایط کسری از جامعه‌ی ناشنوایان که از امکانات فوق استفاده نمی‌کنند در چارچوب فهرستی قرار نمی‌گیرند، لذا به‌منظور برطرف کردن این نقصان از چارچوب ناحیه‌ای که همه‌ی افراد جامعه را در برمی‌گیرد نیز استفاده می‌شود.

۴. دلیل دیگر لزوم استفاده از دو یا بیش از دو چارچوب، حتی در صورت دسترسی به یک چارچوب کامل، تغییرپذیری بسیار زیاد متغیر یا متغیرهای مورد مطالعه در

جامعه است که منجر به چولگی شدید می‌شود. مثلاً در طرح‌های کشاورزی ممکن است تعداد کمی از بهره‌برداران کشاورزی، مالک بهره‌برداری‌های بسیار بزرگ و تعداد بسیار زیادی از بهره‌برداران، مالک بهره‌برداری‌های کوچک باشند. لذا در صورت شناسایی و فهرست کردن بهره‌برداری‌های بزرگ، با تلفیق دو چارچوب ناحیه‌ای و فهرستی می‌توان از مزایای هر دو چارچوب استفاده کرده و کارایی طرح را بالا برد.

۵. گاه ممکن است یک چارچوب فهرستی کامل در دسترس باشد اما عملاً با گذشت زمانی نسبتاً طولانی به‌دلیل بروز تغییرات فراوان در آن، منبعی برای بروز خطاهای غیرنمونه‌گیری شود. از طرفی چون یک فهرست ناحیه‌ای، کم‌تر در معرض تغییرات است، با ترکیب این دو چارچوب می‌توان از مزایای هر دو چارچوب بهره برد.

در سال‌های اخیر به‌دلیل کاربرد روزافزون مصاحبه‌های تلفنی، مطلوبیت طرح‌های چارچوب چندگانه افزایش چشمگیری پیدا کرده است. همان‌طور که اشاره شد مصاحبه‌های تلفنی عموماً ارزان‌تر از مصاحبه‌های غیرتلفنی بوده و همچنین تمرکز پرسش‌گران در یک مکان، امکان نظارت بیش‌تری را فراهم می‌نماید. بنابراین با وجود این‌که چارچوب‌های تلفنی اغلب، کل جامعه‌ی آماری را پوشش نمی‌دهند، دارای مزایای زیادی هستند. به‌عنوان مثال، طرح ملی آمارگیری تندرستی که هر ساله توسط مرکز ملی آمارهای تندرستی ایالت‌های متحده از طریق تلفن اجرا می‌شود و در آن جامعه‌ی آماری شامل همه‌ی خانوارهای آمریکایی است. چارچوب مورد استفاده در این طرح چارچوب شماره‌ی تلفن‌های ثابت است. البته افرادی که در خانوارهای فاقد تلفن ثابت زندگی می‌کنند از شمول طرح خارج می‌شوند. در این طرح، پوشش کامل را می‌توان با استفاده از یک طرح با چارچوب دوگان، متشکل از یک چارچوب تلفنی و یک چارچوب ناحیه‌ای تأمین نمود.

۴- آمارگیری دو چارچوبی

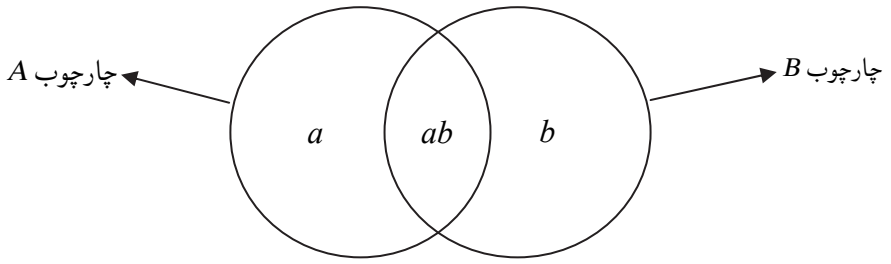
دو چارچوب A و B را در نظر گرفته و فرض می‌کنیم که از هر یک از آن‌ها نمونه‌ای استخراج شده است. می‌توان برای هر یک از چارچوب‌ها، طرح نمونه‌گیری متفاوتی استفاده نمود اما لازم است فرض‌های زیر برقرار باشند:

۱- کامل بودن، یعنی هر واحد در جامعه‌ی مورد مطالعه حداقل باید متعلق به یکی از چارچوب‌ها باشد.

۲- شناسایی‌پذیری، یعنی بتوان تشخیص داد که هر کدام از واحدهای انتخاب‌شده در نمونه به کدام یک از چارچوب‌ها تعلق دارند.

هنگامی که یکی از چارچوب‌ها، چارچوب ناحیه‌ای باشد، فرض کامل بودن معمولاً برقرار است. جوامعی مانند مزارع (در طرح‌های کشاورزی) به‌طور طبیعی با قطعات زمین که واحدهای چارچوب ناحیه‌ای را تشکیل می‌دهند، مرتبط هستند. با وجود ساده بودن فرض نظری شناسایی‌پذیری، اجرای یک طرح آمارگیری مبتنی بر چارچوب چندگانه دارای مشکلات زیادی می‌باشد. تعیین این‌که آیا یک واحد نمونه‌گیری‌شده در یک چارچوب به چارچوب دیگری نیز تعلق دارد، کار ساده‌ای نیست.

هارتلی [۹] نظریه‌ی بنیادی نمونه‌گیری مبتنی بر چارچوب چندگانه را توسعه داد و [۴] این نظریه را گسترش داد. هارتلی جامعه را با استفاده از چارچوب‌های نمونه‌گیری و تداخل بین آن‌ها به حوزه‌های به‌طور توأم ناسازگار تقسیم کرد. برای مثال، دو چارچوب نمونه‌گیری A و B، سه حوزه‌ی ممکن را به‌صورت زیر تشکیل می‌دهند:



شکل ۱- مثالی از پوشش جامعه توسط دو چارچوب

در شکل ۱:

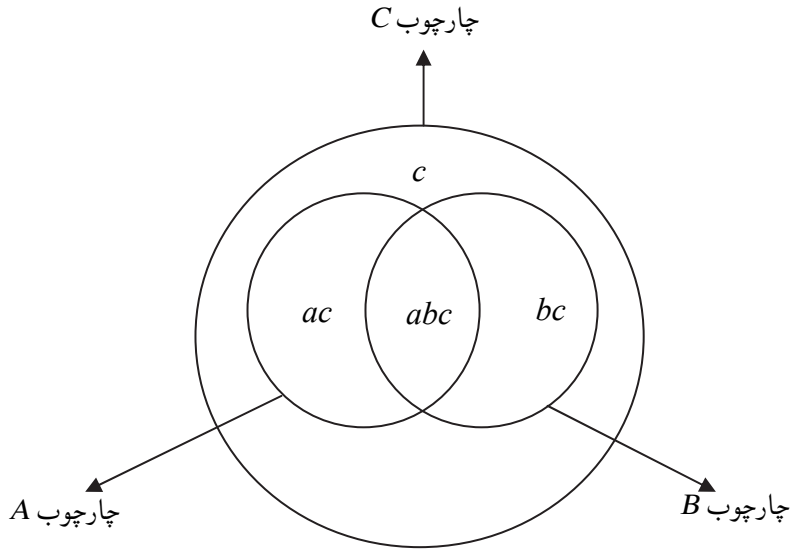
۱- حوزه‌ی a شامل واحدهایی است که فقط به چارچوب A تعلق دارند،

۲- حوزه‌ی b شامل واحدهایی است که فقط به چارچوب B تعلق دارند و

۳- حوزه‌ی ab شامل واحدهایی است که به هر دو چارچوب تعلق دارند.

فرض کنید می‌خواهیم یک بررسی در مورد آماردانان ایالت‌های متحده انجام دهیم. یک روش، استفاده از فهرست اعضای انجمن آمار آمریکا (ASA) به‌عنوان چارچوب نمونه‌گیری و گرفتن یک نمونه‌ی احتمالاتی از فهرست اعضا می‌باشد. اکثر افرادی که در

این بررسی آمارگیری می‌شوند، آماردان هستند اما خیلی از آماردانان در ایالت‌های متحده عضو ASA نیستند پس اگر تنها از این فهرست به‌عنوان چارچوب استفاده شود با کم پوششی جامعه‌ی آماردانان مواجه خواهیم شد. برای بهبود پوشش چارچوب باید یک نمونه‌ی تکمیلی از چارچوب نمونه‌گیری دیگری نظیر فهرست اعضای انجمن آمار ریاضی (IMS) گرفته شود. بنابراین ما دو نمونه گرفته‌ایم؛ یک نمونه‌ی احتمالاتی از چارچوب A (فهرست ASA) و یک نمونه‌ی احتمالاتی از چارچوب B (فهرست اعضای IMS) که مستقل از نمونه‌ی چارچوب A است. دو چارچوب متداخل می‌باشند، چون خیلی از آماردانان عضو هر دو انجمن هستند. همان‌طور که در شکل ۱ نشان داده شده است ۳ حوزه‌ی a، b و ab در جامعه وجود دارند. چارچوب A و B هر دو با هم پوشش بهتری برای جامعه‌ی آماردانان ایالت‌های متحده ایجاد می‌کنند اما هنوز شامل همه‌ی آماردانان نیستند. در این حالت، یک چارچوب سومی (C) می‌تواند استفاده شود که شامل جمعیت افراد بزرگسال است. ساختار جامعه با ۳ چارچوب در شکل ۲ نشان داده شده است.



شکل ۲- پوشش جامعه‌ی آماردانان آمریکا توسط سه چارچوب A، B و C

چارچوب C کل جامعه را پوشش می‌دهد و دو چارچوب متداخل A و B را نیز در بر دارد. در این طرح، حوزه‌ی c بخشی از جامعه در چارچوب کامل C است که در دو فهرست A و B نیستند.

چون چارچوب C کل افراد جامعه را در بر دارد امکان طرح این سئوال وجود دارد که با وجود کامل بودن پوشش چارچوب C، چرا تنها از آن به‌عنوان چارچوب نمونه‌گیری استفاده نمی‌شود. لازم به ذکر است که چارچوب C شامل کل جامعه‌ی بزرگسالان است ولی نمونه‌گیری از آن پرهزینه است. چون اکثر بزرگسالان، آماردان نیستند و بسیاری از نمونه‌های حاصل از چارچوب C آماردان نیستند. نمونه‌گیری از چارچوب‌های A و B خیلی ارزان‌تر است اما آن‌ها کل جامعه‌ی مورد مطالعه را پوشش نمی‌دهند. با ترکیب نمونه از هر سه چارچوب می‌توان از کم هزینه بودن دو چارچوب A و B و کامل بودن چارچوب C سود برد. دلایل مختلفی برای استفاده از چارچوب‌های چندگانه در جمع‌آوری داده‌ها وجود دارد. چارچوب‌های چندگانه به‌خصوص در مواردی همچون نمونه‌گیری از جوامع کمیاب به‌طور چشمگیری کارایی جمع‌آوری داده‌ها را بهبود می‌بخشد. یک جامعه‌ی کمیاب یک زیرگروه از جامعه‌ی مورد مطالعه است که تنها بخش کوچکی از جامعه (معمولاً ۱۰ درصد یا کم‌تر) را در بر دارد. در مثال بالا، آماردانان بخش کوچکی از جامعه‌ی بزرگسالان ایالت‌های متحده هستند. اگر نمونه‌های حاصل از فهرست‌های اعضا را به‌عنوان مکمل نمونه از کل جامعه در نظر بگیریم، تعداد آماردانان در مجموعه‌ی داده‌ها افزایش می‌یابد. فرض کنید می‌خواهیم افراد با یک بیماری خاص را مورد بررسی قرار دهیم. در این حالت می‌توان با نمونه‌گیری از چارچوب‌های اضافی که شامل نسبت بالایی از این افراد هستند نمونه‌ی بزرگ‌تری بدست آورد. البته برای نمونه‌گیری از جوامع کمیاب روش‌های دیگری نیز وجود دارد که از این بحث خارج است.

در صورت استفاده از آمارگیری‌های چندچارچوبی می‌توان از روش‌های مختلفی برای جمع‌آوری داده‌ها استفاده کرد. به‌عنوان مثال در شکل ۱، چارچوب A ممکن است شامل شماره‌های تلفن ثابت و چارچوب B شامل شماره‌های موبایل باشد.

هارتلی [۹] هنگام معرفی نظریه‌ی چارچوب چندگانه، دلیل اصلی بررسی آمارگیری چندچارچوبی را کاهش هزینه‌های جمع‌آوری داده‌ها معرفی کرد. «در یک آمارگیری ممکن است یک چارچوب با پوشش کامل داشته باشیم اما آمارگیری از آن پرهزینه است در

حالی که چارچوب‌های دیگری موجود هستند که نمونه‌گیری از آن‌ها ارزان‌تر است و بخشی از جامعه‌ی هدف را پوشش می‌دهند».

حالت دیگری که اغلب در نمونه‌گیری چندچارچوبی مورد استفاده قرار می‌گیرد و توسط هارتلی بررسی شده در شکل ۳ نشان داده شده است. در این حالت، چارچوب A یک چارچوب ناحیه‌ای است که پوشش کاملی از جامعه دارد و چارچوب B اغلب یک چارچوب فهرستی است.

چارچوب ناحیه‌ای (مجموعه‌ی نواحی جغرافیایی)، چارچوبی شامل واحدهای ناحیه‌ای است و هر عنصر جامعه متعلق به تنها یکی از ناحیه‌ها است و آن عنصر بعد از ارتباط با آن واحد ناحیه‌ای می‌تواند مشخص شود. به‌علت کامل بودن چارچوب‌های ناحیه‌ای، می‌توان از آن‌ها در موارد بسیاری استفاده کرد. برای مثال زمانی که:

۱- یک چارچوب کامل دیگر در دسترس نباشد.

۲- واحدهای نمونه‌گیری در چارچوب به‌سرعت تغییر کنند.

۳- چارچوب موجود، کهنه شده باشد.

۴- چارچوب موجود، از یک سرشماری با پوشش ضعیف بدست آمده باشد.

۵- یک چارچوب چندمنظوره برای براورد صفت‌های مختلف (کشاورزی، زیست محیطی و غیره) مورد نیاز باشد.

در طرح‌هایی که مبتنی بر چارچوب ناحیه‌ای است می‌توان بدون انجام مصاحبه، در ارتباط با خصوصیات قابل مشاهده بر روی زمین، براوردهایی به‌دست آورد. به‌همین علت، خطاهای غیرنمونه‌گیری از جمله خطای بی‌پاسخی در مصاحبه‌ها کاهش می‌یابد. البته طرح‌های نمونه‌گیری مبتنی بر چارچوب ناحیه‌ای دارای معایبی به شرح ذیل هستند:

۱- پرهزینه بودن انجام و اجرای نمونه‌گیری؛

۲- ضرورت استفاده از تجهیزات کارتوگرافیکی؛

۳- حساسیت به داده‌های دورافتاده؛

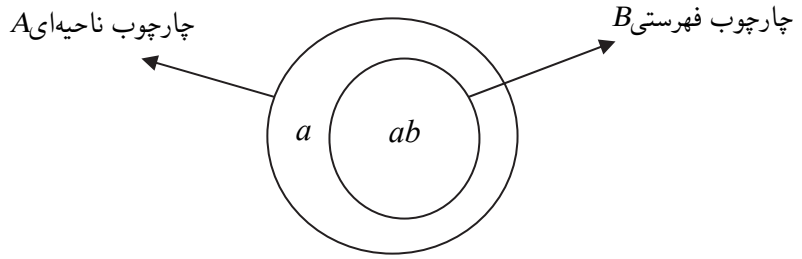
۴- بی‌ثباتی و نااستواری براوردها.

متداول‌ترین راه برای اجتناب از بی‌ثباتی براوردها و بهبود دقت آن‌ها استفاده‌ی هم‌زمان از چارچوب‌های ناحیه‌ای و فهرستی است. مسأله‌ی اصلی در این روش تعیین واحدهای نمونه‌گیری ناحیه‌ای است که در فهرست نیز موجود می‌باشند. وقتی نمونه‌های انتخاب‌شده از چارچوب ناحیه‌ای را نتوان در فهرست تعیین کرد، براوردگر مجموع جامعه

دچار اریبی بالا می‌شود. نقش و وظیفه‌ی چارچوب ناحیه‌ای در چارچوب‌های چندگانه حل مشکلات مربوط به کامل نبودن فهرست و همچنین براورد آن می‌باشد.

به‌عنوان مثالی از این نوع چارچوب‌های چندگانه در آمارگیری‌های کشاورزی، فرض کنید چارچوب B فهرست زمین‌های کشاورزی بزرگ در یک آبادی است. این فهرست ممکن است کهنه شده باشد و یا شامل زمین‌هایی که اخیراً به‌وجود آمده‌اند، نباشد. علاوه بر آن، این فهرست زمین‌های کشاورزی کوچک را هم در بر ندارد. هر چند که روش‌های کارایی برای جمع‌آوری داده می‌تواند در چارچوب B استفاده شود. اگر فقط از چارچوب B استفاده کنیم پوشش جامعه ناقص است، اما اگر از یک چارچوب ناحیه‌ای مثل چارچوب A استفاده کنیم پوشش کاملی از زمین‌های کشاورزی به وجود خواهد آمد. در یک چارچوب ناحیه‌ای، کشور به نواحی جغرافیایی تقسیم می‌شود و بخش‌هایی در این نواحی نمونه‌گیری می‌شوند. آمارگیران به هر بخش انتخاب‌شده رفته و از هر فعالیت کشاورزی درون این محدوده، آمارگیری می‌کنند. چارچوب ناحیه‌ای شامل همه‌ی زمین‌های کشاورزی کشور است اما جمع‌آوری داده از آن پرهزینه‌تر از جمع‌آوری داده از یک چارچوب فهرستی است.

آمارگیری از هزینه‌های مشتری در آمریکا (The U.S. survey of consumer finances) مثالی از یک آمارگیری دو چارچوبی مطابق با شکل ۳ است که در آن یک نمونه (نمونه از چارچوب A) از جامعه‌ی آمریکا با استفاده از چارچوب ناحیه‌ای با نمونه‌گیری طبقه‌بندی چندمرحله‌ای انتخاب می‌شود. ولی این نمونه ممکن است شامل همه‌ی افراد مورد مطالعه نباشد بنابراین یک نمونه دومی از چارچوب فهرستی (چارچوب B) گرفته می‌شود (چارچوب B شامل رکوردهای مالیاتی است).



شکل ۳- مثالی از دو چارچوب ناحیه‌ای و فهرستی

۵- براوردگرهای مجموع در طرح‌های چارچوب دوگانه

در این بخش، روش‌های براورد برای طرح‌های چارچوب دوگان، هنگامی که نمونه‌های حاصل از دو چارچوب به‌طور مستقل استخراج شوند، مورد بررسی قرار می‌گیرند. فرض کنید که مجموعه‌ی U دارای N واحد باشد. مجموع جامعه برای ویژگی y، به‌صورت $Y=\sum_{i=۰}^{N} y_i$ است. اگر در یک طرح نمونه‌گیری احتمالاتی برای یک نمونه‌ی S که از یک چارچوب استفاده می‌شود، $\pi_i = p\ (i \in S)$ احتمال شمول واحد i در نمونه‌ی S باشد، براوردگر هورویتز-تامپسون مجموع جامعه برابر است با:

$$\hat{Y}_{HT} = \sum_{i \in S} w_i\, y_i$$

که در آن $w_i = \frac{۱}{\pi_i}$ ، وزن نمونه‌گیری است.

اکنون حالتی را در نظر می‌گیریم که دو نمونه‌ی مستقل S_A و S_B از دو چارچوب متداخل A و B همانند شکل ۱ استخراج شود. چارچوب A شامل N_A واحد و چارچوب B شامل N_B واحد می‌باشند. واحدهای جامعه در حوزه‌ی متداخل ab می‌تواند از واحدهای یک چارچوب یا هر دو چارچوب انتخاب شوند.

فرض می‌کنیم

$$\delta_i(a)=\begin{cases} ۱ & \text{اگر واحد } i \text{ متعلق به حوزه‌ی } a \text{ باشد} \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

$$\delta_i(b)=\begin{cases} ۱ & \text{اگر واحد } i \text{ متعلق به حوزه‌ی } b \text{ باشد} \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

و

$$\delta_i(ab)=\begin{cases} ۱ & \text{اگر واحد } i \text{ متعلق به حوزه‌ی } ab \text{ باشد} \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

در این صورت می‌توان مجموع جامعه را به‌عنوان مجموع سه حوزه به‌صورت زیر نوشت:

$$Y=Y_a+Y_b+Y_{ab}=\sum_{i=1}^{N}\delta_i(a)\,y_i+\sum_{i=1}^{N}\delta_i(b)\,y_i+\sum_{i=1}^{N}\delta_i(ab)\,y_i$$

براوردگر مجموع جامعه $\hat{Y}$، می‌تواند با مجموع براوردگرهای $\hat{Y}_a$، $\hat{Y}_b$ و $\hat{Y}_{ab}$ بدست آید. فرض کنید احتمال شمول واحد iام در نمونه‌ی S_A برابر $\pi_i^A=p\{i\in S_A\}$ بوده و S_A شامل n_A واحد نمونه باشد. به‌طور متناظر برای چارچوب B احتمال شمول واحد jام در نمونه‌ی S_B برابر $\pi_j^B=p\{j\in S_B\}$ و شامل n_B واحد نمونه می‌باشد. فرض می‌کنیم w_i^A وزن نمونه‌گیری از چارچوب A و w_i^B وزن نمونه‌گیری از چارچوب B باشند. این وزن‌ها هم می‌توانند عکس احتمال‌های π_i^A و π_i^B و هم می توانند وزن‌های هایک [۸] با $w_i^A=N_A\,[\pi_i^A\sum_{j\in S_A}(1/\pi_j^A)]^{-1}$ و $w_i^B=N_B\,[\pi_i^B\sum_{j\in S_B}(1/\pi_j^B)]^{-1}$ باشند. با استفاده از این وزن‌ها، براوردگرهای حوزه‌های a و ab از چارچوب A را به‌صورت زیر تعریف می‌کنیم:

$$\hat{Y}_a^A=\sum_{i\in S_A}w_i^A\delta_i(a)y_i$$
$$\hat{Y}_{ab}^A=\sum_{i\in S_A}w_i^A\delta_i(ab)y_i$$

به‌طور متناظر، براوردگرهای حوزه‌های b و ab از نمونه‌ی چارچوب B به‌صورت زیر هستند:

$$\hat{Y}_b^B=\sum_{i\in S_B}w_i^B\delta_i(b)y_i$$
$$\hat{Y}_{ab}^B=\sum_{i\in S_B}w_i^B\delta_i(ab)y_i$$

تحت نتایج نظریه‌ی نمونه‌گیری استاندارد، هر کدام از براوردگرهای مجموع در حوزه‌های a، b و ab برای پارامتر میانگین جامعه تقریباً ناریب هستند. اندازه‌ی جامعه در هر یک از حوزه‌های a، b و ab با مقداردهی $y_i=1$ برای همه‌ی واحدها براورد می‌شود و این براوردگرها را با $\widehat{N}_a^A$، $\widehat{N}_{ab}^A$، $\widehat{N}_b^B$ و $\widehat{N}_{ab}^B$ نشان می‌دهیم.

در یک طرح چارچوب دوگان که واحدهای مشترک از یک چارچوب غربال می‌شوند، حوزه‌ی ab تهی است و می‌توان چارچوب‌های A و B را به‌عنوان دو طبقه‌ی مجزا فرض کرد. بنابراین براورد مجموع جامعه در یک طرح چارچوب دوگان غربالی به‌صورت زیر است:

$$\hat{Y}=\hat{Y}_a^A+\hat{Y}_b^B$$

در ادامه، روش‌هایی که برای براورد مجموع جامعه، Y ، در حالت چارچوب دوگانه بیان شده، به‌صورت مختصر تشریح می‌شود.

۱-۵- براوردگر هارتلی

هارتلی اولین کسی بود که روش‌شناسی سیستماتیکی را برای تحلیل طرح‌های آمارگیری چندچارچوبی پایه‌گذاری کرد. در آمارگیری‌های چارچوب دوگان، $\hat{Y}_{ab}^A$ و $\hat{Y}_{ab}^B$، هر دو Y_{ab} را براورد می‌کنند. بنابراین به‌منظور اجتناب از چندبارگی (Multiplicity) براورد، هارتلی در [۹] یک براوردگر موزون به‌صورت زیر پیشنهاد داد:

$$\hat{Y}(\theta)=\hat{Y}_a^A+\theta\hat{Y}_{ab}^A+(1-\theta)\hat{Y}_{ab}^B+\hat{Y}_b^B, \quad (1)$$

که در آن θ پارامتر آمیختگی است؛ $0\leq\theta\leq 1$.
این براوردگر، وزن هر واحد نمونه‌گیری‌شده در حوزه‌ی مشترک ab را به‌منظور جبران چندبارگی کاهش می‌دهد. وزن‌های جدید را به‌صورت زیر تعریف می‌کنیم:

$$\tilde{w}_i^A=\delta_i(a)\,w_i^A+\theta\,\delta_i(ab)\,w_i^A$$
$$\tilde{w}_i^B=\delta_i(b)\,w_i^B+(1-\theta)\,\delta_i(ab)\,w_i^B$$

بنابراین

$$\hat{Y}(\theta)=\sum_{i\epsilon S_A}\tilde{w}_i^A y_i+\sum_{i\epsilon S_B}\tilde{w}_i^B y_i\ .$$

هر کدام از براوردگرهای مجموع در هر حوزه تقریباً ناریب است، بنابراین $\hat{Y}(\theta)$ یک براوردگر تقریباً ناریب از مجموع جامعه‌ی Y است. چون چارچوب‌های A و B به‌طور مستقل نمونه‌گیری می‌شوند و θ ثابت است، واریانس براوردگر برابر است با:

$$\mathrm{Var}\,[\hat{Y}(\theta)]=\mathrm{Var}\,[\hat{Y}_a^A+\theta\hat{Y}_{ab}^A]+\mathrm{Var}\,[(1-\theta)\hat{Y}_{ab}^B+\hat{Y}_b^B]. \quad (2)$$

براوردگر واریانس رابطه‌ی (۱) برای محاسبه آسان است اما ممکن است نسبت به براوردگرهای دیگر کارایی کم‌تری داشته باشد. θ در رابطه‌ی (۱) به گونه‌ای انتخاب می‌شود که واریانس $\hat{Y}(\theta)$ را مینیمم کند ([۹] و [۱۰]). به‌دلیل این‌که نمونه‌گیری از چارچوب‌ها

به‌طور مستقل انجام می‌شود، واریانس $\hat{Y}(\theta)$ برابر است با مجموع واریانس دو چارچوب که در رابطه‌ی (۲) نشان داده شده است. مقدار بهینه‌ی θ که واریانس فوق را مینیمم کند برابر است با:

$$\theta_{opt}=\frac{V(\hat{Y}_{ab}^{B})+\mathrm{Cov}(\hat{Y}_{b}^{B},\hat{Y}_{ab}^{B})-\mathrm{Cov}(\hat{Y}_{a}^{A},\hat{Y}_{ab}^{A})}{V(\hat{Y}_{ab}^{A})+V(\hat{Y}_{ab}^{B})} \qquad (۳)$$

با θ_{opt}، واریانس مینیمم‌شده به‌صورت زیر حاصل می‌شود:

$$\mathrm{Var}\,[\hat{Y}(\theta)]=\mathrm{Var}\,(\hat{Y}_{a}^{A}+\hat{Y}_{b}^{B}+\hat{Y}_{ab}^{B})-\theta_{opt}^{۲}[\mathrm{Var}\,(\hat{Y}_{ab}^{A})+\mathrm{Var}(\hat{Y}_{ab}^{B})]$$

با توجه به رابطه‌ی (۳)، بزرگ‌تر بودن $\mathrm{Var}(\hat{Y}_{ab}^{B})$ نسبت به $\mathrm{Var}(\hat{Y}_{ab}^{A})$ منجر به بزرگ‌تر شدن θ_{opt} می‌شود. نکته آن‌که اگر مقدار قدر مطلق $\mathrm{Cov}(\hat{Y}_{b}^{B},\hat{Y}_{ab}^{B})$ یا $\mathrm{Cov}(\hat{Y}_{a}^{A},\hat{Y}_{ab}^{A})$ بزرگ باشد در این صورت، این امکان وجود دارد که θ_{opt} کوچک‌تر از صفر یا بزرگ‌تر از یک شود. هنگامی که دو چارچوب A و B یکسان باشند (حوزه‌های aو b تهی هستند)، θ_{opt} بین صفر و یک است.

در عمل، واریانس‌ها و کوواریانس‌ها در رابطه‌ی (۳) نامعلوم بوده و باید براورد شوند. فرض کنیم $\hat{\theta}_{opt}$ براوردگر θ_{opt} باشد. با جایگذاری آن در وزن‌های تعدیل‌شده داریم:

$$\tilde{w}_{i,H}^{A}=\delta_i(a)\,w_i^{A}+\hat{\theta}_{opt}\delta_i(ab)\,w_i^{A}$$
$$\tilde{w}_{i,H}^{B}=\delta_i(b)\,w_i^{B}+(۱-\hat{\theta}_{opt})\delta_i(ab)\,w_i^{B}$$

۲-۵- براوردگر فولر-بورمیستر (Fuller-Burmeister)

فولر و بورمیستر براوردگر هارتلی را با افزودن اطلاعات بیش‌تر راجع به براورد N_{ab} تعدیل و پیشنهاد کردند[۶]. این براوردگر به‌صورت زیر است:

$$\hat{Y}_{FB}(\beta)=\hat{Y}_{a}^{A}+\hat{Y}_{b}^{B}+\beta_{۱}\hat{Y}_{ab}^{A}+(۱-\beta_{۱})\hat{Y}_{ab}^{B}+\beta_{۲}\,(\hat{N}_{ab}^{A}-\hat{N}_{ab}^{B}). \qquad (۴)$$

همانند براوردگر هارتلی، پارامترهای $\beta_{۱}$ و $\beta_{۲}$ به گونه‌ای انتخاب می‌شوند که واریانس $\hat{Y}_{FB}(\beta)$ را مینیمم کنند. مقادیر بهینه‌ی $\beta_{۱}$ و $\beta_{۲}$ عبارت‌اند از:

$$\begin{bmatrix}\beta_{۱}\\ \beta_{۲}\end{bmatrix}=\begin{bmatrix}\mathrm{Var}(\hat{Y}_{ab}^{A}-\hat{Y}_{ab}^{B}) & \mathrm{Cov}(\hat{Y}_{ab}^{A}-\hat{Y}_{ab}^{B},\hat{N}_{ab}^{A}-\hat{N}_{ab}^{B})\\ \mathrm{Cov}(\hat{Y}_{ab}^{A}-\hat{Y}_{ab}^{B},\hat{N}_{ab}^{A}-\hat{N}_{ab}^{B}) & \mathrm{Var}(\hat{N}_{ab}^{A}-\hat{N}_{ab}^{B})\end{bmatrix}^{-۱}$$

$$\times \begin{bmatrix} \mathrm{Cov}(\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^B, \hat{Y}_{ab}^A - \hat{Y}_{ab}^B) \\ \mathrm{Cov}(\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B) \end{bmatrix}$$

واریانس‌ها و کوواریانس‌ها مانند براوردگر هارتلی بهینه، باید از روی داده‌ها براورد شوند. این نتایج در یک مجموعه‌ی متفاوت از وزن‌ها برای هر متغیر پاسخ استفاده می‌شود.

۳-۵- براوردگر تک‌چارچوبی (Single-frame Estimator)

بانکیر [۲] و [۱۱] روش‌های تک‌چارچوبی را پیشنهاد دادند که مشاهدات را در یک مجموعه داده‌ها ترکیب می‌کند و سپس وزن‌ها در حوزه‌ی مشترک تعدیل می‌شوند. یک واحد مشاهده شده‌ی i در حوزه‌ی ab می‌تواند هم در S_A و هم در S_B انتخاب شود، بنابراین تعداد دفعات مورد انتظار انتخاب واحد i در حوزه‌ی ab برابر $\pi_i^A + \pi_i^B$ است. بنابراین براوردگر تک‌چارچوبی [۱۱] از $1/(\pi_i^A + \pi_i^B)$ به‌عنوان وزن واحدهای مشاهده‌شده در حوزه‌ی ab استفاده می‌کند. بنابراین، اگر $w_i^A = 1/\pi_i^A$ و $w_i^B = 1/\pi_i^B$، در این صورت وزن تعدیل‌شده برای واحدهای نمونه‌گیری‌شده در چارچوب A به‌صورت زیر است:

$$\widetilde{w}_{i,S}^A = \begin{cases} w_i^A & i \in a \\ (1/w_i^A + 1/w_i^B)^{-1} & i \in ab \end{cases}$$

وزن‌های تعدیل شده برای واحدهای نمونه‌گیری شده در چارچوب B به طور مشابه به صورت زیر تعریف می‌شود:

$$\widetilde{w}_{i,S}^B = \begin{cases} w_i^B & i \in b \\ (1/w_i^A + 1/w_i^B)^{-1} & i \in ab \end{cases}$$

بنابراین براوردگر تک‌چارچوبی به‌صورت زیر است:

$$\hat{Y}_S = \sum_{i \in S_A} \widetilde{w}_{i,S}^A y_i + \sum_{i \in S_B} \widetilde{w}_{i,S}^B y_i \qquad (۵)$$

هنگامی که هر نمونه خودوزن است، همه‌ی وزن‌ها برای واحدهای نمونه‌گیری‌شده در S_A برابر w^A و همه‌ی وزن‌ها برای واحدهای نمونه‌گیری‌شده در S_B برابر w^B است، بنابراین

$\hat{Y}_S$ یک حالت خاص از براوردگر رابطه‌ی (۱) بوده و می‌توان آن را به‌صورت زیر بازنویسی کرد:

$$\hat{Y}_S=\hat{Y}_a^A+\theta_s\hat{Y}_{ab}^A+(1-\theta_s)\hat{Y}_{ab}^B+\hat{Y}_b^B$$

به‌طوری که

$$\theta_s=\frac{\pi_i^A}{(\pi_i^A+\pi_i^B)}=[w^A(\frac{1}{w^A}+\frac{1}{w^B})]^{-1}$$

برخلاف براوردگرهای هارتلی و فولر-بورمیستر، براوردگرهای تک‌چارچوبی به براورد کوواریانس‌های متغیر پاسخ بستگی ندارد. این براوردگر از مجموعه‌ی وزن‌های یکسانی برای هر پاسخ مورد نظر، استفاده می‌کند. محاسبه‌ی وزن‌ها در حوزه‌ی ab، نیازمند اطلاعات راجع به احتمال‌های شمول در دو چارچوب (نه فقط چارچوبی که یک واحد از آن انتخاب شده است) می‌باشد. اگر یکی از نمونه‌ها (برای مثال نمونه‌ی چارچوب A) خودوزن نباشد در این صورت π_i^A ممکن است برای یک واحد انتخاب‌شده در S_B نامعلوم باشد.

۴-۵- براوردگر ماکسیمم درستنمایی نما

هنگامی که N_A و N_B معلوم باشند، [۱۶] یک براوردگر دیگر که تعدیل براوردگر فولر-بورمیستر است را برای نمونه‌های تصادفی ساده به‌منظور دستیابی به براوردگر ماکسیمم درستنمایی نما (PML) که می‌تواند در طرح‌های پیچیده استفاده شود، پیشنهاد دادند.
براوردگر PML از مجموعه وزن‌های یکسانی برای همه‌ی متغیرهای پاسخ استفاده می‌کند و به‌صورت زیر است:

$$\hat{Y}_{PML}=\frac{N_A-\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a^A}\hat{Y}_a^A+\frac{N_B-\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b^B}\hat{Y}_b^B+\frac{\hat{N}_{ab}^{PML}(\theta)}{\theta\hat{N}_{ab}^A+(1-\theta)\hat{N}_{ab}^B}\left[\theta\hat{Y}_{ab}^A+\left(1-\theta\right)\hat{Y}_{ab}^B\right], \quad (۶)$$

که در آن $\hat{N}_{ab}^{PML}(\theta)$ کوچک‌ترین ریشه‌ی معادله‌ی درجه‌ی دو زیر است:

$$\left[\frac{\theta}{N_B}+\frac{1-\theta}{N_A}\right]x^2-\left[1+\frac{\theta\hat{N}_{ab}^A}{N_B}+\frac{(1-\theta)\hat{N}_{ab}^B}{N_A}\right]x+\theta\hat{N}_{ab}^A+(1-\theta)\hat{N}_{ab}^B=0$$

فولر و بورمیستر استدلال کردند که وقتی یک نمونه‌ی تصادفی از هر یک از چارچوب‌ها گرفته می‌شود، براوردگر رابطه‌ی (۶) می‌تواند با استفاده از اصول ماکسیمم درستنمایی حاصل شود و بنابراین به طور مجانبی کارا است [۶] و [۱۵]. براوردگر PML، براوردگرهای سازگار از N_{ab}، Y_{ab}، Y_a و Y_b برای کمیت‌های متناظر در براوردگر ماکسیمم درستنمایی که با استفاده از نمونه‌گیری تصادفی ساده استخراج می‌شود، جایگذاری می‌کند. براوردگر PML برخلاف براوردگر رابطه‌ی (۴)، تحت طرح‌های نمونه‌گیری پیچیده سازگار است. این براوردگر به واریانس‌های $\hat{Y}_{ab}^{A}$ و $\hat{Y}_{ab}^{B}$ بستگی ندارد.

اسکینر و رائو [۱۶] استفاده از مقدار $\theta = \theta_P$ که واریانس مجانبی $\hat{N}_{ab}^{PML}(\theta)$ را مینیمم می‌کند، پیشنهاد دادند:

$$\theta_P = \frac{N_a N_B \mathrm{Var}(\hat{N}_{ab}^{B})}{N_a N_B \mathrm{Var}(\hat{N}_{ab}^{B}) + N_b N_A \mathrm{Var}(\hat{N}_{ab}^{A})} \qquad (۷)$$

براوردگر رابطه‌ی (۶) براوردگرهای مجموع سه حوزه یعنی Y_a، Y_b و Y_{ab} را با براوردگر بهینه‌ی N_{ab} تعدیل می‌کند. نکته آن‌که محاسبه‌ی θ_P در رابطه‌ی (۷) نیازمند این است که هر سه حوزه‌ی a، b و ab ناتهی بوده و واریانس‌های $\hat{N}_{ab}^{A}$ و $\hat{N}_{ab}^{B}$ مثبت باشند. در عمل، N_a، N_B، $\mathrm{Var}(\hat{N}_{ab}^{A})$ و $\mathrm{Var}(\hat{N}_{ab}^{B})$ از روی داده‌ها براورد می‌شوند که در نتیجه‌ی آن یک براوردگر $\hat{\theta}_P$ از θ_P در رابطه‌ی (۶) جایگزین می‌شود. در این صورت وزن‌های تعدیل یافته به‌صورت زیر هستند:

$$\tilde{w}_{i,S}^{A} = \begin{cases} \frac{N_A - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_a^{A}} w_i^{A} & i \in a \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^{A} + (۱ - \hat{\theta}_P)\hat{N}_{ab}^{B}} \hat{\theta}_P w_i^{A} & i \in ab \end{cases}$$

و

$$\tilde{w}_{i,S}^{B} = \begin{cases} \frac{N_B - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_a^{B}} w_i^{B} & i \in b \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^{A} + (۱ - \hat{\theta}_P)\hat{N}_{ab}^{B}} (۱ - \hat{\theta}_P) w_i^{B} & i \in ab \end{cases}$$

$\hat{\theta}_P$ تنها به براورد واریانس اندازه‌ی حوزه‌ی متداخل بستگی دارد. بنابراین در براوردگر PML از مجموعه‌ی یکسان وزن‌ها برای هر متغیر پاسخ استفاده می‌شود. اسکینر و رائو

در [۱۶] دریافتند که براوردگر PML دارای میانگین توان دوم خطای کوچک است و در طرح‌های آمارگیری خوب عمل می‌کند.

۶- شبیه‌سازی

فرض می‌کنیم که جامعه‌ی تحت مطالعه نامتناهی باشد. یک نمونه‌ی خوشه‌ای دو مرحله‌ای با $\tilde{n}_A$ خوشه و m واحد از هر خوشه به‌عنوان نمونه‌ی حاصل از چارچوب A $(n_A=\tilde{n}_A \times m)$ و یک نمونه‌ی تصادفی ساده با n_B واحد به‌عنوان نمونه‌ی حاصل از چارچوب B تولید می‌شود. با فرض نامتناهی بودن جامعه، $\frac{N_a}{N}$ و $\frac{N_b}{N}$ را با γ_a و γ_b جایگزین می‌کنیم. نمونه‌ی حاصل از چارچوب A شامل مقادیر

$$\{(y_{ij}, m_{ai}),\ i = ۱,۲, \dots, \tilde{n}_A \quad j = ۱, \dots, m\}$$

می‌باشد که در آن m_{ai} تعداد عناصر نمونه‌گیری شده از iامین خوشه‌ی نمونه، متعلق به حوزه‌ی a و y_{ij} نیز مقدار مرتبط با jامین عنصر نمونه در iامین خوشه‌ی نمونه می‌باشد. نمونه‌ی حاصل از چارچوب B شامل مقادیر نمونه‌ای

$$\{(y_j, n_b),\ j = ۱, \dots, n_B\}$$

است که در آن n_b تعداد عناصر نمونه‌ای متعلق به حوزه‌ی b و y_j مقدار مرتبط با jامین عنصر نمونه است. برای بدست آوردن مقادیر نمونه‌ای حاصل از چارچوب A، ابتدا m_{ai} را از یک توزیع بتا-دوجمله‌ای به‌صورت زیر تولید می‌کنیم.

۱) $\tilde{\gamma}_{ai}$، نسبت مورد انتظار مشاهداتی از خوشه‌ی iام که در حوزه‌ی a قرار دارد و از توزیع بتا با پارامترهای $b_۱$ و $b_۲=b_۱(۱-\gamma_a-\gamma_b)/\gamma_a$ تولید می‌شود بنابراین

$$E(\tilde{\gamma}_{ai}) = \frac{\gamma_a}{۱-\gamma_b} \qquad \text{و} \qquad \mathrm{Var}(\tilde{\gamma}_{ai}) = \frac{\gamma_a^{۲}(۱-\gamma_a-\gamma_b)}{\left[b_۱(۱-\gamma_b)+\gamma_a\right](۱-\gamma_b)^{۲}}$$

در نتیجه مقدار کوچک‌تر $b_۱$، تغییرپذیری $\tilde{\gamma}_{ai}$ از یک خوشه به خوشه‌ی دیگر را افزایش می‌دهد.

۲) به ازای مقدار $\tilde{\gamma}_{ai}$، m_{ai} را از یک توزیع دوجمله‌ای با پارامترهای m و $\tilde{\gamma}_{ai}$ تولید می‌کنیم.

۳) مراحل ۱ و ۲ را مستقلاً تکرار کرده تا $\{m_{ai}, i = ۱,۲, \dots, \tilde{n}_A\}$ به دست آید.

برای هر خوشه به ازای مقدار m_{ai} مربوط، مقادیر y_{ij} از مدل خطای آشیانی (Nested Error Model) زیر تولید می‌شوند:

$$y_{ij} = \mu_a + \alpha_{ai} + \varepsilon_{ij} \qquad j = ۱,۲, \dots, m_{ai}$$
$$y_{ij} = \mu_{ab} + \alpha_{abi} + \varepsilon_{ij} \qquad j = m_{ai} + ۱, \dots, m \qquad (۸)$$

$(\alpha_{ai}, \alpha_{abi})$ به‌ترتیب اثر تصادفی خوشه‌ی iام در حوزه‌های a و ab می‌باشد که از یک توزیع نرمال دو متغیره با میانگین صفر و ماتریس کوواریانس ۲×۲ با عناصر قطری $\rho\sigma^۲$ و عناصر غیر قطری $\delta\rho\sigma^۲$ بدست می‌آیند. ε_{ij}ها هم متغیرهای تصادفی مستقل و هم‌توزیع با $\varepsilon_{ij} \sim N(0, \sigma^۲(۱ - \rho))$ می‌باشند که مستقل از $(\alpha_{ai}, \alpha_{abi}, m_{ai})$ هستند. این مدل یک همبستگی یکسان $\sigma\rho$ را بین حوزه‌ها تضمین می‌کند.

برای بدست آوردن نمونه‌ی حاصل از چارچوب B ابتدا n_b را از یک توزیع دوجمله‌ای با پارامترهای n_B و $\gamma = \frac{\gamma_b}{۱-\gamma_a}$ تولید می‌کنیم. به ازای مقدار n_b حاصل، مقادیر نمونه‌ای $\{y_j,\ j = ۱, \dots, n_B\}$ از چارچوب B را از مدل زیر تولید می‌کنیم:

$$y_j = \mu_b + \delta_j \qquad j = ۱,۲, \dots, n_b$$
$$y_j = \mu_{ab} + \delta_j \qquad j = n_b + ۱, \dots, n_B \qquad (۹)$$

δ_jها متغیرهای تصادفی مستقل و هم‌توزیع با $\delta_j \sim N(0, \sigma^۲)$ می‌باشند. در اینجا از مقادیر $\mu_a = ۹$، $\mu_{ab} = ۱۰$، $\mu_b = ۱۱$، $\rho = ۰/۱$، $\delta = ۰/۵$ و $\sigma^۲ = ۱$ استفاده می‌کنیم.

برای شبیه‌سازی، نمونه‌گیری را $R = ۱۰۰۰۰$ بار تکرار کرده و از هر نمونه براوردگرهای $\hat{Y}_H(\hat{\theta}_H)$، $\hat{Y}_{FB}(\hat{\beta}_{FB})$، $\hat{Y}_{PML}(\hat{\theta}_P)$ و $\hat{Y}_S(\hat{\theta}_S)$ را با استفاده از مقادیر بهینه‌ی $\hat{\theta}_H$، $\hat{\beta}_{FB}$، $\hat{\theta}_P$ و $\hat{\theta}_S$ (این پارامترها در بخش ۵ آورده شده‌اند) و به ازای مقادیر مختلف $b_۱$، $b_۲$، γ_a، γ_b، $\tilde{n}_A$ و n_B محاسبه می‌کنیم. سپس میانگین توان دوم خطای تجربی هر براوردگر، متوسط توان دوم انحراف براوردگر از مقدار واقعی ($EMSE$) را به‌صورت زیر محاسبه می‌کنیم:

$$EMSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2 \qquad (10)$$

که در آن $\hat{Y}_r$ عبارت است از $\hat{Y}$ برای rامین تکرار نمونه‌گیری و R تعداد تکرارهای نمونه‌گیری می‌باشد. در رابطه‌ی (۱۰) چون Y معلوم و در دسترس نمی‌باشد در این صورت از امید ریاضی آن به‌صورت زیر استفاده شده است:

$$\frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - \bar{\hat{Y}})^2 \qquad (11)$$

در این شبیه‌سازی، $EMSE$ به ازای $R = 10000$ نمونه محاسبه شده است. واریانس مونت کارلوی توان دوم خطا به‌صورت زیر محاسبه می‌شود:

$$s^2(EMSE) = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{Z}_r - \bar{\hat{Z}})^2 \qquad (12)$$

که در آن $\hat{Z}_r = (\hat{Y}_r - Y)^2$ و $\bar{\hat{Z}} = \sum_r \hat{Z}_r/R$ می‌باشد. با گرفتن جذر از رابطه‌ی فوق، خطای معیار مونت کارلو را می‌توان محاسبه کرد. تمام محاسبات با استفاده از نرم‌افزار S-plus صورت گرفته است.

جدول (۱) مقادیر $EMSE$ و خطای معیار مونت کارلوی متناظر را به ازای $\tilde{n}_A = 10$، $m = 30$، $n_B = 100$ و $\tilde{n}_A = 20$، $m = 30$، $n_B = 200$ و $\tilde{n}_A = 20$، $m = 30$، $n_B = 100$ و سه مقدار مختلف (b_1, b_2) و نیز مقادیر معین γ_a و γ_b نشان می‌دهد.

در جدول ۱ میانگین توان دوم خطای مونت کارلو ($EMSE$) برای ۱۰۰۰۰ بار تکرار نمونه‌گیری و در داخل دو کمان انحراف معیار مونت کارلو ($s(EMSE)$) هر براوردگر داده شده است.

همان‌طور که در جدول ۱ مشاهده می‌شود براوردگر ماکسیمم درستنمایی نما (PML) همواره دارای $EMSE$ کوچک‌تری نسبت به براوردگرهای دیگر است. در صورتی که $\gamma_a \leq \gamma_b$ یا n_A خیلی بزرگ‌تر از n_B باشد، براوردگر تک‌چارچوبی (S) در مقایسه با براوردگرهای دوچارچوبی (H,FB,PML) افزایش قابل ملاحظه‌ای را در $EMSE$ نشان می‌دهد.

جدول ۱- میانگین توان دوم خطا و خطای معیار مونت کارلوی براوردگرهای چارچوب دوگان (میانگین توان دوم خطا (EMSE)×۱۰۰)

γ_b	γ_a	*H*	*FB*	*PML*	*S*
		$n_B = ۱۰۰\ m = ۳۰\ \tilde{n}_A = ۱۰\ n_A = ۳۰۰\ \ b_1 = ۲,۲/۵,۳$			
۰/۱	۰/۱	۴/۷(۰/۰۶۵)	۵/۱۳(۰/۰۸۱)	۴/۴(۰/۰۶۱)	۶/۶۸(۰/۰۹)
۰/۱	۰/۲	۵/۰۵(۰/۰۷۱)	۵/۱۵(۰/۰۷۸)	۴/۵۷(۰/۰۶)	۱۱/۴۲(۰/۱۵۵)
۰/۲	۰/۱	۵/۷(۰/۰۸۱)	۵/۹(۰/۰۸۶)	۵/۴۶(۰/۰۷۹)	۵/۸(۰/۰۸۳)
		$n_B = ۲۰۰\ m = ۳۰\ \tilde{n}_A = ۲۰\ n_A = ۶۰۰\ \ b_1 = ۲,۲/۵,۳$			
۰/۱	۰/۱	۲/۳۷(۰/۰۳۳)	۲/۴۷(۰/۰۳۹)	۲/۱۳(۰/۰۳۱)	۳/۵۲(۰/۰۴۸)
۰/۱	۰/۲	۲/۳۷(۰/۰۳۴)	۲/۴۳(۰/۰۳۷)	۲/۳(۰/۰۳۳)	۵/۴۵(۰/۰۷۶)
۰/۲	۰/۱	۲/۷۵(۰/۰۳۹)	۲/۸۷(۰/۰۴۱)	۲/۷(۰/۰۳۸)	۲/۹۸(۰/۰۴۲)
		$n_B = ۱۰۰\ m = ۳۰\ \tilde{n}_A = ۲۰\ n_A = ۶۰۰\ \ b_1 = ۲,۲/۵,۳$			
۰/۱	۰/۱	۳/۲۱(۰/۰۴۶)	۳/۱۲(۰/۰۴۴)	۲/۸۵(۰/۰۴۱)	۷/۸۳(۰/۱۲)
۰/۱	۰/۲	۳/۰۷(۰/۰۴۳)	۲/۸۳(۰/۰۴)	۲/۷۴(۰/۰۴۱)	۱۳/۳۱(۰/۱۹)
۰/۲	۰/۱	۴/۲۵(۰/۰۵۹)	۴/۲۶(۰/۰۵۹)	۳/۹۲(۰/۰۵۵)	۶/۶(۰/۰۹۵)

برای مثال در جدول ۱ برای حالتی که $\gamma_a = ۰/۱$ ، $\gamma_b = ۰/۲$ و $n_A/n_B = ۳$ داریم:

$EMSE\ (\hat{Y}_{FB}) = ۲/۴۳$ $EMSE\ (\hat{Y}_{PML}) = ۲/۳$ $EMSE\ (\hat{Y}_S) = ۵/۴۵$

$EMSE\ (\hat{Y}_H) = ۲/۳۷$

و وقتی که n_A/n_B به شش افزایش می‌یابد، داریم:

$EMSE\ (\hat{Y}_{FB}) = ۲/۸۳$ $EMSE\ (\hat{Y}_{PML}) = ۲/۷۴$ $EMSE\ (\hat{Y}_S) = ۱۳/۳۱$

$EMSE\ (\hat{Y}_H) = ۳/۰۷$

حال اگر $\gamma_a = ۰/۲$ ، $\gamma_b = ۰/۱$ و $n_A/n_B = ۳$ کارایی حاصل زیاد نخواهد بود. برای مثال:

$EMSE\ (\hat{Y}_{FB}) = ۲/۸۷$ $EMSE\ (\hat{Y}_{PML}) = ۲/۷$ $EMSE\ (\hat{Y}_S) = ۲/۹۸$

$EMSE\ (\hat{Y}_H) = ۲/۷۵$

و وقتی که n_A/n_B به شش افزایش می‌یابد، داریم:

$EMSE\ (\hat{Y}_s)$ =۶/۶ $EMSE\ (\hat{Y}_{PML})$ =۳/۹۲ $EMSE\ (\hat{Y}_{FB})$ =۴/۲۶
$EMSE\ (\hat{Y}_H)$ =۴/۲۵

حال می‌خواهیم اثر افزایش اندازه‌های نمونه‌ای مختلف را بر میانگین توان دوم خطای براوردگرهای دوچارچوبی H,FB,PML و براوردگر تک‌چارچوبی (S) بررسی کنیم.

الف- زمانی که n_B ثابت و n_A افزایش یابد برای تمامی حالت‌های مختلف γ_a و γ_b ($\gamma_a=\gamma_b$، $\gamma_a<\gamma_b$ و $\gamma_a>\gamma_b$)، میانگین توان دوم خطا برای براوردگرهای دوچارچوبی PML، FB و H کاهش و برای براوردگر S افزایش می‌یابد.

ب- زمانی که n_B افزایش و n_A ثابت باشد برای تمامی حالت‌های مختلف γ_a و γ_b، میانگین توان دوم خطای تمامی براوردگرها کاهش می‌یابند. در این حالت $EMSE$ براوردگر PML کم‌تر از $EMSE$ براوردگرهای S، FB و H می‌باشد.

پ- زمانی که هر دو n_B و n_A افزایش یابند برای تمامی حالت‌های مختلف γ_a و γ_b، میانگین توان دوم خطای تمامی براوردگرها کاهش می‌یابند. در این حالت نیز $EMSE$ براوردگر PML کم‌تر از $EMSE$ براوردگرهای S ، FB و H می‌باشد.

در نهایت، اثر افزایش γ_a و γ_b را بر میانگین توان دوم خطای براوردگرهای دوچارچوبی PML، FB، H و براوردگر تک‌چارچوبی (S) بررسی کنیم.

الف- زمانی که γ_a ثابت و γ_b افزایش یابد، دراین صورت برای حالتی که $n_A = ۳۰۰$ و $n_B = ۱۰۰$ است، میانگین توان دوم خطای تمامی براوردگرها افزایش می‌یابند، برای حالتی که $n_A = ۶۰۰$ و $n_B = ۱۰۰$ است، میانگین توان دوم خطای براوردگرهای دوچارچوبی PML، FB و H کاهش و براوردگر S افزایش می‌یابد و برای حالتی که $n_A = ۶۰۰$ و $n_B = ۲۰۰$ است، میانگین توان دوم خطای براوردگرهای PML، FB و H تقریباً نزدیک به هم هستند اما میانگین توان دوم خطای براوردگر S افزایش داشته است. در این حالت‌ها کم‌ترین EMSE مربوط به براوردگر PML است.

ب- زمانی که γ_a افزایش و γ_b ثابت باشد، دراین صورت برای حالت‌های $n_A = ۳۰۰$ و $n_B = ۱۰۰$، $n_A = ۶۰۰$ و $n_B = ۱۰۰$ و $n_A = ۶۰۰$ و $n_B = ۱۰۰$ میانگین توان دوم خطای براوردگرهای دوچارچوبی PML، FB و H افزایش و براوردگر S کاهش می‌یابد. در این حالت با وجود آن‌که $EMSE$ براوردگر S کاهش یافته ولی مقدار آن

بزرگ‌تر از $EMSE$ براوردگرهای PML، FB و H است و کم‌ترین $EMSE$ مربوط به براوردگر PML می‌باشد.

۷ -بحث و نتیجه‌گیری

آمارگیری‌های چندچارچوبی هنگامی که جوامع کمیاب هستند یا چارچوب کامل موجود برای نمونه‌گیری پرهزینه باشد، بسیار مفید و سودمند است.

در آمارگیری‌های چندچارچوبی، براوردگرهای مختلفی برای براورد مجموع جامعه پیشنهاد داده شده است. اما انتخاب براوردگر بستگی به طراحی و پیچیدگی طرح آمارگیری دارد.

براوردگر ماکسیمم درستنمایی نما به‌دلیل این‌که از وزن‌های یکسانی برای همه‌ی متغیرها استفاده می‌کند، نسبت به براوردگرهای هارتلی و فولر-بورمیستر در بررسی‌هایی با متغیرهای متعدد، ارجحیت دارد.

با افزایش هم‌زمان اندازه‌های نمونه در دو چارچوب (n_A و n_B)، برای تمامی حالت‌های مختلف γ_a و γ_b ($\gamma_a=\gamma_b$، $\gamma_a<\gamma_b$ و $\gamma_a>\gamma_b$)، میانگین توان دوم خطای تمامی براوردگرها (PML، FB، H و S) کاهش می‌یابند.

مرجع‌ها

[۱] نورینی، مرجان (۱۳۸۲). طرح‌های نمونه‌گیری چندچارچوبی. پایان نامه‌ی کارشناسی ارشد. دانشگاه صنعتی اصفهان، اصفهان.

[2] Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association,* **81**, 1074–1079.

[3] Casady, R.J., Snowden, C.B. and Sirken, M.G. (1981). A study of dual frame estimators for the national health interview survey. Proceedings of the Survey Research Section, *American Statistical Association*, 444–447.

[4] Cochran, R.S. (1964). Multiple-frame Surveys, Proceedings at American Statistical Association Meeting, Univ. Wyoming, Dec.

[5] Ford, B.L. and Bosecker, R.R. (1979). Multiple frame estimation with stratified overlap domain, in proceeding of the Social Statistics Section. *American Statistical Association*, 219–224.

[6] Fuller, W.A. and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames, in proceeding of the social statistics section. *American Statistical Association*, 245–249.

[7] Groves, R.M. and Lepkowski, J.M. (1986). A mean squared error model for dual frame mixed model survey design. *Journal of the American Statistical Association*, **81**, 930–937.

[8] Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

[9] Hartley, H.O. (1962). Multiple frame survey, In Proceeding of the Social Statistics Section, *American Statistical Association*, 203–206.

[10] Hartley, H.O. (1974). Multiple frame methodology and selected application. *Sankhya*, Ser. C, **36**, 99–118.

[11] Kalton, G. and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Series A **149**, 65–82.

[12] Lohr, S.L. and Rao, J.N.K. (2006). Estimation in multiple–frame surveys. *Journal of the American Statistical Association,* **101**, 1019–1030.

[13] Lund, R.E. (1968). Estimators in multiple frame surveys, in proceeding of the social statistics section. *American Statistical Association*, 282–288.

[14] Sirken, M.G. and Cassady, R.J. (1982). Nonresponse in dual frame survey based on area/list and telephone frames. *American Statistical Association*, 151–153.

[15] Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association,* **86**, 779–784.

[16] Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual frame survey with complex designs. *Journal of the American Statistical Association*, **91**, 349-356.

[17] Vogel, F.A. (1975). Survey with overlapping frame-problems in application, in proceedings of the social statistics. *American Statistical Association*, 694-699.

مرجان نورینی
کارشناس ارشد آمار
تهران، خیابان دکتر فاطمی، نبش رهی معیری، پلاک ۱، مرکز آمار ایران.
رایانشانی: mar_noorini@yahoo.com

انور قیطولی
کارشناس ارشد آمار
تهران، خیابان دکتر فاطمی، نبش رهی معیری، پلاک ۱، مرکز آمار ایران.
رایانشانی: a.ghaitoly@gmail.com