
بسم الله الرحمن الرحيم

دانشگاه علامه طباطبائی
دانشکده‌ی اقتصاد
گروه آمار، ریاضی و کامپیوتر

پایان‌نامه برای دریافت درجه‌ی کارشناسی ارشد **آمار اجتماعی-اقتصادی**

عنوان

جانهی داده‌های گمشده در آمارگیری‌های پانلی با استفاده از الگوریتم EM

پژوهشگر

آسیه رشیدی نژاد

استاد راهنما

دکتر حمیدرضا نواب‌پور

استاد مشاور

دکتر مجتبی گنجعلی

بهمن ۸۸

تأیید پایان‌نامه‌ی کارشناسی ارشد توسط دانشجو

عنوان پایان‌نامه: جانهی داده‌های گمشده در آمارگیری‌های پانلی با استفاده از الگوریتم EM

نام دانشجو: آسیه رشیدی نژاد

شماره‌ی دانشجویی: ۸۶۱۱۱۶۱۰۲

استاد راهنما: دکتر حمیدرضا نواب‌پور

این‌جانب **آسیه رشیدی نژاد** دانشجوی کارشناسی‌ارشد رشته‌ی **آمار اجتماعی-اقتصادی** دانشکده‌ی اقتصاد دانشگاه علامه طباطبائی گواهی می‌نمایم پژوهش‌های ارائه شده در پایان‌نامه با عنوان مذکور توسط شخص این‌جانب انجام شده است و درستی مطالب نگارش یافته مورد تأیید می‌باشد. همچنین گواهی می‌نمایم مطالب مندرج در پایان‌نامه تاکنون برای دریافت هیچ نوع مدرک یا امتیازی توسط این‌جانب یا فرد دیگری در هیچ کجا ارائه نشده است و در نگارش متن پایان‌نامه شیوه‌ی نگارش مصوب دانشکده‌ی اقتصاد را به‌طور کامل رعایت نموده‌ام. چنان‌چه در هر زمان خلاف آن‌چه گواهی نموده‌ام مشاهده گردد خود را از آثار حقیقی و حقوقی ناشی از دریافت مدرک کارشناسی ارشد محروم می‌دانم و هیچ‌گونه ادعایی نخواهم داشت.

تاریخ: **امضا دانشجو:**

اهدایه

تقدیم به همه ی آن هایی که

می خواهند بیشتر بدانند

سپاس‌گزاری

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست. بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می‌کنم وجود مقدس‌شان را که در این سردترین روزگاران، بهترین پشتیبان من بودند.

در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر حمیدرضا نواب‌پور، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید.

از جناب آقای دکتر مجتبی گنجعلی که زحمت مطالعه و مشاوره این پایان‌نامه را تقبل فرمودند و به نحو احسن اینجانب را مورد راهنمایی قرار دادند، کمال امتنان را دارم.

از جناب آقای دکتر بامنی مقدم و جناب آقای دکتر پورطاهری که داوری این پایان‌نامه را بر عهده داشتند قدردانی می‌کنم.

این پایان‌نامه از حمایت‌های مادی و معنوی پژوهشکده‌ی آمار ایران برخوردار بوده است، لذا بر خود لازم می‌دانم که از توجه این پژوهشکده در حمایت از پایان‌نامه‌های مرتبط با روش‌شناسی آمارگیری تشکر کنم.

آسیه رشیدی نژاد

بهمن ۱۳۸۸

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

چکیده

در اقتصاد و سایر علوم اجتماعی، پژوهش‌گران اغلب تمایل به مدل‌بندی داده‌های پانلی که در آن واحدهای نمونه‌ای به طور مکرر در مقاطع زمانی مختلف مشاهده می‌شوند، دارند. یکی از کاربردهای داده‌های پانلی براورد نرخ تغییر میانگین متغیر پاسخ در طی زمان است. در انواع آمارگیری‌ها به ویژه آمارگیری‌های پانلی، بی‌پاسخی یک مشکل اساسی است که در داده‌های علوم اجتماعی و پزشکی به وفور رخ می‌دهد. این نوع مطالعه‌ها معمولاً با کاهش پاسخگو در دوره‌های دوم به بعد تولید داده‌ها مواجه هستند. این امر که منجر به نمونه‌ی کاهیده می‌شود سبب کاهش کارایی براوردگرها و غالباً نیز سبب اریبی آن‌ها می‌شود. برای برخورد با این مشکل در آمارگیری پانلی روش‌های «جانهی» و «وزن‌دهی» گوناگونی وجود دارد که از جمله‌ی این روش‌های جانهی، جانهی با الگوریتم EM می‌باشد.

الگوریتم EM یک الگوریتم مکرر برای براورد ماکسیمم درستنمایی مسئله‌ی داده‌های گمشده یا ناتمام می‌باشد. با توجه به این‌که ساختار گمشدگی تصادفی فرض شده است لذا این الگوریتم برای جانهی مناسب می‌باشد.

در این پایان‌نامه پس از معرفی مفهوم‌های اولیه آمارگیری پانلی، انواع گمشدگی در آمارگیری‌های پانلی و ساختارهای گمشدگی، الگوریتم EM به عنوان روشی برای جانهی داده‌های گمشده معرفی می‌شود و به دلیل مشکلاتی که در محاسبات این الگوریتم ممکن است وجود داشته باشد چهار نوع از این الگوریتم معرفی می‌شود. سرانجام با استفاده از داده‌های آمارگیری پانلی خانواری انگلیس، توزیع نمونه‌گیری شبیه‌سازی شده و روش جانهی با الگوریتم EM با دو روش جانهی با میانگین مشاهده‌های مشابه و جانهی با نمونه‌ی جدید از نظر معیارهای مختلف (با توجه به اثر اندازه‌ی نمونه، همبستگی بین دوره‌ها و نرخ بی‌پاسخی دوره) مقایسه می‌شوند. نتایج این مطالعه نشان می‌دهد که جانهی متغیر تحت بررسی در آمارگیری پانلی خانواری انگلیس با استفاده از الگوریتم EM وقتی که همبستگی بین دو دوره زیاد باشد، عملکرد بهتری دارد.

واژه‌های کلیدی: *آمارگیری پانلی، بی‌پاسخی، کاهش پاسخگو، داده‌های ناتمام، داده‌های کامل، جانهی، گمشدگی تصادفی، الگوریتم EM، جانهی با نمونه‌ی جدید، جانهی با میانگین مشاهده‌های مشابه.*

فصل ۱
کلیات

۱.۱ مقدمه

امروزه نقش آمارگیری به ویژه آمارگیری در طی زمان که در چند دهه‌ی اخیر آغاز شده است اساسی و پر کاربرد می‌باشد، از جمله در مسئله‌های اجتماعی، سیاسی، اقتصادی و فرهنگی از این آمارگیری‌ها به طور مکرر استفاده می‌شود. یکی از این نوع آمارگیری‌ها، آمارگیری طولی است. معمولاً از آمارگیری طولی برای بررسی تغییر پارامتر مورد نظر در طول زمان استفاده می‌شود برای مثال از این نوع آمارگیری در زمینه‌های پویایی درآمد خانوار، بازار کار و رفتارهای جمعیت‌شناختی و مسائل سیاسی استفاده می‌شود. همچنین برای برنامه‌ریزی در سطوح مختلف یک کشور دانستن نرخ تغییر برخی پارامترهای اجتماعی، اقتصادی و ... از اهمیت بالایی برخوردار است، چنان‌چه بی‌توجهی به آن نمی‌تواند مشکلاتی اعم از بیکاری، وضعیت نامناسب اقتصادی و ... را رفع کند. مثلاً برای بهبود در وضعیت حمل و نقل عمومی در یک جامعه، نمونه‌ای از افراد آن جامعه گزینش شده و برای بررسی این‌که آیا نظر افراد این جامعه با گذشت زمان در مورد کیفیت حمل و نقل تغییر کرده است و اقدامات مسئولان برای بهبود آن در تغییر نگرش افراد موثر بوده است در طی زمان به آن‌ها مراجعه می‌شود. گاهی تغییر بعضی پارامترها، وضعیت اجتماعی، اقتصادی یا سیاسی جامعه را مشخص می‌کند. به عنوان مثال تعیین نرخ بیکاری می‌تواند نشان دهنده‌ی وضعیت اجتماعی و اقتصادی جامعه باشد و در صورت

بالا بودن این نرخ، مسئولان ذی‌ربط باید به گونه‌ای برنامه‌ریزی کنند تا این نرخ کاهش یابد. بنابر این بررسی تغییرها در مورد پارامترهای مختلف جامعه ضرورت پیدا می‌کند.

برای براورد تغییر پارامترها در طی زمان نیاز به انتخاب روش مناسب آمارگیری است، لذا دو نوع روش آمارگیری مکرر (آمارگیری در طی زمان) در نظر گرفته می‌شود. این دو روش عبارتند از:

- آمارگیری طولی، و
- آمارگیری مقطعی.

ویژگی اصلی آمارگیری طولی این است که اندازه‌های مکرر در طول زمان برای واحدهای نمونه‌ای یکسان گرداوری می‌شوند. در آمارگیری طولی خصیصه یا خصیصه‌های واحدهای نمونه‌ای یکسان در طول زمان به طور مکرر اندازه‌گیری می‌شود. آمارگیری طولی تغییر خصیصه را برای واحدهای نمونه‌ای در طول زمان نشان می‌دهد. منظور از طول زمان دوره‌هایی است که در آن‌ها خصیصه‌ی مورد نظر از واحدهای نمونه‌ای اندازه‌گیری می‌شود. اما در آمارگیری مقطعی خصیصه یا خصیصه‌ها در مقاطع زمانی مختلف از واحدهای نمونه‌ای متفاوت اندازه‌گیری می‌شوند. در این شیوه نمونه‌ها در مقاطع مختلف مستقل از هم انتخاب می‌شوند. در این نوع آمارگیری در واقع هدف شناخت یک وضعیت در یک زمان خاص می‌باشد و این آمارگیری توان اندازه‌گیری نرخ تغییر پارامتر مورد نظر در طی زمان را ندارد بلکه تغییر بین واحدهای نمونه‌ای را اندازه‌گیری می‌کند.

به کار بردن این نوع از آمارگیری‌های مکرر، امکان بررسی تغییرهای مختلف در سطح جامعه را فراهم می‌کند؛ همچنین محققان می‌توانند به بررسی تک تک افراد بپردازند. آمارگیری‌های مکرر که در دوره‌های زمانی با فاصله‌ی یکسان صورت می‌گیرند، می‌توانند روند تغییرها را نیز نشان دهند. در آمارگیری طولی تاثیر عامل زمان مورد سنجش و ارزیابی قرار می‌گیرد. با استفاده از این نوع آمارگیری نه تنها می‌توان تغییر را در پاسخ واحدهای نمونه‌ای شناخت، بلکه می‌توان علل تغییرهای به دست آمده در عقیده‌ها و رفتارهای واحدهای نمونه‌ای جامعه تحت مطالعه را مورد شناسایی قرار داد. به عنوان مثال، در یک نظر سنجی از افراد یک جامعه درباره شرکت در انتخابات، افرادی که در دور اول نظر مثبتی به شرکت در انتخابات دارند ممکن است در دور دوم آمارگیری تغییر نظر داده و بعضی دیگر به دلیل‌های مختلف از پاسخ دادن اجتناب کنند.

۲.۱ روش‌های گرداوری داده‌های طولی

راه‌های متفاوتی برای گرداوری داده‌های طولی وجود دارد:

۱) آمارگیری گذشته‌نگر،

۲) آمارگیری پانلی، و

۳) اتصال داده‌ها.

۱.۲.۱ آمارگیری گذشته‌نگر

در آمارگیری گذشته‌نگر؛ فقط یک مرتبه مصاحبه صورت می‌گیرد و درباره‌ی گذشته اطلاعاتی گرداوری می‌شود. از فایده‌های این روش سادگی و ارزان بودن آن می‌باشد، زیرا فقط یک مرتبه مصاحبه صورت می‌گیرد و نیازی به ردیابی واحدهای نمونه‌ای نمی‌باشد. همچنین اطلاعات مربوط به داده‌های طولی فوراً در دسترس می‌باشد، زیرا نیاز به انتظار برای مصاحبه‌های بعدی برای تغییر اندازه‌گیری وجود ندارد. عیب اصلی روش گرداوری داده‌های طولی از طریق آمارگیری گذشته‌نگر این است که اطلاعات مربوط به گذشته است و وابسته به میزان یادآوری اتفاق‌ها توسط پاسخگو است در نتیجه صحت و درستی این روش مورد سؤال می‌باشد. مثلاً میزان یادآوری افراد در مورد سطوح درآمدی آن‌ها دور از انتظار است.

به عنوان مثال آمارگیری زنان و اشتغال در انگلیس WES[1] (مارتین و رابرتس، ۱۹۸۴) گذشته‌نگر است. این آمارگیری شامل یک نمونه‌ی احتمالاتی از تمام زنان انگلیس که در محدوده سنی (۱۶-۵۹) در سال ۱۹۸۰ کار می‌کردند، می‌باشد. هدف اصلی این آمارگیری بررسی اثر وضعیت اشتغال در زندگی زنان بود، که تصویر کاملی از گذشته‌ی زنان و فعالیت و بازار کار فعلی آن‌ها ارائه شد. در این آمارگیری ۵۵۸۸ پاسخگو حضور داشتند و نرخ پاسخ ۸۰% بود و تاریخ‌هایی مربوط به اشتغال، ازدواج و باروری گرداوری شد، به علاوه داده‌های زیادی مربوط به نوع شغل، ساعت کار، وضعیت کار، درآمد، جستجوی شغل و نگرش آن‌ها درباره پرداخت حقوق نیز به دست آمد.

۲.۲.۱ آمارگیری پانلی

در این نوع آمارگیری طولی نمونه‌ای از واحدها در طول زمان دنبال می‌شود و داده‌ها از تکرار مصاحبه‌ها گرداوری می‌شوند (دوره‌ها). حالت‌های متعددی با این تعریف به وجود می‌آید، اما تمایز اصلی بین حالت‌های زیر است:

۱) آمارگیری که شامل پانل تکی در مدت نامحدود می‌باشد، و

۲) آمارگیری که شامل پانل متداخل چندگانه در مدت ثابت می‌باشد (پانل چرخشی).

تمایز دیگر مربوط به انواع اطلاعات طولی که توسط آمارگیری پانلی گرداوری شده است به ویژه گستردگی داده‌ها درباره‌ی مدت بین دوره‌ها و به علاوه درباره‌ی اطلاعات همزمان می‌باشد. حالت‌های مختلف در این آمارگیری در بخش ۳.۱ بیان می‌شوند.

[1] Woman and Employment Survey

آمارگیری پویایی درآمد و نیروی کار کانادا [۲] (SLID) نمونه‌ای از پانل چرخشی می‌باشد. این آمارگیری یک آمارگیری اجتماعی است که هدف آن تعیین الگوی فعالیت در بازار کار و تغییر در درآمد است (وبر، ۱۹۹۴). نمونه‌ی اولیه این پانل در سال ۱۹۹۳ شامل ۱۵۰۰۰ خانوار است که در آن از روش نمونه‌گیری با احتمال‌های یکسان، استفاده شده است. هدف‌های دیگر این آمارگیری از جمله پویایی اشتغال و بیکاری، تحول در نیروی کار، کیفیت شغل، پویایی اقتصاد خانواری، مسائل مربوط به جمعیت‌شناختی و آموزش می‌باشد.

۳.۲.۱ اتصال داده‌ها

در این حالت داده‌های طولی بدون مصاحبه حضوری توسط اتصال داده‌های واحدها به یکدیگر از منابع داده‌ای که وجود دارند، گرداوری می‌شوند. این مجموعه داده‌ها ممکن است داده‌های ثبتی باشند که برای هدف‌های اداری تهیه شده‌اند. از فایده‌های این روش اندازه‌ی نمونه‌ی بزرگ و خطای نمونه‌گیری کوچک می‌باشد. علاوه بر این چون مصاحبه‌ای انجام نمی‌شود هزینه سربار یا خطای یادآوری که توسط پاسخگو ایجاد می‌شود وجود ندارد. به هر حال مشکلاتی هم وجود دارد. یکی از مشکلات این است که اتصال رکوردها به راحتی امکان‌پذیر نمی‌باشد. علاوه بر این تحلیل با متغیرهایی که محدود به آمارگیری اصلی شده است مشکل می‌باشد. به عنوان مثال داده‌های مالیات فقط به افرادی اشاره دارد که مالیات خود را پرداخت کرده‌اند و اطلاعاتی در مورد افراد با درآمد پایین وجود ندارد. همچنین دامنه‌ی متغیرهای گرداوری شده محدود می‌باشند.

مثالی از این حالت، پرونده‌ی سرشماری طولی فنلاند[۳] می‌باشد. مجموعه داده‌های طولی با اتصال از سرشماری‌های فنلاند در سال‌های ۱۹۷۰، ۱۹۷۵، ۱۹۸۰، ۱۹۸۵، ۱۹۹۰ تولید شده‌اند (استارک، ۱۹۹۴). پرونده شامل تمام افرادی است که حداقل در یکی از پنج سرشماری حضور داشته‌اند. در مقابل فایده‌های مهم این روش، می‌توان به این نکته اشاره کرد که فاصله پنج ساله بین مشاهده‌ها برای تحلیل مناسب نمی‌باشد، یعنی این‌که تغییرات در فاصله‌ی زمانی بین دوره‌ها زیاد است.

۳.۱ انواع آمارگیری پانلی

در این پایان‌نامه روش گرداوری داده‌های طولی به صورت آمارگیری پانلی در نظر گرفته شده است. علت ارزیابی طرح پانلی این است که معیاری برای سنجش داده‌ها در زمان‌های مختلف فراهم می‌کند (به عنوان مثال مجموع و میانگین مقادیری که ممکن است به طور مداوم تغییر کنند مانند درآمد یا شاخص‌های مربوط به اتفاق‌های برجسته). دلیل اصلی به کارگیری آمارگیری پانلی این است که این‌گونه پانل‌ها قادر به اندازه‌گیری تغییرهای عمده در سطوح مختلف زمانی مربوط به یک واحد می‌باشند.

[۲] Survey of Labour and Income Dynamics
[۳] The Finnish Longitudinal Census File

در کل پنج نوع طرح امارگیری پانلی مشخص شده است، که به طور خلاصه توضیح داده می‌شوند:

پانل ثابت: در این حالت داده‌ها در طول زمان از واحدهای ثابت گرداوری می‌شوند یعنی بعد از انتخاب نمونه‌ی اولیه هیچ واحدی به آن اضافه نمی‌شود. تنها عیب این حالت کاهش نمونه به دلیل خروج برخی واحدها از مطالعه می‌باشد. مطالعه هم‌گروهی انگلیس (۱۹۷۰)، [۴](BCS۷۰) مثالی از این حالت است.

پانل ثابت به اضافه زاد: تقریباً شبیه پانل ثابت می‌باشد به جز این‌که زادهایی که در طول دوره رخ می‌دهند به نمونه اضافه می‌شوند، یعنی در هر دوره‌ی گرداوری داده‌ها، حداقل یک واحد نمونه‌ای به دور قبل اضافه می‌شود. اگر تعداد زادها در طول عمر پانل زیاد نباشد، در این حالت پانل ثابت در نظر گرفته می‌شود. بیشتر آمارگیری‌های پانلی خانواری به این صورت است.

پانل مکرر: این طرح شامل یک سری آمارگیری پانلی است که ممکن است در طول زمان متداخل یا مستقل باشند. در واقع هر پانل طوری طراحی می‌شود که جمعیت معادل را نشان دهد یعنی تعریف جمعیت یکسانی در نقاط مختلف زمان به کار برده می‌شود. مطالعه هم‌گروهی جوانان [۵](YCS) در انگلیس مثالی از این حالت است که پانل‌ها شامل نمونه‌هایی است که در رده‌ی سنی ۱۶-۱۷ می‌باشند و در سال‌های مختلف انتخاب شده‌اند. هر پانل شامل حداقل سه دوره در طول حداقل سه سال است.

پانل چرخشی: با استفاده از نسبت‌های از قبل تعیین شده، واحدهای نمونه‌ای در هر دوره جایگزین می‌شوند، یعنی هر واحد در نمونه برای تعداد دوره‌های یکسان در پانل باقی خواهد ماند. کالتون و سیترو (۱۹۹۳) پانل چرخشی را حالت خاصی از پانل مکرر با تداخل دانستند، زیرا الگوی تداخل ثابت است و هر پانل در هر نقطه از زمان جمعیت یکسانی را نشان می‌دهد. پانل چرخشی وقتی استفاده می‌شود که هدف اصلی براوردهای مقطعی و کوتاه مدت است. برای آمارگیری نیروی کار در بیشتر کشورها از پانل چرخشی استفاده می‌شود (استیل، ۱۹۹۷).

پانل خردشده: این پانل شامل ترکیبی از نمونه‌های پانلی و مقطعی است به این صورت که تعدادی از واحدهای نمونه‌ای به طور ثابت در تمام دوره‌های پانل حضور دارند و در هر دوره تعدادی واحد نمونه‌ای به صورت مقطعی به مطالعه اضافه می‌شوند به طوری که در ترکیب با طرح پانلی تنها در یک دوره حضور دارند. طرح معمولی آن توسط کیش (۱۹۸۷) بررسی شد که شامل یک نمونه‌ی پانلی از داده‌هایی است که در وضعیت‌های مختلف گرداوری شده است و یک نمونه مقطعی هم به آن اضافه می‌شود.

به طور کلی استفاده از طرح پانلی ثابت وقتی است که نیاز به اطلاعات جزیی تغییرها در مورد ویژگی‌های مورد نظر از نمونه انتخاب شده باشد. پانل چرخشی و مکرر برای سنجش تغییرها در براوردهای طولی مفید

[۴] The 1970 British Cohort Study
[۵] Youth Cohort Study

می‌باشند زیرا شامل پانل‌های چندگانه هستند. برای مطالعه پدیده‌هایی که آرام تغییر می‌کنند یا اثرهای طولانی مدت دارند به پانل‌های طولانی مدت نیاز است در صورتی که تغییرهای سریع یا اثرهای کوتاه مدت یا میان مدت مورد نظر باشند از پانل‌های کوتاه مدت یا پانل‌های چرخشی استفاده می‌شود.

۴.۱ مزیت‌ها و ضعف‌های آمارگیری پانلی

یکی از مزیت‌های داده‌های طولی این است که در آن داده‌های هر خوشه دارای همبستگی می‌باشند، این همبستگی معمولاً مثبت است. اگر براورد نرخ تغییر میانگین مد نظر باشد، واریانس تغییر میانگین در آمارگیرهای پانلی دو دوره‌ای به صورت زیر محاسبه می‌شود:

$$Var(\bar{Y}_2 - \bar{Y}_1) = Var(\bar{Y}_2) + Var(\bar{Y}_1) - 2Cov(\bar{Y}_2, \bar{Y}_1)$$

که در آن $\bar{Y}_t$ میانگین متغیر پاسخ در دوره‌ی tام $(t = 1,2)$ آمارگیری است. با توجه به این‌که معمولاً کوواریانس بین دو دوره مثبت است، لذا واریانس تفاضل کوچک‌تر از حالتی است که همبستگی صفر باشد یعنی کمتر از حالتی است که آمارگیری مقطعی باشد. مزیت دیگر آمارگیرهای طولی کاهش در هزینه‌های آمارگیری است. البته کار با داده‌های طولی پیچیده‌تر از داده‌های مقطعی است بنابر این هزینه‌ی کار با آن‌ها زیادتر است. مهم‌ترین عیب این‌گونه داده‌ها این است که مشاهده‌ی تمام واحدهای نمونه‌ای در طول دوره‌ی مطالعه میسر نیست زیرا ممکن است واحدهای نمونه‌ای مطالعه را ترک کنند. این حالت کاهش پاسخگو نامیده می‌شود و داده‌های تولید شده از این کاهش را نمونه‌ی کاهیده می‌نامند. به عنوان مثال در بررسی مطالعه پانلی پویایی درآمد[۶](PSID) امریکا در دوره‌ی اول (۱۹۶۸) نرخ بی‌پاسخی ۲۴% بود در حالی که در دوره‌ی دوم (۱۹۸۵) این نرخ به ۵۰% رسید. این مسئله معمولاً باعث اریبی براوردها می‌شود.

آمارگیری پانلی دارای مزیت‌هایی نسبت به دیگر روش‌های آمارگیری مکرر است. موزر و کالتون (۱۹۷۹) دریافتند که آمارگیری پانلی در براورد تفاضل دو یا چند پارامتر نسبت به دیگر روش‌های آمارگیری در طی زمان دقیق‌تر عمل می‌کند. در این نوع آمارگیری خطاهای غیرنمونه‌گیری مانند بی‌پاسخی را می‌توان تعدیل کرد. لاوتون و پاس (۱۹۹۶) نشان دادند که نمونه‌گیری پانلی در مقایسه با دیگر روش‌های نمونه‌گیری در طی زمان می‌تواند تا ۵۰% در هزینه‌های نمونه‌گیری در دور دوم صرفه‌جویی کند. هنشر (۱۹۸۶) این نوع آمارگیری را راهی برای دستیابی به یک بانک اطلاعاتی به روز و کارامد معرفی کرد. دانکن و همکاران (۱۹۸۷) آمارگیری پانلی را به عنوان بهترین روش برای بررسی خصوصیت‌های اجتماعی معرفی کردند.

موزر و کالتون (۱۹۷۹) در بررسی‌های خود اظهار داشتند که انتخاب نمونه‌ی مناسب در آمارگیرهای

[۶] Panel Study of Income Dynamics

پانلی نسبت به دیگر روش‌های آمارگیری دشوارتر است، زیرا اگر واحدهای نمونه‌ای مردم یک جامعه باشند تمایلی به این‌که رفتار آن‌ها در بازه‌های زمانی مورد بررسی قرار گیرد، نداشته و نسبت به این امر واکنش نشان می‌دهند. این امر ممکن است سبب بی‌پاسخی در دوره‌های بعدی شود. از طرف دیگر در آمارگیری‌هایی که تعداد دوره‌های آن‌ها زیاد باشد سبب می‌شود که واحدهای نمونه‌ای در هر دوره نسبت به دوره‌ی قبل کم شوند و با کاهش پاسخگو رو به روست. هنشر (۱۹۸۶) اظهار داشت که نتیجه‌گیری اصلی این نوع آمارگیری تا اتمام کامل آن امکان پذیر نیست و تا اتمام کار نمی‌توان نتایج محکمی از آن استخراج کرد. یکی دیگر از ضعف‌های این آمارگیری، مشکل بودن طراحی صحیح آمارگیری و دیگر ابزار اجرای طرح است.

۵.۱ بی‌پاسخی

یکی از حالت‌هایی که داده‌ی گمشده در مجموعه داده‌های آمارگیری ایجاد می‌کند بی‌پاسخی می‌باشد. بی‌پاسخی وقتی رخ می‌دهد که واحدهای نمونه‌ای اطلاعات خواسته شده را در اختیار آمارگیر قرار نمی‌دهند یا این‌که اطلاعات آن‌ها بی‌ارتباط است. بی‌پاسخی‌ها می‌توانند به سه دلیل رخ دهند. اول؛ افراد از مشارکت در آمارگیری اجتناب می‌کنند که اجتناب از پاسخگویی ممکن است به دلیل مزاحمت‌های خصوصی یا خسته‌کننده بودن مصاحبه رخ دهد. دوم؛ به دلیل وجود واحدهای غیر قابل دسترس بی‌پاسخی رخ می‌دهد. مثلاً افرادی که در خانه نیستند یا به نقاط دیگر نقل مکان کرده‌اند. سوم؛ بی‌پاسخی وقتی رخ می‌دهد که افراد از لحاظ فیزیکی یا روانی قادر به مشارکت نباشند، همچنین مشکل زبان هم می‌تواند بی‌پاسخی ایجاد کند.

یک راه برای به دست آوردن اطلاعات بیشتر از بی‌پاسخی‌ها این است که نمونه‌ای از افرادی که پاسخ نداده‌اند، انتخاب شود و با آن‌ها مصاحبه‌ای صورت گیرد. این روش بیشتر برای واحدهایی است که پس از چند بار مراجعه در خانه نباشند؛ همچنین برای افرادی که از پاسخ دادن اجتناب کرده‌اند استفاده می‌شود. نمونه‌گیری از افرادی که پاسخ نداده‌اند یک روش پرهزینه و وقت‌گیر است، لذا اغلب از این روش استفاده نمی‌شود.

شیوه‌ی سؤال پایه‌ای یک روش کم هزینه و ساده است که می‌تواند برای واحدهایی که از پاسخگویی اجتناب کرده‌اند و همچنین برای آن واحدهایی که در دسترس نبوده‌اند، مورد استفاده قرار گیرد. اگر واحد نمونه‌ای از مشارکت اجتناب کند شانس دیگری برای انجام یک مصاحبه کامل در زمان دیگر وجود ندارد، تنها راه این است که در همان زمان مصاحبه اول جواب یک یا چند سؤال مهم را به دست آورد. برای واحدهایی که در دسترس نمی‌باشند این سؤال‌ها را می‌توان توسط تلفن یا پست الکترونیکی پرسید.

به دلیل بی‌پاسخی اندازه‌ی نمونه کوچک‌تر از اندازه‌ی از پیش تعیین شده است. این وضعیت واریانس براوردگر را افزایش می‌دهد، به خصوص وقتی که نرخ بی‌پاسخی زیاد باشد. مشکل اصلی که از بی‌پاسخی

ناشی می‌شود وجود براوردهای اریب است. بزرگی نرخ بی‌پاسخی به دلیل عواملی از جمله موضوع آمارگیری، جامعه هدف، دوره‌ی زمانی، زیاد بودن سؤال‌های پرسشنامه، کیفیت مصاحبه و ... می‌باشد. به عنوان یک قانون نرخ بی‌پاسخی در آمارگیری پانلی با دوره‌های متوالی گرداوری داده افزایش می‌یابد بنابر این مخاطره‌ی اریبی در براوردهای آماری افزایش می‌یابد.

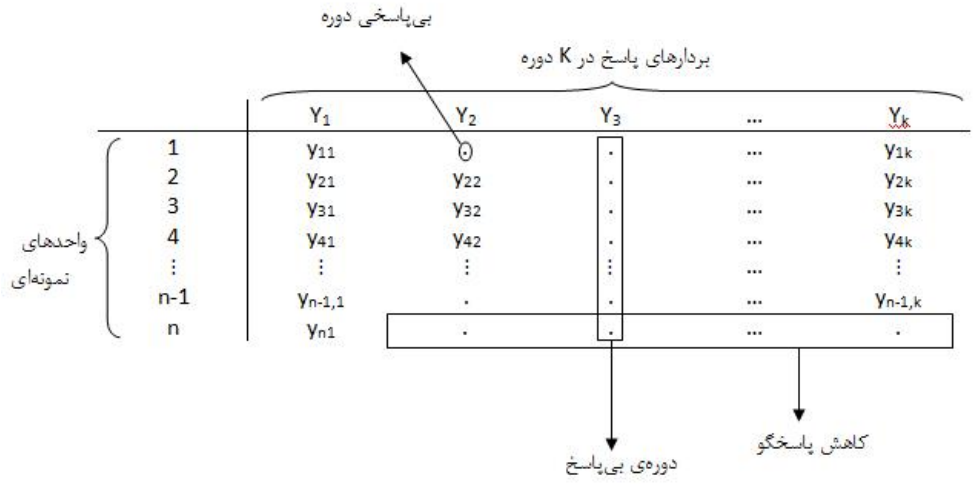
۱.۵.۱ انواع بی‌پاسخی در آمارگیری پانلی

بی‌پاسخی در آمارگیری پانلی را می‌توان به صورت زیر رده‌بندی کرد:

- بی‌پاسخی دوره،
- دوره‌ی بی‌پاسخ، و
- کاهش پاسخگو.

در آمارگیری پانلی نوعی بی‌پاسخی وجود دارد که به عنوان بی‌پاسخی دوره شناخته می‌شود به این صورت که از واحدهای نمونه‌ای در یک یا چند دوره‌ی آمارگیری و نه در تمام دوره‌ها هیچ‌گونه پاسخی دریافت نشود. دوره‌ی بی‌پاسخ به این معنی است که در مطالعه‌های پانلی واحدهای نمونه‌ای برای تعدادی از دوره‌های آمارگیری حضور دارند و برای بقیه دوره‌ها گمشده هستند، یعنی در یک یا چند دوره‌ی آمارگیری از هیچ‌یک از واحدهای نمونه‌ای اطلاعاتی دریافت نشود. مثلاً در یک دوره‌ی آمارگیری به سبب حادثه‌ی غیر مترقبه مثل زلزله، سیل یا آتش‌سوزی و ... واحدهای نمونه‌ای منطقه‌ی آمارگیری را ترک کرده باشند. کاهش پاسخگو حالت خاصی از بی‌پاسخی دوره است به این معنی که بعضی از واحدهای نمونه‌ای که در دور اول انتخاب می‌شوند از یک دوره تا انتهای مطالعه قابل دسترسی نباشند. در شکل ۱.۱ حالت‌های بی‌پاسخی دوره، دوره‌ی بی‌پاسخ و کاهش پاسخگو نشان داده شده است، که در آن $\mathbf{Y}_t$، $t = 1,2,\ldots,k$، متغیر پاسخ است که در دوره‌ی tام آمارگیری اندازه‌گیری می‌شود.

لیتل و دیوید (۱۹۸۳)، سه نوع بی‌پاسخی دوره را تشخیص دادند که شامل کاهش پاسخگو، ورود مجدد و ورود با تأخیر می‌باشند. کاهش پاسخگو که در قبل توضیح داده شد. ورود مجدد یعنی این‌که یک واحد نمونه‌ای برای یک یا چند دوره‌ی آمارگیری از مطالعه خارج می‌شود و مجدداً در دوره‌های بعد به آمارگیری بازمی‌گردد. ورود با تأخیر وقتی اتفاق می‌افتد که واحد نمونه‌ای در دور اول از مشارکت اجتناب می‌کند ولی در دوره‌های بعد به آمارگیری باز می‌گردد. اگر X نشان‌دهنده‌ی پاسخ در یک دوره باشد و O نشان‌دهنده‌ی خروج از مطالعه باشد آن‌گاه در یک پانل سه دوره‌ای وضعیت XOO و XXO کاهش پاسخگو، وضعیت XOX ورود مجدد و وضعیت OXX ورود با تأخیر را نشان می‌دهند.



شکل ۱.۱: انواع بی‌پاسخی در آمارگیری پانلی

۶.۱ نمادگذاری

فرض کنید که Y_{it} پاسخ فرد iام در دوره‌ی tام باشد به طوری که $i = 1,2,\ldots,n$ و $t = 1,2,\ldots,k$، این مجموعه داده‌ها را مجموعه داده‌های طولی (پانلی) می‌نامند. داده‌های طولی (که حداقل شامل دو مشاهده باشد) برای هر فرد تشکیل یک خوشه می‌دهند. این خوشه را می‌توان برای واحد نمونه‌ای iام به صورت یک بردار $k \times 1$، $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik})'$ نشان داد.

۷.۱ الگوی داده‌های گمشده

بیشتر مجموعه داده‌ها شکل مستطیلی یا ماتریسی دارند که در آن سطرها نشان‌دهنده واحدهای نمونه‌ای و ستون‌ها نمایانگر متغیرها یا قلم‌های اطلاعاتی هستند. الگوی داده‌های گمشده به چندین رده تقسیم می‌شود:

در شکل۲.۱ (الف) مقادیر گمشده برای Y که متغیر پاسخ است، وجود دارد؛ اما مجموعه‌ی متغیرهای کمکی $X_1, X_2, \ldots, X_p$ کاملاً مشاهده شده‌اند. این وضعیت الگوی یک متغیره نامیده می‌شود یعنی این‌که Y یک قلم اطلاعاتی است که برای بعضی واحدها مشاهده نمی‌شود.

در شکل ۲.۱ (ب) قلم‌های اطلاعاتی $Y_1, \ldots, Y_p$ این‌گونه هستند که اگر Y_j (مقادیر گرداوری شده در jامین دوره) برای یک واحد نمونه‌ای گمشده باشد پس $Y_{j+1}, \ldots, Y_p$ نیز برای آن واحد گمشده است این حالت الگوی یکنواخت نامیده می‌شود. در مطالعات طولی این حالت را کاهش پاسخگو گویند.

شکل ۲.۱ (ج) نشان‌دهنده‌ی الگوی اختیاری می‌باشد که متغیرها برای هر واحد دلخواه ممکن است گمشده باشند. در این پایان‌نامه، الگوی داده‌های گمشده با توجه به کاهش پاسخگو به صورت یکنواخت می‌باشد.