امکان سنجی به کارگیری روش نوین جانهی در سرشماری عمومی نفوس و مسکن

مجری طرح: آسیه عباسی

حسن رنجی

فریباالسادات بنی‌هاشمی

نعیمه آبی

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

زمستان ۱۳۹۰

به نام خداوند جان و خرد

پیش‌گفتار

از آن‌جا که وجود خطا در داده‌های حاصل از آمارگیری‌ها امری اجتناب‌ناپذیر است و این خطاها به‌صورت اطلاعات گزارش‌نشده، اطلاعات غیرمجاز، اطلاعات ناسازگار، و ... بروز می‌کنند، بنابراین شناسایی این خطاها و اصلاح آن‌ها به بهترین شیوه‌ی جانهی امری ضروری است. نتیجه‌ی جانهی مناسب، دستیابی به مجموعه داده‌هایی حاوی رکوردهای قابل قبول و سازگار برای تمامی اقلام اطلاعاتی ناسازگار است.

روش‌های مختلفی برای جانهی وجود دارد که بسته به نوع آمارگیری، داده‌های حاصل از آن و ... مورد استفاده قرار می‌گیرند. روش نوین جانهی که حالت خاصی و مبتکرانه‌ای از روش جانهی نزدیک‌ترین همسایه (NIM) است بیشتر برای متغیرهایی که دارای ساختار به هم پیوسته باشد، مورد استفاده قرار می‌گیرد. این روش که نخستین بار در سرشماری ۱۹۹۶ کانادا استفاده شد، هم‌اکنون متداول‌ترین روش مورد استفاده برای جانهی داده‌های سرشماری‌هاست که در ایران، برای اولین بار در سرشماری عمومی نفوس و مسکن ۱۳۹۰ استفاده شده است. با توجه به فقدان پژوهشی جامع در زمینه‌ی امکان استفاده از روش نوین جانهی در سرشماری عمومی نفوس و مسکن، پژوهشکده آمار در سال ۱۳۹۰، طرح پژوهشی «امکان‌سنجی به کارگیری روش نوین جانهی در سرشماری عمومی نفوس و مسکن ۱۳۹۰» را در دستور کار خود قرار داد که در آن با استفاده از داده‌های آزمایش سرشماری ۱۳۸۹ به مقایسه روش جانهی NIM و روش «اگر-آن‌گاه» که روش مورد استفاده در اغلب آمارگیری‌های مرکز آمار ایران است، پرداخته شده است. مجری این طرح خانم آسیه عباسی و اعضای گروه آقای حسن رنجی و خانم‌ها فریبا سادات بنی‌هاشمی و نعیمه آبی از مرکز آمار ایران بوده‌اند که بدین‌وسیله از زحمات یکایک این افراد تشکر و قدردانی می‌شود. داوری این طرح به عهده‌ی سرکار خانم‌ها دکتر زهرا رضایی قهرودی و زهره فلاح محسن‌خانی اعضای هیات علمی پژوهشکده‌ی آمار بوده است که بدین وسیله از نظرات و راهنمایی‌های آنان سپاسگزاری می‌شود. نشریه حاضر، حاصل تلاش گروه مطالعاتی فوق می‌باشد که امید است مورد توجه و استفاده‌ی دست‌اندکاران ذیربط قرار گیرد.

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

فهرست مطالب

فصل اول: کلیات

فصل دوم: روش جانهی NIM

فصل سوم: کار کاربردی

فصل اول

کلیات

۱

۱-۱ مقدمه

هدف از ویرایش داده‌های آمارگیری‌ها، کشف از قلم‌افتادگی‌ها و ناسازگاری‌ها در رکوردها است و جانهی برای اصلاح آن‌ها به‌کار می‌رود. در تمام سرشماری‌ها و آمارگیری‌ها به دلایلی مانند اطلاعات گزارش‌نشده، اطلاعات غیرمجاز و اطلاعات ناسازگار، رکوردهای دارای خطا وجود دارد. این خطاها می‌تواند به دلیل اشتباه مأمور آمارگیر، پاسخ‌گو، نحوه‌ی ورود اطلاعات و ... رخ دهد. اما به هر دلیل که خطا رخ دهد باید در گام اول آن‌ها را شناسایی و در گام بعدی به بهترین شیوه، اقدام به اصلاح خطاها کرد.

اصلاح دستی نتایج سرشماری ممکن است ماه‌ها یا حتی سال‌ها طول بکشد و در آن امکان خطاهای انسانی نیز وجود دارد. هنگامی‌که در سرشماری‌ها و آمارگیری‌ها حجم عظیمی از اطلاعات گرداوری می‌شود، کارکنان برای تصحیح اشتباهات همیشه نمی‌توانند به مدارک اصلی رجوع کنند. حتی در صورت در دسترس بودن پرسش‌نامه‌ها، اطلاعات ثبت‌شده ممکن است دارای خطا باشد در حالی‌که در ویرایش ماشینی امکان تصحیح یا تغییر بلافاصله‌ی داده‌های نادرست و تهیه‌ی گزارش برای تمامی خطاهای پیدا شده و تغییرات صورت گرفته وجود دارد.

تا همین اواخر تنها راه اصلاح داده‌ها تغییر دادن اشتباهات به صورت دستی بوده است. اگر مجموعه داده‌ها حجم کمی داشته باشد، شیوه‌های اصلاح دستی در بسیاری موارد کارایی خواهد داشت. مزایای این روش آن است که اگر اطلاعات ثبت‌شده در پرسش‌نامه‌ها کامل و درست باشد و با نگاه به آن‌ها بتوان ناسازگاری‌ها را رفع کرد، کیفیت نتایج سرشماری یا آمارگیری اندکی بهبود می‌یابد. در واقع اصلاح و جانهی داده‌ها به ندرت کیفیت داده‌های گرداوری شده را بهتر می‌کند بلکه تنها داده‌های دارای خطا را تغییر می‌دهد.

وقتی پاسخ‌گو اطلاعی را به هر دلیل پاسخ نمی‌دهد، اصلاح دستی کارایی ندارد و بررسی پرسش‌نامه مشکلی را حل نمی‌کند و در این حالت اختصاص دادن آن متغیر به طبقه‌ی اظهار نشده‌ها مفیدتر است. روش‌های اصلاح بسته به قلم اطلاعاتی متفاوت خواهند بود. در بعضی از موارد می‌توان از سایر اطلاعات مرتبط فرد کمک گرفت و در بعضی دیگر از اطلاعات سایر رکوردها استفاده می‌شود. اصلاح داده‌ها به هر روشی که صورت گیرد باید چند نکته را در درون خود رعایت کند:

- کمترین تغییرات ممکن در داده‌های ثبت‌شده صورت گیرد،

- ناسازگاری‌های اطلاعات رفع شود، و
- برای اقلام نادرست یا گم‌شده با استفاده از سایر اطلاعات مقدار مجازی درج شود در غیر این‌صورت طبقه‌ی اظهار نشده تعریف شود.

نتیجه‌ی جانهی، یک مجموعه‌ی داده‌های حاوی رکوردهای قابل قبول و سازگار برای تمامی اقلام اطلاعاتی است.

۱-۲ تعریف جانهی و لزوم به‌کارگیری آن در آمارگیری‌ها

هر قدر که پرسشنامه‌ی آمارگیری خوب طراحی شده باشد و هر قدر روش‌های به‌کار گرفته شده در گردآوری داده‌ها کارا باشد، وجود بی‌پاسخی مقدارها در هر آمارگیری امری اجتناب‌ناپذیر است. مثال‌هایی از موارد بی‌پاسخی عبارت‌اند از عدم تماس با پاسخ‌گو و امتناع پاسخ‌گو از پاسخ دادن به پرسش‌ها که به این موارد بی‌پاسخی واحد آماری گفته می‌شود. مثال‌های دیگری از قبیل امتناع پاسخ‌گو از پاسخ دادن به برخی پرسش‌ها، قادر نبودن وی در پاسخ دادن به بعضی پرسش‌ها، موفق نبودن پرسش‌گر در پرسیدن پرسش یا ثبت نادرست پاسخ یا ناسازگار بودن پاسخ در مرحله‌ی ویرایش داده‌ها و ... وجود دارند که این قبیل موارد را بی‌پاسخی پرسش یا قلم آماری گویند.

کلی‌ترین تعریفی که از جانهی می‌توان ارایه کرد «اختصاص داده به یک فیلد به دلیل بی‌پاسخی یا برای تعویض یک داده‌ی ثبت‌شده که بر اساس مجموعه‌ای از قواعد ویرایشی، ناسازگار تشخیص داده شده است»، می‌باشد (پیرچالا، ۱۹۹۵).

هدف از جانهی را می‌توان در دو مورد زیر خلاصه کرد (رابین، ۱۹۹۶):

- دادن امکان به کاربران نهایی داده‌ها با دستورالعمل‌ها و خروجی‌های استاندارد در استفاده از ابزار تحلیلی موجود برای هر مجموعه‌ای از داده‌ها که شامل داده‌های گم‌شده نیز می‌باشند.
- ارائه‌ی استنباط‌های آماری معتبر برای داده‌های شامل بی‌پاسخی از طریق به‌کارگیری روش‌های جانهی.

در طی سال‌های متمادی روش‌های متعددی برای جانهی مقدارها گم‌شده پیشنهاد شده‌اند و نرم‌افزارهای مربوط تهیه شده و توسعه پیدا کرده‌اند. هر یک از این روش‌ها در شرایط خاص خود خوب عمل می‌کنند. روش‌های مورد استفاده ضمن این که تابع عوامل مختلفی است به نوع بی‌پاسخی نیز مربوط است.

از عواقبی که با وجود بی‌پاسخی (داده‌های گم‌شده)، ممکن است در تحلیل‌های آماری رخ دهد، می‌توان به موارد ذیل اشاره کرد:

- کاهش کارایی استنباط‌های آماری
- پیچیدگی در پردازش و تحلیل داده‌ها
- ایجاد اریبی ناشی از اختلاف بین داده‌های مشاهده نشده و داده‌های مشاهده شده.

با توجه به اثرات سوء بی‌پاسخی در برآوردهای یک آمارگیری و عدم امکان پرهیز از آن، یکی از راه‌های مقابله با این مشکل، جانهی داده‌های گم‌شده است.

۱-۳ انواع خطا و روش‌های برخورد با آن

قبل از این‌که روش‌های مختلف جانهی را توضیح دهیم، لازم است انواع خطا و انواع بی‌پاسخی که یکی از متداول‌ترین خطاها در آمارگیری‌هاست بیان شود. زیرا بی‌پاسخی‌ها و خطاهای متفاوت منجر به استفاده از روش‌های جانهی متفاوت می‌شود.

۱-۳-۱ انواع خطا

کیش (۱۹۶۵) خطای کل یک آمارگیری را به دو دسته‌ی خطای نمونه‌گیری و خطای غیر نمونه‌گیری تقسیم‌بندی کرد. خطای نمونه‌گیری از به‌کارگیری داده‌های مربوط به بخشی از واحدهای جامعه‌ی آماری (به‌عنوان نمونه) برای استنباط در مورد کل جامعه‌ی مورد بررسی ناشی می‌شود که در سرشماری مقدار این خطا صفر است.
خطای غیر نمونه‌گیری، خطایی است که به نوع آمارگیری اعم از سرشماری یا آمارگیری نمونه‌ای بستگی ندارد و به‌دلیل اشتباهات یا کمبودهای سیستم، هنگام طراحی، گرداوری و پردازش داده‌ها روی می‌دهد. این خطا ممکن است در هر یک از مراحل آمارگیری رخ دهد.
خطای غیر نمونه‌گیری را می‌توان به خطاهای زیر تقسیم کرد:

- **خطای تشخیص:** هنگامی رخ می‌دهد که مفهوم به‌کار رفته در پرسش‌های آمارگیری با مفهومی که باید اندازه‌گیری شود اختلاف داشته باشد. با بروز این خطا، استنباط در مورد براورد مورد نظر به‌درستی صورت نمی‌گیرد. دلیل بروز خطای تشخیص، ارتباط ضعیف بین طراح پرسش‌نامه با کاربر، تحلیل‌گر یا مجری آمارگیری است.

- **خطای پوشش یا چارچوب:** در آمارگیری‌ها، مجموعه‌ی واحدهایی که قرار است مطالعه بر روی آن‌ها صورت گیرد جامعه‌ی هدف را تشکیل می‌دهند و مجموعه‌ای از اعضای جامعه‌ی هدف که شانس انتخاب شدن در نمونه را دارند، جامعه ی چارچوب نام دارد (گراوز و همکاران، ۲۰۰۴). خطای پوشش یا چارچوب از تفاوت بین جامعه‌ی هدف و جامعه‌ی چارچوب ناشی می‌شود.
- **خطای اندازه‌گیری:** تفاوت بین مقدار واقعی و نامعلوم پاسخ و مقدار ثبت شده برای آن است که در مرحله‌ی جمع‌آوری داده‌ها رخ می‌دهد. خطای اندازه‌گیری از چهار منبع اولیه ناشی می‌شود:
  - وسیله‌ی گرداوری اطلاعات و کیفیت آن (برای مثال، طراحی پرسش‌نامه و ادبیات آن)
  - روش گرداوری داده‌ها (مصاحبه‌ی رودررو، مصاحبه‌ی تلفنی، پستی و ...)
  - آمارگیر یا پرسشگر به‌عنوان مطرح‌کننده‌ی پرسش‌ها و درخواست‌ها
  - پاسخ‌گو به‌عنوان عرضه‌کننده‌ی اطلاعات درخواست شده
- **خطای پردازش:** ناشی از اشتباهات رخ داده در مراحل مختلف پردازش داده‌ها و محاسبه‌ی براوردها است. (مراحل پردازش داده‌ها شامل فعالیت‌هایی مانند ورود داده‌ها، کدگذاری، ادیت، جانهی و وزن‌دهی است)
- **خطای فرض‌های مدل:** ناشی از اشتباه در شیوه‌ها و مدل‌های مورد استفاده برای براورد پارامترها در آمارگیری است. خطاهای ناشی از فرضیات مدل هنگام انتخاب روش‌هایی از قبیل به‌کارگیری متغیرهای کمکی برای براوردهای نسبتی یا تعدیل‌های فصلی رخ می‌دهد.
- **خطای انتشار:** از مشکلات موجود در انتشار نتایج آمارگیری ناشی می‌شود مانند خطای حاصل از اشتباه در تنظیم جداول انتشاراتی و اشتباه در تایپ اطلاعات.

همان‌طور که از تعریف این خطاها مشخص می‌شود، خطا می‌تواند از طراحی پرسشنامه شروع شده و تا زمانی که داده‌ها به صورت جداول استخراج می‌شوند، به اشکال مختلف رخ دهد. بخشی از این خطاها را در فرایند سرشماری می‌توان کنترل کرد و برخی دیگر پس از تهیه فایل خام داده‌های حاصل، ایجاد می‌شود. خطاهای موجود در فایل داده‌ها را با توجه به اعدادی که در متغیرها ثبت می‌شوند نیز می‌توان به شرح زیر تقسیم‌بندی کرد: (راهنمای آمارگیری در کشورهای در حال توسعه، ۲۰۰۵)

- **کدهای غیرمجاز:** به عنوان مثال جنس افراد عدد ۱ (مرد) یا ۲ (زن) می‌تواند بگیرد. اعدادی غیر از ۱ و ۲ و سایر کاراکترها، کد غیرمجاز محسوب می‌شوند.

- **پرش‌های غیرمجاز:** به عنوان مثال، برای افراد کمتر از ۶ سال، وضع سواد تکمیل نمی‌شود بنا بر این اگر برای یک فرد ۵ ساله این متغیر دارای هر اطلاعی حتی اطلاع مجاز باشد، خطاست. در حقیقت باید از این متغیرها پرش کند یا این‌که اگر شماره ردیف مادر برای فردی تکمیل می‌شود که مادر وی عضو خانوار نیست، خطاست. عکس این حالت هم ممکن است رخ دهد مثلا فردی که ۱۰ ساله است نمی‌تواند از متغیر وضع فعالیت پرش کند.
- **خطای دامنه:** هرگاه اعداد یک متغیر از یک دامنه که از قبل معرفی می‌شوند، خارج شوند. برای متغیرهای رسته‌ای می‌توان خطای دامنه را با کدهای غیرمجاز ترکیب کرد و در یک گروه قرار داد. خطای دامنه برای متغیرهای پیوسته مانند سن افراد، درامد خانوار و یا تعداد اتاق‌های یک واحد مسکونی تعریف می‌شود.
- **ناسازگاری:** هرگاه اعداد متغیرهای یک فرد با یکدیگر سازگاری نداشته باشند مثلا، فرد همسر سرپرست است اما وضعیت تأهل وی، هرگز ازدواج نکرده باشد. به این حالت که ناسازگاری بین اطلاعات یک فرد با خودش بررسی می‌شود به آن، ناسازگاری داخل رکوردی می‌گویند و اگر اطلاعات فرد با سایر افراد در خانوار ناسازگاری داشته باشد به آن ناسازگاری بین‌رکوردی می‌گویند. مثلاً یک زوج هم‌جنس باشند و یا این‌که اختلاف سنی مادر و فرزند کمتر از ۱۰ سال باشد.
- **داده‌های پرت و نقاط دورافتاده:** این‌ها اعدادی هستند که در محدوده‌ی مجاز قرار دارند اما به دلیل تفاوت فاحشی که با سایر اعداد دارند ممکن است اشتباه باشند. مثلا اگر تعداد فرزندان به‌دنیا آمده برای یک فرد ۲۵ باشد و بدتر این‌که هر ۲۵ فرزند نیز مرده باشند دور از ذهن است.

از دیدگاهی دیگر می‌توان گفت فایل‌های داده‌های خام در سرشماری، حاوی خطاهای گوناگونی هستند که به دو گروه کلی تقسیم می‌شوند (توصیه‌های سازمان ملل، ۲۰۰۱):

- خطاهایی که مانع پیشروی در پردازش اطلاعات می‌شوند.
- خطاهایی که نتایج غیر قابل اطمینان یا ناسازگار تولید می‌کنند، بدون آنکه باعث ایجاد وقفه در جریان منطقی عملیات بعدی پردازش اطلاعات شوند.

۱-۳-۲ انواع بی‌پاسخی

کوکران (۱۹۷۷) انواع بی‌پاسخی را در چهار گروه به شرح ذیل تقسیم‌بندی کرده است:

۱. عدم موفقیت در یافتن واحد آماری و یا تماس با آن
۲. عدم دسترسی به افراد یا اشخاص نمونه در منزل
۳. عدم توانایی پاسخ‌گو به پاسخ دادن به بعضی یا تمامی پرسش‌های پرسشنامه
۴. امتناع از پاسخ‌گویی به تمامی یا بعضی پرسش‌ها

البته از جهت دیگر نیز می‌توان انواع بی‌پاسخی را طبقه‌بندی کرد:

۱. بی‌پاسخی واحد آماری،
۲. بی‌پاسخی قلم یا بعضی از اقلام (نه همه‌ی اقلام پرسشنامه) آماری.

بی‌پاسخی زمانی رخ می‌دهد که یک واحد از واحدهای نمونه جامعه آماری، به تمام یا قسمتی از پرسش‌های پرسشنامه‌ی آمارگیری پاسخ ندهد و این خود یکی از منابع بالقوه در ایجاد خطا در آمارگیری است که در سال‌های اخیر بسیار مورد توجه قرار گرفته است به گونه‌ای که نرخ بی‌پاسخی در برخی آمارگیری‌ها به علت پژوهش‌های گسترده در بررسی علل بی پاسخی و معرفی راهکردهای مواجهه با آن رو به کاهش است (گراوز و همکاران، ۲۰۰۲).

دلایل بی‌پاسخی را می‌توان به صورت زیر برشمرد:

- **موضوع آمارگیری:** در صورتی که موضوعات مطرح در آمارگیری مورد علاقه‌ی پاسخ‌گو باشند می‌توان افزایش نرخ پاسخ‌گوئی و در نتیجه کاهش نرخ بی‌پاسخی را انتظار داشت. در حالت کلی به پرسش‌های عمومی در مقایسه با پرسش‌های شخصی و خصوصی بیشتر پاسخ داده می‌شود.
- **نوع مصاحبه:** نوع مصاحبه نیز از جمله موارد قابل توجه است که روی درصد بی‌پاسخی تاثیر دارد. مصاحبه‌های حضوری با استفاده از مصاحبه‌گرهای آموزش دیده، معمولاً نسبت به آمارگیری‌های پستی، تلفنی و یا پیام‌نگارها (Email) از میزان پاسخ‌گویی بیشتری برخوردار می‌باشند. در مصاحبه‌های از نوع تلفنی و یا پستی انتظار می‌رود نرخ بی‌پاسخی واحد آماری افزایش یابد.
- **فرد پاسخ‌گو:** در برخی آمارگیری‌های خانواری، فرد پاسخ‌گو سرپرست خانوار و یا فرد عموما مطلع از خانوار است. در این آمارگیری‌ها، نرخ بی‌پاسخی قلم آماری کم است و برعکس انتظار می‌رود نرخ بی‌پاسخی واحد آماری

بیشتر باشد. در حالی‌که در برخی آمارگیری‌های دیگر از میان همه افراد واجد شرایط برای پاسخ‌گویی، یک فرد پاسخ‌گو به صورت تصادفی برگزیده می‌شود. در این صورت ممکن است نرخ بی‌پاسخی قلم آماری به علت اطلاعات کم فرد انتخاب شده، افزایش یابد اما انتظار می‌رود نرخ بی‌پاسخی واحد آماری کاهش یابد.

- **خطا در خواندن داده‌ها:** ممکن است فرد یا ماشین وارد کننده‌ی داده‌ها، قلم یا واحدی را جا بیندازد.

۱-۳-۳ روش‌های برخورد با داده‌های دارای خطا

روش‌های تحلیل داده‌ها با مقدارهای گم‌شده (بی‌پاسخی قلم) یا ناسازگار را می‌توان به ۳ دسته‌ی کلی تقسیم کرد.

- **روش‌های مبتنی بر رکوردهای به‌طور کامل ثبت‌شده**

اگر برخی متغیرهای مربوط به یک یا چند واحد آماری ثبت نشده باشند، روش ساده آن است که کلیه‌ی اطلاعات مربوط به آن واحد و یا واحدهای آماری حذف و تحلیل تنها روی مشاهدات آن واحدهای آماری صورت پذیرد که اطلاعات تمامی متغیرها برای آن‌ها ثبت شده باشد. این روش که معمولاً از لحاظ اجرایی ساده است، در صورتی می‌تواند نتایج رضایت‌بخش ارائه دهد که تعداد واحدهای دارای موارد گم‌شده کم باشند (نای و همکاران، ۱۹۷۵). اما این روش موجب اریبی جدی می‌شود و معمولاً خیلی کارا نیست، مخصوصاً آن‌که اگر نتیجه‌گیری آماری بخواهد در سطوح کوچکتر صورت پذیرد.

- **روش‌های مدل‌مبنا**

تعداد زیادی از روش‌های جانهی به‌وسیله‌ی تعریف مدلی برای داده‌های مشاهده‌شده و پایه‌گذاری استنباط‌ها روی درستنمایی یا توزیع پسین تحت آن مدل، انجام می‌شوند. در این روش‌ها پارامترها با استفاده از روش‌هایی مثل ماکسیمم درستنمایی براورد می‌شوند. برخی از مزایای این روش‌ها عبارت‌اند از: انعطاف‌پذیری، پرهیز از روش‌های موردی و خاص، امکان نمایش و ارزیابی روش‌های حاصل که مبتنی بر فرض‌های اساسی مدل هستند و در دسترس بودن براوردهای واریانس که نواقص موجود در داده‌ها در آن‌ها لحاظ می‌شود.

- **روش‌های مبتنی بر جانهی**

در این روش‌ها مقدارهای بی‌پاسخ، جانهی و تکمیل می‌شوند و داده‌های کامل شده، از طریق روش‌های متعارف، تحلیل می‌شوند. روش‌های جانهی هر کدام با توجه به نوع آمارگیری، داده‌های کمکی موجود و میزان دقت آن‌ها، روش تحلیل داده‌ها، تصادفی بودن خطاها یا گم‌شدگی، تعداد نمونه‌ها و ... در آمارگیری‌های مختلف به‌کار می‌روند.

روش‌های اول و دوم، برای سرشماری‌ها مناسب نیستند زیرا اطلاعات هیچ‌یک از افراد را در سرشماری نمی‌توان حذف کرد، این خلاف اصول سرشماری است. در روش دوم، معمولا فرض بر این است که برای جانهی یک قلم سایر اقلام اطلاعاتی کمکی صحیح و معتبر هستند در حالی‌که نمی‌توان این اطمینان را در مورد داده‌های حاصل از آمارگیری یا سرشماری داشت. بنا بر این شاید بتوان گفت مطمئن‌ترین روش برای برخورد با داده‌های گم‌شده یا دارای خطا استفاده از روش‌های جانهی است.

روش‌های جانهی از دید کلی به دو دسته تقسیم می‌شوند:

۱- **جانهی قطعی:** که در آن هر بار (در جانهی ساده یک بار و در جانهی چندگانه چند بار) تنها یک مقدار ثابت برای پرسش بی‌پاسخ جانهی می‌شود که به دو دسته‌ی جانهی قطعی ساده و مدل‌مبنا تقسیم می‌شود.

۲- **جانهی تصادفی:** که در آن برای هر یک از موارد بی‌پاسخی مقدارهای تصادفی از روی مقدارها مشاهده شده یا از یک توزیع پیش‌بینی شده جانهی می‌شوند که به دو دسته‌ی جانهی تصادفی ساده و مدل مبنا تقسیم می‌شود.

در زیر توضیحات بیشتری راجع به این گروه‌ها داده می‌شود و مهم‌ترین روش‌های جانهی هر گروه معرفی می‌شوند.

۱-۳-۳-۱ جانهی قطعی ساده

جانهی قطعی ساده از چند روش تشکیل شده است که در این قبیل روش‌ها معمولاً توزیع داده‌ها به‌هم می‌خورد و منجر به کم‌براورد شدن واریانس می‌گردد (بجز روش جانهی استنتاجی که در زیر توضیح داده می‌شود). با این وجود این روش‌ها، بدلیل سادگی‌شان هنوز در عمل بسیار مورد استفاده قرار می‌گیرند. معروف‌ترین روش‌های این گروه عبارت‌اند از: جانهی استنتاجی یا بادرنگ، جانهی بی‌درنگ قطعی، جانهی بی‌درنگ سنتی، جانهی جورشدگی چندمتغیره، جانهی جورشدگی تابع فاصله و جانهی با میانگین. این روش‌ها را به طور مختصر در ادامه شرح می‌دهیم.

- **جانهی استنتاجی یا بادرنگ**

در بعضی موارد بی‌پاسخی، ارتباط شناخته شده‌ی منطقی بین متغیرهای بدون پاسخ و سایر متغیرها که دارای پاسخ هستند وجود دارد. در روش‌های جانهی استنتاجی با استفاده از متغیرهای دارای پاسخ و با قاطعیت بالا جواب پرسش‌های بی‌پاسخ جانهی می‌شوند.

در این روش مقدارهای گم‌شده از اطلاعات موجود اقلام مشابه از آمارگیری‌های قبلی و یا اقلام مرتبط از آمارگیری جاری و غیره جایگزین می‌شوند. جانهی بادرنگ را می‌توان از این نوع دانست که در آن از اطلاعات آمارگیری‌های قبلی برای جانهی پرسش‌های بی‌پاسخ استفاده می‌شود. معمولاً غیر ممکن است که بتوان به کمک روش‌های استنتاجی، اطلاعات کافی برای جانهی تمامی پرسش‌های بی‌پاسخ پیدا کرد. بلکه از این روش‌ها می‌توان برای جانهی بعضی پرسش‌های بی‌پاسخ استفاده کرد. در شرایطی که امکان‌پذیر باشد باید از این روش‌ها در مقایسه با سایر روش‌ها، برای جانهی استفاده کرد چرا که مقدارهای تقریباً دقیقی برای داده‌های گم‌شده ارائه می‌دهد. باید توجه داشت که توان به‌کارگیری این روش‌ها به منابع موجود بستگی دارد. کاربرد عمده این روش در آمارگیری‌های دوره‌ای است.

- **جانهی بی‌درنگ قطعی**

این روش به دلیل سادگی و دارا بودن معنا و مفهوم برای افرادی که در کارهای آمارگیری مشارکت دارند، اما زمینه‌ی آماری قوی ندارند، یکی از معروف‌ترین انواع روش‌های جانهی است که در آن از هیچ مدل آماری صریحی استفاده نمی‌شود. عیب عمده‌ی این روش آن است که نمی‌تواند مقدارهای مشخصه برای افرادی را پوشش دهد که دارای خصیصه‌های معینی بوده و هیچ یک از افراد دارنده‌ی این خصیصه به پرسش مورد نظر پاسخ نداده باشند. در این قبیل جانهی روش‌های متعددی به‌کار گرفته می‌شوند که در زیر معروف‌ترین آن‌ها شرح داده می‌شود. یکی از تفاوت‌های عمده‌ی این روش‌ها با روش‌های جانهی بادرنگ در استفاده‌ی این روش از اطلاعات طرح آمارگیری جاری است تا از آمارگیری‌های قبلی.

بعضی از معروف‌ترین روش‌های جانهی بی‌درنگ عبارت‌اند از:

- **جانهی بی‌درنگ سنتی**

این روش دارای دو مرحله است. مرحله‌ی اول استفاده از بعضی متغیرهای کمکی معمولاً از نوع رسته‌ای برای تعیین رده‌های جانهی است. در مرحله‌ی دوم در داخل هر رده تنها یک مقدار مثل میانگین رده یا مقدارهای از

قبل تعیین شده به عنوان نقطه شروع در نظر گرفته می‌شود. سپس رکوردهای داخل فایل داده‌ها به صورت دنباله‌ای مورد بررسی قرار می‌گیرند. اگر رکوردی دارای پاسخ برای متغیر مورد نظر باشد این مقدار جایگزین مقدار ذخیره شده‌ی قبلی برای رده‌ی جانهی می‌شود. در غیر این صورت برای آن رکورد مقداری که در آن زمان به عنوان جانهی در آن کلاس ذخیره شده است، برای آن جانهی می‌شود.

جذابیت عمده‌ی این روش در صرفه‌ی محاسباتی آن است، چون که تمامی جانهی‌ها با یک ملاحظه در سراسر فایل داده‌ها انجام می‌شود. یکی از معایب این روش آن است که می‌تواند به راحتی منجر به استفاده چند باره از یک مقدار جانهی شود که کاهش دقت براوردگرهای آمارگیری را به‌دنبال خواهد داشت (کالتن و کاسپرزیک، ۱۹۸۲).

- **جانهی جورشدگی چندمتغیره**

در این روش دهندگان و گیرندگان اطلاعات بر اساس چندین متغیر کمکی از قبل تعیین شده جور می‌شوند. برای هر مورد گم‌شده در یک رده‌ی جور شده، نزدیک‌ترین دهنده‌ی اطلاعات برای جانهی مورد استفاده قرار می‌گیرد. اگر در رده‌ای دهنده‌ی اطلاعات موجود نباشد آن رده با رده‌های دیگر ادغام می‌شود به طوری که در رده‌ی ادغامی، دهندگان اطلاعات وجود داشته باشد.

گرچه استفاده از این روش به کمک برنامه‌های کامپیوتری ساده نیست. الگوریتم‌های تقریباً معادل را می‌توان جایگزین کرد. در این روش ابتدا فایل داده‌ها به کمک متغیرهای کمکی یکسان، مرتب می‌شود و سپس نزدیکترین مقدار پاسخ برای هر مورد گم‌شده جایگزین می‌شود. این روش جایگزینی از لحاظ اجرایی خیلی ساده است. گیرندگان و دهندگان اطلاعات برای تمامی متغیرهای کمکی جور می‌شوند به شرطی که چنین دهندگان اطلاعاتی وجود داشته باشند. در غیر این‌صورت این روش به طور خودکار دهندگان اطلاعاتی را پیدا می‌کند که براساس برخی متغیرهای کمکی جور می‌شوند و این معادل آن است که رده‌های جور شده را به‌هم زده و رده‌های جور شده‌ی جدیدی ساخته می‌شود.

- **جانهی جورشدگی تابع فاصله**

یک روش بسیار عمومی، تعریف یک تابع فاصله است که برای اندازه‌گیری فاصله‌ی بین واحدها بر اساس مقدارها متغیر کمکی و سپس مقدارها جانهی شده از واحدهای پاسخ داده شده که به واحد با مقدار گم‌شده نزدیک‌اند انتخاب می‌شوند. برای مثال فرض کنید $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ مقدارها k متغیر کمکی مناسب

برای یک واحد i که y_i آن گم شده است باشد، اگر این متغیرها برای تشکیل سلول‌ها استفاده شوند، متریک زیر را داریم:

$$d(i,j)=\begin{cases}۰ & \text{در سلول‌های مشابه باشند } i,j\\ ۱ & \text{در سلول‌های متفاوت باشند } i,j\end{cases}$$

می‌توانیم توابع فاصله را به صورت زیر انتخاب کنیم

ماکسیمم انحراف : $d(i,j)=\max_k |x_{ik}-x_{jk}|$

و یا

فاصله‌ی ماهالانوبیس : $d(i,j)=(x_i-x_j)^T S_{xx}^{-۱}(x_i-x_j)$

که S_{xx} براورد ماتریس کوواریانس x است.

لزومی ندارد که متریک پررتبه باشد، مثلاً برای (i,j) وقتی $x_i=x_j$ باشد فاصله صفر است. برای مثال، متریک زیر را در نظر بگیرید:

میانگین پیش بینی شده : $d(i,j)=[\hat{y}(x_i)-\hat{y}(x_j)]^۲$

که $\hat{y}(x_i)$ مقدار پیش‌بینی شده‌ی Y از رگرسیون Y روی x است که با استفاده از حالت داده‌های کامل محاسبه شده است. مقدار جانهی شده برای y_i را از آن j هایی انتخاب می‌کنیم که برای آن‌ها:

(۱) $y_j, x_{j۱}, \ldots, x_{jk}$ مشاهده شده‌اند، و

(۲) $d(i,j)$ کمتر از مقداری مثل d_o باشد.

تغییر مقدار $d_{o.}$ می‌تواند تعداد کاندیدهای j را کنترل کند (کوکران و رابین، ۱۹۷۳؛ رابین و توماس، ۱۹۹۲).

- **جانهی با میانگین**

گرچه این روش ساده‌ترین نوع جانهی است اما شاید بتوان گفت غیر جاذب‌ترین نوع جانهی نیز می‌باشد. جانهی تمامی مقدارها بر اساس میانگین، موجب ایجاد اختلال در توزیع داده‌ها و کم براورد شدن واریانس می‌شود.

در این روش یا میانگین کل مشاهدات هر متغیر و یا میانگین‌های سلولی جایگزین پرسش‌های بی‌پاسخ می‌شود.