امکان‌سنجی به‌کارگیری روش ادیت اتوماتیک در طرح‌های آمارگیری

مجری طرح:

ندا مصطفوی

همکاران:

فریباسادات بنی‌هاشمی

آسیه عباسی

حسن رنجی

نعیمه آبی

پژوهشکده‌ی آمار

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

تابستان ۱۳۹۱

به نام خداوند جان و خرد

پیش‌گفتار

استخراج اطلاعات درست از داده‌های گردآوری شده در سرشماری‌ها یا آمارگیری‌های نمونه‌ای به دلیل کاربرد وسیع آن‌ها در برنامه‌ریزی‌های اجتماعی و اقتصادی و بررسی‌های علمی از اهمیت زیادی برخوردار است. از آنجایی‌که درستی داده‌ها مستقیماً بر کیفیت اطلاعات حاصل از آن‌ها تأثیرگذار است، اطمینان از درستی داده‌های گردآوری‌شده پیش از استخراج اطلاعات امری ضروری است. در اغلب سرشماری‌ها و آمارگیرهای نمونه‌ای، ادیت و جانهی داده‌ها با هدف شناسایی و اصلاح خطاهای احتمالی موجود در داده‌های گردآوری‌شده انجام می‌گیرد تا مجموعه داده‌هایی با کیفیت مطلوب برای استخراج اطلاعات درست فراهم شود.

پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی با هدف افزایش کیفیت آمارها، اجرای طرح پژوهشی «امکان‌سنجی به‌کارگیری روش ادیت اتوماتیک در طرح‌های آمارگیری» را در دستور کار خود قرار داد، که اجرای آن از تیر ۱۳۹۰ آغاز شد و گزارش نهایی آن اکنون در دسترس علاقه‌مندان قرار گرفته است. این پژوهش در گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی پژوهشکده‌ی آمار با همکاری خانم ندا مصطفوی، مجری طرح، خانم فریباسادات بنی‌هاشمی، خانم آسیه عباسی، خانم نعیمه آبی و آقای حسن رنجی (از همکاران مرکز آمار ایران) به‌عنوان همکاران اصلی طرح پژوهشی به انجام رسیده است، که بدین وسیله از همکاری ایشان، صمیمانه تشکر و قدردانی می‌شود.

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

پژوهشکده‌ی آمار

فهرست

۱

کلیات

۱-۱ مقدمه

آمارگیری‌ها بخش عظیمی از اطلاعات مورد نیاز برای برنامه‌ریزی‌های یک کشور را تأمین می‌کنند. یک آمارگیری خواه سرشماری، خواه آمارگیری نمونه‌ای وقتی موفق به اتمام می‌رسد که داده‌های با کیفیت، تولید و استخراج شود، اما هرگز نمی‌توان امید داشت داده‌های یک آمارگیری، عاری از خطا باشد. مجموعه داده‌های یک آمارگیری به دلایل مختلفی مانند خطای مأمور آمارگیر، خطای داده‌آما و سایر خطاهای غیرنمونه‌گیری، اغلب شامل داده‌های گم‌شده، ناسازگار، دورافتاده و کدهای غیرمجاز است. از آنجایی که هر قلم آماری می‌تواند دارای خطا باشد، شاید بتوان گفت اطلاعات حاصل از سرشماری با خطای غیرنمونه‌گیری بیشتری نسبت به آمارگیری نمونه‌ای مواجه است، زیرا در سرشماری اقلام آماری بیشتری از پاسخگویان پرسیده می‌شود.

بنا بر این برای اطمینان از تولید داده‌های سازگار در سرشماری‌ها کمیته‌ای به نام «**کمیته‌ی ادیت و جانهی**» تشکیل می‌شود. این کمیته، یکی از اصلی‌ترین کمیته‌های تخصصی یک سرشماری است که وظیفه‌ی اصلی آن شناسایی و کشف خطاها و تصحیح آن‌ها است. این کمیته موظف است پس از گرداوری داده‌ها، در کمترین زمان ممکن، داده‌ها را به بهترین شیوه پاکسازی کند و برای استخراج اطلاعات به کمیته‌ی جداول تحویل دهد.

کمیته‌ی ادیت و جانهی علاوه بر پاکسازی داده‌ها، باید بـه کشـف خطـا و کـاهش آن در پـایین‌تـرین مراحـل آمارگیری توجه و ابزاری را برای این منظور معرفی نماید. بخشی از این ابزار، نیروهای انسـانی ماننـد بـازبین‌هـا در استان‌ها و بخشی دیگر ابزار ماشینی هستند مانند نرم‌افزار بازبینی ماشینی.

سرشماری عمومی نفوس و مسکن در ایران، از سال ۱۳۳۵ هر ده سال یک‌بار به طور رسمی انجام شده اسـت. پس از سرشماری سال ۱۳۸۵، مسئولان کشور به دلایلی تصمیم گرفتـه‌انـد کـه سرشـماری‌هـا در فواصـل زمـانی پنج‌ساله انجام شود. در همین‌راستا به منظور آزمایش جنبه‌های مختلف سرشماری ۱۳۹۰، دو سرشماری آزمایشی در ابعاد کوچک برای سال‌های ۱۳۸۸ و ۱۳۸۹ در نظر گرفته شد. کمیته‌ی ادیت و جـانهی سرشـماری ۱۳۹۰ نیـز کـار خود را با بررسی روش‌های جدید ادیت و جانهی آغاز و این روش‌هـا را بـر روی داده‌هـای سرشـماری‌هـای قبـل و سرشماری آزمایشی ۱۳۸۹ آزمون کرد.

برای ادیت و جانهی داده‌ها باید از روش‌هایی استفاده شود که بسیار سریع داده‌های با کیفیت را شناسایی کنند. به طور کلی، اهداف کمیته‌ی ادیت و جانهی با سه دیدگاه شکل گرفته است:

آ) بررسی موضوعی و کارشناسی انواع خطا در سرشماری‌های قبل.

ب) بررسی روش‌های مورد استفاده در سرشماری‌های قبل و نگاهی به روش‌های مورد استفاده در سایر کشورها در گذشته و آینده.

ج) تعیین محدوده‌ی ادیت و جانهی و همچنین انتخاب بهترین و کاراترین روش برای ادیت و جـانهی سرشـماری ۱۳۹۰.

کمیته‌ی ادیت و جانهی سرشماری ۱۳۹۰ در زمان بررسی روش‌های مورد استفاده برای پاک‌سازی داده‌هـا در سـایر سرشماری‌ها، روش‌های نوین ادیت و جانهی، به خصوص روش مورد استفاده در کشـور کانـادا را بررسـی کـرد کـه موضوع اصلی این گزارش است. نرم‌افزار طراحی شده در کشور کانادا بر اساس مقاله‌ی فلگی- هولت (۱۹۷۶) است. پس از نگارش این مقاله، اداره آمار کانادا در مطالعات سرشماری آزمایشی ۱۹۹۹ به توسعه ایـن روش پرداخـت کـه حاصل این مطالعات نرم‌افزار CANCIES است. (بنکی‌یر و دیگران، ۲۰۰۲)

نرم‌افزار CANCIES، از چند بخش کلی تشکیل شده است:

- ادیت اتوماتیک فلگی- هولت (داخل رکوردی و بین رکوردی)،
- جانهی قطعی که در آن یک سری از خطاهای سیستماتیک را با قطعیت بالا اصلاح می‌کند.
- جانهی به روش NIM برای متغیرهای جمعیت‌شناختی (بستگی با سرپرست، جنس، سن و وضع زناشویی سرپرست)

در انتها خاطر نشان می‌شود، فرایند ادیت و جانهی بخش عمده‌ای از هزینه‌های سرشماری را به خود اختصاص می‌دهد و در عین حال مسئولیت سنگینی در انجام موفقیت‌آمیز سرشماری را به دوش می‌کشد. هزینه‌ی مربوط به ادیت در آمارگیری‌های خانواری در اوایل دهه‌ی ۱۹۹۰ در حدود ۲۰ درصد بودجه‌ی کل سرشماری‌ها در سراسر دنیا بوده است (گرانکیست و کوار، ۱۹۹۷)، این رقم نشان‌دهنده‌ی حساس بودن انجام فرایند ادیت و جانهی در سرشماری‌هاست.

کمیته‌ی ادیت و جانهی، چند ملاک برای انتخاب روش‌های خود در نظر گرفته است:

- هزینه‌ی کمتر،
- دخالت کمتر نیروی انسانی،
- روش‌های سریع‌تر،
- وجود کمترین خطا در فایل داده‌های بعد از ادیت و جانهی،
- اعمال کمترین تغییرات و
- مستندسازی سوابق پس از انجام هریک از مراحل ادیت و جانهی.

در این فصل ابتدا، به تاریخچه‌ی مختصری از روش‌های ادیت در سایر کشورها و مرکز آمار ایران اشاره می‌شود. پس از آن انواع خطا و انواع ادیت به همراه مزیت‌ها و عیب‌های هر کدام معرفی می‌شوند و ضرورت بررسی ادیت اتوماتیک شرح داده می‌شود. در انتهای فصل، اهداف تحقیق را عنوان می‌کنیم.

۲-۱ پیشینه‌ی تحقیق

تاریخچه‌ی ادیت و جانهی وابسته به شروع به‌کارگیری رایانه در طرح‌های آماری است. قبل از بـه‌کـارگیری رایانـه، تعداد زیادی از کارمندان به استخدام در می‌آمدند تا فرم‌های مجزا از یکدیگر را ویرایش کنند. با این همه بـه علـت پیچیدگی ارتباطات بین اقلام پرسش‌نامه، بازبینی‌های ساده قادر به پوشش تمامی ناسازگاری‌های احتمالی در داده‌ها نبودند. هر یک از کارمندان، قواعد موجود در این زمینه را به صورت‌های مختلف تفسیر می‌کردند و این امکان وجود داشت که حتی یک فرد نیز عملکرد متفاوتی داشته باشد.

در سال‌های ابتدایی ورود رایانه، به دلیل مشکلات سرعت رایانه و محدودیت اصلاح خطاها استفاده از رایانه‌هـا برای انجام ادیت با استقبال خوبی روبرو نشد (نوردباتن، ۱۹۶۳). اما با توسعه‌ی رایانه‌ها نتایج حاصل از سرشماری به نحو چشمگیری دچار دگرگونی شد. در دهه‌ی ۱۹۸۰ با ورود رایانه‌های شخصی به اداره‌های ملی سرشماری و آمـار، پیشرفت بزرگی در زمینه ادیت و جانهی داده‌ها حاصل شد. امروزه ویرایش رایانه‌ای می‌تواند حتی همزمـان بـا ورود اطلاعات به رایانه صورت گیرد و در شیوه‌های گردآوری داده‌ها مانند CAPI[۱] می‌تـوان خطاهـای شناسـایی‌شـده را هم‌زمان با مصاحبه با خانوار اصلاح کنند.

از رایانه برای ادیت و جانهی به صورت‌های مختلفی استفاده می‌شده است. یکـی از ابتـدایی‌تـرین شـیوه‌هـای کشف خطا رسم فلوچارت و تبدیل آن‌ها به برنامه‌هایی برای کشف خطا بود. این روش تا مدت‌ها استفاده مـی‌شـد و در کشورهایی مانند ایران نیز همچنان مورد استفاده قرار می‌گیرد. در سال ۱۹۷۶ فلگی- هولت روشی جدیدی برای کشف خطا معرفی کردند که هم‌اکنون پایه و اساس روش‌های نوین ادیت و جانهی است و شاید بتوان گفت بعـد از ورود رایانه، روش فلگی- هولت نقطه‌ی عطفی برای فرایند ادیت و جانهی به شمار می‌رود. روش فلگی- هولـت در ابتدای معرفی آن مورد اقبال قرار نگرفت، بلکه کشور کانادا برای نخستین بار در سال ۱۹۹۹ بـرای انجـام مطالعـات مربوط به سرشماری خود از این شیوه کمک گرفت. از آن زمان تا کنون نرم‌افزارهـای مختلفـی بـا توسـعه‌ی روش فلگی- هولت و توسعه‌ی آن ساخته شده است.

---

[۱] Computer Assisted Personal Interviewing

در مرکز آمار ایران نیز تا کنون به‌جز ویرایش‌های چشمی و اصلاحات دستی، روش مورد استفاده بـه خصـوص در سرشماری‌ها رسم فلوچارت و تهیه برنامه کشف خطا و جانهی خطاها بوده است. در حقیقت نتیجه‌ی این گزارش نخستین تجربه‌ی مرکز آمار ایران در زمینه‌ی استفاده از شیوه‌ی فلگی– هولت است.

۱-۳ انواع خطا در آمارگیری‌ها

فعالیت‌های کمیته‌ی ادیت و جانهی زمانی معنی پیدا می‌کند که خطا رخ دهد. بخشـی از فعالیـت‌هـای کمیتـه‌هـای سرشماری در جهت جلوگیری از بروز خطا و مابقی در جهت کشف و اصلاح این خطاها است. حال ایـن سـؤال‌هـا مطرح می‌شوند که مفهوم خطا در آمارگیری‌ها چیست؟ و در کدام‌یک از مراحل آمارگیری می‌تواند رخ دهـد؟ کشـف کدام‌یک از انواع خطا به عهده‌ی کمیته‌ی ادیت و جانهی است؟ کدام‌یـک از خطاهـا را نمـی‌تـوان کنتـرل کـرد؟ و سـؤالاتی از این قبیل. در این بخش ابتدا به بررسی مفهوم خطا از دیدگاه کیش می‌پردازیم.

کیش (۱۹۶۵) خطای کل یک آمارگیری را به دو دسـته‌ی خطـای نمونـه‌گیـری و خطـای غیـر نمونـه‌گیـری تقسیم‌بندی کرد. خطای نمونه‌گیری از به‌کارگیری داده‌های مربوط به بخشی از واحدهای جامعه‌ی آماری (به‌عنـوان 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، جمع‌آوری و پردازش داده‌ها روی می‌دهد. این خطا ممکن است در هر یک از مراحل آمارگیری رخ دهد.

خطای غیر نمونه‌گیری را می‌توان به انواع زیر تقسیم کرد:

- **خطای تشخیص:** هنگامی رخ می‌دهد که مفهوم به‌کار رفته در سؤالات آمارگیری با مفهـومی کـه بایـد اندازه‌گیری شود اختلاف داشته باشد. با بروز این خطا، استنباط در مورد براورد مورد نظر به‌درسـتی صـورت نمی‌گیرد. دلیل بروز خطای تشخیص، ارتباط ضعیف بین طراح پرسش‌نامه با کاربر، تحلیـل‌گـر یـا مجـری آمارگیری است.
- **خطای پوشش یا چارچوب:** در آمارگیری‌ها، مجموعه‌ی واحدهایی که قرار است مطالعه بـر روی آن‌هـا صورت گیرد جامعه‌ی هدف را تشکیل می‌دهند و مجموعه‌ای از اعضای جامعه‌ی هدف که شانس انتخـاب

شدن در نمونه را دارند، جامعه‌ی چارچوب نام دارد (گرووز و دیگران، ۲۰۰۴). خطای پوشش یا چارچوب از تفاوت بین جامعه‌ی هدف و جامعه‌ی چارچوب ناشی می‌شود.

- **خطای بی‌پاسخی:** منبع عمده‌ی خطای غیر نمونه‌گیری است که بازتاب تلاشی ناموفق در به‌دست آوردن اطلاعات لازم از یک واحد واجد شرایط آمارگیری است. بی‌پاسخی می‌تواند هم مربوط به واحد آماری باشد و هم قلم اطلاعاتی.
- **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 گردآوری اطلاعات و کیفیت آن (برای مثال، پرسش‌نامه و ادبیات آن)
  - روش گرداوری داده‌ها (مصاحبه‌ی رودررو، مصاحبه‌ی تلفنی، پستی و ...) به‌عنوان شیوه‌ی درخواست اطلاعات
  - آمارگیر یا پرسشگر به‌عنوان مطرح‌کننده‌ی سؤالات 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 نمونه‌ای 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 به‌صورت 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 تقسیم‌بندی کرد: (راهنمای آمارگیری در کشورهای در حال توسعه)

- **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 مثلاً فردی که ۱۰ ساله است نمی‌تواند از متغیر وضع فعالیت پرش کند.
- **بی‌پاسخی:** اگر متغیری سفید باشد باید بررسی شود که آیا این سفید بودن منطقی است یا نه؟ در این قسمت می‌توان با تعریف پرش‌های منطقی، داده‌های گم‌شده را شناسایی کرد.
- **خطای دامنه**: هرگاه اعداد یک متغیر از یک دامنه که از قبل معرفی می‌شوند، خارج شوند خطای دامنه رخ می‌دهد. برای متغیرهای رسته‌ای می‌توان خطای دامنه را با کدهای غیرمجاز ترکیب کرد و در یک گروه قرار داد. خطای دامنه برای متغیرهای پیوسته مانند درامد خانوار تعریف می‌شود.
- **ناسازگاری:** هرگاه اعداد متغیرهای یک فرد با یکدیگر سازگاری نداشته باشند مثلاً، فرد همسر سرپرست است اما وضعیت زناشویی 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 دورافتاده:** این داده‌ها اعدادی هستند که در محدوده‌ی مجاز قرار دارند اما به دلیل تفاوت فاحشی که با سایر اعداد دارند ممکن است اشتباه باشند. مثلاً 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: (سازمان ملل متحد، ۲۰۰۱)

۱. خطاهایی که مانع پیشروی در پردازش اطلاعات می‌شوند. (خطاهای نوع اول)
۲. خطاهایی که نتایج غیر قابل اطمینان یا ناسازگار تولید می‌کنند، بدون آن‌که باعث ایجاد وقفه در جریان منطقی عملیات بعدی پردازش اطلاعات شوند. (خطاهای نوع دوم)

با تعریف خطا و سه نوع دسته‌بندی برای آن می‌توان به این نتیجه رسید که خطا هر چه باشد و هر کجا رخ دهد باید ابتدا شناسایی و سپس تا حد ممکن اصلاح شود. همان‌گونه که در نشریه‌ی «راهنمای بازبینی و اصلاح داده‌ها در سرشماری نفوس و مسکن (مطالعه در روش‌ها)» (سازمان ملل متحد، ۲۰۰۱) خاطر نشان گردیده است، همه‌ی خطاهای نوع اول و تا حد امکان، خطاهای نوع دوم باید اصلاح شوند.

۱-۴ انواع ادیت در آمارگیری‌ها

قبل از این‌که وارد تعریف و بیان روش‌های مختلف ادیت شویم باید به یک سؤال دیگر پاسخ دهیم. چرا لازم است داده‌ها را پاک‌سازی کنیم و یا به عبارتی چرا داده‌ها را ادیت می‌کنیم؟

جواب ساده است چون اولاً بدون ادیت داده‌ها، نتایج ناقص است و ثانیاً اعتبار اداره‌ی سرشماری در گروی ارائه‌ی داده‌های با کیفیت است. اگر داده‌های حاصل از سرشماری ادیت نشوند، مشکلاتی از قبیل نتایج ساختگی، برداشت اشتباه از داده‌ها، نبودن ضمانت کافی برای کاربران و ... رخ می‌دهند. اگر اداره‌های ملی سرشماری/ آمار نتایج حاصل از سرشماری‌ها و آمارگیری‌ها را ادیت نکنند، احتمالاً نشریه‌های سرشماری حاوی میزان قابل توجهی از داده‌های بی‌معنی خواهد بود. ادیت، براوردهای مخدوش را کاهش داده، باعث تسهیل در پردازش اطلاعات می‌شود و اطمینان کاربران را به آمارهای در دسترس افزایش می‌دهد. علاوه بر این، بنا به اظهارات پولوم و دیگران (۱۹۸۶) دستاورد اولیه‌ی حاصل از ادیت یا پاک‌سازی، کشف این نکته است که آیا پاسخ‌های گوناگون با یکدیگر و همچنین چارچوب اساسی ابزار آمارگیری سازگارند یا نه؟

با این حال، زمانی که اعتبار یک اداره‌ی ملی سرشماری/آمار در خطر است، اجرای یک سرشماری با ورودی‌های بدون نقص و بدون ناسازگاری ضرورت دارد. همان‌گونه که بنیستر (۱۹۸۰) اظهار می‌دارد، مقامات

سازمان‌های سرشماری می‌توانند مواردی از نگارش مقاله‌های طنزآمیز روزنامه‌نگاران یا نامه‌نگاری پر از خشم شهروندان به مقامات سرشماری را در مورد جداول انتشاراتی نقل کنند که در آن‌ها پدر بزرگ‌های سه‌ساله دیده می‌شوند یا افراد عازم محل کار، سوار قطارهایی می‌شوند که اصلاً وجود ندارند.

بنا بر این هدف اساسی از ادیت نتایج سرشماری در مرحله‌ی پردازش اطلاعات، شناسایی خطاها تا سر حد امکان و انجام دادن تغییرات لازم در مجموعه‌ی داده‌ها به نحوی است که اقلام اطلاعاتی معتبر و با یکدیگر سازگاری داشته باشند. با این حال مرحله‌ی پردازش اطلاعات قادر به تصحیح تمامی خطاها نخواهد بود؛ از جمله پاسخ‌هایی که ناسازگاری ندارند، اما در واقع نمونه‌هایی از جواب‌های نادرست پاسخ‌گویان یا ثبت نادرست مأموران آمارگیری در هنگام مصاحبه هستند.

حال که ضرورت انجام ادیت را متذکر شدیم، به تعریف دقیق ادیت می‌پردازیم.

ادیت، فرایند بررسی، تصحیح یا تغییر سیستماتیک (قاعده‌مند) پاسخ‌ها بر اساس قواعد از پیش تعیین شده است.

برخی از عملیات مربوط به ادیت به‌صورت دستی و به وسیله‌ی انسان و برخی دیگر به‌صورت الکترونیکی و از طریق رایانه انجام می‌شود.

ادیت به‌طور کلی، به دو گروه تقسیم می‌شود:

۱. ادیت تعیین‌کننده که خطاها را به‌صورت قطعی شناسایی می‌کند، و
۲. ادیت پرسش‌گونه که اقلام اطلاعاتی مشکوک را شناسایی می‌کند (گرانکیست و کوار، ۱۹۹۷).

به عبارت دیگر، ادیت نوع اول، اقلامی را شناسایی می‌کند که به طور حتم اشتباه هستند، در حالی که ادیت نوع دیگر اشاره به داده‌هایی دارد که ممکن است دور از ذهن و خلاف معمول باشند.

خطاهای قطعی که از طریق ادیت نوع اول کشف می‌شوند، شامل ورودی‌های غیر قابل قبول یا گم‌شده، همچنین خطاهای ناشی از ناسازگاری‌ها هستند. در مقابل، ادیت نوع دوم داده‌هایی را شناسایی می‌کند که در خارج از محدوده‌ی نظری غالب در ادیت قرار می‌گیرند؛ اقلامی که ارقام‌شان در مقایسه با سایر داده‌ها در یک پرسش‌نامه، نسبتاً بالا یا پایین است، یا سایر ورودی‌های مشکوک. به منظور ایجاد اطمینان از صحت نتایج، به‌ویژه هنگامی که

اداره‌ی ملی سرشماری/آمار تصمیم به انتشار فایل داده‌ها دارد، در فرایند ادیت، خطاهای قطعی باید کشف و در مورد آن‌ها اقدامات لازم انجام شود. شایان ذکر است که اصلاح خطاهای مشکوک، مشکل‌تر از اصلاح خطاهای ذکر شده از نوع دیگر است و کشف آن‌ها نه تنها فایده‌ی چندانی ندارد بلکه بر هزینه کلی پردازش اطلاعات می‌افزاید.

گروه ادیت باید به طور مداوم فعال باشد تا تعیین کند چه چیزی نتیجه‌بخش است یا نیست؟ این گروه‌ها همچنین باید جنبه‌هایی را که نتیجه‌بخش هستند و می‌توان توسعه داد و کارامدتر ساخت تعیین کنند تا داده‌ها سریع‌تر به دست کاربران برسند، علاوه بر این اداره‌های ملی سرشماری/آمار هر اندازه زودتر در فرایند سرشماری، خطاها را پیدا کنند، احتمال بیشتری برای تصحیح آنها وجود دارد.

ادیت بیش از اندازه باعث تأخیر در ارائه‌ی نتایج سرشماری می‌شود. اگرچه کارکنان درگیر در سرشماری‌ها و آمارگیری‌ها برای تجربه‌ی خود در این زمینه مدرک مستند ندارند، نتایج یک پژوهش توسط پولوم و دیگران (۱۹۸۶) نشان داد که ادیت ماشینی (در طرح آمارگیری جهانی باروری) انتشار نتایج را با یک سال تأخیر روبرو ساخت. بنا بر این شاید بهتر باشد که اداره‌های ملی آمار در وهله‌ی اول، بودجه‌ی خود را صرف دستیابی اطلاعات با کیفیت بالا در جریان اجرای سرشماری یا آمارگیری کنند.

از آن‌جا که رایانه می‌تواند به ویژگی‌های زیادی توجه کند، در فرایند ادیت باید از این مزیت بهره‌برداری شود. بدین ترتیب، روش‌های ادیتی که مستلزم کنترل بسیاری از ویژگی‌های مرتبط با یکدیگرند، احتمالاً بیش از روش‌های ساده منجر به جانهی پاسخ‌های مناسب‌تر می‌شوند. از طرف دیگر، ادیتی که به‌صورت ضعیفی طراحی شده باشد، به احتمال زیاد منجر به تولید داده‌های سرشماری با کیفیت پایین می‌شود. از این رو گروه ادیت باید شامل کارشناسان موضوعی با تجربه از رشته‌های ذی‌ربط و هم‌چنین پردازشگران داده‌ی مجرب باشد. اعضای گروه باید با دقت، متغیرهایی را که در آزمون‌های مربوط به سازگاری به‌منظور تعیین مشخصه‌های ادیت مورد بررسی قرار می‌گیرند انتخاب کنند. خروجی‌های برنامه باید شامل درصد پاسخ‌هایی باشد که تغییر یافته یا جانهی شده‌اند. در این صورت، تحلیل‌گران برای قضاوت درباره‌ی کیفیت اطلاعات در وضعیت بهتری خواهند بود. برای مثال، درصد بالای جانهی، هشداری است برای این که داده‌ها با احتیاط مورد استفاده قرار گیرند.