

کدگذاری با کمک رایانه (CAC)

عبدالحمید حقیقی*، محمدرضا اناری، زهرا نجفی، سمیه رجبی، الهه عباسی

مرکز آمار ایران

چکیده. کدگذاری داده‌های آماری یکی از مراحل حساس و با اهمیت در فرایند فعالیت‌های آماری است. در این مرحله، بخشی از اطلاعات کلیدی و مهم طی فرایند خاصی به کد تبدیل می‌شوند تا ورود اطلاعات و استخراج نتایج با سرعت و دقت بیش‌تری انجام شود. این مقاله به اهمیت کدگذاری داده‌ها، رایج‌ترین روش‌های کدگذاری در کشورها و نقاط ضعف و قوت روش‌های کدگذاری و برخی راه‌کارها برای ارزیابی کدگذاری اشاره می‌کند. در ادامه نیز به سابقه استفاده از کدگذاری با کمک رایانه (CAC) (Computer Aided Coding) در مرکز آمار ایران و نحوه استفاده آن در طرح سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ و روش ارزیابی آن پرداخته می‌شود.

۱- مقدمه

سیاست‌گذاران و برنامه‌ریزان کشورها عموماً برای تصمیم‌گیری‌های خود نیاز به داده‌های واقعی دارند که در زمان مناسب به دست آن‌ها برسد و بتوانند از آن‌ها در تصمیم‌گیری‌ها استفاده نمایند. این داده‌ها معمولاً از طریق نظام‌های طبقه‌بندی و کدگذاری ارقام پرسشنامه به دست می‌آید. طبقه‌بندی و کدگذاری داده‌ها سبب بهبود بهره‌برداری از آمارها و افزایش کارایی تصمیم‌گیری‌ها و برنامه‌ریزی‌ها می‌شود. علاوه بر آن کدگذاری داده‌ها ضمن ایجاد یک زبان مشترک، مقایسه‌پذیری آمارها در سطح ملی و بین‌المللی را نیز

واژگان کلیدی: کدگذاری؛ کدگذاری با کمک رایانه؛ کدگذاری خودکار؛ کدگذاری دستی.

* نویسنده عهده‌دار مکاتبات

فراهم می‌آورد.

کدگذاری اطلاعات و داده‌های آماری یکی از مراحل حساس و با اهمیت در فرایند فعالیت‌های آماری است. در این مرحله بخشی از اطلاعات کلیدی و مهم، طی فرایند خاصی به کد تبدیل می‌شوند تا ورود اطلاعات و استخراج نتایج با سرعت و دقت بیش‌تری انجام شود. کد، مجموعه‌ای از نمادها است که باعث ایجاد زبان مشترک در بین کاربران داده‌ها می‌شود و معمولاً شامل یک یا چند کاراکتر الفبایی، عددی یا الفبایی/ عددی است که به هر یک از رده‌های طبقه‌بندی نسبت داده می‌شود. هر کد منحصرأ برای یک رده خاص داخل طبقه‌بندی به کار گرفته می‌شود و در صورتی که آن رده خاص تغییر کند، کد نیز باید تغییر یابد.

کدگذاری، فرایند کدهمی به هر یک از اقلام با استفاده از طبقه‌بندی‌ها است. بهره‌گیری از طبقه‌بندی‌ها و سیستم کدگذاری‌ها در سطوح ملی و بین‌المللی کارکردهای متفاوتی از جمله صرفه‌جویی در وقت و هزینه، افزایش درجه صحت عملیات کدگذاری، امکان همکاری چند جانبه بین کشورها از طریق ایجاد زبان مشترک، تسریع در تبادل اطلاعات و شناسایی سریع‌تر اقلام، استفاده از کمک‌ها و مشاوره‌های فنی و تخصصی سازمان‌های بین‌المللی ارائه‌دهنده طبقه‌بندی‌ها و سیستم‌های کدگذاری دارد. عمده‌ترین هدف از ایجاد سیستم‌های کدگذاری، افزایش بهره‌وری، کنترل و جلوگیری از بروز خطا، افزایش دقت و سرعت عملیات کدگذاری و صرفه‌جویی اقتصادی است.

فعالیت آماری از مرحله گردآوری داده‌ها تا انتشار نهایی نتایج، فرایندهای متفاوتی را شامل می‌شود که هر کدام با خطاهایی روبه‌رو است. خطاهای پردازش با فرصت‌هایی که فناوری فراهم کرده است مانند استفاده از رایانه، تصویربرداری و... می‌توانند به حداقل برسند، هر چند که اگر همین فناوری‌ها به‌دقت مدیریت نشوند خود می‌توانند خطاهایی را به‌وجود آورند. با برخورداری از شیوه‌های جدیدتر گردآوری و ورود داده‌ها، خطاهای داده‌آمایی به‌طور قابل ملاحظه‌ای کاهش می‌یابند. برای کمک به کاهش خطای کدگذاری نیز می‌توان از روش کدگذاری با کمک رایانه و یا روش خودکار استفاده کرد.

۲- روش‌های کدگذاری

در گذشته، نتایج پاسخ‌های متنی طرح‌های آماری، معمولاً به صورت دستی کدگذاری می‌شد که این کار، به‌خصوص برای حجم زیاد داده‌ها خیلی وقت‌گیر، هزینه‌بر و مستعد خطا بود. به‌همین علت، برای این‌که عملیات کدگذاری تا حد امکان ساده‌تر باشد، مراکز آماری بسیاری تصمیم به استفاده از نرم‌افزارهای کدگذاری گرفتند. قسمت اساسی روش‌های کدگذاری رایانه‌ای، مبتنی بر فرهنگ داده‌ها (Data Dictionary) است که شامل کلمات یا اصطلاحات مربوط به کدگذاری می‌باشد. فرهنگ داده‌ها، شامل شرح کدهای طبقه‌بندی‌های رسمی، به‌علاوه پاسخ‌های تجربی که از طرح‌های گذشته یا مطالعات آزمایشی به دست آمده‌اند می‌باشد که در تعیین کدها مورد استفاده قرار می‌گیرد. در شصت سال اخیر کمیته سرشماری ایالات متحده، سیستم‌های کدگذاری مختلفی را ایجاد کرده‌اند و فرهنگ داده‌ها را بر اساس نمونه بزرگی از پاسخ‌های تشریحی که به‌وسیله کارشناسان به صورت دستی کدگذاری شده‌اند تهیه کرده است. یکی از سیستم‌های کدگذاری مورد استفاده، به این صورت است که رایانه فایل فرهنگ داده‌ها را به‌منظور یافتن پاسخ‌های متنی پرسشنامه‌ها جستجو می‌کند و در صورت یافتن، کدهای منحصر به فرد آن‌ها را نمایش می‌دهد. کدگذاری به‌وسیله الگوریتم دیگری که الگوریتم توزین نامیده می‌شود نیز انجام می‌شود. در این الگوریتم برای هر کلمه ورودی، وزنی در نظر گرفته می‌شود که این وزن میزان اهمیت کلمه را نشان می‌دهد. محاسبه وزن بر اساس فراوانی رخ دادن هر کلمه در فرهنگ داده‌ها است. سپس رایانه برای پاسخ تشریحی ورودی، فرهنگ داده‌ها را جستجو کرده و اگر کلمه تطبیقی مناسبی پیدا نکرد شرح کدهایی را که شبیه به کلمه ورودی است تحلیل می‌کند و موردی را که دارای وزن زیادتری است انتخاب می‌کند و به‌عنوان انطباق جزئی می‌پذیرد. امروزه کدگذاری بسته به میزان استفاده از رایانه در طی فرایند کد دهی، معمولاً به دو روش زیر انجام می‌شود.

الف) کدگذاری خودکار (AC Automated Coding)؛

ب) کدگذاری به کمک رایانه (CAC).

- در کدگذاری خودکار، رایانه برای پاسخ‌های تشریحی، کد تعیین می‌کند. در این تکنیک نمی‌توان انتظار داشت که برای همه شرح کدها، کد تعیین شود. پس یک روش کدگذاری دستی یا کدگذاری به کمک رایانه نیاز است که بعد از این مرحله به پاسخ‌های کدگذاری نشده، کد مناسب اختصاص دهد؛
- در کدگذاری به کمک رایانه، کدگذار با همکاری متقابل رایانه کد مناسب را تعیین می‌کند. به طوری که رایانه، کدگذار را برای جستجوی کد عبارات ورودی در فرهنگ داده‌ها، هدایت می‌کند. برای مثال کدگذار شرح کد را در رایانه وارد می‌کند. رایانه برای کدگذار تمام توضیحات موجود در فرهنگ داده‌ها را که می‌تواند با عبارت ورودی منطبق باشد نشان می‌دهد (اگر یک انطباق درست وجود داشته باشد فقط یک توضیح نشان داده می‌شود) و در غیر این صورت، کدگذار از میان کدهای پیشنهادی، کد مناسب را انتخاب می‌کند. می‌توان گفت که مهم‌ترین ویژگی سیستم CAC ترکیب توانایی ذهنی و قدرت رایانه است.

هدف AC استخراج یک کد واحد از فرهنگ داده‌ها منطبق با عبارت ورودی است. اما CAC کدهای مختلف (با تفاوت‌های کمی از یکدیگر) را نشان می‌دهد. شایان ذکر است که کدگذار به صورت متقابل با رایانه کار می‌کند و به سوی کدهای نشان داده شده هدایت می‌شود و موردی را که مناسب‌تر است انتخاب می‌کند.

مرحله کدگذاری داده‌ها می‌تواند در زمان‌های مختلف انجام شود. AC می‌تواند بعد از مصاحبه انجام شود یعنی هنگامی که گردآوری داده‌ها به پایان رسیده است. اما CAC می‌تواند در طول مصاحبه نیز انجام شود یعنی به صورت قدم به قدم با گردآوری داده‌ها. تصمیم‌گیری درباره این که کدام یک از روش‌های کدگذاری مناسب‌اند به عوامل زیادی بستگی دارد، که عبارتند از:

- | | |
|--|------------------------|
| } به کمک رایانه با کدگذار
} به کمک رایانه بدون کدگذار
} روش دستی و قدیمی | ۱- روش گردآوری اطلاعات |
|--|------------------------|

۲- حجم داده‌های کدگذاری شده } تعداد زیاد
کم

۳- طول مصاحبه } مصاحبه کوتاه
مصاحبه بلند

۴- ساختار طبقه‌بندی در رابطه با } ساختار طبقه‌بندی ساده
تغییرپذیری پاسخ‌های تشریحی } ساختار طبقه‌بندی پیچیده و تغییرپذیری بالای پاسخ‌های تشریحی

ساختار طبقه‌بندی می‌تواند به صورت درختی یا شاخه‌ها، زیرشاخه‌ها و برگ‌ها نشان داده شود. شاخه‌ها سطوح کلی طبقه‌بندی است که به صورت سلسله مراتبی، بالاتر از زیرشاخه‌ها و برگ‌ها که سطوح جزئی‌تر طبقه‌بندی هستند، قرار دارد. بنا بر این ساختار یک طبقه‌بندی ساده به صورت درختی با شاخه و بدون زیرشاخه یا با زیرشاخه کم و بدون برگ است. در حالی که ساختار طبقه‌بندی پیچیده به صورت درختی با همه ترکیبات است. مثال‌هایی از ساختار طبقه‌بندی‌های ساده و پیچیده به ترتیب طبقه‌بندی کشورها و طبقه‌بندی مشاغل است.

داده‌ها معمولاً به کمک رایانه C.A.P.I., C.A.T.I., C.A.S.I یا به وسیله روش دستی با کاغذ و قلم P.A.P.I گردآوری می‌شوند و در پایگاه داده‌ها ذخیره می‌شوند. در این حالت، حجم داده‌هایی که کدگذاری می‌شوند نقش اساسی در تعیین روشی که انتخاب می‌شود بازی می‌کنند به طوری که:

- برای حجم زیاد داده‌ها توصیه می‌شود که از AC استفاده شود و برای مواردی که کدگذاری نشده‌اند از CAC استفاده شود؛
- برای حجم کم داده‌ها و طبقه‌بندی ساده بهتر است که از AC استفاده شود؛
- برای حجم کم داده‌ها و طبقه‌بندی پیچیده استفاده از CAC مناسب‌تر است.

جدول زیر حالت‌های بالا را خلاصه می‌کند.

کیفیت داده‌ها		ساختار طبقه‌بندی
تعداد کم	تعداد زیاد	
AC	AC+CAC	ساده
CAC	AC+CAC	پیچیده

۳- کدگذاری در برخی کشورها

۱- ۳- ایتالیا

اداره آمار ایتالیا کدگذاری به کمک رایانه و کدگذاری خودکار را به منظور غلبه بر مشکلات مربوط به کدگذاری دستی پاسخ‌های متنی پرسشنامه‌ها پذیرفته است. کدگذاری دستی بسیار وقت‌گیر، پر هزینه و مستعد خطا است. نرم‌افزارهای انتخابی این کشور ACTR V۳ و BLAISE که بر اساس دو فلسفه مختلف کدگذاری یعنی کدگذاری خودکار (AC) و کدگذاری به کمک رایانه (CAC) طراحی شده‌اند. دو روش کدگذاری اهداف مختلفی دارند به طوری که هدف AC ماکسیمم کردن فراوانی کدهای یکتایی است که برای پاسخ‌های تشریحی تعیین می‌شود در حالی که هدف CAC ارائه حداکثر کمک ممکن به کدگذار، می‌باشد.

۲- ۳- ایرلند

برای اولین بار، در کدگذاری مشاغل سرشماری نفوس سال ۱۹۹۶ ایرلند تقریباً ۶۰۰۰۰۰ نوع شغل، که ۲/۱ میلیون فرد (۵۸ درصد کل جمعیت) را پوشش می‌داد در ۳۴۶ گروه متمایز و با ترکیبی از روش‌های کاملاً خودکار و کدگذاری با کمک رایانه کدگذاری شد. آموزش کدگذاری با کمک رایانه، در طی ۶ هفته، در ۳ مرحله و هر مرحله به طول ۲ هفته اجرا شد. در دو هفته اول کدگذاران به کدگذاری با روش‌های دستی پرداختند. در ۲ هفته دوم کدگذاران به کدگذاری CAC با داده‌های آزمایشی پرداختند. در نهایت در ۲ هفته آخر به کدگذاران، بخشی از فایل کدگذاری نشده برای کدگذاری داده شد. در تمام مراحل آموزش هر شرح کدی که اشتباهاً کد می‌خورد توسط ناظر به کدگذار برگردانده می‌شد. زمانی که تمام ۶ هفته آموزش به اتمام رسید تقریباً همه کدگذاران صلاحیت

کدگذاری با رایانه را کسب کرده بودند و هیچ کدام از آن‌ها به آموزش بیش‌تری نیاز نداشتند. با این حال نیز در نهایت میانگین خطای کدگذاری، ۱۱ درصد بوده است (با ۸۹ درصد دقت) که این مقدار خطا بین ۶ درصد تا ۱۸ درصد متغیر بود.

۴- ارزیابی کیفیت کدگذاری

عملیات کدگذاری مانند سایر فعالیت‌های طرح‌های آماری عاری از خطا نیست. خطای کدگذاری یکی از خطاهای پردازش و در کل نوعی خطای غیر نمونه‌گیری محسوب می‌شود.

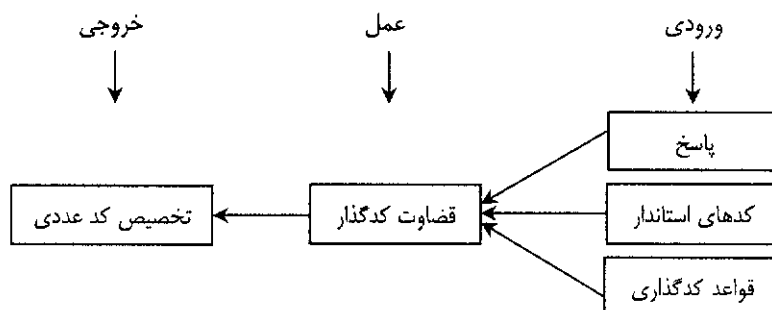
۵- مشکلات کدگذاری

عملیات کدگذاری سه مولفه ورودی دارد:

- پاسخ یا شرح قلم ثبت‌شده؛
- مجموعه‌ای از کدهای از پیش تعیین شده عددی یا کدهای استاندارد؛
- ابزار یا قواعد (راهنمای) کدگذاری؛

کدگذار بر اساس ورودی‌های فوق یک کد عددی به شرح قلم ثبت شده، تخصیص می‌دهد.

مسائل و مشکلاتی که در هنگام کدگذاری رخ می‌دهند زیاد هستند و همیشه توسط سازمان‌های آماری شناخته نمی‌شوند، برخی از این مسائل عبارتند از:



بیشتر کدگذاران مستعد خطا هستند، خطاها به دلیل استفاده نامناسب از راهنمای کدگذاری یا نقص آن رخ می‌دهند. مثلاً کدهای استاندارد نمی‌توانند تمام پاسخ‌های باز پاسخگویان را پوشش دهند. کدگذاران ماهر نیز اغلب در مورد تخصیص کد عددی مناسب توافق ندارند.

افزایش و ارتقای کیفیت عملیات کدگذاری مشکل است زیرا کدگذاری می‌تواند جنبه موضوعی داشته باشد. گاهی اوقات نیز پوشش مصادیق کامل نیست و کدگذاران باید از نظرات و قضاوت‌های خودشان استفاده نمایند. از طرف دیگر ارتقای مهارت کدگذاران نیز به زمان زیادی احتیاج دارد. مدیریت عملیات کدگذاری در آمارگیری‌های مهم و چند منظوره مشکل است. کنترل خطا در چنین عملیاتی بسیار چالش‌برانگیز است. کدگذار معمولاً با چندین کتاب کد برای اقلام مختلف پرسشنامه‌ها سر و کار دارد. به همین دلیل ممکن است برخی کدگذارها در اقلام ویژه‌ای مانند شغل یا فعالیت خبره شوند، اما این کار، کار خسته‌کننده‌ای است و می‌تواند منبع مهمی از خطا باشد. برای اجتناب از منابع جدید خطا، کدگذارها نباید تنها به حافظه‌شان تکیه کنند. آن‌ها باید در مواقع لزوم از کتاب‌های کد استفاده نمایند.

۶- کیفیت نظام‌های طبقه‌بندی

کیفیت نظام‌های طبقه‌بندی مانند طبقه‌بندی‌های استاندارد و بین‌المللی و ملی بر کیفیت عملیات کدگذاری تأثیرگذار است. دو بعد کیفی که در این زمینه می‌توان در نظر گرفت، ابعاد مناسب و درستی هستند. دو مسئله مهم در رابطه با این ابعاد عبارتند از:

آیا عبارات نوشته شده از قواعد و دستورالعمل‌های پذیرفته شده برای نوشتن شرح قلم پیروی می‌کند؟ آیا شرح قلم به‌طور مناسبی مفاهیم بیان‌شده در طبقه‌بندی را شرح می‌دهند؟ (درستی)

آیا رده‌ها، گروه‌ها یا مفاهیم در طبقه‌بندی مورد استفاده دو به دو ناسازگار هستند (همپوشانی ندارند) و تمام مصادیق اعم از مشاغل یا فعالیت‌ها را به‌طور جامع پوشش می‌دهند؟ (مناسبت).

علاوه بر این باید رده‌ها و گروه‌های طبقه‌بندی به خوبی به کارشناسان کدگذاری شرح داده شود. در صورتی که کدگذاری به صورت خودکار انجام می‌شود، طراحان نرم‌افزار نیز باید قادر به درک تفاوت‌های بین رده‌ها باشند تا بتوانند کدهای صحیح را تولید نمایند. عملیات کدگذاری نیازمند ایجاد و بهنگام‌سازی مجموعه کدهای تخصصی از پاسخ‌ها (مصادیق) است. به این فایل‌ها مجموعه‌های واقعی هم گفته می‌شود. در آموزش کدگذاران، از سیستم‌های کدگذاری با کمک کامپیوتر و خودکار استفاده می‌شود. برای هر یک از این مجموعه‌ها، ابعاد کیفیت زیر در نظر گرفته شده و توصیه می‌شود در فراداده‌های مربوط به کیفیت لحاظ شوند:

- توسعه مصادیق تا حدی که مجموعه واقعی معرف جامعه مورد بررسی باشد (مناسبت)؛
- توسعه مصادیق تا حدی که موارد کافی درون هر رده از کدهای مجموعه واقعی موجود باشند (درستی)؛
- توسعه مصادیق تا حدی که مجموعه داده واقعی مانند پاسخ‌های گردآوری شده در آمارگیری یا سرشماری باشد (درستی).

علاوه بر نظام‌های طبقه‌بندی، کیفیت نرم‌افزار مورد استفاده و پاسخ‌های حاصل از آمارگیری نیز بر کیفیت کدگذاری تأثیرگذار است.

۷- کنترل دستی خطای کدگذاری

دو روش مختلف برای کنترل دستی خطای کدگذاری وجود دارد: بازبینی وابسته و بازبینی مستقل. در بازبینی وابسته، کدگذار اول کدی به شرح قلم تخصیص می‌دهد. کد تخصیص یافته، توسط کدگذار دوم یا بازبین کد، مرور می‌شود. در واقع بازبین کد درباره درستی کد اختصاص یافته تصمیم‌گیری می‌نماید.

در صورتی که کد از نظر بازبین کد صحیح باشد، بدون تغییر باقی می‌ماند و در غیر این صورت بازبین کد، کد صحیح را به شرح قلم مربوطه اختصاص می‌دهد. بازبینی وابسته بسیار ناکارا است. تجربه نشان می‌دهد که در این روش، در حدود ۵۰ درصد خطاها

اصلاح می‌شوند. برخی مطالعات نشان می‌دهد که نرخ اصلاح خطا حتی کمتر از این حد نیز می‌تواند باشد. دلیل نرخ پایین تغییر یا اصلاح آن است که قضاوت بازیکن کد به شدت از کد عددی تخصیص یافته توسط کدگذار تأثیر می‌پذیرد. در این روش گرایش و جهت‌گیری کاملاً آشکار است و تنها خطاهای خالی از ابهام اصلاح می‌شوند.

خطاهای کمتر آشکار و مبهم، بدون تغییر باقی می‌ماند زیرا بازیکن کد اغلب این دلیل را می‌پذیرد که کدهای اولیه کاملاً صحیح هستند و نباید تغییر نمایند. به عبارت دیگر تمایل به حمایت و پذیرفتن قضاوت اولیه کدگذار وجود دارد. در بازیکنی مستقل، مبنای چنین مشکلی در بخش بازیکنی کنار گذاشته می‌شود. یعنی بازیکن کد، به کدهای عددی تخصیص یافته اولیه دسترسی ندارد. با بازیکنی مستقل، کدگذار اول کد عددی را به شرح قلم مربوط تخصیص می‌دهد. همان شرح قلم توسط کدگذار دوم یا بازیکن کد، دوباره کدگذاری می‌شود. از آنجایی که کدگذار و بازیکن کد به‌طور مستقل عمل می‌نمایند، هیچ‌کدام از کد عددی دیگری، اطلاعی ندارند. دو کد عددی با یکدیگر مقایسه می‌شوند و از قاعده تصمیم‌گیری زیر استفاده می‌شود.

اگر هر دو کد یکی باشند، آن گاه کد تخصیص یافته توسط کدگذار، کد نهایی محسوب می‌شود. در صورتی که هر دو کد یکسان نباشد، شرح قلم در اختیار کدگذار سوم قرار می‌گیرد و او نیز مستقل از کدگذار قبلی، کد مناسب را به شرح قلم مربوط تخصیص می‌دهد:

- اگر کدهای کدگذار اول و کدگذار سوم یکی باشد، آن گاه کد تخصیص یافته توسط کدگذار اول، کد نهایی محسوب می‌شود؛
- اگر کدهای کدگذار دوم (بازیکن کد) و کدگذار سوم یکی باشد، آن گاه کد تخصیص یافته توسط بازیکن کد، کد نهایی تلقی می‌شود؛
- در صورتی که هر سه کد متفاوت باشد، شرح قلم توسط کدگذار چهارم کد زده می‌شود و اوست که تصمیم نهایی را مشخص می‌کند و کد تخصیص یافته توسط این کدگذار، کد نهایی محسوب می‌شود.

این سیستم، سیستم بازیکنی دو طرفه یا داوری نامیده می‌شود. نوع دیگری از بازیکنی،

بازبینی سه طرفه مستقل است که در آن سه کدگذار مستقل به شرح قلم مربوط، کد تخصیص می‌دهند و کد نهایی با استفاده از کد اکثریت تعیین می‌شود و کدگذار چهارم در صورتی هر سه کد اول متفاوت باشند، در مورد کد تصمیم‌گیری می‌نماید. با وجود این، مطالعات نشان می‌دهد بازبینی دو طرفه مستقل با داوری نتایج مشابه و هزینه‌های کمتری نسبت به بازبینی سه طرفه مستقل با قضاوت دارد.

سوالی که مطرح می‌شود آن است که آیا بازبینی مستقل بهتر از بازبینی وابسته است؟ فرض اصلی در طرح بازبینی آن است که کد نهایی حاصل از قاعده تصمیم‌گیری، کد صحیح است زیرا که احتمال یکسان بودن دو یا سه کد به‌طور مستقل تخصیص یافته و خطا کاملاً کوچک است. با وجود این، مهم است که کدگذاران تقریباً از مهارت یکسانی در کدگذاری برخوردار باشند. در غیر این صورت کدها به‌ندرت یکسان خواهد بود و بسیاری از موارد باید به داوری ارجاع داده شود و به موجب آن حجم کار داور و خطای داوری افزایش خواهد یافت. به‌علاوه ممکن است دو کدگذار ضعیف به سادگی کد یک کدگذار خوب را نپذیرند، زیرا کدگذاران ضعیف، راهنمای کدگذاری را به‌خوبی درک نکرده‌اند.

بازبینی مستقل پرهزینه‌تر از بازبینی وابسته است اما کدهای درست بیش‌تری را نسبت به بازبینی وابسته تولید می‌کنند. سیستم مستقل دو طرفه حداقل هزینه را در بین طرح‌های مستقل دارد و در بسیاری از سازمان‌های آماری مورد استفاده قرار می‌گیرد.

در اداره آمار کانادا، بازبینی مستقل در چارچوب سیستم کنترل کیفیت نمونه‌گیری برای پذیرش انجام می‌شود. در این سیستم بازبینان، سطح مهارت متفاوتی دارند. بازبینان سطح اول با تجربه‌تر از کدگذاران هستند و بازبینان کد سطح دوم با تجربه‌تر از بازبینان کد سطح اول هستند. بازبینان کد سطح اول به‌طور مستقل به نمونه‌ای از کدهای تخصیص یافته توسط کدگذاران، دوباره کد اختصاص می‌دهند. مغایرت‌ها در اختیار بازبین سطح دوم قرار می‌گیرد. در این‌جا نیز از قاعده اکثریت استفاده می‌شود. از آن‌جایی که دو بازبین کد، با تجربه‌تر از کدگذار هستند وضعیتی که قبلاً شرح داده شد، رد کد کدگذار خبره توسط دو کدگذار ضعیف، در این‌جا اصلاً رخ نمی‌دهد.

۸- سنجش کیفیت برای روش‌های AC و CAC

کیفیت روش AC را می‌توان به‌وسیله دو پارامتر فراخوانی و دقت سنجید. اولین پارامتر درصد پاسخ‌های تشریحی است که در مجموع موارد، به صورت خودکار کدگذاری شده‌اند. این پارامتر می‌تواند به صورت نرخ کدگذاری نیز نشان داده شود. از آن جایی که بعضی تعاریف کدگذاری نشده باقی می‌مانند نیاز به کدگذاری به صورت دستی نیز است.

دومین پارامتر، درصدی را نشان می‌دهد که به صورت درست کدگذاری شده‌اند (در مقایسه با مواردی که به صورت دستی کدگذاری شده‌اند). هدف روش AC ماکسیمم کردن هر دو پارامتر زیر به صورت هم‌زمان است. این کار نیاز به یک فرهنگ داده‌ها دارد که به صورت پیوسته به‌روز شود.

- ماکسیمم کردن نرخ فراخوانی، بدون توجه به نرخ دقت، تعداد کدهای تعیین شده را افزایش می‌دهد در حالی که در مجموع کیفیت را پایین می‌آورد؛
- از طرف دیگر رسیدن به دقت بالا، در مجموع کیفیت را بالا می‌برد ولی تعداد مواردی را که کدگذاری می‌شود را به شدت کاهش می‌دهد.

بنا بر این برای اجرای توازن در روش کدگذاری خودکار، پارامترهای فراخوانی و دقت را باید با هم بهبود بخشید. برای روش CAC نیز اندازه‌گیری موارد فوق را می‌توان به کار برد. با این تفاوت که به جای پارامتر فراخوانی، «تعداد کدهای تعیین شده به‌وسیله کدگذار» به کار می‌رود. در حالی که پارامتر دقت مانند گذشته به صورت کدهایی که درست انجام شده‌اند تعریف می‌شود و داشتن یک فرهنگ داده‌ها به‌روز، به‌عنوان شرط اصلی بهبود کیفیت، باقی می‌ماند.

۹- نمونه‌گیری برای پذیرش

نمونه‌گیری برای پذیرش، یکی از روش‌های مورد استفاده در چرخه دمینگ PDCA (Planning, Do, Check and Act)، کنترل فرایند و بهبود مداوم کیفیت است. با استفاده از این نوع نمونه‌گیری در مورد کیفیت محصول یا خدمت مثلاً بر اساس میزان خطا یا درصد اقلام معیوب تصمیم‌گیری می‌شود و هدف از اجرای آن برآورد میزان

کیفیت یا خطا نیست، نمونه‌گیری برای پذیرش توسط دمی‌نگ و راج معرفی و اولین بار در چین جنگ جهانی دوم در صنایع نظامی ارتش امریکا مورد استفاده قرار گرفت. این روش مبتنی بر اجرای یک آزمون فرض با پارامترهای مربوط است و به صورت دو مرحله‌ای نیز قابل اجرا است. امروزه یکی از متداول‌ترین سیستم‌های نمونه‌گیری برای پذیرش مشخصه‌های کیفی وصفی $10.5E$ MIL STD است و امکان سه نوع نمونه‌گیری را فراهم می‌سازد: یک مرحله‌ای، دو مرحله‌ای و چند مرحله‌ای. استاندارد $10.5E$ MIL STD بر اساس سطح کیفیت قابل قبول (Acceptance Quality Level) AQL پایه‌ریزی شده است. وقتی که طرح‌ها بر اساس نسبت اقلام معیوب طراحی می‌شوند، دامنه AQL از 0.1% درصد تا 10% درصد در نظر گرفته می‌شود. در این استاندارد اندازه نمونه به وسیله اندازه انباشته و سطح بازرسی (حساب‌شده‌تر، نرمال و تعدیل‌شده) تعیین می‌شود. از بازرسی نرمال در ابتدای فعالیت‌های بازرسی استفاده می‌شود. بازرسی حساب‌شده‌تر به مراتب مشکل‌تر از بازرسی نرمال است. بازرسی تعدیل‌شده زمانی استفاده می‌شود که سابقه کیفیت محصولات اخیر تولیدکننده به طور قابل ملاحظه‌ای بهبود یافته باشد.

به‌طور کلی اندازه نمونه تحت شرایط بازرسی تعدیل‌شده، کمتر از اندازه نمونه‌ای است که تحت شرایط بازرسی نرمال استفاده می‌شود. این استاندارد روشی را برای تغییر سطح بازرسی از نرمال به تعدیل‌شده و حساب‌شده‌تر فراهم می‌سازد و در زمانی که حساس می‌شود که کیفیت محصول تغییر یافته می‌توان از آن استفاده نمود. برای اطلاعات بیشتر در این زمینه و آشنایی به جدول‌های مربوط به این استاندارد به {نورالنساء، رسول، (۱۳۷۶)، کنترل کیفیت آماری، ویرایش سوم، دانشگاه علم و صنعت ایران، تهران} مراجعه نمایید. همچنین با توجه به این که هدف برآورد خطا یا کیفیت نیست، ممکن است اندازه نمونه به میزان قابل توجهی کاهش یابد.

علاوه بر این سطح کیفیت قابل قبول به طرح نمونه‌گیری بستگی ندارد و توسط تولیدکننده تعیین می‌شود. از طرف دیگر مصرف‌کننده نیز طرح نمونه‌گیری خود را طوری طراحی می‌نماید که احتمال پذیرش در نقطه AQL حداکثر شود.

برای استفاده از این روش در کنترل کیفیت کدگذاری، لازم است سطح کیفیت قابل

قبول و یا حداکثر خطای قابل قبول و متوسط تعداد کد در واحد کار کدگذاری (حوزه) و سطح بازرسی تعیین شوند. علاوه بر این لازم است اصول زیر در عملیات کدگذاری رعایت شوند:

- کنترل کدگذاری (بازبینی کد) در سطح هر حوزه به تفکیک نوع کد و مستقل از فعالیت کدگذاری انجام شود. یعنی کدگذار و بازبین کد، در هر حوزه و برای هر نوع کد، افراد متفاوتی باشند و کدهای اختصاص یافته توسط کدگذار در اختیار بازبین کد قرار داده نشود (شرط مستقل بودن)؛
- بازبینی کدگذاری تا هنگامی که کدگذاری یک نوع کد تأیید نشود ادامه خواهد یافت و کدگذاری پس از عدم تأیید کدگذاری یک نوع کد، پایان کار محسوب نمی‌شود.

۱۰- سابقه استفاده از کدگذاری به روش CAC در مرکز آمار ایران

در سرشماری عمومی کارگاهی سال ۱۳۸۱ برای اولین بار ترکیبی از روش کدگذاری خودکار (AC) و کدگذاری به کمک رایانه (CAC) برای کدگذاری فعالیت‌های اقتصادی مورد استفاده قرار گرفت. قبل از آن، کدگذاری و بازبینی کد به روش دستی و با استفاده از دفترچه‌های کدها (که به صورت الفبایی و موضوعی تدوین می‌شد) انجام می‌شد. در سرشماری کارگاهی سال ۱۳۸۱ نرم‌افزار کدگذاری، ابتدا تمامی پرسشنامه‌های هر واحد کار را خوانده و به پرسشنامه‌هایی را که متن نوشته شده در «شرح فعالیت اقتصادی عمده» آن‌ها را عیناً در فایل مصادیق «طبقه‌بندی فعالیت‌های اقتصادی» پیدا می‌کرد به‌طور خودکار، کد اختصاص می‌داد. سپس پرسشنامه‌هایی که از طریق فوق‌الذکر قابل شناسایی نبود و فاقد کد بود به‌منظور اختصاص کد در اختیار کدگذار قرار می‌گرفت. در این مرحله، نرم‌افزار کدگذاری، اقدام به تفکیک کلمات «شرح فعالیت اقتصادی عمده» پرسشنامه نموده و با ترکیب‌های مختلف کلمات، فایل مصادیق «طبقه‌بندی فعالیت‌های اقتصادی» را جستجو می‌کرد. علاوه بر آن امکان کدگذاری بر اساس محتوا نیز وجود داشت.

۱۱- کدگذاری سرشماری عمومی نفوس و مسکن سال ۱۳۸۵

برای سرشماری عمومی نفوس و مسکن سال ۱۳۸۵، مطالعاتی به منظور انتخاب روش کدگذاری در مرکز آمار ایران انجام شد و در نهایت کدگذاری به روش CAC پیشنهاد و در آزمایش سرشماری (۱۳۸۴) اجرا شد. در این روش پس از تصویربرداری از پرسشنامه‌ها، تصویر قسمت‌هایی از پرسشنامه‌ها که باید کدگذاری شوند بر روی صفحه نمایشگر رایانه پدیدار می‌شود و کدگذاری ارقام با روش CAC به یکی از شیوه‌های زیر انجام می‌شود:

الف) کدگذاری به کمک فهرست الفبایی کدها

در این روش، ابتدا تصویر شرح نوشته شده بر روی صفحه نمایشگر رایانه پدیدار می‌شود، با تایپ حروف آن به ترتیب، مصادیق و کدهای مربوطه که با حروف تایپ شده شروع می‌شوند ظاهر می‌شود و کدگذار می‌تواند کد درست را از فهرست الفبایی انتخاب نماید. در این روش معمولاً با تایپ حروف بیش‌تر شرح کد مورد نظر، گزینه‌های انتخاب شده پیشنهادی محدودتری بر روی صفحه نمایشگر نمایان می‌شود و انتخاب کد سریع‌تر و دقیق‌تر انجام می‌شود.

ب) کدگذاری با جستجوی واژه‌ها

در این روش با توجه به تصویر شرح نوشته شده در پرسشنامه که بر روی صفحه نمایشگر رایانه پدیدار می‌شود، کدگذار یک واژه کلیدی یا بیش از یک واژه را تایپ نموده و سپس با استفاده از دستور « جستجو » عبارت مورد نظر را در فرهنگ داده‌ها جستجو و کد درست را انتخاب می‌نماید. در این روش نیز معمولاً با تایپ عبارت کامل‌تر شرح مورد نظر، فهرست ارائه شده توسط ماشین محدودتر و انتخاب کد مناسب، دقیق‌تر انجام می‌شود. به‌کارگیری این روش هنگامی مناسب است که با استفاده از شیوه نخست (فهرست الفبایی) و تایپ حروف اول عبارت مورد نظر، انتخاب شرح و کد مربوط به آن از فهرست الفبایی میسر نباشد.

پ) کدگذاری با دسته‌بندی مفاهیم

هر یک از نظام‌های طبقه‌بندی دارای سطح‌بندی ویژه‌ای از کدهاست که کدها با یک نظم منطقی از کدهای ۱ رقمی وارد سطوح تفصیلی‌تر می‌شود. در این روش با توجه به تصویر شرح نوشته شده در پرسشنامه بر روی صفحه نمایشگر و تشخیص رده اصلی مربوط به آن، از نخستین سطح طبقه‌بندی وارد سطوح پایین‌تر شده و کدگذار کد درست را در آخرین سطح آن انتخاب می‌نماید. با استفاده از این روش می‌توان انتخاب کد را به‌روش ترکیبی نیز انجام داد، به این شیوه که کدگذار پس از ورود به سطوح پایین‌تر طبقه‌بندی و رده مورد نظر، با تایپ حروف مربوط به شرح و جستجوی فهرست الفبایی، کد مورد نظر را انتخاب می‌نماید.

ت) کدگذاری مستقیم^۴

در این روش، کدگذاری به صورت مستقیم و با تکیه بر حافظه کدگذار انجام می‌شود. در صورتی که کدگذار کد مربوط به شرح یک اطلاع را به حافظه سپرده باشد می‌تواند کد مورد نظر را تایپ نماید. با تایپ کد، شرح آن کد از فهرست ماشینی برروی نمایشگر پدیدار می‌شود و در صورت یکسان بودن شرح اطلاع آماری با شرح نمایش داده شده در ماشین، کدگذار کد مربوط به آن را تأیید می‌کند.

۱۲- کنترل کیفیت عملیات کدگذاری در سرشماری‌های اخیر مرکز آمار ایران

در سرشماری‌های مرکز آمار ایران تا کنون کنترل عملیات کدگذاری (کنترل فرایند)، توسط کارشناسان کدگذاری در حین عملیات کددهی دستی یا با کمک رایانه، انجام شده است. اما از نمونه‌گیری، پس از پایان کدگذاری برای پذیرش یا رد واحد کار کدگذاری شده (کنترل محصول) توسط کدگذار استفاده نشده است.

نتیجه گیری

کدگذاری یکی از مهم‌ترین مراحل فرایند استخراج نتایج طرح‌های آماری محسوب می‌شود و خطاهای ناشی از آن جزء خطاهای غیر نمونه‌گیری است. انتخاب نوع فرایند کدگذاری ارتباط مستقیمی با روش گردآوری اطلاعات دارد. با توجه به استفاده از روش مصاحبه حضوری در سرشماری‌های عمومی نفوس و مسکن سال ۱۳۸۵ و تغییر روش ورود اطلاعات و استفاده از تکنولوژی (ICR (Intelligent Character Recognition)، و نیز در نظر گرفتن دو عامل سرعت و دقت در کدگذاری و تنوع اقلامی که باید کدگذاری شوند، روش کدگذاری با کمک رایانه (CAC) در سرشماری عمومی نفوس و مسکن سال ۱۳۸۵، پیشنهاد می‌شود.

همچنین برای تأیید عملیات کدگذاری در هر حوزه یا هر واحد کار در سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ علاوه بر کنترل فرایند در حین عملیات کددهی توسط کارشناسان کدگذاری، باید از یک طرح نمونه‌گیری برای پذیرش در هر واحد کار، که بهره جستن از آن در فعالیت‌های استخراج سرشماری‌ها توسط سازمان ملل نیز توصیه شده است، با توجه به میزان خطاهای کدگذاری اقلام شغل، فعالیت اصلی محل کار، رشته تحصیلی، پایه، دوره و مدرک تحصیلی، شهر، شهرستان و کشور، به‌تفکیک نوع قلم استفاده شود.

توضیحات

^۱ Computer Aided Self Interviewing

^۲ Computer Aided Telephone Interviewing

^۳ Computer Aided Personal Interviewing

^۴ Paper and Pencil Interviewing

^۵ این روش به دلیل دارا بودن درصد بالای خطا، در سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ مورد استفاده قرار نخواهد گرفت و فقط در آزمایش سرشماری آزمایش شد.

مرجع‌ها

- [۱] Stefania Macchia, Manuela Murgia, Coding Of textual Responses: various issues on automated coding and computer assisted coding, ISTAT, Rome, Italy, 2006.
- [۲] Gerard Keogh, Automatically Coding Occupation Descriptions from the 1996 Census of Population of Ireland, CSO.
- [۳] Frederick Conard, Using Expert Systems to Model and Improve Survey Classification Processes, U.S. Bureau of Labor Statistics.
- [۴] The third International Roundtable on Business Survey Frames, October 31 to November 3, 1998. New Zealand.
- [۵] Lyberg, L. Biemer, P.(2003), Introduction to Survey Quality, WILEY Series in Survey Methodology.
- [۶] Cathryn S. Dippo, Frederick G. Conard, and Daniel W.Gillman, (2000), Metadata and Data Quality. U.S. Bureau of Labor Statistics.
- [۷] نورالنساء، رسول، (۱۳۷۶)، کنترل کیفیت آماری، ویرایش سوم، دانشگاه علم و صنعت ایران، تهران.
- [۸] مستندات مربوط به سرشماری‌های نفوس و مسکن و سرشماری‌های کارگاهی مرکز آمار ایران.
- [۹] مستندات مربوط به کمیته‌ی طبقه‌بندی سرشماری نفوس و مسکن سال ۱۳۸۵.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی