طراحی آماری یک روش نمونه‌برداری در کنترل کیفیت داده‌های پژوهشی

محمدجواد ارشادی [1]

دانشیار، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران

مدیریت اطلاعات

دوره ۸، شماره ۱

بهار و تابستان ۱۴۰۱

چکیده: در مدارک علمی، نمایه‌سازی و کنترل کیفیت، فرایندهایی کلیدی وجود دارد که در صورت انجام درست آنها، امکان بازیابی مناسب در موتورهای جست‌وجو فراهم می‌آید. در منابع علمی، به روش‌های نمونه‌برداری در محصولات فیزیکی به اندازهٔ کافی پرداخته شده است؛ اما در حوزه داده‌ها، به‌ویژه داده‌های پژوهشی، کارهای اندکی انجام شده است. در این پژوهش، چارچوبی برای نمونه‌برداری فرایندهای کنترل کیفیت داده فراهم شده است. به‌عنوان مطالعه موردی، داده‌های پژوهشی پایگاه اشاعه اطلاعات پایان‌نامه‌ها/ رساله (پارسا)های دانش‌آموختگان کل کشور (گنج) انتخاب شده است. بر اساس نتایج، با توجه به کیفیت پذیرفتنی بسیاری از اقلام اطلاعاتی پارسا، پس از ثبت، نمونه‌برداری کاری حیاتی برای ارتقای کارایی واحد سازمان‌دهی و تحلیل اطلاعات است. منحنی OC برای طرح‌های گوناگون نشان می‌دهد که طرح‌های ارائه‌شده برای ارزیابی سطح کیفیت داده‌های پژوهشی، از کارایی مناسبی برخوردارند. چارچوب ارائه‌شده در این پژوهش، برای سازمان‌های گوناگون داده‌محور، به‌ویژه کسب‌وکارهای مبتنی بر داده، قابلیت بومی‌سازی دارد.

کلیدواژه‌ها: کیفیت داده، نمونه‌برداری، کنترل کیفیت، منحنی OCOC، سازمان‌دهی، تحلیل اطلاعات.

---

۱. نویسندهٔ مسئول: mjershadi@gmail.com

مقدمه

امروزه سازمان‌های گوناگون با رشد فزاینده پایگاه‌های داده و اطلاعات مواجه هستند. به مدارک و مستندات به‌عنوان بخشی از این پایگاه‌ها با توجه به نقش ویژه‌ای که در بقا و توسعه دانش سازمانی دارند، به‌عنوان یک پایگاه ارزشمند داده بیش از پیش توجه شده است. سوابق و مستندات بی‌کیفیت، منسوخ و حتی کم‌اهمیت در صورتی که در قالب یک فرایند کنترل کیفیت غربال‌گری نشوند و سطح کیفیت آنها ارتقا پیدا نکند، در آینده سازمان‌ها را با حجم فزاینده داده بی‌کیفیت مواجه خواهند کرد که به‌شدت چابکی و پویایی آنها را با تهدید روبه‌رو خواهد کرد. باید به پایگاه‌های اطلاعاتی مانند گنجینه‌های پایان‌نامه/ رساله (پارسا)ها که در دسترس عموم قرار می‌گیرند، بر پایه همین اصل و قاعده از جنبه‌های کیفیت، بهره‌وری و تعالی توجه شود تا جمعیت زیاد کاربران آنها سردرگم و گمراه نشوند و در نگرشی کلان، جامعه به رشد و تعالی برنامه‌ریزی‌شده دست یابد.

از سوی دیگر، در بهبود کیفیت مدارک و مستندات نیز همچون سایر حوزه‌ها، اصل ۸۰-۲۰ که برای نخستین بار جوران آن را توسعه داد، بسیار تأثیرگذار است (Wilkinson, 2006). بر پایه این اصل در یک جامعه ۸۰ درصد از معلول‌ها فقط از ۲۰ درصد علت‌ها ناشی می‌شوند. در بازیابی مدارک و مستندات تعداد واژه‌های زیاد کتاب‌شناختی بازیابی‌شده، کیفیت بازیابی را کاهش می‌دهد. به بیان دیگر، در مواردی که نتیجه یک جست‌وجو توسط کاربر تعداد زیادی مدرک باشد، نارضایتی را در پی خواهد داشت. از این رو، کیفیت بازیابی به‌عنوان بعدی مهم در حوزه کیفیت مدارک و مستندات معلول علل و عوامل گوناگونی است که لازم است در قالب اصل ۸۰-۲۰ واکاوی و تحلیل شود.

اگرچه پیاده‌سازی اصول تضمین کیفیت در فرایندهای مرتبط با پایگاه‌های اطلاعاتی مدارک علمی کاری کلیدی به حساب می‌آید، شناسایی و برطرف کردن خطاهای مشاهده‌شده در مراحل ثبت، نمایه‌سازی و ویراستاری مدارک علمی به ارتقای سطح کیفیت داده‌های اشاعه داده‌شده کمک شایانی خواهد کرد.

در پژوهش‌های پیشین، به مباحث حوزه تضمین کیفیت در سامانه‌ها و پایگاه‌های اطلاعاتی و همچنین سامانه‌های اطلاعاتی پژوهشی پرداخته شده و به جنبه‌های گوناگون آن نیز توجه شده است. از سوی دیگر، از آنجا که شناسایی و برطرف کردن ناهم‌خوانی‌های داده‌ها مستلزم طراحی روش نمونه‌برداری مناسب در حوزه کنترل کیفیت داده است، بومی‌سازی مفاهیم کنترل کیفیت آماری در داده‌های پژوهشی کاری اجتناب‌ناپذیر است. بر این اساس، نوآوری پژوهش جاری طراحی روشی مبتنی بر کنترل کیفیت آماری در حوزه داده‌های پژوهشی است تا به‌کمک آن چارچوبی برای بررسی حجم زیادی از داده‌ها در فرایندهای گوناگون پردازش داده‌های پژوهشی فراهم شود که بتوان به آن اعتماد کرد. از آنجا که مفهوم نمونه‌برداری با مفهوم کیفیت آمیخته بوده و زمینه پژوهش در حوزه داده و اطلاعات است، در ادامه به مرور برخی مفاهیم پایه‌ای در خصوص کیفیت اطلاعات پرداخته خواهد شد.

کیفیت اطلاعات

اطلاعات در کوتاه‌ترین تعریف، «داده‌های پردازش شده» است (Price & Shanks, 2016). از دیدگاه دیگر، این مفهوم آگاهی‌های به‌دست‌آمده از عنصرها و رویدادهای جهان هستی است (Batini & Scannapieco, 2016). اطلاعات از دو بعد ساختاری و معنایی نیز ارزیابی و تعریف می‌شوند (اثنی عشری و اسدی، ۱۳۹۴). ساختار اطلاعات از جنبه‌های معنایی آن مستقل بوده و ممکن است ایستا یا پویا ارزیابی شود. زمانی که اطلاعات مانند اطلاعات کتاب، تصویر، صفحه فشرده تغییرناپذیر باشند، از نوع ایستا هستند. زمانی‌که اطلاعات را کالا تلقی کنیم، مانند هر فرآورده‌ای قابلیت تولید، ذخیره‌سازی و انتقال دارد و می‌تواند آفریده شود، رشد یابد و نیست شود (Cárdenas-García, De Mesa & Castro, 2019).

از دو دیدگاه می‌توان به بعد معناشناختی اطلاعات توجه کرد. در دیدگاه نخست، اطلاعات بر پایه زمان و مکان در حال تغییر هستند، یک پدیده ممکن است در یک زمان یا مکان اطلاعات باشد و همان پدیده در زمان یا مکانی دیگر ارزش اطلاعاتی چندانی نداشته باشد. در دیدگاه دوم، اطلاعات در ذات خود جنبه فرایندی دارند (این دیدگاه با رویکرد قبلی متفاوت است و تا حدی در مقابل آن است). در این دیدگاه اطلاعات چیزی جز فرایند نیست، بنابراین، ماهیت پویا دارد. آنچه از جنبه ساختاری ماهیت ایستا دارد، در دو سوی این فرایند قرار می‌گیرد. از یک سو، داده (سوی آغازین) و سوی دیگر آن دانش است. اطلاعات برخلاف داده یا دانش تغییرپذیر نیستند، بلکه داده‌ها که همان مواد خام اطلاعات هستند و دانش به‌عنوان صورت هدفمند قابلیت تغییر دارند (اثنی عشری و اسدی، ۱۳۹۴).

مارچاند[1] (۱۹۹۰) هشت بُعد کیفیت اطلاعات را به‌عنوان چارچوبی برای تجزیه و تحلیل مشخص کرده است:

۱. ارزش واقعی محصول یا خدمت اطلاعاتی که ممکن است برای یک کاربر اطلاعات داشته باشد.
۲. ویژگی‌های اطلاعات: ابعادی مانند صحت یا کامل بودن در این بخش گنجانده می‌شوند.
۳. قابلیت اطمینان محصول یا خدمات اطلاعات.
۴. زمان‌مند بودن اطلاعات: معنای اطلاعات باید در چرخه حیات خود تفسیر شود.
۵. ارتباط (مرتبط بودن) اطلاعات: درجه و میزان هم‌خوانی اطلاعات با استانداردها یا معیارهای کاربر. برای طراح سیستم‌های اطلاعاتی، ارتباط ممکن است این هم‌خوانی هماهنگی اطلاعات با ویژگی‌های سامانه باشد، در حالی که برای کاربر این ممکن است میزان فایده و استفاده در زمان مناسب باشد.
۶. اعتبار اطلاعات: روایی چگونگی جمع آوری یا تجزیه و تحلیل اطلاعات، شهرت کسی که آن را جمع‌آوری می‌کند یا چگونگی ارائه نتایج میزان اعتبار اطلاعات را مشخص می‌کند.
۷. زیبایی‌شناسی اطلاعات: ویژگی‌های ذهنی کاربران اطلاعات است که به چگونگی ارائه و اشاعه اطلاعات وابسته است (Lau & Moere, 2007).
۸. ارزش درک‌شده از اطلاعات.

1. Marchand

در تمام زنجیره تولید داده، از تولیدکننده پایگاه داده تا مصرف‌کننده نهایی آن، ایجاد همکاری و تعهد همه‌جانبه در سیستم، برای بهبود پیوسته کیفیت ضروری و اجتناب‌ناپذیر است (Russell, Chamberlain & Azzopardi, 2018). از میان مراحل مختلف تولید، کنترل و اشاعه اطلاعات در پایگاه‌های اطلاعاتی علمی، کنترل کیفیت اطلاعات نقش ویژه‌ای در پردازش اطلاعات بر عهده دارند که در این پژوهش به این بخش خواهیم پرداخت. همچنین، با توجه به اینکه به داده‌های پژوهشی فراداده به‌عنوان مهم‌ترین بخش توجه شده است، در بخش بعدی تعریف عمیق‌تری از آن ارائه خواهیم کرد.

فراداده

فراداده به داده‌هایی گفته می‌شود که جزئیات داده‌های دیگر را توصیف می‌کنند. این مفهوم در متون علمی با نام داده‌نما، متادیتا و ابرداده نیز شناخته می‌شود. فراداده شریان حیاتی هر سامانه کتابخانه‌ای است و در دنیایی که برای نمونه فقط در انگلستان سالانه ۴۵۰۰۰ کتاب جدید تولید می‌شود، تنها ابزاری است که جست‌وجوی دانشگاهی را ممکن می‌کند (دیوید و توماس[1]، ۲۰۱۵). استاندارد مارک[2] با وجود محدودیت‌هایی که به‌ویژه در حوزه کیفیت فراداده دارد، امروزه به‌عنوان یک استاندارد پایه برای فراداده‌های کتابخانه‌های دانشگاهی استفاده می‌شود (دیوید و توماس، ۲۰۱۵). با توجه به حجم زیاد داده در بانک‌های اطلاعاتی برای اطمینان از کیفیت داده‌ها نمونه‌برداری و کنترل کیفیت نمونه‌ها و همچنین قضاوت کیفیت داده بر پایه کیفیت نمونه برداشته‌شده کاری اجتناب‌ناپذیر است. اگرچه روی میزان کامل بودن یا سازگاری محتوای داده/ فراداده‌های بانک‌های اطلاعاتی پژوهشی برخی کنترل‌های خودکار وجود دارد، اما دستی بودن برخی گام‌های فرایند به‌ویژه در کار نمایه‌سازی/ ویراستاری داده‌ها اهمیت کنترل کیفیت فراداده را با توجه به اثرگذاری بر کیفیت پژوهش در جامعه علمی بیشتر نمایان می‌کند.

اهداف پژوهش

همان‌گونه که در مقدمه این بخش عنوان شد، اقدامات انجام‌شده در مراحل گوناگون پردازش داده‌های پژوهشی از ورود و ثبت تا نمایه‌سازی، ویراستاری و اشاعه در پایگاه‌های اطلاعاتی، در کیفیت نهایی تأثیرگذار هستند. در صورتی که کنترل کیفیت اثربخش و کارایی در این مراحل انجام شود، می‌تواند در کیفیت داده و فراداده اشاعه‌داده‌شده در پایگاه‌های اطلاعاتی تأثیر بسزایی داشته باشد. در این پژوهش به چارچوب فرایند کنترل کیفیت داده‌ها در مراحل گوناگون پرداخته خواهد شد. با توجه به حجم داده‌های موجود در سامانه‌های اطلاعاتی، تعیین چارچوب مشخصی برای نمونه‌برداری از داده‌ها که بتواند سطح کیفیت نهایی داده را در سامانه‌های اطلاعاتی در وضعیت مطلوبی حفظ کند، نوآوری اصلی این مقاله به حساب می‌آید. اهداف مشخص این پژوهش را می‌توان در قالب موارد زیر خلاصه کرد:

1. David & Thomas
2. Mark

۱. تعیین پارامترهای طرح نمونه‌برداری و روش تصمیم‌گیری در فرایند کنترل کیفیت داده‌های پژوهشی بر اساس تعداد کل مدارک اختصاص داده‌شده به هرکاربر.
۲. تعیین روش تصمیم‌گیری در صورت مشاهده داده‌های ناهم‌خوان در فرایند کنترل کیفیت.

مرور ادبیات

تنوپیر[1] (۱۹۹۲) در مقاله خود با نام «یک روز در زندگی تولیدکننده بانک اطلاعاتی» تجربه خود را در بازدید و مشاهده همه مراحل ایجاد و اشاعه در سه شرکت بزرگ تولید بانک اطلاعاتی ارائه می‌کند. وی این مراحل را که از تأمین‌کننده مدرک آغاز شده و به مشتری (کاربر) مدرک ختم می‌شود، در سه رابطه خلاصه می‌کند. نخستین رابطه میان کسانی است که تعیین می‌کنند چه عناوین (قلم‌های اطلاعاتی) در بانک اطلاعاتی گنجانده شود و افرادی که باید آنها را به دست آورند. رابطه بعدی بین فراهم‌آوران و ناشران وجود دارد و به‌دنبال آن رابطه‌ای بین ناشران و بخش دریافت‌کننده برقرار می‌شود. رابطه سوم (نهایی) میان مشتریان بخش دریافت‌کننده (نمایه‌سازها) و دریافت‌کنندگان است. نمایه‌سازها کاربران، ویراستاران و فهرست‌نویسانی هستند که فهرستی از عناوین موضوع، کلیدواژگان و نمایه‌ها را تهیه می‌کنند. در این فرایند، تأییدکننده‌ها، انسان یا ماشین، دریافت‌کنندگان خروجی نمایه‌سازی هستند. تنوپیر (۱۹۹۲) گروهی از متخصصان کنترل کیفیت را که بر کیفیت کلی مراحل مختلف نظارت دارند و خطاها را تصحیح می‌کنند، شناسایی می‌کند و نقش آنها را در کیفیت نهایی بسیار تأثیرگذار می‌داند. از این رو، به نظر می‌رسد روند کنترل کیفیت برای تولیدکنندگان پایگاه داده به بازرسی بسیار وابسته است. برخی بازرسی‌ها، مانند چک کردن کامل بودن اقلام داده، خطاهای تایپی و فرمت داده‌ها و همچنین دوبارگی‌ها، بر پایه ماشینی‌سازی یا کنترل خودکار به‌کمک نرم‌افزار است. برخی دیگر، مانند کنترلی که کارشناسان نمایه‌ساز روی نمایه‌ها و همچنین اطلاعات کتاب‌شناختی می‌کنند، به‌شکل کامل به‌صورت چشمی و بر پایه کنترل انسانی انجام می‌شود.

از سوی دیگر، متخصصان کنترل کیفیت در نهایی کردن مدرک پیش از اشاعه و همچنین تصحیح خطاهای معرفی‌شده توسط سایر کاربران و کارشناسان نقش ویژه‌ای بر عهده دارند. بنابراین، تعیین و طراحی فرایندهایی برای نظارت بر کیفیت ورودی داده‌ها و درست کردن پیوسته خطاها کاری ضروری است، حتی اگر این فرایندها همیشه به بهبود کیفیت ختم نشود. رویکردهای تضمین کیفیت نیز می‌توانند هم‌زمان با طراحی مناسب فرایندهای کنترل کیفیت مستندات، استفاده از نرم‌افزارها و الگوریتم‌های از پیش طراحی‌شده خطایاب و نیز ممیزی‌های درست، یک رویکرد بهبود مستمر را در سازمان نهادینه کنند. در تضمین کیفیت مستندات می‌توان به شاخص‌هایی مانند کامل بودن، دقت، صحت و سازگاری توجه کرد که در یک فرایند ارزیابی عملکرد پایگاه داده به‌شکل مستمر اندازه‌گیری خواهند شد.

1. Tenopir

متأسفانه، در خصوص فراداده‌ها و کیفیت آنها در محیط دیجیتال پژوهش‌های کمی انجام شده و حتی در مواردی که نیاز به کیفیت فراداده توسط پژوهشگر پیشنهاد شده است، راه‌کارهای اجرایی و خطوط راهنمای مناسبی ارائه نشده است (دیوید و توماس، ۲۰۱۵). در حقیقت، برای تعریف کیفیت فراداده و روش‌هایی که بتوان آنها را ارزیابی و اندازه‌گیری کرده و بهبود بخشید، روش استاندارد پذیرفته‌شده‌ای وجود ندارد (دیوید و توماس، ۲۰۱۵). با این حال، تعریف‌های مفیدی وجود دارد که چارچوب عمومی را برای ارزیابی کیفیت فراداده ارائه می‌دهد. یکی از این تعاریف نتیجه پژوهش بروس و هیلمن[1] (۲۰۰۴) است که ابعاد مختلف کیفیت فراداده‌ها را بررسی کرده‌اند. آنها تعریفی از کیفیت فراداده و ارزیابی آن به‌عنوان یک چارچوب متشکل از هفت بعد کامل بودن، صحت، دقت، اعتبار، سازگاری با نیازها، زمان‌مند بودن و دسترس‌پذیری ارائه کردند. استویلا، گاسر و ویدال[2] (۲۰۰۷) از مدلی عمومی‌تر استفاده کرده و ارزیابی‌های کیفیت را در زمینه (بافت) استفاده از اطلاعات، با تکیه بر کاربردپذیری اطلاعات، ارائه می‌کنند. آنها در فرایند ارزیابی انواع فعالیت‌های کاربران استفاده‌کننده از اطلاعات و هنجارها و ارزش‌های جامعه کاربر اطلاعات را قضاوت می‌کنند. استویلیا و گاسر[3] (۲۰۰۸) ترکیبی از رویکردهای تحلیلی و تجربی ارائه می‌کنند که ارزش ایجادشده از تغییر کیفیت فراداده به‌عنوان پایه و خط مبنا از دیدگاه کاربران نهایی شفاف باشد.

به‌طور مشخص، پارک، توساکا، مازاروس و لو[4] (۲۰۱۰) دریافتند که رایج‌ترین معیارهای ارزیابی کیفیت فراداده از دیدگاه کارشناسان این زمینه، صحت و سازگاری بوده و هر دو نسبت به شاخص کامل بودن اطلاعات کتاب‌شناختی اولویت بالاتری داشتند. پالویتسینیس و همکاران[5] (۲۰۱۴) سازوکارهای تضمین کیفیت و تأثیر آنها را در پایگاه‌های یادگیری دیجیتال اندازه‌گیری کردند. مطالعه موردی آنها نشان داد که قرار دادن نقاط کنترلی مناسب در فرایندهای ایجاد فراداده‌ها در چرخه عمر یک پایگاه داده می‌تواند تا حد زیادی به بهبود منجر شود.

در دنیای نوظهور دیجیتال که در آن کتابخانه‌های دیجیتال نقش کلیدی را هم در مقام انتشاردهنده و هم جمع‌آوری‌کننده بازی می‌کنند، نه فقط کنترل‌کیفیت فراداده‌ها، بلکه ارزیابی آنها نیز مشکل است (دیوید و توماس، ۲۰۱۵). همان‌گونه که هیلمن[6] (۲۰۰۸) می‌نویسد «از آنجا که ساختار قدرتمند بر پایه کیفیت در دنیای استاندارد مارک (شامل استاندارد بالغ، مستندسازی نهادینه‌شده و اطلاعات کتاب‌شناختی تسهیل‌شده) در دنیای فراداده شامل نقص‌هایی است، کاربران فراداده تصمیم به بهبود کیفیت فراداده در کتابخانه‌های خود گرفته‌اند تا بتوانند فرایندهایی را توسعه بدهند که یک همکاری مشترک سودمند را به‌همراه داشته باشد.».

1. Bruce & Hillmann
2. Stvilia & Gasser & Twidale
3. Stvilia & Gasser
4. Park, Tosaka, Maszaros & Lu
5. Palavitsinis et al
6. Hillmann

مشاهده آنچه در صفحه نخست یک مدرک علمی ظاهر می‌شود این موضوع را گواهی می‌دهد که نگرش کاربر نهایی این اطلاعات در دهه گذشته یا قبل‌تر تا به‌امروز تغییر یافته است. همان‌گونه که در بیانیه فدراسیون بین‌المللی انجمن‌های کتابخانه‌ها و مؤسسه‌ها (IFLA)[1] درباره اصول بین‌المللی فهرست‌نویسی گفته شده است، کیفیت فراداده با آسان‌سازی کشف، شناسایی، انتخاب و استفاده از منابع اطلاعاتی مورد نیاز کاربران نهایی تعیین‌شده ارتباط زیادی دارد (Day, Guy & Powell, 2004). با این حال، سخنرانان در وبیناری در سازمان استانداردهای ملی اطلاعات (NISO)[2] در خصوص کتاب‌های الکترونیکی تأکید کردند که فراداده دیگر فقط ویژه استفاده ناشران و کتابداران نیست، بلکه این برای خوانندگان و کاربران نهایی است. کیفیت خوب فراداده خرید و گردش عناوین کتاب‌های الکترونیکی را ترویج می‌کند. خوانندگان می‌خواهند قبل از تصمیم‌گیری برای خواندن آن کتاب الکترونیکی درباره آن بدانند. با نگاه کردن به جلد، دسترسی به فهرست مطالب و کشف حقایقی در خصوص آن نسخه خاص، کاربران قبل از خرید یا قرض گرفتن آن، در عمل کتاب را دنبال می‌کنند. شناخت این واقعیت بدان معنا است که حتی برای ارتقای فروش عناوین کتاب‌های الکترونیکی نیز، فراداده باکیفیت باید هم‌خوان با روش‌های مشخص ساخته و نگهداری شود (پارک و همکاران، ۲۰۱۰). در این راستا، باید به کنترل کیفیت و ابزارهای آماری مرتبط با آن باید به‌شدت توجه شود تا ارزیابی‌های دقیق و صحیح و در چارچوب مشخص بتواند کیفیت خروجی‌های فرایند تولید و اشاعه محتواها را در سطح مطلوبی حفظ کند. از آنجا که موضوع نمونه‌برداری برای پذیرش بخشی مهم و جدایی‌ناپذیر در فرایند کنترل کیفیت است، در ادامه و در بخش سوم به مفاهیم پایه‌ای این حوزه خواهیم پرداخت.

مروری بر مفاهیم نمونه‌برداری برای پذیرش

کنترل کیفیت آماری (SQC)[3] به روش‌های مختلفی انجام می‌شود که یکی از انواع این روش‌ها، نمونه‌گیری برای پذیرش است (Montgomery, 2009). نمونه‌گیری برای پذیرش داده از انباشته‌هایی با اندازه مشخص تشکیل می‌شود که پس از نمونه‌گیری از هر انباشته، در خصوص کیفیت آن انباشته قضاوت می‌شود (Bhave & Sadhwani, 2021). یکی از استانداردها به‌منظور استفاده از نمونه‌گیری برای پذیرش استاندارد Mil-STD-105-E است. در این روش، داده‌ها، در انباشته‌هایی که تعداد محدودی دارند، دسته‌بندی می‌شوند (Montgomery, 2009).

روش‌های نمونه‌گیری برای پذیرش روش‌هایی برای ارزیابی انباشته هستند، نه برآورد یا ایجاد کیفیت انباشته. همچنین چون فقط به بازرسی می پردازند، کیفیت ایجاد نمی‌کنند. از استانداردهای مهمی که در این زمینه استفاده می‌شوند، استاندارد Mil-STD-105-E است که با توجه به سطح کیفیت قابل قبول (AQL) طراحی شده است (Schilling & Neubauer, 2009). به‌منظور تقسیم‌بندی طرح‌های نمونه‌گیری

1. International Federation of Library Associations and Institutions
2. National Information Standards organization
3. Statistical Quality Control

برای پذیرش، روش‌های مختلفی وجود دارد. یکی از انواع این تقسیم‌بندی‌ها شامل طرح‌های یک بار، دوبار و چندبار نمونه‌گیری است.

در طرح‌های یک بار نمونه‌گیری معیار پذیرش انباشته، نتایج فقط بر اساس یک نمونه تصادفی با اندازه مشخص است، در حالی که طرح‌های دوبار یا چندبار نمونه‌گیری می‌توانند بر اساس بازرسی دو یا چند نمونه انجام شوند. در واقع، یک طرح یک بار، روشی برای ارزیابی کیفیت انباشته است که با اندازه نمونهٔ n تایی از انباشته و عدد پذیرش c مشخص می‌شود (Schilling & Neubauer, 2009). با استفاده از این روش ابتدا یک نمونه n تایی به‌صورت تصادفی انتخاب می‌شود، اگر به تعداد عدد پذیرش c یا کمتر، محصول معیوب در نمونه مشاهده شد، انباشته پذیرفته و اگر بیش از عدد c مشاهده شد، انباشته رد می‌شود.

با افزایش تعداد مراحل نمونه‌برداری، از یک به دو و چند بار نمونه‌برداری، اگرچه به‌طور میانگین اندازه نمونه کاهش می‌یابد، اما پیچیدگی فرایند نمونه‌برداری نیز از جنبه‌های کاربردی و عملیاتی افزایش خواهد یافت. از سوی دیگر، از آنجا که تصمیم‌گیری در خصوص کیفیت انباشته در داده‌ها به‌دلیل تنوع ویژگی‌ها و اقلام اطلاعاتی در ذات خود کار دشواری است چند مرحله کردن فرایند نمونه‌برداری دشواری و پیچیدگی مراحل کنترل کیفیت را دوچندان خواهد کرد. از این رو، در این پژوهش از روش‌های یک بار نمونه‌برداری بهره خواهیم برد.

در طراحی روش‌های نمونه‌برداری ریسک مصرف‌کننده و ریسک تولیدکننده از پارامترهای کلیدی هستند که باید به‌درستی تعریف شوند. ریسک مصرف‌کننده احتمال پذیرش یک بهر بد یا غیر قابل قبول است. مقدار این ریسک معمولاً ۰/۱ تعیین می‌شود. در رابطه با ریسک مصرف‌کننده، یک تعریف عددی بهر معیوب وجود دارد که LTPD یا درصد رواداری اقلام معیوب بهر نامیده می‌شود. احتمال پذیرش بهر با LTPD درصد معیوب، βدرصد خواهد بود. ریسک تولیدکننده α احتمال رد شدن یک بهر خوب یا قابل قبول است که اغلب برای این ریسک مقدار ۰/۰۵ در نظر گرفته می‌شود. در رابطه با ریسک تولیدکننده یک تعریف عددی درصد اقلام معیوب بهر خوب که AQL نامیده می‌شود، وجود دارد. AQL حداکثر درصد اقلام معیوبی است که می‌تواند برای نمونه‌گیری به‌منظور پذیرش رضایت‌بخش باشد. α درصد شانس رد شدن محموله با کیفیت AQL خواهد بود. تابع مشخصهٔ عملکرد (OC)[1] احتمال پذیرش را به‌عنوان تابعی از p (در اینجا نسبت ناهم‌خوانی جامعه) نشان می‌دهد. نمودار این تابع به ما کمک می‌کند تا احتمال رد یا قبول محموله‌ای را که نسبت اقلام ناهم‌خوان خاصی دارد، پیدا کنیم.

می‌دانیم اگر X تعداد اقلام معیوب نمونه nتایی باشد که از جامعه‌ای نامتناهی گرفته شده است (N را خیلی بزرگ در نظر می‌گیریم) آنگاه، متغیر تصادفی X از توزیع بینم (دوجمله‌ای) پیروی خواهد کرد. اگر احتمال هم‌خوان بودن یک نمونه را p و احتمال ناهم‌خوان بودن آن را q بنامیم، آنگاه احتمال مشاهده x ناهم‌خوان در نمونه nتایی به‌شرح زیر خواهد بود.

1. Operating Characteristic

رابطهٔ ۱)

$$X \sim b(n,p)$$
$$P(X=x) = \binom{x}{n} \times p^x \times (1-p)^{n-x}$$

در این صورت احتمال پذیرش عبارت است از:

رابطهٔ ۲)

$$P_a = P(x \le c) = \sum_{x=0}^{c} \binom{n}{x} p^x q^{n-x}$$

حال با داشتن مقادیر احتمال پذیرش، می‌توان منحنی OC را به‌ازای نسبت‌های ناهم‌خوانی گوناگون به‌دست آورد. این تابع را می‌توان به‌عنوان تابعی از p در نظر گرفت، یعنی n وc را معلوم فرض کرده P_a را بر حسب p رسم کرد. در نمودار OC یک پارامتر مهم دیگر LTPD[1] است. این پارامتر پایین‌ترین سطح کیفیت را نشان می‌دهد که مصرف‌کننده در یک بهر مجاز و قابل قبول می‌داند. LTPD را نسبت به اقلام ناهم‌خوان مجاز محموله می‌نامند و آن را با RQL[2] به‌معنای سطح کیفیت قابل رد و LQL[3] به‌معنای سطح کیفیت حدی هم می‌دانند. برای یافتن یک طرح نمونه‌برداری مناسب پیش از یافتن متغیرهای مناسب لازم است برخی پارامترهای کلیدی مشخص شوند. در ادامه به این پارامترها اشاره خواهیم کرد.

- خطای نوع I: در چارچوب ارائه‌شده در این بخش این خطا میزان احتمال بازگرداندن انباشته‌ای از پارسا است که دارای سطح AQL است. در منابع علمی خطای نوع I را برابر ۰/۰۵در نظر می‌گیرند (Montgomery, 2009).
- خطای نوع II: در چارچوب ارائه‌شده در این بخش این خطا میزان احتمال پذیرش انباشته‌ای از پارسا است که دارای سطح LTPD است. در منابع علمی خطای نوع II را برابر ۰/۱ در نظر می‌گیرند (Montgomery, 2009).

در این مقاله در قالب یک مطالعه موردی روشی برای نمونه‌برداری وکنترل کیفیت داده‌های پژوهشی (مدارک علمی) بر پایه طراحی‌های آماری ارائه خواهیم کرد. در ادامه روش و گام‌های اصلی این پژوهش را معرفی خواهیم کرد.

ارائه روش پیشنهادی برای بازرسی انباشته‌های داده‌ای

نمونه‌برداری برای پذیرش، بخش اصلی و ویژه‌ای از ماهیت کنترل کیفیت را به خود اختصاص می‌دهد. اگرچه معمولاً از نمونه‌برداری به‌منظور پذیرش مواد دریافتی (ورودی کالا به سیستم) استفاده می‌شود، اما موارد دیگری نیز برای استفاده از آن وجود دارد. برای نمونه، ممکن است یک تولیدکننده کالای نیمه‌ساخته‌ای را به مرحله بعدی تولید ارسال کند، در حالی که محصولات نیمه‌کاره در این مرحله مردود شده، دوباره‌کاری روی آنها انجام شده یا دور ریخته شوند. از این رو، نمونه‌برداری ابزاری مناسب برای کنترل کیفیت یا ارزیابی سطح کیفیت در مراحل گوناگون شکل‌گیری محصول است.

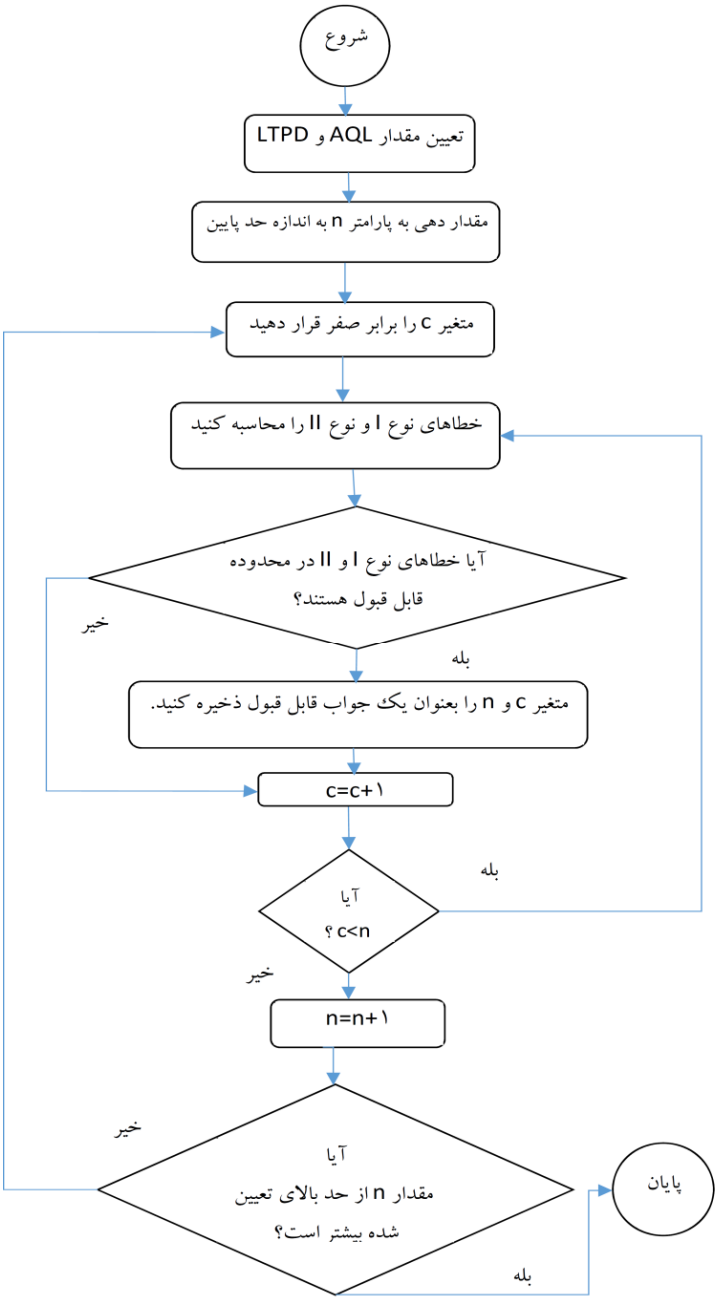
1. Lot Tolerance Percent Defective
2. Rejectable Quality Level
3. Limiting Quality Level

در بخش سوم ریسک مصرف‌کننده و همچنین ریسک تولیدکننده، به‌عنوان دو عامل پایه‌ای در طراحی روش‌های نمونه‌برداری معرفی شدند. این دو پارامتر در کنترل کیفیت آماری با عنوان خطاهای نوع I و نوع II شناخته می‌شوند. از سوی دیگر، دو پارامتر کلیدی با نام‌های AQL و LTPD معرفی شدند و روابط ریاضی تأثیرگذار در محاسبه آنها ارائه شد. با در نظر گرفتن این پارامترها و مقداردهی به آنها می‌توان چارچوب کلان مراحل یافتن یک طرح نمونه‌برداری را ارائه کرد. در این بخش و در قالب هفت گام مراحل یادشده معرفی خواهند شد (شکل ۱).

شکل ۱ . گام‌های اصلی یافتن طرح نمونه‌برداری در فرایندهای نمایه‌سازی و ویراستاری

در شکل ۲ به گام‌های الگوریتمی اشاره شده است که به تعیین پارامترهای طرح نمونه‌برداری می‌انجامد. گام‌های این الگوریتم بر پایه روش مک‌ویلیامز و همکاران[1] توسعه داده شده است.

1. McWilliams et al



شکل ۲. الگوریتم تعیین پارامترهای طرح نمونه‌برداری

همان‌گونه که توضیح داده شد، پیش از اجرای این الگوریتم لازم است برخی پارامترها مقداردهی شوند که این کار بر پایه بازخوردهای دریافت‌شده و گزارش‌ها تهیه شده و بر پایه سطح کیفیت مورد انتظار انجام خواهند شد. به بیان دیگر، پس از مشخص شدن AQL و LTPD از یک سو و مقداردهی خطای نوع I و خطای نوع II به‌ترتیب ۰/۰۵ و ۰/۱ است. از سوی دیگر، شروع به اجرای الگوریتم خواهیم کرد. شایان ذکر است، همان گونه که در شکل ۳ مشاهده می‌شود، باید در هر مرحله، پس از تنظیم مقادیر n و c خطاهای نوع I و نوع II حساب شوند. برای این کار بر پایه توضیحات ارائه‌شده در این بخش از توزیع بینم استفاده خواهد شد.

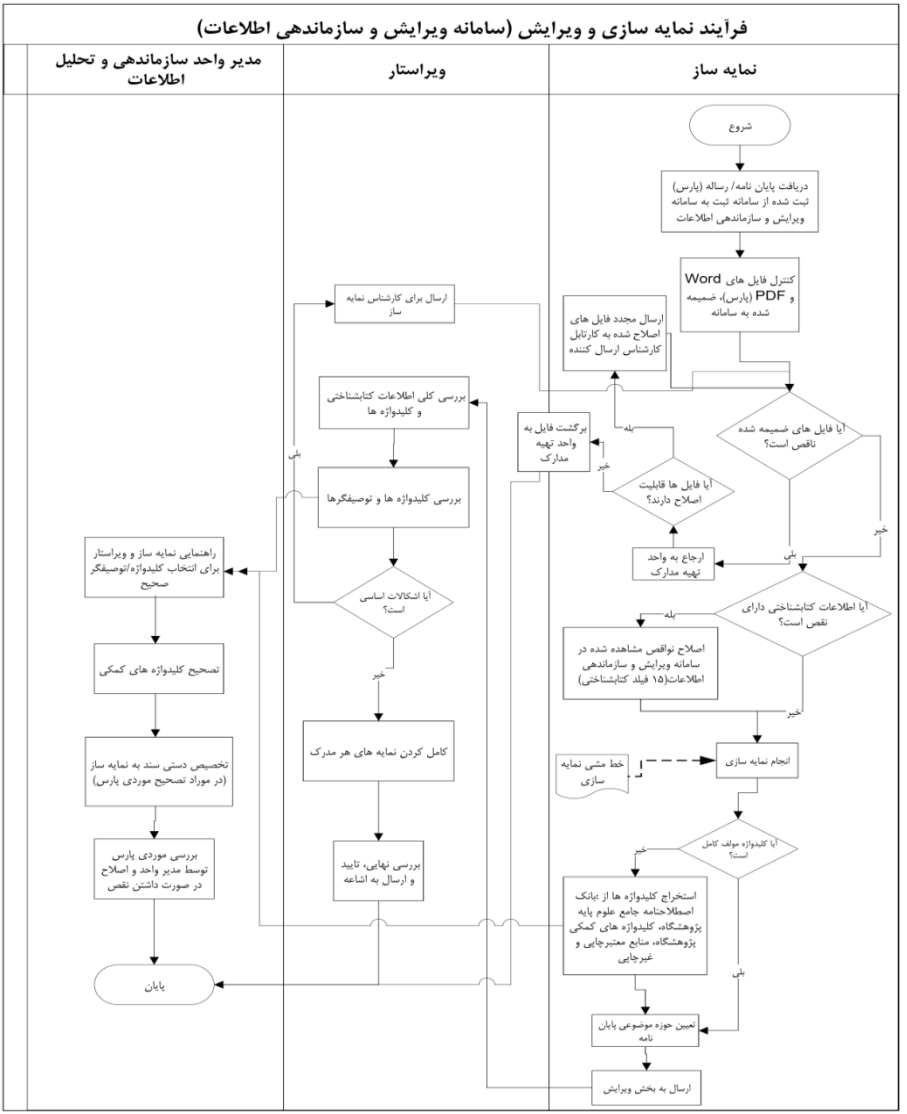
پیش از اجرای الگوریتم ارائه‌شده در شکل ۲، باید پارامترهای مسئله تعیین و مقداردهی شوند. پس از مقداردهی به پارامترها می‌توان الگوریتم‌های تعیین اندازه نمونه را اجرا کرد. گفتنی است که در پایان اجرای این الگوریتم جواب‌های گوناگونی به‌دست خواهد آمد. در نهایت، پاسخی انتخاب خواهد شد که با خطاهای نوع اول و دوم از پیش تنظیم‌شده، نزدیک‌ترین فاصله را داشته باشد.

مطالعه موردی

پایگاه گنج، گنجینه‌ای با ارزش از اطلاعات علمی و فراداده‌ای اساتید، دانشجویان و پژوهشگران ایرانی است. این پایگاه با وجود کارایی و اثربخشی در ثبت و اشاعه اطلاعات علمی به کاربران، با معایب و اشکالاتی مواجه است و به بهینه‌سازی نیاز دارد. بخشی از نابسامانی‌ها در این سامانه ناشی از خطاهای انسانی است که به‌هنگام نمایه‌سازی و ورود اطلاعات در بخش سازمان‌دهی اطلاعات انجام شده است (ارشادی، رجبی، شیرانی و رضایی، ۱۳۹۵). بخشی دیگر از ایرادها، منشأ رایانه‌ای و سیستمی دارند که به‌مرور زمان و با تغییر نرم‌افزارها و سخت‌افزارها و کاراکترها و در هنگام تبدیل‌های مختلف و ورود ماشینی اطلاعات در پایگاه به وجود آمده است. فرایند فراهم‌آوری، سازمان‌دهی و اشاعه اطلاعات علمی که به تولید پایگاه گنج منجر می‌شود، فرایندی جاری، پیوسته و در حال انجام است (ارشادی و همکاران، ۱۳۹۵). موفقیت یا شکست پایگاه‌های اطلاعاتی تا حد زیادی با کیفیت داده‌های موجود به‌عنوان مبنایی برای برنامه‌های کاربردی آن پایگاه‌ها در ارتباط است (Makeleni & Cilliers, 2021). بنابراین، بخش جدایی‌ناپذیری از برنامه‌های بهبود هر پایگاه، ادغام داده‌های حاصل از سیستم‌های عامل است (Vliegen, Moroff & Riehl, 2020). قبل از شروع فرایند ادغام یک سیستم منبع، به تحلیل غنی داده‌های منبع نیاز است. ارشادی و احترامی (۱۳۹۵) در پژوهش خود فرایند کنترل کیفیت (نمایه‌سازی و ویرایش) را که یکی از مراحل مهم در چرخه تولید و اشاعه پارساها به حساب می‌آید، مستند کردند (شکل ۳).

در فرایند کنترل کیفیت روی داده‌های پژوهشی، کنترل‌های مختلفی می‌شود که در وضعیت کنونی برای نمونه‌برداری‌های انجام‌شده و همچنین اقدامات بعدی آن رویه مشخصی وجود ندارد. نداشتن خط‌مشی مشخص در نمونه‌برداری، کنترل‌های انجام شده و تصمیم‌گیری‌های بعدی به این منجر خواهد شد که کارایی کنترل‌ها کاهش پیدا می‌کند، در نتیجه، برخی خطاها با وجود کنترل‌های پی‌درپی در سامانه گنج مشاهده می‌شود. همچنین، مخاطره دیگری که نمونه‌برداری‌های ۱۰۰ درصد و پیاپی دارد،

اطمینان کارشناس مرحله قبلی به عملکرد مرحله بعدی و در نتیجه، افت کارایی کنترل در مراحل قبلی خواهد بود. از این رو، وجود طرح‌های نمونه‌برداری مشخص می‌تواند کارایی کنترل‌های هر مرحله را افزایش دهد. در ادامه و در بخش ششم، نتایج پژوهش ارائه خواهد شد.



شکل۳. فرایند نمایه‌سازی و ویرایش پارسا

نتایج

در این بخش، با توجه به اینکه گوناگونی در ناهم‌خوانی‌ها در هر قلم اطلاعاتی داده پژوهشی شایان توجه بود، برای هر قلم اطلاعاتی اشاره‌شده در بخش قبلی چارچوبی برای دسته‌بندی ناهم‌خوانی‌ها ارائه شد. در ادامه، در جدول ۱ به‌صورت جداگانه به‌ازای هر قلم اطلاعاتی دسته‌بندی ناهم‌خوانی‌ها ارائه شده است. (گام نخست از شکل ۱).

جدول ۱. ناهم‌خوانی‌های بالقوه در فرایند کنترل کیفیت مدارک

ردیف	نام قلم اطلاعاتی	دسته ناهم‌خوانی	چگونگی ثبت و کنترل ناهم‌خوانی
۱	عنوان فارسی	مشکل ویرایشی/ نگارشی	ناهم‌خوانی ویرایشی و نگارشی: مانند نیم‌فاصله، د به‌جای ذ، ر به جای ز، فاصله اضافه و ترکیب شیمیایی، فرمول ریاضی (توان و علائم ریاضی)
		بحرانی (صحت/ ناهمخوانی با متن)	ایراد اصلی مانند ناهم‌خوانی با متن پایان‌نامه و ...
۲	عنوان لاتین	مشکل ویرایشی/ نگارشی	در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		بحرانی (صحت/ ناهم‌خوانی با متن)	در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
۳	پدیدآوران	دانشگاه/ دانشکده	در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		دانشجو/ اساتید	
۴	تاریخ دفاع	-	در صورت هم‌خوانی با صفحه عنوان، درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
۵	رشته	-	در صورت هم‌خوانی با صفحه عنوان، درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
۶	گرایش	-	در صورت هم‌خوانی با صفحه عنوان، درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
۷	چکیده فارسی	ناهم‌خوانی ویرایشی	ناهم‌خوانی ویرایشی و نگارشی: مانند نیم‌فاصله، د به‌جای ذ، ر به‌جای ز، فاصله اضافه و ....
		ناهم‌خوانی نگارشی	در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		وجود کلمات اضافی	در صورت مشاهده کلمات اضافی (مانند چکیده، کلیدواژه، اطلاعات دانشجو و غیره) در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		ناهم‌خوانی با متن پایان‌نامه	در صورت تأیید درج عبارت «هم‌خوانی دارد» در صورت ناهم‌خوانی درج عبارت «هم‌خوانی ندارد»

ردیف	نام قلم اطلاعاتی	دسته ناهم‌خوانی	چگونگی ثبت و کنترل ناهم‌خوانی
		اشکال در اعشار/ فرمول‌ها و روابط ریاضی/ شیمی	در صورت مشاهده اشکال درج عبارت «مشکل دارد» و در صورت ناهم‌خوانی درج عبارت «مشکل ندارد»
۸	چکیده لاتین	ناهم‌خوانی ویرایشی	ناهم‌خوانی ویرایشی و نگارشی: مانند نیم‌فاصله، د به‌جای ذ، ر به‌جای ز، فاصله اضافه و ....
		ناهم‌خوانی نگارشی	در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		وجودکلمات اضافی	در صورت مشاهده کلمات اضافی (مانند چکیده، کلیدواژه، اطلاعات دانشجو و غیره) در صورت تأیید درج عبارت «مشکل ندارد» و در صورت ناهم‌خوانی درج عبارت «مشکل دارد»
		ناهم‌خوانی با متن پایان‌نامه	در صورت تأیید درج عبارت "هم‌خوانی دارد" و در صورت ناهم‌خوانی درج عبارت «هم‌خوانی ندارد»
		اشکال در اعشار/ فرمول‌ها و روابط ریاضی/ شیمی	در صورت مشاهده اشکال درج عبارت «مشکل دارد» و در صورت هم‌خوانی درج عبارت «مشکل ندارد»
۹	فهرست مندرجات	-	در صورت هم‌خوانی با متن و همچنین مشاهده نشدن مشکل عبارت «مشکل ندارد» درج شود. در صورت خالی بودن قلم اطلاعاتی درج عبارت «موجود نیست». در صورت ناقص بودن قلم اطلاعاتی در مقایسه با متن پارسا درج عبارت «ناقص است» و همچنین در صورت به‌هم چسبیده بودن درج عبارت «به‌هم‌چسبیده» درج شود. عبارت «جابه‌جایی فهرست فارسی و لاتین» در صورت مشاهده ناهم‌خوانی جابه‌جایی میان فهرست‌های فارسی و لاتین.
۱۰	فهرست منابع فارسی	-	
۱۱	فهرست منابع لاتین	-	

همان‌گونه که در روش پژوهش توضیح داده شد، برای تعیین پارامترهای روش نمونه‌برداری، افزون بر مقداردهی به پارامتر خطای نوع I و خطای نوع II پارامترهای AQL و LTPD نیز تأثیرگذارند (گام دو). این مقادیر توسط خبرگان حوزه علم داده که به فرایندهای پردازش، نمایه‌سازی و اشاعه داده‌های پژوهشی تسلط کافی دارند، تعیین شد (جدول ۲). اهمیت هر مشخصه کیفی نیز در تعیین پارامترهای کلیدی نمونه‌برداری تأثیرگذار است. مقادیر اهمیت هر مشخصه شامل مقادیر A، B و C بوده و به‌صورت زیر تعریف می‌شوند.

مشخصه نوع A

- این ناهم‌خوانی بحرانی بوده و به صحت پارسا مرتبط است. همچنین ممکن است در نمایش یک داده بی‌کیفیت که تأثیرگذاری شدید بر دیدگاه کاربران دارد، در سامانه گنج منجر شود.

مشخصه نوع B

- این ناهم‌خوانی مهم بوده و به دقت پارسا مرتبط است. همچنین ممکن است در نمایش یک داده بی‌کیفیت که تأثیرگذاری شدید بر دیدگاه کاربران دارد، در سامانه گنج منجر شود.

مشخصه نوع C

- این ناهم‌خوانی جزئی است یا تأثیر زیادی در رضایت کاربران در سامانه گنج ندارد.

جدول ۲. مقداردهی به پارامترهای AQL و LTPD بر پایه نوع مشخصه کیفی

ردیف	نام قلم اطلاعاتی	دسته ناهم‌خوانی	نوع مشخصه کیفی	AQL	LTPD
۱	عنوان فارسی	مشکل نگارشی	A	۰/۰۰۵	۰/۰۵
		مشکل اصلی (صحت/ ناهم‌خوانی با متن و ...)	A	۰/۰۰۵	۰/۰۵
۲	عنوان لاتین	مشکل ویرایشی/ نگارشی	B	۰/۰۱	۰/۰۸
		مشکل اصلی (صحت/ ناهم‌خوانی با متن و ...)	B	۰/۰۱	۰/۰۸
۳	پدیدآوران	دانشگاه	A	۰/۰۰۵	۰/۰۵
		دانشکده	B	۰/۰۱	۰/۰۸
		دانشجو	A	۰/۰۰۵	۰/۰۵
		اساتید راهنما	A	۰/۰۰۵	۰/۰۵
		اساتید مشاور	A	۰/۰۰۵	۰/۰۵
۴	تاریخ دفاع	-		۰/۰۳	۰/۱
۵	رشته	-	B	۰/۰۱	۰/۰۸
۶	گرایش	-	B	۰/۰۱	۰/۰۸
۷	چکیده فارسی	وجودکلمات اضافی (مانند کلمه چکیده در ابتدا)	A	۰/۰۰۵	۰/۰۵
		ناهم‌خوانی با متن پایان‌نامه	A	۰/۰۰۵	۰/۰۵
۸	چکیده لاتین	وجودکلمات اضافی (مانند کلمه چکیده در ابتدا)	B	۰/۰۱	۰/۰۸
		ناهم‌خوانی با متن پایان‌نامه	B	۰/۰۱	۰/۰۸
۹	فهرست مندرجات	-	C	۰/۰۳	۰/۱
۱۰	فهرست منابع فارسی	-	C	۰/۰۳	۰/۱
۱۱	فهرست منابع لاتین	-	C	۰/۰۳	۰/۱

در ادامه و بر پایه گام سوم پژوهش، حدود قابل قبول برای پارامتر n با توجه به بررسی و کنترل کیفیت ۷۰ مدرک در روز و ۱۰ روز کاری، معادل ۷۰۰ مدرک (داده پژوهشی) تعیین شد. اکنون با داشتن همهٔ پارامترهای پیشین یک طرح نمونه‌برداری، می‌توان الگوریتم شکل ۲ را پیاده‌سازی کرد (گام چهارم). در ادامه و در جدول ۳، نتایج طرح نمونه‌برداری مشاهده خواهد شد (گام پنجم).

جدول ۳. طرح نمونه‌برداری مناسب برای مراحل گوناگون ثبت و سازمان‌دهی اطلاعات

ردیف	نام مرحله	مرحله کنترل‌کننده	نوع مشخصه	AQL	LTPD	n	c	r
	ثبت مدارک پژوهشی	کنترل کیفیت داده	A	۰/۰۰۵	۰/۰۵	۱۰۵	۲	۳
			B	۰/۰۱	۰/۰۸	۶۵	۲	۳
			C	۰/۰۳	۰/۱	۱۱۶	۷	۸

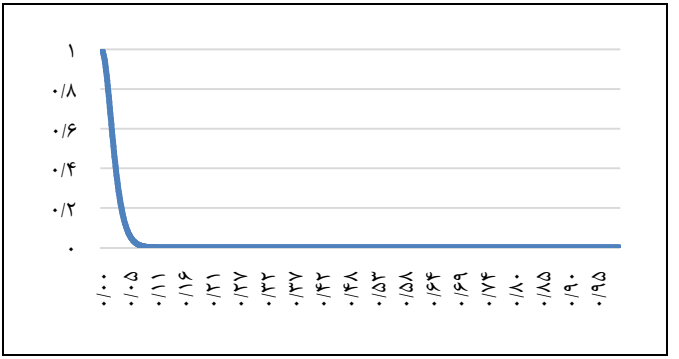
همان گونه که در جدول ۳ مشاهده می‌شود، به‌ازای هر گروه از مشخصه‌های کیفی طرح نمونه‌برداری تعیین شده است. برای نمونه، در مرحلهٔ کنترل کیفیت داده، کارشناس از ۱۰ روز کاری پیشین، ۱۰۵ نمونه انتخاب خواهد کرد. در صورت مشاهده بیشینهٔ ۲ ناهم‌خوانی در هر مشخصهٔ نوع A، حجم کار انجام‌شده تأیید خواهد شد و در صورت مشاهدهٔ ۳ یا بیش از ۳ ناهم‌خوانی، کار تأیید نخواهد شد. عددهای پذیرش و رد در مشخصه‌های نوع B، به‌ترتیب ۲ و ۳ در مشخصه‌های نوع C، به‌ترتیب ۷ و ۸ خواهد بود.

منحنی OC برای طرح‌های نمونه‌برداری

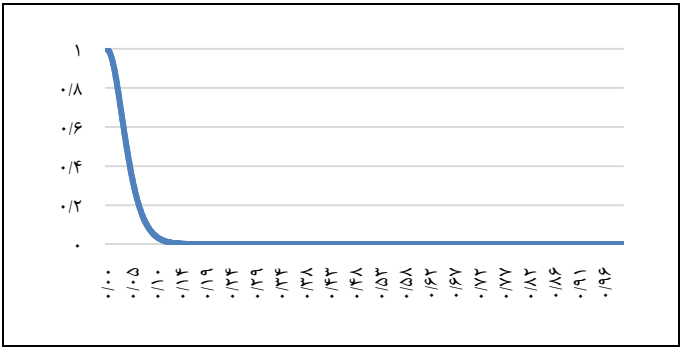
در این بخش منحنی OC برای سه دسته اصلی A، B و C طراحی‌شده در فرایند نمونه‌برداری ارائه شده است. در ادامه و به‌ترتیب در شکل‌های ۴ تا ۶ منحنی OC را برای گروه‌های مشخصه A تا C می‌توان مشاهده کرد.

همان‌گونه که در این شکل‌ها دیده می‌شود، منحنی OC شیب مطلوبی دارد، یعنی در فاصله بین AQL تا LTPD این منحنی با شیب مناسبی و به‌سرعت به پایین حرکت می‌کند. این نشان می‌دهد که طرح نمونه‌برداری ارائه‌شده می‌تواند بچ‌های مناسب را از بچ‌های نامناسب تمیز دهد.

در این بخش عنوان شد که مقادیر AQL و LTPD برای مشخصه‌های کیفی گروه A به‌ترتیب عبارت‌اند از ۰/۰۰۵ و ۰/۰۵. همان گونه که در شکل ۴ می‌بینید، برای مقادیر کمتر از ۰/۰۰۵ احتمال پذیرش مدارک ارائه‌شده تا میزان شایان توجهی بالاست. از سوی دیگر، برای مقادیر بیش از ۰/۰۵ احتمال پذیرش به صفر میل می‌کند. این نشان می‌دهد که طرح نمونه‌برداری ارائه‌شده برای اقلام اطلاعاتی و مشخصه‌های کیفی گروه A از اعتبار لازم برخوردار است.



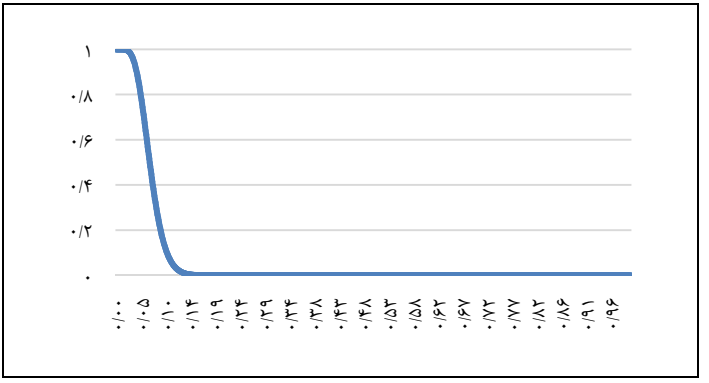
شکل ۴. منحنی OC برای طرح نمونه‌برداری مشخصه‌های نوع A



شکل ۵. منحنی OC برای طرح نمونه‌برداری مشخصه‌های نوع B

همچنین مقادیر AQL و LTPD برای مشخصه‌های کیفی گروه B به‌ترتیب عبارت‌اند از: ۰/۰۱ و ۰/۰۸. همان گونه که در شکل ۵ می‌بینید، برای مقادیر کمتر از ۰/۰۱ احتمال پذیرش مدارک ارائه‌شده به میزان شایان توجهی بالاست. از سوی دیگر، برای مقادیر بیش از ۰/۰۸ احتمال پذیرش به صفر میل می‌کند. این نشان می‌دهد که طرح نمونه‌برداری ارائه‌شده برای اقلام اطلاعاتی و مشخصه‌های کیفی گروه B از اعتبار لازم برخوردار است.

همچنین مقادیر AQL و LTPD برای مشخصه‌های کیفی گروه C به‌ترتیب عبارت‌اند از: ۰/۰۳ و ۰/۱. همان گونه که در شکل ۶ می‌بینید، برای مقادیر کمتر از ۰/۰۳ احتمال پذیرش مدارک ارائه‌شده به میزان شایان توجهی بالاست. از سوی دیگر، برای مقادیر بیش از ۰/۱ احتمال پذیرش به صفر میل می‌کند. این نشان می‌دهد که طرح نمونه‌برداری ارائه‌شده برای اقلام اطلاعاتی و مشخصه‌های کیفی گروه C از اعتبار لازم برخوردار است.



شکل ۶. منحنی OC برای طرح نمونه‌برداری مشخصه‌های نوع C

نتیجه‌گیری و ارائه رهنمودهای کاربردی

در این پژوهش بر پایه روش‌ها و تحلیل‌های آماری، یک چارچوب نمونه‌برداری در فرایند کنترل کیفیت داده ارائه شده است. همان‌گونه که در توضیح‌های ارائه‌شده در بخش قبل نیز عنوان شد، با توجه به سطح کیفیت قابل قبولی که مدارک پس از ثبت دارند وجود یک طرح نمونه‌برداری می‌تواند کارایی فرایند را به سطح مطلوب نزدیک‌تر کند. بر این اساس، طرح نمونه‌برداری به‌گونه‌ای ارائه شد که خطاهای نمونه‌برداری را کاهش داده و از کیفیت مطلوب مدارک اطمینان مناسب حاصل شود. در زمان پیاده‌سازی طرح نمونه‌برداری ارائه‌شده باید به نکات زیر توجه شود.

- در طراحی روش نمونه‌برداری برای ناهم‌خوانی‌ها سه سطح اهمیت در نظر گرفته شد. این سه سطح، از مشخصه‌های کیفی بحرانی (A) تا مشخصه‌های اصلی (B) و ناهم‌خوانی‌های جزئی (C) دسته‌بندی شد. دسته‌بندی ارائه‌شده ممکن است در گذر زمان توسط کارگروه کیفیت بازبینی شده و در مشخصه‌های هر دسته جابه‌جایی انجام شود. پیشنهاد می‌شود این کار سالیانه انجام شود.
- میزان AQL و LTPD تعیین‌شده به هر دسته از مشخصه‌های کیفی در کارگروه کیفیت پارسا و بر پایه اهمیت هر دسته در کیفیت نهایی مدارک اشاعه داده‌شده تعیین شد. از آنجا که مقادیر تعیین‌شده در اندازه نمونه نهایی، عدد پذیرش و همچنین عدد رد نقش بسزایی دارد، در صورت بازبینی دسته‌بندی‌های A تا C میزان مقادیر AQL و LTPD متناظر با آنها نیز باید دوباره بررسی شود.

از آنجا که هرگونه بازبینی در مقادیر AQL و LTPD در خطاهای نوع I و II اثرگذار است، لازم است الگوریتم‌های توسعه داده‌شده در این پژوهش برای مقادیر تازه AQL و LTPD نیز اجرا شود. بدیهی است، نتایج اجرای الگوریتم مانند اندازه نمونه، عدد پذیرش و نیز عدد رد در این فرایند به‌روزرسانی خواهند شد.

فهرست منابع

اثنی عشری، حمیده و اسدی، غلامحسین (۱۳۹۴). معیارهای سنجش کیفیت اطلاعات حسابداری و بازده اضافی مطلق. *دانش حسابداری مالی*، ۲(۴)، ۴۷- ۷۰.

ارشادی، محمدجواد و احترامی، تینا (۱۳۹۵). *طرح‌ریزی و استقرار نظام تضمین کیفیت در سامانه‌های گردآوری و ثبت، سازمان‌دهی و اشاعه اطلاعات پایان‌نامه‌ها/رساله‌های دانش‌آموختگان داخل کشور*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران.

ارشادی، محمدجواد؛ رجبی، تقی؛ شیرانی، فرهاد و رضایی، نسا. (۱۳۹۵). کاربرد تکنیک تحلیل ریشه در حل مشکلات کیفی سامانه‌های اطلاعاتی تحقیقاتی: مطالعه موردی سامانه اشاعه اطلاعات پایان نامه‌ها/رساله‌های دانش‌آموختگان داخل کشور (گنج). *مدیریت اطلاعات*، ۱(۱-۲)، ۷۵-۸۹.

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Cham, Switzerland: Springer International Publishing.

Bhave, P. P., & Sadhwani, K. (2021). Sampling in environmental matrices: a critical review. *Environmental Forensics*, 23(1-2), 75- 92.

Bruce, T. R., & Hillmann, D. I. (2004). *The continuum of metadata quality: defining, expressing, exploiting*. ALA editions.

Cárdenas-García, J. F., De Mesa, B. S., & Castro, D. R. (2019). Understanding Globalized Digital Labor in the Information Age. *Perspectives on Global Development and Technology,* 18(3), 308-326.

David, R.H. & Thomas, D. (2015) Assessing Metadata and Controlling Quality in Scholarly Ebooks. *Cataloging & Classification Quarterly*, 53(7), 801-824.

Day, M., Guy, M., & Powell, A. (2004). *Improving the quality of metadata in eprint archives*. Ariadne, 38.

Hillmann, D. I. (2008). Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, 46(1), 65-80.

Lau, A., & Moere, A. V. (2007, July). Towards a model of information aesthetics in information visualization. *In 2007 11th International Conference Information Visualization (IV'07)* (pp. 87-92). IEEE.

Makeleni, N., & Cilliers, L. (2021). Critical success factors to improve data quality of electronic medical records in public healthcare institutions. *South African Journal of Information Management,* 23(1), 1-8.

Marchand, D. (1990). Managing information quality. *In: Wormell I.(ed.) Information Quality: Definitions and Dimensions.* Taylor Graham, Londres.

McWilliams, T. P., Saniga, E. M., & Davis, D. J. (2001). On the design of single sample acceptance sampling plans. *Stochastics and Quality Control*, 16(2), 193-198.

Montgomery, D.C. (2009). *Design and analysis of experiments* (7th ed.). John Wiley & Sons, Inc., New York.

Palavitsinis, N., Manouselis, N., & Sanchez-Alonso, S. (2014). Metadata quality in learning object repositories: a case study. *The Electronic Library*, 32(1).

Park, J. R., Tosaka, Y., Maszaros, S., & Lu, C. (2010). From metadata creation to metadata quality control: Continuing education needs among cataloging and metadata professionals. *Journal of education for library and information science*, 158-176.

Price, R., & Shanks, G. (2016). A semiotic information quality framework: development and comparative analysis. I*n Enacting Research Methods in Information Systems* (pp. 219-250). Palgrave Macmillan, Cham.

Russell, R.T., Chamberlain, J. & Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management,* 54(6), 1042-1057.

Schilling, E. G., & Neubauer, D. V. (2009). *Acceptance sampling in quality control*. Chapman and Hall/CRC.

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. Library & *Information Science Research*, 30(1), 67-74.

Stvilia, B., Gasser, L. & Twidale, M. B. (2007). Metadata quality problems in federated collections. *In Challenges of Managing Information Quality in Service Organizations* (pp. 154-186). IGI Global.

Tenopir, C. (1992). A Day in the Life of a Database Producer. *Database,* 15(3), 15-17.

Vliegen, L., Moroff, N. U., & Riehl, K. (2020, September). Evaluation of data quality in dimensioning capacity. *In Hamburg International Conference of Logistics (HICL)* 2020 (pp. 355-394).

Wilkinson, L. (2006). Revising the Pareto chart. *The American Statistician*, 60 (4), 332- 334.

Statistical Design of a Sampling Method in Quality Control of Research Data

Mohammad Javad Ershadi[1]
Associate Professor of Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran

Abstract

In the scientific literature, indexing and quality control are key processes that, if done correctly, can be properly retrieved by search engines by researchers. On the other hand, the use of mechanisms such as infallibility and empowerment of users has made research organizations 100% free from quality control. Also, the restriction on the use of specialized organizational human resources has doubled the importance of paying attention to sampling methods. Although in scientific sources, sampling methods in physical and tangible products have been well and adequately addressed, but in the field of data, especially research data, little work has been done. In this research, a framework for sampling in data quality control processes is provided. Also, an algorithm has been developed for statistical design to minimize type I and II errors. As a case study of research data, the information dissemination database of dissertations / dissertations (pious) of graduates of the whole country (Ganj) has been selected and the research method has been implemented in this database. The results of this study showed that, considering the acceptable quality of many pious information items after registration, sampling is a vital task in improving the efficiency of the information organization and analysis unit. The classification of information items into three categories is critical, main and partial, and determining the number and method of sampling for each category is another result of this research. The framework presented in this research can be localized for various data-driven organizations, especially businesses based on research data. Since any revision of AQL and LTPD values affects type I and II errors, it is necessary to apply the algorithms developed in this research to new AQL and LTPD values as well. Obviously, the results of the algorithm implementation such as number of samples, acceptance number and rejection number will be updated in this process.

Keywords: Data quality, Sampling, Quality Control, OC curve, Organize, Analyze information.

1. mjershadi@gmail.com