انتشار داده‌های آماری و کنترل افشای اطلاعات فردی

حمیدرضا نواب‌پور[۱]   محمد بردبار عشرت‌آبادی[۲]

چکیده

با افزایش تقاضا برای آمارها در عرصه‌های جدید، مانند آمارهای زیست‌محیطی، آمارهای فقر، آمارهای جنسیتی، آمارهای فرهنگی و ... ، مراکز آمار در کشورهای مختلف اقدام به برنامه‌ریزی برای تولید و انتشار این‌گونه آمارها علاوه بر آمارهایی که به‌صورت ادواری تولید و منتشر می‌شوند، کرده‌اند. برنامه‌ریزان، اقتصاددانان، و پژوهشگران همواره به داشتن داده‌های خرد تمایل داشته‌اند، زیرا این داده‌ها مانور آنها را در تحلیل پدیده‌های اجتماعی، اقتصادی و فرهنگی بیشتر می‌کند. اما طبق قانون، انتشار داده‌های جمع‌آوری شده نباید منجر به افشای هویت واحدهای اطلاع‌گیری شده، شود. این امر همواره بهانه‌ای برای عدم انتشار داده‌های حاصل از آمارگیری‌ها به نحوی که مورد تقاضای کاربران آن‌ها باشد، بوده است و یا در برخی موارد داده‌ها در جدول‌های انتشاراتی ایمن نبوده و امکان افشای هویت واحدهای اطلاع‌گیری شده وجود داشته است. هدف این مقاله ارایه روش‌هایی است که به توسط آنها می‌توان هم تقاضای کاربران آمارهای تولید شده را اجابت کرد و هم از افشای اطلاعات فردی که در بسیاری از کشورها خلاف قوانین مراکز آمار است، جلوگیری کرد.

واژه‌های کلیدی: آمارهای رسمی، محرمانگی، داده‌های‌خرد، متغیر سلسله مراتبی، فراداده.

۱. مقدمه

امروزه مدیریت جامعه‌ها مبتنی بر اطلاعات است، لذا برنامه‌ریزان، اقتصاددانان، و پژوهشگران برای تصمیم‌سازی نیازمند اطلاعات آماری دقیق، روزآمد، و به‌موقع هستند. از این رو کشورها با تصویب قانون عمومی آمار، اقدام به تأسیس مراکز آمار با وظیفه اصلی تولید و انتشار آمارهای رسمی نموده‌اند. معمولاً قانون‌های عمومی آمار آحاد ملت را ملزم به ارایه اطلاعات صحیح به پرسشگران می‌کنند و همچنین به آن‌ها اطمینان می‌دهند که اطلاعات آن‌ها حفاظت شده و جز در تهیه آمارهای کلی به‌کار برده نمی‌شود. اصل محرمانگی اطلاعات فردی در همه‌ی نظام‌های آماری پذیرفته شده است، لذا مراکز آمار برای رعایت این اصل، محدودیت‌هایی را برای انتشارداده‌های حاصل از آمارگیری‌ها اعمال می‌کنند به‌طوری که بعضاً نیازهای پژوهشگران به برخی اطلاعات آماری تامین نمی‌شود. گاهی نیز امکان افشای هویت‌ها از جدول‌های انتشاراتی وجود دارد. در طول سه دهه گذشته ایجاد تعادل بین تأمین نیازهای آماری کاربران و جلوگیری از افشای هویت پاسخگویان یکی از دغدغه‌های مهم مراکز آمار کشورها بوده است.

عموماً دو نوع افشا درنظر گرفته می‌شود: افشای هویت و افشای صفت. در نوع اول که در واقع مهم‌ترین نوع افشا است، ابتدا فرد مورد نظر شناسایی شده و سپس بر اساس هویت فرد، اطلاعات مربوط به او از داده‌ها استخراج می‌گردد. اما در برخی موارد تنها دانستن این‌که یک پاسخگو عضوی از یک گروه است، بدون این‌که معلوم شود که کدام یک از آن‌ها است، برای افشای اطلاعات در مورد وی کفایت می‌کند. این نوع افشا را افشای صفت[۳] می‌نامند. عمل افشا به هر دلیل که صورت پذیرد موجب بی‌اعتمادی عمومی نسبت به مرکز آمار به عنوان حافظ اطلاعات شخصی پاسخگویان شده و کاهش یا عدم همکاری پاسخگویان را در طرح‌های آمارگیری به‌همراه خواهد داشت.

۲. مرور نوشتگان

موضوع کنترل افشای آماری[۴] اولین بار توسط دالینوس در سال ۱۹۷۷ مطرح شد و سپس سایر آمارشناسان در کشورهای آمریکایی و اروپایی بحث و تحقیق در این موضوع را شروع کردند.

کاکس [۲] روش‌های پنهان‌سازی مقدماتی و مکمل را بحث کرده است. روش‌های گردکردن تصادفی توسط کاکس[۳]، روش پاسخ

۱- استادیار دانشگاه علامه طباطبائی Email:h.navvabpor@sci.org.ir
۲- کارشناس مرکز آمار ایران Email: bord_b@yahoo.com

۳- Attribute Disclosure
۴- Statistical Disclosure Control

پس‌تصادفی‌شده[1] (PRAM) توسط گوویلیو و همکاران [۵] بررسی شده‌اند. ویلن‌بورگ و دوال [۷] روی روش‌های بازکدگذاری عام[2] و پنهان‌سازی موضعی[3] کار کرده‌اند. پیشرفت‌های حاصل در کنترل افشای داده‌های آماری را می‌توان در دو گزارش کمیته فدرال روش‌شناختی آماری، در سال‌های ۱۹۷۸ و ۱۹۹۴ و نیز در مقاله‌های شماره‌های ویژه سال‌های ۱۹۹۳ و ۱۹۹۸ مجله آمار رسمی[4] مشاهده نمود.

در ایران تا به‌حال هیچ پژوهشی در موضوع کنترل افشای آماری به‌جز پایان‌نامه کارشناسی ارشد [۱] صورت نگرفته است و نیز هیچ تجربه‌ای در ایمن‌سازی اطلاعات آماری منتشرشده وجود ندارد.

۳. روش‌های محدودسازی افشا در داده‌های‌خرد

هرگاه یک سازمان آماری بخواهد یک مجموعه داده‌های‌خرد را منتشر کند، متغیرهای شناسایی مستقیم (مانند نام، شماره تلفن، ...) را از داده‌ها حذف می‌کند اما مثال بعد نشان می‌دهد که این کار کافی نیست. فرض کنید یک مجموعه داده‌های‌خرد، شامل اطلاعاتی در مورد محل سکونت، شغل و سابقه جنایی پاسخگویان منتشر شود. علاوه بر این فرض کنید یک رکورد با ترکیب مقادیر زیر در مجموعه داده‌های‌خرد وجود داشته باشد: «محل سکونت: شهر A، شغل: شهردار، سابقه جنایی: یک مورد». اگر چه فرض بر این است که نام یا آدرس پاسخگو منتشر نشود، اما افراد زیادی پی می‌برند که پاسخگو چه کسی است. به‌ویژه می‌توانند نتیجه بگیرند که این پاسخگو یعنی شهردار شهر A دارای یک سابقه جنایی است. چنین ترکیب‌هایی از مقادیر متغیرها که در جامعه به تعداد اندک اتفاق می‌افتند، ترکیب‌های ناایمن نامیده می‌شوند. به‌طور کلی هر پاسخگویی که در یک ترکیب ناایمن قرار بگیرد، ممکن است با توجه به اطلاعات عمومی یا داده‌های بیرونی موجود در معرض خطر شناسایی باشد.

پس از مشخص شدن رکوردها و ترکیب‌های ناایمن برای انتشار، توسط روش‌های محدودسازی خطر افشای مربوط به داده‌های‌خرد، یک فایل داده‌های‌خرد ناایمن به یک فایل داده‌های‌خرد ایمن قابل انتشار با خطر افشایی که در حد قابل قبولی پائین است، تبدیل می‌شود. این روش‌ها به دو گروه عمده تقسیم می‌شوند. گروه اول شامل روش‌های ناپرشیده[5] است که به سبب اعمال آن‌ها روی فایل داده‌های‌خرد، خطائی به داده‌ها افزوده نمی‌شود و فقط اندکی از میزان اطلاعات موجود در داده‌ها کاسته می‌شود. گروه دوم روش‌های پرشیده[6] نام دارند که به واسطه اعمال این روش‌ها، مقادیر اصلی با مقادیر دیگری به گونه‌ای تعویض می‌شوند که امکان انجام تحلیل‌های آماری رایج روی فایل پرشیده شده امکان‌پذیر است.

۳. ۱. روش بازکدگذاری عام و روش پنهان‌سازی موضعی

روش‌های بازکدگذاری عام و پنهان‌سازی موضعی در گروه روش‌های ناپرشیده قرار می‌گیرند و اغلب به‌صورت توأم و برای متغیرهای رسته‌ای به‌کار می‌روند. در روش باز کدگذاری عام چندین رسته از یک متغیر مانند A با یکدیگر ترکیب شده و رسته جدیدی را پدید می‌آورند. مقادیر A متناظر با رسته‌های جدید، مجدداً کدگذاری می‌شوند. این کار روی تمام فایل داده‌های خرد اعمال می‌شود. در روش پنهان‌سازی موضعی، مقدار یک متغیر مانند A در یک ترکیب ناایمن با یک مقدار گمشده تعویض می‌گردد. در حالی که بازکدگذاری عام روی کل فایل داده‌های‌خرد تأثیر می‌گذارد، پنهان‌سازی موضعی تنها برای یک مقدار به‌خصوص در یک رکورد ناایمن به‌کار می‌رود.

هر دو روش موجب از دست رفتن مقداری از اطلاعات موجود در داده‌های‌خرد می‌شوند. علاوه براین پنهان‌سازی موضعی ممکن است موجب برآوردگرهای اریب گردد. به همین دلیل سعی می‌شود تا از آن در مقیاسی کوچک استفاده شود. در عمل یک توازن بین استفاده از این دو روش به‌وجود می‌آید. اغلب در ابتدا بعضی از متغیرها بازکدگذاری عام می‌شوند و در مورد باقیمانده رکوردهای ناایمن از پنهان‌سازی موضعی برای حداقل مقادیر ممکن، استفاده می‌شود. مینیمم‌سازی تعداد پنهان‌سازی‌ها یک موضوع اساسی است و می‌توان با آن به عنوان یک مساله بهینه‌یابی مشابه مسائل مطرح در بحث برنامه‌ریزی خطی برخورد کرد.

۳. ۲. روش پس‌تصادفی‌شده (PRAM)

پرام(PRAM) روشی پرشیده است که برای متغیرهای رسته‌ای به‌کار می‌رود. در این روش برای هر رکورد، امتیاز تعدادی از متغیرها متناظر با یک مکانیسم احتمالاتی معین، تغییر می‌کند، در نتیجه شناسایی رکوردهای متناظر با اشخاصی معین، دشوار می‌شود. از طرفی تا وقتی‌که مکانیسم احتمالاتی استفاده شده معلوم باشد، مشخصه‌های داده‌های اصلی می‌تواند از فایل داده‌های پرشیده شده

۱- Post Randomized Response Method (PRAM)

۲- Global Recoding

۳- Local Suppression

۴- Journal of Official Statistics

۵- Non-Perturbative Methods

۶- Perturbative Methods

برآورد شوند. یعنی به‌کار بردن تمام تحلیل‌های آماری امکان‌پذیر است.

فرض کنید ξ متغیری رسته‌ای در فایل اصلی و X همان متغیر در فایل پرشیده شده و دارای K رسته، $k=۱,۲,\cdots,K$ باشند. اگر $P_{kl}=P(X=l\,|\,\xi=k)$ $(l,k=۱,۲,\cdots,K)$ احتمال این باشد که امتیاز $\xi=k$ به $X=l$ تغییر یابد، آن‌گاه $P=\{p_{kl}\}$ ماتریسی $K\times K$ با اعضای p_{kl} است. P یک ماتریس مارکوف است (P۱=۱ که در آن ۱ برداری $K\times$۱ از ۱هاست). فرض می‌شود که P معکوس‌پذیر است. مثال زیر تأثیر پرام بر محدودسازی افشا را تشریح می‌کند.

مثال: فرض کنید فایل داده‌های‌خرد شامل یک نمونه‌ی تصادفی ساده به اندازه n از جامعه‌ای با اندازه N باشد. فایل دقیقا شامل یک زن جراح است. پرام، روی متغیر جنسیت و برای هر رکورد مستقل از رکوردهای دیگر به‌کار رفته است. امتیاز جنسیت با احتمال‌های $P_{۲۲}=۰/۹$ ، $P_{۲۱}=۰/۱$ ، $P_{۱۲}=۰/۱$ ، و $P_{۱۱}=۰/۹$ تغییر می‌کند و ξ متغیر جنسیت دارای دو رده‌ی ۱= مرد و ۲= زن (در فایل اصلی) و X متغیر جنسیت دارای دو رده‌ی ۱= مرد و ۲= زن (در فایل پرشیده‌شده) می‌باشند. حال فرض کنید که شخص متخلف بداند که جامعه دارای ۱ زن جراح و ۹۹ مرد جراح است. احتمال این‌که وی پی ببرد، زن جراح در فایل پرشیده شده، در واقع زن جراح در جامعه می‌باشد برابر است با:

$$P(\xi=۲|X=۲)=\frac{P_{۲۲}\,P(\xi=۲)}{P_{۱۲}P(\xi=۱)+P_{۲۲}P(\xi=۲)}=\frac{۰/۹\times۰/۰۱}{۰/۱\times۰/۹۹+۰/۹\times۰/۰۱}\approx۰/۰۸$$

مشاهده می‌شود که مقدار این احتمال (۰/۰۸)، بسیار اندک است. بنابراین داده‌های پرشیده شده به اندازه‌ی کافی ایمن هستند. حال فرض کنید متغیر جنسیت با احتمال ۰/۹۹۹۹ بدون تغییر بماند و با احتمال ۰/۰۰۰۱ تغییر کند، آن‌گاه احتمال شناسایی زن جراح برابر ۰/۹۹ است که احتمال بالایی است و داده‌های پرشیده شده ایمن به‌نظر نمی‌رسند.

نسبت مورد انتظار به‌عنوان معیاری برای میزان عدم اطمینان ایجادشده توسط پرام در فایل داده‌های خرد به‌کار می‌رود. فرض کنید $\xi^{(r)}$ $(X^{(r)})$ امتیاز ξ (X) برای rامین رکورد در فایل داده‌های‌خرد باشد. نسبت مورد انتظار امتیاز k $(k=۱,...,K)$ یعنی $ER(k)$ به‌صورت

$$ER(k)=\frac{P_{kk}T_\xi(k)}{\sum P_{lk}T_\xi(l)}\qquad k=۱,...,K$$

معرفی می‌شود که در آن $T_\xi(k)$ تعداد رکوردهای فایل اصلی برای $\xi^{(k)}=k$ است. مقدار کوچک ER(k) نشان دهنده‌ی این است که بیشتر احتمال می‌رود که یک رکورد برای $X^{(r)}=k$ در اصل متعلق به این امتیاز نبوده و بنابراین ایمن‌کننده‌ی فایل پرشیده‌شده است. می‌توان نتایجی که از تحلیل‌های معینی روی فایل پرشیده‌شده حاصل می‌شود را به نتایجی که می‌توانست از فایل اصلی حاصل گردد، تبدیل نمود. برای مثال فرض کنید که بخواهیم یک تحلیل رگرسیونی روی متغیر عددی وابسته y و متغیر رسته‌ای مستقل ξ انجام دهیم و پرام برای ξ با K رسته استفاده شده است. برای هر رکورد، متغیرهای ظاهری $\delta_۱,\delta_۲,\cdots,\delta_K$ به‌صورت مقابل تعریف می‌شوند:

$$\delta_k=\begin{cases}۱\longrightarrow\xi^{(r)}=k\\ ۰\longrightarrow\xi^{(r)}\neq k\end{cases}\qquad,\qquad k=۱,...,K$$

فرض کنید:

$$T_\xi^y(k)=\sum y^{(r)}I_{\{\xi(r)=k\}}\quad \text{و}$$

$$T_\xi^y=(T_\xi^y(۱),...,T_\xi^y(K))،$$

که در آن $y^{(r)}$ برابر مقدار y برای rامین رکورد و I بر تابع نشانگر[۱] دلالت دارد و $T_\xi^y(k)$ مجموع مقادیر y برای رکوردهایی است که دارای امتیاز k می‌باشند. کویمان [۶] نشان داد که T_ξ^y به‌طور ناریب توسط $\hat{T}_\xi^y=(p^{-1})^t T_X^y$ برآورد می‌شود. اکنون رگرسیون y روی ξ با رگرسیون y روی $\delta_۱,...,\delta_K$ یکسان است. ضریب‌های رگرسیونی توسط $(D^tD)^{-۱}D^ty$ داده می‌شوند که در آن D ماتریسی $n\times K$ است (D^t ترانهاده D است) که عضو (r,j)ام آن برابر مقدار δ_j برای r امین رکورد در فایل داده‌های اصلی است. (D^tD) توسط $\hat{T}_\xi^y$ (توجه شود که (D^tD) ماتریسی قطری است که عضو (k,k)ام آن $T_\xi(k)$ است) و D^ty توسط $\hat{T}_\xi^y$ به‌صورت ناریب، برآورد می‌شوند. در نتیجه برآوردگر رگرسیونی نتیجه شده سازگار است. این نتیجه برای تمام فنون تحلیل آماری بر پایه‌ی گشتاورهای مرتبه‌ی دوم داده‌ها از قبیل تحلیل تشخیص[۲] و تحلیـل واریانس[۳] صادق است. در عمل ابتدا باید تصمیم گرفت که برای کدام متغیرها باید از پرام استفاده کرد و

۱- Indicator Function

۲- Discriminant Analysis

۳- Analysis of Variance

این‌که کدام رسته از این متغیرها می‌تواند به کدام رسته و با چه احتمالی تغییر کند.

۳.۳. سایر روش‌ها

روش ریزانبوهش[۱] : روش ریزانبوهش از روش‌های پرشیده برای متغیرهای کمی است و می‌تواند بر روی یک یا چند متغیر اعمال شود. ایده‌ی اصلی این است که قاعده‌های محرمانگی این اجازه را می‌دهد که اگر در یک مجموعه داده‌ها، پاسخگویان در گروه‌های k عضوی یا بیشتر قرار گیرند و هر مقدار در یک گروه با مقدار متوسط گروه تعویض شود و هیچ پاسخگویی نتواند از روی مقدارش شناسایی شود، آن‌گاه آن مجموعه داده‌ها امکان انتشار دارد.

بازکدگذاری به بالا یا پایین : در این روش که حالت خاصی از بازکدگذاری عام است مقادیر بالای یک متغیر(بیشتر از یک مقدار آغازین معین) با هم تشکیل یک رسته جدید می‌دهند. چنین عملی برای مقادیر پایین(کمتر از یک مقدار آغازین معین) نیز به‌کار می‌رود.

افزودن نوفه : برای متغیرهای پیوسته به‌کار می‌رود و توسط آن مقادیری تصادفی از یک توزیع احتمال پیوسته مانند توزیع نرمال به مقادیر اصلی افزوده می‌شود. برای جلوگیری از اریبی برآوردهای خطی، میانگین توزیع، صفر منظور می‌شود و واریانس بزرگ‌تر مبین پرشیدگی بیشتر است.

٤. روش‌های محدودسازی افشاء در جدول‌های انتشاراتی

٤.١. داده‌های جدولی

جدول‌ها رایج‌ترین محصولات سازمان‌های آماری می‌باشند که به دو صورت جدول‌های مقداری و جدول‌های فراوانی منتشر می‌شوند. به‌عنوان مثال میزان سرمایه‌گذاری کارخانجات براساس ناحیه و نوع فعالیت، یک جدول مقداری و تعداد کارخانجات برحسب ناحیه و نوع فعالیت یک جدول فراوانی است.

٤.٢. تشخیص خانه‌های حساس

وقتی داده‌ها جمع‌بندی شده‌اند به‌نظر می‌رسد که خطر افشایی وجود ندارد، اما همیشه این‌گونه نیست. در جدول‌های مقداری هنگامی‌که میزان زیادی از مقدار یک خانه جدول توسط تعداد معدودی از پاسخگویان ارایه شده باشد، امکان برآورد تقریبی مقدار هر پاسخگو وجود دارد. برای تشخیص این خانه‌ها معمولا از قواعدی استفاده می‌شود.

در مورد جدول‌های مقداری، عمومی‌ترین ملاکی که به‌کار می‌رود، قاعده‌ی تسلط (n,k)[۲]است که به‌صورت زیر تعریف می‌شود : یک خانه‌ی جدول حساس است اگر مجموع مقادیر عرضه شده توسط n عدد از بزرگ‌ترین پاسخگویان به مقدار کل خانه، بیشتر از k درصد مقدار کل خانه را شامل شود. قاعده‌ی دیگری که در مورد جدول‌های مقداری به‌کار می‌رود قاعده‌ی پیشین-پسین(p,q)[۳] است. اگر هر پاسخگو قبل از انتشار جدول بتواند مقدار عرضه شده توسط پاسخگوی دیگر به یک خانه‌ی جدول را با اختلاف کمتر یا برابر q درصد از مقدار واقعی آن، برآورد نماید و پس از انتشار جدول، این امکان برای برخی پاسخگویان به‌وجود آید که مقدار پاسخگویی دیگر به خانه‌ی جدول را با اختلاف کمتر یا برابر p درصد ($p<q$) مقدار اصلی آن، برآورد نمایند، این خانه از جدول، حساس شناخته می‌شود.

در مورد جدول‌های فراوانی قاعده‌ای که غالباً برای تشخیص خانه‌های حساس به‌کار می‌رود بدین صورت است که نباید تعداد پاسخگویان در هر خانه‌ی جدول از عدد مشخصی (مقدار آغازین) مثلا سه کمتر باشد. به هر حال این قاعده در برخی موقعیت‌ها کافی نیست.

٤.٣. روش‌های بازطراحی جدول و پنهان‌سازی خانه‌ای

ممکن است در یک جدول، در یک سطر یا ستون، خانه‌های زیادی، حساس تشخیص داده شوند. در این حالت در طبقه‌بندی متغیرهای تبیینی[۴] تجدیدنظر می‌شود. در این صورت جزئیات اطلاعات آماری ارایه شده در جدول کاهش می‌یابد اما از تعداد خانه‌های حساس به مقدار زیادی کاسته می‌شود، سپس می‌توان از روش‌های کنترل افشای دیگری مانند پنهان‌سازی خانه‌ای استفاده نمود. جدول زیر را در نظر بگیرید. فرض کنید که خانه‌ی متناظر با فعالیت دو و ناحیه‌ی C طبق قاعده‌ی تسلط به‌کار رفته، حساس تشخیص داده شده و نمی‌تواند منتشر شود. مقدار این خانه پنهان می‌شود.

۱- Microaggregation

۲- (n,k) Dominance Rule

۳- (p,q) Prior- Postrior rule

۴- Explanatory Variables

جدول (٤ـ١). میزان سرمایه‌گذاری کارخانجات به‌تفکیک ناحیه و نوع فعالیت (پس از پنهان‌سازی مقدماتی)

فعالیت \ ناحیه	A	B	C	کـل
۱	۲۰	۵۰	۱۰	۸۰
۲	۸	۱۹	X	۴۹
۳	۱۷	۳۲	۱۲	۶۱
کل	۴۵	۱۰۱	۴۴	۱۹۰

به‌طور کلی، پنهان کردن خانه حساس، به‌تنهایی کافی نیست. زیرا خانه‌ی پنهان شده می‌تواند با استفاده از جمع‌های حاشیه‌ای به‌آسانی محاسبه شود. پس لازم است خانه‌های دیگری که حساس نیستند، پنهان شوند. این عمل، پنهان‌سازی خانه‌ای مکمل و خانه‌ها، پنهان‌شده‌های مکمل نامیده می‌شوند. یک الگوی پنهان‌سازی خانه‌ای مکمل در جدول (۴ـ۱)، پنهان کردن خانه‌های با فعالیت دو و سه در ناحیه‌ی A و فعالیت سه در ناحیه‌ی C است. به هر حال در جدول‌های بزرگ‌تر هنگامی که هدف مینیمم‌سازی تعداد پنهان سازی‌ها یا مینیمم‌سازی مقادیر پنهان‌شده به طوری‌که خانه‌های پنهان شده به‌صورت تقریبا دقیق قابل برآورد نباشند، است یافتن پنهان‌سازی‌های مکمل بسیار مشکل خواهد بود. در این حالت نیز از فنون بهینه‌یابی مطرح در برنامه‌ریزی خطی استفاده می‌شود و استفاده از یک نرم‌افزار ویژه ضروری است.

٤.٤. معاوضه داده‌ها[1]

این شیوه بر روی داده‌های خرد اعمال می‌شود و سپس این داده‌ها برای ساختن جدول‌ها به‌کار می‌روند. در مرحله اول یک نمونه از رکوردها انتخاب می‌شوند. سپس برای این رکوردها در برخی نواحی جغرافیایی دیگر روی یک مجموعه مشخص از صفات مهم، همانندیابی می‌شود. در مرحله بعد تمام صفات رکوردهای همانندشده با یکدیگر تعویض می‌گردند. از داده‌های خرد نتیجه‌شده می‌توان برای ساختن جدول‌های مورد نیاز استفاده نمود.

٥.٤. روش گردکردن

در این روش هر مقدار از خانه‌ها به یکی از دو نزدیک‌ترین مضرب مقدار پایه به‌طوری که جدول جمع‌پذیر باشد، گرد می‌شوند. خطای گردکردن دارای امید ریاضی صفر است. به این صورت که یک عدد مانند a که می‌تواند به‌صورت $a = kb + r \quad \circ \le r < b$ نوشته شود که در آن k و b عددی صحیح و پایه‌ی صحیح گرد کردن

۱- Data swapping

است، به a^* که مقادیر kb و $(k+۱)b$ را با احتمال‌های $p(a^* = kb) = ۱ - \frac{r}{b}$ و $p(a^* = (k+۱)b) = \frac{r}{b}$ اختیار می‌کند گرد می‌شود. در نتیجه:

$$E(a^*) = (k+۱)b \times \frac{r}{b} + kb \times (۱ - \frac{r}{b})$$
$$= kr + r + kb - kr = kb + r = a$$

در حال حاضر در کشور ما با وجود ارایه‌ی اطلاعات آماری به شکل بسیار کلی باز هم می‌توان مواردی از افشای دقیق یا تقریبی اطلاعات شخصی را به‌دست آورد. اما در صورت اعمال روش‌های ذکرشده در این مقاله بر روی داده‌ها می‌توان در عین ارایه مفصل‌تر اطلاعات آماری، جنبه‌ی محرمانگی اطلاعات شخصی را نیز رعایت نمود.

٥. کاربرد

تاکنون اطلاعات طرح آمارگیری از کارگاه‌های صنعتی ده نفر کارکن و بیشتر[2] که هر ساله به‌صورت سرشماری انجام می‌پذیرد برای حفظ محرمانگی اطلاعات فقط در سطح استان منتشر شده‌اند، در حالی که می‌توان هم زمان با انتشار اطلاعات در سطوح جغرافیایی کوچک تر، جنبه‌ی محرمانگی آن‌ها را نیز رعایت کرد.

٥. ١. چندنمونه از اطلاعات قابل افشا از جدول‌های انتشاراتی

در مطالعه مثال‌های زیر توجه به دو نکته ضروری است: برای حفظ محرمانگی از آوردن نام واحدی که هویت آن قابل افشاء است، خودداری شده است، و دیگر این‌که برای اختصار تنها به شماره جدول‌ها از نشریه «نتایج آمارگیری از کارگاه‌های صنعتی ده نفر کارکن و بیشتر ۱۳۷۹» ارجاع داده شده است.

مثال ۱. قسمتی از جدول شماره (۵-۱) (تعداد کارگاه‌های صنعتی ده نفر کارکن و بیشتر برحسب تعداد شاغلان و نوع فعالیت: ۱۳۷۹) از نشریه نتایج آمارگیری از کارگاه‌های صنعتی ده نفر کارکن و بیشتر سال ۱۳۷۹ به‌صورت زیر است:

در این جدول کدهای متغیر تعداد شاغلان به‌صورت زیر تعریف شده است: ۱ :۱۹-۱۰ ، ۲ :۲۹-۲۰ ، ۳ :۳۹- ۳۰ ، ۴ :۴۹ ـ ۴۰ ، ۵ :۹۹ـ ۵۰ ، ۶ :۴۹۹ ـ ۱۰۰ ، ۷ :۹۹۹ـ ۵۰۰ ، ۸ : ۱۰۰۰ و بیشتر.

۲- مشخصات این طرح را می‌توان در نشریه «نتایج آمارگیری از کارگاههای صنعتی ۱۰ نفر کارکن و بیشتر ۱۳۷۹» از انتشارات مرکز آمار ایران یافت.

همان‌طور که از این جدول مشخص است تنها یک کارگاه تولید محصولات از توتون و تنباکو و سیگار در کل کشور وجود دارد که تعداد کارکنانش بیشتر از ۱۰۰۰ نفر است. به علت یکتا بودن این کارگاه در جامعه، به‌راحتی تمام اطلاعات مربوط به این کارگاه خاص قابل دسترسی است. اگر رضایت این کارگاه برای انتشار اطلاعات خاص آن کسب نشده باشد، یک نمونه از افشای کامل اطلاعات فردی اتفاق افتاده است.

مثال ۲. کارگاه‌های تولید وسایل نقلیه موتوری را در نظر بگیرید. در ابتدا خطر افشایی احساس نمی‌شود، در حالی که این چنین نیست. به‌عنوان مثال در جدول شماره (۴)، تعداد شاغلان این ۱۴ کارگاه تولید وسایل نقلیه موتوری ۳۲۵۱۰ نفر ذکر شده است. تعداد شاغلان ۳ عدد از بزرگ‌ترین کارگاه‌ها به ترتیب برابر ۱۷۴۱۷ نفر، ۵۱۱۳ نفر و ۲۷۶۱ نفر است که جمعا حدود ۷۷ درصد کل شاغلان این بخش را به خود اختصاص داده‌اند. فرض کنید $x_1,...,x_{14}$ مقادیر متناظر ۱۴ کارگاه باشد. طبق قاعده پیشین- پسین (p,q) فرض کنید دومین پاسخگوی بزرگ قبل از انتشار جدول بتواند هر یک از مقادیر دیگر پاسخگویان را با اختلاف کمتر یا برابر ۵۰ درصد برآورد کند $(q=50)$. در صورت اتحاد دومین و سومین کارگاه بزرگ، آن‌ها پس از انتشار جدول می‌توانند پی ببرند که x_1 حداکثر برابر

$$x_1 + \left(\frac{q}{100}\right) \sum_{i=4}^{14} x_i = 21027/65 \approx 21027$$

و حداقل برابر

$$x_1 + \left(\frac{q}{100}\right) \sum_{i=4}^{14} x_i = 13807/75 \approx 1380$$

است. یعنی تعداد شاغلان بزرگ‌ترین کارگاه با اختلافی حدود ۲۰ درصد قابل برآورد است. به‌همین صورت در جدول شماره (۲۴) این نشریه، ارزش کل مواد خام و اولیه، لوازم بسته‌بندی، ابزار و وسایل کار کم‌دوام مصرف‌شده توسط این کارگاه‌ها ۱۳۲۰۴۱۷۶ میلیون ریال ذکر شده است. با توجه به این‌که مقدار این متغیر برای ۳ عدد از بزرگ‌ترین کارگاه‌ها به ترتیب برابر۷۰۸۰۰۰۰ میلیون ریال، ۳۳۰۰۰۰۰ میلیون ریال و۶۲۶۰۰۰ میلیون ریال است، این سه جمعا حدود ۸۳ درصد مقدار کل این خانه را ارایه کرده‌اند. حال فرض کنید دومین کارگاه بزرگ قبل از انتشار جدول، میزان این متغیر برای هر کارگاه را بتواند با اختلاف کمتر یا برابر ۵۰ درصد مقدار واقعی آن‌ها برآورد کند، در این‌صورت پس از انتشار جدول طبق مطالب مطرح شده در بالا، این کارگاه با همکاری سومین کارگاه بزرگ مقدار این متغیر برای بزرگ‌ترین کارگاه را می‌تواند در بازه {۸۱۷۹۰۸۸ و ۵۹۸۰۹۱۲} یعنی با تفاوتی حدود ۱۵ درصد مقدار واقعی برآورد نماید.

۵ .۲. ایمن‌سازی اطلاعات آماری انتشاراتی

به‌منظور ایمن‌سازی داده‌های‌خرد از نرم‌افزار τ_ARGUS و برای ایمن‌سازی داده‌های جدولی از نرم‌افزار μ_ARGUS که حاصل پروژه $CASC$[1] است استفاده می‌شود. این پروژه با هدف تبیین زمینه‌های علمی و تهیه ابزارهای عملی برای کنترل افشای آماری توسط تعدادی از کشورهای اتحادیه اروپا از سال ۲۰۰۱ به مدت سه سال، سازماندهی و اجرا شده است. با توجه به‌این‌که داده‌های سرشماری از کارگاه‌های صنعتی دارای ده نفر کارکن و بیشتر به‌صورت جدول منتشر می‌شوند، لذا در این بخش به نحوه‌ی ایمن‌سازی جدول‌های انتشاراتی طرح مذکور در استان تهران در سال ۱۳۷۹ با استفاده از نرم‌افزار $ARGUS\_\tau$ پرداخته می‌شود.

به‌منظور ساختن جدول‌های انتشاراتی ایمن، به فایل داده‌های‌خرد و فایل فراداده- حاوی مشخصات متغیرهای موجود در فایل داده‌های خرد- نیاز است. در صورتی که فایل فراداده وجود نداشته باشد، توسط نرم‌افزار امکان ایجاد آن وجود دارد. پس از آن‌که فایل‌ها توسط نرم‌افزار خوانده شد، می‌توان جدول‌های مورد نیاز را توسط پنجره‌ای که نرم‌افزار ارایه می‌دهد مشخص کرد. در این مرحله متغیرهای سازنده‌ی جدول، متغیر مورد استفاده در قاعده‌ی تسلط، متغیر مورد استفاده برای می‌نیمم‌سازی فقدان اطلاعات ناشی از اعمال روش‌های کنترل خطر افشا و قاعده‌ی تسلط برای تشخیص خانه‌های حساس تعیین می‌شوند. وقتی جدول‌ها، جدول‌های فراوانی هستند، حفاظت‌کننده‌ی جدول ممکن است به‌سادگی خانه‌هایی که دارای فراوانی‌ای کمتر از یک مقدار آغازین معین هستند را به‌عنوان خانه‌های ناایمن درنظر بگیرد. اما در یک جدول مقداری باید برای تشخیص خانه‌های حساس از قاعده‌های تسلط استفاده کرد. در اینجا جدول مقداری میزان موجودی انبار در پایان اسفند به تفکیک نوع فعالیت و تعداد شاغلان انتخاب می‌شود. قاعده‌ی تسلط مورد استفاده، قاعده‌ی p ٪ با $p = 10$ است. نرم‌افزار جدول درخواستی را به‌صورت زیر نمایش می‌دهد.

در این حالت تعداد ۱۱۵۵ خانه‌ی ایمن، ۴۶۳۷ خانه‌ی ناایمن اولیه و ۶۹۱۵۰ خانه‌ی خالی وجود دارد. اکنون می‌توان این جدول را به لحاظ محرمانگی حفاظت کرد. یکی از روش‌های قابل به‌کارگیری توسط این نرم‌افزار، روش بازکدگذاری با انتخاب گزینه‌ی $Recode$ است. در اینجا متغیر نوع فعالیت برای بازکدگذاری انتخاب شده است. این متغیر از نوع سلسله‌مراتبی است و با حذف ارقام، از میزان

۱- Computational Aspects of Statistical Confidentiality

جزئیات متغیر کاسته می‌شود. با این عمل تعداد خانه‌های ایمن به ۵۴۲، خانه‌های ناایمن اولیه به ۱۶۶۸ و تعداد خانه‌های خالی به ۸۱۷۸ خانه تغییر می‌یابد. می‌توان روی متغیر تعداد شاغلان نیز به صورت زیر بازکدگذاری را انجام داد: ۱ :۱۹-۱۰ ، ۲ :۲۹-۲۰ ، ۳ :۳۹، ۳۰ ، ۴ :۴۹، ۴۰ ، ۵ :۹۹، ۵۰ ، ۶ :۴۹۹، ۱۰۰ ، ۷ :۹۹۹، ۵۰۰ ، ۸ : ۱۰۰۰ و بیشتر. در این حالت تعداد خانه‌های ایمن به ۱۷۰، خانه‌های ناایمن اولیه به ۴۲ و تعداد خانه‌های خالی به ۶۸ خانه تغییر می‌یابد. با ادغام رسته‌های ۷ و ۸ تعداد خانه‌های ایمن به ۱۶۶، تعداد خانه‌های ناایمن به ۲۹ و تعداد خانه‌های خالی به ۵۷ خانه تغییر می‌یابد. پس از پایان بازکدگذاری متغیرها، باقی‌مانده‌ی خانه‌های ناایمن را می‌توان توسط روش پنهان‌سازی محافظت نمود. در این مرحله به منظور حفاظت خانه‌های ناایمن باید تعدادی خانه‌ی اضافه به عنوان پنهان‌سازی مکمل انتخاب شوند. این عمل با توجه به می‌نیمم‌سازی تعداد خانه‌های پنهان‌شده، مقادیر پنهان‌شده یا تعداد پاسخگویان ارائه‌دهنده‌ی مقدار خانه صورت می‌گیرد و همزمان بازه‌ی قابل برآورد مقادیر پنهان‌شده به اندازه‌ی کافی بزرگ باشد. جدول نهائی در صفحه بعد آمده است.

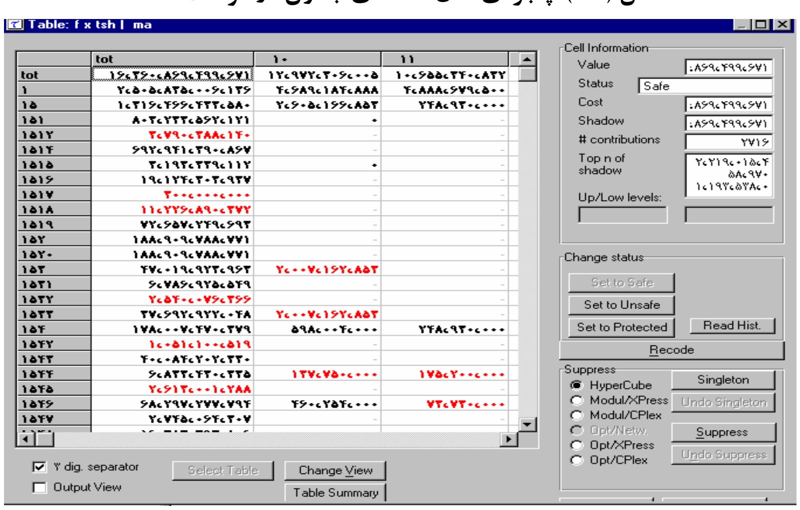
۶. نتیجه‌گیری

یکی از مهم‌ترین وظایف نظام‌های آماری کشورها تولید آمارهای مورد نیاز کاربران آن‌ها با مشخصات دقیق بودن، به‌موقع بودن، پوشش کامل به‌لحاظ موضوعی و جغرافیایی داشتن، ارزان بودن، و به‌نحو مناسب اطلاع‌رسانی شدن، است. به‌منظور حفظ محرمانگی اطلاعات فردی، نظام‌های آماری موظفند که به شیوه‌ای اطلاع‌رسانی آماری کنند که هویت افراد (واحدها) افشاء نشود. این امر که غالبا جنبه‌ی قانونی نیز دارد، بهانه‌ای شده است که برخی از سازمان‌های ملی آمار که وظیفه تولید و اطلاع‌رسانی آماری دارند از انتشار داده‌های‌خرد جلوگیری کنند و یا داده‌های حاصل از آمارگیری‌ها را به گونه‌ای منتشر کنند که از آن‌ها هویت افراد قابل افشا باشد. به‌منظور انتشار ایمن داده‌ها می‌توان از روش‌های کنترل افشای داده‌های آماری استفاده نمود. این روش‌ها هم از افشای هویت‌ها جلوگیری می‌کنند و هم به مراکز ملی آمار اجازه می‌دهند تا به‌صورت گسترده به‌اطلاع‌رسانی آماری مبادرت ورزند. برای این منظور لازم است از نرم‌افزارهای موجود همانند μ-ARGUS و τ-ARGUS استفاده شود.

جدول (۵ـ۱). قسمتی از جدول شماره ۱ نشریه نتایج آمارگیری از کارگاه‌های صنعتی ده نفر کارکن و بیشتر: ۱۳۷۹

فعالیت / تعداد شاغلان	جمع	۱	۲	۳	٤	٥	٦	۷	۸
تولید محصولات از توتون، تنباکو و سیگار	۱	۰	۰	۰	۰	۰	۰	۰	۱
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
تولید وسایل نقلیه موتوری	١٤	۰	۰	۰	۱	۰	٤	٤	٥

شکل (۵ـ۲). پنجره‌ی نشان دهنده‌ی جدول درخواست شده



جدول (۵-۳). میزان موجودی انبار در پایان اسفند به تفکیک نوع فعالیت و تعداد شاغلان

	جمع	1	2	3	4	5	6	7
جمع	16360869499671	188797020707	245766178935	233495337725	299580034342	1037508071901	3520116268838	10835606587223
1	2505835006136	50890802013	33607505069	33965092576	78269537028	183104504153	425673868045	1700323697252
15	1316466433580	25932086250	17797712906	21584004208	59549637353	92327488544	199737449948	899538054371
16	x	-	-	-	-	-	-	X
17	540827776649	15624291039	9315038533	5669876800	11491679957	x	159749055758	X
18	34979712895	4754670000	3271428330	x	2185193870	x	18793910538	-
19	x	4579754724	3223325300	x	5043025848	x	47393451801	X
2	6216406596988	101544289479	144771576985	166706626726	194216577302	619289723820	2022734390424	2967143412252
20	10157604202	2194938568	x	x	0	x	x	-
21	172490436193	387349654	8324202830	28384482810	12496623684	33894724941	89003052274	-
22	332029411823	4388621041	2015559064	x	x	x	x	235482400728
23	365658209205	x	x	x	-	x	-	358541046892
24	1517837224069	23259408438	38852172795	50946746092	37286150905	195071891220	751935581342	420485273277
25	455816432116	12897591112	12303608667	14650941506	35283609568	53851678180	107003306422	219825696661
26	462903015686	11414435418	15323612453	8454667047	16727461676	18608898656	133876027094	258497913342
27	695024661294	x	x	x	x	10541209346	127237037375	538954928194
28	688678787985	18495447092	18698562291	7944266409	16566853875	133924336029	195382105110	297667217179
29	1515810814415	25773140339	46140440895	45629338083	59335170419	152857805343	548385983357	637688935979
3	7638627896547	36361929215	67387096881	32823618423	27093920012	235113843928	1071708010369	6168139477719
30	x	x	9862947008	x	5693968260	45525685173	x	-
31	349887312645	11710886516	8254081501	x	9153424829	52218232306	201114349158	X
32	530483728175	7382769622	x	x	-	x	x	X
33	149768914095	x	x	4100330415	x	x	x	X
34	6415389679000	2139128955	5131712981	6346289545	7168066013	55494321800	565294126859	5773816032847
35	22370700834	x	x	-	x	x	18308668546	-
36	103415295210	6943040372	x	4583339455	x	37361819862	46657143116	-
37	x	x	-	x	-	-	-	-

x نشان‌دهنده‌ی خانه‌های پنهان‌شده‌ی اولیه یا مکمل است.

– نشان‌دهنده‌ی خانه‌هایی است که در آن‌ها پاسخگویی تحت شرایط تعریف‌شده وجود نداشته است.

مراجع

[۱] بردبار عشرت‌آبادی، محمد ،۱۳۸۲، فنون محدودسازی افشای داده‌های آماری و کاربرد آنها، پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

[2] Cox, L. H., 1980, Suppression methodology and statistical disclosure Control, *Journal of the American Statistical Association*, 75, 337-385.

[3] Cox L. H. 1987, A constructive procedure for unbiased controlled rounding, *Journal of the American Statistical Association*, 82, 520-524.

[4] Fienberg, S. E., Markov, U. G., and Steel, R. J. 1998, Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics*, 14, 485-502.

[5] Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. J., and De Wolf, P. P., 1998, Post randomisation for statistical disclosure control: Theory and implementation, *Journal of Official Statistics*, 14, 463-474.

[6] Kooiman, P., Willenborg, L. C. J. and Gouweleeuw, J. M., 1997, PRAM: A Method for Disclosure Limitation of Microdata, *Statistics Netherlands,* The Netherlands, Research Paper No. 97٠5.

[7] Willenborg, L. C. J. and De Waal, T., 1998, *Statistical Disclosure Control in Practice*, Springer-Verlag, New York.