بسمه تعالی

گزارش نهایی

عنوان طرح:

تعیین ریسک افشا و نحوه‌ی انتشار ایمن داده‌های طرح آمارگیری از کارگاه‌های صنعتی

مجری طرح:

محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

مرداد ۱۳۸۷

عنوان طرح:

تعیین ریسک افشا و نحوه‌ی انتشار ایمن داده‌های طرح آمارگیری از کارگاه‌های صنعتی

مجری طرح: محسن محمدزاده

همکاران طرح: ربیع‌الله رحمانی، محسن ملانوری، محمد بردبار عشرت آبادی

مشاور طرح: علی‌رضا زاهدیان

چکیده

یکی از وظایف مهم هر سازمان آماری جمع آوری اطلاعات و انتشار آنها به شکل داده است. معمولاً گاهی تمام یا بخشی از اطلاعات جمع آوری شده به جنبه‌های خصوصی یا تجاری پاسخگویان بستگی دارد، که از نظر آنان حساس و محرمانه هستند. انتشار کامل اینگونه داده‌ها با خطر جدی افشای اطلاعات محرمانه پاسخگویان و نارضایتی آنان مواجه است. لذا برای حفظ محرمانگی داده‌ها و رعایت حریم خصوصی پاسخگویان معمولاً از انواع شیوه‌های دسترسی محدود به داده‌ها استفاده می‌شود. از طرفی کاربران و به خصوص محققان نیاز به دسترسی اطلاعات کافی دارند و محدودسازی دسترسی به اطلاعات باعث کاهش کارایی یا محدودیت در انواع تحلیل‌های آماری مورد نظر کاربران می‌شود. لذا بایستی در انتشار داده‌ها حقوق کاربران نیز در مورد استفاده وسیع از داده‌ها مورد توجه قرار گیرد. بنابراین هر سازمان آماری از یک سو بنا به وظیفه قانونی متعهد به انتشار گسترده اطلاعات به شکل داده‌های آماری است و از سویی دیگر بایستی از اطلاعات خصوصی پاسخگویان مراقبت نموده و این تضمین را بدهد که ضمن ارایه بیشترین اطلاعات به جامعه، حریم شخصی پاسخگویان در حد معقول حفظ می‌شود. بنابراین برای جلوگیری از تضییع حقوق پاسخگویان و رعایت حقوق کاربران، لازم است بین این دو اصل تعادل ایجاد شود. لازمه برقراری این تعادل، شناسایی و اندازه‌گیری ریسک افشای اطلاعات خصوصی در داده‌ها و میزان اطلاعات موجود در آنها است.

در این تحقیق ضمن بیان مفاهیم مربوط به مساله افشا، معیارهایی برای اندازه‌گیری و ارزیابی ریسک افشا به تفکیک شکل و وضعیت انتشار داده‌ها معرفی شده و تکنیک‌هایی برای کاهش ریسک افشا ارائه شده‌اند. سپس نحوه انتشار ایمن انواع داده‌ها نشان داده شده است. در انتها با استفاده از معیارهای ارائه شده و راههای کاهش ریسک افشای اطلاعات، شکل ایمن هر یک از جداول مربوط به طرح سرشماری عمومی از کارگاه‌های صنعتی که توسط مرکز آمار ایران در سال ۱۳۸۴ اجرا گردیده، ارائه شده‌اند.

فهرست مندرجات

لیست جداول

فصل ۱

مفاهیم افشای اطلاعات

۱.۱ مقدمه

حفاظت از اطلاعات شخصی موضوعی است که هم در بخش خصوصی و هم در بخش دولتی مطرح و مورد توجه است. سازمان‌ها و مراکزی از قبیل بانک‌ها، بیمارستان‌ها، موسسات بیمه و ... اطلاعات گسترده‌ای را در مورد مشتریان خود جمع‌آوری می‌کنند که بخشی از آنها از حساسیت زیاد و متقابلاً از حمایت قانون برخوردارند. سازمان‌های آماری نیز اطلاعات بسیار زیادی را در مورد جامعه آماری جمع‌آوری می‌کنند که این جمع‌آوری می‌تواند به صورت مصاحبه حضوری، مصاحبه تلفنی، مصاحبه پستی و ... باشد، یا اینکه برای کاهش بار پاسخگویان و کمتر کردن هزینه جمع‌آوری داده‌ها، از رکوردهای اداری یا ثبتی استفاده شود. این اطلاعات به هر طریقی که جمع‌آوری شوند، بخشی یا تمام آن به جنبه‌هایی از زندگی شخصی یا تجاری پاسخگویان مربوط می‌شود، که از نظر آنان این جنبه‌ها حساس و محرمانه تلقی می‌شود. مساله این است که با توجه به روند روز افزون بهره‌گیری از اطلاعات جمع‌آوری شده، خطر آشکار شدن اطلاعات محرمانه موجود در آنها نیز افزایش می‌یابد. در مباحث مربوط به افشای داده‌های آماری، چهار شخصیت زیر در نظر گرفته می‌شوند:

الف— گردآورنده اطلاعات، مانند سازمان آماری،

ب— پاسخگو، مانند فرد یا خانوار،

ج— کاربر، مانند محقق دانشگاهی،

د— متخلف[۱]، فردی که بر خلاف قانون در صدد دسترسی به اطلاعات محرمانه است.

در سال‌های اخیر رشد قابل ملاحظه‌ای در ایجاد رکوردهای اداری در سازمان‌ها و دستگاه‌های مختلف مشاهده می‌شود و تمرکز روی جمع‌آوری داده‌های آماری برای امور تحقیقاتی بویژه در علوم اجتماعی و پزشکی توسط گردآورندگان داده‌ها نیز افزایش یافته است. جای شگفتی نیست اگر به طور هم زمان علاقه به دسترسی به اطلاعات آماری از جانب بخش دولتی، صنعت، محققان دانشگاهی و کلاً کاربران نیز افزایش یافته باشد. به راستی جمع‌آوری اطلاعات آماری و عرضه آنها به بخش دولتی و خصوصی، یک صنعت اطلاعاتی جدید را به وجود آورده است. برای مثال در بعضی کشورهای صنعتی پیشرفته، دولت‌ها از انتشار داده‌های آماری به عنوان روشی برای کمک به جبران هزینه‌های جمع‌آوری داده‌ها حمایت می‌کنند. تقاضا برای داده‌های آماری نه تنها از جانب حوزه‌های سیاستگزاری وجود دارد، بلکه امروزه بانک‌ها، موسسات تجاری و حتی محققان خصوصی و روزنامه نگاران نیز متقاضی استفاده از داده‌های آماری هستند. نکته مهم در انتشار داده‌های آماری، وجود خطر افشای اطلاعات شخصی موجود در داده‌ها توسط متخلف است. این شخص ممکن است علاقه‌مند به دستیابی به اطلاعاتی در مورد افرادی خاص باشد، یا تلاش کند تا با افشای یک سری اطلاعات، گردآورنده آنها را بی اعتبار سازد، یا از این طریق هوش خود را به اثبات برساند. این عمل به هر دلیل صورت پذیرد، موجب بی اعتمادی عمومی نسبت به سازمان آماری به عنوان حافظ اطلاعات خصوصی پاسخگویان شده و کاهش یا عدم همکاری آنها را در سرشماری‌ها و طرح‌های تحقیقاتی آینده در پی دارد. این اتفاق به هیچ وجه مورد قبول سازمان آماری و جامعه نیست. بنابراین سازمان آماری با این مساله مواجه است که چگونه بین انتشار هر چه وسیع‌تر داده‌های آماری برای استفاده‌های مشروع کاربران و افشای اطلاعات شخصی و محرمانه پاسخگویان از طریق داده‌های منتشر شده، تعادل برقرار نماید. به عبارت دیگر، سازمان آماری از یک سو بنا به وظیفه قانونی و رشد روز افزون تقاضا برای اطلاعات جمع‌آوری شده، خود را متعهد به انتشار گسترده و با کیفیت اطلاعات به شکل داده‌های آماری می‌بیند و از سویی دیگر با توجه به نگرانی عمومی نسبت به افشای اطلاعات شخصی و وظیفه قانونی این سازمان‌ها برای مراقبت از اطلاعات خصوصی پاسخگویان، باید این تضمین را بدهد که ضمن ارایه بیشترین اطلاعات به جامعه، حریم شخصی پاسخگویان در حد معقول حفظ شود. برای برخورد با این مساله معمولاً سازمان‌های آماری از روش‌هایی تحت عناوین دسترسی محدود شده و داده‌های محدود شده استفاده می‌کنند. در شیوه دسترسی محدود شده، روی اینکه چه کسی، چگونه و

[۱] Intruder

برای چه هدفی و برای کدام متغیرها به داده‌ها دسترسی پیدا کند، شرایطی گذاشته می‌شود. این روش شامل ضعف‌هایی می‌باشد. به این صورت که شخص کاربر نمی‌تواند هر نوع تحلیل آماری دلخواه خود را روی داده‌ها انجام دهد. همچنین تعداد کاربرانی را که می‌توانند به داده‌ها دسترسی پیدا کنند، محدود می‌سازد. علاوه بر این کمبود اطلاعات در یک زمینه معین می‌تواند فاجعه آفرین باشد، زیرا با ورود به عصر اطلاعات، تمایل افراد و سازمان‌ها برای بدست آوردن اطلاعات بیشتر به شدت افزایش یافته است. چون داده‌های آماری با هزینه دولت جمع‌آوری می‌شوند، عدم انتشار آنها به منزله تضییع حقوق افراد جامعه بوده و در نتیجه روش دسترسی محدود شده فقط باید در صورتی استفاده شود، که حفظ جنبه محرمانه اطلاعات، از راه‌های دیگر غیر ممکن باشد. در روش داده‌های محدود شده، داده‌های آماری تعدیل می‌شوند به گونه‌ای که امکان تجزیه و تحلیل آماری آنها وجود داشته و در عین حال زیانی متوجه پاسخگو نشود. بعضی از سازمان‌ها برای کاهش خطر افشای اطلاعات، تنها نمونه‌هایی از اطلاعات سرشماری یا زیر نمونه‌هایی از نمونه‌گیری‌های گسترده خود را گزارش می‌کنند. با اتخاذ چنین سیاست‌هایی در کنار یک سری اقدامات قانونی، شمار روز افزونی از سازمان‌ها به روش‌هایی تحت عنوان فنون محدود سازی ریسک افشا، که در اصطلاح آماری تکنیک‌های کنترل افشای آماری[۲] (SDC) نامیده می‌شوند، روی می‌آورند. نوع و نحوه بکارگیری این روش‌ها، به نوع داده‌هایی که باید منتشر شوند بستگی دارد.

با توجه به مطالب بیان شده، سازمان آماری باید بین میزان اطلاعات موجود در داده‌ها و ریسک افشای متناظر آنها، تعادل برقرار کند. اگر سازمان آماری داده‌ها را به تفصیل و با جزئیات کامل منتشر کند، آنگاه حجم زیادی از اطلاعات را در اختیار متقاضیان داده‌ها قرار داده است. اما انتشار داده‌ها به این شکل، ریسک افشا را به شدت افزایش می‌دهد. به طور کلی در مورد اطلاعات جمع‌آوری شده توسط سازمان آماری و نحوه انتشار آنها، سه وضعیت انتشار به شرح زیر وجود دارد.

۱ — داده‌ها مربوط به همه افراد جامعه بوده و سازمان آماری می‌خواهد همه آنها را منتشر کند.

۲ — داده‌ها مربوط به همه افراد جامعه بوده و سازمان آماری می‌خواهد نمونه‌ای از آنها را منتشر کند.

۳ — داده‌ها مربوط به نمونه‌ای از جامعه بوده و سازمان آماری می‌خواهد همه آنها را منتشر کند.

هر یک از این وضعیت‌های سه‌گانه انتشار می‌توانند از طریق ریسک افشا مورد بررسی قرار گیرند. یک روش مناسب ارزیابی، تعیین سطح قابل پذیرش برای ریسک افشا است. به این ترتیب که در بین اشکال مختلف داده‌ها با ریسک افشای قابل پذیرش، آن شکلی برای انتشار انتخاب شود، که دارای

[۲] Statistical disclosure control

بیشترین حجم اطلاعات است.

افشای داده‌های آماری و محرمانگی[۳] اطلاعات شخصی موضوع‌هایی هستند که در چهار دهه اخیر مورد توجه سازمان‌های آماری و محققین قرار گرفته‌اند. بحث درباره این موضوع با مقالات دالینیوس (۱۹۷۷) و دالینیوس و ریس (۱۹۷۸) و سمیناری شروع شد که در سال ۱۹۷۸ توسط کمیته فدرالی روش‌شناسی آماری[۴] ($FCSM$) در آمریکا برگزار گردید و مقالات آن از طریق سایت $http://www.fcsm.gov/working-papers/sw۲.html$ قابل دسترسی هستند. هدف این سمینار کمک به سازمان‌های آماری آمریکا برای حفظ محرمانگی داده‌ها بود. لذا در مقالات ارایه شده عموماً به تعریف و تبیین مفاهیم مرتبط باافشا و لزوم بررسی آن پرداخته شد. از آن پس مطالعات در زمینه افشا با سرعت چشمگیری در آمریکا و اروپا گسترش یافت. کشورهای آمریکا، کانادا، هلند، انگلستان و ایتالیا از جمله کشورهای پیشرو در زمینه تحقیق در این موضوع هستند. بیشتر تحقیقات قبل از سال ۱۹۹۰ به اساس مساله افشا و چگونگی رخ دادن آن محدود بود. از تحقیق‌های مهم آن دوره می‌توان به مقالات کیم (۱۹۸٦)، گریفین و همکاران (۱۹۸۹)، کلر و همکاران (۱۹۸۹)، بتلهم و همکاران (۱۹۹۰) و گرینبرگ و واشل (۱۹۹۰) اشاره کرد.

اندازه‌گیری ریسک افشا و ارزیابی آن، موضوع تحقیقات سال‌های بعد از ۱۹۹۰ بوده است، که می‌توان به مقالاتی مانند گرینبرگ و زایتز (۱۹۹۲)، لامبرت (۱۹۹۳) اشاره نمود. در سال ۱۹۹۲ کمیته‌ای با عنوان ریسک افشا در سازمان برنامه و بودجه آمریکا تشکیل شد. اهداف این کمیته، ارایه معیارها و راهنمایی‌هایی برای انتخاب و اعمال تکنیک‌های SDC مناسب و استفاده از نرم‌افزارهای تخصصی در زمینه افشا بود. این کمیته در سال ۱۹۹۴ سمیناری برگزار کرد که مقالات آن نیز از طریق سایت $http://www.fcsm.gov/working-papers/spwp۲۲-rev.pdf$ قابل دسترسی هستند.

نتایج ادامه تحقیق‌ها در زمینه اندازه‌گیری ریسک افشا و ارزیابی آن در چن و کلر (۱۹۹۸)، فینبرگ و ماکو (۱۹۹۸)، الیوت (۲۰۰۰)، اسکینر و الیوت (۲۰۰۲) و پولتینی و استاندر (۲۰۰۴) آمده است. در زمینه افشا، تاکنون کنفرانس‌های بین‌المللی متعددی در کشورهای مختلف برگزار شده است، که کنفرانس سال ۱۹۹۹ در یونان تحت عنوان محرمانگی داده‌های آماری و نیز کنفرانس‌های سال ۱۹۹۸ در پرتغال، سال ۲۰۰۱ در مقدونیه و سال ۲۰۰۴ در اسپانیا تحت عنوان حفاظت داده‌های آماری از آن جمله‌اند.

---

[۳] Confidentiality

[۴] Federal Committee on Statistical Methodology

کمیته اقتصادی اتحادیه اروپا در سال ۱۹۹۸ در بعضی از کشورهای اروپای شرقی و کشورهای مستقل مشترک‌المنافع مساله افشا را مورد بررسی قرار داد. نتایجی که از این بررسی حاصل شد، نشان داد که در بیشتر کشورهای اروپای شرقی مساله افشا برای سازمان‌های آماری بسیار حائز اهمیت بوده و در این کشورها بستر قانونی برای محدودسازی ریسک افشا وجود دارد. در حالیکه در کشورهای مستقل مشترک‌المنافع حتی الزام قانونی برای محدودسازی ریسک افشا وجود نداشت و سازمان‌های آماری این کشورها تنها به جنبه‌های ریاضی ریسک افشا، آن هم به صورت محدود می‌پرداختند. کمیته اقتصادی اتحادیه اروپا در سال ۲۰۰۰ بررسی خود را در همان کشورها تکرار کرد. نتایجی که از این بررسی حاصل شد، نمایانگر افزایش توجه همه کشورها به مساله افشا بود. سرشماری جمعیتی سال ۲۰۰۰، سرشماری‌های بخش کشاورزی، افزایش اختلاف درآمد ثروتمند و فقیر و افزایش جرم و جنایت از جمله دلایلی بودند که به عنوان علل این افزایش بیان شدند. همچنین نتایج این بررسی نشان داد که سازمان‌های آماری بیشتر به جنبه قانونی و اجرایی افشا پرداخته و جنبه ریاضی مساله افشا کمتر مورد توجه آنها بوده است.

اهمیت مساله افشا برای سازمان‌های آماری به حدی زیاد است که این سازمان‌ها کمیته‌های ویژه‌ای برای بررسی این مساله تشکیل می‌دهند. به عنوان مثال، در دفتر سرشماری آمریکا کمیته‌ای تحت عنوان کمیته اجرایی نظارت داده($DSEP$[۵]) به همین منظور تاسیس شده است. در سال‌های اخیر محققین به استفاده از روش‌های رایج آماری برای ارزیابی ریسک افشا روی آورده‌اند. اگانیان و فرر (۲۰۰۳) با استفاده از آنتروپی ریسک افشا را برآورد کردند. رینوت (۲۰۰۳)، السايد (۲۰۰۴) و فارستر و وب (۲۰۰۵) با استفاده از روش‌های بیزی ریسک افشا را مورد بررسی قرار دادند.

در بخش دوم این فصل شکل‌های مختلف انتشار اطلاعات آماری معرفی می‌شوند. سپس به چگونگی رخ دادن افشا و معرفی انواع افشا و عوامل موثر بر افشا پرداخته می‌شود. معیارهای اندازه‌گیری ریسک افشادر فصل دوم و تکنیک‌های کاهش ریسک افشادر فصل سوم ارائه خواهند شد. نحوه انتشار ایمن داده‌ها موضوع فصل چهارم می‌باشد. در فصل پنجم طرح آمارگیری از کارگاه‌های صنعتی معرفی و جداولی که ریسک افشای آنها بر اساس معیارهای ارائه شده ارزیابی و برای انتشار ایمن شده اند ارایه می‌شوند.

[۵] Data Stewardship Executive Policy Committe

۲.۱ شکل‌های انتشار و افشای اطلاعات آماری

در این بخش ابتدا به شکل‌های مختلف انتشار اطلاعات آماری و مسایل افشای آنها پرداخته می‌شود. سپس دلایل رخ دادن افشا، انواع افشا و عوامل موثر بر آن‌ها بیان می‌گردند.

۱.۲.۱ نحوه انتشار اطلاعات آماری

سازمان‌های آماری معمولاً داده‌ها را به دو شکل جدولی[٦] و خرد[۷] منتشر می‌کنند. داده‌های جدولی شامل داده‌های جمع‌بندی شده‌ای هستند، که به دو شکل جدول‌های مقداری[۸] و جدول‌های فراوانی[۹] منتشر می‌شوند. هر دو نوع جدول بر اساس یک، دو یا چند متغیر رسته‌ای[۱۰] ساخته می‌شوند، که در جدول مقداری، متغیر نمایش داده شده در خانه‌های جدول، متغیری پیوسته بوده و مقدار آن در هر خانه برابر مجموع مقادیر آن متغیر برای پاسخگویانی است که در شرایط آن خانه صدق می‌کنند. ولی در جدول فراوانی عدد نمایش داده شده در هر خانه برابر تعداد پاسخگویانی است که در شرایط آن خانه صدق می‌کنند. به عنوان مثال میزان سرمایه گذاری بر اساس ناحیه و نوع فعالیت، یک جدول مقداری و تعداد کارخانجات بر حسب ناحیه و نوع فعالیت، یک جدول فراوانی را تشکیل می‌دهد.

شکل دیگر انتشار داده‌ها، به صورت مجموعه داده‌های خرد است. این مجوعه داده‌ها که "فایل داده‌های خرد" نامیده می‌شود، دارای رکوردهایی است که شامل اطلاعات زیادی در مورد پاسخگویان است. به عبارت دیگر هر رکورد شامل مقادیر تعدادی متغیر برای یک پاسخگو است. بسته به اینکه پاسخگوی مورد نظر چه باشد، متغیرهای مشمول در داده‌های خرد می‌توانند محل سکونت، شغل، میزان تولید، نوع فعالیت و ... باشند. هر چند فایل داده‌های خرد در حقیقت داده‌های خامی هستند، که معمولاً برای تهیه جدول‌ها به کار می‌رود، اما به دلیل تقاضای کاربران به شکل قابل قبولی نیز منتشر می‌شوند.

امروزه اطلاعات آماری نسبت به سه دهه گذشته در حوزه‌های وسیع‌تری جمع آوری می‌شوند، که متقابلاً درخواست بیشتری را برای انتشار اطلاعات آماری در پی داشته است. همچنین وجود

---

[٦] Tabular data
[۷] Microdata
[۸] Magnitude tables
[۹] Frequency tables
[۱۰] Categorical variable

کامپیوترهای پیشرفته و نرم افزارهای متعدد موجب شده است که بسیاری از محققان بتوانند تحلیل‌های متنوع و مفصلی را روی فایل‌های حجیم داده‌ها انجام دهند. لذا آنها به جدول‌های استانداردی که معمولاً سازمان آماری به عنوان خلاصه‌ای از اطلاعات جمع آوری شده منتشر می‌کند، نیاز مبرم ندارند، و تمایل دارند با استفاده از داده‌های خرد، جدول‌های مورد نیاز خود را تولید نمایند.

جدول‌ها که رایج‌ترین محصولات سازمان آماری هستند، شامل داده‌های جمع‌بندی شده به عنوان مقادیر خانه‌های جدول می‌باشند. چون داده‌ها مستقیماً پاسخ افراد پاسخگو نیستند، به نظر می‌رسد که خطر افشای اطلاعات شخصی وجود ندارد، در حالی که لزوماً این گونه نیست. برای روشن‌تر شدن مطلب، جدول ۱.۱ که نشان دهنده میزان فروش تعدادی موسسه به تفکیک نوع و ناحیه فعالیت می‌باشد، ارایه شده است. در نگاه اول این جدول برای انتشار، مناسب به نظر می‌رسد. زیرا هر خانه

جدول ۱.۱: میزان فروش شرکت‌ها به تفکیک نوع و ناحیه فعالیت

ناحیه فعالیت / نوع فعالیت	A	B	C	جمع
۱	۱۱	۴۷	۵۸	۱۱۶
۲	۱	۱۵	۳۳	۴۹
۳	۲	۳۱	۲۰	۵۳
جمع	۱۴	۹۳	۱۱۱	۲۱۴

جدول تنها شامل داده خلاصه شده‌ای است که مربوط به موسسه خاصی نمی‌باشد. اما این نتیجه‌گیری عجولانه است. برای مثال فرض کنید فقط یک موسسه در ناحیه B با فعالیت نوع دو وجود دارد. با انتشار این جدول مشخص می‌شود که میزان فروش این موسسه برابر ۱۵ است. پس اگر قرار باشد که میزان فروش این موسسه حفظ گردد، جدول نباید منتشر شود. بنابراین برای این مساله، جدول‌هایی که دارای خانه‌هایی شامل داده‌های مربوط به فقط یک پاسخگو هستند، نباید منتشر شوند. ولی موضوع همیشه به این سادگی نیست. اگر در ناحیه B و فعالیت دو به جای یک موسسه، دو موسسه وجود داشته باشد، آنگاه هر کدام از این دو موسسه به راحتی می‌تواند به میزان فروش دیگری پی ببرد. ممکن است منتشر کننده داده‌ها به جای وجود حداقل دو عضو در هر خانه جدول، خواستار وجود سه عضو یا بیشتر باشد. متأسفانه حتی این حالت نیز رضایت بخش نیست. فرض کنید در ناحیه B، ده موسسه دارای فعالیت نوع دو باشند و میزان فروش یکی از آنها ۹۵ درصد مقدار کل فروش آن خانه باشد. در این حالت اگر متخلف بداند که میزان فروش این موسسه خیلی بالاست، می‌تواند برآورد نسبتاً خوبی از میزان فروش موسسه مذکور به دست آورد.

چون مجموعه داده‌های خرد محصولات نسبتاً جدید سازمان‌های آماری هستند، مسایل کنترل افشای مربوط به آنها نیز جدیدند. معمولاً فایل داده‌های خرد به دو صورت فایل داده‌های خرد برای محققان و فایل داده‌های خرد برای عموم منتشر می‌شود، که نوع اول نسبت به نوع دوم دارای اطلاعات بیشتری است. هرگاه سازمان آماری بخواهد مجموعه داده‌های خرد را منتشر کند، متغیرهایی از قبیل نام، آدرس و شماره تلفن را حذف می‌کند. ظاهراً به نظر می‌آید انتشار چنین داده‌هایی فاقد خطر افشا است، اما در حالت کلی این امر صحیح نیست. به عنوان مثال، اگر سازمان آماری مجموعه‌ای از داده‌های خرد شامل اطلاعاتی در مورد محل سکونت، شغل و سابقه جنایی پاسخگویان را منتشر کند، که در آن رکوردی به صورت "محل سکونت: لندن، شغل: شهردار، سابقه جنایی: یک مورد" وجود داشته باشد، آنگاه به سادگی می توان پی برد که پاسخگو چه کسی است. به ویژه می‌توان نتیجه گرفت که شهردار لندن دارای یک مورد سابقه جنایی است. به عنوان مثالی دیگر اگر در شهری کوچک فقط یک نانوا وجود داشته باشد و علاوه بر شغل و محل سکونت، اطلاعات دیگری در مورد او که بخشی از آنها ممکن است محرمانه باشد در مجموعه داده‌های افراد منتشر شوند، خطر افشا وجود دارد. افرادی که می‌دانند تنها یک نانوا در این شهر است، می‌توانند اطلاعات محرمانه او را بدست آورند. اگرچه در اینجا تنها متغیرهای شغل و محل سکونت مثال زده شدند، اما ترکیب‌های دیگری از متغیرها نیز می‌تواند به همین راحتی موجب افشا شوند. بویژه مقادیر خیلی بزرگ یا خیلی کوچک برخی متغیرها می‌توانند موجب افشا گردند. به عنوان مثال، شخصی با ۱۰ فرزند یا شخصی با درآمد سالیانه ۱۰۰ میلیون تومان در منطقه‌ای معین از این نوع هستند. رکوردهایی که بیشتر در معرض خطر افشا قرار دارند، ناایمن[۱۱] گفته می‌شوند. تشخیص اینکه در فایل داده‌های خرد کدام رکوردها ناایمن هستند، بستگی به تعریف سناریوی افشا[۱۲] دارد که در بخش‌های بعدی معرفی خواهد شد. مساله اصلی این است که پس از تشخیص رکوردهای ناایمن، فایل داده‌های خرد چگونه تعدیل شود که با از دست دادن کمترین اطلاعات، انتشار ایمن داده‌ها امکان پذیر شود.

۲.۲.۱ نحوه افشای اطلاعات

نهادها یا اشخاص اطلاعات را از سازمان آماری برای تجزیه و تحلیل آماری درخواست می‌کنند. معمولاً برای کاربران اطلاعات، خود داده‌ها مهم هستند نه اینکه چه داده‌ای متعلق به چه کسی است.

---

[۱۱] Unsafe

[۱۲] Disclosure scenario

بنابراین سازمان آماری در پاسخ به درخواست نهادها و اشخاص، فقط داده‌ها را منتشر می‌کند. در این حالت متخلف ممکن است با ابزارهایی بتواند تشخیص دهد که چه داده‌ای متعلق به چه کسی است و افشا رخ دهد. افشای داده‌های آماری قبل از انتشار آنها باید بررسی شود، زیرا

الف— فاش شدن حریم خصوصی افراد، ناقض اصول اخلاقی است.

ب— انتشار اطلاعات بدون امکان افشا وظیفه قانونی سازمان آماری است.

ج— افشا موجب بی اعتمادی جامعه نسبت به سازمان آماری می‌گردد.

اطلاعات جمع آوری شده در مورد هر فرد، می‌تواند شامل مقادیر متغیرهایی تحت عنوان متغیرهای شناسایی[۱۳] و حساس[۱۴] باشد. متغیرهای شناسایی به دو دسته مستقیم و غیرمستقیم، تقسیم می‌شوند. یک متغیر شناسایی مستقیم، ممکن است به تنهایی منجر به افشا شود، در حالی که متغیر شناسایی غیرمستقیم، معمولاً در ترکیب با سایر متغیرها می‌تواند منجر به افشا شود. نام، آدرس، شماره تلفن و شماره ملی از جمله متغیرهای شناسایی مستقیم و سن، جنس، محل سکونت، محل کار، وضعیت شغل و وضعیت تاهل مثال‌هایی از متغیرهای شناسایی غیرمستقیم هستند. به منظور حفظ حریم خصوصی پاسخ دهندگان، سازمان آماری باید از انتشار مقادیر متغیرهای شناسایی مستقیم، خودداری کند. بنابراین در این تحقیق منظور از متغیر شناسایی، متغیر شناسایی غیرمستقیم می‌باشد. متغیرهای حساس به متغیرهایی گفته می‌شود که پاسخ دهندگان از افشای مقادیر آنها نگران هستند. سابقه جنایی و درآمد از جمله متغیرهای حساس می‌باشند. در مباحث افشای آماری، شخصی که متخلف به دنبال افشای اطلاعات شخصی او است، فرد هدف نامیده می‌شود. در عمل ممکن است متخلف قبل از انتشار داده‌ها، فقط از مقادیر برخی متغیرهای شناسایی فرد هدف مطلع باشد، اما برای حفظ حریم خصوصی افراد، سازمان آماری باید اینگونه فرض کند که شخص متجاوز از مقادیر همه متغیرهای شناسایی فرد هدف مطلع است. اگر متخلف بتواند به درستی داده مربوط به فرد هدف را تشخیص دهد، موفق به شناسایی آن شخص شده است. مسلماً متخلف بدون داشتن اطلاعات اولیه راجع به فرد هدف، قادر به شناسایی او نخواهد شد. با توجه به مفاهیم شناسایی و افشا مشخص است که شناسایی پیش‌نیاز افشا است. پس برای محافظت از داده‌ها در مقابل افشا، باید چگونگی اتفاق افتادن شناسایی در عمل مشخص شود.

متغیر حاصل از ترکیب رده‌های متغیرهای شناسایی، متغیر راهنما گفته می‌شود. تعداد سطوح راهنما

---

[۱۳] Identifying variable

[۱۴] Sensitive variable

برابر با حاصلضرب تعداد رده‌های متغیرهای شناسایی است. به عنوان مثال متغیر راهنمای حاصل از ترکیب رده‌های دو متغیر شناسایی وضعیت تأهل و اشتغال متغیری چهار سطحی با سطوح متأهل — شاغل، متأهل — بیکار، مجرد — شاغل، مجرد — بیکار می‌باشد. هرگاه شخصی در جامعه یا نمونه متعلق به سطحی از متغیر راهنما با فراوانی یک باشد، به ترتیب یکتا در جامعه[۱۵] یا یکتا در نمونه[۱۶] نامیده می‌شود. با توجه به مطالب بیان شده، واضح است که یکتایی منجر به شناسایی و آن نیز منجر به افشا خواهد شد.

۳.۲.۱ انواع افشا

افشای اطلاعات موجب استفاده نابجا (نامناسب) از اطلاعات موجود در داده‌های مربوط به یک شخص یا سازمان می‌گردد. به طور کلی دو نوع افشای اطلاعات تحت عناوین افشای هویت[۱۷] و افشای صفت[۱۸] وجود دارد. در افشای هویت، که در واقع مهمترین نوع افشا است، ابتدا فرد هدف شناسایی شده، سپس بر اساس هویت آن، اطلاعات محرمانه وی از داده‌ها استخراج می‌شود. اما همواره تعیین هویت کامل، شرط لازم برای افشای اطلاعات محرمانه پاسخگو نیست. گاهی دانستن اینکه یک پاسخگو عضوی از یک گروه خاص است، برای افشای یک سری اطلاعات وی کفایت می‌کند. این نوع افشا در اصطلاح "افشای صفت" نامیده می‌شود. به عنوان مثال، اگر معلوم شده باشد که شخص I در گروه G قرار دارد و از طرفی از طریق داده‌های منتشر شده مشخص شود که درآمد هر یک از افراد گروه G بیشتر از مبلغ T است، می‌توان نتیجه گرفت که درآمد شخص I بیش از مبلغ T است. در این تحقیق همواره منظور از افشا، همان افشای هویت می‌باشد.

۴.۲.۱ عوامل موثر بر افشا

افشای اطلاعات ممکن است تحت سناریوهای مختلف صورت پذیرد. هر سناریو مدلی است که شخص متخلف در صدد دستیابی به اطلاعات محرمانه افراد به پیروی از آن است. اطلاع از سناریوی شخص متخلف می‌تواند سازمان آماری را در انتخاب روش‌های مناسب کنترل افشای آماری کمک

[۱۵] Population unique
[۱۶] Sample unique
[۱۷] Identity disclosure
[۱۸] Attribute disclosure

کند. یکی از عوامل تاثیرگذار بر افشا، فاصله بین زمان جمع آوری اطلاعات تا زمان انتشار آنها است. هر چه این فاصله زمانی بیشتر باشد، خطر افشا کمتر می‌شود، زیرا با گذشت زمان مقدار متغیرهای حساس پاسخ دهندگان تغییر کرده و در نتیجه انگیزه متخلف برای دستیابی به آن مقادیر کم می‌شود. ریسک افشا در مورد داده‌های حاصل از یک بررسی مبتنی بر نمونه‌گیری تقریباً کمتر از ریسک افشا در مورد داده‌های حاصل از یک سرشماری می‌باشد. با کاهش کسر نمونه‌گیری یکتاهای کمتری در نمونه ظاهر می‌شوند، لذا احتمال اینکه متخلف با استفاده از داده‌های نمونه به اطلاعات محرمانه آنها دست پیدا کند، کاهش می‌یابد.

فصل ۲

معیارهای اندازه‌گیری ریسک افشا

چون داده‌های مندرج در جدول‌ها مستقیماً پاسخ‌های افراد نیستند، در نظر اول فاش شدن اطلاعات محرمانه با انتشار جدول‌ها بعید به نظر می‌رسد. در حالیکه انتشار آنها نیز می‌تواند به راحتی موجب افشای اطلاعات خصوصی پاسخگویان شود. جدول‌ها به دو دسته فراوانی و مقداری تقسیم می‌شوند. همه متغیرهای جدول فراوانی رسته‌ای بوده و عدد مربوط به هر خانه، تعداد افرادی است که در شرایط آن خانه صدق می‌کنند. در حالی که معمولاً یکی از متغیرهای جدول مقداری پیوسته بوده و عدد مربوط به هر خانه، مجموع مقادیر متغیر مذکور برای افراد آن خانه است. علیرغم تشابه ساختار این دو نوع جدول، ارزیابی ریسک افشای آنها متفاوت است. خانه‌ای از جدول که اطلاعات خصوصی افراد موجود در آن بیشتر در معرض خطر افشا قرار دارد، حساس نامیده می‌شود. در این فصل معیارهای اندازه‌گیری ریسک افشا و نحوه کاهش آنها برای داده‌های خرد در بخش اول ارائه شده‌اند. سپس معیارهای اندازه‌گیری ریسک افشای جدول‌های فراوانی و جدول‌های مقداری به ترتیب در بخش‌های دوم و سوم ارئه گردیده‌اند.

۱.۲ ریسک افشای داده‌های خرد

فایل داده‌های خرد، فهرستی از رکوردها حاوی مقادیر متغیرهای شناسایی و حساس پاسخ دهندگان است. ریسک افشای داده‌های خرد، در هر کدام از وضعیت‌های انتشار سه‌گانه باید ارزیابی شود. همان

طور که در فصل اول بیان شد، افشا با شناسایی و آن نیز با یکتایی امکان پذیر می‌گردد. بنابراین تعداد رکوردهای یکتا می‌تواند مبنای ارزیابی ریسک افشا قرار بگیرد. برای روشن‌تر شدن مطلب، مثال زیر ارایه می‌شود.

مثال ۱.۲ : فرض کنید از افراد A_1، A_2، ...، A_{12} در مورد متغیرهای شناسایی جنس، سن و وضعیت تاهل و متغیر حساس درآمد سوال شده و پاسخ‌های آنها به صورت زیر کدگذاری و در جدول ۱.۲ خلاصه شده باشد، که در آن رکورد iام مربوط به فرد A_i است.

جنس: مرد $=$ ۱ و زن $=$ ۰،

سن: ۰—۲۵ $=$ ۰، ۲۵—۴۰ $=$ ۱، ۴۰—۶۰ $=$ ۲ و ۶۰ به بالا $=$ ۳،

وضعیت تاهل: متاهل $=$ ۱ و مجرد $=$ ۰،

درآمد: کم $=$ ۰، متوسط $=$ ۱ و زیاد $=$ ۲،

ابتدا فرض کنید جدول ۱.۲ داده‌های خرد مربوط به جامعه بوده و A_7 فرد هدف باشد. با توجه به

جدول ۱.۲: فایل داده‌های خرد

شماره رکورد	جنس	سن	وضعیت تاهل	درآمد
۱	۱	۱	۱	۰
۲	۰	۳	۱	۲
۳	۱	۱	۰	۱
۴	۱	۰	۰	۲
۵	۱	۱	۰	۱
۶	۰	۲	۰	۰
۷	۱	۳	۱	۲
۸	۱	۰	۰	۱
۹	۰	۳	۱	۱
۱۰	۱	۲	۱	۲
۱۱	۰	۱	۰	۰
۱۲	۱	۱	۰	۲

مفاهیم شناسایی و یکتایی تشخیص زیاد بودن درآمد A_7 برای متخلف کار ساده‌ای است، زیرا فراوانی رکورد مربوط به او یعنی "مرد، ۶۰ به بالا، متاهل" در فایل داده‌ها یک می‌باشد. در مورد افراد A_1، A_6، A_{10} و A_{11} وضعیت به همین صورت است. پس رکوردهای یکتا امنیت داده‌ها را به شدت تهدید می‌کنند. تشخیص درآمد فرد هدف با افزایش فراوانی رکورد مربوط به او مشکل‌تر می‌شود. البته رکوردهای با فراوانی کم، حتی در صورت یکتا نبودن می‌توانند منجر به افشا شوند. به عنوان مثال، با

انتشار جدول داده‌های خرد معلوم می‌شود که درآمد A_2 و A_4 متوسط یا زیاد می‌باشد، زیرا فراوانی رکوردهای مربوط به آنها یعنی "زن، ۶۰ به بالا، متاهل" و "مرد، ۰—۲۵، مجرد" دو بوده و درآمد متناظر با این رکوردها متوسط یا زیاد می‌باشد. نوع دیگری از افشا در مورد رکوردهای با فراوانی دو در این داده‌ها به این ترتیب است که افراد A_2 و A_9 و همچنین دو فرد A_4 و A_8 از درآمد یکدیگر مطلع می‌شوند. چون مقدار متغیرهای شناسایی A_2 و A_9 یکسان بوده و A_2 از درآمد خودش مطلع است، تشخیص متوسط بودن درآمد A_9 برای او کار ساده‌ای است. پس اگر فایل داده‌ها مربوط به جامعه باشد، رکوردهای با فراوانی کم به سادگی منجر به افشا می‌شوند. بنابراین سازمان آماری فایل داده‌های خرد را باید به گونه‌ای منتشر کند که تعداد این رکوردها در آن کم باشد.

امنیت داده‌ها در حالتی که پاسخ دهندگان نمونه‌ای از جامعه هستند، بیشتر است. به عنوان مثال اگر فایل داده‌های جدول ۱.۲ مربوط به نمونه باشد، متخلف از زیاد بودن درآمد A_7 مطمئن نبوده و فقط زمانی می‌تواند اطمینان حاصل نماید، که بداند فراوانی رکورد هفتم در جامعه نیز یک است. شکل دیگر افشا در این حالت به این صورت می‌تواند باشد که فرد دیگری که جزء افراد نمونه نبوده و مقدار متغیرهای شناسایی او برابر مقدار متغیرهای شناسایی A_7 است، به سادگی از میزان درآمد A_7 مطلع می‌شود. به طور کلی در حالتی که فایل داده‌های منتشر شده مربوط به نمونه است، رکوردهای با فراوانی کم در صورتی که فراوانی آنها در جامعه نیز کم باشد، می‌توانند منجر به افشا شوند.

پس ریسک افشارا می‌توان به سه مولفه ریسک افشای حاصل از رکوردهای با فراوانی یک، رکوردهای با فراوانی دو و رکوردهای با فراوانی سه تجزیه کرد. سازمان آماری می‌تواند از اندازه همه این مولفه‌ها یا اندازه تعدادی از آنها برای ارزیابی ریسک افشا استفاده کند. اگر اندازه همه این مولفه‌ها کمتر از سطوح قابل پذیرش سازمان آماری باشد، داده‌ها می‌توانند منتشر شوند. در غیر اینصورت قبل از انتشار باید برای کاهش ریسک افشای آنها اقدام شود.

۱.۱.۲ ارزیابی ریسک افشای داده‌های خرد

در این بخش ابتدا نمادهای مورد نیاز تعیین و سپس معیارهای اندازه‌گیری ریسک افشا در سه وضعیت انتشار اول، دوم و سوم معرفی می‌شوند.

فرض کنید سازمان آماری نمونه S به حجم n را از جامعه $\mathcal{U}$ به حجم N با یکی از روش‌های مرسوم نمونه‌گیری انتخاب و اطلاعات آن را جمع‌آوری کرده و می‌خواهد در مورد نحوه انتشار داده‌ها

تصمیم‌گیری کند. نمادهای مورد نیاز در ادامه تعریف شده‌اند.

U: متغیر راهنمای جامعه

S: متغیر راهنمای نمونه

J: تعداد سطوح متغیر راهنما

j: سطحی دلخواه از متغیر راهنما

U_i: مقدار متغیر راهنمای واحد iام جامعه $\quad i = ۱, ۲, ..., N$

S_i: مقدار متغیر راهنمای واحد iام نمونه $\quad i = ۱, ۲, ..., n$

F_j: تعداد واحدهای جامعه که U متناظر آنها برابر j است.

f_j: تعداد واحدهای نمونه که S متناظر آنها برابر j است.

N_r: تعداد سطوحی از U که فراوانی اعضای جامعه در آنها r است.

n_r: تعداد سطوحی از S که فراوانی اعضای نمونه در آنها r است.

با استفاده از نمادهای بالا و تابع نشانگر $I(.)$، F_j و f_j به ترتیب برابر $\sum_{i\in\mathcal{U}} I(U_i = j)$ و $\sum_{i\in\mathcal{S}} I(S_i = j)$ می‌باشند. همچنین N_r برابر $\sum_{j\in J} I(F_j = r)$ و n_r برابر $\sum_{j\in J} I(f_j = r)$ می‌باشد. به عنوان مثال، $N_۱$ تعداد یکتاهای جامعه و $N_۲$ تعداد زوج‌هایی است که مقدار متغیرهای شناسایی هر دو فرد، یکسان است. با توجه به نمادهای معرفی شده، فرد iام یکتا در جامعه است هرگاه $U_i = j$ و $F_j = ۱$ باشد. به طور مشابه فرد iام یکتا در نمونه است هرگاه $S_i = j$ و $f_j = ۱$ باشد.

بدیهی است که در وضعیت انتشار اول فقط از نمادهای مربوط به جامعه استفاده می‌شود، زیرا در این وضعیت سازمان آماری اطلاعات مربوط به افراد جامعه را جمع‌آوری کرده و می‌خواهد همه آنها را منتشر کند. بنابراین در این وضعیت رکوردهایی می‌توانند منجر به افشا شوند که فراوانی آنها در جامعه کم باشد. پس $N_۱$، $N_۲$ و $N_۳$، یعنی تعداد سطوحی از U که فراوانی اعضای جامعه در آنها به ترتیب یک، دو و سه است، می‌توانند در ارزیابی ریسک افشا موثر واقع شوند. اینکه سازمان آماری از کدامیک از این کمیت‌ها استفاده کند، تابع ضابطه خاصی نبوده و بستگی به نظر سازمان آماری دارد. به عنوان مثال ممکن است سازمان آماری سه سطح قابل پذیرش $\alpha_۱$، $\alpha_۲$ و $\alpha_۳$ را در نظر گرفته و هرگاه شرط‌های

$$\frac{N_i}{J} < \alpha_i \qquad i = ۱, ۲, ۳$$

همزمان برقرار باشند داده‌ها را منتشر کند و در غیر اینصورت رکوردهای ناایمن را تشخیص داده و تکنیک‌های SDC را اعمال کند. همچنین ممکن است سازمان آماری فقط یکتاهای جامعه را در نظر