افق‌های جدید آمار:
تحلیل داده‌های نمادین

محمدرضا مشکانی [۱]

چکیده

پیشرفت شگفت‌انگیز در محاسبات آماری با رایانه‌های سریع، موجب تحولاتی ژرف در اکثر علوم و از آن جمله در آمار شده است. تا چند دهه پیش، تحلیل مجموعه داده‌هایی که شامل بردارها یا ماتریس‌های با ابعاد زیاد بودند، مشکلی عمده به شمار می‌آمد، ولی امروزه چنین محاسباتی امری عادی به شمار می‌روند، اما تحلیل چنین داده‌های حجیم با ابزارها و روشهای سابق تا حدودی به معنای نادیده انگاشتن اطلاعات جالب توجهی است که در آنها نهفته‌اند. بنابراین، از چندی قبل، ضرورت ابداع روشهایی جدید که از این پیشرفت‌های محاسباتی بهره برگیرند، احساس شده است.

برای بهره‌گیری از این امکانات جدید، باید نگرش ما به مجموعه داده‌ها تغییر یابد و مفاهیمی جدید تعریف شوند تا بر اساس آنها روشهای جدید بنا نهاده شوند. با تعریف داده‌های نمادین که می‌توانند رده‌هایی از بردارها یا ماتریس‌هایی اندازه‌گیری شده از هر فردی آماری باشند، به تعمیم جالبی می‌توان دست یافت.

در این مقاله، ابتدا ضرورت استفاده از داده‌هایی نمادین و تحلیل آنها را مورد بحث قرار می‌دهیم. سپس با معرفی مفاهیم و اصطلاحات مورد استفاده در این رشته به اجمال و همراه با مثال روشهایی از تحلیل داده‌های نمادین را عرضه می‌کنیم.

واژه‌های کلیدی : داده‌های نمادبن، متغیرهای نمادین، داده کاوی، رده‌بندی، خوشه‌بندی، متغیرهای بازه-مقدار، چند مقداره.

۱. مقدمه

با اختراع رایانه‌ها، انواع پایگاههای داده‌ای بزرگ و بسیار بزرگ امری عادی شده است. چیزی که عادی نیست، نحوه تحلیل داده‌های موجود در این پایگاهها و یا چگونگی استخراج اطلاعات مفید از درون این پایگاههای گسترده است، اما روشن است که در مواردی هم که نظریه و روش‌شناسی موجود به نظر می‌رسد که قابل کاربست باشد، استفاده روزمره از این گونه روشهای آماری اغلب مناسب نیست. دلیل‌های زیادی وجود دارند. به طور کلی، یکی از دلایل عمده آن است که این گونه مجموعه‌های داده‌ها واقعاً نمونه‌ای از یک جامعه نیستند. زیرا در اغلب موارد، داده‌ها یک «کل» را تشکیل می‌دهند. مثل سوابق حسابهای کارتهای اعتباری همه کسانی که از کارت استفاده می‌کنند. سؤال مربوط دیگر آن است که آیا داده‌های یک نقطه زمانی مشخص را می‌توان از همان جامعه‌ای دانست که در زمان دیگر گردآوری می‌شوند. مثلاً، آیا داده‌های کارتهای اعتباری این هفته، از همان جامعه داده‌های هفته «بعد»اند؟ یا اگر دو مجموعه داده‌ها در هم ادغام شوند، باز هم متعلق به همان جامعه اولیه‌اند؟ دلیل عمده دیگر که روشهای شناخته شده را نمی‌توان به کار برد، صرفاً اندازه مجموعه داده‌هاست. برای مثال، فرض کنید n مشاهده در مورد p متغیر از هر فرد

[۱] دانشکده علوم ریاضی، گروه آمار، دانشگاه شهید بهشتی

آماری داریم. در اغلب موارد، ماتریس داده‌های X که از بعد n×p است، به صورت $(X'X)^{-1}$ که از بعد p×p است، در محاسبه وارد می‌شود. اگر n از بعد صدها هزار و p از بعد صد یا بیشتر باشد، در عین حال که از حیث نظری، چنین کاری ممکن است، از حیث عملی کاری بسیار سنگین خواهد. حتی وقتی توانایی رایانه‌ها (مثلاً برای وارون کردن ماتریس‌ها در مدت زمانی معقول) افزایش یابند، در نتیجه، مجموعه داده‌های بسیار بسیار بزرگتر تولید خواهند شد و مساله تقریباً به همین شکل باقی خواهد ماند. بنابراین، در حالی که روشهای سنتی در مورد مجموعه‌هایی کوچک از داده‌ها که در گذشته وجه غالب بوده‌اند، به خوبی کار کرده است. اکنون انتظار از آمارشناسان تحلیل داده‌ها آن است که شیوه‌هایی را ابداع کنند که در مورد مجموعه‌های بزرگ داده‌های جدید خوب کار کنند. این شیوه‌ها، باید چنان باشند که ما را از اطلاعات زیر ساختی (یا دانش) نهفته در داده‌ها آگاه سازند.

یک رویکرد آن است که مجموعه‌های بزرگ داده‌ها را به طریقی خلاصه کنیم که مجموعه داده‌های خلاصه اندازه‌ای کوچک داشته باشد تا بتوانیم از عهده‌اش برآییم. بدین ترتیب، در مثال کارت اعتباری به عوض صورت زیر عملیات بانکی برای هر فرد (یا هر کارت) در طی زمان، خلاصه‌ای از این عملیات برای هر کارت (یا برای هر واحد زمان مثل هفته، ماه) تهیه کنیم. یک چنین خلاصه‌ای می‌تواند گروه‌بندی معاملات بر حسب میزان پول خرج شده (مثلاً ۱۰۰-۱۰ هزار ریال، غیره)، یا بر حسب خرید نوع کالا (مثلاً غذا، لباس، بنزین، غیره)، یا بر حسب پول خرج شده و نوع کالا (مثل «غذا-۲۰-۱۵»، «لباس ۳۰-۱۰»، ...) یا هر نوع خلاصه‌سازی دیگر باشد.

در هر یک از این مثالها، داده‌ها دیگر مقداری تک مثل داده‌های سنتی از قبیل: ۱٦٤۰۰ ریال کل هزینه کارت اعتباری، یا ۵۲ مورد معامله، برای هر نفر در هر واحد زمان ندارند. این گونه داده‌ها، مثالهایی از داده‌های نمادین‌اند. بدیهی است که کاربست روشی ساده مثل گرفتن میانگین حسابی در مورد آنها چندان مناسبت ندارد. برای چنین داده‌های نمادین، باید از روشهای تحلیل داده‌های نمادین استفاده کرد.

در حالی که خلاصه‌سازی مجموعه‌هایی بزرگ از داده‌ها می‌تواند مجموعه‌هایی کوچک از داده‌ها مشتمل بر داده‌های نمادین را تولید کند، داده‌های نمادین اصالتاً در مورد هر اندازه از مجموعه داده‌ها، خواه کوچک یا بزرگ متمایزند. مثلاً، نامعقول نیست که داده‌هایی داشته باشیم که مشتمل بر متغیرهایی‌اند که در دامنه‌ای ثبت شده‌اند، مانند نبض یا تپش قلب (مثل «۷۲ و ٦۰»)، فشار خون سیستولیک («۱۳۰ و ۱۲۰») برای مثلاً n=۱۲۰ بیمار (یا مثلاً برای $n=۱/۲\times۱۰^{٦}$ بیمار). یا، ممکن است n=۲۰ دانشجو داشته باشیم که به وسیله بافتنگار یا توزیع نمره‌هایشان برای چند متغیر ریاضی، فیزیک، آمار و غیره مشخص شده باشند. ممکن است پرندگان بر حسب رنگشان مشخص شوند. مثلاً پرنده ۱ = (سیاه)، پرنده ۲ = (زرد و آبی)، پرنده ۳ = (نیمی زرد و نیمی قرمز) ... یعنی، متغیر «رنگ» تنها یک رنگ ممکن را برای هر پرنده اختیار نمی‌کند، بلکه فهرستی از رنگ‌ها را اختیار می‌کند، یا حتی فهرستی را با نسبت‌های متناظر از هر رنگ را برای هر پرنده اختیار می‌کند. از سوی دیگر، نقطه داده‌ای {سیاه} ممکن است گردایه‌ای از پرندگان را که همگی سیاه‌اند، نشان دهد و نقطه {قرمز (۰/٦) و زرد (۰/٤)} گردایه‌ای از پرندگانی خواهد بود که ٤۰٪ زرد و ٦۰٪ قرمزند یا گردایه‌ای که از آنها ٤۰٪ کلاً زرد و ٦۰٪ کلاً قرمزند. مثالهایی بی‌شمار را می‌توان ذکر کرد. در راستایی دیگر، ممکن است پرنده خاصی را مدنظر نداشته باشیم، بلکه به مفهوم پرنده‌ای سیاه، یا پرنده‌ای زرد و قرمز علاقه‌مند باشیم. به همین ترتیب، می‌توانیم یک شرکت مهندسی را با پایگاه علمی آنکه مشتمل بر تجارب کارمندانش است، مشخص کنیم. این گونه تجارب بیشتر با مفاهیم توصیف می‌شوند تا با داده‌های استاندارد و بدین طریق، مثالهایی از داده‌های نمادین‌اند. در مورد مجموعه‌های کوچک از داده‌های نمادین سؤال این است که تحلیل چگونه صورت می‌گیرد. در مورد مجموعه‌های بزرگ، سؤال نخست این است که رویکرد انتخابی برای خلاصه‌سازی چیست تا لزوماً به مجموعه‌ای کوچکتر خلاصه شوند. برخی روشهای خلاصه‌سازی الزاماً شامل داده‌های نمادین و تحلیل نمادین در قالبی از انواع قالبها هستند، در حالی که در برخی دیگر لازم نیست که آن گونه عمل شود.

در این مقاله، سعی بر آن است که مفاهیم و روشهای گوناگون ابداع شده تحت عنوان تحلیل داده‌های نمادین یا سایر عنوان‌ها را مرور کنیم. در حقیقت، این روشها تاکنون محدود به ابداع روش شناسیهایی بوده‌اند که داده‌ها را در قالبهایی معقول و قابل اداره سازمان دهند، کاری تا حدودی مشابه با تشکیل جدول‌های توزیع فراوانی و بافتنگارها و دیگر آماره‌های خلاصه پایه‌ای در دوره قبل از سال ۱۹۰۰ است. همین اقدامات خلاصه‌سازی بود که بنیاد آمار استنباطی را در قرن بیستم میلادی پایه‌گذاری کرد. در زیر مروری کوتاه بر روشهای آماری نمادین موجود می‌آوریم. موضوعی که سریعاً آشکار می‌شود این است

که تاکنون روش‌شناسی آماری اندکی برای صورت‌های مختلف داده‌های نمادین ابداع شده‌اند. اما به تعبیری دیگر، تحلیل اکتشافی داده‌ها که از سوی توکی و همکارانش (توکی، ۱۹۹۷) ابداع شدند، نویدی است از آنچه هم اکنون در حال بسط و گسترش است.

تحلیل اکتشافی داده‌ها، داده‌کاوی، کشف دانش در پایگاههای داده‌ها، آماره‌ها، داده‌های نمادین و مانند آنها اصطلاحاتی هستند که امروزه به فراوانی به چشم می‌خورند. تحلیل داده‌های نمادین اندیشه‌های به کار رفته در تحلیل اکتشافی سنتی را به داده‌های کلی‌تر و پیچیده‌تر گسترش می‌دهد.

سایبس (۱۹۹۸) سعی دارد که داده‌کاوی را به عنوان گامی که طی آن الگوهای موجود در داده‌ها به طور خودکار (مثلاً با استفاده از الگوریتمهای محاسباتی) کشف می‌شوند، معرفی کند. در حالی که کشف دانش نه تنها مرحله داده‌کاوی، بلکه گامهای پیش‌فرآوری (از قبیل پاک‌سازی داده‌ها) و گامهای پس‌فرآوری (نظیر تفسیر نتایج) را نیز در بردارد. روشن است که همین مرحله پس‌فرآوری است که نقش سنتی آمارشناسان بوده است. الدر و پریگیبون (۱۹۹۶) دورنمایی از کشف دانش در پایگاههای داده‌ها را عرضه می‌کنند. هند و همکاران (۲۰۰۰) داده‌کاوی را به عنوان «تحلیل ثانی پایگاههای داده‌ای بزرگ با هدف یافتن روابط غیر منتظره که مورد علاقه صاحبان پایگاهها هستند و برایشان ارزشمندند» تعریف می‌کنند.

غالباً اندازه پایگاه داده‌ای چنان بزرگ است که تحلیل‌های اکتشافی داده‌ها به روش کلاسیک برای آنها نارسا هستند. چون برخی مسائل موجود در داده‌کاوی و کشف دانش در پایگاههای داده‌ای به طور طبیعی به صورت‌های داده‌های نمادین می‌انجامند، تحلیل‌های داده‌های نمادین در اینجا نیز نقشی برای ایفا دارند. بدین ترتیب، همکاری تیمهای علمی بین رشته‌ای در هنگام کار با مجموعه‌های بزرگ داده‌ها (مثل متخصصان رایانه و آمارشناسان) به امری اساسی تبدیل می‌شود.

هدف این مقاله، آن است که مفاهیم داده‌های نمادین و شیوه‌های تحلیل آنها را که در حال حاضر در آثار مکتوب موجودند، مرور کند. بنابراین، داده‌های نمادین که گاهی به اصطلاح «اتم‌های دانش» خوانده می‌شوند، در بخش ۲ تعریف شده با داده‌های کلاسیک مقایسه می‌شوند. ساختن رده‌هایی از اشیاء، نمادین که وقتی اندازه داده‌ها برای تحلیل‌های کلاسیک بسیار بزرگ باشد، یا هنگامی که به عوض داده‌های استاندارد دانش به عنوان درونداد داده شود، از مقدمات ضروری برای تحلیل‌های آماری است، در بخش ۳ بررسی می‌شود. در بخش ۵، به اجمال روشهای موجود برای تحلیل داده‌های نمادین را ارائه می‌کنیم. آنچه که از این بحث‌ها آشکار می‌شود آن است که داده‌های نمادین به طور فزاینده‌ای شایع می‌شوند و بنابراین نیازی مبرم به ابداع روش شناسی‌های آماری برای این گونه داده‌ها وجود دارد. همچنین، ملاحظه خواهد شد که در حال حاضر، روشهایی معدود وجود دارند و حتی برای آنهایی که وجود دارند، نیاز به تاسیس مبانی آماری-ریاضی مطرح است.

۲. مجموعه داده‌های نمادین

یک مجموعه داده‌ای از همان ابتدا ممکن است به صورت مجموعه‌ای از داده‌های نمادین تشکیل شود. به شق دیگر، ممکن است از ابتدا به صورت داده‌های کلاسیک فراهم آید، اما بعداً به صورت داده‌های نمادین سازمان داده شود تا به قالبی بیشتر قابل اداره درآید، به ویژه هنگامی که در ابتدا دارای اندازه‌ای بسیار بزرگ باشد.

در این بخش، مثالهایی از هر دو نوع داده‌های کلاسیک و داده‌های نمادین را عرضه می‌کنیم.

همچنین، نوعی نمادگذاری را معرفی می‌کنیم که مجموعه‌های داده‌های نمادین آماده برای تحلیل را توصیف می‌کند. این فرآیند وضعیت‌هایی را دربرمی‌گیرد که مثلاً دو یا چند مجموعه از داده‌ها در هم ادغام می‌شوند یا گاهی که بر ویژگی‌های مختلفی از داده‌ها باید تأکید شود.

فرض کنید مجموعه‌ای از داده‌ها را داریم که مشتمل بر سوابق پزشکی افراد در کشور است. فرض کنید برای هر فرد، سابقه‌ای از متغیرهای مکان جغرافیایی، مانند ناحیه (شمال، شمال شرق، جنوب، ...)، شهر (بابلسر، سبزوار، کرمان، ...)، شهر یا روستا (بلی، خیر) و غیره وجود دارد. متغیرهای جمعیت شناختی از قبیل جنس، وضع تأهل، سن، اطلاعات درباره پدر و مادر (هنوز زنده یا نه)، برادر و خواهر، تعداد فرزندان، کارفرما، عرضه کننده خدمات تندرستی و غیره نیز در کار خواهند بود. متغیرهای پایه‌ای پزشکی می‌توانند شامل وزن، تپش نبض، فشار خون و غیره باشند. دیگر متغیرهای تندرستی (که برای آنها فهرست متغیرهای ممکن بی‌پایان است) شامل بروز ناراحتی‌ها و بیماریهای معین خواهند بود. همچنین، برای شیوع یا تشخیص یک

بیماری معین معالجات و سایر متغیرهای مربوط به آن بیماری ثبت می‌شوند. برای مثال، مجموعه‌ای نوعی از این گونه داده‌ها مثل جدول ۱ خواهد بود.

گیریم p تعداد متغیرهای مربوط به فرد i، به ازای $i \in \Omega = \{۱,...,n\}$ باشد، روشن است که p و n می‌توانند عددهایی بزرگ یا حتی فوق‌العاده بزرگ باشند و گیریم Y_j به ازای $j = ۱,...,p$ نشانگر متغیر jام باشد. گیریم $Y_j = x_{ij}$ مقدار خاصی باشد که Y_j برای فرد iام در آرایش کلاسیک اختیار می‌کند و به عنوان ماتریس n×p شامل کل مجموعه داده‌ها بنویسیم ($X = (x_{ij}$ گیریم قلمرو Y_j عبارت باشد از y_j به طوری که $X = (Y_۱,...,Y_p)$ مقادیری را در فضای $\mathrm{X} = \underset{j=۱}{\overset{p}{\times}} y_j$ اختیار کند.

چون بود یا نبود مقدارهای گمشده در حال حاضر اهمیتی ندارد، گیریم همه مقادیر وجود دارند؛ اگر چه در مورد مجموعه‌های بزرگ داده‌ها محتمل‌ترین وضع آن است که داده گمشده داشته باشیم.

متغیرها می‌توانند کمی باشند، مثل سن با $\mathrm{Y}_{age} = \{x \geq ۰\} = \mathrm{Y}_+$ به عنوان متغیر تصادفی پیوسته، یا به صورت $\mathrm{Y}_{age} = \{۰,۱,۲,...\} = \eta_۰$ به عنوان متغیری تصادفی گسسته. متغیرها ممکن است رسته‌ای باشند، مثل شهر با $\mathrm{Y}_{city} = \{$... اصفهان، تهران$\}$ یا رمزیده باشند، به صورت $\mathrm{Y} = \{۱,۲,...\}$. متغیرهای بیماری ممکن است به صورت رسته‌هایی (رمزیده یا نارمزیده) از یک متغیر با قلمرو $\{$...، کبد، سرطان، قلب$\} = \mathrm{Y}$ یا به طور محتمل‌تر به صورت متغیری نشانگر، مثل سرطان =Y با قلمرو $\{$بلی، خیر$\} = \mathrm{Y}$ یا $\{۱$و$۰\}$ یا با سطوح رمزیده دیگر که نشانگر مراحل بیماری است، ثبت شوند.

به همین ترتیب، برای ثبت انواع بسیار زیاد سرطان‌های ممکن، هر نوع سرطان ممکن است با متغیری مثل Y نمایش داده شود، یا با رسته‌ای از متغیر سرطان نمایش داده شود.

ماهیت دقیق توصیف متغیرها چندان مهم نیست. آنچه که اساسی است این است که در آرایش کلاسیک برای هر x_{ij} موجود در X، دقیقاً تنها یک مقدار محقق ممکن وجود دارد. یعنی، مثلاً مشخصات یک فرد عبارتند از $\mathrm{Y}_{age}=۲۴$، یا کاشان=Y_{city}، بلی=Y_{cancer}، $\mathrm{Y}_{pulse}=۲۴$ و مانند آن. بدین ترتیب یک نقطه داده‌ای کلاسیک تک نقطه‌ای است در فضای p بعدی X.

بر عکس، یک نقطه داده‌ای نمادین ممکن است یک ابر مکعب در فضای p بعدی با حاصلضرب دکارتی توزیعها باشد.

درایه‌های موجود در مجموعه داده‌های نمادین (که با ξ_{ij} نمایانده می‌شوند) مقید به مقادیر مشخص تک نیستند. بدین ترتیب، سن را می‌شود به صورتی که در یک بازه واقع باشد، ثبت کرد. مثل، [۱۰و۰]، [۲۰و۱۰] و ... این وضع وقتی می‌تواند رخ دهد که نقطه داده‌ای نمایشگر سن یک خانواده، یا گروهی از افراد باشد که سن آنها به طور دستجمعی در بازه‌ای قرار دارد (مانند زن و شوهری مسن [۶۰،۷۰]) یا داده‌ها ممکن است مربوط به یک فرد تنها باشد که سن دقیقش معلوم نیست، جز آنکه می‌دانیم در یک بازه سنی قرار دارد، یا سن او در طی زمان همراه با اجرای آزمایشی که داده‌ها را ایجاد کرده است، تغییر کرده است، یا ترکیبی از این حالتها و انواع مختلف آنها باشد، که داده‌های با دامنه بازه‌ای را به وجود می‌آورند.

در راستایی متفاوت، ممکن است اندازه‌گیری مشخصه‌ای دقیقاً به صورت مقداری تک میسر نباشد، نتوان اندازه گرفت که نبض ۶۴ است، بلکه بتوان اندازه گرفت که مقدار آن (۱±۶۴) یا به طور کلی اندازه متغیری (x±δ) است.

شخصی ممکن است x≤۲ یا x>۲ خواهر و برادر (یا فرزند یا ...) داشته باشد. متغیر فشار خون ممکن است به صورت مقادیر [بالا، پایین] آن، مثلاً $\xi_{ij} = [۷۸,۱۲۰]$ ثبت شود. این متغیرها، متغیرهای نمادین بازه-مقدارند.

نوعی متفاوت از متغیر می‌تواند متغیری مربوط به سرطان باشد که ممکن است قلمرو $\{$...،کبد، سینه، استخوان، ریه$\} = \mathrm{Y}$ را داشته باشد که فهرستی از همه سرطان‌های ممکن است و شخص معینی مثلاً دارای مقادیر خاص $\{$کبد، ریه$\} = \xi_{ij}$ است. در مثالی دیگر، فرض کنید متغیر Y_j نمایشگر نوع خودرویی باشد که خانواده‌ای داراست با $\mathrm{Y}_j = \{$...،تویوتا، بنز، پژو، پیکان$\}$. خانواده خاص i ممکن است مقدار $\xi_{ij} = \{$بنز، پیکان$\}$ را دارا باشد. چنین متغیرهایی، متغیرهای چند مقداره نامیده می‌شود.

نوع سوم متغیر نمادین متغیر مدی است. متغیرهای مدی متغیرهای چند وضعی با فراوانی، احتمال یا وزن منسوب به هر یک از مقادیر خاص در داده‌هااند. یعنی، متغیر مدی Y نگاشتی است مانند

$$Y(i) = \{U(i), \pi_i\}, \quad i \in \Omega$$

که در آن π_i اندازه‌ای نامنفی یا توزیعی روی قلمرو Y متشکل از مقادیر مشاهده‌ای ممکن و $U(i) \subseteq Y$ تکیه گاه π_i است. مثلاً اگر سه نفر از خواهران و برادران شخصی مرض قند داشته باشند و یک نفر نداشته باشد، متغیر توصیف گر استعداد مرض قند می‌تواند مقدار خاص $\xi_{ij} = \{\frac{3}{4}$ مرض قند، $\frac{1}{4}$ سالم$\}$ را اختیار کند. به طور کلی‌تر، ξ_{ij} می‌تواند یک بافتنگار، توزیع تجربی، توزیع احتمال، یک مدل یا مانند آنها باشد. در واقع شرایتزر (۱۹۸٤) نظر داد که «توزیعها عددهای آینده‌اند». در حالی که در این مثال، وزن‌های $(\frac{1}{4}, \frac{3}{4})$ ممکن است معرف فراوانی‌های نسبی باشند، انواع دیگر وزن‌ها از قبیل «گنجایش‌ها»، «باورمندی‌ها»، «بایستگی‌ها»، «امکانات» و غیره می‌توانند به کار گرفته شوند. در اینجا «گنجایش» را به تعبیر شوکه (۱۹٥٤) به عنوان احتمال اینکه دست کم یک فرد در رده مذکور دارای مقدار Y معین (مثلاً مرض قند) است، تعریف می‌کنیم و «باورمندی» به تعبیر شیفر (۱۹۷٦) تعریف می‌شود به احتمال اینکه هر فرد موجود در رده دارای آن مشخصه باشد (دیدی، ۱۹۹٥).

پس به طور کلی، بر خلاف داده‌های کلاسیک که برای آنها هر نقطه داده‌ای مشتمل بر تک مقدار (رسته‌ای یا کمی) است، داده‌های نمادین می‌توانند تغییراتی درونی را در برداشته باشند و می‌توانند ساختار یافته باشند. وجود همین تغییرات درونی است که نیاز به فنون جدید برای تحلیل را لازم می‌نماید. این فنون به طور کلی با فنون مربوط به داده‌های کلاسیک تفاوت خواهد داشت.

از نظر نمادگذاری، مجموعه‌ای پایه‌ای از اشیاء را داریم که اعضاء یا هستی‌هایی هستند، $E = \{1,\ldots,N\}$ مجموعه اشیاء خوانده می‌شود. این مجموعه اشیاء می‌تواند معرف کیهانی از افراد $E=\Omega$ (مثل بالا) باشد که در آن حالت $N=n$ یا اگر داشته باشیم $N \leq n$، هر یک مجموعه از اشیای مذکور زیر مجموعه‌ای از Ω است. همچنین، چنانکه در داده کاوی یا تحلیل‌های نمادین کرراً رخ می‌دهد، اشیاء u در E رده‌های $C_1,\ldots,C_m$ از افراد موجود در Ω هستند، با $E = \{C_1,\ldots,C_m\}$ و $N=m$ بدین ترتیب، مثلاً رده C_1 ممکن مشتمل بر همه افرد موجود در Ω باشد که سرطان داشته بودند. هر شیئی $u \in E$ با p متغیر نمادین Y_j، $j=1,\ldots,p$ با قلمرو $\mathcal{Y}_j$ تعریف می‌شود که Y_j نگاشتی از مجموعه اشیاء E به برد $\mathcal{Y}_j$ است که به نوع متغیر Y_j بستگی دارد. بدین ترتیب، اگر Y_j متغیری کمی کلاسیک باشد، قلمرو B_j زیر مجموعه‌ای از خط حقیقی R است، یعنی $B_j \subseteq R$، اگر Y_j متغیری بازه‌ای باشد، $B_j = \{[\alpha, \beta], -\infty \prec \alpha, \beta \prec \infty\}$؛ اگر Y_j رسته‌ای (نامی، ترتیبی، زیر مجموعه‌ای از یک قلمرو متناهی $\mathcal{Y}_j$) باشد، $B_j=\{B \mid B \subseteq \{$مثلاً فهرست همه سرطان‌ها$\}\}$ و اگر Y_j متغیری مدی باشد، $B_j=M(\mathcal{Y}_j)$ که در آن $M(\mathcal{Y})$ خانواده‌ای از همه اندازه‌های نامنفی روی $\mathcal{Y}$ است. در این صورت، داده نمادین برای مجموعه اشیاء E با ماتریس $X=(\xi_{uj})$ از مرتبه $N \times p$ نمایش داده می‌شود که در آن $\xi_{uj}=Y_j(u) \in B_j$ مقدار نمادین مشاهده شده برای متغیر Y_j، $j = 1,\ldots,p$ در مورد شیئی $u \in E$ است. سطر x'_u از X توصیف نمادین شیئی u خوانده می‌شود. بدین ترتیب برای داده‌های جدول ۲ سطر اول،

x'_1 =([۲۰و۳۰]و[۷۹و۱۲۰]، بوشهر، {سرطان مغز}. {مرد}، [۷۰و۸۰])

نمایشگر مردی است در سنین ۲۰ سالگی که سرطان مغز دارد، با فشار خون ۱۲۰/۷۹، وزنی بین ۸۰-۷۰ کیلوگرم و در بوشهر می‌زید.

شیئی u مربوط به این x'_u می‌تواند مرد خاصی باشد که طی دوره ده ساله‌ای تحت نظارت بوده که در این مدت وزنش بین ۷۰ و ۸۰ کیلوگرم تغییر کرده، یا u می‌توانست گردآیه‌ای از افراد باشد که سنشان از ۲۰ تا ۳۰ سال متغیر بوده و دارای مشخصاتی بوده‌اند که با x'_u توصیف شده است. داده‌های x'_4 در جدول ۲ ممکن است معرف همان فردی باشد که به عنوان فرد i=چهار از جدول ۱ معرفی شده، اما تنها این امر معلوم است که او (با احتمال p) دارای سرطان سینه یا (با احتمال 1-p) دارای سرطان ریه است، ولی معلوم نیست کدام یک. از سوی دیگر، داده‌های فوق می‌توانند معرف مجموعه زنان ٤۷ ساله از بابلسر باشند که از آنها به نسبت p دارای سرطان ریه و (1-p) دارای سرطان سینه‌اند، یا می‌تواند نمایشگر افرادی باشد که دارای هر دو سرطان ریه و سینه‌اند و مانند آن (در مرحله‌ای از کار، اینکه متغیر (در اینجا نوع سرطان) از نوع رسته‌ای، فهرستی، مدی یا هرچه هست، باید به صراحت تعریف شود).

موضوع دیگر، به متغیرهای وابسته مربوط می‌شود که برای داده‌های نمادین به معنای وابستگی منطقی، وابستگی سلسله مراتبی، یا وابستگی تصادفی است. وابستگی منطقی همان گونه که از معنای واژه پیداست به صورت اگر-آنگاه است که در مثال فوق اگر [سن≤۱۰] آنگاه [تعداد فرزندان = ۰]. وابستگی سلسله مراتبی وقتی رخ می‌دهد که

برآمد یک متغیر (مثلاً معالجه سرطان = Y_2) با {...، پرتو درمانی، شیمی درمانی} = Y_2] به برآمد واقعی تحقق یافته متغیری دیگر (مثلاً سرطان دارد = Y_1 با {بلی، خیر} = Y_1) وابسته است. اگر Y_1 دارای مقدار {بلی} باشد، آنگاه {مثلاً شیمی درمانی} = Y_2. در حالی که اگر {خیر} = Y_1 آنگاه روشن است که Y_2 کاربرد ندارد[به منظور توضیح مطلب فرض می‌کنیم که معالجه شیمی درمانی به دلیلی دیگر صورت نگیرد]. در این موارد، متغیر کاربرد ندارد Z را با قلمرو {ک ن}=Z تعریف می‌کنیم. این گونه متغیرها را متغیرهای مادر (Y_1)-دختر(Y_2) نیز گویند.

۳. رده‌ها و طرز تشکیل آنها؛ اشیاء نمادین

مجموعه داده‌های نمادین ممکن است از آغاز کار پیشاپیش از حیث اندازه به قدر کافی کوچک باشند که به منظور انجام هر نوع تحلیل آماری نمادین به طور مستقیم، نیاز به پیش‌فرآوری نباشد. مثالی از این نوع داده‌ها جدول ۳ است که برای تحلیل مولفه‌های اصلی نمادین به کار رفته‌اند. اما به طور کلی‌تر و تقریباً به طرزی گریز ناپذیر قبل از آنکه هر نوع تحلیل نمادین از داده‌ها را بتوان انجام داد، به ویژه برای مجموعه‌های بزرگ داده‌ها، نیاز خواهد بود که درجات مختلفی از داده‌آمایی برای سازمان‌دهی اطلاعات در رده‌هایی که مناسب برای سؤال مورد نظر باشند، صورت گیرند.

در برخی موارد، اشیاء موجود در E (یا Ω) پیشاپیش در رده‌هایی جمع شده‌اند، حتی در این موارد هم سؤال‌هایی خاص ممکن است سازمان‌دهی مجدد در رده‌بندی متفاوتی را ضروری سازند، صرفنظر از اینکه مجموعه داده‌ها کوچک است یا بزرگ. مثلاً یک مجموعه از رده‌های $C_1, ..., C_m$ ممکن است افراد را که بر حسب m نوع مختلف از بیماریهای مهم رسته‌بندی شده‌اند، نمایش دهد. در حالی که تحلیل دیگر ممکن است ضروری نماید که ساختار رده‌ها بر حسب شهر، جنس یا بر حسب جنس و سن و غیره باشد. مرحله ابتدایی دیگر وقتی است که داده‌ها در ابتدا برای پایگاههای داده‌های آمار کلاسیک یا علوم رایانه برای هر فرد $i \in \Omega = \{1, ..., n\}$ با n فوق‌العاده بزرگ ثبت شده‌اند. همین طور است وقتی که پایگاههای داده‌ای نمادین بسیار بزرگ باشد. در این صورت، این مرحله از تحلیل داده‌های نمادین متناظر است با انبوهیدن n شیئی در m رده که m بسیار کوچکتر از n است و طوری طراحی شده که قالبهای قابل اداره‌تری را پیش از هر تحلیل آماری به دست دهد. توجه کنید که این ساخت ممکن است از رده‌های حاصل از یک شیوه خوشه‌بندی فرق داشته باشد، اما لزوماً چنین نیست. توجه داشته باشید که این m رده انبوهیده ممکن است معرف m الگوی به دست آمده از یک شیوه داده‌کاوی باشد.

این توصیف ما را به مفهوم شیئی نمادین رهنمون می‌شود که در سلسله‌ای از مقالات دیدی و همکارانش بسط یافته است (مثلاً دیدی، ۱۹۸۷، ۱۹۸۹، ۱۹۹۰، باک و دیدی، ۲۰۰۰ و استفن و همکاران، ۲۰۰۰). این مطلب را ابتدا با چند مثال معرفی می‌کنیم و در پایان بخش تعریفی دقیق ارائه می‌شود.

چند مثال

فرض کنید به مفهوم «شمال خاوری» علاقه‌مندیم. بدین ترتیب، توصیفی مانند d داریم که معرف مقادیر خاص {... و سایر شهرهای شمال خاوری، ...، نیشابور، مشهد} در قلمرو Y_{city} است و رابطه‌ای مانند R (در اینجا ∈) داریم که متغیرهای Y_{city} را با توصیف خاص مورد نظر ربط می‌دهد. این مطلب را به صورت مثلاً

$$a = [Y_{city} \in \{\text{مشهد، نیشابور، ...، سایر شهرهای ش-خ}\}]$$

می‌نویسیم. در این صورت، هر فرد i در $\Omega = \{1, ..., n\}$ یا یک فرد شمال خاوری است یا نیست. یعنی، a نگاشتی است از Ω بر {۰,۱} که برای فردی مانند i که در شمال خاوری زندگی می‌کند $a(i) = 1$ و گرنه $a(i) = 0$ به ازای $i \in \Omega$. پس، اگر فرد i در مشهد زندگی کند (یعنی، مشهد=$Y_{city}(i)$)، آنگاه

$$a(i) = [\text{مشهد} \subseteq \{\text{مشهد، نیشابور، ...، سایر شهرهای ش-خ}\}] = 1$$

مجموعه $i \in \Omega$ که برای آنها $a(i) = 1$، گستره a در Ω نامیده می‌شود. سه گانه $S = (a, R, d)$ یک شیئی نمادین است که در آن R رابطه‌ای بین توصیف $Y(i)$ از متغیر (بی صدای) Y و توصیف d و a نگاشتی از Ω به L است که L به R و d بستگی دارد (در مثال شمال خاوری، $L = \{0, 1\}$). توصیف d می‌تواند توصیف قصدی باشد. مثلاً همان طور که از نامش پیداست، قصد داریم که مجموعه افراد موجود در Ω را که در «شمال خاوری» زندگی می‌کنند، بیابیم. بدین ترتیب، مفهوم «شمال خاوری» تا حدی شبیه به مفهوم جامعه کلاسیک است و گستره موجود در Ω متناظر با نمونه افراد از شمال خاوری در مجموعه داده‌های واقعی است، اما به خاطر بیاورید که همان طور که در بخش ۲

ذکر شده Ω ممکن است پیشاپیش خود «جامعه» باشد یا ممکن است به معنای نمونه‌گیری آمار کلاسیک یک «نمونه» باشد.
اشیاء نمادین به یکی از سه طریق در محدوده عمل تحلیل‌های نمادین داده‌ها نقش ایفا می‌کند. نخست، یک شیئی نمادین ممکن است به کمک قصد آن (مثلاً توصیفش و راهی برای محاسبه گستره‌اش) نمایشگر مفهومی باشد و می‌توان به عنوان درونداد تحلیل داده‌های نمادین به کار گرفته شود. بدین ترتیب، مفهوم «شمال خاوری» را می‌توان با یک شیئی نمادین نمایش داد که قصد آن از طریقی توصیفی سرشت‌نما و راهی برای پیدا کردن گستره‌اش که مجموعه افرادی است که در شمال خاوری زندگی می‌کنند، تعریف می‌شود.

مجموعه‌ای از این گونه ناحیه‌ها و اشیای نمادین همراه با آنها می‌توانند درونداد یک تحلیل داده‌های نمادین را تشکیل دهند. دوم، می‌توان آن را به عنوان بروندادی از یک تحلیل داده‌های نمادین مورد استفاده قرار داد، مثل وقتی که تحلیل خوشه‌ای حاکی از آن است که شمال خاوری متعلق به خوشه‌ای خاص است که خود خوشه را به عنوان یک مفهوم می‌توان تلقی کرد و به وسیله یک شیئی نمادین نمایش داد. وضعیت سوم زمانی که فردی جدید (i') را داریم که دارای توصیف d' است و می‌خواهیم بدانیم که آیا این فرد (i') با شیئی نمادینی که توصیفش d است، تطبیق می‌کند یا خیر. یعنی d و d' را به کمک R مقایسه می‌کنیم تا $[d'Rd] \in L = \{۰,۱\}$ را بدهد که در آن $[d'Rd] = ۱$ به معنای آن است که ارتباطی بین d' و d وجود دارد. این فرد «جدید» ممکن است فردی «قدیم» اما با داده‌های روزآمد شده باشد یا ممکن است فردی جدید باشد که به پایگاه داده‌ها اضافه شده و امکان دارد که او با یکی از رده‌های قبلاً موجود اشیاء نمادین «همسان» باشد یا نباشد (مثلاً آیا باید برای این شخص پوشش بیمه‌ای خاص را فراهم ساخت؟).

در خصوص انبوهیدن داده‌هایمان در تعداد کمی از رده‌ها، اگر قرار باشد که افراد موجود در Ω را بر حسب شهر، یعنی بر حسب مقدار متغیر u_{city} بیانبوهیم، رده‌های مربوطه C_u با $u \in \{۱,...,m\}$ متشکل از آن افرادی در Ω است که در گستره نگاشت متناظر a_u قرار دارند. تحلیل آماری متعاقب می‌تواند یکی از دو راستای وسیع را در پیش گیرد. یا آنکه برای هر رده بر حسب اقتضا، به طور جداگانه داده‌های کلاسیک یا نمادین را برای افراد موجود در C_u به عنوان نمونه‌ای مرکب از n_u مشاهده تحلیل می‌کنیم، یا داده‌های هر رده را خلاصه می‌کنیم تا مجموعه جدیدی از داده‌ها را به یک «مشاهده» برای هر رده به دست بدهد. در این حالت اخیر، صرفنظر از اینکه مقادیر اولیه داده‌هایی کلاسیک یا نمادین باشند، مقادیر مجموعه داده‌ها به احتمال زیاد داده‌های نمادین خواهند بود. مثلاً حتی اگر هر فرد موجود در Ω به صورت دارای سرطان و یا بدون سرطان (بلی، خیر= Y_{cancer})، یعنی به صورت یک مقدار داده‌ای کلاسیک ثبت شده باشد، وقتی این متغیر با رده مربوط به شهر ارتباط داده شود، به صورت {(۰/۹) خیر، (۰/۱) بلی}، یعنی ۱۰٪ دارای سرطان و ۹۰٪ بدون سرطان در خواهند آمد. پس اکنون متغیر Y_{cancer} متغیری مدی است.

به همین نحو، رده‌ای که از گستره یک شیئی نمادین ساخته می‌شود، نوعاً به وسیله یک مجموعه داده‌های نمادین توصیف می‌شود. مثلاً، فرض کنید که علاقه ما متوجه است به «کسانی که در مشهد زندگی می‌کنند» یعنی، $a = [Y_{city} =$ مشهد$]$ و فرض کنید که متغیر Y_{child} تعداد فرزندان هر فرد $i \in \Omega$ است که مقادیر ممکن {۳≤ و۲و۱و۰} ر دارد. فرض کنید مقدار داده‌ای برای هر i مقداری کلاسیک است (تبدیل آنها به یک داده نمادین از قبیل اینکه فرد i دارای ۱ یا ۲ فرزند است.، یعنی $\xi_i = \{۱,۲\}$ به آسانی نتیجه می‌شود). در این صورت، شیئی که معرف همه کسانی است که در مشهد زندگی می‌کنند، اکنون دارای متغیر نمادین Y_{child} خواهد بود با مقدار خاص،

$$Y_{child} = \{(۰, f_۰), (۱, f_۱), (۲, f_۲), (\geq ۳, f_۳)\}$$

که در آن f_i، $i = ۰,۱,۲,\geq ۳$، فراوانی نسبی افراد این رده است که دارای i فرزندند.

حالتی خاص از یک شیئی نمادین عبارت است از یک حکم. حکم‌ها، مخصوصاً هنگام انبوهیدن افراد درون رده‌ها از یک پایگاه داده‌ای آغازین (رابطه‌ای)، دارای اهمیت‌اند.

گیریم توصیف لازم مورد نظر از یک فرد یا از یک مفهوم w را با $Z = (Z_۱,...,Z_p)$ بنمایانیم. در اینجا، Z_j می‌تواند هستی تک مقداره کلاسیک x_j یا هستی نمادین ξ_j باشد. یعنی در حالی که x_j معرف مقدار داده‌ای کلاسیک تحقیق یافته‌ای است و ξ_j معرف مقدار داده‌ای نمادین تحقیق یافته‌ای است، Z_j مقداری است که مشخصاً به دنبال آن هستیم یـ مقدار تصریح شده است. بدین ترتیب، مثلاً فرض کنید که به شـ نمادینی علاقه‌مندیم که معرف کسانی است که در شمال خاوری

زندگی می‌کنند. در این صورت، Z_{city} مجموعه شهرهای شمال خاوری است. این موضوع را با حکم،

(۱) $a = [Y_{city} \in \{$مشهد، نیشابور، ...، سایر شهرهای ش-خ$\}]$

که در آن a نگاشتی از Ω بر $\{0,1\}$ است، به قسمی است که برای فرد یا شیئی w، داریم $a(w) = 1$، اگر $Y_{city}(w) \in \{$مشهد، ...، سایر شهرهای ش-خ$\}$ درست باشد.

به طور کلی، هر حکمی صورت زیر را اختیار می‌کند

(۲) $$a = [Y_{j1} R_{j1} Z_{j1}] \wedge [Y_{j2} R_{j2} Z_{j2}] \wedge \ldots \wedge [Y_{jv} R_{jv} Z_{jv}]$$

به ازای $1 \le j_1, \ldots, j_v \le p$ که در آن «$\wedge$» به معنی «و» ضریب منطقی است و R رابطه تصریح شده بین متغیرهای نمادین Y_j و مقدار توصیفی Z_j است. برای هر فرد $i \in \Omega$ وقتی که حکم برای آن فرد درست (یا غلط) باشد، داریم $a(i) = 1$ (یا ۰). به طور دقیق‌تر بگوییم، هر حکم عبارت از ترکیب عطفی v پیشامد $[Y_K R_K Z_K]$, $k = 1, \ldots, v$ است. مثلا حکم‌های

(۳) $a = [Y_{cancer}=$بلی$]$ و $a = [Y_{age} \ge 60]$

به ترتیب، نمایشگر افراد دارای سرطان و افراد ۶۰ سال به بالا هستند. حکم

(٤) $a = [Y_{cancer}=$بلی$] \wedge [Y_{age} \ge 60]$

نمایشگر همه بیماران سرطانی است که ۶۰ سال به بالا هستند، در حالی که حکم

(۵) $a = [Y_{age} < 20] \wedge [Y_{abe} > 70]$

در جستجوی همه افرادی است که زیر ۲۰ سال و بالای ۷۰ سال‌اند. در هر مورد، با مفهومی مانند «افراد ۶۰ سال به بالا»، «افراد ۶۰ سال به بالای دارای سرطان» و غیره سروکار داریم و a افراد خاص حاضر در Ω را بر فضای $\{0,1\}$ می‌نگارد.

اگر به عوض ثبت متغیر سرطان به صورت متغیر رسته‌ای $\{$بلی، خیر$\}$ به صورت متغیری مانند $\{$ریه، کبد، سینه، ...$\}$ ثبت می‌کردیم، حکم

(٦) $a = [Y_{cancer} \in \{$ریه، کبد$\}]$

رده افرادی را که دارای سرطان کبد، یا ریه یا هر دو هستند، توصیف می‌کرد.

همچنین، یک حکم می‌تواند صورت زیر را اختیار کند.

(۷) $a = [Y_{age} \subseteq [20, 30]] \wedge [Y_{city} \in \{$مشهد$\}]$

یعنی، این حکم کسانی را که در مشهد زندگی می‌کنند و از لحاظ سنی در سنین بیست سالگی هستند، توصیف می‌کند.

رابطه‌های R می‌تواند هر یک از صورت‌های $=$، $\neq$، $\in$، $\le$، $\subseteq$ و غیره را اختیار کند یا می‌توانند رابطه‌ای تطبیقی (مانند مقایسه‌ای از توزیعهای احتمال) یا دنباله‌ای ساختار یافته و غیره باشند. این روابط پیوند بین متغیر نمادین Y و توصیف خاص مورد نظر Z را تشکیل می‌دهند. قلمرو شیئی نمادین را می‌توان به صورت زیر نوشت.

$$D = D_1 \times \ldots \times D_p \subseteq X = \underset{j=1}{\overset{p}{\times}} Y_j$$

که در آن $D_j \subseteq Y_j$. هر p گانه $(D_1 \ldots D_p)$ از مجموعه‌ها یک نظام توصیف خوانده می‌شود و هر زیر مجموعه D یک مجموعه توصیف مشتمل بر بردارهای توصیف $Z = (Z_1, \ldots, Z_p)$ است. در حالی که ترکیب عناصر $Z_j \in Y_j$ و مجموعه‌های $D_j \subseteq Y_j$ مثلا مثل آنچه در (۷) دیده می‌شود، صرفا یک توصیف است. وقتی قیودی در مورد هر یک از متغیرها وجود داشته باشد (مانند مواردی که وابستگی‌های منطقی وجود دارند)، آنگاه فضای D در خود «سوراخی» دارد که متناظر با مقادیری است که با قیود تطبیق می‌کنند. کل همه توصیف‌های D عبارت است از فضای توصیف D. بنابراین، حکم را می‌توان چنین نوشت.

$$a = [Y \in D] \equiv \underset{k=1}{\overset{v}{\wedge}} [Y_{jk} R_{jk} Z_{jk}] = [YRZ]$$

که در آن $R = \underset{k=1}{\overset{v}{\wedge}} R_{jk}$ رابطه حاصلضربی خوانده می‌شود.

دقت کنید که اگر حکمی تلویحا متغیری خاص مانند Y_w را دربرنداشته باشد، قلمرو مربوط به آن متغیر در Y_w بی تغییر باقی می‌ماند. مثلا، اگر داشته باشیم $p = 3$ و شهر $= Y_1$ و سن $= Y_2$ و وزن $= Y_3$ آنگاه حکم $a = [Y_1 \in \{$مشهد، نیشابور$\}]$ دارای قلمرو $D = \{$نیشابور، مشهد$\} \times Y_2 \times Y_3$ است و در جستجوی همه آن کسانی است که (صرفنظر از سن و وزن) فقط در مشهد و نیشابورند.

تعریف‌های رسمی

اکنون مفهوم‌های زیر را به طور رسمی تعریف می‌کنیم.

تعریف: وقتی حکمی درباره شیئی خاص $i \in \Omega$ بیان می‌شود، اگر آن حکم درباره آن شیئی صادق باشد، مقدار راست ($a=1$) را و اگر دروغ باشد ($a=0$) را اختیار می‌کند. می‌نویسیم

$$a(i)=[Y(i)\in D]=1, \quad Y(i)\in D \qquad (۸)$$
$$=0, \quad o.w.$$

تابع $a(i)$ تابع راستی خوانده می‌شود و نمایشگر نگاشتی از Ω بر $\{0,1\}$ است.

مجموعه همه مقادیر $i\in\Omega$ که برای آنها حکم صادق است، گسترش در Ω خوانده می‌شود که با $Ext(a)$ یا Q نمایانده می‌شود،

$$Ext(a)=Q=\{i\in\Omega \mid Y(i)\in D\}=\{i\in\Omega \mid a(i)=1\} \qquad (۹)$$

نوعا، هر رده با شیئی نمادینی که توصیفش می‌کند، شناسایی می‌شود. مثلا حکم (۷) متناظر است با شیئی نمادین «۲۰-۳۰ ساله‌هایی که در مشهد زندگی می‌کنند». گسترش این حکم، $Ext(a)$، رده متشکل از همه اشیاء $i\in\Omega$ را که با این توصیف a تطبیق دارند، تولید می‌کند.

یک رده ممکن است، مثل بالا، با جستجوی گسترش یک حکم a ساخته شود. به شق دیگر، ممکن است مطلوب آن باشد که همه آن افراد (دیگر) در Ω را که با توصیف $Y_i(u)$ مربوط به فردی معین $u\in\Omega$ تطبیق دارند، پیدا کنیم. بدین ترتیب در این حالت، حکم عبارت است از

$$a_i = \bigwedge_{j=1}^{p} \left[Y_j = Y_j(u)\right]$$

که در آن اکنون $Z_j=Y_j(u)$ و دارای گسترش زیر است،

$$Ext(a_i)=\{i\in\Omega \mid Y_j(i)=Y_j(u), j=1,\ldots,p\}$$

به طور کلی‌تر، حکمی ممکن است با مقداری مدی (از قبیل یک احتمال) صادق باشد، یعنی $0\le a(i)\le 1$ که نمایشگر درجه‌ای از تطبیق یک شیئی i با یک حکم a است.

در این حالت‌ها، نگاشت a بر بازه $[0,1]$ است، یعنی، $a:\Omega\to[0,1]$ در این صورت، گسترش a دارای سطح α، $0\le\alpha\le1$ است. که حال

$$Ext\alpha(a)=Q\alpha=\{i\in a(i)\ge\alpha\}.$$

دیدی و امیلیون (۱۹۹۶) و دیدی و همکاران (۱۹۹۶) اشیاء نمادین مدی را بررسی می‌کنند و برخی از زیر ساخت‌های نظری آن را نیز عرضه می‌دارند. اکنون تعریف رسمی یک شیئی نمادین را به صورت زیر داریم.

تعریف: یک شیئی نمادین عبارت است از سه گانه $S=(a,R,d)$ که در آن a نگاشتی است $a:\Omega\to L$ که افراد $i\in\Omega$ را بر فضای L می‌نگارد و این فضا به رابطه R بین توصیف‌ها و توصیف d بستگی دارد. وقتی $L=\{0,1\}$ یعنی، هنگامی که a نگاشتی دودویی است، S یک شیئی نمادین بولی است. هنگامی که $L=[0,1]$، آنگاه S یک شیئی نمادین مدی است. یعنی، یک شیئی نمادین مدلی ریاضی از یک مفهوم است (ن.ک. به دیدی، ۱۹۹۵). اگر داشته باشیم $[a,R,d]\in\{0,1\}$، آنگاه S یک شیئی نمادین بولی و اگر داشته باشیم $[a,R,d]\in[0,1]$، آنگاه S یک شیئی نمادین مدی است. مثلا در (۷) رابطه $R=(\subseteq,\in)$ و توصیف d=([۲۰و۳۰]و{مشهد}) را داریم. قصد آن است که «۲۰-۳۰ ساله‌هایی را که در مشهد زندگی می‌کنند» بیابیم. گستره متشکل از همه افراد موجود در Ω که با این توصیف تطبیق دارند، یعنی آن iهایی که برای آنها $a(i)=1$.

در حالی که اذعان به ضرورت ابداع روشهایی برای تحلیل نمادین و ابزارهایی برای توصیف اشیاء نمادین ابداع نسبتا جدید است، اندیشه بررسی واحدهای سطح بالاتر مثل مفاهیم در واقع امری باستانی است. ارغنون ارسطو در سده چهارم قبل از میلاد مسیح (ارسطو، IVBC، ۱۹۹٤) به روشنی افراد مرتبه نخست (از قبیل یک اسب یا یک انسان) را که معرف واحدهایی در جهان (جامعه آماری)اند از افراد مرتبه دوم (از قبیل اسب یا انسان) که به عنوان واحدهایی از رده‌ای از افراد معرفی می‌شوند، متمایز می‌سازد.

بعدها آرنو و نیکول (۱۶۶۲) مفهوم را به کمک پنداشته‌های قصد و گستره (که به معنای آنها در ۱۶۶۲ با معانی آنها در این مقاله تطبیق می‌کنند) به شرح زیر تعریف کردند.

اکنون در اندیشه‌های کیهانی دو چیز وجود دارند که تمایز کامل آنها اهمیت دارد؛ فراگیری و گسترش (به جای «قصد» و «گستره»). فراگیری یک اندیشه را صفاتی می‌نامیم که آن اندیشه در بردارد و بدون از بین بردن اندیشه نمی‌توان آنها را از آن گرفت. بدین ترتیب

فراگیری اندیشه یک مثلث، در حد ظاهر، شامل شکل، سه خط، سه زاویه، برابری مجموع این سه زاویه با دو زاویه قائمه و غیره است. گسترش این اندیشه را اشیایی می‌نامیم که این اندیشه درباره آنها مصداق دارد. این اشیاء زیرین‌های یک گزاره کیهانی نیز نامیده می‌شوند و آن گزاره زبرین آنهاست. بدین ترتیب، اندیشه یک مثلث به طور کلی به همه انواع مختلف مثلث گسترش می‌یابد.

سرانجام، کاربست محاسباتی ایجاد رده‌های مقتضی را می‌توان به وسیله احکام مورد استفاده در موتورهای کاوش اجرا کرد. انواع متعددی از این گونه الگوریتمها وجود دارند، مانند نرم‌افزار (SQL) standard query language یا C^{++} یا JAVA. نمونه‌هایی از به کارگیری این الگوریتمها در تعیین رده‌ها در مراجع ذکر شده‌اند.

٤. تحلیل داده‌های نمادین

با در دست داشتن یک مجموعه از داده‌های نمادین، گام بعد انجام تحلیل‌های آماری مقتضی است. مثل داده‌های کلاسیک در اینجا نیز امکانات بی‌پایان هستند. برخلاف داده‌های کلاسیک که یک قرن سعی و کوشش، کتابخانه‌ای از روش‌شناسی‌های تحلیلی یا آماری را فراهم کرده است، تحلیل‌های آماری برای داده‌های نمادین جدید بوده و از نظر تعداد بسیار اندک‌اند. در زیر برخی از آنها را مرور می‌کنیم.

٥. آماره‌های تک متغیره توصیفی

٥-١- مقدمات

آماره‌های توصیفی پایه‌ای عبارتند از بافتنگارهای فراوانی و میانگین و واریانس نمونه. مشابه‌های این آماره‌ها را برای داده‌های نمادین مربوط به متغیرهای چند مقداره و بازه مقدار قاعده‌مند و متغیرهای مدی در نظر می‌گیریم. در حین اینکه این آماره‌ها را بسط می‌دهیم، مثال زیر را در نظر داشته باشیم که نشان می‌دهد باید بین سطوح در داده‌های نمادین و داده‌های کلاسیک تمایز قایل شد. این نکته به ویژه هنگام تشکیل بافتنگار این دو نوع داده آشکارتر می‌شود.

فرض کنید جزیره‌ای دور افتاده دارای هزار پنگوئن و هزار شتر مرغ است که هر دو گونه‌هایی بی‌پرواز از پرندگان‌اند و چهار هزار کبوتر دارد که گونه‌ای پرنده با پرواز است. فرض کنید به متغیر «پرواز» علاقه‌مندیم که در اینجا دو مقدار ممکن «بلی» یا «خیر» را می‌گیرد.

در این صورت، بافتنگار استاندارد مبتنی بر پرندگان طوری است که فراوانی «بلی» دو برابر فراوانی «خیر» است. شکل ۱ الف را ببینید. بر عکس، اگر افراد آماری را گونه‌ها در نظر بگیریم، چون دو گونه بی‌پرواز و یک گونه با پرواز وجود دارند، فراوانی «خیر» اکنون دو برابر فراوانی «بلی» است. شکل ۱ ب را ببینید.

دقت کنید که بافتنگار گونه‌ها صرفا یک بافتنگار است. چیزی که این مثال نشان می‌دهد این است که نسبت به سطح افراد آماری (در اینجا پرندگان) و نسبت به مفهوم (در اینجا گونه‌ها) بافتنگارهای کاملا متفاوتی را به دست می‌آوریم. در اینجا گونه‌ها واقعا در درون خود سطح دیگری از اطلاعات متناظر با نوع پرنده و فراوانی هر یک را دارا هستند. یعنی، «گونه‌ها» خود در بردارنده ساختاری در سطحی دیگر از متغیر گونه‌اند. به نحوی که در تولید بافتنگار شکل ۱ ب به کار رفته است. بافتنگاری نمادین (که در زیر تعریف خواهد شد) این ساختار را به حساب می‌آورد.

آماره‌های تک متغیری را (برای داده‌های $p \geq 1$ متغیر) در نظر بگیریم. نظایر آنها آماره‌های دو متغیری هستند که به خاطر نبود فضا و دقت فعلا آنها را مورد بحث قرار نمی‌دهیم. در مورد متغیرهای با مقدار عدد صحیح و بازه-مقدار از رویکردی که برتران و گوپیل (۲۰۰۰) پیش گرفته‌اند، پیروی می‌کنیم. دوکاروا (۱۹۹۴-۱۹۹۵) و چوآکریا و همکاران (۱۹۹۸) روشهای متفاوت، ولی معادل را برای یافتن بافتنگار و احتمال‌های بازه‌ای برای متغیرهای بازه-مقدار به کار گرفته‌اند (رابطه (۲٦) را در زیر ببینید).

پیش از توصیف این کمیت‌ها، ابتدا نیاز داریم که مفهوم گسترش‌های واقعی را معرفی کنیم. از بخش ۳ به یاد بیاورید که توصیف نمادین یک شیئی $u \in E$ (یا به طور معادل $i \in \Omega$) با بردار توصیف

$$d_u = \left(\xi_{u۱},\ldots,\xi_{up}\right), \qquad u = ۱,..,m$$

یا به طور کلی‌تر با $d \in \left(D_۱,\ldots,D_p\right)$ در فضای $D = \times_{j=۱}^{p} D_j$ داده شد که در آن در هر حالت خاص تحقق Y_j می‌تواند یک x_j برای داده‌های کلاسیک یا یک ξ_i برای داده‌های نمادین باشد. توصیف‌های انفرادی، که با x نمایانده شده، همان توصیف‌هایی‌اند که برای آنها هر D_j یک مجموعه دارای تنها یک مقدار است. یعنی،

$$x \equiv d = (\{x_1\},\ldots\{x_p\}), x \in X = \underset{j=1}{\overset{p}{\times}} Y_j$$

محاسبه بافتنگار فراوانی نمادین شامل شمارش تعداد توصیف‌های انفرادی است که با وابستگی‌های منطقی تلویحی موجود در داده‌ها تطبیق می‌کند. وابستگی منطقی را می‌توان با قاعده‌ای مانند v نمایش داد،

$$v = [x \in A] \Rightarrow [x \in B]$$ (۱۰)

با ازای A⊆D، B⊆D و x∈X و اینکه v نگاشتی از X بر $\{0,1\}$ با $v(x) = 0(1)$ اگر x در قاعده صدق نکند (صدق کند). مثلا فرض کنیم $x = (x_1, x_2) = (10,0)$ توصیفی انفرادی از سن = $Y_1^{(d)}$ و تعداد فرزندان = $Y_2^{(d)}$ برای $u^i \in \Omega$ باشد، فرض کنیم A={سن≤۱۲} و B={ ۰ }. در این صورت، قاعده‌ای که تصریح کند فردی که سنش کمتر از ۱۲ سال است، دال بر این است که فرزندی نداشته است منطقا راست است، در حالی که فردی که سنش کمتر از ۱۲ سال است مستلزم آن باشد که دارای دو فرزند است، منطقا راست نیست. نتیجه می‌شود که بردار توصیف انفرادی x در قاعده v صدق می‌کند اگر و تنها اگر x∈A∩B یا x∉A. این فرمول‌بندی قاعده وابستگی منطقی برای مقاصد محاسبه آماره‌های توصیفی پایه‌ای کافی است. ورده و دوکاروالو (۱۹۹۸) انواع گوناگون قواعد مرتبط (از قبیل هم ارزی منطقی، استلزام منطقی، وابستگی‌های چند گانه، وابستگی‌های سلسله مراتبی و مانند آن را) مورد بحث قرار می‌دهند. پس تعریف رسمی زیر را داریم

تعریف: توصیف واقعی بردار توصیف d عبارت است از مجموعه همه بردارهای توصیف انفرادی x که در همه قواعد (وابستگی منطقی) v در x صدق می‌کنند. این تعریف را به صورت زیر می‌نویسیم،

$$v\,ir(d) = \{x \in D, v(x) = 1, \forall v\ in V_x\}$$ (۱۱)

که در آن V_x مجموعه همه قواهد v است که روی x عمل می‌کنند.

۵-۲- متغیرهای چند مقداره - آماره‌های تک متغیره

فرض کنید می‌خواهیم توزیع فراوانی را برای متغیر نمادین چند مقداره خاص $Y_j \equiv Z$ که مقادیر خاص $\xi \in Z$ را اختیار می‌کند، بیابیم. این مقادیر می‌توانند مقادیر رسته‌ای (مثل انواع سرطان) یا هر نوعی از متغیرهای تصادفی گسسته باشند. فراوانی مشاهده شده ξ را به صورت

$$O_Z(\xi) = \sum\nolimits_{u\in E} \pi_Z(\xi;u)$$ (۱۲)

تعریف می‌کنیم که در آن مجموع یابی روی $u \in E = \{1,\ldots,m\}$ است و

$$\pi_Z(x;u) = \frac{|\{x \in vir(d_u) | x_Z = \xi\}|}{|vir(d_u)|}$$ (۱۳)

درصد بردارهای توصیف انفرادی در $vir(d_u)$ است به طوری که داشته باشیم $x_Z = \xi$ و $|A|$ تعداد توصیف‌های انفرادی موجود در فضای A است. در مجموع‌یابی (۱۲) هر u که برای آن $vir(d_u)$ تهی باشد، نادیده گرفته می‌شود. ملاحظه می‌کنیم که این فراوانی مشاهده شده یک عدد مثبت حقیقی است و مثل مورد داده‌های کلاسیک لزوما عددی صحیح نیست.

در حالت کلاسیک، $|vir(d_u)| = 1$ و بنابراین حالتی خاص از (۱۳) است. به آسانی می‌توان نشان داد که

$$\sum\nolimits_{\xi\in Z} O_Z(\xi) = m'$$ (۱۴)

که $m' = (m - m_0)$ و m_0 تعداد uهایی است که برای آنها $|vir(d_u)| = 0$.

برای یک متغیر نمادین چند مقداره Z، که مقادیر $\xi \in Z$ را اختیار می‌کند، **توزیع فراوانی تجربی** عبارت است از مجموعه زوج‌های $[\xi, O_Z(\xi)]$ به ازای $\xi \in Z$ و توزیع **فراوانی نسبی** یا **بافتنگار فراوانی** عبارت است از مجموعه

$$[\xi, (m')^{-1} O_Z(\xi)]$$ (۱۵)

تعریف‌های زیر بلافاصله نتیجه می‌شود.

تابع توزیع تجربی Z به صورت زیر تعریف می‌شود.

$$F_Z(\xi) = (m')^{-1} \sum\nolimits_{\xi_k \le \xi} O_Z(\xi_k)$$ (۱۶)

هنگامی که مقادیر ممکن ξ برای متغیر نمادین چند مقداره $Y_j=Z$ کمی باشند، می‌توانیم به شرح زیر میانگین، واریانس و میانه را تعریف کنیم. میانگین نمونه نمادین برابر است با

$$\bar{Z} = (m')^{-1} \sum\nolimits_{\xi_k} O_Z(\xi_k)$$ (۱۷)

واریانس نمونه نمادین عبارت است از

$$S_Z^{2} = (m')^{-1}\sum_{\xi_k} O_Z(\xi_k)\left[\xi_k - \bar{Z}\right]^{2} \qquad (۱۸)$$

و میانه نمادین آن مقدار از ξ است که به ازای آن

$$F_Z(\xi) \ge \frac{1}{2} \quad , \; F_Z(\xi) \le \frac{1}{2} \qquad (۱۹)$$

مثال

برای توضیح مطالب بالا، فرض کنید مجموعه‌ای بزرگ از داده‌ها (مشتمل بر بیمارانی که از سوی تامین اجتماعی، خدمات بهداشتی و درمانی دریافت می‌کنند) به طریقی انبوهیده شده‌اند که داده‌های جدول ۳ را داده‌اند. این داده‌ها برآمدهای مربوط به وجود سرطان Y_1 با {۱=بلی، ۰=خیر} = Y_1 و تعداد معالجات مرتبط با سرطان Y_2 با {۰،۱،۲،۳} = Y_2 را درباره $m = 9$ شیئی نمادین نشان می‌دهند.

بدین ترتیب، مثلا، $d_1 = (\{0,1\},\{2\})$ بردار توصیف برای شیئی مندرج در سطر ۱ جدول ۳ است. پس، برای افرادی که با این توصیف نمایش داده می‌شوند، مشاهده $Y_1 = \{0,1\}$ بیان می‌کند که یا بعضی افراد سرطان دارند و برخی ندارند یا آنکه تشخیص خیر/بلی دقیق را برای افراد رده‌بندی شده در اینجا نمی‌دانیم، در حالی که مشاهده $Y_2 = \{2\}$ به ما می‌گوید که همه افراد معرفی شده با d_1 دو مورد معالجه مرتبط با سرطان داشته‌اند. برعکس، $d_7 = (\{1\},\{2,3\})$ معرف افرادی است که همه آنان دارای تشخیص سرطانی بوده‌اند $(Y_1 = 1)$ و آنهایی که ۲ یا ۳ معالجه داشته‌اند، $Y_2 = \{2,3\}$. افزون بر آن فرض کنید یک وابستگی منطقی

$$\nu = y_1 \in \{0\} \Rightarrow y_2\{0\} \qquad (۲۰)$$

وجود داشته باشد، یعنی اگر سرطانی تشخیص داده نشده است، نباید معالجات سرطانی در بین باشد. دقت کنید که

$$y_1 \in \{0\} \Rightarrow A = \{(0,0),(0,1),(0,2),(0,3)\}$$

و

$$y_2 \in \{0\} \Rightarrow B = \{(0,0),(1,0),(2,0),(3,0)\}$$

از (۲۰) نتیجه می‌شود که یک قاعده توصیف انفرادی x که در این قاعده صدق می‌کند، عبارت است از $x \in A \cap B = \{(0,0)\}$ یا $x \notin A$ یعنی، $x \notin \{(1,0),(1,1),(1,2),(1,3)\}$. گیریم همه موارد ممکنی را که در این قاعده صدق می‌کنند با $x = \{(0,0),(1,0),\ldots,(1,3)\}$ نشان دهیم. رابطه (۲۰) را در هر مورد هر d_u، $u = 1,\ldots,m$ در داده‌ها به کار خواهیم بست تا گسترش‌های واقعی $vir(d_u)$ را بیابیم.

بدین ترتیب، برای توصیف اول d_1 داریم،

$$\{vir(d_1) = \{x \in \{1,0\}\times\{2\}\} : \nu(x) = 1\}$$

توصیف‌های انفرادی $x \in \{1,0\}\times\{2\}$ عبارتند از (۲و۰) و (۲و۱) که از آنها تنها یکی، $x = (1,2)$ در فضای C نیز هست. بنابراین، $vir(d_1) = \{(1,2)\}$.

در واقع، این عملیات عبارت است از به اصطلاح پالایش داده‌ها از نظر ریاضی از راه شناسایی فقط آن داده‌هایی که (با صدق کردن در قاعده وابستگی منطقی رابطه (۲۰)) معنای منطقی دارند. بدین ترتیب، مقادیر داده‌ای $(Y_1, Y_2) = (0,2)$ که این اطلاع را ثبت می‌کنند که در عین حال که سرطان وجود نداشته دو معالجه مرتبط با سرطان صورت گرفته (تحت شرایط موجودی که با قاعده ۷ مشخص می‌شوند) به عنوان مقادیر داده‌ای غلط شناسایی می‌شوند و بنابراین برای محاسبه میانگین در این تحلیل به کار نمی‌روند. در حالی که سعی در انجام چنین کاری را در اینجا نداریم، این شناسایی مانع از آن نمی‌شود که شیوه‌هایی دیگر را که متعاقبا به کار خواهند رفت، تا آنچه را که این داده‌ها می‌توانستند باشند جانشین آنها کنند، در تحلیل بگنجانیم. در مورد مجموعه‌های کوچک داده‌ها، ممکن است میسر باشد که داده‌ها را به طور بصری (یا به نوعی دیگر) «تصحیح» کنیم. برای مجموعه‌های بسیار بزرگ داده‌ها، این امر همیشه ممکن نیست، از این رو یک قاعده وابستگی منطقی برای انجام عمل تصحیح به طور ریاضی/محاسباتی امری اساسی است.

به طور مشابه برای توصیف دوم d_2 می‌توانیم به دست آوریم.

$$vir(d_2) = \{x \in \{0,1\}\times\{0,1\} = \nu(x) = 1\}$$

در اینجا، بردارهای توصیف انفرادی $\{x \in \{0,1\}\times\{0,1\}\}$ عبارتند از (۰و۰)(۱و۰)(۰و۱)و(۱و۱) که از آنها $x = (0,0)$ و $x = (1,1)$ و در C هستند. پس $vir(d_2) = \{(0,0),(1,0),(1,1)\}$. گسترش‌های واقعی برای همه d_u، $u = 1,\ldots,9$ در جدول ۳ ارائه شده‌اند. توجه کنید $vir(d_5) = \phi$ مجموعه تهی است، زیرا مقدار داده‌ای d_5 نمی‌تواند با وجود قاعده ۷ منطقا درست باشد.

اکنون می‌توانیم توزیع فراوانی را بیابیم. فرض کنید ابتدا این توزیع را برای Y_1 پیدا کنیم. بنابه تعریف (۱۲)، فراوانی‌های مشاهده شده

$$O_{Y_1}(0) = \sum_{u \in E'} \frac{\left|\{x \in vir(d_u) \,|\, x_{Y_1} = 0\}\right|}{\left|\{vir(d_u)\}\right|}$$

$$=\frac{۰}{۱}+\frac{۱}{۳}+\frac{۰}{۱}+\frac{۰}{۲}+\frac{۱}{۱}+\frac{۰}{۲}+\frac{۰}{۲}+\frac{۰}{۲}=\frac{۴}{۳}$$

و به همین ترتیب $O_{Y_۱}(۱)=\frac{۲۰}{۳}$ که در آنها $E'=E-(u=۵)$ به طوری که $|E'|=۸=m'$. بنابراین، توزیع فراوانی نسبی برای $Y_۱$ طبق رابطه (۱۵) عبارت است از

فراوانی نسبی $Y_۱$:

$$[(۰,O_{Y_۱}(۰)/m'),(۱,O_{Y_۱}(۱)/m')]=\left[\left(۰,\frac{۱}{۶}\right),\left(۱,\frac{۵}{۶}\right)\right]$$

به طور مشابه، فراوانی‌های مشاهده شده برای مقادیر ممکن $Y_۲$، یعنی $\xi=۰,۱,۲,۳$ را به ترتیب چنین داریم

$$O_{Y_۲}(۰)=\sum_{u\in E'}\frac{|\{x\in vir(d_u)|x_{Y_۲}=۰\}|}{|vir(d_u)|}$$

$$=\frac{۰}{۱}+\frac{۲}{۳}+\frac{۰}{۱}+\frac{۰}{۲}+\frac{۱}{۱}+\frac{۰}{۲}+\frac{۰}{۲}+\frac{۰}{۲}=\frac{۵}{۳}$$

$O_{Y_۲}(۲)=۲/۵$ و $O_{Y_۲}(۳)=۲/۵$

$$O_{Y_۲}(۱)=\frac{۴}{۳}$$

و از این رو، فراوانی نسبی $Y_۲$ طبق رابطه (۱۵) عبارت است از،

فراوانی نسبی $Y_۲$: $\left[\left(۰,\frac{۵}{۲۴}\right),\left(۱,\frac{۱}{۶}\right),\left(۲,\frac{۵}{۱۶}\right),\left(۳,\frac{۵}{۱۶}\right)\right]$

تابع توزیع تجربی برای $Y_۲$ طبق رابطه (۱۶) چنین است،

$$F_{Y_۲}(\xi)=\begin{array}{ll}\frac{۵}{۲۴} & \xi<۱\\ \frac{۳}{۸} & ۱<\xi\le ۲\\ \frac{۱۱}{۱۶} & ۲\le\xi\le ۳\\ ۱ & \xi\ge ۳\end{array}$$

از روی رابطه (۱۷) میانگین نمونه نمادین $Y_۱$ و $Y_۲$ به ترتیب برابر است با $\bar{Y}_۱=\frac{۵}{۶}$ و $\bar{Y}_۲=\frac{۸۳}{۸۴}$. از روی رابطه (۱۸) واریانس نمونه نمادین $Y_۱$ و $Y_۲$ به ترتیب $S_۱^۲=۰/۱۲۳۹$ و $S_۲^۲=۱/۱۹۱۵$ است و میانه $Y_۲$ برابر ۲ است.

سرانجام ملاحظه می‌کنیم که می‌توانیم فراوانی‌های موزون، میانگین‌های موزون و واریانس‌های موزون را به جای (۱۲) با استفاده از

$$O_Z(\xi)=\sum_{u\in E}W_u\prod Z(\xi,u) \qquad (۲۱)$$

حساب کنیم که $w\ge ۰$ و $\sum w_V=۱$. مثلا اگر شیئی $u\in E$ رده‌های C_u متشکل از افرادی از مجموعه افراد $\Omega=\{۱,...,n\}$ باشند، یعنی $C_u\subseteq\Omega$ یک وزن ممکن عبارت است از $w_u=|C_u|/|\Omega|$ که متناظر با اندازه‌های نسبی رده‌های C_u، $u=۱,...,m$ است یا به طور کلی‌تر، اگر هر بردار توصیف انفرادی x را به صورت واحدی ابتدایی از شیئی u در نظر بگیریم، می‌توانیم برای $u\in E$ وزن‌های

$$w_u=\frac{|vir(d_u)|}{\sum_{v\in E}|vir(d_u)|} \qquad (۲۲)$$

را به کار ببریم.

مثلا، اگر uهای جدول ۳ رده‌های C_u با اندازه $|C_u|$ را نمایش دهند، با وجود $|\Omega|=۱۰۰۰$ همانگونه که در جدول ۳ نشان داده شده اگر از وزن $w_u=|C_u|/|\Omega|$ استفاده کنیم می‌توانیم نشان دهیم که

$$O_{Y_۱}=(۰)=\frac{۱}{۱۰۰۰}\left[۱۲۸\times\frac{۰}{۱}+۷۵\times\frac{۱}{۳}+۲۴۹\times\frac{۰}{۱}+\cdots+۱۲\times\frac{۰}{۲}\right]$$

$$=۰/۲۲۹$$

$$O_{Y_۱}=(۱)=۰/۷۷۱$$

و از آنجا که فراوانی نسبی $Y_۱$ عبارت است از،

فراوانی نسبی $Y_۱$: $[(۰,۰/۲۲۹),(۱,۰/۷۷۱)]$

به همین ترتیب،

$O_{Y_۲}(۲)=۰/۲۳۹۵$ ، $O_{Y_۲}(۳)=۰/۴۰۹۵$

$O_{Y_۲}(۰)=۰/۲۵۴۰$ ، $O_{Y_۲}(۱)=۰/۰۹۷۰$

و بنابراین فراوانی نسبی $Y_۲$ چنین است.

فراوانی‌نسبی $Y_۲$:

$$[(۰,۰/۲۵۴۰),(۱,۰/۰۹۷۰),(۲,۰/۲۳۹۵),(۳,۰/۴۰۹۵)]$$

تابع توزیع تجربی $Y_۲$ تبدیل می‌شود به

$$F_{Y_۲}(\xi)=\begin{array}{ll}۰/۲۵۴۰ & \xi<۱\\ ۰/۳۵۱۰ & ۱<\xi\le ۲\\ ۰/۵۹۰۵ & ۲\le\xi\le ۳\\ ۱ & \xi\ge ۳\end{array}$$

به طور مشابه، می‌توانیم نشان دهیم که میانگین نمونه نمادین $Y_۱$ برابر است با $\bar{Y}_۱=۰/۷۷۱۰$ و از آن $Y_۲$ عبارت است از $\bar{Y}_۲=۱/۸۰۴۵$ و واریانس نمونه نمادین $Y_۱$ برابر است با $S_۱^۲=۰/۱۷۶۰$ و از آن $Y_۲$ برابر است با $S_۲^۲=۱/۴۸۴۳$.

۵-۳- متغیرهای بازه مقدار - آماره‌های تک متغیره

آماره‌های توصیفی متناظر برای متغیرهای بازه-مقدار به طرز مشابه به دست می‌آیند. برتران و گوپیل (۲۰۰۰) را ببینید. گیریم به متغیر خاص $Y_j \equiv Z$ علاقه‌مند باشیم و فرض کنیم که مقدار مشاهده شده مربوط به شیئی u عبارت است از بازه $Z(u) = [a_u, b_u]$، به ازای $u \in E = \{1,...,m\}$.

بردارهای توصیف انفرادی $x \in vir(d_u)$ فرض می‌شوند که به طور یکنواخت بر روی بازه $Z(u)$ توزیع شده‌اند. بنابراین، نتیجه می‌شود که برای هر ξ،

$$P\{x \le \xi \mid x \in vir(d_u)\} = \begin{cases} 0 & \text{،} \quad \xi < a_u \\ \dfrac{\xi - a_u}{b_u - a_u} & \text{،} \quad a_u \le \xi < b_u \\ 1 & \text{،} \quad \xi \ge b_u \end{cases}$$

بردار توصیف انفرادی به طور یکجا در $U_{u\in E} vir(d_u)$ اختیار می‌کند. به علاوه فرض بر آن است که هر شیئی با احتمال $1/m$ و به طور همشانس مشاهده می‌شود. بنابراین، تابع توزیع تجربی $F_Z(\xi)$، تابع توزیع آمیزه‌ای از m توزیع یکنواخت $\{Z(u), u = \{1,...,m\}\}$ است. بنابراین بنابر (۲۳)،

$$F_Z(\xi) = \frac{1}{m}\sum_{u\in E} P\{x \le \xi \mid x \in vir(d_u)\}$$
$$= \frac{1}{m}\left\{\sum_{\xi\in Z(u)} \frac{\xi - a_u}{b_u - a_u} + |(u \mid \xi \ge b_u)|\right\}$$

از آنجا که از راه مشتق‌گیری نسبت به ξ، **تابع چگالی تجربی** Z را به صورت زیر به دست می‌آوریم.

$$f(\xi) = \frac{1}{m}\sum_{u:\xi\in z(u)}\left(\frac{1}{b_u - a_u}\right) \tag{۲۴}$$

دقت کنید که مجموع‌یابی موجود در (۲۴) تنها روی آن اشیاء u است که برای آنها $\xi \in Z(u)$. می‌توانیم(۲۴) را به صورتی دیگر بنویسیم،

$$f(\xi) = \frac{1}{m}\sum_{u\in E}\frac{I_u(\xi)}{\|Z(u)\|} \quad \text{،} \quad \xi \in \Re \tag{۲۵}$$

که در آن $I_u(\cdot)$ تابع نشانگر است که ξ در بازه $Z(u)$ هست یا نیست و $\|Z(u)\|$ طول آن بازه است. شباهت با (۱۵) و (۱۶) آشکار است. برتران و گوپیل (۲۰۰۰) با استفاده از قانون عددهای بزرگ، خاطرنشان می‌سازند که توزیع حدی واقعی Z وقتی $m \to \infty$ فقط با توزیع دقیق $f(\xi)$ مذکور در (۲۵) تقریب زده می‌شود، زیرا این توزیع به درستی توزیع یکنواخت درون هر بازه بستگی دارد.

برای ساختن بافتنگار، گیریم $I = \left[\min_{u\in E} a_u, \max_{u\in E} b_u\right]$ بازه‌ای باشد که همه مقادیر مشاهده شده Z در X ایجاد می‌کنند و فرض کنید که I را به r زیر بازهٔ $I_g = (\xi_{g-1}, \xi_g)$، $g = 1,...,r-1$ و $I_r = (\xi_{r-1}, \xi_r)$ افراز کنیم. آنگاه **بافتنگار** Z نمایش نموداری توزیع فراوانی $\{(I_g, p_g), g = 1,...,r\}$ است که در آن

$$p_g = \frac{1}{m}\sum_{u\in E}\frac{\|Z(u)\cap I_g\|}{\|Z(u)\|} \tag{۲۶}$$

یعنی، p_g احتمال آن است که یک بردار توصیف انفرادی دلخواه x در بازه I_g بیفتد. اگر بخواهیم بافتنگار را با ارتفاع f_g روی بازه I_g رسم کنیم، به طوری که «مساحت» برابر p_g باشد، آنگاه

$$p_g = (\xi_g - \xi_{g-1}) \times f_g. \tag{۲۷}$$

زیر ساخت‌های ریاضی موجود مربوط به بافتنگار در دیدی (۱۹۹۵) با استفاده از قانون قوی عددهای بزرگ و مفاهیم t-نرم‌ها و t-همنرم‌ها که از سوی شوایتزر و اسکلر (۱۹۸۳) ابداع شده‌اند، بسط یافته‌اند.

میانگین نمونه نمادین

برای یک متغیر بازه-مقدار Z با

$$\bar{Z} = \frac{1}{2m}\sum_{u\in E}(b_u + a_u) \tag{۲۸}$$

داده می‌شود. برای تحقیق (۲۸) به خاطر بیاورید که میانگین تجربی $\bar{Z}$ بر حسب تابع چگالی تجربی عبارت است از،

$$\bar{Z} = \int_{\infty}^{\infty}\xi f(\xi)\, d\xi$$

پس از جایگذاری مقادیر از (۲۵) داریم،

$$\bar{Z} = \frac{1}{m}\sum_{u\in E}\int_{\infty}^{\infty}\frac{I_u(\xi)}{\|Z(u)\|}\quad \xi d\xi$$
$$= \frac{1}{m}\sum_{u\in E}\frac{1}{b_u - a_u}\int_{\xi\in Z(u)}\xi\, d\xi$$
$$= \frac{1}{2m}\sum_{u\in E}\frac{b_u^2 - a_u^2}{b_u - a_u} = \frac{1}{2m}\sum_{u\in E}(b_u + a_u)$$

به همین ترتیب، می‌توانیم واریانس نمونه نمادین را به دست آوریم.

(۲۹)

$$S^{۲} = \frac{۱}{۴m}\sum_{u\in E}(b_u + a_u)^{۲} - \frac{۱}{۴m^{۲}}\sum_{u\in E}(b_u + a_u)^{۲}$$

در مورد متغیرهای چند مقداره، اگر شیئی u دارای ناسازگاری‌های درونی نسبت به یک قاعده منطقی باشد، یعنی اگر u چنان باشد که $|vir(d_u)| = ۰$، آنگاه مجموع‌یابی (۲۸) و (۲۹) فقط روی uهایی است که برای آنها $|vir(d_u)| \neq ۰$ یعنی بر روی $u \in E'$ است، و به جای m عدد m' گذاشته شود (که برابر با تعداد اشیاء u موجود در E' است).

در زیر قرار می‌گذاریم که m و E به آن uهایی اشاره کنند که برای آنها این قواعد صادق‌اند.

مثال: برای توضیح مطلب، داده‌های راجو (۱۹۹۷) را که در جدول ٤ آمده‌اند، در نظر بگیرید.

در آنجا نبض $(Y_۱)$، فشار خون سیستولیک $(Y_۲)$ و دیاستولیک $(Y_۳)$ به صورت متغیرهای بازه-مقدار به ازای هر $u = ۱,...,۱۰$ بیمار ثبت شده‌اند.

گیریم داده‌های نبض $(Y_۱)$ را برداریم. مجموعه کل داده‌ها بازه $I = [۴,۱۱۲]$ را ایجاد می‌کند که در آن

$$\min_{u\in E} a_u = ۴۴ \qquad , \qquad \max_{u\in E} b_u = ۱۱۲$$

فرض کنید می‌خواهیم بافتنگاری را روی $r = ۸$ بازه $I_۱ = [۴۰,۵۰],...,I_۸ = [۱۱۰,۱۲۰]$ بسازیم. با استفاده از رابطه (٢٦) می‌توانیم احتمال p_g را که بردار توصیف انفرادی دلخواه x در بازه I_g، $g = ۱,...,۸$ بیفتد، حساب کنیم. مثلا وقتی $g = ۴$ احتمال آنکه یک x در بازه $I_۴ = [۷۰,۸۰]$ بیفتد، برابر است با

$$p_۴ = \frac{۱}{۱۰}\left\{۰ + \frac{۲}{۱۲} + \frac{۱۰}{۳۴} + \frac{۱۰}{۴۲} + \frac{۲}{۱۸} + \frac{۱۰}{۳۰} + \frac{۸}{۲۸} + \frac{۴}{۲۲} + ۰ + ۰\right\} = ۰/۱۶۱۱$$

از آنجا، بنابر رابطه (۲۷) می‌توانیم ارتفاع f_g از بافتنگار رسم شده را برای آن بازه برابر با

$$f_g = p_g / (\xi_g - \xi_{g-۱})$$

یعنی،

$$f_۴ = (۰/۱۶۱۱)/۱۰ = ۰/۱۶۱۱$$

حساب کنیم. خلاصه‌ای از مقادیر محاسبه شده P_g و f_g در جدول ٥ داده شده‌اند و نمودار بافتنگار در شکل ۲ رسم شده است. جدول ٥ همچنین بازه‌ها و احتمال‌های مربوط به متغیر بازه-مقدار $Y_۲$ را که معرف فشار خون سیستولیک است و $Y_۳$ را که نمایشگر فشار خون دیاستولیک است، برای همان ۱۰ بیمارستان نشان می‌دهد.

با استفاده از روابط (۲۸) و (۲۹) به ترتیب میانگین و واریانس نمونه نمادین را به دست می‌آوریم. بدین ترتیب، میانگین نبض $\overline{Y}_۱ = ۷۹/۱$ با واریانس $S_۱^{۲} = ۱۶۲/۲۹$؛ میانگین فشار خون سیستولیک $\overline{Y}_۲ = ۱۳۱/۳$ با واریانس $S_۲^{۲} = ۴۹۵/۴۱$ و میانگین فشار خون دیاستولیک $\overline{Y}_۳ = ۸۴/۶$ با واریانس $S_۳^{۲} = ۱۸۲/۴۴$.

نظیر مفاهیم و روشهای محاسباتی بالا را می‌توان در مورد متغیرهای مدی، چند مقداره، و بازه-مقدار نیز تعریف کرد. همچنین می‌توان این گونه مفاهیم و روشها را به آماره‌های توصیفی دو متغیری مربوط به متغیرهای چند مقداره و بازه-مقدار تعمیم داد.

راستایی دیگر برای تحویل مشاهدات به مجموعه‌هایی کوچکتر استفاده از روشهای تحویل داده‌ها نظیر تحلیل مؤلفه‌های اصلی و خوشه‌بندی نمادین است. کوشش‌هایی برای گسترش به داده‌های سه طرفه نیز به عمل آمده‌اند که برای اطلاع از آنها به بیلارد و دیدی (۲۰۰۱) مراجعه شود.

٦. نتیجه‌گیری

با وجود قالبهای داده‌ای و اندازه‌های مجموعه‌های داده‌ها، نیاز به ابداع روشهای آماری به منظور تحلیل آنها اهمیت روزافزون دارد. جهان آمار گنجینه‌ای از روش‌شناسی‌ها را در طی قرن بیستم میلادی ابداع کرده است. روشهایی که (در مقام قیاس) عمدتا محدود به مجموعه‌های کوچک داده‌ها و قالبهای داده‌ای کلاسیک (اسکالری، برداری و ماتریسی)اند. این مقاله به اجمال فرمول‌بندی و ساخت اشیاء و رده‌های نمادین را مرور کردیم. به کوتاهی به روشهای تحلیل نمادین اشاره کرده، با مثالهایی طرز کار را بیان کردیم. نیاز به پژوهش و ابداع بیشتر به منظور انجام استنباطهای آماری درباره داده‌های نمادین آشکار بوده، افقی است که اخیرا پیش چشم آمارشناسان گشوده شده است.

جدول ۱

i	Y_1	Y_2	Y_3	…	Y_8	…	Y_{28}	Y_{29}	Y_{30}	شرح متغیرها
۱	بوشهر	M	۲۴		۰		N	N	۰	y_1 : شهر محل سکونت
۲	بوشهر	M	۵۶		۲		N	N	۱	y_2 : جنس(F,M)
۳	ایلام	F	۴۸		۲		y	N	۱	y_3 : سن به سال
۴	ایلام	F	۴۷		۱		Y	۰	۰	y_4 :نژاد(سفیدW، سیاهB، دیگرO)
۵	آبادان	M	۷۹		۴		N	۰	۰	y_5 : تأهل(مجردS، متأهلM)
۶	لامرد	M	۱۲		۰		N	۰	۰	y_6 : تعداد والدین زنده(۰، ۱، ...)
۷	لامرد	M	۶۷		۰		Y	۶	۰	y_7 : تعداد برادران و خواهران(۰، ۱، ...)
۸	لامرد	F	۷۳		۴		N	۰	۲	y_8 : تعداد فرزندان (۰، ۱، ...)
۹	شیراز	M	۲۹		۲		N	۰	۰	.
۱۰	کرمان	F	۴۴		۳		y	۰	۰	.
.	.	.	.	.	.	.	.	.	.	y_{28} : تشخیص سرطان (بلیy، خیرN)
.	.	.	.	.	.	.	.	.	.	y_{29} : تعداد دفعات معالجه سرطان پستان (۱،۲،...،N= مصداق ندارد)
.	.	.	.	.	.	.	.		.	y_{30} : تعداد دفعات معالجه سرطان ریه (۱،۲،...)

جدول ۲- نمونه‌ای از داده‌های نمادین

U	سن	فشار خون	شهر	نوع سرطان	جنس	وزن
۱	[۲۰،۳۰]	(۷۹،۱۲۰)	بوشهر	{سرطان مغزی}	{زن}	[۷۰،۸۰]
۲	[۵۰،۶۰]	(۹۰،۱۳۰)	بوشهر	{کبد، ریه}	{مرد}	[۶۰،۷۵]
۳	[۴۵،۵۵]	(۸۰،۱۳۰)	ایلام	{پروستات}	{مرد}	[۶۰،۶۵]
۴	[۴۷،۴۷]	(۸۶،۱۲۱)	ایلام	{ریه $(1-p)$، پستان p}	{زن}	[۵۵،۶۵]
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

جدول ۳- مجموعه‌ای از داده‌های نمادین

u	$Y_۱$	$Y_۲$	$vir(d_u)$	$\|vir(d_u)\|$	$\|C_u\|$
۱	{۰،۱}	{۲}	{(۱،۲)}	۱	۱۲۸
۲	{۰،۱}	{۰،۱}	{(۰،۰)،(۱،۰)،(۱،۱)}	۳	۷۵
۳	{۰،۱}	{۳}	{(۱،۳)}	۱	۲۴۹
۴	{۰،۱}	{۲،۳}	{(۱،۲)،(۱،۳)}	۲	۱۱۳
۵	{۰}	{۱}	ϕ	۰	۲
۶	{۰}	{۰،۱}	{(۰،۰)}	۱	۲۰۴
۷	{۱}	{۲،۳}	{(۱،۲)،(۱،۳)}	۲	۸۷
۸	{۱}	{۱،۲}	{(۱،۱)،(۱،۲)}	۲	۲۳
۹	{۱}	{۰،۳}	{(۱،۱)،(۱،۳)}	۲	۱۲۱

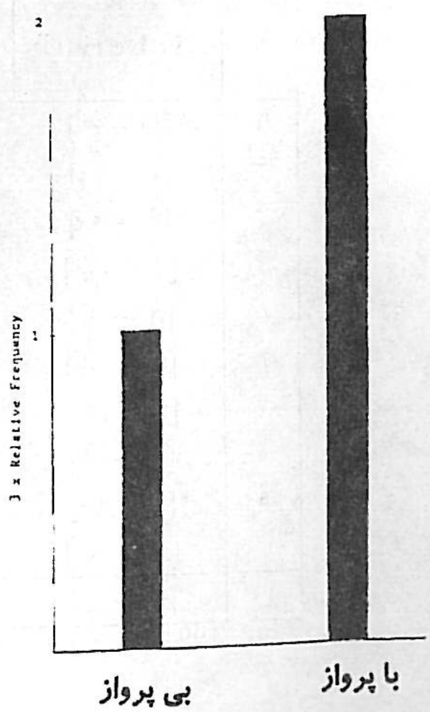
جدول ٤- سه متغیر نمادین

u	$:Y_۱$ **Pulse rate**	$:Y_۲$ **Systolic Pressure**	$:Y_۳$ **Diagnostic ressure**
۱	[۴۴ − ۶۸]	[۹۰ − ۱۱۰]	[۵۰ − ۷۰]
۲	[۶۰ − ۷۲]	[۹۰ − ۱۳۰]	[۷۰ − ۹۰]
۳	[۵۶ − ۹۰]	[۱۴۰ − ۱۸۰]	[۹۰ − ۱۰۰]
۴	[۷۰ − ۱۱۲]	[۱۱۰ − ۱۴۲]	[۸۰ − ۱۰۸]
۵	[۵۴ − ۷۲]	[۹۰ − ۱۰۰]	[۵۰ − ۷۰]
۶	[۷۰ − ۱۰۰]	[۱۳۴ − ۱۴۲]	[۸۰ − ۱۱۰]
۷	[۷۲ − ۱۰۰]	[۱۳۰ − ۱۶۰]	[۷۶ − ۹۰]
۸	[۷۶ − ۹۸]	[۱۱۰ − ۱۹۰]	[۷۰ − ۱۱۰]
۹	[۸۶ − ۹۶]	[۱۳۸ − ۱۸۰]	[۹۰ − ۱۱۰]
۱۰	[۸۶ − ۱۰۰]	[۱۱۰ − ۱۵۰]	[۷۸ − ۱۰۰]
۱۱	[۶۳ − ۷۵]	[۶۰ − ۱۰۰]	[۱۴۰ − ۱۵۰]

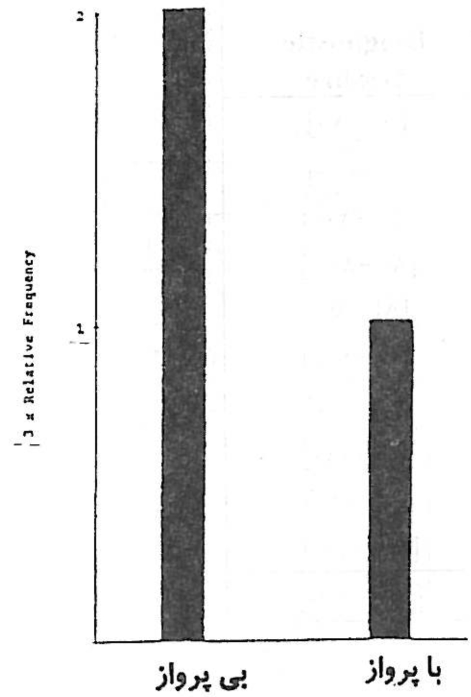
جدول ۵

$Y_{۱}$: *Pulse Rate*		$Y_{۲}$: *Systolic pressure*		$Y_{۳}$: *Diagnostic pressure*	
I_g	p_g	I_g	p_g	I_g	p_g
[۴۰,۵۰)	۰/۰۲۵۰	[۹۰,۱۰۰)	۰/۱۷۵۰	[۵۰,۶۰)	۰/۱۰۰۰
[۵۰,۶۰)	۰/۰۸۶۸	[۱۰۰,۱۱۰)	۰/۰۷۵۰	[۶۰,۷۰)	۰/۱۰۰۰
[۶۰,۷۰)	۰/۲۰۱۶	[۱۱۰,۱۲۰)	۰/۰۹۳۸	[۷۰,۸۰)	۰/۱۱۲۷
[۷۰,۸۰)	۰/۱۶۱۱	[۱۲۰,۱۳۰)	۰/۰۹۳۸	[۸۰,۹۰)	۰/۲۶۰۹
[۸۰,۹۰)	۰/۲۳۶۳	[۱۳۰,۱۴۰)	۰/۱۸۱۸	[۹۰,۱۰۰)	۰/۲۸۹۵
[۹۰,۱۰۰)	۰/۲۶۰۶	[۱۴۰,۱۵۰)	۰/۱۵۰۹	[۱۰۰,۱۱۰)	۰/۱۳۶۹
[۱۰۰,۱۱۰)	۰/۰۲۳۸	[۱۵۰,۱۶۰)	۰/۰۹۴۶		
[۱۱۰,۱۲۰)	۰/۰۰۴۸	[۱۶۰,۱۷۰)	۰/۰۶۱۳		
		[۱۷۰,۱۸۰)	۰/۰۶۱۳		
		[۱۸۰,۱۹۰)	۰/۰۱۲۵		
Mean	$\overline{Y}_{۱} = ۷۹/۱$	$\overline{Y}_{۲} = ۱۳۱/۱$		$\overline{Y}_{۳} = ۸۴/۶$	
Variance	$S_{۱}^{۲} = ۱۶۲/۲۹$	$S_{۱}^{۲} = ۱۶۲/۲۹$		$S_{۳}^{۲} = ۱۸۲/۴۴$	
Covariance	$S_{۱۲} = ۱۹۴/۱۷۰$	$S_{۲۳} = ۲۵۷/۹۲۰$		$S_{۱۳} = ۱۴۱/۰۴۰$	
Correlation	$r(Y_{۱},Y_{۲}) = ۰/۶۸۵$	$r(Y_{۲},Y_{۳}) = ۰/۸۵۸$		$r(Y_{۱},Y_{۳}) = ۰/۸۲۰$	

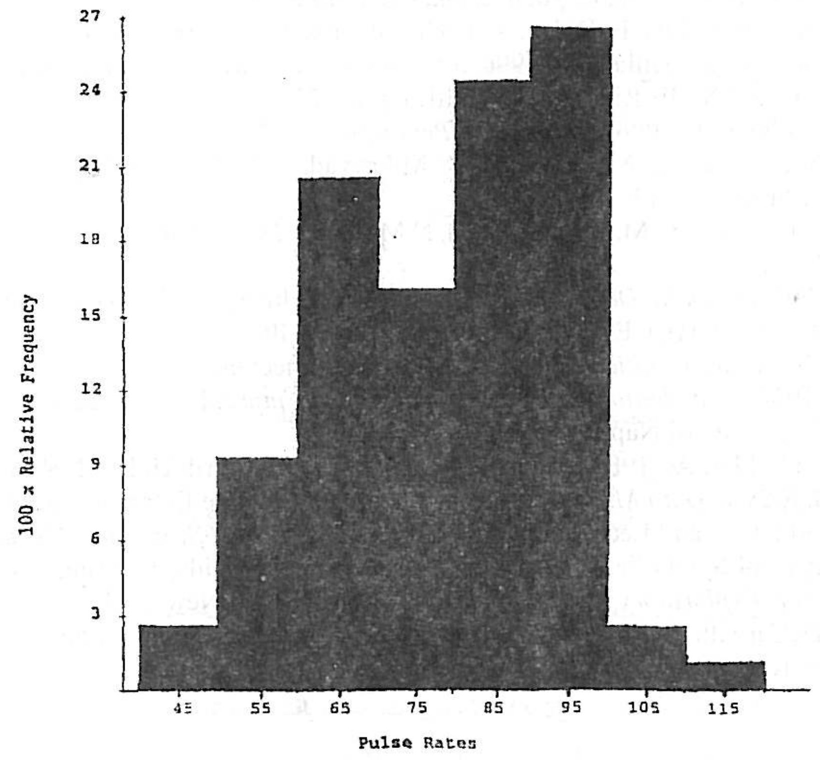
شکل ۱-الف بافتنگار گونه‌های باپرواز/ بی‌پرواز



شکل ۱-ب بافتنگار پرندگان باپرواز/ بی‌پرواز



شکل ۲- بافتنگار داده‌های راجو



مراجع

[1] Aristotle (IVBC), 1994. *Des Categories De l'Interpretation*, Organ, Librarie Philosophique Journal Vrin.
[2] Arnault, A. and Nicole, P., 1662. *La Logic ou l'Art Depensur*, Reprinted by Forman, Stuttgart (1965).
[3] Bertrand, P., 1995. *Structural Properties of Pyramidal Clustering In: Partitioning Data Sets*, (eds. I. Corc, P. Hansen and B. Julesz), American Mathematical Society, 19, pp. 352-353.
[4] Bertrand, P. and Goupil, F., 2000. *Descriptive Statistics for Symbolic Data*, In: Analysis of Symbolic Data (eds. H.H. Bock and E. Diday), Springer-Verlag, pp. 103-124.
[5] Billard, L. and Diday, E., 2001. *From the Statistics of Data to the Statistics of knowledge: Symbolic Data Analysis*, Manuscript.
[6] Bock, H.H. and Diday, E., 2000. *Symbolic Objects*, In: Analysis of Symbolic Data, (eds. H.H. Bock and E. Diday), Springer-Verlag, pp. 54-77.
[7] Choquet, G., 1954. *Theory of Capacities*, Annals de l'Institute Fourier, 5, pp. 131-295.
[8] Chouakria, A., 1998. *Extensions des methodes d'analyse factorielle a des donnee de type intervalle*, Ph.D. Thesis, University of Paris.
[9] DeCarvalho, F.A.T., 1994. *Proxinity Coefficients Boolean Symbolic Objects*, In: New Approaches in Classification and Data Analysis, (eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy). Springer-Verlag, pp. 387-394.
[10] DeCarvalho, F.A.T., 1995. *Histograms in Symbolic Data Analysis*, Annala of Operation Research. 55, pp.299-322.
[11] Diday, E., 1987. *Introduction a l'Approach Symbolique en Analyse des Donnees*, Premere Jouneles Symbolique-Namerique, CEREMADE, Universite Paris-Dauphine, pp. 21-56.

[12] Diday, E., (ed.), 1989. *Data Alalysis, Learning Symbolic and Numercal Knowledge*, Nova Science, Antibes.
[13] Diday, E., 1990. *Knowledge Repressentation and Symbolic Data Analysis*, In: Knowledge Data and Computer Assisted Decisions, (eds. M. Schader and W. Gaul). Springer-Verlag, pp. 17-34.
[14] Diday, E., 1995. *Probabilistic, Possibilistic and Belief Objects for Knowledge Analysis*, Annala of Operation Research, 55, pp. 227-276.
[15] Diday, E. and Emilion, R., 1996. *Capacities and Credibilities in Analysis of Probabilistic Objects*, In: Ordinal and Symbolic Data Analysis, (Eds. E. Diday, Y. Lechevallier and O. Opitz), Springer-Verlag, pp. 13-30.
[16] Diday, E., Emilion, R. and Hillali, Y., 1996. *Symbolic Data Analysis of Probablistic Objects by Capacities and Credibilities*, Atti Della XXXVIII Riunione Scientifica, pp. 5-22.
[17] Elder, J. and Pregibon, D., 1996. *A Statistical Perspective on Knowledge discover in Databases*, In: Advances in Knowledge Discovery and Data Mmining (eds. U.M.Fayyad, G. Piatetesky-Shapiro, P. Symth and R. Uthurusamy), AAAI Press, 83-113.
[18] Hand, D.J., Blunt, G., Kelly, M.G. and Adams, N.M., 2000. *Data Mining for Fun and Profit*, Statistical Science, 15, pp. 111-131.
[19] Raju, S.R.K., 1997. *Symbolic Data Analysis in Cardiology*, In: Symbolic Data Analysis and its Applications (eds. E. Diaay and K.C. Gowda), CEREMADE, Paris, pp. 245-249.
[20] Schafer, G., 1976. *A Mathematical Theory of Evidence*, Princeton.
[21] Schweizer, B., 1984. *Distributions are the Numbers of the Future*, In: Proceedings The Mathematics of Fuzzy Systems Meeting,. University of Naples, pp. 137-149.
[22] Schweizer, B. and Sklar, A., 1983. *Probabilitic Metric Spaces*, North Holland, New York.
[23] Siebs, A., 1998. *KESO: Data Mining and Statictics*, In: Knowledge Extraction from Statistical Data, *pp.* 1-13.
[24] Stephan, V., Herbial, G. and Lechevallier, Y., 2000. *Generation of Symbolic Objects from Relational Databases*, In: Analysis of Symbolic Data, (Eds. H. –H. Bock and E. Diday). Springer-Verlag, pp. 78-105.
[25] Tukey, J.W., 1977. *Exploratory Data Analysis*, Addison-Wesley, New York.
[26] Verbe, R. and DeCarvallo, F.A.T., 1998 *Dependence Rules Influence on Factorial Representation of Boolean Symbolic Objects*, In: Knowledge Extraction from Statistical Data, pp. 14-23.