دانشگاه حکیم سبزواری
انجمن علمی آمار

پارامتر

نشریه علمی-تخصصی پارامتر
نشریه انجمن علمی آمار دانشگاه حکیم سبزواری

شماره اول. بهمن ماه ۱۴۰۰

در این شماره می‌خوانید ؛

بسم الله الرحمن الرحیم

صاحب امتیاز: انجمن علمی آمار دانشگاه حکیم سبزواری

استاد مشاور: دکتر محمد بلبلیان قالیباف

مدیر مسئول: محمد صانعی

سردبیر: زهرا نصیری

ویراستار: زهرا نقدی، محدثه محمدی

صفحه آرا: فاطمه علی‌محمدی

طراح جلد: محمد مرادی

طراح لوگو: مرتضی صانعی

هیات تحریریه: وحیده حاجی حسنی، فاطمه رفیعی، محمد صانعی، امیررضا عباسی، فاطمه علی‌محمدی، فاطمه کلمیشی، زهرا نصیری

نشانی: خراسان رضوی – سبزوار – توحید شهر – دانشگاه حکیم سبزواری

فهرست مطالب

سرمقاله

سخن سردبیر

علم نیاز بشر است و دانستن ، برنامه ی او .

در واقع بشر امروز به خوبی این حقیقت را دریافته است که بدون علم ، زندگی اش یکنواخت و گاه سخت میشود . شاید اولین وآخرین دلیل آدمی برای در پی علم رفتن رفاه او باشد . اما نباید غافل از این واقعیت باشیم که علم و دانش سلاح امروز بشر است و در یک نگاه واقع بینانه اگر بگوییم امروزه علم سلاح هر کشوریست پس بیراه نگفته ایم چرا که این علم است که هم نشان از پویایی یک جامعه میدهد و هم ثمره ای از جنس بی نیازی را بدنبال دارد . در نشریه حاضر ، با اتکا بر توانایی وعزم راسخ اعضای هیئت تحریریه آن سعی شده است مطالب بروز و مفید جهت استفاده ی شما دانشجویان عزیز و اساتید گرامی فراهم شود . امید است توانسته باشیم گامی هر چند کوچک در جهت ارتقاء سطح علمی شما عزیزان برداشته باشیم انشاالله .

اکنون که از برکت الطاف الهی و با تکیه بر نیروی جوانی این فرصت بدست آمده است ، باید به خداوند و باور علمی خود تکیه کنیم و با تشکر از زحمات همه کسانی که نقشی هر چند کوچک در انتشار مجله داشته اند ، با کسب دانش و تجربه ، در جهت تقویت و استحکام محتوای مجله همت کرده و در جهت افزایش سطح کیفی و اعتبار آن بکوشیم .

زهرا نصیری

دانشجوی کارشناسی ارشد آمار دانشگاه حکیم سبزواری

تاریخچه‌ای کوتاه از آمار

فاطمه علی‌محمدی
کارشناس ارشد آمار دانشگاه علامه طباطبایی

واژه‌ی آمار برای اولین بار توسط یک پژوهش گر آلمانی گوتفرید آخنوال در اواسط قرن هجدهم به عنوان علم حکومت‌داری در مورد جمع آوری و استفاده از داده‌ها توسط دولت استفاده شد. در این جا تاریخچه‌ی کوتاه آمار را می‌بینیم.

کلمه‌ی آمار از کلمه‌ی لاتین "Status" یا کلمه‌ی ایتالیایی "Statistia" یا کلمه‌ی آلمانی "Statistik" یا کلمه‌ی فرانسوی "Statistique" گرفته شده است. به معنای یک دولت سیاسی، و در اصل به معنای اطلاعات مفید برای دولت است، مانند اطلاعات مربوط به اندازه‌ی جمعیت (انسان، حیوانات، محصولات و ...) و نیروهای مسلح.

به گفته‌ی آماردان پیشگام یول، کلمه‌ی آمار در ابتدایی‌ترین زمان در کتاب «عنصر دانش جهانی» نوشته‌ی بارون (۱۷۷۰) آمده است. در سال ۱۷۸۷ تعریف گسترده‌تری توسط E.A.W. زیمرمن در «بررسی سیاسی وضعیت کنونی اروپا» به کار برده شد.

این کتاب در سال ۱۷۹۷ در دایره المعارف بریتانیکا ظاهر شد و توسط سر جان سینکلر در بریتانیا در مجموعه‌ای از مجلدات منتشر شده بین سال‌های ۱۷۹۱ و ۱۷۹۹ استفاده شد و گزارشی آماری از اسکاتلند ارائه کرد. در قرن نوزدهم، کلمه‌ی آمار معنای گسترده‌تری پیدا کرد که داده‌های عددی تقریباً هر موضوعی را پوشش می‌داد و همچنین تفسیر داده‌ها را از طریق تحلیل مناسب پوشش می‌داد. این همه در مورد تاریخچه‌ی کوتاه آمار است. حال بیایید ببینیم که امروزه چگونه از آمار به معانی مختلف استفاده می‌شود.

اکنون آمار در معانی مختلفی به کار می‌رود.

- آمار (statistics) به "واقعیت‌های عددی که به صورت سیستماتیک در قالب جداول یا نمودارها و ... مرتب شده‌اند" اشاره دارد. به عنوان مثال آمار قیمت‌ها، تصادفات جاده‌ای، جنایات، تولدها، موسسات آموزشی و ... .

- کلمه‌ی آمار (statistics) کمیت‌های عددی محاسبه شده از مشاهدات نمونه است. یک کمیت منفرد محاسبه شده از مشاهدات نمونه‌ی آماری مانند میانگین نامیده می‌شود. در این‌جا آمار کلمه‌ی جمع است.

- واژه‌ی آمار (statistics) به عنوان رشته‌ای تعریف می‌شود که شامل رویه‌ها و تکنیک‌هایی است که برای جمع‌آوری، پردازش و تحلیل داده‌های عددی برای استنتاج و تصمیم‌گیری مناسب در موقعیت‌های عدم قطعیت استفاده می‌شود (عدم قطعیت به ناقص بودن اشاره دارد، به معنای ناآگاهی نیست). در این معنا کلمه‌ی آمار در معنای مفرد به کار می‌رود. این علم مبتنی بر تصمیم‌گیری بر روی داده‌های عددی است.

"We compute statistics from statistics by statistics"

ما اطلاعات آماری را از روی داده‌های آماری و آماره‌ها محاسبه می‌کنیم.

مرجع

https://itfeature.com/statistics/a-short-history-of-statistics

داده‌های بزرگ چیست؟

مقدمه، انواع، ویژگی‌ها، مثال

فاطمه علی‌محمدی

کارشناس ارشد آمار دانشگاه علامه طباطبایی

داده چیست؟

مقادیر، کاراکترها یا نمادهایی که عملیات بر روی آن‌ها توسط رایانه انجام می‌شود، که ممکن است به شکل سیگنال‌های الکتریکی ذخیره و ارسال شوند و بر روی رسانه‌های ضبط مغناطیسی، نوری یا مکانیکی ضبط شوند.

اکنون، بیایید تعریف داده‌های بزرگ را یاد بگیریم...

داده‌های بزرگ چیست؟

داده‌های بزرگ مجموعه‌ای از داده‌ها است که حجم آن بسیار زیاد است، اما با گذشت زمان به طور تصاعدی در حال رشد است. این یک داده با اندازه و پیچیدگی بسیار بزرگ است که هیچ یک از ابزارهای مدیریت داده سنتی نمی‌توانند آن را ذخیره یا پردازش کنند. داده‌های بزرگ نیز یک داده است اما با اندازه بزرگ.

در این آموزش تحلیل داده‌های بزرگ، یاد خواهید گرفت،

- داده چیست؟
- داده‌های بزرگ چیست؟
- نمونه‌ای از داده‌های بزرگ چیست؟
- انواع داده‌های بزرگ
- ویژگی‌های داده‌های بزرگ
- مزایای پردازش داده‌های بزرگ

نمونه‌ای از داده‌های بزرگ چیست؟

در زیر چند نمونه از داده‌های بزرگ آورده شده است...

بورس نیویورک نمونه‌ای از داده‌های بزرگ است که روزانه حدود یک ترابایت داده تجاری جدید تولید می‌کند.

رسانه‌های اجتماعی

این آمار نشان می‌دهد که روزانه بیش از ۵۰۰ ترابایت داده جدید به پایگاه داده‌های سایت رسانه اجتماعی فیس‌بوک وارد می‌شود. این داده‌ها عمدتاً از نظر آپلود عکس و ویدیو، تبادل پیام، گذاشتن نظرات و ... تولید می‌شوند.

یک موتور جت می‌تواند ۱۰+ ترابایت داده را در ۳۰ دقیقه زمان پرواز تولید کند. با هزاران پرواز در روز، تولید داده‌ها به تعداد زیادی پتابایت می‌رسد.

انواع داده‌های بزرگ

در ادامه انواع داده‌های بزرگ آورده شده است:

- ساختاریافته
- غیر ساختاریافته
- نیمه ساختاریافته

ساختار یافته

هر داده‌ای که می‌تواند در قالب ثابت ذخیره، دسترسی و پردازش شود به عنوان داده "ساختار یافته" نامیده می‌شود. در طول مدت زمان، استعدادهای علوم کامپیوتر در توسعه تکنیک‌های کار با این نوع داده‌ها (که قالب آن از قبل شناخته شده است) و همچنین استخراج ارزش از آن، موفقیت بیشتری کسب کرده است. با این حال، امروزه ما مشکلاتی را پیش‌بینی می‌کنیم که اندازه چنین داده‌هایی تا حد زیادی افزایش می‌یابد، اندازه‌های معمولی در تغیر چندین زتابایت قرار دارند.

غیر ساختاریافته

هر داده‌ای با شکل یا ساختار ناشناخته به عنوان داده‌های غیر ساختاریافته طبقه‌بندی می‌شود. علاوه بر بزرگ بودن اندازه، داده‌های بدون ساختار چالش‌های متعددی را در زمینه پردازش آن برای استخراج ارزش از آن ایجاد می‌کند. یک مثال معمولی از داده‌های غیر ساختاریافته، یک منبع داده ناهمگن است که حاوی ترکیبی از فایل‌های متنی ساده، تصاویر، ویدئوها و ... است. امروزه سازمان‌های امروزی داده‌های زیادی را در اختیار دارند، اما متأسفانه، نمی‌دانند چگونه از آن ارزش

استخراج کنند. این داده‌ها به صورت خام یا غیر ساختاریافته هستند.

داده‌های نیمه ساختاریافته

داده‌های نیمه ساختاریافته اطلاعاتی هستند که در یک پایگاه داده رابطه‌ای قرار نمی‌گیرند، اما دارای برخی ویژگی‌های سازمانی هستند که تحلیل آن را آسان‌تر می کند. با برخی از فرایندها، می‌توانید آن‌ها را در پایگاه داده رابطه ذخیره کنید (برای نوعی از داده‌های نیمه ساختاریافته می‌تواند بسیار سخت باشد)، اما نیمه ساختاریافته برای تسهیل فضا وجود دارد. مثال: داده‌های XML.

ویژگی‌های داده‌های بزرگ

داده‌های بزرگ را می‌توان با ویژگی‌های زیر توصیف کرد:

- حجم
- تنوع
- سرعت
- تغییرپذیری

(i) **حجم** – نام داده‌های بزرگ به خودی خود مربوط به اندازه‌ای است که بسیار زیاد است. اندازه داده‌ها نقش بسیار مهمی در تعیین ارزش داده‌ها دارد. همچنین، این‌که آیا یک داده خاص واقعاً می‌تواند به عنوان یک داده بزرگ در نظر گرفته شود یا خیر، به حجم داده‌ها بستگی دارد. از این رو، "حجم" یکی از ویژگی‌هایی است که باید در هنگام برخورد با راه حل های کلان داده در نظر گرفته شود.

(ii) **تنوع** – جنبه بعدی داده‌های بزرگ تنوع آن است.

تنوع به منابع ناهمگن و ماهیت داده‌ها، چه ساختاریافته و چه غیر ساختاریافته اشاره دارد. در روزهای قبل، صفحات گسترده و پایگاه‌های داده تنها منابع داده‌ای بودند که توسط اکثر برنامه‌ها مورد توجه قرار می‌گرفت. امروزه داده‌هایی به صورت ایمیل، عکس، فیلم، دستگاه‌های مانیتورینگ، پی‌دی‌اف، صوت و ... نیز در برنامه‌های آنالیز مورد توجه قرار می‌گیرند. این تنوع داده‌های غیر ساختاریافته مسائل خاصی را برای ذخیره‌سازی، استخراج و تحلیل داده‌ها ایجاد می‌کند.

(iii) **سرعت** - اصطلاح "سرعت" به سرعت تولید داده‌ها اشاره دارد. سرعت تولید و پردازش داده‌ها برای پاسخ‌گویی به نیازها، پتانسیل واقعی در داده‌ها را تعیین می‌کند.

سرعت داده‌های بزرگ با سرعتی که داده‌ها از منابعی مانند فرایندهای تجاری، گزارش برنامه‌ها، شبکه‌ها و سایت‌های رسانه‌های اجتماعی، حسگرها، دستگاه‌های تلفن همراه و ... وارد می‌شوند، سروکار دارد. جریان داده‌ها عظیم و پیوسته است.

(iv) **تغییرپذیری** - این به ناهماهنگی اشاره دارد که می‌تواند توسط داده‌ها در مواقعی نشان داده شود، بنابراین فرایند توانایی مدیریت داده‌ها را به طور موثر مختل می‌کند.

مزایای پردازش داده‌های بزرگ

توانایی پردازش داده‌های بزرگ مزایای متعددی را به همراه دارد، از جمله:

- کسب و کارها می‌توانند هنگام تصمیم‌گیری از هوش بیرونی استفاده کنند

دسترسی به داده‌های اجتماعی از موتورهای جستجو و سایت‌هایی مانند فیس‌بوک، توییتر، سازمان‌ها را قادر می‌سازد تا استراتژی‌های تجاری خود را تنظیم کنند.

- بهبود خدمات مشتری

سیستم‌های سنتی بازخورد مشتری با سیستم‌های جدیدی که با فناوری‌های داده‌های بزرگ طراحی شده‌اند جایگزین می‌شوند. در این سیستم‌های جدید، داده‌های بزرگ و فناوری‌های پردازش زبان طبیعی برای خواندن و ارزیابی پاسخ‌های مصرف‌کننده استفاده می‌شوند.

- شناسایی زود هنگام خطر برای محصول/ خدمات، در صورت وجود
- بهره‌وری عملیاتی بهتر

فناوری‌های داده‌های بزرگ را می‌توان برای ایجاد یک منطقه مرحله‌بندی یا منطقه فرود برای داده‌های جدید قبل از شناسایی این‌که چه داده‌هایی باید به انبار داده منتقل شوند، استفاده کرد. علاوه بر این، چنین ادغامی از فناوری‌های داده‌های بزرگ و انبار داده به سازمان کمک می‌کند تا داده‌هایی را که به‌ندرت به آن‌ها دسترسی پیدا می‌کند، بارگیری کند.

خلاصه

- تعریف داده‌های بزرگ: داده‌های بزرگ به معنی داده‌ای است که اندازه آن بزرگ است. داده‌های بزرگ اصطلاحی است که برای توصیف مجموعه‌ای از داده‌ها استفاده

می شود که از نظر اندازه بزرگ هستند و در عین حال با گذشت زمان به طور تصاعدی در حال رشد هستند.

- نمونه‌های تحلیل داده‌های بزرگ شامل بورس‌ها، سایت‌های رسانه‌های اجتماعی، موتورهای جت و ... است.
- داده‌های بزرگ می‌تواند ۱) ساختاریافته، ۲) غیر ساختاریافته، ۳) نیمه ساختاریافته باشد.
- حجم، تنوع، سرعت و تغییرپذیری چند ویژگی داده‌های بزرگ هستند.
- بهبود خدمات مشتری، بهره‌وری عملیاتی بهتر، تصمیم‌گیری بهتر از مزایای داده‌های بزرگ هستند.

مرجع‌ها

https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/

https://www.guru99.com/what-is-big-data.html

کنترل فرایند آماری چیست؟

فاطمه علی‌محمدی

کارشناس ارشد آمار دانشگاه علامه طباطبایی

کنترل فرایند آماری (SPC) به عنوان استفاده از تکنیک‌های آماری برای کنترل فرایند یا روش تولید تعریف می‌شود. ابزارها و رویه‌های SPC می‌توانند به شما در نظارت بر رفتار فرایند، کشف مسائل در سیستم‌های داخلی و یافتن راه‌حل برای مسائل تولید کمک کنند. کنترل فرایند آماری اغلب به جای کنترل کیفیت آماری (SQC) استفاده می‌شود.

- ابزارهای SPC
- SQC در مقابل SPC
- هفت ابزار کنترل کیفیت
- هفت ابزار تکمیلی
- تاریخچه‌ی SPC
- منابع SPC

ابزارهای SPC

یک ابزار محبوب SPC، نمودار کنترل است که در ابتدا توسط والتر شوهارت در اوایل دهه ۱۹۲۰ توسعه یافت. یک نمودار کنترلی به فرد کمک می‌کند تا داده‌ها را ثبت کند و به شما امکان می‌دهد زمانی که یک رویداد غیرعادی، مانند مشاهده بسیار

زیاد یا کم در مقایسه با عملکرد فرایند "معمولی" رخ می‌دهد، مشاهده کنید.

نمودارهای کنترل سعی می‌کنند بین دو نوع تنوع فرایند تمایز قائل شوند:

۱. تنوع علت رایج، که ذاتی فرایند است و همیشه وجود خواهد داشت
۲. تغییرات علت خاص، که از منابع خارجی ناشی می‌شود و نشان می‌دهد که فرایند خارج از کنترل آماری است

آزمایش‌های مختلف می‌توانند به تعیین زمان وقوع یک رویداد خارج از کنترل کمک کنند. با این حال، با آزمایش‌های بیش‌تر، احتمال هشدار نادرست نیز افزایش می‌یابد.

SQC در مقابل SPC

کنترل کیفیت آماری (SQC) به عنوان استفاده از ۱۴ ابزار آماری و تحلیلی (۷-QC و ۷-SUPP) برای نظارت بر خروجی‌های فرایند (متغیرهای وابسته) تعریف می‌شود. کنترل فرایند آماری (SPC) استفاده از همان ۱۴ ابزار برای کنترل ورودی‌های فرایند (متغیرهای مستقل) است. اگرچه هر دو اصطلاح اغلب به جای یکدیگر استفاده می‌شوند، SQC شامل نمونه‌گیری پذیرش است که در آن SPC چنین نیست.

هفت ابزار کنترل کیفیت (۷-QC)

در سال ۱۹۷۴، دکتر کائورو ایشیکاوا مجموعه‌ای از ابزارهای بهبود فرایند را در متن راهنمای خود برای کنترل کیفیت گرد هم آورد. در سرتاسر جهان به عنوان هفت ابزار کنترل کیفیت (۷-QC) شناخته می‌شوند، آن‌ها عبارتند از:

۱. نمودار علت و معلولی (همچنین نمودار ایشیکاوا یا نمودار استخوان ماهی نیز نامیده می‌شود)
۲. برگه کنترل
۳. نمودار کنترلی
٤. هیستوگرام
٥. نمودار پارتو
٦. نمودار پراکندگی
۷. طبقه‌بندی

هفت ابزار تکمیلی (۷-SUPP)

علاوه بر ابزارهای پایه ۷-QC، برخی ابزارهای کیفیت آماری اضافی نیز وجود دارد که به عنوان هفت ابزار تکمیلی (۷-SUPP) شناخته می‌شوند:

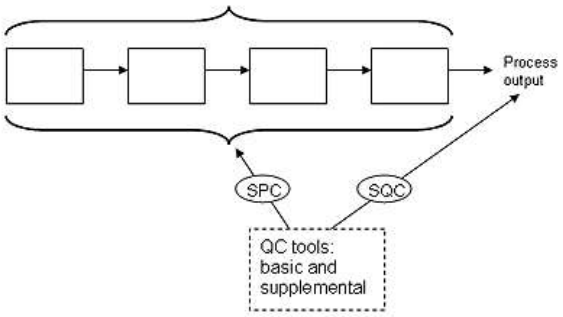
۱. طبقه‌بندی داده‌ها
۲. نقشه‌های نقص
۳. گزارش رویدادها
٤. نمودارهای جریان فرایند
٥. مراکز پیشرفت
٦. تصادفی‌سازی

۷. تعیین اندازه‌ی نمونه

رابطه‌ی بین کنترل کیفیت آماری و کنترل فرایند آماری

طراحی آزمایشات (DOE) و

تحلیل واریانس (AOV یا ANOVA)



تاریخچه‌ی SPC

افزایش قابل توجهی در استفاده از نمودارهای کنترلی در طول جنگ جهانی دوم در ایالات متحده برای اطمینان از کیفیت مهمات و سایر محصولات مهم استراتژیک رخ داد. استفاده از روش‌های SPC پس از جنگ تا حدودی کاهش یافت، اگرچه متعاقباً با تأثیرات زیادی در ژاپن مورد استفاده قرار گرفت و تا به امروز ادامه دارد. (برای اطلاعات بیش‌تر، تاریخچه‌ی کیفیت را ببینید.)

بسیاری از تکنیک‌های SPC در سال‌های اخیر توسط سازمان‌ها در سراسر جهان به‌کار گرفته شده‌اند، به‌ویژه به‌عنوان جزئی از طرح‌های بهبود کیفیت مانند شش سیگما. استفاده‌ی گسترده از روش‌های ترسیم نمودار کنترلی به‌وسیله‌ی بسته‌های نرم‌افزاری آماری و سیستم‌های پیچیده جمع‌آوری داده‌ها کمک زیادی کرده است.

ابزارهای اضافی نظارت بر فرایند عبارتند از:

- نمودارهای جمع انباشته (CUSUM): مجموع هر نقطه رسم شده نشان‌دهنده‌ی مجموع جبری مجموع قبلی و جدیدترین انحرافات از هدف است.
- نمودارهای میانگین متحرک وزنی نمایی (EWMA): هر نقطه‌ی نمودار نشان‌دهنده‌ی میانگین وزنی مقادیر فعلی و همه‌ی زیرگروه‌های قبلی است که وزن بیش‌تری به تاریخچه فرایند اخیر می‌دهد و وزن داده‌های قدیمی را کاهش می‌دهد.

منابع SPC

Peña-Rodríguez, M. E. (2013). *Statistical process control for the FDA-regulated industry*. Quality Press.

Ott, E. R., Schilling, E. G., Neubauer, D. V. (2005). Process Quality Control Troubleshooting and Interpretation of Data, Fourth Edition. Quality Press.

Gupta, B. C., & Walker, F. H. (2007). *Statistical Quality Control for the Six Sigma Green Belt*. ASQ Quality Press.

مرجع

https://asq.org/quality-resources/statistical-process-control

کاربرد طرح‌های آزمایشی در مدیریت مالی

وحیده حاجی حسنی، دانشجوی دکتری مدیریت و عضو باشگاه پژوهشگران جوان و نخبگان واحد قزوین

طرح آزمایش را می‌توان با نقشه‌های ساختمانی که یک مهندس ساختمان برای احداث یک خانه ارائه می‌دهد، مقایسه کرد. با وجود این‌که مهندس می‌تواند با به کارگیری خلاقیت خود، نقشه‌های متنوعی ارائه دهد، اما موظف است نیازهای اساسی ساکنین آیندهٔ ساختمان یا کارفرما را برآورده سازد. بدین منظور، او چندین طرح مختلف ارائه داده و از میان آن‌ها با توجه به تمام جوانب امر، طرح موردنظر انتخاب می‌شود. در طرح یک آزمایش، مهندس در نقش طراح و صاحب آینده ساختمان یا کارفرما در نقش آزمایش‌گر است و تصمیم نهایی را او دربارهٔ آزمایش اتخاذ می کند. بدون وجود یک طرح آزمایش مناسب نمی‌توان فرض‌های بالقوه سودمند را با درجه قابل قبولی از دقت آزمود. قبل از رد کردن فرضی مربوط به زمینه‌های تحقیقاتی باید ساختار آزمایش را مورد بررسی قرار داد و اطمینان حاصل کرد که آزمایش، یک آزمون واقعی از فرض را فراهم ساخته باشد. به‌معنای واقعی کلمه، آزمایش یک آزمون است. آزمایش طرح شده، یک آزمون یا دنباله ای از آزمون‌ها است که در آن‌ها تغییرات مورد نظر در متغیرهای ورودی فرایند یا سیستم اعمال می‌شوند به گونه‌ای که بتوانیم علت‌های تغییرات در پاسخ خروجی را مشاهده و مشخص کنیم(شاهکار و بزرگ نیا، ۱۳۸۱).

برخی کاربردهای طرح‌های آزمایشی

روش‌های طراحی آزمایشی، کاربرد وسیعی در بسیاری از رشته‌های علمی یافته است. حقیقت این است که می‌توان آزمایش‌گری را به منزله بخشی از فرایند علمی و یکی از راه‌های آموختن دربارهٔ چگونگی عمل سیستم‌ها یا فرایندها درنظر گرفت. به‌طورکلی، از طریق مجموعه ای از فعالیت‌ها که درجریان آن‌ها حدس‌هایی دربارهٔ فرایند می‌زنیم، آزمایش‌هایی را برای دستیابی به داده‌هایی دربارهٔ فرایند انجام می‌دهیم و سپس اطلاعات حاصل از آزمایش را برای نهادن حدس‌های جدید که به آزمایش‌های جدید می‌انجامد و به همین ترتیب ادامه می‌یابد، به کار می‌گیریم. طراحی آزمایشی، ابزاری با اهمیت در دنیای مهندسی برای بهبود عملکرد فرایند ساخت است. این طرح، کاربرد وسیعی نیز در پیدایش فرایندهای جدید دارد. استفاده از طراحی آزمایشی در این زمینه‌ها ممکن است به تولید محصولاتی بیانجامد که ساخت آسان‌تر، عملکرد میدانی بهتر، دوام بیشتر، هزینهٔ تولید کمتر و زمان طراحی و توسعه محصول کوتاه‌تری دارد(مونتگمری، ۱۳۹۵).

کاربرد طرح‌های آزمایشی در مباحث مدیریت مالی

طرح‌های آزمایشی در پژوهش‌های مرتبط با رشته‌های آمار، مهندسی، علوم اجتماعی، کشاورزی، روانشناسی، علوم تربیتی و علوم دامی کاربرد دارند و تاکنون پژوهش‌های فراوانی در داخل و خارج از کشور در زمینه‌های ذکر شده انجام گرفته است اما اولین بار، طرح‌های آزمایشی از جمله طرح مربع لاتین، طرح‌های عاملی تکرار شده و طرح کرت خرد شده توسط حاجی‌حسنی در مدیریت مالی به کار برده شد و نتایج این پژوهش‌ها در قالب طرح پژوهشی، مقاله‌های علمی پژوهشی و ارائه در کنفرانس، منتشر شد. از جمله پژوهش‌های انجام شده در این زمینه توسط محقق، پژوهشی است با عنوان "کاربرد مدل طرح‌های عاملی تکرارشده در عوامل مؤثر بر عملکرد صنایع" که بهار سال ۱۳۹۹در فصل‌نامه علمی پژوهش‌های راهبردی بودجه و مالی و در دانشگاه امام حسین(ع) منتشر شد. جامعهٔ آماری این پژوهش، شرکت‌هایی با فعالیت‌های جنبی واسطه‌گری‌های مالی و شرکت‌های حمل و نقل انبارداری و ارتباطات پذیرفته در بورس بودند و حجم نمونه برابر با جامعهٔ موجود در نظر گرفته شد. عوامل موردبررسی، نسبت‌های مالی شرکت‌های مورد بررسی، نوع صنعت و زمان بودند. روش پژوهش، توصیفی پیمایشی و بر اساس نوع هدف کاربردی درنظرگرفته شده و نتایج پژوهش نشان می دهد که اثر عامل صنعت و همچنین اثر متقابل زمان، سودآوری و صنعت، بر عملکرد شرکت‌ها معنی‌دار است و همچنین محقق مجری طرح پژوهشی با عنوان "بررسی عوامل موثر بر نقدینگی شرکت‌های ماشین‌آلات کشاورزی پذیرفته شده در بورس براساس مدل طرح کرت خرد شده" با همکاری دکتر یدالله رجایی در دانشگاه آزاد اسلامی واحد ابهر در سال ۱۳۹۳ انجام داده که نتایج در قالب یک مقاله ISI و یک مقالهٔ کنفراسی در دومین همایش ملی مدیریت کسب‌وکار در همدان به چاپ رسید. پژوهش دیگری نیز توسط محقق در این قالب با عنوان "کاربرد مدل طرح مربع لاتین در بررسی عوامل موثر بر کارایی صنعت انبوه سازی" در سال ۱۳۹۲ انجام گرفت که در فصلنامهٔ علمی دانش سرمایه گذاری که توسط انجمن مهندسی مالی منتشر می‌شود به چاپ رسید که در این پژوهش از یکی‌دیگر از طرح‌های آزمایشی با عنوان طرح مربع لاتین استفاده شد و این پژوهش در صنعت انبوه‌سازی و در ارتباط با کارایی شرکت‌ها مورد بررسی قرار گرفت. با وجود پژوهش‌های انجام یافته در زمینهٔ طرح‌های آزمایشی، به محققان، دانشجویان و اساتید گرامی پیشنهاد می‌شود تا مطالعات و پژوهش‌های بیشتری در این زمینه انجام داده تا شاهد کاربردهای دیگر طرح‌های آزمایشی در علوم مختلف با هدف حل مشکلات کشور باشیم.

لینک اسناد پژوهش‌های انجام گرفته به پیوست ارائه می‌شود:

کاربرد طرح کرت خردشده در بررسی عوامل موثر بر نقدینگی شرکت‌های ماشین آلات کشاورزی((civilica.com

کاربرد مدل طرح‌های عاملی تکرارشده در عوامل مؤثر بر عملکرد صنایع((ihu.ac.ir

Using Model of Split-Plot Design in the Study of Factors Affecting on Accepted in Stock Exchange Agricultural Machinery Companies Liquidity(researchgate.net)

مرجع‌ها

۱. حاجی حسنی، وحیده، رجایی، یدالله. ۱۳۹۹، کاربرد مدل طرح‌های عاملی تکرارشده در عوامل مؤثر بر عملکرد صنایع ، فصلنامه علمی پژوهش‌های راهبردی مالی و بودجه، دانشگاه امام حسین، تهران ، دوره ۱، شماره۱، ۱۴۶-۱۲۹.

۲. حاجی حسنی، وحیده، رجایی، یدالله. ۱۳۹۳، کاربرد طرح کرت خردشده در بررسی عوامل موثر بر نقدینگی شرکت‌های ماشین آلات کشاورزی، دومین همایش ملی مدیریت کسب و کار، همدان.

۳. حاجی حسنی، وحیده، رجایی، یدالله. ۱۳۹۳، بررسی عوامل مؤثر بر نقدینگی شرکت‌های ماشین آلات کشاورزی پذیرفته شده در بورس بر اساس مدل طرح کرت خردشده، طرح پژوهشی، باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی ابهر.

٤. حاجی حسنی، وحیده،۱۳۹۲،کاربرد مدل طرح مربع لاتین در بررسی عوامل موثر بر کارایی صنعت انبوه‌سازی، فصلنامه علمی پژوهشی دانش سرمایه گذاری، ویژه نامه کنفرانس مهندسی مالی تهران،۱۸۸-۱۷۹.

٥. شاهکار، غلامحسین. بزرگ نیا، ابوالقاسم. ۱۳۹۱، طرح آزمایشهای ۱، انتشارات دانشگاه پیام نور.

٦. مونتگمری، داگلاس سی.۱۳۹۵، طراحی و تحلیل آزمایش‌ها، کحالزاده، عباس، جباری، علیرضا، انتشارات مرکز نشر دانشگاهی.

7. Hajihassani.V., Yadollah Rajaei, 2015, Using Model of Split-Plot Design in the Study of Factors Affecting on Accepted in Stock Exchange Agricultural Machinery Companies Liquidity, Indian Journal of Science and Technology, Vol 8(S9), 530-533.

نگاهی به نرم‌افزار آماری R و SPSS

فاطمه رفیعی، دانشجوی کارشناسی ارشد آمار
دانشگاه حکیم سبزواری

نرم‌افزار آماری SPSS

SPSS مخفف عبارت Statistical Packge "For Social Science" بوده و به معنای بسته نرم‌افزاری آماری برای علوم اجتماعی می‌باشد که یکی از توانمندترین و جامع‌ترین نرم‌افزارهای آماری برای تحلیل داده است. این نرم‌افزار ساخت کمپانی IBM کشور آمریکا است که ۱۰۹ سال از تأسیس آن می‌گذرد. نرم‌افزار آماری SPSS با توجه به سادگی کار و سایر خصوصیات آن امروزه پرکاربردترین نرم‌افزار آماری در کشور محسوب می‌شود. به‌طور کلی زمانی به استفاده از نرم‌افزار SPSSبرای تحلیل پرسشنامه‌ها و براورد روش تحقیق روی می‌آوریم که استفاده از روش‌های آماری مد نظر باشد، یعنی بخواهیم داده‌ها را به‌صورت یک مجموعه خلاصه کرده و توصیف کنیم.

اجزای مهم و اصلی در SPSS

هنگام ورود به صفحهٔ اصلی SPSS با صفحه‌گستردهٔ Data Editor مواجه می‌شوید که در بالا و پایین آن دو نوار وجود دارد. نوار بالای صفحه شامل ابزارهای مختلف برای انجام دستورهای مختلف بر طبق نیاز است و نوار پایین با داشتن دو گزینهٔ Variable View و Data View محل ویرایش و وارد کردن اطلاعات پرسشنامه در SPSS است. نوار ابزار که در بالای صفحه قرار دارد دارای ۱۰ منوی اصلی برای تحلیل پرسشنامه است. این منوها عبارتند از : منوی فایل، منوی ویرایش، منوی نمایش، منوی داده‌ها، منوی تبدیل و انتقال، منوی آنالیز، منوی نمودارها و گراف‌ها، منوی امکانات، منوی ویندو و منوی کمک. هرکدام از این منوها برحسب نیاز کاربردهای مخصوص خودشآن‌را دارند.

چگونه اطلاعات پرسشنامه‌ها را درSPSS وارد کنیم؟

برای وارد کردن اطلاعات پرسشنامه‌ها در نرم‌افزار SPSS هر یک از پرسشنامه‌ها به‌عنوان یک نمونه و اطلاعات داخل آن متغیرهایی هستند که هر کدام دارای ماهیت‌های متفاوتی می‌باشند. بعد از باز کردن نرم‌افزار SPSS با صفحه‌گسترده‌ای روبرو می‌شویم که برای وارد کردن اطلاعات پرسشنامه باید از نوار پایین صفحه قسمت Variable View را انتخاب کرده و شروع به وارد کردن متغیرها کنیم و با توجه به هدف خواسته شده ماهیت متغیرها را تعیین کنیم.

پنجره Variable

در این پنجره می‌توانیم متغیرها را به نرم‌افزار معرفی کنیم. این پنجره شامل ستون‌هایی است که هر کدام یک ویژگی از متغیر را نشان می‌دهند.

Nam

در این قسمت می‌توانیم برای متغیر، اسم تعریف کنیم. نام متغیر می‌تواند شامل حروف انگلیسی یا فارسی، اعداد و یا کاراکترهای خاص مانند(. (نقطه) و _ (اندرلاین)) باشد. توجه داشته باشید که نام متغیر حتماً باید با یک حرف شروع شود و نمی‌تواند با عدد یا کاراکترهای خاص آغاز شود. همچنین نام متغیر نمی‌تواند شامل کاراکتر فاصله باشد.

Type

در این قسمت می‌توانیم نوع متغیر را به نرم‌افزار معرفی کنیم. برخی از انواع متغیر عبارتند از : تاریخ(Data)، رشته‌ای(String) و عددی(Numeric). قالباً متغیرهای مورد بررسی از نوع عددی هستند و حتی زمانی که متغیر کیفی است برای راحتی کار می‌توانیم سطوح آن‌را کدبندی کنیم و آن‌را به‌عنوان یک متغیر عددی در نظر بگیریم.

Width

در این ستون طول متغیر یعنی تعداد کاراکترهای اختصاص داده شده به متغیر تعیین می‌شود. نرم‌افزار به‌صورت خودکار این تعداد را برابر با ۸ در نظر می‌گیرد که این عدد حداکثر تعداد کاراکترهای متغیر را نشان می‌دهد. به جز مواردی که طول متغیر بیشتر از ۸ کاراکتر است بهتر است این عدد را تغییر ندهیم.

Decimals

در این ستون تعداد رقم‌های اعشار متغیرهای عددی تعیین می‌شود. نرم‌افزار به‌صورت خودکار دو رقم اعشار برای متغیرها در نظر می‌گیرد که می‌توانیم با کلیک کردن در خانهٔ مربوط به متغیر مورد نظر این تعداد را تغییر دهیم.

Lable

در این ستون می‌توانیم توضیحی برای متغیر اضافه کنیم. همان‌گونه که عنوان شد برای معرفی نام متغیر محدودیت‌هایی وجود دارد که در این قسمت چنین محدودیت‌هایی وجود ندارد.

Values

در صورتی که بخواهیم متغیر کیفی را به‌صورت عددی معرفی کنیم، می‌توانیم مقادیر عددی متناظر با سطوح متغیر کیفی را در این قسمت ثبت کنیم. برای این کار در پنجرهٔ مربوطه، کد عددی را در قسمت Value و سطوح متغیر کیفی را در قسمت Lable اضافه می‌کنیم و پس از معرفی هر سطح روی گزینهٔ Add کلیک می‌کنیم.

Missing

در این قسمت می‌توانیم نماد مورد استفاده برای نمایش مشاهدات گم شده را مشخص کنیم. نرم‌افزار به‌صورت خودکار از نماد "." استفاده می‌کند؛ اما

می‌توانیم نمادهای دیگری را نیز برای این کار در نظر بگیریم.

Columns

در این قسمت می‌توانیم عرض ستون مربوط به متغیر مورد نظر را تغییر دهیم.

Align

در این قسمت می‌توانیم نحوهٔ چینش داده‌ها در خانه‌های جدول را به‌صورت چپ چین، وسط چین و یا راست چین تغییر دهیم.

Measure

در این قسمت می‌توانیم مقیاس متغیر را به نرم‌افزار معرفی کنیم که در سه دسته اسمی(Nominal)، ترتیبی(Ordinal) و فاصله‌ای یا نسبتی (Scale) دسته بندی می‌شود.

Role

در این قسمت می‌توانیم نقش متغیر را به نرم‌افزار معرفی کنیم. بعد از وارد کردن متغیرها و اطلاعات مربوط به آن‌ها، از نوار پایین صفحه گزینهٔ Data Viewرا انتخاب می‌کنیم. در این قسمت تمام متغیرهای وارد شده در نرم‌افزار SPSS نشان داده می‌شوند؛ به این صورت که هر ستون از آن مربوط به یک متغیر بوده و هر سطر بیانگر پاسخ شرکت کنندگان در پرسشنامه است.

تحلیل داده‌ها در SPSS به دو صورت انجام می‌پذیرد:

۱) آمار توصیفی که شامل رسم جداول فراوانی، ترسیم نمودارها و محاسبهٔ شاخص‌های آماری است.

۲) آمار استنباطی که شامل آزمون فرضیه‌ها است.

آموزش تحلیل آماری پایان نامه با نرم‌افزار SPSS

برای انجام تحلیل‌های آماری با استفاده از نرم‌افزار آماری SPSS می‌توان به جدول‌های فراوانی نمودارهای آماری و مقایسه‌ای و مقایسه شاخص‌ها دست پیدا کرد. به دست آوردن قابلیت اطمینان پرسشنامه‌ها، آزمون‌های پارامتری مرتبط با میانگین جامعه، سنجش همبستگی بین متغیرها در رگرسیون، اکتشاف و خوشه بندی از جمله فرایندهایی است که در تحلیل آماری پایان نامه با استفاده از نرم‌افزار SPSS قابل انجام است. انجام هرکدام از آن‌ها با وارد کردن داده‌های مورد نظر و انتخاب گزینه‌های مناسب صورت می‌گیرد.

به‌طور مثال برای تحلیل عاملی اکتشافی انتخاب Analyze از نوار بالای صفحهٔ نرم‌افزار در پنجرهٔ باز شده گزینهٔ DeScriptive Statistics و سپس Explore را انتخاب می‌کنیم. این فرایند برای داده‌های کمی‌که از طرح‌های مستقل جمع‌آوری شده‌اند کاربرد دارد. همچنین این دستور امکان توصیف داده‌ها به تفکیک گروه‌های مختلف را فراهم می‌کند. خلاصه این تحلیل‌ها با استفاده از ابزاری مانند Plot و میانگین و فاصله‌های تعیین شده به دست می‌ایند که در نهایت خروجی تحلیل آماری

از پرسشنامه را در قالب جدول فراوانی و Plot نمایش می‌دهد.

چگونه در SPSS خروجی بگیریم؟

بعد از این‌که دستور مورد نظردر SPSS انجام شد، نتایج حاصل در پنجره‌ایبه نام "Output View" نمایش داده می‌شوند. این نمایش شامل نتایج و تحلیل تمام دستورات خواسته شده از SPSS است که به‌صورت جدول‌های فراوانی و انواع نمودارها نمایش داده می‌شود. در صورت نیاز می‌توان از طریق منوی File گزینهٔ Print را برای چاپ فیزیکی نتایج استفاده کرد. همچنین این امکان وجود دارد که برای استفاده از نتایج تحلیل‌ها در نرم‌افزار آماری SPSS، خروجی آن‌ها را به نرم‌افزارهای دیگری مانند Word یا Excel انتقال داد. یکی از ساده‌ترین روش‌ها برای انتقال خروجی از SPSS به Word و یا Excel، استفاده از دستور Copy و Paste است؛ به این صورت که با انتخاب خروجی مورد نظر در پنجره Output و استفاده از کلیک راست و دکمه Edit دستور Copy Special را انتخاب کرده و سپس در برنامه Wordیا Excel عملیات Paste را انجام داده و یا با استفاده از کلیدهای ترکیبی Ctrl+V خروجی را در محل مورد نظر پیاده می‌کنیم.

نرم‌افزار R

نرم‌افزار R یک زبان برنامه‌نویسی ریاضی شیءگرا می‌باشد که شباهت زیادی با S_Plus داشته و برای انجام محاسبات آماری طراحی شده است. پروژهٔ Rاز سال ۱۹۹۵ در گروه آمار دانشگاه اوکلند شروع شد و به سرعت توانست مخاطبان زیادی بدست آورد. در حال حاضر این زبان توسط یک تیم بین‌المللی نگهداری می‌شود.

آشنایی مقدماتی

ابتدا فایل اجرایی "win.exe۲.o.۳R" را از سایت R دانلود نموده و با دوبار کلیک روی آن نصب می‌کنیم. پس از پایان نصب، بر روی دسکتاپ کامپیوتر یک آیکون به شکل حرف R قرار می‌گیرد. اگر روی آن دوبار کلیک کنیم صفحه‌ای باز می‌شود که به آن صفحهٔ کنسول(R Console) می‌گویند. برای اجرای دستورها آن‌ها را در صفحهٔ کنسول تایپ و با رفتن به خط بعدی آن‌ها را اجرا می‌کنیم. دستورات پس از اجرا قابل ویرایش نیستند؛ برای ویرایش دستورها می‌توانیم قبل از اجرا، آن‌ها را در صفحهٔ Script بنویسیم و آن‌ها را با دستورRun Line و یا Ctrl+R اجرا کنیم. پس از اجرا در صفحهٔ Script آن‌ها را ویرایش می‌کنیم و تغییرات مورد نظر را اعمال کرده و سپس دوباره اجرا می‌کنیم. همچنین برای پاک کردن صفحهٔ

کنسول از دستور Clear Console یا Ctrl+L، برای ایجاد صفحهٔ Script جدید از دستور New Script یا Ctrl+N و برای پیدا کردن و باز کردن برنامهٔ نوشته شده از دستور Open Script یاCtrl+o استفاده می‌کنیم. تعداد زیادی بستهٔ نرم‌افزاری یا پکیج(Package) می‌تواند روی نرم‌افزار R نصب شود که زمینه‌های مختلف آماری را در بر می‌گیرد و دارای توابع پیش‌ساخته بسیاری است که از این حیث به این نرم‌افزار قدرت شگرفی بخشیده است.

نصب و فراخوانی پکیج

برای نصب و فراخوانی پکیج‌ها به ترتیب می‌توان از دستورهای زیر استفاده کرد:

```
install.packages(" ")
library( )
```

توجه شود که برای به کارگیری پکیج مورد نظر یک بار نصب کردن آن کافی است ولی برای هربار استفاده باید آن‌را فراخوانی کرد.

دایرکتوری

مکانی در حافظهٔ کامپیوتر است که متغیرهایی که در R تعریف می‌کنیم در آن ذخیره می‌شوند. پوشهٔ دایرکتوری یا پوشهٔ مدیریت، پوشه‌ای است که به‌عنوان پیش‌فرض برای باز کردن یا ذخیره کردن برنامه‌ها انتخاب می‌کنیم. البته پوشه دایرکتوری قابل تغییر است. برای تغییر پوشهٔ دایرکتوری از منوی File به گزینهٔ Change Directory بروید و پوشهٔ مورد نظر را انتخاب کنید. این پوشه هر بار پس از بستن نرم‌افزار R و باز کردن آن به حالت پیش‌فرض برمی‌گردد. برای انتخاب پوشه دایرکتوری و ثبت دائمی یک پوشه به‌عنوان پوشهٔ پیش‌فرض، پس از انتخاب پوشهٔ مورد نظر از منوی File به گزینهٔ Workspace بروید و فضای کار خود را با فرمت RData ذخیره کنید؛ سپس آن‌را در دسترس قرار دهید و هر بار برای باز کردن نرم‌افزارR آن‌را اجرا کنید.

روش‌های آماری در ستاره‌شناسی

امیررضا عباسی، دانشجوی کارشناسی آمار دانشگاه اصفهان

در این‌جا مروری خواهیم داشت به انواع داده‌ها و روش‌های آماری که به‌صورت متداول در زمینه ستاره‌شناسی استفاده می‌شود. هدف از ارائه چنین شیوه‌ای، این است که مقدمه بر استفادهٔ کاربردی از علم آمار در ستاره‌شناسی برای آماردانان و دانشمندان علوم کامپیوتر ارائه شود. تمرکز ما بر بررسی ماهیت پیچیده و سلسله مراتبی ستاره‌شناسی خواهد بود و بسیاری از مشکلات استنتاجی این علم را برجسته می‌کنیم و به چالش‌های موجود در این راه پاسخ خواهیم داد.

۱. مقدمه

ستاره‌شناسی تاریخی طولانی در بهره بردن از مشاهدات رصدی برای تخمین متغیرها و تعیین کمیت عدم قطعیت در مدل‌های فیزیکی دارد. مسائل ستاره‌شناسی باعث پیشرفته شدن بسیاری از تکنیک‌های آماری از براورد حداقل مربعات کلاسیک تا روش‌های نوین مانند نمونه‌گیری تو در تو و انطباق آن‌ها با این علم شده‌است.

پیشرفت‌های اواخر قرن بیستم در جمع‌اوری داده‌ها، مانند خودکار کردن تلسکوپ‌ها و استفاده از دوربین‌های "CDD" منجر به افزایش چشم‌گیر اندازه و پیچیدگی داده‌ها شد که نتیجه آن، استفاده بیش‌تر اخترشناسان از روش‌های آماری و انطباق و توسعه آن در ستاره‌شناسی بود. ستاره‌شناسان از این مجموعه داده‌ها برای گسترهٔ گوناگونی از اهداف علمی، مانند مدل‌سازی شکل‌گیری کهکشان‌ها، یافتن سیارات شبیه به زمین، تخمین انبساط متریک فضا و طبقه‌بندی گذرا استفاده می‌کنند.

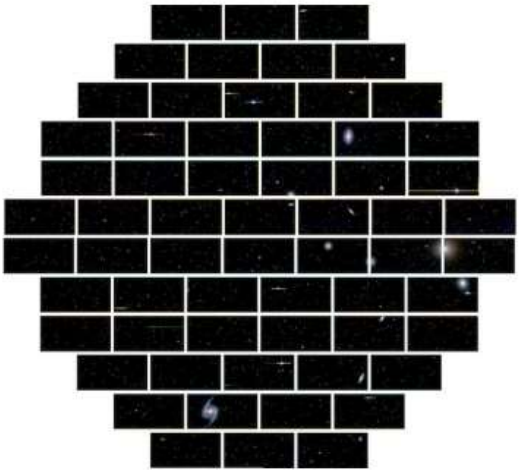
در این مقاله قصد داریم انواع داده‌های رایج و روش‌شناسی آماری که در حال حاضر در ستاره‌شناسی مورد استفاده قرار می‌گیرد را مورد بررسی قرار دهیم. این کار به منظور آشنایی بیش‌تر کاربردهای نجومی برای آماردانان انجام می‌شود. لازم به ذکر است که مجموعهٔ ارائه شده بسیار مختصر بوده و خوانندگان برای مطالعه بیش‌تر می‌توانند به ارجاعات پایان مقاله مراجعه کنند تا از دیدگاه‌های تاریخی و روش‌شناختی آمار در ستاره‌شناسی آگاه شوند. در بخش ۲ ما سه نوع داده از داده‌های ستاره‌شناسی رایج را بررسی می‌کنیم: تصاویر، طیف‌ها و سری‌های زمانی. در بخش ۳ برخی از روش‌های آماری مورد استفاده در نجوم را

مورد بحث قرار می‌دهیم. بسیاری از این روش‌ها هم‌اکنون در کارگاه‌های تحقیقاتی آمار و علوم کامپیوتر درحال توسعه هستند. در بخش ۴ با شرح یک چالش ستاره‌شناسی مرتبط با آمار، نقشه برداری از حلقهٔ نور کهکشان راه شیری با ستاره‌های " RR Lyrae" و ابزارهای آماری مختلف لازم برای پرداختن به این کار را شرح می‌دهیم.

۲. انواع داده‌های ستاره‌شناسی

الف. داده‌های تصویری

تلسکوپ‌ها از آسمان شب عکس می‌گیرند. نمایهٔ S1 تصویری را نشان می‌دهد که توسط دوربین انرژی تاریک(DECam) به‌عنوان بخشی از بررسی انرژی تاریک(DES) گرفته شده است. DES تقریباً ۴۰۰ تصویر یک گیگابایتی در هر شب می‌گیرد. تصاویر اخترشناسی معمولاً با یک فیلتر فوتومتریک گرفته می‌شوند که طول موج‌های نور خاصی را مسدود می‌کند. خط لوله فوتومتریک اجسام را در تصاویر شناسایی کرده و روشنایی آن‌ها را تخمین می‌زند. این خطوط لوله حاوی بسیاری از ابزارهای آماری هستند مانند الگوریتم‌های یادگیری ماشین(به بخش ث نگاه کنید) و مدل‌های سلسله مراتبی(به بخش پ توجه کنید). خط لوله، کاتالوگی حاوی موقعیت اجرام، درخشندگی‌ها و طبقه‌بندی‌ها(ستاره، کهکشان، سیارک و غیره) است. مطالعه و مدل‌سازی داده‌های کاتالوگ معمولاً بسیار ساده‌تر از داده‌های تصویر خام است، بنابراین بیش‌تر تحلیل‌های بعدی بر روی آن‌ها انجام می‌شود.

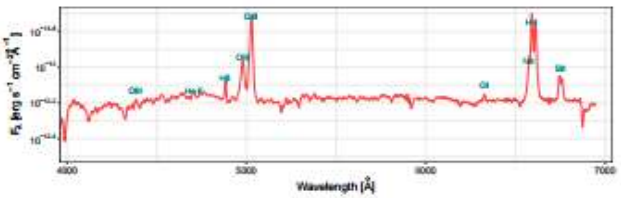


نمایه ۱. تصویری از آسمان شب که توسط DECam گرفته شده است. خطوط سفید شکاف میان CCDهای آشکارساز هستند. شناسایی، طبقه‌بندی و تخمین روشنایی اجرام در تصاویر یک چالش آماری بزرگ در نجوم است.

ب. داده‌های طیفی

یک طیف، شدت نور را در طول موج‌های مختلف نشان می‌دهد و اطلاعات بسیار بیش‌تری را نسبت به چیزی که به‌صورت مستقیم از داده‌های تصویری قابل استنباط است، ارائه می‌دهد. نمایهٔ S2 طیف کهکشان مسی‌یر ۷۷(Messier۷۷) را نشان می‌دهد که به‌صورت یک مارپیچ میله‌ای در صورت فلکی سیتوس قرار دارد. طیف‌ها حاوی اطلاعاتی درباره برخی از مهم‌ترین خواص فیزیکی اجرام نجومی مانند دما و ترکیب شیمیایی هستند. علاوه بر این،

جابه‌جایی ویژگی‌های طیفی به سمت طول موج‌های بلند(معروف به انتقال قرمز) ممکن است برای تخمین فاصله جسم مورد استفاده قرار گیرد، در نتیجه ابزار ارزشمندی برای درک کامل جهان را ارائه می‌دهد. چندین بررسی نجومی وجود دارد که اطلاعات طیفی را مانند آزمایش سرعت شعاعی جمع‌اوری می‌کنند؛ یکی از بزرگ‌ترین بررسی‌های طیف‌سنجی ستاره‌های کهکشان راه شیری است که در اختیار همگان قرار گرفته است. این مطالعه، بررسی صرف و تاریخچه پیشین کهشکان راه شیری از طریق مشاهدات طیف‌سنجی ستاره‌ای و پایگاه‌های اطلاعاتی نجومی را امکان‌پذیر می‌سازد. بررسی "SDSSIV MaNGA 10000" اندازه‌گیری طیفی را بر روی کهکشان‌های نزدیک جمع‌اوری می‌کند و امکان ساخت نقشه‌های دو

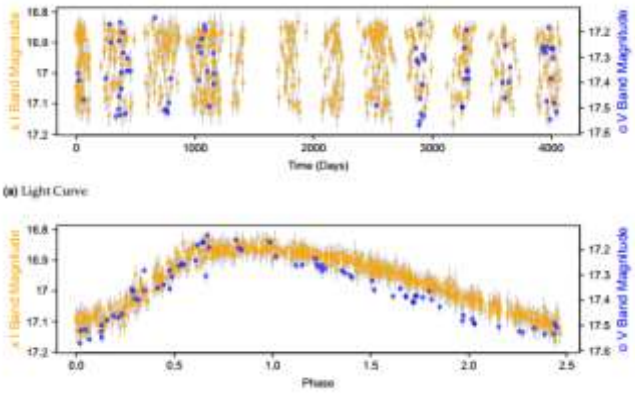


نمایه ۲. نمونه‌ای از طیف کهکشان از مسی‌یر ۷۷

پ. سری‌های زمانی و داده‌های عملکردی

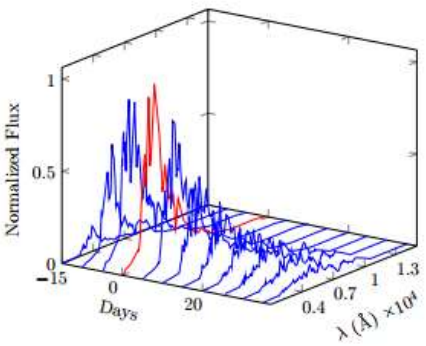
تصاویر و طیف‌ها دو شکل رایج از داده‌های ستاره‌شناسی خام را نشان می‌دهند که همراه با اطلاعات فوتومتریک، شار یکپارچه از طریق یک فیلتر معین، مبنایی برای استخراج چندین نوع داده فراهم می‌کند. برای نمونه، روشنایی بسیاری از منابع نور بر حسب زمان متفاوت است. بررسی‌های اخترشناسی که در طول زمان، یک منطقه از آسمان را به‌صورت پیوسته تصویربرداری می‌کنند، یک سری زمانی یا منحنی نوری برای هر جسم ایجاد می‌کنند که امکان تجزیه و تحلیل تغییرات روشنایی زمانی را فراهم می‌سازد.

نمایه الف-S3 یک سری زمانی برای یک ستارهٔ مشاهده شده توسط آزمایش عدسی گرانشی نوری(OGLE) را نشان می‌دهد. داده‌ها در دو فیلتر جمع‌اوری شده که با صلیب‌های نارنجی و دایره‌های آبی در طول تقریباً ۱۰ سال نشان داده شده است. آهنگ یا فاصله زمانی میان رصدها نامنظم است که امری معمول در داده‌های نجومی است. OGLE تقریبا ۴۰۰۰۰۰ منحنی نوری را جمع‌اوری کرده است که همه آن‌ها در دسترس عموم هستند. چالش‌های آماری با این داده‌ها شامل مدل‌سازی تغییرات شکل و طبقه‌بندی منابع بر اساس دلیل اخترفیزیکی تغییر روشنایی است. به‌عنوان مثال، ستاره‌ها در نمایه الف-S3 در طول زمان به شکل متناوب درخشندگی متفاوتی دارند. از این داده‌ها می‌توان برای تخمین یک دورهٔ بزرگ‌تر و نموداری گسترده‌تر مانند نمایه ب-S3 استفاده کرد.



نمایهٔ S3. منحنی نور یک ستاره متغیر مشاهده شده توسط OGLE. مدل‌هایی از سری‌های زمانی و ادبیات تحلیل داده‌های عملکرد اغلب برای مطالعه این اشیاء استفاده می‌شوند. در شکل پایین(ب) داده‌ها به‌صورت منحنی نمایش داده شده‌اند و در تصویر بالا(الف) از نور تولیدی متناوب توسط یک ستاره استفاده شده است. از داده‌های الف می‌توان یک منحنی نور ۲.۴۸ روزه را تخمین زد.

روش‌های آماری برای تخمین این دوره‌های گسترده دراین نوع داده‌ها در حال توسعه هستند. برای مقایسه، در نمایهٔ S4 یک طیف ابر نواختر را به‌عنوان تابعی از زمان نشان می‌دهیم که دورهٔ حداکثر روشنایی با رنگ قرمز مشخص شده است.



نمایهٔ S4. نمونه‌ای از یک طیف ابر نواختر به‌عنوان تابعی از روزهای پس از حداکثر روشنایی

۳. روش‌شناسی آماری در ستاره‌شناسی

ستاره شناسان از طیف گسترده‌ای از روش‌های آماری برای تجزیه و تحلیل مجموعه داده‌های پیچیده خود استفاده می‌کنند. در ادامه، چندین زمینه از روش‌شناسی آماری را با کاربردهای اخیر در نجوم مورد بحث قرار می‌دهیم.

الف. مدل‌های اندازه‌گیری

به‌طور معمول فرض بر برابری واریانس‌ها است(یکسان بودن واریانس خطاها). اغلب فرض می‌شود که متغیرهای پیش‌بینی کننده(مستقل) در مدل‌های رگرسیون بدون خطا اندازه‌گیری می‌شوند. با این‌حال در نجوم، خطاهای برابر نبودن واریانس‌ها معمول هستند. علاوه بر این، طبیعی است که تخمین‌هایی از واریانس‌های خطای اندازه‌گیری از طریق مدل‌سازی عدم قطعیت‌های ذاتی روش تشخیص در دسترس باشند. در مواردی که خطای اندازه‌گیری بزرگ است، مدل‌های صریح خطا در متغیرها برای جلوگیری از تخمین‌های اریب، به‌ویژه در مدل‌های رگرسیونی ضروری هستند. این مدل‌ها معمولاً دارای یک ساختار سلسله مراتبی هستند که در آن مقادیر پیش‌بینی کننده واقعی به‌عنوان متغیر در نظر گرفته می‌شوند. روش‌هایی که معمولاً در نجوم برای حل این مشکل استفاده می‌شود، شامل خطاهای همبسته دومتغیره و مدل پراکندگی ذاتی و

مدل‌های بیزی سلسله مراتبی است. توسعهٔ یک مدل مخصول گاوسی برای تخمین چگالی از مشاهدات مشروط به خطای اندازه‌گیری، به حساب آوردن عدم قطعیت‌های شار در طبقه‌بندی احتمالی اختر روش‌ها و استفاده از یک مدل بیزی سلسله مراتبی برای رسیدگی به عدم قطعیت‌های اندازه‌گیری گسسته در یک مدل دو جمله‌ای منفی برای بررسی جمعیت خوشه‌های کروی در کهکشان‌ها نمونه‌هایی از کاربردهای بعدی آن هستند.

ب. تحلیل بقا

بررسی‌های نجومی، از نظر ساخت، توانایی به‌دست آوردن نمونه‌های بی‌طرفانه از جمعیت اجرام را ندارند. بررسی‌ها اغلب با محدودتی انجام می‌شوند، یعنی اجسام روشن‌تر به احتمال زیاد شناسایی می‌شوند. به این دلیل محدودیت تلسکوپ منجر به برش داده‌ها می‌شود که در ستاره‌شناسی به آن سوگیری مالم‌کوییست می‌گویند. در موقعیت‌های دیگر ما می‌دانیم که یک شئ وجود دارد، اما برخی ویژگی‌های آن به‌قدری ضعیف اند که شناسایی آن برای ما ناممکن است و همین باعث سانسور داده می‌شود. در آمار، راه‌حل چنین مشکلاتی در دامنهٔ کلی تحلیل بقا بررسی می‌شوند. در نجوم، به سانسور و کوتاه کردن، اثرات انتخاب گفته می‌شود. چالش‌های تحلیل بقا در نجوم ممکن است شامل داده‌های چندمتغیره، تخمین چگالی ناپارامتری با برش یا اثرات انتخاب در مدل رگرسیونی باشد.

پ. مدل‌های بیزی و محاسبات

استفاده از روش بیزی به شکل قابل توجهی رشد پیدا کرده است. حوزه‌های فعال تحقیقات بیزی شامل مدل‌های سلسله‌مراتبی، نمونه‌گیری‌های پسین و مدل‌هایی برای انواع داده‌های پیچیده مانند تصاویر و توابع است. در مدل‌های بیزی سلسله مراتبی(HBM) برای مشکلاتی که در آن پارامترها و جمعیت شئ منفرد استفاده می‌شود پارامترها ناشناخته هستند. مدل‌های بیزی سلسله مراتبی برای بسیاری از انواع داده‌ها در ستاره‌شناسی استفاده می‌شوند؛ برای مثال شناسایی و توصیف کهکشان‌ها در تصاویر ستاره‌شناسی با استفاده از مدل‌های بیزی سلسله مراتبی با استنتاج متغیر برای تقریب خلفی، مدل‌سازی منحنی‌های نور ابرنواخترها و متناسب کردن داده‌های پرتوهای کیهانی. محاسبات بیزی تقریبی(ABC) با شبیه‌سازی مجموعهٔ داده‌ها و مقایسهٔ فاصلهٔ بین داده‌های شبیه‌سازی شده و داده‌های واقعی، از ارزیابی احتمال‌گران محاسباتی جلوگیری می‌کند. محاسبات بیزی در نجوم برای تقریب و استنتاج متغیرهای کیهانی و کاوش در تکامل کهکشان‌ها استفاده می‌شود. استفاده روزافزون از محاسبات بیزی تقریبی منجر به توسعه بسته‌های نرم‌افزاری مانند "Cosmoabc"،

نمونه‌بردار محاسبات بیزی تقریبی از طریق جمعیت مونته‌کارلو برای کابردهای عمومی نجومی شده است. شاخه‌ها و انواع مختلفی از نمونه‌گیرهای زنجیرهٔ مارکوف مونت کارلو توسط اخترشناسان توسعه داده شده است. از جمله اجرای یک نمونهٔ گروهی غیر متغیر وابسته و نمونه‌برداری تو در توی پراکنده، به‌عنوان مدل تعمیم‌یافتهٔ نمونه‌گیری تو در تو.

ت. مدل‌های تعمیم‌یافتهٔ خطی

مدل رگرسیون خطی تعمیم‌یافته وابسته بر تعدادی مفروضات توزیعی است، در زمانی‌که داده‌ها از توزیع‌های خانوادهٔ نمایی غیرگوسی به‌دست می‌ایند. مدل‌های خطی تعمیم یافته(GLM)، از طریق یک تابع پیوند، یک رابطه خطی بین متغیر پاسخ y و مجموعه‌ای از پیش‌بینی‌کننده‌های x را براورد می‌کنند. مسائل زیادی در ستاره‌شناسی نیازمند استفاده از مدل‌های تعمیم‌یافته و تعمیم‌های آن‌ها است. مانند مدل‌سازی کسر کهکشان‌ها از نظر محیطی(برنولی)، جمعیت خوشه‌های کروی به‌عنوان تابعی از ویژگی‌های کهکشان میزبان(دو جمله‌ای منفی) و فاصله کهکشان‌ها به‌عنوان تابعی از رنگ آن‌ها(گاما).

ث. یادگیری ماشین

برای بسیاری از مسائل ستاره‌شناسی، پیش‌بینی و تشخیص الگو مهم‌تر از تخمین پارامتر است. درحال حاضر برای حل این مشکلات، به‌طور گسترده‌ای از فناوری یادگیری ماشین(ML) استفاده می‌شود. درحالی‌که روش‌های یادگیری ماشین نامتداول گاهی اوقات برای حل مسائلی کاربرد دارند، اما مسائل نجومی چالش‌های خاص خودشان را دارند. چالش‌هایی نظیر مجموعه آزمایش‌های اریب، استخراج ویژگی‌های فشردهٔ محاسباتی یا طبقه‌بندی زمان واقعی که ممکن است با استفاده از روش‌های معمول قابل انجام نباشند. در ادامه، برخی از این چالش‌ها را شرح می‌دهیم.

یادگیری ماشین در طبقه‌بندی منحنی‌های نور منبع متغیر کاربرد گسترده‌ای داشته است(برای تعریف این موضوع به بخش پ مراجعه کنید). اما ستاره‌شناسان بیش‌تر از تخمین پارامترهای منبع، علاقه دارند کلاس منبع را طبقه‌بندی کنند. از آن‌جایی که منحنی‌های نور به شکل توابع ظاهر می‌شوند، این موضوع یک مشکل جدی برای طبقه‌بندی کردن به وجود می‌اورد. یک رویکرد رایج برای حل این مشکل، ساختن مجموعهٔ آموزشی از اشیاء کلاس شناخته شده(اغلب با استفاده از سطح طبقه‌بندی انسانی)، استخراج ویژگی‌های این اشیاء و سپس آموزش یک

طبقه‌بندی کنندهٔ یادگیری ماشین بر روی این داده‌ها است. سپس طبقه‌بندی کننده می‌تواند برای طبقه‌بندی اجرام جدید در بررسی‌های دیگر مورد استفاده قرار گیرد. ابزارهای یادگیری ماشین برای چند مسئلهٔ نجومی دیگر از جمله شناسایی منابع در تصاویر، خوشه‌بندی داده‌های طیفی و تخمین جابه‌جایی نوری نیز مورد استفاده قرار می‌گیرند. یک مسئلهٔ مهم که معمولاً نادیده گرفته می‌شود، عدم وجود نمایندگی بین نمونه‌های طیف‌سنجی و فوتومتریک است. تجربه نشان داده است که معیارهای عملکرد اعتبارسنجی متقابل در این وضعیت گمراه‌کننده‌اند. عدم تطابق میان نمونه‌های آموزشی و آزمایشی منحصر به مسائل نجومی نیست. روش‌شناسی توسعه‌یافته توسط جامعهٔ یادگیری ماشین برای رسیدگی به این چالش در چندین مشکل ستاره‌شناسی از جمله تطبیق دامنه، یادگیری فعال و ترکیبی از هر دو در دامنه تکنیک‌های یادگیری تطبیقی استفاده شده است. با این‌حال، چالش‌ها، از جمله گنجاندن خطای اندازه‌گیری ویژگی، داده‌های گم شده، سانسور و برش در الگوریتم‌های یادگیری ماشین همچنان باقی است.

ج. تجسم اطلاعات

روش‌های تجسم از سیستم بینایی انسان برای بهینه‌سازی بینش بصری در ساختار داده استفاده می‌کنند. درحالی‌که نقش تجسم به پایهٔ تحلیل نجومی تعلق دارد، نمونه‌های جدید برای تجسم داده‌های چند بعدی هنوز به‌صورت کامل مورد استفاده قرار نگرفته‌اند. الگوها و همبستگی‌های غیر پیش پا افتاده‌ای که ممکن است در داده‌های مبتنی بر جدول شناسایی نشوند، در صورت استفاده از ابزارهای مناسب، می‌توانند آشکار شوند. از جمله روش‌هایی که برای تسهیل اکتشاف داده‌های نجومی چند متغیره ایجاد شده‌اند می‌توان به نمودار درختی فیلوژنتیک، نمودارها، آکوردها و نمودارهای ستاره دریایی اشاره کرد. نمایهٔ S5 مجموعه داده‌ای از شبیه‌سازی کیهان شناسی با روش "N-body/Hyedro" را نشان می‌دهد که با سه تکنیک متفاوت مجسم شده است.

نمایهٔ S5. سه مجسم‌سازی(از چپ به راست): نمودار، کلادوگرام و نمودار وتر) از یک کاتالوگ کهکشان که توسط روش N-body/Hydro شبیه‌سازی شده است.

٤. چالش‌های استنتاج پیچیده در نجوم؛ یک مثال

فرایند تبدیل داده‌ها به کشف دانش علمی معمولاً نیازمند استفاده از بسیاری از ابزارهای آماری و عموماً به شکل نواورانه است. برخی از چالش برانگیزترین سوالات آماری که در نجوم مطرح می‌شود، مربوط به چگونگی ادغام این ابزارها در خط لوله تجزیه و تحلیل داده است که استنتاج‌های آماری معتبر را درصورتی‌که که از نظر محاسباتی قابل انجام باشد، امکان‌پذیر می‌سازد. به‌عنوان مثال، چالش نقشه‌برداری از حلقهٔ نور کهکشان راه شیری را در نظر بگیرید. منطقه‌ای از فضا که کهکشان ما را احاطه کرده است. این مشکل اخیراً توجه بسیاری را به‌خود جلب کرده است. ستاره‌شناسان مایل هستند نقشه‌هایی از مکان ستارگان در حلقهٔ نور تهیه کنند و ساختارهایی مانند مجموعه‌ای ستارگان گرانشی را شناسایی کنند. این موضوع اثرات مهمی در چارچوب کنونی درک ما از چگونگی تولد و توسعه جهان و شناخت کیهانی ماده تاریک سرد(LCDM) دارد. ایجاد نقشه حلقهٔ نور دشوار است چراکه تعیین فاصله تا بیش‌تر ستارگان غیرممکن است. با این‌حال، ما می‌توانیم فاصلهٔ یک زیرمجموعهٔ کوچک از ستارگان به نام " RR Lyrae"(و به اختصار RRL) را تخمین بزنیم. علت دسته‌بندی آن‌ها درخشندگی‌های مشابه در میان همه آن‌ها است. پیدا کردن مکان این ستاره‌ها باعث می‌شود ما بتوانیم ساختار هاله یا حلقهٔ نور را پیدا کنیم. استنباط در مورد حلقهٔ نور راه شیری، نیازمند چند مورد است:

۱- شناسایی ستارگان RRL در میان تمام ستارگان مشاهده شده در یک بررسی ستاره‌شناسی. با یاداوری این نکته که ستارگان متغیر، به‌عنوان داده، توابع نمونه‌برداری نامنظم هستند(نمایهٔ الف-S3 را ببینید)، این یک مشکل بزرگ طبقه‌بندی داده‌های عملکرد است. پس از شناسایی RRL فاصلهٔ آن‌ها تخمین زده می‌شود.

۲- با استفاده از مکان‌های تخمین زده شده از RRL، ما چگالی محلی اشیاء را به منظور شناسایی ساختار تخمین می‌زنیم. زیرا اغلب مکان‌های RRL به‌عنوان نتیجه‌ای از فرایند پواسن در فضای سه‌بعدی در نظر گرفته می‌شوند. خطاهای مرحله قبل، از جمله ستارگان طبقه‌بندی شده در قسمت‌های نادرست و عدم قطعیت در براورد فاصله، بر این تخمین نقشه تاثیر می‌گذارد.

۳- در نهایت، می‌توان ساختار مشاهده‌شده در نقشهٔ حلقهٔ نور را با پیش‌بینی‌های انجام شده توسط مدل کیهانی LCDM مقایسه کرد. ساختارهای حلقه‌های نور گوناگون پارامترهای آزاد در LCDM ارائه می‌دهند. این مقایسه‌ها می‌تواند

اکتشاف یا کمی بیش‌تر باشد(برای نمونه بهینه‌سازی پارامترها در یک شبیه‌سازی کیهان‌شناسی برای تولید ساختار حلقهٔ نور که بیش‌تر شبیه مشاهدات است).

شفر استدلال می‌کند که مسائل استنتاج کیهانی به بهترین وجه به سه مرحله تقسیم می‌شوند. استنتاج در پارامترهای شئ، استنتاج بر روی پارامترهای کلاس و در نهایت استنتاج بر روی پارامترهای کیهان‌شناسی بنیادی. سه مرحلهٔ ذکر شده تقریباً با این مراحل مطابقت دارند. هر مرحله نیاز به تصمیمات آماری زیادی دارد. عدم قطعیت باید در طی مراحل منتشر شود و در حین آن باید تقریب‌هایی انجام شود تا خط لوله تجزیه و تحلیل از نظر محاسباتی امکان‌پذیر باشد. بررسی‌های اخترشناسی آینده، مانند بررسی سیاره فراخورشیدی(TESS)، تلسکوپ فضایی جیمز وب(JWST) و بررسی سینوپتیک بزرگ تلسکوپ(LSST)، مجموعه داده‌های بزرگ‌تری را با مشکلات استنتاجی چالش برانگیزتر نوید می‌دهند. همکاری‌های بین رشته‌ای آماردانان و ستاره‌شناسان برای توسعه روش آماری جدید برای تحقق کامل پتانسیل علمی این پروژه‌ها امری بسیار ضروری است.

مرجع‌ها

جیمز پی.لانگ و رافائل اس ده سوزا

دانشگاه مکانیک و کشاورزی تگزاس، کالج استیشن، تگزاس، ایالات متحده آمریکا

گروه فیزیک و ستاره‌شناسی، دانشگاه کارولینای شمال در شَپل هیل، شپل هیل، کارولینای شمالی، ایالات متحده آمریکا

دانشگاه «MTA Eötvös»، گروه تحقیقات ستاره‌شناسی، بوداپست، مجارستان.

موسسه ستاره‌شناسی، جئوفیسکا، سائوپائولو، برزیل.

https://builtin.com/data-science

علم داده چیست؟

امیررضا عباسی، دانشجوی کارشناسی آمار دانشگاه اصفهان

یک مطالعهٔ پر سر و صدا در سال ۲۰۱۳ گزارش می‌دهد که ۹۰ درصد از کل داده‌های جهان در دو سال گذشته ایجاد شده است. اجازه دهید آن را فراموش کنیم. در دو سال گذشته ما ۹ برابر اطلاعاتی که در ۹۲۰۰۰ سال اخیر گردآوری شده را به دست آورده‌ایم و این روند همچنان با سرعت پیش می‌رود. درحال حاضر ما ۲.۷ زتابایت داده تولید کرده‌ایم و انتظار می‌رود این رقم به زودی به ۴۴ زتابایت برسد.

با این همه داده چه کار می‌توان کرد؟ چه استفادهٔ مفیدی از آن‌ها می‌توان داشت؟ کاربرد حقیقی آن‌ها چیست؟ این‌ها سؤالاتی است که علم داده به آن پاسخ می‌دهد.

هر شرکت می‌گوید که درحال انجام نوعی عملیات مربوط به علم داده است، اما این دقیقاً به چه معناست؟ این رشته به سرعت درحال رشد است و صنایع بی‌شماری را دگرگون کرده است. ارائهٔ یک تعریف واحد برای علم داده دشوار است، اما در کل علم داده مربوط به انجام فرایندهایی روی اطلاعات خام و تمیز برای نتیجه‌گیری است.

در عصر حاضر، از داده‌ها به‌عنوان "نفت قرن بیست و یکم" یاد می‌شود چراکه داده‌ها بیش‌ترین اهمیت را برای ما دارند. این موضوع در تجارت، تحقیقات و زندگی روزمرهٔ ما منافعی غیرقابل چشم‌پوشی دارد. در مسیر حرکت شما به سمت محل کار، آخرین جست‌وجوی شما در گوگل برای نزدیک‌ترین کافه، پست اینستاگرام شما در مورد آن‌چه که خورده‌اید و حتی اطلاعات آمادگی جسمانی شما که توسط دستبندهای سلامتی ثبت می‌شوند، برای دانشمندان حوزهٔ داده به روش‌های گوناگون دارای اهمیت هستند. غربال کردن دریاچه‌های بزرگ داده، پیدا کردن ارتباط و الگو میان آن‌ها، ارائه محصولات جدید، رویکردهای نوین و راحت‌تر کردن زندگی؛ تمام این موارد از مسئولیت‌های علم داده به شمار می‌آید. اما این علم مهم، چگونه کار می‌کند؟

علم داده شامل مجموعه‌ای از رشته‌ها و زمینه‌های تخصصی برای ایجاد نگاهی جامع، کامل و دقیق به داده‌های خام است. دانشمندان علم داده باید تقریباً در همه چیز، از مهندسی داده گرفته تا ریاضی، آمار، محاسبات پیشرفته و تجزیه و تحلیل مهارت داشته باشند تا بتوانند به‌شکل مؤثر کلاف درهم تنیده اطلاعات را غربال کنند و فقط حیاتی‌ترین بخش‌هایی را که به نواوری و کارایی کمک می‌کنند، به یکدیگر متصل نمایند. دانشمندان داده همچنین وابستگی زیادی به هوش مصنوعی در شاخه‌های یادگیری ماشین و یادگیری عمیق دارند. آن‌ها از این ابزار برای ایجاد مدل‌ها و پیش‌بینی با استفاده الگوریتم‌ها استفاده می‌کنند.

علم داده به طور کلی دارای یک چرخهٔ پنج مرحله‌ای است:

۱- **دریافت:** کسب داده‌ها، ورود داده‌ها، دریافت سیگنال، استخراج داده‌ها

۲- **نگهداری:** انبار داده، پاکسازی داده‌ها، مرحله‌بندی داده‌ها، پردازش داده‌ها، معماری داده‌ها

۳- **فرایند:** داده‌کاوی، خوشه‌بندی/طبقه‌بندی، مدل‌سازی داده‌ها، خلاصه‌سازی داده‌ها

٤- **ارتباط:** گزارش داده‌ها، تجسم داده‌ها، هوش تجاری، تصمیم‌گیری

۵- **تجزیه و تحلیل:** اکتشاف و تایید، تحلیل پیش‌بینی‌کننده، رگرسیون، متن‌کاوی، تحلیل کیفی

هر پنج مرحله نیازمند تکنیک‌ها، برنامه‌ها و در برخی موارد مهارت‌های منحصربه‌فرد خود هستند.

کاربرد علم داده

علم داده به ما کمک می کند تا به برخی اهداف بزرگ خود که تا سال‌های قبل امکان‌پذیر نبوده یا رسیدن به آن‌ها نیازمند صرف وقت و انرژی زیادی بودند، دست پیدا کنیم.

موارد استفاده از علم داده:

- تشخیص ناهنجاری‌ها(کلاهبرداری، جرم و جنایت و غیره)
- اتوماتیک کردن کارها و تصمیم‌گیری (بررسی پیشینه، اعتبار و غیره)
- طبقه‌بندی(در سرور ایمیل، این موضوع می‌تواند به معنای طبقه‌بندی ایمیل‌ها به عنوان مهم یا ایمیل‌های غیر ضروری باشد)
- پیش‌بینی(فروش، درآمد و حفظ مشتری)
- تشخیص الگو(الگوهای آب و هوا، الگوهای بازار مالی و غیره)
- تشخیص(چهره، صدا و متن و غیره)
- توصیه‌ها(بر اساس اولویت‌های آموخته شده، الگوریتم‌ها می‌توانند به شما فیلم، رستوران، کتاب و چیزهایی که ممکن است دوست داشته باشید را معرفی کنند.)

در ادامه چند نمونه از نحوهٔ استفادهٔ کسب و کارها از علم داده برای نوآوری در بخش‌های خود، ایجاد محصولات جدید و کارامدتر کردن محیط اطراف آورده شده است.

مراقبت‌های پزشکی و بهداشتی

علم داده منجر به پیشرفت‌های متعددی در صنعت مراقبت‌های بهداشتی شده است. با شبکهٔ وسیعی از داده‌ها که اکنون از EMR تا پایگاه‌های اطلاعاتی بالینی و ردیاب‌های تناسب اندام شخصی در دسترس است. متخصصان پزشکی راه‌های جدیدی برای درک بیماری، پیشگیری از وقوع عارضه، تشخیص سریع‌تر بیماری‌ها و یافتن گزینه‌های درمانی جدید پیدا می‌کنند.

ماشین‌های خودران

تسلا، فورد و فولکس واگن همه درحال پیاده‌سازی تجزیه و تحلیل پیش‌بینی در موج جدید خودروهای خودران هستند. این خودروها از هزاران دوربین و حسگر کوچک برای انتقال اطلاعات در زمان واقعی استفاده می‌کنند. با استفاده از یادگیری ماشین، تجزیه و تحلیل پیش‌بینی‌کننده و علم داده، خودروهای خودران می‌توانند محدودیت‌های سرعت را تنظیم کنند، از تغییر خط خطرناک اجتناب کرده و حتی مسافران را به کوتاه‌ترین مسیر ببرند.

لجستیک

منبع تغذیهٔ اضطراری(UPS) برای به حداکثر رساندن کارایی، هم در داخل و هم در طول مسیرهای تحویل خود، به علم داده احتیاج دارد. ابزار بهینه‌سازی و ناوبری یکپارچه در جاده از مدل‌سازی آماری مبتنی بر علم داده و الگوریتم‌هایی استفاده می‌کند که مسیرهای بهینه را بر اساس آب و هوا، ترافیک، تعمیر جاده و غیره برای رانندگان تحلیل می‌کند. تخمین زده می‌شود که علم داده در حال صرفه‌جویی ۳۹ میلیون گالن سوخت و بیش از ۱۰۰ میلیون مایل تحویل در سال است.

سرگرمی

آیا تا به حال از خود پرسیده‌اید که اسپاتیفای(سیستم پخش موسیقی) چگونه آهنگی که متناسب با حال و هوای شماست، به شما توصیه می‌کند؟ یا چگونه

نتفلیکس(سیستم پخش فیلم و سریال) می‌داند چه چیزی را به شما نمایش دهد؟ با استفاده از علم داده، غول پخش موسیقی می‌تواند قهرست آهنگ‌ها را بر اساس ژانر موسیقی یا گروهی که درحال حاضر به آن علاقه‌مند هستید، به دقت تنظیم کند. واقعاً اخیراً به آشپزی علاقه پیدا کرده‌اید؟ جمع‌اورندهٔ دادهٔ نتفلیکس نیاز شما به الهام از آشپزی را تشخیص می‌دهد و نمایش‌های مربوطه را از روی مجموعهٔ عظیم خود توصیه می‌کند.

دارایی و سرمایه‌گذاری

یادگیری ماشین و علم داده میلیون‌ها دلار در صنعت مالی و در زمان غیرقابل اندازه‌گیری صرفه‌جویی کرده‌اند. به عنوان مثال، بستر اطلاعات قراردادی جی‌پی‌مورگان از پردازش زبان طبیعی برای پردازش و استخراج داده‌های حیاتی از حدود ۱۲۰۰۰ قرارداد اعتبار تجاری در سال استفاده می‌کند. به لطف علم داده، کاری که حدود ۳۶۰۰۰۰ ساعت کار دستی می‌طلبد، اکنون در چند ساعت انجام می‌شود. علاوه بر این، شرکت‌های فین‌تک(اقتصادی) مانند استرایپ و پی‌پال سرمایه‌گذاری زیادی در علم داده برای ایجاد ابزارهای یادگیری ماشینی می‌کنند که به سرعت فعالیت‌های تقلبی را شناسایی و از آن جلوگیری می‌کند.

امنیت سایبری

علم داده در هر صنعتی مفید است، اما ممکن است در امنیت سایبری اوج بهره‌وری را داشته باشد. شرکت بین‌المللی کسپراسکای از علم داده و یادگیری ماشین برای شناسایی روزانه ۳۶۰۰۰۰ نمونه بدافزار جدید استفاده می‌کند. توانایی شناسایی و یادگیری آنی روش‌های جدید جرایم سایبری از طریق علم داده، برای ایمنی و امنیت ما در آینده ضروری است.

کاربرد آمار و احتمال در هواشناسی

محمد صانعی، دانشجوی کارشناسی آمار دانشگاه حکیم سبزواری

متغیرهای هواشناسی

الف- متغیرهای پیوسته: پارامترهایی که مقدارشان می تواند هر عدد صحیح یا اعشاری باشد. مانند مقدار بارندگی، درجه حرارت، فشار هوا و ...

ب- متغیر گسسته: متغییرهایی که مقدارشان فقط اعداد صحیح بوده و در واقع کمیت آنها با فراوانی وقوعشان سنجیده می شود. مانند تعداد روزهای بارانی در سال، تعداد روزهای یخبندان در سال و....

تحلیل فراوانی وقایع در هواشناسی

فراوانی وقوع: تعداد دفعاتی که یک پارامتر مشخص در یک زمان معین به وقوع می پیوندد.

مثال: بارش های اسفند ماه ایستگاه سینوپتیک پلور در یک دوره آماری ۱۰ ساله از سال ۱۳۸۰ تا ۱۳۹۰ به صورت زیر است.

سال	بارندگی اسفند ماه (mm)
1380	53.8
1381	125.3
1382	72
1383	275.3
1384	3.2
1385	109.8
1386	52.5
1387	28
1388	58.5
1389	150.9

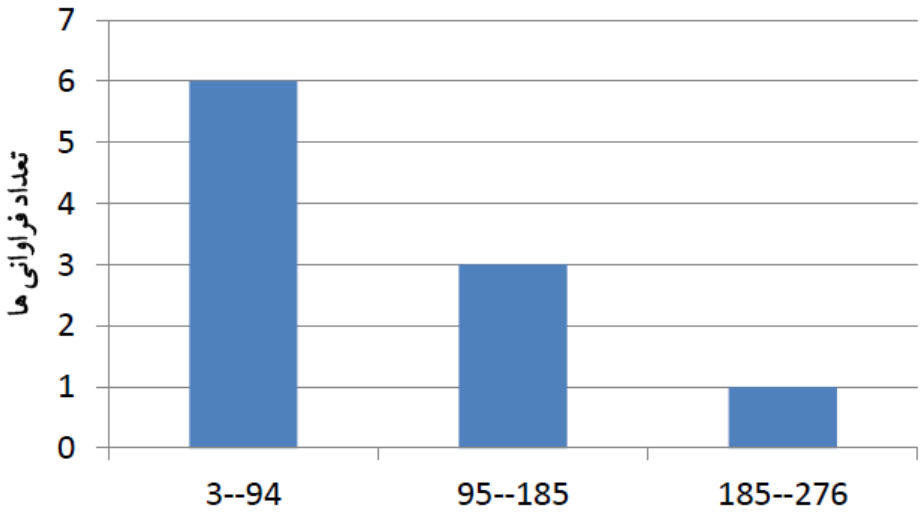
اگر این بارش ها را به ۳ دسته تقسیم کنیم خواهیم داشت:

دسته ها	فراوانی	فراوانی تجمعی	فراوانی تجمعی نسبی(٪)
3-94	6	6	60
95-185	3	9	90
185-276	1	10	100

دو پارامتر پوشش ابر و جهت باد نیز جزء متغییرهای ناپیوسته به حساب می آیند.

تحلیل فراوانی وقایع در هواشناسی

دسته ها	فراوانی	فراوانی تجمعی	فراوانی تجمعی نسبی(٪)
3-94	6	6	60
95-185	3	9	90
185-276	1	10	100



اگر بخواهیم درصد بارندگی های کمتر یا بیشتر از یک مقدار خاص را به دست بیاوریم باید جدول توزیع را بر اساس یکی از روابط احتمالاتی ارائه شده تشکیل داد.

کالیفرنیا	P= m / n
ویبول	P= m / n+ 1
هیزن	P= 2m-1 / 2n
چگادایف	P= m-0.3 / n+0.4
توکی	P=3m-1 / 3n+ 1
بلوم	P= m-0.375 / n+0.25

دوره بازگشت:

دوره بازگشت در واقع عکس احتمال است. به عبارت دیگر تعداد سال هایی می باشد که بین وقوع دو حادثه مشابه وجود دارد.

بنابراین اگر دوره بازگشت T و احتمال وقوع P باشد خواهیم داشت $T=1/P$

مثال: در صورتی که احتمال وقوع بارندگی مساوی یا کمتر از ۶۰ میلیمتر ۵۰ درصد باشد دوره

بازگشت آن را محاسبه کنید.

حل:

$T=1\div0.5=2$

مشخصات آماری داده های هواشناسی

میانگین حسابی یا ریاضی

میانگین تخمینی از متغییر است که احتمال وقوع آن در آینده بیشتر از هر مقدار دیگری می باشد.

محاسن:

- محاسبات ساده
- معمول بودن
- استفاده از تمام گروه های آماری
- قابل استفاده بودن در تمام شرایط
- نیازی به ردیف کردن داده ها نیست

معایب:

- در صورت اشتباه، پیدا کردن نقطه اشتباه دشوار است
- عدم استفاده از حتی یک رقم اثرات زیادی در میانگین بدست آمده دارد.
- وجود ارقام بسیار بزرگ و بسیار کوچک اثرات زیادی بر نتیجه دارد.

میانگین وزنی:

در میانگین حسابی همه اعداد ارزش یکسانی دارند ولی در میانگین وزنی هر عدد دارای ارزش یا وزن خاصی در نظر گرفته می شود.

میانگین هندسی:

در این روش ریشه n ام حاصل ضرب n متغییر محاسبه می شود

محاسن:

- استفاده از تمام داده ها
- محاسبات ساده است.

- ارقام کوچک و بزرگ اثر زیادی در نتیجه ندارند.
- به ارقام بزرگ بهای زیادی داده نمی شود و برعکس به ارقام کوچک ارزش بیشتری داده می شود.

معایب:

- اگر رقمی معادل صفر یا منفی باشد قابل استفاده نیست.

میانگین متحرک:

در شرایطی که بخواهیم تغییرات زمانی داده را مورد بررسی قرار دهیم ممکن است دامنه تغییرات به حدی باشد

که نتوان از آن چیزی درککرد. در این شرایطاز میانگین متحرک استفاده می شود.

انتخاب پایه زمانی مشترک آماری

مشکل همیشگی در تجزیه و تحلیل آمارهای منطقه ای وجود تعداد سال های آماری متفاوت برای

ایستگاه ها می باشد که مربوط به تاسیس آنها در سال های مختلف ، خرابی ایستگاه و .. در طی چند

سال می باشد.

آمار یک ایستگاه ممکن است مربوط به دوره خشکسالی و آمار ایستگاه دیگر مربوط به دوره ترسالی باشد و بنابراین ممکن است تغییراتمکانی و زمانی باهم مخلوط شده و نامشخص گردند.

اقلیم:

به معنی آب و هوا بوده و یک مفهوم احساسی است که می توان آن را متوسط وضعیت هوا در یک منطقه توصیف کرد.

دو پارامتر مهم برای سنجش اقلیم : دما و بارندگی نقش بارز تابش، جنس سطح، باد و فشار و...

ساده ترین طبقه بندی اقلیمی:

- مناطق تروپیکی
- عرض های معتدله
- مناطق قطبی

اساس طبقه بندی ها :

- تجربی
- پیشنهادی
- ژنتیکی

تعیین نوع اقلیم:

۱- اقلیم فراخشک
۲- اقلیم خشک
۳- اقلیم نیمه خشک
٤- نیمه مرطوب
٥- اقلیم مرطوب

آشنایی با مارکوف

فاطمه کلمیشی، کارشناس ارشد آمار دانشگاه حکیم سبزواری

آندری آندریویچ مارکوف در ۱۴ ژوئن ۱۸۵۶ در روسیه متولد شد. او در رشته دانشگاهی خود در بسیاری از موضوعات به جز ریاضیات خیلی ضعیف عمل کرد. بعداً در دانشگاه پترزبورگ تحصیل کرد؛ او تحصیلاتش را در دانشگاه به پایان رساند و بعدها از وی درخواست شد به عنوان یک ریاضیدان مشغول به کار شود. بعدها در دبیرستان تدریس کرد و مطالعات ریاضی خود را ادامه داد. در این زمان برای مهارتهای ریاضی خود یک کاربرد عملی پیدا کرد و به این نتیجه رسید که میتواند از زنجیر برای مدلسازی حروف صدادار و بیصدا در ادبیات روسی استفاده کند.

در ابتدا تمرکزش بر روی نظریه اعداد بود و بعداً توجهش به نظریه احتمالات جلب شد و تا پس از بازنشستگی و تا آخر عمر مشغول به تدریس بود. از دستاوردهای وی تدوین قانون اعداد بزرگ و قضیه حد مرکزی جهت کاربرد در دنباله های معینی از متغیرهای تصادفی وابسته بود که بعدها به عنوان زنجیر مارکوف شناخته شد. از زنجیر مارکوف در بسیاری از علوم از جمله فیزیک، اقتصاد، آمار، زیستشناسی استفاده میشود و مسائلی مانند حرکت براونی، گشت تصادفی و پیجرنک از مشهورترین موارد کاربردهای زنجیر مارکوف هستند . وی به خاطر کارش در فرآیندهای تصادفی مشهور است. موضوع اصلی تحقیقات او بعدها به عنوان زنجیر مارکوف و فرایند مارکوف شناخته شد . در سال ۱۸۷۷ مدال طلا برای راه حل برجسته حل مسئله به وی اهدا شد. سال بعد در آزمون های نامزدی برای استادی قبول شد و در دانشگاه ماند تا برای سمت استادیاری آماده شود. در آوریل ۱۸۸۰ ، مارکوف از رساله فوق دفاع کرد که مورد تشویق قرار گرفت. چهار سال بعد در « اشکال مربع دودویی با تعریف مثبت » لیسانس اش در سال ۱۸۸۴ دفاع کرد « کاربرد خاص جبری بخش متوالی » ، از تز دکترای خود با عنوان . شغل استادی او پس از دفاع از پایان نامه کارشناسی ارشد در پاییز سال ۱۸۸۰ آغاز شد. وی در زمینه دیفرانسیل و انتگرال به تدریس پرداخت. بعدها به طور متناوب درباره ی (معرفی تحلیل)، نظریه احتمال و ریاضیات تفاوتها تدریس کرد. از ۱۸۹۵ تا ۱۹۰۵ نیز محاسبات دیفرانسیل را تدریس کرد . مارکوف و برادر کوچکترش، ولادیمیر نابرابری برادران مارکوف را ثابت کردند. وی در سن ۶۶ سالگی در ۲۰ ژوییه ۱۹۲۲ درگذشت.

راه‌های ارتباط با انجمن علمی آمار دانشگاه حکیم سبزواری

کانال تلگرام:

https://t.me/anjoman_amar_hsu

اینستاگرام:

https://www.instagram.com/hsu.statistics?r=namtag

وبلاگ:

http://anjoman-amar-hakim.blogfa.com/

انجمن علمی آمار دانشگاه حکیم سبزواری

خلاقیت، آموزش، مسابقه

#ارتباط

@anjoman_amar_hsu

دبیر انجمن: محمد صانعی

استاد مشاور: دکتر محمد بلبلیان قالیباف