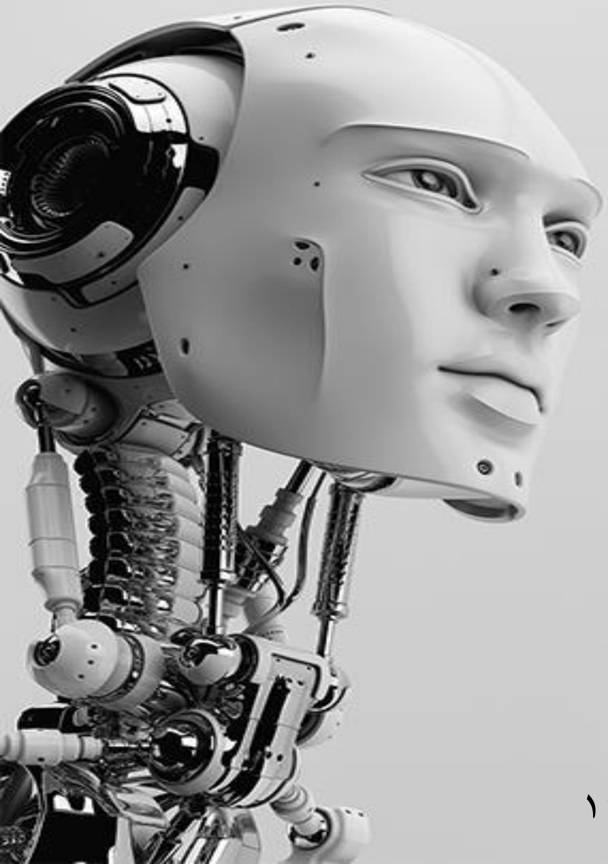
کاربرد آمار مقدماتی و تحلیل اکتشافی داده (EDA) در علم داده

04

سید ایمان غفوریان - عضو هیات علمی دانشگاه آزاد مشهد -
زمستان ۱۴۰۳

بخش دوم » کاربرد آمار مقدماتی و تحلیل اکتشافی داده (EDA) در علم داده



02

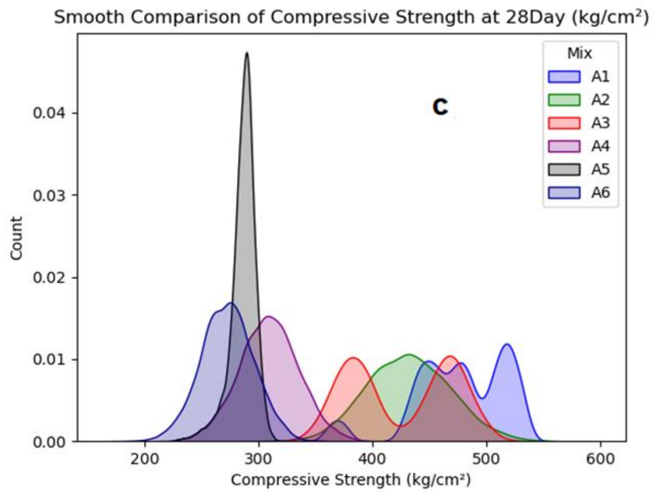
آمار و احتمالات از مباحث مقدماتی تحلیل داده‌ها در علم داده data science، داده‌کاوی data mining یادگیری ماشین machine learning است. فراگیریِ مباحث پایه آمار و احتمالات به ما در درک مفاهیم پایه‌ی الگوریتم‌های مختلف کمک خواهد کرد. همچنین با یادگیریِ این مباحث می‌توانید تحلیل‌های ساده‌ترِ آماری را که نیازی به الگوریتم‌های پیچیده و ابزارهای خاص ندارند به سادگی انجام داده و از نتایج استخراج شده از آن بهره برداری نمایید. در این میان تحلیل اکتشافی داده یا همان exploratory data analysis به این معناست که ما داده‌های موجود را به صورت خلاصه و جمع و جور با انواع روش‌های مختلف جمع‌آوری و آماده‌سازی کرده و آن‌ها را تحلیل کنیم. از قدیم گفته‌اند مشت نمونه خروار است. پس اگر بتوان یک مشت خوب از خروارِ داده برداشت احتمالاً می‌توان با همان مشت، یک مجموعه تحلیلِ درست از روی داده‌ها انجام داد. این مشت از داده‌ها یا نمونه‌های کوچک شده، اطلاعاتی را راجع به داده‌ها به ما می‌دهند که می‌تواند بسیار مفید باشد.

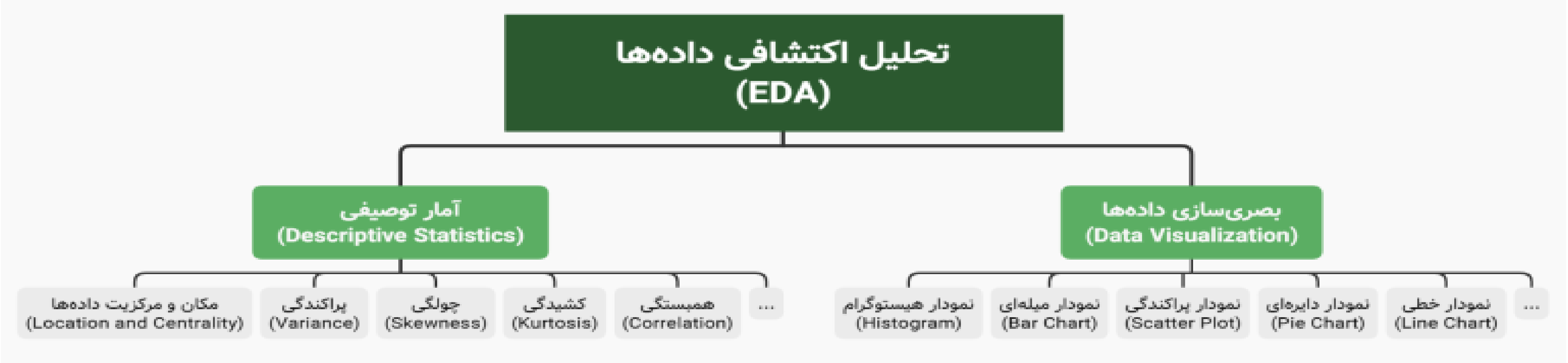
تحلیل اکتشافی داده‌ها Exploratory Data Analysisچیست؟

سن دانشجویان ورودی سال جدید دانشگاه شما چند سال است؟ ۱۸سال
به همین سادگی؟ بله. آیا دقیقاً تمامی دانشجویایی که ورودی امسال بوده‌اند همه آن‌ها ۱۸ساله هستند؟ قطعاً خیر ولی برای سادگی و خلاصه‌سازی می‌توان عدد ۱۸ را پاسخ صحیحی دانست تا یک جوابِ تخمینی و نیمه دقیق به سوال کننده داده شود. اگر سوال کننده بخواهد اطلاعات دقیق‌تر را به دست آورد می‌تواند لیستی از تمامی ورودی‌های سال جدید دانشگاه به همراه سن آن‌ها بیابد. این خلاصه‌سازی عمل اصلی در تحلیل اکتشافی داده‌ها یا همان EDA است که یکی از مباحث مهم در آمار برای متخصصان علم داده، داده‌کاوی و یادگیری ماشین است.
به کمک EDAو تکنیک‌ها و مفاهیم آن شما می‌توانید داده‌ها را خلاصه سازی کنید. یک شمای کلی از داده‌ها به افراد مختلف نمایش دهید و یا با انواع نمودارها و گراف‌ها، خلاصه‌ای از داده‌ها را به افرادی که به آن‌ها نیاز دارند نمایش دهید. در واقع EDAرا می‌توان از جنبه‌ای اولین قدم برای رسیدن به یک تحیل یا همان آنالیز داده‌ها و یک شمای کلی از داده‌ها دانست. ذهن ما انسان‌ها برای پردازش اعداد و ارقام ضعیف است (احتمالاً این‌گونه تکامل یافته‌ایم) ، پس نیاز داریم به جای مشاهده و پردازشِ کل داده‌ها، خلاصه‌ای از داده‌ها را داشته باشیم. مثلاً نقشه جهان را در نظر بگیرید. نقشه جهان در واقع یک خلاصه‌ای از کل جهان است که خیلی از جزئیات در آن وجود ندارد ولی بسیار کاربردی است. یک مدلِ ساده شده و خلاصه شده از جهان که به ما می‌گوید الان در کجای جهان ایستاده‌ایم و کشورهای مجاور ما کدام‌ها هستند.
حال فرض کنید یک سری داده‌ی حجیم و بسیار زیاد دارید که می‌خواهید آن‌ها را تحلیل کنید. آیا بهتر نیست ابتدا یک مدلِ ساده شده و خلاصه شده به صورت شهودی‌تر از داده‌ها داشته باشید تا بتوانید فرضیات خود را برای تحلیل داده بر اساس آن‌ها بچینید؟ مانند این که می‌خواهید به یک مسافرت بروید و یک نقشه که قطعاً خلاصه‌ای از مسیرها و پیچ و خم های جاده است در اختیار داشته باشید تا به کمک آن بتوانید مسیر را به صورت شماتیک بهتر درک کنید. EDA در داده‌ها دقیقاً همین کار را می‌کند. با استفاده از تکنیک‌ها و روش‌های موجود در EDA، خواهید توانست داده‌ها را ساده‌تر در ذهن خود جای دهید و تحلیل و بررسی کنید.

تحلیل اکتشافی داده‌ها Exploratory Data Analysis چیست؟

EDA در داده‌ها دقیقاً همین کار را می‌کند. با استفاده از تکنیک‌ها و روش‌های موجود در EDA، خواهید توانست داده‌ها را ساده‌تر در ذهن خود جای دهید و تحلیل و بررسی کنید. روش‌ها و تکنیک‌های EDA را می‌توان در دو دسته‌ی کلی مانند شکل زیر تقسیم‌بندی کرد:





همان‌طور که در شکل بالا مشخص است، آمار توصیفی یا همان descriptive statistics و بصری‌سازی داده‌ها یا همان data visualization دو بخش است از روش‌ها و تکنیک‌های موجود در EDA هستند. هر کدام از این روش‌ها با استفاده از فرمول‌ها و تکنیک‌هایی می‌توانند داده‌ها را خلاصه‌سازی کرده و به تحلیل کمک کنند.

منظور از داده‌های مستطیلی Rectangular Data چیست؟

در داده‌کاوی معمولا سعی می‌شود که داده‌ها به صورت مستطیلی یا همان Rectangular شکل بگیرد تا بتوانند عملیات بندی مانند طبقه‌بندی یا خوشه‌بندی را بر روی آنها انجام دهند. سطرها در داده‌های مستطیل همان نمونه‌ها Samplesها هستند. برای مثال یک مجموعه داده دانشجویان را در نظر بگیرید که کدام ویژگی‌های مشخصی دارند. مثلا سن، معدل، قد، جنسیت و... . درشکل زیر هر سطر یک نمونه یا Instance یا رکورد Record نامیده می‌شود ،که بیانگر یک دانشجو است. هردانشجو می‌تواند چندین ویژگی (ستون‌ها) داشته باشد. در درس ویژگی یا بعد چیست کامل به این موضوع پرداخته‌ایم. به این ترتیب داده‌ها را می‌توان به شکل مستطیلی درآورد. حتی اگر داده‌ای شکل مستطیلی نداشت می‌توان آن را به حالت مستطیلی درآورد. برای این کار روش‌های مختلفی است که یکی از آنها را برای سادگی اینجا می‌آوریم. One Hot Encoding در واقع روشی برای تبدیل داده‌های غیرعددی به عددی است که می‌توان از آن برای ساخت داده‌هایی با شکل مستطیلی استفاده کرد . برای مثال در همان مثال فوق جنسیت را در نظر بگیرید. اگر بخواهیم این جنسیت را هم به یک ویژگی عددی تبدیل کنیم تا مانند ستون‌های دیگر بتواند مقدار عددی بگیرد به صورت زیر تبدیل می‌شود:چون دو جنس (مرد و زن) داریم باید دو ستون اضافه کنیم و آنجایی که دانشجو مرد است، ستون مرد برابر ۱قرار می‌گیرد و ستون زن برابر ۰ و برای بالعکس برای زن‌ها. به همین راحتی می‌توان تبدیل One Hot Encoding را انجام داد و تمامی ویژگی‌ها را به ویژگی‌های عددی تبدیل کرد.

	سن	معدل	قد	جنسیت
Student #1	19	18.5	185	مرد
Student #2	21	17.0	162	زن
Student #3	22	15.5	177	زن
Student #4	20	18	156	مرد
⋮				

	سن	معدل	قد	مرد	زن
Student #1	19	18.5	185	1	0
Student #2	21	17.0	162	0	1
Student #3	22	15.5	177	0	1
Student #4	20	18	156	1	0
⋮					

داده پرت در داده‌کاوی چیست؟

اگر در Google Translate واژه Outlier را ترجمه کنید، ترجمه به شدت درستی انجام می‌شود: **داده‌های خارج از محدوده**! این همان تعریف برای داده‌های پرت یا همان Outlierها هستند. داده‌های پرت در بعضی مواقع می‌توانند **دردسر ساز** باشند و در بعضی مواقع هم خود مسئله تشخیص داده‌های پرت و به نوعی **تشخیص ناهنجاری** Anomaly Detection است که در آن ما به دنبال یافتن داده‌هایی هستیم که پرت هستند.

فرض کنید می‌گویند میانگین حقوق در یک شرکت ۴میلیون تومان است. آیا این بدان معنی است که اکثر افراد حاضر در آن شرکت ۴میلیون تومان (یا نزدیک به آن) حقوق می‌گیرند. در نگاه یک غیرمتخصص بلی ولی در نگاه یک متخصص آمار و داده‌کاوی قطعا جواب خیر است. ممکن است ۹۵درصد افراد حاضر در آن شرکت حقوق ۱میلیون تومان بگیرند و ۵درصد بقیه حقوقشان ۲۰میلیون تومان باشد.در واقع این ۵درصد نوعی داده پرت هستند که میانگین میانگین Mean را به نفع خود جا به جا کرده‌اند!

فرض کنید بگویند میانگین ثروت افراد حاضر در ماکروسافت ۱۰۰میلیون دلار است. طبیعتا شخصی مانند بیل گیتس که ثروت او به اندازه چندین کشور است، میانگین را به بسیار بالاتر از میانگین واقعی تبدیل کرده‌است و یک متخصص داده‌کاوی باید بتواند این این مسئله را تشخیص دهد. در واقع ثروت بیل گیتس یک داده پرت یا همان Outlier در میان ثروت کارکنان و مدیران ماکروسافت است.

تخمین مکان داده‌ها چیست و انواع مختلف آن کدامند؟

معیارهایی که برای سنجش موقعیت یا همان Location و تخمین آن بپردازیم. بسیاری از این موارد را قبلا در دروسی مانند ریاضیات یا آمار و احتمالات خوانده‌اید ولی تکرار آن‌های می‌تواند به یادگیری کمک کند:

۱. میانگین Mean: برای مثال میانگین سن افراد ورودی یک سال در یک دانشگاه حدودا ۱۸ سال است. جمع تمامی سنین و تقسیم آن به تعداد افراد میانگین سن افراد را به ما می‌دهد. اگر درس داده‌های پرت را خوانده باشید متوجه می‌شوید که میانگین، نسبت به داده‌های پرت انحراف دارد به این معنی که یک داده پرت (مثلا یک سن خیلی زیاد مثلا ۶۰ سال در میان ورودی‌های یک دانشگاه) می‌تواند میانگین را به سمت این سن بالا جا به جا کند. پس به اصطلاح گفته می‌شود که میانگین یک واحد قوی برای محاسبه تخمین مکان داده‌ها نیست.

۲. میانگین وزن دار Weighted Mean: اگر هر کدام از داده‌های ما یک وزن مشخص داشته باشند و بخواهیم میانگین آن‌ها را حساب کنیم ابتدا باید این داده‌ها را در وزن آنها ضرب کنیم. سپس با هم جمع کرده و تقسیم بر تعداد کنیم. این کار را در کارنامه یک ترم دانشگاهی خود حتما دیده‌اید. مثلا درس تربیت بدنی ضریب ۲دارید و درس آمار و احتمالات ضریب ۳ و به همین ترتیب بقیه دروس هر کدام ضریب خود را دارند. حالا وقتی می‌خواهند معدل یک ترم شما را حساب کنند، هر کدام از دروس را در ضریب (یا همان وزن آن) ضرب می‌کنند و سپس این مقادیر را با هم جمع کرده و تقسیم بر تعداد ضرایب می‌کنند تا معدل ترم شما حساب شود.

۳. مُد Mode: مقداری که بیشترین تکرار را در میان داده‌ها دارد. برای مثال فرض کنید سن افراد حاضر در یک کلاس به صورت زیر است:

۱۸, ۱۹, ۱۸, ۱۸, ۲۰, ۶۰, ۲۱, ۲۰, ۱۸, ۱۹ همان طور که می‌بینید بیشترین تکرار را عدد ۱۸ با ۴بار تکرار داشته است. پس مُد برای این داده‌ها عدد ۱۸ است.

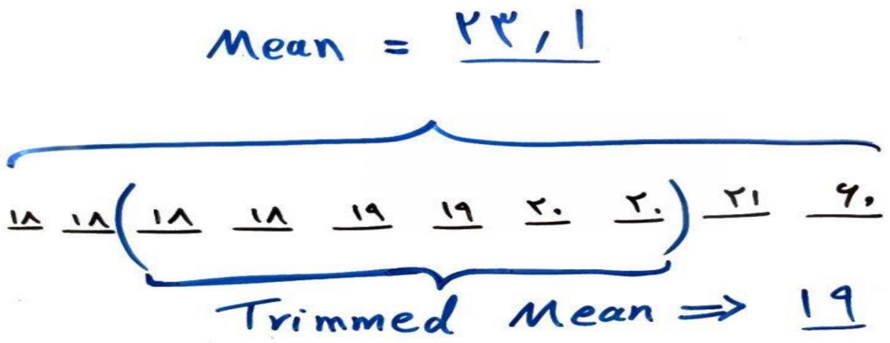
تخمین مکان داده‌ها چیست و انواع مختلف آن کدامند؟

۴. میانه Median: ابتدا داده‌هایی را که دارید به ترتیب مرتب کنید. سپس مقدار وسطی (که نصف داده‌ها از آن بیشتر باشند و نصف داده‌ها از آن کمتر باشند) را انتخاب کنید. این مقدار همان مقدار Median است. اگر تعداد داده‌ها زوج بود، برای محاسبه Median بایستی میانگین دو عدد وسط را حساب کنید. برای مثال اعداد بالا را برای سن افراد حاضر در کلاس به ترتیب چینش می‌کنیم:

۱۸, ۱۸, ۱۸, ۱۸, ۱۹, ۱۹, ۲۰, ۲۰, ۲۱, ۶۰

مقدار Median برای داده‌های فوق برابر ۱۹ است. با توجه به درس داده‌های پرت احتمالا متوجه شده‌اید که، مقدار Median از مقدار میانگین قوی‌تر عمل می‌کند. یعنی سن ۶۰ سال در مثال بالا تقریبا در میان داده‌ها برای مقدار Median حساب نشده است در حالی‌که این مقدار (۶۰ سال) میانگین را به نفع خود خیلی جا به جا می‌کرد. البته میانه وزن‌دار هم داریم که می‌توانید نحوه محاسبه آن را مانند میانگین وزن‌دار به دست آورید.

۵. میانگین برش‌خورده Trimmed Mean: داده‌ها را به ترتیب بچینید. n درصد از بالای داده‌ها و n درصد از پایین داده‌ها را بردارید (برش دهید). حال میانگین مقادیر باقی مانده را محاسبه کنید. به این کار به اصلاح Trimmed Mean می‌گویند. شکل زیر را ملاحظه فرمایید:

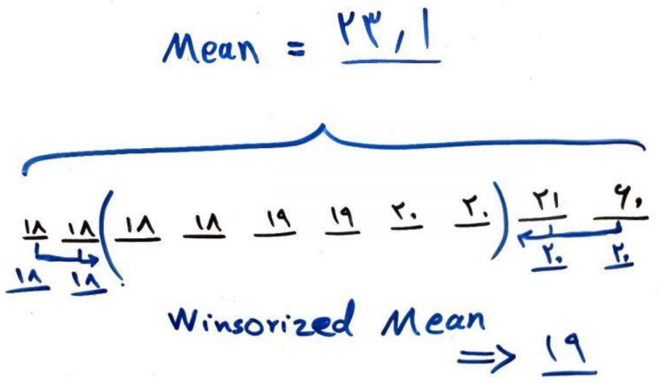


تخمین مکان داده‌ها چیست و انواع مختلف آن کدامند؟

همان‌طور که می‌بینید از میان‌داده‌های بالا، ۲تا از بالا و ۲تا از پایین را برش دادیم. حالا میانگین گرفته می شود. این کار باعث شد که میانگین نسبت به داده‌های پرت قدرتمند Robust باشد. یعنی دیگر مقادیر پرتی مانند سن ۶۰سال نمی‌توانند میانگین را خیلی خراب کنند.

۶. میانه–میانگین Mid–Mean: مانند میانگین برش‌خورده Trimmed Mean است با این تفاوت که ۲۵درصد از بالای داده‌ها و ۲۵درصد از پایین داده‌ها را برداشته و از میان داده‌های باقی مانده در بین این دو (۲۵ تا ۷۵درصد) میانگین را محاسبه می‌کنیم. با این کار باز هم میانگین محاسبه شده نسبت به داده‌های پرت قوی Robust است.

۷. میانگین :Winsorized این میانگین هم مانند میانگین برش خورده Trimmed Mean است با این تفاوت که داده‌هایی که از بالا و پایین قرار است حذف شوند، حذف نمی‌شوند. این داده‌ها تبدیل به بیشتر و کمتر مقدار باقی مانده شده و در محاسبه میانگین حساب می‌شوند. شکل زیر را ملاحظه فرمایید :



در این شکل ما ۶۰ و ۲۱ را به جای اینکه حذف کنیم،به عدد ۲۰تبدیل کردیم و دو ۱۸پایین رابه جای حذف کردن به عدد کوچک‌تر باقی مانده یعنی ۱۸ (سومین ۱۸) تبدیل کردیم. حالا میانگین همه اعداد را محاسبه کردیم.

این میانگین علاوه بر اینکه به نسبت به داده‌های پرت قوی است این داده‌ها را حذف نکرده و به نوعی در میانگین گیری تاثیر می‌دهد.

۸. میانه بازه Mid–Range: کوچک‌ترین مقدار و بزرگترین مقدار را با هم جمع و سپس تقسیم بر ۲ می‌کنیم. این میانگین شدیدا به داده‌های پرت دار حساس و در واقع قدرت کمی نسبت به داده‌های پرت دارد.

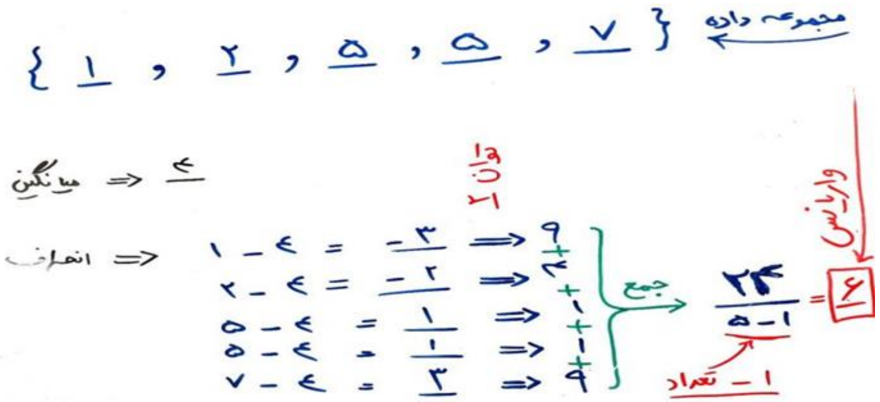
تخمین تنوع و پراکندگی Estimation Of Variabilityو انواع مختلف آن

برای تعریف پراکندگی داده‌ها (یا مجموعه‌ای از یک سری داده‌ها) تعاریف و معیارهای مختلفی موجود است که هر کدام کاربرد خاص خود را دارند. احتمالا اگر به صورت کاربردی یا حتی تئوری علوم داده را خوانده باشید در چند جا به انواع معیارهای پراکندگی برخورده‌اید. چند مورد از مهم‌ترین معیارهای پراکندگی را با هم مرور می کنیم.

داده‌های زیر را فرض کنید: ۱، ۲، ۵، ۵، ۷ میانگین این داده‌ها برابر ۴ است. یعنی فرض کنید تخمین مکان را برای این داده‌ها معیار میانگین در نظر گرفته‌ایم. حال اگر بخواهیم انحراف یا همان **Deviation**را تعریف کنیم مانند شکل زیر است:

در واقع اختلاف هر کدام از اعداد نسبت به میانگین را اینجا انحراف یا همان **Deviation**در نظر گرفته‌ایم. البته برای انحراف می‌توانیم مقدار قدر مطلق یا همان **Absolute** را در نظر گرفته‌ایم. یعنی مقدار منفی‌ها را قدر مطلق گرفته و مثبت کرده‌ایم. حال اگر میانگین این انحرافات را به دست بیاوریم یک معیار استاندارد را حساب کرده‌ایم. این معیار همان میانگین قدر مطلق انحراف یا انحراف مطلق میانگین یا **Mean Absolute Deviation**است.

البته که **Mean Absolute Deviation**تنها معیار برای تخمین پراکندگی نیست. معروف‌ترین معیار در این حوزه شاید همان واریانس **Variance**باشد. شکل زیر واریانس را برای اعداد بالا محاسبه کرده‌است:



تخمین تنوع و پراکندگی Estimation Of Variability و انواع مختلف آن

اگر از **واریانس** رادیکال بگیرید، معیار دیگری به اسم **انحراف استاندارد** یا همان **Standard Deviation** به دست می‌آید که عددی که از انحراف استاندارد به دست می‌آید نسبت به واریانس گویاتر است. زیرا واریانس هر کدام از اختلافات را به توان ۲ رساننده است. **در واقع مثلا وقتی می‌گوییم انحراف استاندارد برابر ۲.۴ است یعنی هر کدام از اعداد تقریبا به اندازه ۲.۴ واحد (به طور میانگین) از معیار میانگین مجموعه اعداد فاصله دارند. ولی وقتی بگوییم واریانس برابر ۶ است، چیزی دستگیرمان نمی‌شود. (در واقع معیار انحراف استاندارد قابل تفسیرتر است) فقط می‌توانیم واریانس دو مجموعه‌ی مختلف را با هم مقایسه کنیم. یعنی اگر یک مجموعه‌ای واریانس بیشتری نسبت به یک مجموعه‌ی دیگر داشت، می‌توانیم بفهمیم که این مجموعه پراکندگی بیشتری دارد.**

دو معیاری **واریانس و انحراف استاندارد** نسبت به **داده‌های پرت** حساس هستند به این معنی که **داده‌های پرت می‌توانند** بر روی آن‌ها **تاثیر زیادی بگذارند**.

انحراف استاندارد

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = انحراف معیار جمعیت

N = تعداد اعضای جمعیت

x_i = اندازه هر عضو از جمعیت

μ = میانگین جمعیت

واریانس

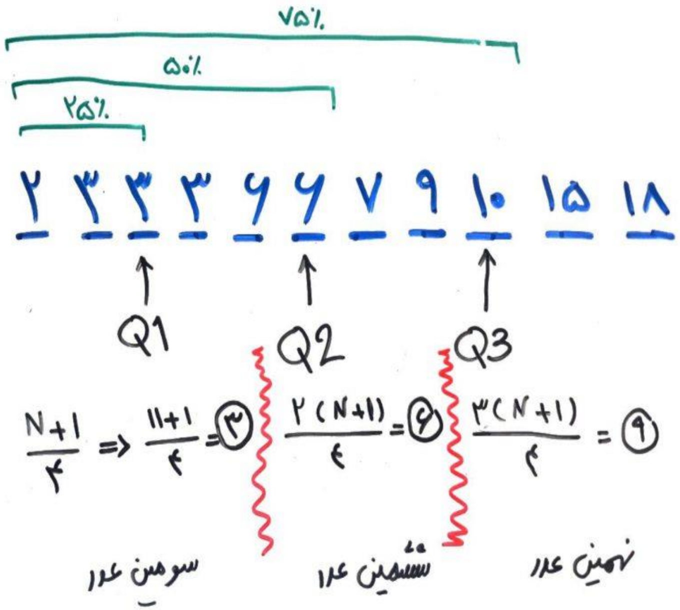
$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

چارک و IQR در داده‌ها و کاربردهای آن

چارک یا همان quartileداده‌ها را به چهار قسمت تقسیم می‌کند به گونه‌ای که هر قسمت ۲۵درصد (یک چهارم) داده‌ها را در خود داشته باشد. فرض کنید داده‌های زیر، امتیازاتی باشد که کاربران مختلف به یک راننده در تاکسی اینترنتی (بین صفر تا بیست) داده‌اند:

۶ – ۳ – ۲ – ۶ – ۳ – ۷ – ۱۵ – ۳ – ۹ – ۱۰ – ۱۸

برای اینکه این داده‌ها را چارک‌بندی کنیم، ابتدا آن‌ها را به ترتیب از کوچک به بزرگ مرتب کرده و داده‌ها به به بخش‌های ۲۵ درصدی تقسیم می‌کنیم. چیزی مانند شکل زیر:فرمول ساده است. تعداد داده‌ها را به علاوه‌ی یک کرده و سپس تقسیم بر چهار می‌کنیم تا به چارک اول برسیم. مثلاً در این مثال ۱۱ + ۱ برابر ۱۲ می‌شود و سپس ۱۲ را تقسیم بر ۴ می‌کنیم که به عدد ۳ می‌رسیم. پس باید داده‌ها سه تا سه تا تقسیم شود.



در کل سه چارک داریم، Q1، Q2 وQ3. چارک اول یا همان Q1، عددی است که ۲۵ درصد داده‌ها کمتر از آن هستند و۷۵ درصد داده‌ها بیشتر از آن هستند. Q2 چارک دوم است، میانه‌ی median داده‌ها را نشان می‌دهد. میانه که در دروس قبلی به آن اشاره کردیم همان عددی است که ۵۰ درصد داده‌ها کمتر از آن هستند و ۵۰ درصد داده‌ها بیشتر از آن. Q3هم که همان‌طور که حدس می‌زنید، داده‌ای است که ۷۵ درصد داده‌ها کمتر از آن و ۲۵ درصد داده‌ها بیشتر از آن هستند.

البته نکته اینجاست که اگر داده‌ها به صورتی باشند که با استفاده از فرمولِ (N+1)/4) به عدد صحیح نرسند، می باید میانگین آن دو عددی که به آن می‌رسیم را به عنوان چارک‌ها انتخاب نمود. برای مثال شکل زیر را نگاه کنید که در آن چارک اول عددی بین ۱۸ و ۲۲ شده است که بایستی میانگین ۱۸ و ۲۲ را محاسبه کرده و آن را به عنوان چارک اول Q1در نظر گرفت.

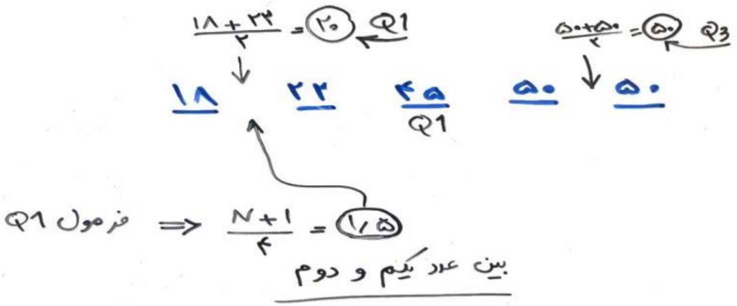
چارک و IQR در داده‌ها و کاربردهای آن

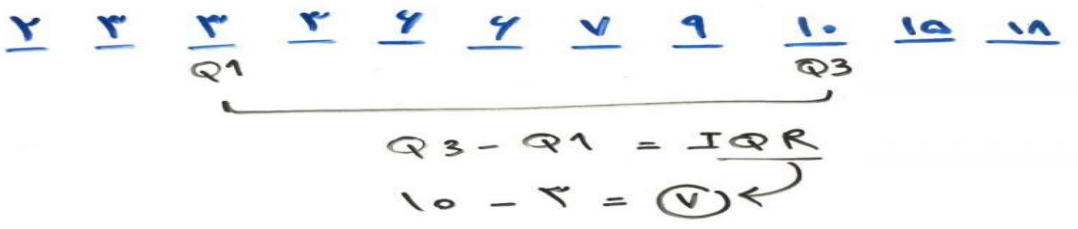
چارک‌ها به شما اجازه می‌دهد که بتوانید بااستفاده از چند عدد،به سادگی داده‌ها را شناسایی کنید.مثلا با استفاده از کمینه، چارک اول، میانه، چارک سوم و بیشینه در یک نگاه می‌توان به توزیع و شکل داده‌ها به صورت تخمینی و حدودی دست پیدا کرد. یعنی به جای اینکه کل داده‌ها را یکی یکی بررسی کنیم می‌توانیم داده‌ها به صورت خلاصه با این پنج عدد مشخص کنیم.

IQR (Interquartile Range)یا همان فاصله‌ی بین چارکی، به فاصله‌ی بین چارک اول و سوم می‌گویند

. IQRبه نوعی پراکندگی داده‌ها را بدون نویز (داده‌های پرت) نشان می‌دهد.

برای مثال در شکل اول همین درس، IQRبرابر ۷ می‌شود. مانند شکل زیر:

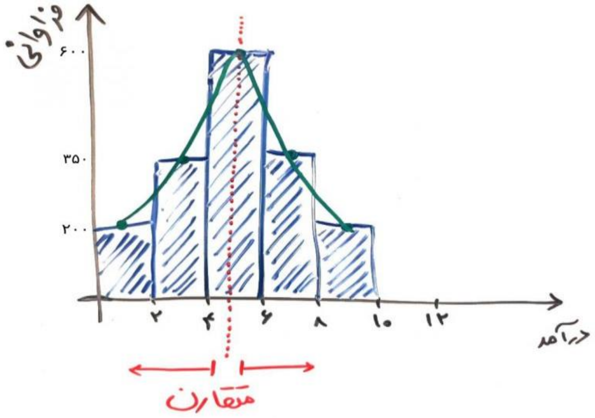




در بسیار از موارد، برای تحلیل پراکندگی داده‌ها می‌توانیم از IQRاستفاده کنیم. IQRداده‌های پرت را حذف می‌کند و به نوعی پراکندگی را به صورت واقع‌بینانه (بدون در نظر گرفتن اتفاقاتی که ممکن است باعث ایجاد آن داده‌ی پرت شده باشند)، بررسی می‌کند. مثلاً وقتی می‌خواهید بدانید که مسافران به یک راننده‌ی تاکسیِ اینترنتی در چه بازه‌ای امتیاز داده‌اند، IQRمی‌تواند گزینه‌ی مناسبی باشد چون داده‌های پرت (مثلاً یکی از مسافران که به دلیل خاصی نمره‌ی خیلی پایین‌تر یا بالاتر نسبت به بقیه را به این راننده داده است)، تاثیری در محاسبه‌ی IQRندارند. همچنین از چارک‌ها می‌توان در تخمین توزیع و شکل و چولگی داده‌ها نیز استفاده کرد.

چولگی Skewness در داده‌ها

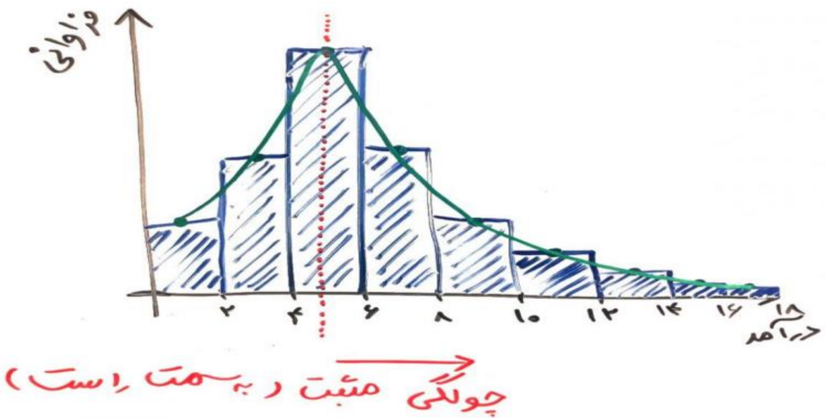
احتمالاً با توزیع نرمال آشنایی دارید. در مورد این توزیع و توزیع‌های دیگر در این دوره بخش کرده‌ایم. اگر توزیع داده‌ها به سمت راست یا چپ کشیده شده باشد، می‌گوییم داده‌ها چولگی یا همان skewness دارند.فرض کنید توزیعِ درآمد ماهیانه برای راننده‌های مختلف یک تاکسی اینترنتی به صورت شکل زیر باشد:



نمودار مقابل یک نمودار هیستوگرام (فراوانی) است.

در شکل مشاهده می‌کنید که توزیع داده‌ها به صورت نرمال(یا همان گوسین) پراکنده شده است. برای مثال تعداد ۶۰۰ راننده تاکسی، درآمدی بین ۴ تا ۶ میلیون تومان دارند.تعداد ۳۵۰ راننده‌ی تاکسی درآمدی بین ۶ تا ۸ میلیون دارند و تعداد۲۴۰ راننده هستند که درآمدی بین ۲ تا ۴ میلیون تومان دارند و همین‌طور برای بقیه‌ی داده‌ها...

نمودار بالا یک نمودار هیستوگرام برای نمایش توزیع داده‌هاست. این نمودار به صورت متقارن است. یعنی سمت چپ و راستِ داده ها، یکسان هستند. ولی فرض کنید هیستوگرام داده‌ها به جای اینکه به صورت بالا باشد، به صورت شکل زیر توزیع شده باشند:

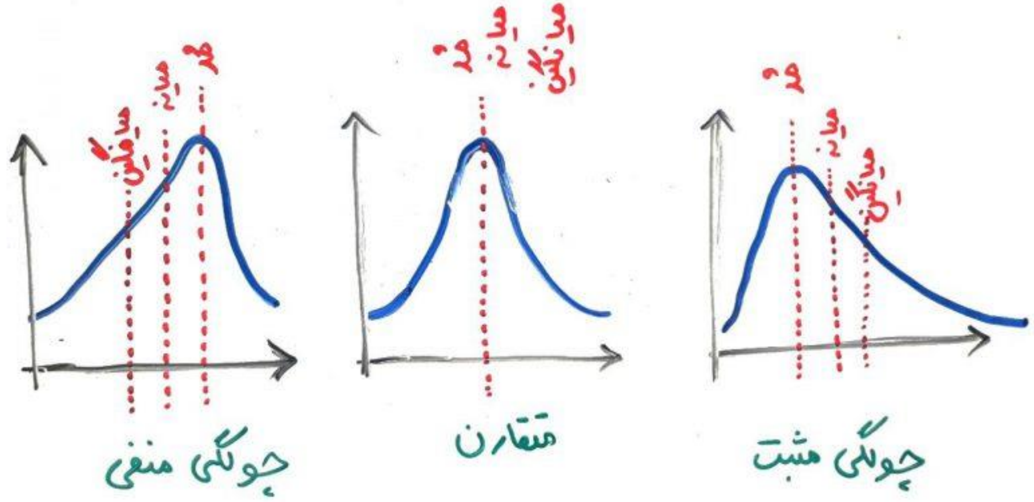


همان‌طور که مشاهده می‌کنید، توزیع داده به سمت راست کشیده شده است. تفسیر شکل این صورت است که تعدادی از راننده‌ها بوده‌اند که درآمدهای بالایی نسبت به عرف راننده‌ها (که همان بین ۴ تا ۶ میلیون است) داشته‌اند.

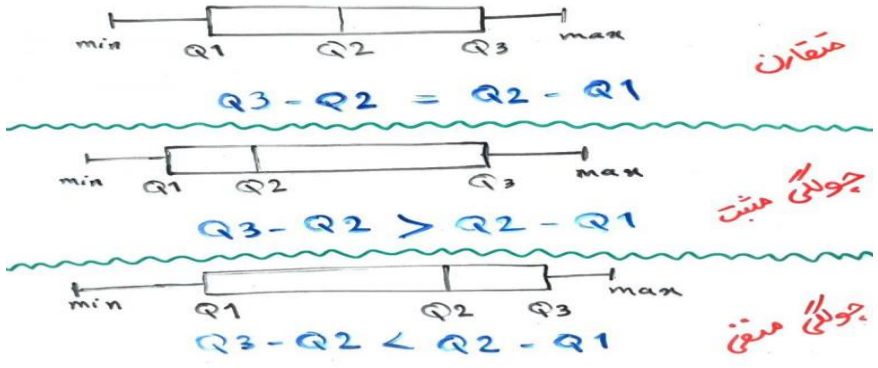
اگر داده‌ها به این صورت پراکنده شده باشند، می‌گوییم توزیع داده به سمت راست (مثبت) چولگی دارد.

چولگیSkewness در داده‌ها

به صورت کلی می‌توان چولگی داده‌ها را نسبت به یک توزیع نرمالِ ایده‌آل به صورت زیر نمایش داد:در حالتی که داده‌های ما به سمت راست چولگی داشته باشند، میانگین **mean**از مُد **mode** بیشتر است و در حالتی که داده‌های ما به سمت چپ چولگی داشته باشند، میانگین از مُد کمتر است. میانه **median** که در هر دو حالت بین میانگین و مُد قرار می‌گیرد.



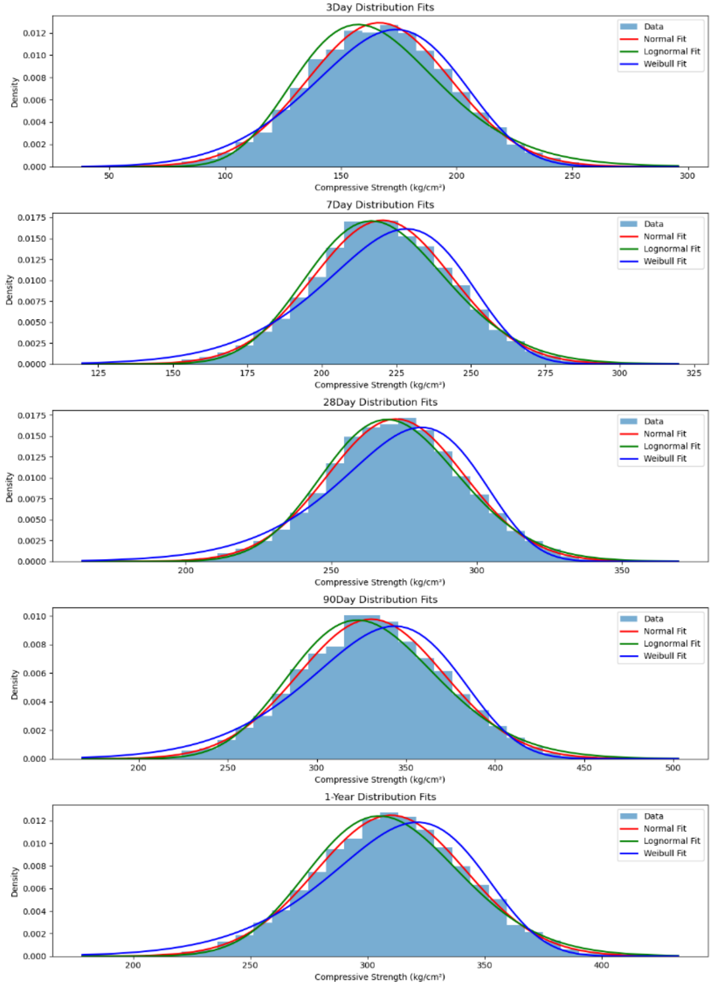
با استفاده از چارک‌ها **quartiles**نیز می‌توان چولگی داده‌ها را تشخیص داد. شکل زیر را در نظر بگیرید: در قسمت اول از شکل مقابل، داده‌ها نرمال هستند، چون چارک‌ها در حالت متوازن قرار دارند، ولی در قسمت دوم، مشخص می‌شود که داده‌ها به سمت راست چولگی دارد. در واقع اگر اختلاف Q3و Q2 از اختلاف Q2و Q1بیشتر باشد، به این نتیجه می‌رسیم که داده‌ها به سمت راست (مثبت) چولگی دارند و اگر مانند قسمت سوم، برعکسِ این اتفاق بیوفتد به این نتیجه می‌رسیم که داده ها به سمت چپ (منفی) چولگی دارند.



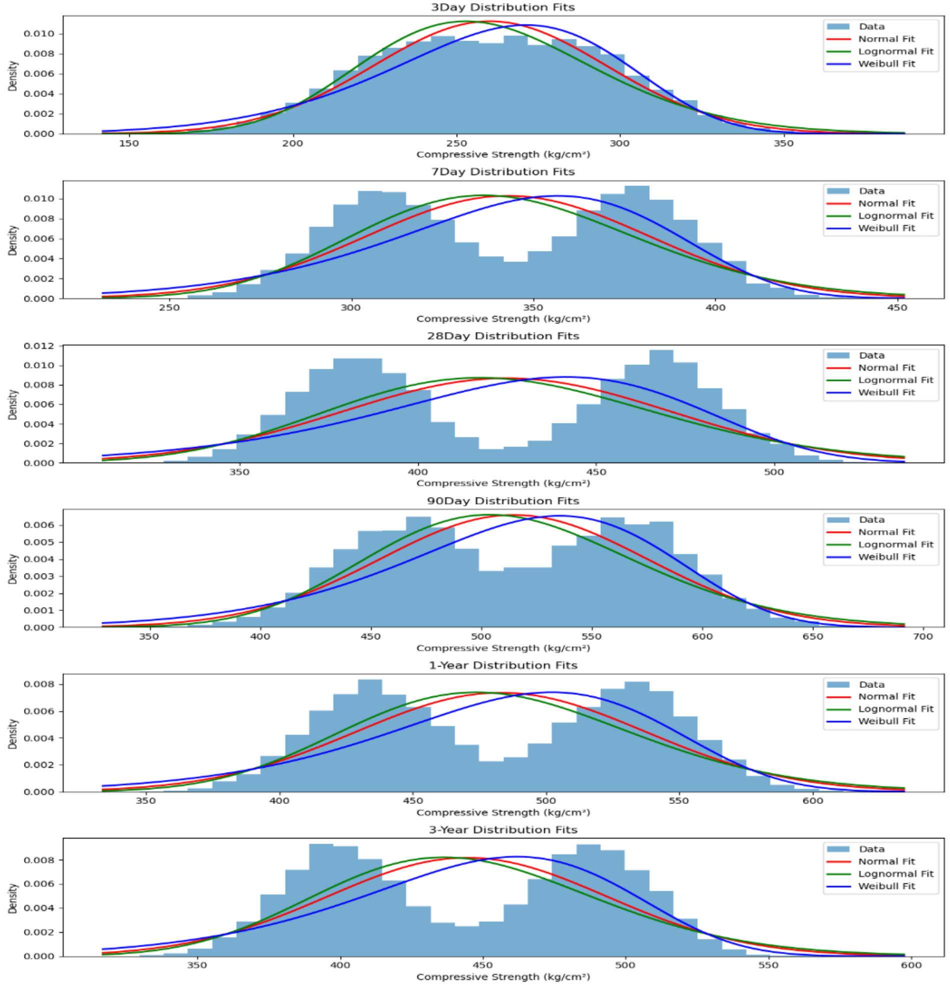
چولگی داده‌ها به ما می‌گوید که داده‌های پرت در کدام سمت (راست یا چپ) نسبت به میانگین قرار دارند. همچنین در فرآیندهای یادگیری ماشین نیز، **داده‌هایی** (ویژگی‌هایی) که **چولگی داشته باشند، می‌توانند منجر به افزایش خطا شوند.** برای همین می‌توان این داده‌ها با استفاده از تبدیل‌هایی مانند تبدیل توان **power transformation**به داده‌های نرمال ایده‌آل تبدیل کرد .

مقایسه آماری سه داده مختلف

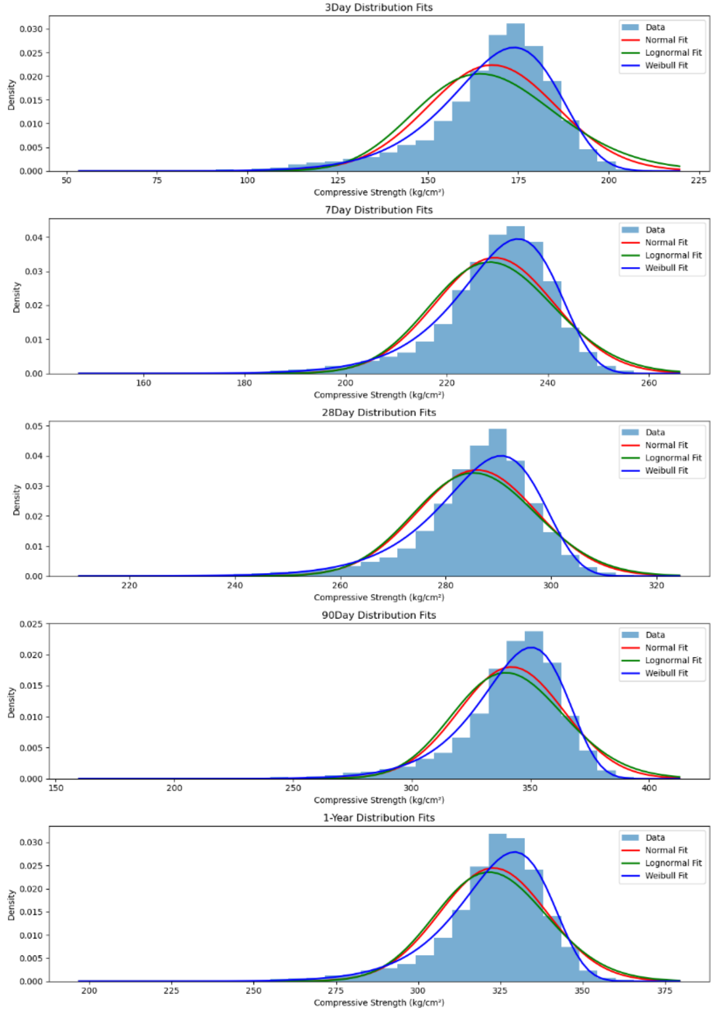
Distribution Fits for A6



Distribution Fits for A3

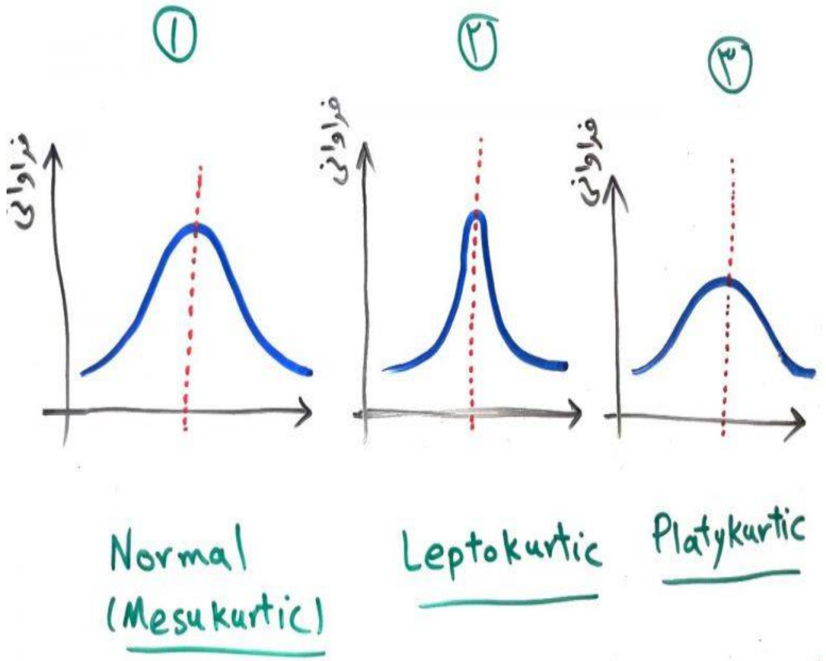


Distribution Fits for A5



کشیدگی یا برجستگی (kurtosis) در داده ها

در مباحث گذشته به معرفی **گشتاور اول** (میانگین)، **گشتاور دوم** (واریانس) و **گشتاور سوم** (چولگی) پرداختیم. در این بخش قصد داریم **گشتاور چهارم** از یک مجموعه‌ی داده را با هم مشاهده کنیم. **کشیدگی یا برجستگی kurtosis** گشتاور چهارم در یک مجموعه‌ی داده است که میزان برجستگی قله را در توزیع یک مجموعه‌ی داده مشخص می‌کند. فرض کنید سه مجموعه‌ی داده داریم با توزیع‌های زیر:



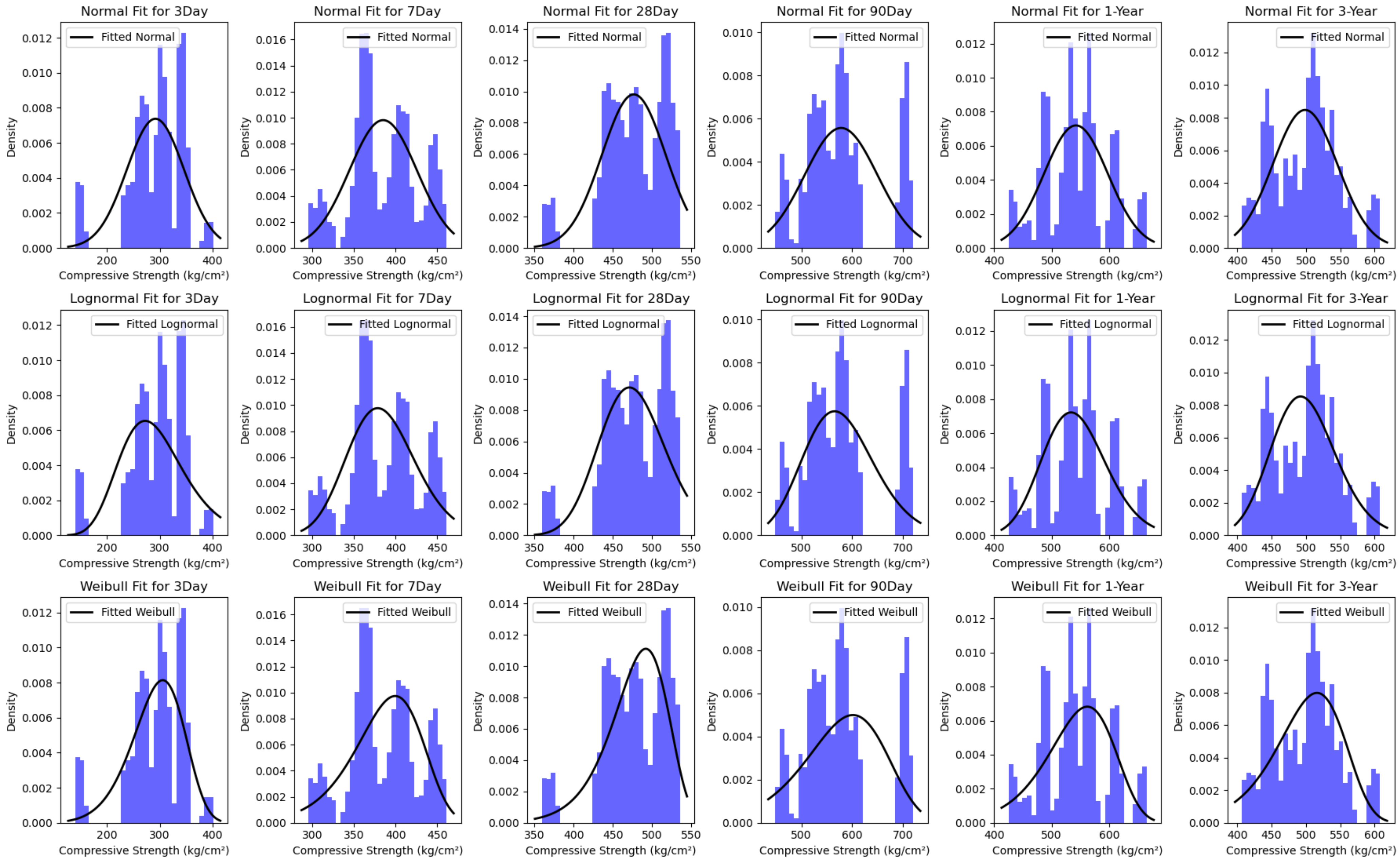
شکل شماره‌ی ۱ (سمت چپ) توزیعِ یک مجموعه‌ی داده‌ی نرمال است. یعنی قله‌ی آن همان مُد – mode خیلی کم یا خیلی زیاد نیست و به سمت راست یا چپ هم چولگی ندارد. اما شکل **شماره‌ی ۲** (وسط) قله یاهمان مُد mode **تیزتر** از حد معمول است.به این حالت که **قله تیزتر**از حد معمول باشد **برجستگی بالا**می‌گویند در حالی که شکل شماره‌ی ۳ (سمت راست) به دلیل اینکه **قله‌ی کوتاهی** دارد، **برجستگی پایینی** هم دارد. پس برجستگی به معنای تیزی، یا برجستگی قله (مُد) است و هر چقدر این مقدار بیشتر باشد، تیزیِ قله نیز بیشتر خواهد شد.

در آمار معمولاً به جای برجستگی kurtosis از برجستگی مازاد excess kurtosis استفاده می‌کنند. **در واقع برجستگی برای داده‌های کاملاً نرمال برابر ۳ است. ولی در برجستگی مازاد یا همان excess kurtosisما یک منهای ۳ (۳−) به فرمول اضافه می‌کنیم** تا برجستگیِ داده‌های نرمال (مانند شکل بالا سمت چپ) برابر با صفر شود. __بنابراین داده‌هایی که از حالت نرمال__ __قله‌ی__ __تیزتری دارند، دارای برجستگیِ مازاد excess kurtosis__ __مثبت__ __هستند و داده‌هایی که از حالت نرمال__ __پهن‌ترند____، برجستگیِ مازاد__ __منفی__ __خواهند داشت__.

مشخصات آماری یک مجموعه داده بتنی برای ۴ طرح اختلاط بتن

	MIX-1						MIX-2					
Age	Std	Min	Max	Standard error	Skewness	kurtosis	Std	Min	Max	Standard error	Skewness	kurtosis
3Day	54	140	402	0.54	-0.033	2.990	49	65	503	0.49	0.048	3.010
7Day	41	295	462	0.41	-0.039	2.988	37	224	540	0.37	0.011	3.002
28Day	41	359	536	0.41	0.065	3.013	37	286	583	0.37	-0.057	2.983
90Day	72	449	720	0.72	-0.063	2.981	66	306	772	0.66	-0.023	2.993
1Year	55	425	665	0.55	-0.095	2.972	50	304	686	0.5	0.067	3.013
3Year	47	407	608	0.47	-0.156	2.953	43	286	630	0.43	-0.025	2.993
	MIX-3						MIX-4					
Age	Std	Min	Max	Standard error	Skewness	kurtosis	Std	Min	Max	Standard error	Skewness	kurtosis
3Day	35	153	376	0.35	0.043	3.009	35	61	318	0.35	-0.016	2.995
7Day	39	242	442	0.39	-0.038	2.989	26	153	354	0.26	0.045	3.009
28Day	46	322	526	0.46	0.020	3.004	26	216	404	0.26	0.001	3.000
90Day	61	345	675	0.61	0.010	3.002	46	171	536	0.46	-0.025	2.993
1Year	54	347	621	0.54	-0.026	2.992	36	205	479	0.36	0.033	3.007
3Year	49	330	585	0.49	0.030	3.006	31	210	439	0.31	-0.042	2.987

Fits for A1


Normal Fit for 3Day
Normal Fit for 7Day
Normal Fit for 28Day
Normal Fit for 90Day
Normal Fit for 1-Year
Normal Fit for 3-Year
Fitted Normal
Lognormal Fit for 3Day
Lognormal Fit for 7Day
Lognormal Fit for 28Day
Lognormal Fit for 90Day
Lognormal Fit for 1-Year
Lognormal Fit for 3-Year
Fitted Lognormal
Weibull Fit for 3Day
Weibull Fit for 7Day
Weibull Fit for 28Day
Weibull Fit for 90Day
Weibull Fit for 1-Year
Weibull Fit for 3-Year
Fitted Weibull
Density
Compressive Strength (kg/cm²)

کشیدگی یا برجستگی (kurtosis) در داده ها

فرمول برجستگیِ مازاد به صورت زیر است:

میانگین

$$\text{excess kurtosis} = n \times \frac{\sum_{i}^{n} (x_i - \bar{x}_i)^4}{\left(\sum_{i}^{n} (x_i - \bar{x}_i)^2\right)^2} - 3$$

در کل با توجه به فرمول بالا، سه نوع برجستگی **kurtosis** در داده‌ها موجود است:

۱. برجستگی :**Mesukurtic**:اگر برجستگیِ مازاد یا همان **excess kurtosis**برابر صفر باشد یعنی قله در توزیع داده‌ها به صورت نرمال باشد، حالت **Mesukurtic**داریم.

۲. برجستگی :**Leptokurtic**:اگر برجستگیِ مازاد یا همان **excess kurtosis**بالاتر از صفر باشد، یعنی قله در توزیع داده‌ها تیزتر از حالت نرمال باشد، برجستگی **Leptokurtic**رخ داده است .

۳. برجستگی :**Platykurtic**:اگر برجستگیِ مازاد یا همان **excess kurtosis**کمتر از صفر باشد، یعنی قله در توزیع داده‌ها پهن‌تر از حالت نرمال باشد، برجستگی **Platykurtic**رخ داده است .

همان‌طور که درابتدای درس اشاره شد،به برجستگی **kurtosis** گشتاور چهارم نیز گفته می‌شود چون درفرمولش توان چهارموجوداست.برجستگی زیاد **leptokurtic**نشان می‌دهد که داده‌های ما به صورت متراکم نزدیک به قله جمع شده‌اند.

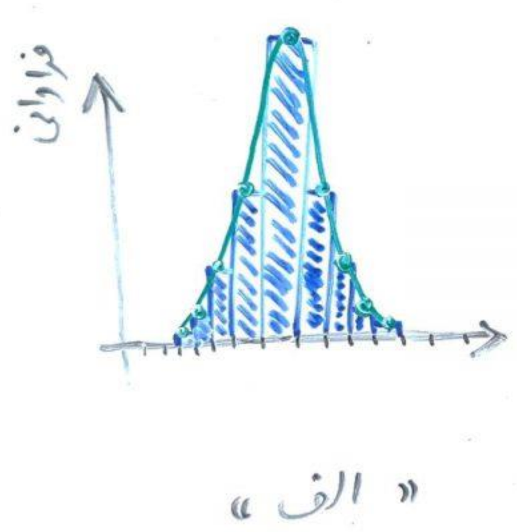
کشیدگی یا برجستگی (kurtosis) در داده ها

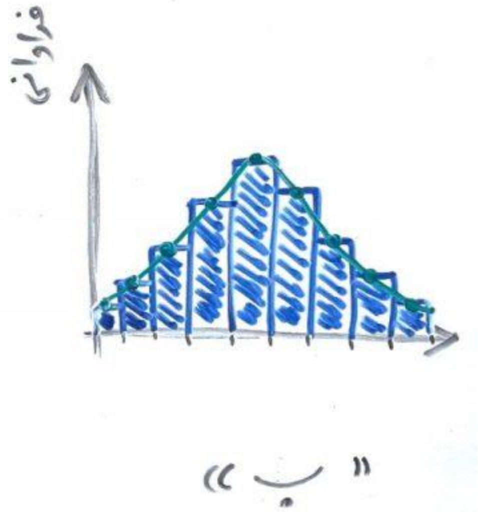
برای مثال فرض کنید شکل زیر، توزیع سود دو شرکت در سالیان مختلف است:

همان‌طور که ملاحظه می شود ، توزیع سود شرکت «الف»، به نوعی برجستگی بالا leptokurticمحسوب می‌شود در حالی که توزیع سود شرکت «ب»، برجستگی پایینی playkurticاست. سرمایه‌گذاران معمولاً با مشاهده‌ی شرکت «ب» درمی یابند که ریسک سرمایه‌گذاری در آن بالاست زیرا قابلیت پیش‌بینی کمتری دارد.

در واقع احتمال اینکه سود خیلی زیاد یا سود خیلی کم از آن بگیرند، بیشتر است.

ولی سرمایه‌گذاری در شرکت «الف» ریسک کمتری دارد چون پیش‌بینی‌پذیر تر است.البته برای مثال یک سرمایه‌گذار می‌تواند بخشی از پولش را در شرکت‌هایی سرمایه‌گذاری کند که سود آن‌ها leptokurticهستند و بخشی دیگر را برای اطمینان در شرکت‌هایی با توزیع سود platykurticسرمایه‌گذاری کند.



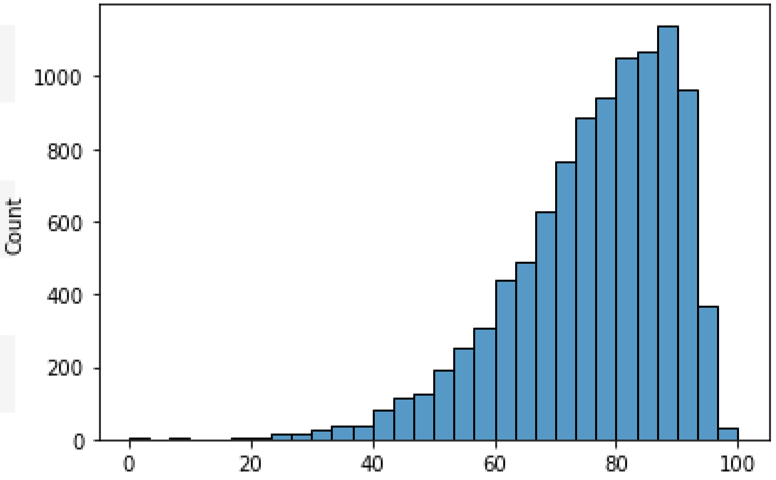


توزیع‌های آماری Statistical Distributions

توزیع آماری به **پراکندگی داده‌ها و فراوانیِ** هر کدام از مقادیر آن‌ها می‌گویند. با استفاده **از توزیعِ آماریِ یک متغیر، می‌توانیم به نحوه‌ی پراکندگی و احتمال هر کدام از قسمت‌های آن متغیر (در بازه‌ی پراکندگی)** پی ببریم.

	Age	Illness
0	42	1
1	87	1
2	75	1
3	77	1
4	83	0
...	...	...
9995	69	1
9996	86	1
9997	73	1
9998	88	1
9999	50	1

در مجموعه‌ی داده‌ی مقابل هر **سطر** یک **نمونه** است. اگر بخواهیم **توزیع** داده‌ها را برای متغیرِ سن Age نمایش دهیم، نموداری شبیه به شکل زیر ساخته می‌شود:



شکل مقابل یک نمودار فراوانی (هیستوگرام) برای متغیرِ سن Age است.این نمودار، توزیعِ متغیر سن Age را نمایش داده و **نشان می‌دهد** که در کدام یک از بازه‌های سنی، فراوانی و احتمال بیشتری داشته‌ایم. برای مثال در نمودار بالا بازه‌ی سن که بین ۰ تا ۱۰۰ سال بوده است را به ۳۰ قسمت مساوی تقسیم کرده محور xو تعدادِ تکرار (فرکانس – frequency) هر کدام از قسمت‌ها را در محور عمودی محور yنمایش داده‌ایم. هر کدام از این ۳۰ قسمت، به اندازه‌ی ۳/۳۳ بازه‌ی سن را نمایش می‌دهند. مثلاً در بازه‌ی سن ۰ تا ۳/۳۳ تعداد خیلی کمی نمونه داریم ولی در **بازه‌ی سنِ ۸۰ تا ۸۳/۳۳** بیش **از هزار نفر** نمونه در مجموعه‌ی داده وجود دارد.هر چقدر تعداد قسمت‌ها بیشتر باشد، نمودار دقیق‌تر می‌شود.برای مثال اگر متغیر سن Ageرا به ۵۰ قسمت تقسیم کنیم هر قسمت نشان دهنده‌ی تعداد تکرار برای بازه‌ی ۲ سال است (چون کل نمونه‌ها از ۰ تا ۱۰۰ سال سن دارند).

چگونه با EDA در داده‌ها اکتشاف کنیم؟

آمار و احتمالات به ما اجازه می‌دهند که داده‌های حجیم را خلاصه‌سازی کرده و به صورت ساده و قابل درک، در ذهن‌مان جای دهیم. با این کار می‌توانیم داده‌ها را درک کرده و با استفاده از درکِ داده‌ها، به ارزشی که در داده‌ها نهفته شده است دست پیدا کنیم. همچنین اکتشاف و درک داده‌ها، برای پیاده‌سازی و استفاده از الگوریتم‌های یادگیری ماشین نیز کمک می‌کند که در بخش های بعدی با آن‌ها بیشتر آشنا خواهیم شد.

فرض کنید می‌خواهید سیستمی بسازید که بتواند با استفاده از نتایج آزمایشِ قلب و عروق (داده‌های برگه‌ی آزمایشِ یک شخص)، پیش‌بینی کند که آیا یک شخص بیماری قلبی عروقی دارد یا خیر. این یادگیری باید از روی یک مجموعه‌ی داده‌ی آموزشی انجام شود. چند سطر از داده‌ها را می‌توانید در زیر مشاهده کنید:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

در مجموعه داده‌ی بالا، هر سطر یک شخص بوده که دارای ستون‌های مختلف است. در هر ستون یک ویژگی از یک شخص در دسترس است . مثلا ستونِ **id**، شناسه‌ی یک شخص است. ویژگیِ **age**،سن شخص (تعداد روز عمر) است. مثلاً شخصِ شماره‌ی صفر (۰)،۱۸۳۹۳ روز عمر کرده است و شخصِ شماره‌ی یک (۱)، ۲۰۲۲۸ روز عمر کرده است. ستونِ بعدی **gender**است که نشان‌دهده‌ی جنسیت (۱=مرد و ۲=زن) است. **height** و**weight** هم که دو ویژگی دیگر هستند که به ترتیب به قد و وزن اشخاص اشاره دارند.

چگونه با EDA در داده‌ها اکتشاف کنیم؟

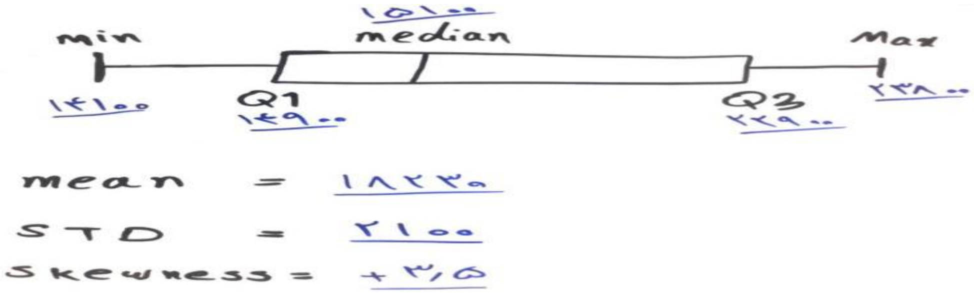
دو ستونِ بعدی در مجموعه‌ی داده‌ی بالا، یعنی ap_hi و ap_lo به ترتیب فشار خون سیستولیک و دیاستولیک (فشار خون اول و دوم) است که توسط پزشک از شما گرفته می‌شود. ستونِ بعدی یا همان cholesterol میزان کلسترل موجود در هر شخص را نشان می‌دهد. ستون بعدی یا همان gluc هم نشان‌دهنده‌ی میزان گلوکز موجود در شخص است. دو ستونِ بعدی یعنی smoke و alco نشان‌دهنده‌ی این است که آیا شخص سیگار/قلیان مصرف می‌کند یا خیر و همچنین آیا شخص الکل مصرف می‌کند یا خیر. ستون بعدی یعنی active هم به این معناست که آیا شخص فعالیت فیزیکی در طول روز دارد یا خیر؟

ستون آخر cardio را پزشک تعیین کرده است و نشان‌دهنده‌ی این است که آیا شخص، بیماریِ قلبی عروقی داشته یا خیر؟ یک الگوریتم یادگیری ماشین، با استفاده از این داده‌ها می‌تواند یاد بگیرد که اشخاص با ویژگی‌های مختلف (ستون‌های اول تا یازدهم)، آیا بیماری قلبی خواهد داشت یا نه. در واقع الگوریتم یادگیری ماشین می‌تواند با مشاهده‌ی این مجموعه، که به آن مجموعه‌ی آموزشی می‌گویند، علمِ پزشک را یاد بگیرد و اگر الگوریتم به دقت مناسبی برسد، دیگر نیازی به پزشک نیست و خود الگوریتم با دیدن نتایج آزمایشات (ستون‌های اول یا یازدهم) می‌تواند تشخیص دهد که آیا یک شخص بیماری قلبی عروقی دارد یا خیر. در واقع الگوریتم با دیدن یازده ستون اول، می‌توانید ستون آخر cardio را پیش‌بینی کند و این کار با استفاده از مجموعه‌ی داده‌ی بالا انجام می‌شود.

اگر مدلِ CRISP را در داده‌کاوی دیده باشید، می‌دانید که یکی از اولین اقدامات برای انجام عملیات داده‌کاوی، درک و اکتشاف در داده‌هاست. پس ما برای فعالیت‌های بعدی، بایستی ابتدا داده‌ها را کشف و شناسایی کنیم. در مثال بالا ما تعدادی ستون داریم که هر کدام حاوی اطلاعاتی از اشخاص است. برای اینکه بتوانیم کل این داده‌ها را کشف کنیم، بایستی ابتدا تک تک ستون‌ها (ویژگی‌ها) را شناسایی کنیم. اما چگونه؟ برای مثال آیا می‌شود تمام داده‌ها در ستون age را در حافظه‌ی ذهن خودمان بسپاریم؟ قطعا خیر.

چگونه با EDA در داده‌ها اکتشاف کنیم؟

برای همین بایستی اطلاعات آماری هر کدام از ستون‌ها را به دست بیاوریم. با استفاده از اطلاعات آماری هر ستون (ویژگی) می‌توانیم آن ستون را خلاصه کرده و تحلیل کنیم. برای مثال، فرض کنید با استفاده از دروس قبلیِ همین دوره، ستونِ age سن را از داده‌های مجموعه‌ی داده‌ی بالا به صورت زیر خلاصه‌سازی کرده‌ایم:



مشاهده می‌کنید که از بین تمامی اشخاصی که در جدول اول قسمتی از آن‌ها را نشان دادیم، میانگین سن برابر ۱۸۲۳۰ شده است. انحراف استاندارد برای age برابر ۲۱۰۰ شده و داده‌ها چولگی skewness به سمت راست (مثبت) داشته‌اند. مشاهده می‌کنید که به این ترتیب با همین چند عدد ساده، متوجه می‌شویم که افرادِ حاضر در این مجموعه‌ی داده حدوداً ۵۰ ساله هستند (۱۸۲۳۰) و بیشتر به سمت سن‌های بالا پراکنده شده‌اند (چولگی مثبت). پس این مجموعه‌ی داده و مدلِ یادگیریِ ماشینی که روی آن ساخته می‌شود، احتمالاً نتواند برای افراد جوان کاربرد داشته باشد.

مشاهده می‌کنید که با همین چند آمار ساده، توانستیم یکی از ستون‌ها (ویژگی‌ها) که همان age بود را کمی بیشتر درک کنیم. در واقع اکتشاف در داده‌ها انجام دادیم. اگر همین‌کار را برای ستون‌های دیگر هم انجام دهیم، خواهیم توانست این مجموعه‌ی داده را برای خود خلاصه‌سازی کرده و بعد از آن تحلیل‌های مناسب را بر روی داده‌ها انجام دهیم. قطعاً با استفاده از معیارهای آماریِ پیچیده‌تر می‌توانیم هر کدام از ستون‌ها را بیشتر کشف کرده و تحلیل‌های پیچیده‌تری ارائه نماییم.

نمونه‌گیری آماری و محاسبه‌ی حداقل تعداد نمونه Min Sample Size

فرض کنید می‌خواهیم نظر مردم شهر اهواز را در مورد شهردار این شهر بدانیم. یک پرسشنامه طراحی می‌کنیم وآن را به تعدادی از شهروندان اهوازی داده تا به آن پاسخ دهند. اولین سوالی که احتمالاً ذهن ما را درگیر خود می‌کند، این است که از چه تعداد از شهروندان بخواهیم پرسشنامه را پاسخ دهند؟ پاسخ واضح است، هر چه بیشتر، بهتر. اما هر چقدر تعداد افرادی بیشتری در پاسخ به پرسشنامه درگیر باشند، زمان و هزینه‌ی بیشتری نیز بایستی صرف نظرخواهی از شهروندان شود. پس به دنبال راهی هستیم که حداقل تعداد نمونه‌ی مناسب که نظر آنها بیان‌گر نظر کل مردم شهر باشد را پیدا کنیم.به این کار محاسبه‌ی حداقل تعداد نمونه **minimum sample siz** می‌گویند که کاربردهای متعددی در پردازش داده‌ها دارد.

پیدا کردن نمونه‌ی مناسب با یک فرمول نسبتاً ساده آماری جمعیت محدود به دست می‌آید (فرمول‌های متعددی برای نمونه‌گیری موجود است که ما به یکی از ساده‌ترین آنها اشاره می‌کنیم):

عدد Z   انحراف استاندارد

$$\text{Sample Size} = \frac{(Z)^2 \times STD \times (1-STD)}{(\text{Confidence Interval})^2}$$

بازه اطمینان

۱. <u>بازه‌ی اطمینان Confidence Interval یا حاشیه‌ی خطا Margin of Error</u>

همان‌طور که از اسم آن مشخص است، <u>بازه‌ی</u> اطمینان، <u>مقدار عدم اطمینان</u> یک نمونه را نسبت به جمعیت کل مشخص می‌کند. <u>به زبان ساده یعنی چقدر به این نمونه شک داریم</u>. به بیان دیگر، مقدار <u>بازه‌ی اطمینان یا حاشیه‌ی خطا به ما می‌گوید که چقدر اطمینان داریم نمونه‌ی گرفته شده، به کل جمعیت قابل تعمیم باشد</u>.

نمونه‌گیری آماری و محاسبه‌ی حداقل تعداد نمونه Min Sample Size

برای مثال فرض کنید در همان مثال بالا (پرسشنامه‌ی میزان رضایت از شهردار اهواز)، بازه‌ی اطمینان را بر روی ۵ درصد تنظیم کنیم و نظر شهروندان (که به صورت نمونه انتخاب شده‌اند) هم ۶۰ درصد رضایت از شهردار باشد. به این ترتیب می‌توانیم بگوییم که نظر کل مردم شهر (جمعیت کل)، بین بازه‌ی ۵۷ تا ۶۳ (پنج‌درصد کمتر (−)از شصت تا پنج‌درصد بیشتر(+) از شصت) درصد رضایت است. در واقع برای تعمیم از نمونه به جمیعت، ۵ درصد خطا داریم.

۲. سطح اطمینان Confidence Level

سطح اطمینان میزان احتمال درستی در تکرارهای زیاد از آزمایش است. برای مثال اگر سطح اطمینان را برابر ۹۵ درصد بگذاریم، در همان مثال قبلی (پرسشنامه‌ی میزان رضایت از شهردار اهواز)، به این معناست که در صد بار تکرار این آزمایش (صد بار که از گروه‌های مختلف شهروندان، نمونه گرفته و پرسش کنیم)، ۹۵ مرتبه، همان عددی به دست خواهد آمد که از جمیعت کل خواهیم گرفت.

هر چقدر این عدد بالاتر باشد، طبیعتاً تعداد نمونه‌های لازم نیست بیشتر است ولی معمولاً این عدد را ۹۵ یا ۹۹ می‌گذارند. در فرمول بالا، سطح اطمینان وجود ندارد. در واقع بایستی با استفاده از جدول Z (که یک جدول معروف آماریست)، سطح اطمینان را به عدد Zتبدیل کنیم. جدول Zدر کتاب‌های آماری وجود دارد و فعلاً نمی‌خواهیم به آن بپردازیم ولی برای همین درس، یک قسمت کوچک تبدیل سطح اطمینان به Zرا در جدول زیر مشاهده می‌کنید:

مقدار Z	سطح اطمینان
۱.۲۸	۸۰٪
۱.۶۵	۹۰٪
۱.۹۶	۹۵٪
۲.۵۸	۹۹٪

برای مثال اگر بخواهیم سطح اطمینان را ۹۵ قرار دهیم، در فرمول بالا به جای Z، عدد ۱/۹۶ قرار می‌گیرد.

نمونه‌گیری آماری و محاسبه‌ی حداقل تعداد نمونه Min Sample Size

۳. جمعیت کل و انحراف استاندارد

جمعیت کل هر چقدر بیشتر باشد، معمولاً تعداد نمونه‌های مورد احتیاج بیشتر است. البته در فرمول گفته شده در بالا، جمعیت در نظر گرفته نشده است. فرمول‌های دیگری وجود دارد که جمعیت را نیز دخیل می‌کند (البته تغییر زیادی در نتیجه حاصل نمی‌شود).

معمولاً انحراف استاندارد برابر ۰/۵ در نظر گرفته می‌شود ولی اگر کسی بخواهد دقت زیادی به خرج دهد می‌تواند انحراف استاندارد نسبت جمعیت را نیز حساب کند.
حال که عناصر مختلف فرمول را متوجه شدیم، اجازه دهید مثال کاربردی خود را در مورد نظر شهروندان اهواز نسبت به شهردار این شهر، ادامه دهیم. فرض کنید

مثال : جمعیت کل شهر ۳۰۰۰۰۰۰ نفر باشد (در اینجا انحراف استاندارد را ۰/۵ می‌گذاریم). بازه‌ی اطمینان یا همان حاشیه‌ی خطا را برابر ۵٪ و سطح اطمینان را برابر ۹۵٪ قرار می‌دهیم. یعنی عدد Z بر اساس جدول بالا، برابر ۱/۹۶ خواهد بود. حال اعداد را درون فرمول جایگذاری می‌کنیم

عدد Z

انحراف استاندارد

$$\text{Sample Size} = \frac{(۱٫۹۶)^{۲} \times ۰/۵ \times (۱-۰/۵)}{(۰/۰۵)^{۲}}$$

۳۸۴٫۱۶

بازه اطمینان

نمونه‌گیری آماری و محاسبه‌ی حداقل تعداد نمونه Min Sample Size

همان‌طور که مشاهده می‌کنید حداقل تعداد نمونه‌های لازم با خطای ۵ درصد و سطح اطمینان ۹۵ درصد، برابر **۳۸۵** نفر است. یعنی اگر از ۳۸۵ شهروند نظرسنجی کنیم، با تعمیم خوبی، می‌توانیم این نظرسنجی را به کل شهروندان شهر اهواز تعمیم دهیم.

البته چند نکته در نمونه‌گیری باید مورد توجه قرار گیرد. اول اینکه نمونه‌های انتخاب شده کاملاً تصادفی باشند. مثلاً در همان مثال شهروندان اهوازی، از یک محله‌ی خاص نظرسنجی نکنیم. در این مثال بایستی احتمال انتخاب یک شهروند برای نظرسنجی دقیقا با احتمال انتخاب هر شهروند دیگر برابر باشد. در واقع نمونه‌گیری کاملاً تصادفی باشد. ثانیاً نمونه‌های انتخاب شده در همان جمعیت باشند. مثلاً از شهروندان شهرکرد، در نمونه‌های ما برای نظرسنجی نباشد (که خوب این کاملاً بدیهی است).نمونه‌گیری آماری خود یک علم (یا حداقل یک زیرحوزه مهم) است که در این درس یک مثال ساده و کاربردی از آن را با هم مرور کردیم. همان‌طور که می‌دانید نمونه‌گیری آماری کاربردهای متعددی (علاوه بر مثال گفته شده در بالا) دارد. مثلاً اگر با داده‌کاوی و یادگیری ماشین آشنا باشید و بخواهید چندین الگوریتمِ مختلفِ طبقه‌بندی را با پارامترهای مختلف بر روی یک مجموعه‌ی حجیم داده، آزمایش کنید، چون حجم داده‌ها بالاست، برای آزمایشِ هر کدام از الگوریتم‌ها زمان زیادی بایستی صرف شود. در این مثال می‌توانیم یک نمونه‌ی مناسب و کوچکتر از داده‌ها به دست بیاوریم و الگوریتم‌هایمان را بر روی این نمونه‌ی کوچک آزمایش کنیم. بعد از نتیجه‌گیری اولیه، الگوریتم‌های کاندید با پارامترهای مناسب را بر روی داده‌های اصلی (با حجم بالا) برازش می‌کنیم.

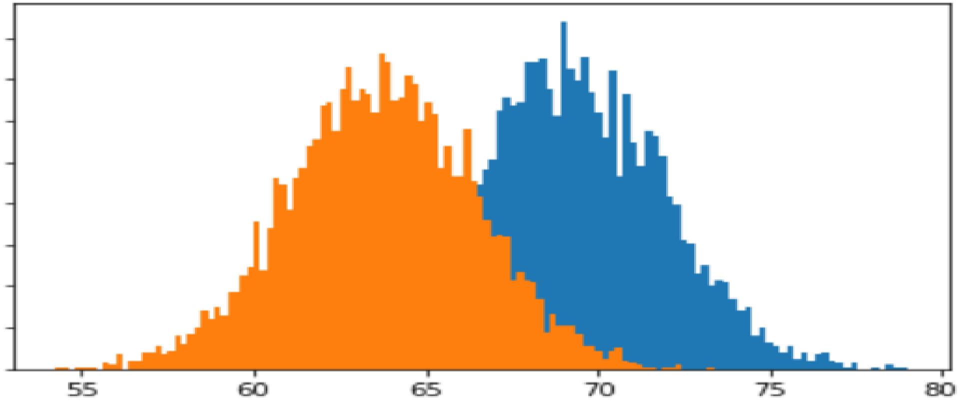
همچنین در آدرس ssc.chistio.ir یک نرم‌افزار آنلاین ایجاده شده که همین پارامترها را گرفته و تعداد حداقل نمونه را به عنوان نتیجه برمی‌گرداند. این نرم‌افزار از یک فرمول دیگر (شبیه به فرمول بالا) استفاده کرده است که تعداد جمعیت کل را نیز برای محاسبه احتیاج دارد (البته الزامی نیست)

فاصله‌ی آماری Statistical Distance و کاربردهای آن

روش‌های متعددی برای مقایسه‌ی دو مجموعه‌ی داده با یکدیگر وجود دارد. مثلاً می‌توان میانگین آنها را با یکدیگر بررسی و یا پراکندگی (واریانس) آنها را مورد مقایسه قرار داد. اما هر کدام از این مقایسه‌ها قسمتی از حقیقت را پنهان می‌کنند. به همین دلیل معیارهایی با نام فاصله‌ی آماری به وجود آمده است که با استفاده از آنها بتوان مجموعه داده‌های مختلف و یا متغیرهای متفاوت را با یکدیگر مقایسه کرد.

فرض کنید دو گروه ورزشکار داریم. یک گروه شناگران و گروه دوم بازیکنان واترپلو هستند. می‌خواهیم وزن هر یک از ورزشکاران را محاسبه کرده و ببینیم که آیا وزن ورزشکاران گروه شناگر با وزن ورزشکاران واترپلو تفاوت دارد یا خیر؟ برای این کار می‌توانیم میانگین وزن شناگران را از میانگین وزن بازیکنان واترپلو کم کرده و اختلاف آنها را به دست بیاوریم. ولی این روش به خوبی نمی‌تواند فاصله‌ی دو مجموعه را نشان دهند. زیرا در این روش ما فقط به یک معیار (میانگین) بسنده کرده‌ایم و از معیارهای دیگر مانند پراکندگی، چولگی و... صرف نظر کرده‌ایم. برای همین بهتر است برای مقایسه‌ی این دو مجموعه از معیارهای فاصله‌ی آماری استفاده کرد.

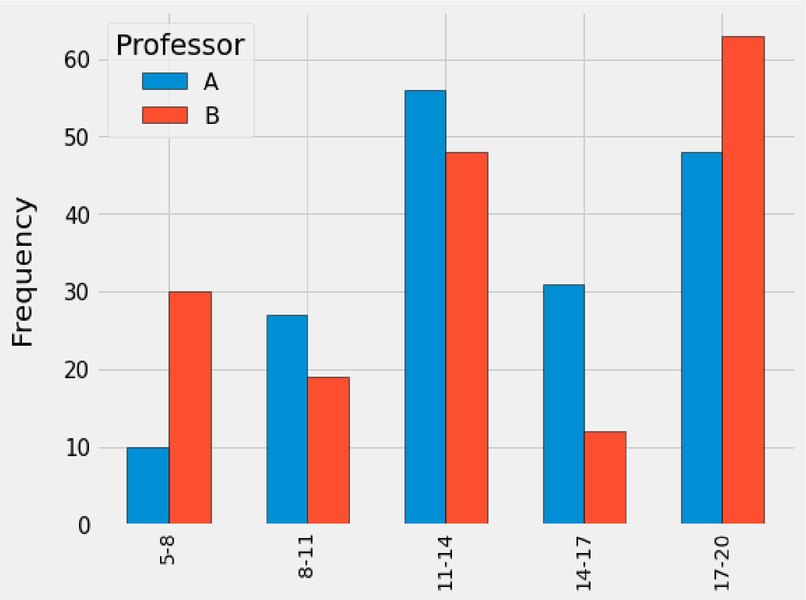
برای محاسبه‌ی فاصله‌ی آماری نیاز به توزیعِ آماریِ هر مجموعه داریم. در درس قبل در مورد توزیع آماری و نحوه‌ی ساخت آن صحبت کردیم. حال فرض کنید وزنِ ورزشکارانِ مثال بالا که در دو گروه قرار داشتند را به صورت توزیع آماری نمایش دهیم. چیزی مانند شکل زیر:



واگرایی کولبک–لیبلر Kullback–Leibler Divergence و کاربرد آن در فاصله‌ی آماری

در این بخش می‌خواهیم یکی از معیارهای فاصله‌ی آماری به نام واگرایی کولبک–لیبلر که به صورت مخفف واگراییِ KL نیز خوانده می‌شود بپردازیم و ببینم که چگونه می‌توان با استفاده از این معیار، فاصله‌ی بین دو مجموعه‌ی داده (شباهت دو توزیع آماری) را به صورت آماری محاسبه کرد.

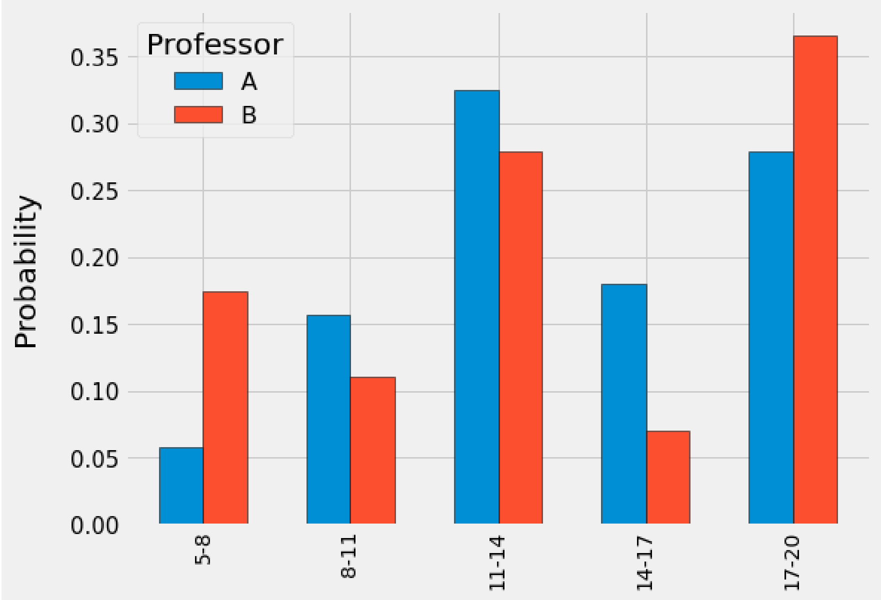
فرض کنید در دانشگاهی هستیم و یک استاد مدعو برای درس مبانی کامپیوتر به دانشگاه می‌آید. می‌خواهیم ببینیم که آیا نمراتی که این این استادِ مدعو به دانشجویان می‌دهد با استادِ اصلی این درس که هیئت علمی دانشگاه بوده و در ترم‌های گذشته درس مبانی کامپیوتر را تدریس می‌کرده، تفاوتی دارد یا خیر. برای این کار دو مجموعه‌ی داده از نمرات درس مبانی کامپیوتر در دو گروه دانشجویان جمع‌آوری می‌کنیم. گروه اول دانشجویانِ استاد اصلی درس $\mathbf{A}$ و گروه دوم دانشجویان استاد مدعو $\mathbf{B}$ در درس مبانی کامپیوتر هستند و توزیع نمرات هر گروه در شکل زیر مشخص شده است:



درشکل مقابل، توزیع نمرات گروه $\mathbf{A}$ استاد اصلی با رنگ آبی و توزیع نمرات گروه $\mathbf{B}$ استاد مدعو با رنگ قرمز مشخص شده است. برای درک بهتر، نمودار هیستوگرام را جدا جدا برای هر بازه رسم شده است. توجه داشته باشید که برای مقایسه‌ی دو توزیع بایستی بازه‌ی تقسیم‌شده برای دو گروه شبیه به هم باشد. برای مثال در شکل بالا نمرات از بازه‌ی ۵ تا ۲۰ قرار داشته‌اند که ما این بازه را به ۵ قسمت تقسیم کرده‌ایم. مثلاً تعداد ۱۰ دانشجو در گروه $\mathbf{A}$ ، نمره‌ای در بازه‌ی ۵ تا ۸ گرفته‌اند و در گروه $\mathbf{B}$ برای همین بازه (۵ تا ۸)، تعداد ۳۰ دانشجو موجود بوده‌اند.

واگرایی کولبک-لیبلر Kullback–Leibler Divergence و کاربرد آن در فاصله‌ی آماری

اگر نمرات بخواهیم دو گروه A و B را با استفاده از واگرایی KL با یکدیگر مقایسه کنیم، ابتدا بهتر است آن‌ها را به حالت احتمالی تبدیل کنیم. یعنی به جای تعداد تکرار frequency در محور عمودی، احتمال هر کدام از قسمت‌ها قرار بگیرد. برای این کار کافیست تعدادِ تکرار هر قسمت در هر گروه را تقسیم بر جمع کل همان گروه کنیم تا شکلِ بالا به شکل زیر تبدیل شود:



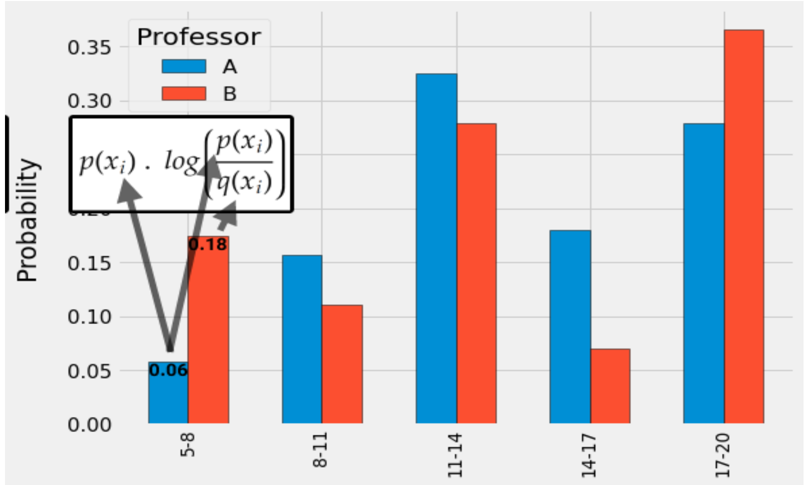
حال بایستی با استفاده از فرمولِ واگراییِ KL، دو توزیع را به صورت نظیر به نظیر در هر قسمت با یکدیگر مقایسه کنیم. فرمول واگرایی KL به صورت زیر است:

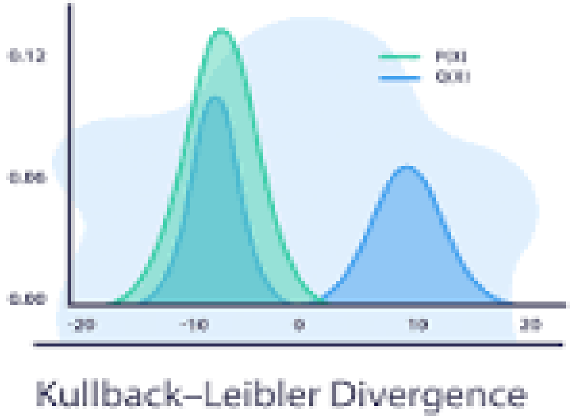
$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \,.\, \log\left(\frac{p(x_i)}{q(x_i)}\right)$$

در فرمول بالا مشخص است که اگر بخواهیم میزان اختلاف بین توزیع p با q را به دست بیاوریم می باید، برای هر کدام از قسمت‌های توزیع p، احتمالِ هر قسمت p را مطابق فرمولِ بالا با احتمال متناظر توزیع q محاسبه کنیم. برای مثال در نمونه‌ی بالا برای قسمت اول، فرمول به صورت زیر محاسبه می‌شود:

واگرایی کولبک–لیبلر Kullback–Leibler Divergence و کاربرد آن در فاصله‌ی آماری

در شکل بالا مشاهده می‌کنید که برای یک قسمت از پنج قسمت به جای احتمال **p** و **q** می‌توانیم مقادیر مناسب را قرار دهیم. در اینجا **p** گروه **A** نمرات استاد اصلی است و **q** گروه **B** نمرات استاد مدعو. همین کار را برای تمامیِ قسمت ها انجام داده و سپس نتیجه را با یکدیگر جمع می‌کنیم. خروجی نهایی، همان معیار واگراییِ **KL** خواهد شد.





هر چقدر عددِ به دست آمده در معیارِ واگرایِ **KL** بزرگتر باشد به این معنی است که توزیع **q** کمتر به توزیع **p** شباهت دارد. مثلاً در این مثال می‌توانیم بگوییم که اگر **KL** بین نمرات استاد مدعو و استاد اصلی درس بیشتر از یک عدد خاص (مثلا ۰/۲۰) شد، نمرات بایستی توسط یک استادِ سوم مورد بازبینی قرار بگیرد.

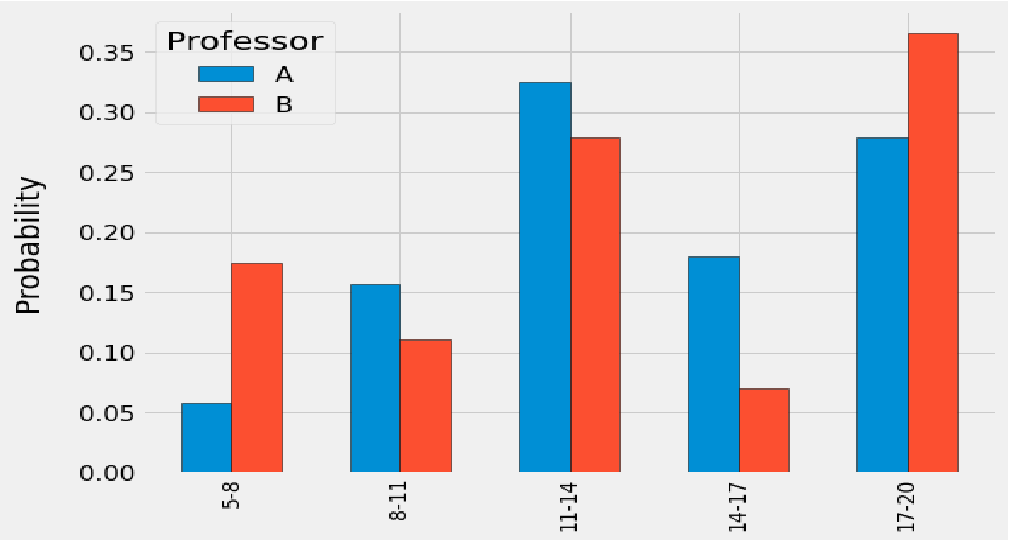
توجه کنید که به معیارِ واگرایی **KL**، نمی‌توانیم معیار فاصله **distance** بگوییم. چون واگرایی **p** از **q** با واگرایی **q** از **p** برابر نیست. به همین دلیل است که به جای واژه‌ی فاصله در فرمول از واژه‌ی واگرایی **divergence** استفاده می‌شود.

البته کاربردهای واگرایی **KL** بسیار زیاد بوده و در تولید الگوریتم‌های یادگیری ماشین و یادگیری عمیق نیز استفاده‌هایی از این معیار می‌شود.

فاصله‌ی جنسون–شنون Jenson–Shannon برای مقایسه‌ی توزیع‌های آماری

در درس گذشته در مورد واگرایی کولبک–لیبلر مطالبی بیان گردید که، این معیار واگرایی می‌تواند شباهت دو توزیع آماری را نمایش دهد. فاصله‌ی جنسون–شنون Jenson–Shannon نیز مانند کولبک–لیبلر رفتار می‌کند با این تفاوت که این فاصله بین دو توزیع آماری، تقارن دارد. یعنی فاصله‌ی توزیعِ اول و دوم برابر با فاصله‌ی توزیع دوم و اول است در حالی که در واگرایی کولبک–لیبلر این خاصیت برقرار نیست.

فرض کنید داده‌هایی شبیه به مثال قبل در درسِ کولبک–لیبلر داریم. در این داده‌ها فرض بر این است که در دانشگاهی هستیم و یک استاد مدعو برای درس مبانی کامپیوتر به دانشگاه می‌آید. می‌خواهیم ببینیم که آیا نمراتی که این استادِ مدعو به دانشجویان می‌دهد با استادِ اصلی این درس که هیئت علمی دانشگاه بوده و در ترم‌های گذشته درس مبانی کامپیوتر را تدریس می‌کرده، تفاوتی دارد یا خیر. برای این کار دو مجموعه‌ی داده از نمرات درس مبانی کامپیوتر در دو گروه دانشجویان جمع‌آوری می‌کنیم. گروه اول دانشجویانِ استاد اصلی درس A و گروه دوم دانشجویان استاد مدعو B در درس مبانی کامپیوتر هستند و توزیع نمرات هر گروه در شکل زیر مشخص شده است:



:

:

فاصله‌ی جنسون-شنون Jenson-Shannon برای مقایسه‌ی توزیع‌های آماری

داده‌های بالا مانند درسِ قبلی به حالت احتمالی تبدیل شده‌اند. حال اگر بخواهیم فاصله‌ی بین توزیعِ نمراتِ کسب شده توسط استادِ اول **A** و استادِ دوم **B** را بسنجیم، علاوه بر فرمولِ کولبک-لیبلر، می‌توانیم از فرمول جنسون-شنون نیز استفاده کنیم. این فرمول از واگرایی کولبک-لیبر استفاده می‌کند:

واگرایی کولبک-لیبلر

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}\left(p \,||\, \frac{p+q}{2}\right) + \frac{1}{2} D_{KL}\left(q \,||\, \frac{p+q}{2}\right)$$

در واقع فاصله‌ی جنسون-شنون برابر با جمعِ وزنیِ واگرایی کولبک-لیبلر است. با این کار، معیار جنسون شنون تبدیل به یک معیار فاصله شده که قابلیت تقارن دارد. یعنی فاصله‌ی بین توزیع الف و ب برابر با فاصله‌ی بین توزیع ب و الف است.

هر کدام از معیارهای جنسون-شنون یا کولبک-لیبلر در مسائل مختلف می‌توانند مورد استفاده قرار بگیرند. معمولاً اگر تعداد ویژگی‌ها (ابعاد) مسئله بالا باشد بهتر است از جنسون-شنون به جای کولبک-لیبلر استفاده شود. برای مثال در شبکه‌های عصبی عمیق مولد متخاصم از معیار جنسون-شنون استفاده می‌شود.