



آمار مقدماتی

در این دوره خواهید دید...

مقدمه

جمعیت

نمونه

اندازه گیری و مقیاس سازی

مقیاس های استیونز

متغیرها

داده ها

سرراست کردن داده پیوسته

جدول های اماری

فراوانی و فراوانی نسبی

فراوانی انباشته و فراوانی نسبی انباشته

جدول فراوانی

نمودار های اماری

نمودار دایره ای

نمودار میله ای

نمودار جعبه ای

نمودار خطی

نمودار نقطه ای

نمودار پشته ای

نقشه گرمایی

هیستوگرام

چندبر فراوانی

چندبر فراوانی انباشته

منحنی های فراوانی و

فراوانی انباشته

منحنی فراوانی نرمال

همچنین...

معیار های تمرکز

میانگین

میان

چندک ها

چندک ها داده های گسسته

چندک ها داده های پیوسته

نما

مقایسه معیار های تمرکز

میانگین اصلاح شده

معیار های پراکندگی

برد

میانگین انحراف ها

میانگین انحرافها

واریانس و انحراف استاندارد

روش تبدیل یا روش کوتاه برای محاسبه

میانگین واریانس

داده های تبدیل شده

داده های استاندارد

ضریب تغییر

نیم برد چارک ها

چولگی و برجستگی

گشتاور و گشتاور مرکزی داده

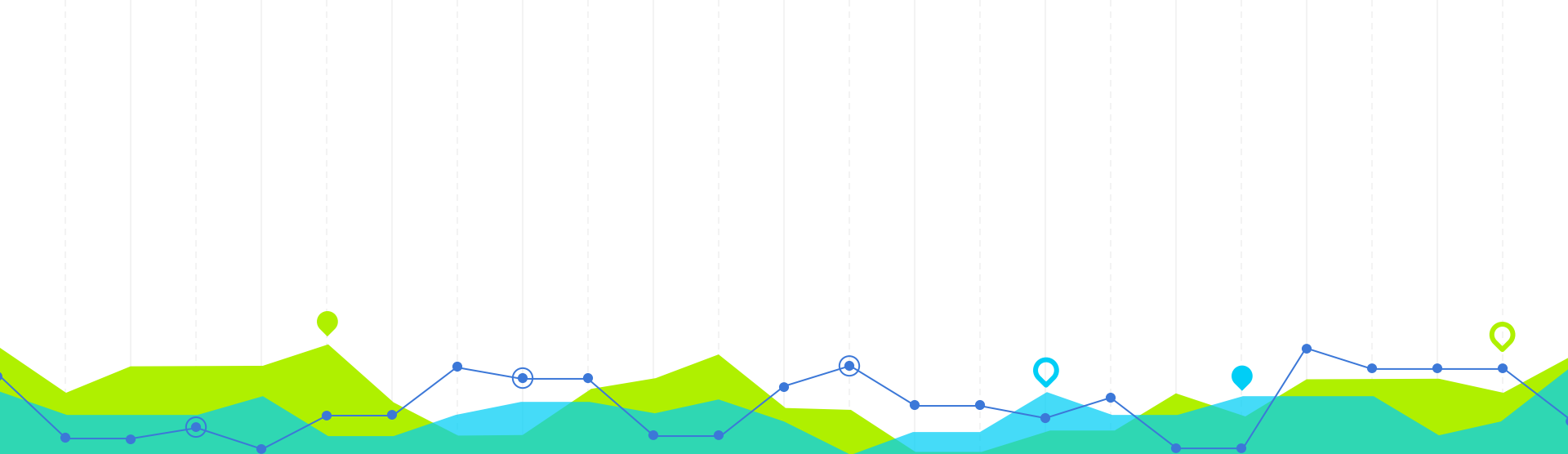
چولگی

برجستگی

چند نمودار جدید

نمودار های ساقه ای

نمودار جعبه ای



1

مقدمه

بیایید با الفبای آمار شروع کنیم..

جمعیت (population)

افراد یا اعضای که در یک یا چند ویژگی مشترک بوده و روی آنها تحقیق میشود



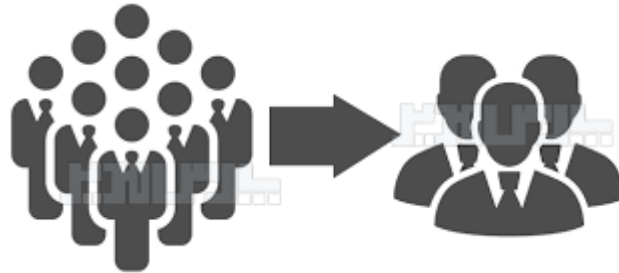
جمعیت میتواند متناهی یا نامتناهی باشد

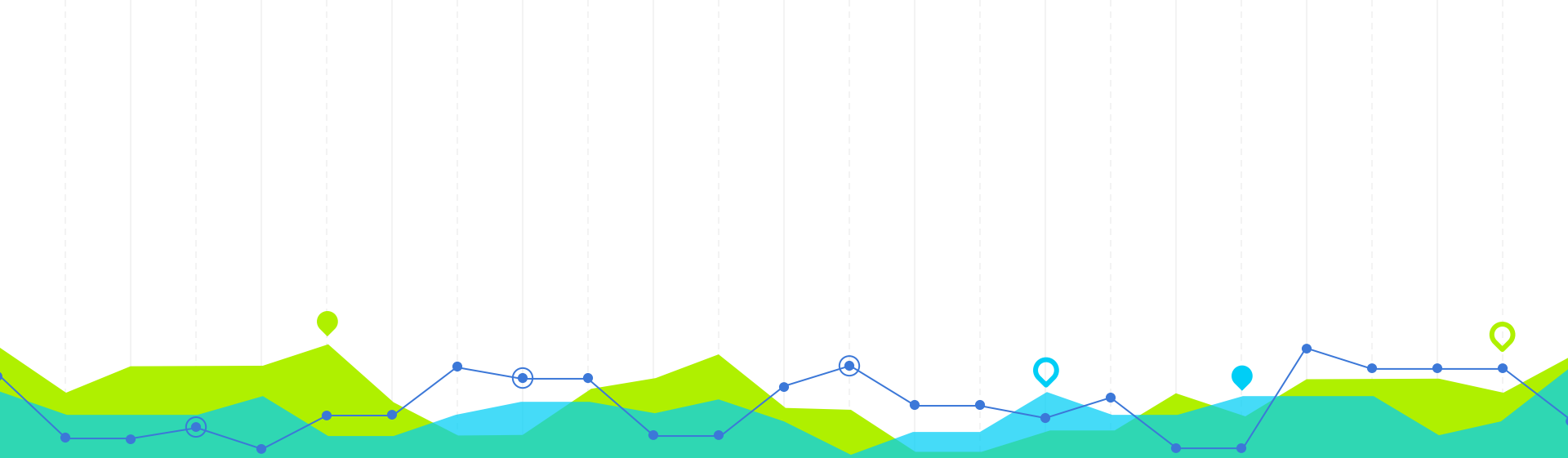


نمونه (Sample)

ان قسمتی از جمعیت که طبق ضوابط انتخاب میشود و مطالعه آن به جای کل جمعیت انجام میشود

مهم است که نمونه بی طرف باشد در نتیجه معمولاً از نمونه های تصادفی (random sample) استفاده میشود



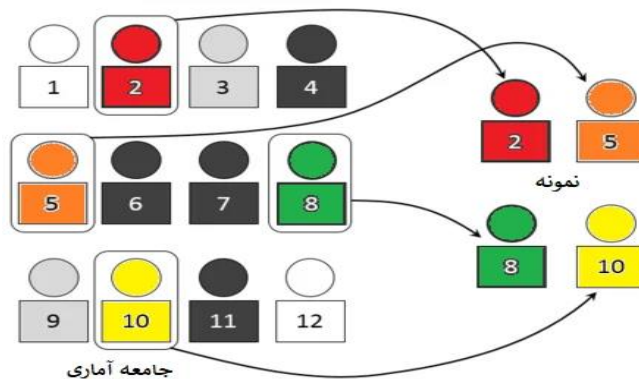


نمونه گیری تصادفی

داده ها شانس برابری برای انتخاب شدن دارند...

نمونه‌گیری تصادفی ساده (Simple Random Sample)

همانطور که از نام آن مشخص است، ساده‌ترین روش نمونه‌گیری تصادفی است. در این روش هر یک از اعضای جامعه برای انتخاب شدن به عنوان نمونه، شانس مساوی دارند. در این روش، ابتدا باید لیستی از افراد جامعه فراهم شود. از معایب این روش نسبت به روش‌های تصادفی دیگر می‌توان به زمان‌بر بودن و احتمال تورش بالا در جوامع با پراکندگی زیاد اشاره کرد. نمونه‌گیری تصادفی ساده به سه شیوه قابل انجام می‌باشد:



نمونه گیری تصادفی ساده با استفاده از قرعه کشی

در این نوع نمونه گیری به هر یک از اعضای جامعه یک شماره یا کد داده می شود. سپس شماره ها در ظرف قرعه کشی قرار می گیرند. در نهایت از بین آن ها، شماره ها انتخاب و ثبت شده و به ظرف بازگردانده می شوند (جایگذاری). علت بازگردادن شماره ها به ظرف، این است که احتمال انتخاب تمامی افراد باهم یکسان باشد. مثلا اگر ۱۰ شماره وجود داشته باشد، احتمال انتخاب همه ی افراد باید مساوی با $1/10$ باشد. در صورتی که اگر انتخاب بدون جایگذاری صورت بگیرد، این احتمال به ترتیب می شود: $1/10$ ، $9/10$ ، $8/10$ و ... شماره های تکراری که از ظرف خارج می شوند پوچ در نظر گرفته شده و دوباره به جای آن ها قرعه کشی انجام می شود. قرعه کشی تا رسیدن به تعداد مورد نیاز ادامه می یابد.

نمونه گیری تصادفی ساده با استفاده از جدول اعداد تصادفی

این جدول مجموعه‌ای از اعداد می‌باشد که بدون نظم یا الگویی مشخص و به صورت کاملاً تصادفی طراحی شده است. پژوهشگر برای انتخاب افراد نمونه از جدول، بطور تصادفی از یک نقطه‌ی جدول در جهت افقی یا عمودی شروع می‌کند. انتخاب نقطه می‌تواند با بستن چشم و گذاشتن انگشت یا نوک قلم روی جدول انجام شود. بعد با توجه به نوع رقم کدها (یک رقمی، دو رقمی، سه رقمی و ...) باید به صورت افقی یا عمودی دنباله‌ی اعداد انتخاب شوند.

نمونه گیری تصادفی ساده با استفاده از سایت‌ها و نرم افزارهای رایانه‌ای

امروزه استفاده از از جدول اعداد تصادفی اصلا پیشنهاد نمی‌شود، چون سایت‌ها و نرم‌افزارهایی در دسترس هستند که با وارد کردن حجم نمونه و تعداد افراد جامعه، می‌توان خیلی ساده‌تر به اعداد تصادفی مورد نیاز رسید. البته اینجا هم باید کدگذاری برای افراد صورت بگیرد.

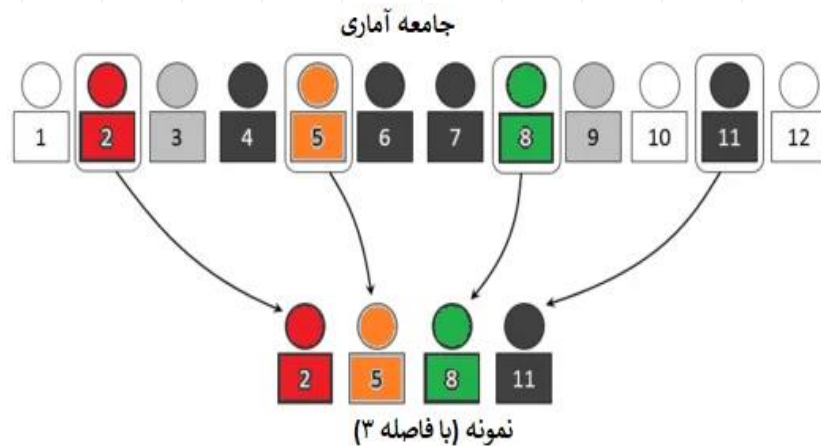
نمونه گیری سیستماتیک یا منظم (Stratified Sampling)

اگر افراد جامعه در لیستی براساس یک ویژگی مرتب شده باشند، می توان نمونه گیری سیستماتیک را به کار برد. این روش مشابه نمونه گیری تصادفی ساده می باشد و به لیستی از افراد جامعه نیز نیاز دارد؛ اما از لحاظ اجرا نسبت به نمونه گیری تصادفی ساده، تا حدودی آسان تر است، سرعت بیشتری دارد و پراکندگی را بهتر در نظر می گیرد.

در این روش، ابتدا به هر یک از افراد جامعه یک شماره از ۱ تا N اختصاص داده می شود. سپس با استفاده از یک عدد تصادفی و همچنین مقداری به عنوان فاصله (k) ، انتخاب اعضا انجام می گیرد. به این صورت که برای انتخاب نمونه ی مورد نیاز $((n)$ از کل جامعه (N) ، ابتدا فاصله ی نمونه گیری (k) با استفاده از فرمول زیر محاسبه می شود.

$$k=N/n$$

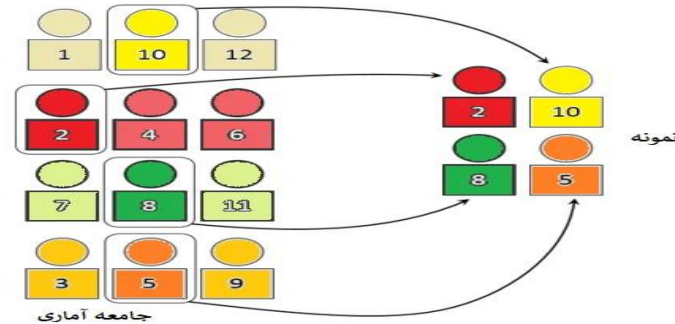
از k سپس از بین اعداد ۱ تا ۹، یک عدد به طور تصادفی انتخاب شده و اعداد بعدی با فاصله‌ی عدد مذکور انتخاب می‌شوند. مثلاً اگر قرار است ۱۰۰ نفر از یک جامعه‌ی ۱۰۰۰ نفری انتخاب اگر فرض کنیم عدد تصادفی انتخاب شده، عدد ۵ باشد، $k=1000/100=10$ می‌شود: k شوند، اعداد بعدی می‌شوند ۱۵، ۲۵، ۳۵، ۴۵ و ...



نمونه‌گیری طبقه‌ای (Stratified Sampling)

در نمونه‌گیری طبقه‌ای ابتدا جامعه به طبقات مختلف با تفاوت‌های زیاد تقسیم می‌شود؛ درحالی که در هر طبقه، افراد از نظر ویژگی مورد نظر همگن و شبیه به هم هستند. سپس از هر یک از این طبقات نمونه‌گیری به روش ساده یا سیستماتیک انجام می‌شود.

زمانی از این روش استفاده می‌شود که رعایت نسبت‌ها درمورد یک یا چند متغیر برای پژوهشگر اهمیت دارد. باید بدانید که تعداد نمونه‌ی انتخابی از هر طبقه متناسب با تعداد افراد آن طبقه است.



مثال

مثلا فرض کنید در یک دانشگاه، ۳۰۰ نفر دانشجوی پزشکی، ۵۰۰ نفر دانشجوی پرستاری و ۲۰۰ نفر دانشجوی علوم آزمایشگاهی وجود دارد و پژوهشگری می‌خواهد بر روی ۱۰۰ نفر از آن‌ها پژوهشی انجام دهد. در این حالت او می‌تواند آن‌ها را با توجه به رشته‌هایشان طبقه‌بندی کند. سپس با روش نمونه‌گیری تصادفی از گروه پزشکی ۳۰ نفر، از گروه پرستاری ۵۰ نفر و از گروه علوم آزمایشگاهی ۲۰ نفر انتخاب نماید. خوب است بدانید طبقه‌بندی بر حسب سن و جنسیت، رایج‌ترین طبقه‌بندی‌ها می‌باشند.

مثال

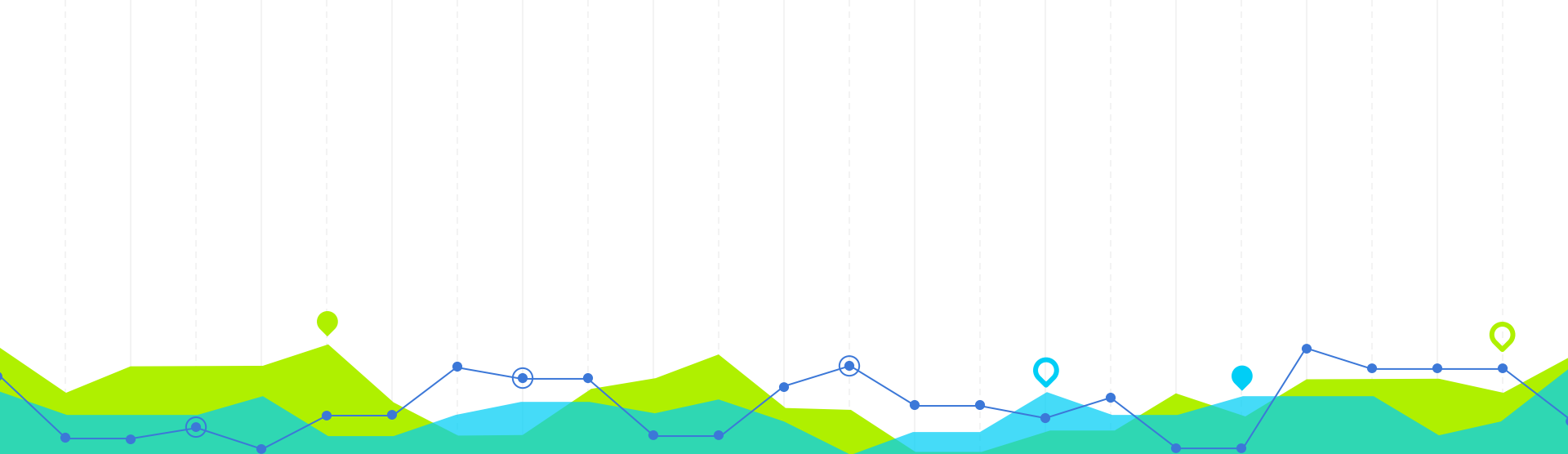
برای مثال اگر ۱۰۰۰ دانشجوی گفته شده، همگی دانشجوی پزشکی از دانشگاه‌های علوم پزشکی مختلف باشند، می‌توان یک یا چند دانشگاه را به طور تصادفی انتخاب نمود. سپس دانشجویان دانشگاه یا دانشگاه‌های انتخاب‌شده را مورد بررسی قرار داد.

نمونه‌گیری چند مرحله‌ای (Multi-stage sampling)

نمونه‌گیری چند مرحله‌ای زمانی استفاده می‌شود که باید کل جامعه پوشش داده شود. در نمونه‌گیری خوشه‌ای مثلاً از بین ۱۰ دانشگاه، ۳ دانشگاه به طور تصادفی انتخاب و دانشجویان آن مورد بررسی قرار می‌گیرند، درحالی که ممکن است در دانشجویان ۷ دانشگاه دیگر تفاوت‌هایی از نظر ویژگی مورد نظر وجود داشته باشد که نتایج پژوهش را زیر سوال ببرد. اما در روش چند مرحله‌ای از هیچ بخشی چشم‌پوشی نمی‌شود.

مثال

مثال: در مطالعه‌ای با عنوان «تعیین میانگین معدل دانش‌آموزان کلاس ششم مدارس شهر تهران» می‌توان یک نمونه‌گیری چند مرحله‌ای را طراحی و اجرا کرد. بدین صورت که پژوهشگر ابتدا شهر تهران را بر اساس منطقه-بندی شهرداری تقسیم نماید (نمونه‌گیری طبقه‌ای) و از هر منطقه مدارس را به صورت نواحی در نظر بگیرد (نمونه‌گیری خوشه‌ای). سپس از هر ناحیه به صورت تصادفی تعدادی مدرسه را به نسبت انتخاب کند (نمونه-گیری تصادفی ساده). در مدارس انتخاب‌شده نیز مجدداً می‌توان هر کلاس را یک خوشه در نظر گرفت و باز به روش تصادفی ساده برخی خوشه‌ها را انتخاب کرد (نمونه‌گیری خوشه‌ای). سپس به روش تصادفی سیستماتیک یا ساده نمونه‌های نهایی انتخاب شوند.

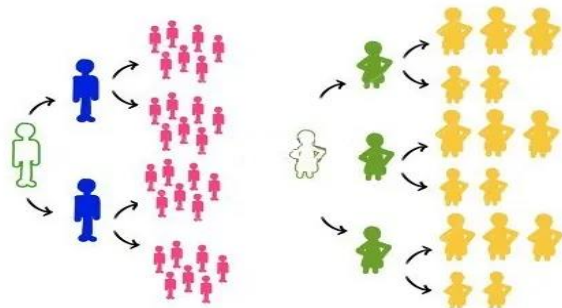


نمونه گیری غیر تصادفی

برخی داده ها شانس بیشتری دارند...

«نمونه‌گیری گلوله برفی» (Snowball Sampling)

در روش نمونه‌گیری گلوله برفی، ابتدا یک یا چند نمونه از کیس مورد نظر شناسایی شده و از آن‌ها درخواست می‌شود که افراد مشابه خود را به پژوهشگر معرفی نمایند. درواقع در این روش از طریق هر نمونه، امکان دسترسی به نمونه‌های بیشتری به‌وجود می‌آید. این روند تا رسیدن به حجم مورد نظر ادامه می‌یابد. مثلاً برای مطالعه بر روی افرادی که به ماده‌ی خاصی اعتیاد دارند، می‌توان با پیدا کردن چند نفر، از آن‌ها خواست تا دوستان یا آشنایان خود را معرفی کنند.



«نمونه‌گیری اتفاقی» (Accidental Sampling)

در این نوع نمونه‌گیری از تمامی افراد جامعه که معیارهای ورود را دارند و در دسترس هستند می‌توان به عنوان نمونه استفاده کرد. این روش بسیار ساده، ارزان و سریع می‌باشد. اما به هیچ‌وجه نشان‌دهنده‌ی جامعه‌ی واقعی نیست. زمانی از این روش استفاده می‌شود که تعداد کیس‌های مورد نظر و به تبع آن حق انتخاب پژوهشگر کم است. مثلاً مطالعه‌ای را فرض کنید که قصد بررسی کیفیت زندگی افراد دیالیزی را دارد. در یک جامعه تعداد افراد دیالیزی کم می‌باشد و امکان انتخاب بین این افراد وجود ندارد یا بسیار زمان‌بر است. در این حالت هر بیمار دیالیزی که به مرکز دیالیز شهر مربوطه مراجعه کرد، مورد پژوهش قرار خواهد گرفت.



«نمونه‌گیری متوالی» (Consecutive Sampling)

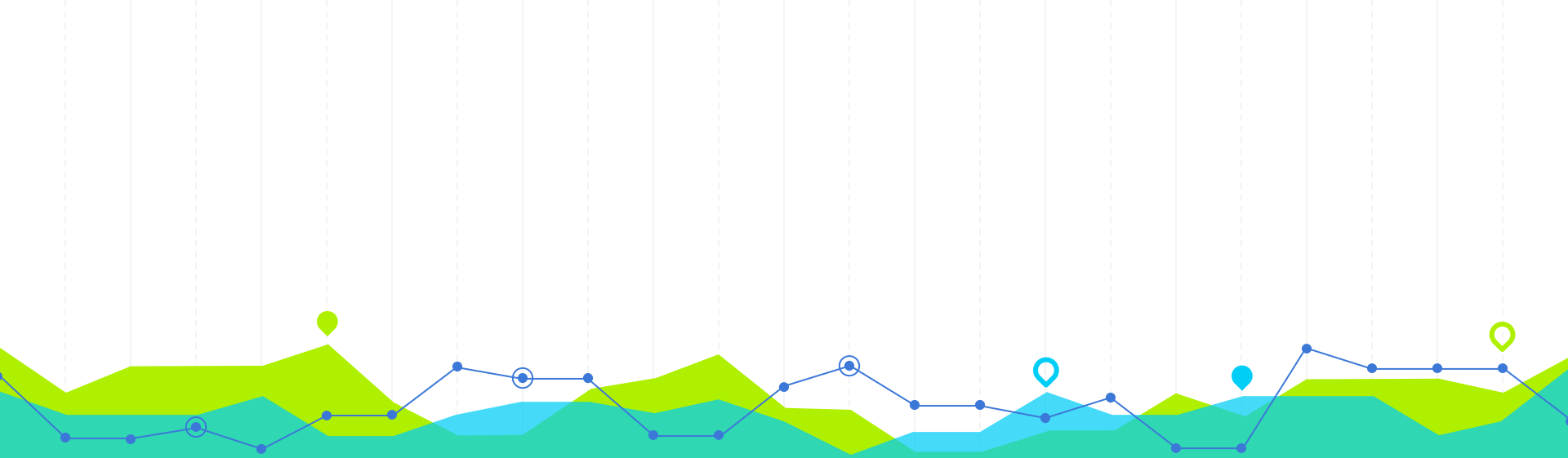
اگر نمونه‌گیری را به صورتی انجام دهیم که با استفاده از یک یا چند شرط اعضای جامعه را محدود کرده و سپس آن را اجرایی کنیم، در اصل روش نمونه‌گیری متوالی را به کار بسته‌ایم. عمل انتخاب اعضای نمونه تا رسیدن به حجم نمونه مورد نیاز ادامه پیدا می‌کند. برای مثال اگر منظور از نمونه‌گیری بررسی تعداد خودروهای قرمز رنگ باشد، می‌توان نمونه را به یک چهار راه محدود و رنگ خودروها را یادداشت کرد. نسبت تعداد خودروهای قرمز رنگ به کل می‌تواند درصد خودروهای قرمز رنگ را تخمین بزند.



«نمونه‌گیری قضاوتی» (Judgmental Sampling)

در این روش، محقق براساس نظر و پیشینه‌ای که در مورد اعضای جامعه آماری دارد، دست به نمونه‌گیری می‌زند. انتخاب یا عدم انتخاب عضوی از جامعه در نمونه بسته به نظر محقق و تجربیات او دارد. معمولاً این روش در جوامع آماری محدود و با حجم کم به کار می‌رود زیرا محقق باید در مورد تک تک اعضا اطلاعات قبلی داشته باشد تا بتواند نمونه حاصل را بهتر انتخاب کند.





اندازه گیری و مقیاس سازی

پیشرفت علم مدیون اندازه گیری است...

2

بیایید با یک مثال شروع کنیم

اگر سیب های یک باغ را جمعیت و a را یکی از سیب ها در نظر بگیریم ویژگی t برای a ممکن است وزن سیب یا تردی آن باشد.

وزن سیب با اعداد قابل بررسی است اما درباره تردی چه؟



مقیاس های استیونز (Stevens scales)

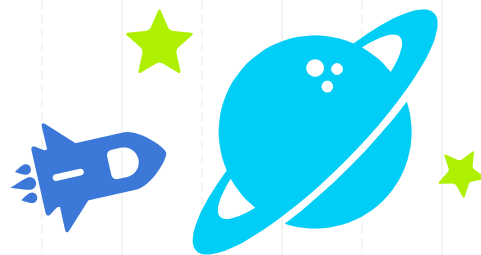
مقیاس اسمی (Nominal scale) : تنها برای شناسایی
نسبت دادن ۱ و ۲ و ۳ به کارگران اصفهان و تهران و شیراز

مقیاس ترتیبی (Ordinal scale) : برای بیان برتری
نسبت دادن ۱ و ۲ و ۳ به متوسط خوب عالی

مقیاس فاصله ای (Interval scale) : نسبت دو تفاضل حفظ شود
۱۰ و ۱۵ و ۲۰ و ۲۵

مقیاس نسبتی (Ratio scale) : نسبت را حفظ کند
برای وزن طول و ...

مقیاس های استیونز



نگاهی دقیق تر

بررسی هریک از مقیاس ها و روابط



مقیاس اسمی (Nominal): (scale)

در این مقیاس افراد یا اشیاء بر اساس یک ملاک معین در طبقه ها که **کیفی هستند و نه کمی**، جایگزین می شوند. در این مقیاس، اندازه گیرنده باید بتواند طبقه ها را از یکدیگر تشخیص دهد و ملاکی را که بر اساس آن افراد یا اشیاء را در طبقه های مختلف جایگزین می کند بشناسد.

در این مقیاس **هیچ گونه همبستگی** یا ارتباطی بین اعداد به کار برده شده وجود ندارد. به عنوان مثال، طبقه ای که عدد یک به آن اختصاص داده می شود، در مقایسه با طبقه ای که به آن **عدد صفر** داده می شود، دارای **هیچ ویژگی اضافه** ای نیست.

مقیاس ترتیبی (Ordinal): (scale)

در این مقیاس وضعیت نسبی اشیاء با افراد **بدون تعیین فاصله** بین آنها بر اساس صفت معینی مشخص می شود. شرط ضروری اندازه گیری در این مقیاس رعایت ملاک رتبه بندی کردن اشیاء یا افراد است، به این معنی که باید روشی را به کار برد که به کمک آن بتوان تعیین کرد که فرد یا شیء مورد اندازه گیری دارای **ارزش بیشتر، کمتر یا مساوی** است.

در صورتی که $B < A$ و $C < B$ باشد، $C < A$ خواهد بود. به بیان دیگر باید ارتباط به گونه ای باشد که اگر A بزرگتر از B و B از C بزرگتر باشد، در نتیجه A بزرگتر از C باشد. البته به جای کلمه بزرگتر می توان از کلمات دیگری مانند قوی تر، پیشرفته تر، بلندتر و غیره استفاده کرد.

مقیاس فاصله ای (Interval): (scale)

مقیاس فاصله ای دارای کلیه ویژگی های مقیاس های اسمی و ترتیبی است و علاوه بر آنها، در این مقیاس **فاصله هر صفت تا مبدأ** آن نیز مشخص است. در این مقیاس نه تنها **ترتیب اشیاء** یا صفتهای مورد اندازه گیری مشخص است، بلکه **فاصله بین واحدهای اندازه گیری** نیز معلوم است. مقیاس فاصله ای نه تنها گروه ها را طبقه بندی و رتبه آنها را نشان می دهد، بلکه مقدار این تفاوت بین گروه ها را نیز اندازه گیری می کند.

مقیاس فاصله ای، فاصله ها، ترتیب تقدم و تساوی مقادیر را بین متغیرها نشان می دهد که نسبت به مقیاس اسمی و ترتیبی قوی تر است.

مقیاس نسبتی (Ratio scale):

مقیاس نسبتی دارای کلیه ویژگی های مقیاس های فاصله ای، ترتیبی و اسمی است. این **مقیاس بالاترین سطح اندازه گیری** است و در آن صفر واقعی وجود دارد. در این مقیاس، برای مقایسه دو ارزش یا دو واحد می توان از نسبت استفاده کرد. متر که برای اندازه گیری طول به کار برده می شود و دارای **مبدأ صفر** است، یک مقیاس نسبی است. بنابراین در این مقیاس می توان گفت ۶ سانتی متر دو برابر ۳ سانتی متر طول دارد. در مقیاس **نسبتی امکان ضرب و تقسیم** هر یک از اندازه ها در یک عدد معین بدون تغییر ویژگی مورد اندازه گیری وجود دارد.



متغیرها (Variables, features)

ویژگی t که در افراد مختلف جمعیت یکسان نیست و معمولاً از فردی به فرد دیگر تغییر میکند.

متغیر {
گروهی: با مقیاس اسمی یا ترتیبی سنجیده میشود و براساس آن جمعیت را گروه بندی می کنند: گروه خونی
عددی: با شمارش بدست می آید و مقیاس فاصله ای و نسبتی : وزن

داده ها (Data)

اگر ویژگی t که معمولا یک متغیر است را از یک جمعیت مطالعه کنیم و آن را درباره تک افراد با مقیاس مناسب اندازه گیری کنیم مجموعه اعدادی بدست می آید که **داده** نام دارد که به خودی خود خام اند

گسسته (Discreet Variable) : (جدا از هم) اندازه گیری با مقیاس اسمی- ترتیبی- شمارشی: تعداد سیب ها

پیوسته (Continuous) : اندازه گیری با مقیاس فاصله ای- نسبتی: وزن افراد

داده ها

تنظیم در جدول

رسم نمودار براساس جدول

اختصار در یک یا چند عدد

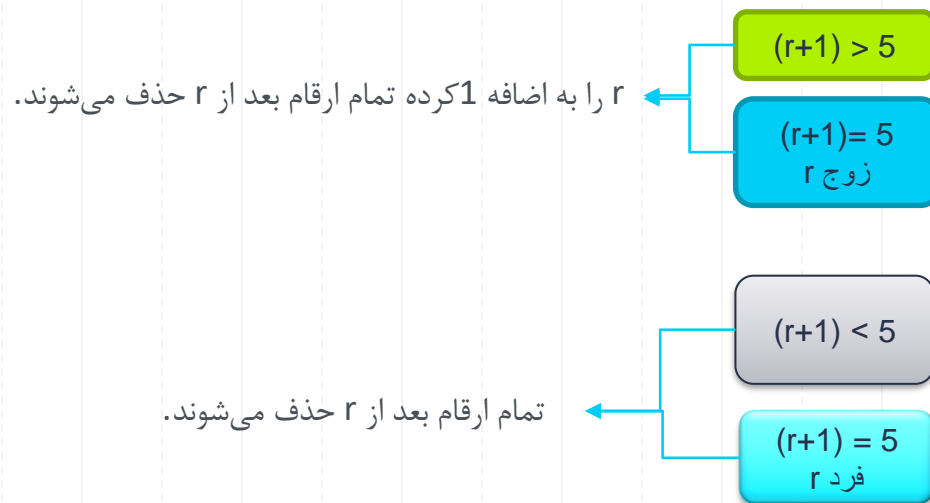
داده خام به پخته



داده های پیوسته (Rounding) سراسر کردن

برای گرد کردن یک عدد به طریق زیر عمل می کنیم:

- برای سراسر کردن یک عدد حقیقی تا r را اعشار داریم:



۶,۳۸۴۹

با دقت ۰,۰۱

۶,۳۸

چون عدد بعد از آخرین رقم باقیمانده ۴ است که کوچکترین از ۵ می باشد، لذا ۴ و ۹ هر دو حذف می شوند

۹,۰۶۵۴۷

با دقت ۰,۰۱

۹,۰۷

اگر رقم بعد از آخرین عددی که باید نگه داشته شود، ۵ و آخرین عددی که باید نگه داشته شود، زوج باشد به آخرین رقم یک واحد اضافه می شود و بقیه اعداد حذف می شوند.

۰,۳۵۰۱

با دقت ۰,۱

۰,۳

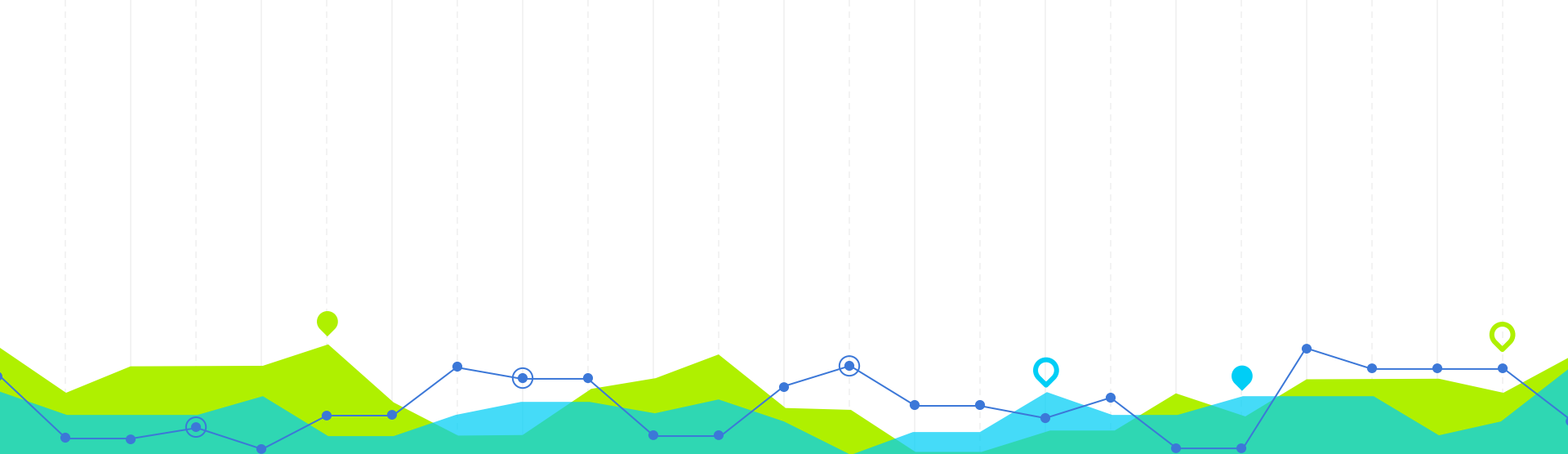
اگر رقم بعد از آخرین عددی که باید نگه داشته شود، ۵ و آخرین عددی که باید نگه داشته شود، فرد باشد آخرین رقم باقی می ماند و بقیه اعداد حذف می شوند.

۴,۸۴۰۶

با دقت ۰,۰۰۱

۴,۸۴۱

اگر رقم بعد از آخرین عددی که باید نگه داشته شود، بیشتر از ۵ باشد یک واحد به آخرین رقم اضافه می شود.



جدول های آماری

بررسی داده را آسان میکند...

3

جدول های آماری

نمایش داده ها با نظم خاصی در چند سطر و ستون بطوریکه بتوان به آسانی پاره ای از دانسته های نهفته در داده ها را از روی آن خواند.

جدولی که از روی تمام داده ها بدست آید **جدول اصلی** و جدولی که از روی جدول اصلی برای بررسی دانسته های ویژه مشتق می شود **جدول فرعی** نام دارد.

در امار برای خلاصه سازی داده از جدول فراوانی استفاده میشود.



فراوانی (Frequency)

هرگاه n چیز از k نوع T_1, T_2, \dots, T_k با فرض $n \geq k \geq 2$ به ترتیب با تعداد f_1, f_2, \dots, f_k تشکیل شده باشند این تعداد را **فراوانی** می نامیم.

۱۴/۵ ۱۷/۵ ۱۳ ۹ ۱۲/۵ ۱۸ ۱۴ ۶/۵ ۱۰/۵ ۱۷
 ۴ ۱۸ ۱۹ ۱۲/۵ ۱۹ ۱۱ ۱۷/۷۵ ۸/۵ ۱۵ ۱۳/۵
 ۱۰ ۱۵ ۷/۵ ۹ ۱۶ ۱۴ ۱۹ ۱۲ ۱۸ ۱۵

| حدود دسته ها | خط نشان | فراوانی | مرکز دسته | فراوانی × مرکز |
|---------------------|--------------|---------|-----------|----------------|
| $4 \leq x < 9$ | //// | ۴ | ۶/۵ | ۲۶ |
| $9 \leq x < 14$ | //// // | ۱۰ | ۱۱/۵ | ۱۱۵ |
| $14 \leq x \leq 19$ | //// // // / | ۱۶ | ۱۶/۵ | ۲۶۴ |

فراوانی نسبی (Relative Frequency)

اگر فراوانی هر رده را به جمع فراوانی‌ها تقسیم کنیم، «فراوانی نسبی» حاصل می‌شود. البته می‌توان مقدار این ستون را به صورت درصدی نیز نمایش داد. برای این کار کافی است حاصل تقسیم را در ۱۰۰ ضرب کنیم و حاصل را با علامت % نشان دهیم. نماد مربوط به فراوانی نسبی رده r_i است.

$$\sum_{i=1}^k f_i = n$$

$$1 \leq f_i \leq n$$

$$\sum_{i=1}^k r_i = 1$$

$$\frac{1}{n} \leq r_i \leq 1$$

فراوانی انباشته و فراوانی نسبی انباشته

باتوجه به فراوانی و فراوانی نسبی برای $i=1,2,\dots,k$

$$s_i = \sum_{j=1}^i r_j$$

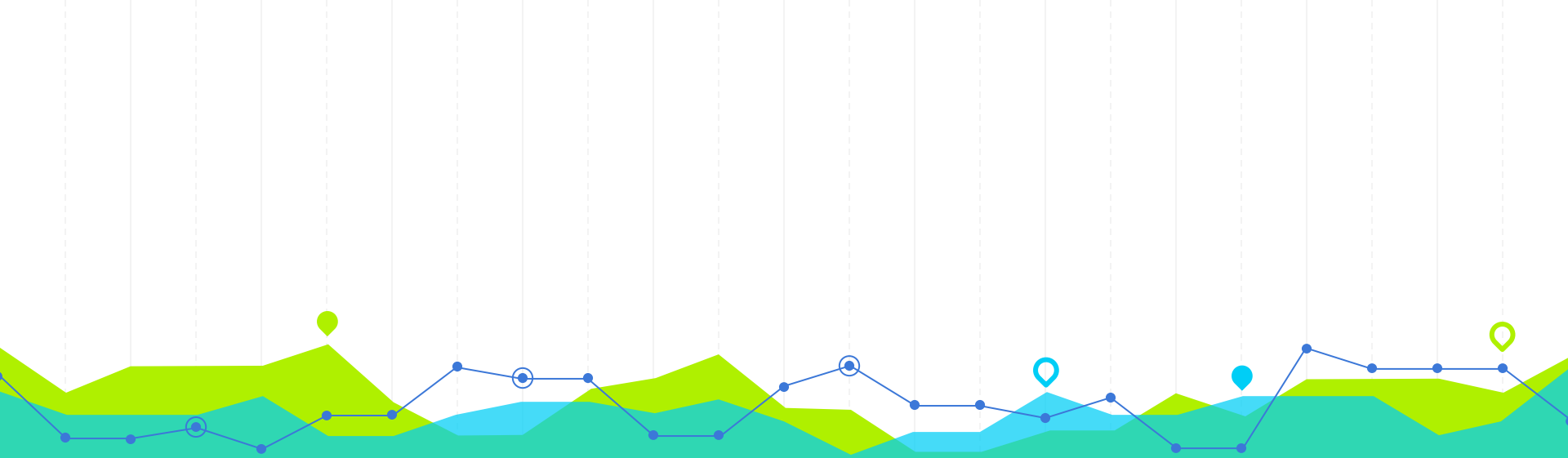
$$g_i = \sum_{j=1}^i f_j$$

را به ترتیب فراوانی انباشته و فراوانی نسبی انباشته میگویند.

جدول فراوانی

جدولی که داده ها را برحسب فراوانی تنظیم میکند جدول فراوانی از چند سطر و ستون تشکیل شده است. هر سطر نشانگر خصوصیات یک طبقه یا رده است.

| فراوانی | دسته |
|---------|---------------------|
| ۵ | $0 \leq x < 4$ |
| ۵ | $4 \leq x < 8$ |
| ۱۰ | $8 \leq x < 12$ |
| ۳۰ | $12 \leq x < 16$ |
| ۵۰ | $16 \leq x \leq 20$ |



نمودار های آماری

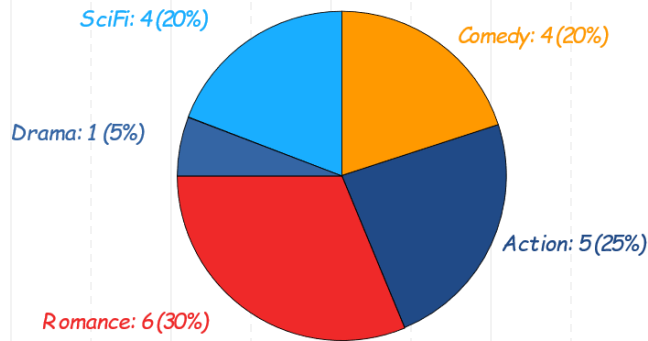
داده ها در یک نگاه...

4

نمودار دایره ای (Pie chart)

نمودار دایره ای نسبت گروه ای از داده ها را به کل داده ها به صورت برش هایی از یک دایره نمایش میدهد. بر روی هر برش نسبت آن گروه به کل داده ها نوشته شده است. مجموع نسبت ها باید ۱۰۰ درصد باشد. نمودار های دایره زمانی که تعداد داده ها کم است بیشترین تاثیر گذاری را دارند.

Favorite Type of Movie



چه زمانی از نمودار دایره ای استفاده کنیم؟

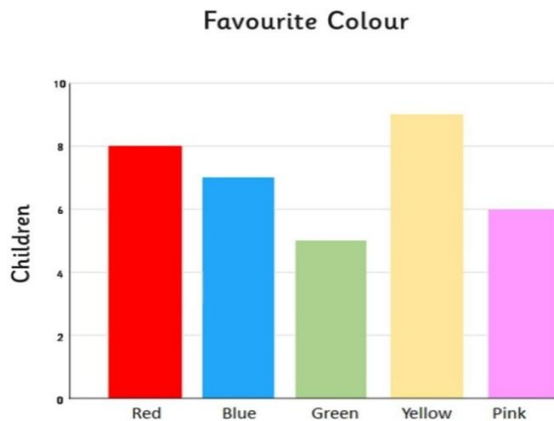
مقایسه بخشی از داده ها با کل داده ها
مشخص کردن کوچک ترین و بزرگترین دسته بندی در داده ها
زمانی که تعداد داده های مقایسه کم است.

نکاتی برای رسم نمودار دایره ای بهتر:

تعداد دسته ها بین ۳ تا ۵ عدد باشد که کیفیت برش ها حفظ شود.
در صورت نیاز گروه های نزدیک به هم را در یک گروه ادغام کنید.
رنگ مهم ترین گروه را به گونه ای انتخاب کنید که جلب توجه کند.
به ترتیب گروه ها اهمیت دهید؛ برای مثال گروه ها را به ترتیب اندازه به صورت ساعت گرد قرار دهید.

نمودار میله ای (Bar chart)

نمودار میله ای یکی از محبوب ترین نمودار هاست و دلیل این محبوبیت، سادگی آن برای تحلیل است. در این نمودار می توان کمترین و بیشترین مقدار ها را به سرعت تشخیص داد. همچنین نسبت دو گروه به راحتی قابل تصور است.



چه زمانی از نمودار میله ای استفاده کنیم؟

زمانی که بیش از ۱۰ گروه برای مقایسه داریم.
زمانی که برچسب (Label) های نمودار طولانی اند.

نکاتی برای رسم نمودار میله ای بهتر:

از یک رنگ برای میله ها استفاده کنید، مگر زمانی که میخواهید توجه مخاطب را به یکی از میله ها جلب کنید.

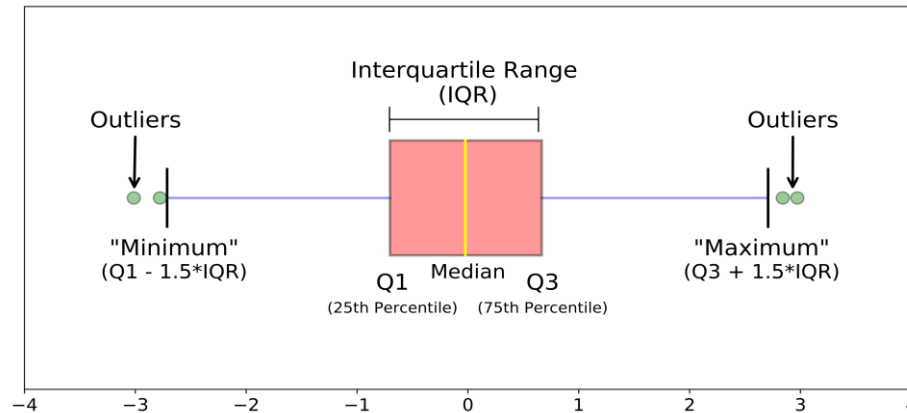
ضخامت میله ها باید بیشتر از فاصله بین آن ها باشد.

برای راحتی مخاطب، نوشته ها را افقی بنویسید (و نه عمودی).

چینش میله ها را بر اساس حروف الفبا یا فراوانی داده ها قرار دهید.

نمودار جعبه ای (Box plot/Whisker plot)

نمودار جعبه ای معیارهای تمرکز در گروه‌ها را در یک نمودار خلاصه می‌کند. موقعیت جعبه نشان می‌دهد بیشتر داده‌ها در کدام بازه قرار گرفته‌اند. این نمودارها معمولاً برای مقایسه یک ویژگی مشترک بین گروه‌های مستقل استفاده می‌شود.

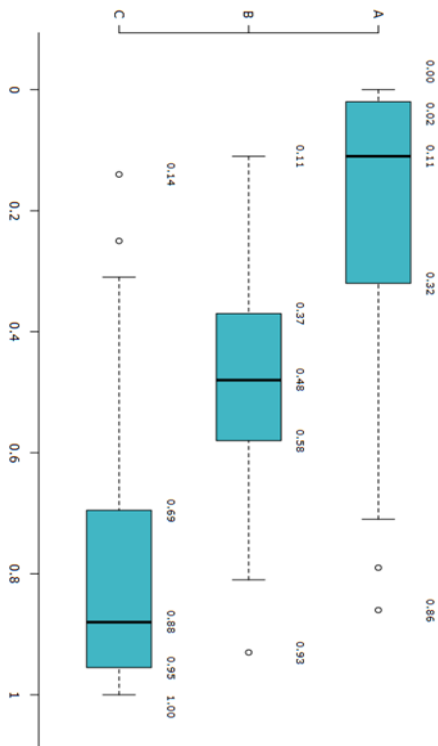


چه زمانی از نمودار جعبه ای استفاده کنیم؟

مقایسه معیارهای تمرکز گروه های بزرگی از داده ها
به دست آوردن ایده کلی از توزیع داده ها در گروه ها
مقایسه داده ها بدست آمده از منابع مختلف

نکاتی برای رسم نمودار جعبه ای بهتر:

از این نمودار برای نمایش ۱ تا ۳ گروه استفاده کنید.
گروه بندی داده ها باید مستقل از مشخصه در حال بررسی باشد.

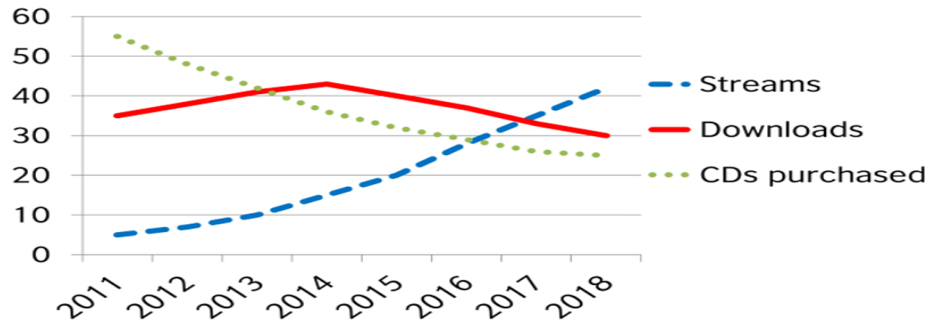


نمودار خطی (Line Graph)

یک نمودار خطی میتواند الگوها یا پیشرفت را در داده‌های پیوسته در مقاطع زمانی نشان دهد. همچنین از این نمودار میتواند برای بررسی تاثیر تغییرات بین دو یا چند گروه استفاده کرد.

Percentage of total music sales by method

Percentage



چه زمانی از نمودار خطی استفاده کنیم؟

بررسی ظهور الگوها در دوره های زمانی
مقایسه تغییرات دو یا چند گروه در دوره زمانی
تعیین تاثیر دو فرایند (مانند روش های بازاریابی) بر روی داده ها

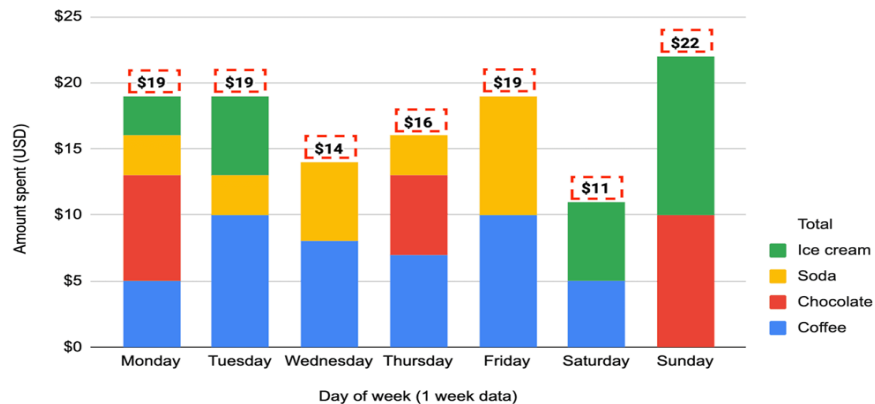
نکاتی برای رسم نمودار خطی بهتر:

از خطوط پر استفاده کنید. نقطه چین ها سبب عدم تمرکز مخاطب میشوند.
بیش از چهار خط در یک نمودار رسم نکنید.
ارتفاع گراف را به گونه ای تنظیم کنید که مرتفع ترین خط از بیش از ۷۰ درصد ارتفاع نمودار بالاتر نرود.

نمودار پشته ای (Stacked Bar Chart)

این نمودار برای مقایسه بین گروه ها و چگونگی تشکیل آن داده ها استفاده میشود. این نمودار ها همچنین میتواند در مقایسه داده های که در طول زمان جمع آوری شده اند بسیار کاربردی باشند.

How I spend money on little luxuries by day of week



چه زمانی از نمودار پشته ای استفاده کنیم؟

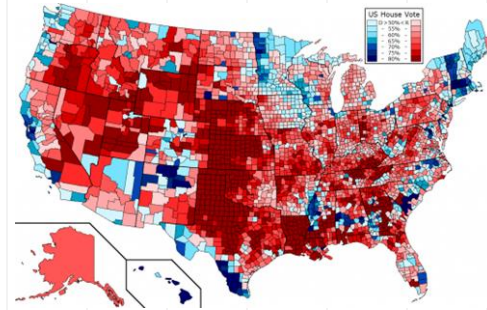
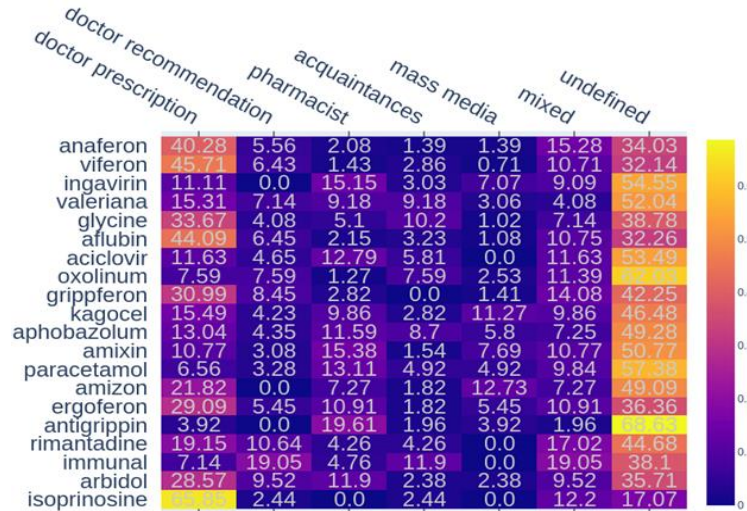
تغییر داده ها در طول زمان
نمایش مقدار زیادی اطلاعات در یک چارت
تشخیص تاثیر وقایع در داده های زمانی
نمایش نسبت زیر گروه ها در گروه های مختلف

نکاتی برای رسم نمودار پشته ای بهتر:

رنگ ها را برای زیر گروه های مشابه تغییر ندهید.
اندازه نمودار را به گونه ای تنظیم کنید که نسبت زیر گروه ها نسبت به هم مشخص باشد.

نقشه گرمایی (Heatmap)

نقشه گرمایی میتواند نسبت دو متغیر به هم را نشان دهد و یک نمره به این نسبت بیافزاید. این نمره میتواند در قالب تغییر رنگ یا سایه در نمودار مشخص شود.



چه زمانی از نقشه گرمایی استفاده کنیم؟

بررسی الگوهایی که به سرعت تغییر میکنند.
نمایش عملی نقاط نظر
نمایش اطلاعات بر روی نقشه های مکانی

نکاتی برای رسم نقشه گرمایی بهتر:

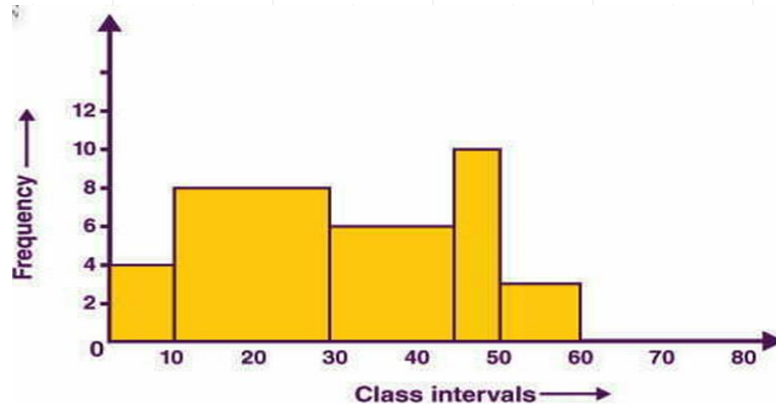
خط مرزی نقشه را به خوبی مشخص کنید.
تنها از یک رنگ و سایه های آن استفاده کنید.
از مشخص کردن تعدادی زیادی الگو بر روی نقشه بپرهیزید.



نمونه استفاده از نقشه گرمایی برای بررسی
نحوه تعامل با یک وبسایت

هیستوگرام (Histogram)

هیستوگرام ها داده های پیوسته رده بندی شده و ترتیب یافته را به نمایش میگذارند. این گروه ها باید دیتاست را بپوشانند اما اشتراک (overlap) نداشته باشند. با این نمودار میتوان تقریبی درباره توزیع داده ها یافت.



چه زمانی از هیستوگرام استفاده کنیم؟

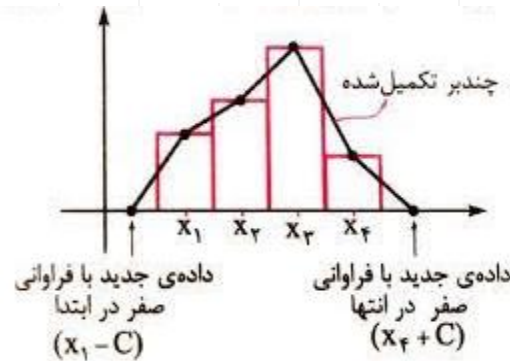
نگاه کلی به شیوه توزیع داده ها
تعیین اتفاقی یا دائمی بودن وقایع در داده ها
تعیین تاثیر دو فرایند (مانند روش های بازاریابی) بر روی داده ها

نکاتی برای رسم هیستوگرام بهتر:

خط شروع را صفر قرار دهید.
از رده های با معنی استفاده کنید.
تعداد رده ها را بیش از حد زیاد نکنید.

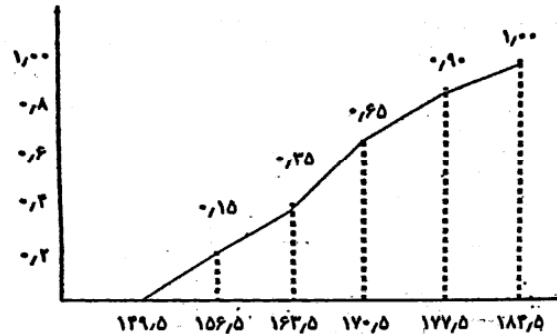
چندبر فراوانی

اگر نقطه های وسط قاعده های بالای مستطیل های هیستوگرام و نقطه های وسط رده هایی را که بلافاصله در دو انتهای هیستوگرام بوده اند و دارای فراوانی صفر هستند به هم بپیوندیم یک خط شکسته بدست می آید که **چندبر فراوانی** نام دارد.



چندبر فراوانی انباشته

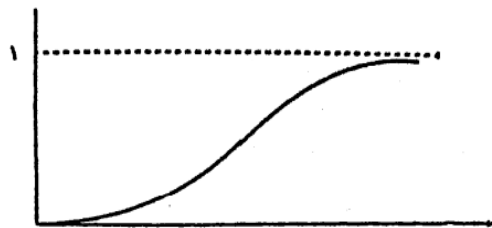
اگر نقطه های که طول آنها مرز رده ها و عرض آنها فراوانی نسبی انباشته تا آن مرز باشد به هم پیوندیم یک خط شکسته بدست می آید که **چندبر فراوانی انباشته** نام دارد.



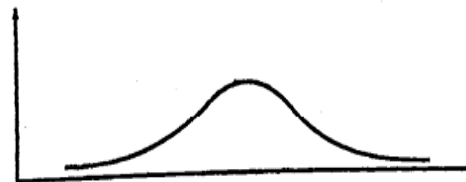
چندبر فراوانی انباشته مربوط به قدما

منحنی های فراوانی و فراوانی انباشته

اگر تعداد داده ها زیاد و طول رده ها کوچک و در نتیجه رده ها زیاد باشد چندبر فراوانی و انباشته دارای اضلاع زیاد خواهد بود و میتوان بر انها منحنی هایی منطبق کرد که به ترتیب **منحنی فراوانی** و **منحنی فراوانی انباشته** نامیده میشوند



منحنی فراوانی انباشته

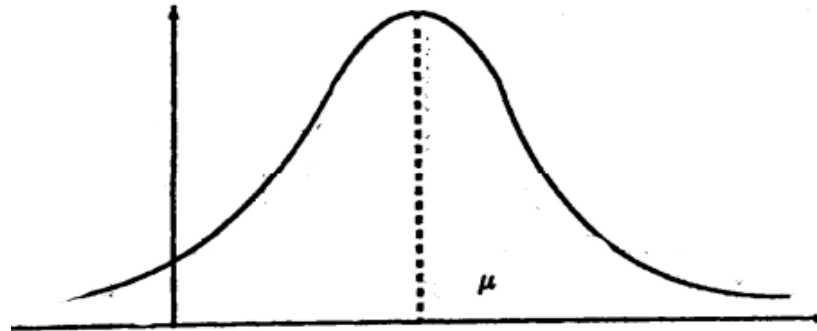


منحنی فراوانی

منحنی فراوانی نرمال

اگر منحنی فراوانی دارای معادله مختصاتی زیر باشد آن را منحنی فراوانی نرمال می نامیم

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}\right)(x-\mu)^2}$$



منحنی نرمال با σ^2 کوچک

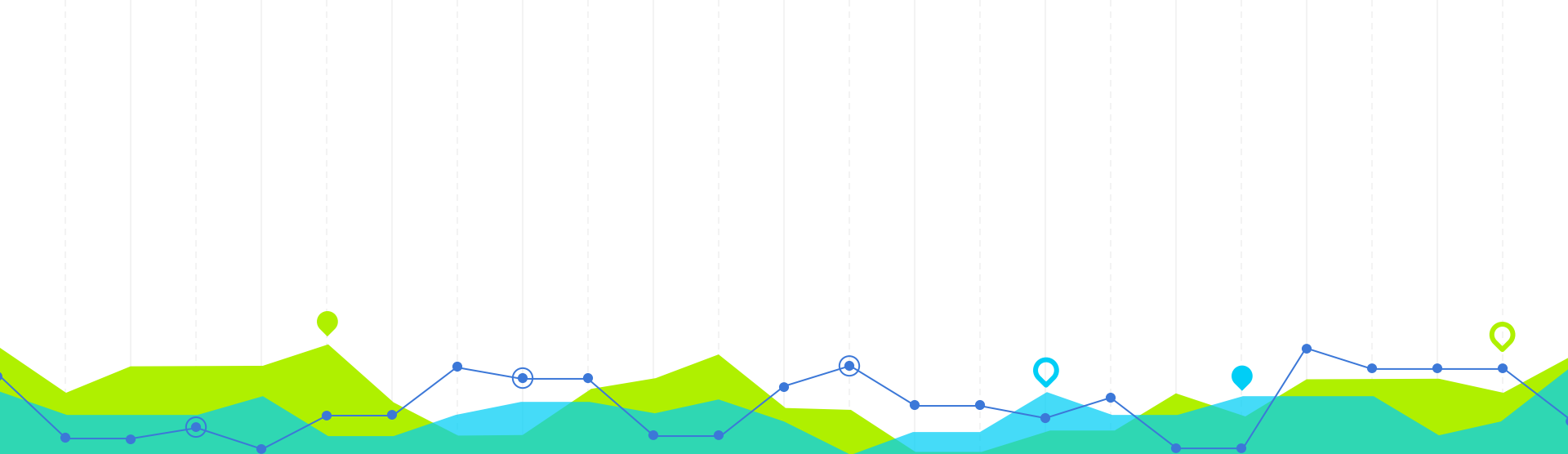
نکات منحنی فراوانی نرمال

نسبت به $x=\mu$ قرینه است

σ^2 هرچه کوچکتر منحنی کشیده تر

$\mu=0, \sigma^2=1$ نرمال استاندارد

دارای ماکزیمم $(\mu, \frac{1}{\sqrt{2\pi}\sigma^2})$



5 معیار های تمرکز

چقدر داده ها بهم نزدیک اند...


5

میانگین حسابی (arithmetic)

اگر مجموع داده ها را بر تعدادشان تقسیم کنیم، میانگین حسابی به دست می آوریم:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

اگر تعداد هر یک از داده ها برابر ۱ باشد به این میانگین در زبان عامیانه **معدل** میگویند. 

اگر $\omega_i = \frac{f_i}{n}$ را وزن داده i ام بنامیم آنگاه خواهیم داشت $\sum_{i=1}^k \omega_i = 1$ و آن را **میانگین وزنی** مینامیم. 

$$\overline{x_\omega} = \sum_{i=1}^k \omega_i x_i$$

میانگین هندسی (Geometric)

اگر مجموعه از داده های مثبت داشته باشیم، میانگین هندسی آن ها برابر است با:

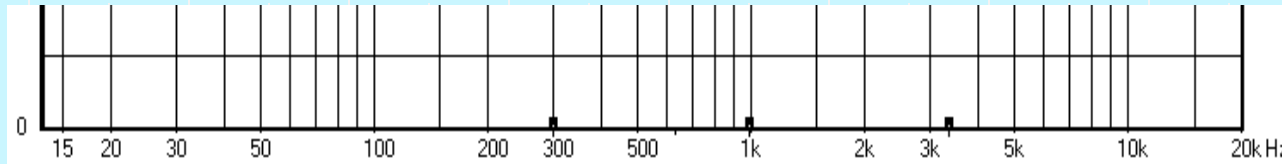
$$G = \left(\prod_{i=1}^k x_i^{f_i} \right)^{\frac{1}{n}}$$

در علوم داده این میانگین در مقایسه اشیا با مشخصه و برد عددی های مختلف کاربرد دارد.



مثال

محدوده فرکانس در خطوط تلفنی بین ۳۰۰ تا ۳۳۰۰ هرتز است. با توجه به لگاریتمی بودن مقیاس فرکانس انتقال خطوط تلفن، میانگین این دو مقدار بر اساس میانگین هندسی محاسبه می‌شود که برابر با $۹۹۵ = ۳۰۰ \times ۹۹۵ \sqrt{۳۳۰۰} = ۳۰۰ \times ۳۳۰۰$ هرتز خواهد بود، در حالیکه میانگین حسابی برای آن‌ها ۱۸۰۰ هرتز است.



با ضرب یا تقسیم داده‌ها در مقدار ثابت، میانگین هندسی نیز تحت تاثیر قرار گرفته و در همان مقدار ضرب یا تقسیم خواهد شد. پس اگر $y = b \cdot x$ باشد، میانگین هندسی y برحسب میانگین هندسی x به صورت زیر خواهد بود.

$$GM_y = b GM_x$$

میانگین توافقی (Harmonic)

اگر مجموعه از داده های ناصفر داشته باشیم، میانگین توافقی آن ها برابر است با:

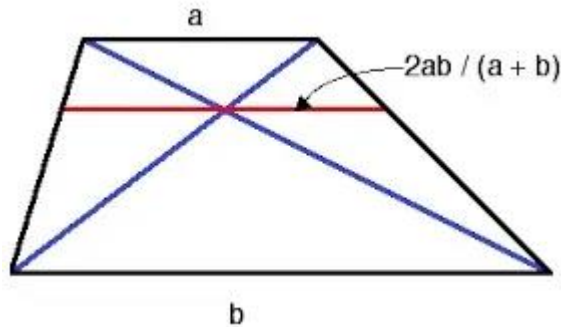
$$\frac{1}{H} = \frac{1}{n} \sum_{i=1}^k \frac{f_i}{x_i}$$

برای محاسبه متوسط متغیر مانند سرعت، نرخ و ... که از نوع نسبت هستند باید از این نوع میانگین استفاده کرد.



تحلیل هندسی

تفسیر هندسی نیز برای میانگین همساز وجود دارد. برای این کار در یک دوزنقه قطرها را رسم کنید. از محل برخورد این قطرهای خطی به موازات قاعده ترسیم کنید تا دوزنقه را قطع کند، طول این خط برابر با میانگین همساز برای دو قاعده خواهد بود



مثال

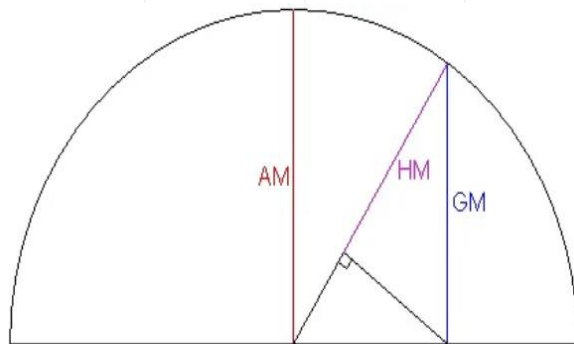
میزان کارکرد هفتگی برحسب ساعت برای چهار کارمند بخش انتشارات یک شرکت طبق جدول زیر ثبت شده است. هر کارمند در سال به میزان ۲۰۰۰ ساعت کار کرده ولی کارکرد هفتگی آن‌ها در هفته‌های مختلف متفاوت است. متوسط زمان کارکرد هفتگی بخش انتشارات این شرکت، براساس میانگین همساز محاسبه می‌شود.

| کارمند | کل زمان کاری | تعداد هفته | متوسط زمان کاری در هفته (ساعت) |
|--------|--------------|------------|--------------------------------|
| ۱ | ۲۰۰۰ | ۴۰ | ۵۰ |
| ۲ | ۲۰۰۰ | ۴۵ | ۴۴.۴۴۴۴ |
| ۳ | ۲۰۰۰ | ۳۵ | ۵۷.۱۴۲۸۶ |
| ۴ | ۲۰۰۰ | ۵۰ | ۴۰ |
| جمع | ۸۰۰۰ | | ۱۹۱.۵۸۷۲۹۷ |

متوسط هفته‌های کاری طبق محاسبه میانگین همساز مقدار ۴۱,۷۵۶۴۲ هفته خواهد شد و بر این اساس از تقسیم ۸۰۰۰ ساعت بر این میانگین (۴۱,۷۵۶۴۲) متوسط ساعت کاری در هفته (۱۹۱,۵۸۷۳۰۵۶ ساعت) نیز استخراج می‌شود.

ترتیب میانگین‌ها

با توجه به شیوه محاسبه این سه نوع میانگین می‌توان نشان داد که میانگین همساز تمایل دارد که به سمت مقادیر کوچکتر نزدیک شود، در نتیجه اگر میانگین حسابی را با AM و میانگین هندسی را با GM و در آخر میانگین همساز را با HM نشان دهیم، رابطه ترتیبی بین این سه میانگین به صورت زیر خواهد بود:



میانه (median)

عدد m را میانه میگویند هرگاه نیمی از داده ها از m کوچک تر باشند.

محاسبه میانه برای متغیرهای گسسته: داده ها را به ترتیب صعودی (یا نزولی) مرتب میکنیم، اگر تعداد داده ها فرد باشد داده میانی و اگر تعداد زوج باشد، میانگین دو داده میانی میانه است:

$$m = x_{\frac{n+1}{2}}$$

$$m = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

چندک ها (quantiles)

چندک ها نقاطی در داده ها هستند که فضای نمونه را به بازه هایی با تعداد مساوی تقسیم میکنند را چندک گفته و با Q_p نمایش میدهیم.

چهارک ها: نقاط Q_1, Q_2, Q_3 که داده ها را به بازه هایی که هر کدام ۲۵ درصد داده ها را در بر میگیرند تقسیم میکنند.

دهک ها: نقاط D_1, D_2, \dots, D_9 که داده ها را به بازه هایی که هر کدام ۱۰ درصد داده ها را در بر میگیرند تقسیم میکنند.

صدک ها: نقاط P_1, P_2, \dots, P_{99} که داده ها را به بازه هایی که هر کدام ۱۰ درصد داده ها را در بر میگیرند تقسیم میکنند.



محاسبه چندک ها برای داده های گسسته

در داده های مرتب، برای به دست آوردن Q_p اگر $(n + 1)p$ برابر عدد صحیح r باشد داده x_r را به عنوان چندک مد نظر اختیار میکنیم.
در غیر این صورت بزرگترین عدد صحیح پیش از $(n + 1)p$ را x_r در نظر گرفته و اختلاف آن با r را ω مینامیم. حال خواهیم داشت:

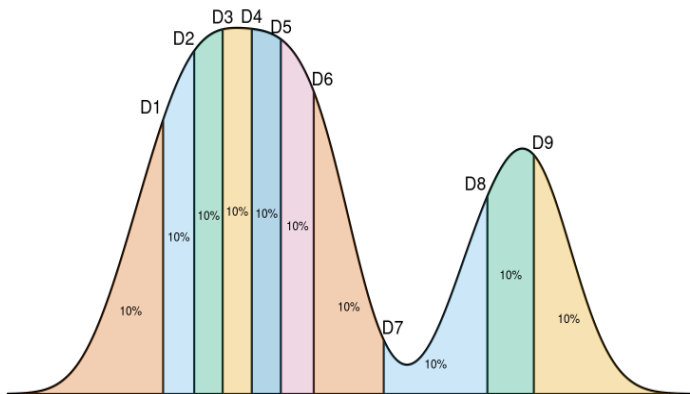
$$Q_p = (1 - \omega)x_r + \omega x_{r+1} + 1$$



محاسبه چنک ها برای داده های پیوسته

در جدول فراوانی ابتدا رده را یافته و سپس چند مد نظر را به فرمول زیر پیدا میکنیم.

$$Q_P = L_p + \frac{(pn - g_p)}{f_p} w$$



نما-مد (mode)

داده ای با بیشترین تکرار در نمونه را نما (مد) میگویند.

به دست آوردن نما برای داده های گسسته: داده ای با بیشترین تکرار را به عنوان نما انتخاب کرده و با M نمایش میدهیم.

به دست آوردن نما برای داده های پیوسته: بازه ای با بیشترین تعداد داده را بازه نمایی در نظر میگیریم. مرز پایینی آن را با a و اختلاف فراوانی نسبی رده میانه با رده قبلی و بعدی و طول رده میانه را نشان میدهند. حال خواهیم داشت:

$$M = L_M + \frac{d_1}{d_1 + d_2} w$$

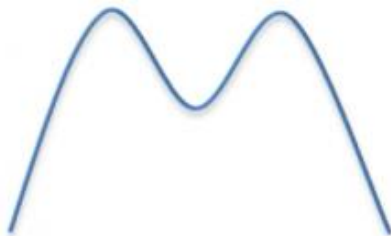
نمونه ها میتوانند بیش از یک مد داشته باشند



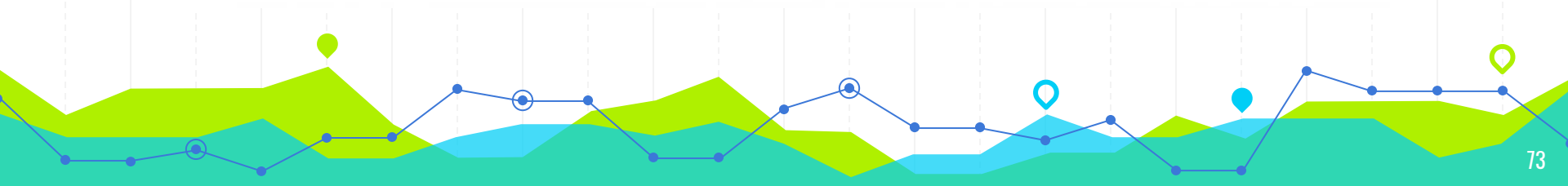
Unimodal



Bimodal



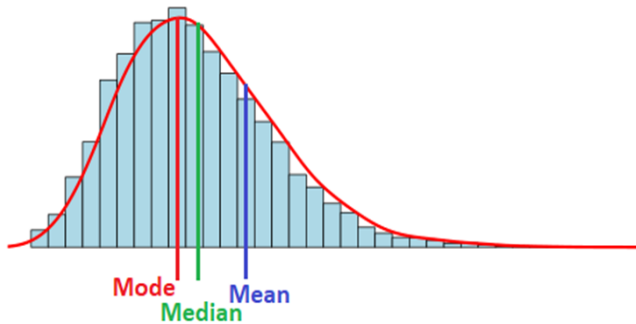
Multimodal



مقایسه معیار های تمرکز (Central Measure)

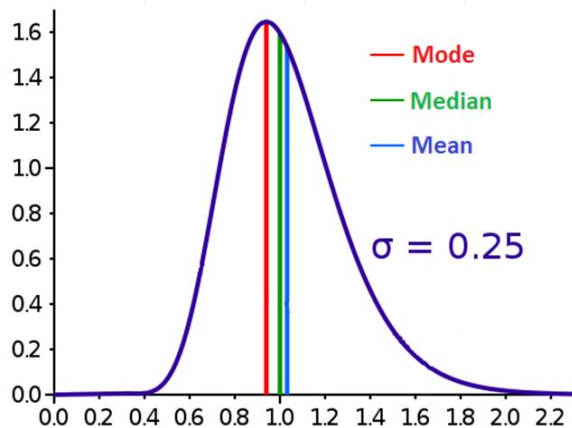
وجود داده های پرت بسیار در یک نمونه میتواند معیاری مانند میانگین را به گونه ای تغییر دهد که دیگر معیار خوبی برای نمایش تمرکز نباشند.
همچنین در داده های ترتیبی میانه و نما معیار بهتری از میانگین هستند.

Positive Skewed



مقایسه معیار های تمرکز

در آمار رابطه به صورت زیر وجود دارد که در آزمایش های فراوان به چشم میخورد اما به اثبات نرسیده است.
این رابطه نشان دهنده توازن داده هاست:

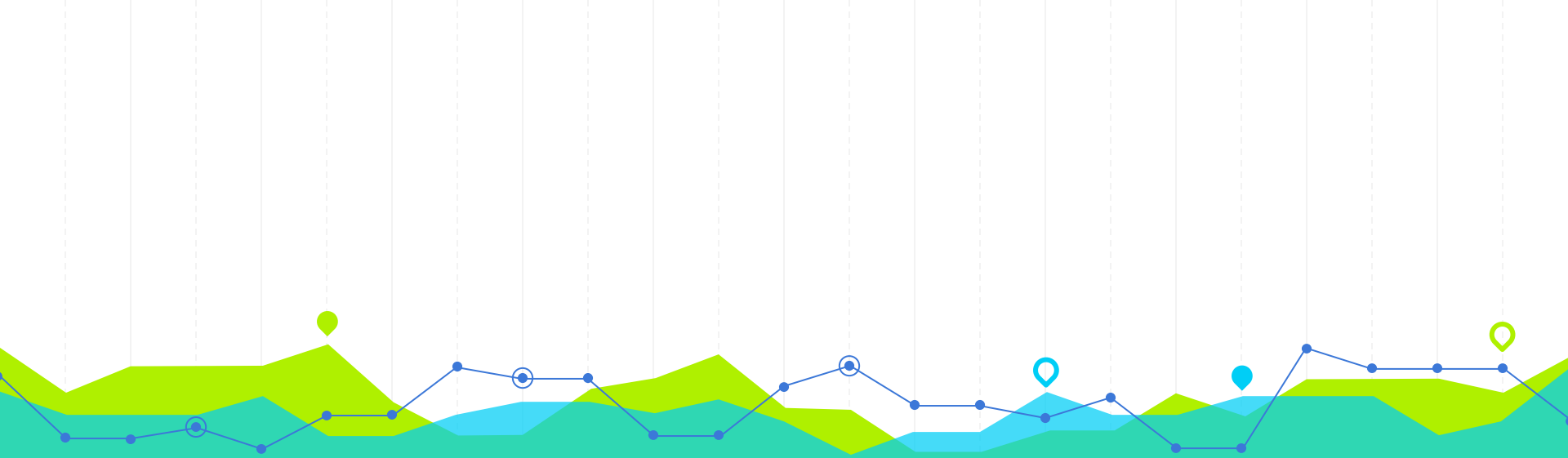


$$\bar{x} - M \approx 3(\bar{x} - m)$$

میانگین اصلاح شده (Trimmed Mean)

همانطور که اشاره شد داشتن داده هایی که تعداد آن ها بسیار کم اما مقدار آن ها بسیار متفاوت از سایر داده هاست میتواند باعث کم شدن دقت معیار میانگین شود به همین دلیل نوع دیگری از میانگین را تعریف میکنیم که در آن ابتدا داده های را مرتبط و سپس میانگین داده ها را پس از حذف k داده از ابتدا و انتهای بازه محاسبه میکنیم.

$$T_k = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_i$$



6

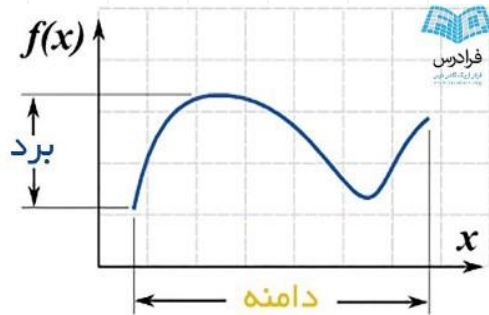
معیار های پراکندگی

به بررسی میزان پراکندگی داده ها می پردازد...

برد (range)

تفاوت مابین کمترین و بیشترین مقادارها است.

برد گاهها ممکن است گمراه کننده باشد هنگامی که اعداد بسیار بزرگ یا بسیار کوچک باشد.



با افزایش داده ها برد میتواند بزرگتر یا ثابت باشد

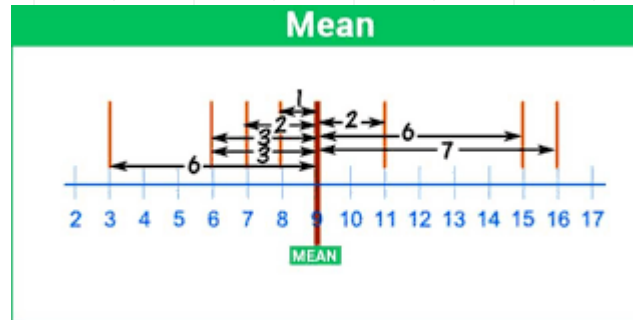


میانگین انحراف ها

قدر مطلق $x_i - x$ را انحراف از میانگین برای داده x_i و عبارت زیر میانگین انحراف ها می نامند

$$d = \frac{\sum_{i=1}^k f_i |x_i - x|}{n}$$

هر قدر داده ها از x دورتر d بزرگتر

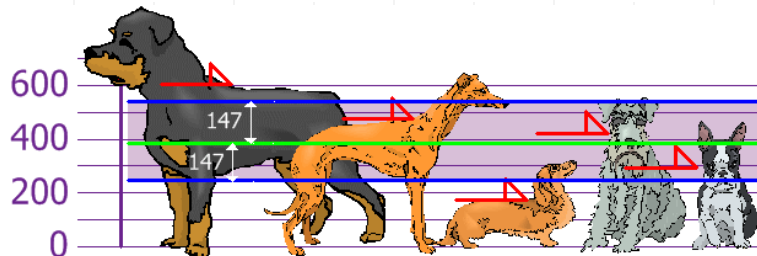


واریانس و انحراف (Deviation) استاندارد

واریانس به صورت «مقدار متوسط مربع اختلاف مقادیر از میانگین» تعریف شده است. برای محاسبه واریانس، باید گام‌های زیر را دنبال کنید:

ابتدا میانگین را پیدا کنید (میانگین ساده اعداد).
سپس برای هر عدد، مقدار میانگین را از آن تفریق کرده و سپس نتیجه را به توان دو برسانید (مربع اختلاف).
و در نهایت میانگین مربع اختلافات به دست آمده را محاسبه کنید.

$$\sigma^2 = \frac{\sum (xi - \bar{x})^2}{N}$$



روش تبدیل یا روش کوتاه برای محاسبه میانگین واریانس

گاهی برای محاسبه میانگین و واریانس میتوان داده ها را به دیگر تبدیل کرده سپس محاسبات را انجام داد

-105

120, 125, 130, 105

کم کردن عددی ثابت از همه اعداد
ترجیحا نزدیک به میانگین

15, 20, 25, 0

$(15+20+25+0)/4$

محاسبه میانگین با اعداد جدید

15

افزودن عدد ثابت به میانگین بدست آمده

$105 + 15 = 120$

داده های تبدیل شده

فرض کنید x_1, x_2, \dots, x_k با فراوانی f_1, f_2, \dots, f_k یکسری داده n تایی و $a, b > 0$ دو عدد مناسب باشند a برای تغییر مبدا اندازه گیری و b تغییر واحد اندازه گیری بکار میبرند

$$y_i = \frac{(x_i - a)}{b}$$

داده های تبدیل شده اند به سادگی نشان میدهم

$$\bar{x} = b\bar{y} + a$$

$$s_x^2 = b^2 s_y^2$$

$$s_x = b s_y$$

داده های استاندارد (Standard Data)

فرض کنید x_1, x_2, \dots, x_k با فراوانی f_1, f_2, \dots, f_k یکسری داده n تایی با میانگین \bar{x} و انحراف استاندارد S با تبدیل زیر

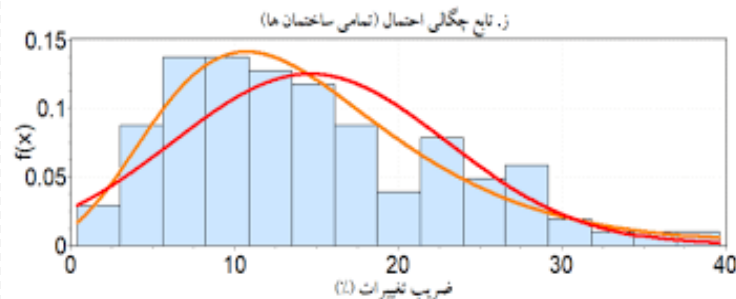
$$Z_i = \frac{(x_i - \bar{x})}{s}$$

Z_i ها را داده های استاندارد می نامند.
میانگین برای این داده ها صفر واریانس برابر ۱ است.

ضریب تغییر (Coefficient of Variance)

نسبت انحراف استاندارد به میانگین که اغلب بصورت درصد بیان میشود

$$V = \frac{s}{\bar{x}}$$



نیم برد چارک ها

نصف برد چارکهای اول و سوم نیم برد چارک ها نام دارد

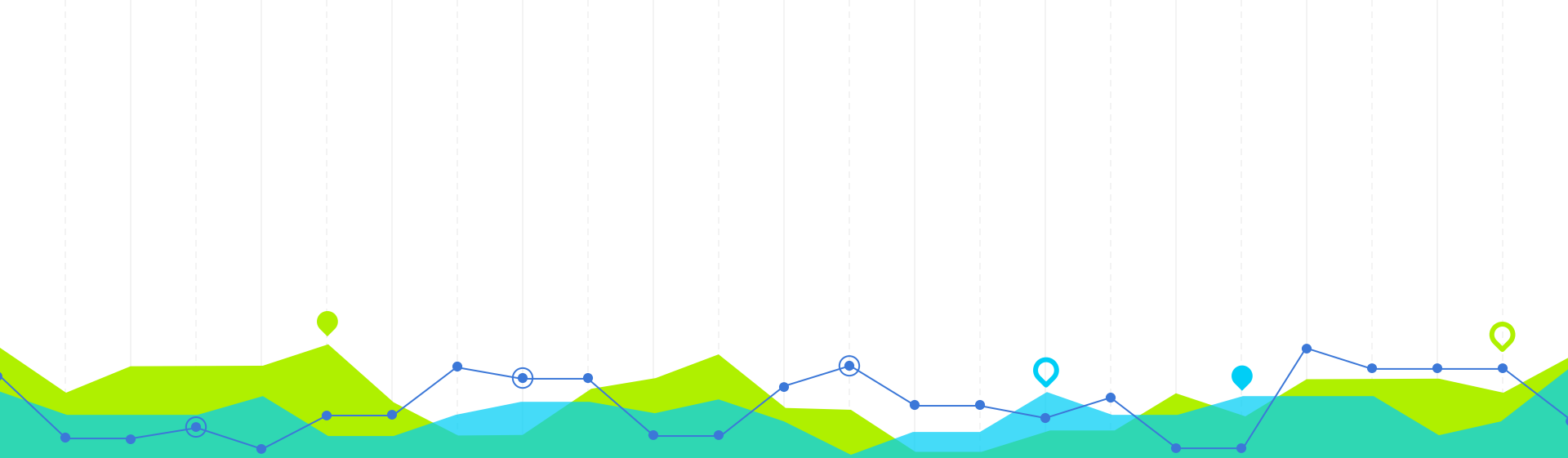
$$Q = \frac{Q_3 - Q_1}{2}$$

$$\hat{Q} = \frac{Q_3 + Q_1}{2}$$

و \hat{Q} میان چارکی نام دارد

میان چارکی نوعی معیار تمرکز است برای توزیع های متقارن $Q_2 = \hat{Q}$
معیار Q برای سنجش پراکندگی است (برای داده های بسیار کوچک و بزرگ)





چولگی و برجستگی

چون تقارن مهم است...

7

گشتاور (Moment) و گشتاور مرکزی داده (Central Moment)

فرض کنید x_1, x_2, \dots, x_k با فراوانی f_1, f_2, \dots, f_k یکسری داده n تایی باشند. میانگین توان r ام x_i ها و $x_i - \bar{x}$ ها یعنی

$$m_r = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^r}{n} \quad m'_r = \frac{\sum_{i=1}^k f_i x_i^r}{n}$$

گشتاور r ام و گشتاور مرکزی r ام داده می نامند m'_1 برابر \bar{x} و m_1 برابر صفر و m_2 برابر s^2 می باشد r معمولاً عددی طبیعی است.

اگر داده ها نسبت به میانگین متقارن باشند گشتاور مرکزی فرد یعنی m_3, m_5, \dots برابر صفر هستند.

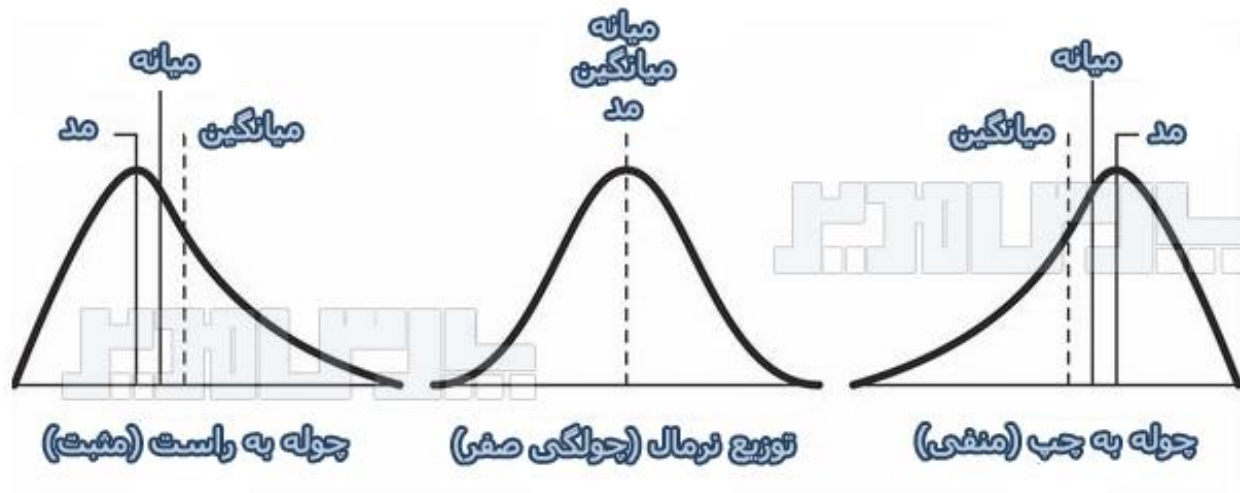
چولگی (Skewness)

چولگی در آمار نشان دهنده میزان عدم تقارن توزیع احتمالی است. اگر داده‌ها نسبت به میانگین متقارن چوله به راست: بزرگتر از صفر
چوله چپ: کوچکتر از صفر

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) \sigma^3} \quad \tilde{\mu}_3 = \frac{m_3}{s^3}$$

در حالت کلی چنانچه چولگی و کشیدگی در بازه $(-2, 2)$ نباشند داده‌ها از توزیع نرمال برخوردار نیستند. باشند، چولگی برابر صفر خواهد بود.

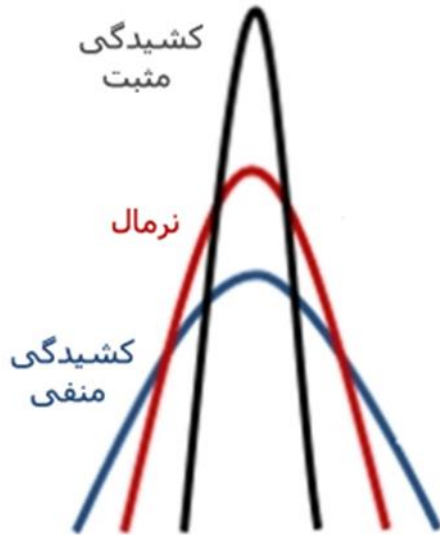




برجستگی (kurtosis)

میزان برجستگی یا پخی یا کشیدگی منحنی فراوانی نسبت به منحنی نرمال استاندارد است فرض کنید m_4 گشتاور مرکزی چهارم و S انحراف استاندارد باشد چون برای داده های نرمال $\frac{m_4}{S^4}$ به عدد ۳ نزدیک است داریم:

$$k = \frac{m_4}{S^4} - 3$$



مثبت: کشیده

منفی: پخ

صفر: نرمال

k

Kurtosis

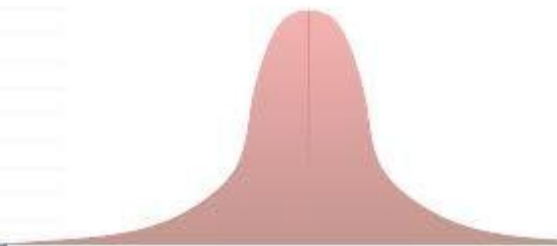
Platykurtic Distribution

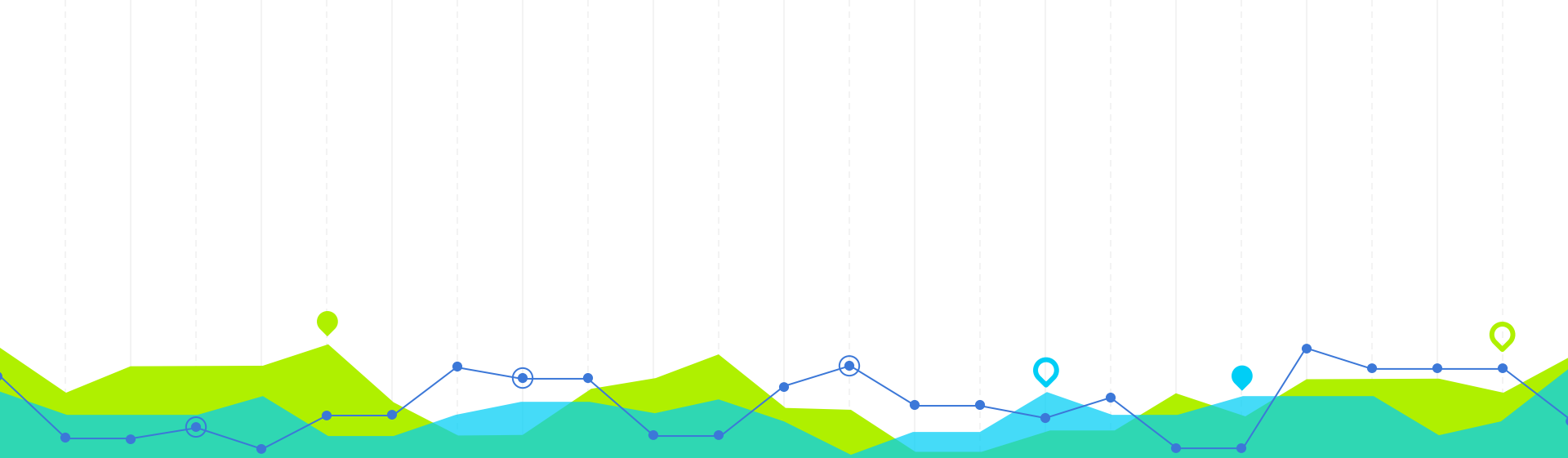


Normal Distribution
Mesokurtic Distribution



Leptokurtic Distribution





چند نمودار جدید

آنالیز داده ها کمی متفاوت تر...

8

نمودار های ساقه ای (Stem Graph)

یکی از نمودارهای آماری و شبیه به هیستوگرام است که برای داده‌نمایی داده‌های کمی به کار می‌رود تا به تصویرسازی از شکل توزیع احتمال کمک کند. این نمودارها در تحلیل کاوشی داده‌ها مفید هستند. بر خلاف هیستوگرام‌ها، نمودارهای ساقه و برگ اصل داده‌ها را دست کم تا دو رقم حفظ می‌کنند. یک نمودار ساقه و برگ ساده شامل دو ستون که با استفاده از یک خط عمومی جدا شده‌اند می‌شود. ستون سمت چپ ساقه‌ها و ستون سمت راست برگ‌ها را در بر می‌گیرد.



مثال

برای نمونه داده‌های زیر مرتب شده‌اند:

۴۴ ۴۶ ۴۷ ۴۹ ۶۳ ۶۴ ۶۶ ۶۸ ۶۸ ۷۲ ۷۲ ۷۵ ۷۶ ۸۱ ۸۴ ۸۸ ۱۰۶

| | | |
|----|--|-----------|
| ۴ | | ۴ ۶ ۷ ۹ |
| ۵ | | |
| ۶ | | ۳ ۴ ۶ ۸ ۸ |
| ۷ | | ۲ ۲ ۵ ۶ |
| ۸ | | ۱ ۴ ۸ |
| ۹ | | |
| ۱۰ | | ۶ |

۶: کلید

۶۳=۶|۳ یکای برگ:

۱۰,۰ یکای ساقه: ۱۰,۰

سپس باید تصمیم گرفت که کدام بخش از اعداد را ساقه و کدام بخش را برگ در نظر بگیریم. معمولاً آخرین رقم هر عدد را برگ، و همهٔ رقم‌های باقی‌مانده را ساقه در نظر می‌گیرند. اگر داده‌ها عددهای خیلی بزرگی باشند ممکن است آن‌ها را تا حد معینی (مثلاً تا صدگان) گرد کنند. در نمونهٔ بالا برگ‌ها را رقم یکان و ساقه را رقم دهگان در نظر می‌گیریم. در هنگام رسم، ساقه‌ها را (بدون پرش از روی عددها) در ستونشان می‌نویسیم و سپس برگ‌ها را در برابر ساقهٔ خودشان و به ترتیب از کم به زیاد قرار می‌دهیم.

نمودار جعبه ای (Box plot)

نمودار جعبه‌ای یک روش استاندارد برای نمایش توزیع داده‌ها است که براساس شاخص‌های آماری

«کوچکترین مقدار» (Minimum)،

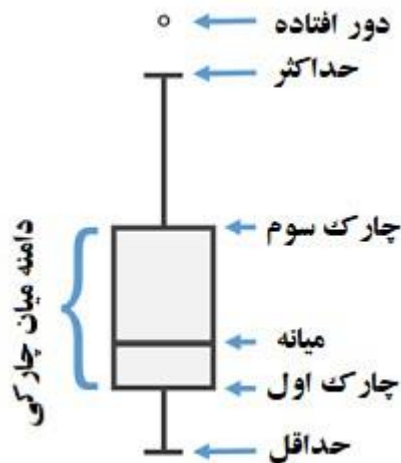
«چارک اول» (First Quartile - Q1)،

«میانه» (Median)،

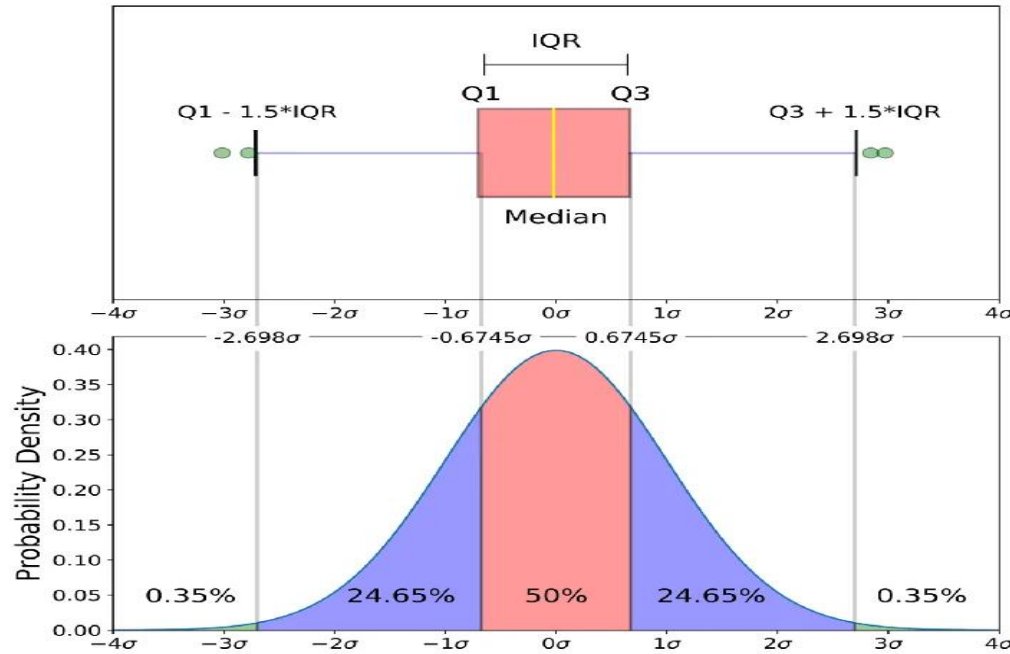
«چارک سوم» (Third Quartile - Q3)

و «بزرگترین مقدار» (Maximum) ساخته شده است.

همچنین این نمودار می‌تواند در مورد وجود داده‌های دورافتاده (Outlier یا پرت، اطلاعاتی به شما بدهد و مقدار آن‌ها را تعیین کند. همچنین نشان دادن تقارن در داده‌ها از کارهایی این نمودار است. شایان ذکر است که میزان تمرکز و حتی چولگی داده‌ها نیز در این نمودار دیده می‌شود.



مقایسه نمودار جعبه‌ای با منحنی احتمال نرمال



این مقایسه به درک شاخص‌های کوچکترین و بزرگترین مقدار و همچنین داده‌های پرت کمک خواهد کرد.



ممنون از نگاهتون

سوالی هست ؟

منابع

آمار و احتمال مقدماتی جواد بهبودیان
فرادرس



تیم ارائه



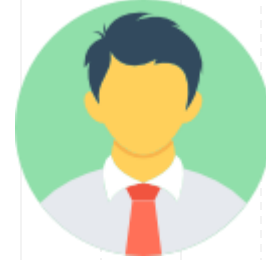
مصطفی عبدالملکی



علیرضا افروزی



سارا سادات نصر



محمد رضا همتی



حسین ضرابی