ماژول اول آمار و احتمال

بخش چهارم - آمار توصیفی

المپیاد هوش مصنوعی

۱ آمار توصیفی

۱ آمار توصیفی

مقدمه

در زندگی واقعی، ما با داده‌هایی کار می‌کنیم که تحت تأثیر تصادفی بودن قرار دارند و نیاز داریم اطلاعات را استخراج کرده و از داده‌ها نتیجه‌گیری کنیم. این تصادفی بودن ممکن است از منابع مختلفی ناشی شود. در اینجا دو نمونه از چنین موقعیت‌هایی آورده شده است:

- فرض کنید که می‌خواهیم نتیجه یک انتخابات را پیش‌بینی کنیم. از آنجایی که نمی‌توانیم کل جمعیت را مورد نظرسنجی قرار دهیم، یک نمونه تصادفی از جمعیت انتخاب کرده و از آن‌ها می‌پرسیم که قصد دارند به چه کسی رأی دهند. در این آزمایش، تصادفی بودن از نمونه‌گیری ناشی می‌شود. همچنین توجه داشته باشید که اگر نظرسنجی ما یک ماه قبل از انتخابات انجام شود، یک منبع دیگر تصادفی بودن این است که ممکن است افراد در طول این یک ماه نظر خود را تغییر دهند
- در یک سیستم ارتباط بی‌سیم، یک پیام از فرستنده به گیرنده منتقل می‌شود. با این حال، گیرنده نسخه‌ای خراب‌شده (یک نسخه پر از نویز) از سیگنال ارسالی را دریافت می‌کند. گیرنده باید پیام اصلی را از نسخه نویزی دریافت‌شده استخراج کند. در اینجا، تصادفی بودن ناشی از نویز است.

استنباط آماری

نمونه‌هایی مانند این فراوان هستند. پردازش چنین موقعیت‌هایی موضوعی در حوزه استنباط آماری است.

استنباط آماری

استنباط آماری $(Statistical\ \ inference)$ مجموعه‌ای از روش‌ها است که به نتیجه‌گیری از داده‌هایی که در معرض تغییرات تصادفی هستند، می‌پردازد.

استنباط آماری

هنگام کار بر روی مسائل استنباط آماری، از دانش نظریه احتمال استفاده می‌کنیم. با این حال، تفاوت بزرگ در اینجا این است که باید با داده‌های واقعی کار کنیم. مسائلی که تاکنون در مورد احتمال دیده‌ایم، به وضوح تعریف شده بودند و مدل‌های احتمالی از پیش به ما داده شده بودند. به عنوان مثال، ممکن است با مسئله‌ای مانند این روبه‌رو شده باشید:

- به عنوان مثال، فرض کنید یک سکه‌ی متقارن داریم و می‌دانیم که احتمال شیر آمدن آن 0.5 است. حال اگر این سکه را ۱۰ بار پرتاب کنیم، می‌توانیم احتمال مشاهده‌ی دقیقاً ۶ بار شیر را می توانیم محاسبه کنیم:

$$P(۶ = \text{number of heads}) = \binom{10}{6} p^6 (1-p)^4 = \binom{10}{6} 0.5^{10}$$

استنباط آماری

استنباط آماری

در زندگی واقعی، ممکن است این مقدار احتمال p را ندانیم، بنابراین باید داده‌هایی جمع‌آوری کنیم و از این داده‌ها نتیجه بگیریم که احتمال شیر آمدن p چقدر است.
یک راه کلی برای یک مسئله استنباط آماری داریم: یک کمیت ناشناخته وجود دارد که می‌خواهیم آن را تخمین بزنیم. ما داده‌هایی را جکع آوری می کنیم. از داده‌ها، کمیت مورد نظر را تخمین می‌زنیم.

آمار توصیفی

بعد از گرد آوری داده ها، به تنظیم، رده بندی و خلاصه کردن آنها می پردازیم. به این منظور می توان از روش های زیر استفاده نمود:

- تنظیم وطبقه بندی داده ها در یک جدول به نام جدول فراوانی
- رسم کردن نمودارهای مختلف براساس مقادیر جدول فراوانی

جدول فراوانی

مثال: فرض کنید رنگ کلاه ۱۰ دانش آموز را که در صبح یک روز سرد زمستانی به مدرسه می روند را مشاهده می کنید:
زرد،قرمز،زرد،زرد،آبی،آبی،آبی،قرمز،زرد،آبی،قرمز
جدول فراوانی به صورت زیر خواهد بود:

رنگ کلاه	شماره کلاه	فراوانی یا تعداد کلاه	تعداد کل کلاه ها/فراوانی
زرد	۱	۴	0.4
قرمز	۲	۳	0.3
آبی	۳	۳	0.3
تعداد کل کلاه ها		۱۰	1

جدول: جدول فراوانی

نمودار های فراوانی

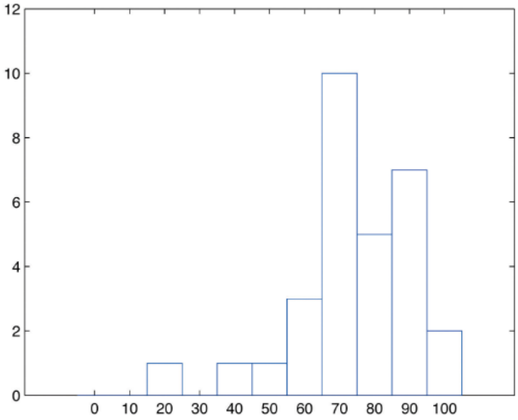
- یکی از نمودارهایی که به منظور نمایش میزان فراوانی داده ها استفاده می کنیم هیستوگرام فراوانی نشان می دهیم .
- فرض کنید داده های زیر را که مربوط به نمرات دانش آموزان یک کلاس است را در اختیار دارید:

86	80	25	77	73	76	100	90	69	93
90	83	70	73	73	70	90	83	71	95
40	58	68	69	100	78	87	97	92	74

نمودارهای فراوانی

- برای ۳۰ نمره‌ی امتحان، منطقی است که نمرات را بر اساس مقیاس استاندارد ده‌تایی گروه‌بندی کرده و تعداد نمرات در هر گروه را بشماریم. به این ترتیب، دو نمره‌ی ۱۰۰ ، هفت نمره در بازه‌ی ۹۰، شش نمره در بازه‌ی ۸۰ و ... وجود دارد. سپس نموداری مانند شکل صفحه بعد را ترسیم می‌کنیم که در آن برای هر گروه یا کلاس، یک میله‌ی عمودی رسم می‌شود که طول آن برابر با تعداد مشاهدات در آن گروه است.

نمودار های فراوانی



شکل: نمودار هیستوگرام

نمودار های فراوانی

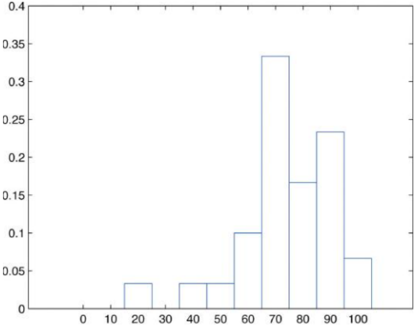
- در مثال ما، میله‌ای که با ۱۰۰ برچسب‌گذاری شده است، دارای طول ۲ واحد است، میله‌ی مربوط به ۹۰ دارای طول ۷ واحد است و به همین ترتیب ادامه دارد. در این روش، مقادیر دقیق داده‌ها از بین می‌روند، اما تعداد نمرات در هر کلاس مشخص است. این تعداد را فراوانی کلاس می‌نامند، و به همین دلیل، این نمودار هیستوگرام فراوانی نامیده می‌شود.

نمودارهای فراوانی

- همین روش را می‌توان برای هر مجموعه‌ای از داده‌های عددی اعمال کرد. مشاهدات در چندین کلاس گروه‌بندی می‌شوند و فراوانی (تعداد مشاهدات) هر کلاس ثبت می‌شود. این کلاس‌ها به ترتیب روی محور افقی قرار می‌گیرند و برای هر گروه یک میله‌ی عمودی رسم می‌شود که طول آن برابر با تعداد مشاهدات در آن گروه است. نمایش حاصل، هیستوگرام فراوانی داده‌ها خواهد بود.

نمودارهای فراوانی

هیستوگرام فراوانی نسبی برای داده‌ها است و در شکل زیر نشان داده شده است. این نمودار دقیقاً مشابه هیستوگرام فراوانی است، با این تفاوت که محور عمودی در هیستوگرام فراوانی نسبی به جای فراوانی مطلق، فراوانی نسبی را نمایش می‌دهد.



نمودارهای دایره ای

برای رسم این نمودار به هر مقدار از متغیر قسمتی از دایره را متناسب با فراوانی نسبی آن نسبت می دهیم.

تعاریف

- داده ها: واقعیت هایی دربارهٔ یک شیء یا فردند که در محاسبه، برنامه ریزی و پیش بینی به کار می روند.
- متغیر: هر ویژگی از اشیا یا اشخاص، که در اعضای جامعه یکسان نیست و معمولاً از یک عضو به عضو دیگر تغییر می کند را متغیر می گویند و عددی که به آن ویژگی یک عضو نسبت داده می شود را مقدار متغیر، یا مشاهده می گویند.
- فراوانی یک داده: تعداد دفعاتی که هر داده مشاهده می شود را فراوانی آن داده می گویند
- فراوانی نسبی یک داده: با تقسیم فراوانی هر داده به تعداد کلّ داده ها، فراوانی نسبی آن داده به دست می آید. اگر فراوانی نسبی داده ها در ۱۰۰ ضرب شود، آن گاه درصد داده ها به دست می آید.

معیار های گرایش به مرکز

- در این بخش می خواهیم به این سوال جواب دهیم که مرکز داده های داده شده کجا قرار دارد؟

میانگین

- اولین معیاری که برای مکان مرکز معرفی میکنیم مایانگین است که تعریف آن احتمالا برای شما روشن است و به صورت زیر تعریف می شود:

میانگین نمونه

میانگین نمونه ای که از n داده تشکیل شده است برابر است با :

$$\bar{x} = \frac{\sum x}{n}$$

میانگین

◀ مثال: فرض کنید که داده های زیر که بیانگر معدل دانش آموزان در یک کلاس ده نفره است را در اختیار دارید:
۵.۹ ، ۱۵ ، ۵۶.۱۲ ، ۵۵.۱۸ ، ۶.۱۰ ، ۸.۸ ، ۵۵.۱۳ ، ۹۵.۶ ، ۲۰ ، ۶۵.۱۶
در این صورت میانگین نمرات دانش آموزان این کلاس برابر است با:

$$\bar{x} = \frac{\sum x}{n} = \frac{9.5 + 15 + 12.56 + 18.55 + 10.6 + 8.8 + 13.55 + 6.95 + 20 + 16.65}{10} = 13.225$$

میانگین

- مثال: یک نمونه تصادفی داریم که سن تعدادی دانش آموز می باشد.که در آن x سن آن ها و f فراوانی آنها می باشد.میانگین نمونه را بیابید:

x	13	14	15	16	17
f	3	6	6	3	1

در این مثال، داده‌ها با استفاده از یک جدول فراوانی داده‌ها ارائه شده‌اند. هر عدد در سطر اول جدول، عددی است که در مجموعه داده‌ها ظاهر شده است؛ عدد زیر آن نشان‌دهنده تعداد دفعات وقوع آن مقدار است. بنابراین ۱۳ سه بار، ۱۴ شش بار و ...

میانگین

می توان گفت نمونه داده شده به صورت :

$$13, 13, 13, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 17$$

بنابراین میانگین داده ها برابر است با :

$$\bar{x} = \frac{\sum x}{n} = \frac{13 \times 3 + 14 \times 6 + 15 \times 6 + 16 \times 3 + 17 \times 1}{3 + 6 + 6 + 3 + 1} = 17.631$$

میانه

- برای اینکه دلیل استفاده از این معیار را درک کنید این مثال را در نظر بگیرید: فرض کنید می‌خواهیم میانگین درآمد سالانه کارکنان یک شرکت بزرگ را بررسی کنیم. یک نمونه تصادفی از هفت کارمند می‌گیریم و داده‌های نمونه را (هزار دلار در سال) به دست می‌آوریم:

$$24.8, 22.8, 24.6, 192.5, 25.2, 18.5, 23.7$$

میانگین این داده ها برابر است با: $\bar{x} = 4.74$
اما بیان اینکه "میانگین درآمد کارکنان این شرکت ۴۷،۴۰۰ دلار است" قطعاً گمراه‌کننده خواهد بود. این مقدار تقریباً دو برابر درآمد شش نفر از هفت کارمند نمونه است و به هیچ وجه به درآمد واقعی آن‌ها نزدیک نیست.
مشخص است که مشکل از کجاست: وجود یک مدیر اجرایی در نمونه، که درآمد او بسیار بیشتر از سایرین است، باعث شده که صورت کسر در فرمول میانگین نمونه بیش از حد بزرگ شود و مقدار میانگین بسیار زیاد باشد. در حالی که میانگین باید در حدود ۲۴،۰۰۰ دلار یا ۲۵،۰۰۰ دلار باشد. عدد ۵.۱۹۲ به صورت قابل توجهی از بقیه مقادیر تفاوت دارد.

میانه

حال عدد میانی یعنی 24.6 را در نظر بگیرید.
میانه ویژگی مهمی دارد: تقریباً نیمی از داده‌ها بزرگ‌تر از آن و نیمی دیگر کوچک‌تر از آن هستند. به همین دلیل، میانه مرکز واقعی داده‌ها را بهتر نشان می‌دهد.
اگر تعداد مشاهدات زوج باشد، در این صورت دو مقدار میانی وجود خواهد داشت، که در این شرایط، میانگین این دو مقدار به‌عنوان میانه در نظر گرفته می‌شود.

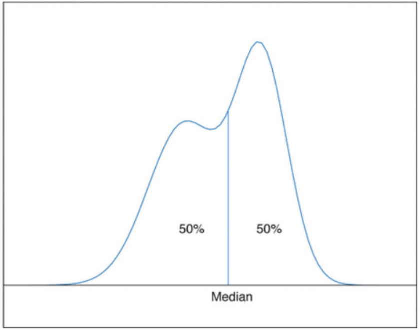
میانه

میانه

- میانه نمونه از داده ها $\tilde{x}$ که تعداد مشاهدات آن فرد است، مقدار میانی داده‌ها پس از مرتب‌سازی عددی آن‌ها است.
- میانه نمونه از داده ها $\tilde{x}$ که تعداد مشاهدات آن زوج است، برابر با میانگین دو مقدار میانی پس از مرتب‌سازی عددی داده‌ها است.

میانه

میانه مقداری است که مشاهدات یک مجموعه داده را به دو بخش تقسیم می‌کند، به‌طوری که ٪۵۰ داده‌ها در سمت چپ آن و ٪۵۰ دیگر در سمت راست آن قرار دارند. مطابق با شکل زیر، در منحنی‌ای که توزیع داده‌ها را نشان می‌دهد، یک خط عمودی که در مقدار میانه رسم شود، ناحیه را به دو بخش تقسیم می‌کند: ۵.۰ (یا ٪۵۰ از کل ناحیه که برابر با ۱ است) در سمت چپ و ۵.۰ (یا ٪۵۰ از کل ناحیه که برابر با ۱ است) در سمت راست:



میانه

- مثال:میانه اعداد زیر را حساب کنید:

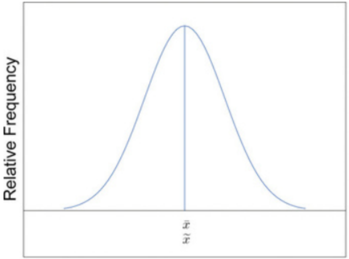
$$1.39, 1.76, 1.90, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00$$

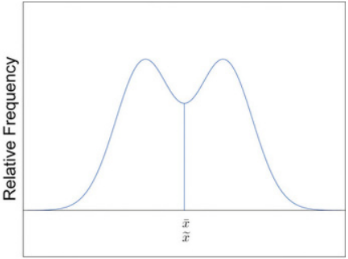
جواب: ۱۰ عدد داریم و طبق تعریف میانه میانه اعداد داده شده برابر است با :

$$\tilde{x} = \frac{2.53 + 2.71}{2} = 2.62$$

میانه و میانگین

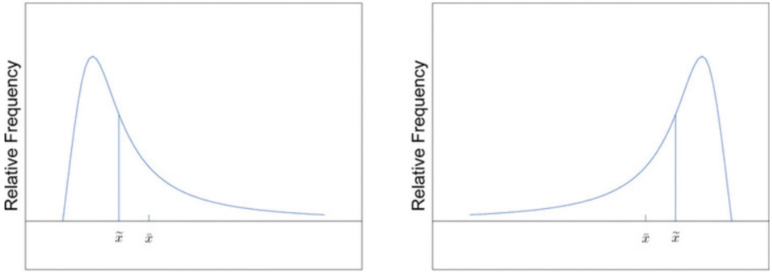
توزیع متقارن





میانه و میانگین

توزیع نامتقارن



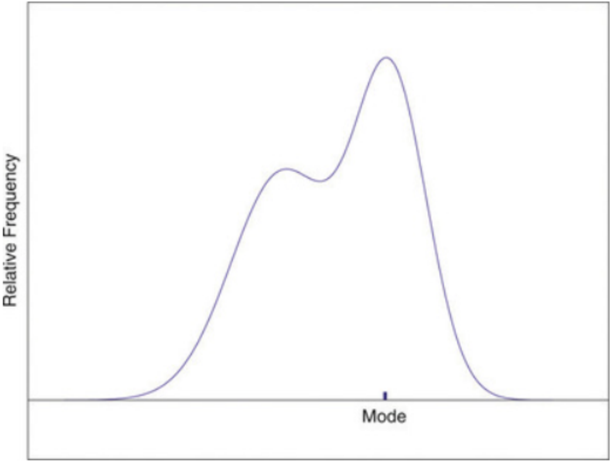
مد

ممکن است شنیده باشید که «میانگین تعداد اتاق‌های یک خانه ۶.۳ است» و این تصور برایتان جالب باشد که چگونه می‌توان ۶.۰ از یک اتاق را در نظر گرفت. در چنین مواردی، معیار مد برای تعیین مکان مرکزی داده‌ها ممکن است منطقی‌تر به نظر برسد.

مد

- مد نمونه در یک مجموعه داده، مقداری است که بیشترین فراوانی را دارد.
- در یک هیستوگرام فراوانی نسبی، بلندترین نقطه‌ی هیستوگرام نشان‌دهنده‌ی مد مجموعه داده است.

مد



مد

- برای هر مجموعه داده، همیشه دقیقاً یک میانگین و دقیقاً یک میانه وجود دارد. اما این موضوع برای مد صادق نیست.
ممکن است چند مقدار مختلف بیشترین فراوانی را داشته باشند، همان‌طور که خواهیم دید. حتی در برخی موارد، ممکن است همه مقادیر با فراوانی یکسانی ظاهر شوند، که در این صورت، مفهوم مد چندان معنادار نخواهد بود.

مد

- مثال: مد مجموعه داده های زیر را پیدا کنید:

$$-1, 0, 1, 0, 1, 1$$

- داده ی ۱ بیشترین فراوانی را دارد بنابراین مد برابر با ۱ است.

مد

- برای داده های زیر مد را حساب کنید:

x	13	14	15	16	17
f	3	6	6	3	1

دومقدار ۱۴ و ۱۵ بیشترین فراوانی را دارند بنابراین مد مجموعه ای از دو مقدار است: $\{14, 15\}$
مد به‌عنوان یک معیار برای مکان مرکزی در نظر گرفته می‌شود، زیرا در بیشتر مجموعه داده‌های واقعی، تعداد بیشتری از مشاهدات در مرکز دامنه داده‌ها قرار دارند و تعداد کمتری در انتهای و بالای دامنه ظاهر می‌شوند. مقداری که دارای بیشترین فراوانی است، اغلب در نزدیکی مرکز دامنه داده قرار دارد.

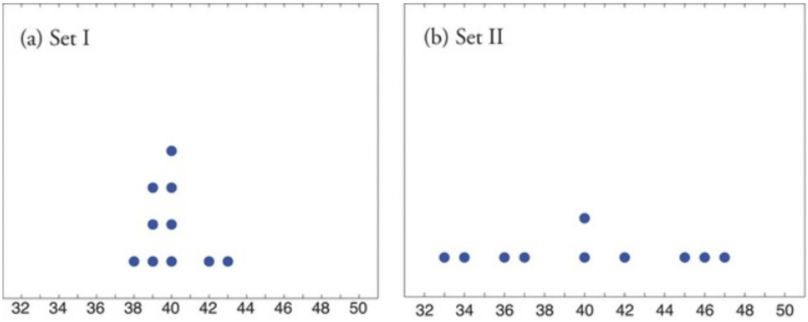
معیار های پراکندگی

دو مجموعه داده ی زیر را درنظر بگیرید:

$Set1$	40	38	42	40	39	39	43	40	39	40
$Set2$	46	37	40	33	42	36	40	47	34	45

- هر دو مجموعه داده شامل ده مقدار هستند و مقدار مرکزی آن‌ها یکسان است: هر دو میانگین، میانه و مد ۴۰ دارند. با این حال، با یک نگاه به شکل مشاهده می‌شود که این دو مجموعه کاملاً متفاوت هستند. در مجموعه داده I، مقادیر فقط اندکی از مرکز فاصله دارند، در حالی که در مجموعه داده II، مقادیر اختلاف زیادی با مقدار مرکزی دارند.

معیار های پراکندگی



معیار های پراکندگی

همان‌طور که قبلاً برای یافتن مقدار مرکزی داده‌ها از اعداد استفاده کردیم، اکنون می‌خواهیم برای هر مجموعه داده، عددی را تعیین کنیم که نشان دهد داده‌ها چقدر از مرکز پراکنده شده‌اند یا چقدر به آن نزدیک هستند. این مقادیر جدید معیارهای پراکندگی نامیده می‌شوند که سه مورد از آن‌ها را بررسی می کنیم.

دامنه

دامنه(Range)

- دامنه (R) یک مجموعه داده برابر است با تفاضل بزرگ‌ترین و کوچک‌ترین مقدار در آن:

$$R = x_{max} - x_{min}$$

که x_{max} بزرگترین مقدار و x_{min} کوچکترین مقدار می باشد.

- دامنه یکی از معیارهای پراکندگی است، زیرا نشان می‌دهد داده‌ها در چه بازه‌ای توزیع شده‌اند. دامنه کوچک‌تر نشان‌دهنده پراکندگی کمتر (داده‌ها به هم نزدیک‌تر هستند)، در حالی که دامنه بزرگ‌تر نشان‌دهنده پراکندگی بیشتر (داده‌ها فاصله بیشتری از هم دارند) است.

دامنه

- مثال: برای مجموعه داده های داده شده در اسلایدهای قبل دامنه را بدست می آوریم:
  - برای مجموعه داده یک بزرگترین مقدار برابر با ۴۳ و کوچکترین مقدار برابر با ۳۸ است بنابر این دامنه برابر است با :

$$R = 43 - 38 = 5$$

  - برای مجموعه داده دوم بزرگترین مقدار برابر با ۴۷ و کوچکترین مقدار برابر با ۳۳ است بنابراین دامنه برابر است با:

$$R = 47 - 33 = 14$$

دامنه

پس میتوانیم بگوییم:

- محدوده یک معیار برای سنجش پراکندگی است، زیرا نشان می‌دهد که داده‌ها در چه بازه‌ای توزیع شده‌اند.
- هرچه محدوده کوچکتر باشد، پراکندگی کمتری میان داده‌ها وجود دارد و برعکس، محدوده بزرگ‌تر نشان‌دهنده‌ی پراکندگی بیشتر داده‌هاست.

واریانس و انحراف معیار

- دامنه (Range) ساده‌ترین معیار پراکندگی است که فقط اختلاف بین بزرگ‌ترین و کوچک‌ترین مقدار را در یک مجموعه داده نشان می‌دهد. این معیار به ما می‌گوید که داده‌ها در چه بازه‌ای قرار دارند، اما اطلاعاتی درباره نحوه پخش شدن داده‌ها بین این دو حد ارائه نمی‌دهد.
- مثلاً اگر بیشتر داده‌ها نزدیک به هم باشند اما فقط یک مقدار خیلی دور از بقیه باشد، دامنه همچنان زیاد خواهد بود و ما را گمراه می‌کند.
- در مقابل، واریانس یک معیار دقیق‌تر و عمیق‌تر برای سنجش پراکندگی داده‌هاست. واریانس با بررسی فاصله‌ی هر مقدار از میانگین مجموعه، نشان می‌دهد که داده‌ها چقدر به‌طور کلی از میانگین فاصله دارند. بنابراین، واریانس تحت تأثیر تمام داده‌ها قرار می‌گیرد، نه فقط دو عدد انتهایی.

واریانس و انحراف معیار

واریانس

- واریانس نمونه برای یک مجموعه داده‌ی نمونه‌ای با اندازه ی n که با σ^2 نشان داده می شود به صورت زیر تعریف می شود:

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

یا با انجام کمی عملیات جبری:

$$\sigma^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n}$$

واریانس و انحراف معیار

- تا اینجا با واریانس آشنا شدیم و فهمیدیم که معیاری برای اندازه‌گیری پراکندگی داده‌ها نسبت به میانگین است. اما مشکلی وجود دارد: واریانس واحدی متفاوت از واحد داده‌ها دارد. مثلاً اگر داده‌ها بر حسب «کیلوگرم» باشند، واریانس بر حسب «کیلوگرم مربع» است که از نظر تفسیر برای ما طبیعی و قابل لمس نیست.
- برای رفع این مشکل، از ریشه‌ی دوم واریانس استفاده می‌کنیم که به آن انحراف معیار گفته می‌شود.

$$\sigma = \sqrt{\sigma^2}$$

- با این کار، نه‌تنها اطلاعات واریانس را حفظ می‌کنیم، بلکه نتیجه‌ای به‌دست می‌آوریم که هم‌واحد با داده‌های اصلی است و درک آن ساده‌تر خواهد بود.

واریانس و انحراف معیار

مثال: اگر واریانس وزن دانش‌آموزان ۹ کیلوگرم مربع باشد، انحراف معیار برابر خواهد بود با:

$$\sigma = \sqrt{9kg^2} = 3kg$$

یعنی به‌طور میانگین، وزن دانش‌آموزان حدود ۳ کیلوگرم با میانگین فاصله دارند — که خیلی راحت‌تر می‌توان آن را تفسیر کرد.

واریانس و انحراف معیار

انحراف معیار

- ریشه دوم σ^2 که با σ نمایش داده می‌شود، انحراف معیار نمونه نام دارد و به صورت زیر محاسبه می‌شود:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

واریانس و انحراف معیار

- مثال: واریانس و انحراف معیار را برای داده های زیر بدست آورید:

$$1.90, 3.00, 2.53, 3.71, 2.12, 1.76, 2.71, 1.39, 4.00, 3.33$$

- پاسخ:

$$\sum x = 1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33 = 26.45$$

$$\sum x^2 = 1.902+3.002+2.532+3.712+2.122+1.762+2.712+1.392+4.002+3.332 = 76.7321$$

$$\sigma^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n} = \frac{76.7321 - \frac{(26.45)^2}{10}}{10} = 0.677185$$

$$\sigma = 0.822912$$