بسمه تعالی

موضوع : مبانی آمار و احتمال برای ورود به علم داده و تحلیل داده

بخش هشتم:

- **شکل توزیع داده‌ها:**
  - چارک
  - توزیع متقارن
  - چولگی
  - کشیدگی

در این پارت از اموزش به بررسی **شکل توزیع ها** می پردازیم .

در اموزش های قبلی درباره آمار **معیار پراکندگی** صحبت کردیم .

یاد گرفتیم که میانگین، میانه و نما مرکز داده‌ها را نشان می‌دهند و انحراف معیار نیز میزان پراکندگی آن‌ها را مشخص می‌کند.

اما آیا این اطلاعات برای شناخت کامل داده‌ها کافی است؟

فرض کنید دو مجموعه داده داریم که **میانگین و انحراف معیار یکسانی دارند.**

آیا می‌توانیم نتیجه بگیریم که این دو مجموعه کاملاً مشابه هستند؟

پاسخ خیر است.

زیرا ممکن است شکل توزیع داده‌ها با یکدیگر متفاوت باشد.

به همین دلیل در آمار علاوه بر مرکز و پراکندگی، به **شکل توزیع داده‌ها** نیز توجه می‌کنیم.

انواع شکل توزیع ها :

۱- چارک
۲- متقارن
۳- چولگی مثبت یا منفی
۴- کشیدگی مثبت یا منفی

۱– چارک:

چارک‌ها (Quartiles) مقادیری هستند که داده‌های مرتب‌شده را به **چهار قسمت** تقسیم می‌کنند.

یعنی تقریباً:

- ٪۲۵ داده‌ها در قسمت اول
- ٪۲۵ در قسمت دوم
- ٪۲۵ در قسمت سوم
- ٪۲۵ در قسمت چهارم

قرار می‌گیرند.

سه چارک اصلی داریم:

- **چارک اول (Q1)**
- **چارک دوم (Q2)**
- **چارک سوم(Q3)**

چارک اولQ1

چارک اول یا Q1 مقداری است که تقریباً ۲۵ درصد داده‌ها پایین‌تر از آن قرار دارند.

مثلاً اگر Q1 سن مشتریان برابر با ۳۰ سال باشد، یعنی تقریباً ۲۵ درصد مشتریان سن کمتر یا مساوی ۳۰ سال دارند.

چارک دومQ2 :

چارک دوم همان **میانه (Median)** است.

یعنی:

Q2 = Median

در نتیجه تقریباً ۵۰ درصد داده‌ها پایین‌تر و ۵۰ درصد بالاتر از Q2 قرار می‌گیرند.

مثلاً اگر:

Q2 = 38سال

باشد، می‌توان گفت میانه سن مشتریان ۳۸ سال است.

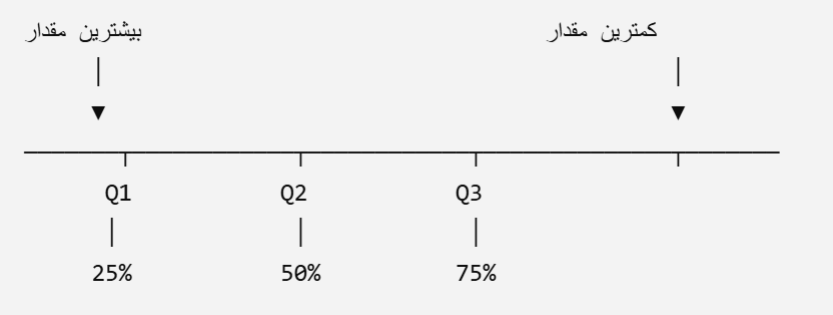
چارک سومQ3

چارک سوم یا **Q3**مقداری است که تقریباً **۷۵ درصد داده‌ها پایین‌تر از آن** قرار دارند.

مثلاً اگر:

Q3 = 50سال

باشد، تقریباً ۷۵ درصد مشتریان سن کمتر یا مساوی ۵۰ سال دارند.



۲– متقارن :

که به آن نرمال هم گفته می شود. ابتدایی ترین و ساده ترین نوع توزیع داده هاست . که در آن داده ها به صورت متقارن و کاملا قرینه نسبت به میانگین پراکنده شده اند.





دراین حالت مقادیر آماره ی میانگین / میانه و مد با هم برابر است.

به عنوان مثال: سن مشتریان یک بان را اندازه می گیریم و مشاهده می شه که داده ها مقداری نزدیک به میانگین ۳۵ سال را دارند و تعداد افراد کم سن یا با سن بالا در داده ها کم است در این صورت می توان گفت داده ها از توزیع متقارن یا نرمال پیروی می کنند.

۳– چولگی :

در این حالت داده ها به یک سمت بیشتر میل می کنند یا کشیده می شود. در توزیع چوله، داده‌ها به صورت متقارن در اطراف مرکز توزیع قرار نگرفته‌اند و یک دُم توزیع نسبت به سمت دیگر کشیده‌تر است.

اگر دُم توزیع به سمت مقادیر بزرگ‌تر و در نتیجه به سمت راست کشیده شود، به آن چولگی به راست یا چولگی مثبت (Positive Skewness) می‌گوییم.



نما < میانه < میانگین

اگر دُم توزیع به سمت مقادیر کوچک‌تر و در نتیجه به سمت چپ کشیده شود، به آن چولگی به چپ یا چولگی منفی (Negative Skewness) می‌گوییم.



میانگین > میانه > نما

به عنوان مثال:

برای مثال اگر سن مشتریان یک بانک را در نظر بگیریم:

حالت اول — چولگی مثبت:

فرض کنید بیشتر مشتریان سن‌های نسبتاً پایین دارند، مثلاً بین ۲۰ تا ۴۰ سال، اما تعداد کمی مشتری با سن‌های بسیار بالا، مثلاً ۷۰ یا ۸۰ سال، وجود دارند.

این چند مقدار بزرگ باعث ایجاد دُم در سمت راست می‌شوند.

بنابراین:

چولگی = مثبت (راست)

حالت دوم — چولگی منفی:

اگر بیشتر مشتریان سن‌های نسبتاً بالا داشته باشند، مثلاً ۵۰ تا ۷۰ سال، اما تعداد کمی مشتری بسیار جوان، مثلاً ۱۸ یا ۲۰ ساله، وجود داشته باشند، دُم توزیع به سمت مقادیر کوچک‌تر کشیده می‌شود.

بنابراین:

چولگی = منفی (چپ)

۴– کشیدگی :

تیزی یا پخی نمودار را نشان می دهد.

اگر کشیدگی مثبت باشد ، قله بلندتر و نمودار باریک تر و احتمال مشاهده مقادیر دورتر از میانگین بیشتر است.

اگر کشیدگی منفی باشد، قله کوتاه و نمودار پهن تر خواهد بود و داده ها یکنواخت تر پخش شده و تمرکز اطراف میانگین کمتر است.

