Ep 4 - Descriptive Statistics 2

Nov 29, 2023

Percentiles, Quartiles and Box-plots

Percentile

A **percentile** is a value below which a certain percentage of observations lie.

- Percentile from value:

$$P_p = \frac{k}{n} \times 100$$

- Value from Percentile:

$$P_k = \frac{k}{100}(n+1)$$

> **Example**
>
> If your percentile rank in an exam is 85%, it means that you have scored more than 85% of people who gave the exam with you.

Quartiles

Quartiles are values that divide a dataset into four equal parts, each containing approximately 25% of the data. They are used to understand the spread and distribution of data, particularly in statistics and data analysis. The three quartiles, which divide the data into four parts, are:

1. **First Quartile (Q1):** Also known as the lower quartile, it *is the 25th percentile* of the data. Q1 divides the lowest 25% of the data from the rest.
2. **Second Quartile (Q2):** This is equivalent to the median, which divides the data into two equal halves, with 50% of the data below and 50% above it. Q2 *is the 50th percentile*.
3. **Third Quartile (Q3):** Also known as the upper quartile, it *is the 75th percentile* of the data. Q3 divides the lowest 75% of the data from the highest 25%.

An Explanation of Percentiles and Quartiles

Box-Plot

We make use 5 points to make sense of the shape of the data, namely:

- Minimum
- First Quartile (Q1)
- Median
- Third Quartile (Q3)
- Maximum

We can also detect outliers using it. To remove outliers we use a concept of **upper and lower fence** to get the boundaries of the data. We use IQR (Inter Quartile Range) to calculate the fences.

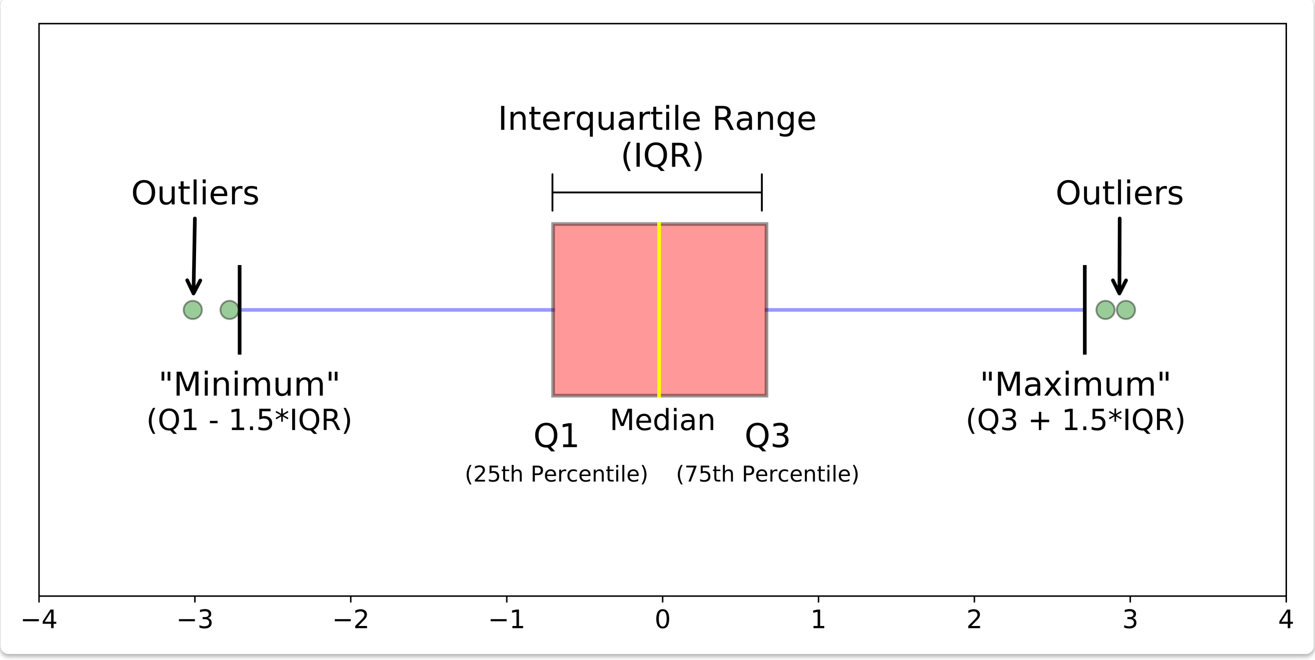
$$IQR = Q3 - Q1$$

$$\text{Lower Fence } = Q1 - 1.5(IQR)$$

$$\text{Upper Fence } = Q3 + 1.5(IQR)$$

Tip

Best way to view and manage the outliers are by using percentiles and quartiles. We make use of 5-number summary as well and also construct a box-plot with the help of it.



A Short Video on Box-Plots

Covariance and Correlation

Covariance

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures *how much two variables change together*, such that when one variable increases, does the other variable also increase, or does it decrease?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. A covariance of zero indicates that the variables are not linearly related.

$$\text{Population Covariance: } \sigma(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

$$\text{Sample Covariance: } s(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Drawback of Covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.

Explanation of What Covariance Is
A Visual Explanation of Covariance

Tip

Covariance of a variable with itself is equal to the variance of the variable.

Correlation

Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.

$$\text{Correlation Coefficient: } \rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

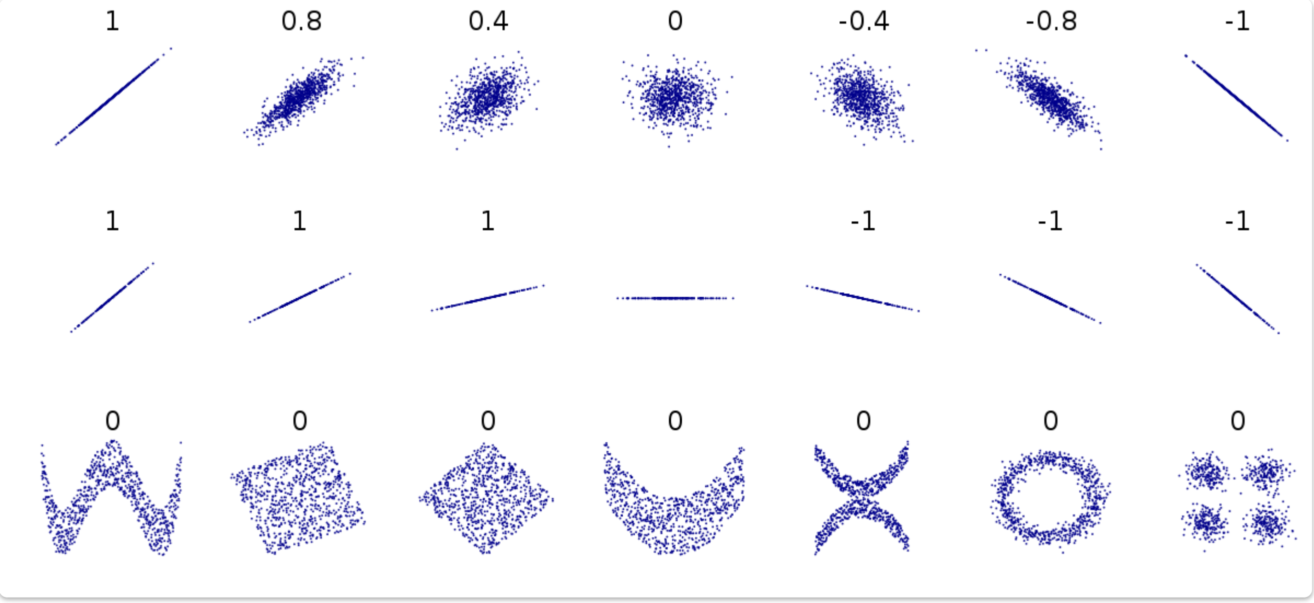
$$\text{Pearson Correlation Coefficient: } r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1.

$$\text{Corr } = -1 \implies \text{Strong Negative Correlation}$$

$$\text{Corr } = 0 \implies \text{No Correlation}$$

$$\text{Corr } = 1 \implies \text{Strong Positive Correlation}$$



⚠ Correlation Does Not Imply Causation

just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other variable's behaviour.

Tip

While correlations can provide valuable insights into how different variables are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.

Pearson's Correlation