Measure of spread

Histogram

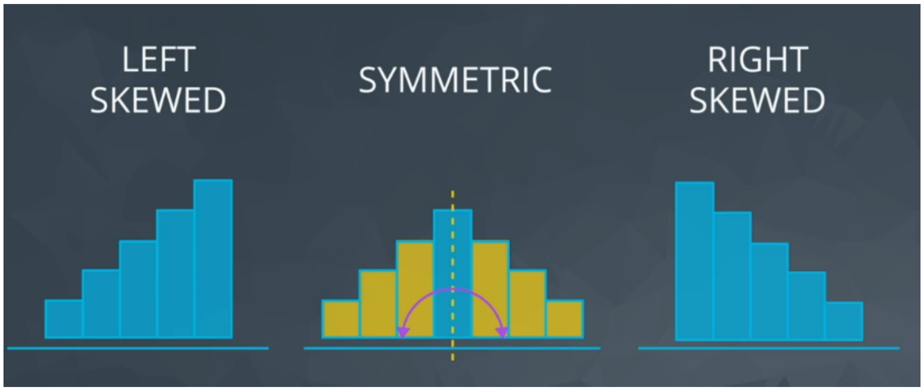
The most common way to visualize quantitative data.

5 number summary gives values for calculative range and interquartile range.

- First quartile value Q_1 is the median of the first half
- Third quartile value Q_3 is the median of the second half
- Second quartile value Q_2 is the median of all values
- When the dataset has odd number of values the middle number (Q_3) is not considered in the first nor the second half
- Range = max − min
- Interquartile range IQR $= Q_3 - Q_1$

We can use theses 5 numbers to visualize what is called **box plot**.

Special shapes of histogram



- one of the well-known symmetric histograms is the normal distribution which is also known as bell curve
- symmetric histogram has also a symmetric box plot
- left skewed histogram is a result of median < mean
- right skewed histogram is a result of median > mean

Standard deviation

The most common way to measure the spread which tells us on average **how much each point varies from the mean** of the points.

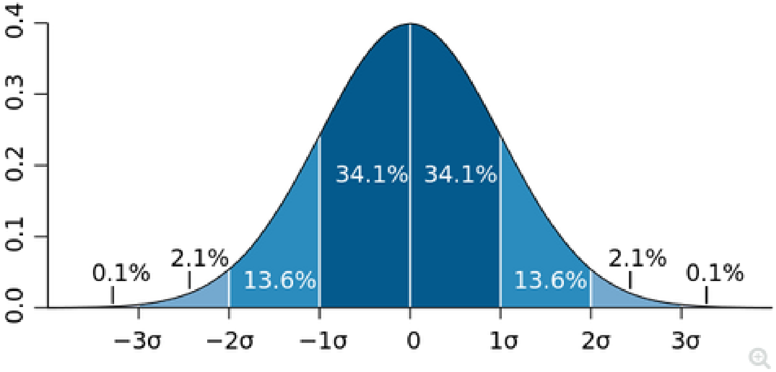
We use std deviation to describe spread with **only one** number.

- used to compare the spread of different groups
- higher standard deviation of stock prices means higher risk

Normal distribution

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

- mathematical equation of normal distribution (bell curve)
- μ is the mean
- σ is the standard deviation



Measure of outliers

Outliers are data points that fall very far from the rest of the values in our dataset.

Outliers significantly increase mean and standard deviation.

Steps to analyze a dataset:

1. Plot data and try to handle outlier (remove them)
2. If the data is normally distributed (bell-shaped), mean and std deviation give lots of information about the dataset
3. If the data is skewed 5 number summary gives more useful information