W2 - Basic Statistics

Mean - most commonly used average.

Calculation: Sum divided by count

Intuitive: Represents the number you could replace all individual values with to get the same total

Easily skewed by outliers

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

$\bar{x}$: Mean
$\sum x$: Sum of all values
N : Number of values

Median - middle value.

Calculation:

N is odd, find the middle number when ordered

N is even, find the average of the two middle numbers when ordered

Handles outliers well.

$$\tilde{x} = \left(\frac{N+1}{2}\right)^{th} term$$

$$\tilde{x} = \frac{\left(\frac{N}{2}\right)^{th} term + \left(\frac{N+2}{2}\right)^{th} term}{2}$$

$\tilde{x}$: Median
N : Number of values

Mode - most frequently occurring value.

Other averages

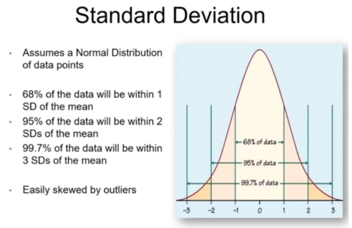
- **Geometric Mean**: percentages, areas, volumes..
- **Harmonic Mean**: used for rates mph.
- **Weighted Average**: a type of mean where some values count for more than others, often used in course gradings.

Standard Deviation - how spread your data is.

Larger SD = More speed

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

σ : Standard Deviation
Σ: Sum
x : Each value in the dataset
$\bar{x}$: Mean
N: Number of values



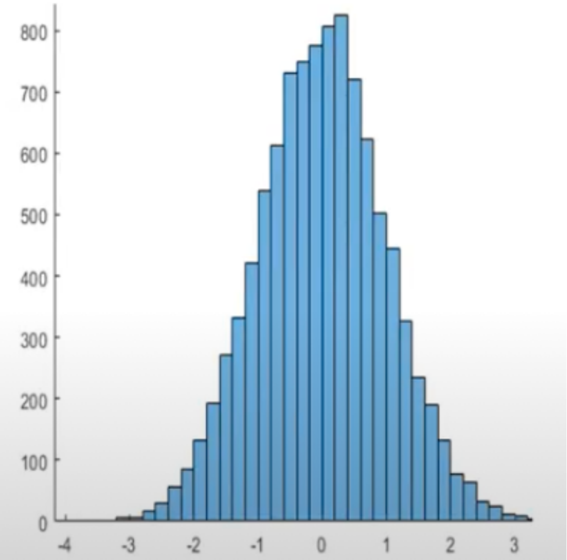
Standard Deviation

- Assumes a Normal Distribution of data points
- 68% of the data will be within 1 SD of the mean
- 95% of the data will be within 2 SDs of the mean
- 99.7% of the data will be within 3 SDs of the mean
- Easily skewed by outliers

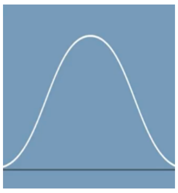
68% of data
95% of data
99.7% of data
-3
-2
-1
0
1
2
3

Distributions - how data is distributed.

Histograms - chart that plots the distribution of numeric variables as a series of bars.

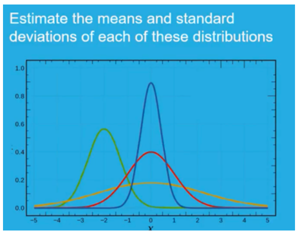


Normal Distribution - Bell curve/ Gaussian Distribution



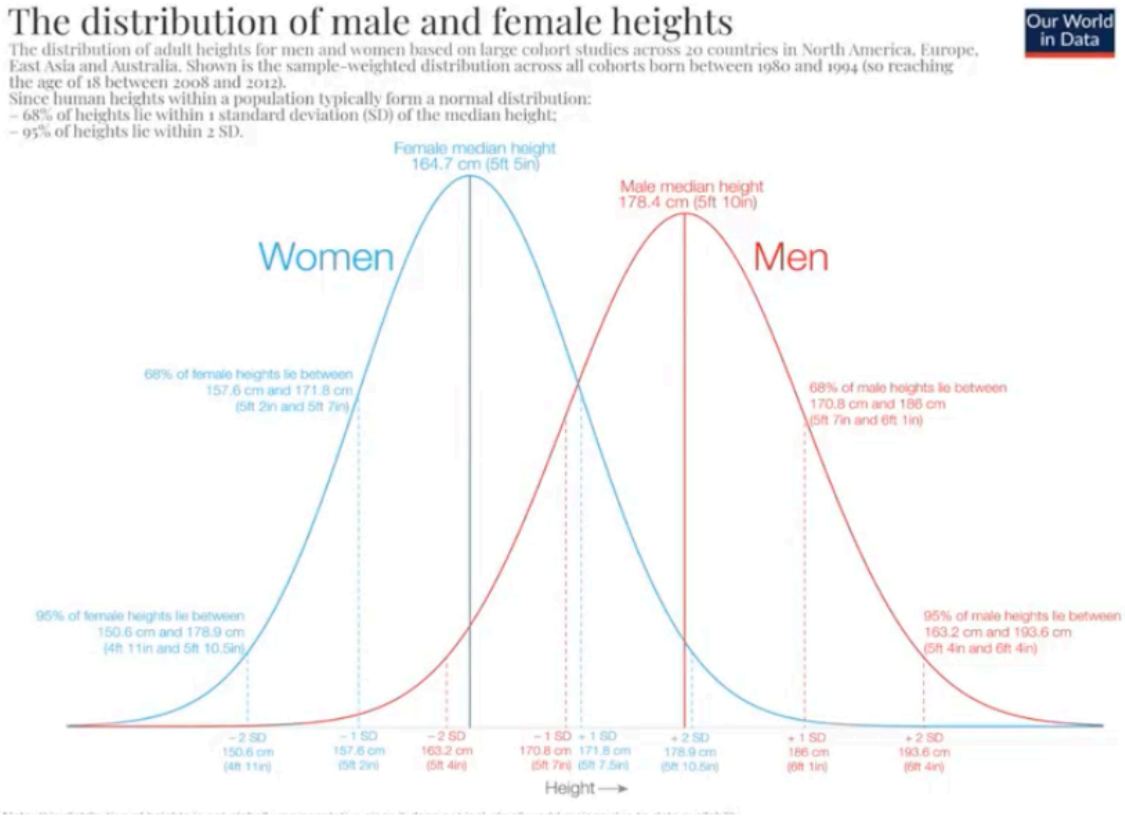
Data here distributed symmetrically with mean and median in the middle.

Kernel Density Estimates - y axis showing probabilities, not counts, x-axis=mean



Estimate the means and standard deviations of each of these distributions

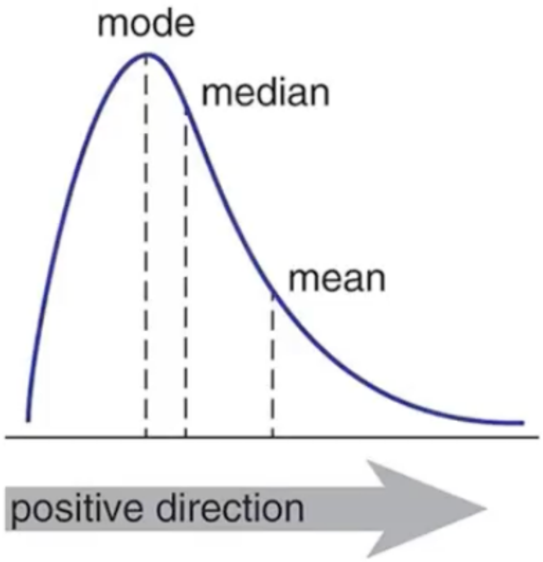
Example:



The distribution of male and female heights

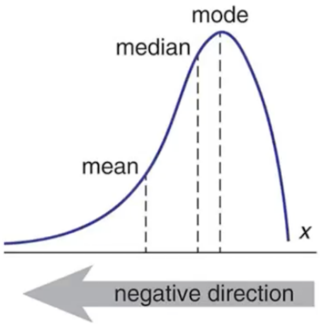
The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.

Positive Skew - tail is in the direction of positive, larger values



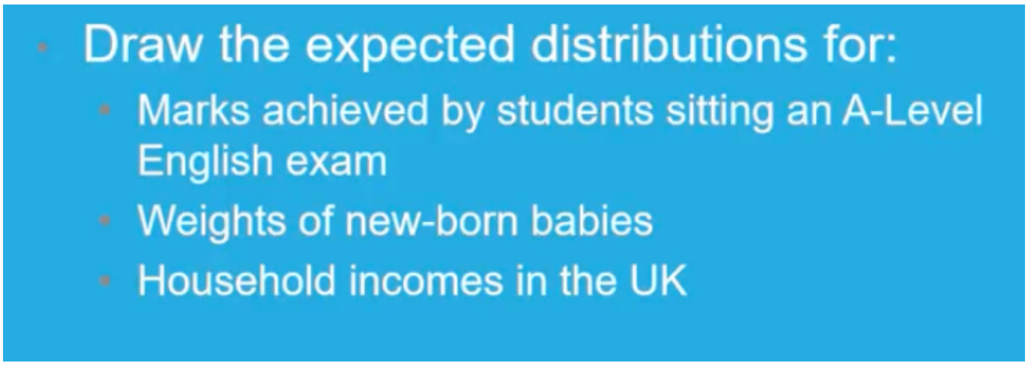
Mean > Median > Mode

Negative Skew - tail is in the direction of negative, smaller values



Mean < Median < Mode

Exercise:



- Draw the expected distributions for:
  - Marks achieved by students sitting an A-Level English exam
  - Weights of new-born babies
  - Household incomes in the UK

Characteristics of Normally Distributed Data:

- Symmetrical around the **mean**
- Mean = Median = Mode
- Data within:
  - ±1 standard deviation ≈ 68% of values
  - ±2 standard deviations ≈ 95%
  - ±3 standard deviations ≈ 99.7%

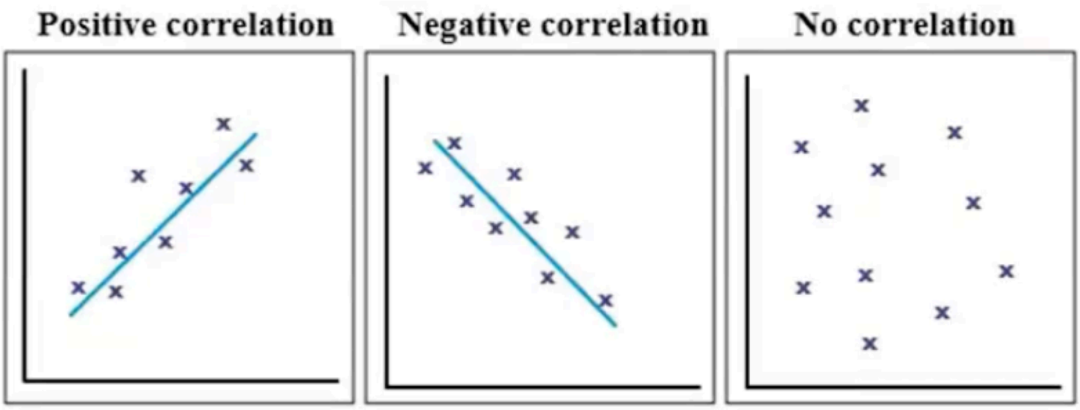
Data that often follow a normal distribution:

- Heights of people in a large population
- Test Scores in large groups
- Daily Temperature Variations

Outliers - a small number of very extreme values can have a great effect on your mean & SD.

What to do with outliers? obvious error or problematic values should be removed.

Correlations - linear relationship between 2 continuous variables



🎯 Think of it like this:

- **Discrete = Counting apples** 🍎: 1, 2, 3, never 2.7 apples.
- **Continuous = Pouring water** 💧: You can measure 1.5L, 1.53L, 1.532L... it never ends.

Examples:

📈Height vs Weight (Taller people tend to weight more)

📚 Study Time vs Test Scores

Correlation Coefficient - r = how closely two variables are related

Pearson's Correlation Coefficient

r = 0 no correlation

r = 1 positive correlation

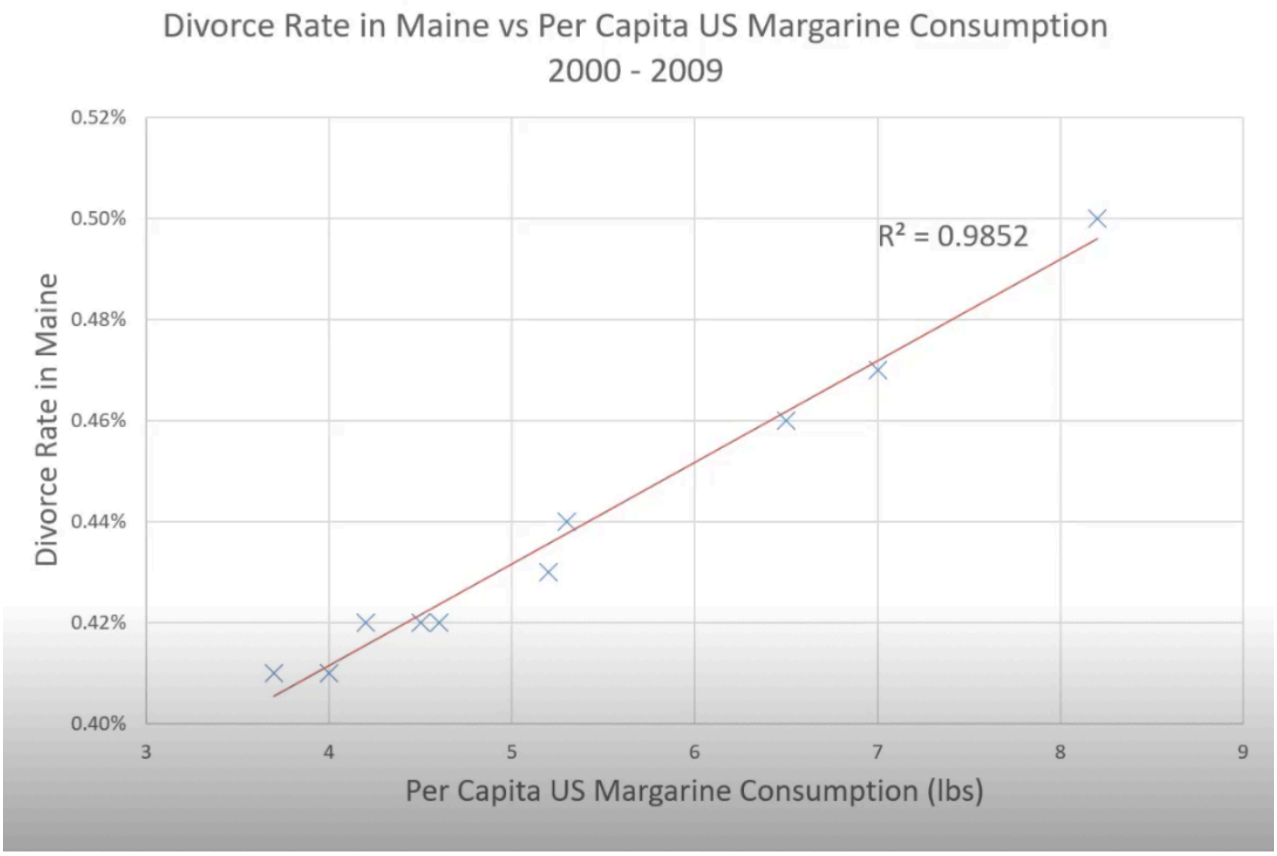
r = -1 negative correlation

R-Squared R^2 shows how well the data fit the regression model (the goodness of fit) - **scalar(distance) not vector (often used instead of r)**

Calculated based on distance of each point from a line of best fit. Amount of variation in one variable that can be explained by another.

R^2 = 75% means 75% of changes in Y can be accounted for by changes in X

0<=R^2<=1



These 2 are correlated even if they're unrelated, so there is a 3rd factor which is almost aways time. No causation at all.