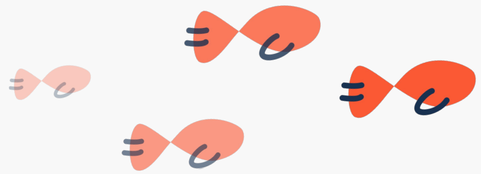
neue fische

School and Pool for Digital Talent

Descriptive Statistics



Today's Objective

Descriptive Statistics

Why?

- Data is huge and must be summarised to be understandable
- Descriptive statistics allows us to boil many data into small amount of information
- Aggregation and summation has to be done carefully to ensure the information remains truthful and useful

What we aim for today:

- Lay the foundation of statistical concepts
- Introduction to the basic charts used to display statistics
- Practice

Type of Statistics

- Descriptive Statistics
- Inferential Statistics
- Segmentation / Classification
- Predictive Statistics

Examples of Business Questions

- Descriptive Statistics
  - How much are our customers spending with us?
  - Are all our customers the same?
- Inferential Statistics
  - Can we group customers based on how valuable they are?
- Segmentation / Classification
  - What are the common characteristics of the customers in these groups?
- Predictive Statistics
  - Will this new customer become a ‘more valuable’ customer?

Agenda

Today’s topics

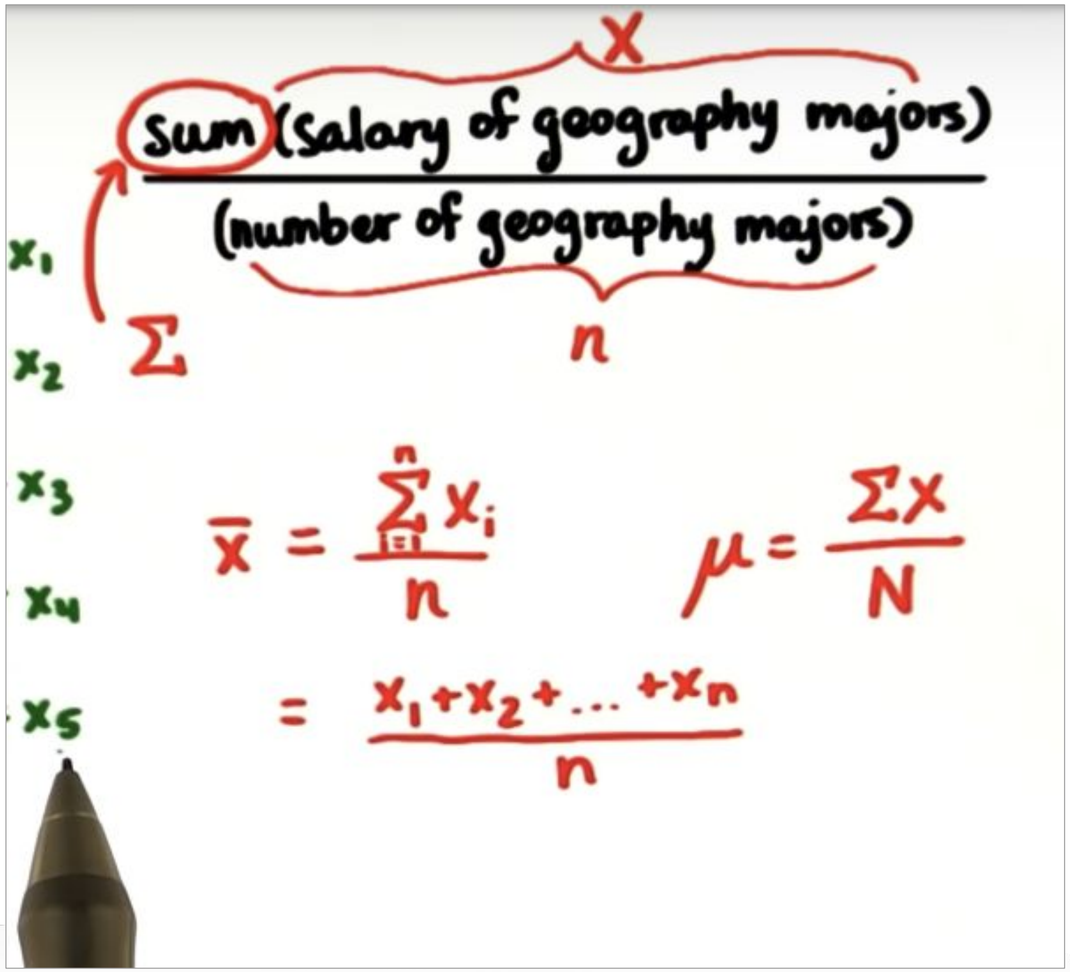
Types of Data
Basic aggregation
Measures of Central Tendency
Measures of Variation
Frequency distributions

It's all Greek to me… but it saves a lot of effort

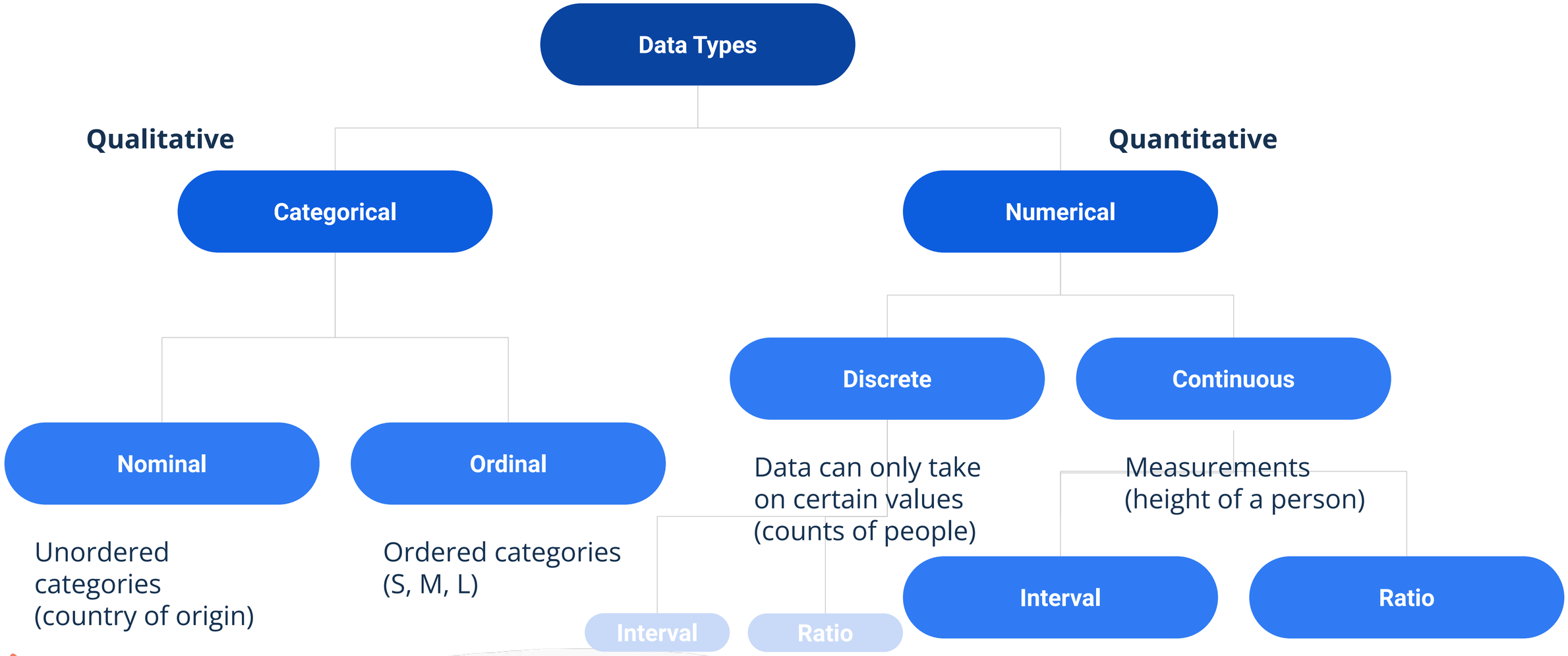
Important letters

- sigma: σ
- Sigma: Σ
- Pi: Π
- Theta: θ
- Delta: **Δ**

- X / i
- n / N
- Mu: μ
- x-bar: x̄
- y-hat: ŷ



Types of data



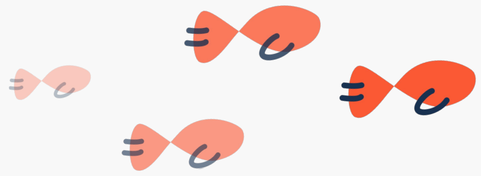
Source: On the Theory of Scales of Measurement Author(s): S. S. Stevens

Types of data

	Nominal	Ordinal	Interval	Ratio
Categories and labels Count	✓	✓	✓	✓
Ranked Categories Rank and order		✓	✓	✓
Equal intervals Add and Substract			✓	✓
True zero Multiply and divide				✓

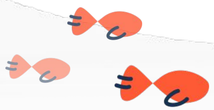
Source: On the Theory of Scales of MeasurementAuthor(s): S. S. Stevens

Descriptive Statistics



Mathematical aggregation of your data

- Sum
- Count
- Minimum
- Maximum
- Distinct Count



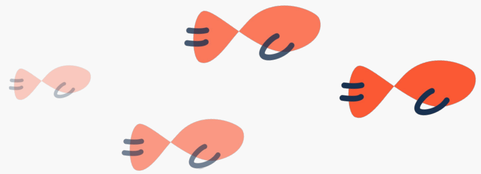
Counting:
Frequency distribution

Number of times a data value occurs.

Shown in a frequency table

Data Value	Frequency
1	4
2	2
3	3
4	3
5	3
6	5

Measures of central tendency



Measures of central tendency describe a sample

- One of a set of summary statistics
- A single value describes entire sample
- Identifies the central position within that set of data
- Some measures of central tendency are better to use than others under certain circumstances

Mean

Sample mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean is particularly susceptible to the influence of outliers

Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Median

The median is less susceptible to the influence of outliers

Middle score for a set of data

Method:

1. Bring data into order
2. Median is value in the middle (odd number of scores)
3. Median is average of the two middle scores (even number of scores)

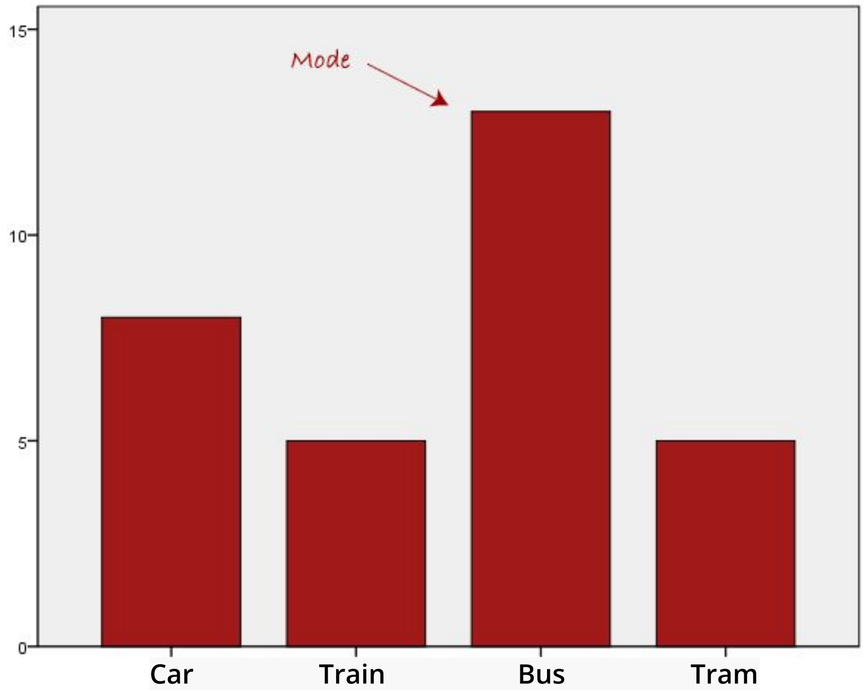
65	55	89	56	35	14	56	55	87	45	92

14	35	45	55	55	**56**	56	65	87	89	92

14	35	45	55	**55**	**56**	56	65	87	89

Mode

Most frequent score in our data set



Often used for categorical data - which is the most common category?

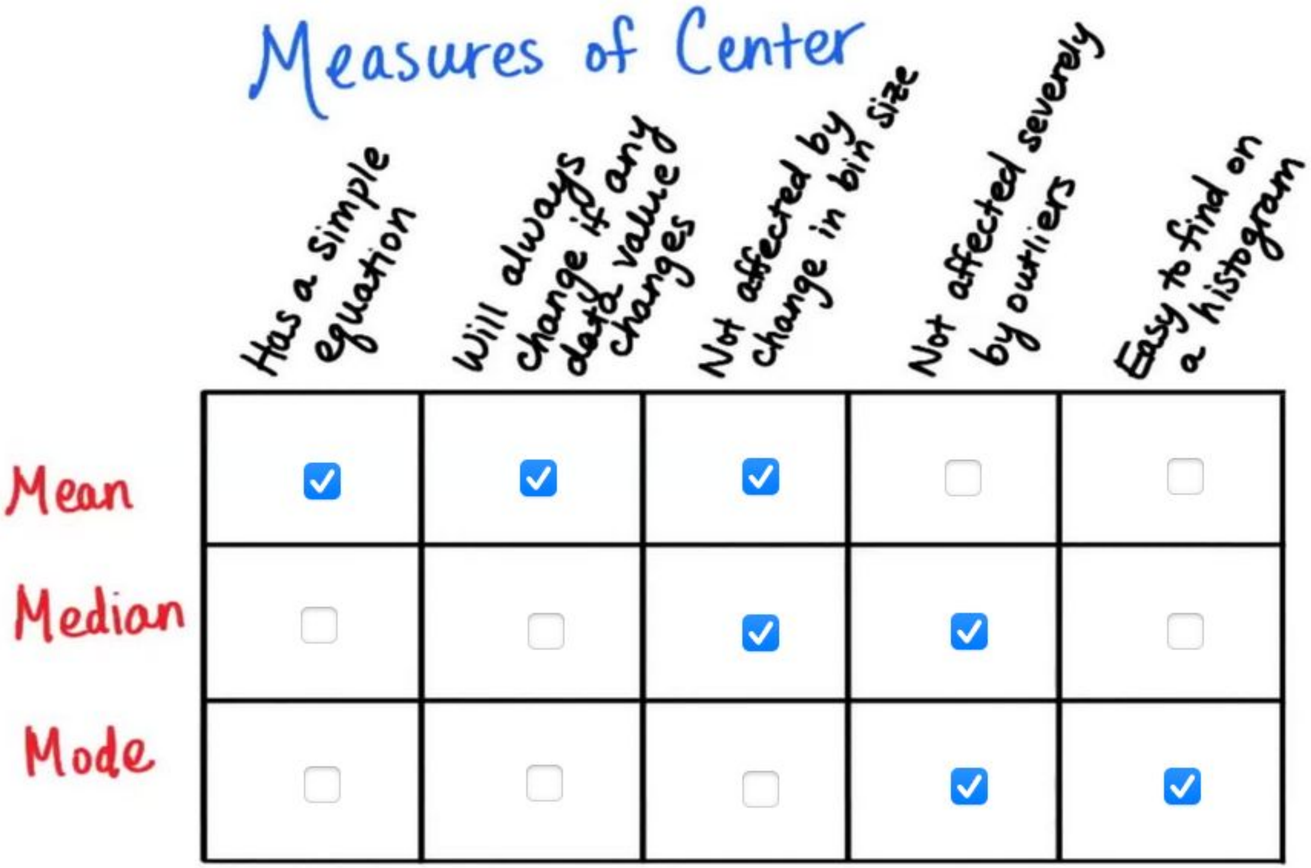
Mode is not unique!

Mode can be far away from the rest of the data in the data set

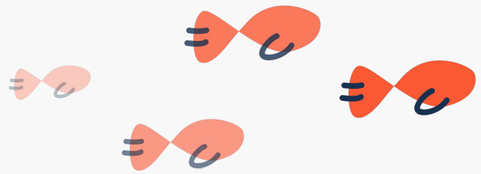
Mean versus Median versus Mode

Choice of metric makes a difference!

Outliers or skewed data affect central tendency measures differently!

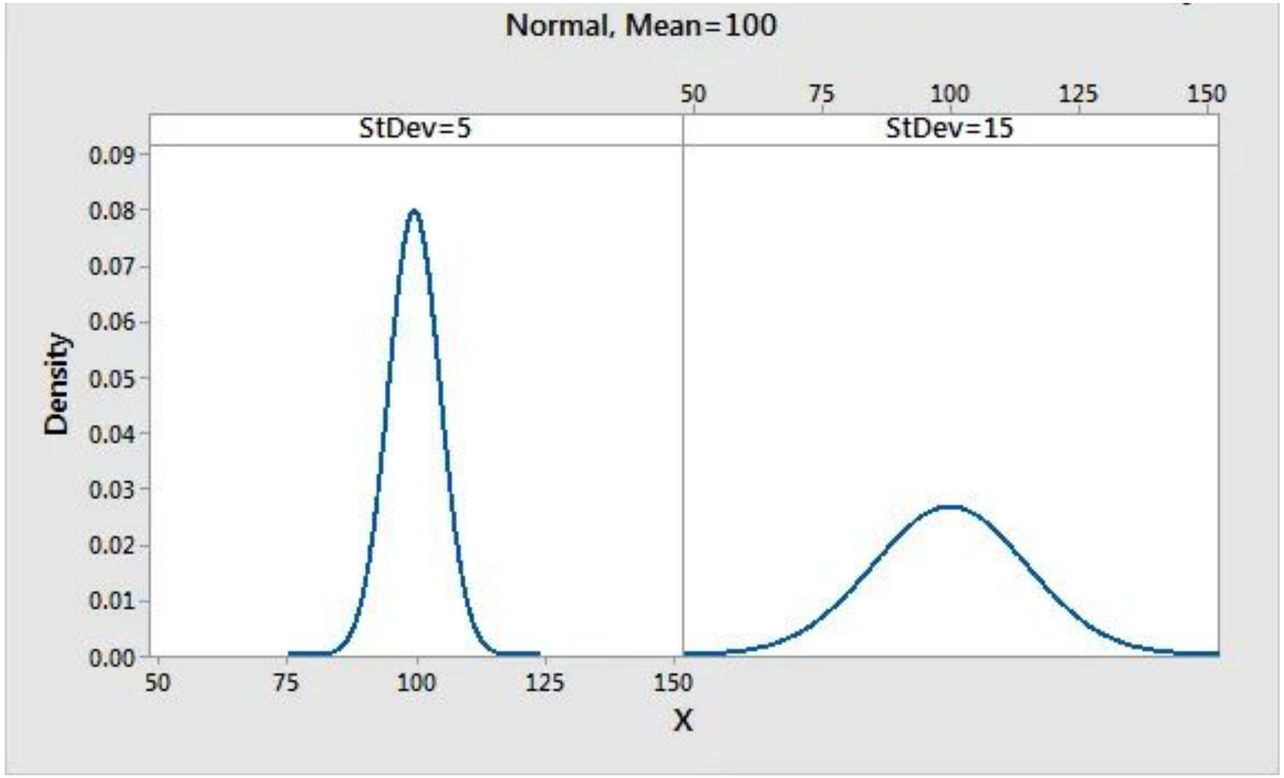


Measures of variation



Measures of variation answer the question: How spread out is my data?

- variability = spread = dispersion



Range

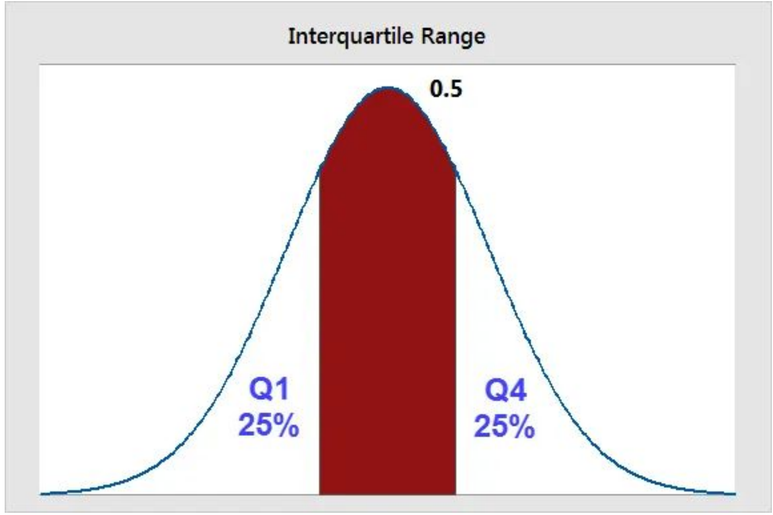
equal to the distance
from the smallest to largest value

Which dataset has the higher range?

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

Interquartile Range (IQR)

Middle half of the data: 75th percentile - 25th percentile



Percentile?

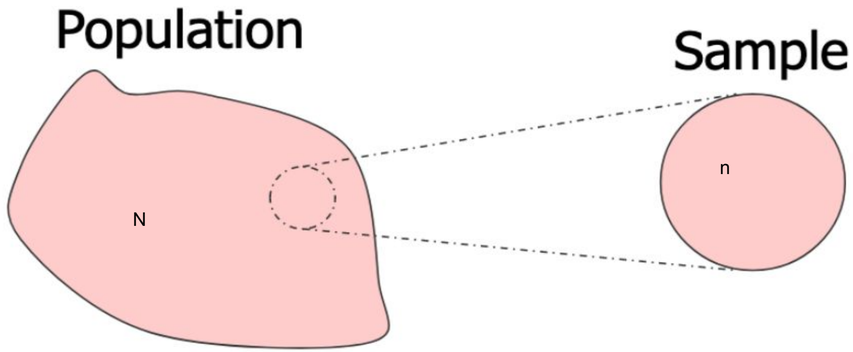
- pth percentile: value of x such that p% of the data is less than or equal to x
- Top 10% = 90th Percentile
- Special Percentiles:
  - Max: 100th percentile
  - Min: 0th percentile
  - Median: 50th percentile
  - Quartiles: 25th and 75th percentiles

IQR is like range but not affected by outliers and explains a broad central range

Source of Data: Samples versus population

Observations vs Aggregations
Measurement method is important

Samples vs Populations
The observed cases vs all possible cases



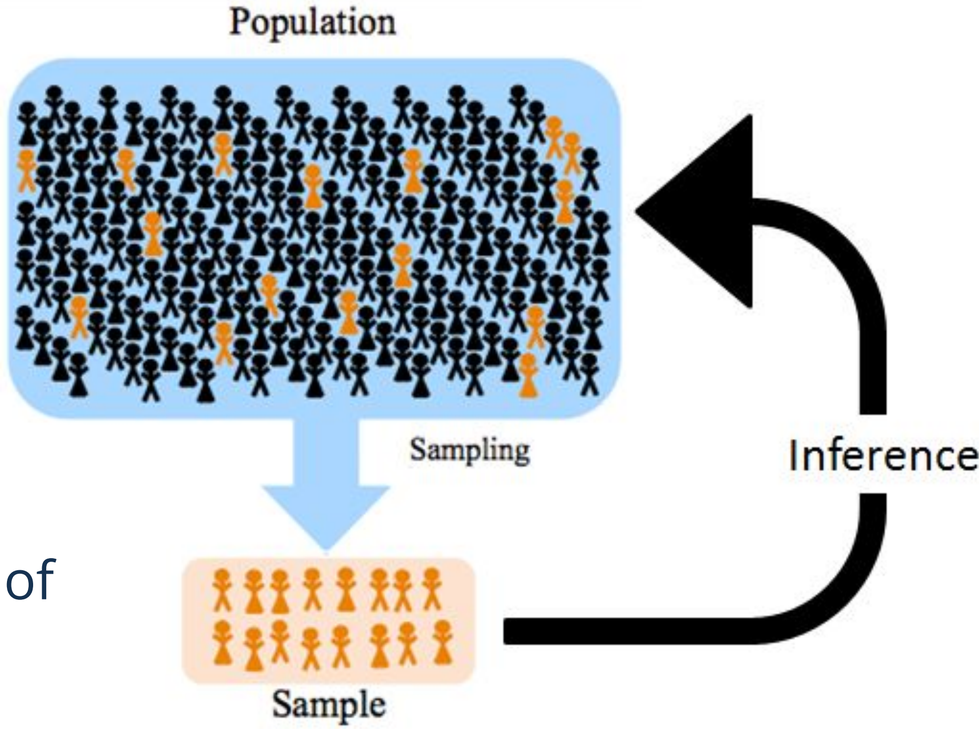
Populations and Samples

Inferential statistics lets us draw conclusions about a population by looking at subsets of the population.

Why: 1. The population is too large
2. The total population is unknown
3. The population is difficult to measure

What: Using a subset of the population that is made via *random and unbiased* sampling.
i.e. every member of the population has an equal chance of being selected.

How: Using statistical hypothesis testing or building confidence intervals



Source: https://statsandr.com/blog/what-is-the-difference-between-population-and-sample/

Variance is in squared units rather the original units of the data

Variance

Variance is the *average squared difference* of the values *from the mean*

Sample variance

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

- M: sample mean
- N-1: corrects for the tendency of a sample to underestimate the population variance

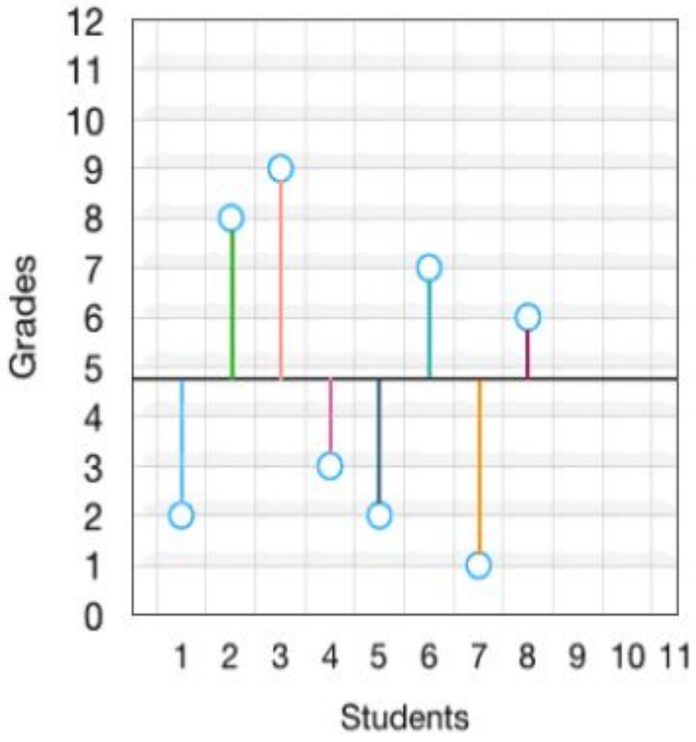
Population variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

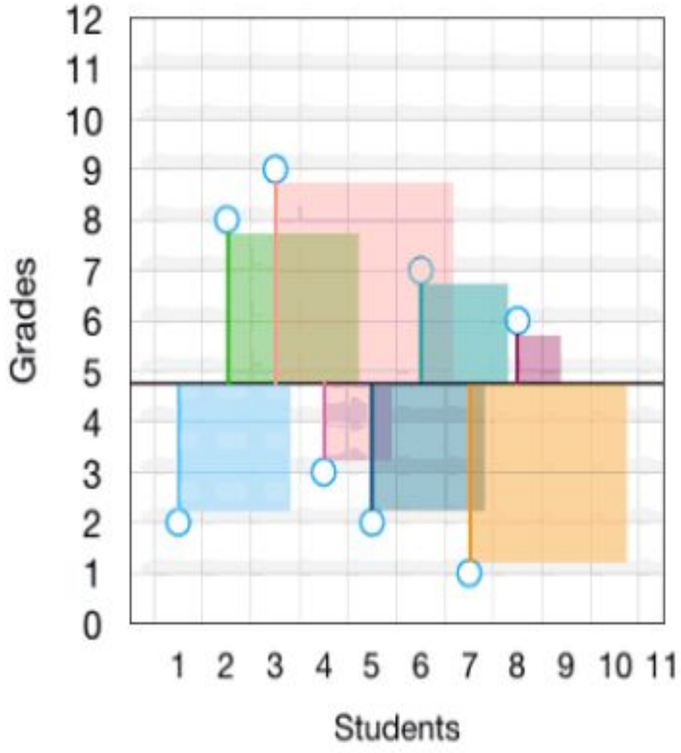
- μ: population mean
- N: number of data points, which should include the entire population

Variance is average squared distance to mean

Distance to mean $(X-\mu)$

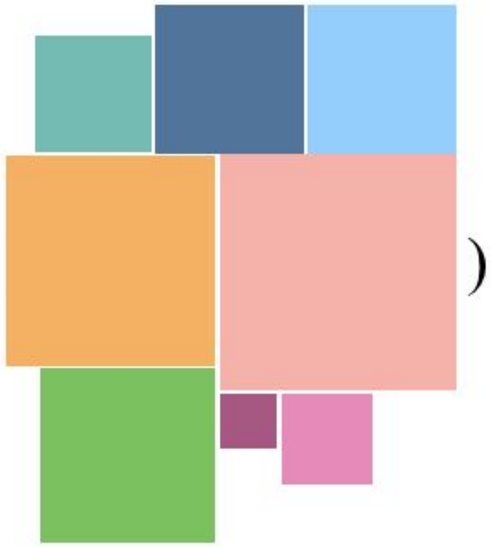


Squared Distance to mean $(X-\mu)^2$



Variance $\frac{\sum(X-\mu)^2}{N}$

Mean ( )



Standard Deviation is in the original units of the data

Standard Deviation

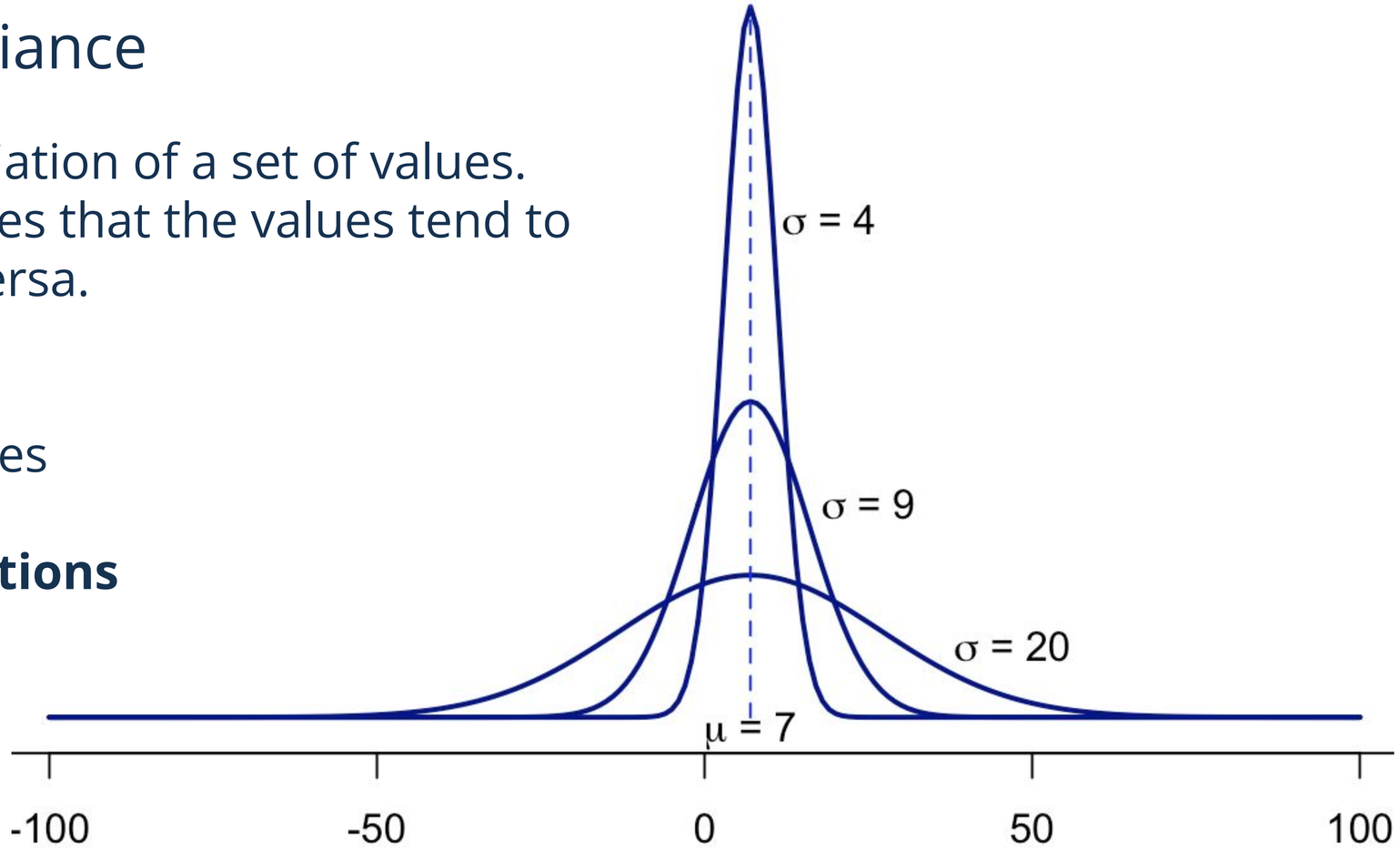
the square root of the variance

A measure of the amount of variation of a set of values.
A low standard deviation indicates that the values tend to be close to the mean and vice versa.

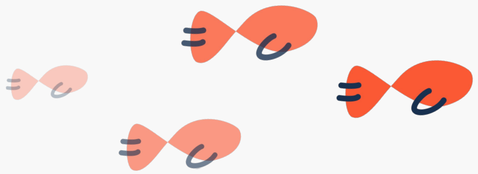
The graph shows

- three normal distribution curves
- with the **same mean**
- but **different standard deviations**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$



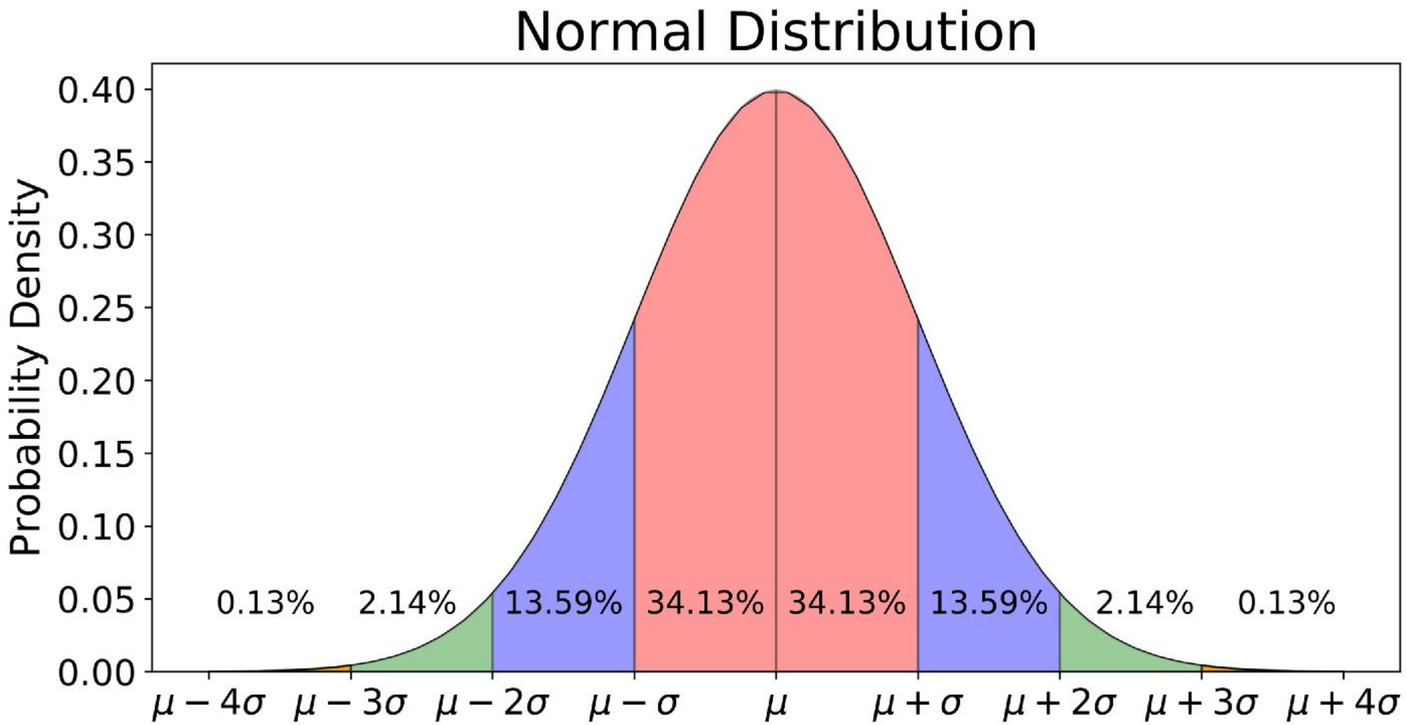
Frequency distribution



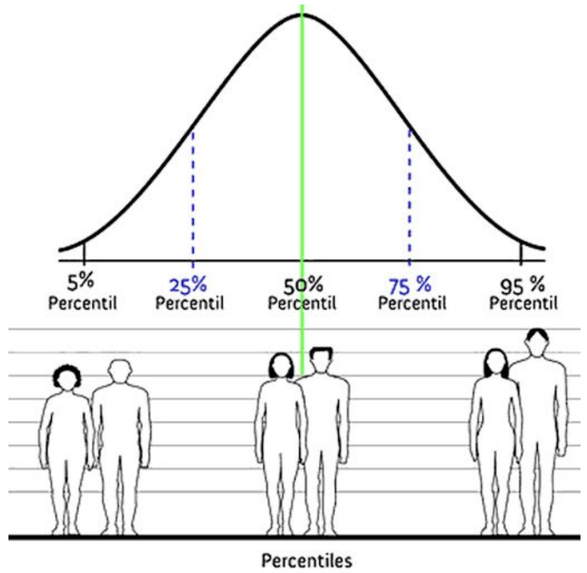
Normal Distribution

Symmetric probability distribution with bell-shaped curve

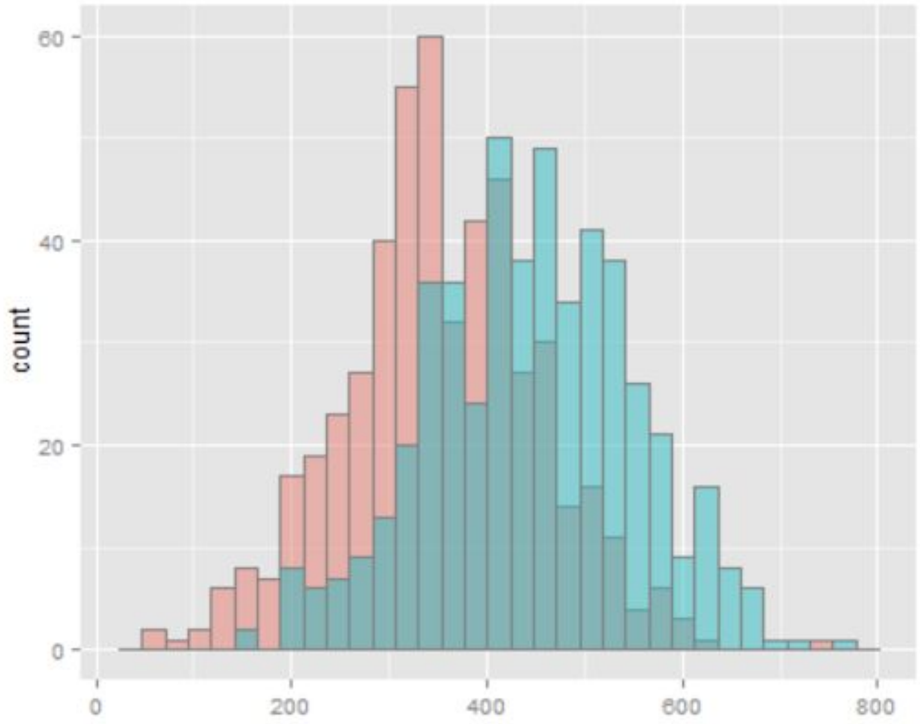
Found in many natural situations

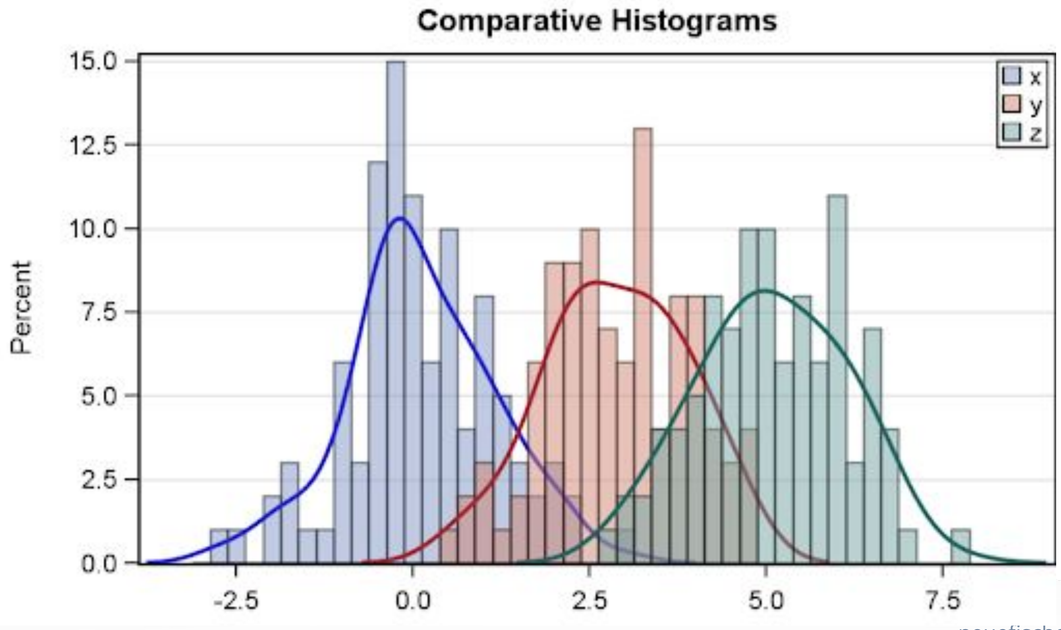


Visualising and comparing data distributions



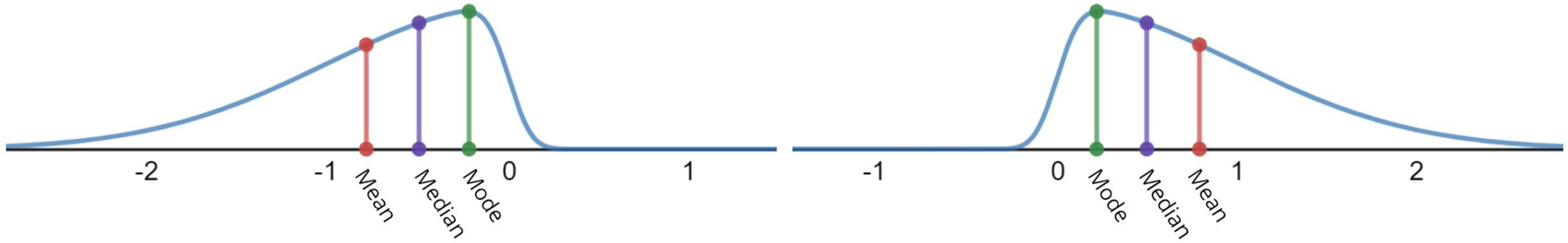
Histograms - plotting the frequency distribution
How many times does each score occur?





Deep Dive into distribution: Skewness

Skewness is a statistical measure that quantifies the asymmetry of a distribution.



Left/negative skewed

- longer tail on the left side
- concentration on the right side
- typically: mean less than median

Right/positive skewed

- longer tail on the right side
- concentration on the left side
- typically: mean greater than median

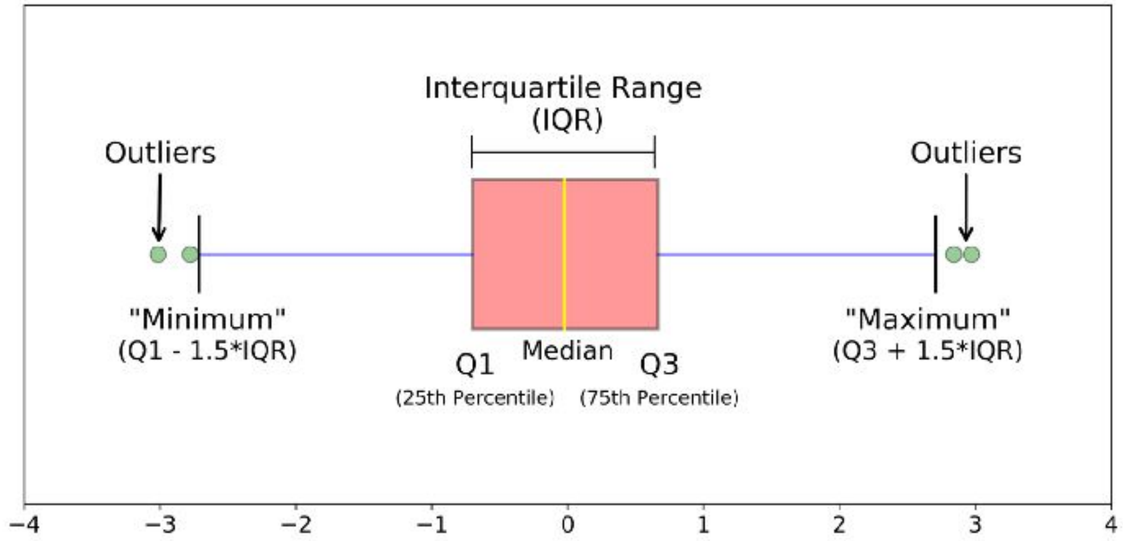
https://www.expii.com/t/normal-distribution-right-and-left-skewed-graphs-5338

Excursion: Boxplots

Boxplots as tool to display distribution of data

Tell you about

- Skewness
- Dispersion
- Outliers



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

References

- Field, A. P. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)*. Los Angeles [i.e. Thousand Oaks, Calif.: SAGE Publications.
- Fahrmeir, Ludwig & Künstler, Rita & Pigeot, Iris & Tutz, Gerhard. (2004). Statistik: Der Weg zur Datenanalyse. 10.1007/3-540-35037-3.
- https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php
- https://statisticsbyjim.com/basics/variability-range-interquartile-variance-standard-deviation/
- http://onlinestatbook.com/2/summarizing_distributions/variability.html
- https://towardsdatascience.com/data-types-in-statistics-347e152e8bee
- https://www.universalclass.com/articles/math/statistics/frequencies.htm
- https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Practice

Let’s test our knowledge by doing a quiz on github

https://github.com/neuefische/descriptive_statistics_practice