Statistics - gathering, describing and analyzing data.
Statistics is a process of collecting information/ data (most of the time it's numbers of some kind) and presenting it in lots of different ways and drawing conclusions from that data.
Statistics - the actual numeric description of sample data.

Data statistics deal with:
- **Population** (**N**) - the collections of all items of interest to our study.
- **Sample** (**n**) - a subset of the population.

Sample collection (Survey):
- *Simple random sampling* (eg. used for Exit Poll) (issues: non-response bias, voluntary response)
- *Stratified sampling* (splits N into *non-overlapping groups - strata*) (eg. male/female; could be used for household survey)
- Cluster sampling (randomly select few clusters from naturally occurring clusters)
- Snowball sampling (current respondent are asked to help recruit people they know from N of interest; eg. people with rare genetic disease)
- *Systematic sampling* (N $\rightarrow n^{th}$ member; drug testing (could be also combined with convenience sampling for eg.)
- *Convenience sampling* (eg. only people interested in Data Science will be in survey)
- Census (survey of an entire population)

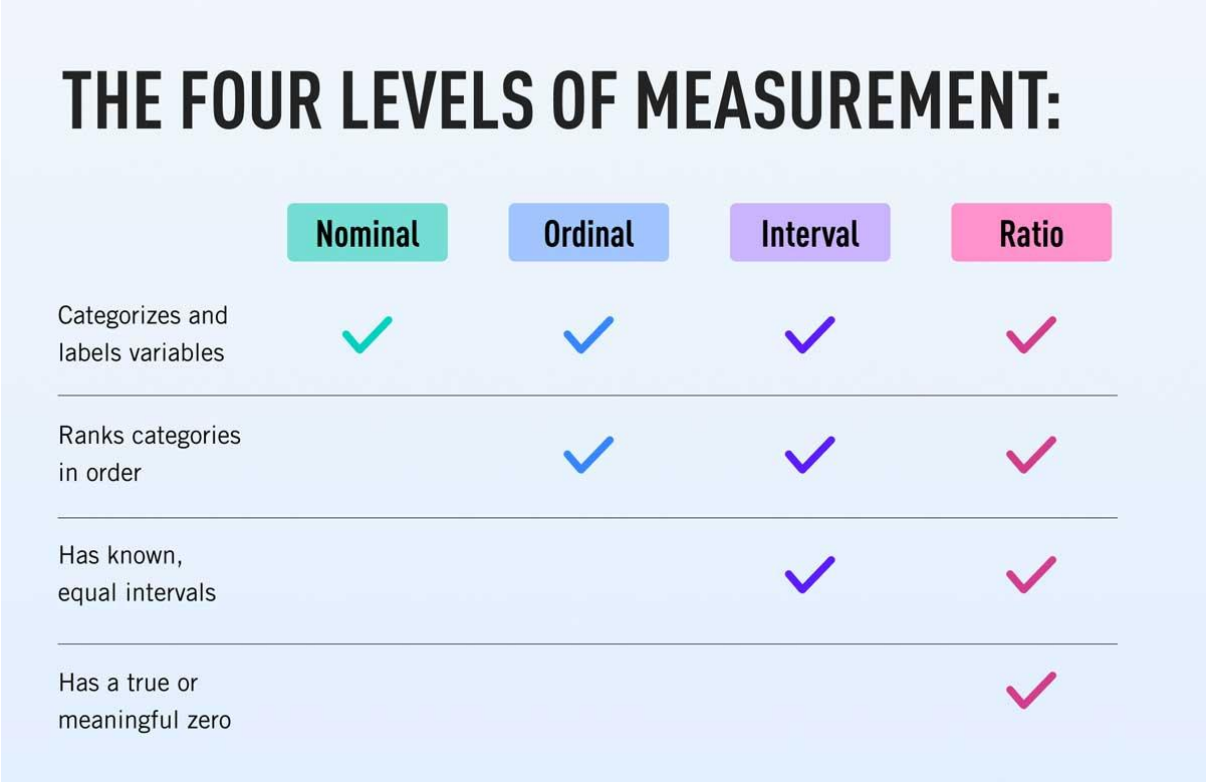
Population is hard to define and observe in real life. Observing samples on the other hand is less time consuming and less costly.
Samples must be **random** and **representative**.

Randomness means collected when each member of the sample is chosen from the population strictly by chance.
Representative means a subset of the population accurately reflects the members of the entire population.

Types of statistics:
- **Descriptive** (include measures of central tendency, measure of dispersion or how spread out the data are) - organizing and summarizing data
- **Inferential** (let test an idea or hypothesis) - technique where in the measured data is used to form the conclusion

Types of Data:
- **Categorical** (eg. yes/no, Audi/BMW/Mercedes)
- **Numerical**:
- ***Discrete*** (eg. number of children)
- ***Continuous*** (eg. hight, area, time)

Measurement levels:
- **Qualitative/**:
- ***Nominal*** (eg. winter/spring/summer/autumn)
- ***Ordinal*** (- → +, order matter but not value, eg. rank in competition)
- **Quantitative** (measured numerically):
- ***Interval*** (eg. t° in °C or °F) continuous variable,does not have true 0
- ***Ratio*** (have true 0, eg. 0°K = -273.15°C = -459.67°F)



Frequency distribution - table that divides data into groups (classes) and shows how many data values occur in each group. (eg. 10 lillies of 50 flowers)
Relative frequency is the percentage of the data set that falls in a class. (eg. 20% = 10 lillies of 50 flowers)

$Relative\ frequency\ = \frac{class\ frequence}{\sum frequencies}$, $\sum frequencies$ - sample size (n)

Categorical variables visualization technique:
- Frequency distribution tables
- Bar charts (for discrete variables)
- Histogram (for continuous values)
- Pie charts
- Pareto diagrams
- Stem and Leaf

Measures of central tendency

Mean (**μ**) - average value of the data.

Sample mean ($\bar{x}$) - mean of sample values collected.

Sample mean

$$\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} \equiv \frac{\sum_{i=1}^{n} x_i}{n}$$

Population mean

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

Weighted mean - when each value is not equally important.

Median - the middle number in an ordered dataset.

$\frac{n+1}{2}$-position (median)

Mode - most common value (if each value presents only once, there is no mode).

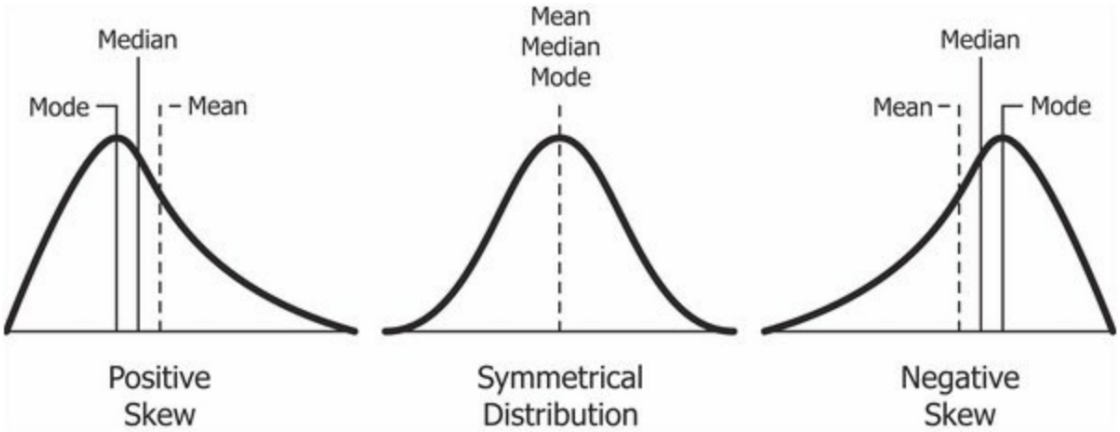
Bimodal data is an example of 'multimodal' data which has many values that are similarly common. Usually multimodal data results from two or more underlying groups all being measured together.

5-number summary of Data:

Min, Q1,Q2(median),Q3, max

Interquartile range (IQR) - middle 50% of the data (Q1↔Q3)

Skewness



Variance - squared deviation of a random variable from its population mean or sample mean (measures the dispersion of dataset points around the mean).

For each data point: deviation from mean = $x_i - \bar{x}$

Sample formula	Population formula
$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

Standard deviation measures the amount of variation, or dispersion, in a set of values.

Sample formula	Population formula
$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

Coefficient of Variation measures the ratio between the standard deviation and the mean (relative standard deviation).

Sample formula	Population formula
$\hat{c}_x = \frac{s}{x}$	$c_v = \frac{\sigma}{\mu}$

Covariance

Sample formula	Population formula
$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$	$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$

Correlation coefficient (-1↔1)

$$\frac{Covariance(x,y)}{Stdev(x)*Stdev(y)}$$

Chebyshev's Theorem - the % of the data that lies within 'K' standard deviation is at least.

$$1 - \frac{1}{K^2}, \; for \; K > 1$$

Eg. for K=2, $1 - \frac{1}{2^2} = \frac{3}{4}$ = 75% of the data lie within 2σ of the mean

---

Distribution, Standardization, Normalization
Normal (Gaussian) Distribution

Imperial formula: within 1s(σ) 68% of data, 2s(σ) 95%, 3s(σ) 99.7%

68 - 95- 99.7%

StandardScaler standardizes a feature by subtracting the mean and dividing by standard deviation. Results in a distribution with a standard deviation equal to 1.

z-Score how distributed data points are around the mean with respect to the stdev

$$\frac{x - \bar{x}}{s}$$

Standard normal distribution derived from Gaussian distribution by applying z-score formula to each point of a data. The process is called **Standardization**.

Lower fence:	**Upper fence**:
Q1-1.5(IQR)	Q3+1.5(IQR)

Eg. for z-score application:
mean=100
stdev=15
% ↔ x1=90 & x2=120?
z- score
(90-100)/15=-0.7

(120-100)/15= 1.3
difference
90.82-25.46=65.36%

MinMaxScaler (in sklearn default range is 0 to 1) subtract the minimum value in the feature and divide by the range. The range is the difference between the original maximum and original minimal. MinMaxScaler preserves the original distribution shape and doesn't reduce the importance of outliers.

Application: CNN Image classification or object detection. Each pixel lies in range 0-255, applying MinMaxScaler the range could be specified 0-1. It's done by dividing each pixel by 255.

RobustScaler transforms the feature vector by subtracting the median and dividing by IQR(75%values -25% values). Recommended to use to reduce the effect of outliers.

Normalizer works on the rows not the columns!It can use l2 or l1 normalization.

Sources:

1. 365 careers, Udemy course
https://www.udemy.com/course/the-data-science-course-complete-data-science-bootcamp/
2. Marth and Science
https://www.youtube.com/channel/UCYgL81Ic7DOLNhneI1_J6Vg
3. Wikipedia
https://en.wikipedia.org/wiki/Variance
4. Krish Naik
https://www.youtube.com/watch?v=11unm2hmvOQ
5. What is Ratio Data, Will Hillier
https://careerfoundry.com/en/blog/data-analytics/what-is-ratio-data/
6. Scale, Standardize, or Normalize with Scikit-Learn, Jeff Hale
https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02