# Data Analysis

Week 1

# Research Process

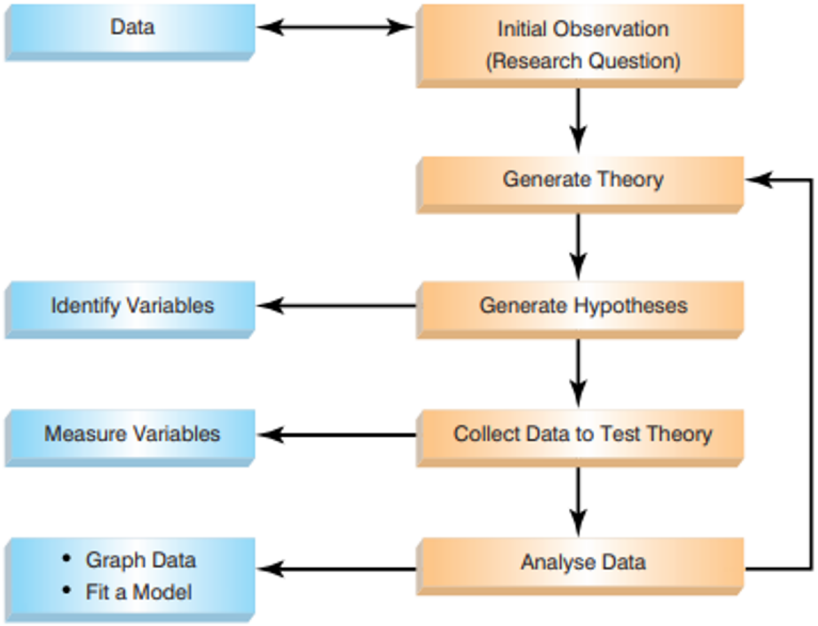



Table of content

# Variables

Variable: a characteristic that can change or take on different values

- Independent vs. Dependent
- Predictor vs. Outcome
- Discrete vs. Continuous (level of measurement)

Discrete variable: A variable consists of indivisible, separate categories. No values can exist between two neighbouring categories.

Continuous variable: A variable that can be divided into smaller units without limits.

Table of content

# Variables

**Categorical** (entities are divided into distinct categories):

- Nominal variable: There are more than two categories (e.g. whether someone is an omnivore, vegetarian, vegan, or fruitarian).
- Ordinal variable: The same as a nominal variable but the categories have a logical order (e.g. whether people got a fail, a pass, a merit or a distinction in their exam).

**Continuous** (entities get a distinct score):

- Interval variable: Equal intervals on the variable represent equal differences in the property being measured (e.g. the difference between 6 and 8 is equivalent to the difference between 13 and 15).
- Ratio variable: The same as an interval variable, but the ratios of scores on the scale must also make sense (e.g. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8)

Table of content

# Variables

## Types of data on the basis of measurement

| Scale | True Zero | Equal Intervals | Order | Category | Example |
|---|---|---|---|---|---|
| Nominal | No | No | No | Yes | Marital Status, Sex, Gender, Ethnicity |
| Ordinal | No | No | Yes | Yes | Student Letter Grade, NFL Team Rankings |
| Interval | No | Yes | Yes | Yes | Temperature in Fahrenheit, SAT Scores, IQ, Year |
| Ratio | Yes | Yes | Yes | Yes | Age, Height, Weight |

Table of content

# Validity and Reliability

There will often be a discrepancy between the numbers we use to represent the thing we're measuring and the actual value of the thing we're measuring (i.e. the value we would get if we could measure it directly). This discrepancy is known as measurement error.

**Validity:** whether an instrument actually measures what it sets out to measure.

**Reliability:** whether an instrument can be interpreted consistently across different situations

Table of content

# Frequency Distributions

Organized display of the number of individuals in each category on the scale of measurement.

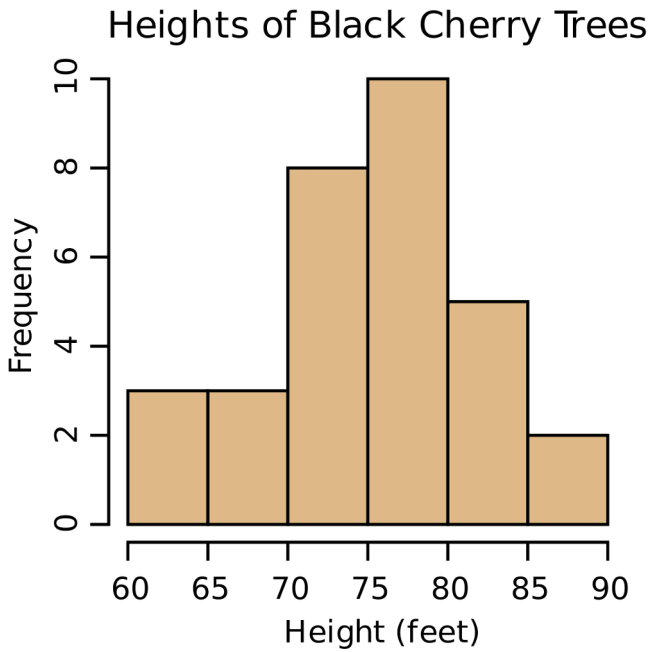
Frequency distribution graphs are different for variables.

- Histogram and Polygons are used for interval and ratio scale data

- Bar graphs are used for nominal or ordinal scale data.

# Frequency Distributions

**Histogram:** A graph showing bar above each score or interval so that the height of the bar corresponds to the frequency and width extends to the real limits.

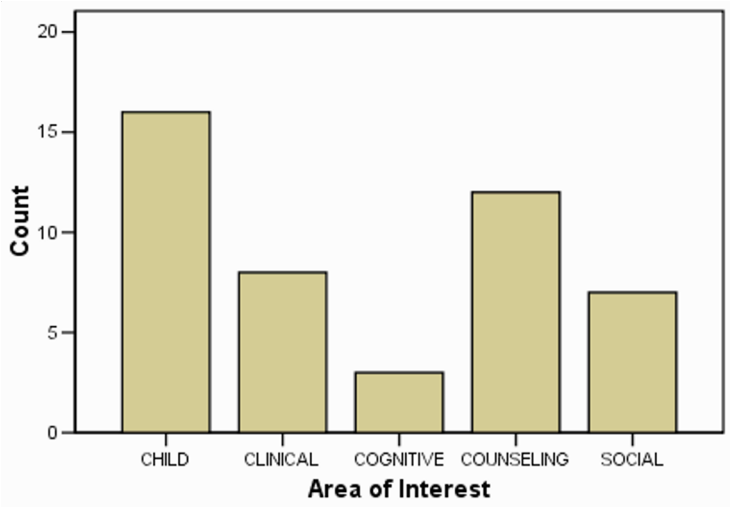


# Frequency Distributions


**Polygon:** A graph consisting of a line that connects a series of dots. A dot is placed above each score or interval so that the height of the dot correponds to the frequency

# Frequency Distributions

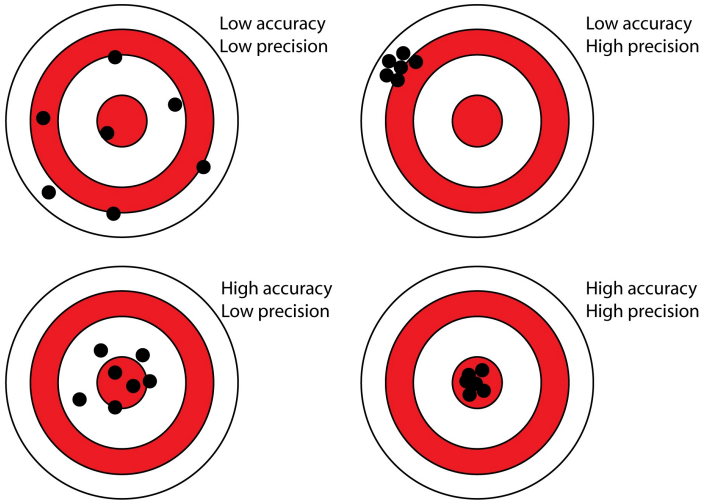
**Bar Graph:** A graph showing a bar above each category so that the height of the bar corresponds to the frequency. A space is left between adjacent bars.



# Normal Distribution

- The average of the distance between bull's eye and the place where the target was hit is **ACCURACY**
- The amount of variation in these distances is **PRECISION**

- NORMAL DISTRIBUTION:
  - MEAN is accuracy
  - STANDARD DEVIATION is precision



# Normal Distribution

- 68.27% chance that the arrow will land anywhere between +/- 1SD from the average distance from the bull's eye
- 95.45% chance that the arrow will land anywhere between +/- 2SD from the average distance from the bull's eye
- 99.73% chance that the arrow will land anywhere between +/- 3SD from the average distance from the bull's eye

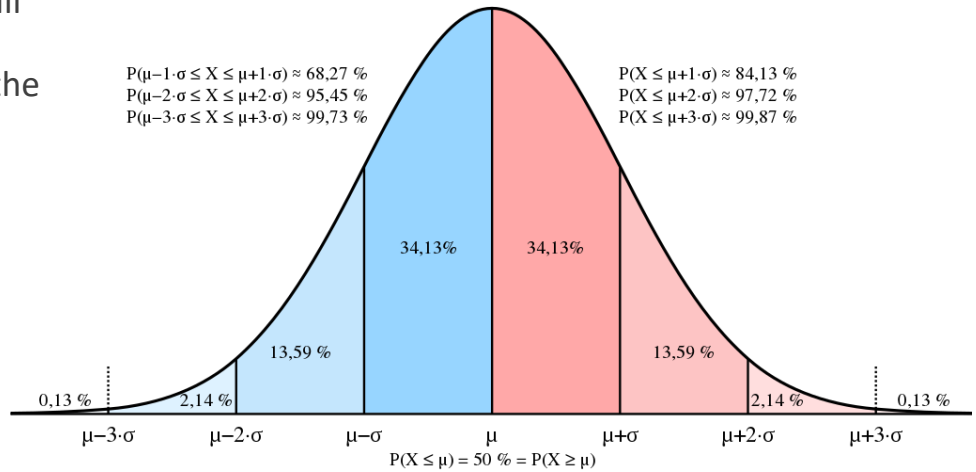


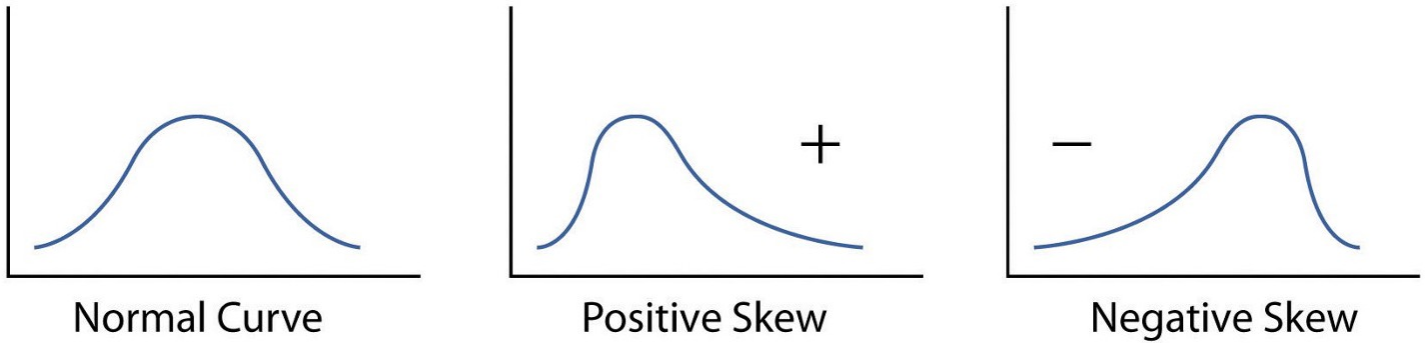

Table of content

# Testing for Normality

The population or the two populations from which samples are selected must be normal.

- **Skewness** – measure of asymmetry of the probability distribution
- **Kurtosis** – tells you the height and sharpness of the central peak, relative to that of a standard bell curve
- Tests:
  - Kolmogorov- Smirnov
  - **Shapiro Wilk**
- Data is normally distributed if the significance in these tests is > 0.05

# Symmetrical vs. Skewed Distribution



# Central Tendencies

Central tendency is a statistical measure that identifies a single score (usually a central value) as representative of an entire population.

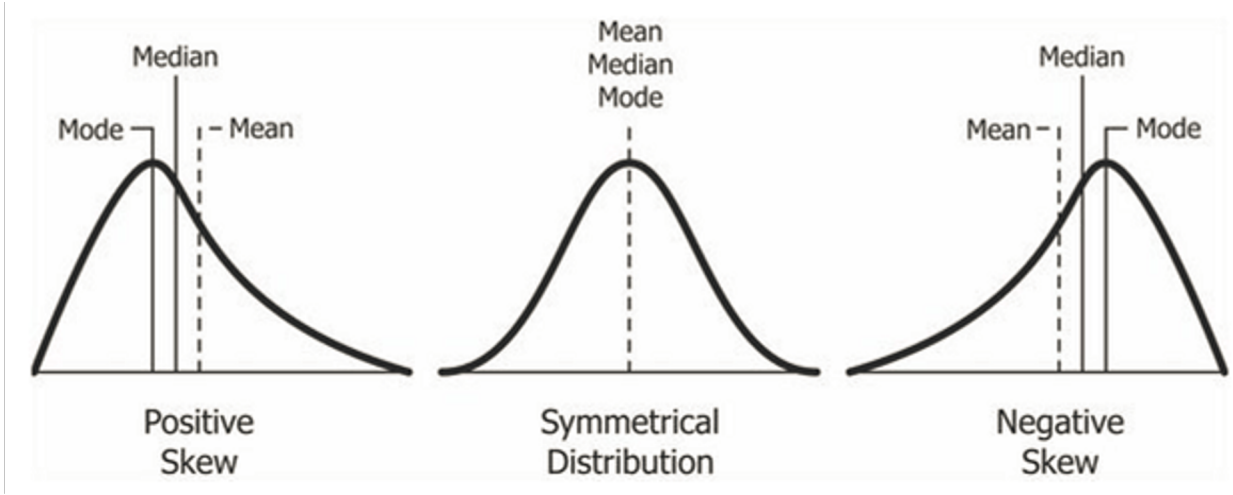
Mean: Arithmetic average

Median: The core that divides a distribution exactly in half

Mode: The score that has the greatest frequency

Q: 9 people with salary of $5 000 and one person with salary of $50 000. What is the average salary?

# Mean vs. Median vs. Mode



# Variability

Variability provides a quantatative measure of the degree to which scores in a distribution are spread out or clustered together.

- Range
- Variance
- Standard Deviation

Table of content

# Range

Distance between the largest and the smallest score in the distribution.

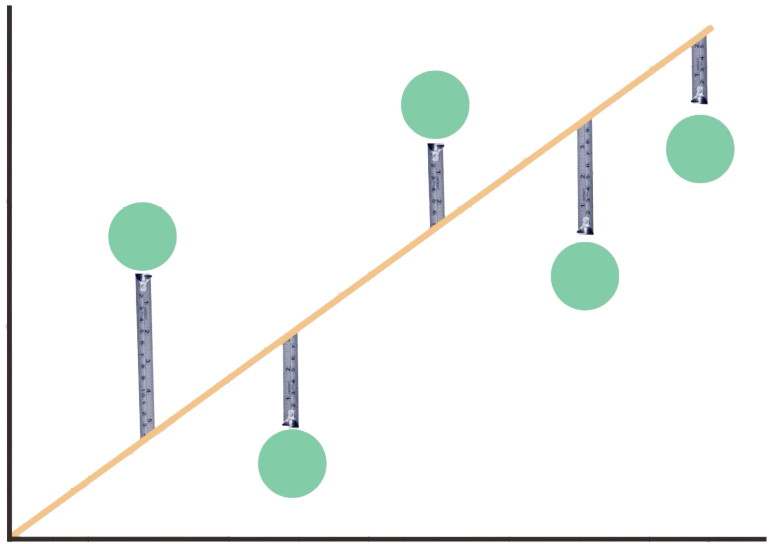
Note that range is completely determined by two extreme scores and does not consider all scores in a distribution.

# Standard Deviation and Variance

The most commonly used and most important measures of variability. They use the mean as a reference point and measure variability by considering the distance between each score and the mean.

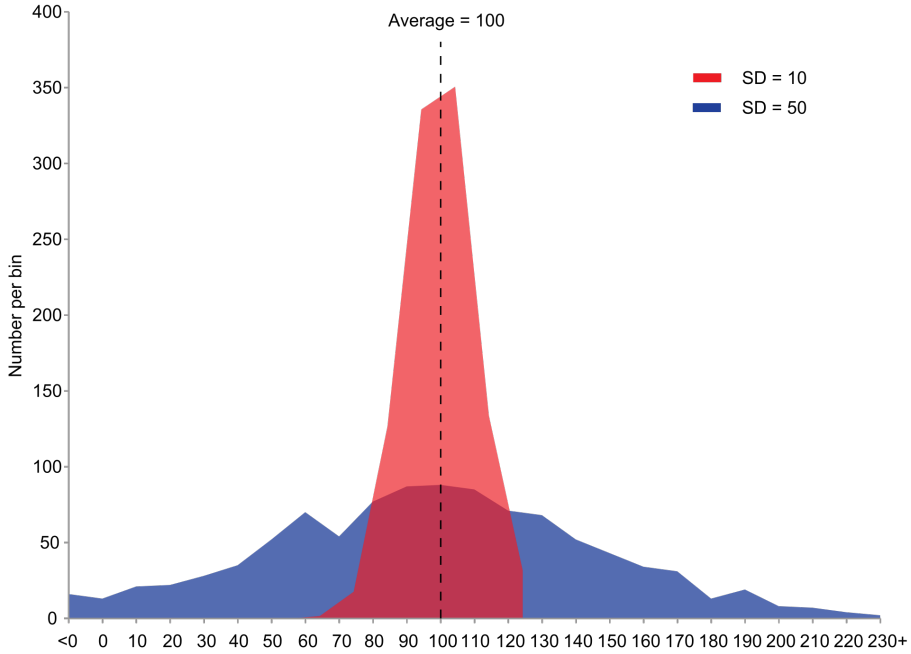
The SD approximates the average distance from the mean.

Deviation is the distance from the mean.



# Standard Deviation and Variance

- Low STANDARD DEVIATION indicates that the values tend to be close to the mean (expected value)
- High STANDARD DEVIATION indicates that the values tend to be far from the mean (expected value) or, in other words, the values tend to be dispersed from the mean



# Probability

- Normal distribution can help us make meaningful predictions about the future:
  - How likely is it that the arrow will land +/- 1 SD from the average distance from the bull's eye
  - How likely is it that the arrow will land less or more than X cm from the bull's eye?

- Z – score or the standard score is the number of standard deviations by which the value of a raw score is above or below the mean value of what is being measured or observed

$$Z = \frac{X - \mu}{\sigma}$$

# Hypothesis Testing

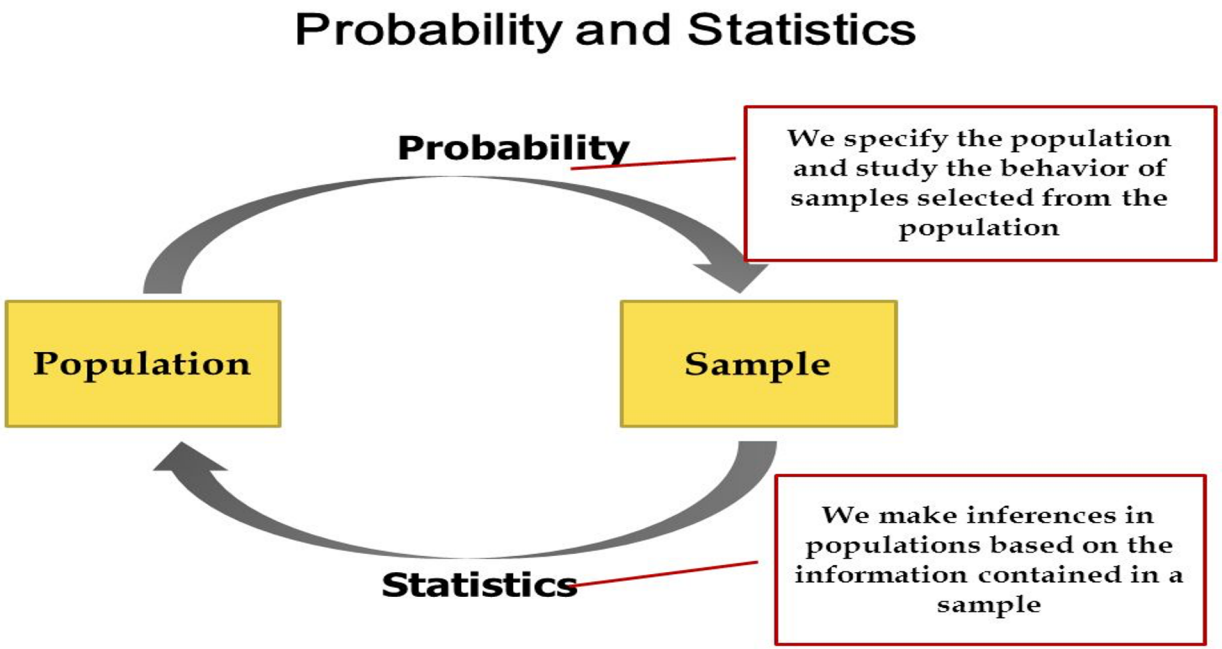



Table of content

# Hypothesis Testing

Statistical procedure that uses sample data to evaluate hypothesis about a population parameter.

In statistics, hypothesis is a claim or statement about a property of a population.

Table of content

# Core logic of hypothesis testing

1. We state a hypothesis about a population. It is usually based on existing research literature, on observation, or through reasoning.

2. We collect data through a random sample from the population.

3. We compare the sample data with the prediction that was made from the hypothesis. If consistent, the hypothesis is reasonable; if not, the hypothesis is wrong.

# Basic Steps in Hypothesis Testing

1. State the hypothesis
2. Set the criteria
3. Collect data and compute the sample statistics
4. Make a decision

# State the Hypothesis

The null hypothesis states that exposure to cigarettes has no effect on birth weight

$$H_0: \mu = 2.0$$

Alternative hypothesis states that exposure to cigarettes has effect on birth weight.

$$H_a: \mu \neq 2.0$$

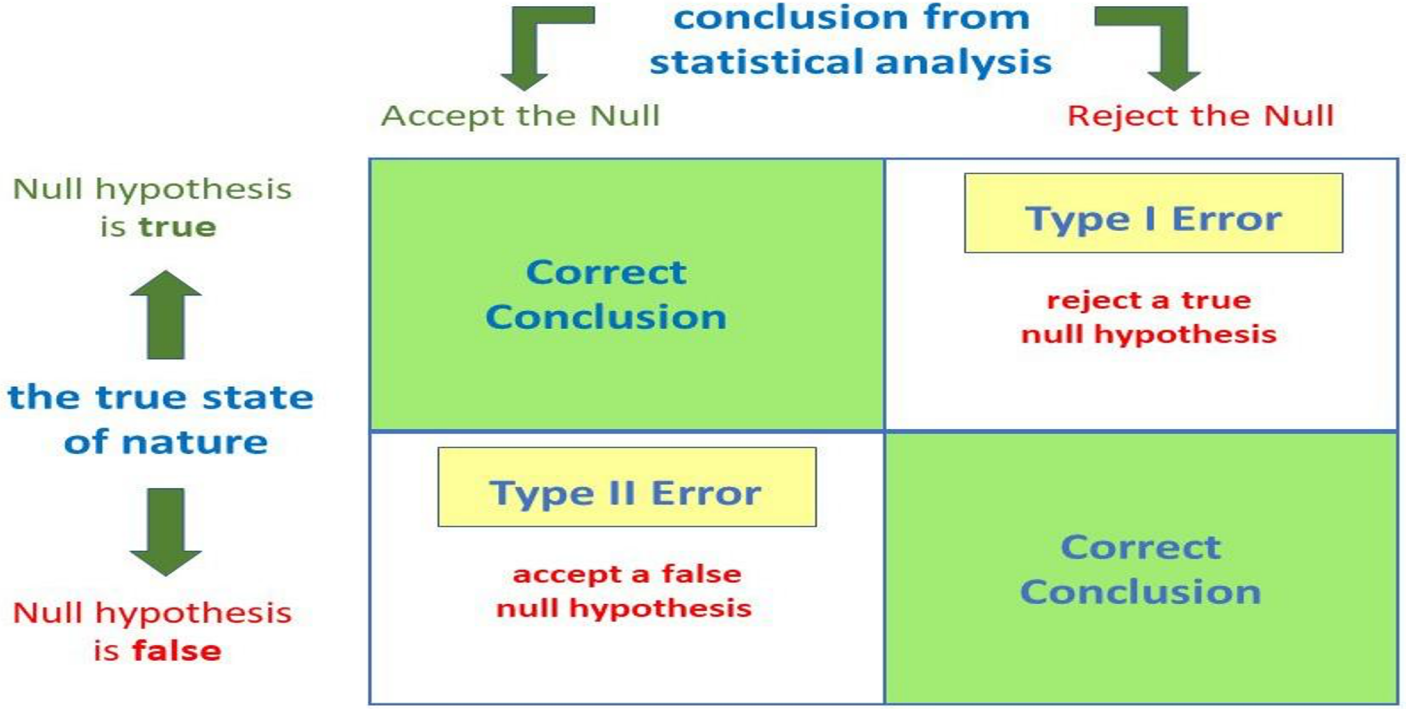
# Set the Criteria – type I error

- The alpha level – usually denoted by the Greek letter alpha ( $\alpha$ ) or level of significance is a probability of rejecting the null hypothesis given that it is true.
  - Courtroom example of type I error: convicting an innocent defendant.
- Commonly used alpha levels are: .05 (5%), .01 (1%), .001 (0.1%).
- A significance level of .05 implies that it is acceptable to have a 5% probability of incorrectly rejecting the true null hypothesis.

# Set the Criteria – type II error

- Usually denoted by the Greek letter beta ( β )
- The probability that the test correctly rejects the null hypothesis ($H_0$) when a specific alternative hypothesis ($H_1$) is true.
  - Courtroom example of type II error: acquitting a criminal.
- The power of a test = 1 – β. In other words, as the probability for a false negative (type II error) decreases, the power of the test increases.

# Type I vs Type II Error



# Type I vs type II Error

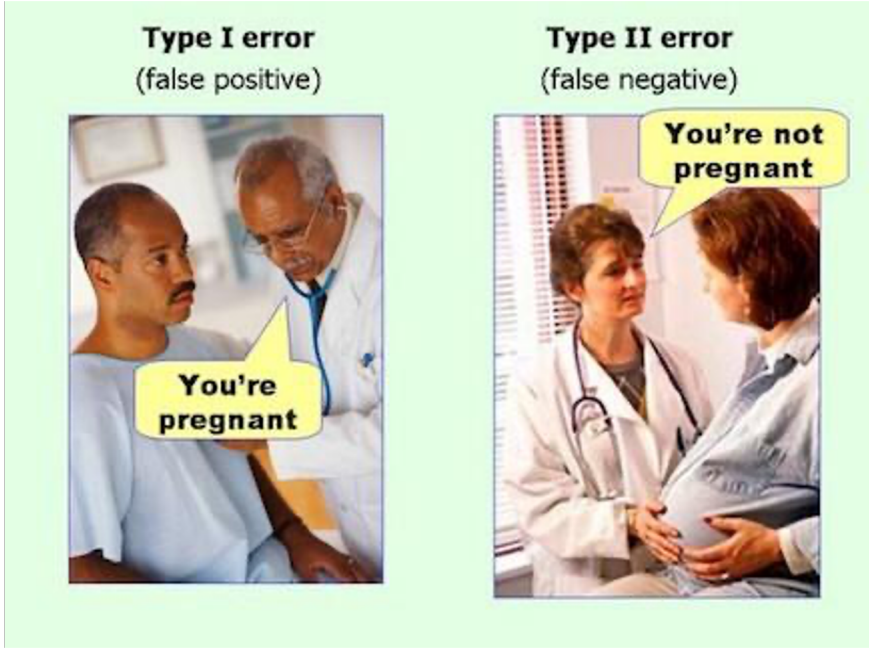



Table of content

# Hypothesis Testing

## One-tailed vs. Two-tailed test

One-tailed: In directional hypothesis, the statistical hypothesis specify either decrease or increase in the population mean score.

Two-tailed: Non directional test.

Table of content