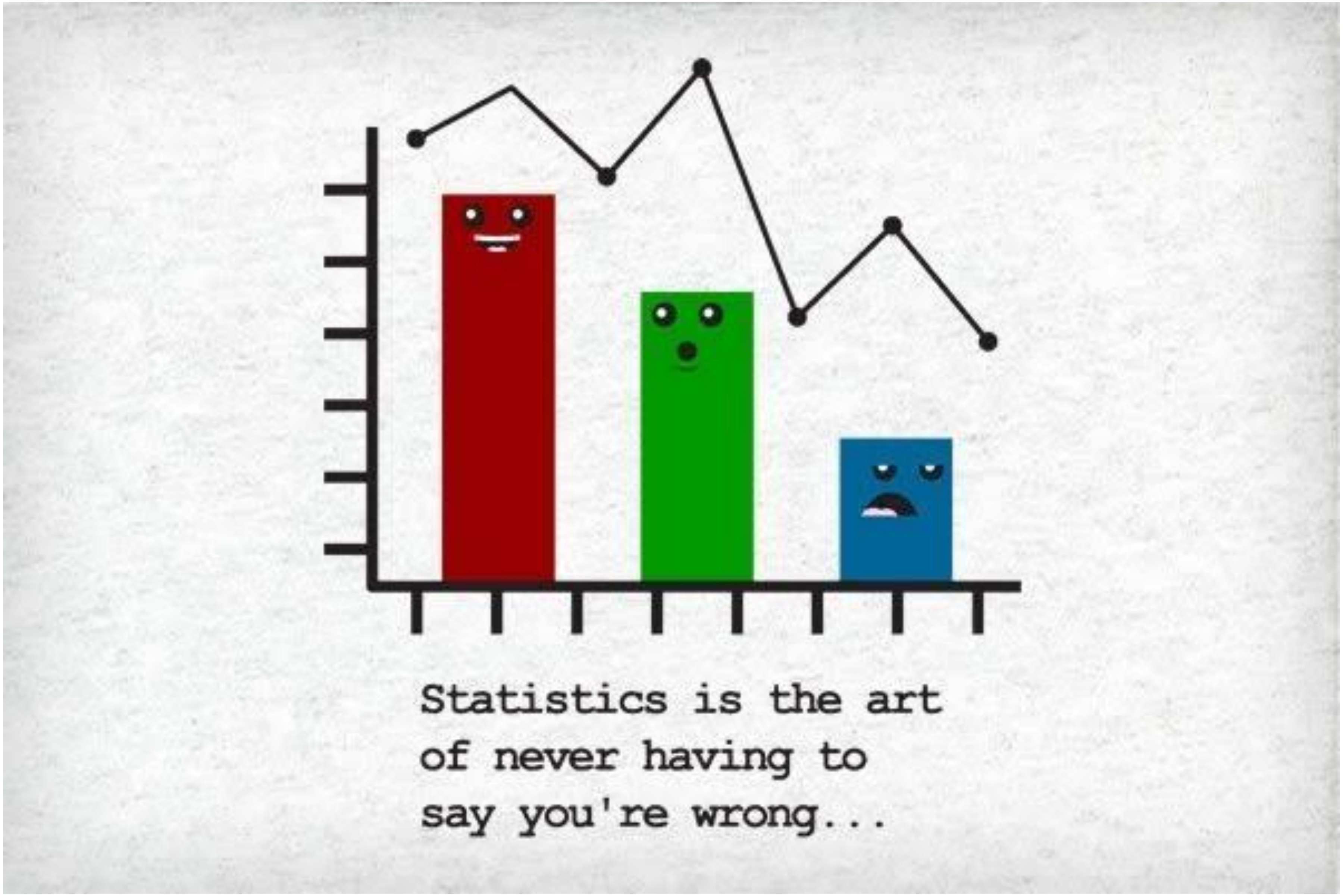
Basics of Biostatistics

Prof. Dr. Mohammed S. El-Awady

Department of Pharmacology & Toxicology,
Faculty of Pharmacy,
Mansoura University

➢ *Statistics*



➢ Statistics

"I can prove it or disprove it! What do you want me to do?"

➢*Descriptive Statistics*

I- Organize Data

Tables / Graphs (Diagrams)

II- Summarize Data

A- Central Tendency:

B- Variation (Dispersion):

➢*Measurement of Central Tendency*

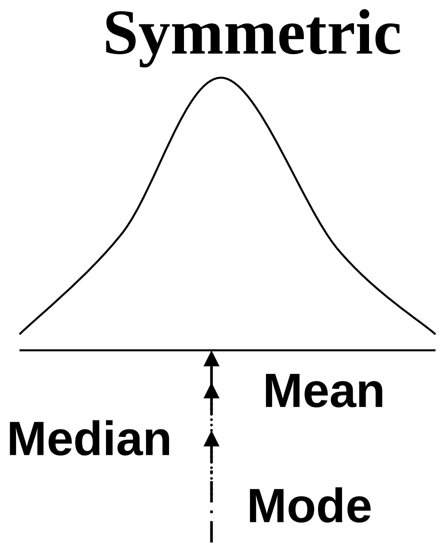
It is the measurement of groups' middle values

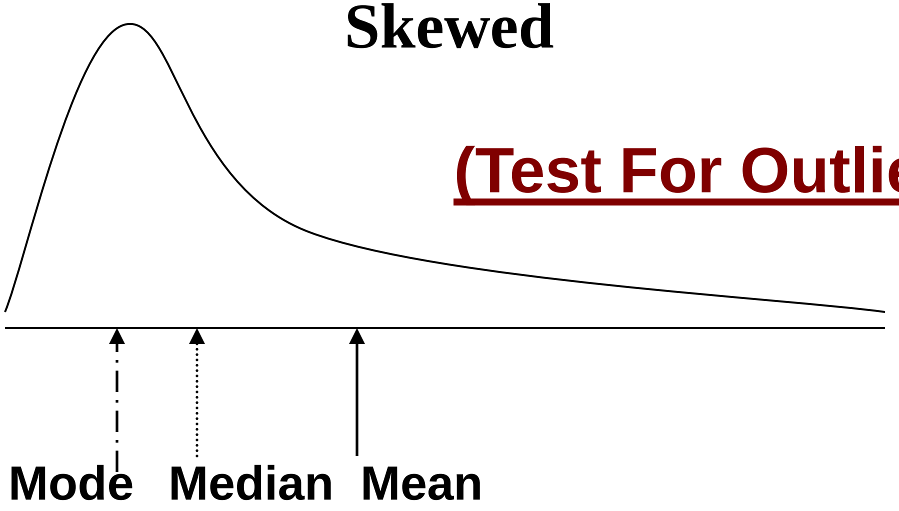
It involves:

1- The Mean (Average) **(very sensitive to outliers).**

2- The Median (central value)

3- The Mode (most common value):





(Test For Outliers)

➢*Measurement of Variation (Dispersion)*

It is the summary of differences within groups:

1- Range
2- Variance
3- Standard Deviation (SD)
4 - Standard Error of the Mean (SEM)
5- Coefficient of Variation
6- Confidence Intervals/limits

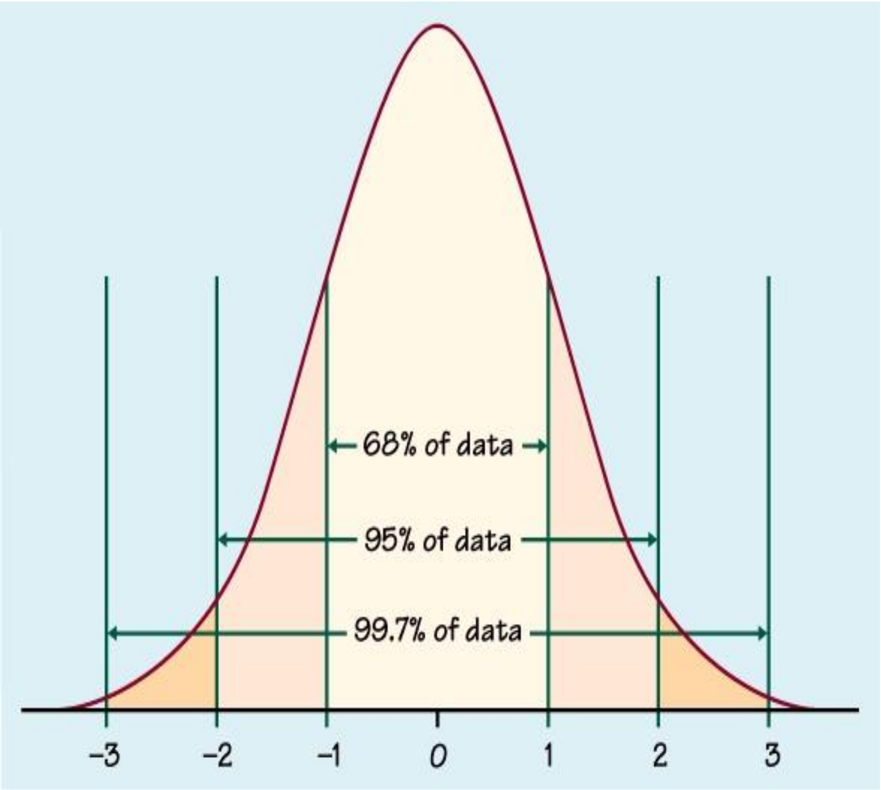
$$S.E.M. = \frac{S.D.}{\sqrt{n}}$$

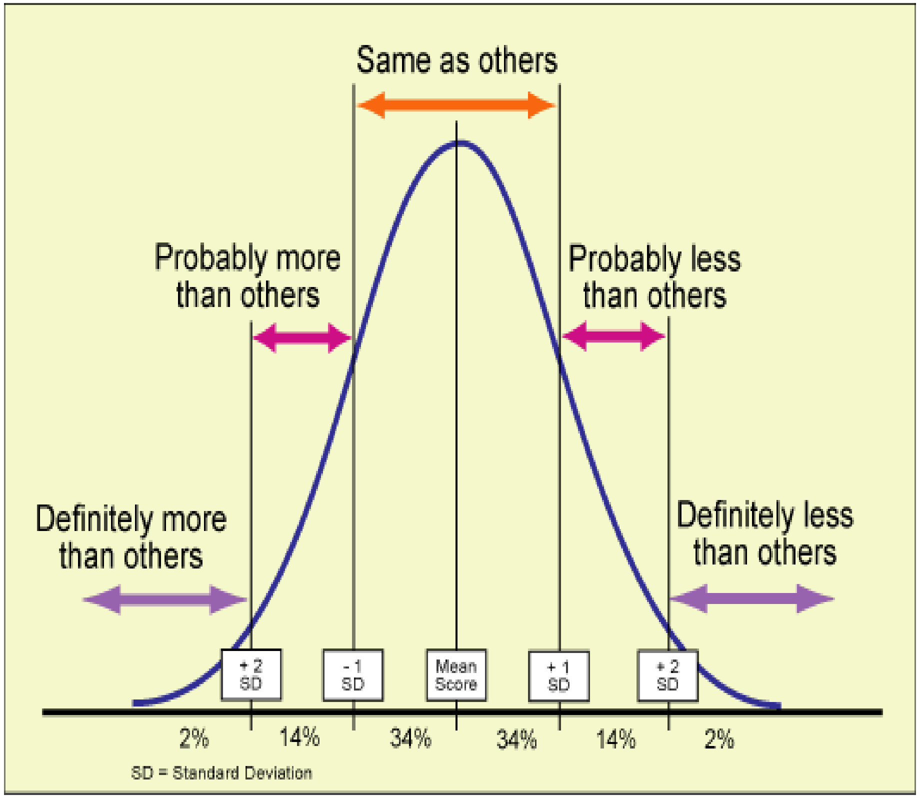
$$S.D. = \sqrt{Variance} = \sqrt{\frac{\sum d^2}{n-1}}$$

➢ *Standard Deviation (S or S.D.)*

It represents the average deviation of the observations from the mean

Normal Distribution





Percent of items included between certain values of S.D.

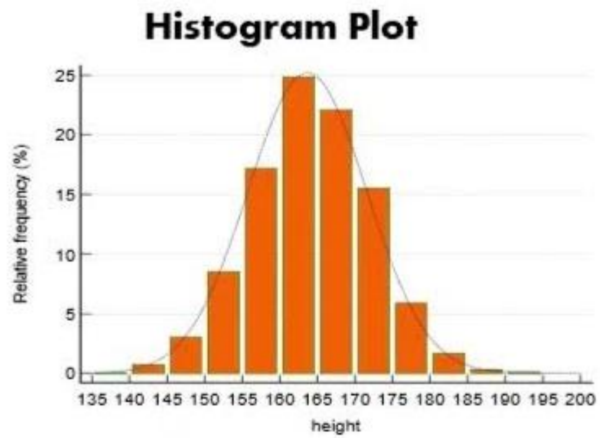
➢ *Methods For Normality Testing*

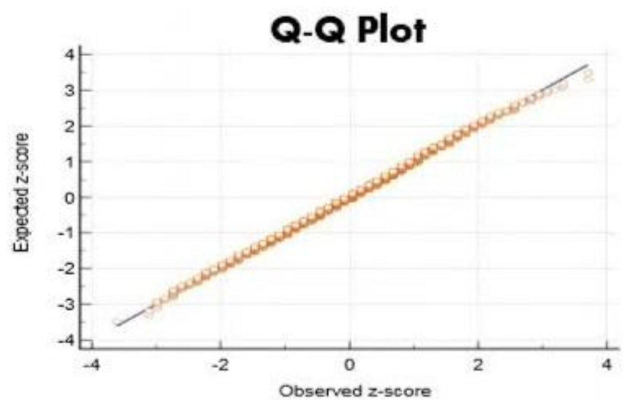
Graphical Methods:

- **Histogram Plot**
- **Quantile-Quantile Plot [Q-Q Plot]**

Statistical Tests:

- **Shapiro-Wilk Test**
- **D'Agostino K^2 Test**
- **Anderson-Darling Test**





Disadvantage of the analytical tests for normal distribution

Unfortunately, the analytical method has a major drawback, which is the calculated p-value is affected by the size of the sample (increasing sample size decreases your p value).

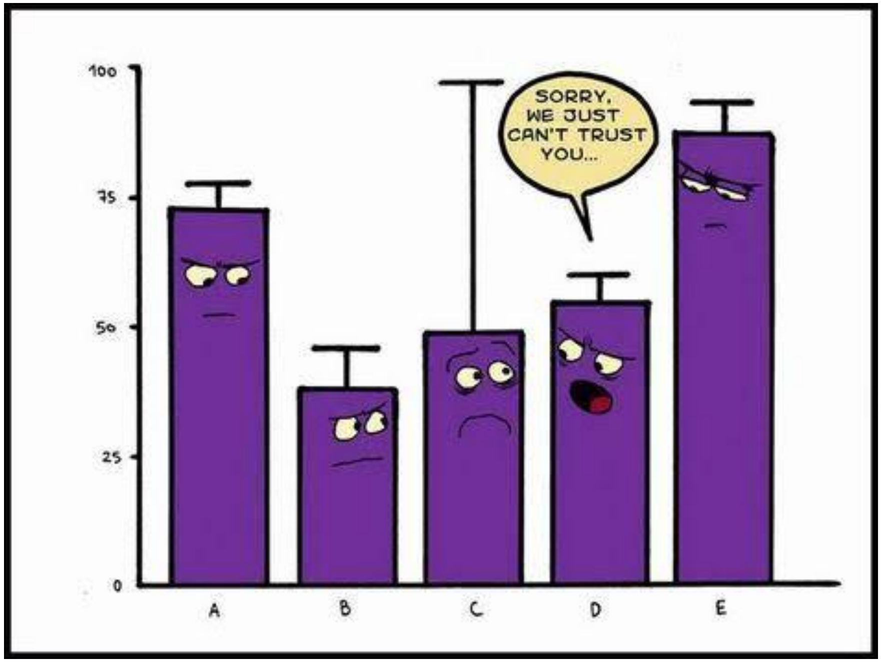
➢ *Standard Error of the Mean (S.E.M.)*

The standard error of the mean (S.E.M.) is an estimate of how close your sample mean to the population mean.

Standard error should decrease with larger sample sizes, as the estimate of the population mean improves.

Standard deviation will be unaffected by sample size

The Mean is Valid if it is $\geq$ 2.5 X S.D. OR $\geq$ 10 X S.E.M.



➢ *Confidence Intervals or limits (C.I.)*

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter. Confidence intervals are typically stated at the 95% confidence level.

$$C.I. = \overline{X} \pm t \times S.E.$$

C.I. = Confidence Interval

$\bar{X}$ = Mean.

S.E. = standard error.

t = a value obtained from a distribution table using P (Probability = 0.05) and df (= degree of freedom = n-1).

➢*Inferential Statistics (Hypothesis testing)*

➢ **These are methods for using sample data to make general conclusions (inferences) about populations.**

➢**Steps:**

- ➢ **1- Determine your data type**
- ➢ **2- Choose the suitable statistical test**

➢Types of Data/Variables

	Data type	Definition	Examples
Discrete	**Category data - without order** (also called **Nominal**)	Data has a name only	Street, Road, Way Male, Female
	Category data - ordered (also called **Ordinal**)	Data has order, but does not have a numerical scale	Very happy, Happy, Unhappy, Very unhappy
	Whole number data	Data can have any whole number value	Day 1, Day 2, Day 3, … (simple time series data) Number of books (0, 1, 2, 3, …)
Continuous	**Measurement data**	Data can take any numerical value	Length of a pencil. It can be 8 cm, 9.1 cm, 9.48m. Time in seconds.

➢*Types of Data/Variables*

VARIABLES

QUANTITATIVE

RATIO
Pulse rate
Height

INTERVAL
36^{o}-38^{o}C

QUALITATIVE

ORDINAL
Social class

NOMINAL
Gender
Ethnicity

➢Levels of Measurement

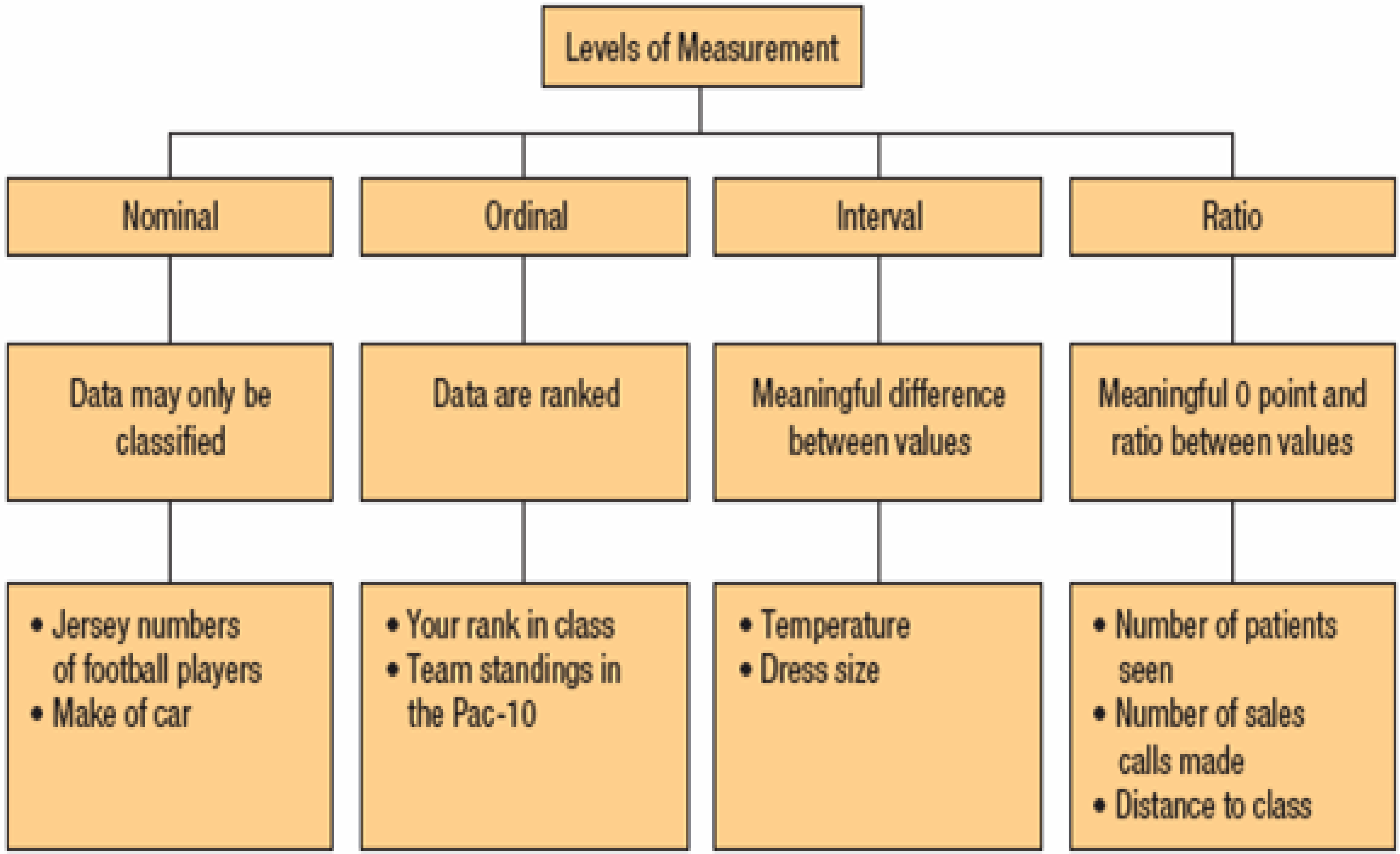


CHART 1–3 Summary of the Characteristics for Levels of Measurement

➢*Examples of Nominal Data*

I- Nominal (Description with no ranking):

- Gender (male, female)
- Eye color (blue, brown, green, hazel)
- Surgical outcome (dead, alive)
- Blood type (A, B, AB, O)
- Prevalence (Infected or non-infected)
- Convulsion or no convulsion
- Cured or not-cured
- Discharged from hospital or not

Note: **When only two possible categories exist, the variable is sometimes called dichotomous, binary, or binomial.**

➢*Examples of Ordinal Data*

II- Ordinal (ranking):

- **Stage of cancer (stage I, II, III, IV)**
- **Pathological scores**
- **Education level (elementary, secondary, college)**
- **Pain level (mild, moderate, severe)**
- **Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)**
- **Agreement level (strongly disagree, disagree, neutral, agree, strongly agree)**

➢*Examples of Interval Data*

III- Interval:

➢ Measurements with usually whole numbers (Discrete).

➢ They are measured and have constant, equal distances between values, but the zero point is arbitrary (isn't meaningful).

➢ Example: intelligence or IQ test. We know that the scoring difference between a 100 and a 110 is equal to the scoring distance between 120 and 130, but there is no true zero on this test and an IQ of 140 is not twice as high as an IQ of 70.

➢ **IQ test**
➢ **Number of children in school or house**
➢ **Temperatures (Celsius OR Fahrenheit)**
➢ **Number of admission to hospital**

➢*Examples of Ratio Data*

IV- Ratio:

➢ They are measurable continuous variables and can theoretically take on an infinite number of values - the accuracy of the measurement is limited only by the measuring instrument.

➢Variables have equal intervals between values, the zero point is meaningful, and the numerical relationships between numbers is meaningful.

➢**Examples:**

➢ **Height, weight, heart rate, blood pressure, serum cholesterol, age, temperature (K)**

➢*Changing level of measurements*

Ratio > Interval > Ordinal > Nominal:

➢**Higher level variables can always be expressed at a lower level, but the reverse is not true.**

➢**For example, Body Mass Index (BMI) is typically measured at an interval-level such as 23.4.**

➢**BMI can be collapsed into lower-level Ordinal categories such as:**

- **>30: Obese**
- **25-29.9: Overweight**
- **<25: Underweight**

➢**OR Nominal categories such as:**

- **Overweight**
- **Not overweight**

➢*How to express your data*

➢The most important factor to consider is the level of measurement or type of variable

➢**Nominal variables:**

- ➢ the **mode** is the only appropriate measure of central tendency.

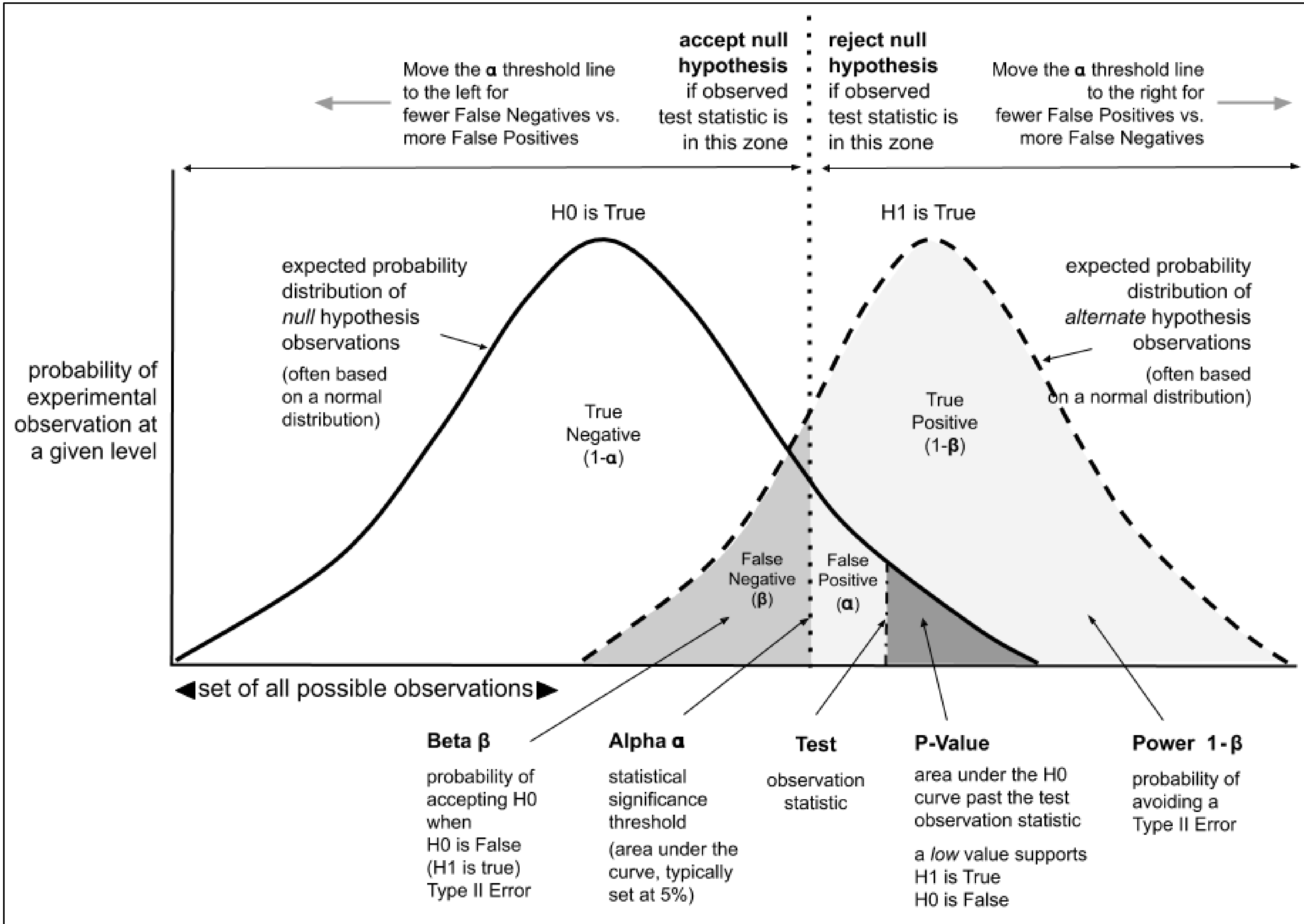
➢**Ordinal variables:**

- ➢ both the **mode and median** are used.

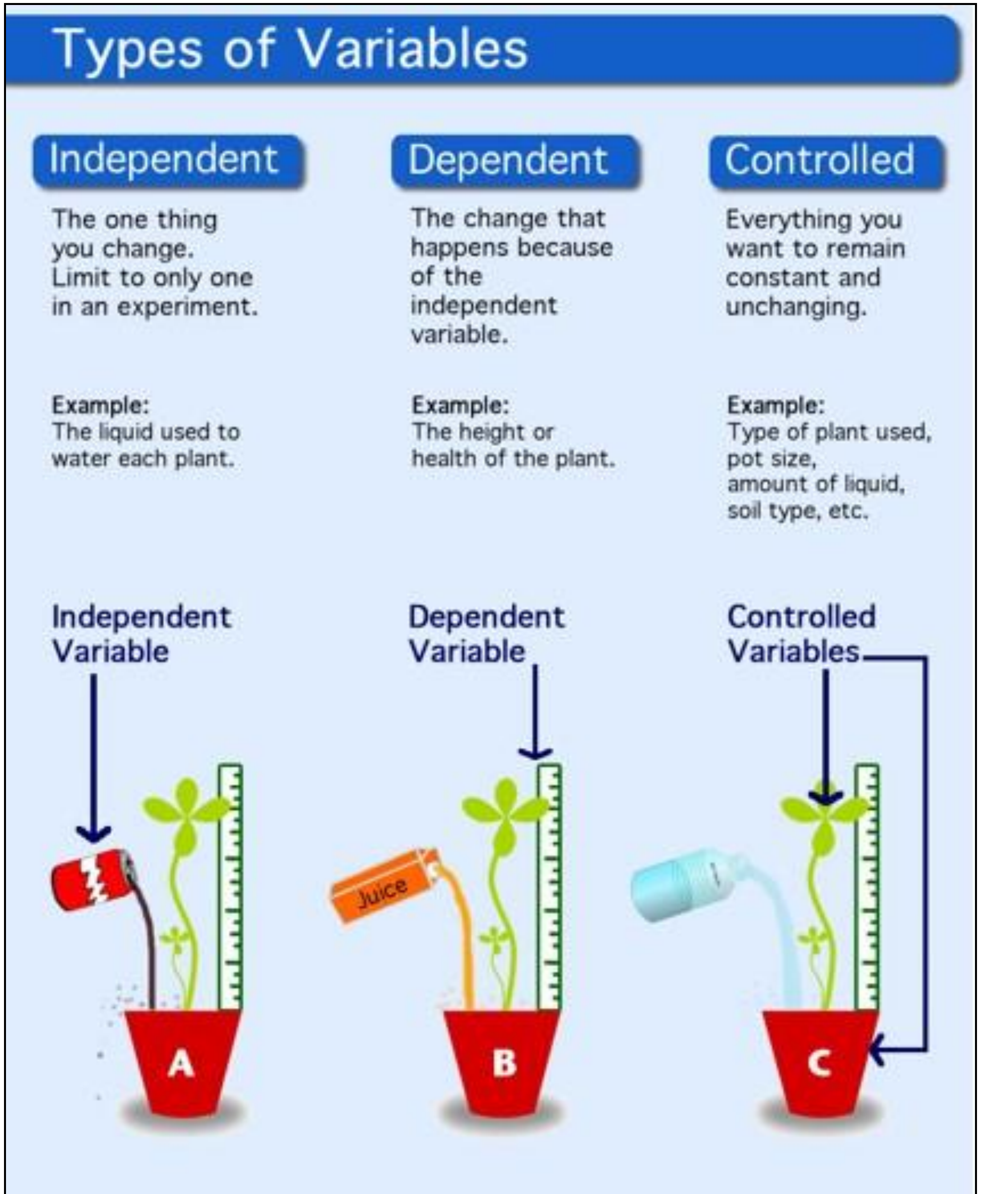
➢**Quantitative variables:**

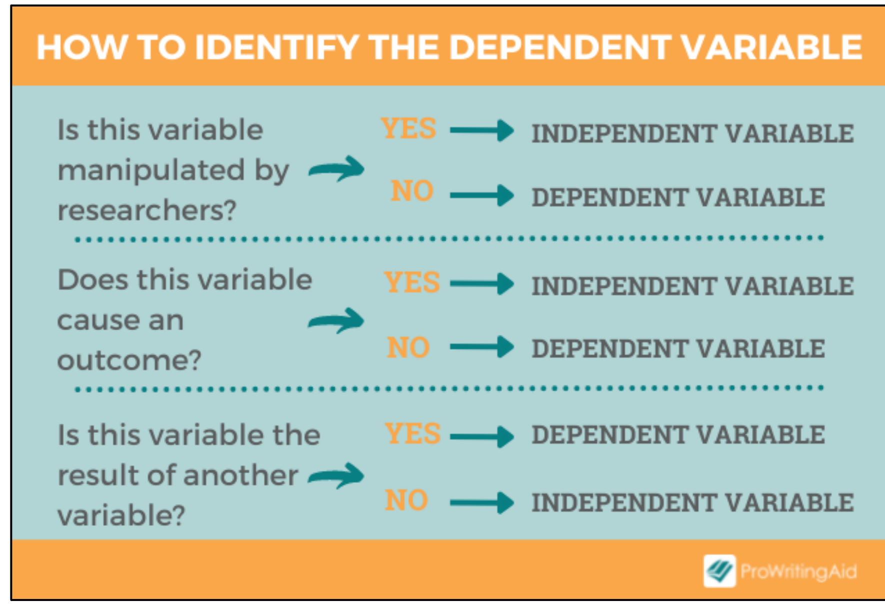
- ➢ **mode, median, and mean** can all be used. Mean is usually the measure of choice because it is a unique value, it uses all values in the data set, and it can be used in subsequent analyses. However, the mean is subject to extreme (high or low) values (outliers). If the data set contains extreme values, calculate both median & mean.

➢ Null Hypothesis



➢ *Types of Variables*





➢ *Types of Statistical Errors*

	Alpha Error (False Positive)	Beta Error (False Negative)
What it is	The error of concluding that there <u>is</u> a Statistically Significant difference, change, or an effect when, in reality, there is not.	The error of concluding that there <u>is</u> <u>not</u> a Statistically Significant difference, change, effect -- when, in reality, there is.
In Hypothesis Testing	The error of Rejecting the Null Hypothesis when it is true.	The error of Failing to Reject the Null Hypothesis when it is false.

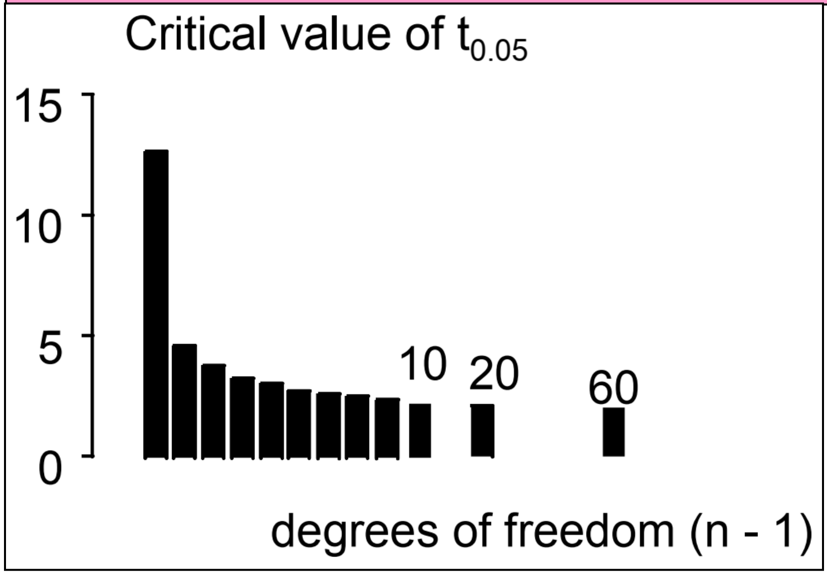
➢ **<u>*β* =</u> probability of a Type II error, known as a "false negative"**

➢ **<u>*1 − β* =</u> probability of a "true positive", i.e., correctly rejecting the null hypothesis. <u>"1 − *β*" is also known as the power of the test.</u>**

➢ **<u>*α* =</u> probability of a Type I error, known as a "false positive"**

➢ **<u>*1 − α* =</u> probability of a "true negative", i.e., correctly not rejecting the null hypothesis. <u>It is also called confidence level</u>**

➢ *What is the optimal size of an experiment?*



➢ **For comparisons between two samples, a minimum of 6 experimental units should be employed. Also, degrees of freedom greater than 10 will not yield a substantial reduction in the critical value of t.**

➢ **The following equation can help to ascertain the number of 'experimental units' required in complex experiments which involve more than two samples:**

$$df = (N - 1) - (T - 1) - (B - 1)$$

df = degrees of freedom of experimental error
N = the number of experimental units
T = the number of treatment combinations
B = the number of blocks and covariates

➢*Parametric vs Non-Parametric*

➢ **Parametric is used for ratio or interval while Non-parametric is used for nominal or ordinal**

➢ **Also parametric tests assumes normal distribution (mean is not significantly different from the median and the variances are homogenous).**

➢ **Sometimes, a skewed distribution can be made sufficiently normal to apply parametric statistics by transforming the variable (by taking its square root, squaring it, taking its log, etc)**

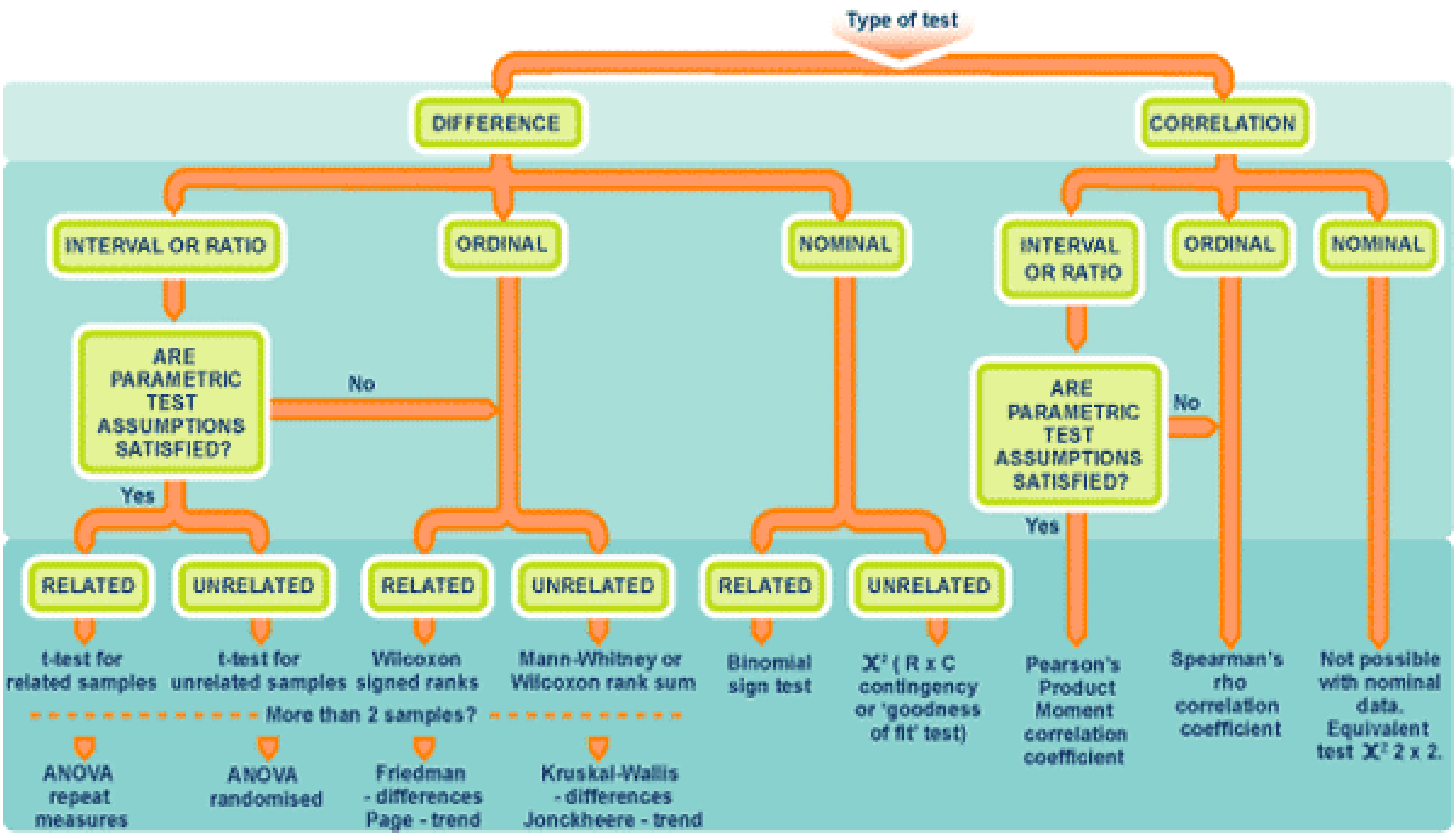
➢ **Non-parameteric statistical procedures are less powerful and give less conclusions, but have little assumptions (conditions to apply), they are better for small sample sizes ($n < 30$).**

➢ **Large data sets present no problems. It is usually easy to tell if the data come from a Gaussian population, but it doesn't really matter because the nonparametric tests are so powerful and the parametric tests are so robust. Small data sets present a dilemma. It is difficult to tell if the data come from a Gaussian population, but it matters a lot. The nonparametric tests are not powerful and the parametric tests are not robust.**

➢Parametric vs Non-Parametric

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

➢Choosing a statistical test



➢ *Degree of Freedom (DF)*

➢ **Simply, in t-test for example, Degrees of Freedom = n – 1**

➢ **DF is the number of values that are free to vary in a data set.**

➢ **Example using the mean (average):**

➢ **Q. Pick a set of 3 numbers that have a mean (average) of 10.**

➢ **A. Some sets of numbers you might pick: 9, 10, 11 or 8, 10, 12 or 5, 10, 15.**

➢ **Once you have chosen the first two numbers in the set, the third is fixed. In other words, you can't choose the third item in the set. The only numbers that are free to vary are the first two. You can pick 9 + 10 or 5 + 15, but once you've made that decision you must choose a particular number that will give you the mean you are looking for. So degrees of freedom for a set of three numbers is TWO.**

➢ *Effect Size*

➢ **Effect size is a statistical concept that measures the strength of the relationship between two variables on a numeric scale.**

➢ **Statistic effect size helps us in determining if the difference is real or if it is due to a change of factors.**

➢ **In hypothesis testing, effect size, power, sample size, and critical significance level are related to each other.**

➢ **In Meta-analysis, effect size is concerned with different studies and then combines all the studies into single analysis.**

➢**In statistics analysis, the effect size is usually measured in three ways:**

- ➢ **(1) standardized mean difference**
- ➢ **(2) odd ratio**
- ➢ **(3) correlation coefficient.**

➢ *Regression OR Correlation*

➢ **Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations.**

➢ **Calculate linear correlation if you measured both X and Y in each subject and wish to quantity how well they are associated.**

➢ **Select the <u>Pearson (parametric)</u> correlation coefficient if you can assume that both X and Y are sampled from Gaussian populations. Otherwise choose the <u>Spearman (nonparametric)</u> correlation coefficient.**

➢ **Don't calculate the correlation coefficient (or its confidence interval) if you manipulated the X variable. Definitely choose linear regression if you manipulated the X variable.**

➢ **Calculate linear regressions only if one of the variables (X) is likely to precede or cause the other variable (Y).**

➢ **It makes a big difference which variable is called X and which is called Y, as linear regression calculations are not symmetrical with respect to X and Y. If you swap the two variables, you will obtain a different regression line.**

➢ **In contrast, linear correlation calculations are symmetrical with respect to X and Y. If you swap the labels X and Y, you will still get the same correlation coefficient.**

➢*Paired vs Unpaired t-test*

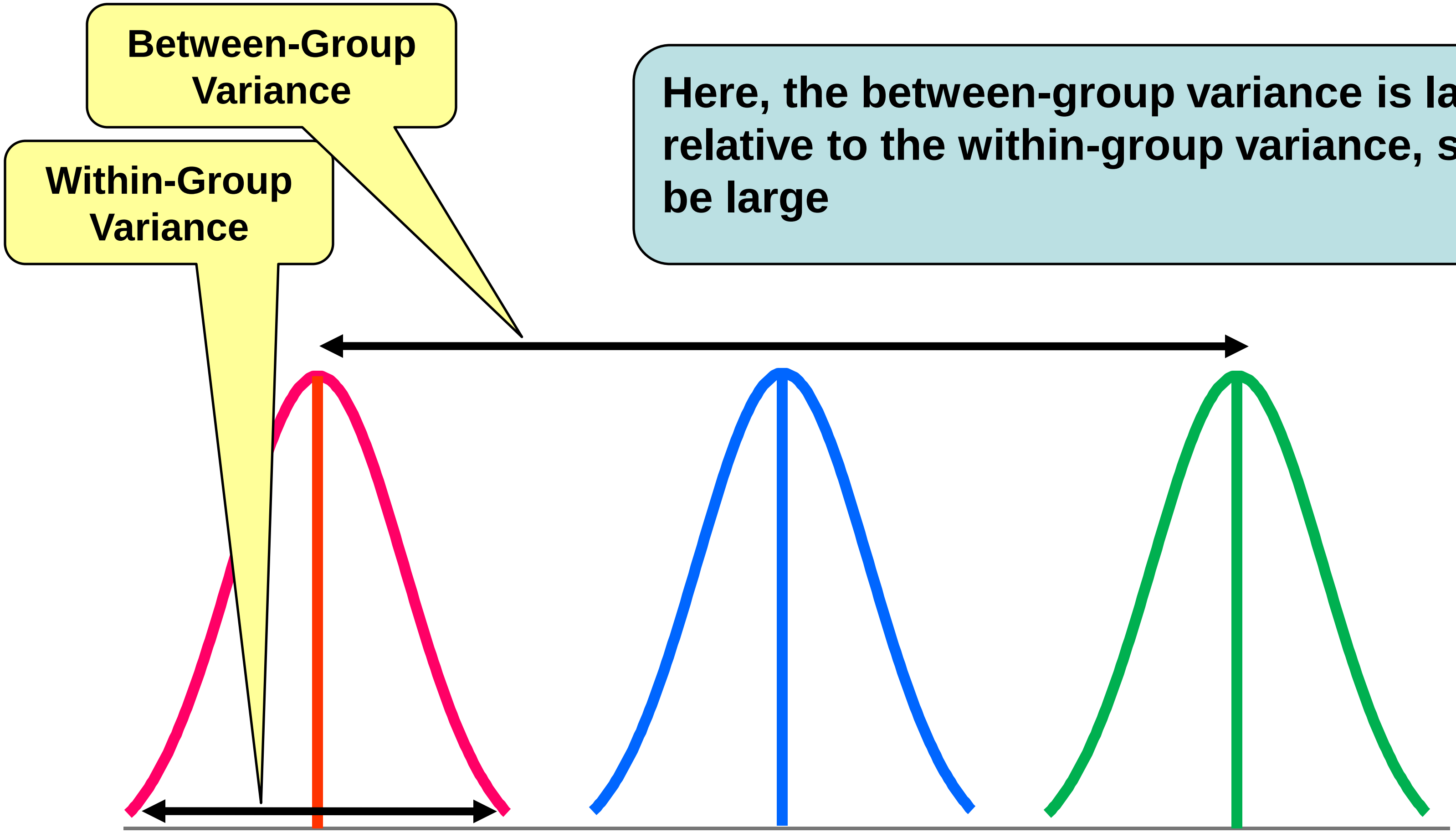
Important: **in paired t-test we need individual data, but in unpaired we can use only means , SE of means and n to calculate t.**
Also DF is (n-2) in unpaired and (n-1) in paired

Subject	A	B
1	10	11
2	0	3
3	60	65
4	27	31

- **Note that the variation between subjects (in rows) is much wider than that within subjects (in columns).**
- **Treating A and B as entirely separate, t=-0.17, p=0.89**
- **Treating the values as paired, t=3.81, p=0.03**
- **Thus paired t-test is more sensitive**

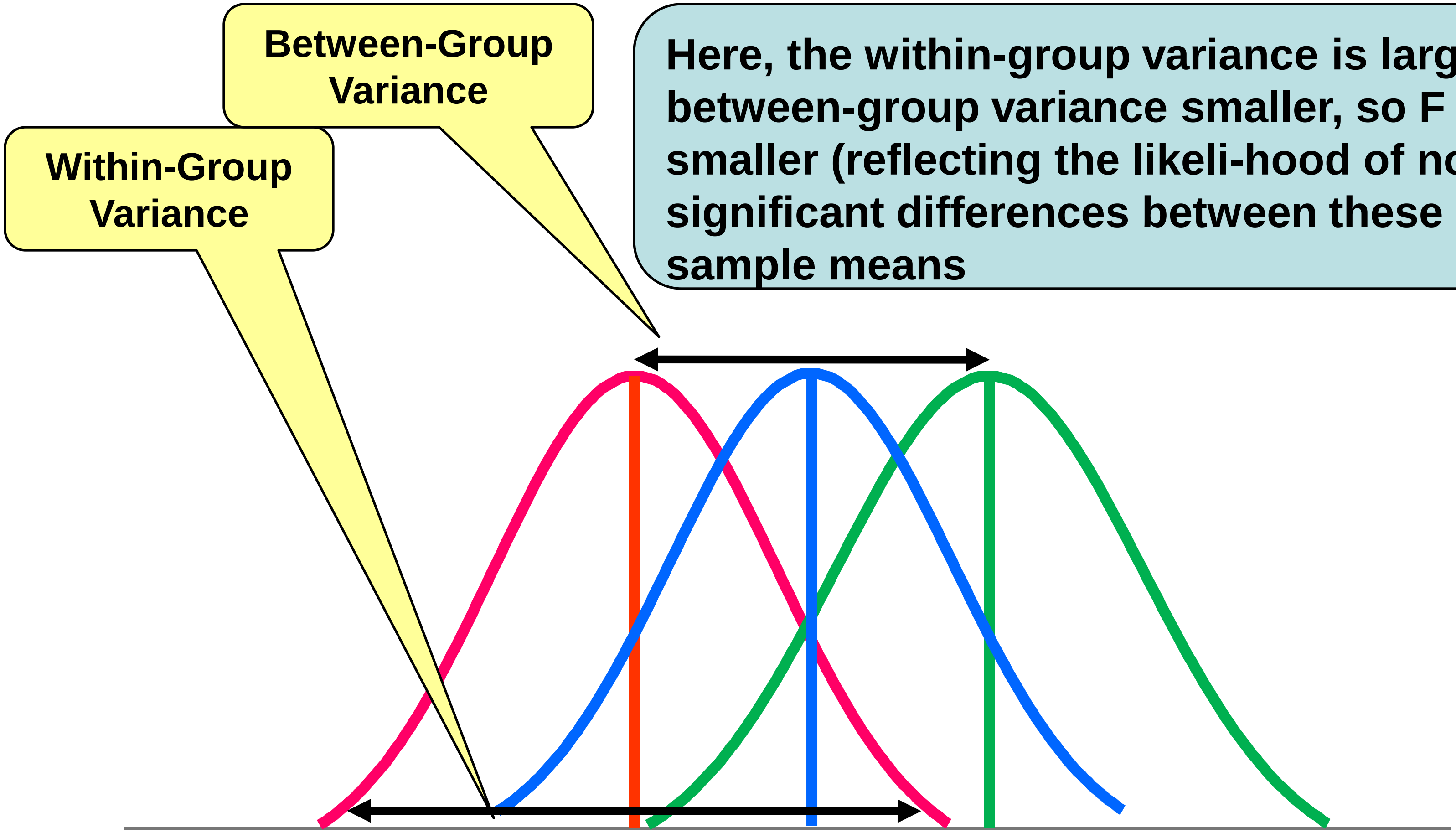
➢ *Analysis of Variance (ANOVA)*

Here, the between-group variance is large relative to the within-group variance, so F will be large

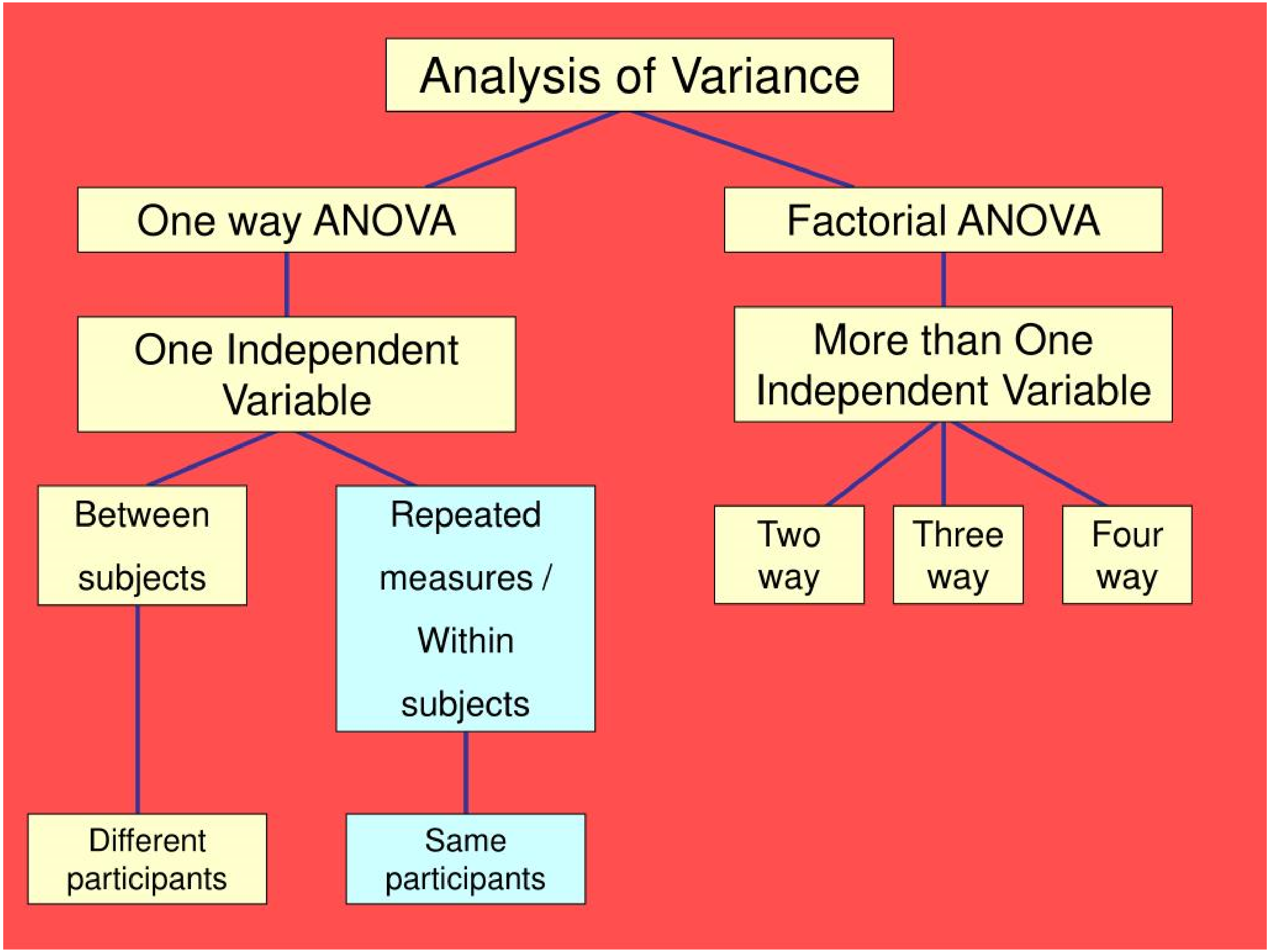


➢ *Analysis of Variance (ANOVA)*

Here, the within-group variance is larger, and the between-group variance smaller, so F will be smaller (reflecting the likeli-hood of no significant differences between these three sample means



➢ *Types of ANOVA*

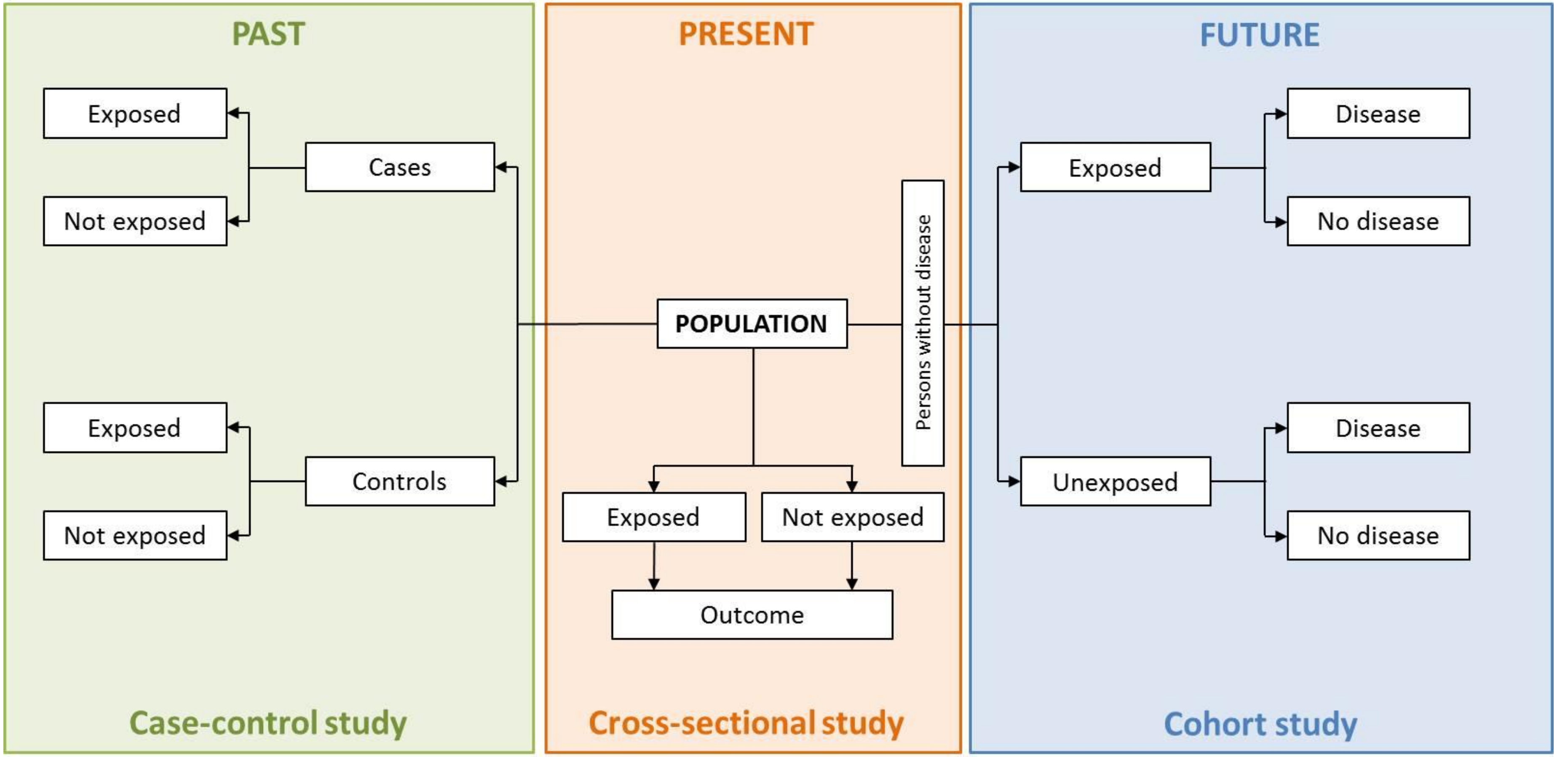


➢ Types of Study

Study Design	Key Feature	Example
Experimental Design		
1- Randomized blinded trial	Two groups created by a random process, and a blinded intervention	The investigator randomly assigns women to receive hormone or identical placebo, then follows both treatment groups for several years to observe the incidence of heart attacks.
Observational Designs		
1- Cohort study	A group followed over time	The investigator examines a cohort of women yearly for several years, observing the incidence of heart attacks in hormone users and non-users.
2- Case-control study	Two groups, based on the outcome	The investigator examines a group of women with heart attacks (the "cases") and compares them with a group of healthy women (the controls) asking about hormone use.
3- Cross-sectional study	A group examined at one point in time	The investigator examines the group of women once, observing the prevalence of a history of heart attacks in hormone users and non-users.

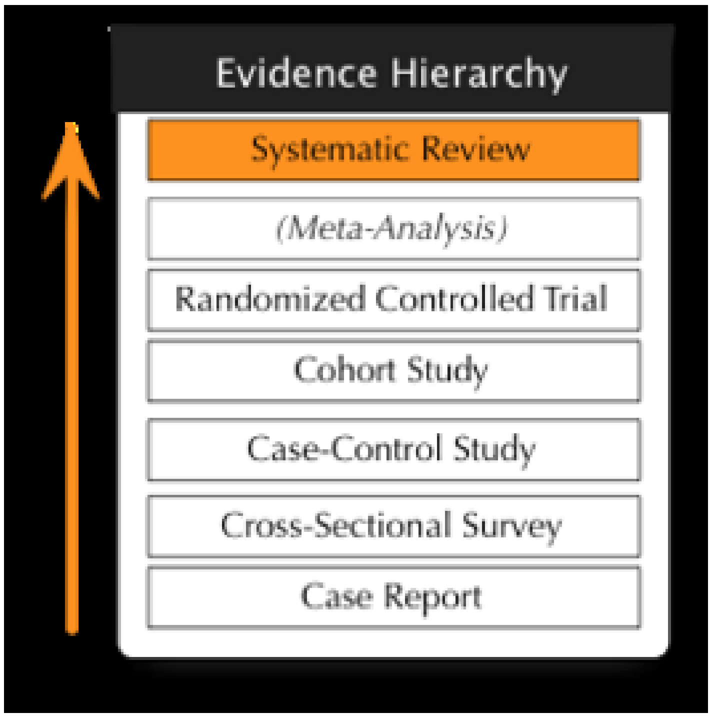
➢ Types of Study

➢ Types of Study



Retrospective OR Prospective?

➢ Types of Study



Objective	Common design
Prevalence	Cross sectional
Incidence	Cohort
Cause (in order of reliability)	Cohort, case-control, cross sectional
Prognosis	Cohort
Treatment effect	Controlled trial

Thank you