Planning and Performing data analysis

Dr Khalifa Elmusharaf

Senior Lecturer in Public Health

Contact information

Training Course in Sexual and Reproductive Health Research
Geneva Workshop
October 2016

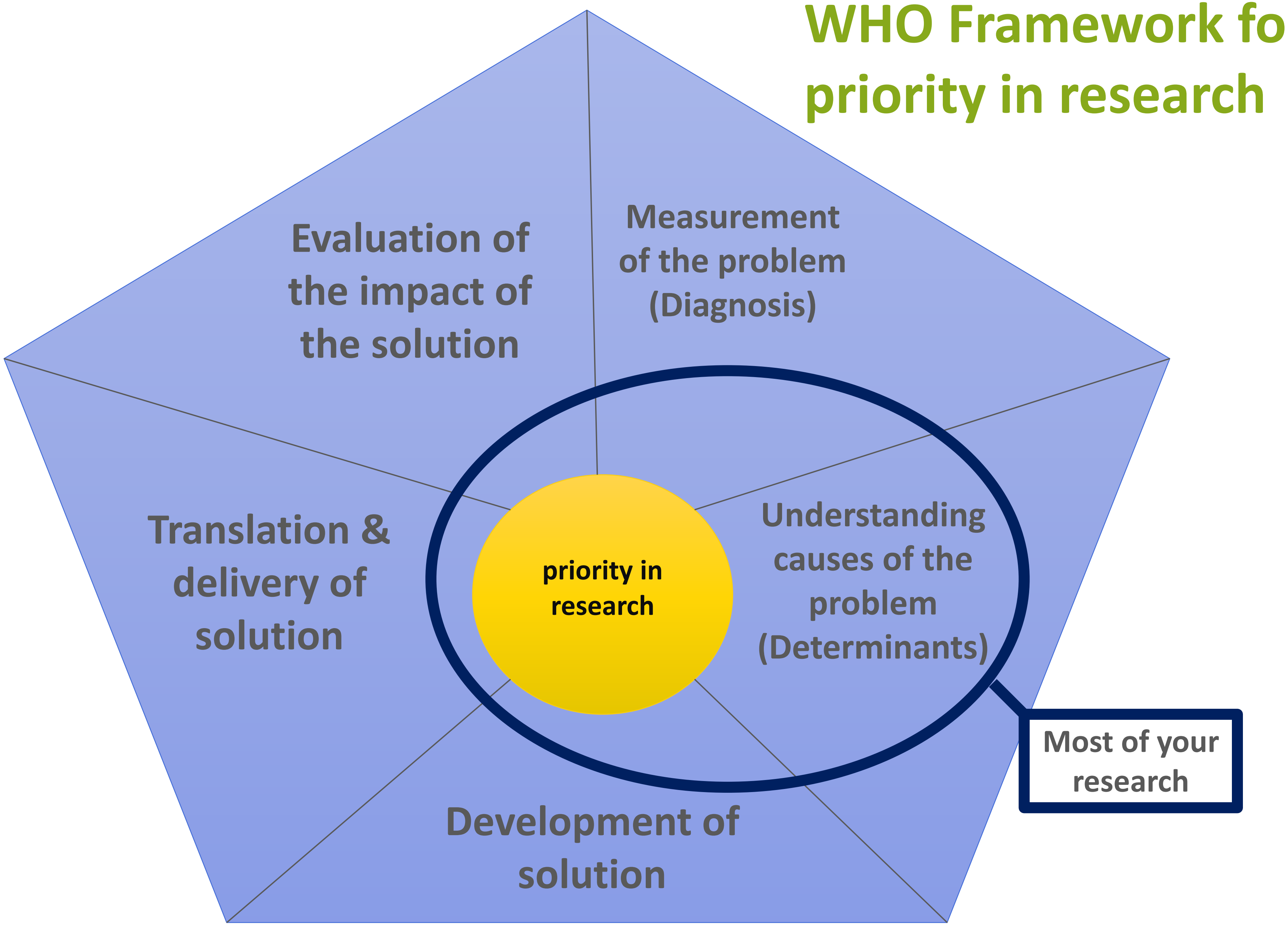
UNIVERSITY of LIMERICK
OLLSCOIL LUIMNIGH

GRADUATE ENTRY MEDICAL SCHOOL

THE SUNDAY TIMES
UNIVERSITY OF THE YEAR 2015

GENEVA FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH

WHO Framework for describing priority in research



Measurement of the problem (Diagnosis)

Understanding causes of the problem (Determinants)

Development of solution

Translation & delivery of solution

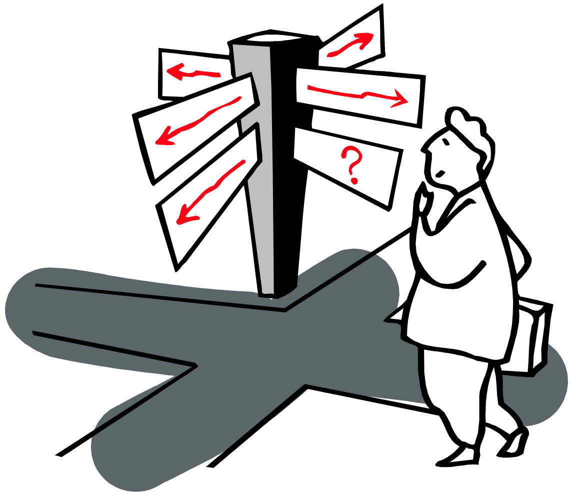
Evaluation of the impact of the solution

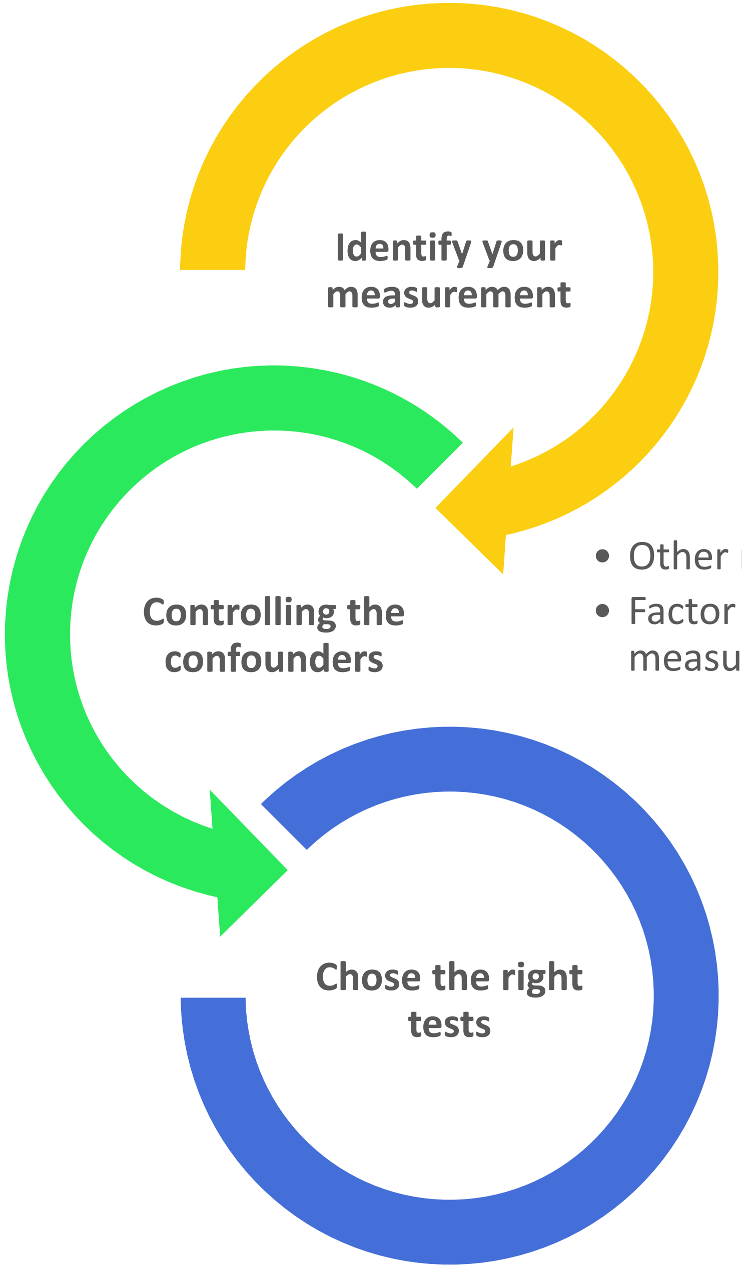
priority in research

Most of your research

Identifying your analysis strategy

- Why A Data Analysis Strategy?
- If you don’t know where you are going, you can wind up anywhere





Identify your measurement

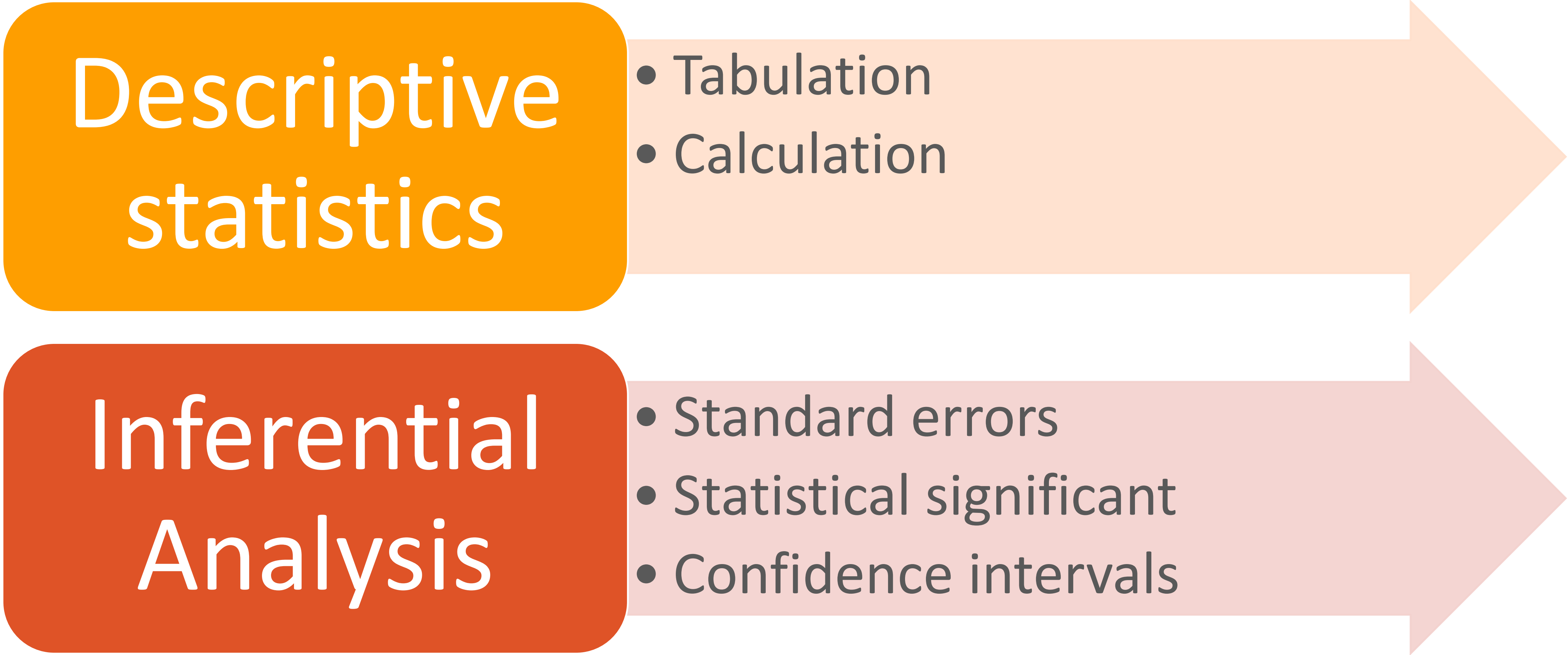
- Operationalization
- Proxy indicators

Controlling the confounders

- Other risk factors
- Factor affect your measurements

Chose the right tests

Descripting & analysing research results



Descriptive statistics

- Tabulation
- Calculation

Inferential Analysis

- Standard errors
- Statistical significant
- Confidence intervals

Tabulation (categorical data)

Frequency distribution tables

Educational Level	Frequency	Percentage
Primary	70	35%
Secondary	80	40%
University	50	25%
Total	200	100%

Cross-tabulation tables

	HCV + ve		HCV - ve		Total
Sex	n	%	n	%	n
Male	33	16.4	168	83.6	201
Female	11	10.3	96	89.7	107
Total	44	14.3	264	85.7	308

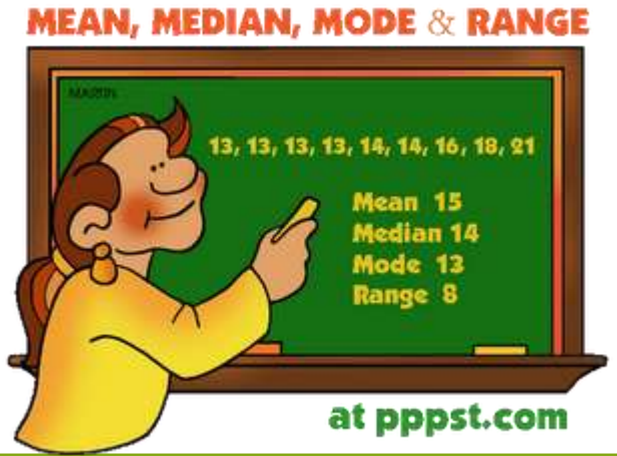
Calculations (Numerical)

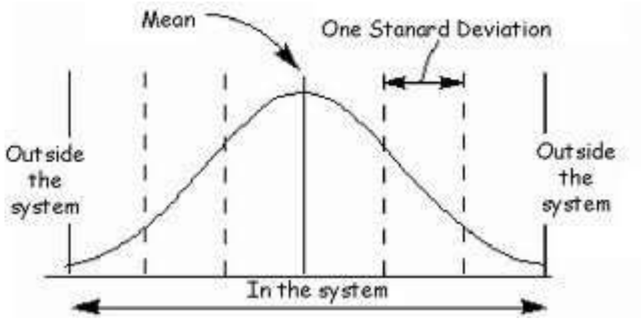
- **Central tendency**
  - The mean
  - The median
  - The mode

- **Variability**
- The range
- The standard deviation (SD)
- The percentiles.

Other calculations

- Ratios
- Rates





25% 50% 75% 100%

Standard errors

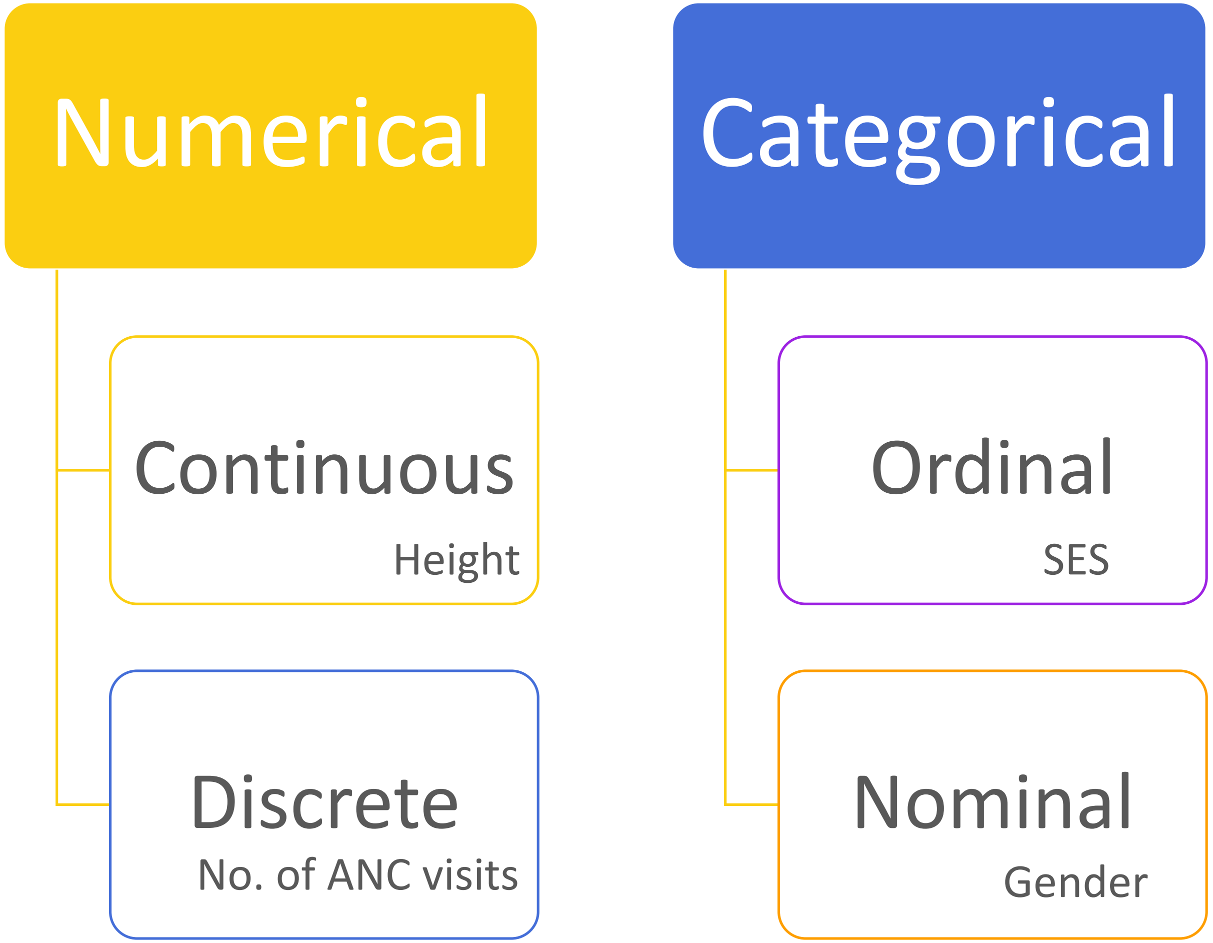
- Measure of the uncertainty in a sample statistic
- Measure the probability that the finding in the sample will reflect the finding in the population
- SE depends on two factors:
  - The size of the sample,
  - The variations of measurements in the sample indicated by the standard deviation
- **Calculated for:**
  - Mean
  - Difference between 2 means
  - Percentage
  - Difference between 2 percentages
  - Correlation coefficient

Selection of statistical test

- **In general, the type of statistical test to be used depends on:**

1. Type of data to be analysed
2. How the data are distributed
3. Type of sample
4. The question to be answered.

1. Type of data



Numerical

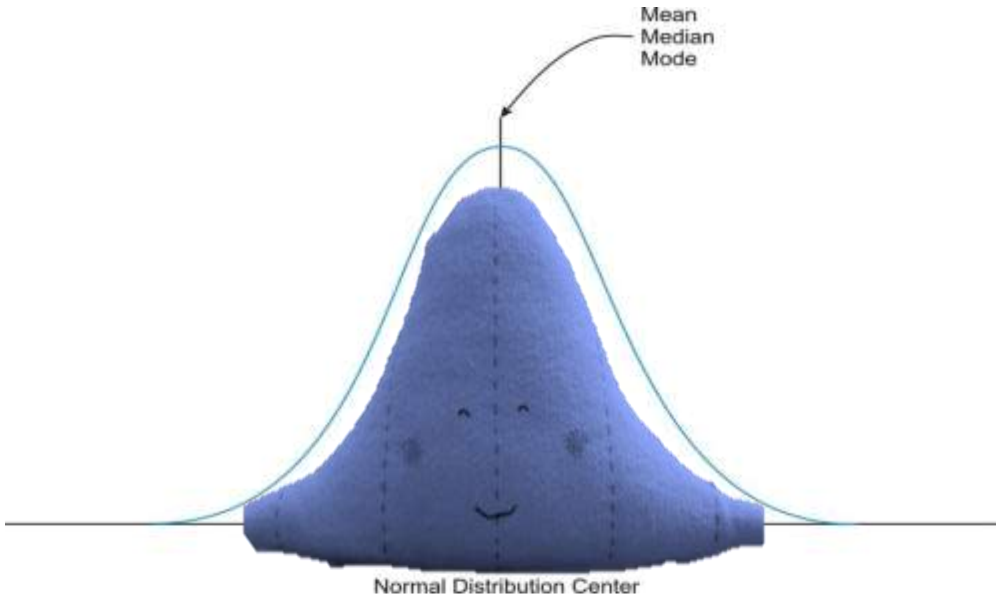
- Continuous
  Height
- Discrete
  No. of ANC visits

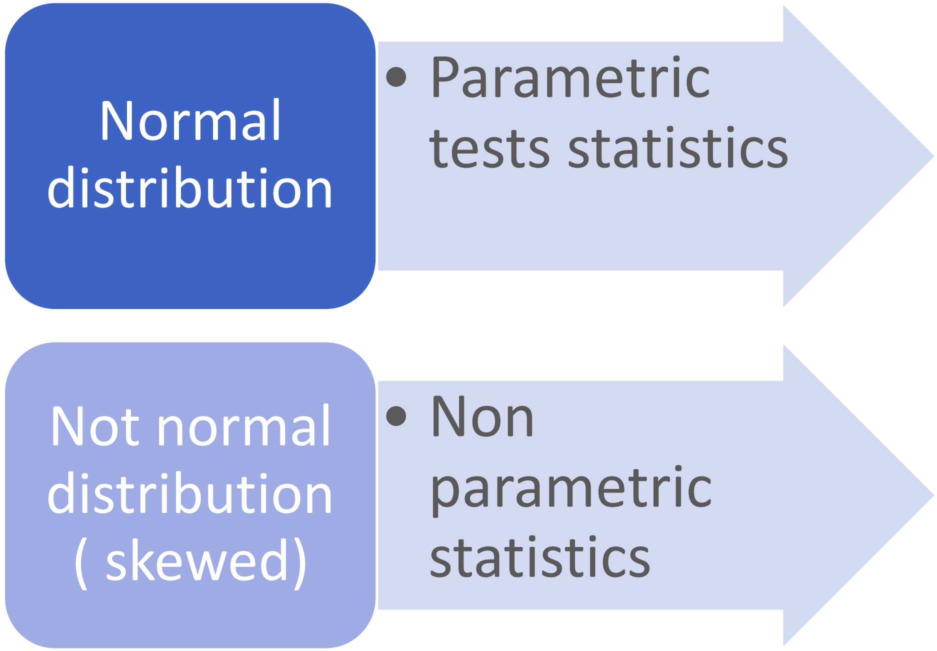
Categorical

- Ordinal
  SES
- Nominal
  Gender

2. Distribution of the data

- Data fall in a normal distribution when they are spread evenly around the mean, and the frequency distribution curve is bell shaped





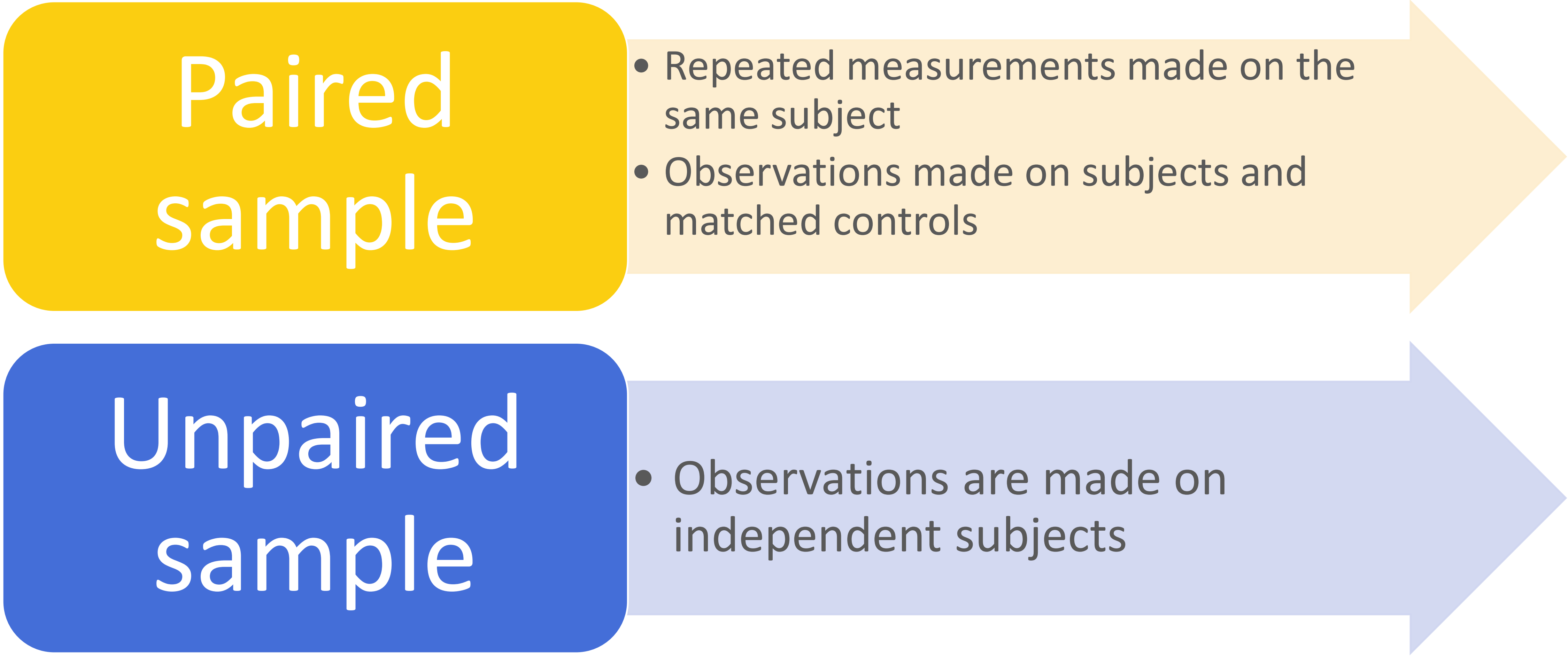
Normal distribution

- Parametric tests statistics

Not normal distribution (skewed)

- Non parametric statistics

3. Type of sample



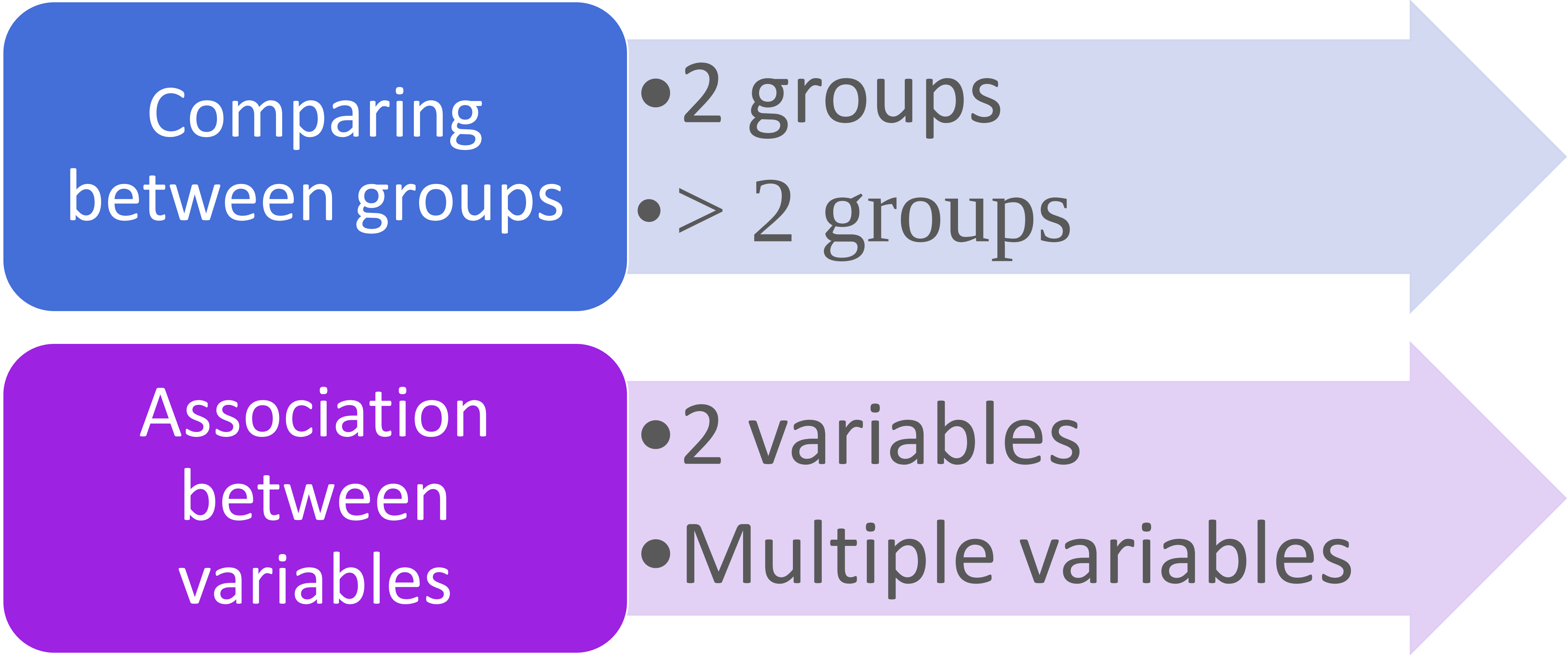
Paired sample

- Repeated measurements made on the same subject
- Observations made on subjects and matched controls

Unpaired sample

- Observations are made on independent subjects

4. Questions to be answered



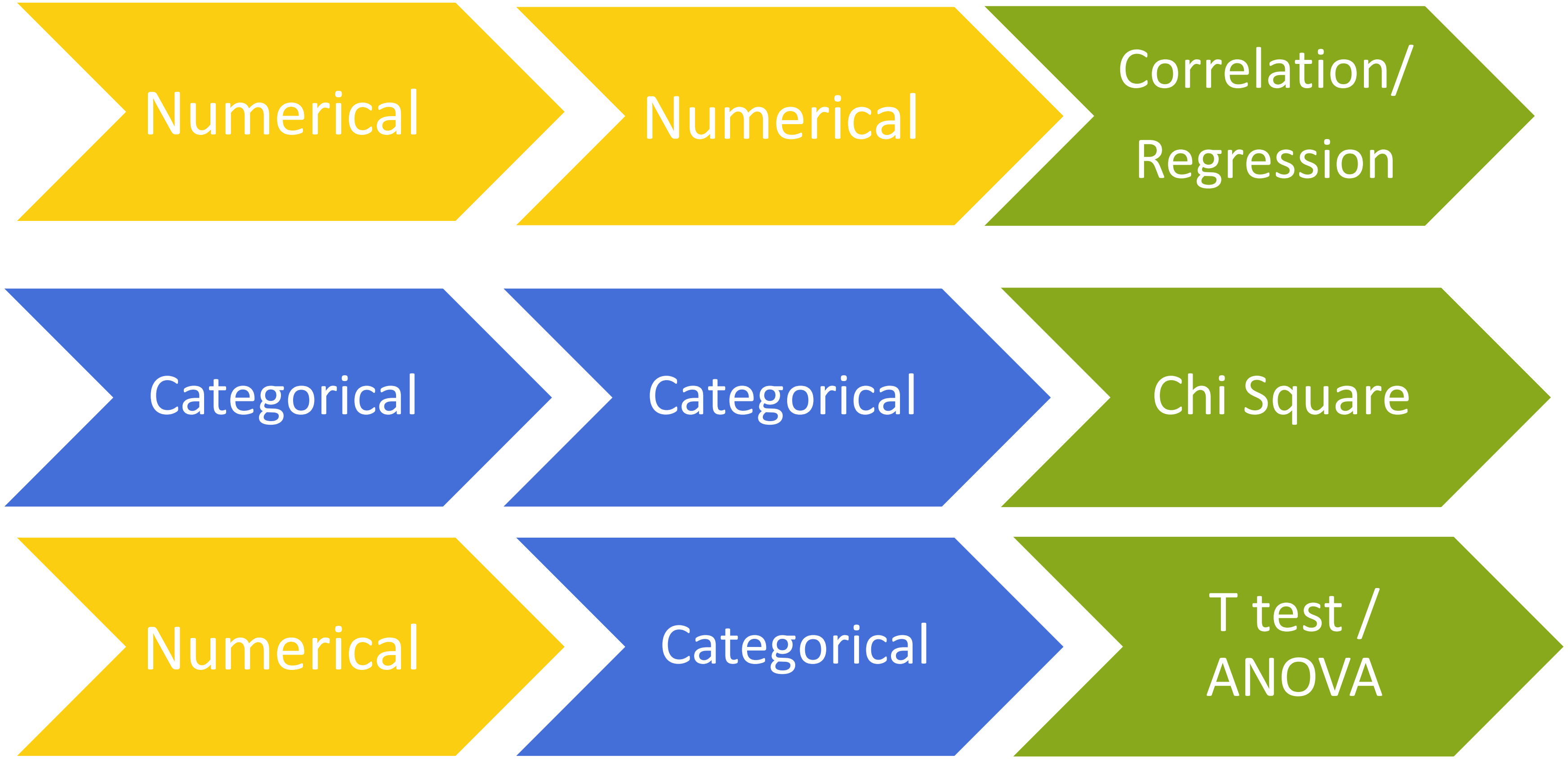
Comparing between groups

- 2 groups
- > 2 groups

Association between variables

- 2 variables
- Multiple variables

Simple way to chose the right test



Numerical → Numerical → Correlation/ Regression

Categorical → Categorical → Chi Square

Numerical → Categorical → T test / ANOVA

Compare between means

Task:

- Determine the association between maternal **smoking** and **baby birth weight**

What to do:

- You need to test if there is any significant difference between the mean birth weight among smokers and non-smokers.

How to do it:

- T test: The t test is used for numerical data to determine whether the difference between the means of two groups can be considered statistically significant.

Linear relationship

Task:

Determine the association between **weight** of mother and birth **weight**

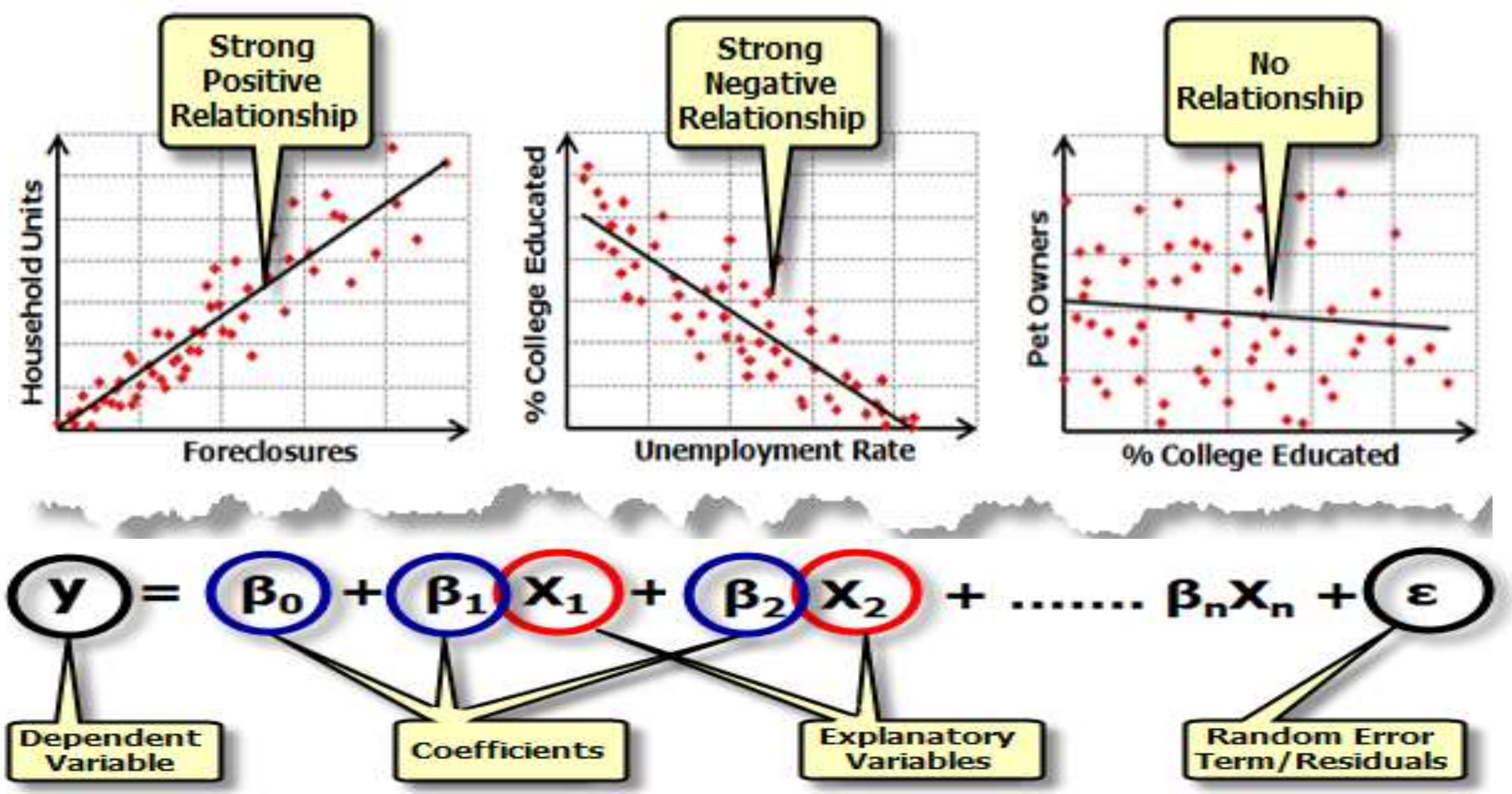
What to do:

- You need to test if there is a significant linear relationship between Weight of mother and Birth weight correlation

How to do it:

- Measure correlation coefficient “r”

Correlation


Strong Positive Relationship
Household Units
Foreclosures
Strong Negative Relationship
% College Educated
Unemployment Rate
No Relationship
Pet Owners
% College Educated
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ \beta_n X_n + \varepsilon$
Dependent Variable
Coefficients
Explanatory Variables
Random Error Term/Residuals

Scatter diagram

Correlation coefficient

Regression equation

Correlation

- **Correlation coefficient “r”**
- When the relationship between two variables can be expressed graphically by a straight line, correlation can be expressed as the correlation coefficient.
- Correlation may be positive or negative. When one variable increases as the other increases, the correlation is positive; when one decreases as the other increases it is negative.
- The correlation coefficient (r) is measured on a scale that varies from +1 through 0 to –1. Complete correlation between two variables is expressed as 1. It should be clear that correlation means association, but does not necessarily mean causation. This conclusion is left to the interpretation of the results.

Compare between proportions

Task:

Determine the association between **prevalence** of Low birth weight and **smoking**

What to do:

- You need to test if the prevalence of Low birth weight statistically significant different between smokers and non-smokers

- **How to do it:**
- **Chi** square test: The Chi-square test is used for categorical data to find out whether observed differences between proportions of events in groups may be considered statistically significant.

Statistical significance and P value

- The likelihood that a relationship is not caused by chance.

- In general, chance less than 5% is acceptable.

- ➔ if P value < 5% = the relationship is not due chance

- A result is considered to be statistically significant (unlikely to be due to chance), if the P value is less than 5% (**P less than 0.05**)

Examples of P value

- P= o.2 (no statistical significant)
- P=0.1 (no statistical significant)
- P=0.06 (no statistical significant)
- P=0.05 (statistical significant)
- P=0.03 (statistical significant)
- P=0.006 (statistical significant)

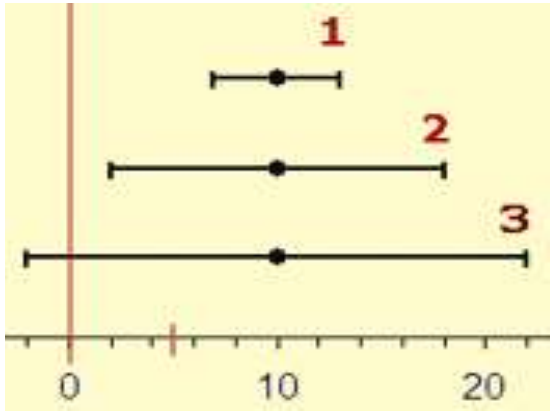
Confidence intervals (CI)

Statistical significance of the result does not give us an indication of the magnitude of that difference in the population from which the sample was studied.

CI provides a range of possibilities for the population value

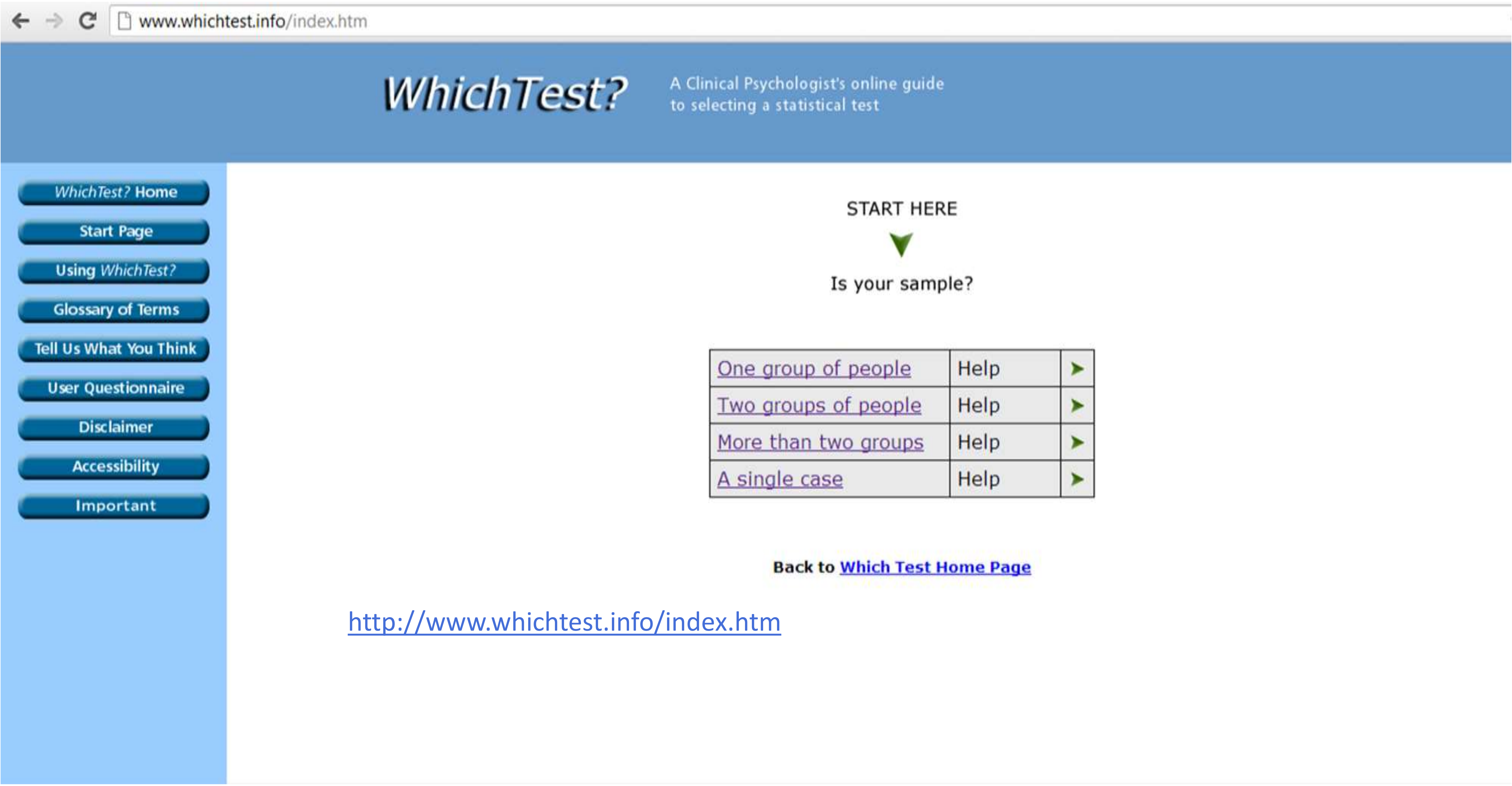
- **CI allows us to estimate the strength of the evidence:**

- Narrow CI indicates **strong evidence.**
- Wide CI indicates greater uncertainty about the true value of a result
- 95% CI doesn't contain a zero difference.



Confidence intervals (CI) Formula:

$$95\%\ \text{Confidence Interval}: \quad \underset{\text{estimate}}{\bar{x}} \underbrace{\pm 1.96 * \left(\frac{\overset{\text{Standard deviation}}{SD}}{\sqrt{n}}\right)}_{\text{Margin of error}}$$



WhichTest?

A Clinical Psychologist's online guide
to selecting a statistical test

START HERE

Is your sample?

One group of people	Help	➤
Two groups of people	Help	➤
More than two groups	Help	➤
A single case	Help	➤

Back to Which Test Home Page

http://www.whichtest.info/index.htm