

Author: PhD. MsC. Bondor Cosmina-Ioana

Hypothesis Testing 1

A ALWAYS

S SEEK

K KNOWLEDGE

Lecture Objectives

- Statistical Tests
- Definition of statistical tests
- Steps involved in conducting a statistical test
  - Null and alternative hypotheses (H_0 and H_1) – how to formulate them.
  - One-tailed and two-tailed tests – understanding different types of tests.
  - Significance level
  - Critical region (rejection region)
  - p-values – how to calculate and interpret them
  - Decision-making in hypothesis testing – rejecting or failing to reject the null hypothesis.
- Errors in statistical hypothesis testing (Type I and Type II errors)
- Power of a statistical test
- Common statistical tests

Aim

Compare two groups
Compare one group with the population
Compare multiple groups

} two approaches:
with confidence intervals
with statistical tests

Objective: to find difference δ (delta)

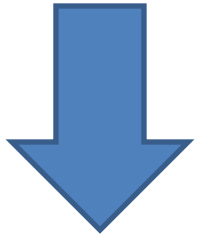
Examples of medical questions

- Treatments
  - Is Losartan more efficient than Furosemide in reducing systolic blood pressure?
- Risk factors
  - Is Borrelia burgdorferi virus a risk factor for neuropathy?
  - Is the weight related to oxidative stress ?

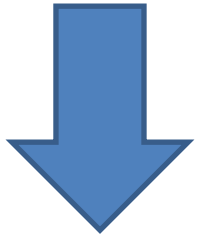
- statistical inference
  - evaluate whether the observed difference is likely to reflect a real difference or random variation

Why Are Inference Important in Medicine?

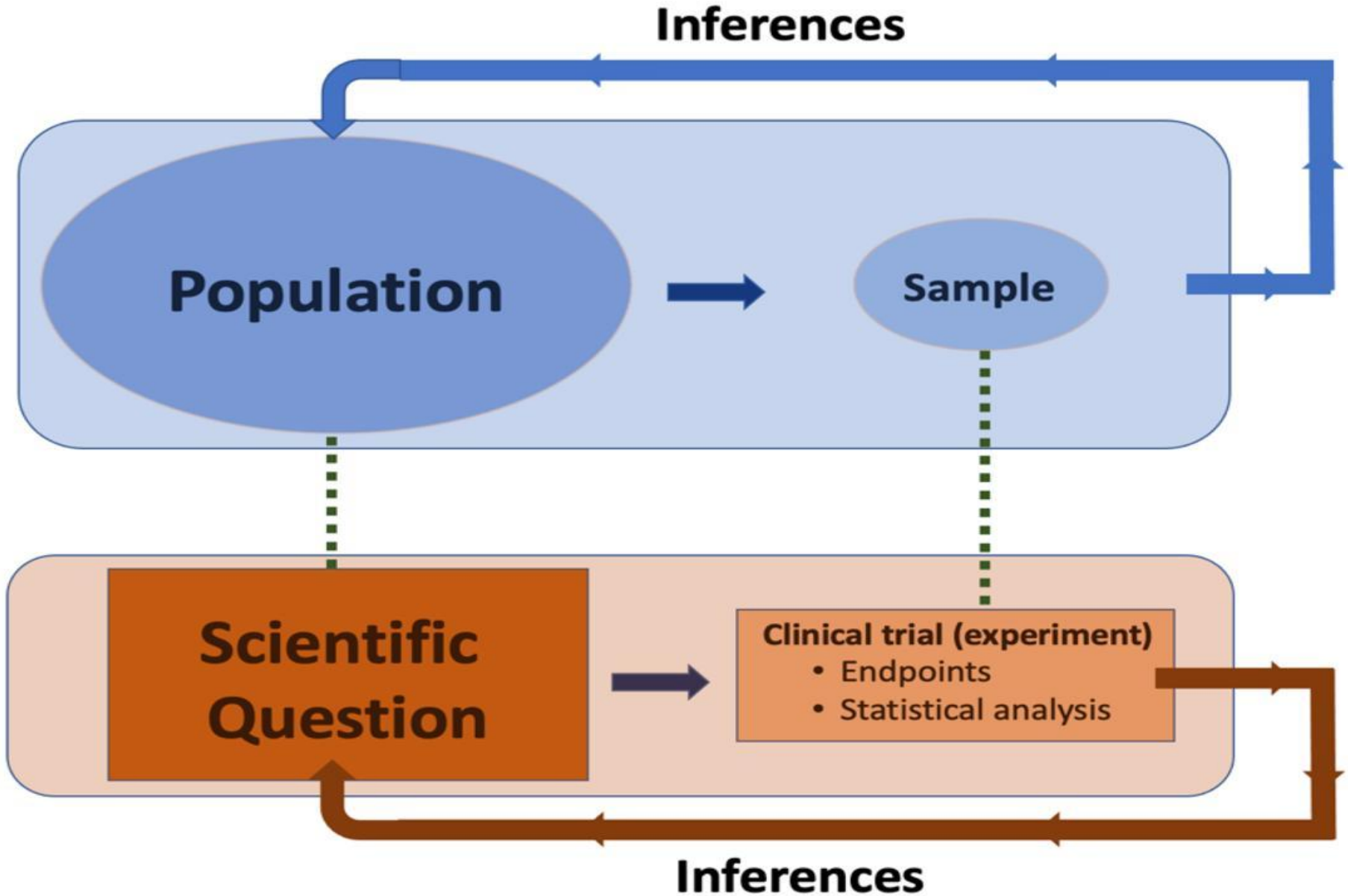
Inference helps answer this questions based on evidence (real life evidence)



the answer is objectively



Support evidence-based medicine


Inferences
Population
Sample
Scientific Question
Clinical trial (experiment)
• Endpoints
• Statistical analysis
Inferences

- Medical question
  - if the premature-born boy's weight during childhood is influenced by the condition of their birth
- we cannot respond exactly to the question
  - but we can compare the average weight of 2 years boys who were born prematurely with the 2 years boys who were not
    - we will compare something with something to see a difference δ
    - what to compare
    - condition of what we compare should be similar except the feature of interest

Formulating the Research Question

- defines what the researcher wants to investigate
- A good research question:
  - Clear
  - Specific
  - Measurable
  - Relevant to clinical practice

Objective: δ

Step 1: establishing the objective

Transform the medical question into statistical hypotheses

ex. Is the infection with Covid-19 virus the cause of depression? → Is the frequency of depression higher in people who tested positive for Covid-19 in the last year compared to those who did not?

Objective of the research: demonstrate that **there is a difference** δ between two groups **which will maintain in majority o**f the studies if we repeat the study on all possible samples (is not random)

Objective: δ – clinically important

Step 1: establishing the difference δ

- the difference δ should be clinically important
- What difference δ (mean δ) should be between two drugs to declare one is more efficient than the other for the treatment of the targeted disease?
- or
- In how many people (frequency δ) the treatment with the drug should be efficient compare with the placebo that one can say that the drug is more efficient than the placebo?

Observation: δ can be difference between arithmetic means, medians, proportions etc.

Step 2: establish the design of the study based on:

Objective: δ – clinically important

Study design

- what you want to compare:
  - 2 frequencies
  - 2 means
  - 2 medians
  - 3 frequencies
  - 3 means
  - etc...

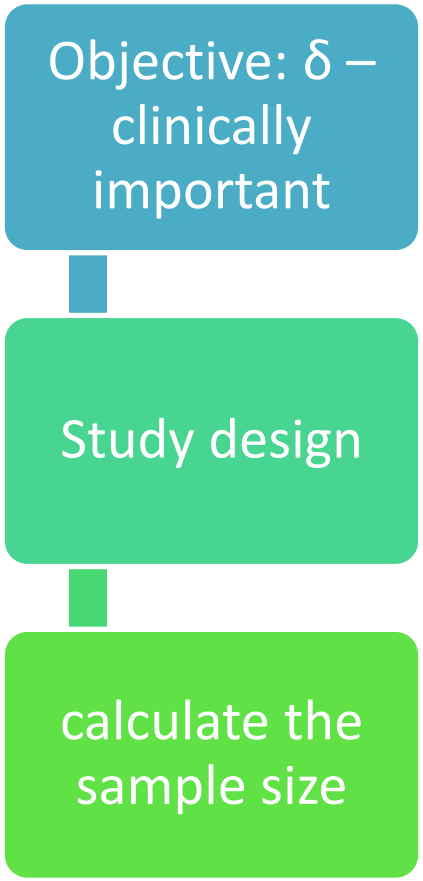
- the groups
  - dependent
  - independent

meaning: finding the right statistical test

Independent or dependent samples

- **Independent samples**
  - there is no connection between subjects in the two samples
- **Dependent samples**
  - subjects are related somehow
    - subjects are paired,
    - subjects are matched
      - age,
      - gender
      - etc.,
    - the same subjects tested two/multiple times
      - after the treatment
      - before the treatment

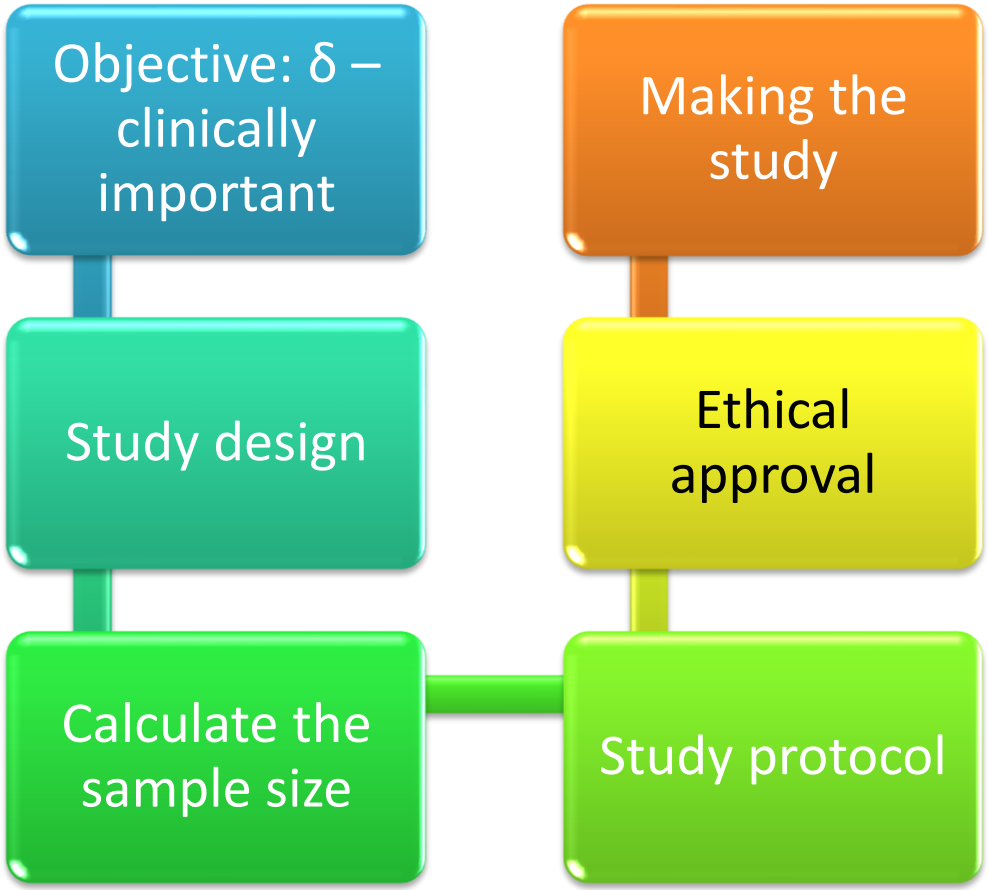
Step 3: compute the needed samples size



- different formulas for different designs
- a lot of possible design: you will need a statistician

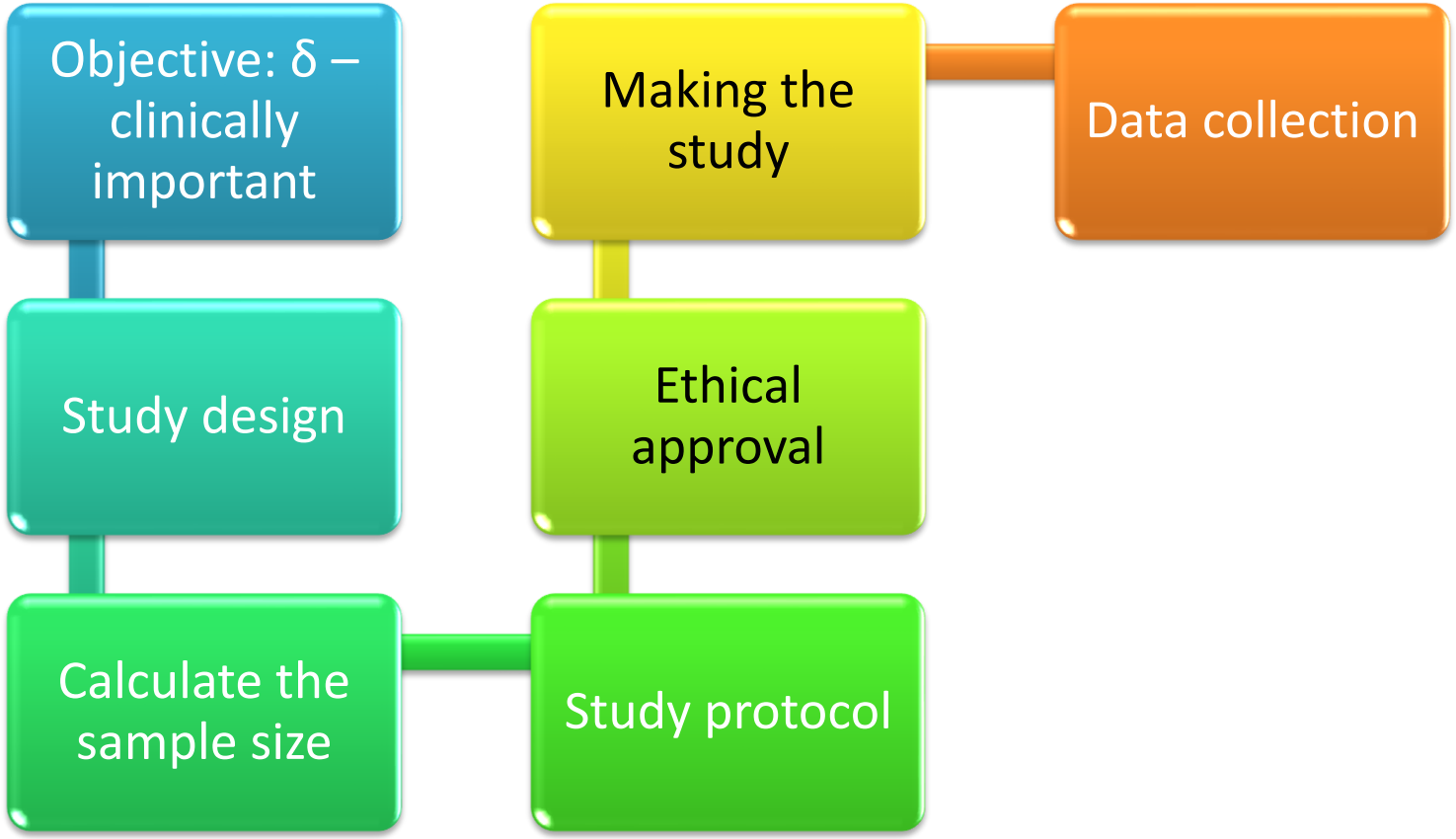
- there are online calculators
- example for two proportion:
- https://statpages.info/proppowr.html

Step 6: carrying out the study



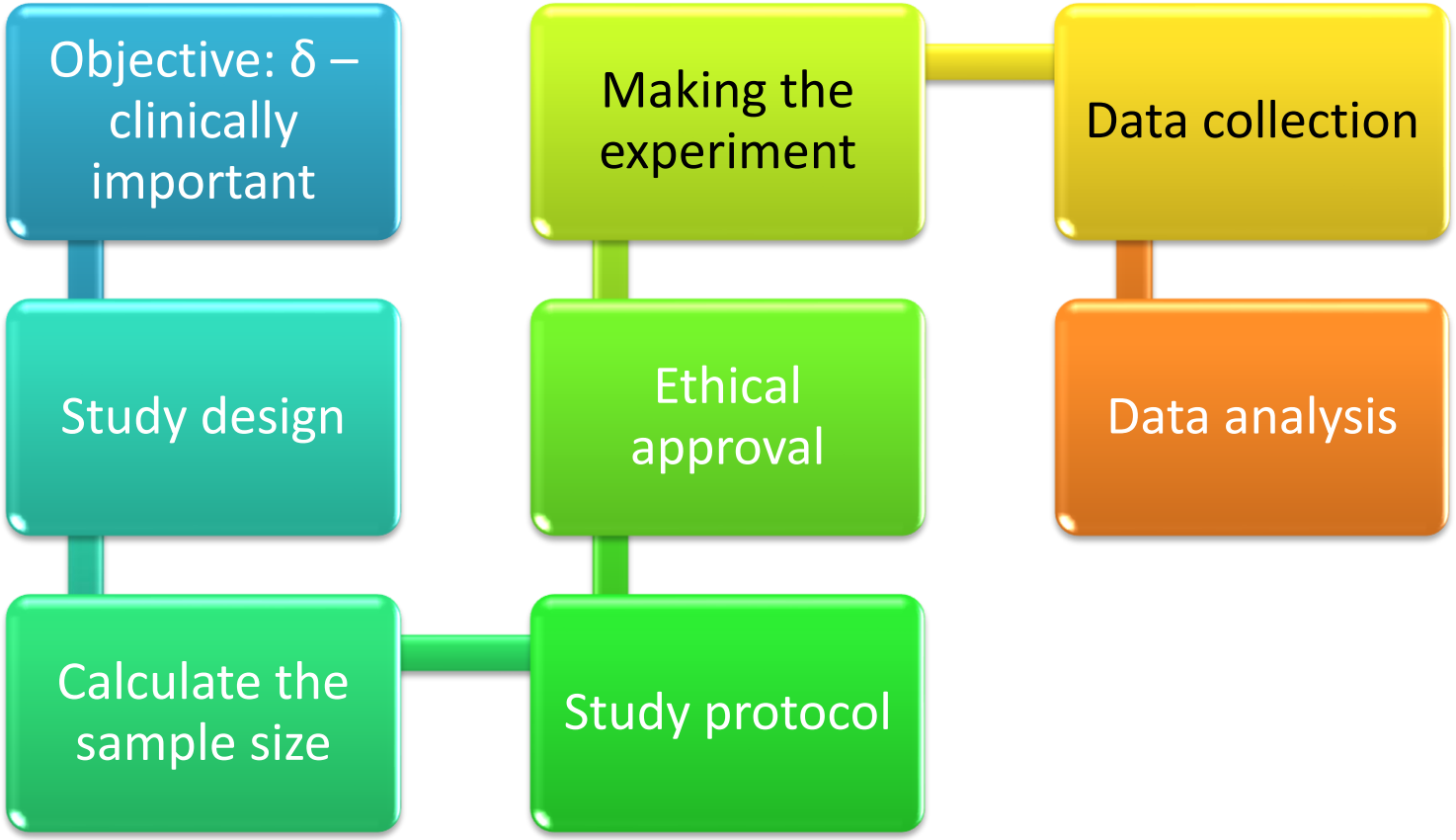
- after you establish the protocol of the study in all details you can start the study

Step 7: collect the data



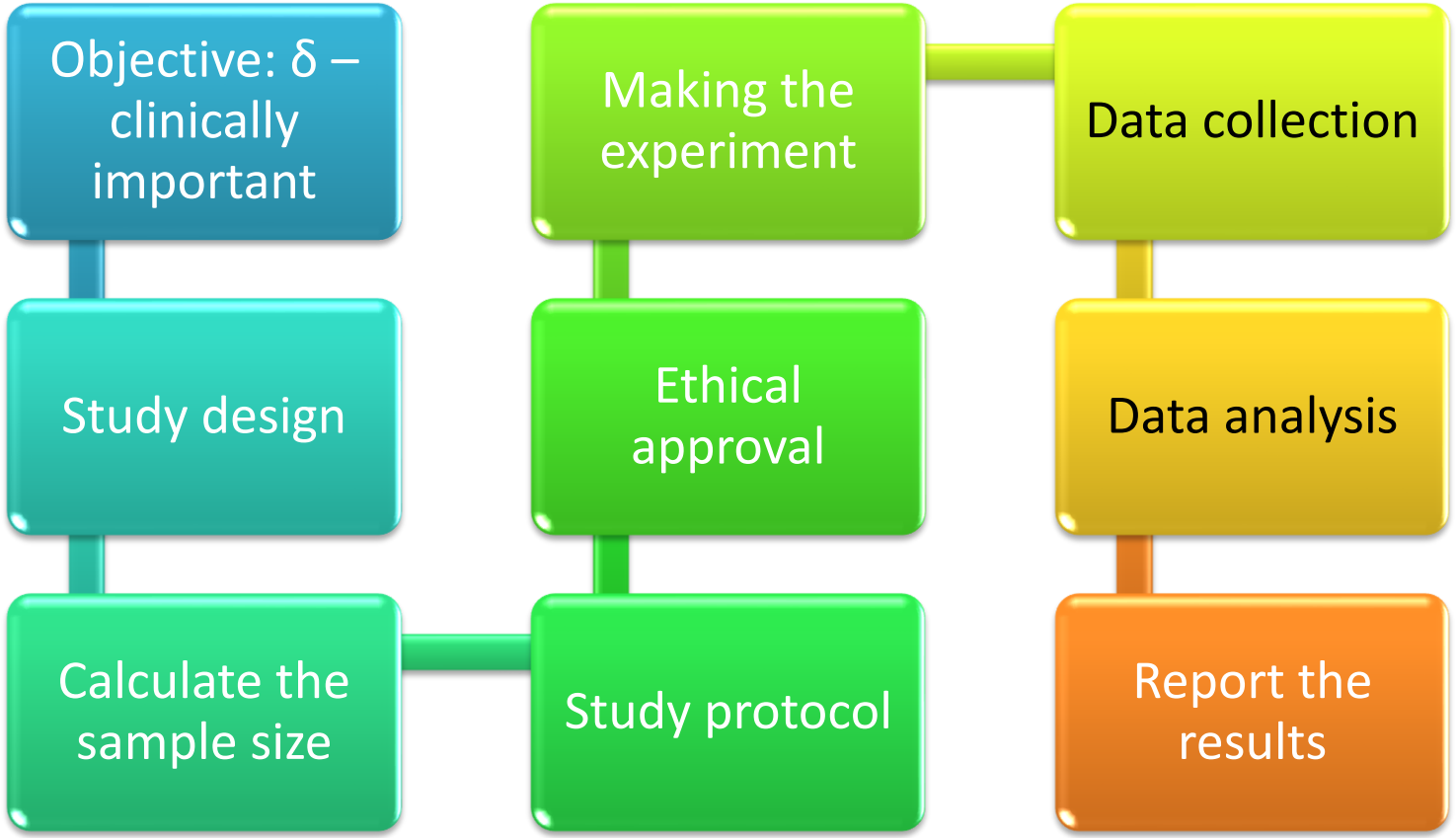
- collect the data about the variables of interest in Excel table or other applications
- data should be deidentified (no personal data should be collected)

Step 8: analyze the data



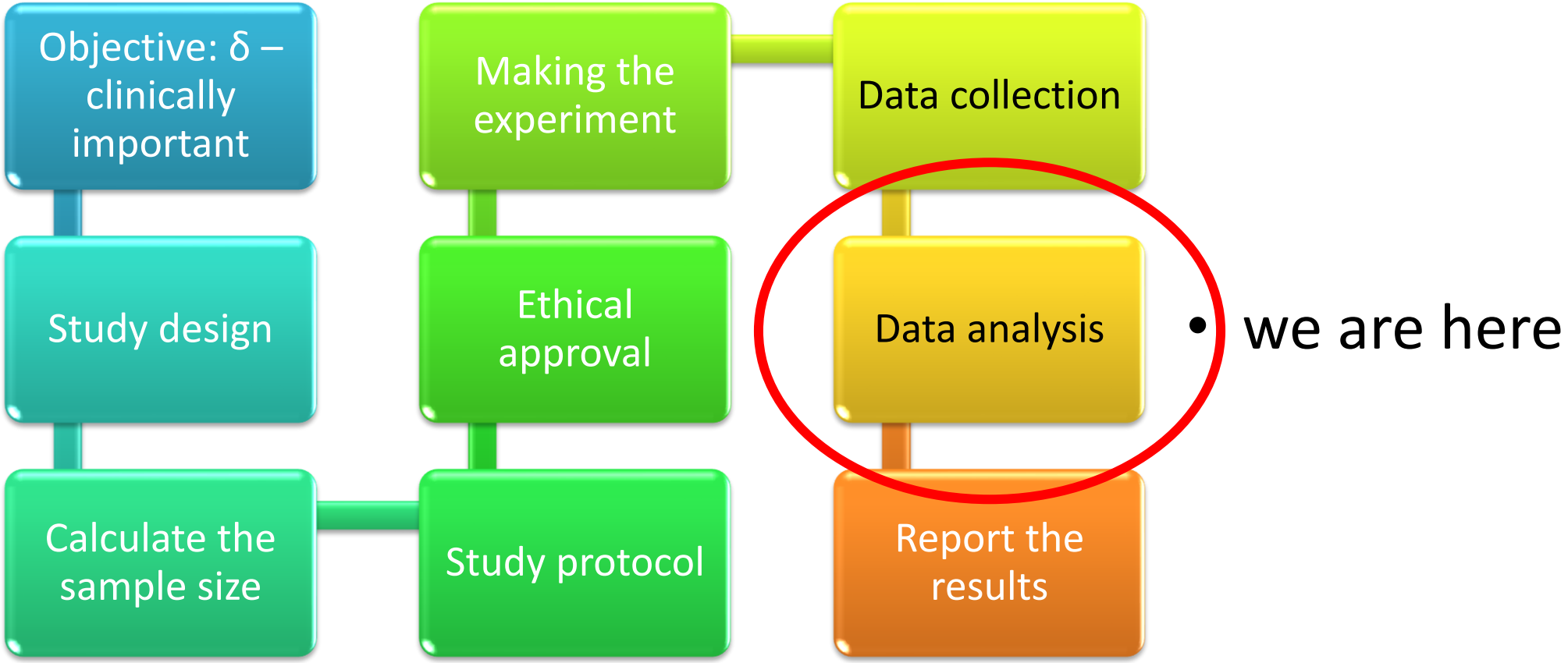
- use an application to analyze the data with statistical tests
  - SPSS
  - SAS
  - STATA
  - STATISTICA
  - R language (Jamovi)
  - etc.
- choose the conclusion for the medical question

Step 9: report the results



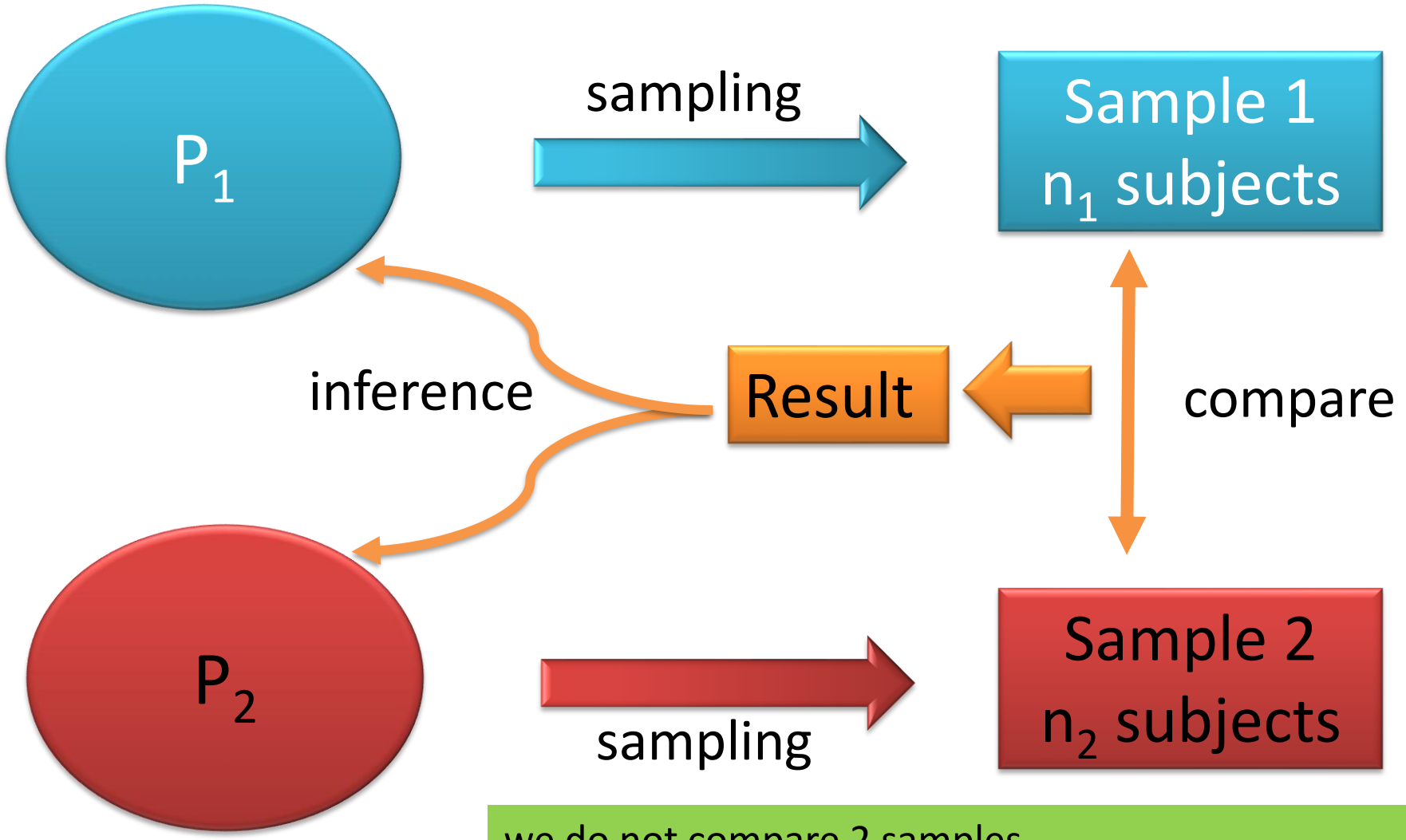
- report the results of the study in:
  - medical conferences
  - mass media
  - journals as original articles
  - for clinical trial in WHO International Clinical Trials Registry Platform

Compare samples



- we are here

the observed samples serve to appreciate what is happening in the populations from which they were extracted



we do not compare 2 samples
inference = compare all possible samples of n_1 respectively n_2 subjects that can be taken from the populations

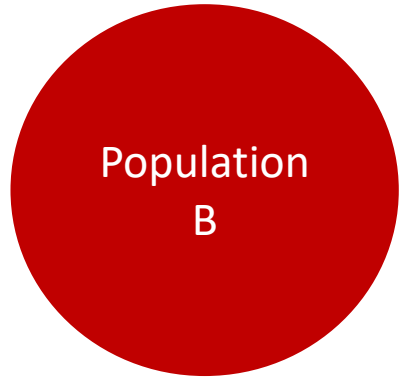
Compare two groups with confidence intervals method

Compare two groups with confidence intervals method

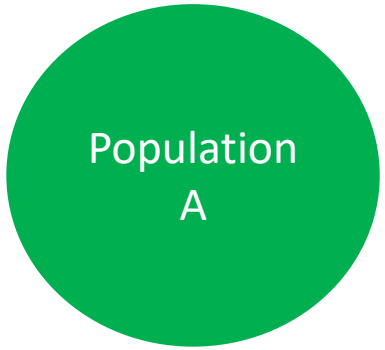
- subjects with hipercholesterolemia

Objectiv: To compare cholesterol

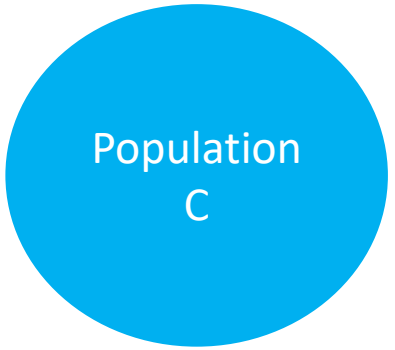
- diet A (mediteranean diet)
- diet B (paleo diet)
- diet C (high-protein diet).



diet B



diet A

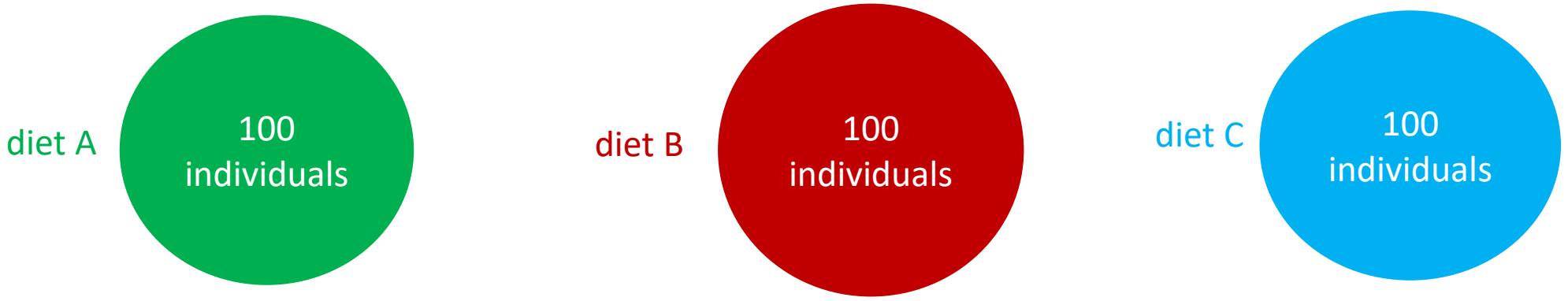


diet C

Compare two groups with confidence intervals method

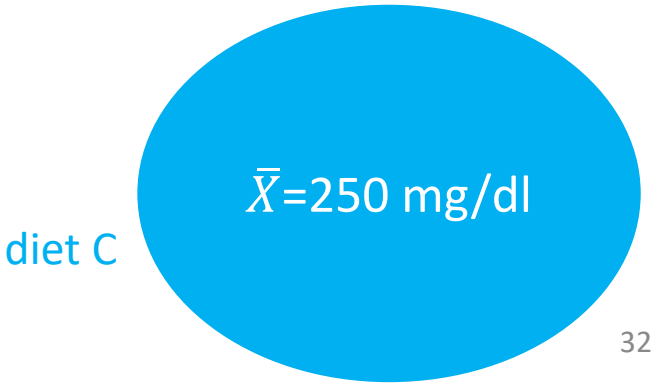
- Objective: To compare **cholesterol** in subjects with hypercholesterolemia who followed **diet A** (Mediterranean diet) versus **diet B** (Paleo diet) versus **diet C** (High-protein diet).

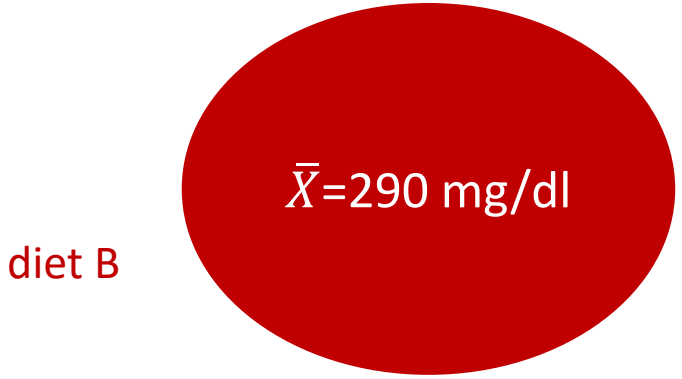
- Method:
- a difference by 10mg/dl we consider it clinically important.
- We design the study to detect a difference of the cholesterol by 10mg/dl or more between the diets in the population with a 95% probability
  - we calculate the needed sample size take into the study 300 subjects randomly selected

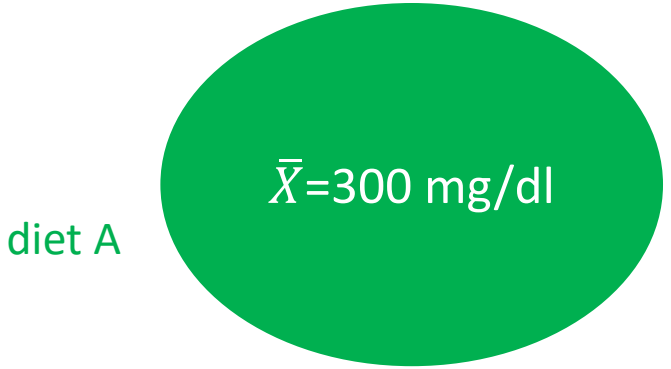


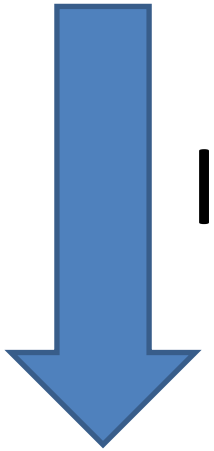
Findings on the samples







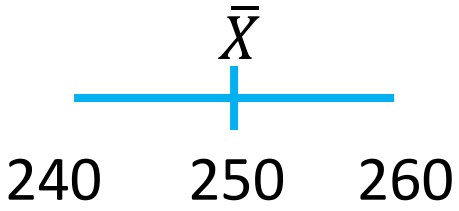


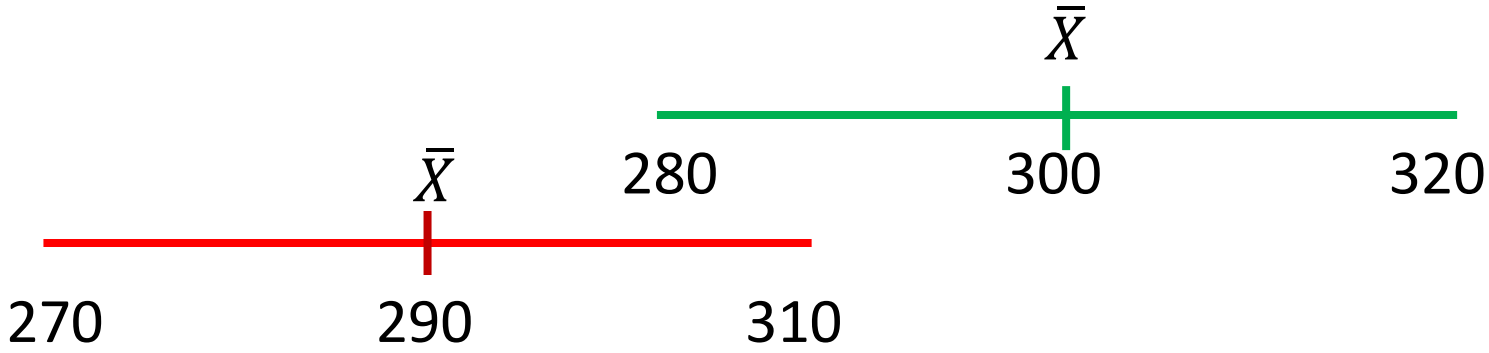


Inference to the population

we estimate 95% confidence intervals

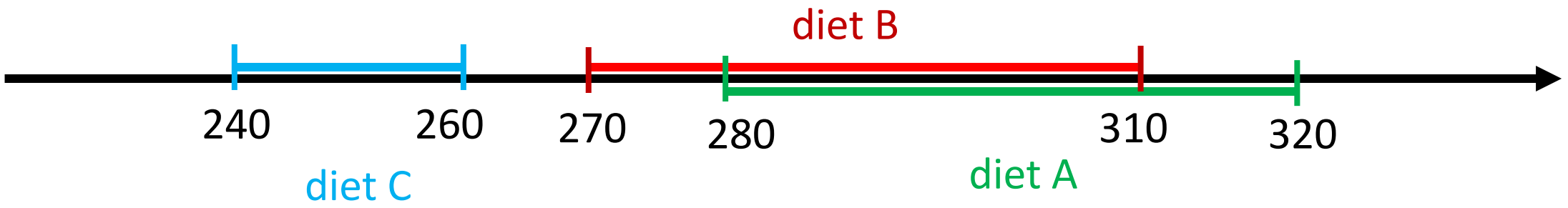
diet A: $\bar{X}$= 300mg/dl, 95%CI 280-320
diet B: $\bar{X}$= 290mg/dl, 95%CI 270-310
diet C: $\bar{X}$= 250mg/dl, 95%CI 240-260





Inference to the population

we estimate 95% confidence intervals



diet A: $\bar{X}$= 300mg/dl, 95%CI 280-320
diet B: $\bar{X}$= 290mg/dl, 95%CI 270-310
diet C: $\bar{X}$= 250mg/dl, 95%CI 240-260



- 95% CI 240-260

Interpretation: the true average cholesterol of population of subjects following diet C (high protein diet) will be somewhere between 240mg/dl and 260mg/dl with 5% error,

(we do not know where exactly in this interval will be.

Observation: if we want more precisse results we should take more subjects into the study).

- In the worst case average cholesterol can be 260mg/dl, in the best case 240mg/dl

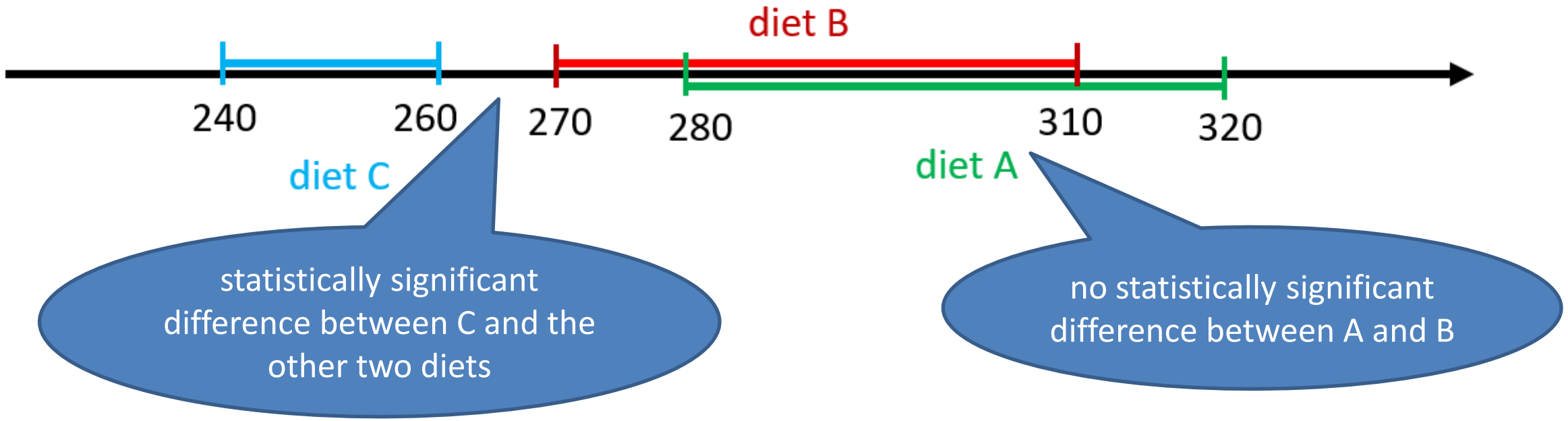


- 95% CI 270-310

Interpretation: the true average cholesterol of population following mediteranean diet will be somewhere between 270mg/dl and 310mg/dl with 5% error,

we do not know where exactly in this interval is.

- In the worst case it can be 310mg/dl, in the best case can be 270mg/dl



- <u>Conclusion 1:</u> this study demonstrate that subjects that follow diet C have low cholesterol than those who follow diet B (with 10mg/dl) and diet A (with 20mg/dl) (95% CI of diet C did not overlap the 95% CI of diet B and A. Any value in the 95% CI of diet C < any value in the 95% CI of diet B and A)

- <u>Conclusion 2:</u> we fail to demonstrate that there is a significant statistically difference between cholesterol in diet B compare with diet A (95% CIs of diet A and B overlap. Observation: The difference between diet A and B is less than 10mg/dl, if we want to know the difference we should repeat the study with more subjects to have a better precision)

Compare two groups arithmetic means with statistical tests

steps of a statistical test exemplified with t test for independent samples in case of unequal variances

What Are Statistical Tests?

Definition:

- a mathematical method used to analyze sample data and determine whether there is sufficient statistical evidence to support or reject a hypothesis about a population

Main purposes:

- Compare groups
- Investigate relationships between variables
- Identify differences
- Identify associations
- Evaluate the effectiveness of medical treatments

Comparing two means

- General:

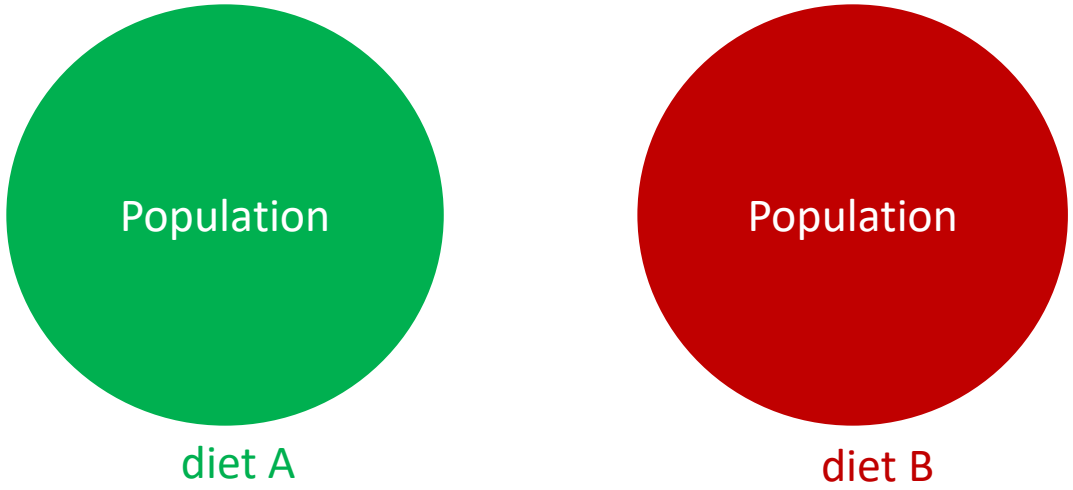
population P_1

population P_2

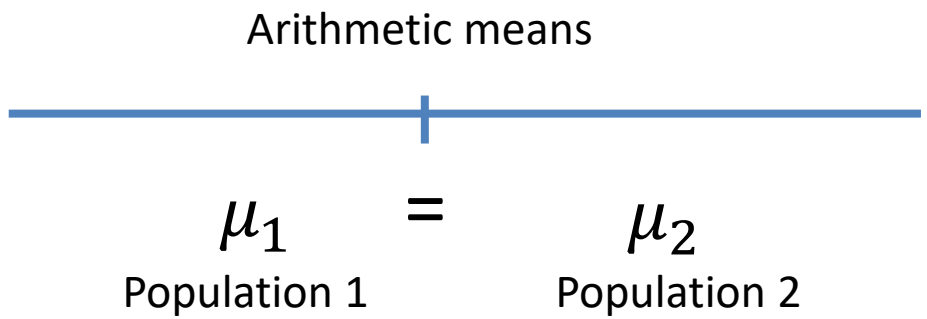
- Example:

patients with

- diet A
- diet B;

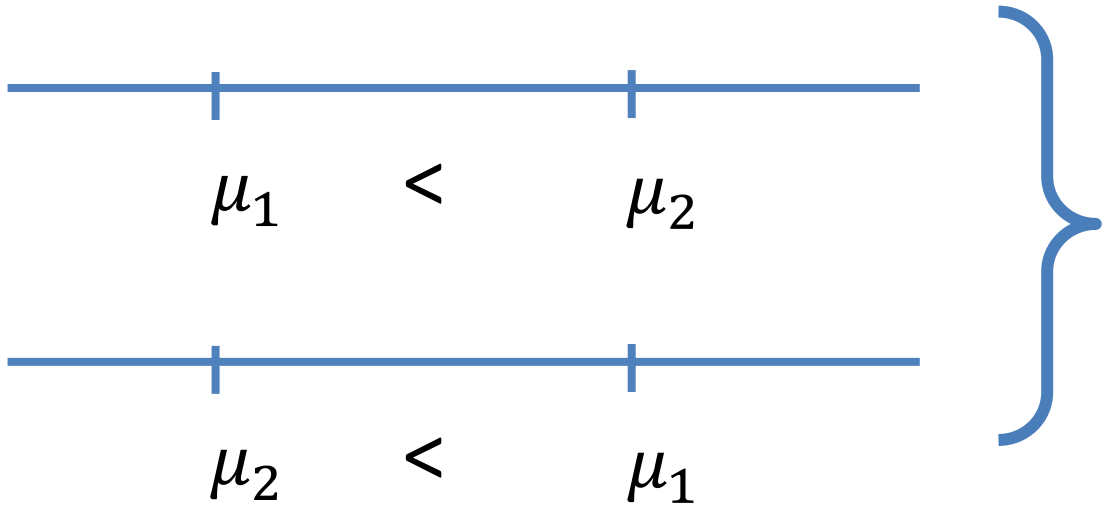


Possible results



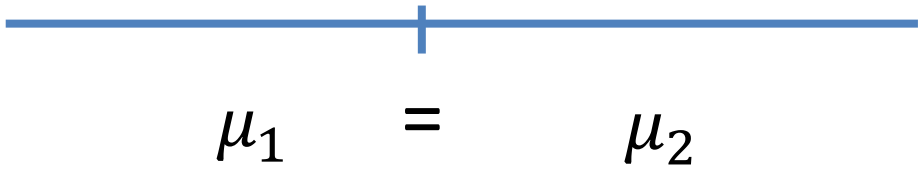
$\mu_1 = \mu_2$(or near equal)

OR



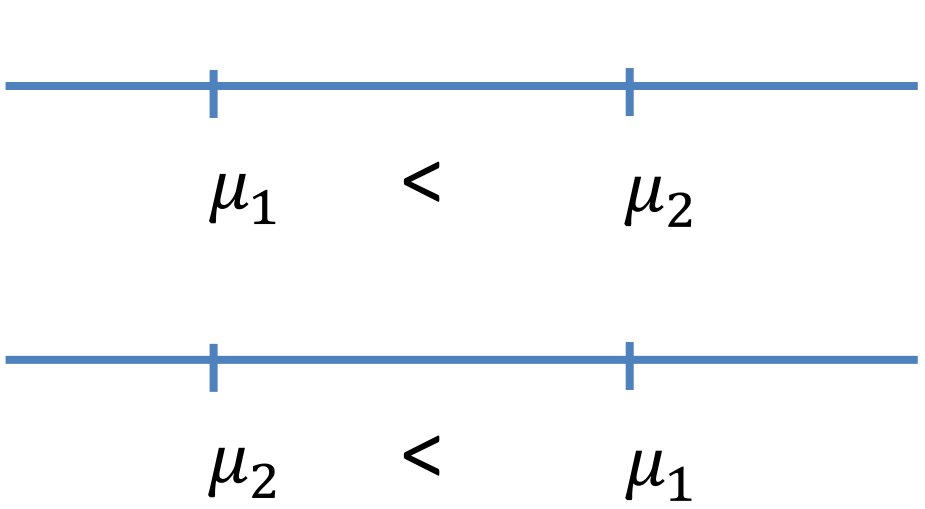
$\mu_1 \neq \mu_2$

Possible results



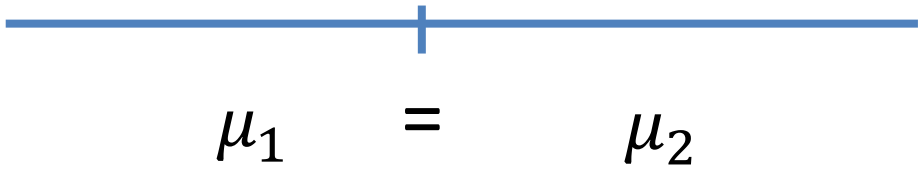
$\mu_1 = \mu_2$ H0: there is no difference

OR



$\mu_1 \neq \mu_2$ H1: there is difference

Possible results

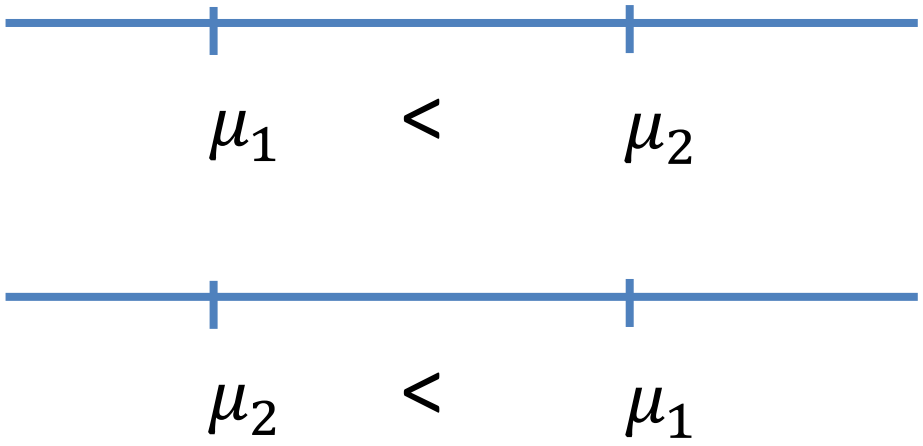


$\mu_1 = \mu_2$ H0: there is no difference

OR

Which one is true?

we will choose H0 or H1 based on a probability



$\mu_1 \neq \mu_2$ H1: there is difference

Statistical hypothesis testing

Involves two competing hypotheses:

1. Null hypothesis (H_0)
   - States that there is no difference, no effect, or no association
   - Represents the default assumption
2. Alternative hypothesis (H_1 or H_a)
   - States that a difference, effect, or association exists
   - Represents what the researcher wants to investigate

Example:

- Research question: Does a new medication reduce blood pressure?
  - H_0: The new medication does not reduce mean blood pressure compared with the control treatment ($\mu_1 \approx \mu_2$)
  - H_1: The new medication reduces mean blood pressure compared with the control treatment ($\mu_1 \neq \mu_2$)

Select samples, make the study, obtain the results on the samples

- General:

population P_1

population P_2

$\bar{X}_1$ = average for first sample,

$\bar{X}_2$ = average for second sample,

s_1^2 = variances for first sample

s_2^2 = variances for second sample

- Example:

patients with

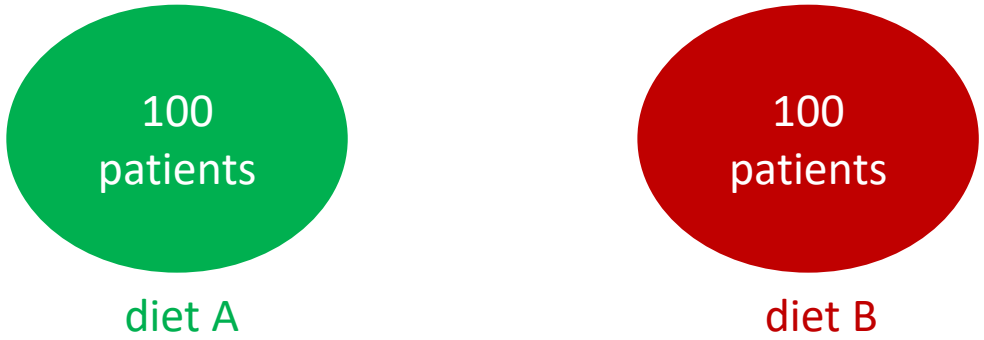
- diet A
- diet B;

= cholesterol $\bar{X}_1$ for first sample 300mg/dl,

= cholesterol $\bar{X}_2$ for second sample 290mg/dl,

= cholesterol s_1^2 for first sample 100

= cholesterol s_2^2 for second sample 100



Step 1 of a statistical test: Formulating Hypothesis

General:

- H0 - null hypothesis: at the population level, there is **no** statistical significant difference between group 1 average and group 2 average

- H1 - alternative hypothesis: at the population level, there is statistical significant difference between group 1 average and group 2 average

Example:

- H0 - null hypothesis: at the population level there is **no** statistically significant difference between the averages of cholesterol for patients with diet A versus diet B

- H1 - alternative hypothesis: at the population level there is statistically significant difference between the averages of cholesterol for patients with diet A versus diet B

Step 1 of a statistical test: Formulating Hypothesis

- H0 - null hypothesis: to demonstrate there is no statistically significant difference between the averages of cholesterol for patients with diet A versus diet B

"statistically significant" = suggest that is at the population level, so the term "population level" is redundant

- H1 - alternative hypothesis: there is no statistically significant difference between the averages of cholesterol for patients with diet A versus diet B

"statistically" – we estimate statistically the cholesterol average in the population of subjects with diet A with the cholesterol average from the sample with diet B

"significant" – we will choose a level of significance, i.e. a probability level for the confidence

Alternative=the other possible situation, when null hypothesis is not true

One-Tailed and Two-Tailed Tests

- Statistical hypothesis tests can be classified according to the direction of the alternative hypothesis.
- Two-tailed test:
  - Investigates whether a difference exists in either direction.
  - The alternative hypothesis uses the symbol ≠.
- One-tailed test:
  - Investigates a difference in a specific direction.
  - The alternative hypothesis uses the symbol greater than (>) or less than (<)

One-Tailed Test

Definition:

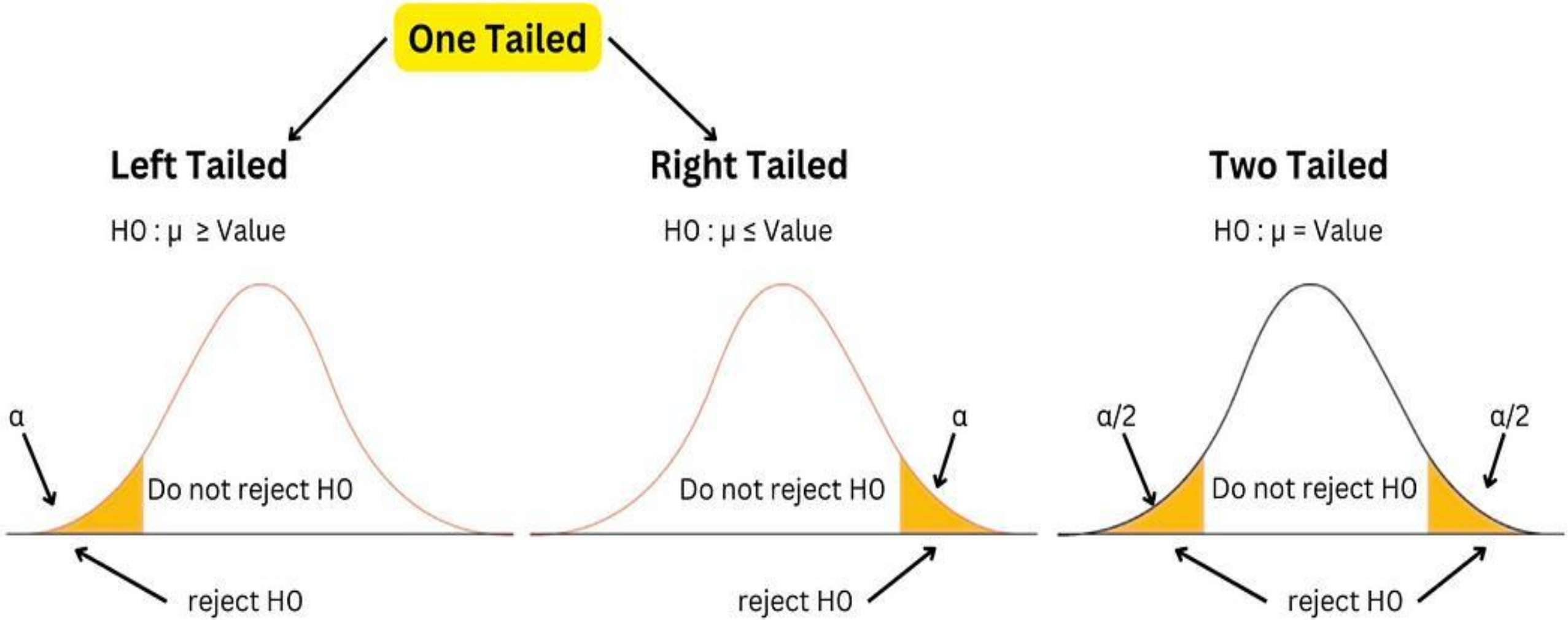
- A one-tailed test determines whether a parameter is significantly greater than or less than a specified value

Two possibilities:

- Upper-tailed test:
  - H_0: $(\mu_1 \approx \mu_2)$
  - H_1: $(\mu_1 < \mu_2)$
- Lower-tailed test:
  - H_0: $(\mu_1 \approx \mu_2)$
  - H_1: $(\mu_1 > \mu_2)$

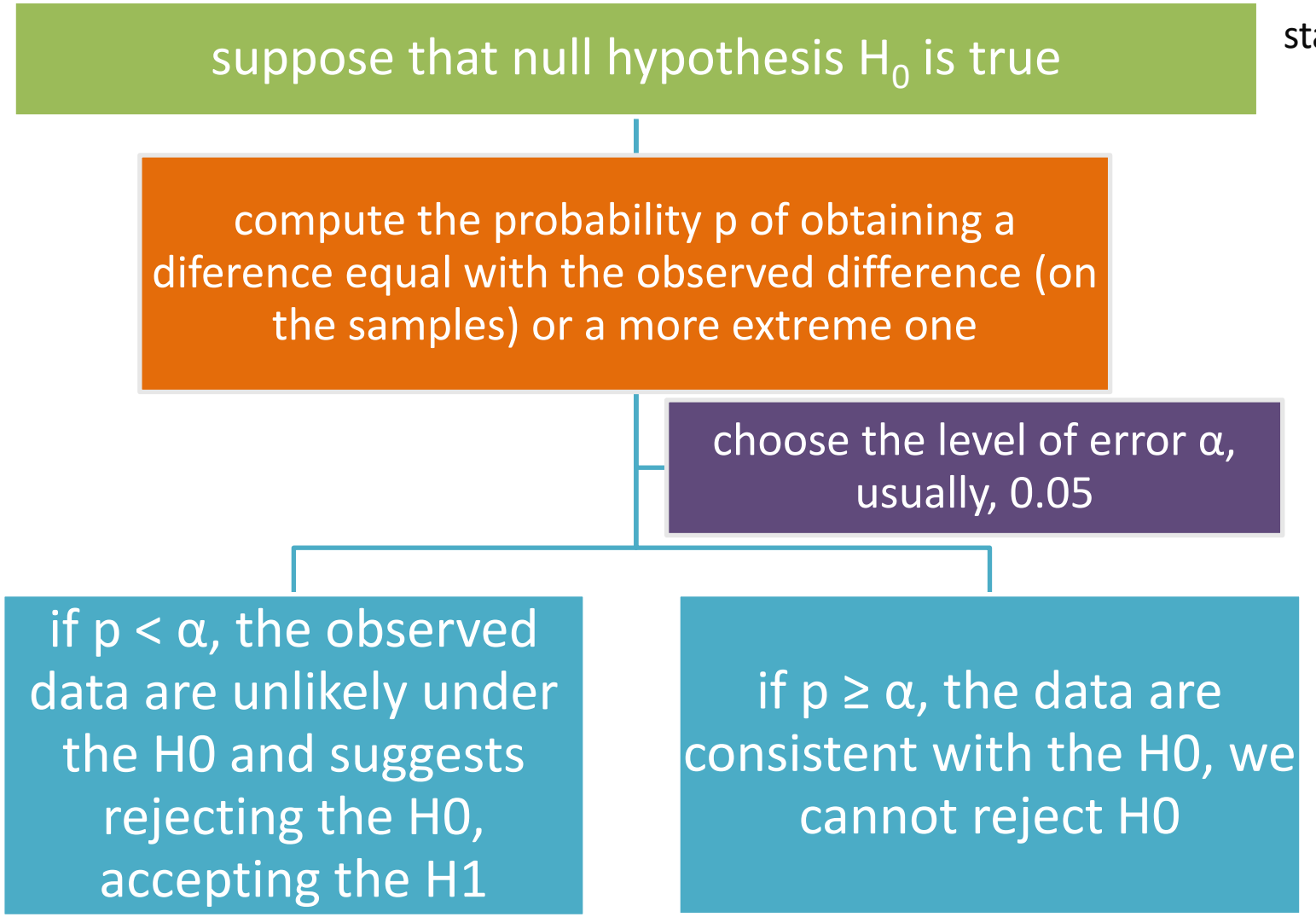
Example:

- A researcher investigates whether a new drug reduces mean systolic blood pressure.
  - H_0: there is no statistically significant reduction between the averages of systolic blood pressure in patients with new drug compare with standard drug
  - H_1: there is statistically significant reduction of the mean systolic blood pressure in patients with new drug compare with standard drug
- The rejection region is located in one tail of the distribution.


One Tailed
Left Tailed
H0 : μ ≥ Value
α
Do not reject H0
reject H0
Right Tailed
H0 : μ ≤ Value
Do not reject H0
α
reject H0
Two Tailed
H0 : μ = Value
α/2
Do not reject H0
α/2
reject H0

Method:

absurd assumption



starting assumption

Step 2 of a statistical test – Finding the proper statistical test to apply by answering to the questions:

General:

- What we compare?
  - means,
  - variances,
  - percentages
  - ….
- How is the distribution?
  - normal
  - non-normal,
  - binomial
  - …
- How are the samples?
  - independent
  - dependent
- How is the sample size?
  - small (<30)
  - large
- How are the variances?
  - equal
  - unequal
- How many samples we want to compare?
  - one,
  - two,
  - three
  - many

- Example:
  - normal distributions
    - we should check the distribution,
      - Result: stress is normally distributed in both groups,
  - unequal variances
    - we should check (not the subject of this lecture, see next lecture),
      - Result: variances are not equal,
  - two arithmetic means are compared,
  - groups are independent
    - between groups - within the sample there are no subjects that are extremely similar to the others (repeated observations on the same subjects, measurements on the left hand and on the right hand of the same subject, …)
    - not the case - between the subjects in the same group
- we will choose:
  - **Student t-test for independent samples in the case of unequal variances**

Step 3. Select the level of significance

General:

- α - level of significance
  - probability of incorrectly rejecting the null hypothesis when it is actually true (Type I error)
- **Usually, α is choose to be 0.05**
  - **(corresponding to Z_{α}=1.96)**
- It does NOT mean:
  - There is a 5% probability that the null hypothesis is true
  - There is a 95% probability that the alternative hypothesis is true
  - The treatment is 95% effective

Example:

- we choose α = 0.05

Step 4. Determine the value the test parameter must attain to be declared significant

General:

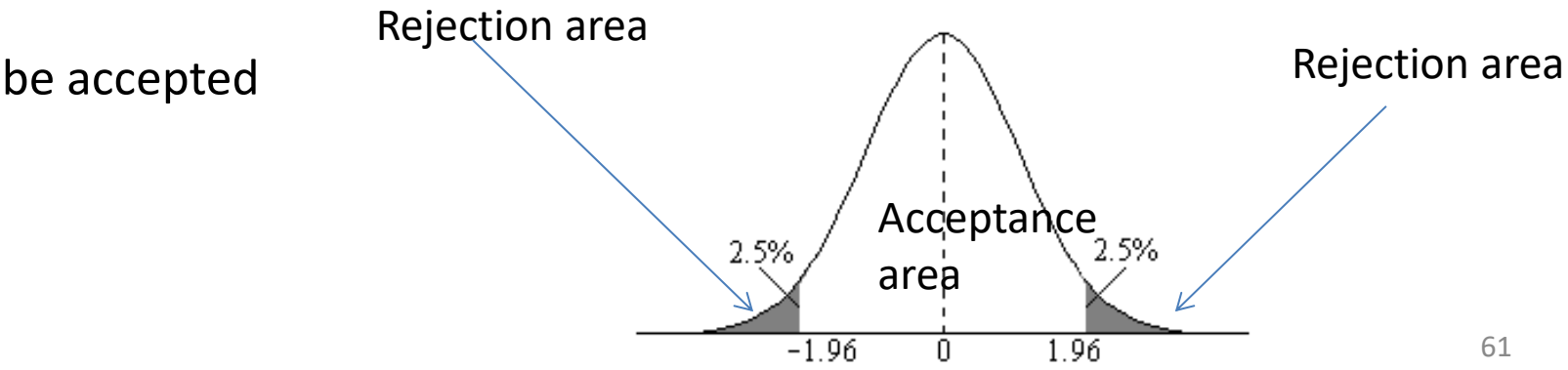
- **critical value**
  - ex. for t-test t_α, for Z test Z_α
- Rejection area – critical interval
  - interval where H_0 should be rejected
  - the interval depend on the distribution of the test, whether the test is one-tailed or two-tailed, and sometimes degrees of freedom
  - ex. $(-\infty, -Z_\alpha) \cup (Z_\alpha, \infty)$
- Acceptance interval:
  - interval where H_0 should be accepted
  - ex. $[-Z_\alpha, Z_\alpha]$
  - for α=0.05, Z_α=1.96

Example:

t test – Student distribution – **critical value** is different for each degree of freedom

- t_α = 1.97 = **critical value for 100+100-2 degree of freedom**
- rejection interval $(-\infty, -t_\alpha) \cup (t_\alpha, \infty)$ = $(-\infty, -1.97) \cup (1.97, \infty)$
- acceptance interval: $[-t_\alpha, t_\alpha]$=[-1.97,1.97]

= an interval around 0, if we found a difference in this interval it means is a small difference and H_0 will be accepted (there is no difference)



Step 5: computing the formula for test parameter

General:

- any statistical test have a test parameter
- test two means, then $\bar{X}_1 - \bar{X}_2$ should be in the formula, divided by standard errors

if parameter of the test is small, then the difference between $\bar{X}_1$ and $\bar{X}_2$ is small, near 0.

Example:

- the parameter of the test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

difference between averages.

material A: $\bar{X}_1$ = 300MPa, s = 100MPa
material B: $\bar{X}_2$ = 290MPa, s = 100MPa

$$t = \frac{300-290}{\sqrt{\frac{100^2}{100}+\frac{100^2}{100}}} = \frac{10}{\sqrt{200}} = \frac{10}{14.14} = 0.71,$$

p=0.48

p can be computed only by computers – have a very complicated formula

p-value interpretation

$p < \alpha$ (small p-value)

- the observed data are unlikely under the null hypothesis and suggests **rejecting the H0, accepting the H1**
- suggests that the observed result is statistically significant and unlikely to have occurred by chance. **We reject H0, we are in favor of H1**
- the result is considered **statistically significant**
- the obtain difference does not automatically mean is a medical relevant result

- $p \geq \alpha$
- the data are consistent with the null hypothesis, **we cannot reject H0**
- suggests that the observed result is not statistically significant and is likely to have occurred by chance. **We cannot reject H0**
- the result is considered **statistically not significant**
- the obtain difference does not automatically mean that is no medical relevant result
- We should not say that we have accepted or proved the null hypothesis

Example:

$-1.97 \leq t = 0.71 \leq 1.97$

we fail to reject null hypothesis,

the average cholesterol in people who have diet A did not differ statistically significant from the average cholesterol in people who have diet B

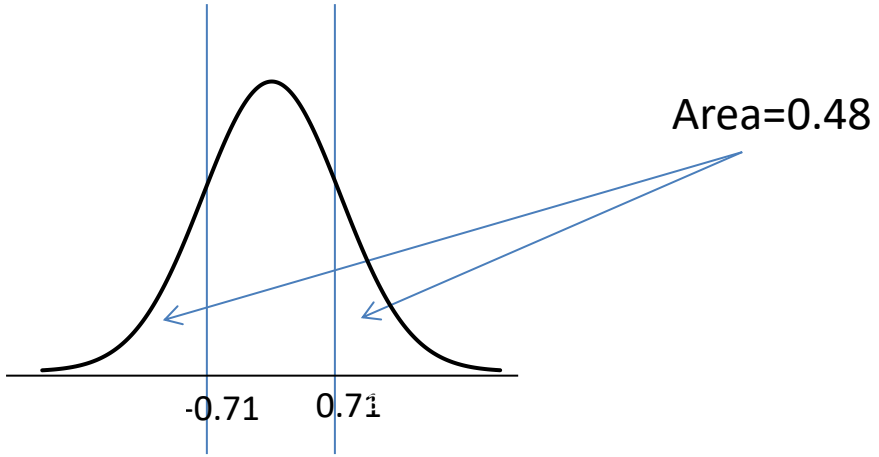
or

Area under the curve for $t = 0.71$ is 0.48, $p = 0.48 > 0.05$ we fail to reject null hypothesis,

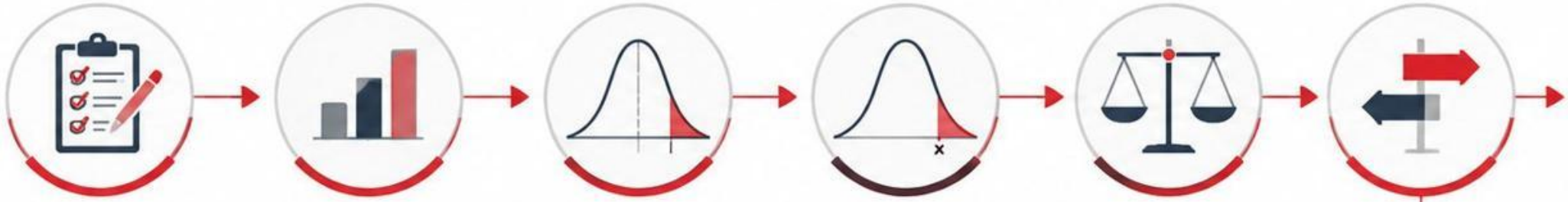
the average cholesterol in people who have diet A did not differ statistically significant from the average cholesterol in people who have diet B

~~We accept the null hypothesis~~

We fail to reject the null hypothesis



UNDERSTANDING HYPOTHESIS TESTING AND THE P-VALUE



1

STATE THE HYPOTHESES

Define the null hypothesis (H_0) and the alternative hypothesis (H_1).

2

CHOOSE THE SIGNIFICANCE LEVEL (α)

Set the threshold for decision making (commonly 0.05).

3

COLLECT DATA & CALCULATE THE TEST STATISTIC

Use your sample data to compute the test statistic.

4

CALCULATE THE P-VALUE

The p-value is the probability of observing data as extreme as (or more extreme than) what you observed, assuming H_0 is true.

5

COMPARE THE P-VALUE TO α

Determine whether the p-value is less than or equal to α.

6

IF P-VALUE ≤ α

REJECT THE NULL HYPOTHESIS (H_0)

There is sufficient evidence to support the alternative hypothesis (H_1).

IF P-VALUE > α

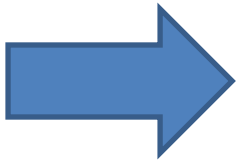
FAIL TO REJECT THE NULL HYPOTHESIS (H_0)

There is not sufficient evidence to support the alternative hypothesis (H_1).

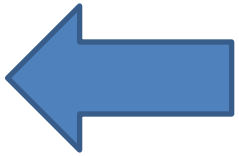
KEY TAKEAWAY

The p-value helps you assess how strong the evidence is against the null hypothesis. A smaller p-value provides stronger evidence to reject the null hypothesis.

Errors in Hypothesis Testing

Statistical decisions are made using sample data  may be incorrect

- types of errors:
  - Type I error (α): Rejecting a true null hypothesis
  - Type II error (β): Failing to reject a false null hypothesis

- These errors are important  they can influence clinical decisions

Type I Error (False Positive)

Definition:

- the null hypothesis is rejected even though it is actually true

i.e.

- The study concludes that a difference exists when, in reality, it does not

Example:

- A study concludes that a new medication reduces blood pressure more effectively than the standard treatment, although there is actually no difference

Probability:

- P(Type I error)=α
- If $\alpha = 0.05$, the Type I error probability is 5% under the null hypothesis.

Type II Error (False Negative)

Definition:

- fail to reject the null hypothesis even though it is actually false

i.e.

- The study fails to detect a real difference

Example:

- A new medication genuinely reduces blood pressure more effectively than the standard treatment, but the study fails to detect the difference

Probability:

- P(Type II error)=β
- The probability of Type II error depends on several factors, including sample size, effect size and variability.

Comparison of Type I and Type II Errors

Remember:

- Type I error: False alarm.
- Type II error: Missed detection.

Characteristic	Type I error	Type II error
Symbol	α	β
Meaning	False positive	False negative
Decision	Reject true H_o	Fail to reject false H_o
Main problem	Detecting a nonexistent effect	Missing a real effect
Example	Drug appears effective when it is not	Effective drug appears ineffective

The Four Possible Outcomes of Hypothesis Testing

Because statistical tests use samples rather than entire populations, there is always a possibility of making an incorrect decision

Reality	Reject H_0	Fail to reject H_0
H_0 is true	Type I error	Correct decision
H_0 is false	Correct decision	Type II error

What Is Statistical Power?

Definition:

- the probability that a statistical test correctly rejects the null hypothesis when the alternative hypothesis is true

i.e.

- It is the probability of detecting a real effect when one actually exists

Formula:

- $\text{Power}=1-\beta$

Where:

β = Probability of Type II error

Example:

- If $\beta = 0.20$:
- $\text{Power}=1-0.20=0.80$
- Power = 80%

Factors Affecting Statistical Power

1. Sample size
   - Larger samples generally increase power
2. Effect size
   - Larger differences between groups are easier to detect
3. Variability
   - Lower variability generally increases power
4. Significance level (α)
   - Increasing α (ex. 10% instead of 5%) generally increases power

Why Is Statistical Power Important?

- Determine the appropriate sample size
- Reduce the risk of missing clinically important effects
- Design reliable study
- Avoid conducting underpowered studies
- Improve the quality of medical evidence

Example:

- A study with only 15 patients may fail to detect a real treatment effect because the sample is too small
- A larger, appropriately designed study may have sufficient power to detect that effect

Common target:

- Medical research studies often aim for minimum **80%** statistical power.

Thank you !!!