STATISTICS & ML WITH R

Statistical inference & hypothesis testing

2024

M. Chiara Mimmi & Luisa M. Mimmi

WORKSHOP SCHEDULE

- 4 days
  - 1. Intro to R and data analysis
  - 2. Statistical inference & hypothesis testing
  - 3. Modeling correlation and regression
  - 4. Machine Learning; MetaboAnalyst; Power Analysis

- Each day will include:
  - Frontal class (MORNING)
  - Practical training with R about the topics discussed in the morning. (AFTERNOON)

DAY 2 – LECTURE OUTLINE

- Purpose and foundations of inferential statistics
  - Probability and random variables
  - Meaningful probability distributions
  - Sampling distributions and Central Limit Theorem

- Getting to know the “language” of hypothesis testing
  - The null and alternative hypothesis
  - The probability of error? (α or “significance level")
  - The p-value probability and tests interpretation
  - Confidence Intervals
  - Types of errors (Type 1 and Type 2)
  - Effective vs statistical significance

- Hypothesis tests **examples**
  - Comparing sample mean to a hypothesized population mean (Z test & t test)
  - Comparing two independent sample means (t test)
  - Comparing sample means from 3 or more samples (ANOVA)

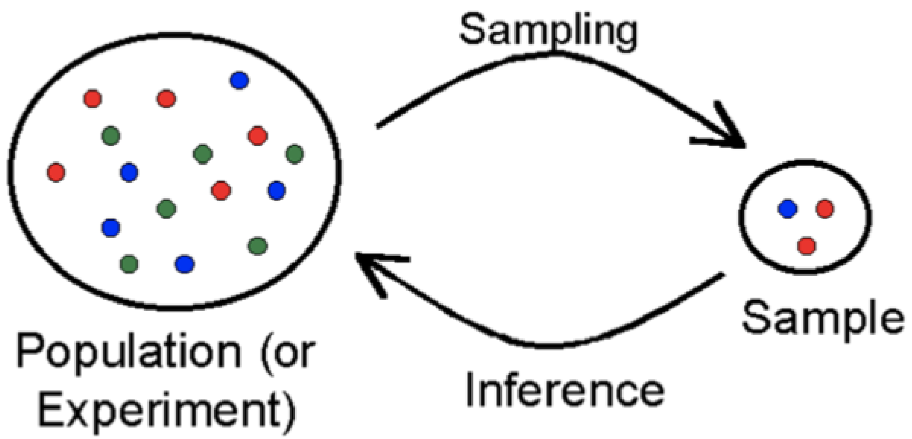
- A closer look at testing assumptions (with **examples**)
  - Testing two groups that are NOT independent
  - Testing if the data are not normally distributed: non-parametric tests
  - Testing samples without homogeneous variance of observations

Inferential Statistics: population and samples

Gathering all data is not always possible due to barriers like time, accessibility, or cost.

Therefore, we often gather information from a smaller subset of the population: a **SAMPLE**.

- **POPULATION** = the universe of all possible observations we are interested in
- **SAMPLE** = a subset of the population from which information is actually collected
- **INFERENTIAL STATISTICS** = a collection of methods for using sample data to make conclusions about a population



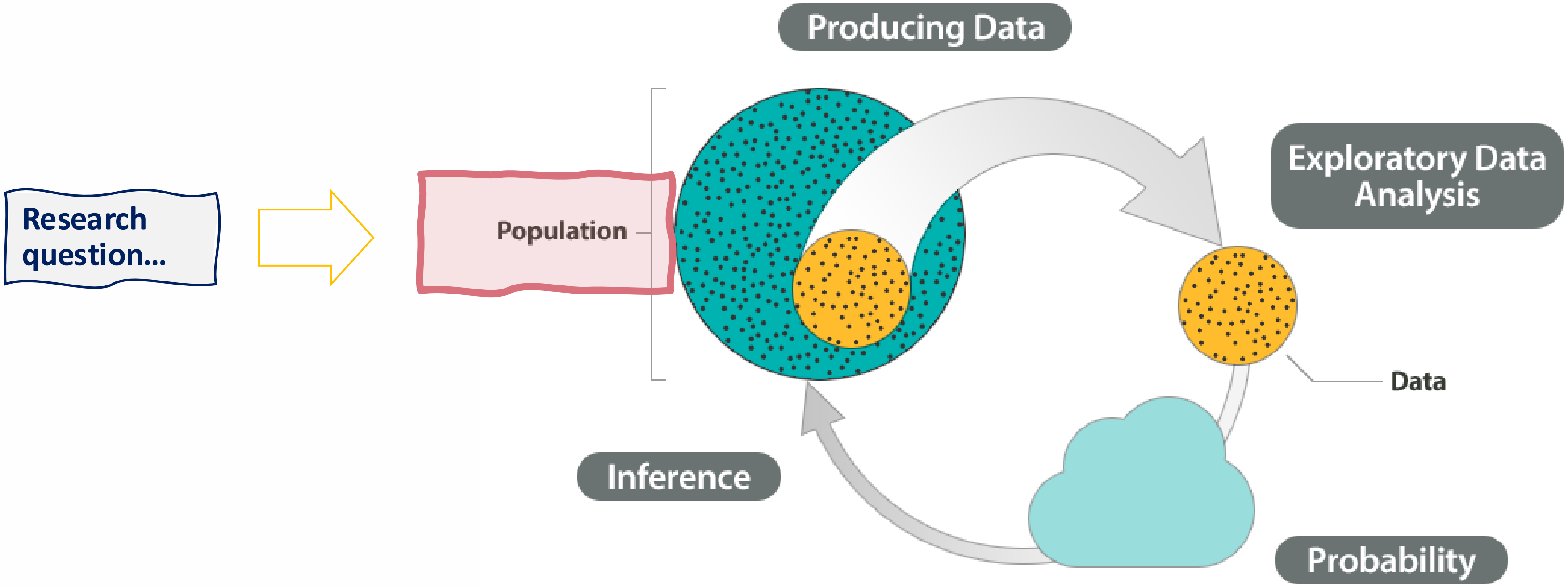
Inferential statistics workflow: defining a population

Decide which population(s) is (are) important

- We may need 2 of them—control and treated groups—according to the experimental design
- **Statistical populations** are something the investigator defines

Decide which attributes of the population(s) need to be measured

- **Variable(s)** to measure
- (… not all relevant information are MEASURABLE)
- often the literature can provide information about the general population we are studying



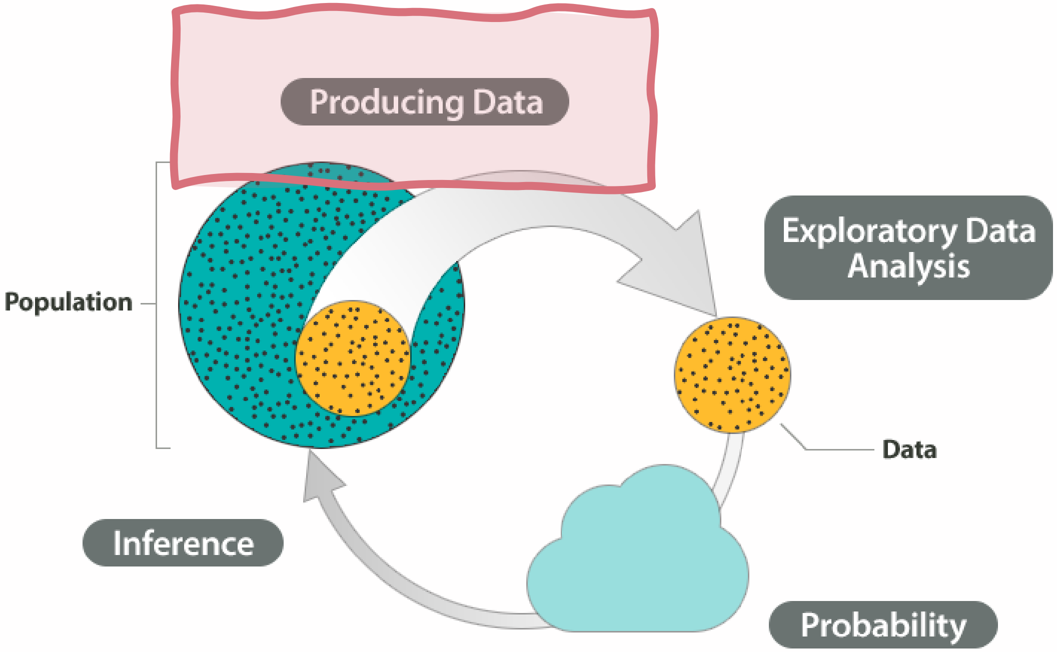
Source image: https://stats.libretexts.org/Courses/Lumen_Learning/

Inferential statistics workflow: producing data

- **Pick a representative sample**
  - A sample is a subset of a population that has been selected to be representative (or unbiased), i.e. it REFLECTS the characteristics of the entire population
    - ideally, a **random sample,** where each individual has a known, non-zero probability of being selected into the sample

- **Estimate the population parameter**
  - From a representative sample, we can calculate a point estimate of the population parameter (unknown)

- **Estimate uncertainty**
  - Sampling error: any point estimate from the sample will be imperfect (it won't exactly match the true population value)



Source image: https://stats.libretexts.org/Courses/Lumen_Learning/

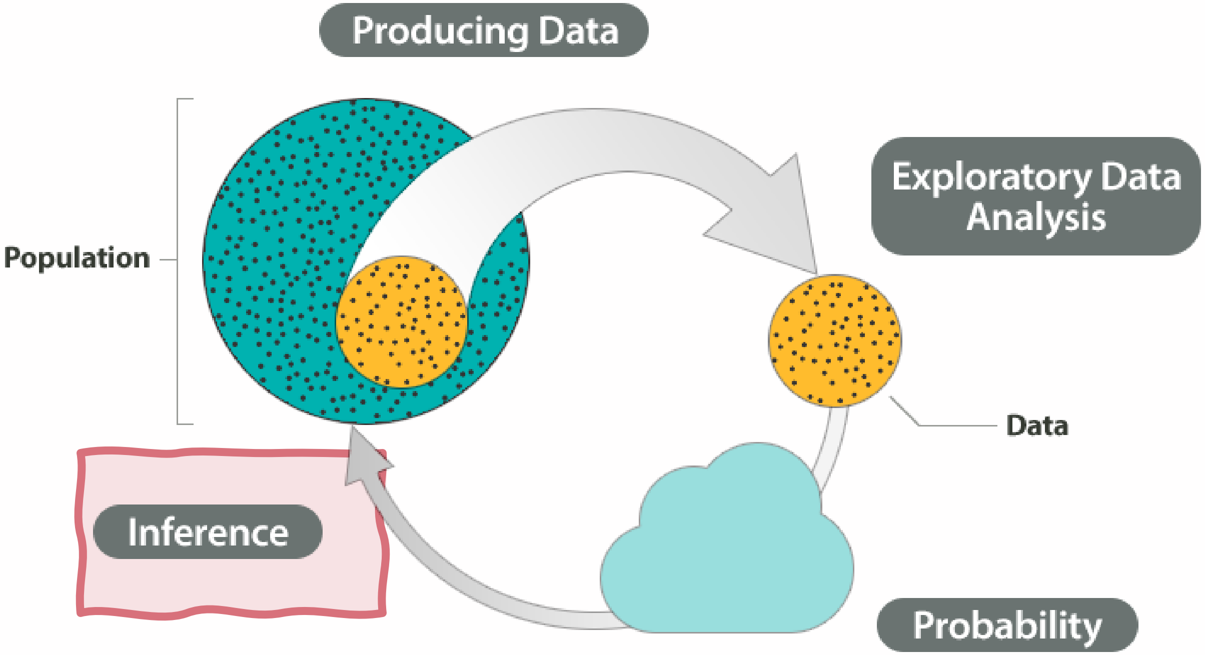
Inferential statistics workflow: making inference based on sample

Formulate hypotheses to test through experiments:

- features of a population
- interaction of dependent and independent variables
- degree of uncertainty and error

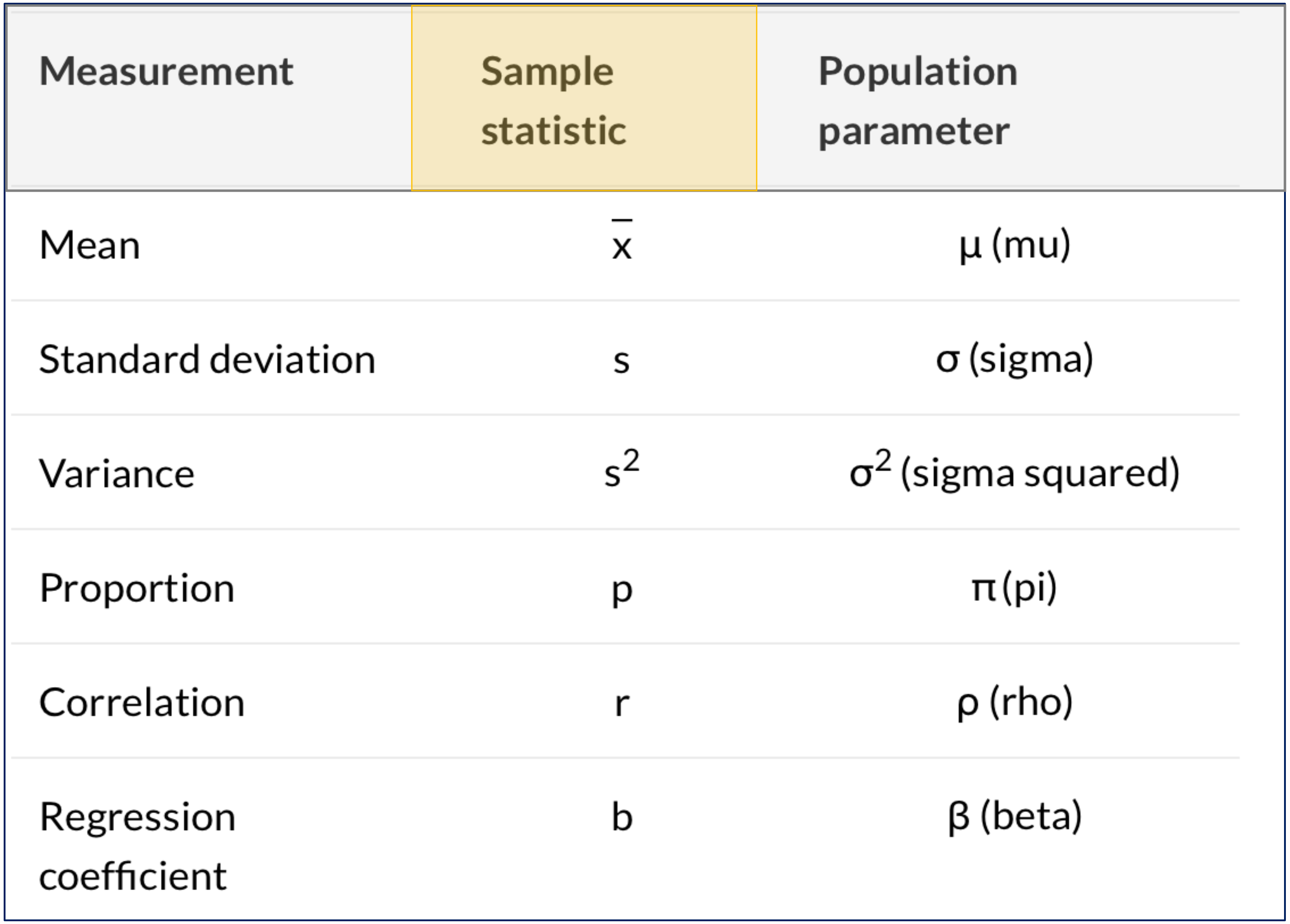
Generalize findings to population of interest, assessing:

- construct validity
- validity of causal relationship
- generalizability



Source image: https://stats.libretexts.org/Courses/Lumen_Learning/

Population versus sample: terms



Measurement	Sample statistic	Population parameter
Mean	$\bar{x}$	μ (mu)
Standard deviation	s	σ (sigma)
Variance	s^2	σ^2 (sigma squared)
Proportion	p	π (pi)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta)

Source: https://www.statology.org/statistic-vs-parameter/

Probability
and
random variables

Probability & statistical models (in *frequentist* approach)

- Data are **observed values** («realizations») of **random variables**.
- The **probability distribution** of these random variables can be used to reason about *properties* (‘parameters’) of the unobserved universe (**inference**).
- Statistical models are probability distributions for observable data constructed to enable inferences to be drawn or decisions made from data
- From **observation** to theoretical **probability**:
  - **Absolute frequency (af)** $= n_{favorable\ events}$ **Relative frequency (rf)** $= \frac{n_{favorable\ events}}{N_{observations}}$
- Relative frequencies also serve as “empirical probabilities”, [between 0 and 1].
  - **Relative frequency** $\approx$ **Probability of an event (p)** $= \frac{n_{favorable\ events}}{N_{possible\ events}}$
- The relationship between Absolute Frequency and Probability is further reinforced by the **Law of Large Numbers**

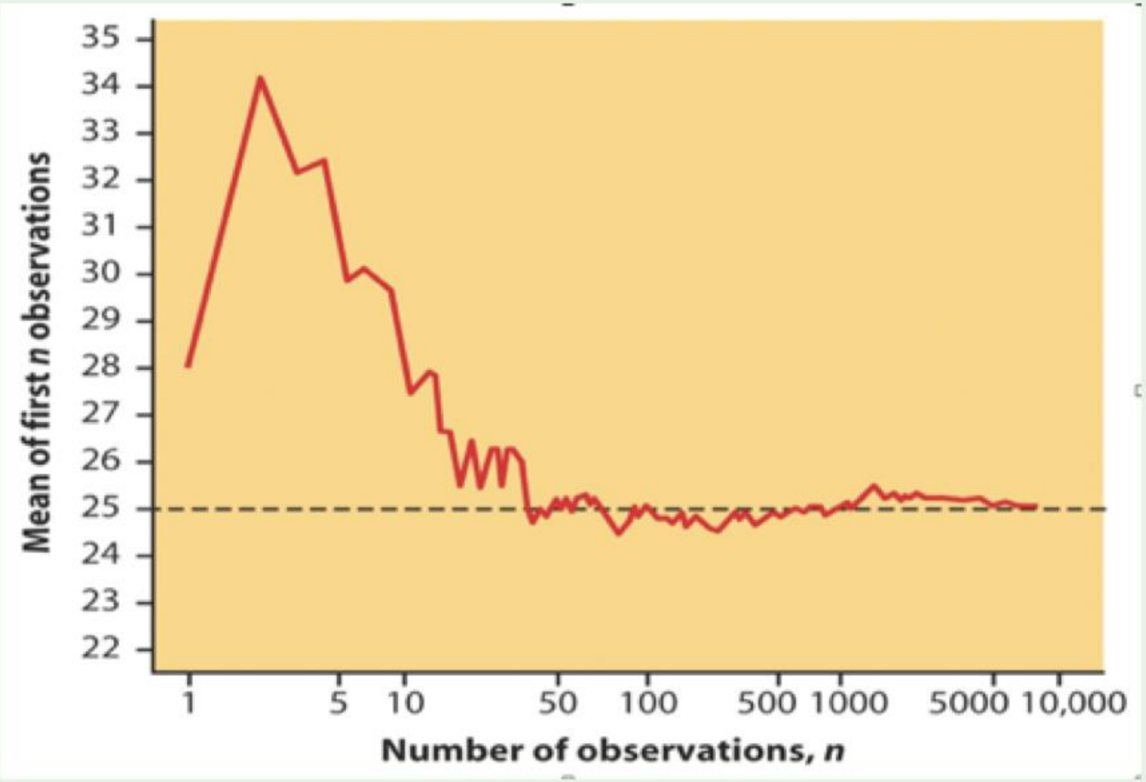
$$\textbf{Relationship (af - probability)} = \mathrm{p} = \lim_{big\ N} \boldsymbol{rf}$$

(By conducting a larger number of trials or observations)
we can derive probabilities from absolute frequency!

The Law of Large Numbers allows us to use probability to predict absolute frequency and viceversa

Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the sample mean of the observed values $\bar{x}$ gets closer and closer to the mean μ of the population.

Example: How sample means approach the population mean ($\mu = 25$).

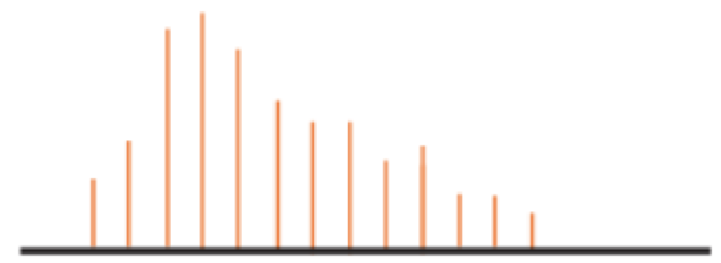


Random Variables and probability distributions

Discrete Random Variables	Continuous Random Variables
• can take a finite number of distinct values • e.g. # of children per family	• can take an infinite (or impossible to count) number of possible values • e.g. weight of a person

2 types of probability distributions

PROBABILITY DISTRIBUTION

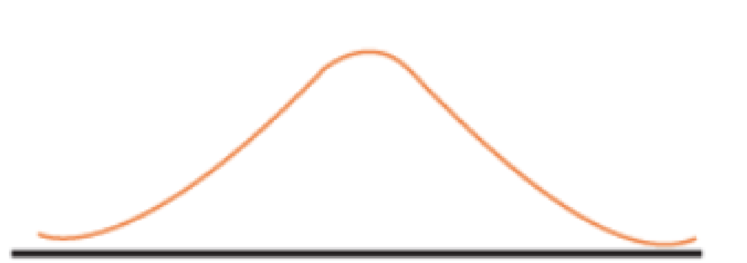


The **probability distribution** describes the likelihood of the variable to take a possible individual value.

It is called: ***Probability Mass Function (PMF)***

$$\boldsymbol{P_X(x_k)} = P(X = x_k),\ for\ k = 1,2,3\ldots$$

PROBABILITY DENSITY FUNCTION



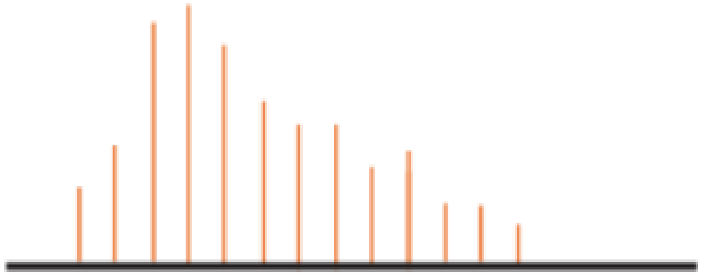
(Since $P(X = x_k) = 0\,,\ for\ all\ x_k$) the **probability distribution** describes the likelihood of the variable to fall within an interval of values.

It is called: ***Probability Density Function (PDF)***

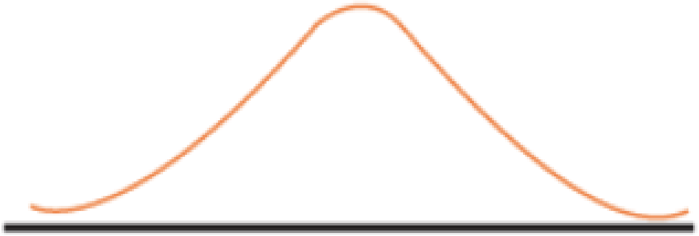
$$\boldsymbol{P[a \leq X \leq b]} = \int_a^b f_X(x)dx$$

Random Variables and probability distributions

Discrete Random Variables



Continuous Random Variables



SOME “FAMOUS” DISTRIBUTIONS

- Bernoulli Random Variable X~Bernoulli(p)
  - applicable to random experiments that can only have 1 trial and only 2 possible results -like above (e.g. pass/fail, head/tail)
- Binomial Random Variable X~Bin(n,p)
  - applicable to Bernoulli experiments (2 possible results), but here we can have 1 or more trials (e.g. probability that # patients will experience side effects from a new medication)
- Poisson Random Variable X~Poisson(λ)
  - used to show how many times an event will occur within a given time period – knowing events occur independently and with a constant mean rate (e.g. # of meteorites striking Earth in a year)

- Normal Random Variable X~(μ,σ2)
  - more on its feature later… (e.g. birthweight of newborn babies, shoe sizes, diastolic blood pressure, …)
- Exponential Random Variable X~Exp(λ)
  - refers to the process in which the event happens at a constant average rate independently and continuously (e.g. , the amount of time until an earthquake occurs, amount of time a car battery lasts)

Meaningful probability distributions for inference

The normal distribution

The normal distribution (special case of continuous distr.)

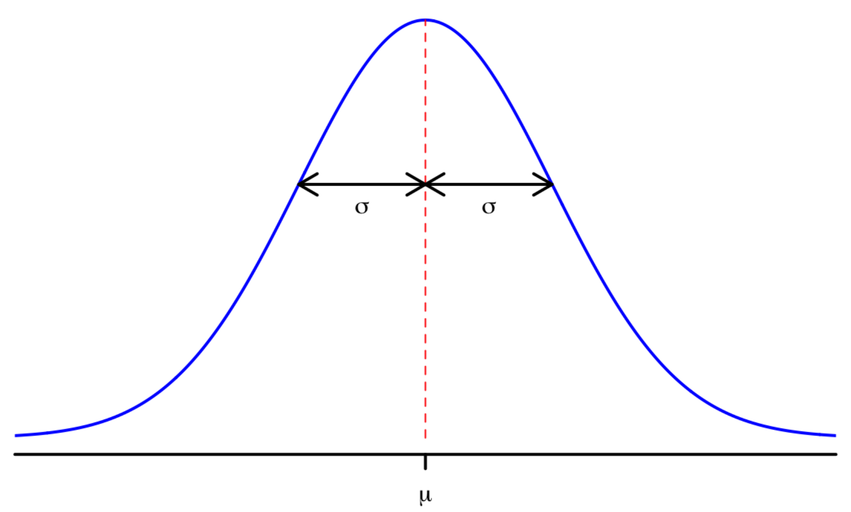
Normally distributed Random Variable

X~ Normal (μ,σ^2)

with

Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

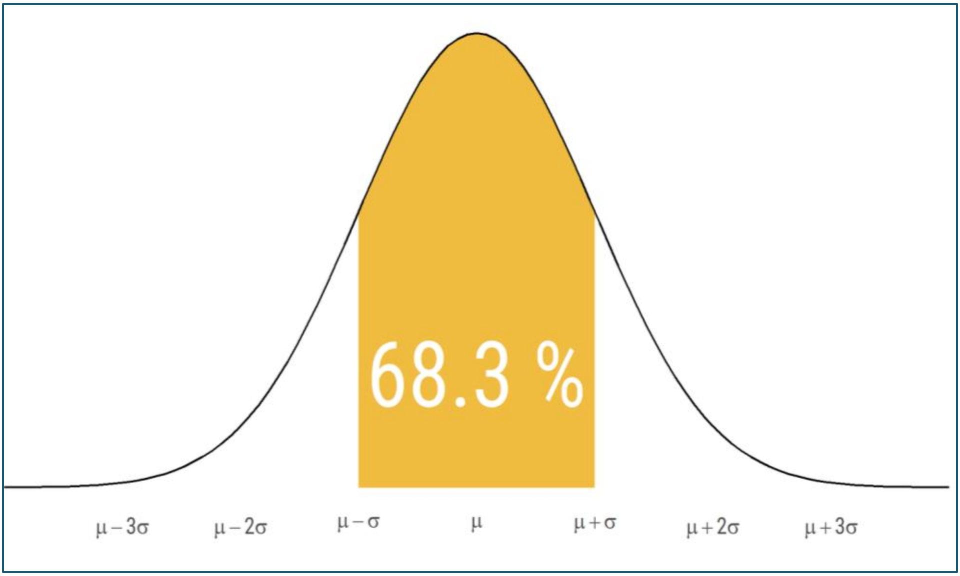


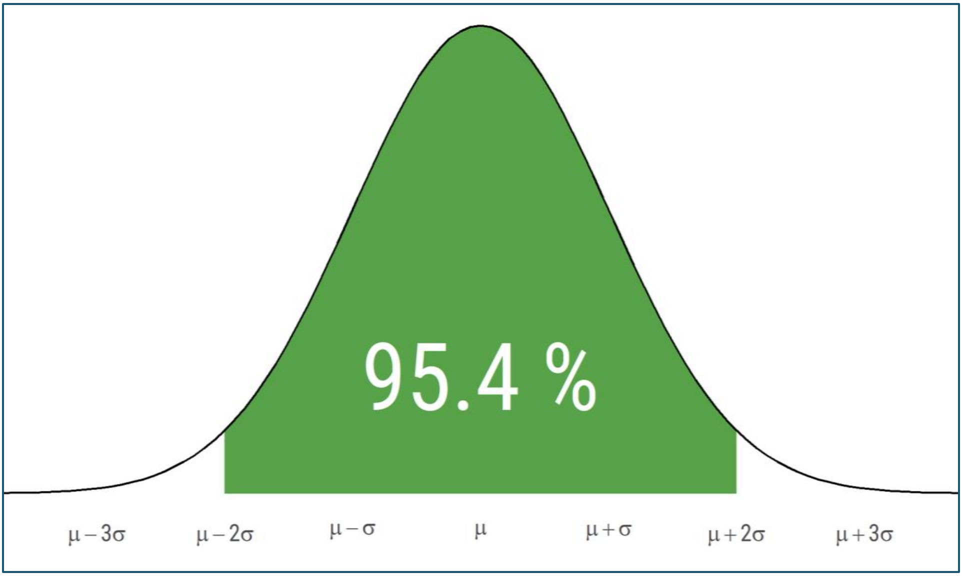
- The **normal distribution** («*bell curve*» or «*Gaussian curve*») is extremely important in statistics because:
  - it provides an excellent summary of an empirical distribution providing 2 parameters: the **mean (μ)** and **standard deviation (σ)**
  - many things in real life (and experimental science) can be approximated with it (e.g. blood pressure, height, weight, age children get disease, standardized test scores, etc.)
  - **Measurement error** with scientific instruments is typically modelled as a normal distribution with expectation $\mu=0$. The more precise the instrument, the lower the value of the variance σ^2.

Features of the normal distribution

- Normal distributions are symmetric around the mean
- The mean, median and mode of a normal distribution are equal
- Normal distributions are denser in the center and less dense in the tails
- The area under the normal curve is = 1
- 68% of the area of a normal distribution is within 1 standard deviation of the mean
- Approximately 95% of the area of a normal distribution is within 2 standard deviation of the man

X~ Normal (μ,σ^2)

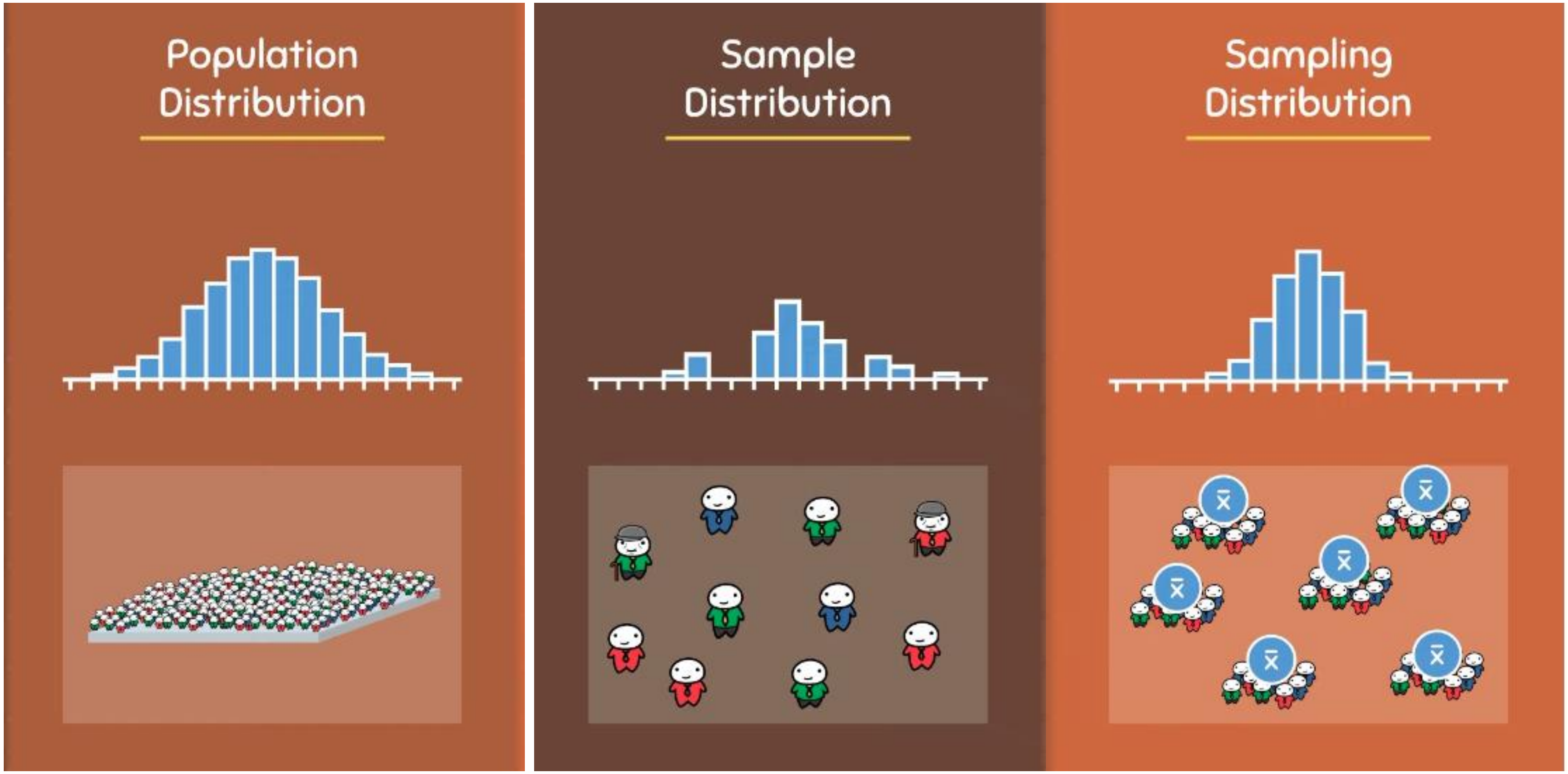




Sampling distributions and the CLT

From population, to sample, to sampling distribution

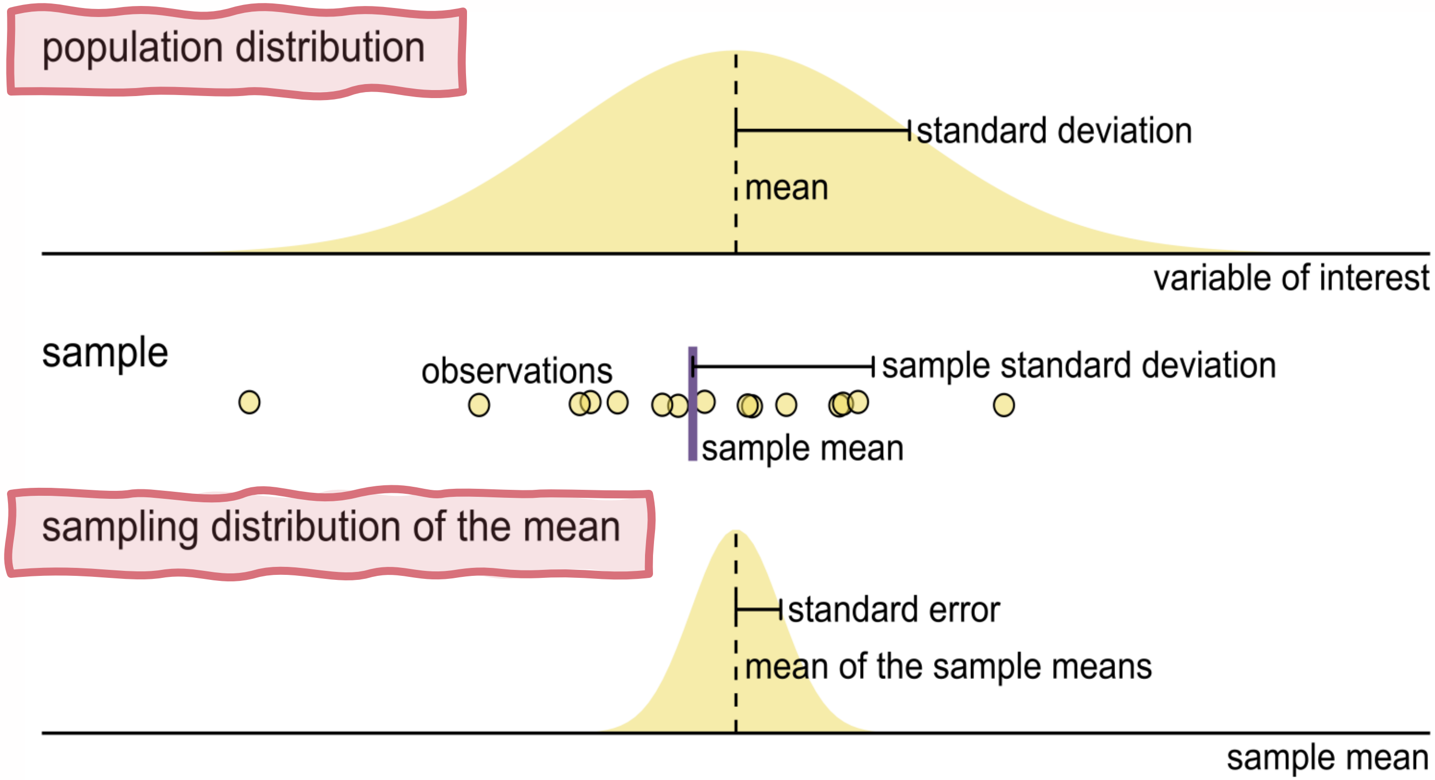
- A **sampling distribution** shows the frequency distribution of n sample means (e.g. $\overline{x}_1, \overline{x}_2, \overline{x}_3, \overline{x}_4 \ldots \overline{x}_n$), assuming that we take several random samples of the population
  - with large n it approximates a normal



Source image: https://www.youtube.com/watch?v=7S7j75d3GM4 https://lulliter.github.io/R4stats/

Dealing with the uncertainty of point estimates

- *The premise of (**frequentist) statistics** is that we can approximate a hypothetical population parameter by studying (or simulating) several samples from from it*
- A **point estimate** (e.g. $\overline{x}$) will always have a degree of **uncertainty:**
  - sampling error = the variability or 'noise' that comes with the process of taking repeated samples from a population of interest (i.e. each sample will be a little different)
  - standard error = a quantitative measure of sampling error variation (the standard deviation of the estimate's sampling distribution)



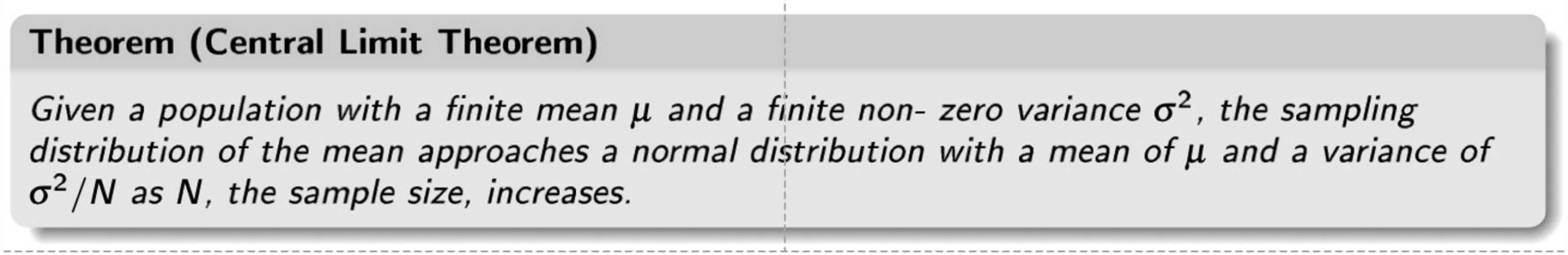
Source image: https://vizdata.org/slides/16/16-uncertainty-I.html#/

Central Limit Theorem and Sampling Distributions

PROBLEM: *How often do we know what the population of values looks like?*

- (i.e. what if the observations in our sample are not normally distributed?)

SOLUTION: the **Central Limit Theorem** provides a bridge between (a) the nice properties of the normal distribution and (b) the fact that the distribution of individual elements of many samples are **not** normally distributed



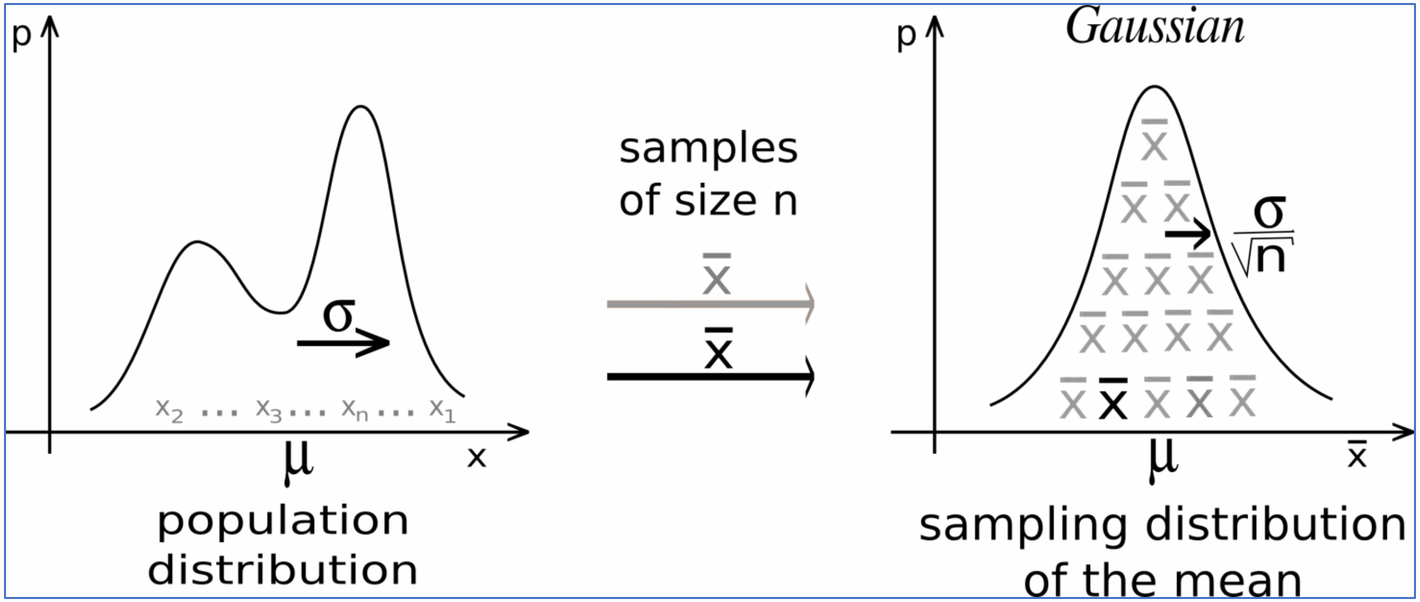
Theorem (Central Limit Theorem)

Given a population with a finite mean μ and a finite non- zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

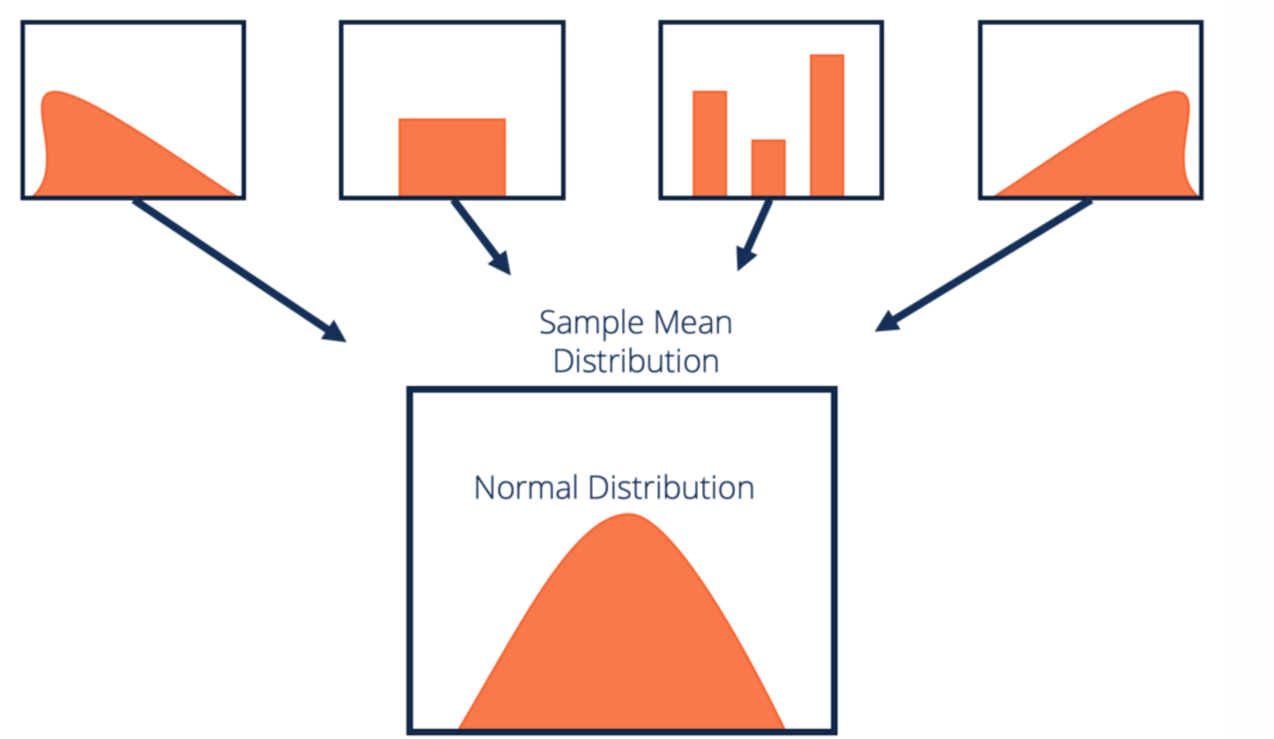
CLT: fundamental for inferences about a population based on a sample

CLT: regardless of the shape of the population distribution, if we take *several random* samples from a population and look at all their mean: $\bar{X} = \bar{x}_1, \bar{x}_2, \ldots \bar{x}_n$:

- by the **Central Limit Theorem**, the distribution of sample means $\bar{X}$ (a "**sampling distribution"**) will have a **normal** (or near normal) shape with mean $E(\bar{X}) = \mu$ (= the population's) and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \text{SE}$ (this is called Standard Error, < than population's)
  - IF $\mathbf{n}$ **is large enough** and
  - IF **samples are taken at random**


p
σ
$x_2 \ldots x_3 \ldots x_n \ldots x_1$
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
p
Gaussian
$\frac{\sigma}{\sqrt{n}}$
μ
$\bar{x}$
sampling distribution of the mean

Important implication of the CLT



Source: https://corporatefinanceinstitute.com/resources/data-science/central-limit-theorem/

Even if the **objective population** has an unknown or non-normal distribution, the mean of all possible casual samples with n sufficiently large (n>30) will have a normal distribution

→ The CLT makes **parametric test** on sample means (or its linear combination) applicable even if the assumption of population normality is violated, provided that the sample size is n>30.

DAY 2 – LECTURE OUTLINE

- Purpose and foundations of inferential statistics
  - Population and samples
  - Probability and random variables &
  - Meaningful probability distributions
  - Sampling distributions and Central Limit Theorem

- Getting to know the “language” of hypothesis testing
  - The null and alternative hypothesis
  - The probability of error? (α or “significance level")
  - The p-value probability and tests interpretation
  - Types of errors (Type 1 and Type 2)
  - Confidence Intervals
  - Effective vs statistical significance

- Hypothesis tests **examples**
  - Comparing sample mean to a hypothesized population mean (Z test & t test)
  - Comparing two independent sample means (t test)
  - Comparing sample means from 3 or more samples (ANOVA)

- A closer look at testing assumptions (with **examples**)
  - Testing two groups that are NOT independent
  - Testing if the data are not normally distributed: non-parametric tests
  - Testing samples without homogeneous variance of observations

Getting to know the “language” of hypothesis testing

(classical approach)

Statistical Hypotheses

- **Any claim made about one or more populations of interest constitutes a statistical hypothesis**
  - These hypotheses usually involve **population parameters**, the nature of the population, the relation between the populations, and so on
- For example, we may hypothesize that:
  - The **mean** of a population, μ, is 2
  - Two populations have the same **variance**
  - A population is **normally distributed**, etc.
- **Procedures leading to either the acceptance or rejection of statistical hypotheses are called statistical tests**
  - The number obtained from the sample to estimate the population parameter is the **point estimate**

Hypothesis testing steps

1. State the hypotheses (the null hypothesis and an alternative hypothesis)
2. Formulate an analysis plan (e.g. the significance level is 0.05, the test method one-sample z-test)
3. Analyze sample data
4. Interpret result and make decision

What are the Null and Alternative hypotheses?

Null Hypothesis H_0

- H_0 is the hypothesis that a sample data statistic occurs purely from chance
  - e.g. there is no difference between the mean pulse rate for people doing physical exercise and the normal pulse rate
- Must contain condition of equality $=, \leq$,or $\geq$
- Test the Null Hypothesis directly: reject H_0 or fail to reject H_0

Alternative Hypothesis H_1 or H_a

- H_1 is the hypothesis that a sample data statistic is influenced by some non-random cause
  - e.g. the mean pulse rate for persons doing the physical exercise is higher than the normal
- Must be true if H_0 is false (corresponding to $=, \leq$,or $\geq$ conditions)
- `opposite' of Null Hypothesis

What is the “significance level”, α?

There is always a certain probability of error: that $\boldsymbol{H_0}$ is rejected even though it is actually true.

This probability of this error (Type I error) is called the **significance level** or **α.**

- Usually, a significance level is set at 5% or 1% (the error you can tolerate). For example, a significance level of **0.05** signifies a 5% risk of deciding that an effect exists (reject $\boldsymbol{H_0}$) when it does not exist (= FALSE POSITIVE).

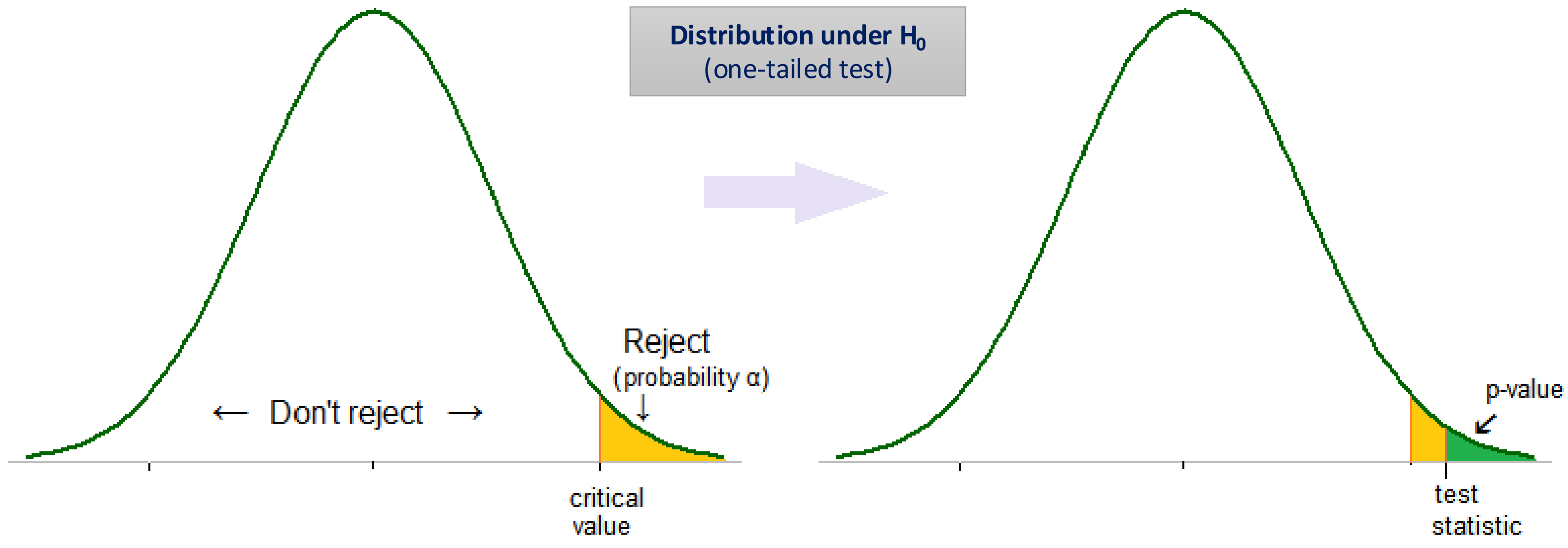
  e.g. To test the hypothesis (there is no difference between the mean pulse rate for people doing physical exercise and the normal pulse rate)

  1) fix the significance level at 5%
  2) measurement of pulse rate conducted over **n** persons +/- physical exercise
  3) calculate the t-statistics for the sample
  4) calculate the p-value associated to the found t

Lower significance levels require stronger sample evidence to be able to reject the null hypothesis

What is the p-value?

- The p-value is **the probability to obtain our test statistic (or a more extreme value) if the null hypothesis were true**
  - i.e. the p-value shows how strongly your sample data contradict the null hypothesis
- Conventionally, we use p-values in conjunction with α to determine whether our data favor the null or alternative hypothesis:
  - $p < 0.05$ means we 'reject the null hypothesis' / the 'effect is statistically significant at 5% level'



Lower p-value represents stronger evidence against the null hypothesis

Another way to think of tests' accuracy and precision

- **Accuracy**

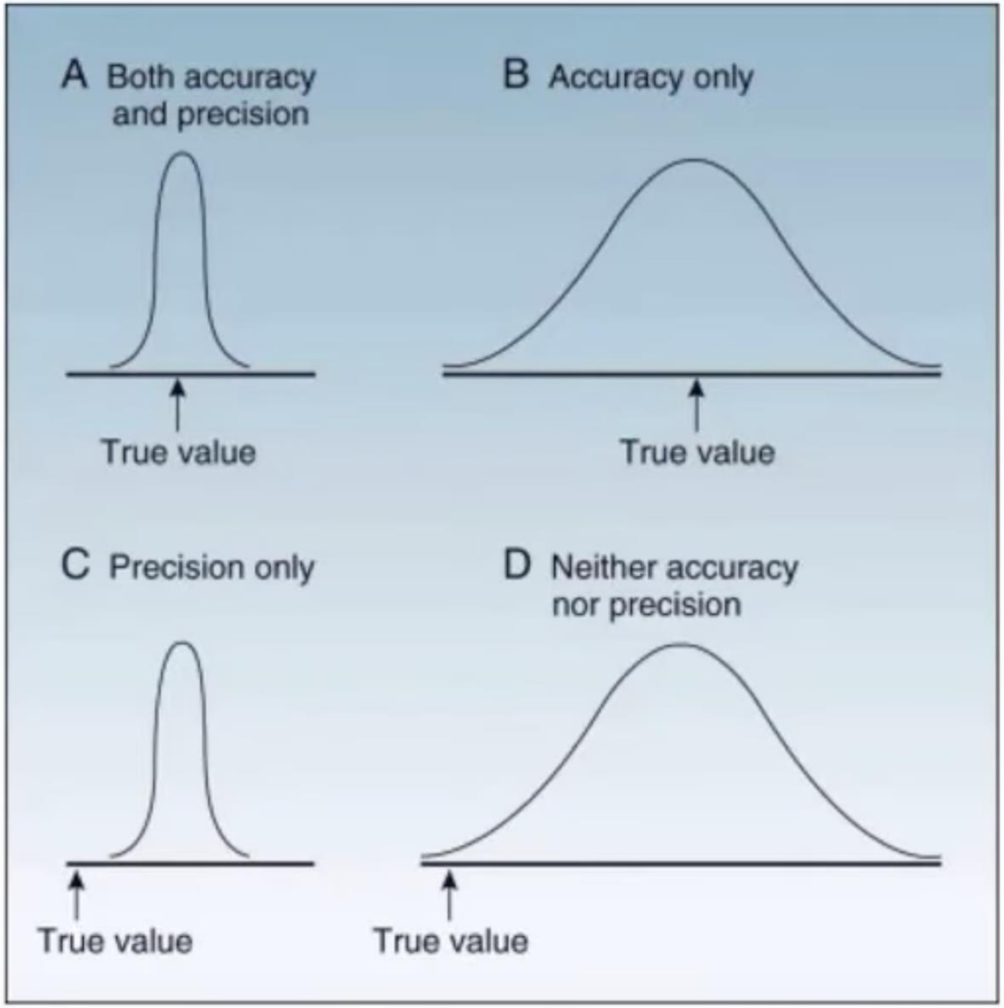
«**accurate tests**» capture what you want them to measure in an unbiased way.

- E.g. if your lab measurement gives you a 3.2 kg for an object that actually weights 10 kg , it is NOT ACCURATE

- **Precision**

«**precise tests**» give reproducible, or reliable results (although *not necessarily accurate*)

- E.g. … if you weight your object 5 times, and you get 3.2 kg each time, your measurement is VERY PRECISE



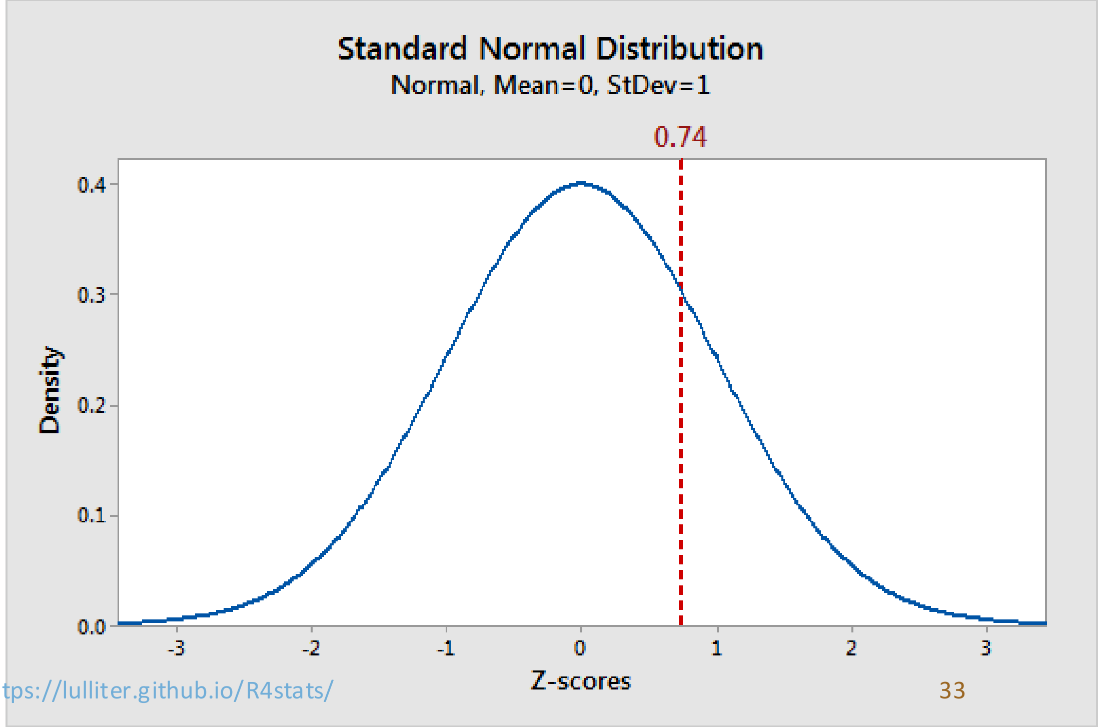
Source image: https://www.youtube.com/watch?v=1Q6_LRZwZrc&t=22249s

The Z-score helps to "*standardize*" an observation relative to its frequency distribution

- Z-score helps us understand where a specific observation falls within a distribution
- Consider the Z-distribution, a standardized Normal distribution with $\mu = 0$ and $\sigma = 1$
- The formula for finding z-scores is the following: $Z = \frac{x - \mu}{\sigma}$
- Where:
  - x represents the data point of interest
  - μ and σ represent the mean and standard deviation for the population from which you drew your sample

EXAMPLE
Using the *standard normal distribution* $N(\mu = 0, \sigma = 1)$, if a one-month-old baby girl weighs 5 kg, how does she compare to others?

A **5 kg** weight equals a Z-score of **0.74**


Standard Normal Distribution
Normal, Mean=0, StDev=1

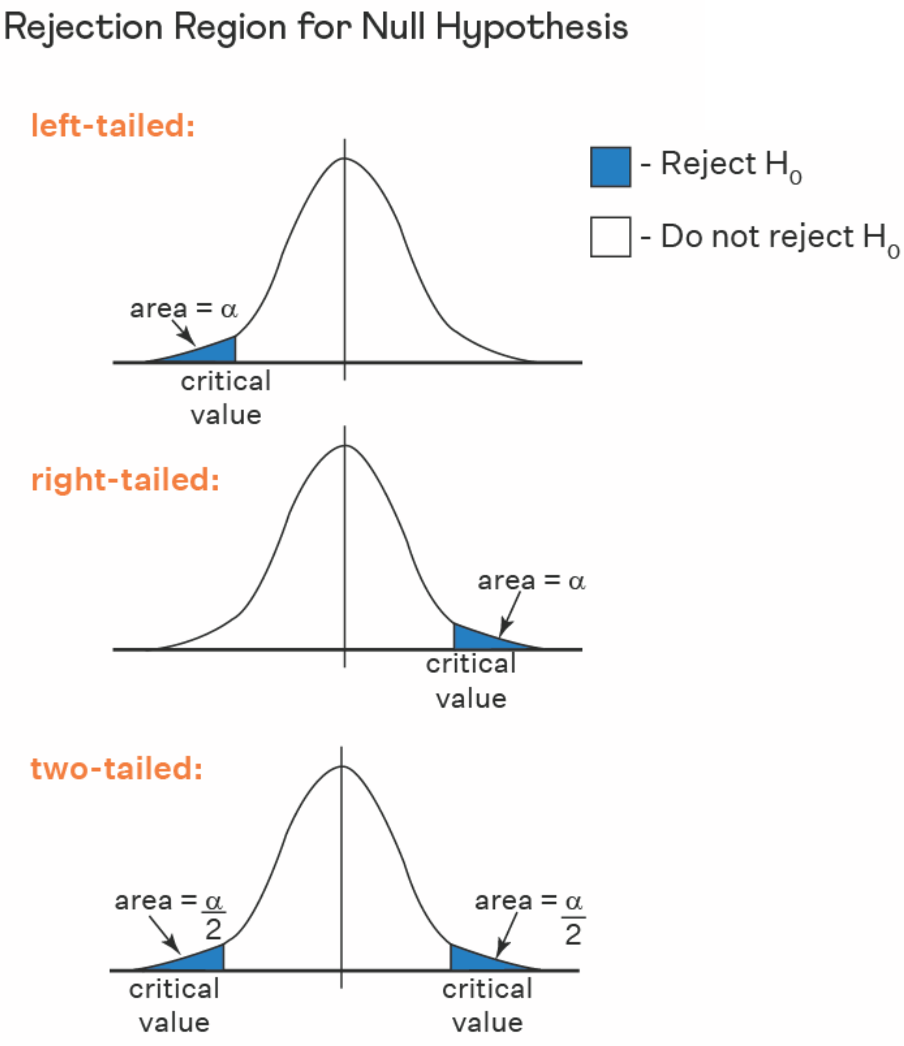
One-tailed and two-tailed tests of hypothesis

- The choice between a one-tailed and two-tailed test depends on our expectations about the reference population:
  - **one-tailed test** → We know in which direction the estimate diverges compared to the population (e.g. there is no difference between the mean pulse rate for people doing physical exercise and the normal pulse rate)

    H_a: $\mu >$ value | H_a: $\mu <$ value

  - **two-tailed test** → We don’t know in which direction the estimate diverges compared to the population

    H_a: $\mu \neq$ value



Source image: https://www.cuemath.com/data/z-test/

Confidence intervals and estimate precision

A possible application of the CLT:

- A confidence interval (CI) is a range of values that is likely to contain a population parameter with a certain level of confidence.

Confidence Interval = point estimate +/- margin of error

$$CI = \bar{x} \pm \varepsilon_i$$

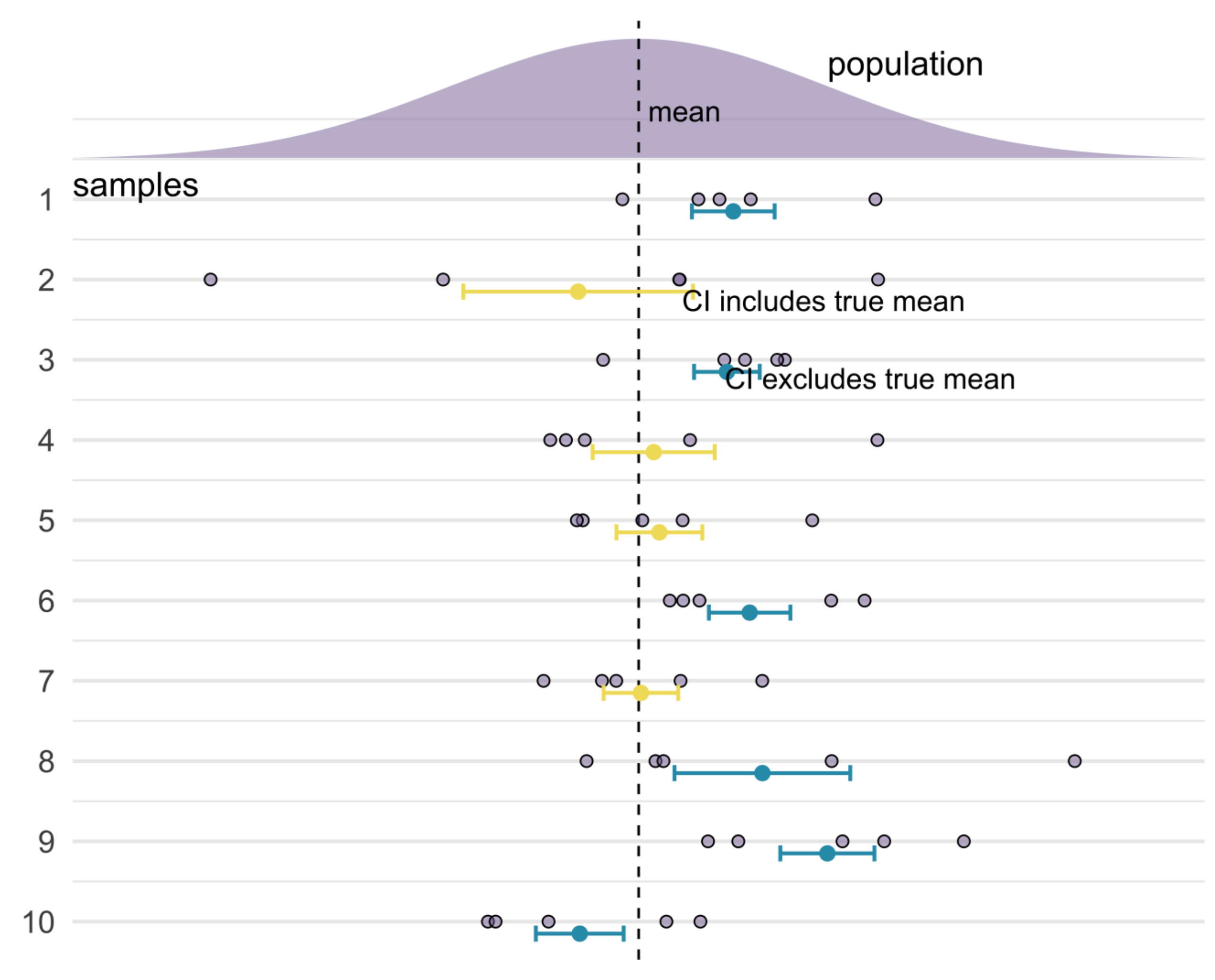
Where ε_i = (critical value)(sd of the statistic)

$$CI = \bar{x} \pm Z * \frac{\sigma}{\sqrt{n}}$$

  - The sample mean $\bar{x}$ is a point estimate of μ
  - Each sample mean $\bar{x}_i$ differs from the next one and from μ only by chance
  - The smaller ε_i , the more precise will be the sample mean $\bar{x}_i$
  - Z is a critical value that depend upon a test statistic

- The estimate precision is evaluated with the CI, which is an interval with a lower bound and an upper bound, which likely contains a population parameter with a certain level of confidence

Frequentist interpretation of a confidence interval



Source image: https://vizdata.org/slides/16/16-uncertainty-I.html#/

Example of constructing a CI for a mean point estimate

What probability level do we want? (that the interval il contain the population parameter). We choose 95% which implies:

- 5% probability of error = α

So, for any single sample we draw, we can calculate a range of values (CI) on either side of the sample mean equal to:

$$\mathbf{CI(95\%)} = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} = \boldsymbol{\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}}$$

such that:

- in 95% of all possible samples of size n , μ will fall in our confidence interval
- in 5% of all possible samples of size n, we would miss μ

Also, increasing n will reduce the margin of error ε for a fixed Z.

Example of constructing a CI for a mean point estimate

Before we were assuming to know the population standard deviation parameter $\boldsymbol{\sigma}$, but we rarely do.

$$\text{CI(95\%)} = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

If we don't know the population sd, we will estimate is with the sample $\boldsymbol{s}$:

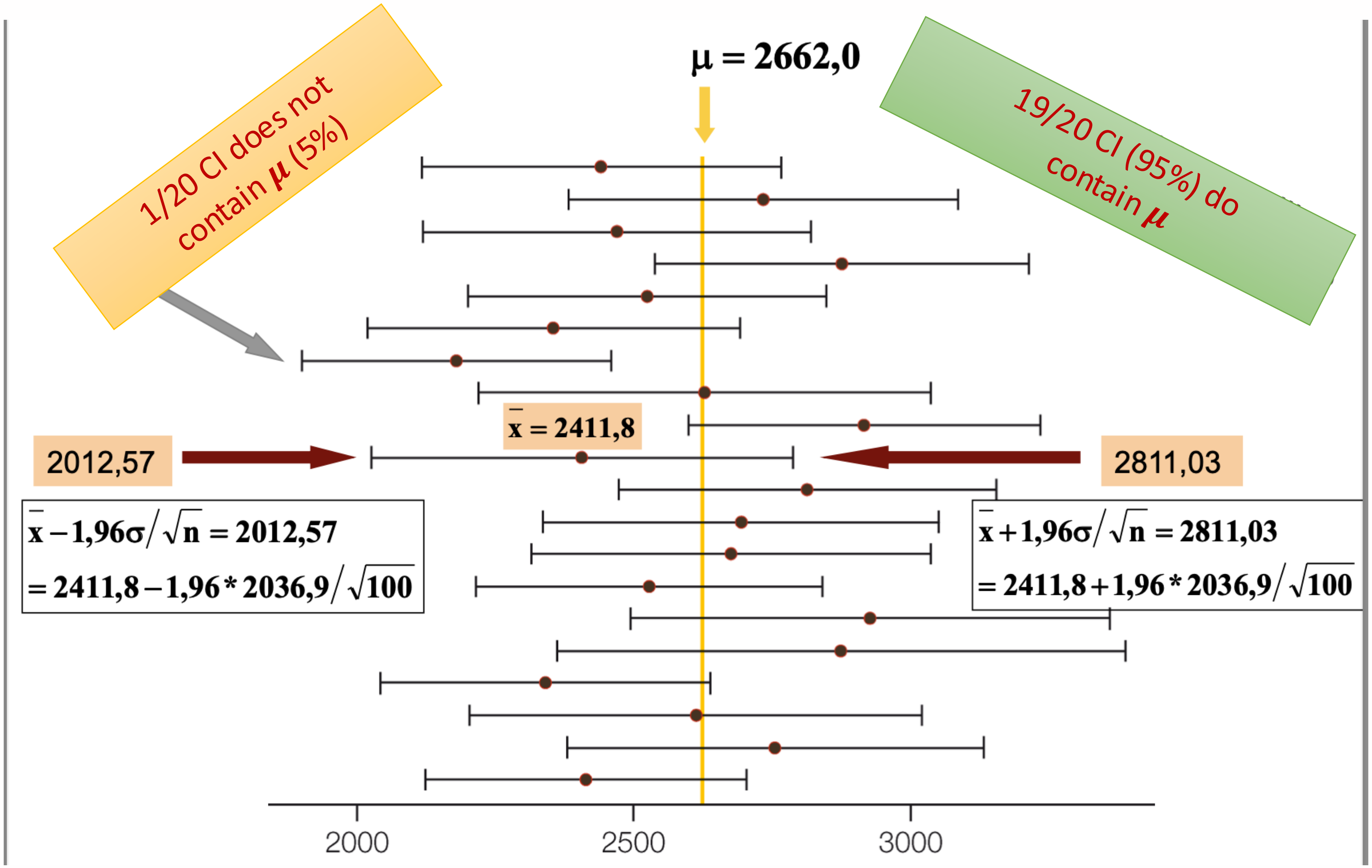
$$\text{CI(95\%)} = \bar{x} \pm z\frac{s}{\sqrt{n-1}} = \bar{x} \pm 1.96\frac{s}{\sqrt{n-1}}$$

- **NOTE**: we can use this formula (with the $\boldsymbol{Z}$ statistics) only when the sample size $\boldsymbol{n}$ is sufficiently large.

Confidence Interval example

CI (95%) of the mean calculated from 20 random samples of n =100 genes from the human genome

Length of genes (number of nucleotides)



$\mu = 2662{,}0$

1/20 CI does not contain μ (5%)

19/20 CI (95%) do contain μ

$\bar{x} = 2411{,}8$

2012,57

2811,03

$$\bar{x} - 1{,}96\sigma/\sqrt{n} = 2012{,}57$$
$$= 2411{,}8 - 1{,}96 * 2036{,}9/\sqrt{100}$$

$$\bar{x} + 1{,}96\sigma/\sqrt{n} = 2811{,}03$$
$$= 2411{,}8 + 1{,}96 * 2036{,}9/\sqrt{100}$$

DAY 2 – LECTURE OUTLINE

- Purpose and foundations of inferential statistics
    - Population and samples
    - Probability and random variables &
    - Meaningful probability distributions
    - Sampling distributions and Central Limit Theorem

- Getting to know the “language” of hypothesis testing
    - The null and alternative hypothesis
    - The probability of error? (α or “significance level")
    - The p-value probability and tests interpretation
    - Types of errors (Type 1 and Type 2)
    - Confidence Intervals
    - Effective vs statistical significance

- Hypothesis tests **examples**
    - Comparing sample mean to a hypothesized population mean (Z test & t test)
    - Comparing two independent sample means (t test)
    - Comparing sample means from 3 or more samples (ANOVA)

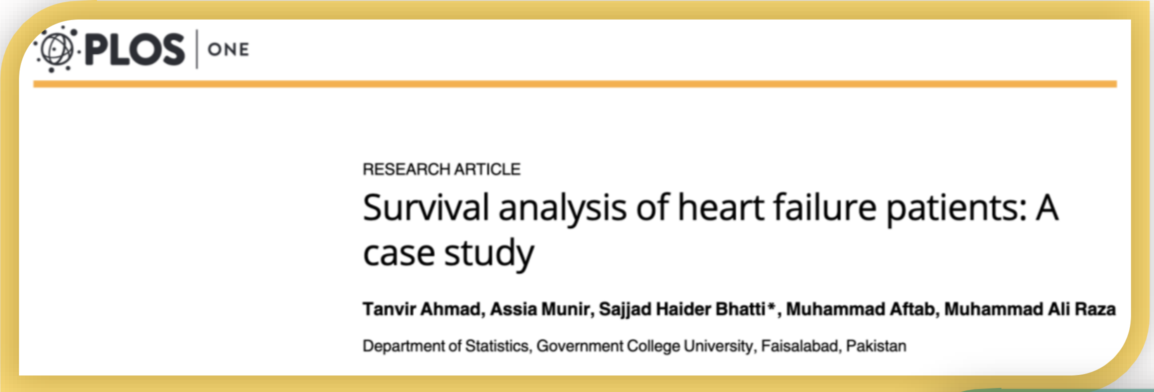
- A closer look at testing assumptions (with **examples**)
    - Testing two groups that are NOT independent
    - Testing if the data are not normally distributed: non-parametric tests
    - Testing samples without homogeneous variance of observations

Comparing a sample mean to a hypothesized population mean

EXAMPLE A: Z-test one-sample hypothesis (for large samples with known population's variance)

Our dataset for the day

- We will be working on a datasets described in 2 recent open access articles on cardiovascular heart diseases
- The original authors (Ahmad, et al., 2017) released also the open access dataset containing the **medical records of 299 heart failure patients collected at a Hospital in Faisalabad (Punjab, Pakistan), in April–December 2015**
    - all patients > 40 years old, having left ventricular systolic dysfunction
    - age, serum sodium, serum creatinine, gender, smoking, Blood Pressure (BP), Ejection Fraction (EF), anemia, platelets, Creatinine Phosphokinase (CPK) and diabetes were recorded and considered for potentially explaining mortality caused by Cardiovascular Heart Disease (CHD)



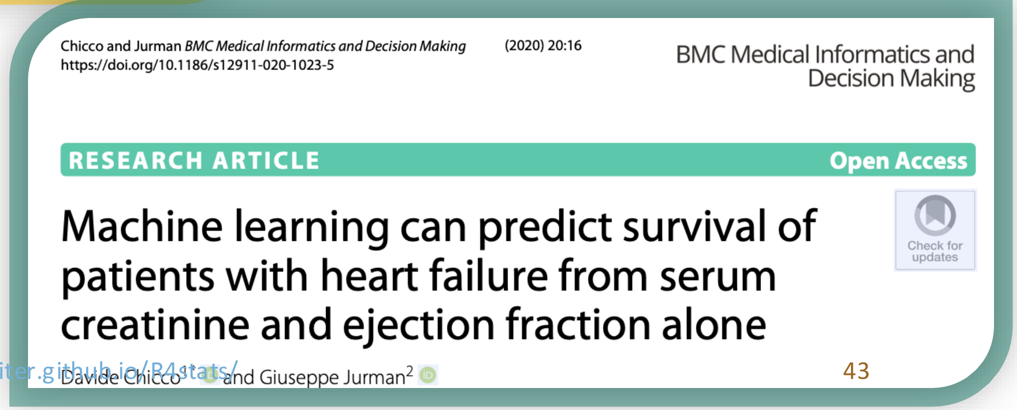
PLOS | ONE

RESEARCH ARTICLE

Survival analysis of heart failure patients: A case study

Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti*, Muhammad Aftab, Muhammad Ali Raza

Department of Statistics, Government College University, Faisalabad, Pakistan



Chicco and Jurman *BMC Medical Informatics and Decision Making* (2020) 20:16
https://doi.org/10.1186/s12911-020-1023-5

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

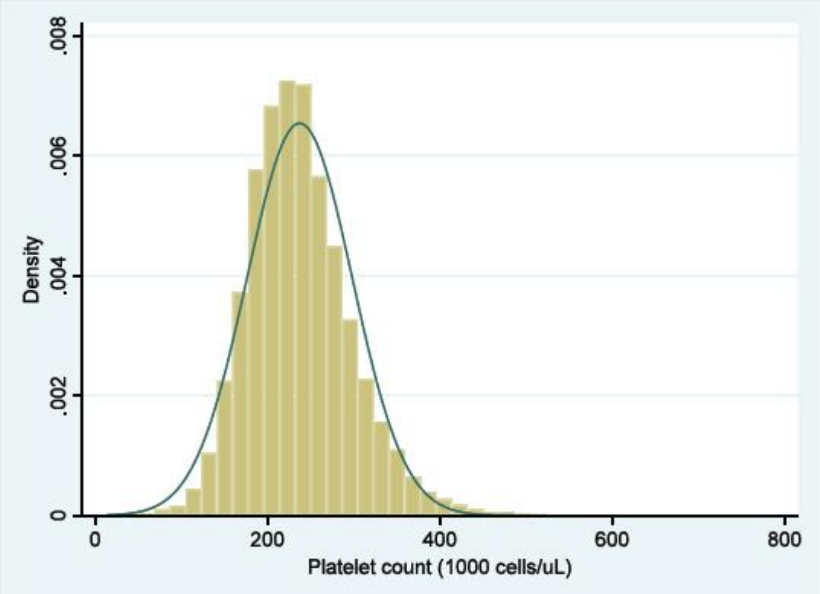
Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone

Davide Chicco[1]* and Giuseppe Jurman[2]

QUESTION: Compare the mean *platelets count* in the patients' sample against a reference distribution - with known mean (µ) and standard deviation (σ)

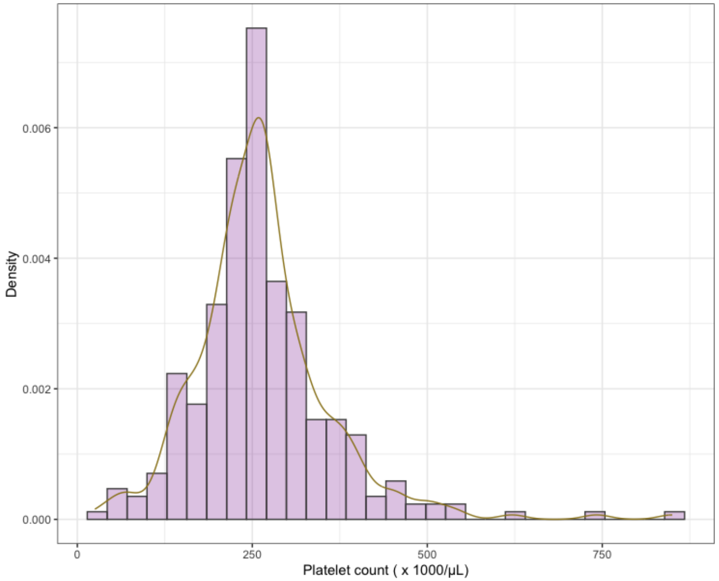
Total Platelet Count distribution
(in *Reference* general Population)

- Large population study conducted in the US between 2011–2016 on 17,969 adults.
- Total Platelet Count (TPC) measurement showed a Normal distribution with
  - $\mu = 236 \times 10^3/\mu L$
  - $\sigma = 59 \times 10^3/\mu L$



Total Platelet Count distribution
(in *Sample* of heart failure patients)

- Sample of 299 patients population study collected in Pakistan in 2015
- Total Platelet Count (TPC) measurement showed a Normal distribution with
  - $\bar{x} = 263 \times 10^3/\mu L$
  - $s = 97 \times 10^3/\mu L$



Source: https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/558/1223 (left) & author calculations (right)

Example A (step 1): Expressing the research question in the form of hypotheses

- I have a **sample mean** ($\bar{x} = 263$) collected measuring **Total Platelet Count (TPL)** for heart failure patients and I wonder if such population differs from the general population (*which I know have mean* $\mu = 236$ *with* $\sigma = 59$) simply by chance (H_0), or sampling variability, OR if (H_a) the sample mean is different from the population's because of some specific effect related to having heart disease.

- Stating the above hypotheses more formally:
  - *What is the population Total Platelet Count (TPC) mean for all people who suffered of heart failure* (μ_HF)*?*
  - H_0 : there is no difference in mean TPC between patients who suffered heart failure and the general population

    $\mu_{HF} = 236$ → hypothesis of no effect or ("no difference")

  - H_a : there is a difference in mean TPC between patients who have suffered heart failure and the general population ("some effect"). This can be formalized as either:

    $\mu_{HF} < 236$ (one-sided test), or

    $\mu_{HF} > 236$ (one-sided test), or

    $\mu_{HF} \neq 236$ (two-sided test)

Example A (step 2): Formulating an analysis plan

I start by the analysis plan, i.e. how to use sample data to evaluate $\boldsymbol{H_0}$

- The evaluation often focuses around a single test statistic
- The analysis plan should specify the following elements
  - Significance level (α): conventionally significance levels are chosen equal to 0.01, 0.05, or 0.10; but *any value between 0 and 1 can be used*.
  - Test method to determine in which direction the hypothesized mean differs significantly from the observed sample mean. Alternative options are:
    - the one-sample z-test
    - the two-sample t-test
    - the two-sample z-test
    - etc.

- Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size, less than 30.
- Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known.

Example A (step 3a): Analyze sample data: Test Statistic

- Find the value of the **test statistic** described in the analysis plan. Here:
    - Test statistic → Z score
    - Significance level → 0.05
- Given the assumptions below:
    - Patients in the HEART FAILURE were independently sampled
    - Large sample with n $\geq$ 30
    - The level of measurement of TPL is interval-ratio
    - the sampling distribution of sample means for heart failure patients is normally distributed

Test method → the one-sample z-test

One-sample test of a mean

- Take a sample $\bar{x}$ of the size $\mathbf{n}$ and a standard error $\boldsymbol{se_{\bar{x}}}$;
- Compute the **z-statistic** for the μ_{HF} $\equiv$ population mean assuming H_0 is true.

Where :

IF n $\geq$ 30 and	Standard Error	Calculated Z score
σ^2 known	$se_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$Z_{calc} = \frac{\bar{x}-\mu}{se_{\bar{x}}}$
σ^2 UNknown	$se_{\bar{x}} = \frac{s}{\sqrt{n-1}}$	$Z_{calc} = \frac{\bar{x}-\mu}{se_{\bar{x}}}$

Example A (step 3b): Analyze sample data: Test Statistic

In this case, we have:

- a large sample (**n > 100**)
- and a **known** $\boldsymbol{\sigma}^2$

We compute the standard error $se_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ and the **Z-statistic**

One-sample hypothesis test of Total Platelet Count mean in a Heart Failure affected population

Assuming H_0 is true:

- Heart Failure population mean would be ≡ general population mean $\mu = 236$
- Given our random sample size n = 299, our sample mean $\overline{x} = 263$, the general population standard deviation σ = 59 , we can compute:

$$se_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{59}{\sqrt{299}} = 3.4120$$

$$Z_{calc} = \frac{\bar{x} - \mu_0}{se_{\bar{x}}} = \frac{263 - 236}{3.4120} = 8.0180$$

Example A (step 4): make a decision using the critical region

Given that:

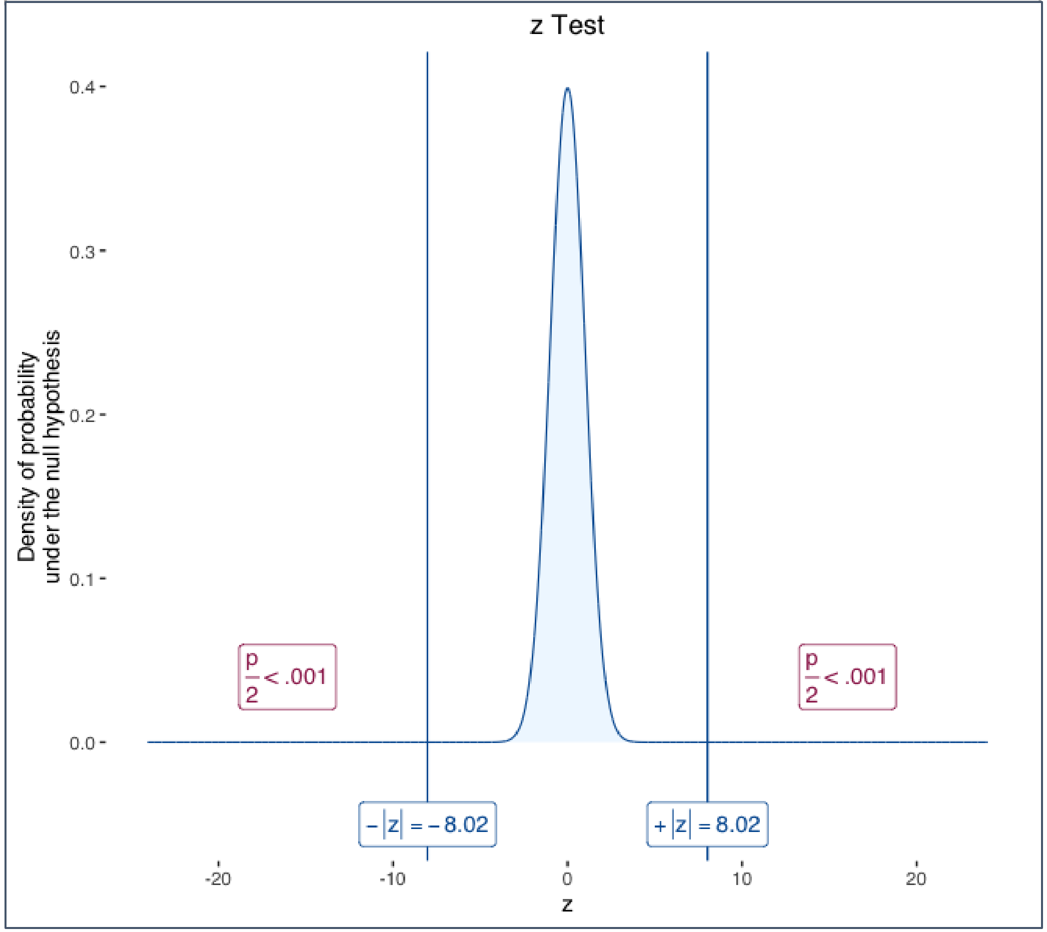
$Z_{critical} = \pm 1.96$ (Z score corresponding to α = 0.05);

Z_{calc} **= 8.0180** actually falls in the CRITICAL REGION (well beyond the critical point)

DECISION: we reject the Null Hypothesis

So the test indicates that indeed ***there is* a difference between heart failure patients and the general population in terms of average platelets count**

Sampling distribution of sample means with $\overline{x} = 263$ noted in **Z scores**

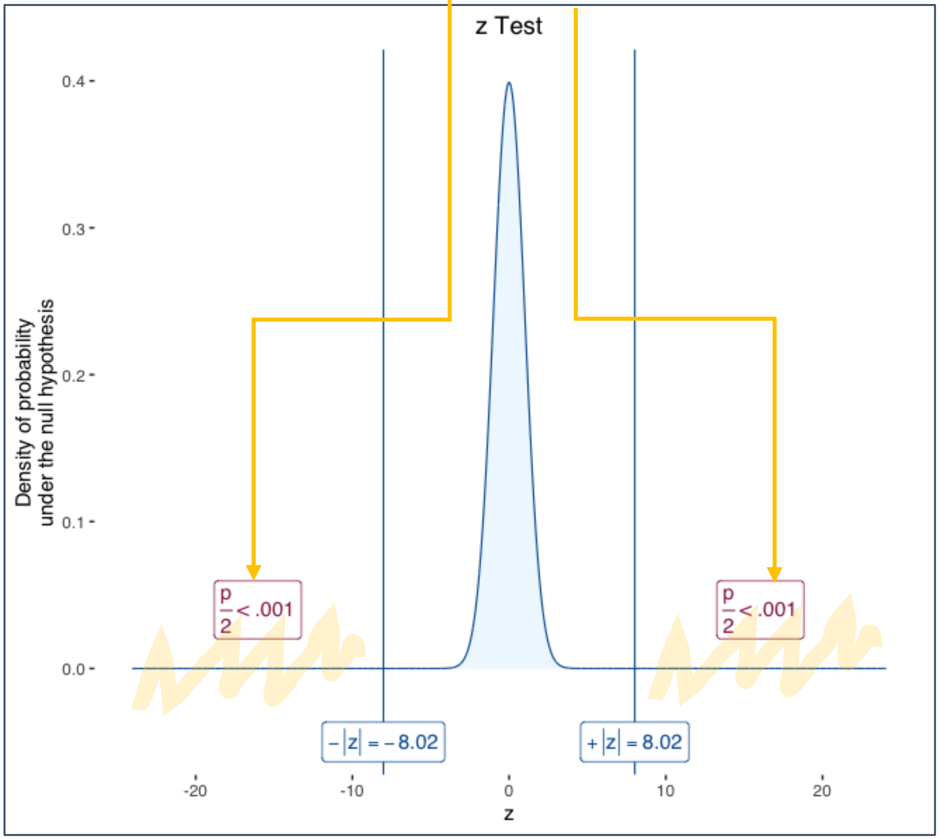


Example (step 4a): make a decision using the p-value

- The p-value answer the question: *"What is the probability of the observed test statistic or one more extreme when H_0 is true?"*
- This is the area under the curve of the Standard Normal distribution beyond the z.
- Convert z statistic to P-value:
  - For H_a : $\mu_{HF} > \mu \Rightarrow$ p = P(Z > z) = area under right-tail beyond z
  - For H_a : $\mu_{HF} < \mu \Rightarrow$ p = P(Z < z) = area under left-tail beyond z
  - For H_a : $\mu_{HF} \neq \mu \Rightarrow$ p = 2 × one-tailed P-value

$\bar{x} = 263$, with $Z_{calc} = \mathbf{8.0180}$

p-value = 0.00000000000000107443 (two-tailed)



DECISION: highly significant evidence against the Null Hypothesis

Interpretation: Thus, smaller and smaller P-values provide stronger and stronger evidence against H_0

(Variation on the theme)

Comparing a sample mean to a hypothesized population mean

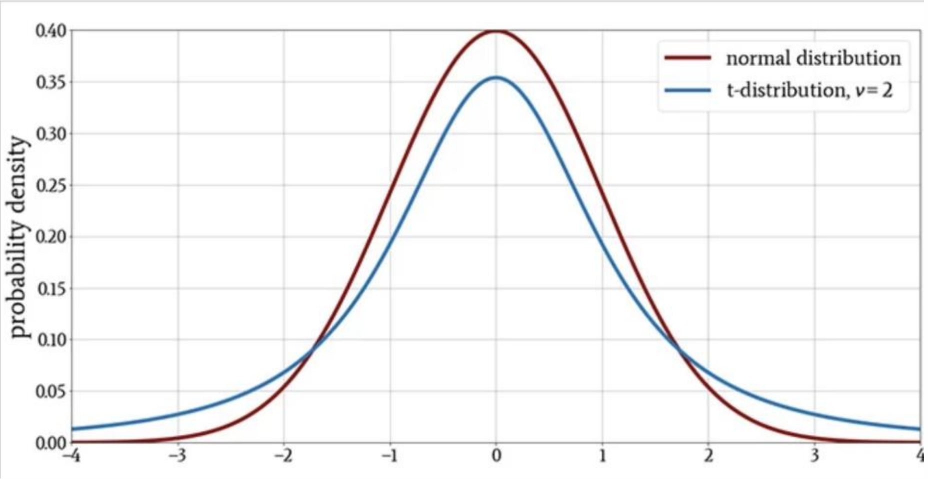
EXAMPLE B: t-test for small samples ($n < 30$) with unknown population’s variance)

[Necessary digression] Student's t-Distribution

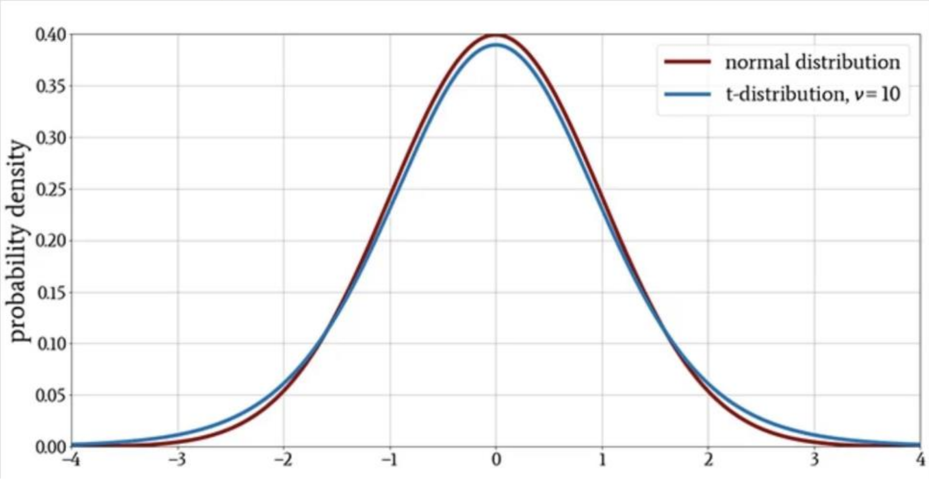
- In some cases we use the t-distribution instead of the normal distribution to model the null hypothesis because:
  1. it provides a more accurate representation of the variability in sample means and
  2. it enables making reliable inferences, even in situations where the standard normal distribution fails
- In particular in cases when dealing with **small sample sizes** or when the **population variance is unknown**
- The shape of the t-distribution changes according to the parameter ν, which denotes degrees of freedom and is determined by the sample size (denoted by n):

$$\nu = n - 1$$

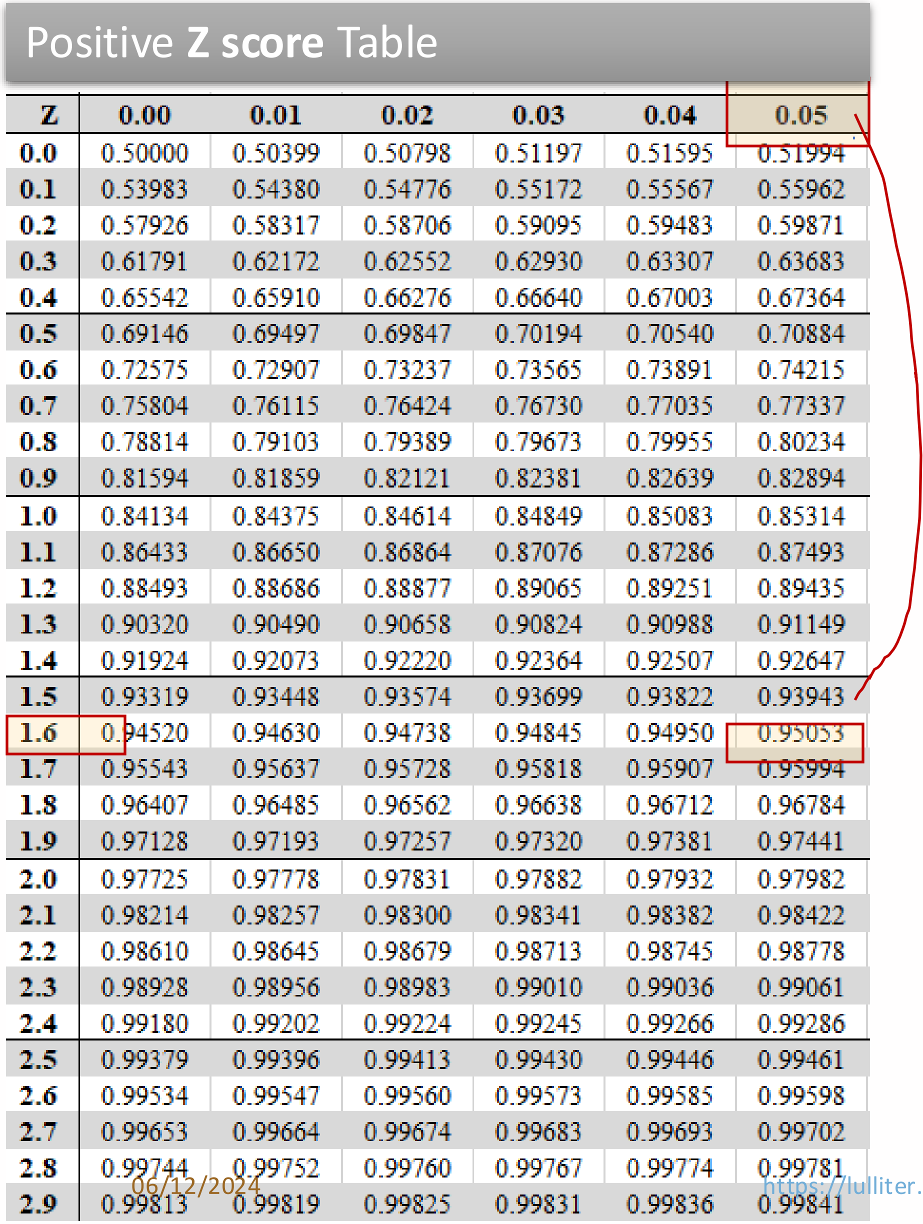
For small n the **t-distribution** is a "flattened" version of normal



For larger n, the t-distribution and the normal distribution are increasingly close

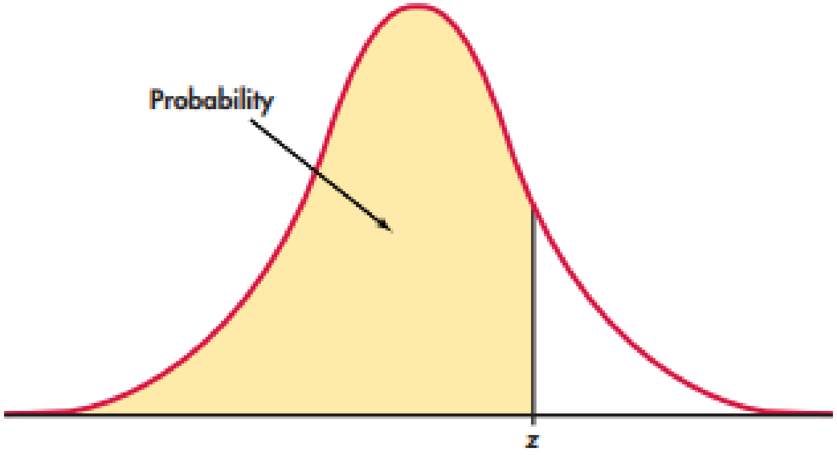


Critical value for the standard normal distribution (z) versus the t-distribution (t)



Positive **Z score** Table

Z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841



In the Z distribution, the area to the left of **z = 1.6** is equal to 0.9505,

(the probability of values falling below this observation is 95.05%)

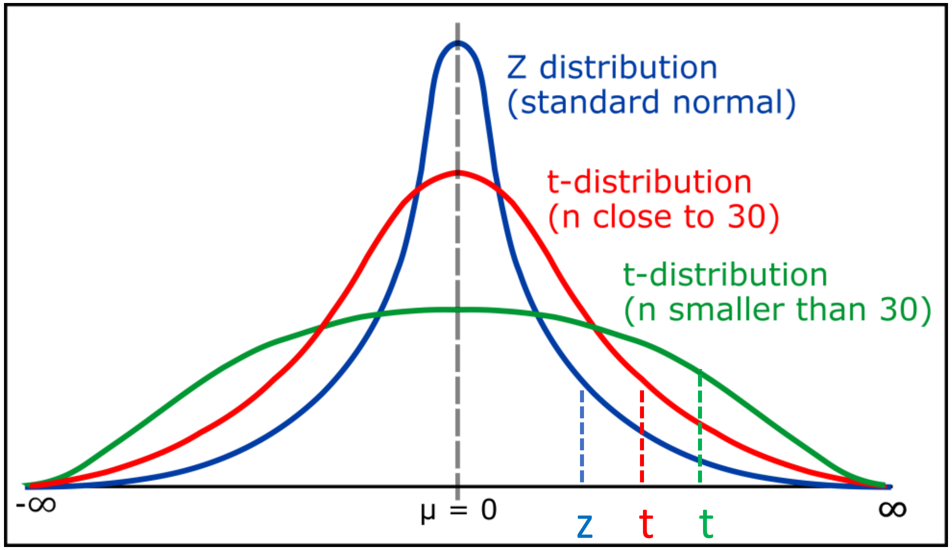
Critical value for the standard normal distribution (z) versus the t-distribution (t)

- In this case, the critical values t depend on the degrees of freedom df
- The higher is n, the closer are t critical values to Z critical values

Excerpt form **t Table**

1-tailed α	**0.025**	**0.005**
2 -taield α	**0.05**	**0.01**

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626



In the t distribution, the t = ? corresponding to an area of 95% depends on the d.f.:

- df = 100 → t = 1.984
- df = 50 → t = 2.009
- df = 10 → t = 2.228
- …

Example B (step 1): Expressing the research question in the form of hypotheses

*[Let's consider a situation in which the available **sample is smaller (n = 23)**, e.g. there are records only for patients who had a follow-up visit in 21days or less.]*

- GOAL: make inference on a "hypothetical" population of patients who suffered heart failure & who have had a follow-up visit within 21 days → μ_{HF_21d}
- QUESTION: the **sample mean** ($\overline{x} = 263$) of **Total Platelet Count (TPC)** for heart failure patients differs from the expected mean of a general population ($\mu = 236, \sigma = 59$) simply by chance (H_0) OR (H_a) the sample mean is different from the population's because of some specific effect of having heart failure.
- MORE FORMALLY:
  - H_0 : there is no difference in mean TPC between patients who suffered heart failure (visited in 21 days) and the mean TPC the general population

    $\mu_{HF\_21d} = 236$ → hypothesis of no effect or ("no difference")

  - H_a : there is a difference in mean TPC between patients who have suffered heart failure and the general population ("some effect"). This can be formalized as:

    $\mu_{HF\_21d} \neq 236$ (two-sided test)

Example B (step 2): Analyze sample data: Test Statistic

- Find the value of the **test statistic** described in the analysis plan. Here:
  - Test statistic → Z-score
  - Significance level → 0.05
- Given the assumptions below:
  - Patients in the HEART FAILURE with follow-up visit in 21 days or less were **independently sampled**
  - “Small” sample with **n** $\leq$ **30**
  - The level of measurement of **TPL is interval-ratio**
  - The **sampling distribution** of sampling means for heart failure patients is **normally distributed**

Test method → the one-sample t-test

One-sample test of a mean

- Take a sample $\bar{x}$ of the size $\mathbf{n}$ and a standard error $se_{\bar{x}}$ (let’s pretend we don’t know the general population variance $\boldsymbol{\sigma}^2$)
- Compute the **t-statistic** for the μ_{HF_30d} ≡ population mean assuming H_0 is true.

Where :

IF n $\leq$ 30 and	Standard Error	Calculated Z score
$\boldsymbol{\sigma}^2$ known	$se_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$Z_{calc} = \frac{\bar{x}-\mu}{se_{\bar{x}}}$
$\boldsymbol{\sigma}^2$ UNknown	$se_{\bar{x}} = \frac{s}{\sqrt{n-1}}$	$t_{calc} = \frac{\bar{x}-\mu}{se_{\bar{x}}}$

Example B (step 3): Analyze sample data: Test Statistic

In this case, we have:

- a small sample (**n < 30**)
- and an **unknown** $\boldsymbol{\sigma}^2$

Hence, we compute the standard error $se_{\bar{x}} = \frac{s}{\sqrt{n-1}}$ and the **t-statistic**

One-sample hypothesis test of Total Platelet Count mean in a Heart Failure affected population (<30 d)

Assuming H_0 is true:

- Heart Failure population mean μ_{HF_21d} would be≡ general population mean $\mu = 236$
- Our random sample has size **n = 23**, sample mean $\bar{x} = 251$, and standard deviation $s = 102$.
- So we compute:

$$se_{\bar{x}} = \frac{s}{\sqrt{n-1}} = \frac{102}{\sqrt{23}} = 21.90$$

$$t_{calc} = \frac{\bar{\mathrm{x}} - \mu_0}{se_{\bar{\mathrm{x}}}} = \frac{251-236}{21.90} = 0.70$$

Example B (step 4a): make a decision using the critical region

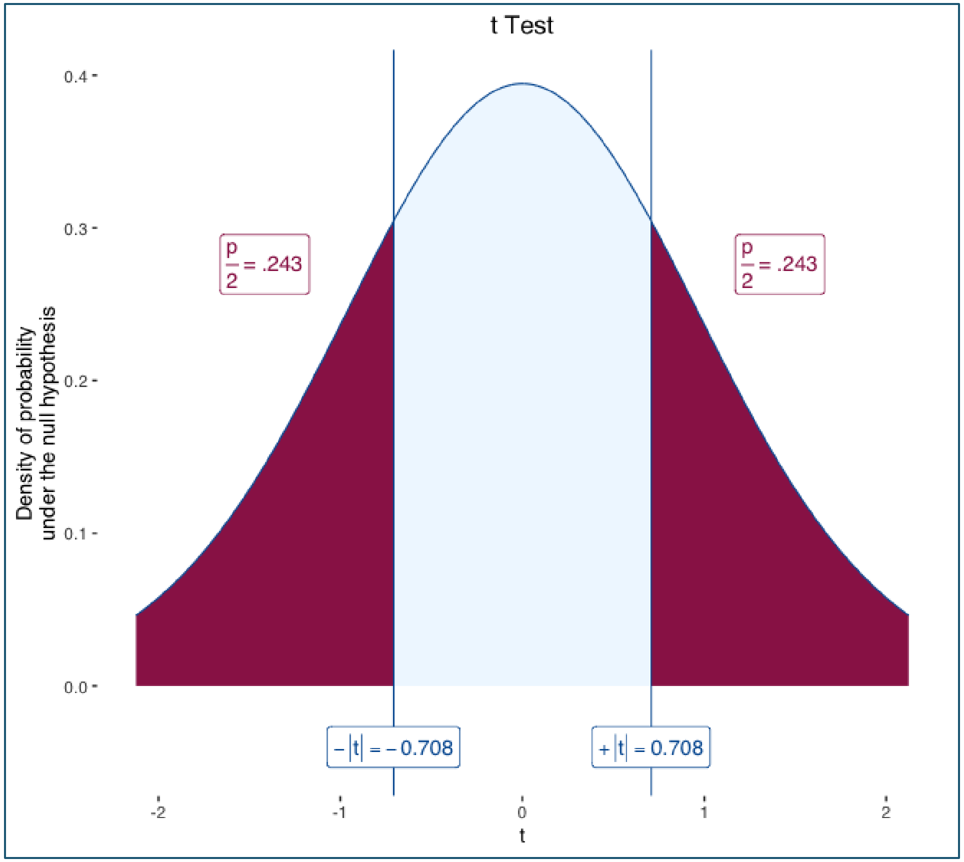
In a two-tailed test with α = 0.05 and d.f. = 23-1 = 22

the $t_{critical} = \pm 2.0738$ (t score corresponding to α = 0.05);

Hence, $t_{calc} = 0.7080$ actually falls in the ACCEPTANCE REGION

DECISION: Since the *t* statistic$\mathbf{0.7080}$ is less than the *t* critical value of 2.07 on 22 degrees of freedom at 95% level (P=0.05): **we FAIL to reject the Null Hypothesis**

Sampling distribution of sample means with $\overline{x} = 251$ noted in **t scores**

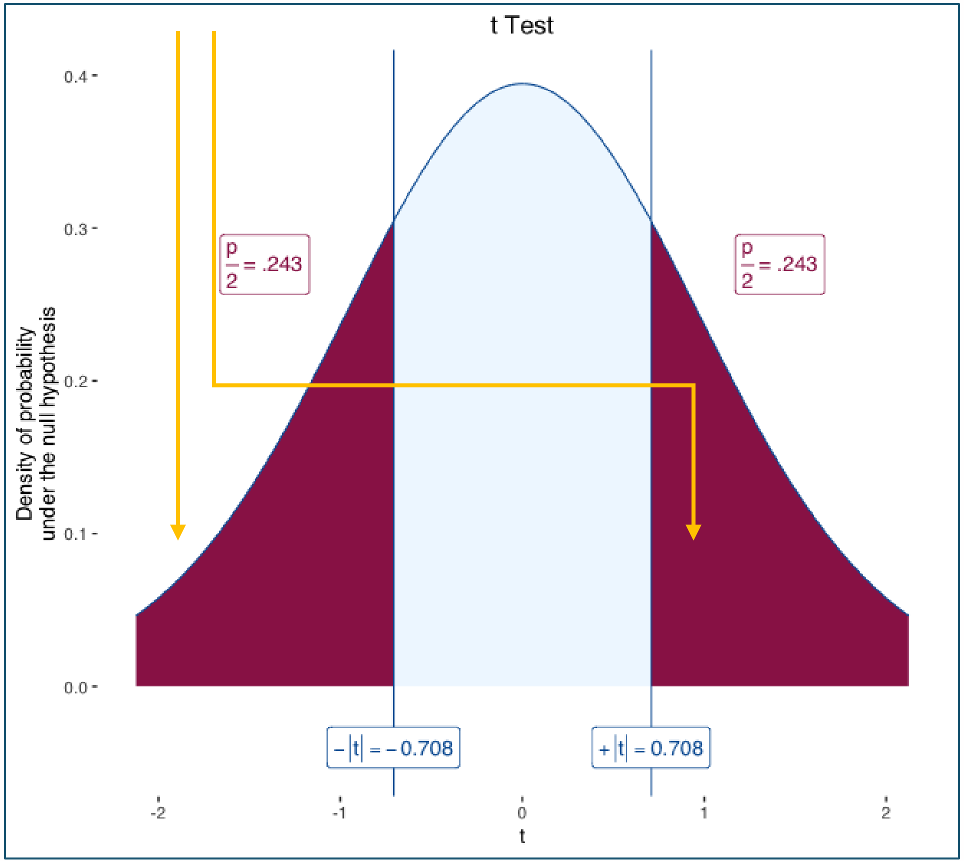


Example B (step 4b): make a decision using the p-value

- The p-value responds to the question: *"What is the probability of the observed test statistic or one more extreme when H_0 is true?"*
- This is the area under the curve of the t distribution beyond the t_{calc}.
- Convert t statistic to P-value:
  - For H_a : $\mu_{HF\_21d} > \mu \Rightarrow$ p = P($t_{calc} > t_{critical}$) = area under right-tail beyond z
  - For H_a : $\mu_{HF\_21d} < \mu \Rightarrow$ p = P($t_{calc} < t_{critical}$) = area under left-tail beyond z
  - For H_a : $\mu_{\mu_{HF\_21d}} \neq \mu \Rightarrow$ p = 2 × one-tailed P-value

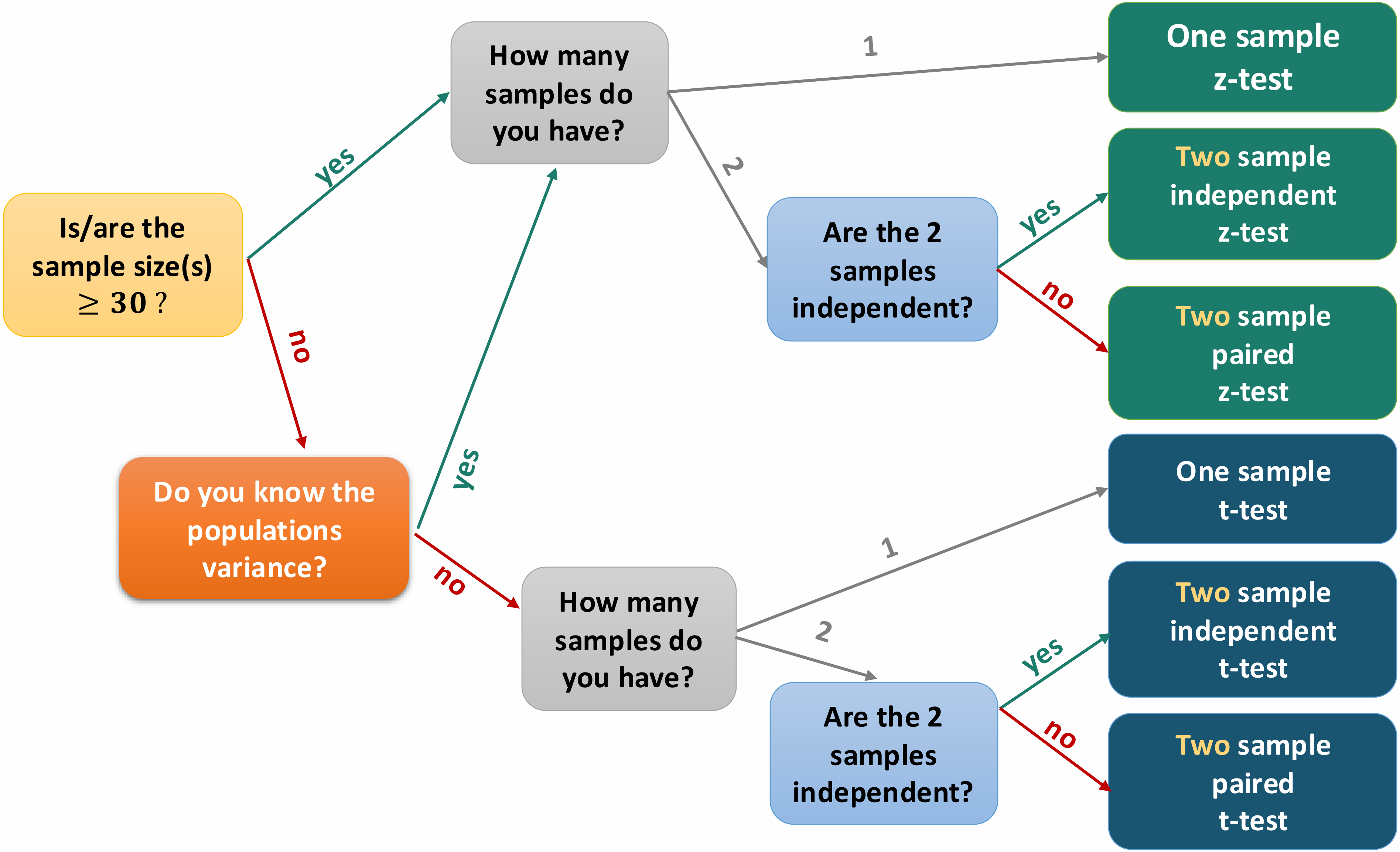
$\bar{x} = 252$, with $t_{calc} = 0.7080$

p-value = *0.48* (two-tailed)



DECISION: Since the obtained p-value is (much) higher than our significance level α = 0.05, **we FAIL to reject the null hypothesis**. We have insufficient evidence proving the difference between the general population mean TPC and mean TPC of HF patients visited within 21 days is statistically significant.

To recap: z test or t test?

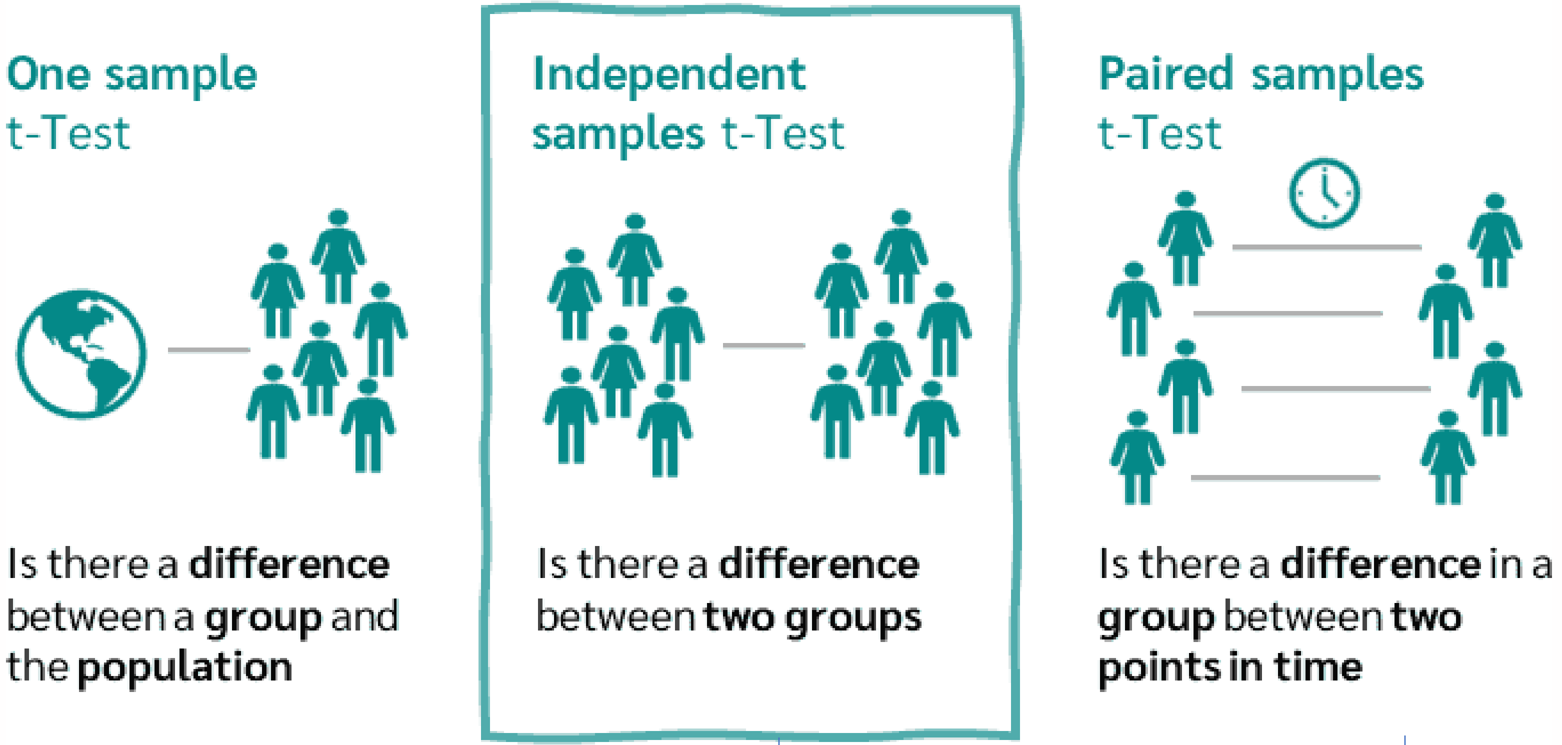

Is/are the sample size(s) ≥ 30 ?
yes
no
How many samples do you have?
1
2
One sample z-test
Are the 2 samples independent?
yes
no
Two sample independent z-test
Two sample paired z-test
Do you know the populations variance?
yes
no
How many samples do you have?
1
2
One sample t-test
Are the 2 samples independent?
yes
no
Two sample independent t-test
Two sample paired t-test

Comparing 2 sample means

EXAMPLE C: two samples independent t-test

What changes in two-sample tests type of problems?

1 sample	2 samples	



One sample t-Test

Is there a **difference** between a **group** and the **population**

Independent samples t-Test

Is there a **difference** between **two groups**

Paired samples t-Test

Is there a **difference** in a **group** between **two points in time**

Independent samples: e.g. patients receiving treatment v. those receiving placebo drug

Dependent samples: e.g. same patients visited twice: before and after surgery

Source image: https://datatab.net/tutorial/one-sample-t-test

Example C (step 1): the research question and the test hypotheses

[This time, I wonder if there a statistically significant difference between the ***Total Platelet Count*** *in the patients who died and the patients who survived heart failure.]*

- GOAL: verify if there is a difference between means μ_1 and μ_0
  - where: died = 1 and alive = 0
- QUESTION: Is there a statistically significant difference between the mean values of two groups?
- MORE FORMALLY:
  - $H_0 : \mu_1 = \mu_0 \Leftrightarrow \mu_1 - \mu_0 = 0$ The two population means are equal
  - H_a : There is a mean difference between the two groups in the population. Possible directional difference formulation (two-tailed, left-tailed, right-tailed) :
    - H_a: $\mu_1 \neq \mu_0 \Leftrightarrow \mu_1 - \mu_0 \neq 0$ (the two population means are not equal)
    - $H_a : \mu_1 < \mu_0 \Leftrightarrow \mu_1 - \mu_0 < 0$(population 1 mean is less than population 0 mean)
    - $H_a : \mu_1 > \mu_0 \Leftrightarrow \mu_1 - \mu_0 > 0$ (population 1 mean is greater than population 0 mean)

Example C (step 2): analyze sample data, test statistic

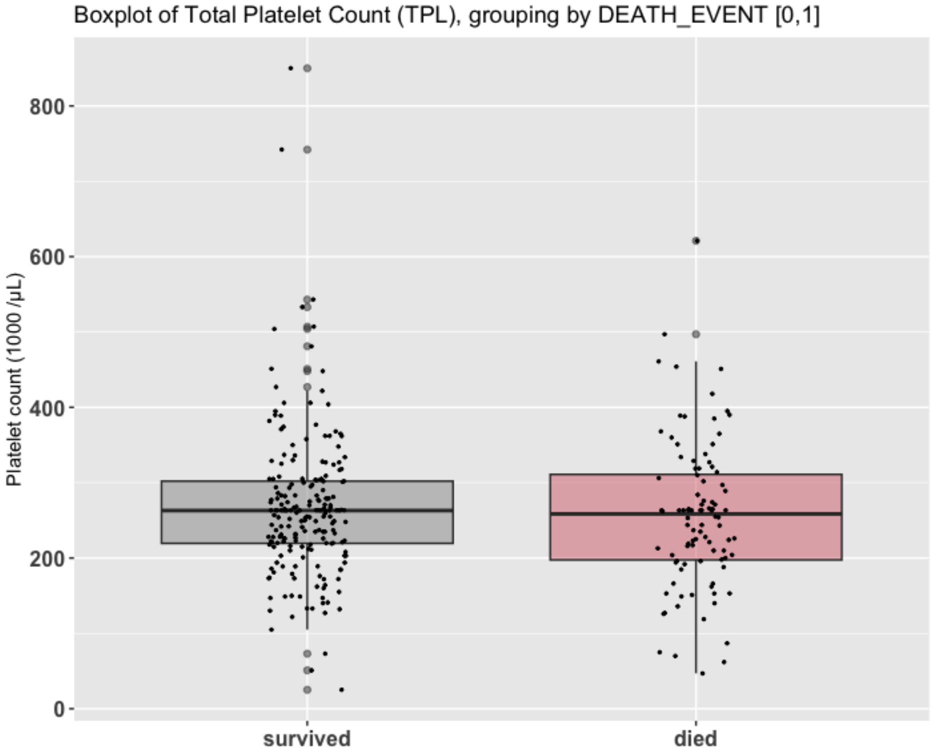
- ANALYSIS PLAN decisions:
  - Data → there is a categorical variable defining 2 groups (DEATH_EVENT = 1 or 0)
  - Test statistic → two samples' means comparison
  - Significance level → 0.05
  - Direction of equality → two-tailed (I don't have an expectation)
  - Test method → ??

- Verify the ASSUMPTIONS for independent (unpaired) t-test:
  1. The 2 samples ("died" and "survived") must be **independent***: i.e. a value in one sample must not influence a value in the other sample ✅
  2. The dependent variable is scaled in **intervals** (Total Platelets Count in $10^3/\mu L$) ✅
  3. The dependent variable is **normally distributed** (Total Platelets Count in $10^3/\mu L$) ✅
     - If the variables are not normally distributed, the Mann-Whitney U test can be used.
  4. The **variance within the two groups** should be similar (F-test or Levene's test, …) ✅
     - If the variances are not equal you should instead perform Welch's t-test (the R default)

  - Test method → **two sample independent t-test**

Example C (step 3): collect data about the two samples

- **Sample 1** (DEATH_EVENT = 1)
  - size n_1 = 96
  - mean TPC $\bar{x}_1$= 256 $(10^3/\mu L)$
  - standard deviation s_1 = 98.5 $(10^3/\mu L)$
- **Sample 2** (DEATH_EVENT = 0)
  - size n_2 = 203
  - mean TPC $\bar{x}_2$ = 267 $(10^3/\mu L)$
  - standard deviation s_2 = 97.5 $(10^3/\mu L)$
- Also the last assumption (equal variances) is verified by the F test of variance equality
  - $F_{calc} = \frac{variance\ group\ 1}{variance\ group\ 2} = \frac{\sigma_1^2}{\sigma_2^2}$
  - with $d.f. = n_1 - 1$ and $n_2 - 1$
  - and H_0: $\sigma_1^2 = \sigma_2^2$



F test statistic	df 1	df2	p
1.0205	95	202	0.8915

Interpretation

H_0 is “equal variances”

p-value = 0.89 > 0.05 → FAIL to reject H_0

Equal variances between groups ✅

Example C (step 4): compute the test statistic for t-Test for independent samples

- Since we verified the required assumptions, the test method is the independent (two-sample) t-test
- The test statistic is computed with this equation, given:
  - the population standard deviation(s) are unknown, but we can assume = variances in 2 groups
  - large sample ($n_1 + n_2 > 100$)
  - *Standard Error of the Difference* obtained as pooled estimate standard deviation of the sampling distribution of the difference

$$t_{calc} = \frac{\textit{Difference Between Sample Means}}{\textit{Standard Error of the Difference}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{(n_1-1)} + \frac{s_2^2}{(n_2-1)}}} \text{ (corrected bias) or } = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{(n_1)} + \frac{s_2^2}{(n_2)}}}$$

where:

- n_1 and n_2 are the sample sizes,
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means,
- s_1^2 and s_2^2 are the sample variances
- $df = n_1 + n_2 - 2$ are the degrees of freedom

Results:

$\text{Difference Sample Means} = -10.28$ $\text{CI }(95\%) = [-34.23, +13.66]$

$t_{calc} = -0.84$

$\text{p-value} = 0.3989$

Example C (step 5): Interpret the results

RECAP:

An independent samples t-test was conducted to compare mean counts of platelets in patients with heart failure who died and patients with heart failure who survived.

RESULTS INTERPRETATION:

To make a statement about the results of the test (whether the H_0 of equal populations means holds or not) one of the following two values is used:

- p-value (2-tailed)
  - The **p-value** $= 0.3989$ correspondent to the test statistic $\mathbf{t_{calc}} = \mathbf{-0.84}$ with $\mathbf{n_1 + n_2 - 2}$ degrees of freedom is MORE than our chosen significance level (0.05), so we CAN NOT reject the null hypothesis.
- lower and upper confidence interval of the difference
  - The magnitude of the differences in the means **Difference Sample Means = -10.28** falls inside the lower and upper bounds of the Confidence Interval **CI (95%) = [-34.23, +13.66]** consistent with the null hypothesis.

DECISION:

We do not have sufficient evidence to say that the mean counts of platelets in between these two populations is different.

Other similar cases we cannot review…

- one-sample Tests on proportions?
- two-sample Tests on proportions?

Comparing sample means from 3 or more samples

EXAMPLE D: using ANOVA test

Extending comparison to 3 or more groups…

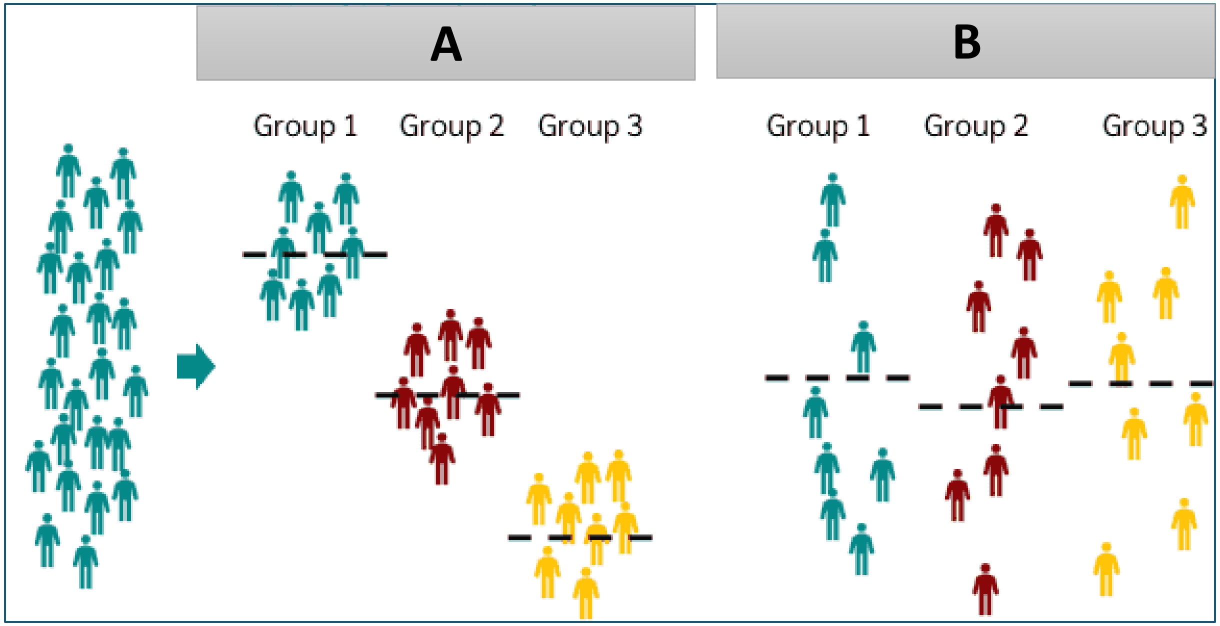
- ANOVA ("Analysis Of Variance") is an extension of the previous hypothesis testing ideas, but examined how means of a variable differ across **three or more groups**
- For this purpose, the means and variances of the respective groups are compared with each other.
- While the t-test serves with a categorical explanatory variable that has two levels, (one-way) ANOVA looks at quantitative outcomes and a single categorical explanatory variable with any number of levels
- There are different types of ANOVA. The most used are:
  - 'one- way' ANOVA (or one-factor ANOVA) if there is only one explanatory variable ("treatment") with two or more levels, and only one level of treatment is applied for a given subject (e.g. ethnicity)
  - 'two-way' ANOVA (or two-factor ANOVA), if the levels of two different explanatory variables are being assigned, and each subject is mapped to one level of each factor (e.g. ethnicity + treatment type)
- Another distinction refers to the selection of groups:
  - with repetition (as in the case the same person interviewed at several points in time)
  - without repetition (independent groups)

Example: an oncologist may be interested in knowing whether patients with different types of cancer have the same average survival times ('one-way' ANOVA) under several different competing cancer treatments ('two-way' ANOVA) .

Formalizing One-Way ANOVA ("Analysis Of Variance")

- The dependent variable is on a metric scale. In the case of the analysis of variance, the independent variable (*factor) h*as at least three levels.
- **Assumptions** for the results of a one-way ANOVA to be valid:
  1. **Independence of observations** – The observations in each group are *independent* of each other and the observations within groups were obtained by a random sample.
  2. **Normally-distributed response variable** – The values of the dependent variable follow a normal distribution.
  3. **Homogeneity of variance** – The variances of the populations that the samples come from are equal.
- Key concept with ANOVA: **"within"** and **"between" variations** in the dependent variable values
  - **Total Variation** = Within Variation + Between Variation

- **Case A** (small variance within the groups, large between the groups)
- **Case B** (large variance within the groups, small between the groups)



Source image: https://datatab.net/tutorial/anova

ANOVA: Research application

Population	Sample Size	Sample Mean	Sample Variance
1	$n1$	$\bar{x}_1$	s_1^2
2	$n2$	$\bar{x}_2$	s_2^2
…	…	…	…
k	nk	$\bar{x}_k$	s_k^2

- MOTIVATION: In general, suppose there are *K* normal populations with possibly different means, $(\mu 1, \mu 2, \ldots, \mu K)$, but all with the same variance σ^2. To perform the test, K independent random samples are taken from the populations to obtain K sample means.
- QUESTION: Is there a statistically significant difference between the mean values of the k populations?
- MORE FORMALLY:
  - $\boldsymbol{H_0}$: $\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \ldots = \boldsymbol{\mu_k}$ all *K* population means are equal
  - $\boldsymbol{H_a}$: not all *K* population means are equal
- To calculate the variation we use the "sum of squares", like so:
  - $SSB = \sum n_k(\bar{X}_k - \bar{X})^2$ "**sum of squares BETWEEN groups**", with *df1=k−1*
  - $SSW = \sum(X_i - \bar{X}_k)^2$ "**sum of squares WITHIN groups**", with *df2=N−k*
  - $SST = \sum(X_i - \bar{X})^2$ "**TOTAL sum of squares**", where *df2=N−1*
- where, $\bar{X}_k$ = mean of each category, $\bar{X}$ = the "grand mean" of the sample, N = total number of observation, n_k = number of observations in each group

Test statistic for ANOVA ("Analysis Of Variance")

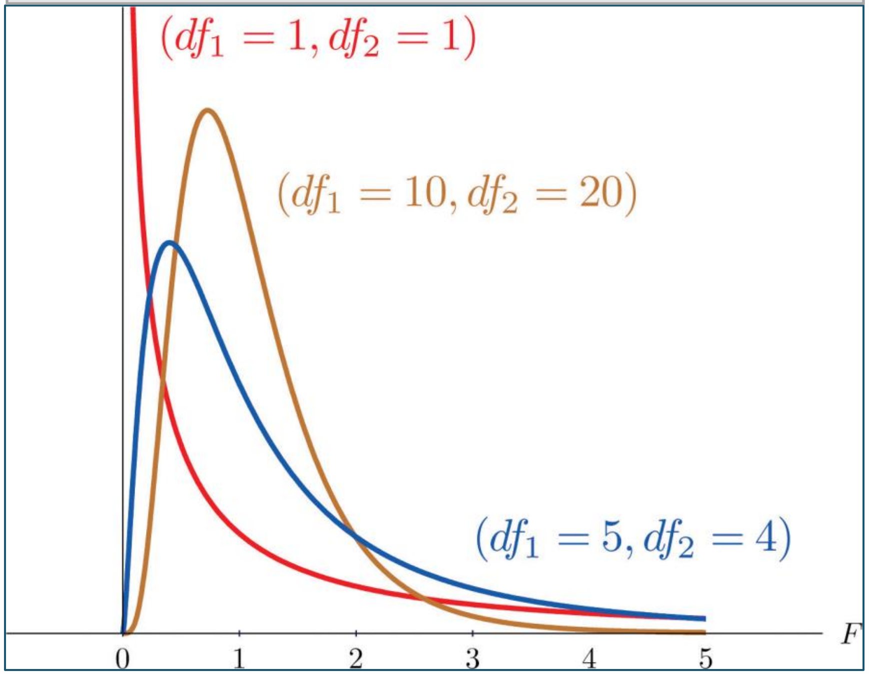
- For ANOVA, the test distribution we use is the **F-distribution**
  - The F test (like the t test) requires the **degrees of freedom** input
- The test is right-tailed: H_0 is rejected at level of significance α if $F_{calc} \geq F_\alpha$
- With:

$$F_{calc} = \frac{Mean\, Square\, Between}{Mean\, Square\, Within} = \frac{MSB}{MSW} = \frac{SSB/\,df1}{SSW\,/\,df2}$$ with $df1=K-1$ and $df2=N-K$

Each *F*-distribution is specified by 2 *degrees of freedom* parameters denoted:

- ***df*1** (numerator d.f.)
- ***df*2** (denominator d.f.)

F-Distribution examples



Example D (problem): research question and test hypotheses

- DATA: A research laboratory developed two treatments aimed at prolonging the survival times of patients with an acute form of thymic leukemia. To evaluate the potential treatment effects 33 laboratory mice with thymic leukemia were randomly divided into 3 groups.
  - The 1st group received Treatment 1, the 2nd group received Treatment 2, and the 3rd group was observed as a control group (survival times of these mice are given below)
- RESEARCH QUESTION: Is there sufficient evidence to confirm the belief that at least one of the two treatments affects the average survival time of mice with thymic leukemia?

Mice survival times in days by group			
Treatment 1		Treatment 2	Control
71	75	77	81
72	73	67	79
75	72	79	73
80	65	78	71
60	63	81	75
65	69	72	84
63	64	71	77
78		84	67
71		91	

Source example data: https://saylordotorg.github.io/text_introductory-statistics/s15-04-f-tests-in-one-way-anova.html

Example D (step 1): analysis plan and hypotheses

- GOAL: Test, at the 1% level of significance, whether the differences between the samples are large enough to reject the null hypotheses and justify the conclusion that *the populations represented by the samples* are different

- Verify the ASSUMPTIONS for One-way ANOVA
  1. Independence of observations ✓ (assignment to groups was done randomly)
  2. Normally-distributed response variable ✓
  3. Homogeneity of variance ✓
     - (more on this in the practice session)

- ANALYSIS PLAN decisions:
  - Data → there is a categorical variable defining 3 groups
  - Test statistic → F distribution
  - Significance level → 0.01
  - Direction of equality → The test is right-tailed: H_0 is rejected at level of significance α if $F_{calc} \geq F_{\alpha}$

- HYPOTHESES formalization:
  - H_0 : $\mu_1 = \mu_2 = \mu_3$ all *K* population means are equal
  - H_a : at least one population mean is different from the rest

$$F_{calc} = \frac{Mean\ square\ between}{MEan\ Square\ Within} = \frac{MSB}{MSW} = \frac{SSB/\ df1}{SSW\ /\ df2}$$ with *df*1=*K*−1 and *df*2=*N*−*K*

Example D (step 2): compute the test statistic for H_0 (i.e. *K* Population Means Are Equal)

- Experiment set up:
  - n = 33, K = 3, so that degrees of freedom ***df*1**=$K-1$ = **2** and ***df*2**=$n-K$ = ***30***
- with these samples' statistics:

Groups	Sample Size	Sample Mean	Sample Variance
Treatment 1	$n_1 = 16$	$\bar{x}_1 = 69.75$	$s_1^2 = 34.47$
Treatment 2	$n_2 = 9$	$\bar{x}_2 = 77.78$	$s_2^2 = 52.69$
Control	$n_3 = 8$	$\bar{x}_3 = 75.88$	$s_3^2 = 30.69$

- The overall sample mean (all 33 observations) is $\bar{X}$ = 73.42
- We compute Means Square Between and Means Square Within:

$$MSB = \frac{\sum n_k(\bar{X}_k - \bar{X})^2}{(K-1)} = \frac{16(69.75-73.42)^2+9(77.78-73.42)^2+8(75.88-73.42)^2}{3-1} = \frac{434.63}{2} = 217.31$$

$$MSW = \frac{\sum(X_i - \bar{X}_k)^2}{(n-K)} = \frac{(71-69.75)^2 + \ldots + (77-77.7)^2 + \ldots + (81-75.87)^2}{33-3} = \frac{1153.4}{30} = 38.45$$

- so that:

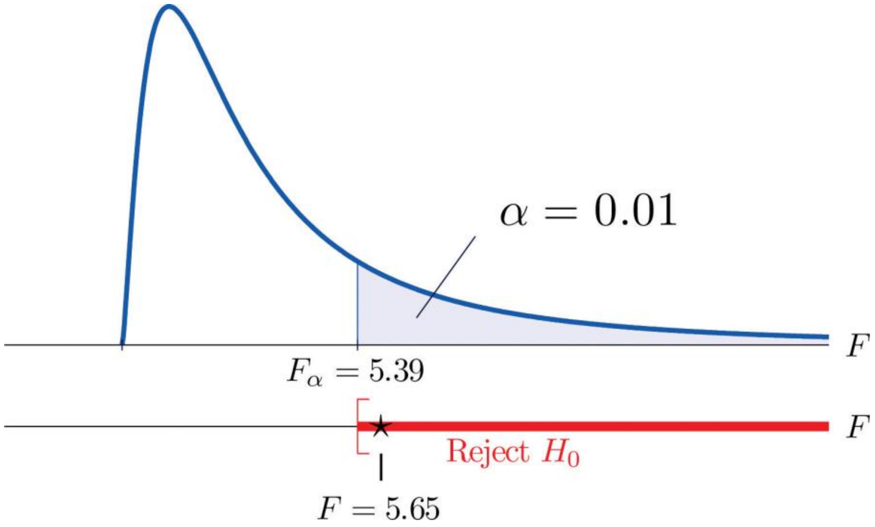
$$F_{calc} = \frac{MSB}{MSW} = \frac{217.31}{38.45} = 5.65$$

Example D (step 3): interpret the results

- The obtained test statistic is:

$$\mathbf{F_{calc}} = \frac{\mathbf{MST}}{\mathbf{MSE}} = \frac{217.31}{38.45} = 5.65$$

- The test is right-tailed. The single critical value is $\mathbf{F_\alpha} = \mathbf{F_{0.01}} = \mathbf{5.39}$, thus the rejection region is $\mathbf{[5.39, \infty)}$



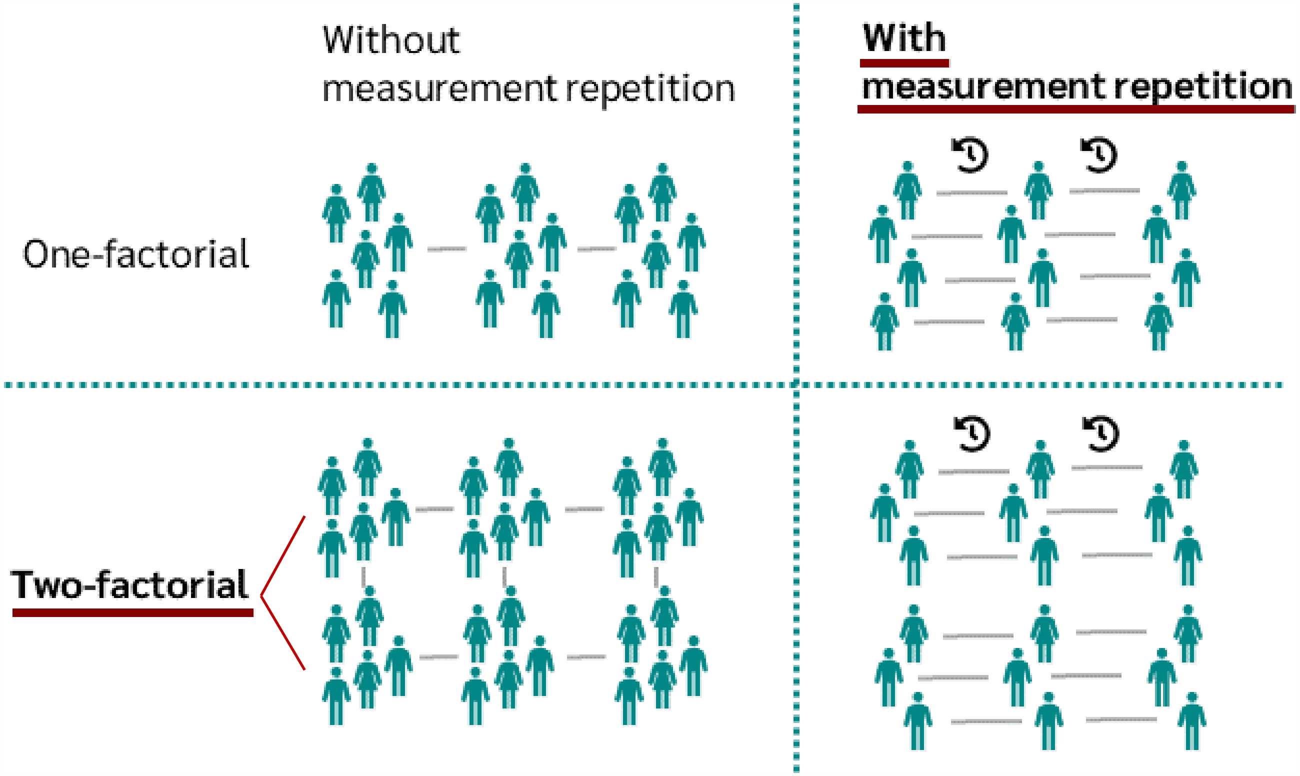
DECISION: Since $\mathbf{F_{calc}} = 5.65 > 5.39$, we reject H_0.

The data provide sufficient evidence, at the 1% level of significance, to conclude that **a treatment effect exists <u>at least for one</u> of the two treatments** in increasing the mean survival time of mice with thymic leukemia.

NOTE: ANOVA does NOT specify *which* population means are different. To determine this, you need to perform post hoc tests, also known as “multiple comparisons” tests.

Other types of ANOVA…

- Depending on:
  - how many explanatory variables ("categories") we consider
  - if samples are assigned with measurement repetition (within-subjects factor) or without measurement repetition (between-subjects factor)



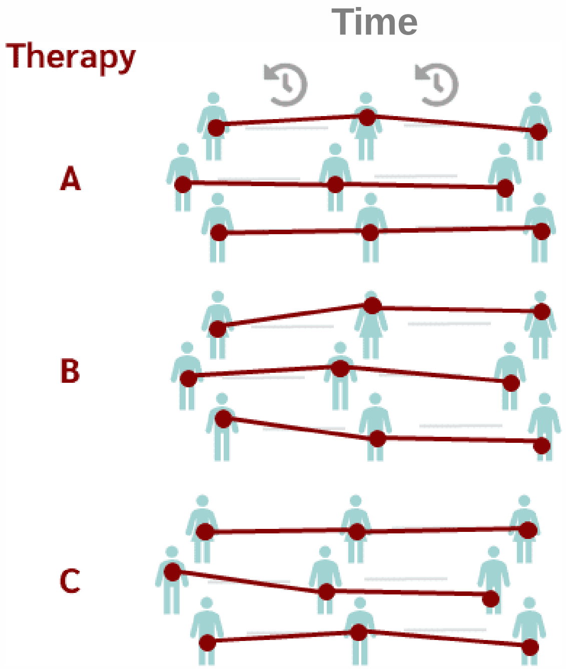
Source image: https://datatab.net/tutorial/two-factorial-anova-with-repeated-measures

Two-way, repeated measures ANOVA

- A repeated measures ANOVA is used to compare mean scores across multiple observations of the same subjects (dependency). It is typically used in two specific situations:
  - Measuring the mean scores of subjects during three or more time points
  - Measuring the mean scores of subjects under three or more different conditions.
- Two-factor ANOVA allows to decompose the total dispersion of the data into four components:
  1. the share attributable to the 1st factor
  2. the share attributable to the 2nd factor
  3. the share attributable to the interaction between 1st factor and 2nd factor
  4. the unexplained, or residual portion

EXAMPLE two-factorial ANOVA with repeated measures

Taking samples of people with high blood pressure for each treatment (1st factor = treatment) & repeat their blood pressure measurement over 3 points in time: before, during and after the treatment (2nd factor = time)



Source image: https://datatab.net/tutorial/two-factorial-anova-with-repeated-measures

Repeated measures ANOVA: applications

- In real life there are two benefits of using the same subjects across multiple treatment conditions:
  1. It’s cheaper and faster for researchers to recruit and pay a smaller number of people to carry out an experiment since they can just obtain data from the same people multiple times
  2. We are able to attribute some of the variance in the data to the subjects themselves, which makes it easier to obtain a smaller p-value («control of confounders»)

- One potential drawback of experimental design is that subjects might get bored or tired if an experiment lasts too long («attrition»), which could skew the results.

DAY 2 – LECTURE OUTLINE

- Purpose and foundations of inferential statistics
  - Population and samples
  - Probability and random variables &
  - Meaningful probability distributions
  - Sampling distributions and Central Limit Theorem

- Getting to know the "language" of hypothesis testing
  - The null and alternative hypothesis
  - The probability of error? (α or "significance level")
  - The p-value probability and tests interpretation
  - Types of errors (Type 1 and Type 2)
  - Confidence Intervals
  - Effective vs statistical significance

- Hypothesis tests **examples**
  - Comparing sample mean to a hypothesized population mean (Z test & t test)
  - Comparing two independent sample means (t test)
  - Comparing sample means from 3 or more samples (ANOVA)

- A closer look at testing assumptions (with **examples**)
  - Testing two groups that are NOT independent
  - Testing if the data are not normally distributed: non-parametric tests
  - Testing samples without homogeneous variance of observations

What if tests assumptions do not hold?

What if the assumptions did not hold?

Revisit the ASSUMPTIONS we verified for independent (unpaired) t-test:

1. The **response variable must be expressed through an interval and ratio scale** (quantitative variable → continuous scale)
2. What if the **2 groups/samples are NOT independent** (e.g. taken via before & after surveys)**?**
   - We use the dependent sample t-test **(or paired t-test)**

$$t_{calc} = \frac{\mathit{Average\ Difference\ measurement}}{\mathit{Standard\ Error\ of\ the\ Difference}} = \frac{\bar{x}_{diff}}{\sqrt{\frac{s^2_{diff}}{n}}} \text{ with } df = n - 1$$

3. What if **dependent variable is NOT normally distributed?**
   - The normality assumption is more important for small sample sizes than for larger sample sizes (but if it is hard to verify, we rely on our domain knowledge)* we use (*nonparametric* tests that doesn't assume normality)
     - For **INDEPENDENT** SAMPLES we can perform Mann-Whitney *U* test (or Wilcoxon rank-sum test as in R) -- best for continuous variables
     - For **PAIRED** SAMPLES we can perform Wilcoxon signed-rank test -- OK with ordinal variables
4. What if **variance within the two groups is NOT similar?** (F-test, Levene's test,…)
   - We can perform Welch's t-test (the R default)

[We will learn the R code in the PRACTICE SESSION]

UNMET ASSUMPTION II)

The two groups are not independent: paired t-tests

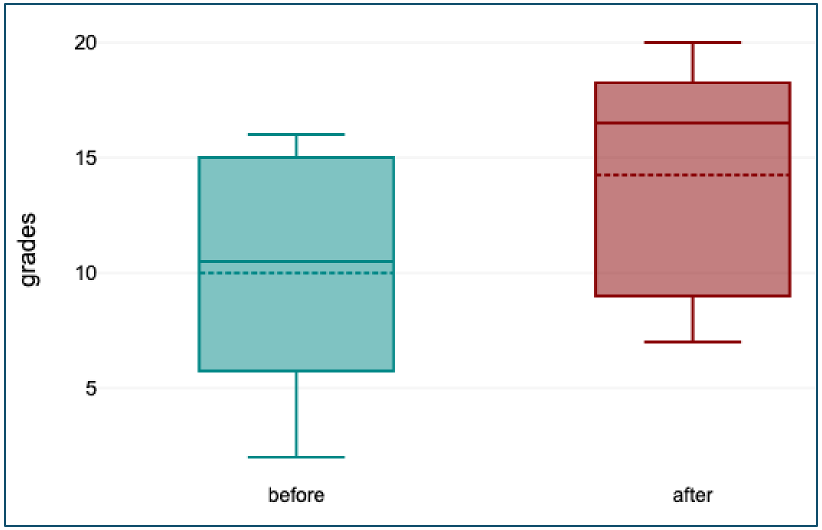
EXAMPLE E: two (paired) small samples

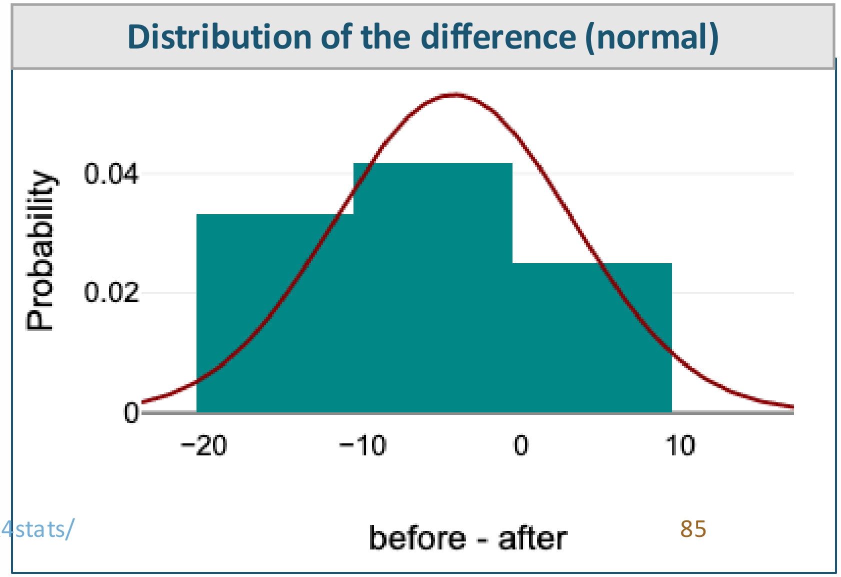
Example E (step 2): Question: Is the difference between two PAIRED samples statistically significant?

[This time, Let's imagine a statistics test is administered to the same group of 12 students before & after attending a workshop 😉]

- Given the assumptions:
  - **outcome variable is interval scaled** ✅
  - the differences of the paired values are **normally distributed** ✅
  - **independence** ❌
    - (observations are paired by design)

- We use the **dependent t-Test** for paired samples, with the following hypotheses:
  - H_0 : the mean grades before and after the workshop are equal
  - H_a : the mean grades before and after the workshop are different





Example E (step 3): Interpret the results

- The **paired dependent t-Test** (to evaluate $H_a : \mathbf{before} \neq \boldsymbol{after}$) basically is executed on the mean of the paired differences $\overline{x}_{diff}$. The test statistic $\boldsymbol{t}$ is:
  - $t_{before-after} = \frac{\overline{x}_{diff} - 0}{se_{\overline{x}}}$
  - with standard error of the mean $se_{\overline{x}} = \frac{s_{diff}}{\sqrt{n}}$
- In this case (two-sided H_a, n = 12, and df = 11, $\alpha = 0.05$), we obtain t = -1.88, with p-value = 0.087.
- Since p-value > 0.05 , this results suggests that there is no statistically significant difference between the before and after means.
- Therefore, at the 5% significance level, **we do not reject the null hypothesis** that the grades are similar before and after the workshop 😭.

	t	df	p
before - after	-1.88	11	.087

UNMET ASSUMPTION III)

If the data are not normally distributed: non-parametric tests

EXAMPLE F: using Wilcoxon Rank Sum Test

Example F (step 1): Compare two independent sample when the data are not normally distributed

*[Let's go back to the HEART FAILURE dataset but looking at the levels of **Creatinine Phosphokinase** (CPK) in the blood, an enzyme that might indicate a heart failure or injury]*

- GOAL: verify if the difference in **CPK levels in the blood** of the survivors versus those who died after heart failure is statistically significant or only due to sampling error.
- From the sample I get:
  - $\overline{x} = 582$ the general sample mean (with $s \quad = 970$)
  - $\overline{x}_{survived} = 670$ sample mean for group of survived patients (with $s_{survived} = 754$)
  - $\overline{x}_{died} = 540$ sample mean for group of dead patients (with $s_{died} = 1,317$)

- MORE FORMALLY: I want to run a test to verify whether my sample' statistics represent an actual difference in the respective hypothetical populations (H_a) or if there is no difference between the two hypothetical populations (H_0)
  - H_0 : there is no difference in mean CPK between patients who suffered heart failure and died versus patients who survived after heart failure

    $\mu_{CPK\_died} = \mu_{CPK\_survived}$ → hypothesis of no effect or ("no difference")

  - H_a : there is a difference in mean CPK between patients who suffered heart failure and died versus patients who survived after heart failure ("some effect").

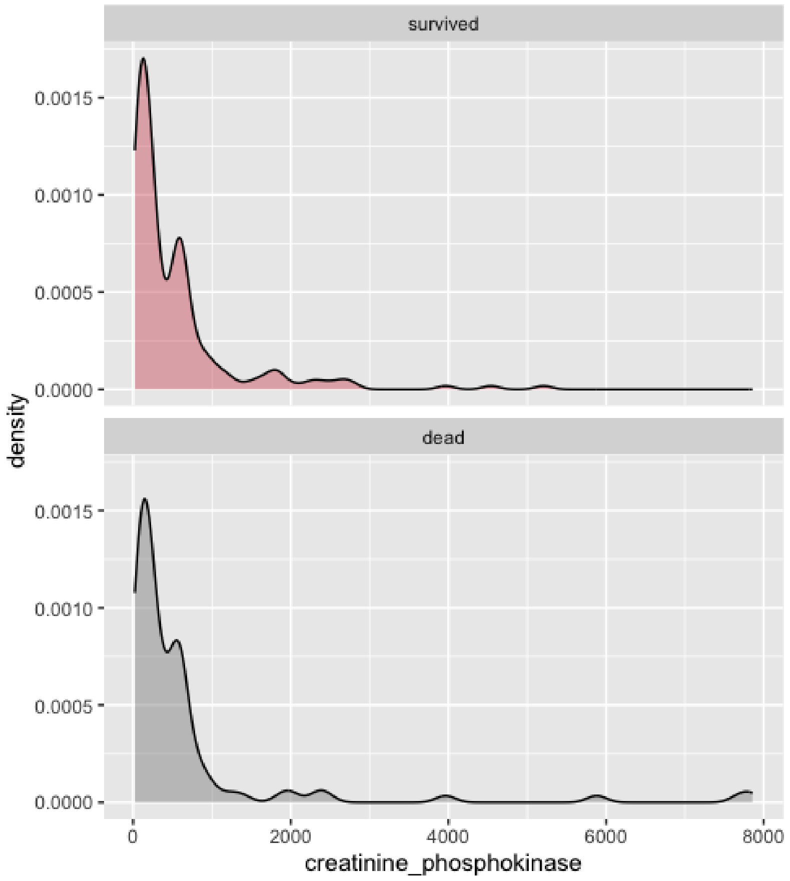
    $\mu_{CPK\_died} \neq \mu_{CPK\_survived}$ (two-sided test) or

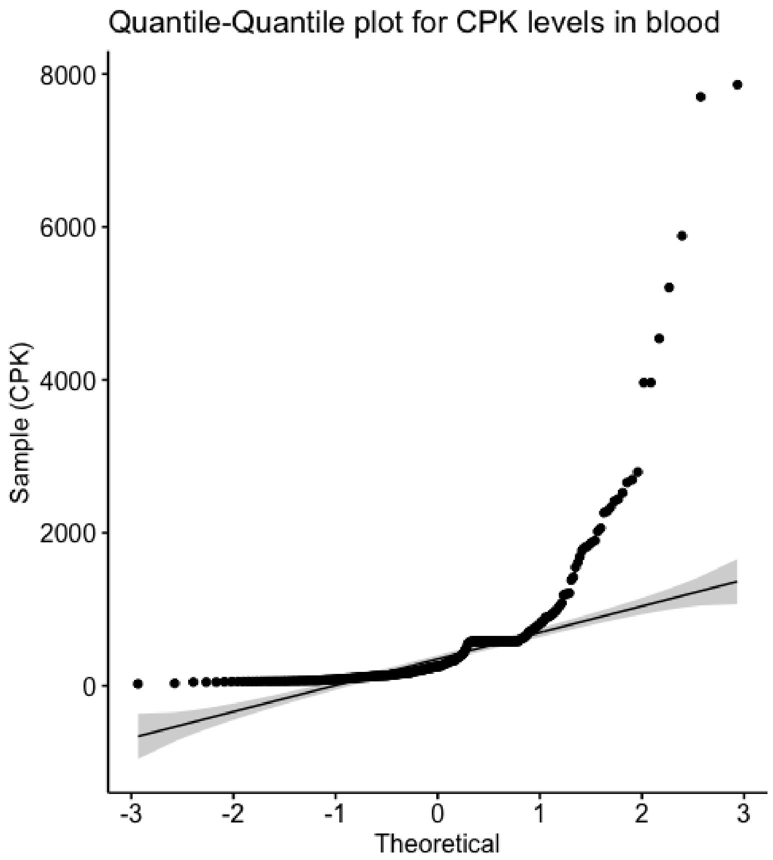
    $\mu_{CPK\_died} < \mu_{CPK\_survived}$

Example F (preliminary check): visually check the "normality" assumption for parametric testing

We can graphically confirm that CPK is not normally distributed by using:

1. the **density plot**, in which we can see the distribution is not bell shaped
2. the **QQ plot** (or Quantile-Quantile plot) for large samples – data points should roughly fall along a straight diagonal line when the dataset follows a normal distribution.





Example F (preliminary check): confirming a “normality” assumption violation with tests

- We can confirm that CPK is not normally distributed by using tests for normality:
  1. **Shapiro-Wilk test**
  2. **Kolmogorov-Smirnov test**
- The null hypotheses are defined as:
  - H_0: CPK distribution is normal
  - H_1: CPK distribution is not normal
- The test has been run over each group:
  - Evidently, we reject the H_0 !

Shapiro-Wilk Normality Test results

variable	DEATH EVENT	statistic	p
creatinine phosphokinase	survived	0.6277141	0.000000000000000000008509006
	died	0.4392427	0.000000000000000019922527779

Example F (step 2): compare independent samples using a non parametric test

- Since we verified that the explained variable (CPK) is NOT normally distributed we will run a Wilcoxon Rank Sum test (equivalent to the Mann-Whitney U test) to compare two independent samples
    - It is considered to be the nonparametric equivalent to the two-sample independent t-test
- ASSUMPTIONS :
    - Ordinal or Continuous dependent variable: the variable you’re analyzing, e.g. CPK levels ✅
    - Independence: All of the observations from both groups are independent of each other ✅
    - Shape: The shapes of the distributions for the two groups are roughly the same ✅

Wilcoxon Rank Sum Test (two-sided alternative)	
DATA	creatinine_phosphokinase by DEATH_EVENT_f
Statistic	W = 9460
	p-value = 0.684

RESULTS: since the test statistic is W = 9460 and the corresponding p-value is 0.684 > 0.05, we fail to reject the null hypothesis.

We do not have sufficient evidence to say that CPK levels for dead patients is less than that of survived patients ($\mu_{CPK\_died} < \mu_{CPK\_survived}$)

UNMET ASSUMPTION IV)

The variances of the two groups are not homogeneous

EXAMPLE G: using t test with the Welch correction

Example G (step 1): Compare two independent sample when the variance is not homogeneous

[Once again we use HEART FAILURE dataset but look at the levels of ***Serum sodium*** *in the blood, a mineral that serves for the correct functioning of muscles and nerves]*

- GOAL: verify if the difference in **Serum sodium levels in the blood** of the survivors versus those who died after heart failure is statistically significant or only due to sampling error.
- From the sample I get:
- $\bar{x} = 136$ milliequivalents per liter (mEq/L) the general sample mean (with $s = 4.41$)
- $\bar{x}_{survived} = 137$ sample mean for group of survived patients (with $s_{survived} = 3.98$)
- $\bar{x}_{died} = 135$ sample mean for group of dead patients (with $s_{died} = 5.00$)

- MORE FORMALLY: I want to run a test to verify whether my sample' statistics represent an actual difference in the respective hypothetical populations (H_a) or if there is no difference between the two hypothetical populations (H_0)
- H_0 : there is no difference in mean serum sodium between patients who suffered heart failure and died versus patients who survived after heart failure

$\mu_{died} = \mu_{survived}$ → hypothesis of no effect or ("no difference")

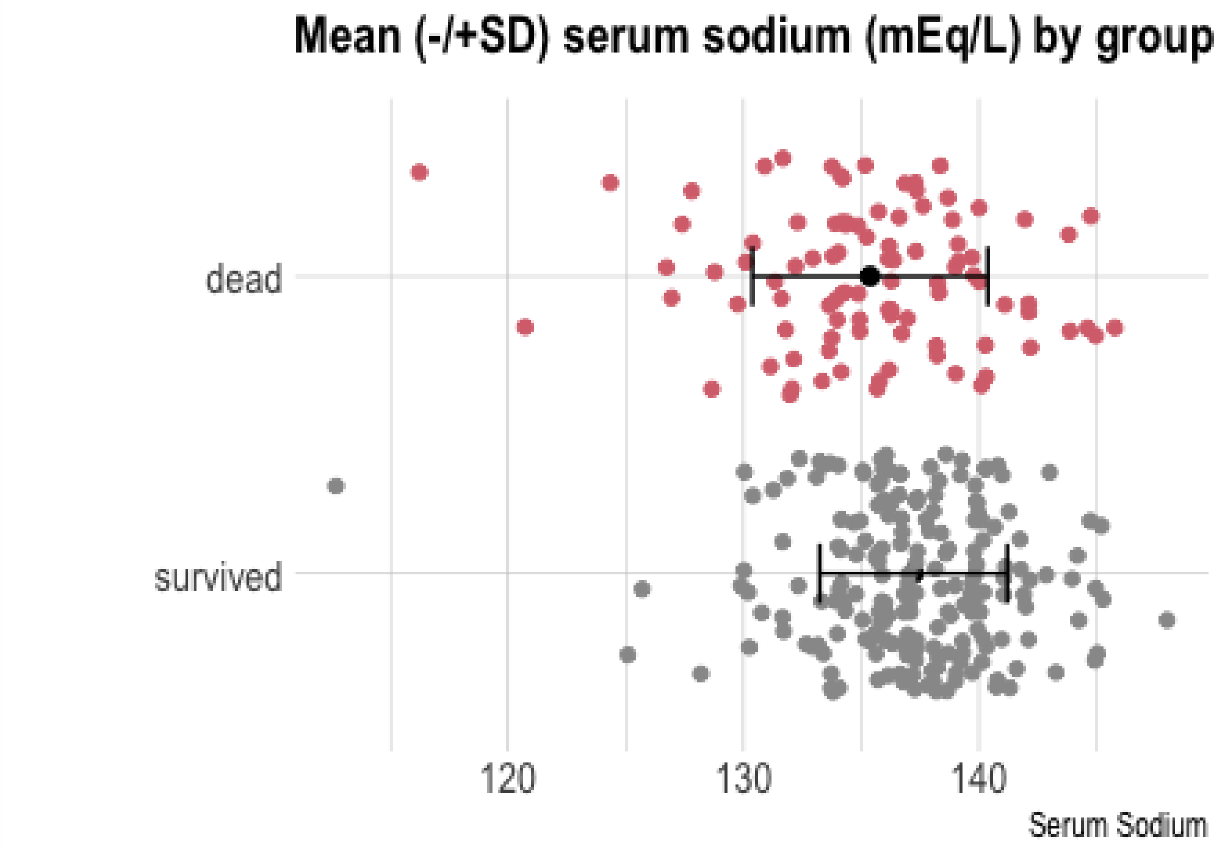
- H_a : there is a difference in mean serum sodium between patients who suffered heart failure and died versus patients who survived after heart failure ("some effect").

$\mu_{died} \neq \mu_{survived}$ (two-sided test) or

$\mu_{died} < \mu_{survived}$

Example G (preliminary check): visually check the "HOV" assumption for parametric testing

- Once again, plotting the data offers a graphical intuition that the variance of observations in the two groups seem not homogenous
  - recall that $\sigma = sd^2$



Example G (preliminary check): confirming a “HOV” assumption violation with tests

- It is always best to use an actual test to confirm this intuition. So in this case we can use the **Fisher's F test** to verify equal variances of Serum Sodium concentration in the two groups
- The null hypotheses are defined as:
  - H_0: $\sigma^2_{died} = \sigma^2_{survived}$ ➔ $\sigma^2_{died}/\sigma^2_{survived} = 1$ (The true ratio of variances is equal to 1)
  - H_1: $\sigma^2_{died} \neq \sigma^2_{survived}$ ➔ $\sigma^2_{died}/\sigma^2_{survived} \neq 1$ (The true ratio of variances *not* equal to 1)

F test to compare two variances			
variable	**groups**	**statistic**	**p-value**
Serum Sodium	survived	F = 1.5769, num df = 95, denom df = 202	p-value = 0.007646
	died		

- I use the P-value as a decision rule, which leads me to reject the null hypothesis since it is lower than 0.05 conventional alpha level

Example G (step 2): compare independent samples using a specification of the t test

- Since we verified that the explained variable (Serum Sodium in the blood) is NOT homogeneous in variance, we will run a t test with the Welch correction to compare two independent samples
  - Unequal variance is compensated by lowering the degrees of freedom

Welch Two Sample t-test	
DATA	serum_sodium by DEATH_EVENT_f

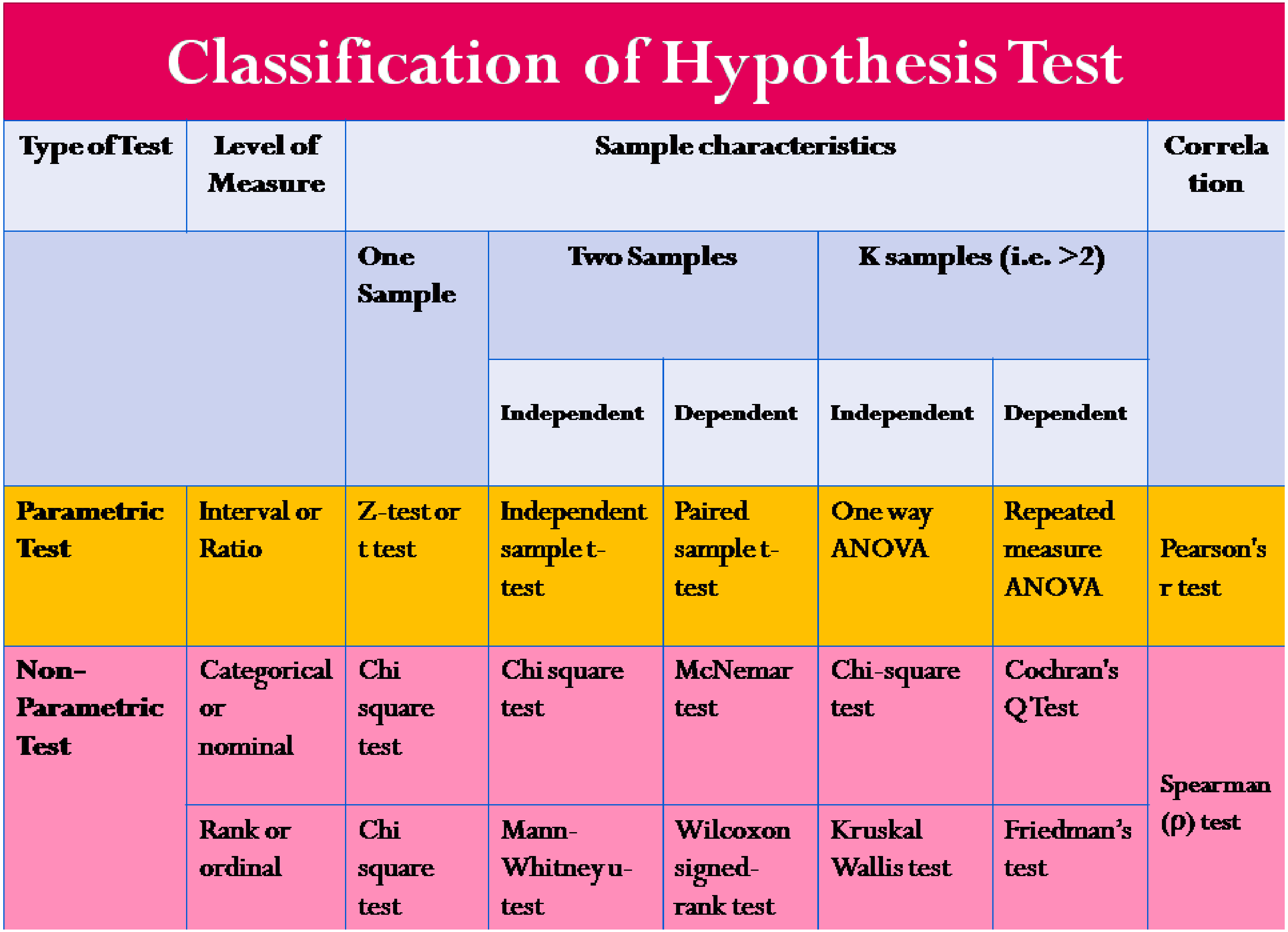
Statistic	t = -3.1645
	df = 154.01

	p-value = 0.001872

RESULTS: since the test statistic is t = -3.1645 (with df = 154.01) and the corresponding p-value is 0.001872 < 0.05, we reject the null hypothesis.

We therefore have sufficient evidence to say that the level of serum sodium levels for dead patients is significantly different than that of survived patients

$$\mu_{serum\_died} \neq \mu_{serum\_survived}$$



Classification of Hypothesis Test

Type of Test	Level of Measure	Sample characteristics					Correlation
		One Sample	Two Samples		K samples (i.e. >2)		
			Independent	Dependent	Independent	Dependent	
Parametric Test	Interval or Ratio	Z-test or t test	Independent sample t-test	Paired sample t-test	One way ANOVA	Repeated measure ANOVA	Pearson's r test
Non-Parametric Test	Categorical or nominal	Chi square test	Chi square test	McNemar test	Chi-square test	Cochran's Q Test	Spearman (ρ) test
	Rank or ordinal	Chi square test	Mann-Whitney u-test	Wilcoxon signed-rank test	Kruskal Wallis test	Friedman's test	