GATE

by moyyn

Data Analyst Module 3

Dr. Kavitha Chetana Didugu

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law
Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

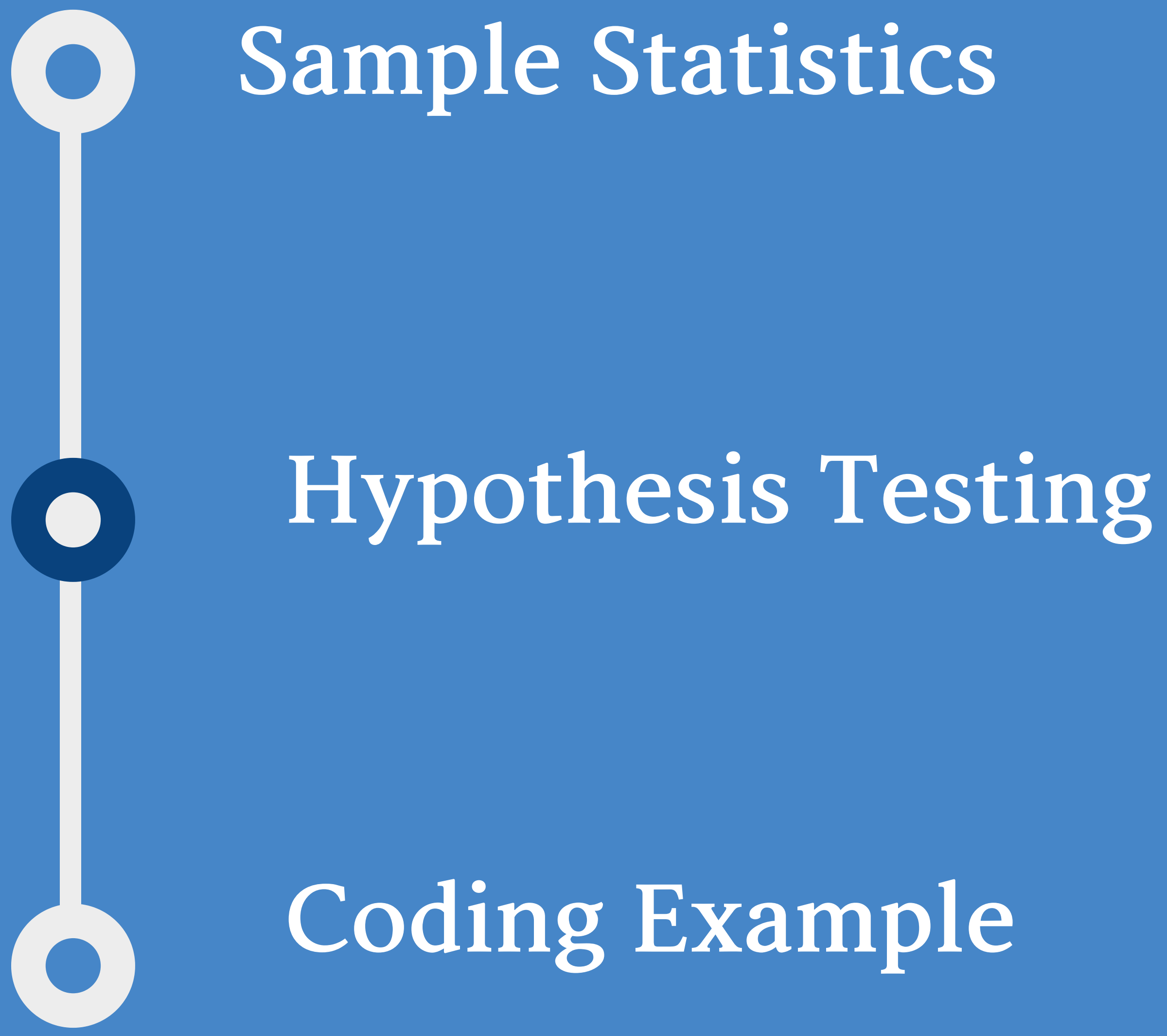
STARTUP SCHOOL
by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin
Der Regierende Bürgermeister von Berlin
Senatskanzlei
Wissenschaft und Forschung

Today's Agenda

Back to Basics



Sample Statistics

Hypothesis Testing

Coding Example

Sample Statistics and Population Parameters

Central Limit Theorem and Confidence Intervals

Need for sampling

Need for sampling:

- Time factor
- Effort factor

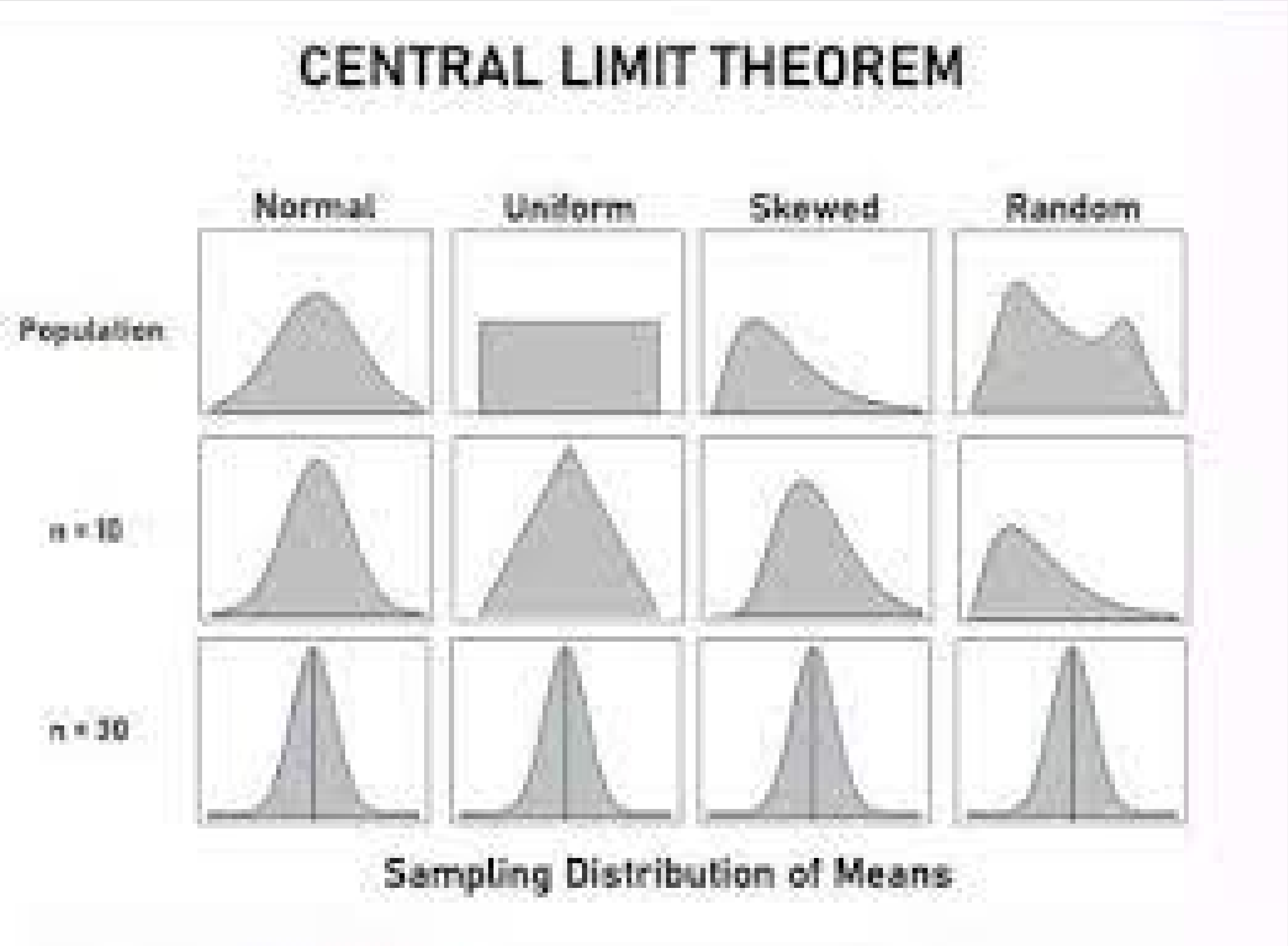
It is usually not feasible to make a complete survey of an entire population because of time and budget constraints. Therefore, a sample of the population is collected, and analysed to make inferences about the whole population.

Ideally, the goal of this type of sampling is to collect data that is representative of the entire population of interest.

But to be able to come to the right conclusions about your population, you need the sample statics to match population parameters as closely as possible

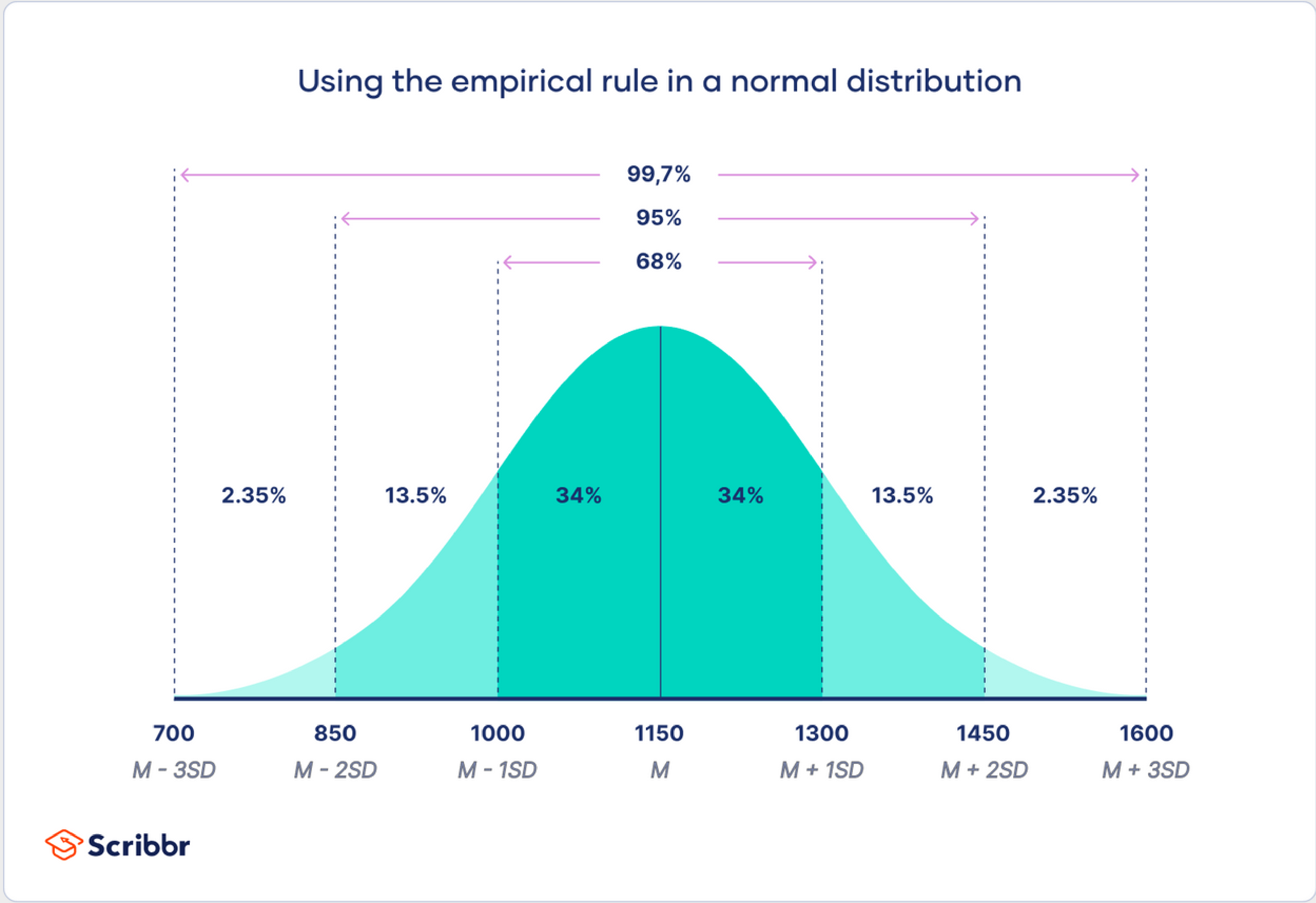
Central Limit Theorem

- Sampling Distribution of the mean of any independent random variable will be normal
- This applies to both discrete and continuous distributions.
- The random variable should have a well defined mean and variance (standard deviation).
- Applicable even when the original variable is not normally distributed.
- Sample size >= 30



Why Normal Distribution?

- Many naturally occurring phenomena follow normal distribution, test scores, height, weight, etc.
- The empirical rule, or the 68-95-99.7 rule: Around 68% of values are within 1 standard deviation (sigma) from mean. Around 95% of values are within 2 sigmas from mean. Around 99.7% of values are within 3 sigmas from mean.
- Owing to ubiquitous normality, we can make use of its properties to compare different samples, or sample vs population


Using the empirical rule in a normal distribution

Hypothesis Testing

Terminology

What is a Hypothesis?

An assumption about certain characteristics of a population

- **Null hypothesis (H0)**

  The hypothesis that does not challenge the status quo

- **Alternative hypothesis (Ha)**

  The hypothesis that challenges the status quo

Test Statistic

Given a sample, you want to know how similar/dissimilar is it to your reference sample/population, based on a particular measure (mean, variance, standard deviation, etc.) This measure is called TEST STATISTIC.

- A test statistic is a random variable that is calculated from sample data and used in a hypothesis test.

- A test statistic describes how closely the distribution of your data matches the distribution predicted under the null hypothesis

Examples for Intuition

In a factory that manufactures nuts, you are testing quality of the nuts your team produced today. Ideally, you are supposed produce nuts of average length of 20mm. Any longer or shorter nuts result in bad quality. Use this sample to test whether the nuts produced

- test statistic: average nut length
- sample: sample of nuts for QC
- reference: length of a nut as per guideline
- H0: average nut length of sample = average nut length as per guideline
- Ha: average nut length of sample <> average nut length as per guideline

Examples for Intuition

You work in an e-commerce company, and you want to improve the conversion rate on your products. You propose to change the purchase button from red to green, because your research teams says having a green button increases customer’s intention to purchase. So you split your incoming traffic between red and green button pages, and observe conversion rate for 2 weeks.

- test statistic: average conversion rate over 2 weeks
- sample 1: conversion rate with green button
- sample 2: conversion rate with red button
- H0: average conversion with green button <= average conversion with red button
- Ha: average conversion with green button > average conversion with red button

Types of Hypothesis Tests

- Single sample
  - You have only one sample, and you are testing it against a reference value
- Two or multiple samples
  - You are testing the closeness of test statistic across two or more different samples
- One tailed or two tailed
  - One-tailed: Greater than, Less than, Greater Than or Equal to, Less than or Equal to
  - Two-tailed: Not equal to
- Tests of Mean, Proportion, or Variance
  - Is the test statistic the mean, the proportion or the variance of the sample

Errors in Hypothesis Testing

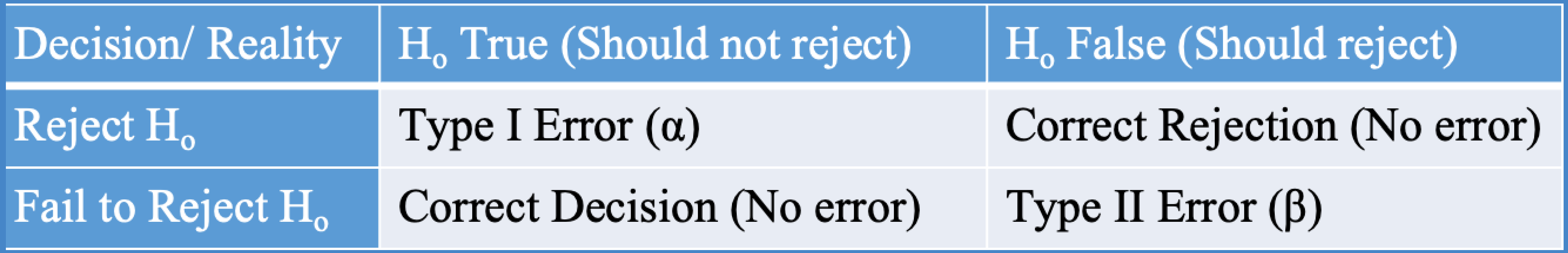
Type I and Type II Errors

Type I Error:

- Rejection of null hypothesis when it should not have been rejected.
- Incorrectly rejecting the null hypothesis.

Type II Error:

- Failure to reject the null hypothesis, when it should have been rejected.
- Incorrectly not rejecting the null hypothesis.



Decision/ Reality	H_o True (Should not reject)	H_o False (Should reject)
Reject H_o	Type I Error (α)	Correct Rejection (No error)
Fail to Reject H_o	Correct Decision (No error)	Type II Error (β)

How are Error, CLT and Hypothesis Testing Related?

Why are we talking about CLT: Confidence Intervals

A Confidence Interval tells you the range of values within which the parameter of interest has the highest chance of falling

If we repeated a certain experiment a large number of times (take 10000 different samples of nuts and do quality control on each), 90% Confidence Interval would tell us between what range of values the mean of the sample would lie within, in 90% of the time.

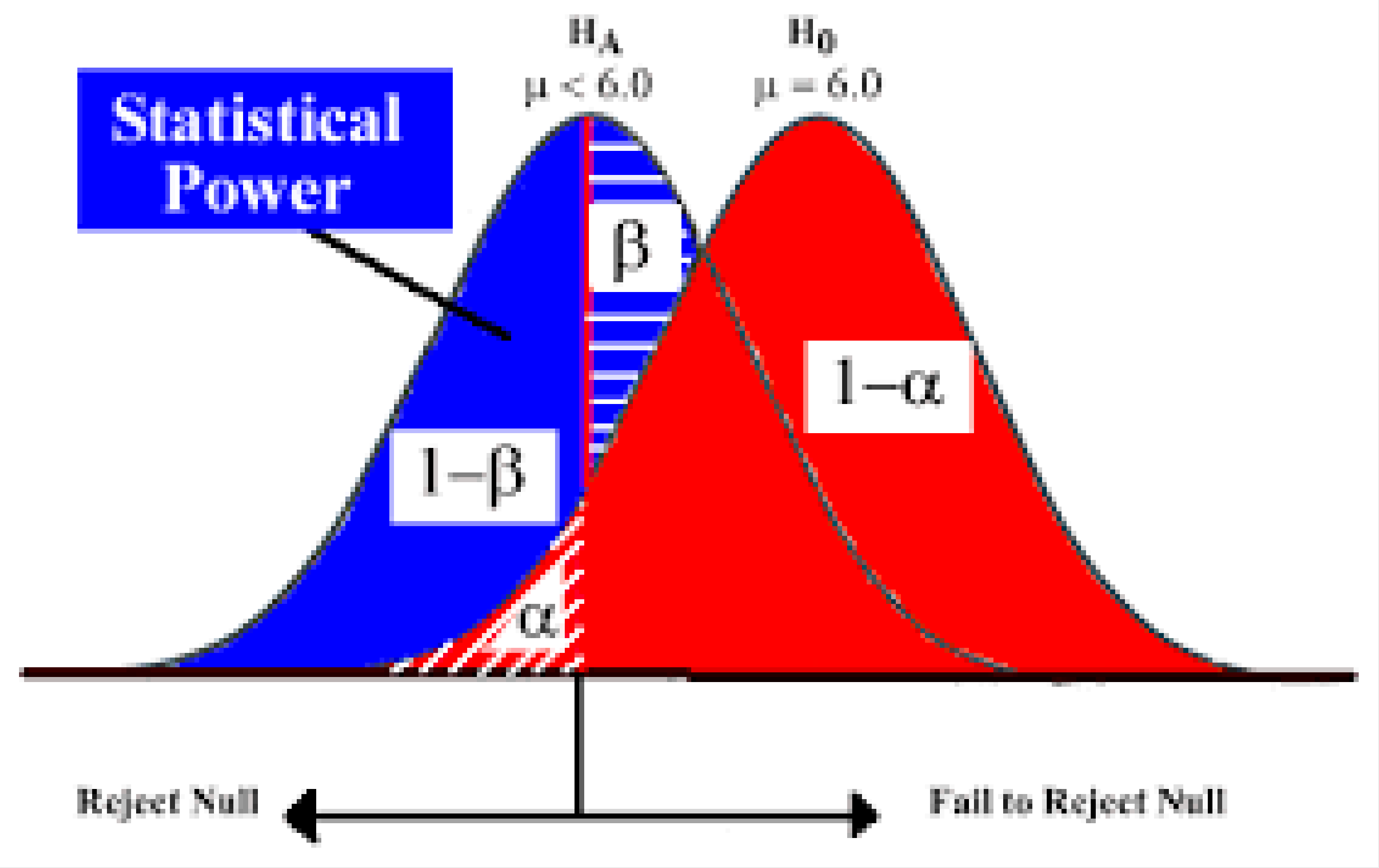
A 95% Confidence interval means that 95% of all sample means ($\bar{x}$) are hypothesized/expected to be in this region

Visualising Error in Hypothesis Test

Ideally, Null Hypothesis and Alternate Hypothesis are opposites of each other: Both cannot be true at once (the average nut length in the sample is either 20mm or not).

Power (1-Beta) refers to the likelihood of a hypothesis test detecting a true effect if there is one. A statistically powerful test is more likely to reject a false negative (Type II error or Beta)

Significance level (Alpha) is the probability of rejecting the null hypothesis even when it is true (Type I error). 5% significance level is a common choice for statistical test.

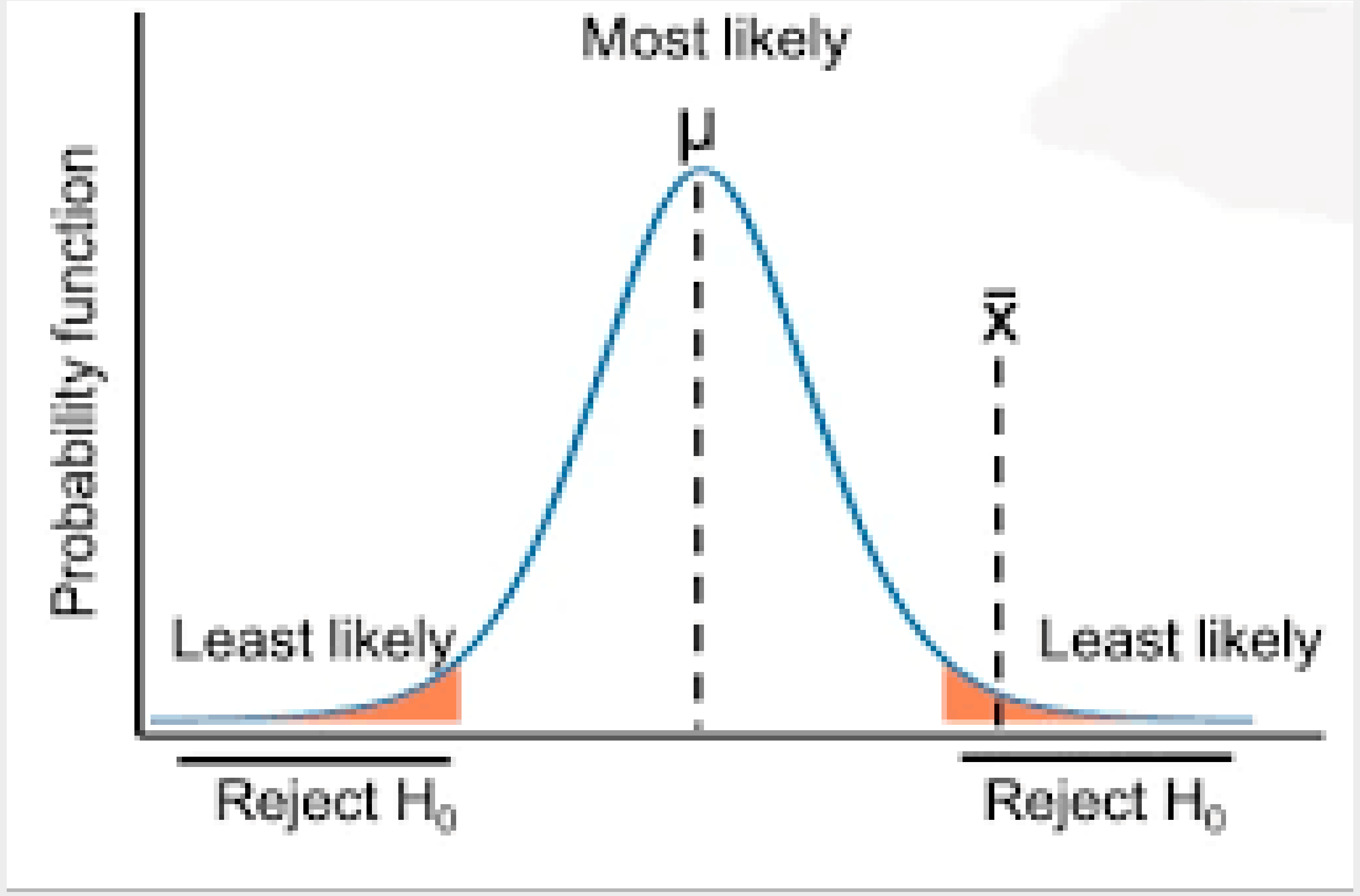


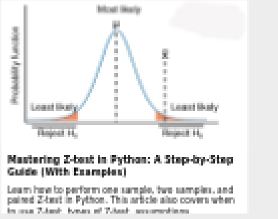
In a Two-tailed Situation

- Here, α = 0.05, the level of significance or our tolerance level towards making a Type I error.

- If the null hypothesis is correct, (α * 100)% of the sample means should lie in the rejection region.

- If sample mean is in the white region, we fail to reject the null hypothesis
- If sample mean is in the orange region, we reject the null hypothesis.



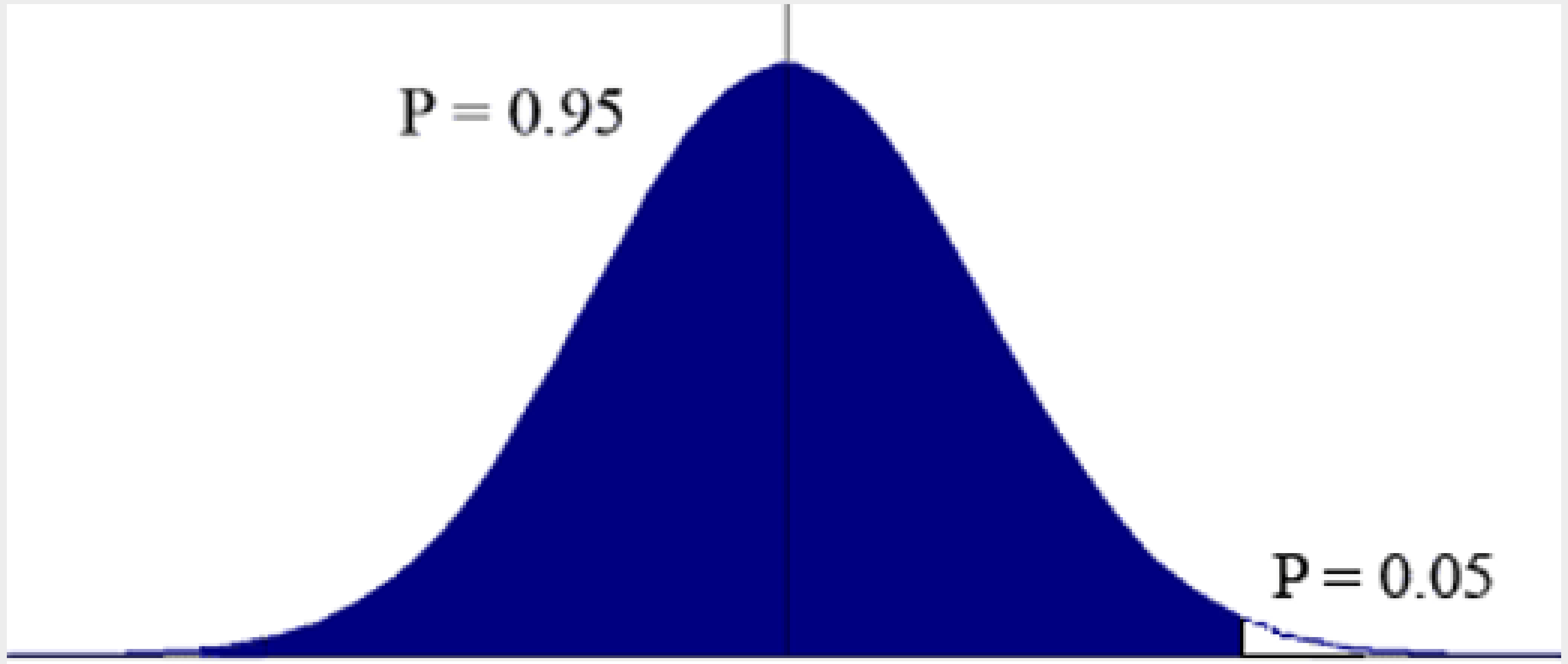


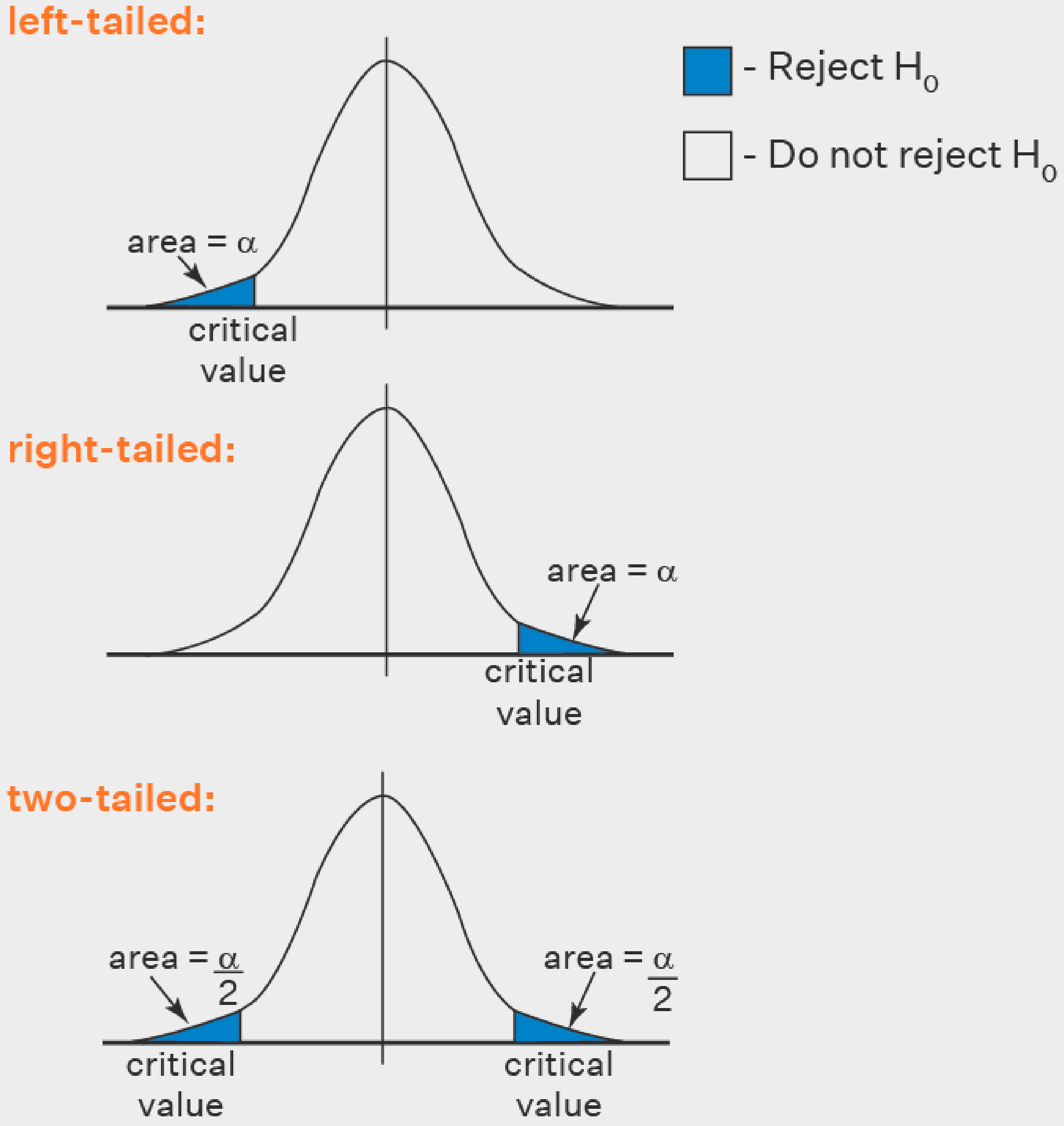
Mastering Z-test in Python: A Step-by-Step Guide (With Examples)

Learn how to perform one sample, two samples, and paired Z-test in Python. This article also covers when to use Z-test, types of Z-test, assumptions

In case of One-tailed situation

- All of α is in one tail or the other, depending on the alternative hypothesis.
- Ha points to the tail, where the critical value and the rejection region are (Ex: when observed mean > reference/hypothesised mean)



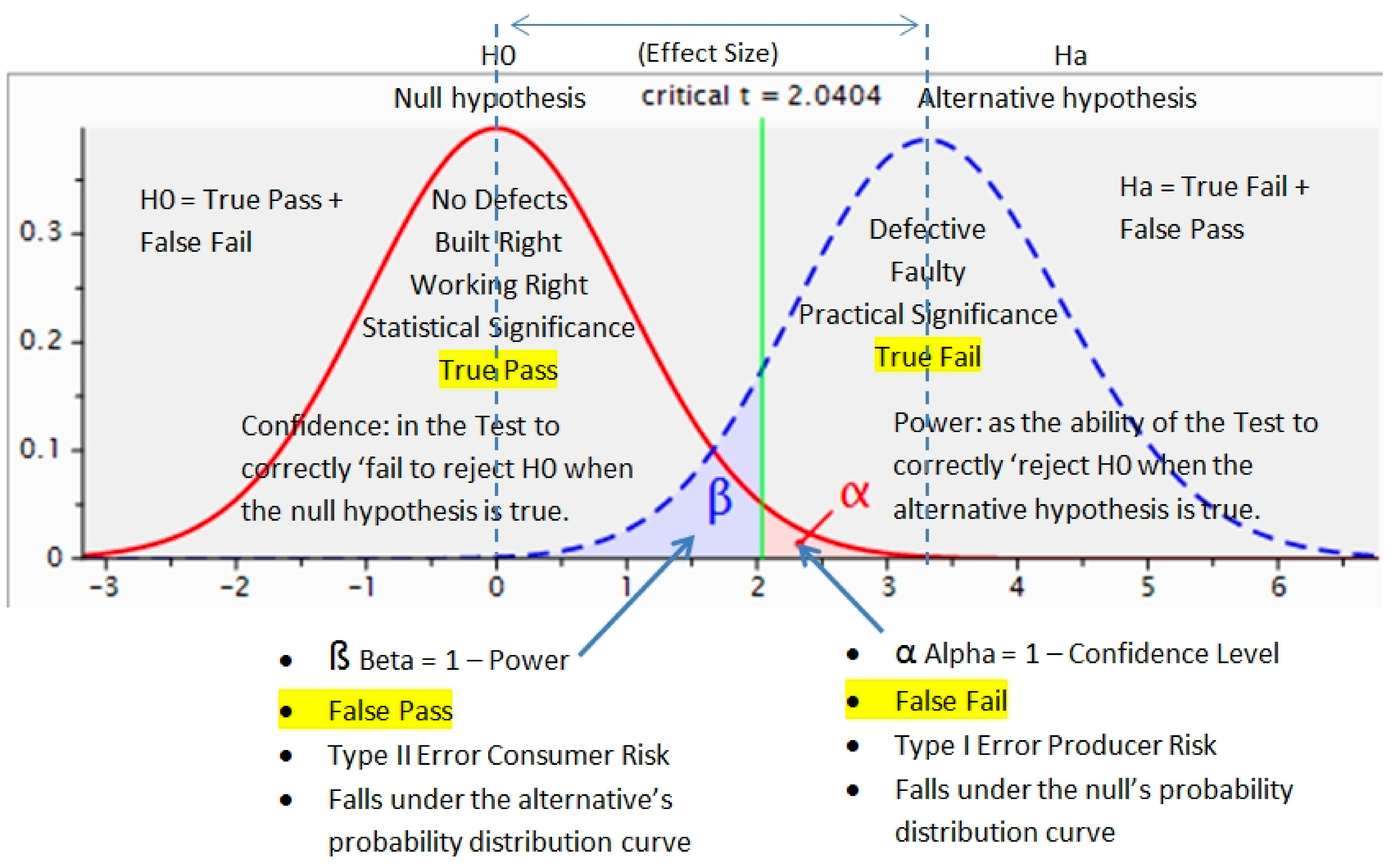


CUEMAT

Z Test - Formula, Definition, Examples, Types

Z test is a statistical test conducted on normally distributed data to check whether the means of two

Visualising Error in Hypothesis Test



H0 = True Pass + False Fail

No Defects

Built Right

Working Right

Statistical Significance

True Pass

Ha = True Fail + False Pass

Defective

Faulty

Practical Significance

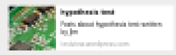
True Fail

Confidence: in the Test to correctly ‘fail to reject H0 when the null hypothesis is true.

Power: as the ability of the Test to correctly ‘reject H0 when the alternative hypothesis is true.

- ß Beta = 1 – Power
- False Pass
- Type II Error Consumer Risk
- Falls under the alternative’s probability distribution curve

- α Alpha = 1 – Confidence Level
- False Fail
- Type I Error Producer Risk
- Falls under the null’s probability distribution curve



Causes of Type I and Type II Errors

Power is the probability of avoiding a Type II error. The higher the statistical power of a test, the lower the risk of making a Type II error.

- By random chance, we may select a sample which is not representative of the population.
- Sampling techniques may be flawed.
- Assumptions in our null hypothesis may be flawed

Sample size - Statistical Significance and Power

If I could measure the entire population, would I have better power & significance?

Power

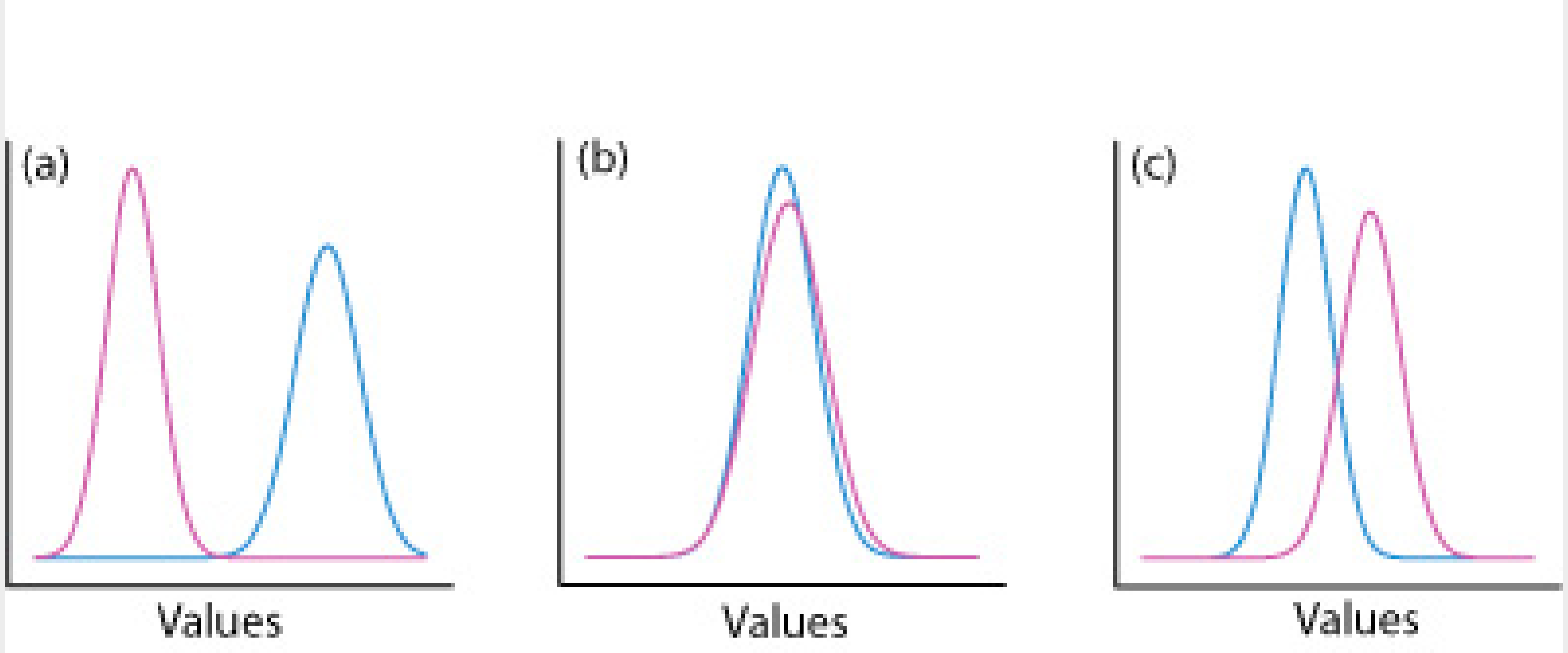
Sample size is positively related to power. A small sample (less than 30 units) may only have low power while a large sample has high power (low Beta or Type II error). Increasing the sample size enhances power, but only up to a point. When you have a large enough sample, every observation that's added to the sample only marginally increases power.

Significance Level

In order to increase significance level (lower Alpha or lower Type I error), you can increase your sample size. However, this is not always true. In fact, The larger the actual difference between the groups (ie. student test scores) the smaller of a sample we'll need to find a significant difference (ie. $p \leq 0.05$).

Sample size - Statistical Significance and Power

You have taken a sample size of 30 units in each case. In which case would you need more samples to make sure you want to reject or fail to reject the null hypothesis?



A small review

A study was done to see the effect of presence of pet dogs on kids (ages 10 to 18). Two groups of kids aged 12-16, one group with those who owned a dog for minimum 5 years and another group who never owned a dog, were asked to complete a survey and scores were computed. High score corresponds to higher happiness and low score corresponds to lower happiness.
Do dogs have a significant effect (either positive or negative) on the cheerfulness of kids?

Dog: 6, 8, 4, 7, 7, 8, 9, 9, 10, 9, 5, 7, 6, 8, 7, 5, 4
No_dog: 9, 3, 1, 2, 8, 9, 6, 7, 5, 8, 6, 7, 6, 8, 7, 6, 9

A small review

- What are the null and alternative hypothesis?
- Is it a right tailed or a left tailed test?
- Is it a one sample or a two sample test?
- Is it a test of mean, proportion or variance?
- Which statistical test do you think is appropriate?

How to perform a Hypothesis Test

When CLT holds (sample size >=30), it is the most straightforward case of hypothesis testing. sample mean follows normal distribution, and also will be close to population mean.

The z-test formula is as follows:

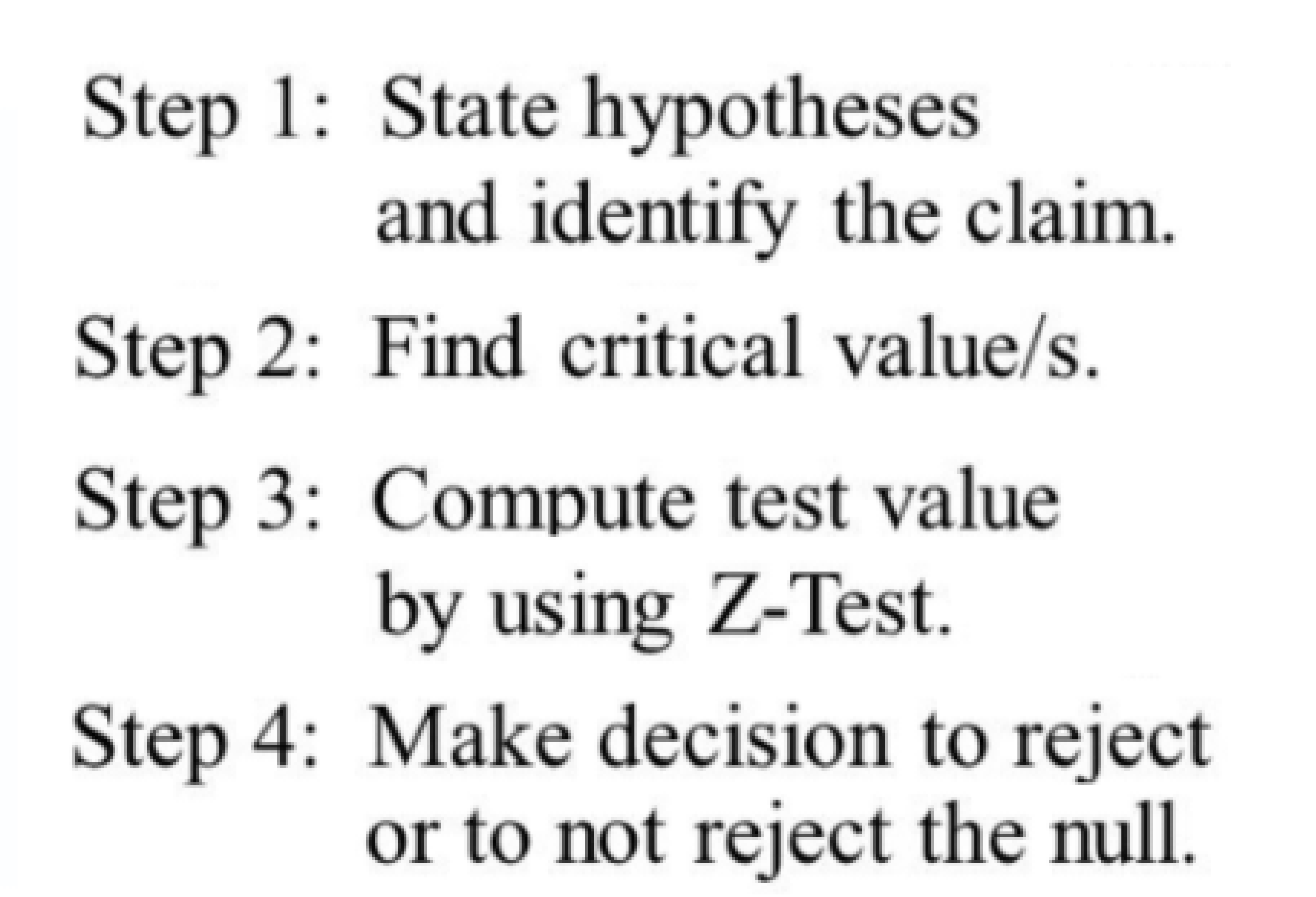
$$Z = (\bar{x} - \mu0) / (\sigma / \sqrt{n})$$

Here, $\bar{x}$ is the <u>sample mean</u>

$\mu0$ is the population mean

σ is the standard deviation

n is the sample size



Step 1: State hypotheses and identify the claim.

Step 2: Find critical value/s.

Step 3: Compute test value by using Z-Test.

Step 4: Make decision to reject or to not reject the null.

Other types of Tests

Z-test vs T-test

So far we have looked at hypothesis testing in the most straight forward case: sample size is big enough (at least 30 data points in the sample). Test statistic was sample mean, which follows a Normal distribution. This type of test is called a Z-test.

But what if we have data that is not ideal?

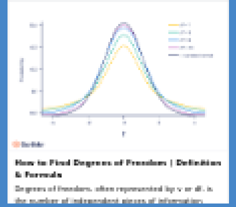
Sample size < 30

We know that if sample size is less than 30, we get a near normal distribution of sample mean, but not an exact normal. A t-test is used when the sample size is less than 30 and the population variance is unknown.

Two-Sample Test and Test of Variance

Two sample tests: What if instead of a test sample and a reference value/population parameter, you have to test whether two samples are similar to or different from each other? We will have to calculate sample variances ourselves and use it to calculate test statistics (AB-test)

Test of Variance: distribution of sample means follows normal distribution, so the z-test, t-test helps.
What about Variance? mean of sample variances follows a chi-square distribution, so we need a test based on chi-square statistics


Scribbr

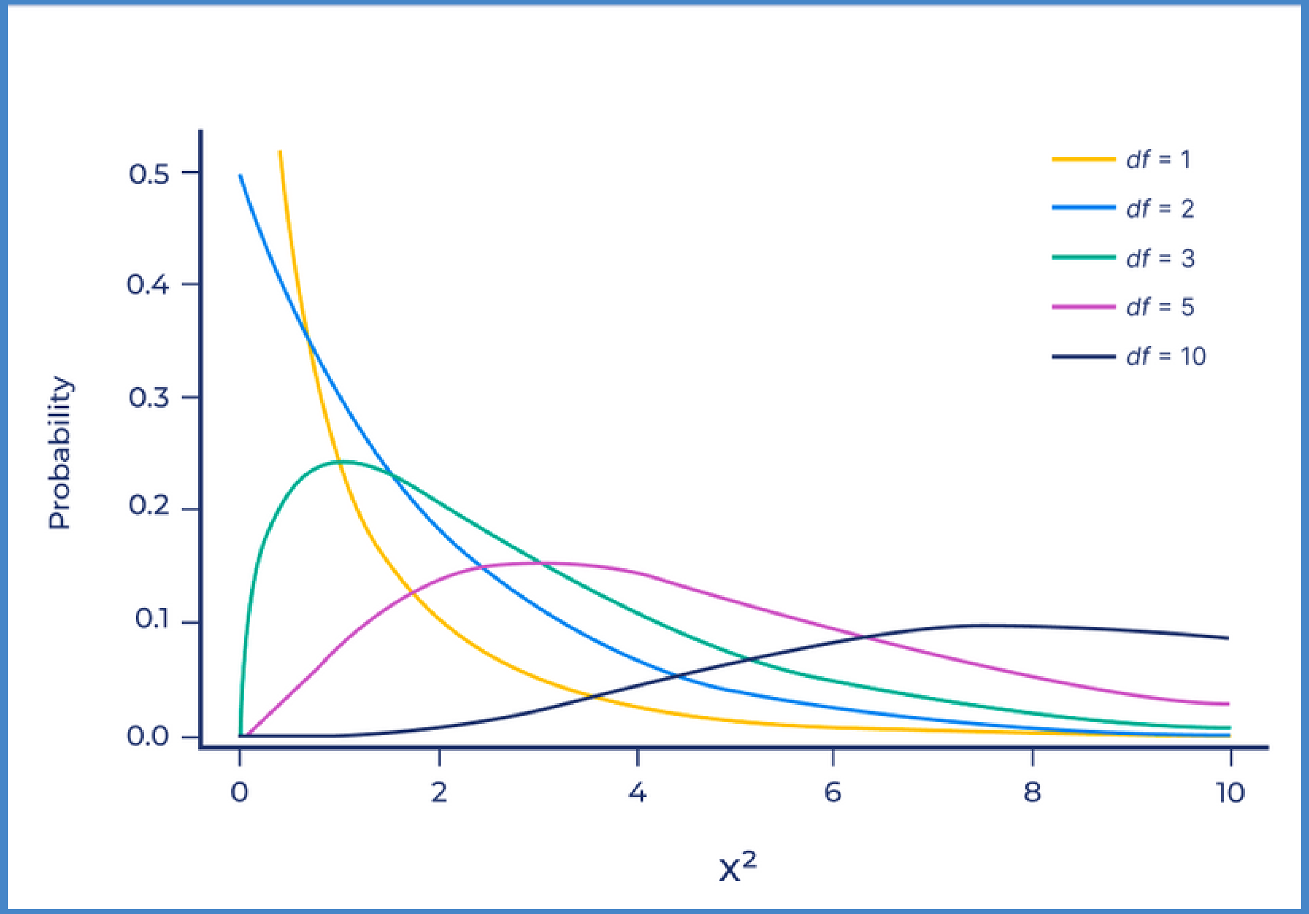
How to Find Degrees of Freedom | Definition & Formula

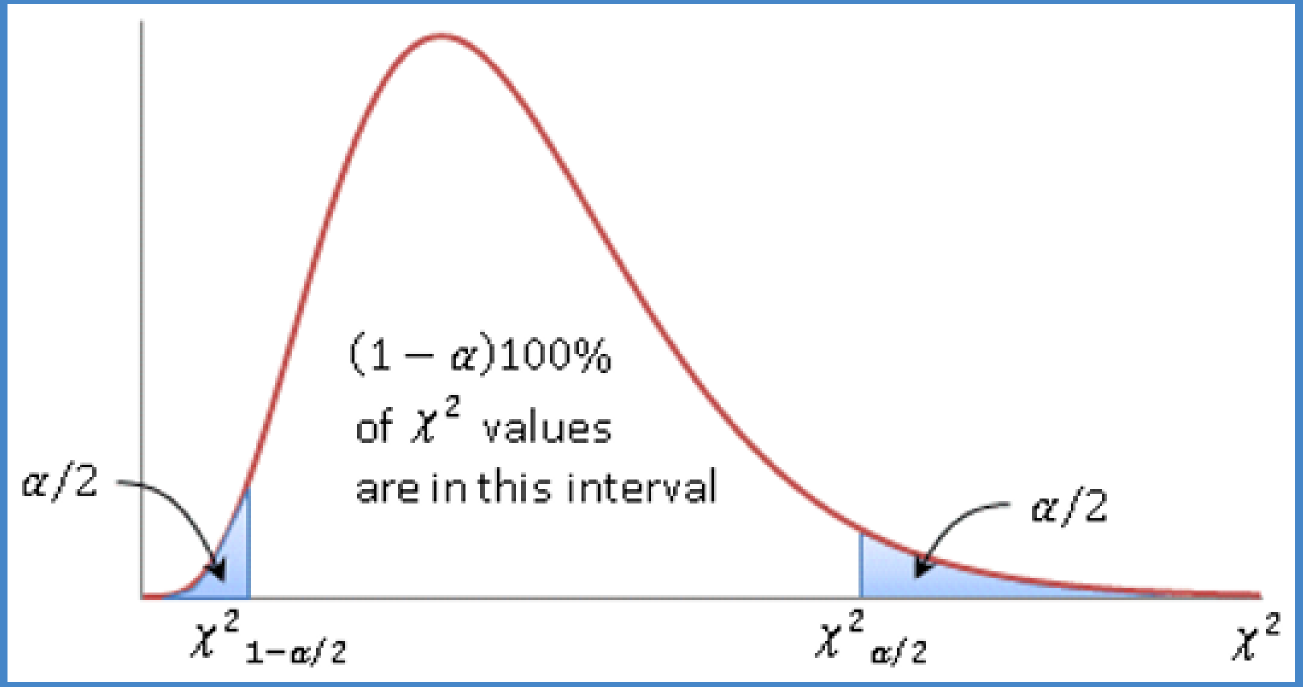
Degrees of freedom, often represented by v or df, is the number of independent pieces of information

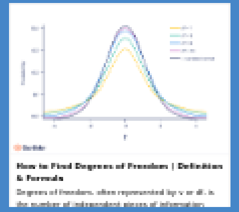
Test of Variance

When we take many samples of the same size from a normal population and find the sample variances, they DO NOT follow a normal distribution; instead they follow a chi-square (χ2) distribution , which is dependent on the degrees of freedom.

- Area under the curve is always 1.
- Cumulative Probability runs from right to left; 1 is towards the left end, while 0 is towards the right.







How to Find Degrees of Freedom | Definition & Formula

Degrees of freedom, often represented by v or df, is the number of independent pieces of information

Chi-square (χ2) Test of Variance

χ2 test compares the population variance, with the hypothesized variance

$\chi2 = (n-1)\ s2$

$\qquad \sigma2$

where, n = sample size

s2 = sample variance and σ2 = population variance (which we wish to test)

At α = 0.05 and n = 5 (df = 4)

p-value: How much of the area is above the test-statistic? (Does test statistic fall in the rejection region?)

If it is less than the specific α, we reject the null hypothesis

One-way ANOVA

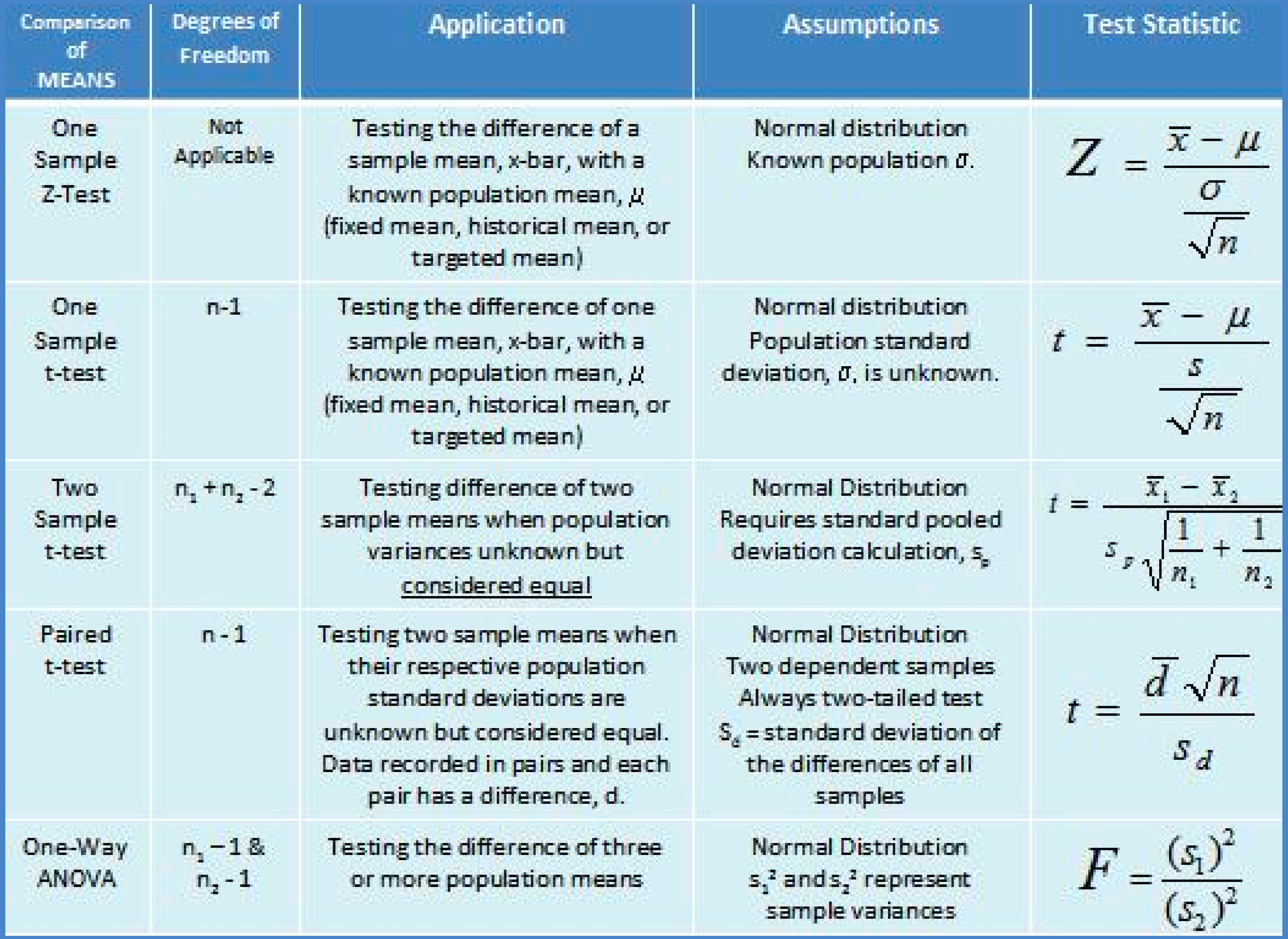
H0：All population means are equal

μ1 = μ2 = μ3 = μ4 = ⋯= μk

Ha：Not all of the population means are equal

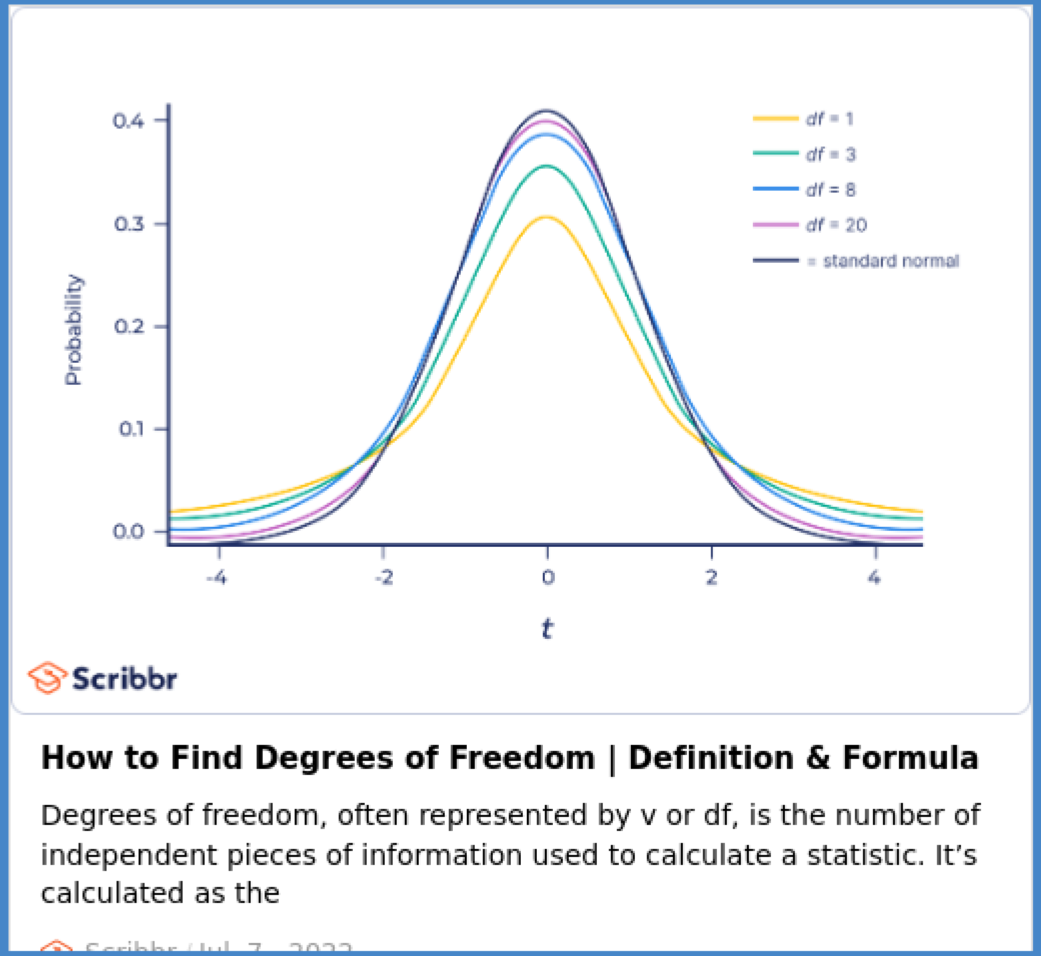
For at least one pair, the population means are unequal.

Data, Test Statistic, and Test



Comparison of MEANS	Degrees of Freedom	Application	Assumptions	Test Statistic
One Sample Z-Test	Not Applicable	Testing the difference of a sample mean, x-bar, with a known population mean, μ (fixed mean, historical mean, or targeted mean)	Normal distribution Known population σ.	$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
One Sample t-test	n-1	Testing the difference of one sample mean, x-bar, with a known population mean, μ (fixed mean, historical mean, or targeted mean)	Normal distribution Population standard deviation, σ, is unknown.	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Two Sample t-test	$n_1 + n_2 - 2$	Testing difference of two sample means when population variances unknown but considered equal	Normal Distribution Requires standard pooled deviation calculation, s_p	$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Paired t-test	n - 1	Testing two sample means when their respective population standard deviations are unknown but considered equal. Data recorded in pairs and each pair has a difference, d.	Normal Distribution Two dependent samples Always two-tailed test S_d = standard deviation of the differences of all samples	$t = \frac{\bar{d}\sqrt{n}}{s_d}$
One-Way ANOVA	$n_1 - 1$ & $n_2 - 1$	Testing the difference of three or more population means	Normal Distribution s_1^2 and s_2^2 represent sample variances	$F = \frac{(s_1)^2}{(s_2)^2}$

Degrees of freedom (df), is the number of independent pieces of information used to calculate a statistic. It’s calculated as the sample size minus the number of restrictions.



How to Find Degrees of Freedom | Definition & Formula

Degrees of freedom, often represented by v or df, is the number of independent pieces of information used to calculate a statistic. It’s calculated as the

Scribbr / Jul 7, 2022

Interaction

Code Example

Feedback Survey

http://www.moyyn.com/gate-feedback

Further Reading and Resources

- https://www.udacity.com/course/ab-testing--ud257
- https://www.scribbr.com/statistics/test-statistic/
- https://michaelminn.net/tutorials/r-hypothesis-tests/index.html
- https://educationalresearchtechniques.com/2016/02/03/type-i-and-type-ii-error/
- https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
- https://www.scribbr.com/statistics/one-way-anova/#:~:text=ANOVA%2C%20which%20stands%20for%20Analysis,ANOVA%20uses%20two%20independent%20variables.
- https://towardsdatascience.com/the-relationship-between-significance-power-sample-size-effect-size-899fcf95a76d#:~:text=hypothesis%20is%20correct.-,Statistical%20Power,p%20%E2%89%A4%200.05)
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.six-sigma-material.com%2Ft-distribution.html&psig=AOvVaw3LS1BvSZovudl83EWorYRB&ust=1699119757751000&source=images&cd=vfe&opi=89978449&ved=0CBUQ3YkBahcKEwiwyaTCsKiCAxUAAAAAHQAAAAAQDw
- https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-quantitative-means/two-sample-t-test-means/v/two-sample-t-test-for-difference-of-means
- https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_power/BS704_Power6.html
- https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/

GATE

by moyyn

German Academy for Technology and Entrpreneurship

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

STARTUP SCHOOL
by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin

Der Regierende Bürgermeister
von Berlin
Senatskanzlei
Wissenschaft und Forschung