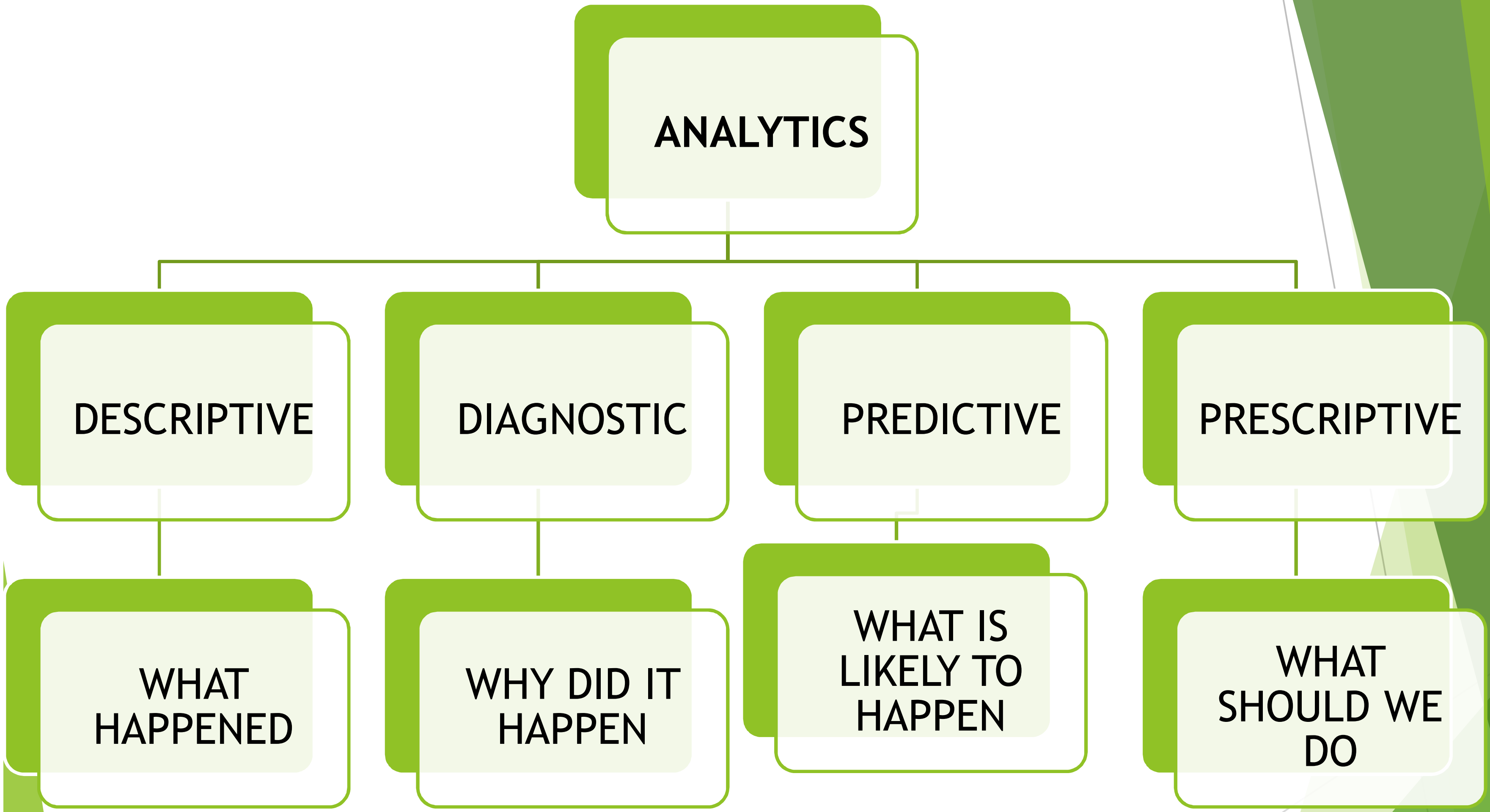
Statistics
for
Data Science and Data Analysis

As it tells the health of Data.



ANALYTICS

DESCRIPTIVE

WHAT HAPPENED

DIAGNOSTIC

WHY DID IT HAPPEN

PREDICTIVE

WHAT IS LIKELY TO HAPPEN

PRESCRIPTIVE

WHAT SHOULD WE DO

Inferential Statistics

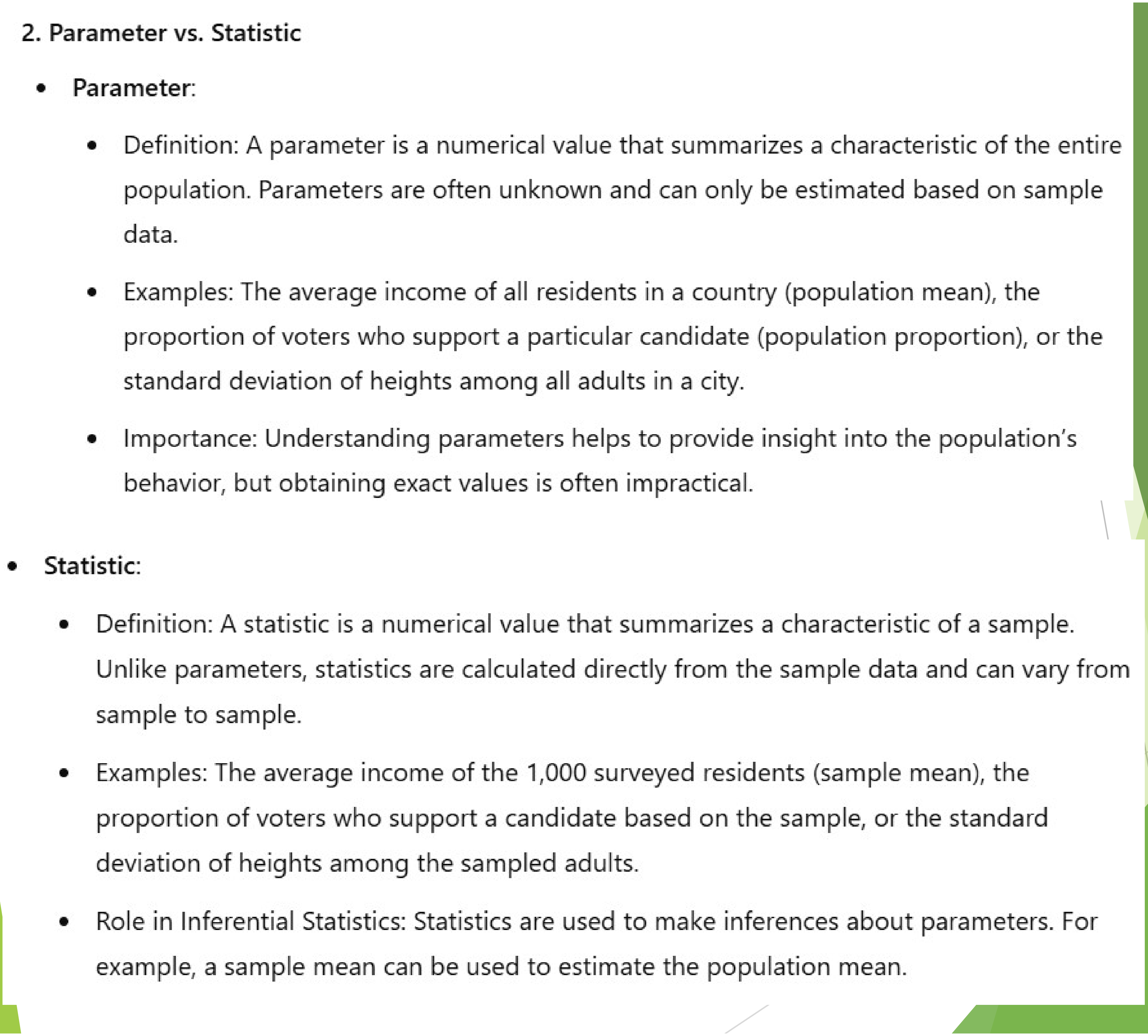
Definition: Inferential statistics is a branch of statistics that focuses on using data collected from a sample to make generalizations, predictions, or conclusions about a larger population. This approach allows researchers to infer characteristics of the whole population without having to collect data from every individual in that group.

Purpose: The primary purpose of inferential statistics is to draw conclusions about a population based on sample data. It is particularly useful when it is impractical or impossible to collect data from the entire population due to constraints such as time, cost, or accessibility.

Key Concepts in Inferential Statistics

1. Population vs. Sample

- **Population**:
  - Definition: The population refers to the entire group of individuals or observations that are of interest in a statistical study. This group can be large and diverse, encompassing every member that meets certain criteria.
  - Examples: All residents of a country, all students in a school, or every manufactured item in a production line.
  - Characteristics: Parameters are used to describe populations. These parameters are often unknown and need to be estimated using sample data.
- **Sample**:
  - Definition: A sample is a subset of the population selected for analysis. The sample is intended to represent the larger population and is used to draw conclusions about it.
  - Examples: A survey of 1,000 residents in a country, a group of 50 students from a school, or 100 items selected from a production line.
  - Characteristics: Statistics are used to describe samples. These statistics are calculated from the data collected in the sample and are used to estimate population parameters.

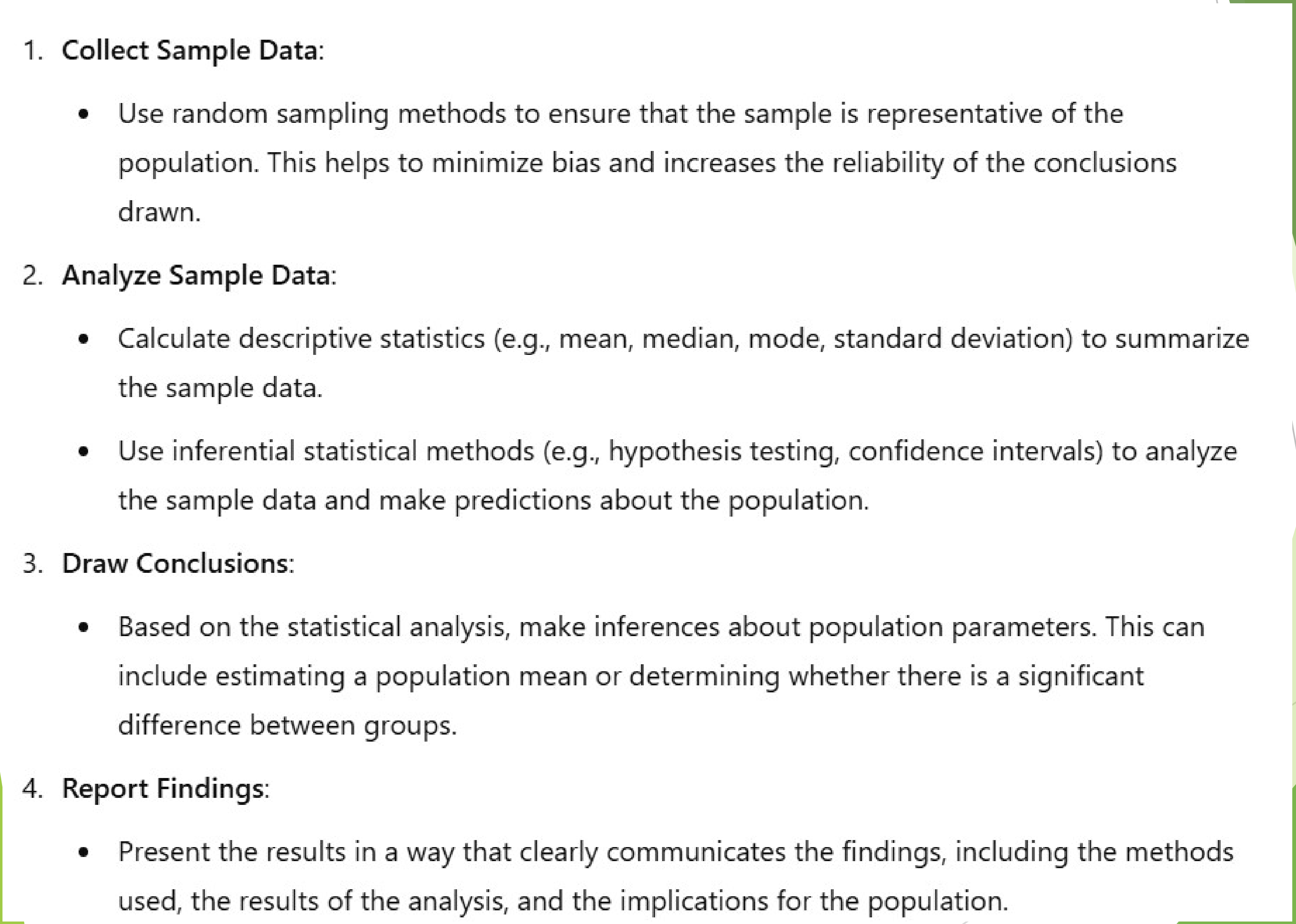


2. Parameter vs. Statistic

- Parameter:
  - Definition: A parameter is a numerical value that summarizes a characteristic of the entire population. Parameters are often unknown and can only be estimated based on sample data.
  - Examples: The average income of all residents in a country (population mean), the proportion of voters who support a particular candidate (population proportion), or the standard deviation of heights among all adults in a city.
  - Importance: Understanding parameters helps to provide insight into the population’s behavior, but obtaining exact values is often impractical.

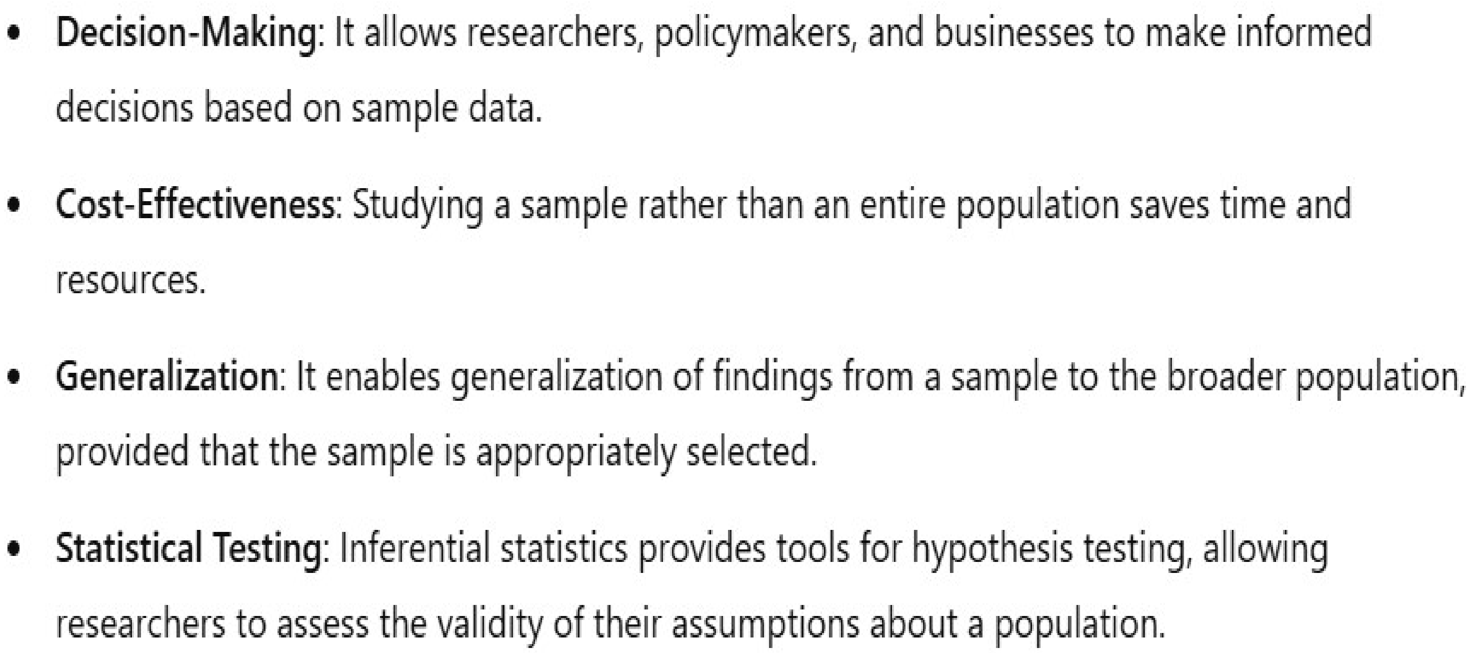
- Statistic:
  - Definition: A statistic is a numerical value that summarizes a characteristic of a sample. Unlike parameters, statistics are calculated directly from the sample data and can vary from sample to sample.
  - Examples: The average income of the 1,000 surveyed residents (sample mean), the proportion of voters who support a candidate based on the sample, or the standard deviation of heights among the sampled adults.
  - Role in Inferential Statistics: Statistics are used to make inferences about parameters. For example, a sample mean can be used to estimate the population mean.

The Process of Inferential Statistics



1. **Collect Sample Data**:
   - Use random sampling methods to ensure that the sample is representative of the population. This helps to minimize bias and increases the reliability of the conclusions drawn.
2. **Analyze Sample Data**:
   - Calculate descriptive statistics (e.g., mean, median, mode, standard deviation) to summarize the sample data.
   - Use inferential statistical methods (e.g., hypothesis testing, confidence intervals) to analyze the sample data and make predictions about the population.
3. **Draw Conclusions**:
   - Based on the statistical analysis, make inferences about population parameters. This can include estimating a population mean or determining whether there is a significant difference between groups.
4. **Report Findings**:
   - Present the results in a way that clearly communicates the findings, including the methods used, the results of the analysis, and the implications for the population.

Importance of Inferential Statistics



- **Decision-Making**: It allows researchers, policymakers, and businesses to make informed decisions based on sample data.
- **Cost-Effectiveness**: Studying a sample rather than an entire population saves time and resources.
- **Generalization**: It enables generalization of findings from a sample to the broader population, provided that the sample is appropriately selected.
- **Statistical Testing**: Inferential statistics provides tools for hypothesis testing, allowing researchers to assess the validity of their assumptions about a population.

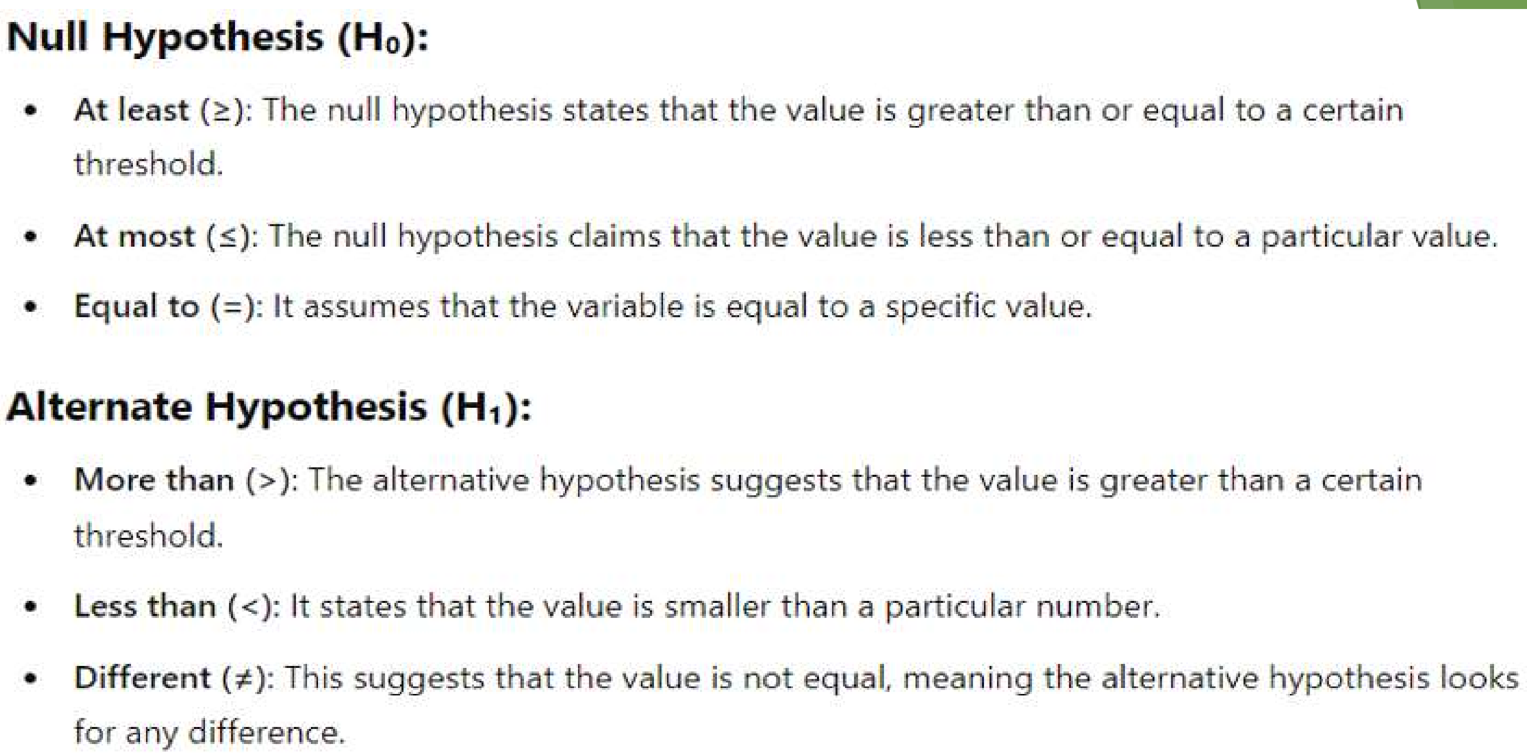
Hypothesis Testing:

- **Hypothesis testing:** In context of Data analytics, It talks about testing claim and make decisions about significance of relation between dependent and independent variables or We can say 'Independent variable' is statistically significant or not with 'dependent variable' is known as hypothesis.
- It starts with taking claim as Alternate hypothesis and any thing against claim as Null hypothesis.
  - **Null Hypothesis (H0):** Independent variable does not influence dependent variable.
  - **Alternative Hypothesis (H1 or Ha):** Independent variable does influence dependent variable.

- Than depending upon nature of variables perform appropriate test to check the validity of claim: Z-test, t-test, Chi-Square test, ANOVA with in certain Confidence interval and corresponding Significance level.
- Than P-value helps to make inference about accepting or rejecting null hypothesis.

- In general, when formulating hypotheses, the **null hypothesis (H0)** is designed to cover **all possibilities except the claim** you're testing.
- This is a crucial concept in hypothesis testing.
- **General Rule:**
- The **null hypothesis** is a statement of **no effect**, **no difference**, or **status quo**, meaning it **includes everything** except the specific outcome you're trying to prove.
- The **alternative hypothesis** is what you're claiming and provide evidence for the same.

- **Like In case of Sun: If I claim Sun rises in the west than**
  - Ho = SUN RISES IN THE EAST(Default belief)
  - Ha = SUN RISES IN THE WEST (My claim)
  - Than I need to give evidence for this claim else it will be rejected.
- The **null hypothesis** should cover the **broadest range of possibilities**, except for the specific claim you're testing.
- The **alternative hypothesis** is focused only on the specific outcome that you want to test or prove.

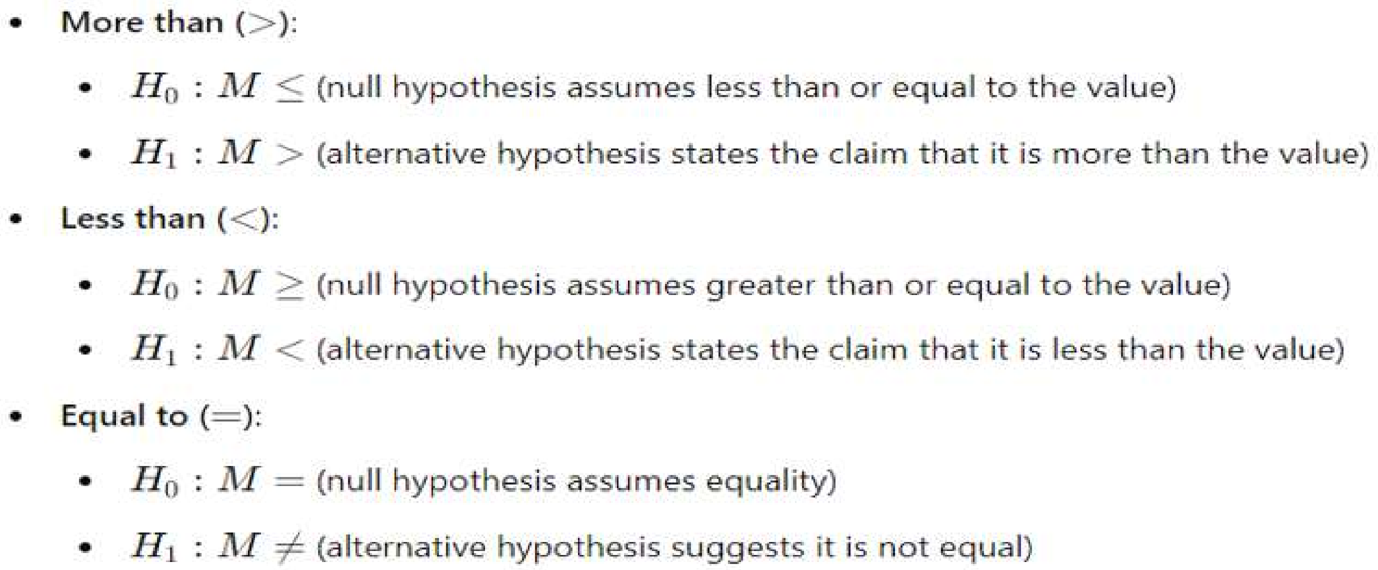


Null Hypothesis (H_0):

- **At least** (≥): The null hypothesis states that the value is greater than or equal to a certain threshold.
- **At most** (≤): The null hypothesis claims that the value is less than or equal to a particular value.
- **Equal to** (=): It assumes that the variable is equal to a specific value.

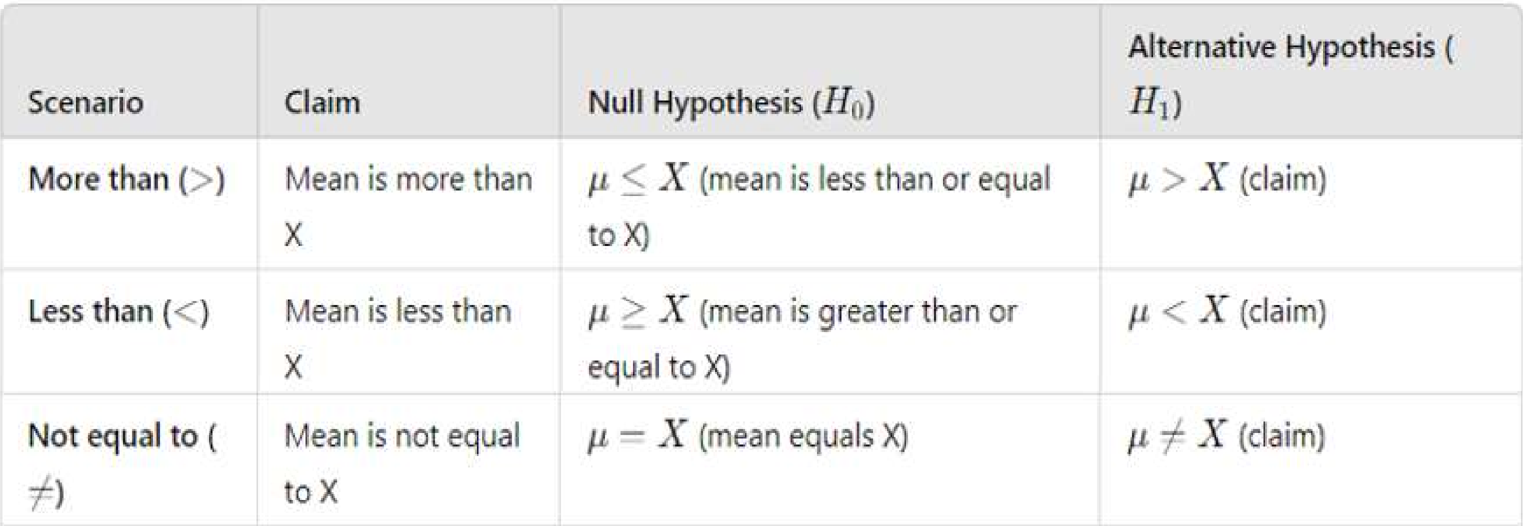
Alternate Hypothesis (H_1):

- **More than** (>): The alternative hypothesis suggests that the value is greater than a certain threshold.
- **Less than** (<): It states that the value is smaller than a particular number.
- **Different** (≠): This suggests that the value is not equal, meaning the alternative hypothesis looks for any difference.



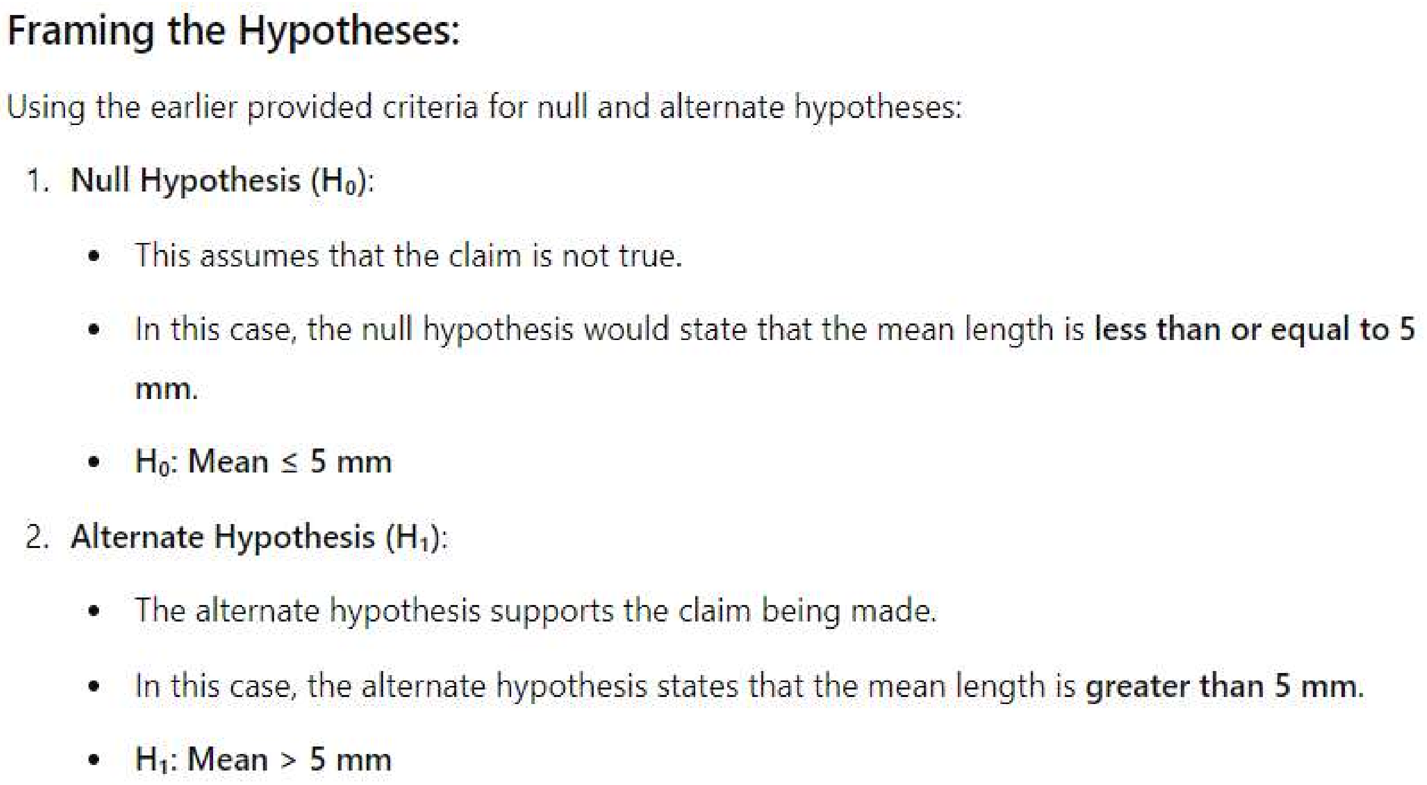
- More than ($>$):
  - $H_0 : M \leq$ (null hypothesis assumes less than or equal to the value)
  - $H_1 : M >$ (alternative hypothesis states the claim that it is more than the value)
- Less than ($<$):
  - $H_0 : M \geq$ (null hypothesis assumes greater than or equal to the value)
  - $H_1 : M <$ (alternative hypothesis states the claim that it is less than the value)
- Equal to ($=$):
  - $H_0 : M =$ (null hypothesis assumes equality)
  - $H_1 : M \neq$ (alternative hypothesis suggests it is not equal)

Guidelines for Framing Hypotheses Based on the Claim:



Scenario	Claim	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
More than ($>$)	Mean is more than X	$\mu \leq X$ (mean is less than or equal to X)	$\mu > X$ (claim)
Less than ($<$)	Mean is less than X	$\mu \geq X$ (mean is greater than or equal to X)	$\mu < X$ (claim)
Not equal to ($\neq$)	Mean is not equal to X	$\mu = X$ (mean equals X)	$\mu \neq X$ (claim)

- **Example: The Claim is a** manufacturer of tubes says that the **mean length is more than 5 mm.**

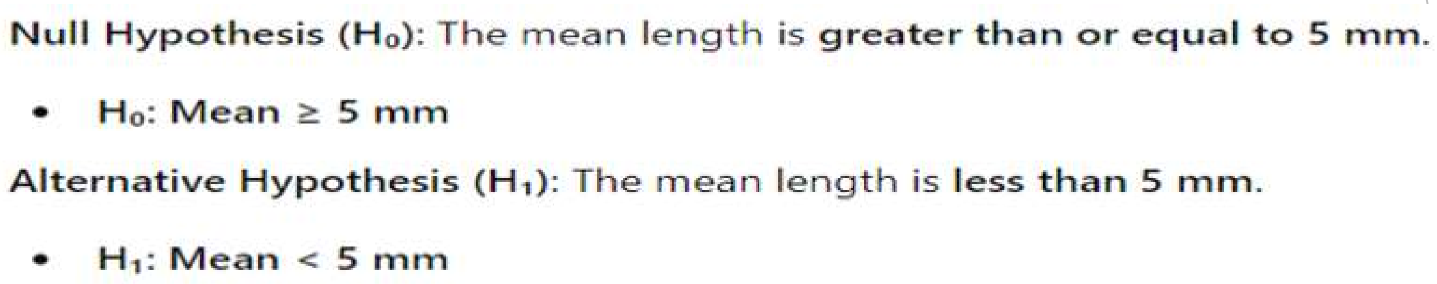


Framing the Hypotheses:

Using the earlier provided criteria for null and alternate hypotheses:

1. **Null Hypothesis (H_0):**
   - This assumes that the claim is not true.
   - In this case, the null hypothesis would state that the mean length is **less than or equal to 5 mm.**
   - **H_0: Mean ≤ 5 mm**
2. **Alternate Hypothesis (H_1):**
   - The alternate hypothesis supports the claim being made.
   - In this case, the alternate hypothesis states that the mean length is **greater than 5 mm.**
   - **H_1: Mean > 5 mm**

- **Example: The Claim is a** manufacturer of tubes says that the **mean length is less than 5 mm.**



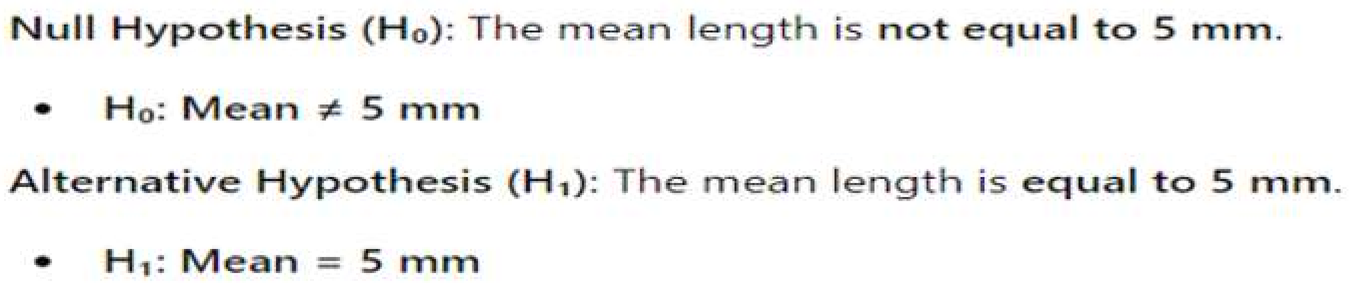
Null Hypothesis (H_0): The mean length is **greater than or equal to 5 mm.**

- H_0: Mean ≥ 5 mm

Alternative Hypothesis (H_1): The mean length is **less than 5 mm.**

- H_1: Mean < 5 mm

- **Example: The Claim is a** manufacturer of tubes says that the **mean length is equal to 5 mm.**



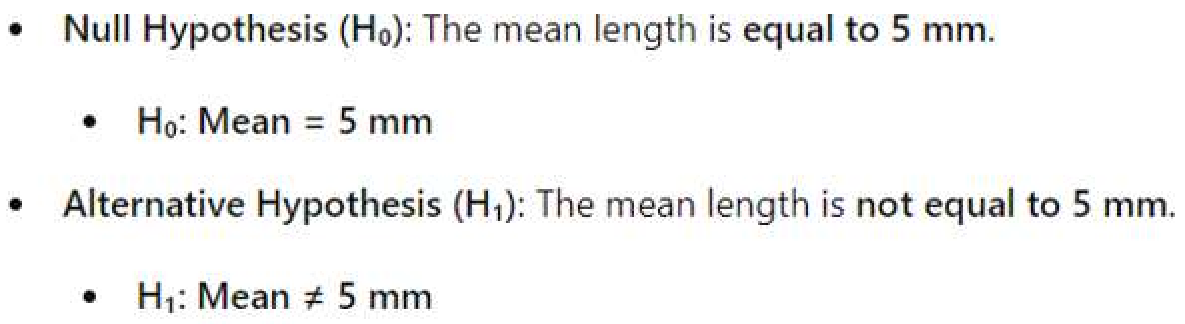
Null Hypothesis (H_0): The mean length is **not equal to 5 mm.**

- H_0: Mean ≠ 5 mm

Alternative Hypothesis (H_1): The mean length is **equal to 5 mm.**

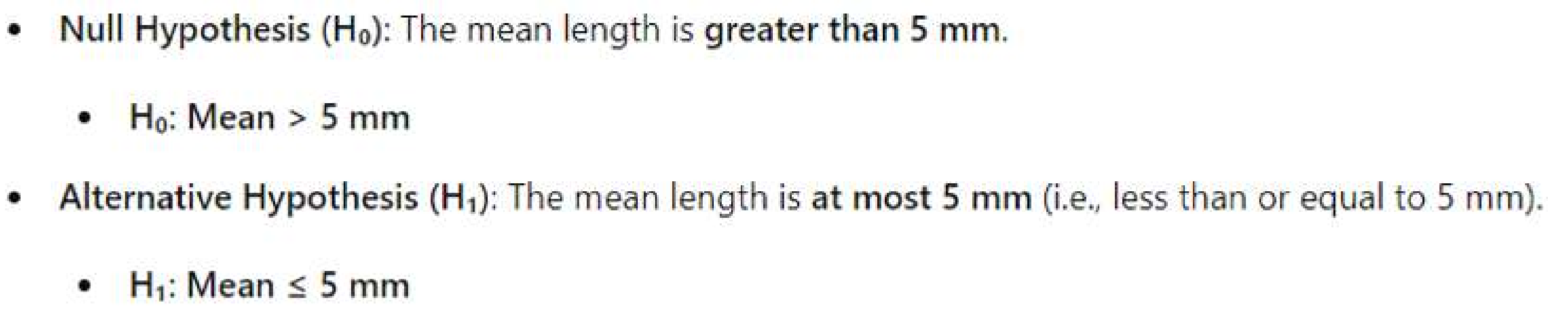
- H_1: Mean = 5 mm

- **Example: The Claim is a** manufacturer of tubes says that the **mean length is not equal to 5 mm.**



- Null Hypothesis (H_0): The mean length is **equal to 5 mm.**
  - H_0: Mean = 5 mm
- Alternative Hypothesis (H_1): The mean length is **not equal to 5 mm.**
  - H_1: Mean ≠ 5 mm

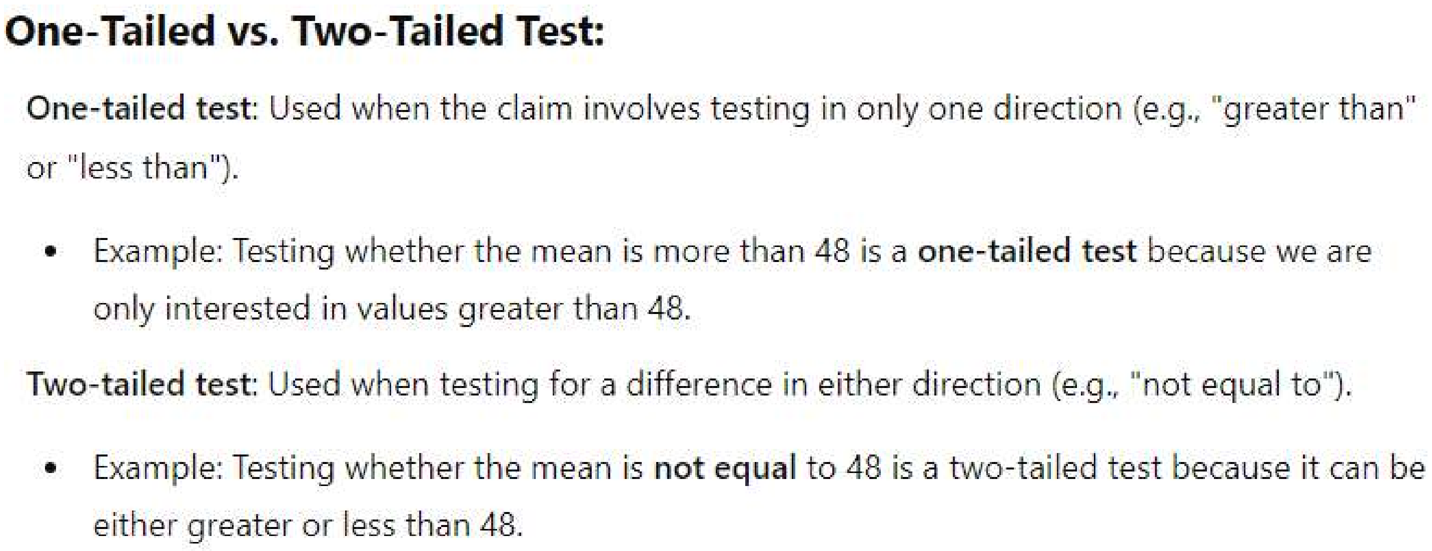
- **Example: The Claim is a** manufacturer of tubes says that the **mean length is at most 5 mm.**



- Null Hypothesis (H_0): The mean length is **greater than 5 mm.**
  - H_0: Mean > 5 mm
- Alternative Hypothesis (H_1): The mean length is **at most 5 mm** (i.e., less than or equal to 5 mm).
  - H_1: Mean ≤ 5 mm

- Thus based on these claim we can decide whether the test is one tailed or two tailed.
- When alternate hypothesis has > or < symbol it is considered to be a one tailed test
- When alternate hypothesis has a “ not equal to “ symbol , it is considered to be two tailed

Set of hypothesis	Tail
Ho: M ≤ Ha: M >	right tailed
Ho: M ≥ Ha: M <	Left tailed
Ho: M = Ha: M ≠	Two tailed



One-Tailed vs. Two-Tailed Test:

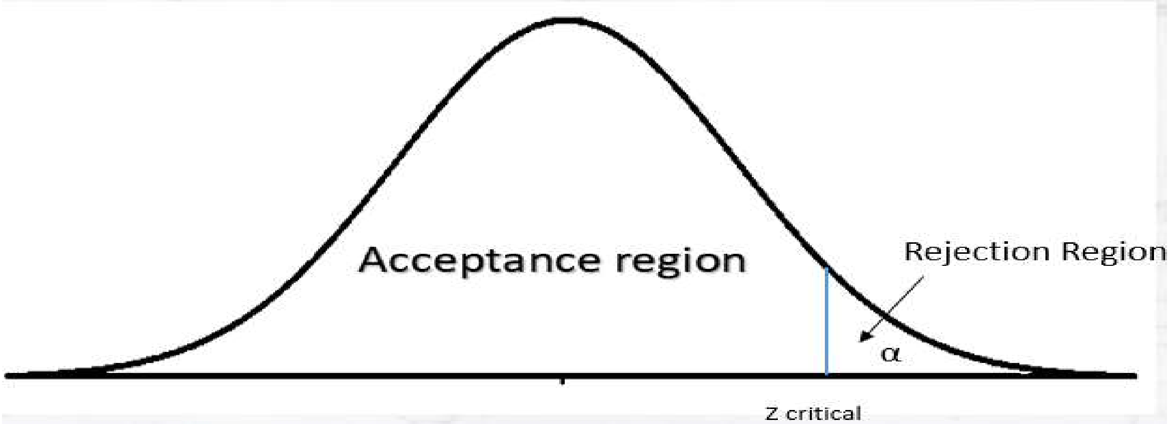
One-tailed test: Used when the claim involves testing in only one direction (e.g., "greater than" or "less than").

- Example: Testing whether the mean is more than 48 is a **one-tailed test** because we are only interested in values greater than 48.

Two-tailed test: Used when testing for a difference in either direction (e.g., "not equal to").

- Example: Testing whether the mean is **not equal** to 48 is a two-tailed test because it can be either greater or less than 48.

Significance level



- **Significance Level (α)** guides our decision-making in hypothesis testing by setting a threshold for accepting or rejecting the null hypothesis based on the observed data.
  - Commonly set at 0.05 or 5% for 95% acceptance region.

Confidence interval

- **Confidence interval** talks about our confidence for rejecting the null hypothesis.(CI = 1oo-SL)
  - Eg. 95%,99%,98% . . . .
  - In simple terms, a confidence interval helps us understand the "uncertainty" or "margin of error" around our estimate and gives us a sense of how confident we can be about our estimate of the rejection of null hypothesis.

Test Statistic:

- **Test statistics** are critical components in hypothesis testing, used to determine whether to reject the null hypothesis.
  - Different test statistics are chosen based on the nature of the data, the sample size, and the specific hypothesis being tested.
  - Here's the logic behind using various test statistics:

- **Chi-Square Statistic (Chi-Square Test)**
  - **When to Use:**
  - For categorical data.
  - To test the independence of two categorical variables (Chi-Square Test of Independence).
  - To test the goodness-of-fit between observed and expected frequencies.
- **Z-Statistic (Z-Test)**
  - **When to Use:**
  - When the sample size is large (n>30).
  - When the population standard deviation is known.
  - For comparing sample means to a population mean or comparing proportions.

- **T-Statistic (T-Test)**
  - **When to Use:**
  - When the sample size is small (n≤30).
  - When the population standard deviation is unknown and must be estimated from the sample.
  - For comparing sample means to a population mean or comparing means from two independent samples.
- **F-Statistic (ANOVA)**
  - **When to Use:**
  - To compare the variances between multiple groups.
  - In Analysis of Variance (ANOVA) when comparing the means of three or more groups.

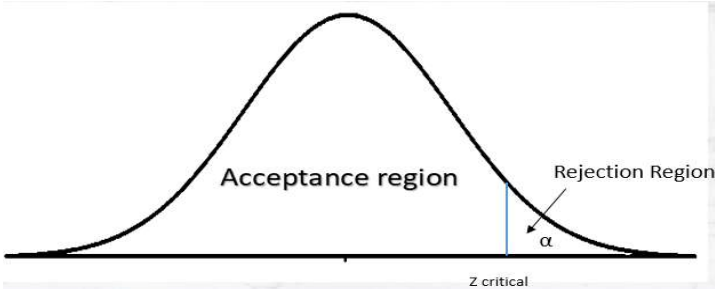
Criterion to make decision

- Acceptance Region
- Critical Value
- P value

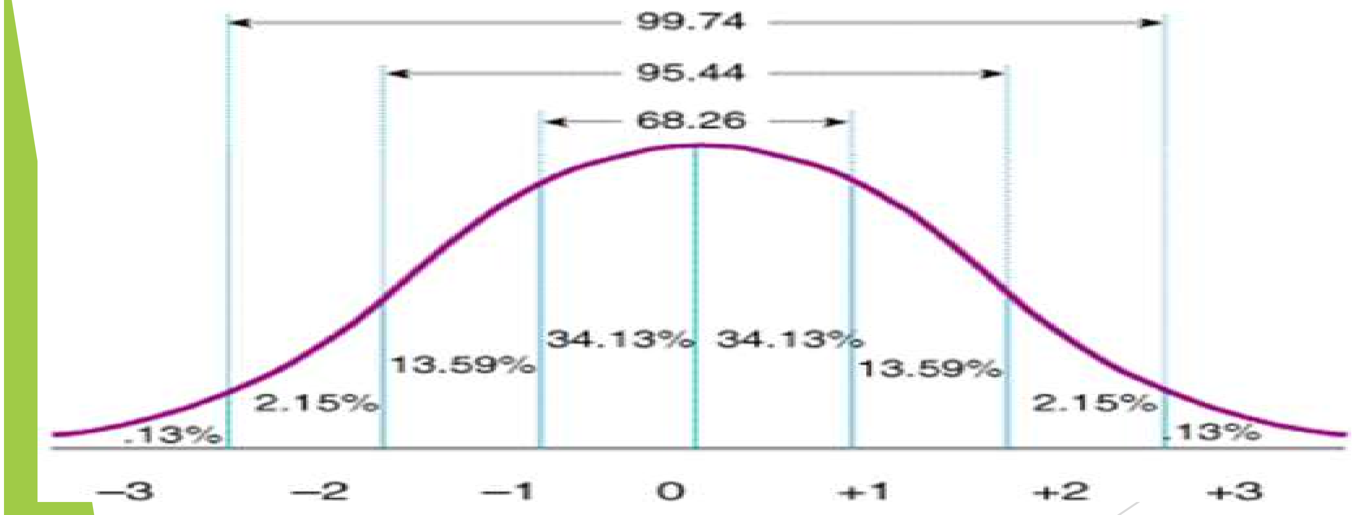
Acceptance Region

- In this approach we defined a range of values as an acceptance region (or critical region) based on the chosen significance level (α).
  - If the test statistic falls within the acceptance region, fail to reject the null hypothesis.
  - If the test statistic falls outside the acceptance region, reject the null hypothesis.

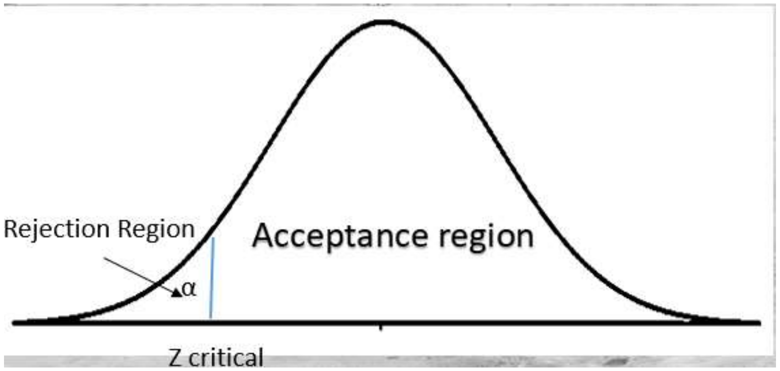
- On standard normal distribution curve where range varies from 3sd on both side of mean value.
- For Z-score at α=0.05 ,Acceptance region is between 1.96sd on both side of mean.(1.64 one tailed)
- If the computed Z-statistic lies within this region, we fail to reject H0.
- If the Z-statistic lies outside this interval, we reject H0.



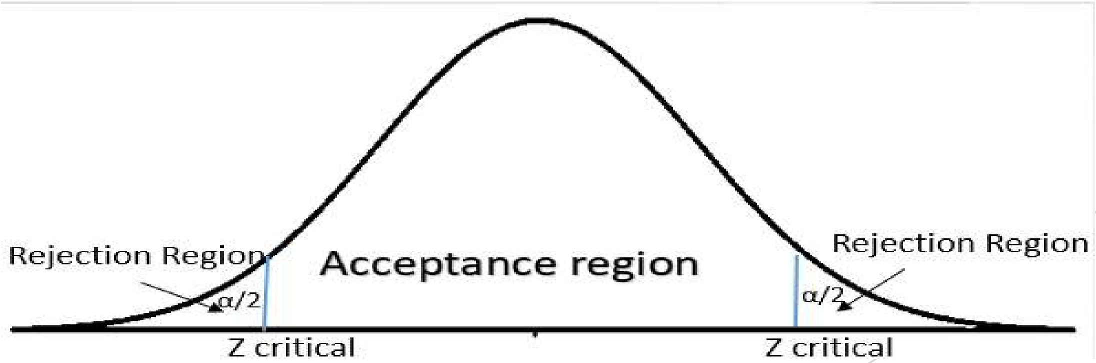
Hypothesis	Tail	Reject
Ho: M ≤		
Ha: M >	right tailed	Z > Z critical



Ho: M ≥		
Ha: M <	Left tailed	Z < Z critical

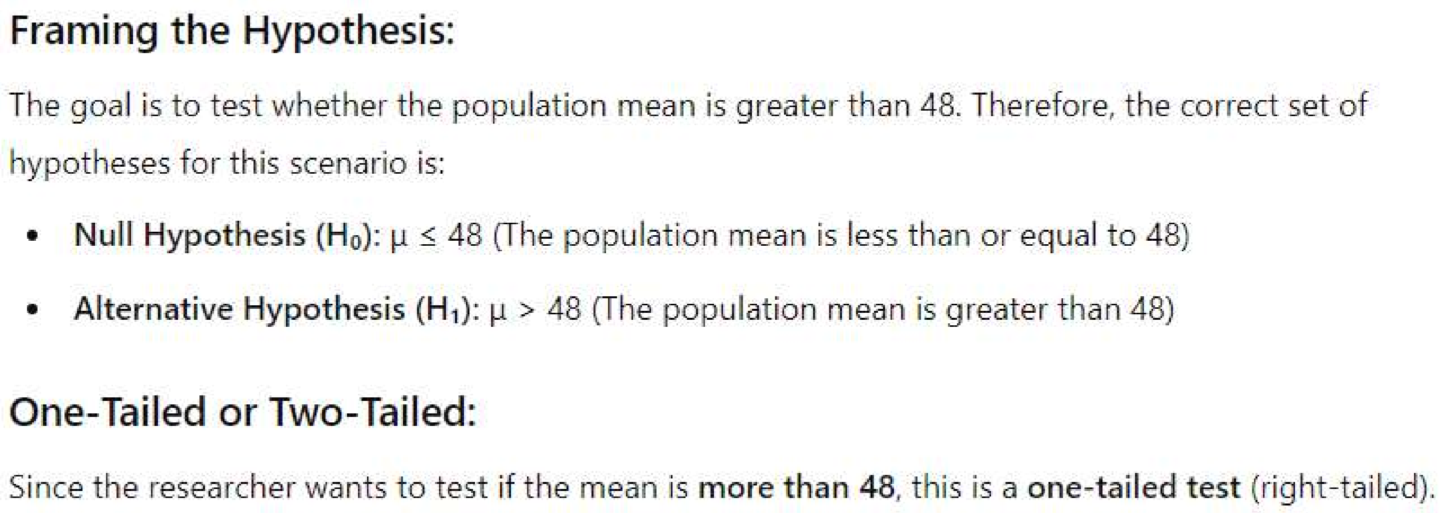


Ho: M =		
Ha: M ≠	Two tailed	|Z| > Z critical



Ques : Mean and standard deviation of the population is 48 and 2. The mean obtained by taking a sample of 12 people is 55. Researcher wishes to test the claim mean is more than 48.

- Frame the correct set of hypothesis Is it one tailed or two tailed? Define the rejection region Draw the graph to show rejection region



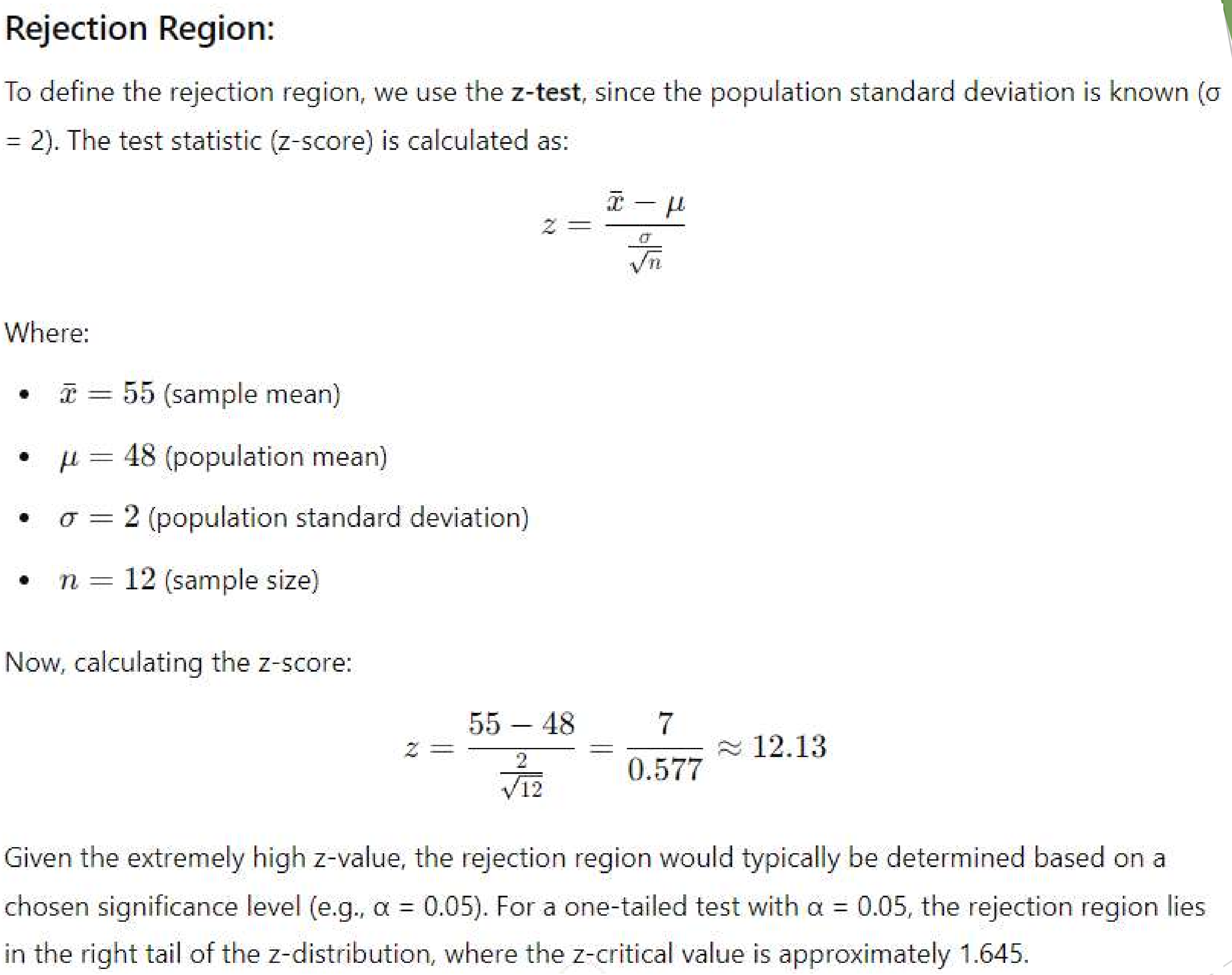
Framing the Hypothesis:

The goal is to test whether the population mean is greater than 48. Therefore, the correct set of hypotheses for this scenario is:

- **Null Hypothesis (H_0)**: $\mu \leq 48$ (The population mean is less than or equal to 48)
- **Alternative Hypothesis (H_1)**: $\mu > 48$ (The population mean is greater than 48)

One-Tailed or Two-Tailed:

Since the researcher wants to test if the mean is **more than 48**, this is a **one-tailed test** (right-tailed).



Rejection Region:

To define the rejection region, we use the **z-test**, since the population standard deviation is known (σ = 2). The test statistic (z-score) is calculated as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

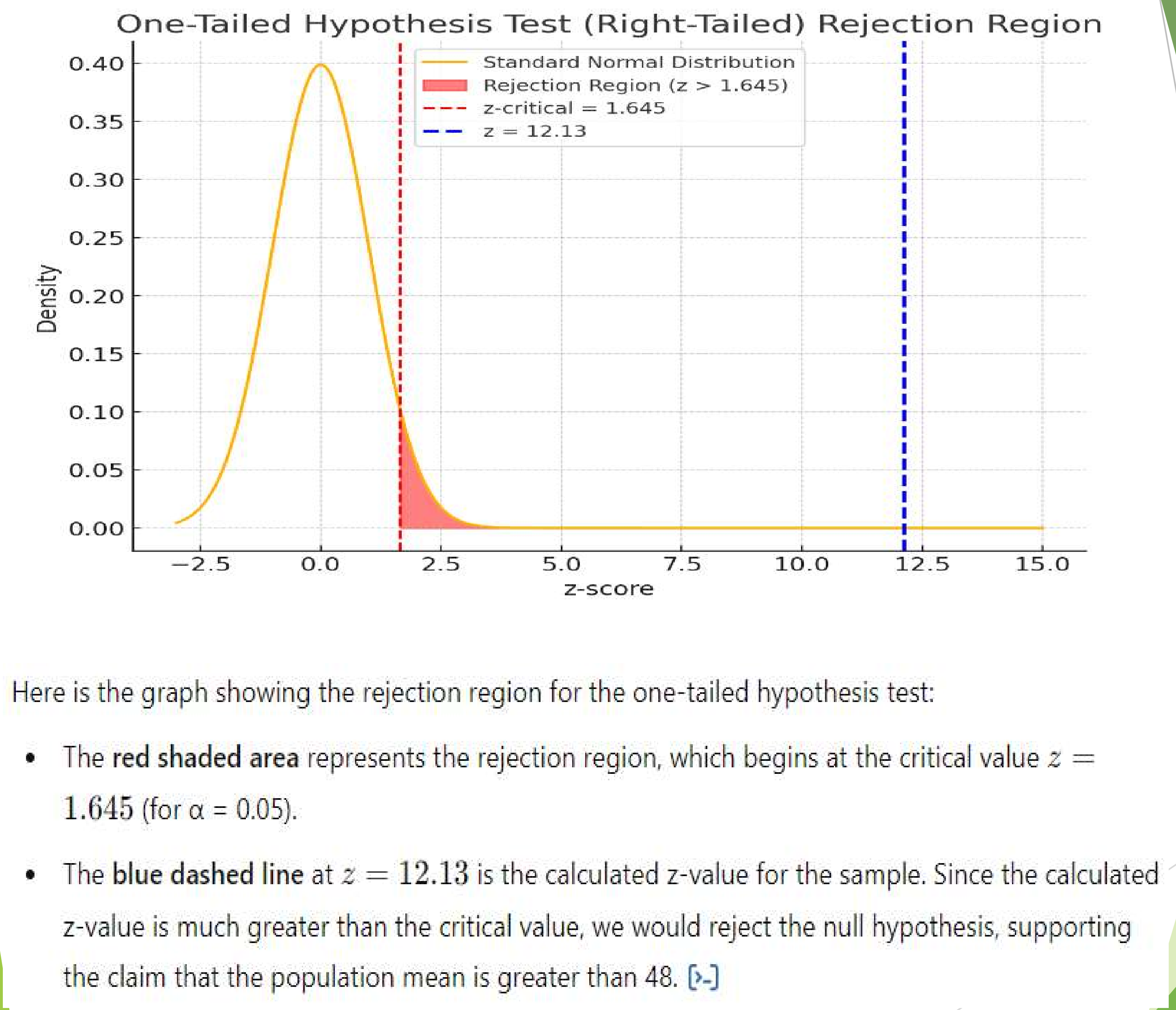
Where:

- $\bar{x} = 55$ (sample mean)
- $\mu = 48$ (population mean)
- $\sigma = 2$ (population standard deviation)
- $n = 12$ (sample size)

Now, calculating the z-score:

$$z = \frac{55 - 48}{\frac{2}{\sqrt{12}}} = \frac{7}{0.577} \approx 12.13$$

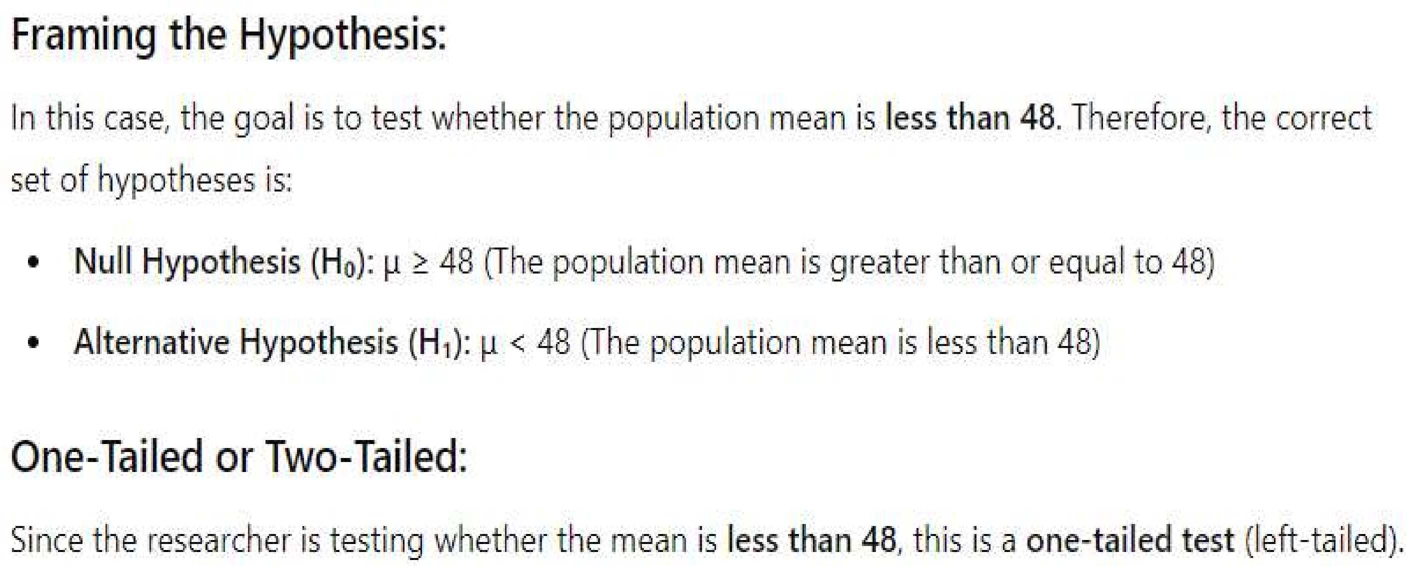
Given the extremely high z-value, the rejection region would typically be determined based on a chosen significance level (e.g., α = 0.05). For a one-tailed test with α = 0.05, the rejection region lies in the right tail of the z-distribution, where the z-critical value is approximately 1.645.


One-Tailed Hypothesis Test (Right-Tailed) Rejection Region
Standard Normal Distribution
Rejection Region (z > 1.645)
z-critical = 1.645
z = 12.13
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
z-score

Here is the graph showing the rejection region for the one-tailed hypothesis test:

- The **red shaded area** represents the rejection region, which begins at the critical value $z = 1.645$ (for α = 0.05).
- The **blue dashed line** at $z = 12.13$ is the calculated z-value for the sample. Since the calculated z-value is much greater than the critical value, we would reject the null hypothesis, supporting the claim that the population mean is greater than 48.

- **Ques:** Mean and standard deviation of the population is 48 and 2. The mean obtained by taking a sample of 12 people is 55. Researcher wishes to test the claim mean is less than 48
- Frame the correct set of hypothesis Is it one tailed or two tailed? Define the rejection region Draw the graph to show rejection region



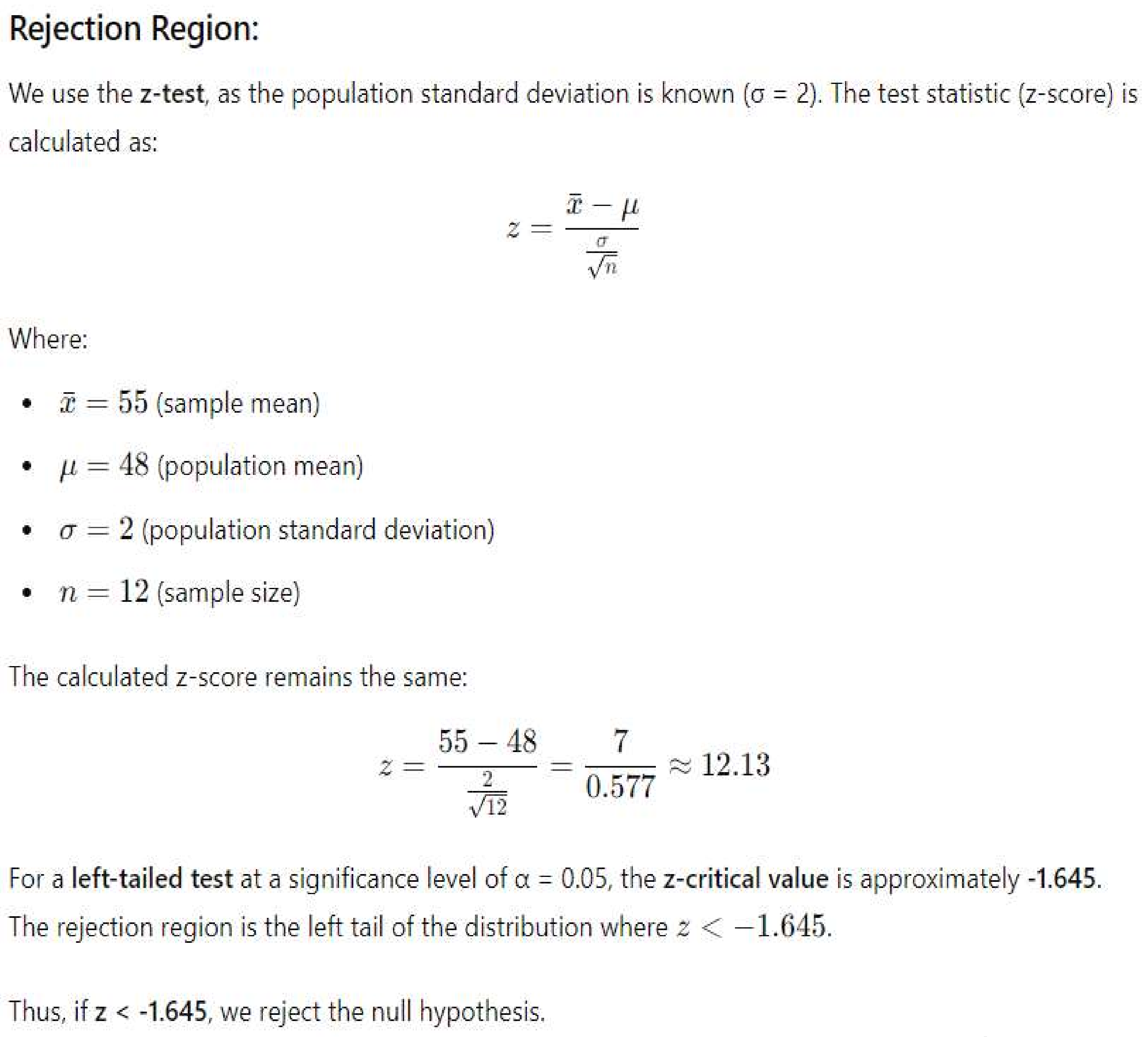
Framing the Hypothesis:

In this case, the goal is to test whether the population mean is **less than 48**. Therefore, the correct set of hypotheses is:

- **Null Hypothesis (H_0):** $\mu \geq 48$ (The population mean is greater than or equal to 48)
- **Alternative Hypothesis (H_1):** $\mu < 48$ (The population mean is less than 48)

One-Tailed or Two-Tailed:

Since the researcher is testing whether the mean is **less than 48**, this is a **one-tailed test** (left-tailed).



Rejection Region:

We use the **z-test**, as the population standard deviation is known (σ = 2). The test statistic (z-score) is calculated as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

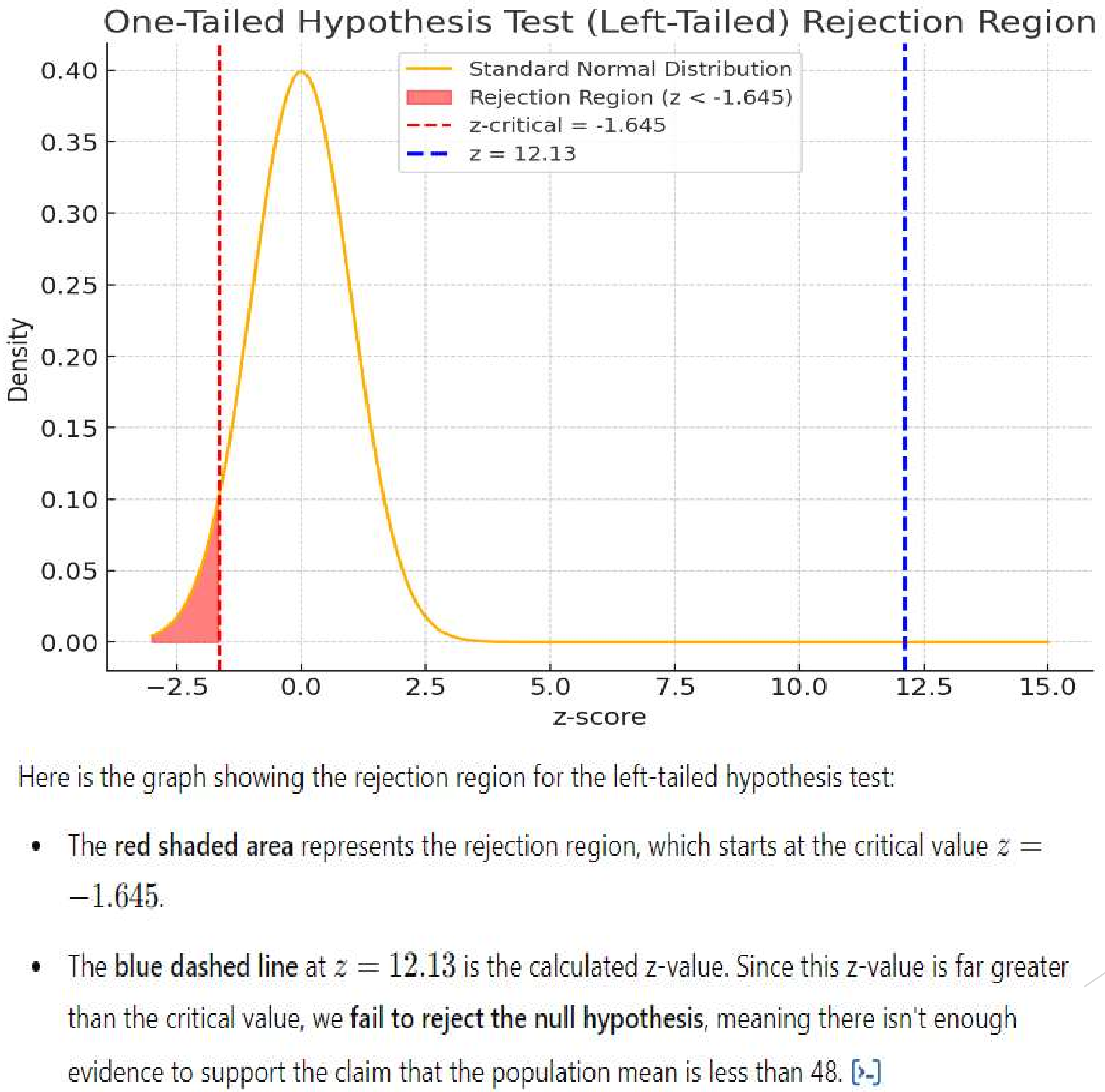
- $\bar{x} = 55$ (sample mean)
- $\mu = 48$ (population mean)
- $\sigma = 2$ (population standard deviation)
- $n = 12$ (sample size)

The calculated z-score remains the same:

$$z = \frac{55 - 48}{\frac{2}{\sqrt{12}}} = \frac{7}{0.577} \approx 12.13$$

For a **left-tailed test** at a significance level of α = 0.05, the **z-critical value** is approximately **-1.645**. The rejection region is the left tail of the distribution where $z < -1.645$.

Thus, if **z < -1.645**, we reject the null hypothesis.

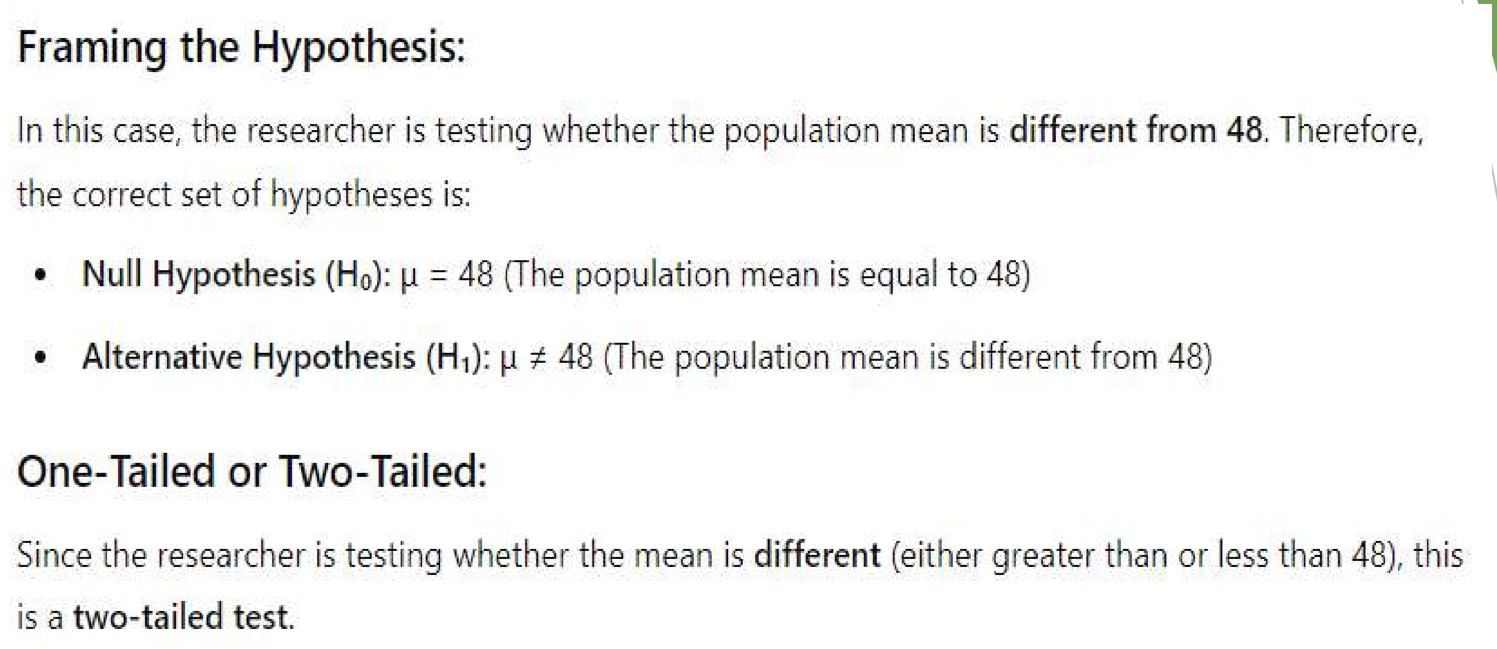

One-Tailed Hypothesis Test (Left-Tailed) Rejection Region
Standard Normal Distribution
Rejection Region (z < -1.645)
z-critical = -1.645
z = 12.13
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
z-score

Here is the graph showing the rejection region for the left-tailed hypothesis test:

- The **red shaded area** represents the rejection region, which starts at the critical value $z = -1.645$.
- The **blue dashed line** at $z = 12.13$ is the calculated z-value. Since this z-value is far greater than the critical value, we **fail to reject the null hypothesis**, meaning there isn't enough evidence to support the claim that the population mean is less than 48. [>_]

Ques: Mean and standard deviation of the population is 48 and 2. The mean obtained by taking a sample of 12 people is 55. Researcher wishes to test the claim mean is different than 48

- Frame the correct set of hypothesis Is it one tailed or two tailed? Define the rejection region Draw the graph to show rejection region



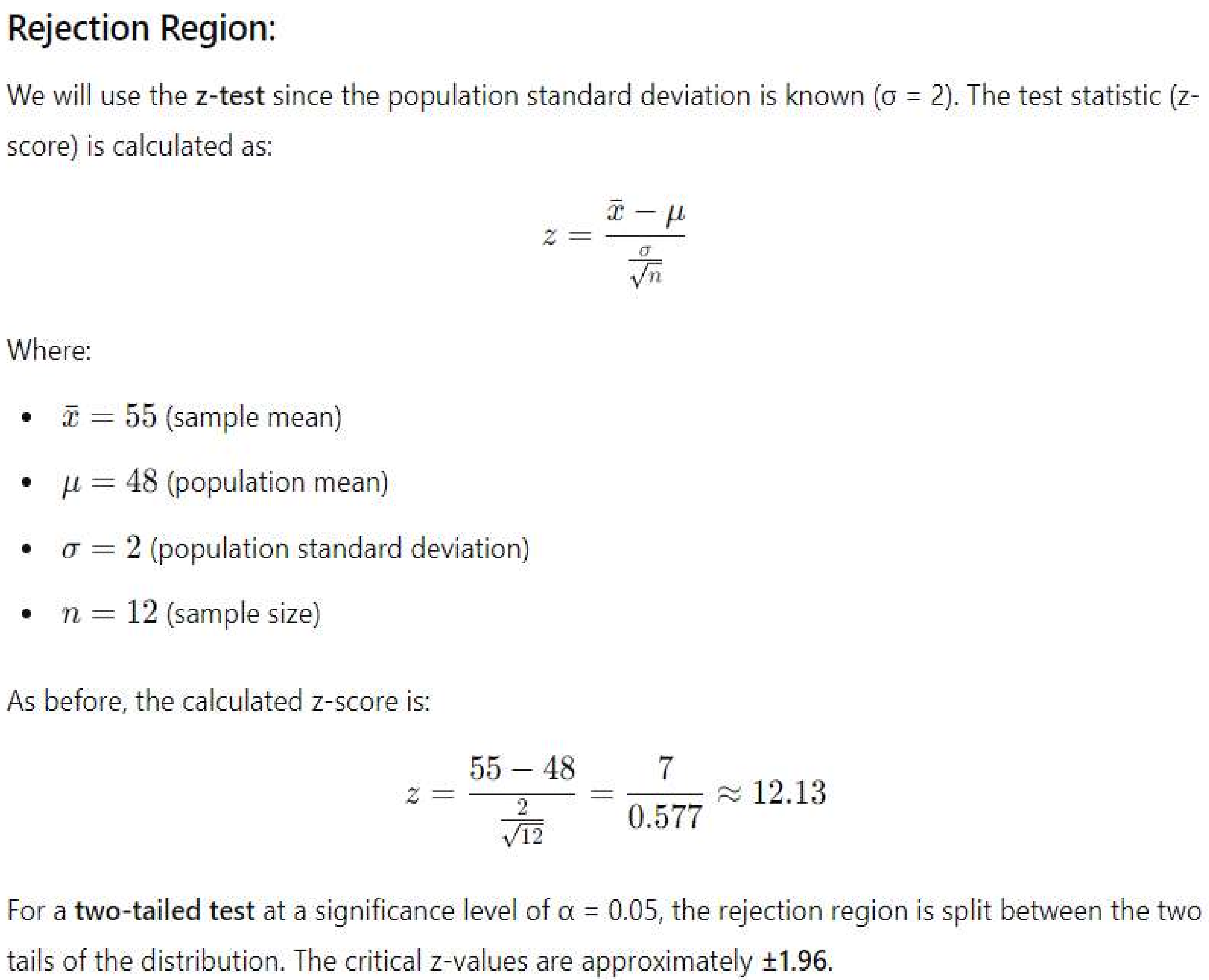
Framing the Hypothesis:

In this case, the researcher is testing whether the population mean is **different from 48**. Therefore, the correct set of hypotheses is:

- **Null Hypothesis (H_0)**: $\mu = 48$ (The population mean is equal to 48)
- **Alternative Hypothesis (H_1)**: $\mu \neq 48$ (The population mean is different from 48)

One-Tailed or Two-Tailed:

Since the researcher is testing whether the mean is **different** (either greater than or less than 48), this is a **two-tailed test**.



Rejection Region:

We will use the **z-test** since the population standard deviation is known (σ = 2). The test statistic (z-score) is calculated as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

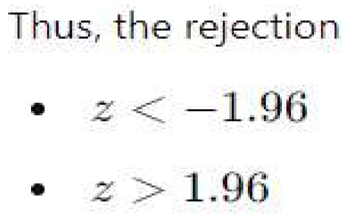
Where:

- $\bar{x} = 55$ (sample mean)
- $\mu = 48$ (population mean)
- $\sigma = 2$ (population standard deviation)
- $n = 12$ (sample size)

As before, the calculated z-score is:

$$z = \frac{55 - 48}{\frac{2}{\sqrt{12}}} = \frac{7}{0.577} \approx 12.13$$

For a **two-tailed test** at a significance level of α = 0.05, the rejection region is split between the two tails of the distribution. The critical z-values are approximately **±1.96.**



Thus, the rejection regions are:

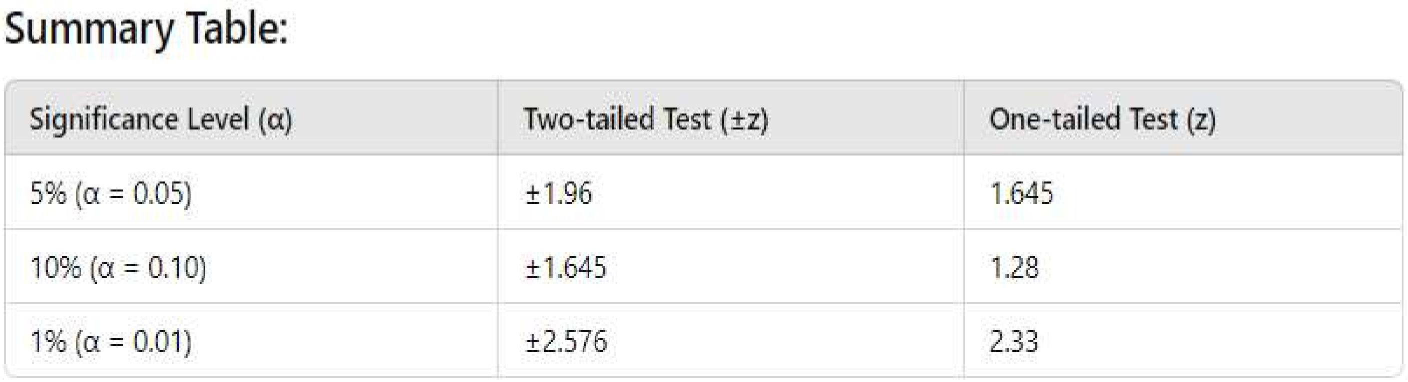
- $z < -1.96$
- $z > 1.96$

Critical Z Value

- Critical Z value is the cutoff value that defines the boundary of rejection region in Z test.
- Compare the Z test statistic(value) to a Z critical value determined by the significance level (α).
- If it is greater than Z critical we reject the null hypothesis.

- **Example:**
  - For a one-tailed Z-test at α=0.05:
    - **Critical value:** 1.645 for a right-tailed test (or −1.645 for a left-tailed test).
    - If the Z-statistic is greater than 1.645(or less than −1.645 for left-tailed), reject H0.
    - If the Z-statistic is less than or equal to 1.645 (or greater than or equal to −1.645 for left-tailed), fail to reject H0.

Z-critical values



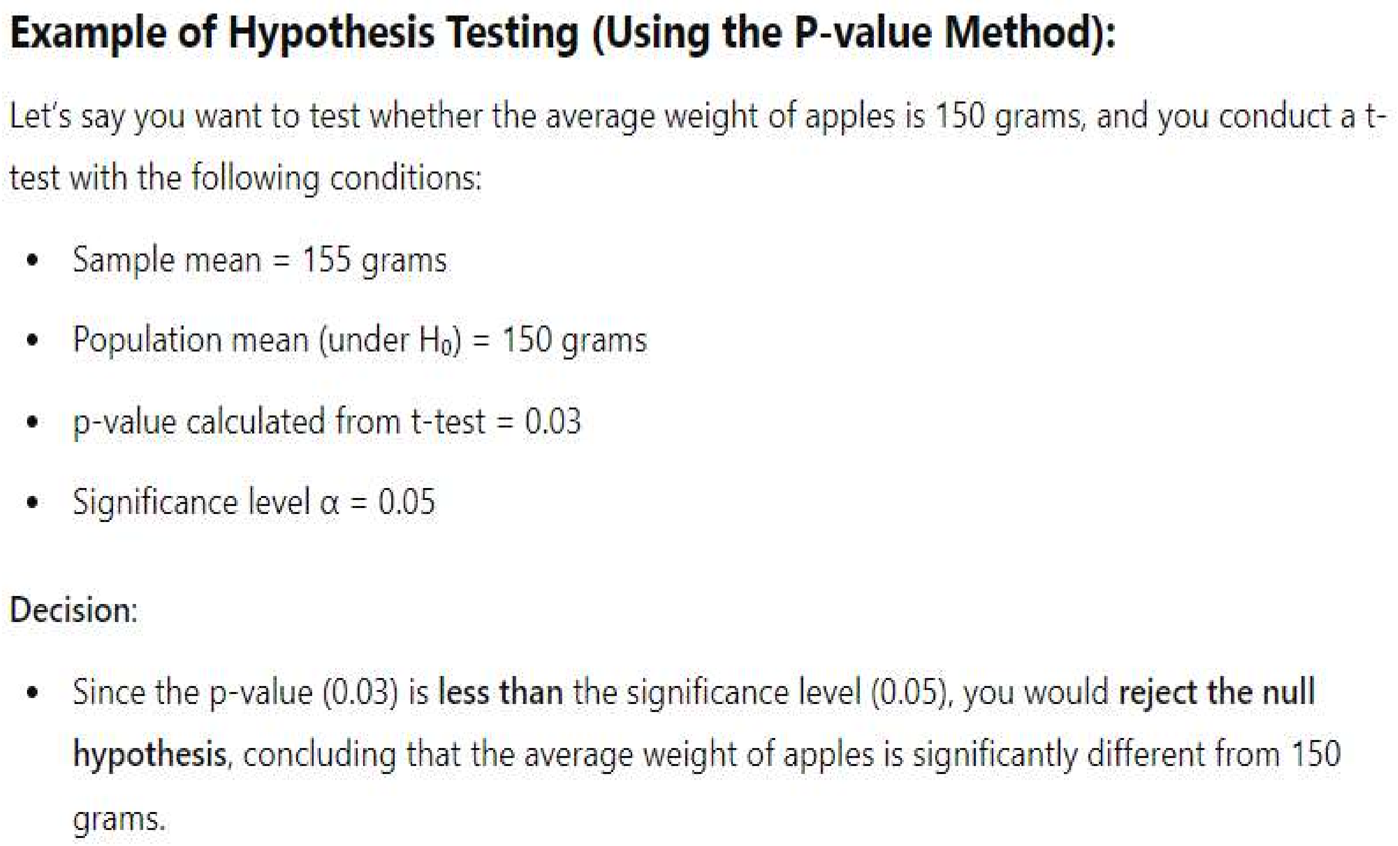
Summary Table:

Significance Level (α)	Two-tailed Test (±z)	One-tailed Test (z)
5% (α = 0.05)	±1.96	1.645
10% (α = 0.10)	±1.645	1.28
1% (α = 0.01)	±2.576	2.33

P-Value:

- Calculate the p-value for Z calculated value using Z .
- Compare this p-value to the significance level (α).
- If p-value ≤ α, reject H0.As we have stronger evidence against Null Hypothesis.
- If p-value > α, fail to reject H0.

- **Example:**
  - For a Z-test with α=0.05:
    - If the p-value is less than or equal to 0.05, reject H0.
    - If the p-value is greater than 0.05, fail to reject H0.



Example of Hypothesis Testing (Using the P-value Method):

Let's say you want to test whether the average weight of apples is 150 grams, and you conduct a t-test with the following conditions:

- Sample mean = 155 grams
- Population mean (under H_0) = 150 grams
- p-value calculated from t-test = 0.03
- Significance level α = 0.05

Decision:

- Since the p-value (0.03) is **less than** the significance level (0.05), you would **reject the null hypothesis**, concluding that the average weight of apples is significantly different from 150 grams.

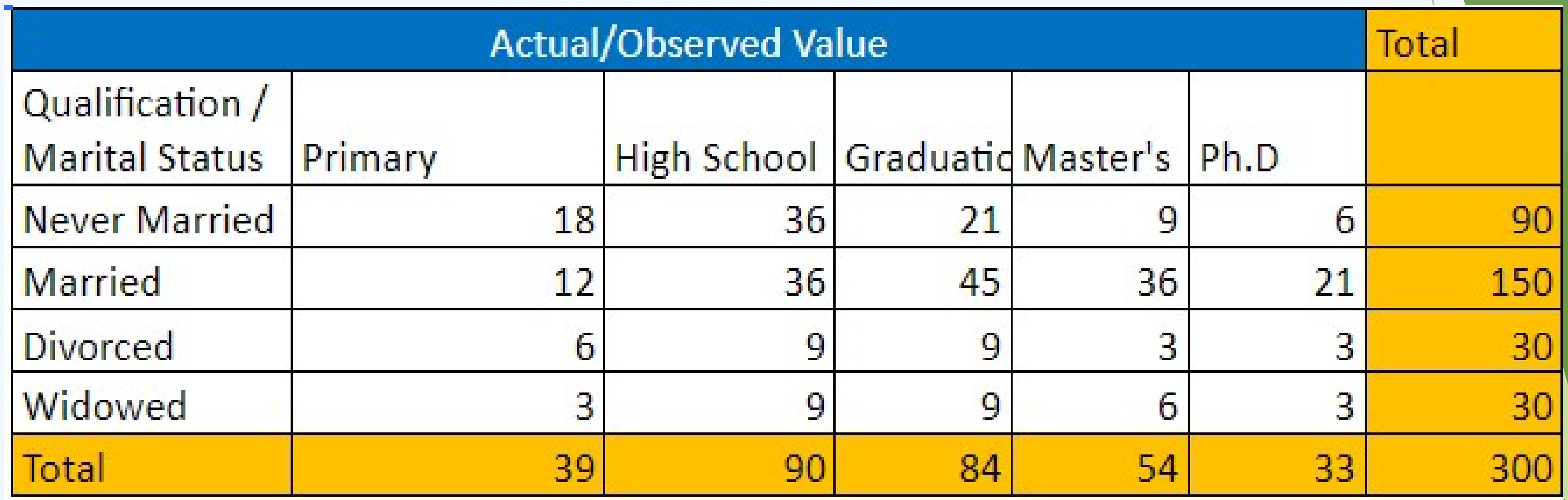
Chi-Square Test

- **Theory:** The Chi-Square test is used to determine whether there is a significant association between two categorical variables.
  - **Example:** Suppose we want to test if there is a significant association between gender (male/female) and smoking status (smoker/non-smoker) in a population.
    - We collect data and create a table showing the counts of males and females who are smokers and non-smokers.
    - We then use the chi-square test to determine if there is a significant relationship between gender and smoking status.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Problem :

Ques: Let us have data about marital status and education, means as "Never married, Married, Divorced, Widowed" and "Primary School, High School, Grad, Master, Ph.D.".



Actual/Observed Value						Total
Qualification / Marital Status	Primary	High School	Graduatic	Master's	Ph.D	
Never Married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

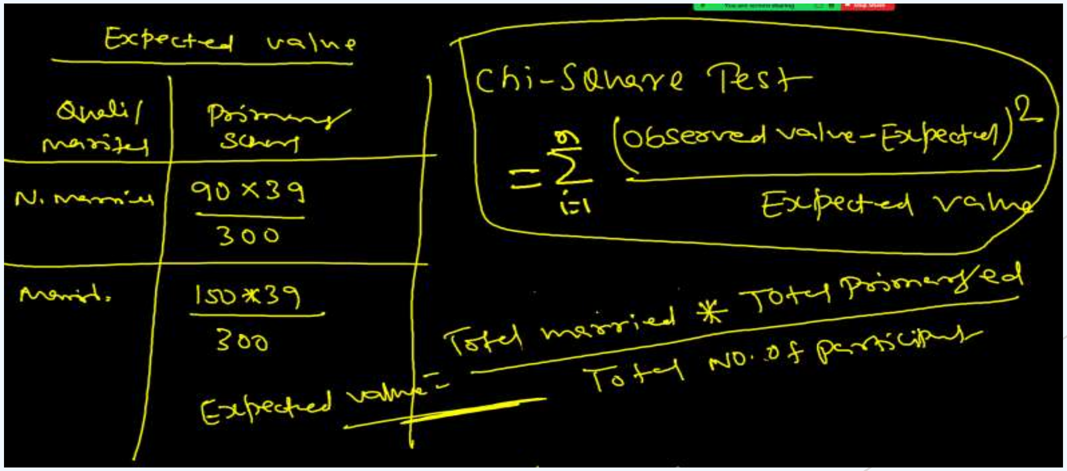
- **Null:** There is no relation between marital status and Education.
- **Alternate:** There is significant relation between marital status and Education.
- Significance Level <= 0.05
- Confidence level: 95%

Table of expected value

Expected Value						Total
Qualification / Marital Status	Primary	High School	Graduatic	Master's	Ph.D	
Never Married	11.7	27	25.2	16.2	9.9	90
Married	19.5	45	42	27	16.5	150
Divorced	3.9	9	8.4	5.4	3.3	30
Widowed	3.9	9	8.4	5.4	3.3	30
Total	39	90	84	54	33	300

Calculate Expected Frequencies:

$$E_{ij} = \frac{(Row\,Total_i) \times (Column\,Total_j)}{Grand\,Total}$$



Expected value

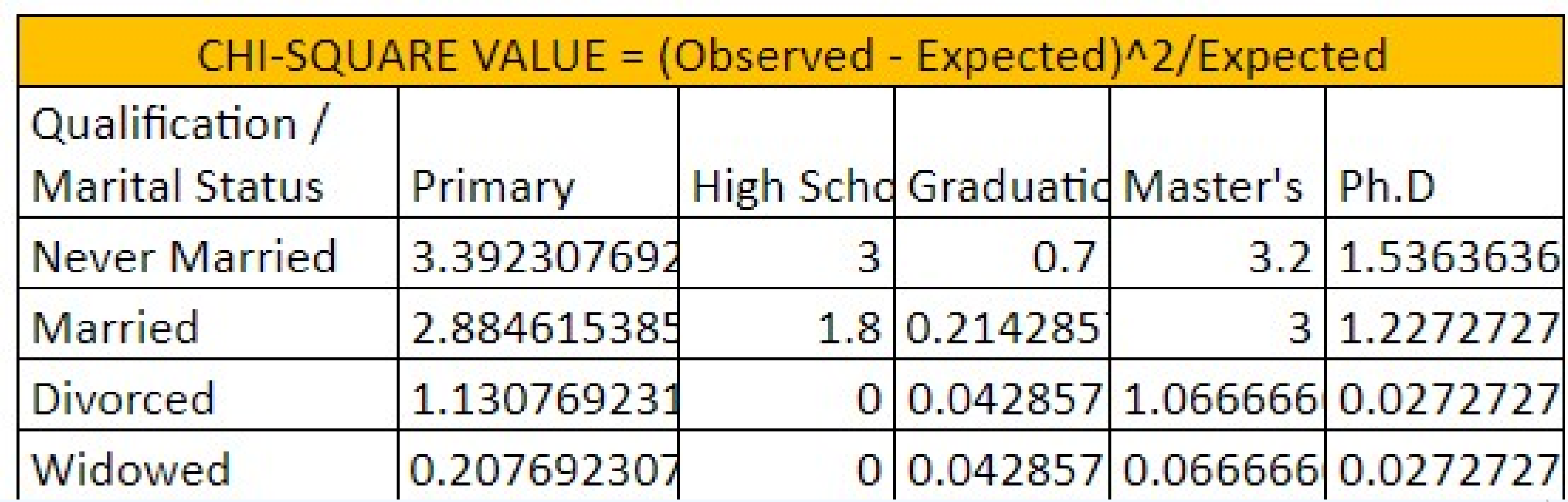
Quali/ married	Primary school
N. married	$\frac{90 \times 39}{300}$
Married.	$\frac{150 * 39}{300}$

Chi-Square Test

$$= \sum_{i=1}^{n} \frac{(\text{Observed value} - \text{Expected})^2}{\text{Expected value}}$$

$$\text{Expected value} = \frac{\text{Total married} * \text{Total Primary ed}}{\text{Total No. of participant}}$$

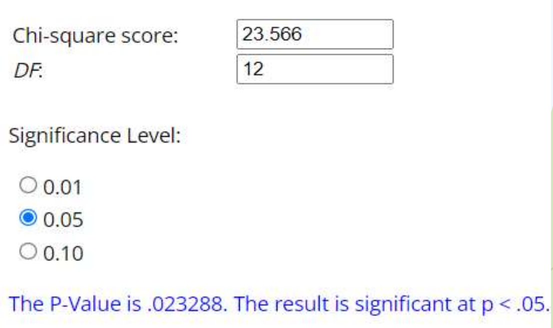
Chi Square values:



CHI-SQUARE VALUE = (Observed - Expected)^2/Expected					
Qualification / Marital Status	Primary	High Scho	Graduatic	Master's	Ph.D
Never Married	3.392307692	3	0.7	3.2	1.5363636
Married	2.884615385	1.8	0.2142857	3	1.2272727
Divorced	1.130769231	0	0.042857	1.066666	0.0272727
Widowed	0.207692307	0	0.042857	0.066666	0.0272727

$x2=\Sigma[(Oij-Eij)**2/Eij]$ = 23.5668
Tabulated Value = 21.06
DOF = (c-1)*(r-1) = (5-1)*(4-1) = 12

Since calculated value >= Tabulated value.
: Reject the null hypothesis.
: Accept the Alternate hypothesis
: Also P<5% so significant level satisfied.

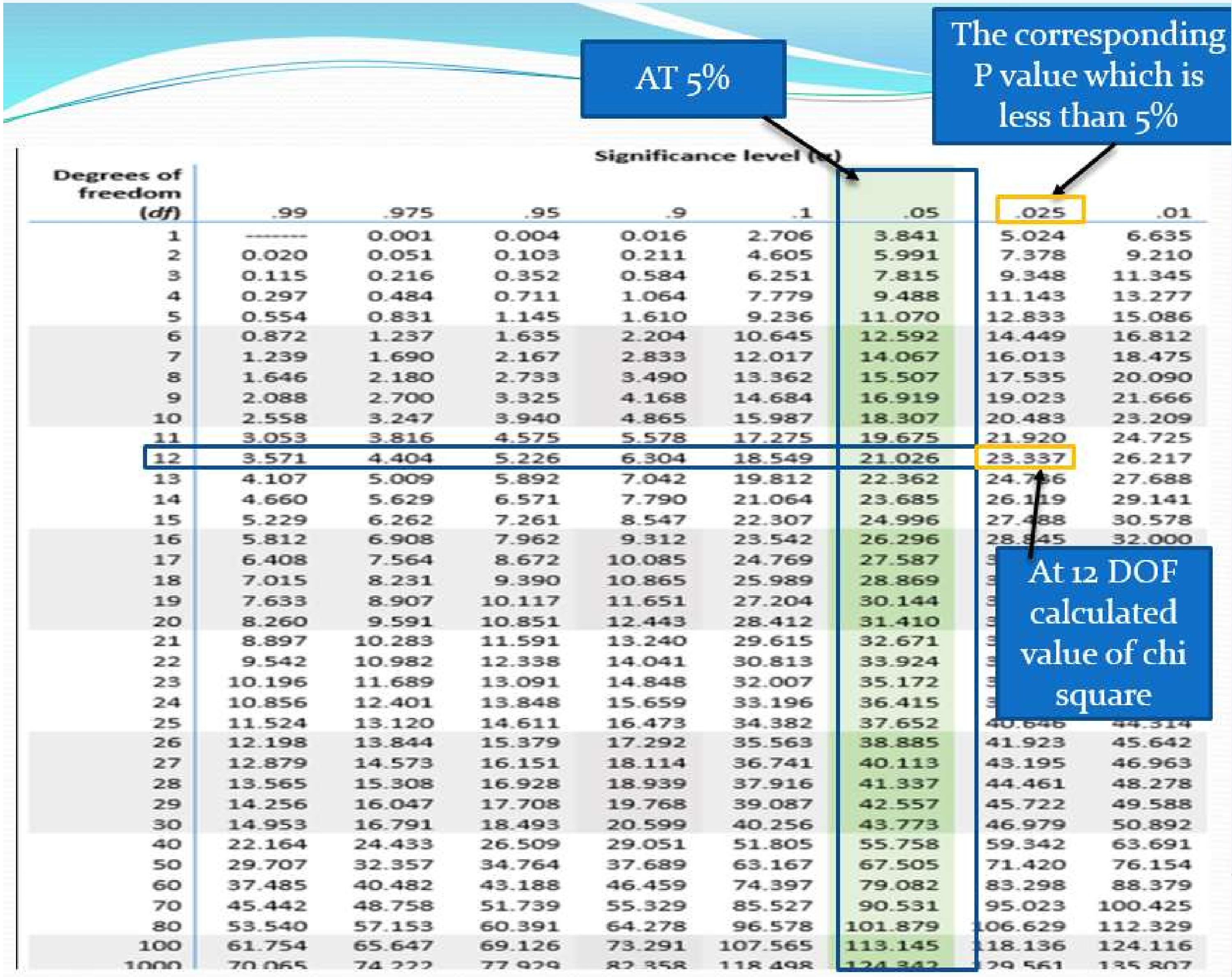


Chi-square score: 23.566
DF: 12

Significance Level:

○ 0.01
● 0.05
○ 0.10

The P-Value is .023288. The result is significant at p < .05.



AT 5%

The corresponding P value which is less than 5%

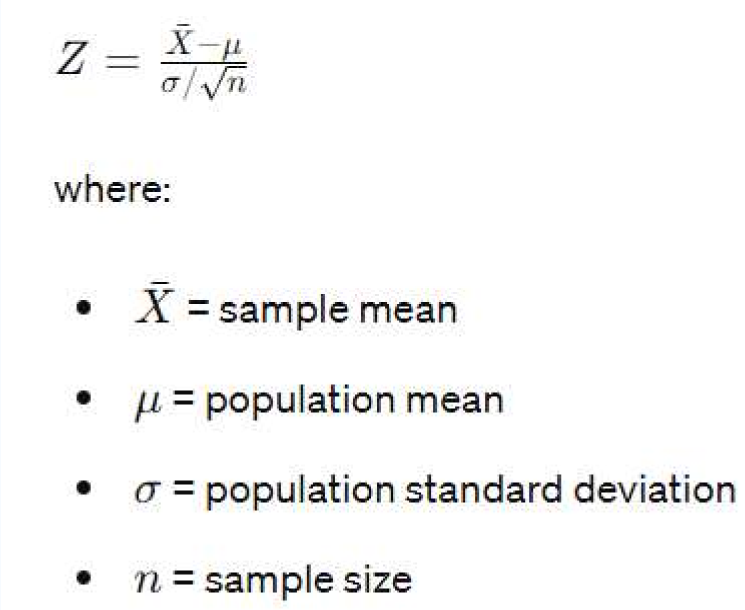
At 12 DOF calculated value of chi square

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	--------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	[illegible]	[illegible]
18	7.015	8.231	9.390	10.865	25.989	28.869	[illegible]	[illegible]
19	7.633	8.907	10.117	11.651	27.204	30.144	[illegible]	[illegible]
20	8.260	9.591	10.851	12.443	28.412	31.410	[illegible]	[illegible]
21	8.897	10.283	11.591	13.240	29.615	32.671	[illegible]	[illegible]
22	9.542	10.982	12.338	14.041	30.813	33.924	[illegible]	[illegible]
23	10.196	11.689	13.091	14.848	32.007	35.172	[illegible]	[illegible]
24	10.856	12.401	13.848	15.659	33.196	36.415	[illegible]	[illegible]
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Z-test

- A Z-test is a statistical test used to determine whether there is a significant difference between sample and population means, or between the means of two samples.
- Z-tests are used to test hypotheses about population means or proportions.
- They help infer whether observed differences are statistically significant or due to random chance.

- Use a Z-test when the population standard deviation (σ) is known.
- The Z-test is more appropriate for large sample sizes (typically n > 30), where the Central Limit Theorem ensures that sample means are normally distributed.
- The Z-test is suitable for analyzing continuous numerical data, such as heights, weights, test scores, etc.



$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

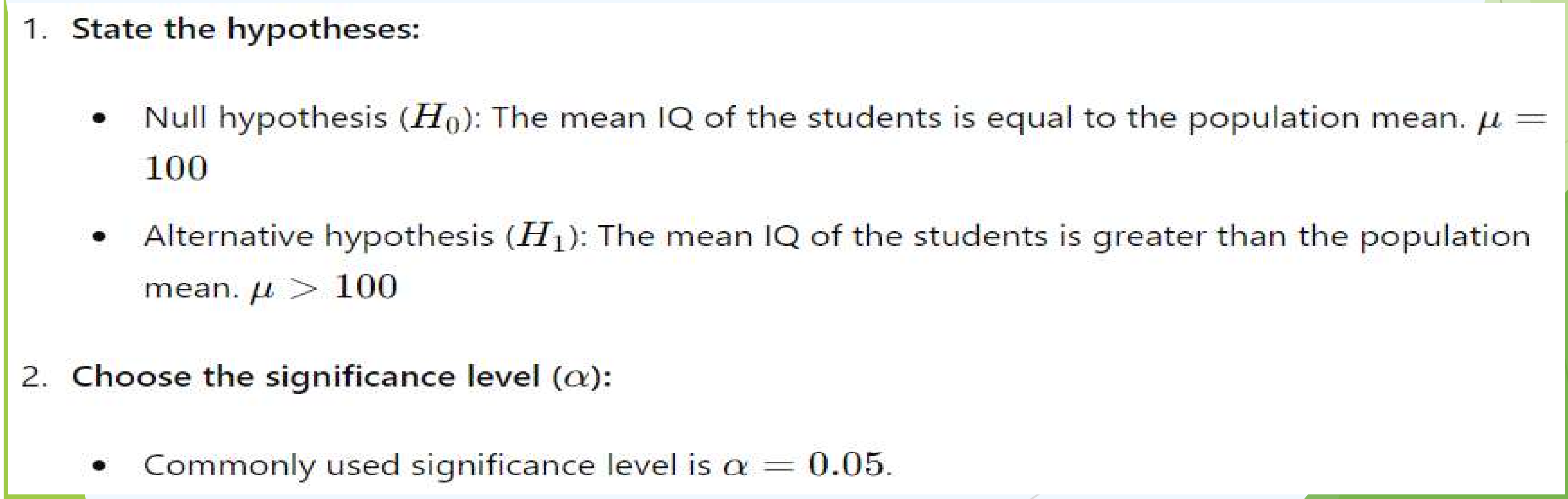
where:

- $\bar{X}$ = sample mean
- μ = population mean
- σ = population standard deviation
- n = sample size

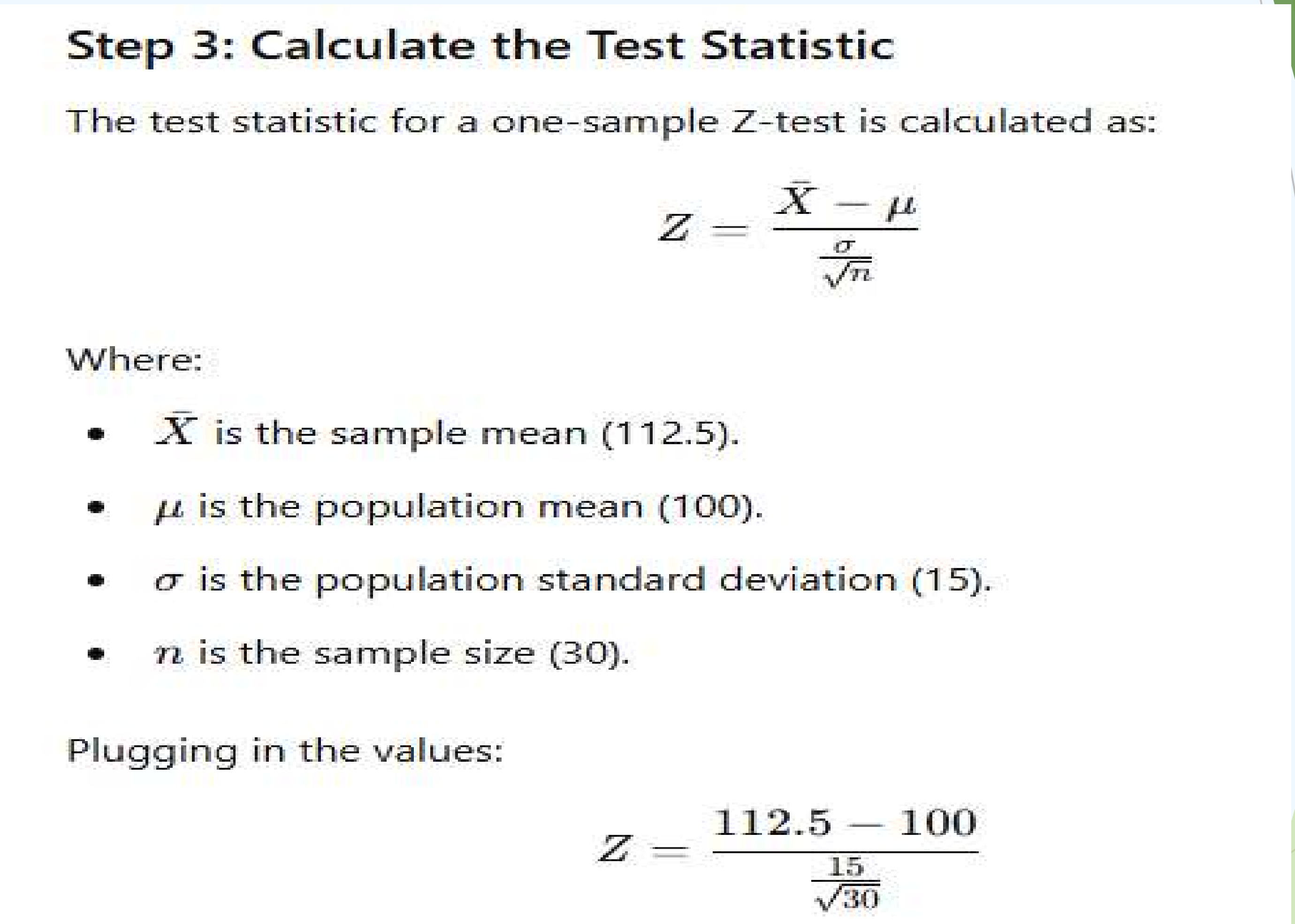
Ques: A principal claims his students have above average IQ.A sample of 30 students and mean112.5. The mean population IQ is 100 with standard deviations of 15. Do hypothesis testing for the claim.

Solution: To test the principal's claim that his students have above-average IQs, we can perform a one-sample z-test because we know the population standard deviation. Here's the step-by-step process:

Given Data: Sample size (n) = 30, Sample mean (x^{-}) = 112.5,Population mean (μ) = 100, Population standard deviation (σ) = 15



1. **State the hypotheses:**
   - Null hypothesis (H_0): The mean IQ of the students is equal to the population mean. $\mu = 100$
   - Alternative hypothesis (H_1): The mean IQ of the students is greater than the population mean. $\mu > 100$
2. **Choose the significance level (α):**
   - Commonly used significance level is $\alpha = 0.05$.



Step 3: Calculate the Test Statistic

The test statistic for a one-sample Z-test is calculated as:

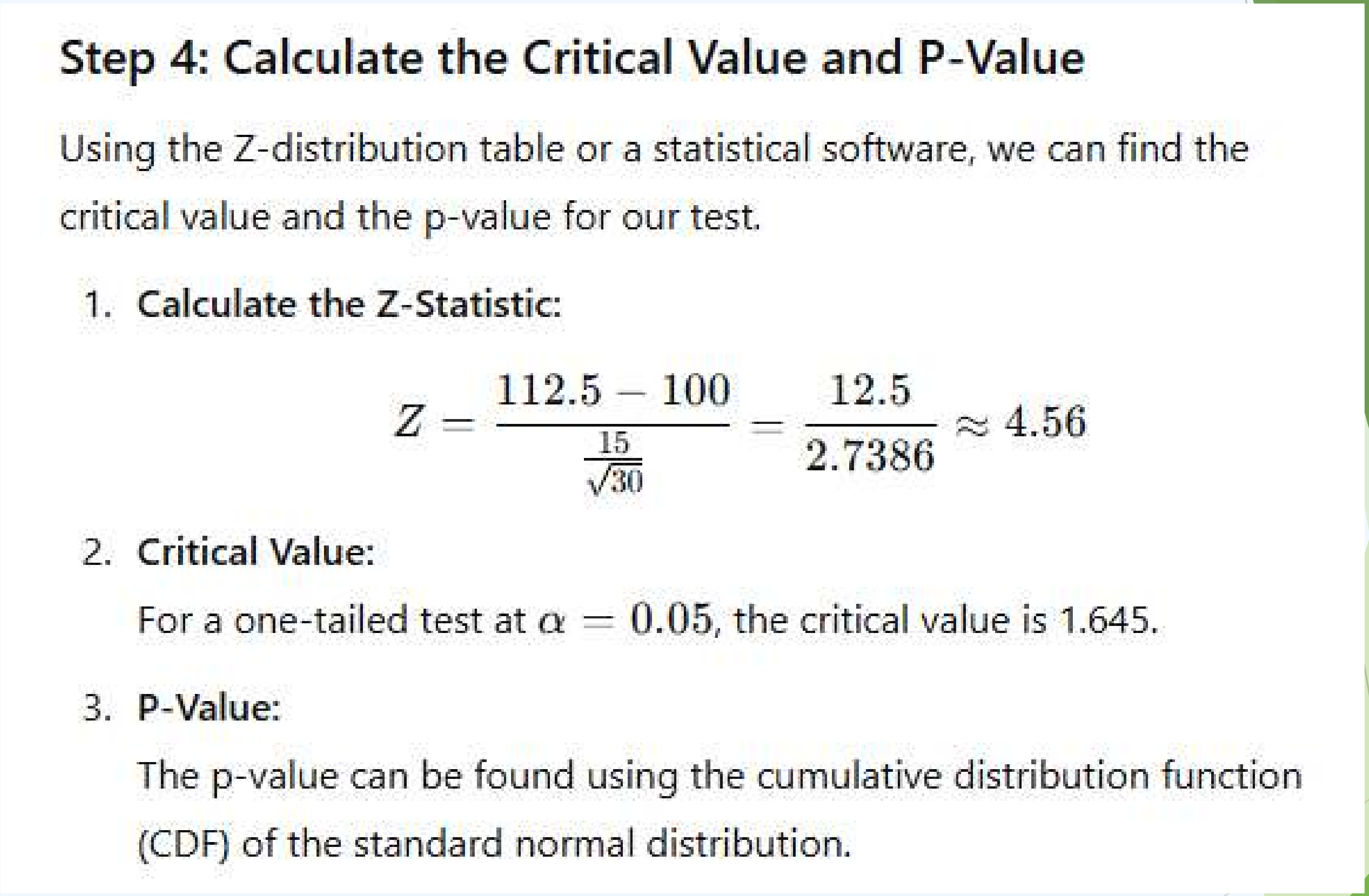
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{X}$ is the sample mean (112.5).
- μ is the population mean (100).
- σ is the population standard deviation (15).
- n is the sample size (30).

Plugging in the values:

$$Z = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}}$$



Step 4: Calculate the Critical Value and P-Value

Using the Z-distribution table or a statistical software, we can find the critical value and the p-value for our test.

1. **Calculate the Z-Statistic:**

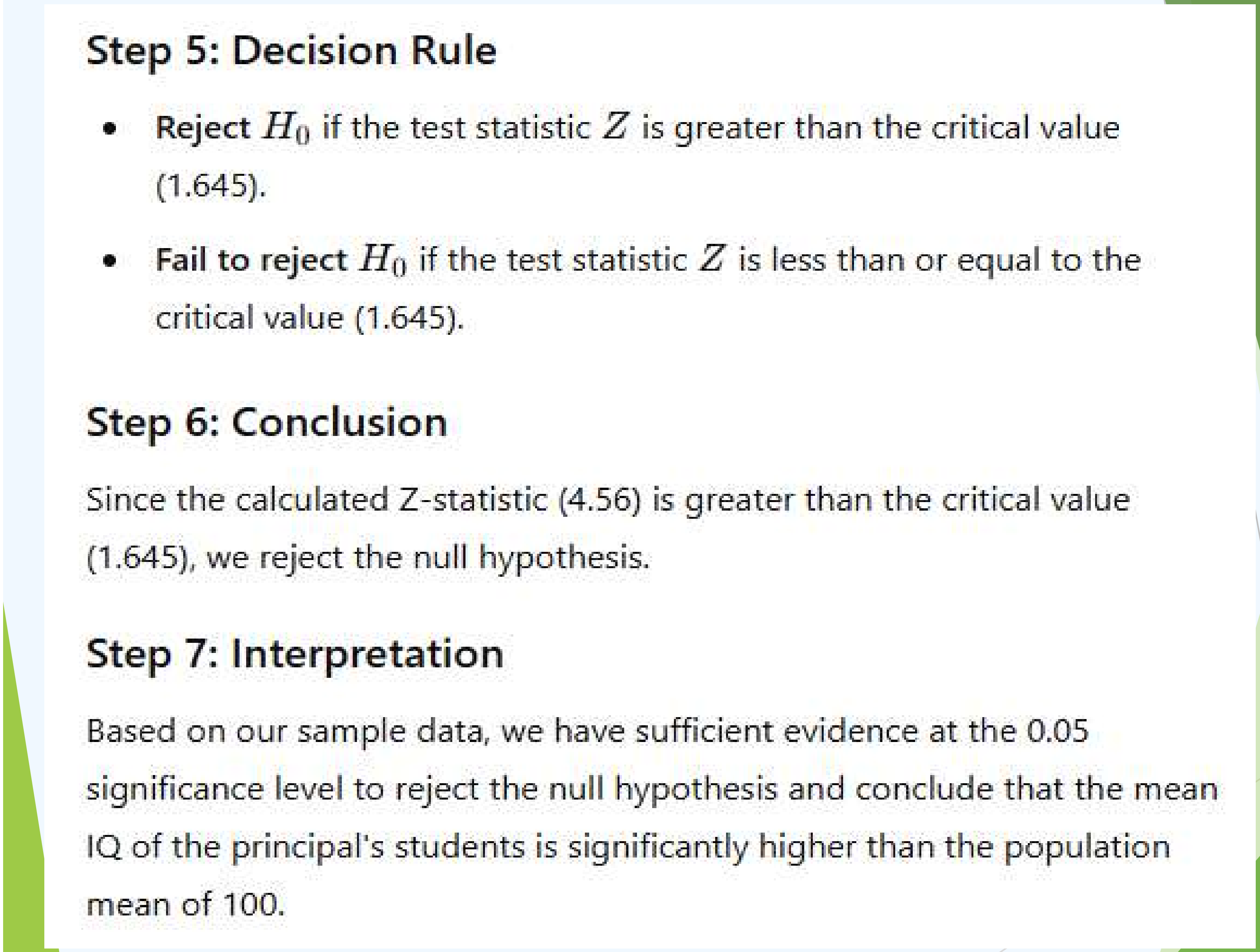
$$Z = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}} = \frac{12.5}{2.7386} \approx 4.56$$

2. **Critical Value:**

   For a one-tailed test at $\alpha = 0.05$, the critical value is 1.645.

3. **P-Value:**

   The p-value can be found using the cumulative distribution function (CDF) of the standard normal distribution.



Step 5: Decision Rule

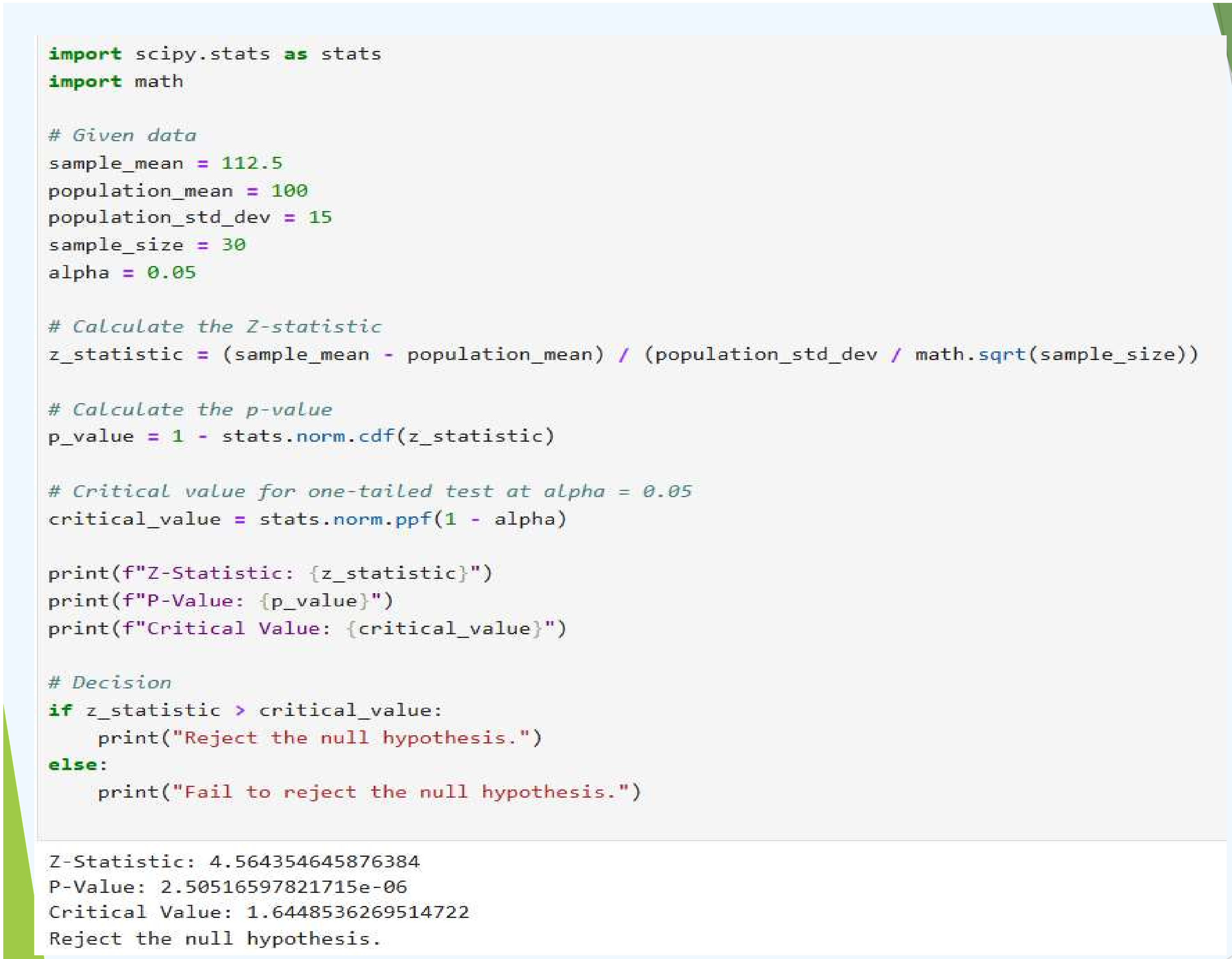
- **Reject** H_0 if the test statistic Z is greater than the critical value (1.645).
- **Fail to reject** H_0 if the test statistic Z is less than or equal to the critical value (1.645).

Step 6: Conclusion

Since the calculated Z-statistic (4.56) is greater than the critical value (1.645), we reject the null hypothesis.

Step 7: Interpretation

Based on our sample data, we have sufficient evidence at the 0.05 significance level to reject the null hypothesis and conclude that the mean IQ of the principal's students is significantly higher than the population mean of 100.



```python
import scipy.stats as stats
import math

# Given data
sample_mean = 112.5
population_mean = 100
population_std_dev = 15
sample_size = 30
alpha = 0.05

# Calculate the Z-statistic
z_statistic = (sample_mean - population_mean) / (population_std_dev / math.sqrt(sample_size))

# Calculate the p-value
p_value = 1 - stats.norm.cdf(z_statistic)

# Critical value for one-tailed test at alpha = 0.05
critical_value = stats.norm.ppf(1 - alpha)

print(f"Z-Statistic: {z_statistic}")
print(f"P-Value: {p_value}")
print(f"Critical Value: {critical_value}")

# Decision
if z_statistic > critical_value:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
```

```
Z-Statistic: 4.564354645876384
P-Value: 2.50516597821715e-06
Critical Value: 1.6448536269514722
Reject the null hypothesis.
```

T-test

- Use a t-test when the population standard deviation (σ) is unknown and must be estimated from the sample.
- The t-test is robust for small sample sizes and is appropriate when dealing with less than 30 samples or when the population standard deviation is unknown.

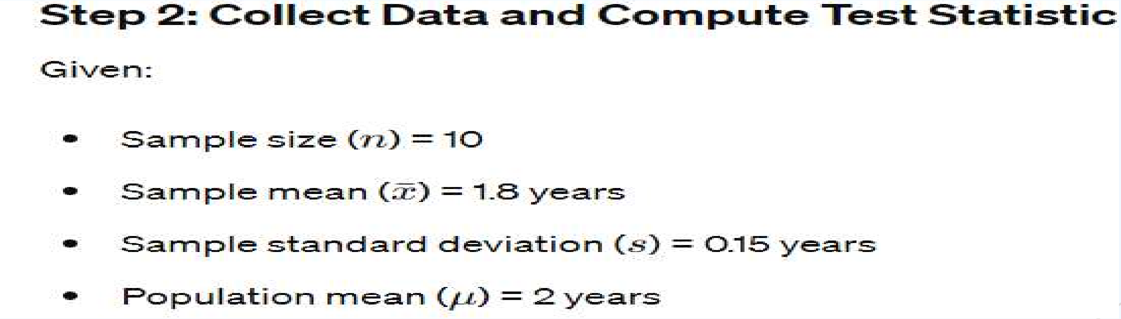
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Ques: To evaluate the engineer's claim that the average life span of car batteries is less than 2 years, we will perform a hypothesis test using the sample data provided.

Solution:

Step 1: State the Hypotheses

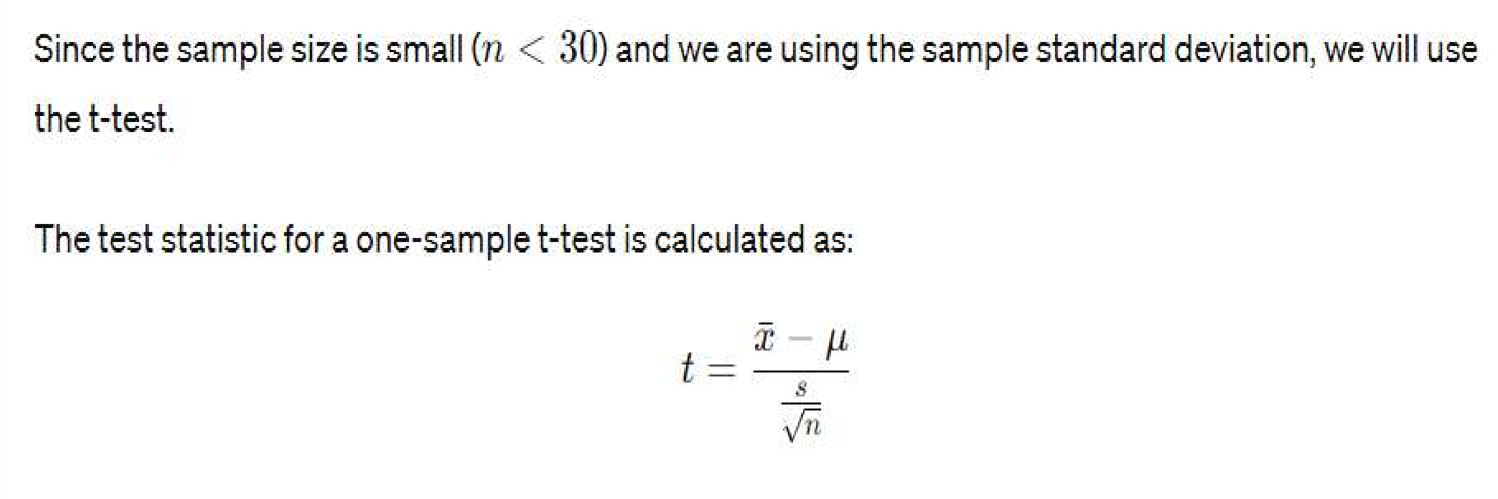
- **Null Hypothesis ($H0$):** The average life span of car batteries is 2 or more years ($\mu \geq 2$).
- **Alternative Hypothesis ($H1$):** The average life span of car batteries is less than 2 years ($\mu < 2$).



Step 2: Collect Data and Compute Test Statistic

Given:

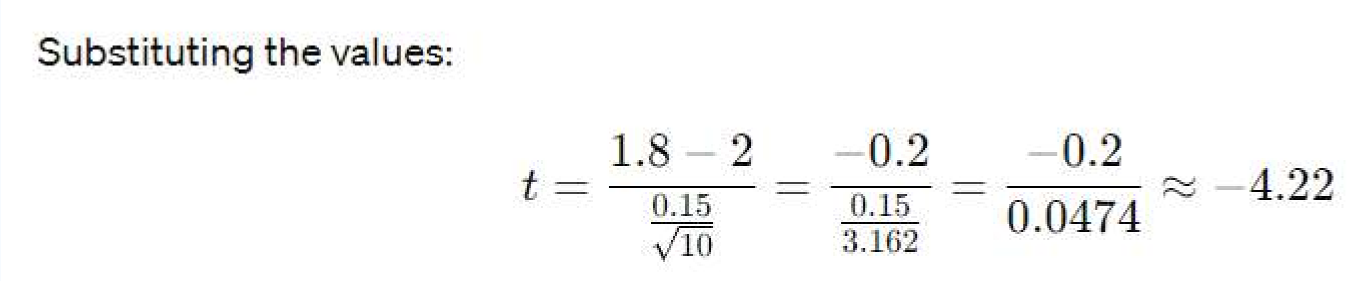
- Sample size (n) = 10
- Sample mean ($\bar{x}$) = 1.8 years
- Sample standard deviation (s) = 0.15 years
- Population mean (μ) = 2 years



Since the sample size is small ($n < 30$) and we are using the sample standard deviation, we will use the t-test.

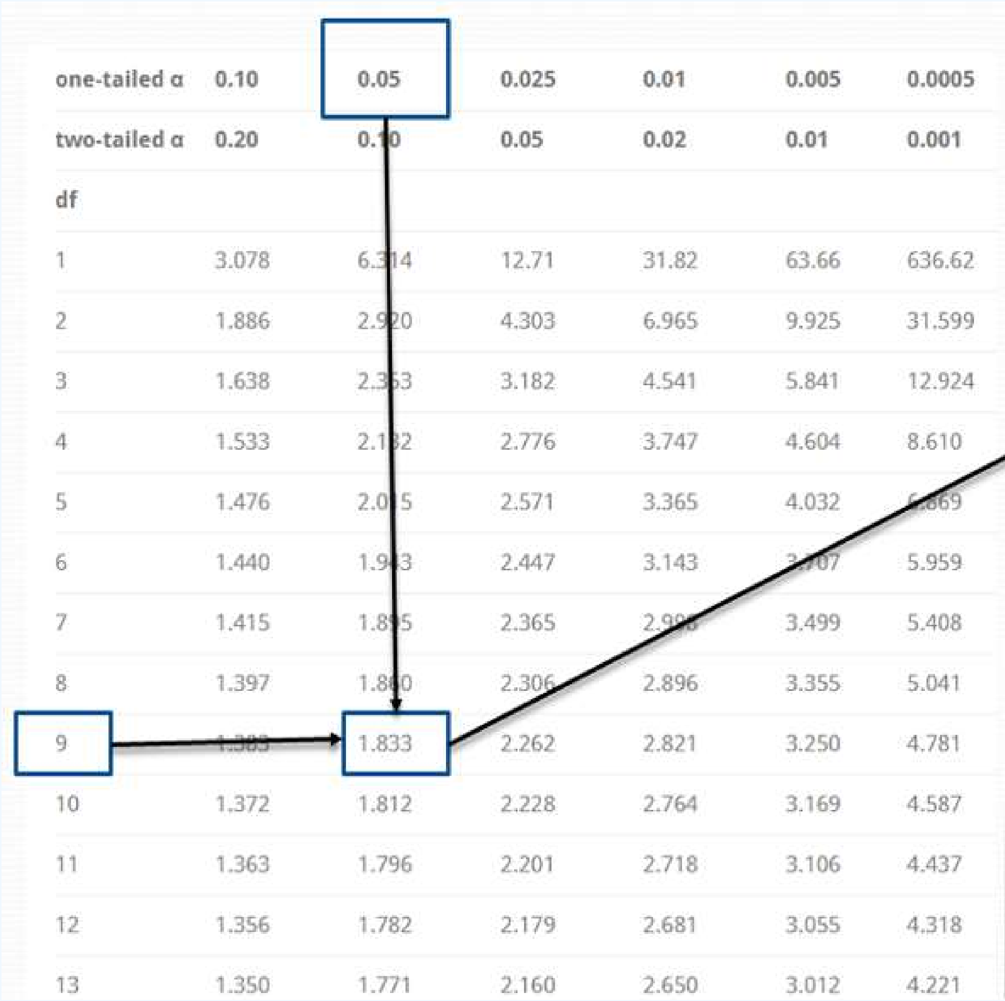
The test statistic for a one-sample t-test is calculated as:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$



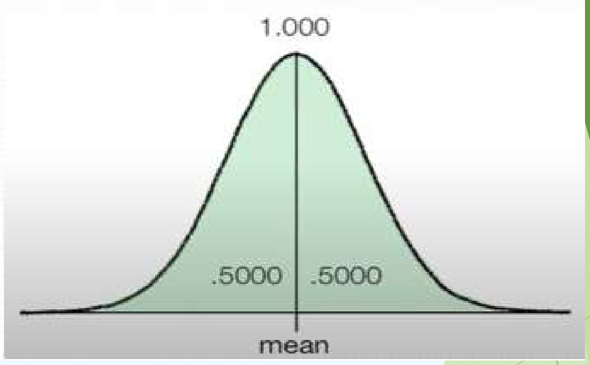
Substituting the values:

$$t = \frac{1.8 - 2}{\frac{0.15}{\sqrt{10}}} = \frac{-0.2}{\frac{0.15}{3.162}} = \frac{-0.2}{0.0474} \approx -4.22$$



one-tailed α	0.10	0.05	0.025	0.01	0.005	0.0005
two-tailed α	0.20	0.10	0.05	0.02	0.01	0.001
df						
1	3.078	6.314	12.71	31.82	63.66	636.62
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221

At 5% significance level for 9 degree of freedom t value is 1.833



- **Step 3: Determine the Critical Value**
  - For a one-tailed test at the 95% confidence level with (n−1=9) degrees of freedom, we use the t-distribution table to find the critical t-value.
  - For α=0.05 (95% confidence level), the critical t-value for 9 degrees of freedom is approximately -1.8331 (negative because it is a one-tailed test to the left).
- **Step 4: Compare Test Statistic to Critical Value**
  - If $t \leq$ critical t-value, we reject the null hypothesis.
  - If $t >$critical t-value, we fail to reject the null hypothesis.
  - In this case, the calculated t-value (-4.22) is less than the critical t-value -1.8331).

- **Step 5: Make a Decision**
  - Since t (-4.22) is less than the critical t-value -1.8331), we reject the null hypothesis.
- **Conclusion**
  - At the 95% confidence interval, there is enough evidence to reject the null hypothesis and support the engineer's claim that the average life span of car batteries is less than 2 years.

Z-Test or T-Test?

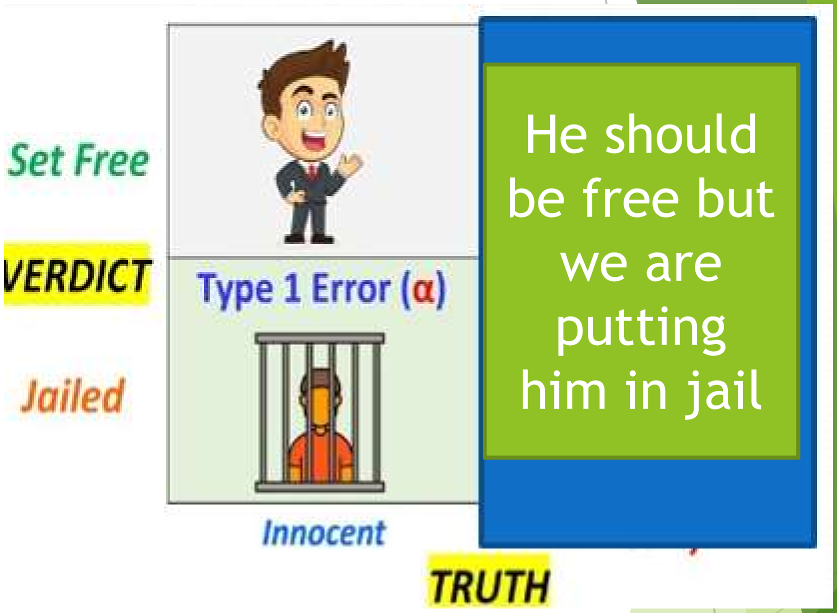
- If you know the **population standard deviation** and have a large sample size **(n > 30)**, you can **use a Z-test** for comparing means.
- If the **population standard deviation is unknown** or the sample size is small **(n < 30), use a t-test** to compare means.
- **When in doubt or when dealing with real-world data, it's often prudent to perform both tests if possible and compare the results.**
- In summary, the choice between a Z-test and a t-test depends on the availability of information about the population standard deviation, the sample size, and the underlying assumptions about the data distribution.

Errors

- Any false decision related to null hypothesis is consider as error.
- As our confidence level will never be 100% so there is always scope of error.
- Like if CL is 95% than there is a possibility that out of 100 there may be 5 times we can make wrong inferences about null hypothesis.
- Thus errors can be classified as Type 1 error and Type II error.

Type I Error

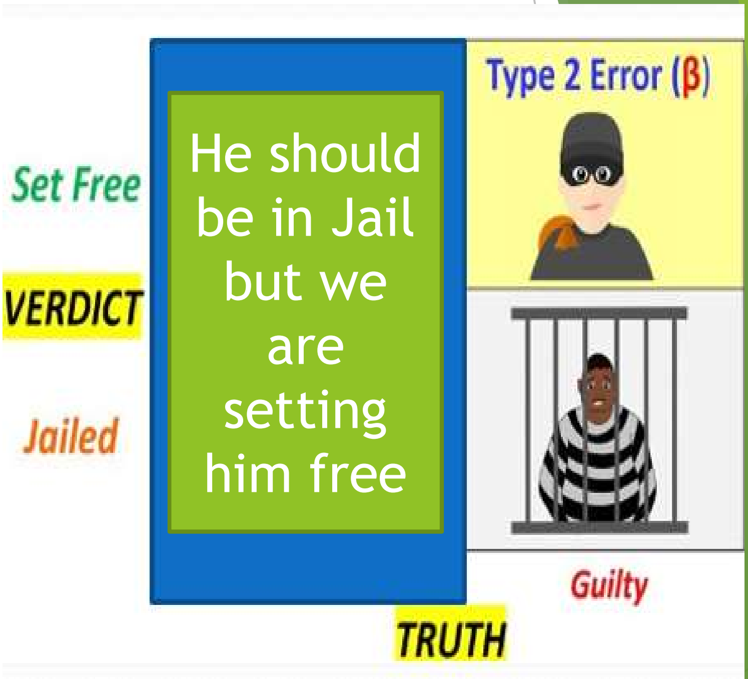
- **Type 1 Error(False Positive):** Rejected Null hypothesis when it is true.
- In other words, it's a false alarm or a false positive result when we conclude that there is a significant effect or difference while actually there isn't one in reality.
- The probability of making a Type I error is denoted by α (alpha) and is typically set as the significance level in hypothesis testing (e.g., α = 0.05 or 0.01).



He should be free but we are putting him in jail

Type II Error

- **Type II Error(False Negative) :** Rejected Alternate hypothesis when it is true.
- It's a miss or a false negative result where we fail to detect a significant effect or difference when it actually exists.
- The probability of making a Type II error is denoted by β (beta).



He should be in Jail but we are setting him free

Relationship with Hypotheses:

- **Type I Error (False Positive):**
  - **Null Hypothesis (H0):** Represents no effect, no difference, or no relationship.
  - **Type I error** occurs when we wrongly reject a true null hypothesis (H0 is true, but we reject it).
- **Type II Error (False Negative):**
  - **Alternative Hypothesis (Ha):** Represents an effect, difference, or relationship.
  - **Type II error** occurs when we fail to reject a false null hypothesis (H0 is false, but we fail to reject it).

- There is often a trade-off between Type I and Type II errors. Lowering the significance level (α) to reduce Type I errors can increase the likelihood of Type II errors and vice versa.

- Power of a test is the probability of correctly rejecting a false null hypothesis (1 - β).
- Higher power means lower chances of Type II errors.

ANOVA (Analysis of Variance)

- **Theory:** ANOVA is used to compare means across multiple groups or treatments.
  - It assesses whether there are statistically significant differences between the means of three or more independent groups.
  - ANOVA tests the **null hypothesis** that all group means are equal.
  - ANOVA is used to check significance of a variable before dropping it.

- **Purpose:**
  - **Hypothesis Testing:** To test if there is a significant difference between group means.
  - **Variance Analysis:** To understand how variance in data is attributed to different sources.
  - **Multiple Comparisons:** To avoid the increased risk of Type I errors when making multiple comparisons.

Assumptions of ANOVA

- **Independence:** Samples must be independent of each other.
- **Normality:** The data in each group should be approximately normally distributed.
- **Homogeneity of Variances:** The variances among the groups should be approximately equal.

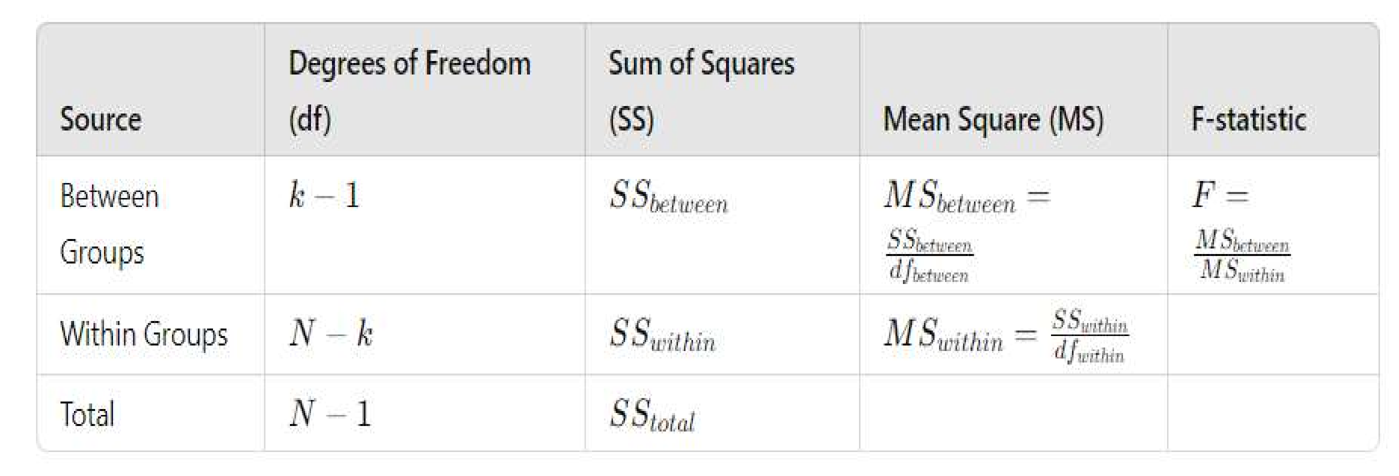
- **Types of ANOVA:**
  - One-Way ANOVA
  - Two-Way ANOVA

One-Way ANOVA:

- **Purpose:** One-Way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if there is a statistically significant difference among them.
- It is called "one-way" because it examines the impact of a single factor or independent variable.

- **Example :**
  - Suppose you are a teacher who wants to compare the **exam scores of students** from **three different teaching methods:** Traditional, Online, and Hybrid. You collect the exam scores from students in each of these three groups.

Elements of ANOVA Table



Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-statistic
Between Groups	$k-1$	$SS_{between}$	$MS_{between} = \frac{SS_{between}}{df_{between}}$	$F = \frac{MS_{between}}{MS_{within}}$
Within Groups	$N-k$	SS_{within}	$MS_{within} = \frac{SS_{within}}{df_{within}}$	
Total	$N-1$	SS_{total}		

- **Source** : Identifies different sources of variance (between groups, within groups, total).
- **Sum of Squares (SS)**: Measures the total variance attributable to each source.
- **Degrees of Freedom (df)**: The number of independent values that can vary in the analysis.
- **Mean Square (MS)**: Average of the sum of squares, calculated as MS=SS/df.
- **F-Statistic (F)**: Ratio of the mean square(variance) between groups to the mean square(variance) within groups.
- N denotes the total number of observations across k groups.

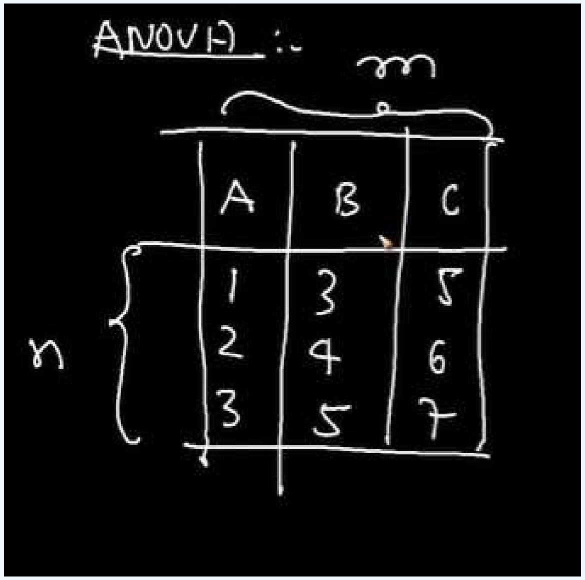
Example: One Way ANOVA

Ques: We have **three drugs** (A, B, and C) and **three observations** for each drug. The data is structured as follows:

Drug A out comes: 1, 2, 3

Drug B out comes: 3, 4, 5

Drug C out comes: 5, 6, 7



ANOVA :-

m

n

A	B	C
1	3	5
2	4	6
3	5	7

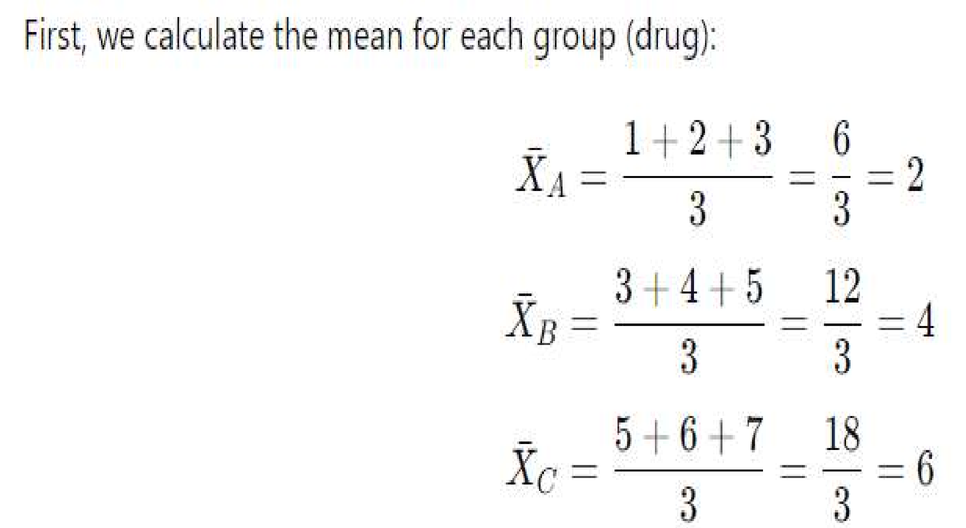
- ➢ Let's perform a one-way ANOVA on the given data.

Problem Statement

- **Research Question:**
  - "Is there a significant difference in the mean outcomes of the three drugs (A, B, and C)?"
- **Null Hypothesis (H0):**
  - There is no significant difference in the mean outcomes of the three drugs.
- **Alternative Hypothesis (H1):**
  - There is a significant difference in the mean outcomes of at least one of the drugs compared to the others.

Step-by-Step One-Way ANOVA Calculation

Step 1: Calculate Group Means

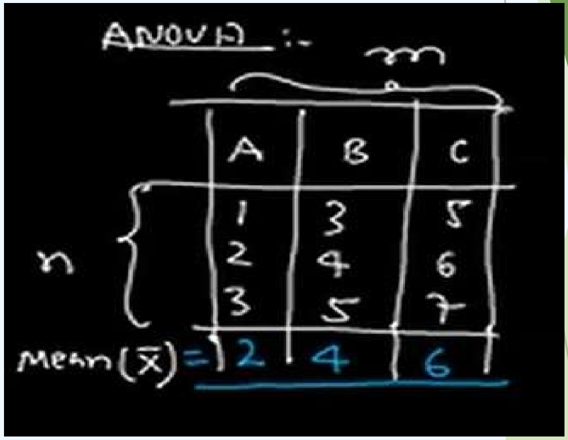


First, we calculate the mean for each group (drug):

$$\bar{X}_A = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

$$\bar{X}_B = \frac{3+4+5}{3} = \frac{12}{3} = 4$$

$$\bar{X}_C = \frac{5+6+7}{3} = \frac{18}{3} = 6$$



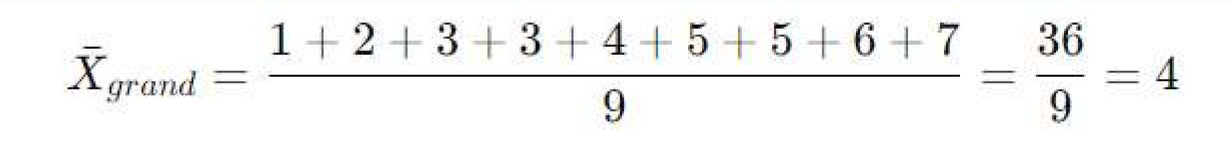
ANOVA :-

m

	A	B	C
n	1	3	5
	2	4	6
	3	5	7
Mean($\bar{x}$) =	2	4	6

Step 2: Calculate the Grand Mean

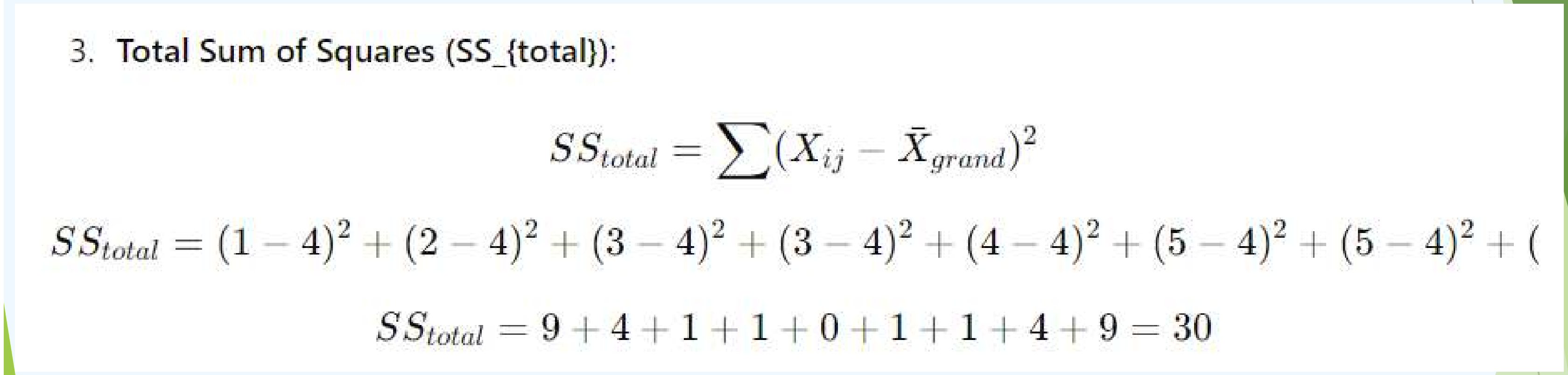
- The grand mean ($X^{-}grand$) is the mean of all observations combined:



$$\bar{X}_{grand} = \frac{1+2+3+3+4+5+5+6+7}{9} = \frac{36}{9} = 4$$

Step 3: Calculate Sum of Squares:

- **Total (SST):** Sum of square of difference between grand mean value and individual element of the table.

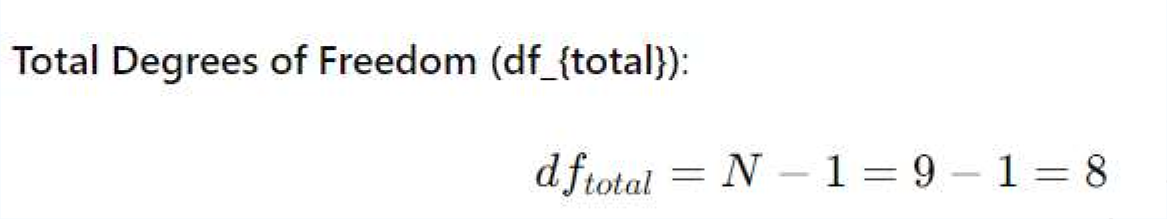


3. Total Sum of Squares (SS_{total}):

$$SS_{total} = \sum (X_{ij} - \bar{X}_{grand})^2$$

$$SS_{total} = (1-4)^2 + (2-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (5-4)^2 + ($$

$$SS_{total} = 9 + 4 + 1 + 1 + 0 + 1 + 1 + 4 + 9 = 30$$

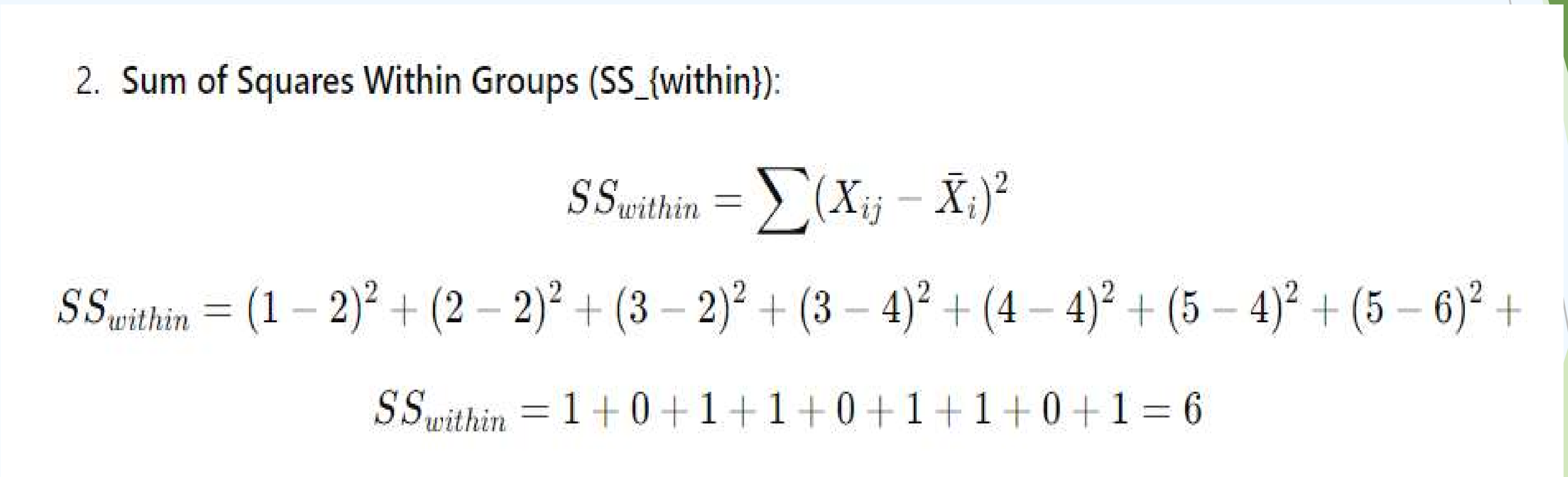


Total Degrees of Freedom (df_{total}):

$$df_{total} = N - 1 = 9 - 1 = 8$$

Sum of Square within rows(SSW)

- Sum of square of difference between mean value of each column and individual element of that column.

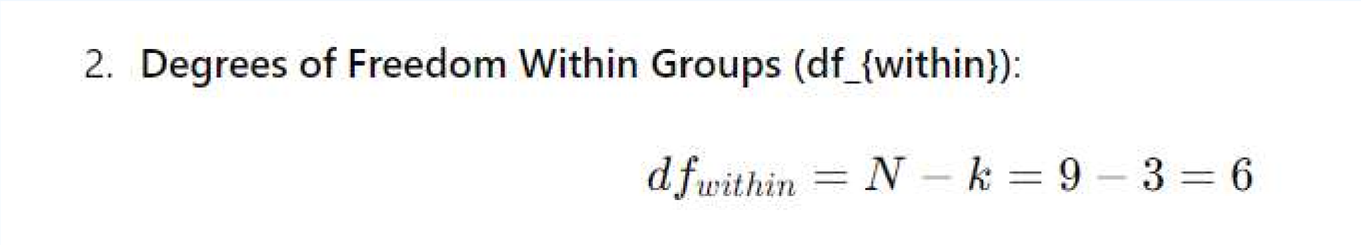


2. Sum of Squares Within Groups (SS_{within}):

$$SS_{within} = \sum (X_{ij} - \bar{X}_i)^2$$

$$SS_{within} = (1-2)^2 + (2-2)^2 + (3-2)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (5-6)^2 +$$

$$SS_{within} = 1 + 0 + 1 + 1 + 0 + 1 + 1 + 0 + 1 = 6$$

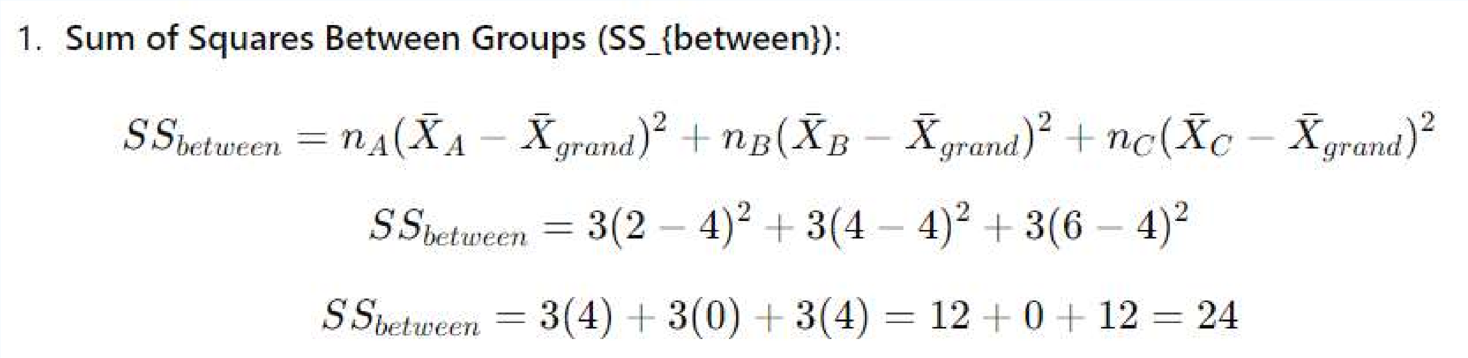


2. Degrees of Freedom Within Groups (df_{within}):

$$df_{within} = N - k = 9 - 3 = 6$$

Calculate Sum of Square between(SSB)

- SSB = n{sum of (Grand Mean - Actual mean)**2}
- Where n = no rows
- For DOF , k = number of columns

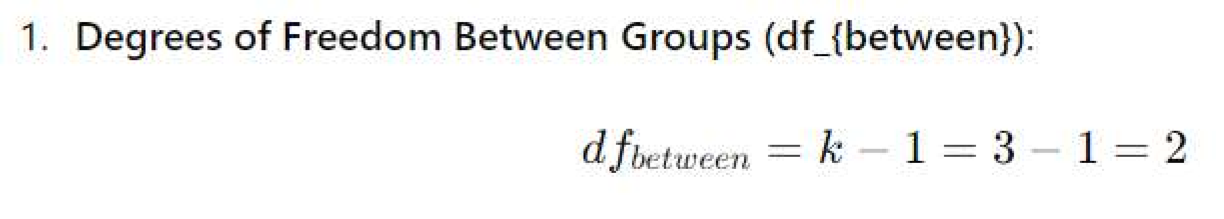


1. Sum of Squares Between Groups (SS_{between}):

$$SS_{between} = n_A(\bar{X}_A - \bar{X}_{grand})^2 + n_B(\bar{X}_B - \bar{X}_{grand})^2 + n_C(\bar{X}_C - \bar{X}_{grand})^2$$

$$SS_{between} = 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2$$

$$SS_{between} = 3(4) + 3(0) + 3(4) = 12 + 0 + 12 = 24$$



1. Degrees of Freedom Between Groups (df_{between}):

$$df_{between} = k - 1 = 3 - 1 = 2$$

- **Step 4: Relationship between SST, SSW, and SSB:**

SST = SSW + SSB

- The equation shows that the total variation (SST) is the sum of the within-group variation (SSW) and the between-group variation (SSB):

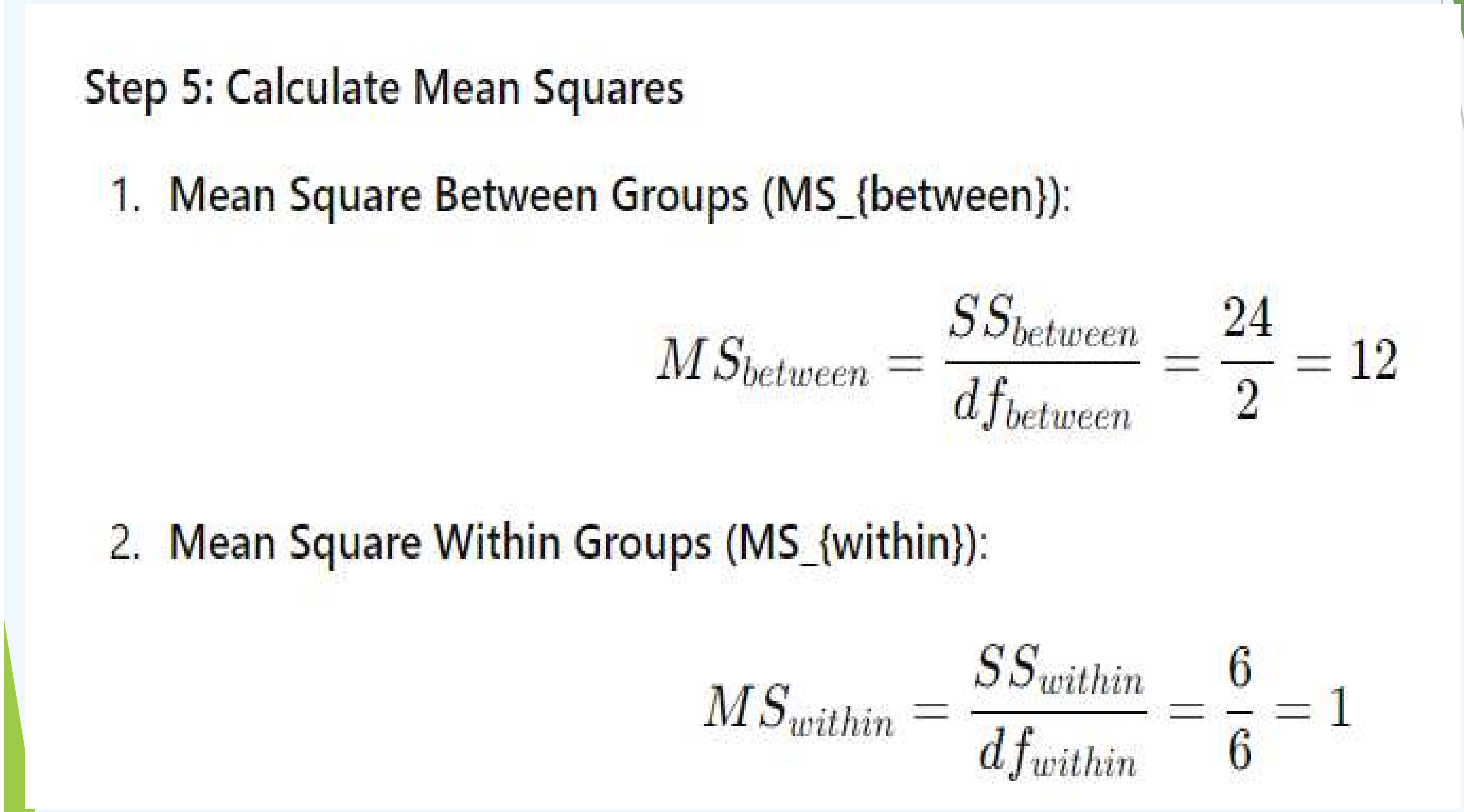
30 = 6 + 24

- Relationship between DOF(T), DOF(W), and DOF(B):

DOF(T) =DOF(W) + DOF(B)

8 = 6 + 2

As LHS = RHS, so this relationship holds true.



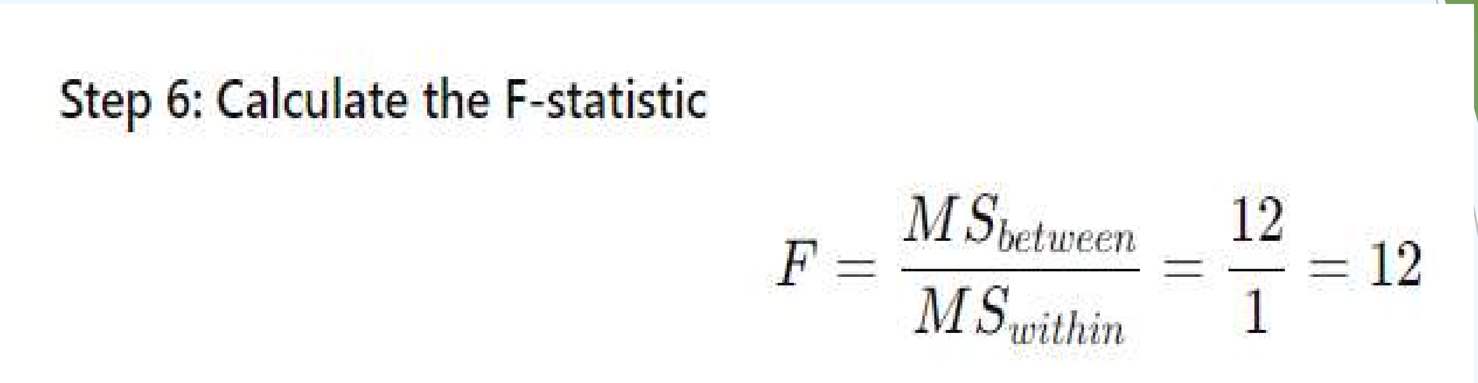
Step 5: Calculate Mean Squares

1. Mean Square Between Groups (MS_{between}):

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{24}{2} = 12$$

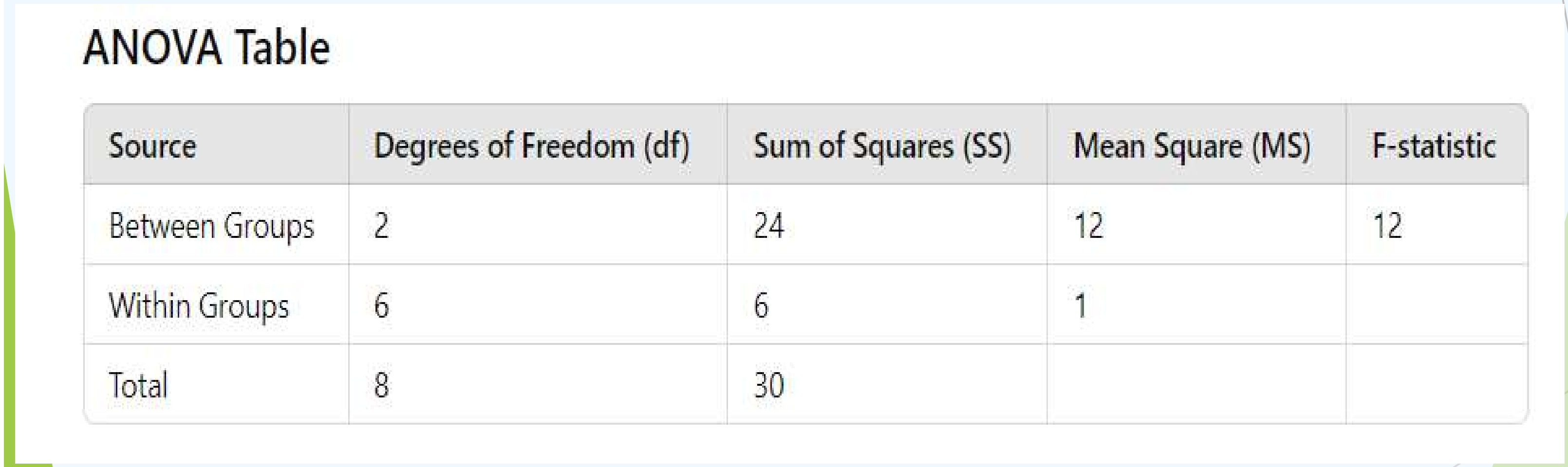
2. Mean Square Within Groups (MS_{within}):

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{6}{6} = 1$$



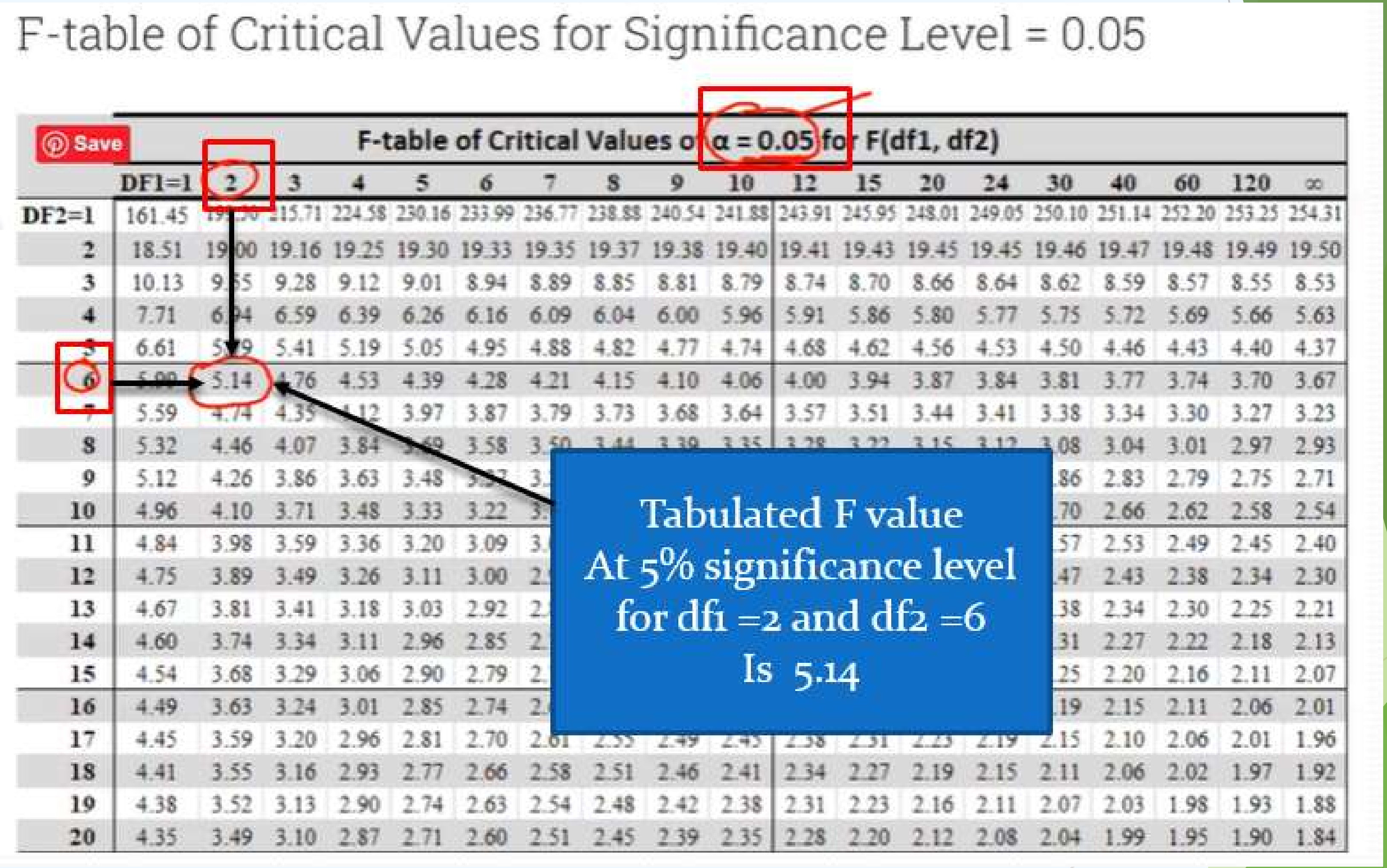
Step 6: Calculate the F-statistic

$$F = \frac{MS_{between}}{MS_{within}} = \frac{12}{1} = 12$$



ANOVA Table

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-statistic
Between Groups	2	24	12	12
Within Groups	6	6	1	
Total	8	30		

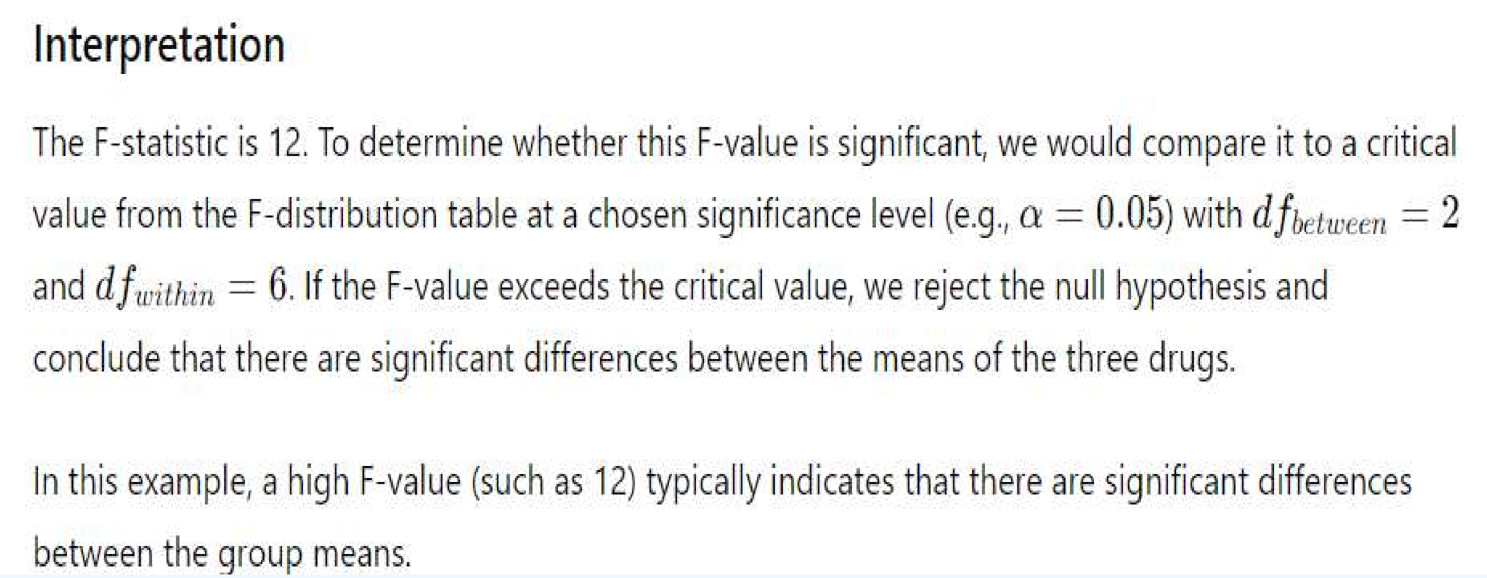


F-table of Critical Values for Significance Level = 0.05

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84

Tabulated F value
At 5% significance level
for df1 =2 and df2 =6
Is 5.14



Interpretation

The F-statistic is 12. To determine whether this F-value is significant, we would compare it to a critical value from the F-distribution table at a chosen significance level (e.g., $\alpha = 0.05$) with $df_{between} = 2$ and $df_{within} = 6$. If the F-value exceeds the critical value, we reject the null hypothesis and conclude that there are significant differences between the means of the three drugs.

In this example, a high F-value (such as 12) typically indicates that there are significant differences between the group means.

- F(calculated) =12
- F(tabulated) = 5.14
- Since F(calculated) > F(tabulated)
- Will go with Alternate hypothesis testing.

Application in ML

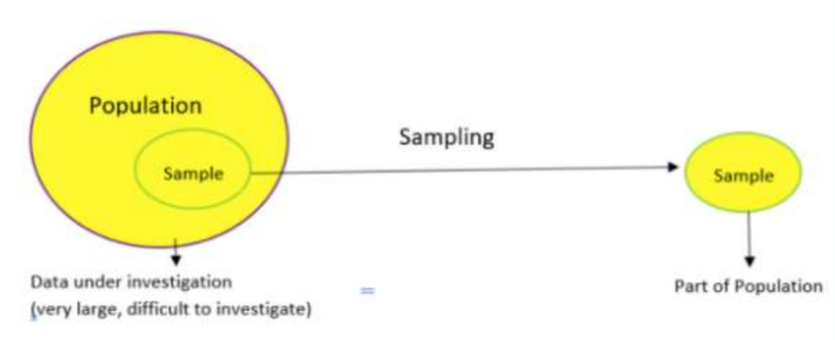
- Feature Selection: t test and chi Sq test
- A/B testing
- Time series: Augmented Dickey Fuller test

Sampling Techniques

Understanding the Basics of Sampling

What is Sampling?

- Sampling is a method to obtain information about a population based on a subset (sample).

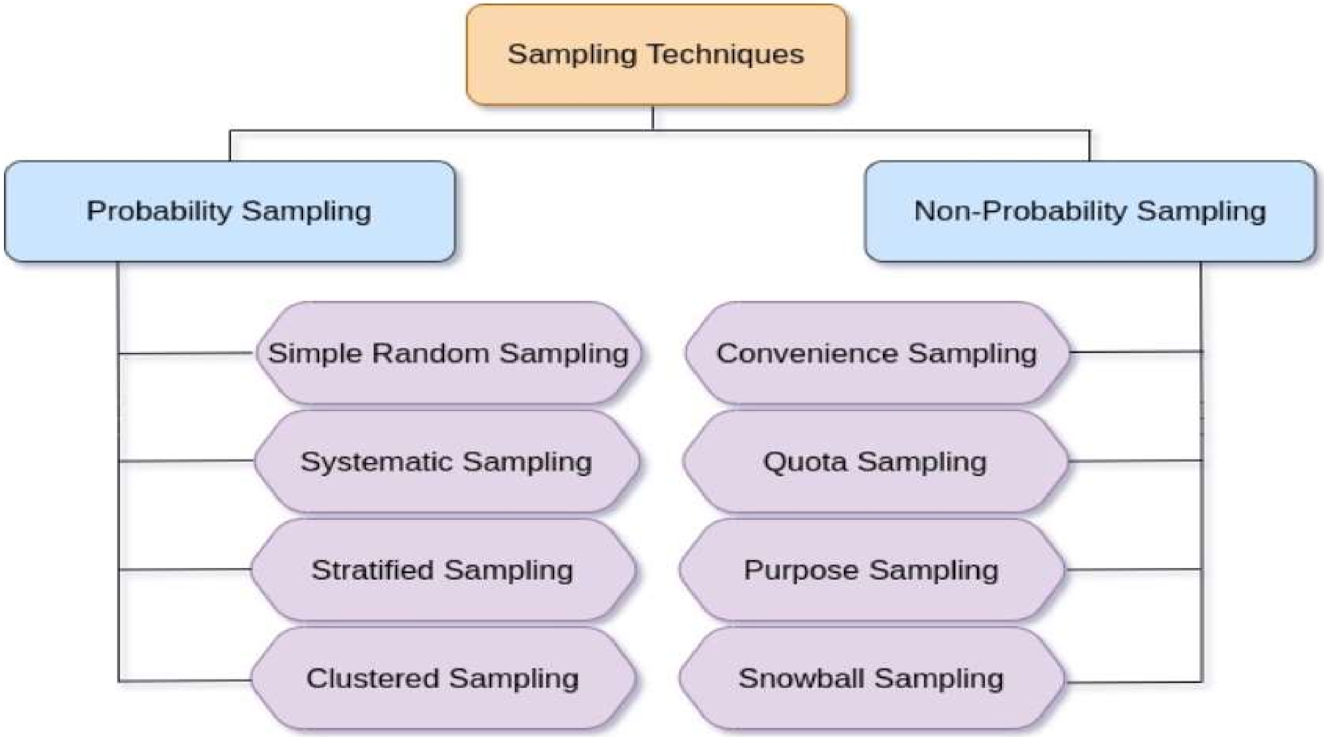


Why Do We Need Sampling?

- **Less time-consuming** than investigating the entire population.
- **Cost-efficient**.
- **Practical and manageable** for analysis.

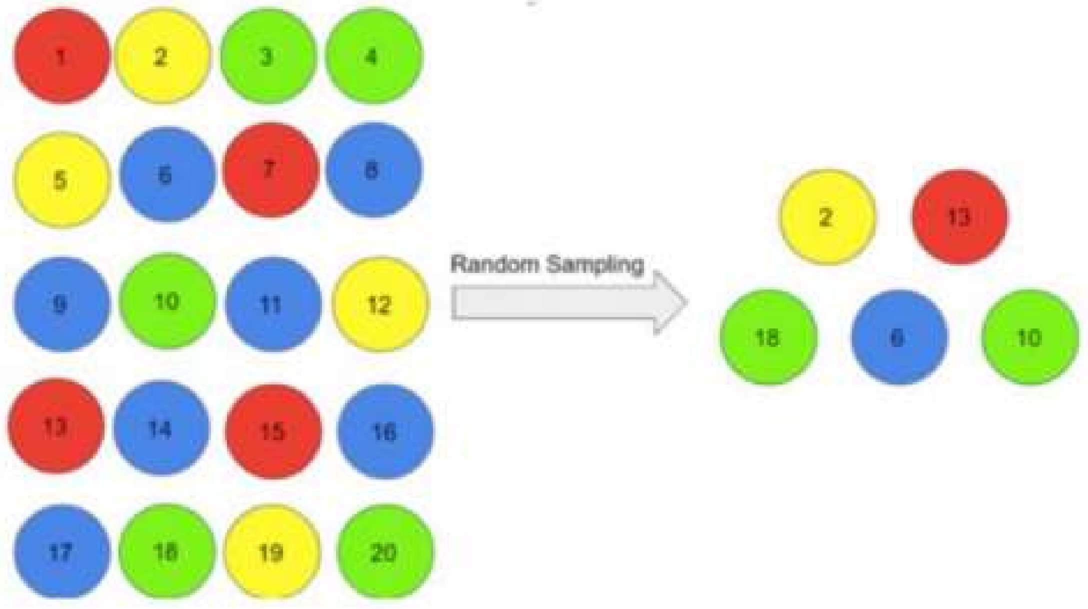
Types of Sampling

- **Probability Sampling**: Every element has an equal chance of selection.
- **Non-Probability Sampling**: Not every element has an equal chance.



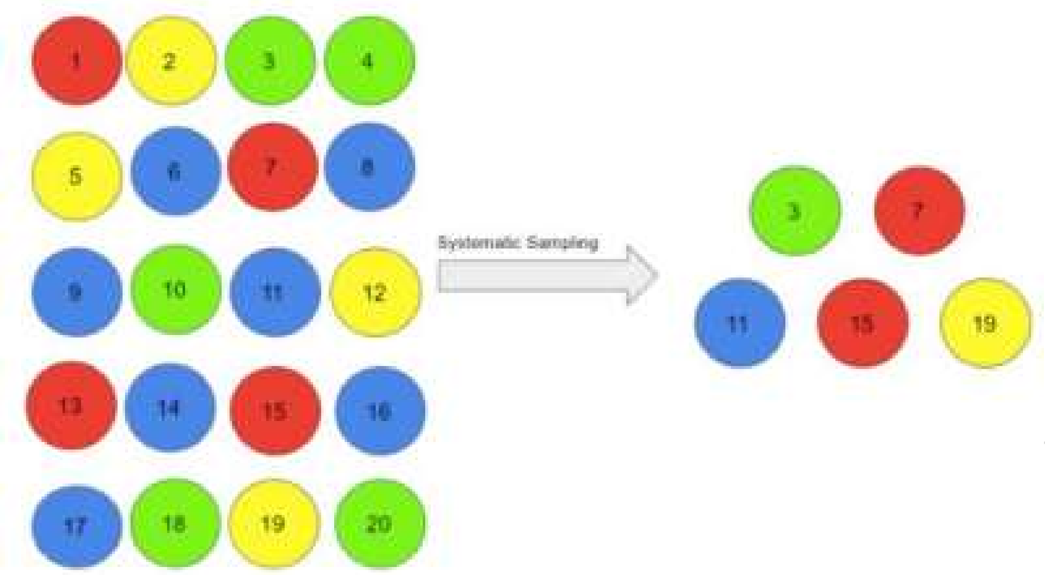
Simple Random Sampling

- **Simple Random Sampling:** Each data point in the population has an equal chance of being selected.
  - **Statistical Context:** Ensures each sample has an equal probability of selection.



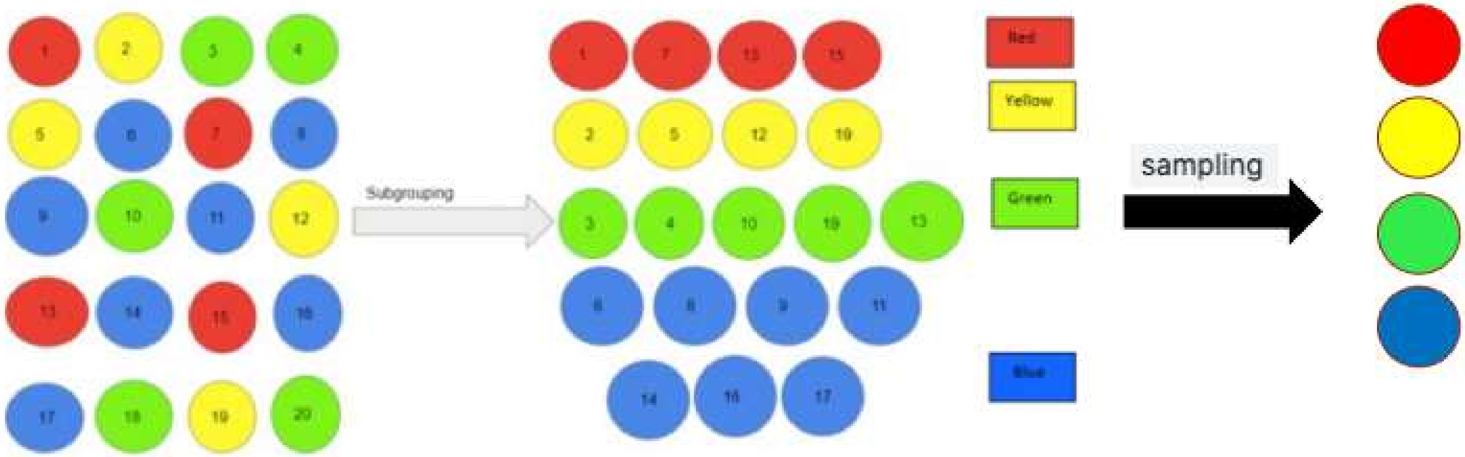
Systematic Sampling

- Every k-th element in the population selected, starting from a randomly chosen element.
- Selecting members at regular intervals from a population.
  - **Example**: Choosing every 10th individual in a population of 5000 people.



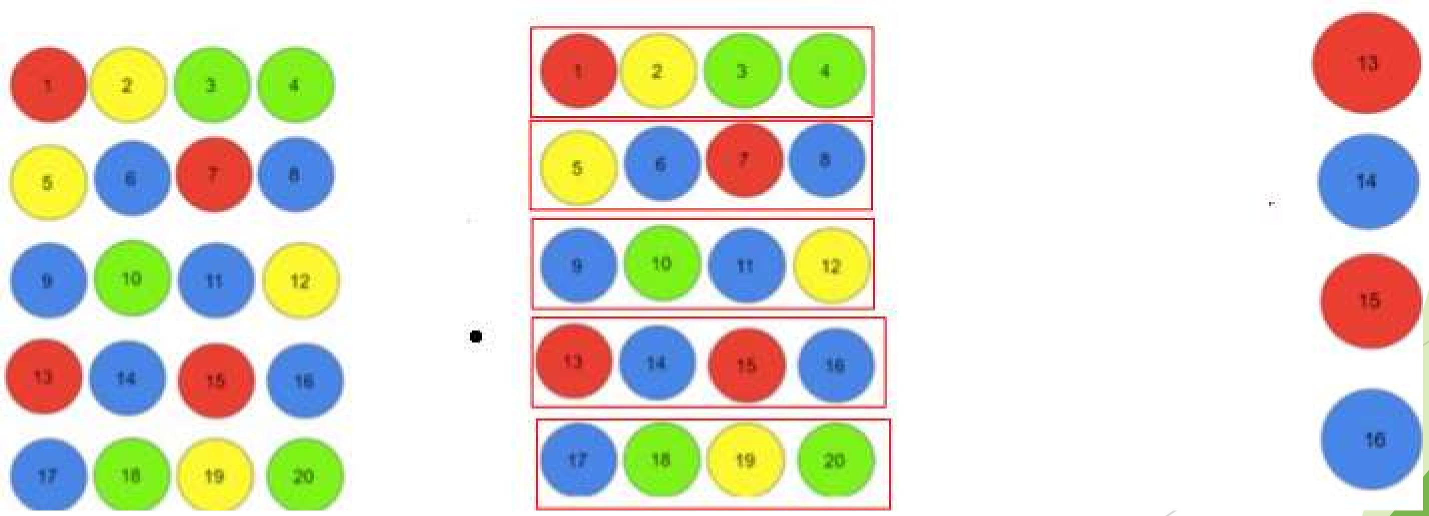
Stratified Sampling

- The population is divided into subgroups based on specific traits (e.g., gender, age), and samples are taken from each group.
- This ensures subgroups are proportionately represented.
- **Example**: Ensuring gender representation by sampling from both male and female strata.



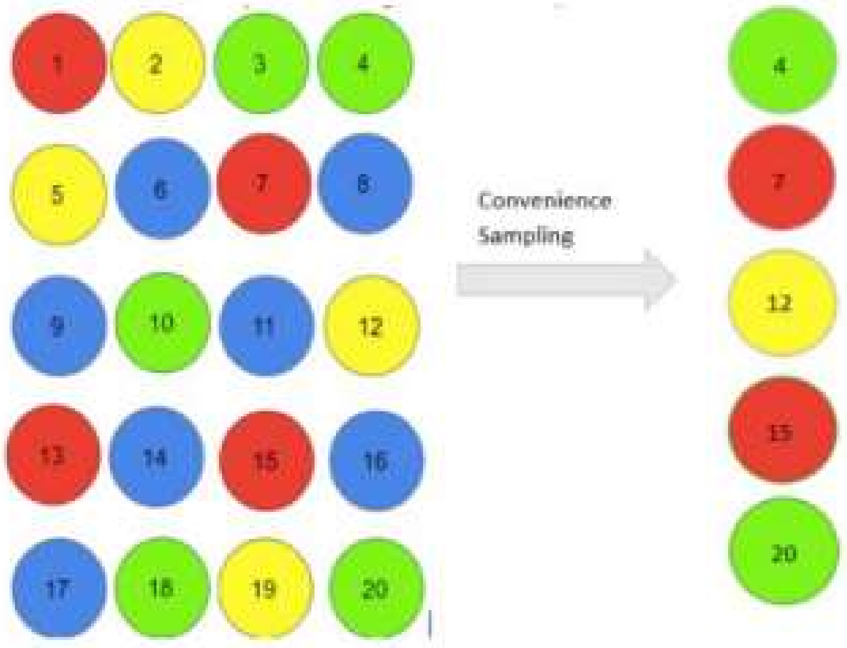
Cluster Sampling

- The population is divided into subgroups (clusters), and a random cluster is selected for study.
- **Example**: Choosing one school from several schools to study the behavior of students.



Non-Probability Sampling

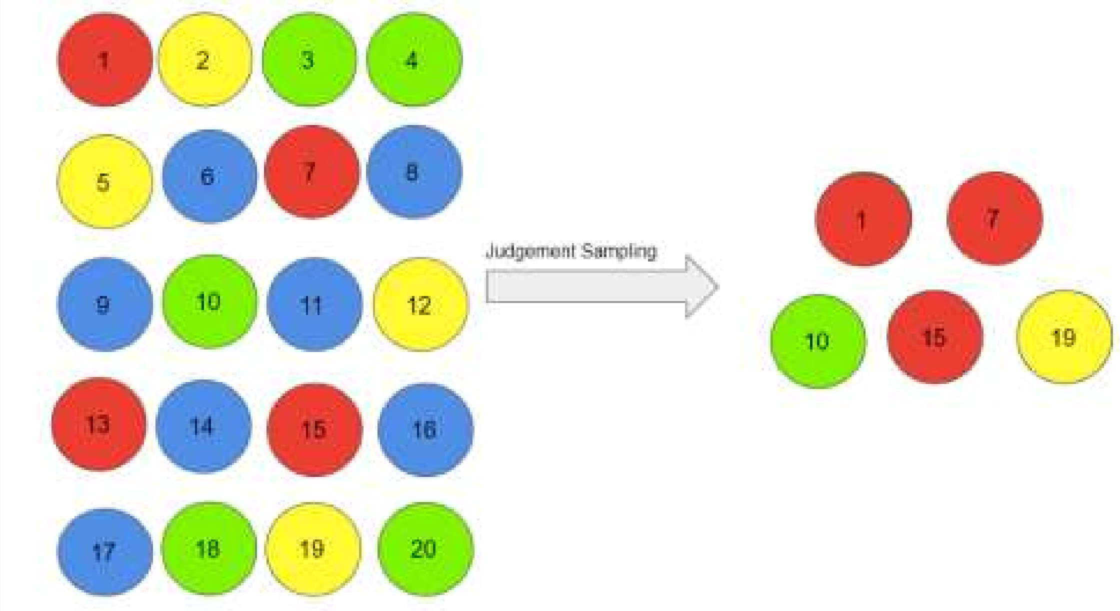
- **Convenience Sampling**: Choosing individuals based on their availability.
- **Example**: Distributing pamphlets at a mall to whoever is present.
- Rarely used due to risk of bias but might be used for quick, preliminary analyses.



- **Judgmental/Purposive Sampling:** Samples are chosen based on the judgment of the researcher.

- Experts choose participants they believe will best represent the population.

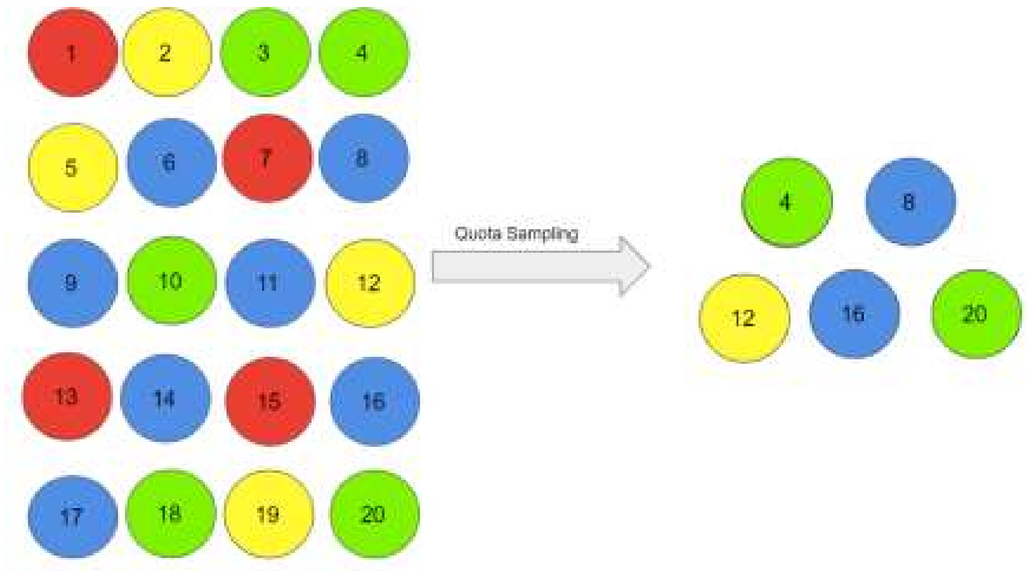
- **Example**: Selecting only individuals with specific expertise in a technical field.



- **Quota Sampling:** Similar to stratified sampling but not random. Samples are chosen until a quota is met.

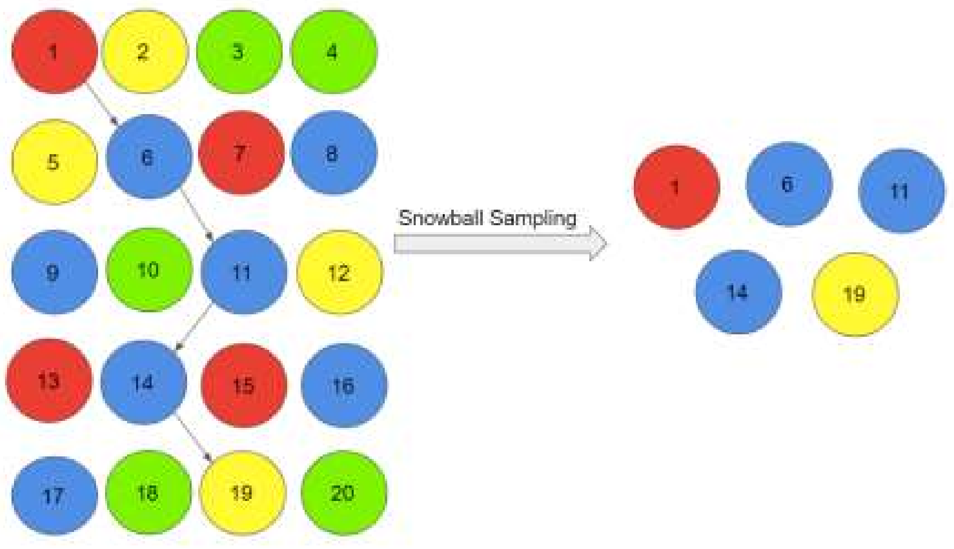
- Dividing the population into subgroups and selecting individuals based on predetermined quotas.

- **Example**: Ensuring that 20% of the sample is from a minority group, even if it may not fully represent the population.



➢**Snowball Sampling:** Existing study subjects recruit future subjects from among their acquaintances

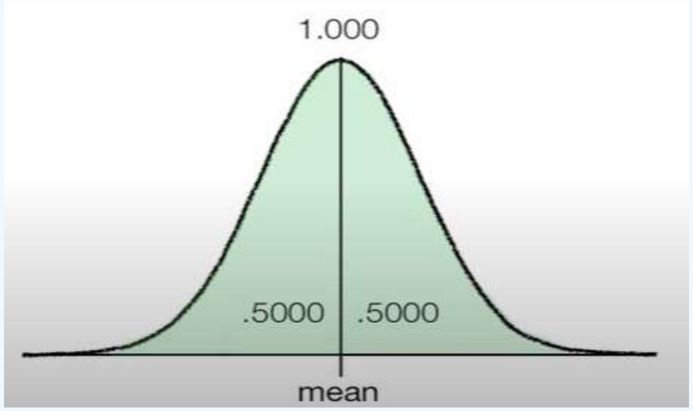
- ➢ Existing participants nominate others, leading to a growing sample.
- ➢ **Example**: In a study on rare diseases, participants recommend others with the same condition.

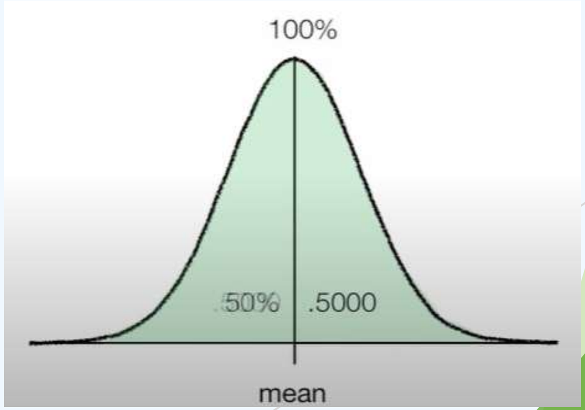


1->6->11->14->19

How to calculate P-value from table

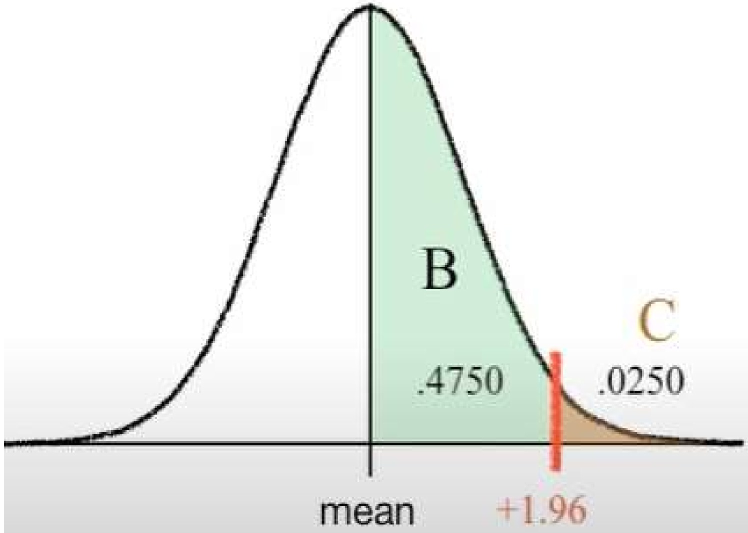
- Calculating a p-value from a table typically involves using statistical tables such as the Z-table (for the standard normal distribution), the t-table (for the Student's t-distribution), or chi-square and F-distribution tables.
- A p-value plot typically shows the distribution of the test statistic under the null hypothesis.





- **Cumulative Distribution Function (CDF)**
  - The CDF of a random variable X gives the probability that X will take a value less than or equal to x. Mathematically, for a continuous random variable, the CDF is defined as: F(x)=P(X≤x).
  - It talks about area under bell curve.

95% Confidence Level



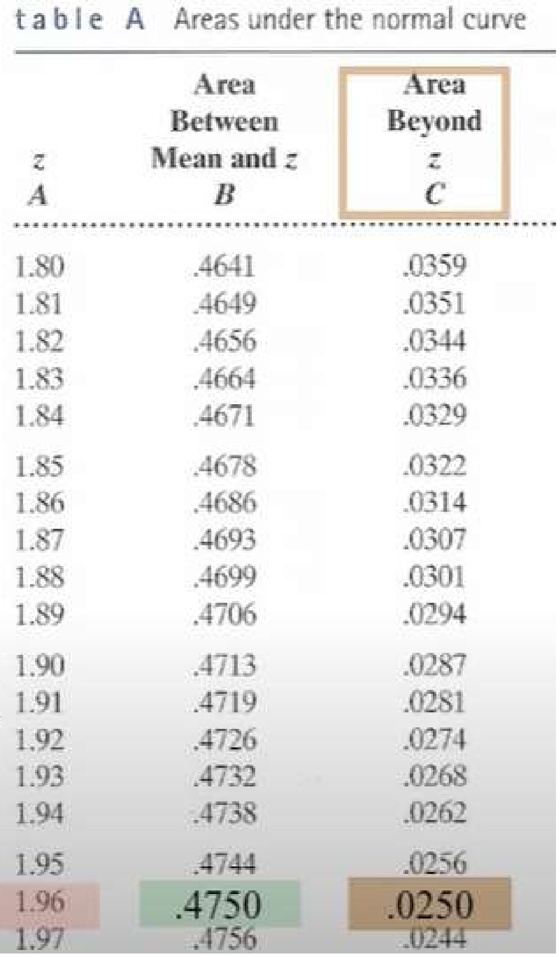
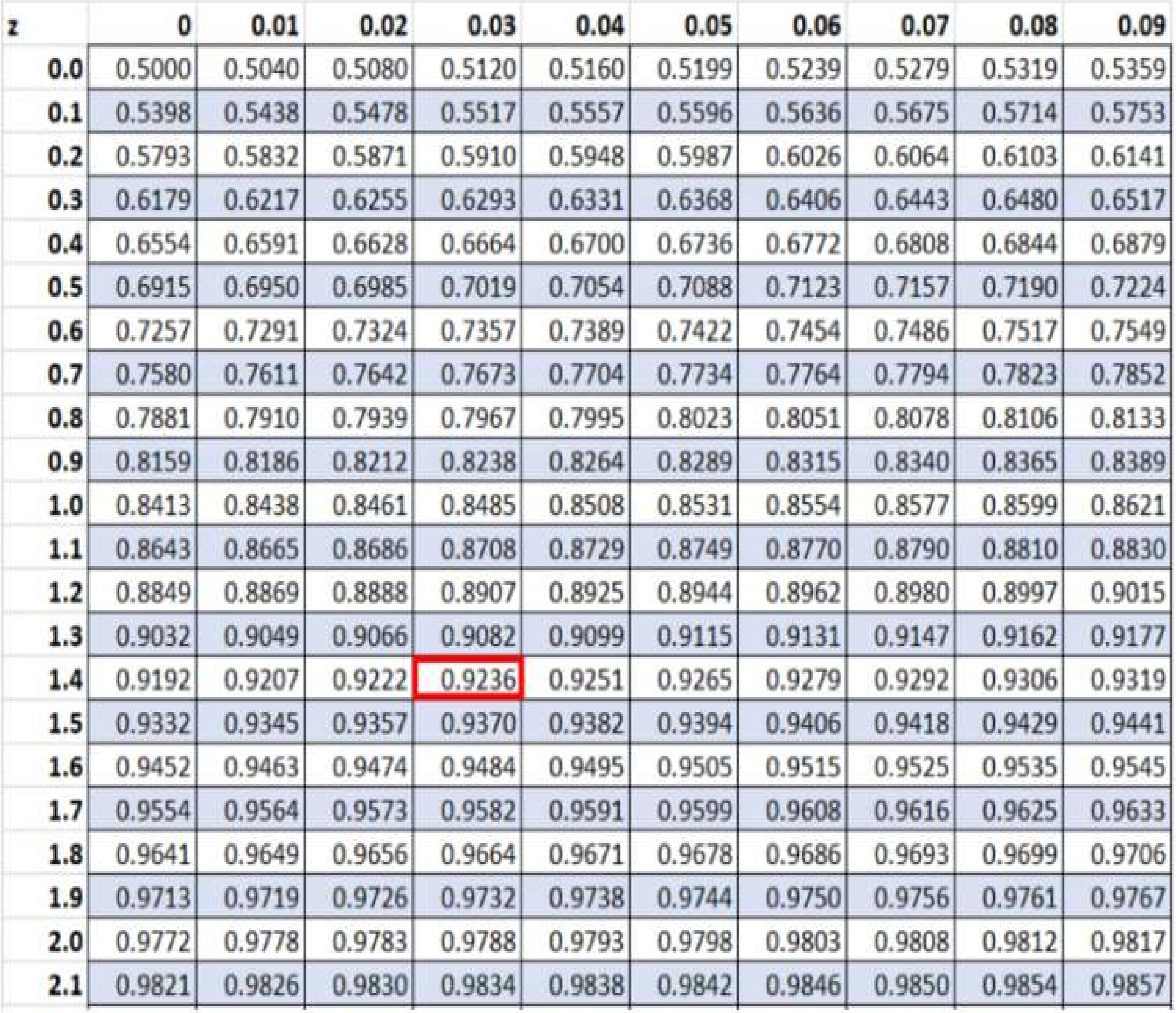
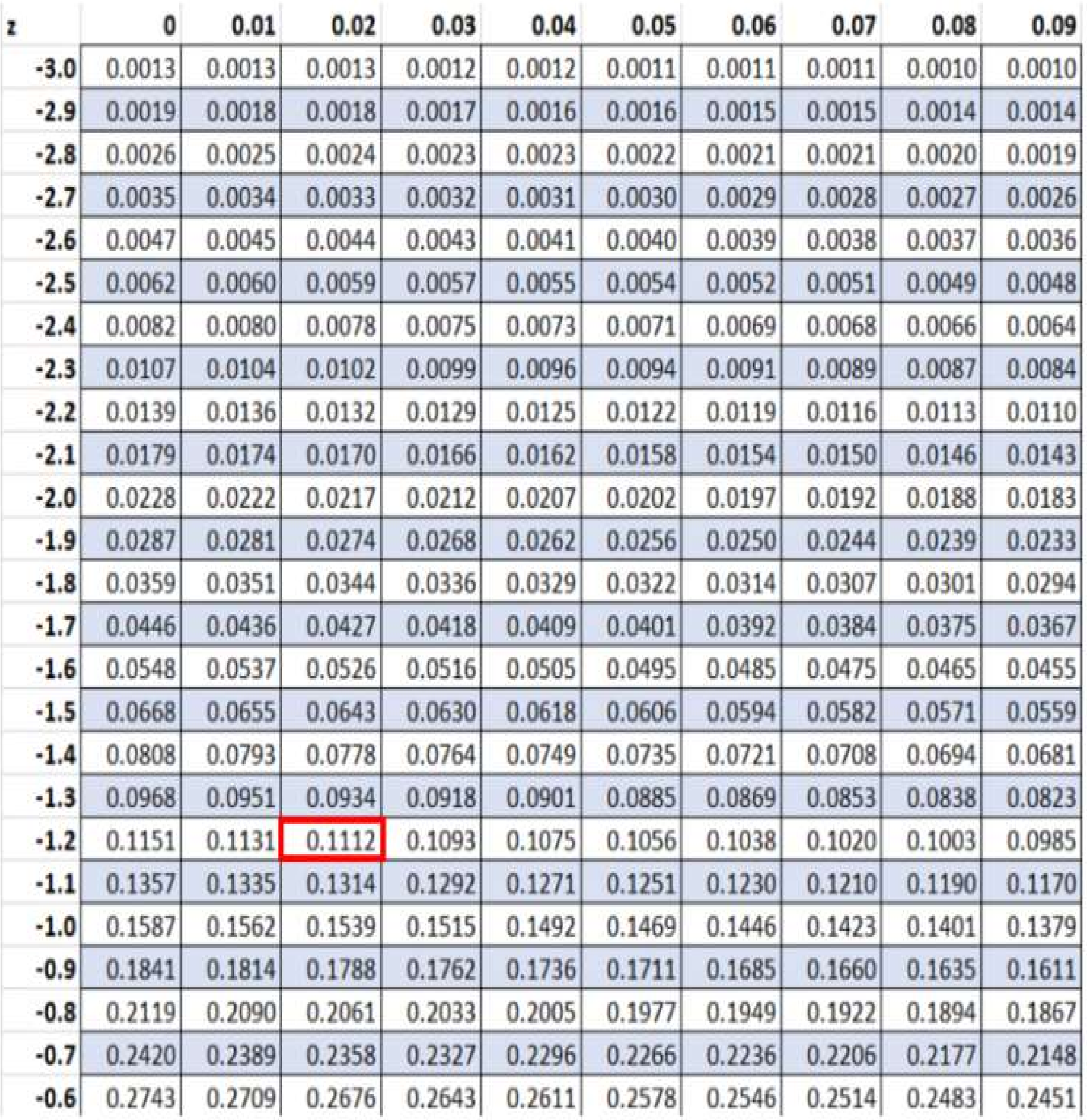


table A Areas under the normal curve

z *A*	Area Between Mean and z *B*	Area Beyond z *C*
1.80	.4641	.0359
1.81	.4649	.0351
1.82	.4656	.0344
1.83	.4664	.0336
1.84	.4671	.0329
1.85	.4678	.0322
1.86	.4686	.0314
1.87	.4693	.0307
1.88	.4699	.0301
1.89	.4706	.0294
1.90	.4713	.0287
1.91	.4719	.0281
1.92	.4726	.0274
1.93	.4732	.0268
1.94	.4738	.0262
1.95	.4744	.0256
1.96	.4750	.0250
1.97	.4756	.0244



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451

Thanks for giving your time…..