STATISTICS

- Statistics is the science, or a branch of mathematics, that involves collecting, classifying, analyzing, interpreting, and presenting numerical facts and data.

- It is especially handy when dealing with populations too numerous and extensive for specific, detailed measurements.

- Statistics are crucial for drawing general conclusions relating to a dataset from a data sample.

DESCRIPTIVE STATISTICS

- Descriptive statistics describe, show, and summarize the basic features of a dataset and presents it in a summary.

- So, this describes the data sample and its measurements.

- It helps analysts to understand the data better.

- Descriptive statistics represent the available data sample and do not include theories, inferences, probabilities, or conclusions.

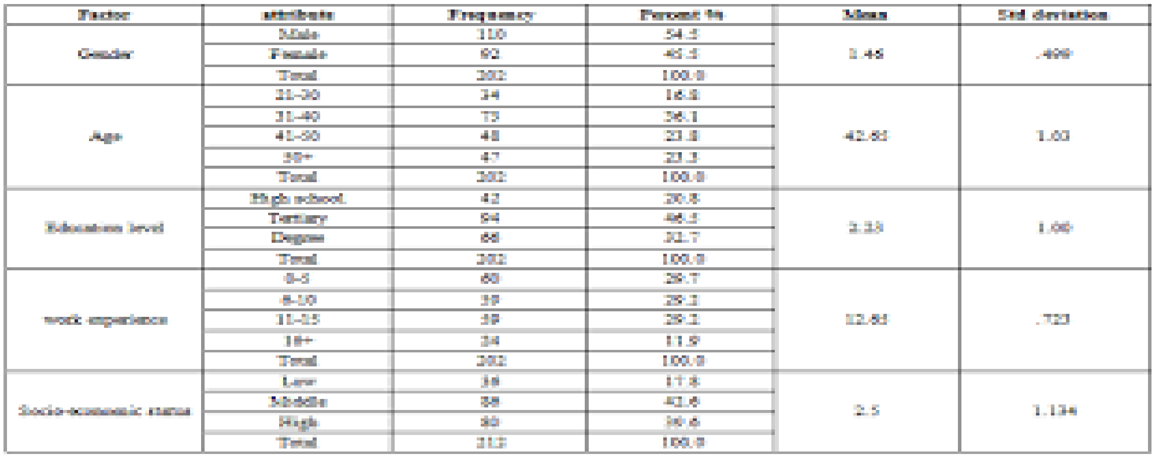
- For eg, A student’s GPA. A GPA collects all the grades that the student scores of different phases of exam and then presents the general idea of the academic performance which is the average of all scores. This is actually a summary of his performance.

- Another eg, 50 people watch a movie and we ask all of them if they like the movie or not, and collect the result and project it in a bar graph. This can be considered as a descriptive statistics.

- Political polling is also considered another example.

TYPES OF DESCRIPTIVE STATISTICS

- **Distribution (Frequency distribution)**
  It has distribution of scores or values. Statisticians use graphs and tables to summarize the frequency of every possible value of a variable, rendered in percentages or numbers.



Factor	attribute	Frequency	Percent %	Mean	Std deviation
Gender	Male	110	54.5	1.46	.499
	Female	92	45.5		
	Total	202	100.0		
Age	21-30	34	16.8	42.65	1.03
	31-40	73	36.1		
	41-50	48	23.8		
	50+	47	23.3		
	Total	202	100.0		
Education level	High school	42	20.8	2.13	1.06
	Tertiary	94	46.5		
	Degree	66	32.7		
	Total	202	100.0		
work experience	0-5	60	29.7	12.65	.723
	6-10	59	29.2		
	11-15	59	29.2		
	16+	24	11.9		
	Total	202	100.0		
Socio-economic status	Low	38	17.8	2.5	1.134
	Middle	88	42.6		
	High	80	39.6		
	Total	212	100.0		

In this table, we can see lots of variables such as gender, age, education, work experience and socio-economic status with different attribute and frequency along with percentage.

- **Measures of Central tendency**

Measures of central tendency estimate a dataset's average or center, finding the result using three methods: mean, mode, and median.

1) Mean: The mean is the average value of a set of numbers. To find the mean, you add up all the numbers and then divide by the total number of values. For example, if we have the numbers 6, 8, 7, 10, 8, 4, and 9, the mean would be calculated by adding them up (52) and dividing by the total number of values (7), which gives you a mean of 7.3.
2) Mode: The mode is the most frequent response value. Datasets may have any number of modes, including "zero." We can find the mode by arranging our dataset's order from the lowest to highest value and then looking for the most common response. So, here from the last part: 4,6,7,8,8,9,10, the mode is eight.
3) Median: The value in the precise center of the dataset. Arranging the values in ascending order and we determine for the number in the set's middle. In this case, the median is eight.

- **Variability (Dispersion)**

  The measure of variability gives the statistician an idea of how spread out the responses are. The spread has three aspects — range, standard deviation, and variance.

  Range: Use range to determine how far apart the most extreme values are. Start by subtracting the dataset's lowest value from its highest value. 4,6,7,8,8,9,10. We subtract four (the lowest) from ten (the highest) and get six. There's your range.

  Standard Deviation: The standard deviation (s) is dataset's average amount of variability, showing y how far each score lies from the mean. The larger standard deviation, the greater dataset's variable. Follow these six steps:

1) List the scores and their means.

2) Find the deviation by subtracting the mean from each score.
3) Square each deviation.
4) Total up all the squared deviations.
5) Divide the sum of the squared deviations by N-1.
6) Find the result’s square root.

Variance: Variance is a measure of how spread out the data in a dataset is. It's calculated by taking the average of the squared differences between each value and the mean.

To calculate variance, we first need to calculate the mean (M) of the dataset. Next, for each value in the dataset, subtract the mean and then square the result. Add up all of the squared differences and divide the total by the number of values minus one (N-1).

This formula is represented as:

Variance = $\Sigma(x-M)^2 / (N-1)$

where Σ represents the sum of all the squared differences, x is a value in the dataset, M is the mean of the dataset, and N is the number of values in the dataset.

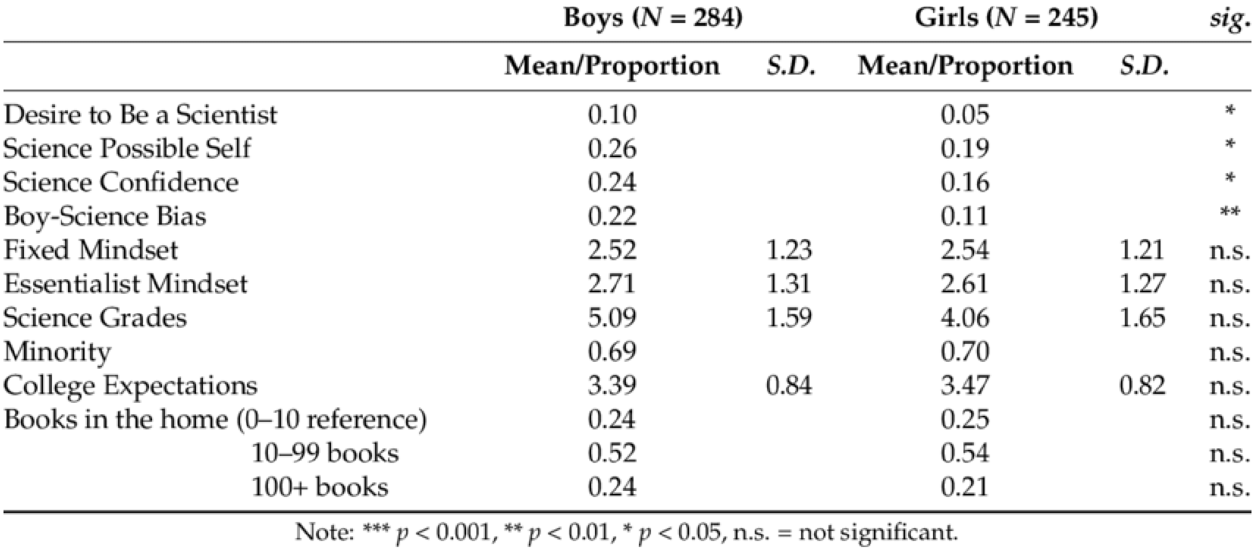
Alternatively, we can calculate the variance by squaring the standard deviation. The standard deviation (s) is the square root of the variance, so by squaring the standard deviation, you'll get the variance.

- **UNIVARIATE DESCRIPTIVE STATISTICS**

Univariate descriptive statistics examine only one variable at a time and do not compare variables. It allows the researcher to describe individual variables. The patterns identified in this sort of data may be explained using the following:

- Measures of central tendency (mean, mode, and median)
- Data dispersion (standard deviation, variance, range, minimum, maximum, and quartiles) (standard deviation, variance, range, minimum, maximum, and quartiles)
- Tables of frequency distribution
- Pie graphs
- Frequency polygon histograms
- Bar graphs

- **BIVARIATE STATISTICS**
  When using bivariate descriptive statistics, two variables are concurrently analyzed (compared) to see whether they are correlated. Generally, by convention, the independent variable is represented by the columns, and the rows represent the dependent variable.



	Boys (N = 284)		**Girls (N = 245)**		*sig.*
	Mean/Proportion	***S.D.***	**Mean/Proportion**	***S.D.***	
Desire to Be a Scientist	0.10		0.05		*
Science Possible Self	0.26		0.19		*
Science Confidence	0.24		0.16		*
Boy-Science Bias	0.22		0.11		**
Fixed Mindset	2.52	1.23	2.54	1.21	n.s.
Essentialist Mindset	2.71	1.31	2.61	1.27	n.s.
Science Grades	5.09	1.59	4.06	1.65	n.s.
Minority	0.69		0.70		n.s.
College Expectations	3.39	0.84	3.47	0.82	n.s.
Books in the home (0–10 reference)	0.24		0.25		n.s.
10–99 books	0.52		0.54		n.s.
100+ books	0.24		0.21		n.s.

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant.

- **USES OF DESCRIPTIVE STATISTICS**

Descriptive statistics serve two main purposes:

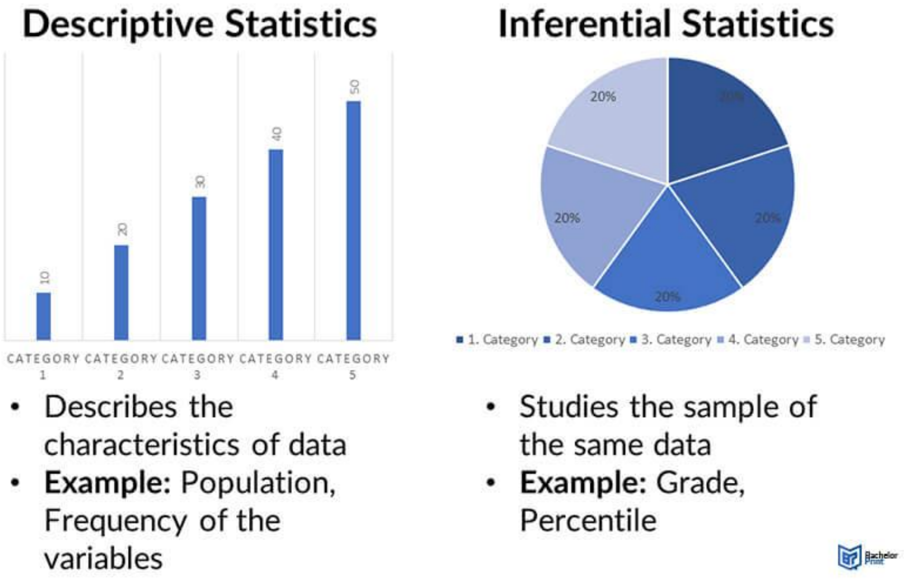
1) providing basic information about variables in a dataset,

2) highlighting potential relationships between variables.

Graphical and pictorial methods are common ways of summarizing data and displaying descriptive statistics. However, it's important to note that descriptive statistics only make statements about the specific dataset used to calculate them and cannot be used to draw conclusions beyond the data**.**

INFERENTIAL STATISTICS

- Inferential statistics are used to make conclusions and predictions about a larger population based on data collected from a sample.
- They serve two main purposes: making estimates about populations, such as the average score of all 11th graders, and testing hypotheses to draw conclusions about populations, such as the relationship between scores and family income.
- Inferential statistics use data from a sample to make guesses about a larger population. Random and unbiased sampling methods are crucial for valid statistical inferences and generalizations.
- Descriptive statistics can only summarize sample characteristics, while inferential statistics allow for reasonable guesses about the larger population.
- **Sampling error:**
  Sampling error is the difference between the true population values and the measured sample values, which arises whenever we use a sample instead of the whole population. Even if the sample is random and unbiased, there will always be some uncertainty in inferential statistics.

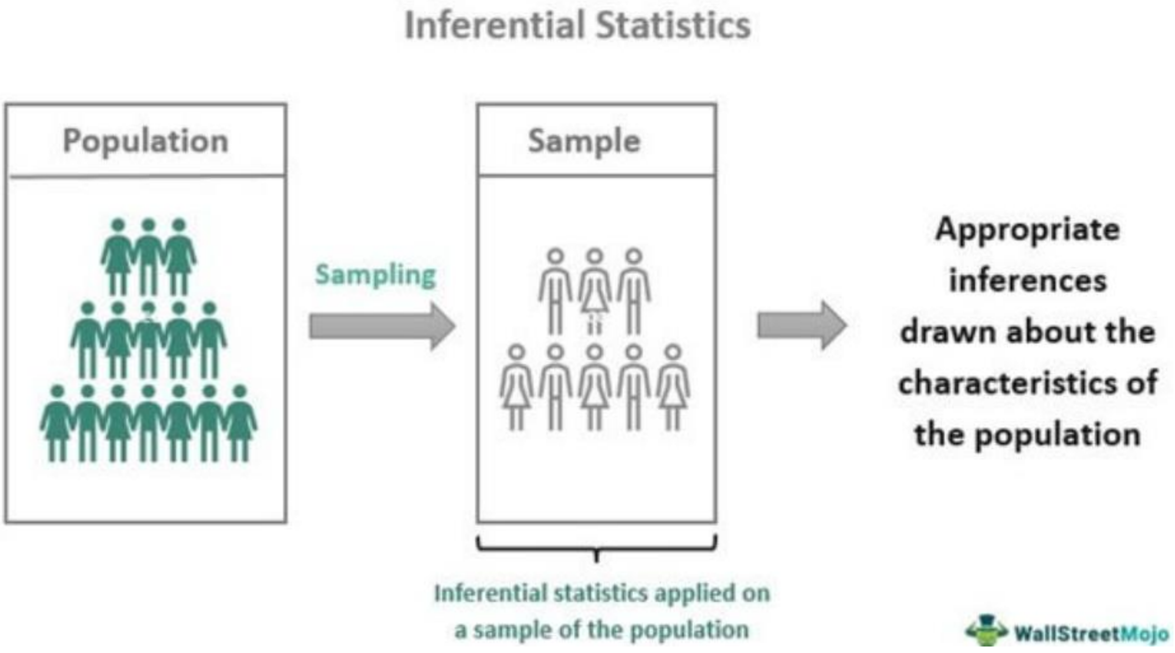


Descriptive Statistics

- Describes the characteristics of data
- **Example:** Population, Frequency of the variables

Inferential Statistics

- Studies the sample of the same data
- **Example:** Grade, Percentile



Inferential Statistics

Population

Sampling

Sample

Appropriate inferences drawn about the characteristics of the population

Inferential statistics applied on a sample of the population

- Regression analysis: This tool measures the relationship between two variables, and how a change in one variable affects the other.

- Hypothesis testing models: These models involve creating a null hypothesis (H0) and an alternate hypothesis (H1) to test the validity of an assumption or claim.

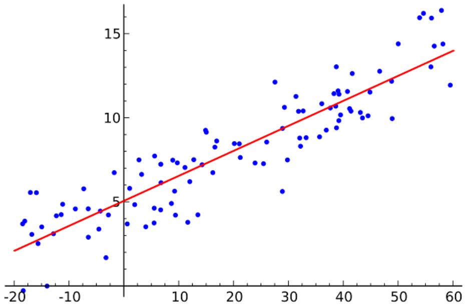
a) Z-test: This test is used when the sample size is 30 or more, and the data set follows a normal distribution. The population variance is known, and the test is used to determine whether a sample mean is significantly different from the population mean.

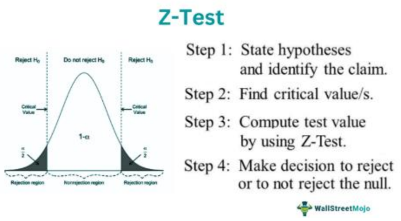
b) T-test: This test is used when the sample size is less than 30, and the population variance is unknown. The test is used to determine whether a sample mean is significantly different from the population mean.

c) F-test: This test is used to compare the variances of two populations or samples to determine if they are significantly different.

d) Confidence interval: This tool suggests the range within which an estimate will fall if the test is conducted on the population. A higher confidence interval means that the sample results are more likely to reflect the behavior of the population.

Inferential statistics can be very useful in making informed decisions, but it's important to use random and unbiased sampling methods to minimize sampling error. It's also crucial to carefully analyze and interpret the results of inferential statistics to draw valid conclusions about the population.





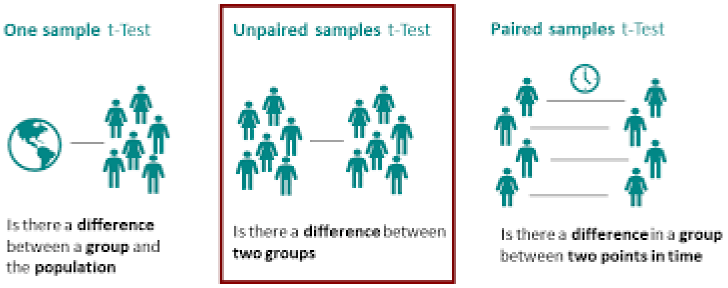
Z-Test

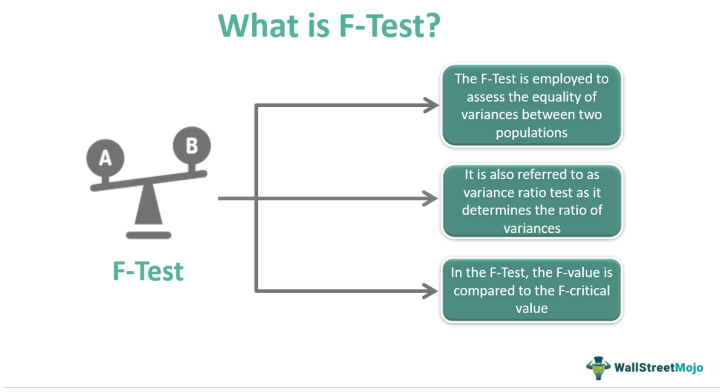
Step 1: State hypotheses and identify the claim.

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.


One sample t-Test
Is there a difference between a group and the population
Unpaired 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



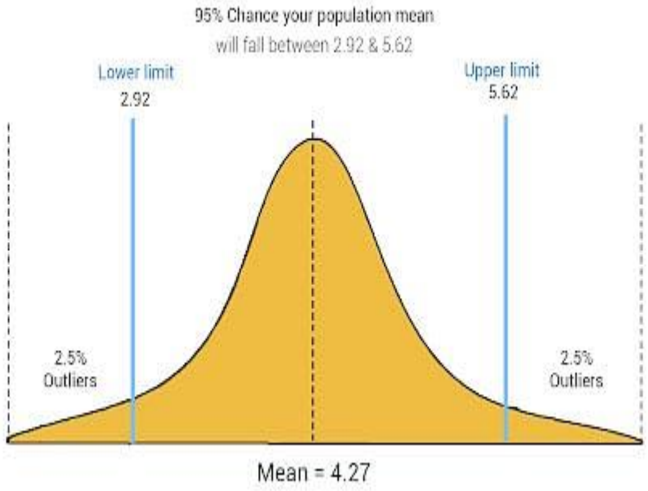
What is F-Test?

F-Test

The F-Test is employed to assess the equality of variances between two populations

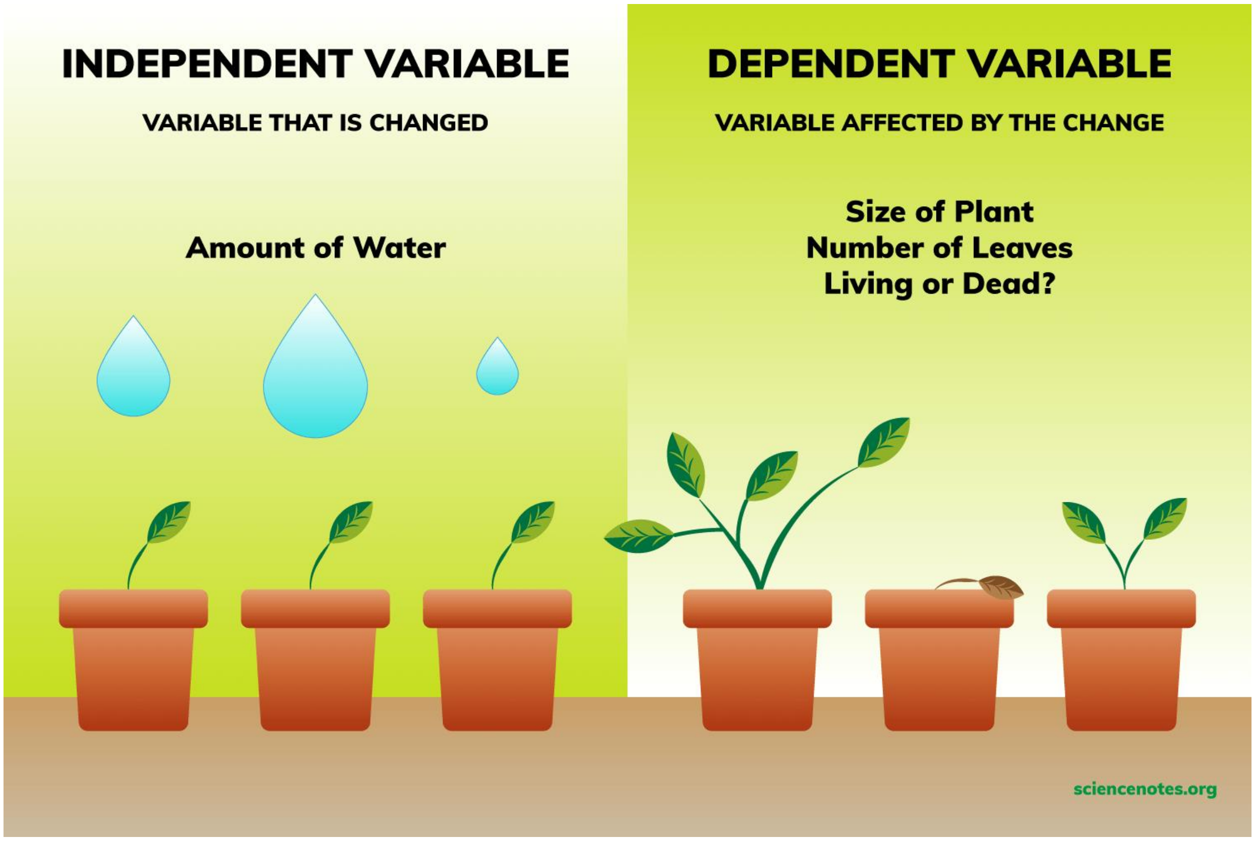
It is also referred to as variance ratio test as it determines the ratio of variances

In the F-Test, the F-value is compared to the F-critical value



Q) Does the independent variable affect the dependent variable?

The independent variable is the factor that is manipulated or changed by the researcher, while the dependent variable is the response or outcome that is measured. The goal of the study is to determine whether changes in the independent variable have an effect on the dependent variable.

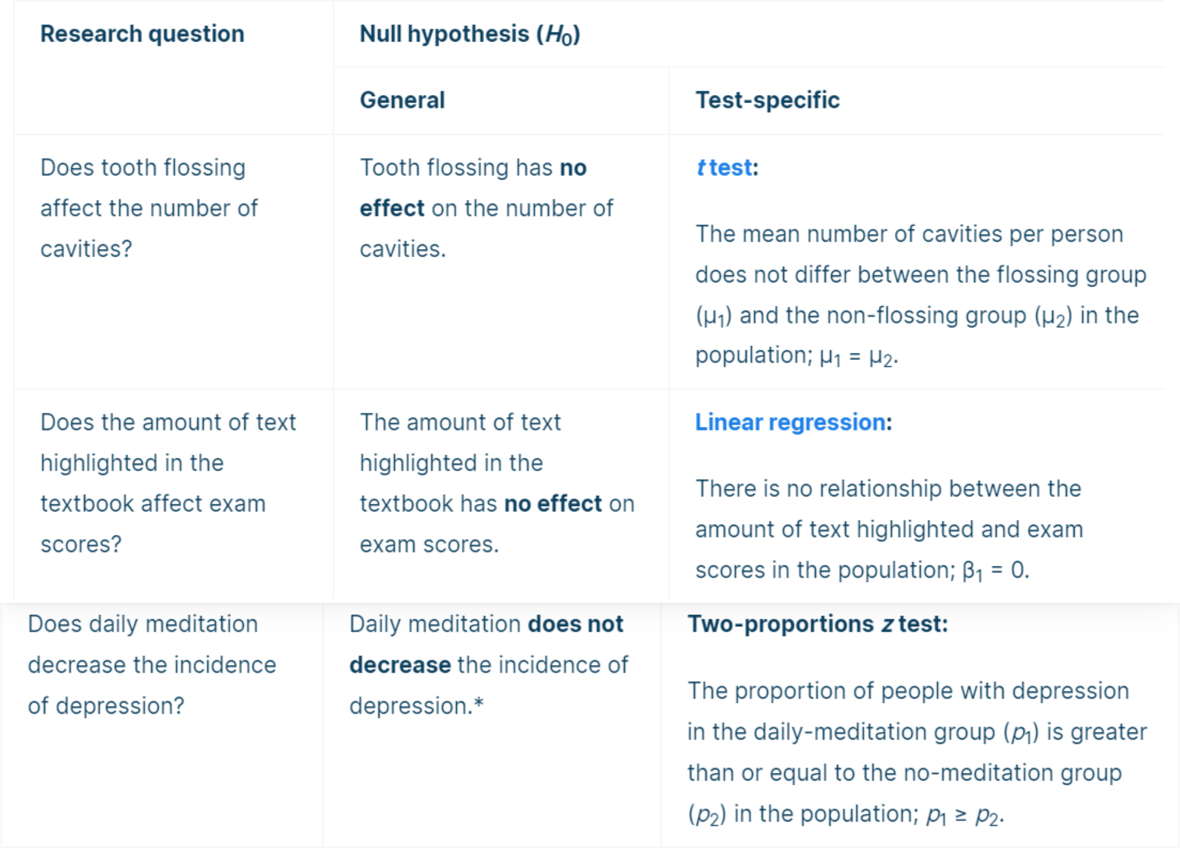


The null hypothesis is the claim that there's no effect in the population.

If the sample provides enough evidence against the claim that there's no effect in the population ($p \leq \alpha$), then we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Although "fail to reject" may sound awkward, it's the only wording that statisticians accept. Be careful not to say you "prove" or "accept" the null hypothesis.

The alternative hypothesis (H_a) claims that there's an effect in the population. Null and alternative hypotheses are exhaustive, meaning that together they cover every possible outcome. They are also mutually exclusive, meaning that only one can be true at a time.



Research question	Null hypothesis (H_0)	
	General	Test-specific
Does tooth flossing affect the number of cavities?	Tooth flossing has **no effect** on the number of cavities.	***t* test:** The mean number of cavities per person does not differ between the flossing group (μ_1) and the non-flossing group (μ_2) in the population; $\mu_1 = \mu_2$.
Does the amount of text highlighted in the textbook affect exam scores?	The amount of text highlighted in the textbook has **no effect** on exam scores.	**Linear regression:** There is no relationship between the amount of text highlighted and exam scores in the population; $\beta_1 = 0$.
Does daily meditation decrease the incidence of depression?	Daily meditation **does not decrease** the incidence of depression.*	**Two-proportions *z* test:** The proportion of people with depression in the daily-meditation group (p_1) is greater than or equal to the no-meditation group (p_2) in the population; $p_1 \geq p_2$.

- **HYPOTHESIS TESTING ON DATA SCIENCE LIFE CYCLE**

Hypothesis testing is an important step in the data science life cycle as it helps in making statistical inferences about the population based on the sample data. In simple terms, hypothesis testing is a process of making an assumption about the population and then testing that assumption with the sample data to see if it holds true.

The hypothesis testing process involves the following steps:

Formulating a hypothesis: The first step is to formulate a hypothesis, which is a statement about the population that we want to test. The hypothesis can be either a null hypothesis or an alternative hypothesis.

Selecting a statistical test: The next step is to select a statistical test based on the type of data and the nature of the hypothesis.

Setting the significance level: The significance level is the probability of rejecting the null hypothesis when it is actually true. It is usually set at 0.05 or 0.01.

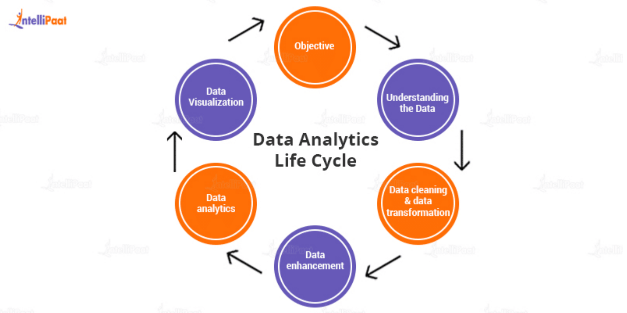
Collecting data: The next step is to collect data from the population or a sample of the population.

Calculating the test statistic: Using the collected data, we calculate the test statistic, which is a measure of how much the sample data deviates from the null hypothesis.

Making a decision: Based on the test statistic and the significance level, we make a decision to either reject or fail to reject the null hypothesis.

Drawing conclusions: Finally, we draw conclusions based on the decision made and interpret the results in the context of the problem being studied.

In summary, hypothesis testing is a crucial step in the data science life cycle as it helps in making statistical inferences about the population based on the sample data. It involves formulating a hypothesis, selecting a statistical test, setting the significance level, collecting data, calculating the test statistic, making a decision, and drawing conclusions.


IntelliPaat
Objective
Understanding the Data
Data cleaning & data transformation
Data enhancement
Data analytics
Data Visualization
Data Analytics Life Cycle