

UNIT-II
Statistical Inference

Dr. Vijay A. Kotkar
Assistant Professor

Department of Computer Engineering
Pimpri Chinchwad College of Engineering & Research,
Ravet, Pune

Need of Statistics in Data Science

- The most important aspect of any Data Science approach is how the information is processed.
- When we talk about developing insights out of data it is basically digging out the possibilities. Those possibilities in Data Science are known as **Statistical Analysis**.
- Most Data Scientists always invest more in pre-processing of data. This requires a good understanding of statistics.
- There are few general steps that always need to be performed to process any data:
 - 1) Identify the importance of **features** by using various statistical tests.
 - 2) Finding the **relationship between features** to eliminate the possibility of duplicate features.
 - 3) Converting the features into the **required format**.
 - 4) **Normalizing and scaling** the data. This step also involves the identification of the distribution of data and the nature of data.
 - 5) Taking the data for further processing by using **required adjustments** in the data.
 - 6) After processing the data identify the **right mathematical approach/model**.
 - 7) Once the results are obtained the results are verified on the different **accuracy measurement scales**.

Need of Statistics in Data Science

Key concepts to understand the fundamentals of statistics for Data Science:

- 1) Probability
- 2) Sampling
- 3) Tendency and Distribution of data
- 4) Hypotheses Testing
- 5) Variations
- 6) Regression

Need of Statistics in Data Science

Ways in which statistics helps in Data Science are:

- 1) **Prediction and Classification:** Statistics help in prediction and classification of data whether it would be right for the clients viewing by their previous usage of data.
- 2) **Helps to create Probability Distribution and Estimation:** Probability Distribution and Estimation are crucial in understanding the basics of machine learning and algorithms like logistic regressions.
- 3) **Cross-validation and LOOCV techniques** are also inherently statistical tools that have been brought into the Machine Learning and Data Analytics world for inference-based research, A/B and hypothesis testing.
- 4) **Pattern Detection and Grouping:** Statistics help in picking out the optimal data and weeding out the unnecessary dump of data for companies who like their work organized. It also helps spot out anomalies which further helps in processing the right data.

Need of Statistics in Data Science

- 5) **Powerful Insights:** Dashboards, charts, reports and other data visualizations types in the form of interactive and effective representations give much more powerful insights than plain data and it also makes the data more readable and interesting.
- 6) **Segmentation and Optimization:** It also segments the data according to different kinds of demographic or psychographic factors that affect its processing. It also optimizes data in accordance with minimizing risk and maximizing outputs.

Need of Statistics in Big Data

- Statistics is fundamental to ensuring meaningful, accurate information is extracted from Big Data.
- The following issues are crucial and are only exacerbated by Big Data:
 - 1) Data quality and missing data
 - 2) Observational nature of data, so that causal questions such as the comparison of interventions may be subject to confounding.
 - 3) Quantification of the uncertainty of predictions, forecasts and models
 - 4) The scientific discipline of statistics brings sophisticated techniques and models to bear on these issues
 - 5) Statisticians help translate the scientific question into a statistical question, which includes carefully describing data structure, the underlying system that generated the data (the model) and what we are trying to assess (the parameter or parameters we wish to estimate) or predict.

Measures of Central Tendency

- A measure of central tendency is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.
- There are three main measures of central tendency:
 - 1) Mode
 - 2) Median
 - 3) Mean
- Each of these measures describes a different indication of the typical or central value in the distribution.

Measures of Central Tendency - Mode

- The mode is the most commonly occurring value in a distribution.
- **Ex.** Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Measures of Central Tendency - Mode

Advantage of the mode:

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the mode:

There are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well. When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the centre of the distribution is 57 years, but the mode is lower, at 54 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

Measures of Central Tendency - Median

- The median is the middle value in distribution when the values are arranged in ascending or descending order.
- The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.
- **Ex.** Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years:

52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Measures of Central Tendency - Median

Advantage of the median:

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median:

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Measures of Central Tendency - Mean

- The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.
- **Ex.** Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values ($54+54+54+55+56+57+57+58+58+60+60 = 623$) and dividing by the number of observations (11) which equals 56.6 years.

Measures of Central Tendency - Mean

Advantage of the mean:

The mean can be used for both continuous and discrete numeric data.

Limitations of the mean:

The mean cannot be calculated for categorical data, as the values cannot be summed.

As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

What else do I need to know about the mean?

The population mean is indicated by the Greek symbol μ (pronounced 'mu'). When the mean is calculated on a distribution from a sample it is indicated by the symbol \bar{x} (pronounced X-bar).

Measures of Central Tendency – Mid-range

- Mid-range in layman terms is the middle of any data set or the simply the average, mean of the data.
- A midrange is a statistical tool which is also known as the measure of center in statistics. Along with the existence of the midrange formula means, medium, average, mode, and range are also known as the measure of central tendency.
- The mid-range of the data set is simply the value between the biggest value and the lowest value.
- In order to find the midrange of the data set the value is then divided by 2 after summing the lowest value present in the data set with the highest value present in the data set.
- Ex. The daily temperature recorded in the city of Colombia is 55, 65, 67, 69, 70, 80, 81, 87, 90. We need to calculate the mid-temperature in Colombia during this period.

$$\text{Midrange} = (90 + 55) / 2$$

$$\text{Midrange} = 145 / 2$$

$$\text{Midrange} = \mathbf{72.5}$$

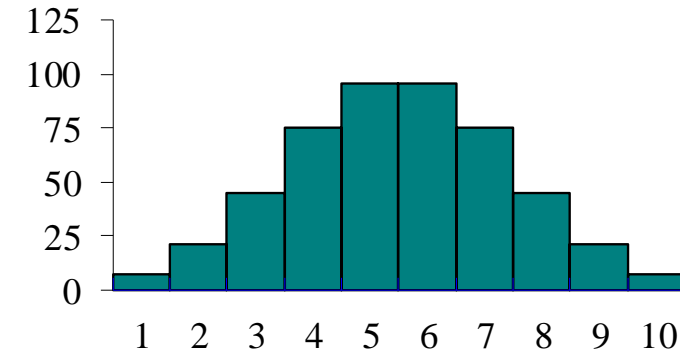
Measures of Dispersion

Measures of dispersion are descriptive statistics that describe how similar a set of scores are to each other.

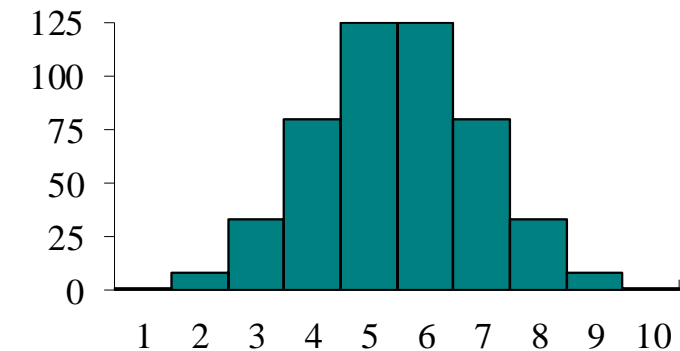
- The more similar the scores are to each other, the lower the measure of dispersion will be.
- The less similar the scores are to each other, the higher the measure of dispersion will be.
- In general, the more spread out a distribution is, the larger the measure of dispersion will be.

Measures of Dispersion

- Which of the distributions of scores has the larger dispersion?



- ✚ The upper distribution has more dispersion because the scores are more spread out
 - ✚ That is, they are less similar to each other



Measures of Dispersion

- These are the measures of dispersion:
 - 1) The range
 - 2) Variance
 - 3) Mean Deviation
 - 4) Standard Deviation

Measures of Dispersion - Range

- The range is defined as the difference between the largest score in the set of data and the smallest score in the set of data, $X_L - X_S$.
- What is the range of the following data:
4 8 1 6 6 2 9 3 6 9
- The largest score (X_L) is 9; the smallest score (X_S) is 1; the range is $X_L - X_S = 9 - 1 = 8$

When to use the Range

- The range is used when
 - you have ordinal data or
 - you are presenting your results to people with little or no knowledge of statistics
- The range is rarely used in scientific work as it is fairly insensitive
 - It depends on only two scores in the set of data, X_L and X_S
 - Two very different sets of data can have the same range:
1 1 1 1 9 vs 1 3 5 7 9

Measures of Dispersion - Variance

- *Variance* is defined as the average of the square deviations:

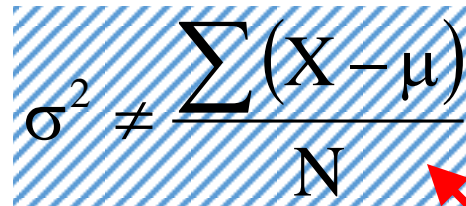
$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

What does the Variance formula mean?

- First, it says to subtract the mean from each of the scores
 - This difference is called a *deviate* or a *deviation score*
 - The deviate tells us how far a given score is from the typical, or average, score
 - Thus, the deviate is a measure of dispersion for a given score

What does the Variance formula mean?

- Why can't we simply take the average of the deviates? That is, why is not variance defined as:


$$\sigma^2 \neq \frac{\sum (X - \mu)}{N}$$



This is not the formula for variance!

What does the Variance formula mean?

- One of the definitions of the *mean* was that it always made the sum of the scores minus the mean equal to 0
- Thus, the average of the deviates must be 0 since the sum of the deviates must equal 0
- To avoid this problem, statisticians square the deviate score prior to averaging them
 - Squaring the deviate score makes all the squared scores positive

What does the Variance formula mean?

- Variance is the mean of the squared deviation scores.
- The larger the variance is, the more the scores deviate, on average, away from the mean.
- The smaller the variance is, the less the scores deviate, on average, from the mean.

Variance of a Sample

- Because the sample mean is not a perfect estimate of the population mean, the formula for the **variance of a sample is slightly different** from the formula for the variance of a population:

$$S^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

✚ s^2 is the sample variance, X is a score, \bar{X} is the sample mean, and N is the number of scores

Measures of Dispersion – Mean Deviation

- The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set.
- The mean deviation of the data values can be easily calculated using the below procedure:

Step 1: Find the mean value for the given data values

Step 2: Now, subtract the mean value from each of the data values given (Note: Ignore the minus symbol)

Step 3: Now, find the mean of those values obtained in step 2.

Measures of Dispersion – Mean Deviation

The formula to calculate the mean deviation for the given data set is given below.

$$\text{Mean Deviation} = [\Sigma |X - \mu|]/N$$

Here,

Σ represents the addition of values

X represents each value in the data set

μ represents the mean of the data set

N represents the number of data values

$| |$ represents the absolute value, which ignores the “-” symbol

Measures of Dispersion – Mean Deviation

Example 1:

Determine the mean deviation for the data values 5, 3, 7, 8, 4, 9.

Solution:

Given data values are 5, 3, 7, 8, 4, 9.

We know that the procedure to calculate the mean deviation.

First, find the mean for the given data:

$$\text{Mean, } \mu = (5+3+7+8+4+9)/6$$

$$\mu = 36/6$$

$$\mu = 6$$

Therefore, the mean value is 6.

Measures of Dispersion – Mean Deviation

Now, subtract each mean from the data value, and ignore the minus symbol if any

(Ignore“-”)

$$5 - 6 = 1$$

$$3 - 6 = 3$$

$$7 - 6 = 1$$

$$8 - 6 = 2$$

$$4 - 6 = 2$$

$$9 - 6 = 3$$

Measures of Dispersion – Mean Deviation

Now, the obtained data set is 1, 3, 1, 2, 2, 3.

Finally, find the mean value for the obtained data set

Therefore, the mean deviation is

$$= (1+3 + 1+ 2+ 2+3) /6$$

$$= 12/6$$

$$= 2$$

Hence, the mean deviation for 5, 3, 7, 8, 4, 9 is 2.

Measures of Dispersion – Standard Deviation

- When the deviate scores are squared in variance, their unit of measure is squared as well
 - E.g. If people's weights are measured in pounds, then the variance of the weights would be expressed in pounds² (or squared pounds)
- Since squared units of measure are often awkward to deal with, the square root of variance is often used instead
 - The standard deviation is the square root of variance

Measures of Dispersion – Standard Deviation

- Standard deviation = $\sqrt{\text{variance}}$
- Variance = $\text{standard deviation}^2$

Bayes' Theorem

- Probability is at the very core of data science algorithms.
- In fact, the solutions to so many data science problems are probabilistic in nature.
- Created by Thomas Bayes – worked in **decision theory** (the field of mathematics that involves probabilities).
- Bayes' theorem is used in Machine Learning to predict the classes.

Prerequisites for Bayes' Theorem

- 1) **Experiment** – planned operation carried out under controlled conditions
- 2) **Sample space** – set of all possible outcomes of an event
- 3) **Event** – set of outcomes of an experiment
- 4) **Exhaustive events** – at least one of the events must occur at any time
- 5) **Independent events** – occurrence of one event does not effect on the occurrence of another event
- 6) **Conditional probability** – probability of an event A, given that another event B has already occurred (i.e. A conditional B)
- 7) **Marginal probability** – probability of an event A occurring, independent of any other event B, i.e. marginalizing the event B.

What is Bayes' Theorem

- In any crime thriller TV show – our beliefs about the culprit change throughout the episode. We process new evidence and refine our **hypothesis at each step**, this is Bayes' theorem in real life.
- Consider that A and B are any two events from a sample space S where $P(B) \neq 0$. Using conditional probability:

$$P(A|B) = P(A \cap B) / P(B)$$

$$\text{Similarly, } P(B|A) = P(A \cap B) / P(A)$$

$$\text{It follows that } P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$$

$$\text{Thus, } P(A|B) = P(B|A) * P(A) / P(B)$$

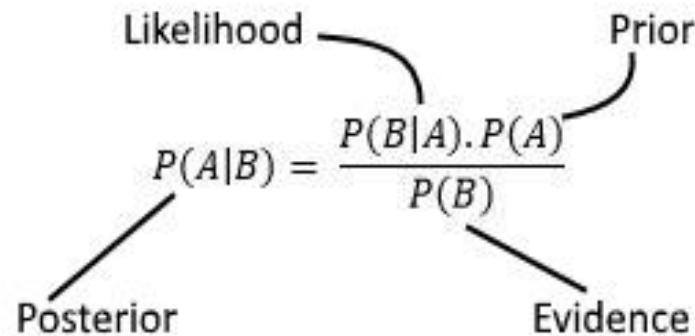
- Here, $P(A)$ and $P(B)$ are probabilities of observing A and B independently of each other. That's why we can say that they are marginal probabilities. $P(B|A)$ and $P(A|B)$ are conditional probabilities.

What is Bayes' Theorem?

- $P(A)$ is called **Prior probability** and $P(B)$ is called **Evidence**.

$$P(B) = P(B|A)*P(A)+P(B|\sim A)*P(\sim A)$$

- $P(B|A)$ is called **Likelihood** and $P(A|B)$ is called **Posterior probability**.



The diagram shows the formula for the posterior probability: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$. Labels with arrows point to the components: 'Likelihood' points to $P(B|A)$, 'Prior' points to $P(A)$, 'Posterior' points to $P(A|B)$, and 'Evidence' points to $P(B)$.

- Equivalently, Bayes Theorem can be written as:

$$\text{posterior} = \text{likelihood} * \text{prior} / \text{evidence}$$

Example of Bayes' Theorem

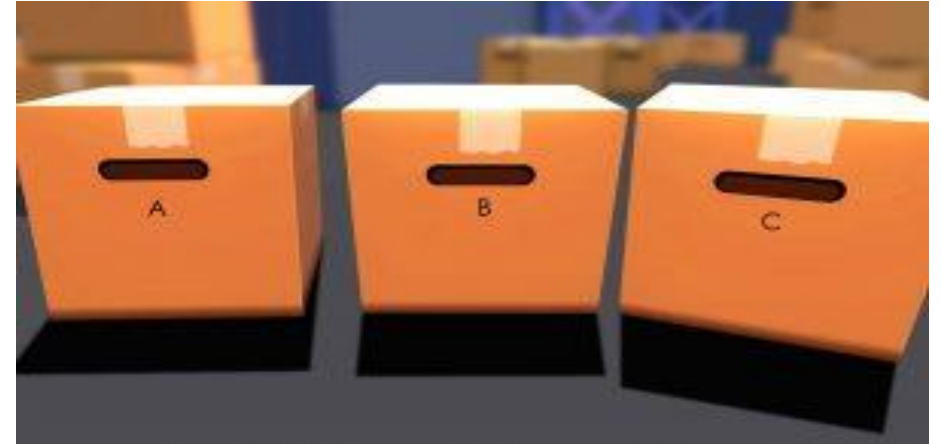
Example:

There are 3 boxes labeled A, B, and C:

Box A contains 2 red and 3 black balls

Box B contains 3 red and 1 black ball

And box C contains 1 red ball and 4 black balls.



The three boxes are identical and have an equal probability of getting picked. Consider that a red ball is chosen. Then what is the probability that this red ball was picked out of box A?

Solution:

Let E denote the event that a red ball is chosen and A , B , and C denote that the respective box is picked. We are required to calculate the conditional probability $P(A|E)$.

Example of Bayes' Theorem

Probability of getting picked.

$P(E|A)$ = Number of red balls in box A / Total number of balls in box A = $2 / 5$

Similarly, $P(E|B) = 3 / 4$ and $P(E|C) = 1 / 5$

$$\begin{aligned}\text{Then evidence } P(E) &= P(E|A)*P(A) + P(E|B)*P(B) + P(E|C)*P(C) \\ &= (2/5) * (1/3) + (3/4) * (1/3) + (1/5) * (1/3) \\ &= 0.45\end{aligned}$$

$$\begin{aligned}\text{Therefore, } P(A|E) &= P(E|A) * P(A) / P(E) \\ &= (2/5) * (1/3) / 0.45 \\ &= \mathbf{0.296}\end{aligned}$$

Applications of Bayes' Theorem

Applications of Bayes' theorem:

- 1) **Naive Bayes' Classifiers** - the features used for classification are independent of each other.
- 2) **Discriminant Functions and Decision Surfaces** - used to “discriminate” its argument into its relevant class.
- 3) **Bayesian Parameter Estimation** - a random variable as opposed to an “unknown but fixed” value

Hypothesis Testing

- When using Machine Learning, we need to be able to trust our models and the predictions they make. We may use sample data to train our models. This sample data may make certain assumptions about a population.
- Assumptions regarding a population parameter:

1) Statistical Hypotheses

- These tests are concerned with how likely the effect is present or absent in the population in consideration.
- Based on identifying the relationships between observations.
- Null and Alternative hypotheses are denoted by H_0 and H_a respectively.

2) Machine Learning Hypotheses

- Approximating target functions and performing mappings of input and output.
- Approximate an unknown target function, which we assume exists.

Steps to test Hypothesis

- A hypothesis test evaluates two statements about a population. These statements are mutually exclusive.
- A test concludes which statement best reflects the sample data.
- A hypothesis test helps us to determine the statistical significance of a finding.

Step 1: Establish Hypotheses

- Establish both null and alternative hypothesis.
- **Null** – assumption made, which may be based on domain experience
- **Alternative** – alternate to null hypothesis

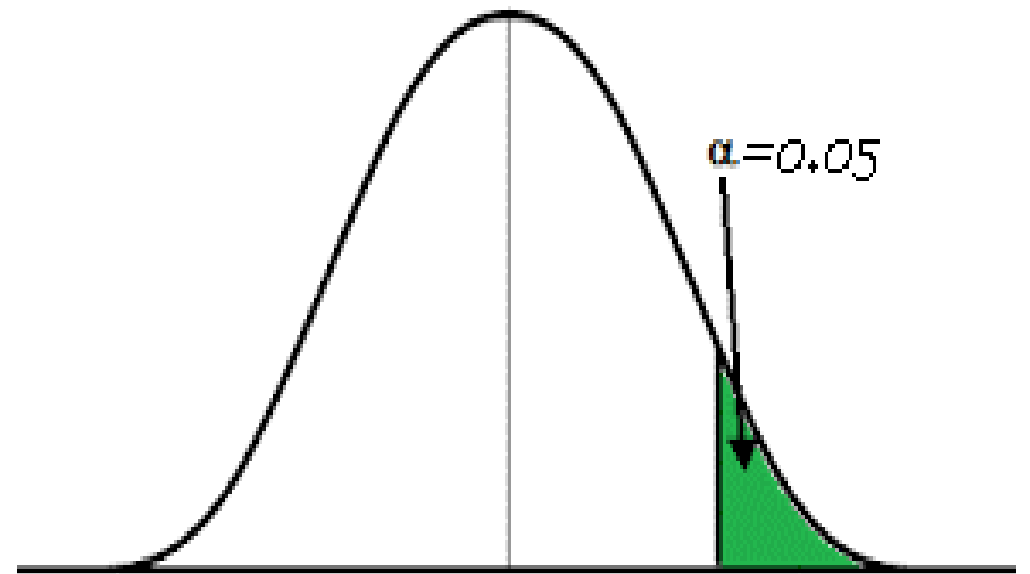
Step 2: Set a significance level

- This is the measure of the influence of the evidence that needs to be available in a sample before rejecting the null hypothesis.

Steps to test Hypothesis

- The significance level is usually 5%. It means that it is probable that the test may suffer a type I error.
- Since the significance level is 5%, our level of confidence becomes 95%. This means that 95% of hypothesis tests won't end in a type I error. You may ask why 5% and not any other value is commonly chosen. It simply is standard practice to use 5%.

Alpha = 5%, denoting significance level



Steps to test Hypothesis

- **Type I error.** This is an error characterized by a scenario where we reject a **true** null hypothesis. The symbol alpha (α) represents it.
- **Type II error.** We can define a type II error in a situation, where we retain a null hypothesis, but it is **false**. It is denoted by beta (β).

Steps to test Hypothesis

Step 3: Find the region of rejection for the null hypothesis



- A region in the sample space where we reject the null hypothesis. The rejection is if a calculated value lies in the region. This region is known as the critical region.

Step 4: Compute p-value

- Assuming the null hypothesis is true, the probability of getting an outcome at least as extreme as the observed outcome of a hypothesis test is what we call the p-value.
- The p-value determines whether there is enough evidence to retain the alternative hypothesis or retain the null hypothesis.
- Given the probability distribution of a specific statistic we are testing, we may use the deviation between an observed value and a selected reference value.

Steps to test Hypothesis

P Value Formula


$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$


Where

- \hat{p} is Sample Proportion
- p_0 is Assumed Population Proportion in the Null Hypothesis
- n is the Sample Size

Steps to test Hypothesis

Step 5: Compare p-value to the significance level to retain or reject the null hypothesis

- Significance level is 5% (or 0.05). The smaller the p-value, the greater the evidence is favoring the alternative hypothesis.
- If the p-value is less than the significance level we selected, then we reject the null hypothesis. This means that if the p-value is less than our 0.05 significance level, we accept that the sample we used supports the alternative hypothesis.

Example:

- a) P-value is 0.3015. If the level of significance is 5%, find if we can reject the null hypothesis.
- b) P-value is 0.0129. If the level of significance is 5%, find if we can reject the null hypothesis.

Steps to test Hypothesis

	A	B	C	
1	P-Value	Level of Significance	Conclusion	
2	0.3015	0.05	We Fail to Reject Null Hypothesis	
3	0.0129	0.05	We Reject Null Hypothesis	
4				

Example of Hypothesis Testing

Hypothesis No.	Hypothesis	Null Hypothesis (H_0)	Alternate hypothesis (H_A)	Accepted Hypothesis after research
Hypothesis 1	Classification of Anomaly	Class ≤ 0	Class ≥ 2	H_A
Hypothesis 2	Statistical Measures	Score _{Acc, P, R, S, F1} ≤ 90	Score _{Acc, P, R, S, F1} > 90	H_A
Hypothesis 3	Result	$R_{SVM, ANN, KNN} \geq R_{RNN_LSTM}$	$R_{RNN_LSTM} > R_{SVM, ANN, KNN}$	H_A

Pearson Correlation

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of **measuring the association between variables of interest** because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson's correlation is used **when you want to see if there is a linear relationship between two quantitative variables.**

Where,

r = Pearson Correlation Coefficient

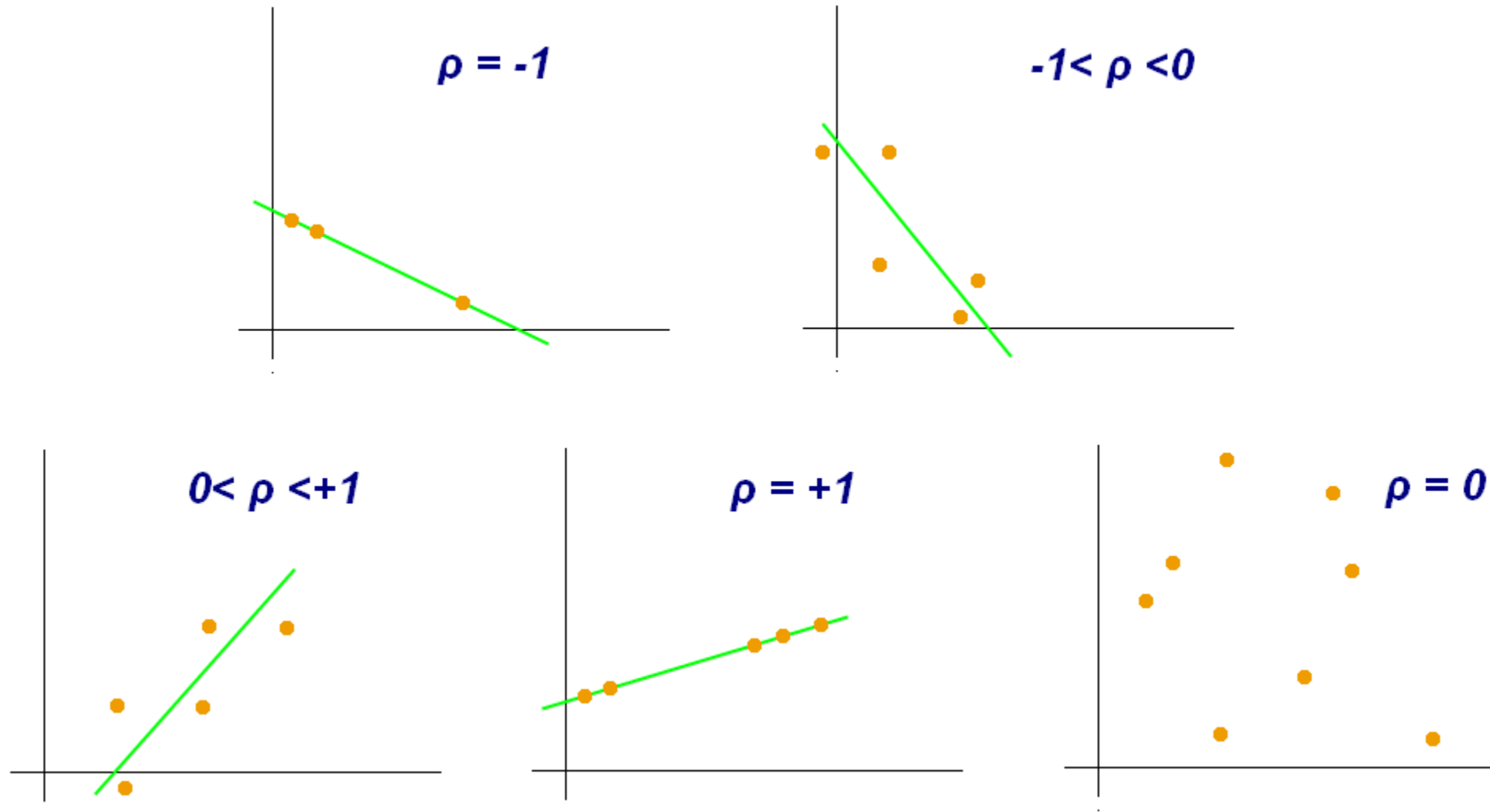
x_i = x variable samples

y_i = y variable sample

\bar{x} = mean of values in x variable

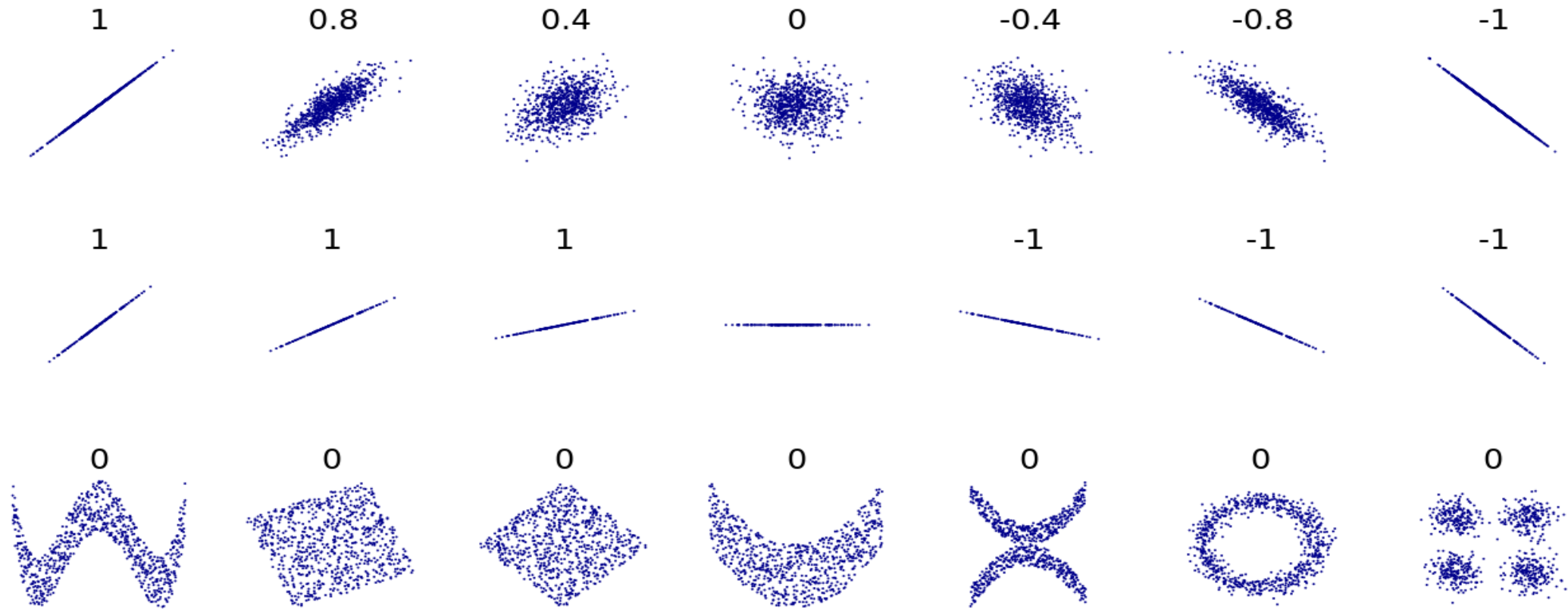
\bar{y} = mean of values in y variable

Pearson Correlation



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Pearson Correlation



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the strength and direction of a linear relationship (**top row**), but not the slope of that relationship (**middle**), nor many aspects of nonlinear relationships (**bottom**).

Pearson Correlation

Correlation Coefficient Value (r)	Direction and Strength of Correlation
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

Pearson Correlation

Assumptions:

- **Independent of case:** Cases should be independent to each other.
- **Linear relationship:** Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
- **Homoscedasticity:** the residuals scatterplot should be roughly rectangular-shaped.

Pearson Correlation

Properties:

- 1) **Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.
- 2) **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
- 3) **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Pearson Correlation

Degree of correlation:

- 1) **Perfect:** If the value is near ± 1 , then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- 2) **High degree:** If the coefficient value lies between ± 0.50 and ± 1 , then it is said to be a strong correlation.
- 3) **Moderate degree:** If the value lies between ± 0.30 and ± 0.49 , then it is said to be a medium correlation.
- 4) **Low degree:** When the value lies below $+ .29$, then it is said to be a small correlation.
- 5) **No correlation:** When the value is zero.

Chi-Square Test

- Statistical test follows a specific distribution known as chi square distribution.
- In general The test we use to **measure the differences between what is observed and what is expected according to an assumed** hypothesis is called the chi-square test.
- There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:
 - 1) A **chi-square goodness of fit test** determines if sample data matches a population. For more details on this type, see: Goodness of Fit Test.
 - 2) A **chi-square test for independence compares** two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
- The formula for the chi-square statistic used in the chi square test is:
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Test Example

Table of Observed Values

Qualification/ Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D.	Total
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 hypothesis: There is no relation between the marital status and educational qualification.

Alternate hypothesis: There is significant relation between the marital status and educational qualification.

Chi-Square Test Example

Table of Expected Values

Qualification/ Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D.
Never Married	$90 \cdot 39 / 300 = 11.7$	$90 \cdot 90 / 300 = 27$	$90 \cdot 84 / 300 = 25.2$	$90 \cdot 54 / 300 = 16.2$	$90 \cdot 33 / 300 = 9.9$
Married	$150 \cdot 39 / 300 = 19.5$	$150 \cdot 90 / 300 = 45$	$150 \cdot 84 / 300 = 42$	$150 \cdot 54 / 300 = 27$	$150 \cdot 33 / 300 = 16.5$
Divorced	$30 \cdot 39 / 300 = 3.9$	$30 \cdot 90 / 300 = 9$	$30 \cdot 84 / 300 = 8.4$	$30 \cdot 54 / 300 = 5.4$	$30 \cdot 33 / 300 = 3.3$
Widowed	$30 \cdot 39 / 300 = 3.9$	$30 \cdot 90 / 300 = 9$	$30 \cdot 84 / 300 = 8.4$	$30 \cdot 54 / 300 = 5.4$	$30 \cdot 33 / 300 = 3.3$

$$\frac{(O - E)^2}{E}$$

Chi-Square Test Example

Observed Values (O)	Expected Values (E)	(O – E)	(O – E) ²	$\frac{(O - E)^2}{E}$
18	11.7	6.3	39.69	3.69
36	27	9	81	3
21	25.2	-4.2	17.64	0.7
9	16.2	-7.2	51.84	3.2
6	9.9	-3.9	15.21	1.53
12	19.5	-7.5	56.25	2.88
36	45	-9	81	1.8
45	42	3	9	0.21
36	27	9	81	3
21	16.5	4.5	20.25	1.22

Observed Values (O)	Expected Values (E)	(O – E)	(O – E) ²	$\frac{(O - E)^2}{E}$
6	3.9	2.1	4.41	1.13
9	9	0	0	0
9	8.4	0.6	0.36	0.04
3	5.4	-2.4	5.76	1.06
3	3.3	-0.3	0.09	0.02
3	3.9	2.1	4.41	1.13
9	9	0	0	0
9	8.4	0.6	0.36	0.04
6	5.4	-2.4	5.76	1.06
3	3.3	-0.3	0.09	0.02
				$\sum (O - E)^2 / E$ $\chi^2 = 23.57$ (Calculated)

Chi-Square Test Example

For Tabular value of Chi-Square:

Degrees of Freedom = (columns - 1) (rows - 1)

$$= (5 - 1)(4 - 1)$$

$$= 4 * 3$$

$$= \mathbf{12}$$

- Refer Percentage point of the Chi-Square Distribution chart

$$\chi^2 \text{ tabular} = 21.03$$

Chi-Square Test Example

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of χ^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Chi-Square Test Example

$$\chi^2 \text{ tabular} = 21.03$$

$$\chi^2 \text{ calculated} = 23.57$$

$$\chi^2 \text{ calculated} > \chi^2 \text{ tabular}$$

We reject Null hypothesis and accept alternate hypothesis.

T-Test

- A t-test is a type of inferential [statistic](#) used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
- The t-test is one of many tests used for the purpose of [hypothesis testing](#) in statistics.
- Calculating a t-test requires three key data values. They include the difference between the **mean values** from each data set (called the mean difference), the **standard deviation** of each group, and the **number of data values** of each group.
- There are several different types of t-test that can be performed depending on the data and type of analysis required.

T-Test

t-Test Formula



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



$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

\bar{x} = Observed Mean of the Sample

μ = Theoretical Mean of the Population

s = Standard Deviation of the Sample

n = Sample Size

Where,

\bar{x}_1 = Observed Mean of 1st Sample

\bar{x}_2 = Observed Mean of 2nd Sample

s_1 = Standard Deviation of 1st Sample

s_2 = Standard Deviation of 2nd Sample

n_1 = Size of 1st Sample

n_2 = Size of 2nd Sample

T-Test

The formula for **one-sample t-test** can be derived by using the following steps:

Step 1: Firstly, determine the observed sample mean, and the theoretical population means specified. The sample mean and population mean is denoted by \bar{x} and μ , respectively.

Step 2: Next, determine the standard deviation of the sample, and it is denoted by s .

Step 3: Next, determine the sample size, which is the number of data points in the sample. It is denoted by n .

Step 4: Finally, the formula for a one-sample t-test can be derived using the observed sample mean (step 1), the theoretical population means (step 1), sample standard deviation (step 2) and sample size (step 3), as shown below.

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

T-Test

The formula for the **two-sample t-test** can be derived by using the following steps:

Step 1: Firstly, determine the observed sample mean of the two samples under consideration. The sample means are denoted by \bar{x}_1 and \bar{x}_2 .

Step 2: Next, determine the standard deviation of the two samples, which are denoted by s_{21} and s_{22} .

Step 3: Next, determine the size of the two samples, which are denoted by n_1 and n_2 .

Step 4: Finally, the formula for a two-sample t-test can be derived using observed sample means (step 1), sample standard deviations (step 2) and sample sizes (step 3) as shown below.

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt { (s_{21} / n_1) + (s_{22} / n_2) }$$

T-Test

