Applied Data Science with Python

Advanced Statistics

Learning Objectives

By the end of this lesson, you will be able to:

- Explain advanced statistical concepts like hypothesis testing
- Describe null hypothesis and alternate hypothesis
- Measure the p-value of a hypothesis
- Examine different hypothesis tests like Z-test and T-test

Learning Objectives

By the end of this lesson, you will be able to:

- Examine Bayes Theorem and interpret confidence levels and intervals
- Create chi-square distribution and F distribution using Python
- Describe the concept of ANOVA
- Determine when and which type of statistical distribution to use

Business Scenario

A manufacturer of electronic devices aims to enhance its product quality and customer satisfaction. It conducts a survey to collect customer feedback but lacks expertise in analyzing the results. Seeking assistance, the company hires a data analyst who suggests using hypothesis testing to draw statistical inferences from the survey data.

The analyst explains that hypothesis testing involves comparing survey results to a null hypothesis and using confidence levels to support or reject it.

With the analyst's guidance, the company effectively analyzes the survey data, identifies areas for improvement, and achieves increased customer satisfaction and product quality.

Distribution

Discussion: Distribution

Duration: 10 minutes

- What are distribution and normal distribution?
- What is a hypothesis, and what are its components?

Distribution

Statistical distribution or probability distribution function (PDF) is a fundamental concept in probability theory and statistics.

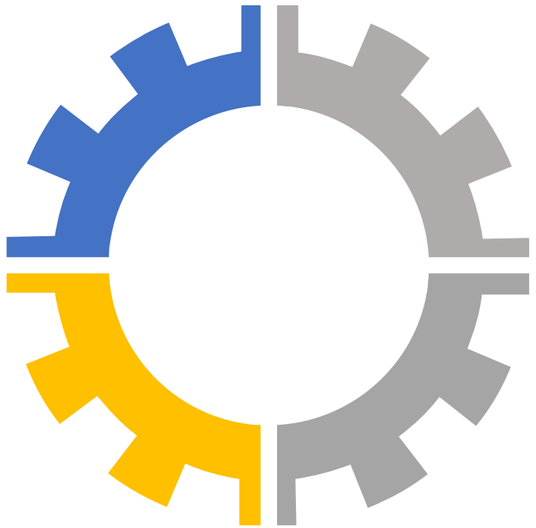
PDF describes all possible values that a random variable can assume in a given range.

Distribution

The possible values of this distribution function are determined by factors such as:

Kurtosis

Mean



Skewness

Standard deviation

Distribution

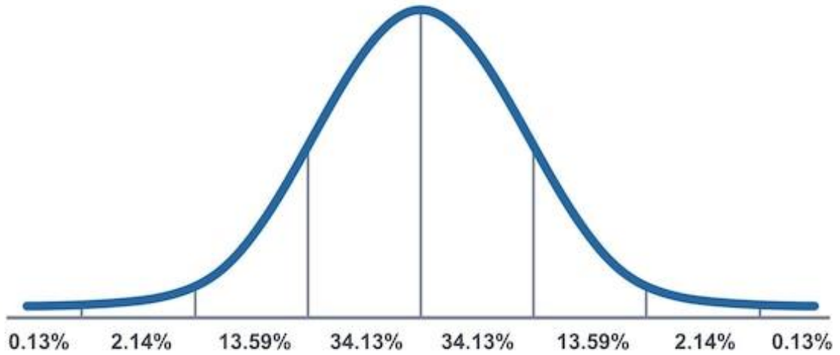
A **probability distribution** is a mathematical function that describes the probability of different variable values.

The most common probability distributions:

- Normal distribution
- Binomial distribution
- Chi-square distribution
- Poisson distribution

Normal Dīstribution

Normal distribution is the most commonly occurring distribution and is also known as Gaussian distribution.



It is used in almost every field like finance, science, and engineering.

Student’s T-Distribution: Skewness and Kurtosis

Skewness and Kurtosis

Real-life data is never distributed normally, even though such a distribution is called a normal distribution.

Two factors that impact the bell curve shape are:

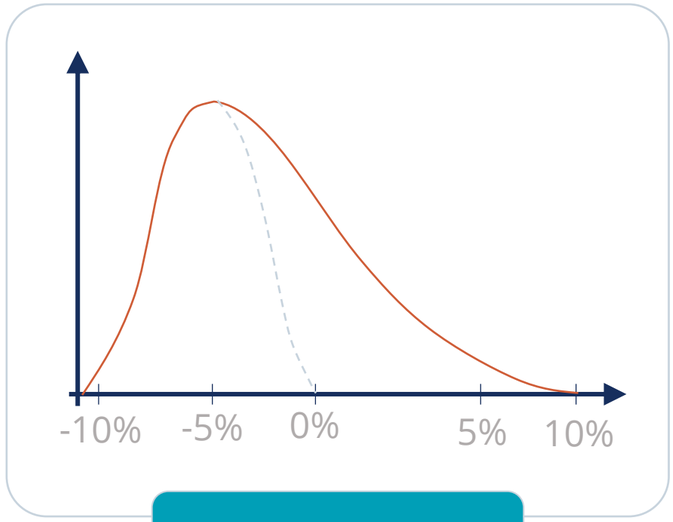
Skewness

Kurtosis

A shifted bell curve is called **skewed**.

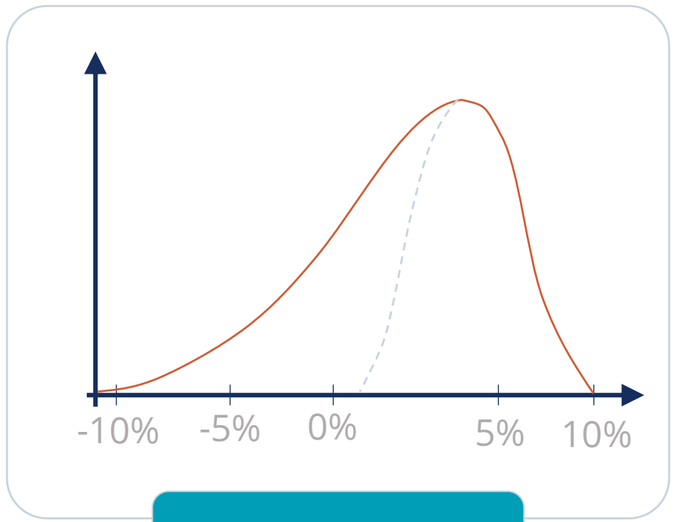
Skewness

Skewness is a numerical measure that is either negative or positive.



Positive skew

Positive skewness shifts the curve left.



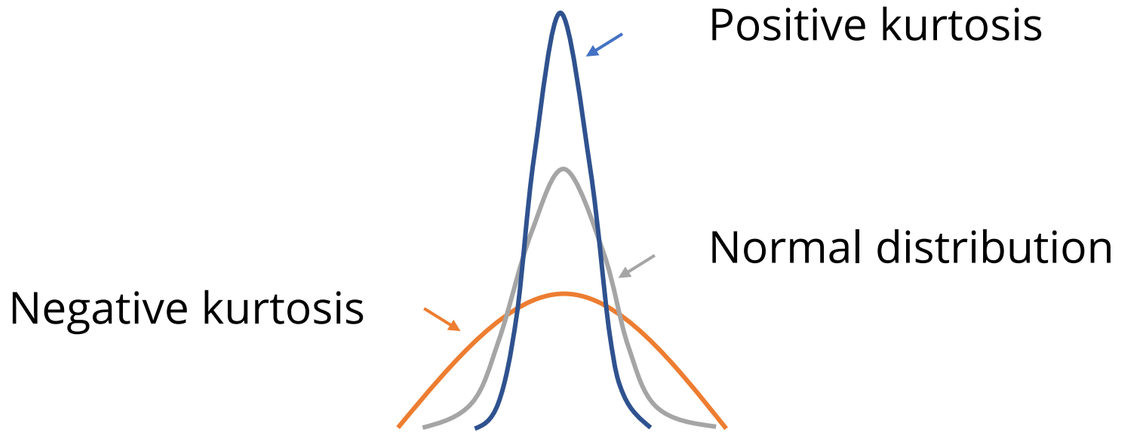
Negative skew

Negative skewness shifts the curve right.

Kurtosis

Kurtosis is a statistical measure used to describe a distribution.

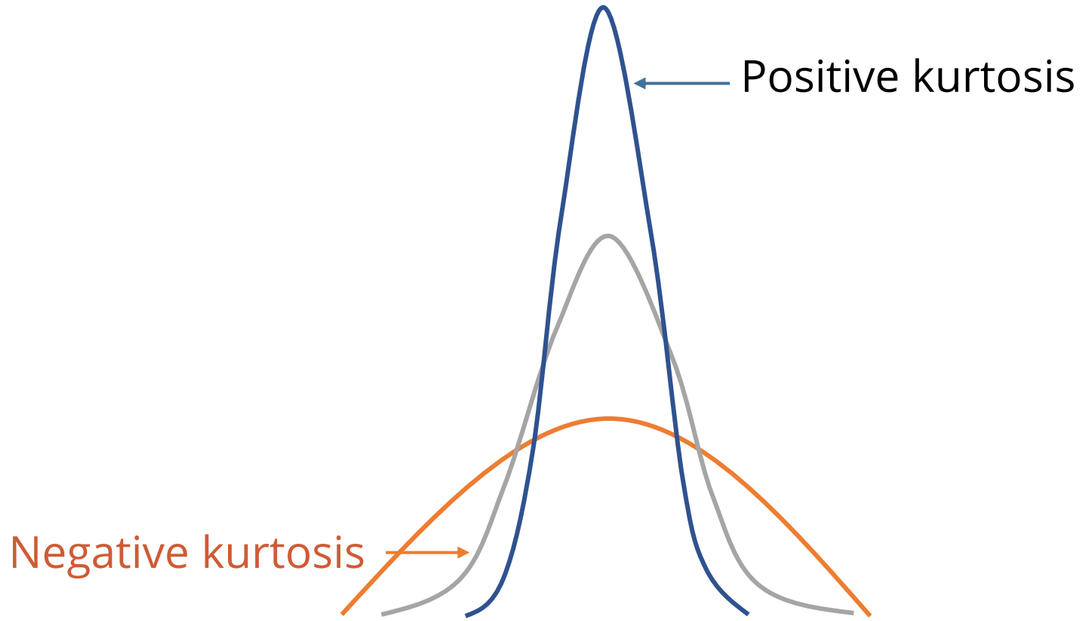
The following image shows a normal distribution superimposed with a positive kurtosis and a distribution with a negative kurtosis.



While skewness indicates the extremity of the normal distribution on one of the tails, kurtosis measures extreme values on both tails of the distribution.

Kurtosis

A positive kurtosis will be narrower and taller around the mean; a negative kurtosis will be flatter and shorter.



Student’s T-Distribution

A Student’s T-Distribution is similar to a normal distribution, like a bell curve, but with a thicker tail.

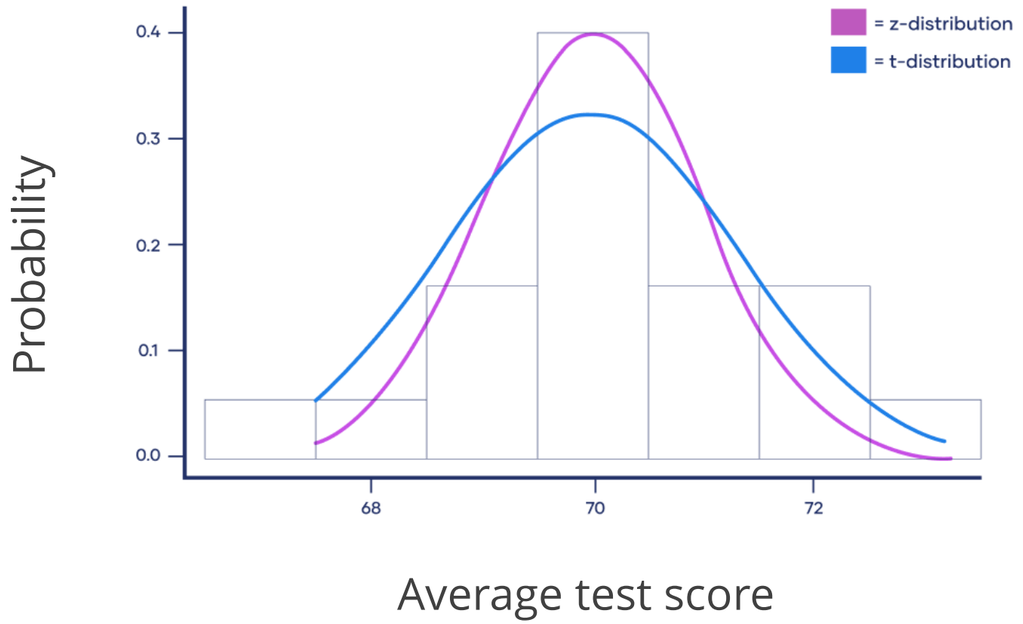
It is characterized by a parameter called degrees of freedom (or DOF), a mean of 0, and a standard deviation of 1.

The tail in a normal distribution qualifies its classification as a Student’s T-Distribution.

They have a greater chance of extreme values than normal distributions, indicated by their flatter tails.

Student's T-Distribution

When measuring the average test score from a sample of only 20 students, the t-distribution should be used to estimate the confidence interval around the mean.



Hypothesis Testing and Mechanism

Hypothesis

It is a statement that proposes a possible explanation for an observed phenomenon or relationship, which can be tested through scientific investigation.

Hypothesis Components

A hypothesis has two components:

- Dependent variable
- Independent variable

Hypothesis Components

Independent variable

It is the cause of a study, and its value influences the value of the dependent variable.

Example

Dosage of drug administered for blood pressure

Dependent variable

It is a variable that is being studied and measured, and its value is influenced by another variable.

Example

Value of blood pressure influenced by the dosage of drug

Hypothesis Components

Consider the following phrase:

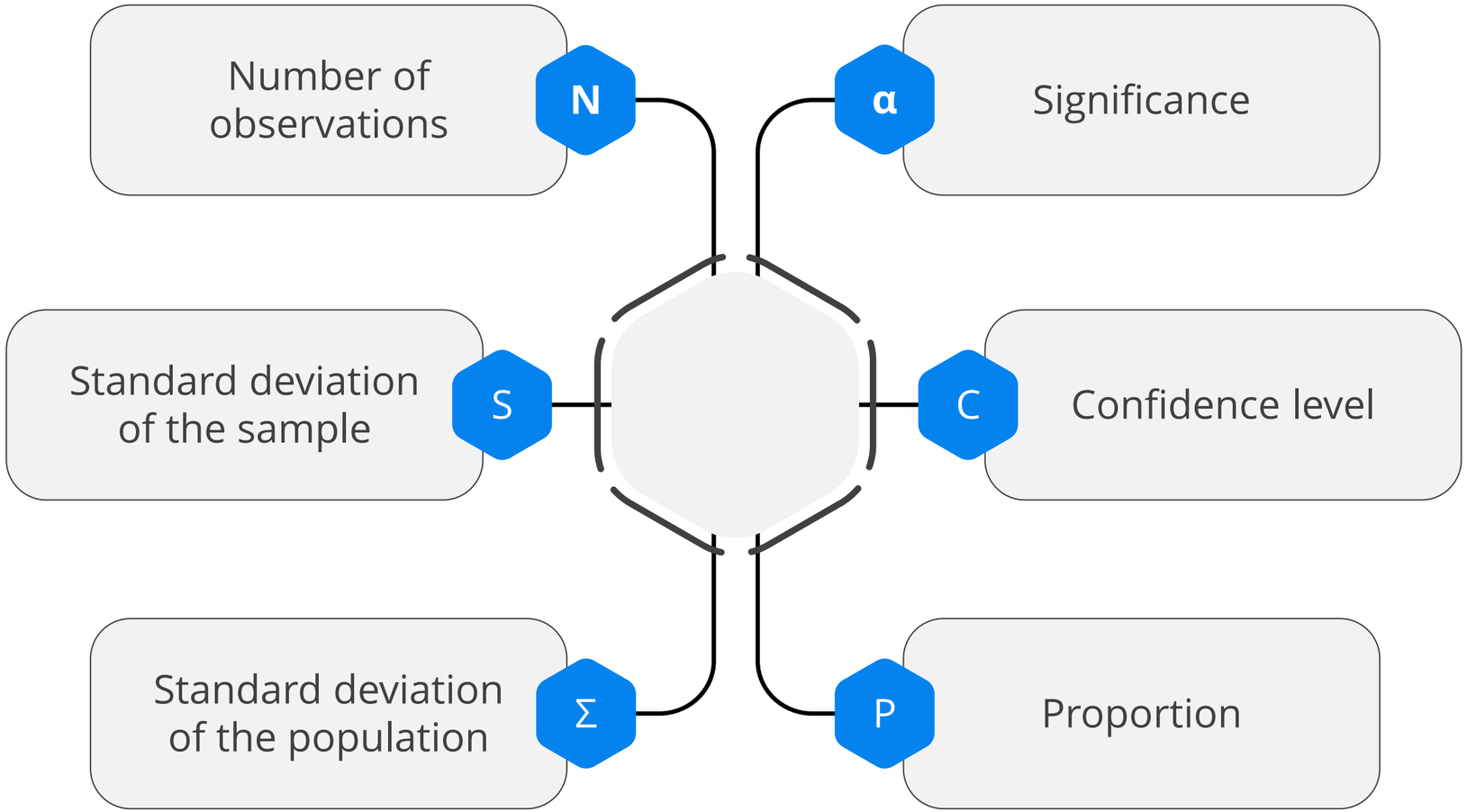
If you do not clean the fish tank once every three days, the fish will probably not survive for more than three months.

‘don't clean ..... day’ is the independent variable

‘fish won’t survive ..... months’ is the dependent variable

Hypothesis Components

The following elements are included in a hypothesis' proposition and expression:



Hypothesis Components

A good hypothesis must:

Explicitly include the independent and dependent variable

Be based on authentic research work

Be testable or amenable for testing

Null and Alternate Hypothesis

Most statisticians typically provide two ways of evaluating hypotheses in hypothesis testing:

Null hypothesis [H_0]

Alternate hypothesis [H_1]

Null and Alternate Hypothesis: Example

Hypothesis: A sales promotion drive will increase the average monthly sales [μ] by 500 units.

Null hypothesis [H_0]: Insignificant impact on the average monthly sales

$$\mu = \mu_0$$

Alternate hypothesis [H_1]: Significant impact on the average monthly sales

$$\mu > \mu_0$$

Null and Alternate Hypothesis: Example

Consider the hourly output of two machines, A and B.

Null hypothesis [H_0]: Average hourly output of machine A (μ_1) differs insignificantly from machine B (μ_2)

$$\mu_1 = \mu_2$$

Alternate hypothesis [H_1]: Average hourly output of machine A (μ_1) is significantly larger than that of machine B (μ_2)

$$\mu_1 > \mu_2$$

A null hypothesis is the negation of the assertion, while an alternative hypothesis is the assertion itself.

Hypothesis Testing

It is a statistical method used to determine whether a hypothesis about a population parameter is supported by the sample data or not.

The sample data may come from a larger population of data, or even from data-generating experimentation.

Hypothesis Testing

The hypothesis testing can be interpreted using confusion matrix as shown below:

Actual class	Predicted class	
	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

Here's a breakdown of the elements in the confusion matrix:

- True Positive (TP): The number of observations correctly predicted as positive.
- False Negative (FN): The number of observations incorrectly predicted as negative.
- False Positive (FP): The number of observations incorrectly predicted as positive.
- True Negative (TN): The number of observations correctly predicted as negative

Hypothesis Testing

A hypothesis test has four steps:

A statement stating which of the two hypotheses (null or alternative) is right

A plan that outlines how to evaluate data

Execution of the plan to physically carry out the analysis

An analysis of the result that rejects the null hypothesis or states that the null hypothesis is plausible

Discussion: Distribution

Duration: 10 minutes

- What are distribution and normal distribution?

Answer:

- A statistical or probability distribution function (PDF) is a central concept in probability theory and statistics.
- A normal distribution, also known as a Gaussian distribution, is the most frequently encountered type of distribution.

- What is a hypothesis, and what are its components?

Answer:

- A hypothesis is a statement suggesting a potential explanation for an observed phenomenon or relationship. It can be tested through scientific investigation.
- The two key components of a hypothesis are the dependent and independent variables.

Hypothesis Testing Outcomes: Type I and Type II Errors

Outcomes of Hypothesis Testing

There are four decisions and outcomes of hypothesis testing, H_0 (Null hypothesis):

- True and is rejected
- True and is accepted
- False and is rejected
- False and is accepted

- The first and fourth outcomes are incorrect inferences and are referred to as type I errors.
- The second and third are correct inferences and are referred to as type II errors.

Probability of Errors

These are represented as:

Type I error

Type II error

It is impossible for both values to be zero simultaneously when inferences are based on samples, but it can be achieved through complete enumeration.

Sampling variability in statistical inference makes it unlikely for both values to be zero simultaneously. However, complete enumeration allows for the determination of both values as zero.

Probability of Errors

If one of them is set to zero, the other becomes one. This implies that:

$$\alpha + \beta \neq 1$$

Alpha and Beta are commonly low values.

In most situations, the value of α is set at 0.05 or 0.01, and large sample sizes are used so that β also has a low value.

Level of Significance

The selected value of α is known as the level of significance

If α = 0.5

Level of significance = 5%

Confidence Interval

Discussion: Confidence Interval

Duration: 10 minutes

- What is a confidence interval?
- What is Bayes' theorem, and how is it helpful?

Confidence Interval

A confidence interval (CI) generally indicates the amount of uncertainty in any distribution.

- It is usually expressed as a number or range of numbers and computed for various distribution statistics.
- CI is the probability that a particular population parameter will fall between a set of values for a certain period of time.
- CIs can take any number of probability limits, the most common being 95%, and in some cases even 99%.

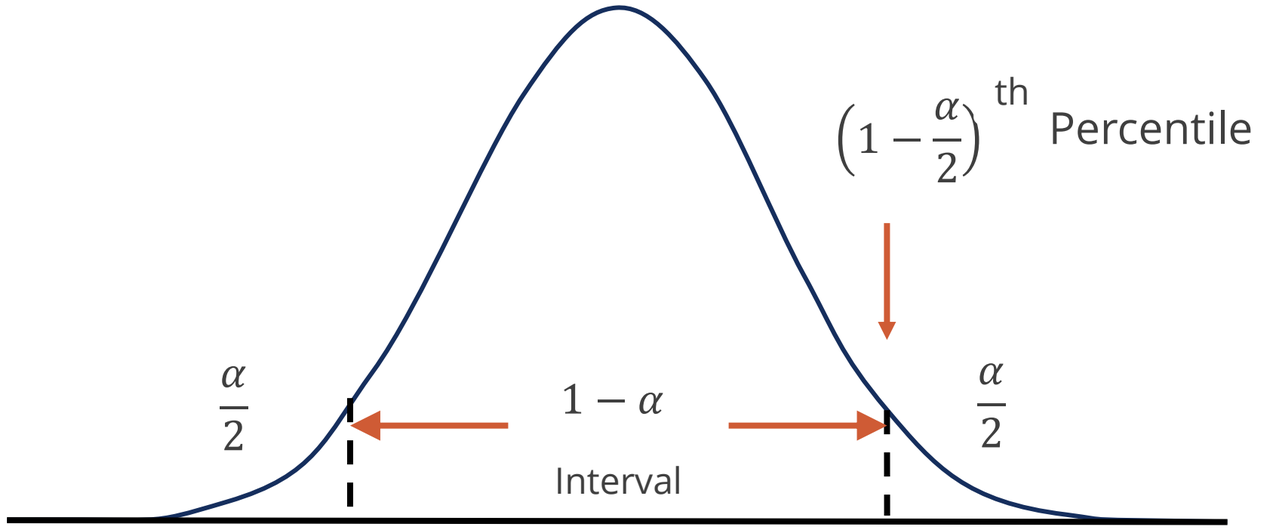
Confidence Interval

When not aware of the behavior of a population, deduce CIs based on sample data.

CI is essentially a range of values that bind the statistic's mean value, which could in turn contain an unknown population parameter.

Confidence Interval

A typical CI in a statistical distribution can be seen below:



An upper limit and a lower limit of CI are marked on either side of the distribution.

Confidence Interval: Example

Consider a survey of a group of car owners to see how many gallons of gas they fill in their car in a year.

Reduce CI value if there is a lack of information about the population characteristics. e.g. from 99% to 90%.

Margin of Errors

Margin of Errors

It shows by how many percentage points the results will differ from the one indicated by population value.

Consider the following statement:

Conduct a poll of a group of automobile owners to see how many gallons of gas they use in a year.

The statistic distribution data is within three percentage points of the real population value, 95% of the time.

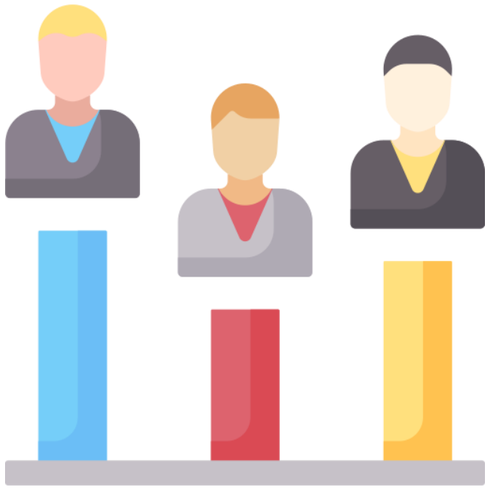
Margin of Errors

The margin of error (MoE) is an important part. Without it one cannot accept inference from statistical analysis.

Lower MoE value implies better accuracy.

Margin of Errors

MoEs are popularly used in pools and elections surveys.



A pool survey's margin of error must be scrutinized before accepting the confidence interval.

Margin of Errors: Example

Example: Gallup poll survey conducted in the 2012 US Presidential elections

The survey indicated 49% voting in favor of Mitt Romney and 47% in favor of Barack Obama, with:

Confidence Interval of 95%

Margin of Error of +/- 2%

However, Barack Obama polled 51%, while Mitt Romney got 47% in the actual election.

The results were even outside the range of the Gallup poll's MoE of +/-2%.

This demonstrates how carefully statistics must be handled while considering CI, CL, and MoE.

Confidence Level

Confidence Levels

It is the percentage of probability or certainty that the confidence interval will contain the true population parameter.

In statistics, confidence levels are expressed as a percentage. Example: 99%, 95%, or 80%

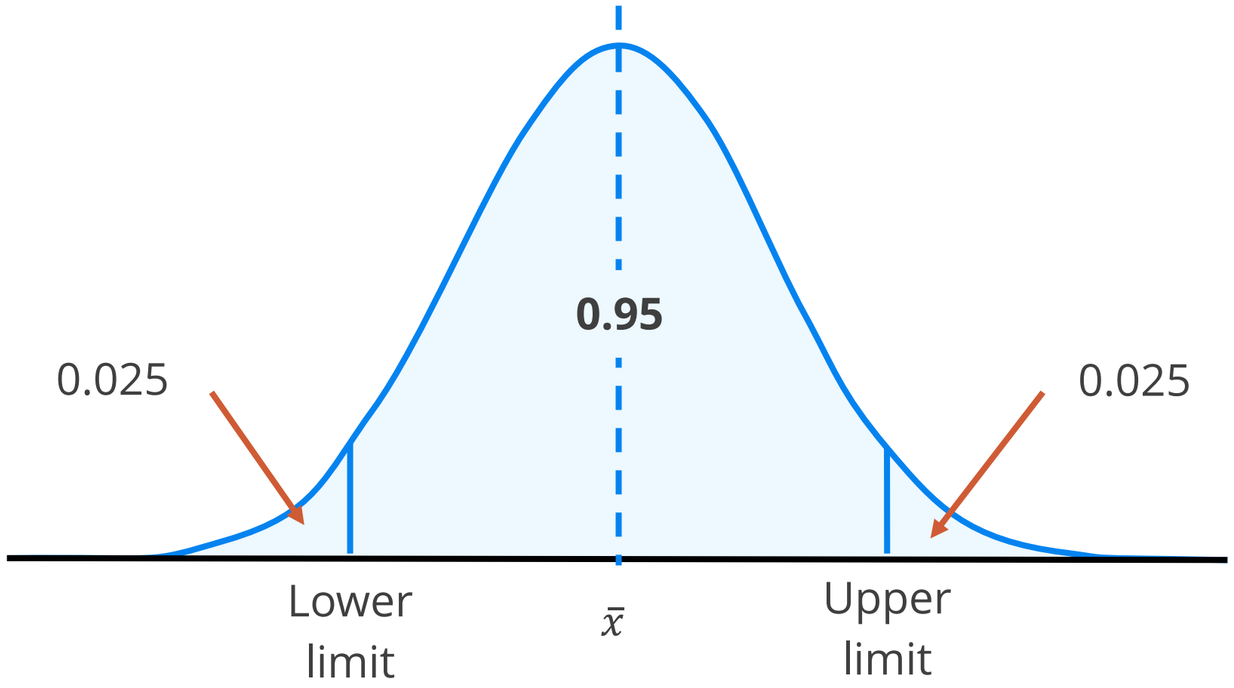
For the purpose of supporting or disposing of the null hypothesis:

Scientists and engineers work with a level of 95% or more

Governmental organizations and departments use 90%

Confidence Levels

A graphical representation of a confidence level of 95% is shown on the screen:



A confidence level of 95% means that when the experiment is repeated, survey results will match the results from a population 95% of the time.

Confidence Levels

Confidence levels vary by field and are determined in by domain experts.

It ensures that the prediction from the statistic is reliable.

Assisted Practices

Let’s understand the topics below using Jupyter Notebooks.

8.10_T-Test and P-Value Using Python
8.11_Z-Test and P-Value Using Python

Note: Please download the pdf files for each topics mentioned above from the Reference Material section.

Comparing and Contrasting T-tests and Z-tests

T-test

It is a statistical test used to compare the means of two groups and determine if the difference is statistically significant.

It is chosen when:

The population variance is unknown.

The sample size is comparatively small (n < 30).

T-test

T-tests are of two types:

One-sample

Standard deviation of the sample is used instead of population's standard deviation

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

Two-sample

Used for comparing the means of two samples

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

$\bar{x}$: Sample mean | **s:** Standard deviation | **μ:** Mean of population | **n:** Sample size

Z-test

It is a statistical test used to determine whether two population means are different when the variances are known, and the sample size is large.

It is chosen when:

The population variance is known.

The population variance is unknown but Sample size is comparatively large (n ≥ 30).

Z-test

Z-tests are of two types:

One-sample

Compares a population mean with the sample mean

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

Two-sample

Compares the mean of two different samples

Z-score $= \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}}$

$\bar{x}$: Sample mean | **σ:** Standard deviation | **μ:** Mean of population | **n:** Sample size

Choosing Between T-test and Z-test

Both T-test and Z-test work with one-sample and two-sample scenarios.

Both use mean and standard deviation, based on the sample or population.

If the sample size is large enough, both T-test and Z-test will correlate with the same results.

For a large sample size, sample variance provides a better estimate of population variance. Therefore, even if population variance is unknown, one can choose Z-test using sample variance.

Choosing Between T-test and Z-test

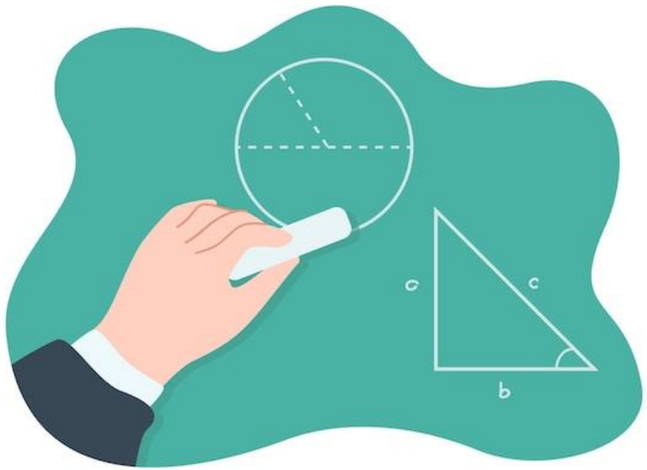
Large sample size has a higher degrees of freedom.

Since T-distribution approaches the normal distribution, the difference between the T-score and Z-score is small.

Bayes' Theorem

Bayes' Theorem

It describes how the conditional probability of every possible cause for a given observed outcome is computed.



It requires knowledge of the probability of each cause and conditional probability of each outcome.

It can be expressed as a mathematical equation.

It allows the calculation of the probability of one event based on its connection with another event.

It is also known as Bayes' law or Bayes' rule.

Applications of Bayes' Theorem

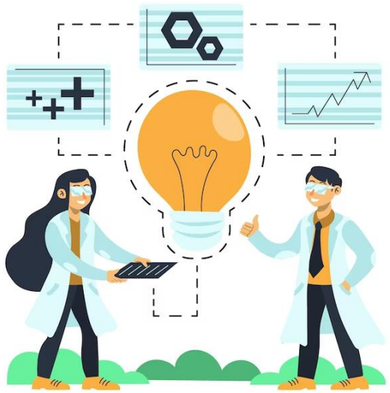
Mathematically, Bayes' theorem is expressed as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- where A and B are events
- P(A|B) is the probability of A happening in case B happens
- P(B|A) is the probability of B happening in case A happens
- P(A) is the independent probability of A
- P(B) is the independent probability of B

Applications of Bayes' Theorem

Bayes' theorem tries to predict one event in case the other is true, as shown below:



Rainy days bring us showers and lightning. Sometimes, there are showers with lightning, and sometimes showers without lightning.

Sometimes there is lightning but no showers.

Bayes' theorem helps find how often there is lightning when there are showers, that is, P(Lightning|Shower).

Applications of Bayes' Theorem

$$P(Lightning|Shower) = P(Shower|Lightning) \times P(Lightning) / P(Shower)$$

P(Lightning) is the probability of lightning

P(Shower) is the probability of showers

P(Shower|Lightning) is the probability of showers when there is lightning

We need to know P(Shower|Lightning), which is the probability of showers when there is lightning.

The formula predicts the **forward** event, P(Lightning|Shower), by knowing the **backward** event, P(Shower|Lightning).

- The problem is seasonal, and the events are related.
- Bayes' theorem helps predict one event when the other is known.

Discussion: Confidence Interval

Duration: 10 minutes

What is the confidence interval?

Answer:

- Answer: A confidence interval (CI) represents a distribution’s uncertainty level.
- It's typically expressed as a range and calculated for various statistical measures within the distribution.

What is Bayes’ theorem and how is it helpful?

Answer:

- Bayes' theorem computes the conditional probability of all potential causes for a given observed outcome.
- For instance, it can calculate the probability of lightning given that there is a shower, denoted as P(Lightning|Shower).

Chi-Squared Distribution

Chi-Squared Distribution

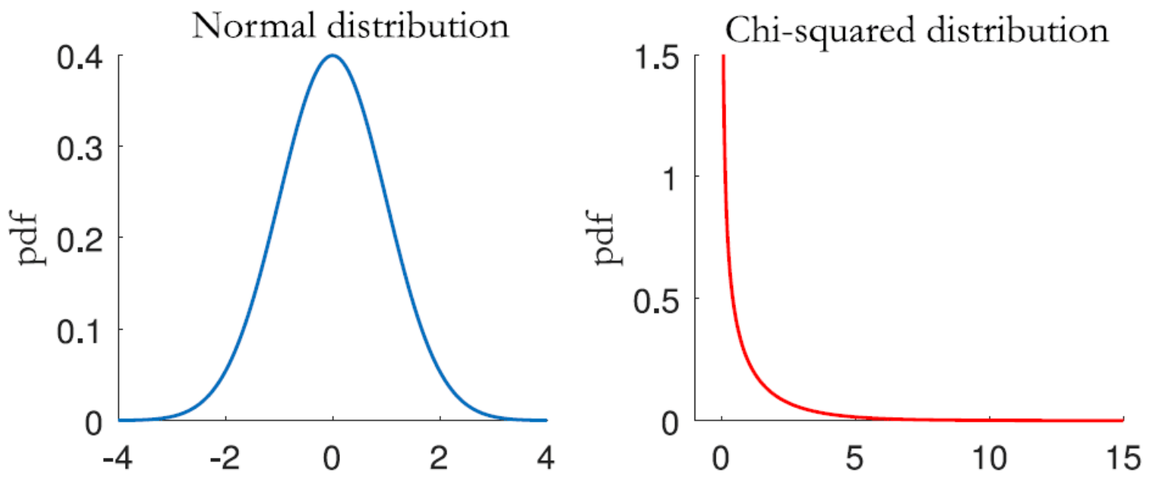
A chi-squared distribution is a continuous probability distribution widely used in statistical inference.

The Greek letter χ is often used. χ2 is termed as chi-squared.

χ2 distribution and standard normal distribution are related.

Chi-Squared Distribution

If a random variable Z has a standard normal distribution, then Z^2 has the $\chi 2$ distribution with one degree of freedom.



Chi-Squared Distribution

Multiple standard random variables are possible.

Mathematically, for k variables, the following can happen:

$Z_1^2 + Z_2^2$ has 2 degrees of freedom.

$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2$ has 6 degrees of freedom.

For k degrees of freedom:

$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2 + ... + Z_k^2$; this has a χ^2 distribution of k degrees of freedom.

Chi-Squared Distribution

The equation for the probability density function (PDF) of the χ2 distribution with k degrees of freedom is:

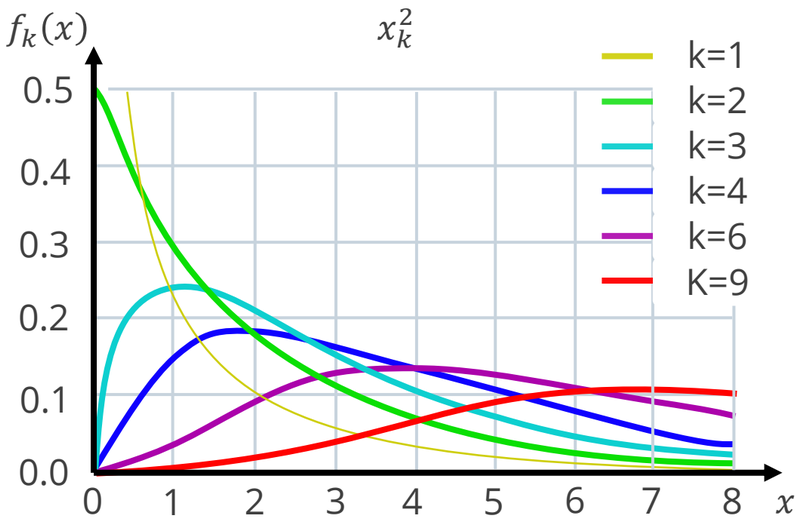
$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}\mathrm{e}^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \\ 0 \end{cases}$$

The function is valid for all positive values of x; k is the number of degrees of freedom.

As the PDF has the gamma function Γ, the χ2 distribution of k degrees of freedom is also a gamma function.

Chi-Squared Distribution

The χ2 distribution of k degrees of freedom, which varies with increasing degrees of freedom, can be seen below :



For k = 1, the PDF is infinity, when χ2 = 0

For k = 2, PDF is 0.5 for χ2 = 0

For k ≥ 3, the χ2 distribution changes to a positively skewed standard normal distribution

With higher degrees of freedom, the skewness and the kurtosis of the χ2 distribution change; the distribution becomes increasingly symmetric.

Chi-Squared Distribution

Note that in any $\chi 2$ distribution, the mean (μ) is k, the number of degrees of freedom, and the variance is 2k.

For example, for k = 3 in the diagram, μ = 3, while the variance is 2 x k, or 6.

The mode of the distribution will occur at k-2 for the distributions with k = 3 and above. So, when k = 4, the mode is at k – 2, which is 2.

Assisted Practices

Let’s understand the topics below using Jupyter Notebooks.

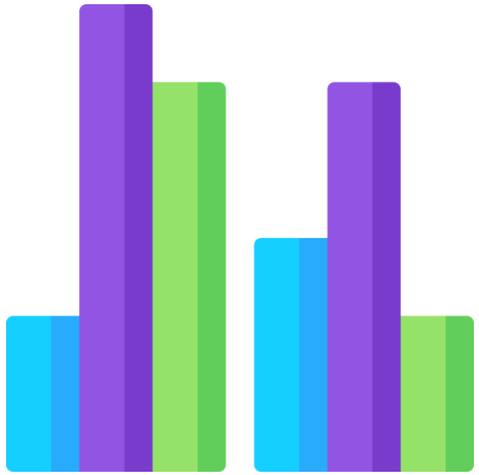
- 8.15_Chi-Squared Distribution Using Python

Note: Please download the pdf files for each topics mentioned above from the Reference Material section.

Chi-Square Test and Goodness-of-Fit

Goodness-of-Fit

It is a statistical test that examines how closely the sample data fits a population with a normal distribution.

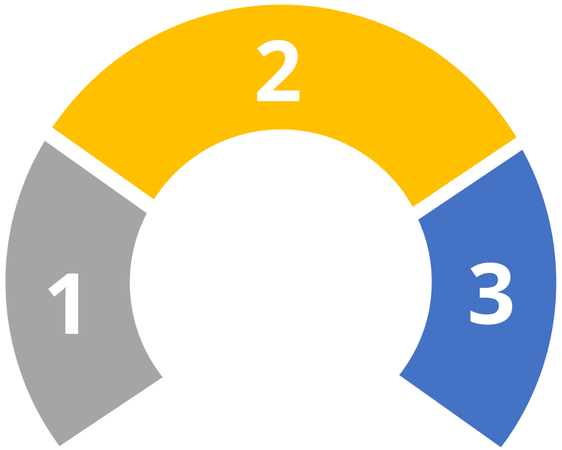


It hypothesizes if a sample is skewed or represents the data found in the actual population.

Goodness-of-Fit

There are multiple ways to determine goodness-of-fit:

Kolmogorov-Smirnov test

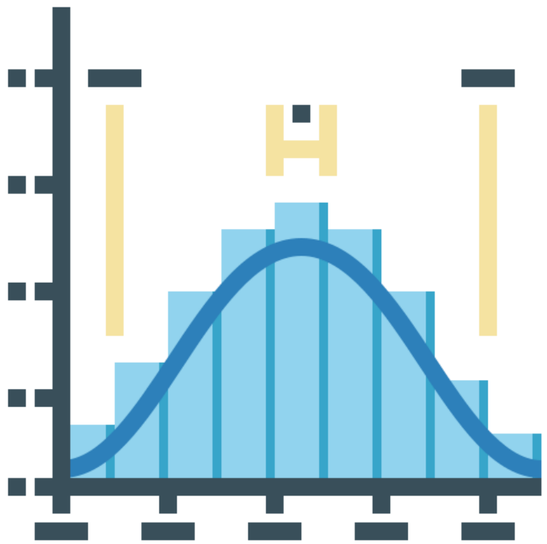


Shipiro-Wilk test

Chi-squared test

Chi-Squared Test

Chi-squared test is the most common and popular goodness-of-fit test.



It determines the relationship between categorical data.

It's a non-parametric test, also known as Pearson's Chi-squared test.

To confirm the goodness-of-fit test, it's important to establish an alpha value, such as the p-value, for the chi-squared test.

p-value refers to the probability of getting results close to the extremes of observed distribution. This assumes that the null hypothesis is correct.

Chi-Squared Test

The chi-squared test is used in data science to determine associations between categorical variables.

It helps in feature selection by identifying the most relevant variables for predictive models.

The test is also used for evaluating model performance and assessing the goodness-of-fit between observed and expected distributions.

Steps for Chi-Squared Test

Step 1: Define the null and alternate hypotheses based on the data.

H_0 implies that the data meets the expected distribution while H_1 implies that it does not.

Step 2: State the alpha value (0.05).

Steps for Chi-Squared Test

Step 3: Calculate the degrees of freedom, k. It depends on the number of categories or groups.

It is usually K – 1, where K is the number of frequencies.

Step 4: State the decision rule and calculate the **decision value** based on the alpha value and degrees of freedom.

Based on this value, either reject H_0 or H_1.

Steps for Chi-Squared Test

Step 5: Calculate the test statistic for χ2 using the formula:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 \Big/ E_i$$

O is observed value, E is expected value from the sample, and K is the degrees of freedom

Step 6: Compare the decision value computed in Step 4 and Step 5 to either accept or reject the null hypothesis.

Step 7: Based on Step 6, conclude the domain observation.

Analysis of Variance or ANOVA

Discussion: ANOVA

Duration: 10 minutes

- What is ANOVA, and how does it work?
- What are the types of ANOVA?

Analysis of Variance or ANOVA

ANOVA (Analysis of Variance) is a collection of statistical models and associated estimation procedures.

It was developed by the statistician Ronald Fisher in 1918.

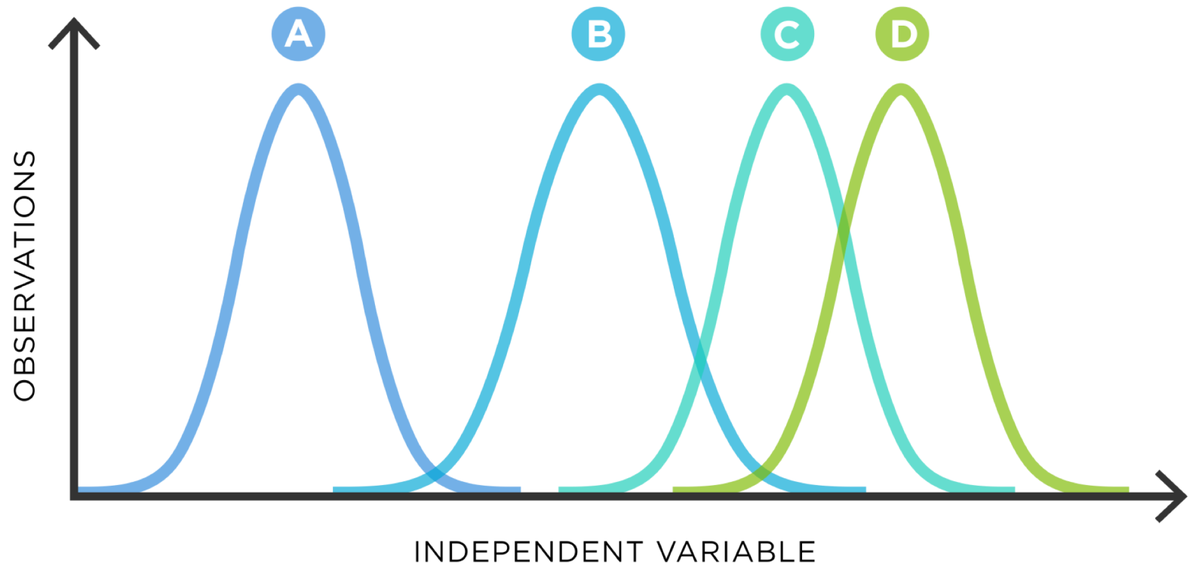
It extends the T and Z tests, as they are constrained to allowing the nominal-level variable to have only two categories.

It is also called Fisher's method and is suitable for performing simultaneous tests on sets of data drawn from different populations.

The term **variance** is used to indicate the variation or the dispersion.

Analysis of Variance or ANOVA

The chart below shows there are four groups, as seen on the screen: A, B, C, and D, and their sample distributions are compared.



Working of ANOVA

ANOVA test compares the means of different groups and reveals any statistical differences between them.

F= MSE / MST

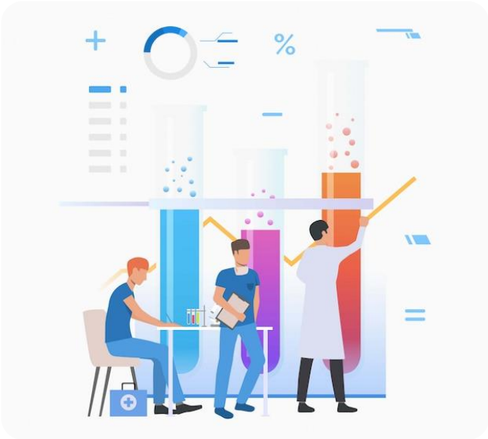
where:
F=ANOVA coefficient
MST=Mean sum of squares due to treatment
MSE=Mean sum of squares due to error

Most of the bio-chemical, pharmaceutical, and science researchers depend heavily on ANOVA for studying the effect on the dependent variable by multiple independent variables.

ANOVA is used in the analysis of complex multivariate situations.

Working of ANOVA: Example

Consider a medical experiment where scientists plan and conduct experiments to understand the relationship between a type of hypertension medicine and the resulting blood pressure.



A set of people or patients is taken as sample population.

Scientists divide the sample population into multiple groups and administer each group with a particular medicine for a trial period.

Hypertension levels are measured for each patient.

Mean hypertension level for each group is calculated.

ANOVA then helps in comparing these group means to find out if they are statistically varying or similar.

ANOVA Terminologies

Some of them are:

Sample mean: It is the average value for a specific group.

Grand mean: It is the mean of all sample means (for multiple groups).

Dependent variable: It is the item or subject under investigation that is supposed to be influenced by many other independent variables.

Independent variable: It can have an impact on or influence the dependent variable.

ANOVA Terminologies

Factor: Independent variables are frequently referred to as factors.

Levels: It denotes the different values of factors that are used in any typical experiment.

F-Statistic: It is also known as F-Ratio and is the outcome of ANOVA.

Fixed-factor model: In this model, experiments use only a discrete set of levels of factors.

Random-factor model: In this model, a random value of level is drawn from multiple possible values of the factor.

Outcome of ANOVA

ANOVA's outcome is known as **F-statistic**.

It is a ratio that shows the difference between the variation in the inter-group and the intra-group.

With the help of this ratio, one can conclude whether the null hypothesis is true and whether to accept or reject it.

Purpose and Procedure of ANOVA

ANOVA is an omnibus test statistic.

The null hypothesis for ANOVA is that there is no significant difference in the **mean** values.

The alternate hypothesis concludes that there are at least some significant differences.

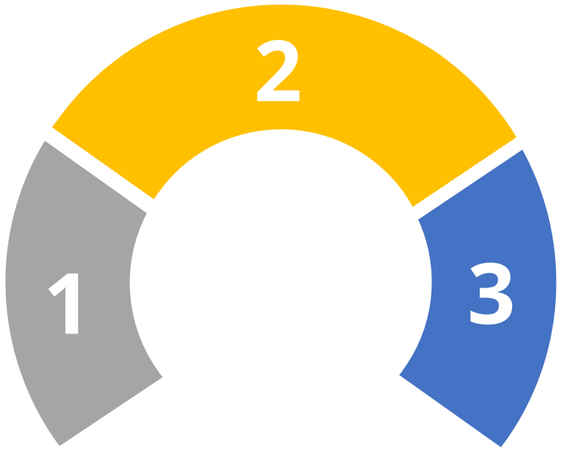
Purpose and Procedure of ANOVA

ANOVA makes the following assumptions about the probability distribution of the responses:

2. Normalcy (distribution of the residuals is normal)

1. Independence of observations

3. Equality of variances (equal variance of data in the groups)



Note that the assumptions also depend on the types of ANOVA used.

Types of ANOVA

There are three types of ANOVA:

One-way ANOVA (one-factor ANOVA)

Two-way ANOVA (two-factor ANOVA)

N-way ANOVA (MANOVA)

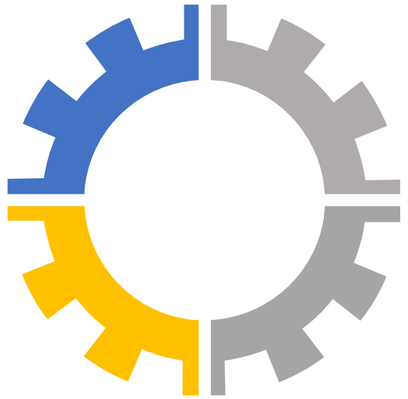
One-Way ANOVA

Also known as simple ANOVA or just ANOVA, it is suitable for experiments with only one factor (independent variable) with two or more levels.

A one-way ANOVA assumes the following:

Continuous data

Independence



Variance

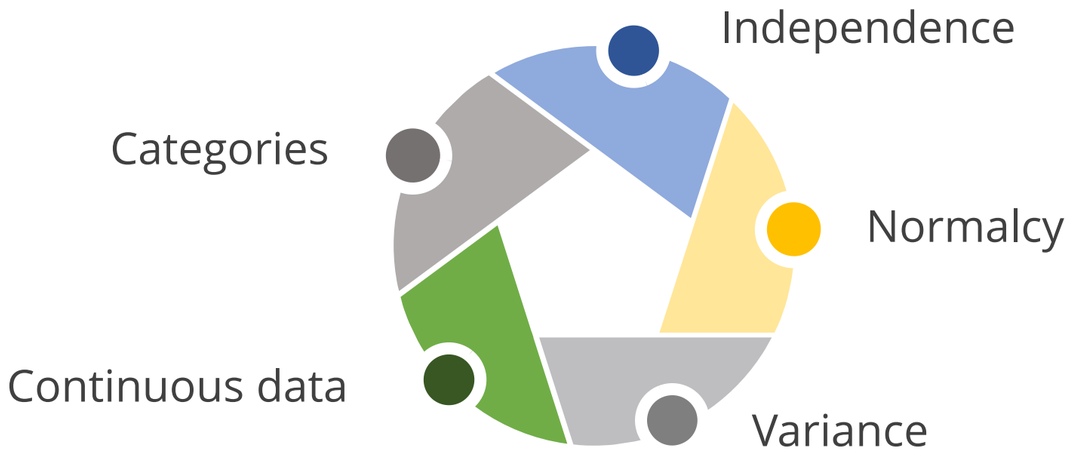
Normalcy

The one-way ANOVA needs to operate on continuous data.

Two-Way ANOVA

It is used when there are two or more independent variables and is also known as full factorial ANOVA or two-factor ANOVA.

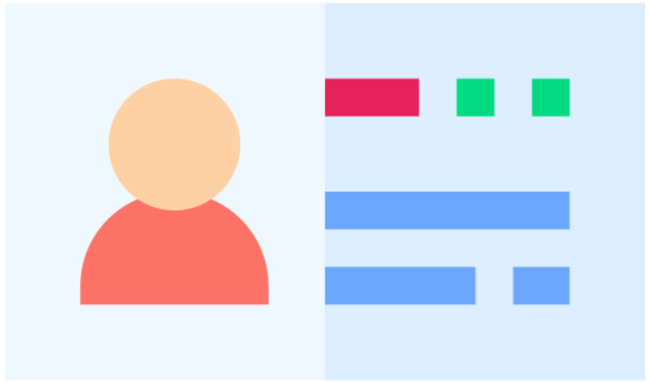
Two-way ANOVA is based on the following assumptions:



For two-way ANOVA, the independent variables should be in separate groups or categories.

N-Way ANOVA

N-Way ANOVA or MANOVA stands for multivariate analysis of variance.



For example, analysis of voter preferences based on gender, age, ethnicity, etc., can be studied using MANOVA.

Discussion: ANOVA

Duration: 10 minutes

- What is ANOVA, and how does it work?

Answer: ANOVA, or Analysis of Variance, is a set of statistical models and estimation procedures. It is used to compare the means of different groups and identify any significant differences.

- What are the types of ANOVA?

Answer: There are three main types of ANOVA: One-way ANOVA, Two-way ANOVA, and N-way ANOVA.

Partition of Variance

Partition of Variance

It is a statistical analysis of the distributions of two or more samples in a population.

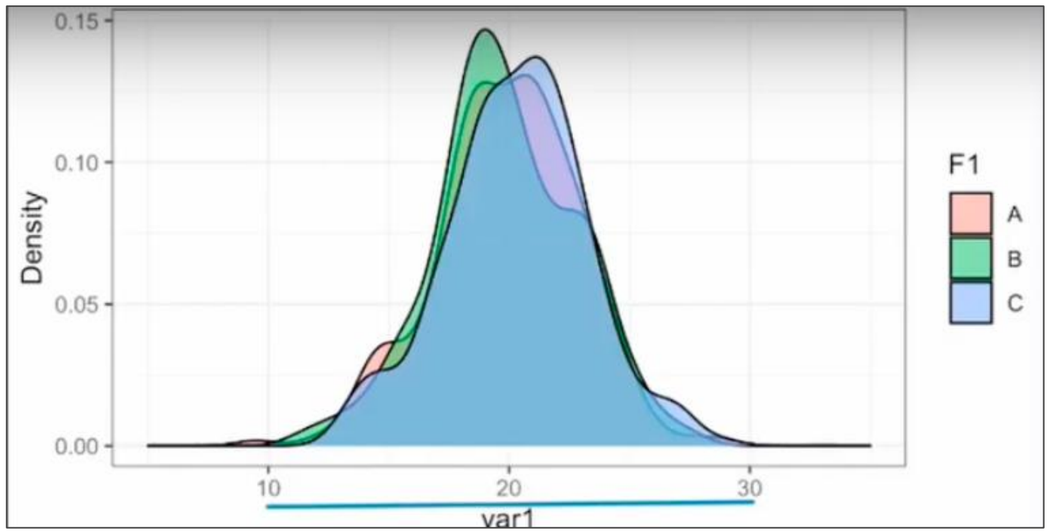


Figure shows the distribution of three different groups, A, B, and C, from a typical population

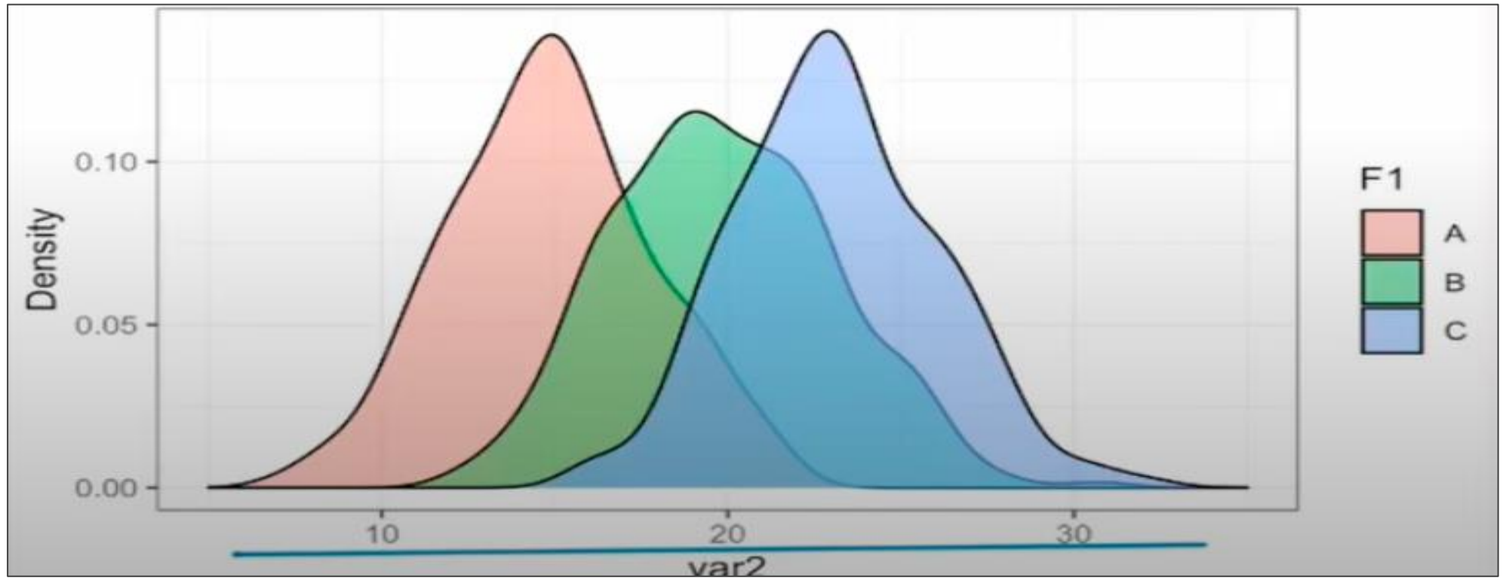
The PDFs are marked using different colours for clarity.

They all have mean = 20 and standard deviation = 3.

The spread in A, B, and C is between 10 and 30, and is denoted by var1.

Partition of Variance

Consider another figure where groups A, B, and C have means of 15, 20, and 25 respectively.



The standard deviation for these three distributions is 3.

Partition of Variance

There are similar distribution groups marked A, B, and C but the variance range, var2, is between 5 and 35.

For var1, the mean is almost the same for A, B and C, which is around 20.

In var2, the means for A, B, and C are approximately 15, 20, and 25 respectively.

Partition of Variance

As the difference between the means of the groups varies, the dispersion range also varies.

The concept is cemented by the variance values of the groups in each case.

Variance for var1 is around 8.69

Variance for var2 is around 20.68

Partition of Variance

Two recognizable components of variances are:

Inter-group (within a group)

Intra-group (between two or more groups).

The former is also called **error** or **residual variance**.

Partitioning

Partitioning variance is done indirectly using the sum of squares.

$$\sigma^2 = \sum (x - x)^2 \,/ (n - 1)$$

Here, the group mean (GM) is considered.

Total Sum of Squares, or $SS_{total} = \Sigma (x - \bar{x})^2$

Partitioning

Similarly, the Error Sum of Squares:

For A, $SS_{error} = \Sigma (x_A - \bar{x_A})^2$

For B, $SS_{error} = \Sigma (x_B - \bar{x_B})^2$

For C, $SS_{error} = \Sigma (x_C - \bar{x_C})^2$

The $SS_{error} = \Sigma (x_A - \bar{x_A})^2 + \Sigma (x_B - \bar{x_B})^2 + \Sigma (x_C - \bar{x_C})^2$

The $SS_{Treatment}$ or the sum of squares between groups and the error is the difference $SS_{total} - SS_{error}$

F-Distribution

F-Distribution

F-distribution is similar and related to $\chi 2$ distribution.

It is also known as F-ratio, Snedecor's F-distribution, and Fisher-Snedecor distribution.

"...F-distribution is essentially a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and other F-tests...".

Unlike chi-square distribution, F-distribution deals with multiple random variables.

F-Distribution

Consider a situation of two independent random variables, R_1 and R_2.

Suppose that R_1 and R_2 have a $\chi 2$ distribution with Degrees Of Freedom (DOF) d_1 and d_2 respectively.

The F-distribution or the F-ratio for this situation is expressed as:

$$F = (R_1/d_1)/(R_2/d_2)$$

Probability Density Function

The probability density function (PDF) of F-distribution can be computed using the following formula:

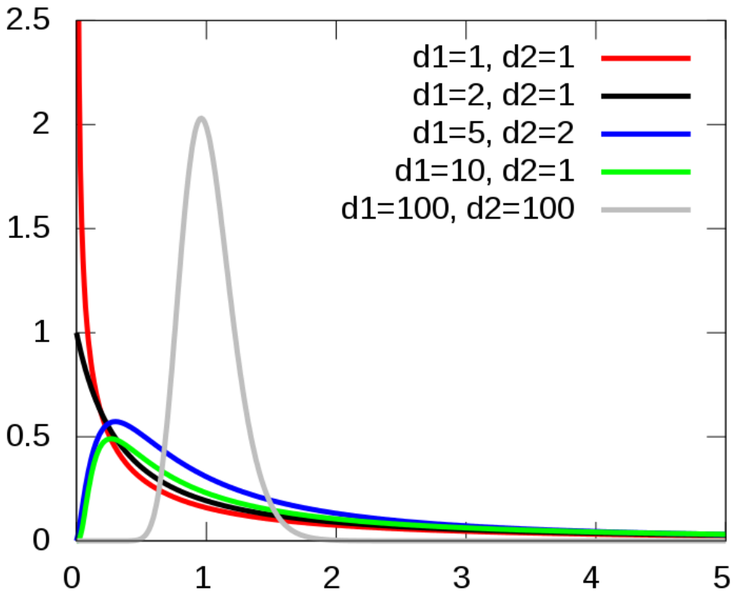
$$f(x; d_1, d_2) = \sqrt{\frac{(d_1 x)^{d1}\, d_2^{d_2} / (d_1 x + d_2)^{d_1 + d_2}}{x B\left({}^{d_1}/_{2}, {}^{d_2}/_{2}\right)}}$$

The formula is valid for all positive values of x and the degrees of freedom d_1 and d_2.

The B in the formula represents the beta function.

Graph of F-distribution's PDF

A typical PDF of F-distribution is as shown below:



The figure suggests that the shape of the F-distribution curve depends on the two degrees of freedom, d_1 and d_2.

Description of F-distribution

Thus, f(6, 8) would refer to an F-distribution with $d_1 = 6$ and $d_2 = 8$ degrees of freedom.

Likewise, $f(8, 6)$ would refer to an F-distribution with $d_1 = 8$ and $d_2 = 6$ degrees of freedom.

Note: the curves represented by f(6, 8) and f(8, 6) are different from each other.

Assisted Practices

Let’s understand the topics below using Jupyter Notebooks.

- 8.22_F-Distribution Using Python

Note: Please download the pdf files for each topics mentioned above from the Reference Material section.

F-Test

F-Test

An F-test is a statistical test where the test statistic has an F-distribution under the null hypothesis.

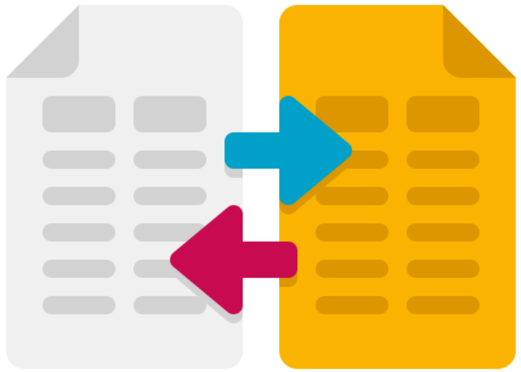
It is used when comparing statistical models fitted to a data set to identify the model that best fits the population.

It always implies the comparison of two variances.

It resembles ANOVA.

Comparison of Two Variances

The F-test compares two variances, s1 and s2, by calculating their ratio.



The result is always positive as variances are always positive.

Comparison of Two Variances

The equation for F-test is:

$$F = s^2_1 / s^2_2$$

When the variances are equal, their ratio is 1.

Example

If there are two data sets with sample 1 (variance of 8) and sample 2 (variance of 8), the ratio will be 64/64 = 1.

Comparison of Two Variances

As a rule of thumb, when doing an F-test, check to see if the population variances are equal.

If the samples are from the same population, the variances are one.

The null hypothesis will be that the variances are equal.

Assumptions for F-test

While performing an F-test, the following assumptions are made:

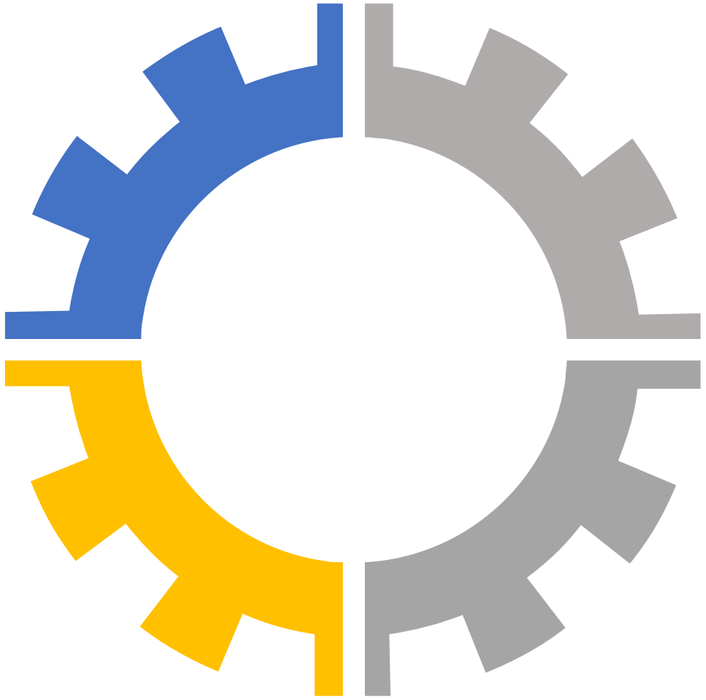
The population distribution must be approximately fitting to the normal distribution.

The samples must be independent of each other.

Assumptions for F-test

If degrees of freedom aren't available in F-table, use the larger critical value.

If standard deviations are available, square them to get the value of variances.



The larger variance is always in the numerator.

For two-tailed tests, alpha is to be taken at half of its value.

Steps to Perform F-test

There are four simple steps for performing an F-test:

Step 1: State the null hypothesis and alternate hypothesis.

Step 2: Compute the F-value using the residual sum of squares, number of restrictions, and number of independent variables.

$$F = (SSE1 – SSE2 / m) / SSE2 / n-k,$$

Steps to Perform F-test

There are four simple steps for performing an F-test:

Step 3: Find F-statistic, the critical value.

F-statistic = (variance of the group means) / (mean of variances within the group)

Step 4: Based on the results, support or reject the null hypothesis.

Rejecting the null hypothesis signifies a significant difference or relationship while failing to reject it suggests insufficient evidence for such distinction or association.

Key Takeaways

- A hypothesis is a statement that introduces a question and proposes an answer.
- A null hypothesis is the negation of the assertion, while an alternative hypothesis is the assertion itself.
- A confidence interval (CI) generally indicates the amount of uncertainty in any distribution.
- Margin of Errors shows by how many percentage points the results will differ from the one indicated by population value.
- Bayes' theorem describes how the conditional probability of every possible cause for a given observed outcome is computed.

Knowledge Check

Knowledge Check

1

What are the four steps of hypothesis testing?

A. Statement, plan, execution, and analysis

B. Statement, observation, analysis, and conclusion

C. Statement, experimentation, analysis, and conclusion

D. Statement, data collection, analysis, and conclusion

Knowledge Check

1

What are the four steps of hypothesis testing?

A. Statement, plan, execution, and analysis

B. Statement, observation, analysis, and conclusion

C. Statement, experimentation, analysis, and conclusion

D. Statement, data collection, analysis, and conclusion

The correct answer is **A**

Statement, plan, execution, and analysis are the four steps of hypothesis testing.

Knowledge Check

2

What is a Margin of Error (MoE)?

A. A measure of the probability that a particular population parameter will fall between a set of values for a certain period of time

B. A range of values that bind the statistic's mean value, which could in turn contain an unknown population parameter

C. A measure of how much the results will differ from the one indicated by the population value

D. A percentage indicating the probability that the confidence interval will contain the true population parameter

Knowledge Check 2

What is a Margin of Error (MoE)?

A. A measure of the probability that a particular population parameter will fall between a set of values for a certain period of time

B. A range of values that bind the statistic's mean value, which could in turn contain an unknown population parameter

C. A measure of how much the results will differ from the one indicated by the population value

D. A percentage indicating the probability that the confidence interval will contain the true population parameter

The correct answer is **C**

It is a measure of how much the results will differ from the one indicated by the population value.

Knowledge Check

3

What is Bayes' theorem used for?

A. To calculate the mean and standard deviation of a sample

B. To calculate the probability of one event based on its connection with another event

C. To determine goodness-of-fit

D. To analyze the differences among means

Knowledge Check

3

What is Bayes' theorem used for?

A. To calculate the mean and standard deviation of a sample

B. To calculate the probability of one event based on its connection with another event

C. To determine goodness-of-fit

D. To analyze the differences among means

The correct answer is **B**

Bayes' theorem is used to calculate the probability of one event based on its connection with another event.

Thank You