BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Statistical Inference-II

Course Code: 23ST11CC

Unit-I

Parametric Tests

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

Basic Statistics

Statistics

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a mathematical discipline to collect, summarize data.

According to statistician Sir **Arthur Lyon Bowley**, statistics is defined as "Numerical statements of facts in any department of inquiry placed in relation to each other".

Scope of Statistics

Statistics is used in many sectors such as psychology, geology, sociology, weather forecasting, probability and much more. The goal of statistics is to gain understanding from the data, it focuses on applications, and hence, it is distinctively considered as a mathematical science.

Methods in Statistics

The methods involve collecting, summarizing, analyzing, and interpreting variable numerical data. Here some of the methods are provided below.

- Data collection
- Data summarization
- Statistical analysis

Types of Statistics

There are two types of statistics.

- Descriptive Statistics
- Inferential Statistics

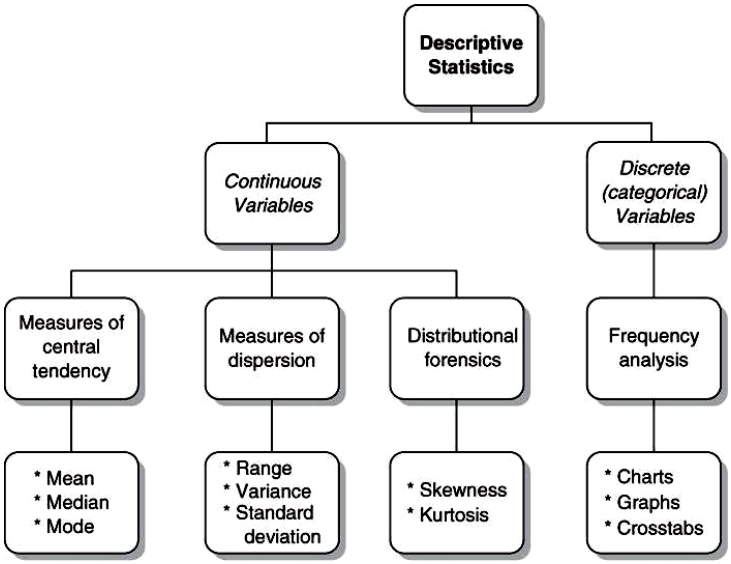
Descriptive statistics

Descriptive Statistics is the branch of statistics that involves organizing, displaying, and describing data.

Types of descriptive statistics

There are 4 main types of descriptive statistics:

- The distribution concerns the frequency of each value.
- The central tendency concerns the averages of the values.
- The variability or dispersion concerns how spread out the values are.
- Skewness and Kurtosis.



Inferential statistics

Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population.

Types of Inferential statistics

There are 2 main types of inferential statistics:

- Estimation
- Hypothesis testing

Estimation

Estimation in statistics is any procedures used to calculate the value of a population drawn from observations within a sample size drawn from that population. A good Estimation is,

- Unbasiedness
- Efficiency
- Consistency
- Sufficiency

Types of estimation

There are two types of estimation.

- Point estimation
- Interval estimation.

UNIT - I

Hypothesis Testing

Hypothesis testing or significance testing is a method for testing a assumption about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

According to **Jim Frost**, Hypothesis Testing is a form of inferential statistics that allows us to draw conclusions about an entire population based on a representative sample

Population

Population is a collection of all the units or elements that possess common characteristics

Sample

Sample is a subgroup of the members of the population

Parameter

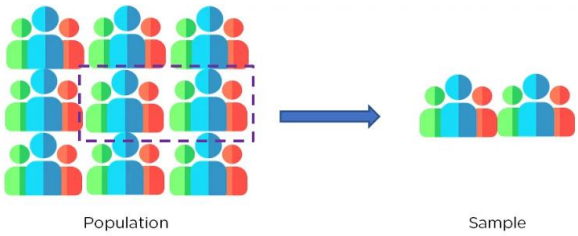
A parameter is a measure that describes the whole population.

Statistic

A statistic is a measure that describes the sample.

Sampling error

A sampling error is the difference between a population parameter and a sample statistic.



Samples are used when

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size. Take the example of a study that documents the results of a new medical procedure. It is unknown how the procedure will affect people across the globe, so a test group is used to find out how people react to it.

Data

Data is defined as a systematic record corresponding to a specific quantity. Basically, data can be summarized as a set of facts and figures which can be used to serve a specific usage or purpose. For instance, data can be used as a survey or an analysis.

Types of Data

Data can be classified into two types.

- Qualitative
- Quantitative

Qualitative data

Qualitative data are also referred to as categorical data. They are an observed phenomenon and cannot be measured with numbers.

Examples: age group, gender, origin, and so on.

Types of Qualitative data

Qualitative data can be subdivided into two types.

- Nominal
- Ordinal

Nominal Data

Nominal data is a type of data that represents discrete units which is why it cannot be ordered and measured.

Examples: Gender (Male, Female)

Hair color (Black, Brown, Gray, etc)

Nationality (Indian, American, Chinese, etc)

Ordinal Data

Ordinal values represent discrete as well as ordered units.

Example: Opinion (agree, mostly agree, neutral, mostly disagree, disagree)

Socioeconomic status (low income, middle income, high income)

Quantitative data

Quantitative data, on the other hand, tells us about the quantities of things or the things we can measure. And, so they are expressed in terms of numbers. It is also known as numerical data and includes statistical data analysis.

Examples: height, water, distance, and so on.

Types of Quantitative data

Two types of quantitative data are discrete data and continuous data.

Discrete and Continuous Data

- **Discrete Data:** These are data that can take only certain specific values rather than a range of values. For example, data on the blood group of a certain population or on their genders is termed as discrete data. A usual way to represent this is by using bar charts.

- **Continuous Data:** These are data that can take values between a certain range with the highest and lowest values. The difference between the highest and lowest value is called the range of data. For example, the age of persons can take values even in decimals or so is the case of the height and weights of the students of your school.

Interval Data

It represents ordered data that is measured along a numerical scale with equal distances between the adjacent units. These equal distances are also referred to as intervals.

Examples: IQ test's intelligence scale

Time if measured using a 12-hour clock

Ratio Data

Like Interval data, ratio data are also ordered with the same difference between the individual units. However, they also have a meaningful zero so they cannot take negative values.

Examples: Temperature on a Kelvin scale (0 degrees represent total absence of thermal energy)

Height (zero is the starting point)

Types of Hypothesis Tests

Hypothesis Tests can be classified into two big families:

- **Parametric Tests:** if samples follow a normal distribution. In general, samples follow a normal distribution if their mean is 0 and variance is 1.

- **Non-Parametric Tests:** if samples do not follow a normal distribution.

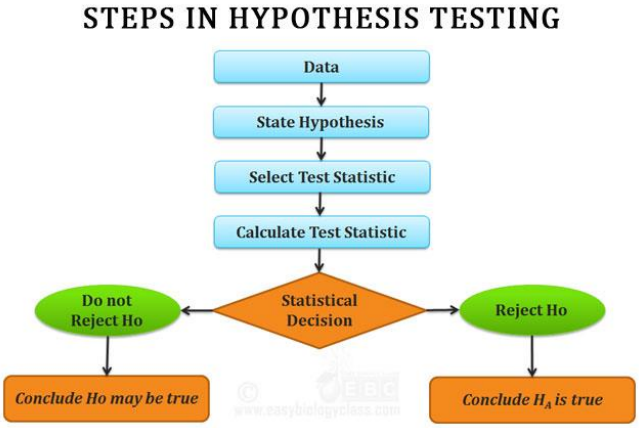
Four Steps to Hypothesis Testing

Step 1: State the hypotheses.

Step 2: Set the criteria for a decision.

Step 3: Compute the test statistic.

Step 4: Make a decision.



Hypothesis

Statement about the Population is called Hypothesis. Hypothesis testing is a statistical procedure in which a choice is made between a null hypothesis and an alternative hypothesis based on information in a sample.

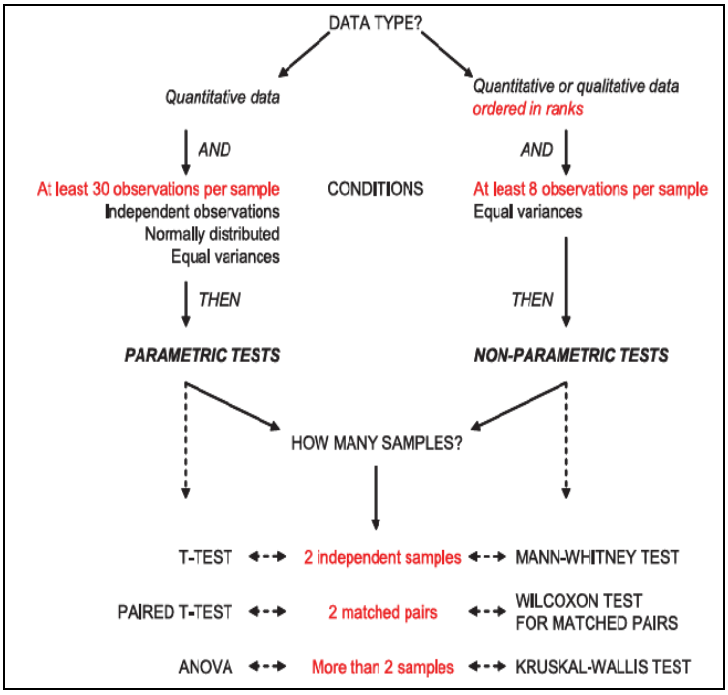
Null Hypothesis

The null hypothesis, denoted H_0, is the statement about the population parameter that is assumed to be true unless there is convincing evidence to the contrary.

Alternative Hypothesis

The alternative hypothesis, denoted Ha, is a statement about the population parameter that is contradictory to the null hypothesis, and is accepted as true only if there is convincing evidence in favor of it.

Choosing the Appropriate Statistical Test



Simple Hypothesis

When a hypothesis specifies all the parameters of a probability distribution, it is known as simple hypothesis. The hypothesis specifies all the parameters, i.e μ and σ of a normal distribution. For Example, the random variable x is distributed normally with mean $\mu=0$ and SD=1 is a simple hypothesis. The hypothesis specifies all the parameters (μ and σ) of normal distributions.

Composite Hypothesis

If the hypothesis specific only some of the parameters of the probability distribution, it is known as composite hypothesis. In the above example if only the μ is specified or only the σ is specified it is a composite hypothesis.

Independent variables

Independent variables are the ones that are manipulated, controlled, or changed in the study. Independent variables are isolated from other factors of the study. For example, the independent variable is whether people in the study wear masks.

Dependent variables

Dependent variables are dependent on other factors of the study and are influenced by what happens to the independent variables. The dependent variable is how many cases of virus emerge among the group studied.

Critical Region

The null hypothesis H_0 is rejected if the observed sample point falls in w and if it falls in, we accept H_0 i.e the region of rejection of H_0 when H_0 is true is that region of the outcome set where H_0 is rejected. If the sample point falls in that region, then it is called critical region.

Test Statistic

A test statistic measures the accuracy of the predicted data distribution relating to the null hypothesis you use when analyzing data samples.

P-value

P-values are the probability that a sample will have an effect at least as extreme as the effect observed in your sample if the null hypothesis is correct.

One tailed

If Hypothesis has the form $\mu < \mu_0$ the test is called a **left-tailed test.**

If Hypothesis has the form $\mu > \mu_0$ the test is called a **right-tailed test.**

Each of the last two forms is also called a **one-tailed test.**

Two tailed

If Hypothesis has the form $\mu \neq \mu_0$ the test is called a **two-tailed test.**

Types of Errors

There are two types of Errors.

- Type I Error
- Type II Error

		Decision From Sample	
		Reject H_0	**Accept H_0**
True State	**H_0 True**	Wrong **(Type I Error)**	Correct
	H_0 False	Correct	Wrong **(Type II Error)**

Type I Error

Rejecting the null hypothesis H_0, when is true is called type I error.

Type II Error

The error of accepting H_0, when it false is called type II error.

Level of Significance

Probability of type I error is known as level of significance of test. It is also called as size of the critical region.

$$\alpha = p[type\ I\ error]$$

$$\alpha = p[x \epsilon w / H_0]$$

$$\alpha = \int_w L_0\ dx$$

Where, L_0 is the likelihood function of the sample observation under H_0.

Power of the test

Probability of type II error is denoted by β. *1* - β is called power function of the hypothesis against the alternative H_1. The value of the power function at a parameter point is called power of the test at that point (i.e).

$$\beta = p[type\ II\ error]$$

$$\beta = p[x \epsilon w / H_1]$$

$$\beta = \int_{\bar{w}} L_1\ dx$$

we have,

$$\int_w L_1\ dx + \int_{\bar{w}} L_1\ dx = 1 \Rightarrow \int_w L_1\ dx + \beta = 1$$

$$\int_w L_1\ dx = 1 - \beta$$

Large sample theory

The sample size n is greater than 30 (n≥30) it is known as large sample. For large samples the sampling distributions of statistic are normal (Z test). A study of sampling distribution of statistic for large sample is known as large sample theory.

Small sample theory

If the sample size n ils less than 30 (n<30), it is known as small sample. For small samples the sampling distributions are t, F and $\chi 2$ distribution. A study of sampling distributions for small samples is known as small sample theory.

Parametric Test for Population Proportion (Small and Large Samples)

Large sample test

Large sample test are

1. Sampling from attributes
2. Sampling from variables

Sampling from attributes

There are two types of tests for attributes.

1. Test for single proportion
2. Test for equality of two proportions

Test for Single Population Proportion

Single Proportion Test

The One Sample Proportion Test is used to estimate the proportion of a population. It compares the proportion to a target or reference value and also calculates a range of values that is likely to include the population proportion. This is also called the hypothesis of inequality.

Assumptions

- The data are simply random values from the population
- The population follows a binomial distribution
- When both mean (np) and variance (np(1-p)) values are greater than 10, the binomial distribution can be approximated by the normal distribution

Test statistic for one sample Z proportions test

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

Where

- z is the test statistic
- $\hat{p}$ is observed proportion
- P_0 is the hypothesized probability
- n is the sample size

Point estimate

We estimate the proportion, p, as:

$$\hat{p} = \frac{x}{n}$$

Where, x is the number in the sample who has the trait or outcome of interest, and n is the size of the sample.

Confidence Intervals

We can calculate confidence intervals for the sample proportion. The upper and lower limits of the confidence interval are given by:

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Test Procedure

In a sample of large size n, we may examine whether the sample would have come from a population having a specified proportion $P=P_o$. For testing we may proceed as follows:

1. Null Hypothesis (H_o)

H_o: The given sample would have come from a population with specified proportion. i.e., $P = P_o$

2. Alternative Hypothesis (H_1)

H_1: The given sample may not be from a population with specified proportion.

$P \neq P_o$ (Two Sided)

$P > P_o$(One sided-right sided)

$P < P_o$(One sided-left sided)

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

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

It follows a standard normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Expected vale or critical value

In case of test statistic Z, the expected value is

$Z_e =$ 1.96 at 5% level, 2.58 at 1% level → Two tailed test

$Z_e =$ 1.65 at 5% level, 2.33 at 1% level → One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we reject the Null Hypothesis H_o otherwise accept it.

Example

A researcher claims that Republican Party will win in the next Senate elections, especially in Florida State. Statistical data reported that 23% voted for Republican Party in the last election. To test the claim a researcher surveyed 80 people and found 22 said they voted for Republican Party in the last election. Is there enough evidence at α=0.05 to support this claim?

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic
- To determine the critical value (from the critical value table)
- Then, define the rejection criteria
- Finally, interpret the result. In the event that the test statistic falls in the critical region, reject the null hypothesis

Calculation
Hypothesis

- H_0: The proportion of the People voted for Republican Party in the last election is 23%. ($p_0 = 0.23$).
- H_1: The proportion of the People voted for Republican Party in the last election is not 23%. ($p_0 \neq 0.23$).

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, $x = 22, n = 80 \text{ and } p_0 = 23\% \text{ or } 0.23$

Under the null hypothesis, the test statistic is

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

$$\hat{p} = \frac{x}{n} = \frac{22}{80} = 0.275$$

$$z = \frac{0.275 - 0.23}{\sqrt{0.23 * 0.77/80}}$$

$$= 0.045 / 0.047 = 0.957$$

Confidence Interval

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$UL = 0.275 + 1.96\sqrt{\frac{0.275(1-0.275)}{80}}$$

$$= 0.275 + 1.96 \times 0.0499$$

$$= 0.275 + 0.097804$$

$$=0.3728$$

$$LL = 0.275 - 1.96\sqrt{\frac{0.275(1-0.275)}{80}}$$

$= 0.275 - 1.96 \times 0.0499$

$= 0.275 - 0.097804$

$= 0.1772$

Table Value

The Critical Region = 1.96

Table Value = 0.9750

Conclusion

Since the calculated value is less than the table value (0.9570 < 0.9750). We accept the null hypothesis. Hence, we conclude that the People voted for Republican Party in the last election is 23%. The 95% confidence interval is 0.1772 to 0.3728.

Two Population Proportion

Two sample Z tests of proportions is the test to determine whether the two populations differ significantly on specific characteristics. In other words, compare the proportion of two different populations that have some single characteristic. It calculates the range of values that is likely to include the difference between the population proportions.

Assumptions

- The data are simple random values from both the populations.
- Both populations follow a binomial distribution.
- Samples are independent of each other.
- Test results are accurate when np and np(1-p) are greater than 5.

Test for Two Population Proportion

Given two sets of sample data of large size n_1 and n_2 from attributes. We may examine whether the two samples come from the populations having the same proportion. We may proceed as follows:

1. Null Hypothesis (H_o)

H_o: The given two samples would have come from a population having the same proportion. i.e., $P_1=P_2$.

2. Alternative Hypothesis (H_1)

H_1 : The given two sample may not be from a population with specified proportion.

$P_1 \neq P_2$ (Two Sided)

$P_1 > P_2$ (One sided-right sided)

$P_1 < P_2$ (One sided-left sided)

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$Z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where

$$p_0 = \frac{X_1 + X_2}{n_1 + n_2}$$

- z is test statistic
- $\hat{p}_1$ and $\hat{p}_2$ are observed proportion of events in the two samples
- n_1 and n_2 are sample sizes
- X_1 and X_2 are number of trails

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Expected vale or critical value

In case of test statistic Z, the expected value is

$Z_e =$ 1.96 at 5% level, 2.58 at 1% level → Two tailed test

$Z_e =$ 1.65 at 5% level, 2.33 at 1% level → One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the difference between two population proportions $p_1 - p_2$:

$$UL = (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$LL = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Example

A vice principal wants to see if there is a difference between the number of students who are late for the first class of the day versus the class of students after lunch. To test their claim that there is a difference in the proportion of late students between first and lunch classes, the vice-principal randomly selects 200 students from first class and records if they are late, then randomly selects 200 students in their class after lunch and records if they are late. 13 students are late for first class and 16 students are late after lunch. At the 0.05 level of significance, can a difference be concluded? Find the 95% confidence interval for the difference in the proportion of late students in their first class and the proportion that are late to their class after lunch.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic
- To determine the critical value (from the critical value table)
- Then, define the rejection criteria
- Finally, interpret the result.

Calculation

Hypothesis

H_0: There is a difference between the proportions of late students. (i.e., $p_1 \neq p_2$)

H_1: There is no difference between the proportions of late students. (i.e., $p_1 = p_2$)

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, $x_1 = 13$, $x_2 = 16$, $n_1 = 200$ and $n_2 = 200$

Under the null hypothesis, the test statistic is

$$\widehat{p_1} = \frac{x_1}{n_1} = \frac{13}{200} = 0.065$$

$$\widehat{p_2} = \frac{x_2}{n_2} = \frac{16}{200} = 0.08$$

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2} = \frac{13 + 16}{200 + 200} = 0.0725$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{p_0(1 - p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{0.065 - 0.08}{\sqrt{0.0725\ (1 - 0.0725)\left(\frac{1}{200} + \frac{1}{200}\right)}}$$

$$Z = \frac{-0.015}{\sqrt{0.0725\ (0.9275)(0.005 + 0.005)}}$$

$$Z = \frac{-0.015}{\sqrt{0.0007}}$$

$$Z = \frac{-0.015}{0.0259}$$

$$Z = |-0.5786|$$

$$Z = 0.5786$$

Confidence Interval

A 95% confidence interval for the difference between two population proportions $p_1 - p_2$:

$$UL = (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$UL = (0.065 - 0.08) + 1.96\sqrt{\frac{0.065 \times 0.935}{200} + \frac{0.08 \times 0.92}{200}}$$

$$= (-0.015) + 1.96\sqrt{0.0003 + 0.0004}$$

$$= (-0.015) + 0.0519$$

$$= 0.0369$$

$$LL = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$LL = (0.065 - 0.08) - 1.96\sqrt{\frac{0.065 \times 0.935}{200} + \frac{0.08 \times 0.92}{200}}$$

$$= (-0.015) - 1.96\sqrt{0.0003 + 0.0004}$$

$$= (-0.015) - 0.0519$$

$$= -0.0669$$

The 95% confidence interval is -0.0669 to 0.369.

Table Value

The Critical Region = 1.96

Table Value = 0.9750

Conclusion

Since the calculated value is less than the table value (0.5792 < 0.9750). We accept the null hypothesis. Hence, we conclude that there is difference between the proportions of late students. The 95% confidence interval is -0.0669 to 0.369.

Z – Test:

A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known. The null hypothesis of a z test can be rejected if the z test statistic is statistically significant when compared with the critical value.

Sampling from variable

In sampling for variables, the tests are as follows

- One sample Z test
- Two sample Z test

One sample Z test

The one-sample z-test is used to test whether the mean of a population is greater than, less than, or not equal to a specific value. Because the standard normal distribution is used to calculate critical values for the test, this test is often called the one-sample z-test. The z-test assumes that the population standard deviation is known.

One-Sample Z-Test Assumptions

The assumptions of the one-sample z-test are:

- The data are continuous (not discrete).
- The data follow the normal probability distribution.
- The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.
- The population standard deviation is known.

Formula for One sample Z test

A one-sample z test is used to check if there is a difference between the sample mean and the population mean when the population standard deviation is known. The formula for the z test statistic is given as follows:

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

Where,

- $\bar{x}$ is the sample mean
- μ is the population mean
- σ is the population standard deviation
- n is the sample size.

Procedure for One sample Z Test

A one-sample z-test is a statistical test used to compare the mean of a sample to a known population mean. We may proceed as follows:

1. Null Hypothesis (H_o)

H_o: There is no difference between the mean of the sample and the known population mean. i.e., $\mu = \mu_0$.

2. Alternative Hypothesis (H_1)

H_1: There is no difference between the mean of the sample and the known population mean. i.e. $\mu \neq \mu_0$.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

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

Where,

- $\bar{x}$ is the sample mean
- μ is the population mean
- σ is the population standard deviation
- n is the sample size.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Expected vale or critical value

In case of test statistic Z, the expected value is

$Z_e =$ 1.96 at 5% level, 2.58 at 1% level → Two tailed test

$Z_e =$ 1.65 at 5% level, 2.33 at 1% level → One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the difference between sample Mean and Population mean:

$$LL = \bar{x} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$UL = \bar{x} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Example

Twenty-five high school students complete a preparation program for taking the SAT test. Here are the SAT scores from the 25 students who completed the SAT prep program:

434, 694, 457, 534, 720, 400, 484, 478, 610, 641, 425, 636, 454, 514, 563, 370, 499, 640, 501, 625, 612, 471, 598, 509 and 531

Test the hypothesis that the Population Mean for SAT scores is 500 with a standard deviation of 100 at the 5% significance level.

Procedure

- Specify the null and alternative hypotheses.
- Select a sample from the population and calculate the sample mean and standard deviation.
- Calculate the test statistic
- Determine the critical value of the test statistic based on the significance level (alpha) of the test.
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no significant deviation between the sample mean and the population mean.

H_1: There is significant deviation between the sample mean and the population mean.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, $\mu = 500$, $\sigma = 100$ and $n = 25$

Under the null hypothesis, the test statistic is

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

Sample mean

$$\bar{x} = \frac{434 + 694 + 457 + 534 + 720 + 400 + 484 + 478 + 610 + 641 + 425 + 636 + 454 + 514 + 563 + 370 + 499 + 640 + 501 + 625 + 612 + 471 + 598 + 509 + 531}{25} = 536$$

$$z = \frac{536 - 500}{100 / \sqrt{25}}$$

$$z = \frac{36}{20}$$

$$z = 1.8$$

Confidence Interval

A 95% confidence interval for the difference between sample Mean and Population mean:

$$LL = \bar{x} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$LL = 536 - 1.96(20)$$

$$LL = 536 - 39.2 = 496.8$$

$$UL = \bar{x} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$UL = 536 + 1.96(20)$$

$$UL = 536 + 39.2 = 575.2$$

The 95% confidence interval is 496.8 to 575.2.

Table Value

The Critical Region = 1.96

Table Value = 0.9750

Conclusion

Since the calculated value is greater than the table value (1.8 > 0.9750). We reject the null hypothesis. Hence, we conclude that there is significant deviation between the sample mean and the population mean. The 95% confidence interval is 496.8 to 575.2.

Two sample Z Test

Two-sample Z-test for means is a statistical hypothesis testing technique that compares two independent samples to determine whether the means of the populations that generated them are different or not. It relies on the assumption that the populations have normal distributions, known population variances or equal variances, and that the samples are randomly and independently drawn from the respective populations. This test is used when the standard deviations (σ) of the two populations are known.

Assumption

- The data from each population are continuous (not discrete).
- Each sample is a simple random sample from the population of interest.
- The data in each population is approximately normally distributed.
- The population standard deviations are known.

Formula for Two sample Z Test

The following is the formula for **z-statistics** for two-sample z-test for means **given the population standard deviation is known**.

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Where,

- $\bar{x}_1$ is the mean of the first sample
- $\bar{x}_2$ is the mean of the second sample
- μ_1 is the mean of the first population
- μ_2 is the mean of the second population
- $(\mu_1 - \mu_2)$ is **hypothesized difference** between the population means
- σ_1 is the standard deviation of the first population
- σ_2 is the standard deviation of the second population
- n_1 is the number of the data points in the first sample
- n_2 is the number of the data points in the second sample

Test Procedure for Two sample Z Test

Two sample Z test compares the means of samples of independent groups taken from a normal population. We may proceed as follows:

1. Null Hypothesis (H_o)

H_o: The two-population means are equal mean. i.e., $\mu_1 = \mu_2$.

2. Alternative Hypothesis (H_1)

H_1: The two-population means are not equal mean. i.e., $\mu_1 \neq \mu_2$.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Where,

- $\bar{x}_1$ is the mean of the first sample
- $\bar{x}_2$ is the mean of the second sample
- μ_1 is the mean of the first population
- μ_2 is the mean of the second population
- $(\mu_1 - \mu_2)$ is **hypothesized difference** between the population means
- σ_1 is the standard deviation of the first population
- σ_2 is the standard deviation of the second population
- n_1 is the number of the data points in the first sample
- n_2 is the number of the data points in the second sample

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Expected vale or critical value

In case of test statistic Z, the expected value is

$Z_e =$ 1.96 at 5% level, 2.58 at 1% level $\longrightarrow$ Two tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the two sample Z-test is:

$$LL = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] - Z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$UL = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] + Z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Example

Glenbrook sports center is planning to compare the ages from a random sample of male and female swimmers in their coaching center. The swimming coordinator collected 62 female swimmers' data, and the mean age is 23.1 with a standard deviation of 3.5. Similarly, collected 46 male swimmers' data and the mean average is 19.2 with a standard deviation of 4.8. Assume the population follows a standard normal distribution. At 5% significance level, test whether there is a significant difference in age between the sexes?

Procedure

- Specify the null and alternative hypotheses.
- Select a sample from the population and calculate the sample mean and standard deviation.
- Calculate the test statistic

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

- Determine the critical value of the test statistic based on the significance level (alpha) of the test.
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no significant difference between the age and the sexes.

H_1: There is significant difference between the age and the sexes.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, $\bar{x}_1 = 23.1$, $\bar{x}_2 = 19.2$, $\sigma_1 = 3.5$, $\sigma_2 = 4.8$, $n_1 = 62$ and $n_2 = 46$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{(23.1 - 19.2) - (0)}{\sqrt{\frac{3.5^2}{62} + \frac{4.8^2}{46}}} = 4.66$$

Confidence Interval

A 95% confidence interval for the difference between sample Mean and Population mean:

$$LL = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] - Z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$= (23.1 - 19.2) - 1.96\left(\sqrt{\frac{12.25}{62} + \frac{23.04}{46}}\right)$$

$$= (3.9) - 1.96\left(\sqrt{0.1976 + 0.5009}\right)$$

$$= 3.9 - 1.6381$$

$$= 2.2619$$

$$UL = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] + Z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$= (23.1 - 19.2) + 1.96\left(\sqrt{\frac{12.25}{62} + \frac{23.04}{46}}\right)$$

$$= (3.9) + 1.96\left(\sqrt{0.1976 + 0.5009}\right)$$

$$= 3.9 + 1.6381$$

$$= 5.5381$$

The 95% confidence interval is 2.2619 to 5.5381.

Table Value

The Critical Region = 1.96

Table Value = 0.9750

Conclusion

Since the calculated value is greater than the table value (4.66 > 0.9750). We reject the null hypothesis. Hence, we conclude that there is significant difference between the age and the sexes. The 95% confidence interval is 2.2619 to 5.5381.

Small Sample Test

One-Sample T-Test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

One-Sample T-Test Assumptions

The assumptions of the one-sample t-test are:

- The data are continuous (not discrete).
- The data follow the normal probability distribution.
- The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Formula for Single mean test

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where,

$$\mathbf{S} = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]$$

- $\bar{x}$ is the sample mean
- μ is the population mean
- s is the sample standard deviation
- n is the sample size.

T-test for single Mean

In a sample of small size n, we examine whether the sample would have come from a Population having a specified mean

1. Null Hypothesis (Ho)

H_o: There is no significance difference between the sample mean. i.e., $\mu=\mu_o$

2. Alternative Hypothesis (H1)

H_1: There is significance difference between the sample mean. i.e., $\mu \neq \mu_o$ or $\mu > \mu_o$ or $\mu < \mu_o$

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where,

$$S = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]$$

- $\bar{x}$ is the sample mean
- μ is the population mean
- s is the sample standard deviation
- n is the sample size.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the t-score in a t-test depend on the degrees of freedom and the significance level of the test.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the difference between sample Mean and Population mean:

$$LL = \bar{x} - t_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$UL = \bar{x} + t_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

Example

A random sample of 10 boys had the following I.Q.'s: 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Do these data support the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: The data are consistent with the assumption of a mean I.Q. of 100 in the population. i.e., $\mu = 100$.

H_1: The data are not consistent with the assumption of a mean I.Q. of 100 in the population. i.e., $\mu \neq 100$.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Calculate mean and SD from the sample values of I.Q. is,

X	$\bar{x} = x - \bar{x}$	$(x - \bar{x})^2$
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
972		**1833.60**

Here, $n = 10$, $\bar{x} = 972/10 = 97.2$ and $S = 1833.60/9 = 14.27$

Under the null hypothesis, the test statistic is

$$t = \frac{97.2 - 100}{14.27 / \sqrt{10}}$$

$$t = \frac{2.8}{4.514}$$

$$= 0.62$$

Confidence Interval

A 95% confidence interval for the difference between sample Mean:

$$LL = \bar{x} - t_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$LL = 97.2 - 2.262(4.514)$$

$$LL = 86.99$$

$$UL = \bar{x} + t_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$UL = 97.2 + 2.262(4.514)$$

$$UL = 107.41$$

The 95% confidence interval is 86.99 to 107.41.

Table Value

The degree of freedom = n – 1

= 10 – 1 = 9df

Table Value = 2.262

Conclusion

Since the calculated value is less than the table value (0.62 < 2.262). We do not reject the null hypothesis. Hence, we conclude that the data are consistent with the assumption of a mean I.Q. of 100 in the population. The 95% confidence interval is 86.99 to 107.41.

Independent Sample Mean Test

The independent samples t-test is used to compare two sample means from unrelated groups. This means that there are different people providing scores for each group. The purpose of this test is to determine if the samples are different from each other.

Assumption

- The observations in one sample should be independent of the observations in the other sample.
- The data should be approximately normally distributed.
- The two samples should have approximately the same variance. If this assumption is not met, you should instead perform Welch's t-test.
- The data in both samples was obtained using a random sampling method.

Formula for Independent sample t-test

The Independent sample t-test formula is given by,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-1}$$

Where,

$$s^2 = \frac{\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right]}{n_1 + n_2 - 1}$$

- $\bar{x}$ is the mean of the first sample
- $\bar{y}$ is the mean of the second sample
- s^2 is the sample standard deviation
- n_1 is the first sample size.
- n_2 is the second sample size.

Test Procedure for Independent sample t-test

The two-sample *t*-test is a method used to test whether the unknown population means of two groups are equal or not.

1. Null Hypothesis (Ho)

H_o: The sample mean from Group 1 is not different from the sample mean from Group 2.

2. Alternative Hypothesis (H_1)

H_1: The sample mean from Group 1 is significantly different from the sample mean from Group 2.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Where,

$$s^2 = \frac{\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right]}{n_1 + n_2 - 1}$$

- $\bar{x}$ is the mean of the first sample
- $\bar{y}$ is the mean of the second sample
- s^2 is the sample standard deviation
- n_1 is the first sample size.
- n_2 is the second sample size.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the t-score in a t-test depend on the degrees of freedom ($n_1 + n_2 - 2$) and the significance level of the test.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the two sample Z-test is:

$$LL = \left[(\bar{x} - \bar{y})\right] - t_{0.05}\left(\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

$$UL = \left[(\bar{x} - \bar{y})\right] + t_{0.05}\left(\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Example

Below are given the gain in weights of pig fed on two diets A and B. gain in weight using:

Diet A: 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B: 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test if the two diets differ significantly regards their effect on increase in weight.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha,
- Compute the test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no significant difference between the mean increase in weight diet A and diet B. i.e., $\mu_x = \mu_y$.

H_1: There is significant difference between the mean increase in weight diet A and diet B. i.e., $\mu_x \neq \mu_y$.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Under the null hypothesis (H_o), the test statistic is,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Here, $n_1 = 12$, and $n_2 = 15$

Calculate mean and SD from the sample values of diet A and diet B,

x	**x-$\bar{x}$**	**(x- $\bar{x}$)^2**	**y**	$y-\bar{y}$	$(y-\bar{y})^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	144
25	3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
336	**0**	**380**	**450**	**0**	**1410**

$$\bar{x} = \frac{336}{12} = 28$$

$$\bar{y} = \frac{450}{15} = 30$$

$$s^2 = \frac{\left[\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2\right]}{n_1 + n_2 - 2}$$

$$s^2 = \frac{380 + 1410}{12 + 15 - 2} = 71.6$$

$$t = \frac{28 - 30}{\sqrt{71.6\left(\frac{1}{12} + \frac{1}{15}\right)}}$$

$$t = \frac{-2}{\sqrt{10.74}}$$

$$t = -0.609$$

$$t = |-0.609|$$

$$t = 0.609$$

Confidence Interval

A 95% confidence interval for the difference between sample Mean:

$$LL = [(\bar{x} - \bar{y})] - t_{0.05}\left(\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

$$LL = [(28 - 30)] - 2.06\left(\sqrt{71.6\left(\frac{1}{12} + \frac{1}{15}\right)}\right)$$

$$LL = -2 - 2.06\sqrt{10.74}$$

$$LL = -2 - 6.75$$

$$LL = -8.75$$

$$UL = [(\bar{x} - \bar{y})] + t_{0.05}\left(\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

$$LL = [(28 - 30)] + 2.06\left(\sqrt{71.6\left(\frac{1}{12} + \frac{1}{15}\right)}\right)$$

$$LL = -2 + 2.06\sqrt{10.74}$$

$$LL = -2 + 6.75$$

$$LL = 4.75$$

The 95% confidence interval is -8.75 to 4.75.

Table Value

The degree of freedom = $n_1 + n_2 - 2$

$= 12 + 15 - 2 = 25$ d.f.

Table Value = 2.06

Conclusion

Since the calculated value is less than the table value (0.609 < 2.06). We do not reject the null hypothesis. Hence, we conclude that there is no significant difference between the mean increase in weight diet A and diet B. The 95% confidence interval is -8.75 to 4.75.

Paired T-test

The dependent samples t-test is used to compare the sample means from two *related* groups. This means that the scores for both groups being compared come from the same people. The purpose of this test is to determine if there is a change from one measurement (group) to the other.

Assumption

- The dependent variable for data should be continuous.
- The observation must be independent of each other that is a random sample of data should be done.
- The paired t-test can be only implemented to related sample or groups. The subject must be the same for each sample or group.
- The dependent variable data used in a paired t-test must be free from outliers.
- The dependent variable should be normally distributed.

Formula for Paired T-test

The formula of the paired t-test is given by,

$$t = \frac{\bar{d}}{SE(d)}$$

Where,

$$\bar{d} = \frac{\sum d_i}{n}, \qquad \sigma_{diff} = \sqrt{\frac{\sum d_i^2 - \bar{d} \times n}{n-1}} \qquad \text{and} \qquad SE(d) = \frac{\sigma_{diff}}{\sqrt{n}}$$

- d_i is the difference of two samples
- $\bar{d}$ is the mean difference of the two samples
- σ is the Standard Deviation of the difference
- SE(d) is the standard error of the two samples
- n is the sample size

Test procedure for Paired T-test

The paired t-test is a method used to test whether the mean difference between pairs of measurements is zero or not.

1. Null Hypothesis (Ho)

H_o: The mean difference between the two groups is not different from 0.

2. Alternative Hypothesis (H_1)

H_1: The mean difference between the two groups is different from 0.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t = \frac{\bar{d}}{SE(d)}$$

Where,

$$\bar{d} = \frac{\sum d_i}{n}, \qquad \sigma_{diff} = \sqrt{\frac{\sum d_i^2 - \bar{d} \times n}{n-1}} \qquad \text{and} \qquad SE(d) = \frac{\sigma_{diff}}{\sqrt{n}}$$

- d_i is the difference of two samples
- $\bar{d}$ is the mean difference of the two samples
- σ is the Standard Deviation of the difference
- SE(d) is the standard error of the two samples
- n is the sample size

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the t-score in a t-test depend on the degrees of freedom and the significance level of the test.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the two sample Z-test is:

$$LL = \bar{d} - t_{0.05}(SE(d))$$
$$UL = \bar{d} + t_{0.05}(SE(d))$$

Example

Memory capacity of 9 students was tested before and after training. State at 5 per cent level of significance whether the training was effective from the following scores:

Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

Use paired t-test.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$t = \frac{\bar{d}}{SE(d)}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: The mean of difference is zero.

H_1: The mean of difference is not zero.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Take the score before training as X and the score after training as Y

Calculate mean and standard deviation of differences,

X	Y	d = X - Y	d^2
10	12	-2	4
15	17	-2	4
9	8	1	1
3	5	-2	4
7	6	1	1
12	11	1	1
16	18	-2	4
17	20	-3	9
4	3	1	1
		-7	**29**

Here, n = 9, d = -7 and $d^2 = 29$

Mean of difference is $\bar{d} = \frac{-7}{9} = -0.778$

And Standard Deviation of difference is

$$\sigma_{diff} = \sqrt{\frac{\sum d_i^2 - \bar{d} \times n}{n-1}}$$

$$\sigma_{diff} = \sqrt{\frac{29 - (-0.778)^2 \times 9}{9-1}} = \sqrt{2.944} = 1.715$$

$$SE(d) = \frac{\sigma_{diff}}{\sqrt{n}} \qquad SE(d) = \frac{1.715}{\sqrt{9}} = 0.572$$

Under the null hypothesis, the test statistic is

$$t = \frac{\bar{d}}{SE(d)}$$

$$t = \frac{-0.778}{0.572} = -1.361$$

$$t = \left|-1.361\right| = 1.361$$

Confidence Interval

A 95% confidence interval is

$$LL = \bar{d} - t_{0.05}(SE(d))$$
$$LL = -0.778 - 2.306(0.572)$$
$$LL = -2.097$$
$$UL = \bar{d} + t_{0.05}(SE(d))$$
$$UL = -0.778 + 2.306(0.572)$$
$$UL = 0.541$$

The 95% confidence interval is -2.097 to 0.541.

Table Value

The degree of freedom = n – 1

= 9 – 1 = 8df

Table Value = 2.306

Conclusion

Since the calculated value is less than the table value (1.361 < 2.306). We do not reject the null hypothesis. Hence, we conclude that the mean of difference is zero. The 95% confidence interval is -2.097 to 0.541.

Significance of Observed Correlation Coefficient

The correlation coefficient *r*, tells us about the strength and direction of the linear relationship between X_1 and X_2.

The sample data are used to compute *r*, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, *r*, is our estimate of the unknown population correlation coefficient.

- ρ = population correlation coefficient (unknown)
- *r* = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient *r* and the sample size *n*.

Assumption

- That both variables are plausibly normally distributed.
- That there is a linear relationship between them.
- The null hypothesis is that there is no association between them.

Formula for Significance of Observed Correlation Coefficient

The Significance of Observed Correlation Coefficient formula is given by,

$$t_c = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

OR

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where,

- *r* is the sample correlation coefficient
- n is the sample size

Test Procedure for Significance of Observed Correlation Coefficient

1. Null Hypothesis (Ho)

H_0: The population correlation coefficient is not significantly different from zero.

(or)

There is no a significant linear relationship between X_1 and X_2 in the population.

2. Alternative Hypothesis (H_1)

H_1: The population correlation coefficient is not significantly different from zero.

(or)

There is no a significant linear relationship between X_1 and X_2 in the population.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t_c = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

OR

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Degrees of freedom = n - 2

Where,

- *r* is the sample correlation coefficient
- n is the sample size

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the t-score in a t-test depend on the degrees of freedom and the significance level of the test.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the Significance of Observed Correlation Coefficient is:

$$LL = r\sqrt{n-2} - r_{0.05}\left(\sqrt{1-r^2}\right)$$

$$UL = r\sqrt{n-2} + r_{0.05}\left(\sqrt{1-r^2}\right)$$

Example

a) A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0·6. Is this significant of correlation in the population?

b) Find the least value of r in a sample of 18 pairs of observations from a bi-variate normal population, significant at 5% level of significance.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

a. Hypothesis

H_0: The observed sample correlation coefficient is not significant of any correlation in the population.

H_1: The observed sample correlation coefficient is significant of any correlation in the population.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, n = 27 and r = 0.6

Under the null hypothesis, the test statistic is

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t_c = \frac{0.6\sqrt{27-2}}{\sqrt{1-0.36}}$$

$$= \frac{3}{\sqrt{0.64}} = 3.75$$

Confidence Interval

A 95% confidence interval is

$$LL = r\sqrt{n-2} - t_{0.05}\left(\sqrt{1-r^2}\right)$$

$$LL = 0.6\sqrt{25} - 2.06\left(\sqrt{1-0.36}\right)$$

$$LL = 3 - 1.648$$

$$LL = 1.352$$

$$UL = r\sqrt{n-2} + t_{0.05}\left(\sqrt{1-r^2}\right)$$

$$UL = 0.6\sqrt{25} + 2.06\left(\sqrt{1-0.36}\right)$$

$$UL = 3 + 1.648$$

$$UL = 4.648$$

The 95% confidence interval is 1.352 to 4.648.

Table Value

The degree of freedom = n – 2

= 27 – 2 = 25df

Table Value = 2.060

Conclusion

Since the calculated value is greater than the table value (3.75 > 2.060). We reject the null hypothesis. Hence, we conclude that the observed sample correlation coefficient is significant of correlation in the population. The 95% confidence interval is 1.352 to 4.648.

b. Calculation

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, n = 18 and table value = (18 – 2)df = 16df is 2.12

Under the null hypothesis, the test statistic is

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$t_c = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| > t_{0.05}$$

$$t_c = \left| \frac{r\sqrt{18-2}}{\sqrt{1-r^2}} \right| > 2.12$$

$$16r^2 > (2.12)^2(1-r^2)$$

$$16r^2 > 4.4944 - 4.4944r^2$$

$$20.4944r^2 > 4.4944$$

$$r^2 > \frac{4.4944}{20.4944} = 0.2192$$

$$r > 0.4682$$

Sample correlation coefficient r is 0.4682.

Significance of Observed Regression Coefficient

Here the problem is to test if a random sample *(x_i, y_i), (i = 1, 2, . . . , n)* has been drawn from a bivariate normal population in which regression coefficient of Y on X is β.

The line of regression of Y on X is:

$$Y - \bar{y} = b(X - \bar{x}), \qquad b = \frac{\mu_{11}}{\sigma_X^2} \qquad \text{---------- (1)}$$

The estimate of Y for a given value x_i of X as given by line (1) is:

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x})$$

Under the null hypothesis, H_0 that the population regression coefficient is β, the test statistic,

$$t = (b - \beta)\left[\frac{(n-2)\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \hat{y}_i)^2} \right]^{1/2}$$

Assumption

- The population is bivariate normal with regression coefficient of Y on X is β.
- β is unknown.
- A random sample of size n is drawn from bivariate normal population and its regression coefficient of Y on X is b.

Formula

The significance of regression coefficient formula is,

$$t = \frac{b - E(b)}{S.E(b)}$$

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim t_{(n-2)}$$

Where,

- *x*, *y* are the sample observation
- n is sample size

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - (\bar{x})^2}$$

$$\hat{y} = \hat{a} + \hat{b}x_i$$

$$a = \bar{y} - b\bar{x}$$

Test procedure for significance of regression coefficient

1. Null Hypothesis (Ho)

H_0: There is no significant difference between sample regression coefficient and population regression coefficient.

2. Alternative Hypothesis (H_1)

H_1: There is significant difference between sample regression coefficient and population regression coefficient

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t = \frac{b - E(b)}{S.E(b)}$$

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim t_{(n-2)}$$

Where,

- *x*, *y* are the sample observation
- n is sample size

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - (\bar{x})^2}$$

$$\hat{y} = \hat{a} + \hat{b}x_i$$

$$a = \bar{y} - b\bar{x}$$

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

From t table, we can find $t_{\alpha/2,\, n-2}$, t from t table for n-2 degrees of freedom.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the Significance of Observed regression Coefficient is:

$$LL = b - E(b) - t_{0.05}\big(SE(b)\big)$$

$$UL = b - E(b) + t_{0.05}\big(SE(b)\big)$$

Example

Test the significance of regression coefficient by X if the following are the values of sample drawn from bivariate normal population.

X	1	2	3	4	5	6
Y	10	12	14	16	14	15

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim t_{(n-2)}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: The regression equation is linear. i.e., $\beta = 0$.

H_1: The regression equation is non-linear. i.e., $\beta \neq 0$.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Under the null hypothesis, the test statistic is

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim t_{(n-2)}$$

Assume $\beta = 0$.

x_i	y_i	$x_i y_i$	$x_i - \bar{x}$	x_i^2	$\hat{y}_i$	$(y_i - \bar{y}_i)^2$	$(x_i - \bar{x})^2$
1	10	10	-2.5	1	11.143	1.3064	6.25
2	12	24	-1.5	4	121.0858	0.0074	2.25
3	14	42	-0.5	9	13.0286	0.9436	0.25
4	16	64	0.5	16	13.9714	4.1152	0.25
5	14	70	1.5	25	14.9142	0.8358	2.25
6	15	90	2.5	36	15.857	0.7344	6.25

$$\bar{x} = 3.5$$

$$\bar{y} = 13.5$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - (\bar{x})^2}$$

$$b = \frac{50 - 47.25}{15.1667 - 12.25}$$

$$b = 0.9428$$

$$a = \bar{y} - b\bar{x}$$

$$a = 13.5 - (0.9428)(3.5)$$

$$a = 10.2002$$

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}}$$

$$t = \frac{0.9428 - 0}{\sqrt{\frac{7.9428}{4(17.5)}}}$$

$$t_{cal} = 2.7985$$

Confidence Interval

A 95% confidence interval is

$$LL = b - E(b) - t_{0.05}\left(SE(b)\right)$$

$$LL = (0.943 - 0) - 2.776\left(\sqrt{\frac{7.943}{4(17.5)}}\right)$$

$$LL = 0.943 - 0.935$$

$$LL = 0.008$$

$$UL = b - E(b) + t_{0.05}\left(SE(b)\right)$$

$$UL = (0.943 - 0) + 2.776\left(\sqrt{\frac{7.943}{4(17.5)}}\right)$$

$$UL = 0.943 + 0.935$$

$$UL = 1.878$$

The 95% confidence interval is 0.008 to 1.878.

Table Value

The degree of freedom = n – 2

= 6 – 2 = 4df

Table Value = 2.776

Conclusion

Since the calculated value is greater than the table value (2.799 > 2.776). We reject the null hypothesis. Hence, we conclude that the regression equation is non-linear. The 95% confidence interval is 0.008 to 1.878.

Test for Significance of Variance

The chi square (χ^2) distribution is used to determine if a population variance meets a specified standard or takes on a particular value.

Assumption

- The population is normal population mean μ and variance σ^2
- μ and σ^2 are unknown.
- A random sample of size n is drawn from normal population with mean μ and variance σ^2.

Formula

The chi-square test formula is given by,

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Where,

- σ_s^2 = variance of the sample
- σ_p^2 = variance of the population
- (n – 1) = degree of freedom
- n is the sample size.

Test Procedure for Significance of Variance

1. Null Hypothesis (Ho)

H_0: There is no a significant association between population variance and the sample variance.

2. Alternative Hypothesis (H_1)

H_1: There is a significant association between population variance and the sample variance.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Where,

- σ_s^2 = variance of the sample
- σ_p^2 = variance of the population
- (n – 1) = degree of freedom
- n is the sample size.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the chi-square table depend on the degrees of freedom and the significance level of the test.

6. Inference

If the observed value of the test statistic Chi-square exceeds the table value of χ^2, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the chi-square test is:

$$LL = (n-1)\sigma_s^2 - \chi^2_{0.05}(\sigma_p^2)$$

$$UL = (n-1)\sigma_s^2 + \chi^2_{0.05}(\sigma_p^2)$$

Example

Weight of 10 students is as follows:

Sl. No.	1	2	3	4	5	6	7	8	9	10
Weight in kg	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent level of significance.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no a significant association between population variance and the sample variance.

H_1: There is a significant association between population variance and the sample variance.

.**Level of Significance**

$\alpha = 0.05$

Test Statistic:

Calculate sample variance

Sl. No.	X_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
1	38	-9	81
2	40	-7	49
3	45	-2	4
4	53	6	36
5	47	0	0
6	43	-4	16
7	55	8	64
8	48	1	1
9	52	5	25
10	49	2	4
	470		***280***

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{470}{10} = 47kgs$$

$$\sigma_s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} = \frac{280}{10-1} = 31.11$$

Under the null hypothesis, the test statistic is

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Here, σ_p^2 = 20 and n = 10

$$\chi^2 = \frac{31.11}{20}(10-1)$$

$$= 13.999$$

Confidence Interval

A 95% confidence interval is

$$LL = (n-1)\sigma_s^2 - \chi^2_{0.05}(\sigma_p^2)$$

$$LL = (10-1)(31.11) - (16.92 \times 20)$$

$$LL = 279.99 - 338.4$$

$$LL = -58.41$$

$$UL = (n-1)\sigma_s^2 + \chi^2_{0.05}(\sigma_p^2)$$

$$UL = (10-1)(31.11) + (19.92 \times 20)$$

$$UL = 279.99 + 338.4$$

$$UL = 618.39$$

The 95% confidence interval is -58.99 to 618.39.

Table Value

The degree of freedom = n – 1

= 10 – 1 = 9df

Table Value = 16.92

Conclusion

Since the calculated value is less than the table value (13.999 < 16.92). We do not reject the null hypothesis. Hence, we conclude that there is no a significant association between population variance and the sample variance. The 95% confidence interval is -58.99 to 618.39.

Variance of a Normal Population

The F Test for Equality of Variances between two groups is a statistical test used to compare the variances of two samples to determine whether they are equal.

Assumption

- The populations are normal
- Samples have been drawn randomly
- Observations are independent
- There is no measurement error

Formula

The formula for Variance of a Normal Population is,

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

Where,

$$\sigma_{s_1}^2 = \frac{\Sigma\left(X_{1i} - \overline{X}_1\right)^2}{(n_1 - 1)}$$

$$\sigma_{s_2}^2 = \frac{\Sigma\left(X_{2i} - \overline{X}_2\right)^2}{(n_2 - 1)}$$

$\sigma_{s_1}^2$ is sample variance of first sample

$\sigma_{s_2}^2$ is sample variance of Second sample

Test procedure for significance of Normal Population

1. Null Hypothesis (Ho)

H_0: The two random samples drawn from two normal populations.

2. Alternative Hypothesis (H_1)

H_1: The two random samples not drawn from two normal populations.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

Where,

$$\sigma_{s_1}^2 = \frac{\Sigma\left(X_{1i} - \overline{X}_1\right)^2}{(n_1 - 1)}$$

$$\sigma_{s_2}^2 = \frac{\Sigma\left(X_{2i} - \overline{X}_2\right)^2}{(n_2 - 1)}$$

$\sigma_{s_1}^2$ is sample variance of first sample

$\sigma_{s_2}^2$ is sample variance of Second sample

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

For two sided test, we find F_1 and F_2 from F table for F((n_1-1),(n_2-1)) degrees of freedom.

6. Inference

If the observed value of the test statistic F exceeds the table value of F, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the Significance of Observed regression Coefficient is:

$$LL = \sigma_{s_1}^2 - F_{0.05}\left(\sigma_{s_2}^2\right)$$

$$UL = \sigma_{s_1}^2 + F_{0.05}\left(\sigma_{s_2}^2\right)$$

Example

Two random samples drawn from two normal populations are:

Sample 1	20	16	26	27	23	22	18	24	25	19		
Sample 2	27	33	42	35	32	34	38	28	41	43	30	37

Test using variance ratio at 5 per cent level of significance whether the two populations have the same variances.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level

- Compute the test statistic

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no a significant difference between the two sample variance.

H_1: There is a significant difference between the two sample variance.

.**Level of Significance**

$\alpha = 0.05$

Test Statistic:

Calculate sample variance

X_{1i}	$(X_{1i} - \overline{X}_1)$	$(X_{1i} - \overline{X}_1)^2$	X_{2i}	$(X_{2i} - \overline{X}_2)$	$(X_{2i} - \overline{X}_2)^2$
20	-9	81	27	-9	81
16	-7	49	33	-7	49
26	-2	4	42	-2	4
27	6	36	35	6	36
23	0	0	32	0	0
22	-4	16	34	-4	16
18	8	64	38	8	64
24	1	1	28	1	1
25	5	25	41	5	25
19	2	4	43	2	4
			30		
			37		
220		***120***	***420***		***314***

$$\overline{X}_1 = \frac{\sum_{i=1}^{n} X_{1i}}{n_1} = \frac{220}{10} = 22$$

$$\overline{X}_2 = \frac{\sum_{i=1}^{n} X_{2i}}{n_2} = \frac{420}{12} = 35$$

$$\sigma_{s_1}^2 = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} = \frac{120}{10-1} = 13.33$$

$$\sigma_{s_2}^2 = \frac{\sum_{i=1}^{n_2}\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1} = \frac{314}{12-1} = 28.55$$

Under the null hypothesis, the test statistic is

$$F = \frac{\sigma_{s_2}^2}{\sigma_{s_1}^2} \qquad \left(\because \sigma_{s_2}^2 > \sigma_{s_1}^2\right)$$

$$F = \frac{28.55}{13.33}$$

$$= 2.14$$

Confidence Interval

A 95% confidence interval is

$$LL = \sigma_{s_2}^2 - F_{0.05}(\sigma_{s_1}^2)$$

$$LL = (28.55) - (2.90 \times 13.33)$$

$$LL = 28.55 - 38.66$$

$$LL = -10.11$$

$$UL = \sigma_{s_2}^2 + F_{0.05}(\sigma_{s_1}^2)$$

$$UL = (28.55) + (2.90 \times 13.33)$$

$$UL = 28.55 + 38.66$$

$$UL = 67.21$$

The 95% confidence interval is -10.11 to 67.21.

Table Value

The degrees of freedom in sample1 = $(n_1 - 1)$

$= 10 - 1 = 9df$

The degrees of freedom in sample2 = $(n_2 - 1)$

$= 12 - 1 = 11df$

As the variance of sample 2 is greater variance, hence

$v_1 = 9$; $v_2 = 11$

Table Value = 2.90

Conclusion

Since the calculated value is less than the table value (2.14 < 2.90). We do not reject the null hypothesis. Hence, we conclude that there is no a significant difference between the two sample variance. The 95% confidence interval is -10.11 to 67.21.

Pearson Correlation

The Pearson correlation coefficient represents the relationship between the two variables, measured on the same interval or ratio scale. It measures the strength of the relationship between the two continuous variables.

Assumption

- The two variables need to be using a continuous scale.
- The two variables of interest should have a linear relationship, which you can check with a scatter plot.
- There should be no spurious outliers.
- The variables should be normally or near-to-normally distributed.

Formula

The Pearson Correlation Coefficient formula is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where,

- r = Pearson Coefficient
- n= number of pairs of the stock
- $\sum xy$ = sum of products of the paired stocks
- $\sum x$ = sum of the x scores
- $\sum y$= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

Test Procedure for Significance of Variance

1. Null Hypothesis (Ho)

H_0: There is no a significant relationship between two or more variables.

2. Alternative Hypothesis (H_1)

H_1: There is a significant relationship between two or more variables.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where,

- r = Pearson Coefficient
- n= number of pairs of the stock
- $\sum$xy = sum of products of the paired stocks
- $\sum$x = sum of the x scores
- $\sum$y= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

The critical values for the Pearson correlation table depend on the degrees of freedom and the significance level of the test.

6. Inference

If the observed value of the test statistic r exceeds the table value of Pearson's correlation, we reject the Null Hypothesis H_o otherwise accept it.

Confidence Interval

A 95% confidence interval for the Pearson Correlation Coefficient is:

$$LL = Z_r - \frac{Z_{(1-\alpha/2)}}{\sqrt{n-3}}$$

$$UL = Z_r + \frac{Z_{(1-\alpha/2)}}{\sqrt{n-3}}$$

Where,

$$Z_r = \frac{\ln(1+r)/(1-r)}{2}$$

Example

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There is no a significant correlation between height of father and son.

H_1: There is a significant correlation between height of father and son.

Level of Significance

$\alpha = 0.05$

Test Statistic:

Here, n = 7, then

x	**y**	**x^2**	**y^2**	**xy**
64	66	4096	4356	4224
65	67	4225	4489	4355
66	65	4356	4225	4290
67	68	4489	4624	4556
68	70	4624	4900	4760
69	68	4761	4624	4692
70	72	4900	5184	5040
469	**476**	**31451**	**32402**	**31917**

Under the null hypothesis, the test statistic is

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

$$r = \frac{(7 \times 31917) - (469 \times 476)}{\sqrt{\left[(7 \times 31451) - (469)^2\right]\left[(7 \times 32402) - (476)^2\right]}}$$

$$r = \frac{175}{\sqrt{196 \times 238}} = \frac{175}{215.98}$$

$$r = 0.810$$

Confidence Interval

A 95% confidence interval is

$$Z_r = \frac{\ln(1+r)/(1-r)}{2}$$

$$Z_r = \frac{\ln(1+0.810)/(1-0.810)}{2}$$

$$Z_r = 1.56$$

$$LL = Z_r - \frac{Z_{(1-\alpha/2)}}{\sqrt{n-3}}$$

$$LL = 1.56 - \frac{1.96}{\sqrt{7-3}}$$

$$LL = 1.56 - 0.98$$

$$LL = 0.58$$

$$UL = Z_r + \frac{Z_{(1-\alpha/2)}}{\sqrt{n-3}}$$

$$UL = 1.56 + \frac{1.96}{\sqrt{7-3}}$$

$$LL = 1.56 + 0.98$$

$$LL = 2.54$$

The 95% confidence interval is 0.58 to 2.54.

Table Value

The degrees of freedom = (n – 2)

= 7 – 2 = 5df

Table Value = 0.754

Conclusion

Since the calculated value is greater than the table value (0.810 > 0.754). We reject the null hypothesis. Hence, we conclude that there is a significant difference between the height of father and son. The 95% confidence interval is 0.58 to 2.54.

Test for Significance of Partial Correlation Coefficient

The partial correlation coefficient is frequently used to measure the correlation of two variables after eliminating the effect of other variable(s) in a set of correlated variables. For example, it may be of interest to know the correlation between intelligence and weight of people after eliminating the effect of age.

Assumption

1. The population is multivariate normal with partial correlation coefficient ρ of order k.
2. A random sample is drawn from a population with the sample partial coefficient r of order k.

Formula

The formula for Significance of Partial Correlation Coefficient is,

$$t = \frac{r - E(r)}{SE(r)}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

Where,

n = Sample size

k = Order of partial correlation coefficient,

r = Partial correlation coefficient of the samples.

Test Procedure for Significance of Partial Correlation Coefficient

1. Null Hypothesis (Ho)

H_0: The population partial coefficient ρ of order k is not significant.

2. Alternative Hypothesis (H_1)

H_1: The population partial coefficient ρ of order k is significant.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

Where,

n = Sample size

k = Order of partial correlation coefficient,

r = Partial correlation coefficient of the samples.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

From t table, we can find $t_{\alpha/2,(n-k-2)}$, from t table for n-k-2 degrees of freedom.

6. Inference

If the observed value of the test statistic t exceeds the table value of t, we reject the Null Hypothesis H_o otherwise accept it.

Example

A sample of size 10 observation from trivariate normal population gave the partial correlation coefficient between first and second variable as 0.3247 is this significant at 5% level.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: The partial correlation coefficient of order 1 is not significant.

H_1: The partial correlation coefficient of order 1 is significant.

Level of Significance

$$\alpha = 0.05$$

Test Statistic:

Here, n = 10, k = 1 and r = 0.3247

Under the null hypothesis, the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

$$t = \frac{0.3247}{\sqrt{\frac{1-(0.3247)^2}{10-1-2}}} = \frac{0.3247}{\sqrt{0.1278}} = 0.9083$$

Table Value

The degrees of freedom = (n – k – 2)

= 10 – 1 – 2 = 7df

Table Value = 2.365

Conclusion

Since the calculated value is less than the table value (0.9083 < 2.365). We do not reject the null hypothesis. Hence, we conclude that the partial correlation coefficient of order 1 is not significant.

Test for Significance of Multiple Correlation Coefficient

Assumption

1. The population is multivariate normal.
2. A random sample of size n is drawn from the population and desired multiple correlation coefficient is obtained $r_{1.234\ldots\ k+1}$.
3. We want to test whether there exist multiple correlation in the population $R_{1.234\ k+1}=0$.

Formula

The formula for Significance of Multiple Correlation Coefficient is,

$$F = \frac{r^2}{1-r^2} . \frac{n-k-1}{k} \sim F_{(k,n-k-1)}$$

Where,

n = Sample size

k = Order of partial correlation coefficient,

r = Partial correlation coefficient of the samples.

Test Procedure for Significance of Multiple Correlation Coefficients

1. Null Hypothesis (Ho)

H_0: The multiple correlation in the population is zero.

2. Alternative Hypothesis (H_1)

H_1: The multiple correlation in the population is not zero.

3. Test statistic

Under the null hypothesis (H_0), the test statistic is

$$F = \frac{r^2}{1-r^2} . \frac{n-k-1}{k} \sim F_{(k,n-k-1)}$$

Where,

n = Sample size

k = Order of Multiple correlation coefficient

r = Partial correlation coefficient of the samples.

4. Level of Significance

The level of significance may be fixed at either 5% or 1%.

5. Critical value

From t table, we can find $F_{\alpha,(k, n-k-1)}$, from F table for (k, n-k-1) degrees of freedom.

6. Inference

If the observed value of the test statistic F exceeds the table value of F, we reject the Null Hypothesis H_o otherwise accept it.

Example

From a 5 variate normal population a random sample of size 20 is taken and multiple correlation coefficient $r_{1.2345}$ is found to be 0.27, tests at 5% level the existence of multiple correlation coefficients in the population.

Procedure

- State the null hypothesis and alternative hypothesis
- State alpha, in other words, determines the significance level
- Compute the test statistic

$$F = \frac{r^2}{1-r^2} \cdot \frac{n-k-1}{k} \sim F_{(k,n-k-1)}$$

- To determine the critical value
- Compare the calculated test statistic to the critical value.

Calculation

Hypothesis

H_0: There does not exist population multiple correlation coefficient.

H_1: There exist population multiple correlation coefficient.

Level of Significance

$$\alpha = 0.05$$

Test Statistic:

Here, n = 20, k = 20/5 = 4 and r = 0.27

Under the null hypothesis, the test statistic is

$$F = \frac{r^2}{1-r^2} \cdot \frac{n-k-1}{k}$$

$$F = \frac{(0.27)^2}{1-(0.27)^2} \cdot \frac{20-4-1}{4}$$

$$F = \frac{1.0935}{3.7084}$$

$$F = 0.2949$$

Table Value

The degrees of freedom = (n – k – 1)

= 20 – 4 – 1 = 15df

= (4, 15) df

Table Value = 3.06

Conclusion

Since the calculated value is less than the table value (0.2949 < 3.06). We do not reject the null hypothesis. Hence, we conclude that there does not exist population multiple correlation coefficient

Normality Test

A normality test determines whether a sample data has been drawn from a normally distributed population. It is generally performed to verify whether the data involved in the research have a normal distribution. Many statistical procedures such as correlation, regression, t-tests, and ANOVA, namely parametric tests, are based on the normal distribution of data.

Types of normality Test

Normal distribution can be tested either statistical tests or graphically. The most common analytical tests to check data for normal distribution are:

- Kolmogorov-Smirnov Test
- Shapiro-Wilk Test
- Anderson-Darling Test

For graphical,

- Histogram
- Q-Q plot

Q-Q stands for quantile-quantile plot, where they actually observed distribution is compared with the theoretically expected distribution.