**Statistical Modeling**

The science of statistics is the study of how to learn from data. It helps you collect the right data, perform the correct analysis, and effectively present the results with statistical knowledge. Statistical modeling is a key to making scientific discoveries, data-driven decisions, and predictions.

**Definition**

*Statistical modeling is the process of describing the connections between variables in a dataset using mathematical equations and statistical approaches. In statistical modeling, we use a collection of statistical methods to investigate the connections between variables and uncover patterns in data.*

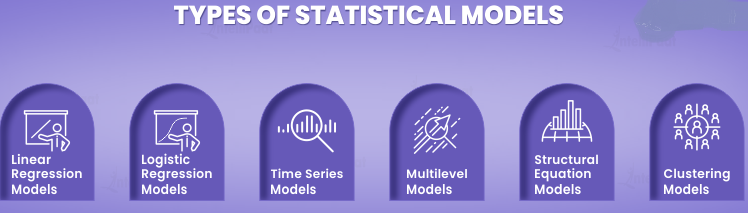
Predicting the number of people who will travel on a specific rail route is an example of statistical modeling. To develop a statistical model, we would collect data on the number of passengers who utilize the train route over time, as well as data on variables that might affect passenger counts, such as time of day, day of the week, and weather.

Then, using statistical approaches such as regression analysis, we can determine the correlations between these factors and the number of passengers utilizing the railway route. For example, we might discover that the number of passengers is larger during rush hour and on weekdays, and fewer when it is raining.

We can apply this data to build a statistical model that forecasts the number of people who would use the railway route depending on the time of day, day of the week, and weather conditions. This model can then be used to anticipate future passenger numbers and make resource allocation choices, such as adding additional trains during rush hour or giving specials during severe weather.

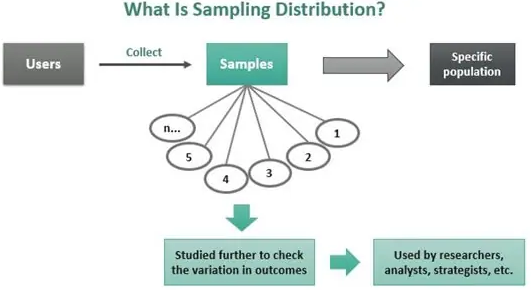
**Types of Statistical Models**

There are several statistical models, each designed to solve a specific research issue or data format. Here are a few common types of statistical models and their applications:



1. **Linear regression models**: These models are used to represent the connection between a continuous result variable and one or more predictor variables. For example, depending on a person’s height, age, and gender, a linear regression model may be used to estimate their weight.
2. **Logistic regression models**: Logistic regression models are used to represent the connection between a binary outcome variable (for example, yes/no) and one or more predictor variables. For example, depending on age, blood pressure, and cholesterol levels, a logistic regression model may be used to predict if a patient would have a heart attack.
3. **Time series models**: Time series models are used to model data that change over time, such as stock prices, weather trends, or monthly sales numbers. These types of models may be applied to data to find trends, seasonal patterns, and other forms of temporal correlations.
4. **Multilevel models**: These models are used to model data having a hierarchical structure, such as pupils in schools or patients in hospitals. Multilevel models can be used to investigate how individual-level and group-level factors impact outcomes, as well as to account for the fact that people in the same group may be more similar to each other than those in different groups.
5. **Structural equation models**: These types of models are used to represent complicated interactions between several variables. Structural equation models can be used to evaluate ideas regarding causal links between variables and to quantify their strength and direction.
6. **Clustering models**: Clustering models are used to bring together comparable observations based on their similarities in terms of features. Clustering algorithms can be used to uncover patterns in data that would be difficult to detect using other approaches.

**Sampling Distribution**

****

Sampling distribution in statistics represents the probability of varied outcomes when a study is conducted. It is also known as finite-sample distribution. In the process, users collect samples randomly but from one chosen population.

However, the data collected is not based on the population but on samples collected from a specific population to be studied. Thus, a sample becomes a subset of the chosen population.

*With sampling distribution*, the samples are studied to determine the probability of various outcomes occurring with respect to certain events.

**Types**

The finite-sample distribution can be expressed in various forms. Here is a list of some of its types:

1. Sampling Distribution of Mean
2. Sampling Distribution of Proportion
3. T-Distribution
4. **Sampling Distribution of the Mean Example**

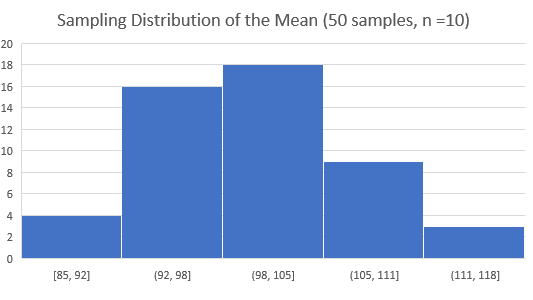
**Example 1:**

Imagine you draw a random sample of 10 apples. Then you calculate the mean of that sample as 103 grams. That’s one sample mean from one sample. However, you realize that if you were to draw another sample, you’d obtain a different mean. A third sample would produce yet another mean. And so on.

With this in mind, suppose you decide to collect 50 random samples of the same apple population. Each sample contains 10 apples, and you calculate the mean for each sample.

**Repeated Apple Samples**

At this point, you have 50 sample means for apple weights. You plot these sample means in the histogram below to display your sampling distribution of the mean.

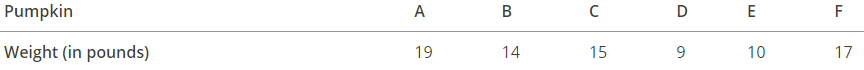


This histogram shows us that our initial sample mean of 103 falls near the center of the sampling distribution. Means occur in this range the most frequently—18 of the 50 samples (36%) fall within the middle bar. However, other samples from the same population have higher and lower means. The frequency of means is highest in the sampling distribution center and tapers off in both directions. None of our 50 sample means fall outside the range of 85-118. Consequently, it is very unusual to obtain sample means outside this range.

The average weight computed for each sample set is the sampling distribution of the mean. Not just the mean can be calculated from a sample. Other statistics, such as the *standard deviation, variance, proportion, and range* can be calculated from sample data. The standard deviation and variance measure the variability of the sampling distribution.

**Example2:**

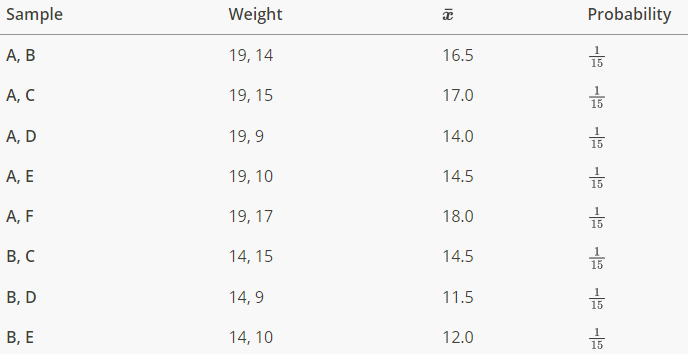
In this example, the population is the weight of six pumpkins (in pounds) displayed in a carnival "guess the weight" game booth. You are asked to guess the average weight of the six pumpkins by taking a random sample without replacement from the population.



Since we know the weights from the population, we can find the population mean.

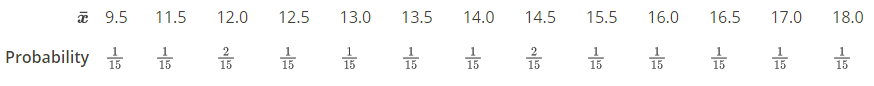


To demonstrate the sampling distribution, let’s start with obtaining all of the possible samples of size n=2 from the populations, sampling without replacement. The table below shows all the possible samples, the weights for the chosen pumpkins, the sample mean and the probability of obtaining each sample. Since we are drawing at random, each sample will have the same probability of being chosen.

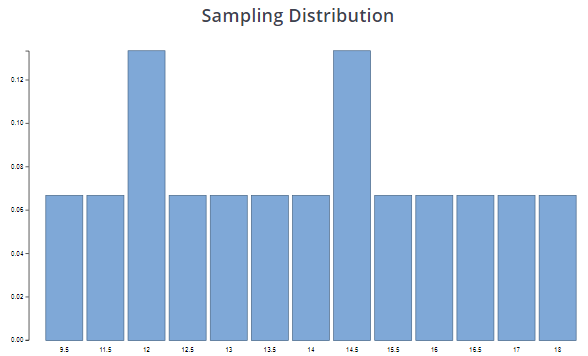




We can combine all of the values and create a table of the possible values and their respective probabilities.



The table is the probability table for the sample mean and it is the sampling distribution of the sample mean weights of the pumpkins when the sample size is 2. It is also worth noting that the sum of all the probabilities equals 1. It might be helpful to graph these values.



One can see that the chance that the sample mean is exactly the population mean is only 1 in 15, very small. (In some other examples, it may happen that the sample mean can never be the same value as the population mean.)

1. **Sampling Distribution of Proportion**

The Sampling Distribution of Proportion measures the proportion of success, i.e. a chance of occurrence of certain events, by dividing the number of successes i.e. chances by the sample size ’n’. Thus, the sample proportion is defined as p = x/n.

**Example:**

An online retailer claims that 90% of all orders are shipped within 12 hours of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within 12 hours.

**Solution:**

The sample proportion is the number x of orders that are shipped within 12 hours divided by the number n of orders in the sample:



1. **T-Distribution**

The t-distribution, also known as the Student’s t-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails. It is used for estimating population parameters for small sample sizes or unknown variances. T-distributions have a greater chance for extreme values than normal distributions, and as a result have fatter tails.

**Example**

ABC Poultry Farms supplies eggs. The company claims its eggs remain fresh for five days if refrigerated. An analyst samples 25 eggs to test this claim. The average freshness of eggs was 4.5 days, with a standard deviation of a day. If the company’s claim is true, find the probability of all selected eggs lasting about 4.5 days.

**Solution:**   
Given:

x̄ = 4.5 days

μ = 5 days

s = 1 day

n = 25

Therefore,  
t = (x̄-µ)/(s/√n)

t = (4.5 – 5)/(1/√25)

t = -0.5/0.2 = -2.5

Since the minus sign is irrelevant here, we get t = 2.5.

Degree of Freedom (df) = n – 1

**df** = 25 – 1 = 24

Thus, according to the t-test, the probability (p-value) of eggs not lasting for more than 4.5 days is 0.01965418.

**Note:** To find the p-value, we have substituted the values of t-score and degree of freedom into an online calculator to get the result: 0.01965418.

**Hypothesis**

A hypothesis is a statement or claim about a property of a population.

**The Fundamentals of Hypothesis Testing**

When conducting scientific research, typically there is some known information, perhaps from some past work or from a long accepted idea. We want to test whether this claim is believable. This is the basic idea behind a hypothesis test:

* State what we think is true.
* Quantify how confident we are about our claim.
* Use sample statistics to make inferences about population parameters.

For example, past research tells us that the average life span for a hummingbird is about four years. You have been studying the hummingbirds in the southeastern United States and find a sample mean lifespan of 4.8 years. Should you reject the known or accepted information in favor of your results? How confident are you in your estimate? At what point would you say that there is enough evidence to reject the known information and support your alternative claim? How far from the known mean of four years can the sample mean be before we reject the idea that the average lifespan of a hummingbird is four years?

**Definition:**

*Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a characteristic of a population.*

*or*

*Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population.*

A hypothesis is a claim or statement about a characteristic of a population of interest to us. A hypothesis test is a way for us to use our sample statistics to test a specific claim.

**Example 1**

The population mean weight is known to be 157 lb. We want to test the claim that the mean weight has increased.

**Example 2**

Two years ago, the proportion of infected plants was 37%. We believe that a treatment has helped, and we want to test the claim that there has been a reduction in the proportion of infected plants.

**Components of a Formal Hypothesis Test**

* The **null hypothesis** is a statement about the value of a population parameter, such as the population mean (µ) or the population proportion (p). It contains the condition of equality and is denoted as H0 (H-naught).

H0 : µ = 157 or H0 : p = 0.37

* The **alternative hypothesis** is the claim to be tested, the opposite of the null hypothesis. It contains the value of the parameter that we consider plausible and is denoted as H1 .

H1 : µ > 157 or H1 : p ≠ 0.37

* The **test statistic** is a value computed from the sample data that is used in making a decision about the rejection of the null hypothesis. The test statistic converts the sample mean (x̄) or sample proportion (p̂) to a Z- or t-score under the assumption that the null hypothesis is true. It is used to decide whether the difference between the sample statistic and the hypothesized claim is significant.
* The **p-value** is the probability of finding the observed t- value, given that the null hypothesis is true. It is the area under the curve to the left or right of the test statistic. It is compared to the level of significance (α).
* The **critical value** is the value that defines the rejection zone (the test statistic values that would lead to rejection of the null hypothesis). It is defined by the level of significance.
* The **level of significance (α)** is the probability that the test statistic will fall into the critical region when the null hypothesis is true. This level is set by the researcher.
* The **conclusion** is the final decision of the hypothesis test. The conclusion must always be clearly stated, communicating the decision based on the components of the test. It is important to realize that we never prove or accept the null hypothesis. We are merely saying that the sample evidence is not strong enough to warrant the rejection of the null hypothesis. The conclusion is made up of two parts:

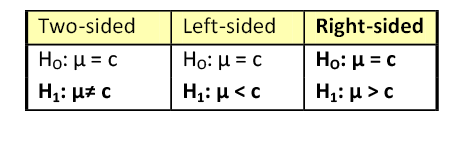
1) Reject or fail to reject the null hypothesis, and 2) there is or is not enough evidence to support the alternative claim.

**Option 1)** Reject the null hypothesis (H0). This means that you have enough statistical evidence to support the alternative claim (H1).

**Option 2)** Fail to reject the null hypothesis (H0). This means that you do NOT have enough evidence to support the alternative claim (H1).

**The Null and Alternative Hypotheses**

There are three different pairs of null and alternative hypotheses:



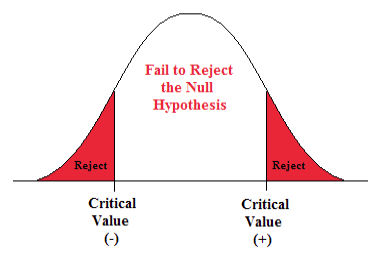
Where, c is some known value.

1. **A Two-sided Test**

This tests whether the population parameter is equal to, versus not equal to, some specific value.

Ho: μ = 12 vs. H1: μ ≠ 12

The critical region is divided equally into the two tails and the critical values are ± values that define the rejection zones.



***Figure 1. The rejection zone for a two-sided hypothesis test.***

**Example 3**

A forester studying diameter growth of red pine believes that the mean diameter growth will be different if a fertilization treatment is applied to the stand.

Ho: μ = 1.2 in./ year

H1: μ ≠ 1.2 in./ year

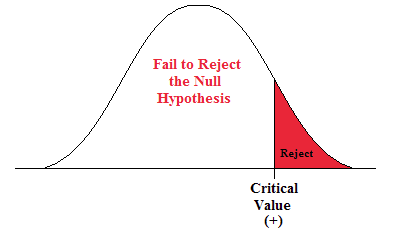
This is a two-sided question, as the forester doesn’t state whether population mean diameter growth will increase or decrease.

1. **A Right-sided Test**

This tests whether the population parameter is equal to, versus greater than, some specific value.

Ho: μ = 12 vs. H1: μ > 12

The critical region is in the right tail and the critical value is a positive value that defines the rejection zone.



***Figure 2. The rejection zone for a right-sided hypothesis test.***

**Example 4**

A biologist believes that there has been an increase in the mean number of lakes infected with milfoil, an invasive species, since the last study five years ago.

Ho: μ = 15 lakes

H1: μ >15 lakes

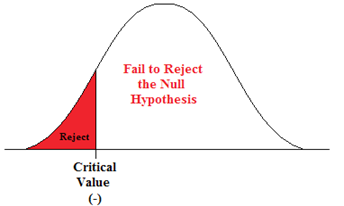
This is a right-sided question, as the biologist believes that there has been an increase in population mean number of infected lakes.

1. **A Left-sided Test**

This tests whether the population parameter is equal to, versus less than, some specific value.

Ho: μ = 12 vs. H1: μ < 12

The critical region is in the left tail and the critical value is a negative value that defines the rejection zone.



***Figure .: The rejection zone for a left-sided hypothesis test.***

**Example 5**

A scientist’s research indicates that there has been a change in the proportion of people who support certain environmental policies. He wants to test the claim that there has been a reduction in the proportion of people who support these policies.

Ho: p = 0.57

H1: p < 0.57

This is a left-sided question, as the scientist believes that there has been a reduction in the true population proportion.

**Significance level**

The significance level is the probability of rejecting the null hypothesis when it is true.

Significance is usually denoted by a p-value, or probability value.

**Statistically Significant**

When the observed results (the sample statistics) are unlikely (a low probability) under the assumption that the null hypothesis is true, we say that the result is statistically significant, and we reject the null hypothesis. This result depends on the level of significance, the sample statistic, sample size, and whether it is a one- or two-sided alternative hypothesis.

**What is Significance level?**

The significance level is the probability of rejecting the null hypothesis when it is true.  Significance is usually denoted by a p-value, or probability value.

Compare your p-value to your significance level. If the p-value is less than your significance level, you can reject the null hypothesis and conclude that the effect is statistically significant. In other words, the evidence in your sample is strong enough to be able to reject the null hypothesis at the population level.

**Test of proportions**

Let us consider the parameter p of the population proportion. For instance, we might want to know the proportion of males within a total population of adults when we conduct a survey.

*A test of proportion will assess whether or not a sample from a population represents the true proportion of the entire population.*

There are 3 tests used in statistics that are tests of proportions including

1. Z-test
2. T Test
3. Chi-square

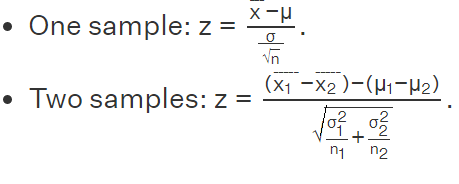
**Types of Test Statistic**

There are four main statistics you can use in a hypothesis test. Which one you use depends on which statistical test you run.

|  |  |  |
| --- | --- | --- |
| **Hypothesis Test** | **Test Statistic** | **Purpose** |
| Z-Test | Z-Score | z-score (also called a standard score) gives you an idea of how far from the mean a data point is. |
| T-Test | T-Score | The t score formula when you don’t know the population standard deviation and you have a small sample (under 30). |
| Chi-Square Test | Chi-square statistic | It tests to see whether distributions of categorical variables differ from each another. |

1. **Z- test**

The **Z-test** is used when comparing the difference in population proportions between 2 groups. A z test is a way of hypothesis testing that is used for a large sample size (n ≥ 30). It is used to determine whether there is a difference between the ***population mean*** and the ***sample mean*** when the population standard deviation is known. It can also be used to compare the mean of two samples. It is used to compute the z test statistic. The formulas are given as follows:



**Note :**

* For a one-tailed test, the critical value is 1.645
* For a two-tailed test, the critical value is 1.96

**Hypothesis Testing Example**

**Example 1:** The best way to solve a problem on hypothesis testing is by applying the 5 steps mentioned in the previous section. Suppose a researcher claims that the mean average weight of men is greater than 100kgs with a standard deviation of 15kgs. 30 men are chosen with an average weight of 112.5 Kgs. Using hypothesis testing; check if there is enough evidence to support the researcher's claim. The confidence interval is given as 95%.

**Step 1:** This is an example of a right-tailed test. Set up the null hypothesis as H0: μ = 100.

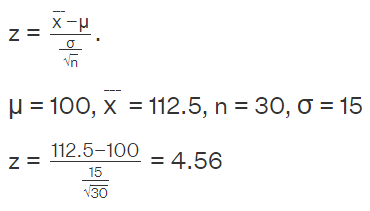
**Step 2:** The alternative hypothesis is given by H1: μ > 100.

**Step 3:** As this is a one-tailed test, α = 100% - 95% = 5%. This can be used to determine the critical value.

1 - α = 1 - 0.05 = 0.95

0.95 gives the required area under the curve. Now using a normal distribution table, the area 0.95 is at z = 1.645. A similar process can be followed for a t-test. The only additional requirement is to calculate the degrees of freedom given by n - 1.

**Step 4:** Calculate the z test statistic. This is because the sample size is 30. Furthermore, the sample and population means are known along with the standard deviation.



**Step 5:** Conclusion. As 4.56 > 1.645 thus, the null hypothesis can be rejected.

**Example 2:**

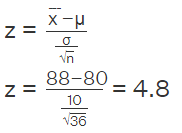
The average score on a test is 80 with a standard deviation of 10. With a new teaching curriculum introduced it is believed that this score will change. On random testing, the score of 38 students, the mean was found to be 88. With a 0.05 significance level, is there any evidence to support this claim?

**Solution:** This is an example of two-tail hypothesis testing.

The z test will be used.

H0: μ = 80, H1: μ≠ 80  
 = 88, μ = 80, n = 36, σ = 10.  
α = 0.05 / 2 = 0.025

The critical value using the normal distribution table is 1.96



As 4.8 > 1.96, the null hypothesis is rejected.  
**Answer:** There is a difference in the scores after the new curriculum was introduced.

**Example 3:** Suppose we want to test whether or not girls, on average, score higher than 600 on the SAT verbal section. We have the information that the standard deviation for girls’ scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our ⍺ value (significance level) to be 0.05. The scores are [650, 730, 510, 670, 480,800, 690, 530, 590, 620,710, 670, 640, 780, 650,490, 800, 600, 510, 700]

**Solution:** Suppose we want to test whether or not girls, on average, score higher than 600 on the SAT verbal section

* Mean Score for Girls is 641
* The size of the sample is 20
* The population mean is 600
* Standard Deviation for Population is 100

Z= =1.8336

Critical value = 1.645

z-score > critical value.

H0:μ <= 600

H1:μ > 600

We can reject the null hypothesis and conclude based on our result that boys on average scored higher than 600.

**Examples using R**

**Ex1.** A random sample of 400 flower stems has an average length of 10.2cms and a standard deviation of 2.25 cm. At 0.05 significance level, can this be regarded as a sample from a large population with mean length of 10cms?

**Solution: R-code with output**

|  |
| --- |
| > n=400;xbar=10.2;sigma=2.25;mu0=10;  > z=(xbar-mu0)/(sigma/sqrt(n))  > z  ***[1] 1.777778*** |

Then the null hypothesis of the lower tail test is to be rejected if Z<=-Zα

We then compute the critical value at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > z.alpha = qnorm(1-alpha)  > z.alpha # critical value  **[1] 1.644854** |

The test statistic 1.77778 is greater than the critical value of 1.6449. Hence, at 0.05 significance level, we reject the claims that sample from a large population with mean length of 10cms

**Ex2.**Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At 0.05 significance level, can we reject the claim by the manufacturer?

**Solution: R-code with output**

|  |
| --- |
| > xbar = 9900 # sample mean  > mu0 = 10000 # hypothesized value  > sigma = 120 # population standard deviation  > n = 30 # sample size  > z =(xbar-mu0)/(sigma/sqrt(n))  > z # test statistic  **[1] -4.564355** |

Then the null hypothesis of the lower tail test is to be rejected if Z<=-Zα

We then compute the critical value at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > z.alpha = qnorm(1-alpha)  > -z.alpha # critical value  **[1] -1.644854** |

The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at 0.05 significance level, we reject the claims that mean lifetime of a light bulb is above 10,000 hours.

**Ex3.** Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At 0.05 significance level, can we reject the claim on food label?

**Solution:** The null hypothesis is that μ≤2. We begin with computing the test statistic.

**R-code with output**

|  |
| --- |
| > xbar = 2.1 # sample mean  > mu0 = 2 # hypothesized value  > sigma = 0.25 # population standard deviation  > n = 35 # sample size  > z = (xbar-mu0)/(sigma/sqrt(n))  > z # test statistic  **[1] 2.366432** |

Then the null hypothesis of the lower tail test is to be rejected if Z<=-Zα

We then compute the critical value at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > z.alpha = qnorm(1-alpha)  > z.alpha # critical value  **[1] 1.644854** |

The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at 0.05 significance level, we reject the claim that there is an atmost 2 gram of saturated fat in a cookie.

**Ex4. Two Tailed Test**

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At 0.05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

|  |
| --- |
| > xbar = 14.6 # sample mean  > mu0 = 15.4 # hypothesized value  > sigma = 2.5 # population standard deviation  > n = 35 # sample size  > z = (xbar-mu0)/(sigma/sqrt(n))  > z # test statistic  **[1] -1.893146** |

Then the null hypothesis of the two-tailed test is to be rejected if z ≤ -zα/2 or  z ≥ zα/2.

We then compute the critical values at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > z.half.alpha = qnorm(1-alpha/2)  > c(-z.half.alpha, z.half.alpha)  **[1] -1.959964 1.959964** |

**Ex5:** A certain restaurant advertises that it puts 0.25 pound of beef in its burgers. A customer who frequents the restaurant thinks the burger actually contain less than 0.25 pound of beef. With permission from the owner, the customer selected a random sample of 60 burgers and found the mean and standard deviation to be 0.22 and 0.07, respectively.

**H0: μ≥0.25**

**Ha: μ<0.25**

|  |
| --- |
| > xbar = 0.22 # sample mean  > mu0 = 0.25 # hypothesized value  > sigma = 0.07 # population standard deviation  > n = 60 # sample size  > z = (xbar-mu0)/(sigma/sqrt(n))  > z # test statistic  **[1] -3.3197** |

Then the null hypothesis of the lower tail test is to be rejected if z≤−zα

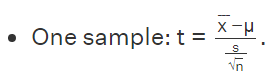
We then compute the critical value at 0.01 significance level.

|  |
| --- |
| > alpha = .01  > z.alpha = qnorm(1-alpha)  > -z.alpha # critical value  **[1] -2.326348** |

Since z=−3.3197<−2.33=z−α, H0 is rejected.At α=0.01 level of significance, the customer has sufficient evidence to claim that the mean amount of beef in burgers the restaurant makes is less than 0.25 lbs.

1. **t Test**

The t test is another method of hypothesis testing that is used for a small sample size (n < 30). A t-test determines whether the mean (average) of a sample population differs significantly from the mean of the larger population. However, the population standard deviation is not known. Instead, the sample standard deviation is known. The mean of two samples can also be compared using the t test.

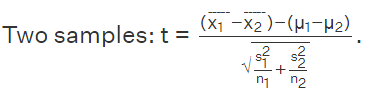


Where,

* T = t-statistic
* = mean of the group



* µ= theoretical mean value of the population
* s = standard deviation of the group
* n = sample size



Where,

* + 1– 2 = means of samples from two different groups or populations



* + n1 and n2 = respective sample sizes
  + s1 and s2 = standard deviation or common variance of two samples

**Definitions and Vocabulary**

**Degrees of freedom:** The degrees of freedom of a t- distribution is given by n-1 for a sample of size n.

**t-score:** Given a sample size n, the t-score tα is a number (any decimal on the horizontal axis) such that the area to the right of tα and between the t-curve (with n-1 degrees of freedom) and the horizontal axis is equal to α.

**critical value:** The critical t-value for a given confidence level c and sample size n is obtained by computing the quantity tα/2 for a t-distribution with n-1 degrees of freedom.

**How to Find the Critical T-value for a Given Confidence Level and Sample Size**

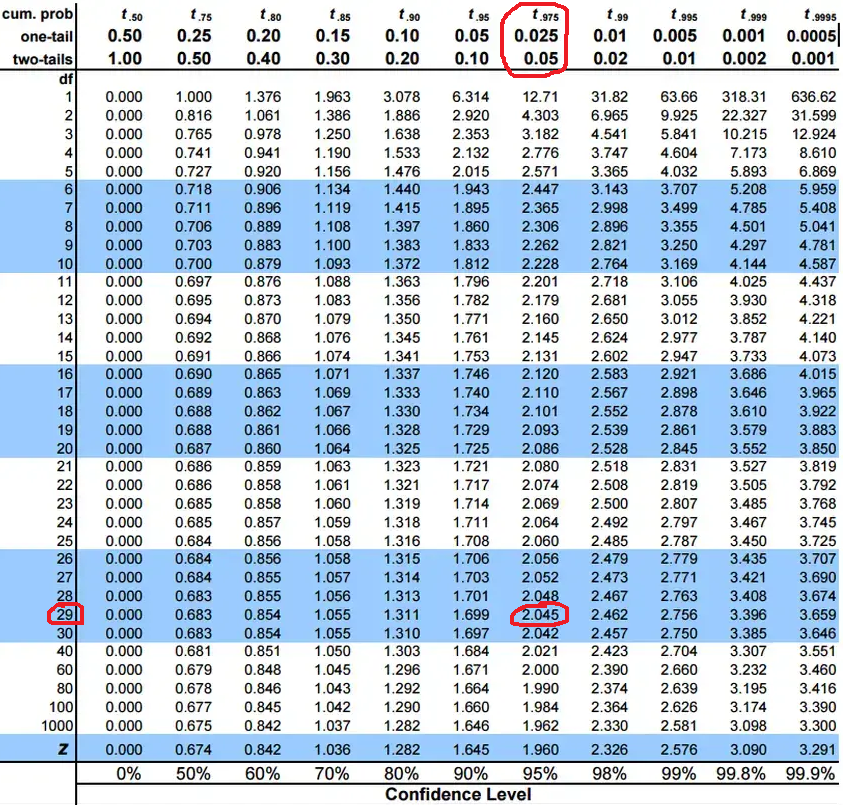
**Step 1:**Express the confidence level as a number (decimal) c with 0<c<1

**Step 2:** Obtain the significance level, denoted α, by α=1-c

**Step 3:** Use the t- table or a calculator to obtain the t- score (critical value) t α/2 where

1. α is from Step 2 and (ii) the degrees of freedom equals n-1, where n is the sample size.

**t-table:**



**Example :** A t-value of 2.35, from a t-distribution with 14 degrees of freedom, has an upper-tail (“greater than”) probability between which two values on the t-table?

**Answer: 0.025 and 0.01**

Using the t-table, locate the row with 14 degrees of freedom and look for 2.35. However, this exact value doesn’t lie in this row, so look for the values on either side of it: 2.14479 and 2.62449. The upper-tail probabilities appear in the column headings; the column heading for 2.14479 is 0.025, and the column heading for 2.62449 is 0.01.

Hence, the upper-tail probability for a t-value of 2.35 must lie between **0.025 and 0.01.**

**Example1:** **95% Confidence Level**

A botanist wants to estimate the mean growth (in millimeters) of all of the plants in a forest after a rainstorm. The botanist tags and measures 30 plants' heights before and after the storm, and the mean growth was 2.1 mm with a standard deviation of 0.4 mm. The botanist wants to construct a 95% confidence interval for the population mean growth increase. Determine the critical t-value needed to construct the confidence interval.

***Step 1:*** Express the confidence level as a number (decimal) c with 0<c<1.

Since the confidence level is 95%, let's take c=0.95.

**Step 2:** Obtain the significance level, denoted α, by α =1−c.

By Step 1, the significance level is given by α =1−0.95=0.05.

**Step 3:** Use the t-table or a calculator to obtain the t-score tα/2 where (i) the α is from Step 2 and (ii) the degrees of freedom equals n−1, where n is the sample size.

The sample size is n=30, so the degrees of freedom is n−1=30−1=29. Thus, we are interested in the quantity

tα/2=t0.05/2=t0.025

for a t-distribution with 29 degrees of freedom. Upon using a t-table or a calculator, we see that the critical t-value for this 95% confidence interval is

**tα/2=2.045**

**Example 2: 99% Confidence Level**

A company wants to estimate the mean price of all used cars in its state. Based on a random sample of 1000 used cars, the mean price was 24,500 dollars with a sample standard deviation of 8500 dollars. The company wants to construct a 99% confidence interval for the population mean used car price. Determine the critical t-value needed to construct this confidence interval.

**Step 1**: Express the confidence level as a number (decimal) c with 0<c<1.

Since the confidence level is 99%, let's take c=0.99.

**Step 2:** Obtain the significance level, denoted , by α =1−c.

By Step 1, the significance level is given by α =1−0.99=0.01.

**Step 3:** Use the t-table or a calculator to obtain the t-score t α /2 where (i) the α is from Step 2 and (ii) the degrees of freedom equals n−1, where n is the sample size.

The sample size is n=1000, so the degrees of freedom is n−1=1000−1=999. Thus, we are interested in the quantity

tα /2=t0.01/2=t0.005

for a t-distribution with 999 degrees of freedom. Upon using a t-table or a calculator, we see that the critical t-value for this 99% confidence interval is

**t0.005=2.581**

**Example 1:** The average weight of a dumbbell in a gym is 90lbs. However, a physical trainer believes that the average weight might be higher. A random sample of 5 dumbbells with an average weight of 110lbs and a standard deviation of 18lbs. Using hypothesis testing check if the physical trainer's claim can be supported for a 95% confidence level.

**Solution:** As the sample size is lesser than 30, the t-test is used.  
H0: μ= 90, H1: μ > 90  
= 110, μ = 90, n = 5, s = 18.  
α = 0.05  
Using the t-distribution table, the critical value is 2.132

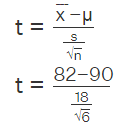


t = 2.484  
As 2.484 > 2.132, the null hypothesis is rejected.  
**Answer:**The average weight of the dumbbells may be greater than 90lbs

**Example 2:** The average score of a class is 90. However, a teacher believes that the average score might be lower. The scores of 6 students were randomly measured. The mean was 82 with a standard deviation of 18. With a 0.05 significance level use hypothesis testing to check if this claim is true.

**Solution:** The t test will be used.  
H0: μ = 90, H1: μ < 90  
 = 82, μ = 90, n = 6, s = 18

The critical value from the t table is -2.015



t = -1.088  
As 1.088 > -2.015, we fail to reject the null hypothesis.  
Answer: There is not enough evidence to support the claim.

**Example 3:** Is there a significant difference in test scores between 25 students who received in-person instruction and 25 students who received online instruction? The mean test score for the in-person group is 80 (SD = 5) and for the online group is 75 (SD = 7).

**Solution:** This is an independent samples t-test problem as the two groups being compared are independent of each other. To perform the t-test, we first calculate the t-value using the t test formula:



Where,

* mean1 is the mean test score for the in-person group,
* mean2 is the mean test score for the online group
* s1 and s2 are the standard deviations for the two groups
* and n1 and n2 are the sample sizes.

Substituting the numbers, we get:

t = (80 - 75) / (sqrt((52/25) + (72/25))) = 2.02

Next, a t-table is used to find the critical t-value for the desired level of significance and degrees of freedom (df = n1 + n2 - 2).

Let us assume a significance level of 0.05 and df = 48. The critical t-value is 2.01.

Since the calculated t-value of 2.02 is greater than the critical t-value of 2.01, we can conclude that there is a significant difference between the test scores of students who receive in-person instruction versus those who receive online instruction.

**Example 4:** A researcher wants to know if there is a significant difference in the weight of newborn babies between two hospitals in a city. The researcher randomly selects 20 newborns from Hospital A and 20 newborns from Hospital B and records their weights in pounds. The mean weight for the Hospital A group is 7.5, with a standard deviation of 0.8. The mean weight for the Hospital B group is 7.1, with a standard deviation of 1.2. Is there a significant difference between the two hospitals?

**Solution:** This is an independent samples t-test problem since the two groups being compared are independent of each other. To perform the t-test, we follow the same steps as in Example 3.

Using the formula for the t-value, we get:



t = (7.5 - 7.1) / (sqrt((0.82/20) + (1.22/20)))  
t = **1.77**

Assuming a significance level of 0.05 and df = 38, the critical t-value is 2.024.

Since the calculated t-value of 1.77 is less than the critical t-value of 2.024, we can conclude that there is not a significant difference in the weight of newborn babies between the two hospitals in the city.

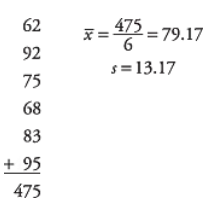
**Assignments:**

**Ques.** A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?

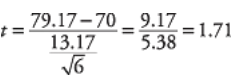
Ans. Null hypothesis: H0: μ = 70

Alternative hypothesis: Ha: μ > 70

First, calculate the mean and standard deviation:



Next, calculate the t-value:



For testing the hypothesis, the computed t-value of 1.71 needs to be compared with the critical value in the t-table. If the calculated t-value is larger than the critical t-value from the table, the null hypothesis will be discarded.

**Ex1.** Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the sample standard deviation is 125 hours. At 0.05 significance level, can we reject the claim by the manufacturer?

**Solution**

The null hypothesis is that μ ≥ 10000. We begin with computing the test statistic.

|  |
| --- |
| > xbar = 9900 # sample mean  > mu0 = 10000 # hypothesized value  > s = 125 # sample standard deviation  > n = 30 # sample size  > t = (xbar-mu0)/(s/sqrt(n))  > t # test statistic  **[1] -4.38178** |

Then the null hypothesis of the lower tail test is to be rejected if t≤−tα

We then compute the critical value at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > t.alpha = qt(1-alpha, df=n-1)  > -t.alpha # critical value  **[1] -1.699127** |

The test statistic -4.3818 is less than the critical value of -1.6991. Hence, at .05 significance level, we can reject the claim that mean lifetime of a light bulb is above 10,000 hours.

**Ex2.** Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the sample standard deviation is 0.3 gram. At 0.05 significance level, can we reject the claim on food label?

**Solution:** The null hypothesis is that μ ≤ 2. We begin with computing the test statistic.

|  |
| --- |
| > xbar = 2.1 # sample mean  > mu0 = 2 # hypothesized value  > s = 0.3 # sample standard deviation  > n = 35 # sample size  > t = (xbar-mu0)/(s/sqrt(n))  > t # test statistic  **[1] 1.972027** |

Then the null hypothesis of the upper tail test is to be rejected if t≥tα

We then compute the critical value at 0.05 significance level.

|  |
| --- |
| > alpha = .05  > t.alpha = qt(1-alpha, df=n-1)  > t.alpha # critical value  **[1] 1.690924** |

The test statistic 1.9720 is greater than the critical value of 1.6991. Hence, at .05 significance level, we can reject the claim that there is at most 2 grams of saturated fat in a cookie.

**Ex3.** Suppose we want to know whether or not the mean weight of a certain species of some turtle is equal to 310 pounds. We go out and collect a simple random sample of turtles with the following weights:

Weights: 300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303

The following code shows how to perform this one sample t-test in R:

|  |
| --- |
| > #define vector of turtle weights  > turtle\_weights <- c(300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303)  >  > #perform one sample t-test  > t.test(x = turtle\_weights, mu = 310)  **One Sample t-test**  **data: turtle\_weights**  **t = -1.5848, df = 12, p-value = 0.139**  **alternative hypothesis: true mean is not equal to 310**  **95 percent confidence interval:**  **303.4236 311.0379**  **sample estimates:**  **mean of x**  **307.2308** |

Since the p-value of the test (0.139) is not less than .05, we fail to reject the null hypothesis.

This means we do not have sufficient evidence to say that the mean weight of this species of turtle is different from 310 pounds.

**Ex4.** Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation is 2.5 kg. At 0.05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

**Solution**

The null hypothesis is that μ = 15.4. We begin with computing the test statistic.

|  |
| --- |
| > xbar = 14.6 # sample mean  > mu0 = 15.4 # hypothesized value  > s = 2.5 # sample standard deviation  > n = 35 # sample size  > t = (xbar-mu0)/(s/sqrt(n))  > t # test statistic  **[1] -1.893146** |

Then the null hypothesis of the two-tailed test is to be rejected if t≤−tα/2or t≥tα/2

We then compute the critical values at .05 significance level.

|  |
| --- |
| > alpha = .05  > t.half.alpha = qt(1-alpha/2, df=n-1)  > c(-t.half.alpha, t.half.alpha)  **[1] -2.032245 2.032245** |

The test statistic -1.8931 lies between the critical values -2.0322, and 2.0322. Hence, at 0.05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

1. **Chi-square test**

The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

**Types of Chi-square test**

There are two types of Pearson’s chi-square tests:

* **Chi-square goodness of fit test**

The **chi-square goodness of fit** **test**is used when you have one categorical variable. It allows you to test whether the frequency distribution of a categorical variable is different from your expectations.

* **Chi-square test of independence**

The chi-square test of independence is used when you have two categorical variables. It allows you to test whether the two variables are related to each other. If two variables are independent (unrelated), the probability of belonging to a certain group of one variable isn’t affected by the other variable.

Chi-square is often written as χ2 and is pronounced “kai-square”. It is also called chi-squared.

**The chi-square formula**

Both of **Pearson’s chi-square tests** use the same formula to calculate the test statistic, chi-square (χ2):

χ2=

Where:

* χ2 is the chi-square test statistic
* Σ is the summation operator (it means “take the sum of”)
* O is the observed frequency
* E is the expected frequency

**Example 1:**

Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Republican** | **Democrat** | **Independent** | **Total** |
| **Male** | 100 | 70 | 30 | 200 |
| **Female** | 140 | 60 | 20 | 220 |
| **Total** | 240 | 130 | 50 | 440 |

To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.

**Step 1:** Define the Hypothesis

**H0:** There is no link between gender and political party preference.

**H1:** There is a link between gender and political party preference.

**Step 2:** Calculate the Expected Values

Now you will calculate the expected frequency.

*Expected Value* =

For example, the expected value for Male Republicans is:

= =109

Similarly, you can calculate the expected value for each of the cells.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Expected Values** |  |  |  |  |
|  | **Republican** | **Democrat** | **Independent** | **Total** |
| **Male** | 109 | 59 | 22.72 | 200 |
| **Female** | 120 | 65 | 25 | 220 |
| **Total** | 240 | 130 | 50 | 440 |

**Step 3:** Calculate (O-E)2 / E for Each Cell in the Table

Now you will calculate the (O - E)2 / E for each cell in the table.

Where, O = Observed Value

E = Expected Value

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **(O-E)2/E** |  |  |  |  |
|  | **Republican** | **Democrat** | **Independent** | **Total** |
| **Male** | 0.74311927 | 2.05847 | 2.332676056 | 200 |
| **Female** | 3.33333333 | 0.384615 | 1 | 220 |
| **Total** | 240 | 130 | 50 | 440 |

**Step 4:** Calculate the Test Statistic X2

X2  is the sum of all the values in the last table

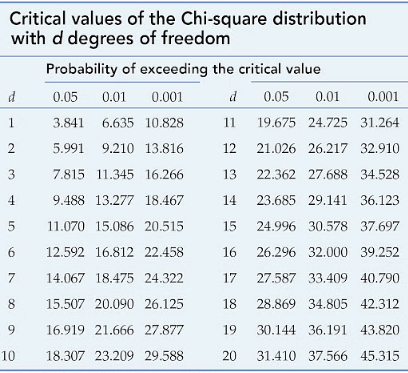
  =  0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1

 = 9.837

Before you can conclude, you must first determine the *critical statistic*, which requires determining our degrees of freedom. The degrees of freedom in this case are equal to the table's number of columns minus one multiplied by the table's number of rows minus one, or (r-1) (c-1). We have (3-1)(2-1) = 2.

Finally, you compare our obtained statistic to the critical statistic found in the chi-square table. As you can see, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, which is less than our obtained statistic of 9.83. You can reject our null hypothesis because the critical statistic is higher than your obtained statistic.

This means you have sufficient evidence to say that there is an association between gender and political party preference.



**Example 2:**

Is gender independent of education level? A random sample of 395 people was surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey are summarized in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **High School** | **Bachelors** | **Masters** | **Ph.D** | **Total** |
| **Female** | 60 | 54 | 46 | 41 | 201 |
| **Male** | 40 | 44 | 53 | 57 | 194 |
| **Total** | 100 | 98 | 99 | 98 | 395 |

**Question:** Are gender and education level dependent at a 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

Here's the table of expected counts:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **High School** | **Bachelors** | **Masters** | **Ph.D** | **Total** |
| **Female** | 50.886 | 49.868 | 50.377 | 49.868 | 201 |
| **Male** | 49.114 | 48.132 | 48.623 | 48.132 | 194 |
| **Total** | 100 | 98 | 99 | 98 | 395 |

So, Working this out, χ2 = +. . . + = 8.006

The critical value of χ2 with 3 degrees of freedom is 7.815. Since 8.006 > 7.815, we reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance.

**Chi-Square Test of Independence in R**

**Ex1:** Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 500 voters and survey them on their political party preference. The following table shows the results of the survey:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Republican** | **Democrat** | **Independent** | **Total** |
| **Male** | 120 | 90 | 40 | 250 |
| **Female** | 110 | 95 | 45 | 250 |
| **Total** | 230 | 185 | 85 | 500 |

|  |
| --- |
| > *#create table*  > data <- matrix(c(120, 90, 40, 110, 95, 45), ncol=3, byrow=TRUE)  > colnames(data) <- c("Rep","Dem","Ind")  > rownames(data) <- c("Male","Female")  > data <- as.table(data)  > *#view table*  > data  **Rep Dem Ind**  **Male 120 90 40**  **Female 110 95 45**  > *#Perform Chi-Square Test of Independence*  > chisq.test(data)  **Pearson's Chi-squared test**  **data: data**  **X-squared = 0.86404, df = 2, p-value = 0.6492**  **Warning message:**  **In chisq.test(mx, correct = T) : Chi-squared approximation may be incorrect** |

* H0: (null hypothesis) The two variables are independent.
* H1: (alternative hypothesis) The two variables are not independent.

Since the p-value (0.6492) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and political party preference.

**Ex2.** Following table provides data with regard to stature(height) of the fathers and their first son at the age of 25 years.

|  |  |  |
| --- | --- | --- |
| **Stature of Sons** | **Stature of Father** | |
|  | **Tall** | **Short** |
| **Tall** | 8 | 2 |
| **Short** | 7 | 6 |

Test whether the stature of sons is independent of the stature of the fathers.

|  |
| --- |
| > #create table  > data <- matrix(c(8,2,7,6),ncol=2, byrow=TRUE)  > colnames(data) <- c("FTall","FShort")  > rownames(data) <- c("STall","SShort")  > #view table  > data  **FTall FShort**  **STall 8 2**  **SShort 7 6**  > #Perform Chi-Square Test of Independence  > chisq.test(data)  **Pearson's Chi-squared test with Yates' continuity correction**  **data: data**  **X-squared = 0.74639, df = 1, p-value = 0.3876**  **Warning message:**  **In chisq.test(data) : Chi-squared approximation may be incorrect** |

* H0: (null hypothesis) The stature of son is independent of stature of father
* H1: (alternative hypothesis) The stature of son is dependent of stature of father

Since the p-value (0.3876) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the stature of son is dependent of stature of father.

**Ex3.** Use the following data to test whether the attributes condition of home and condition of child are independent.

|  |  |  |
| --- | --- | --- |
| **Condition of Child** | **Condition of Home** | |
| **Clean** | **Dirty** |
| **Clean** | 70 | 50 |
| **Fairly Clean** | 80 | 20 |
| **Dirty** | 35 | 45 |

|  |
| --- |
| > #create table  > data <- matrix(c(70,50,80,20,35,45),ncol=2, byrow=TRUE)  > colnames(data) <- c("HomeClean","HomeDirty")  > rownames(data) <- c("Childclean","ChildFairlyClean","ChildDirty")  > #view table  > data  **HomeClean HomeDirty**  **Childclean 70 50**  **ChildFairlyClean 80 20**  **ChildDirty 35 45**  > #Perform Chi-Square Test of Independence  > chisq.test(data)  **Pearson's Chi-squared test**  **data: data**  **X-squared = 25.646, df = 2, p-value = 2.698e-06** |

* H0: (null hypothesis) The attributes condition of home and condition of child are independent
* H1: (alternative hypothesis) The attributes condition of home and condition of child are dependent
* Since the p-value (2.698e-06) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that The attributes condition of home and condition of child are dependent

**Chi-square goodness of fit test**

**Ex1.** A die is tossed 120 times and the following results are obtained.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number turned up :** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Frequency:** | 30 | 25 | 18 | 10 | 22 | 15 |

Test the hypothesis that the die is unbiased (Neutral).

|  |
| --- |
| > f=c(30,25,18,10,22,15)  > chisq.test(f,p=rep(1/6,6))  **Chi-squared test for given probabilities**  **data: f**  **X-squared = 12.9, df = 5, p-value = 0.02433** |

Here level of significance α=0.05 is greater than p-value=0.02433. Hence we reject the null hypothesis. Therefore a die may not be unbiased.

**Ex2.** A shop owner claims that an equal number of customers come into his shop each weekday. To test this hypothesis, a researcher records the number of customers that come into the shop in a given week and finds the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks:** | Mon | Tue | Wed | Thu | Fri |
| **No. Customers:** | 50 | 60 | 40 | 47 | 53 |

|  |
| --- |
| > observed=c(50,60,40,47,53)  > expected <- c(.2, .2, .2, .2, .2) #must add up to 1 i.e, 1/5…1/5  > #perform Chi-Square Goodness of Fit Test  > chisq.test(x=observed, p=expected)  **Chi-squared test for given probabilities**  **data: observed**  **X-squared = 4.36, df = 4, p-value = 0.3595** |

Since the p-value (.35947) is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the true distribution of customers is different from the distribution that the shop owner claimed.

**Ex3:** we collected wild tulips and found the following data

|  |  |  |  |
| --- | --- | --- | --- |
| **Colors:** | Red | Yellow | White |
| **No. Tulips:** | 81 | 50 | 27 |

1. Are these colors equally common?
2. We want to know, if there is any significant difference between the observed proportions and the expected proportions.

**Solution:**

1. *Are these colors equally common?*

If these colors were equally distributed, the expected proportion would be 1/3 for each of the color.

|  |
| --- |
| > tulip <- c(81, 50, 27)  > res <- chisq.test(tulip, p = c(1/3, 1/3, 1/3))  > res  **Chi-squared test for given probabilities**  **data: tulip**  **X-squared = 27.886, df = 2, p-value = 8.803e-07** |

The p-value of the test is 8.80310^{-7}, which is less than the significance level alpha = 0.05. We can conclude that the colors are significantly not commonly distributed with a p-value = 8.80310^{-7}.

1. *We want to know, if there is any significant difference between the observed proportions and the expected proportions.*

Suppose that, in the region where you collected the data, the ratio of red, yellow and white tulip is 3:2:1 (3+2+1 = 6). This means that the expected proportion is:

* 3/6 (= 1/2) for red
* 2/6 ( = 1/3) for yellow
* 1/6 for white

|  |
| --- |
| > tulip <- c(81, 50, 27)  > res <- chisq.test(tulip, p = c(1/2, 1/3, 1/6))  > res  **Chi-squared test for given probabilities**  **data: tulip**  **X-squared = 0.20253, df = 2, p-value = 0.9037** |

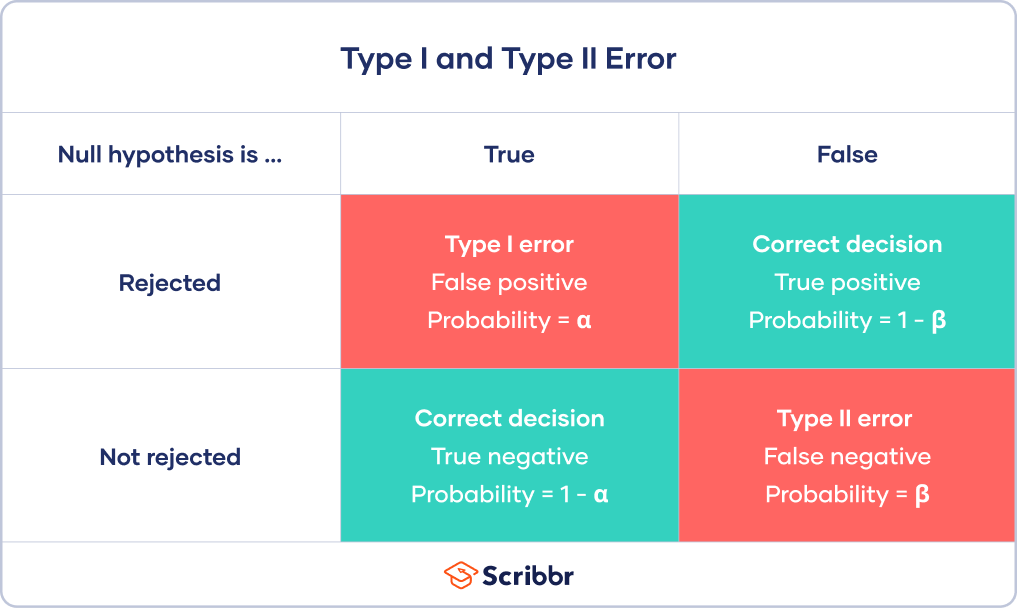
**Statistical hypotheses**

* **Null hypothesis (H0):** There is no significant difference between the observed and the expected value.
* **Alternative hypothesis (Ha):** There is a significant difference between the observed and the expected value.

The p-value of the test is 0.9037, which is greater than the significance level alpha = 0.05. We can conclude that the observed proportions are not significantly different from the expected proportions.

**Types of Errors**

When testing, we arrive at a conclusion of rejecting the null hypothesis or failing to reject the null hypothesis. Such conclusions are sometimes correct and sometimes incorrect (even when we have followed all the correct procedures). We use incomplete sample data to reach a conclusion and there is always the possibility of reaching the wrong conclusion. There are four possible conclusions to reach from hypothesis testing. Of the four possible outcomes, two are correct and two are NOT correct.



*Table 1. Possible outcomes from a hypothesis test*.

A **Type I** **error** is when we reject the null hypothesis when it is true. The symbol **α (alpha)** is used to represent Type I errors. This is the same alpha we use as ***the level of significance.*** By setting alpha as low as reasonably possible, we try to control the Type I error through the level of significance.

A **Type II error** is when we fail to reject the null hypothesis when it is false. The symbol **β (beta)** is used to represent Type II errors.

**Type I vs Type II error**

**Example1:**

You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

* **Type I error** (false positive): the test result says you have coronavirus, but you actually don’t.
* **Type II error** (false negative): the test result says you don’t have coronavirus, but you actually do.

**Example 2:**

A **Type I error** happens when you get false positive results: you conclude that the drug intervention improved symptoms when it actually didn’t. These improvements could have arisen from other random factors or measurement errors.

A **Type II** error happens when you get false negative results: you conclude that the drug intervention didn’t improve symptoms when it actually did. Your study may have missed key indicators of improvements or attributed any improvements to other factors instead.

In general, Type I errors are considered more serious. One step in the hypothesis test procedure involves selecting the significance level (α), which is the probability of rejecting the null hypothesis when it is correct. So the researcher can select the level of significance that minimizes Type I errors. However, there is a mathematical relationship between α, β, and n (sample size).

* As α increases, β decreases
* As α decreases, β increases
* As sample size increases (n), both α and β decrease

**Why do hypothesis tests make errors?**

Hypothesis tests use samples to draw conclusions about entire populations. Researchers use these tests because it’s rarely possible to measure a whole population. So, they’re stuck with samples.

Unfortunately, samples don’t always accurately reflect the population. Statisticians define sampling error as the difference between a sample and the target population. Occasionally, this error can be large enough to cause hypothesis tests to draw the wrong conclusions. Consequently, statistical power becomes a crucial issue because increasing it reduces the chance of errors.

**What Is Statistical Power?**

Statistical power or the power of a hypothesis test is the probability that the test correctly rejects the null hypothesis.

That is, the probability of a true positive result. It is only useful when the null hypothesis is rejected.

The higher the statistical power for a given experiment, the lower the probability of making a Type II (false negative) error. That is the higher the probability of detecting an effect when there is an effect. In fact, the power is precisely the inverse of the probability of a Type II error.

More intuitively, the statistical power can be thought of as the probability of accepting an alternative hypothesis, when the alternative hypothesis is true.

When interpreting statistical power, we seek experiential setups that have high statistical power.

* **Low Statistical Power:** Large risk of committing Type II errors, e.g. a false negative.
* **High Statistical Power:** Small risk of committing Type II errors.

Experimental results with too low statistical power will lead to invalid conclusions about the meaning of the results. Therefore a minimum level of statistical power must be sought.

**Power Analysis**

A statistical power analysis helps determine how large your sample must be to detect an effect.

**Example:** Let's say you are conducting a study to test whether a new drug is effective at lowering blood pressure. You have two groups: a treatment group that receives the new drug, and a control group that receives a placebo. Your null hypothesis (H0) is that there is no difference in blood pressure between the two groups, and your alternative hypothesis (Ha) is that there is a difference, with the drug lowering blood pressure.

This process requires considering several factors:

1. **Effect Size (ES):** This is the size of the difference in means (or proportions) between the treatment and control groups. In our example, it might be the average reduction in blood pressure due to the drug.
2. **Significance Level (α):** This is the probability of making a Type I error, which is rejecting the null hypothesis when it is true. Commonly chosen values are 0.05 or 0.01.
3. **Sample Size (n):** The number of participants in each group. A larger sample size generally leads to higher power.
4. **Variability (Standard Deviation):** The spread of data in each group. A smaller standard deviation leads to higher power.
5. **Type of Test:** The choice of statistical test (e.g., t-test, chi-squared test) can also affect power.

The formula for power (often denoted as β) is:

***β=P (Reject H0​∣ Ha​ is true)***

In our example, if the new drug truly lowers blood pressure (i.e., Ha is true), and you have determined the effect size, significance level, sample size, and standard deviation, you can calculate the power of your test.

A high power (*close to 1*) indicates that your test is likely to correctly detect a true effect if it exists. Conversely, a low power (*close to 0*) means your test may fail to detect a true effect, increasing the risk of a Type II error (failing to reject the null hypothesis when it is false).

For example, if you calculate that the power of your test is 0.8, it means there is an 80% chance that your test will correctly identify the difference in blood pressure if the drug is indeed effective.

**Example**

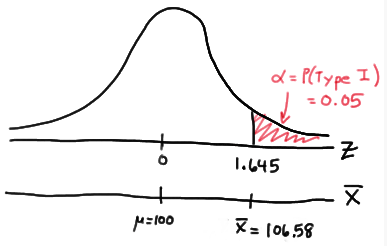
**Calculating the power of a hypothesis test.**

Let X denotes the IQ of a randomly selected adult American. Assume, a bit unrealistically, that X is normally distributed with unknown mean µ and standard deviation 16. Take a random sample of n=16 students, so that, after setting the probability of committing a Type I error at α=0.05, we can test the null hypothesis H0: µ =100 against the alternative hypothesis that H1: µ >100.

What is the power of the hypothesis test if the true population mean were µ =108?

**Solution:**

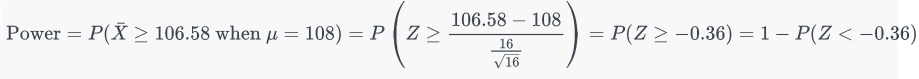
Setting α, the probability of committing a Type I error, to 0.05, implies that we should reject the null hypothesis when the test statistic Z≥1.645, or equivalently, when the observed sample mean is 106.58 or greater:

****

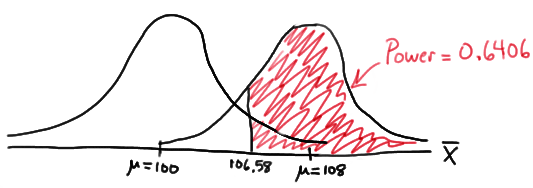
because we transform the test statistic Z to the sample mean by way of:



Now, that implies that the power, that is, the probability of rejecting the null hypothesis, when µ =108 is 0.6406 as calculated here (recalling that Phi (z) is standard notation for the cumulative distribution function of the standard normal random variable):



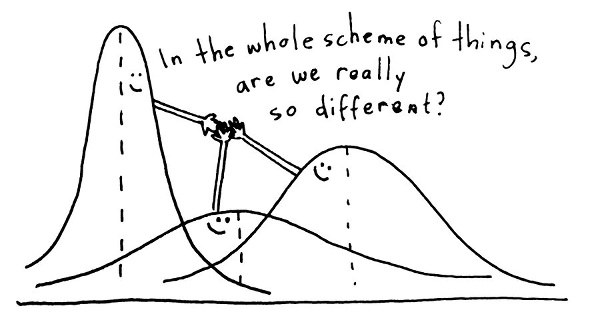
and illustrated here:



In summary, we have determined that we have (only) a 64.06% chance of rejecting the null hypothesis H0: µ =100 in favor of the alternative hypothesis HA: µ >100 if the true unknown population mean is in reality µ =108.

**Analysis of Variance (ANOVA)- A test for Homogeneity of Mean**

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method. ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.



*Analysis of Variance* is a method of splitting the total variation of a data into constituent parts which measures different sources of variations.

or

*Analysis of Variance (ANOVA)* is an analysis tool used to compare mean between two or more items. It’s a statistical method that yields values that can be tested to determine whether a significant relation exists between variables.

Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

**Example:**

* To study the effectiveness of different diabetes medications, scientists design and experiment to explore the relationship between the type of medicine and the resulting blood sugar level. The sample population is a set of people. We divide the sample population into multiple groups, and each group receives a particular medicine for a trial period. At the end of the trial period, blood sugar levels are measured for each of the individual participants. Then for each group, the mean blood sugar level is calculated. ANOVA helps to compare these group means to find out if they are statistically different or if they are similar.
* A car company wishes to compare the average petrol consumption of three similar models of cars and has six vehicles available for each model.  It follows a 6×3 matrix, columns have cars and rows have models. Here, we compare the average petrol consumption.
* A teacher is interested in comparing the average percentage marks attained in the examinations of five different subjects and the marks are available for eight students, who have completed each examination. If the teacher wants to compare the mean average % of marks between all students of five different subjects, for comparing the mean between two entities we use Analysis of Variance.

ANOVA separates the variation in the dataset into 2 parts: between-group and within-group. These variations are called the sums of squares

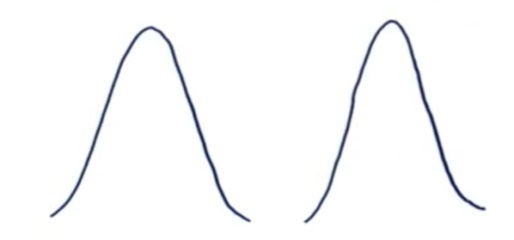
**The total variation is split up into the following two components:**

1. Variation **between** the subgroups of the samples**.**
2. Variation **within** the subgroups of samples.
3. **Between Group Variability**

* Consider the distributions of the below two samples. As these samples overlap, their individual means won’t differ by a great margin. Hence the difference between their individual and grand means won’t be significant enough.

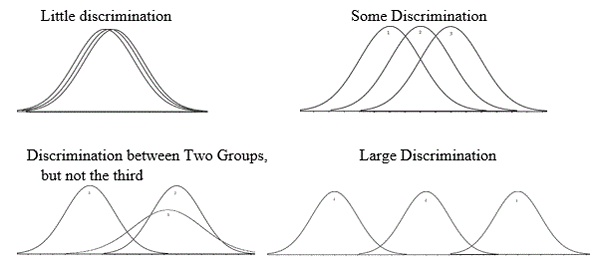


* Now consider these two sample distributions. As the samples differ from each other by a big margin, their individual means would also differ. The difference between the individual means and grand mean would, therefore, also be significant.



Such variability between the distributions is called **Between-group variability**. It refers to variations between the distributions of individual groups (or levels) as the values within each group differ.

Each sample is examined, and the difference between its mean and grand mean is calculated to calculate the variability. If the distributions overlap or are close, the grand mean will be similar to the individual means, whereas if the distributions are far apart, the difference between means and grand mean would be large.

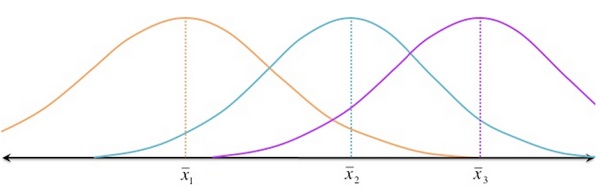


**Sum-of-squares for between-group variability**

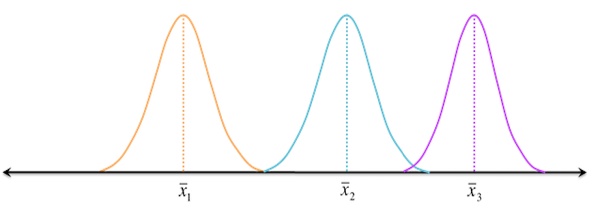


1. **Within Group Variability**

Consider the given distributions of three samples. As the spread (variability) of each sample increases, their distributions overlap, and they become part of a big population.



Now consider another distribution of the same three samples but with less variability. Although the means of samples are similar to those in the above image, they seem to belong to different populations.



Such variations within a sample are denoted by Within-group variation. It refers to variations caused by differences within individual groups (or levels), as not all the values within each group are the same. Each sample is looked at on its own, and variability between the individual points in the sample is calculated. In other words, no interactions between samples are considered.

We can measure Within-group variability by looking at how much each value in each sample differs from its respective sample mean. So first, we’ll take the squared deviation of each value from its respective sample mean and add them up. This is the sum of **squares for within-group variability.**

**SSwithin =∑(Xi1-1)2+∑(Xi2-2)2+∑(Xi3-3)2+………..+=∑(Xik-k)2**

**=∑(Xij-j)2**

**The F ratio**

The outcome of ANOVA is the **‘F statistic’**. This ratio shows the difference between the within group variance and the between group variance, which ultimately produces a figure which allows a conclusion that the null hypothesis is supported or rejected. If there is a significant difference between the groups, the null hypothesis is not supported, and the F-ratio will be larger.

**F = Between-group variability / Within-group variability**

**F=SSC/SSE**

This F-statistic calculated here is compared with the F-critical value for concluding. If the value of the calculated F-statistic is more than the F-critical value (for a specific α/significance level), then we reject the null hypothesis

**ANOVA Terminology**

* **Dependent variable:** This is the item being measured that is theorized to be affected by the independent variables.
* **Independent variable/s:** These are the items being measured that may have an effect on the dependent variable.
* **A null hypothesis (H0):** This is when there is no difference between the groups or means. Depending on the result of the ANOVA test, the null hypothesis will either be accepted or rejected.
* **An alternative hypothesis (Ha):** When it is theorized that there is a difference between groups and means.
* **Factors and levels:** In ANOVA terminology, an independent variable is called a factor which affects the dependent variable. Level denotes the different values of the independent variable that are used in an experiment.
* **Fixed-factor model:** Some experiments use only a discrete set of levels for factors. For example, a fixed-factor test would be testing three different dosages of a drug and not looking at any other dosages.
* **Random-factor model:** This model draws a random value of level from all the possible values of the independent variable.

**Types of ANOVA**

There are two types of ANOVA.

1. **One-Way ANOVA**

The one-way analysis of variance is also known as single-factor ANOVA or simple ANOVA. As the name suggests, the one-way ANOVA is suitable for experiments with only one independent variable (factor) with two or more levels. For instance a dependent variable may be what month of the year there are more flowers in the garden. There will be twelve levels. A one-way ANOVA assumes:

**Independence:** The value of the dependent variable for one observation is independent of the value of any other observations.

**Normalcy:** The value of the dependent variable is normally distributed

**Variance:** The variance is comparable in different experiment groups.

**Continuous:** The dependent variable (number of flowers) is continuous and can be measured on a scale which can be subdivided.

1. **Full Factorial ANOVA (also called two-way ANOVA)**

Full Factorial ANOVA is used when there are two or more independent variables. Each of these factors can have multiple levels. Full-factorial ANOVA can only be used in the case of a full factorial experiment, where there is use of every possible permutation of factors and their levels. This might be the month of the year when there are more flowers in the garden, and then the number of sunshine hours. This two-way ANOVA not only measures the independent vs the independent variable, but if the two factors affect each other. A two-way ANOVA assumes:

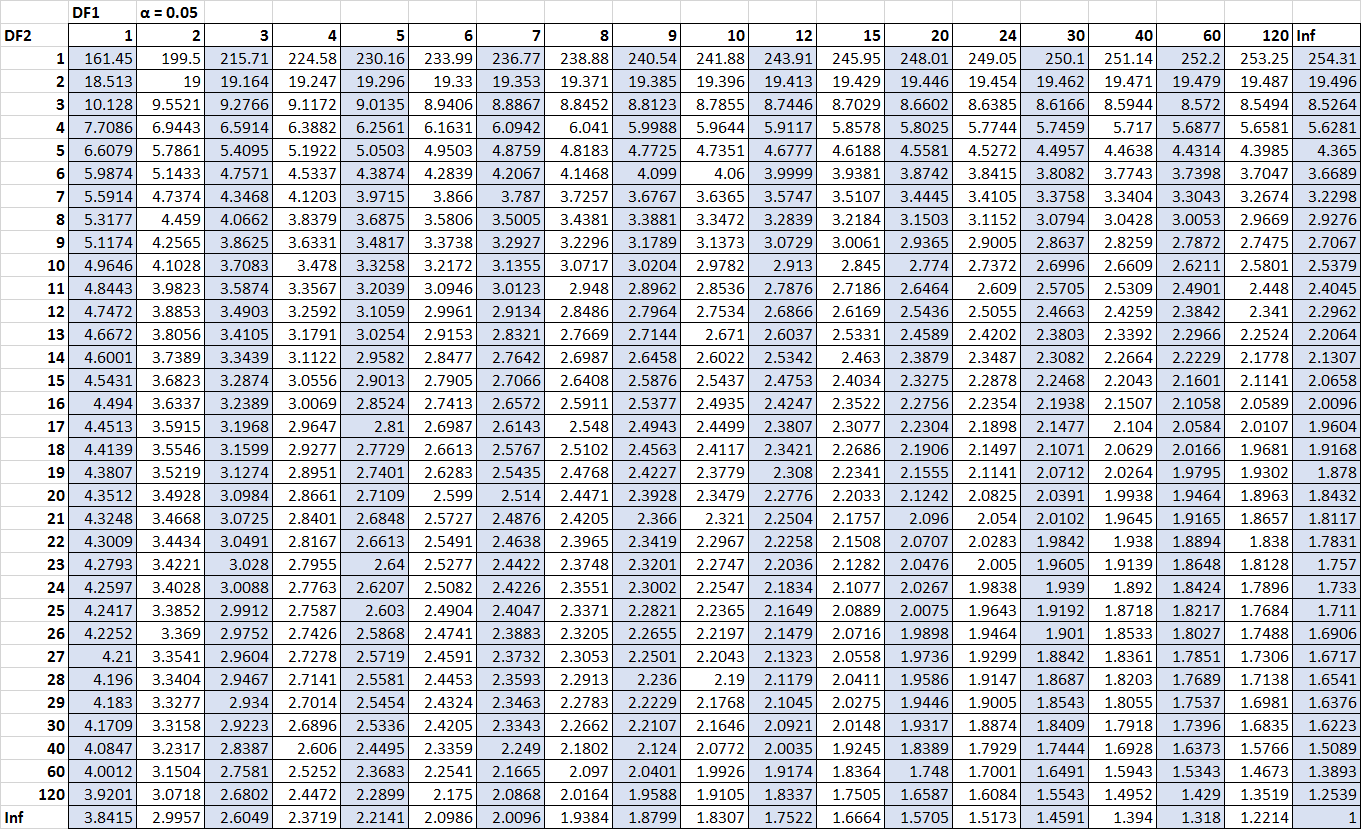
**Continuous:** The same as a one-way ANOVA, the dependent variable should be continuous.

**Independence:** Each sample is independent of other samples, with no crossover.

**Variance:** The variance in data across the different groups is the same.

**Normalcy:** The samples are representative of a normal population.

**Categories:** The independent variables should be in separate categories or groups.



**Example Problem:**

1. The three samples below have been obtained from normal population with equal variances. Test the hypothesis at 5% level that the population means are equal;

|  |  |  |
| --- | --- | --- |
| **Sample 1** | **Sample 2** | **Sample 3** |
| 8 | 7 | 12 |
| 10 | 5 | 9 |
| 7 | 10 | 13 |
| 14 | 9 | 12 |
| 11 | 9 | 14 |

*(The table value of F at 5% level of significance for v1=2 and v2=12 is 3.88)*

**Solution:**

**Null Hypothesis: - H0: There is no significant difference in means of the three samples.**

Using analysis of variance technique we have the following table

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X1** | **X2** | **X3** |
|  | 8 | 7 | 12 |
| 10 | 5 | 9 |
| 7 | 10 | 13 |
| 14 | 9 | 12 |
| 11 | 9 | 14 |
| **Total** | **50** | **40** | **60** |
|  | 1=50/5=**10** | 2=40/5=**8** | 3=60/5=**12** |

**Grand Mean:** = (10+8+12)/3=**10**

**Variance Between samples**

|  |  |  |
| --- | --- | --- |
| **(1-)2** | **(2-)2** | **(3-)2** |
| 0 | 4 | 4 |
| 0 | 4 | 4 |
| 0 | 4 | 4 |
| 0 | 4 | 4 |
| 0 | 4 | 4 |
| **0** | **20** | **20** |

**SSC= Sum of squares between samples=0+20+20=40**

**Variance within samples**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **X1** | **(X1-1)2** | **X2** | **(X2-2)2** | **X3** | **(X3-3)2** |
| 8 | 4 | 7 | 1 | 12 | 0 |
| 10 | 0 | 5 | 9 | 9 | 9 |
| 7 | 9 | 10 | 4 | 13 | 1 |
| 14 | 16 | 9 | 1 | 12 | 0 |
| 11 | 1 | 9 | 1 | 14 | 4 |
|  | **30** |  | **16** |  | **14** |

**SSE = Sum of squares within samples = 30+16+14 = 60**

**ANOVA Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source of Variation** | **Sum of squares** | **V** | **Mean squares** |
| SSC= Between samples | 40 | 2 | 40/2=**20** |
| SSE = Within samples | 60 | 12 | 60/12=**5** |
| **Total** | **SST=100** | **14** |  |

**Test Statistic:** F = SSC/SSE

F = 20/5 **= 4**

For v1=2, v2=12 and for =0.05, the table value of **F0.05 = 3.80**

**Since the calculated value of F=4 is greater than the tabled value F0.05 = 3.80, so that the null hypothesis is rejected. Hence there is a significant difference in the means of the three given samples.**

**Problem 2:** Four salesmen were posted in different areas by a company. The number of units of commodity X sold by them are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | 20 | 23 | 28 | 29 |
| B | 25 | 32 | 30 | 21 |
| C | 23 | 28 | 35 | 18 |
| D | 15 | 21 | 19 | 25 |

On the basis of his information, can it be concluded that there is a significant difference in the performance of the four salesmen? (Given for v1=3 and v2=12, F0.05=3.24).

**Ex1. (One way classification)**

The following table shows the lives(in hours) of four batches of electric lamps:

|  |  |
| --- | --- |
| **Batches** | **Life of bulbs in Hours** |
| 1 | 1600,1610,1650,1680,1700,1720,1800 |
| 2 | 1580,1640,1640,1700,1750 |
| 3 | 1460,1550,1600,1620,1640,1660,1740,1820 |
| 4 | 1510,1520,1530,1570,1600,1680 |

Perform an analysis of variance of these data.

**Solution:** H0:α1=α2=…….=αk=0 i.e, Batch means do not differ significantly.

|  |
| --- |
| > x1=c(1600,1610,1650,1680,1700,1720,1800) #x1 is vector of observations of class1  > x2=c(1580,1640,1640,1700,1750)  > x3=c(1460,1550,1600,1620,1640,1660,1740,1820)  > x4=c(1510,1520,1530,1570,1600,1680)  > d=stack(list(b1=x1,b2=x2,b3=x3,b4=x4)) #stack returns two variables  > names(d) #To give names to list  [1] "values" "ind"  > av1=aov(values ~ind,data=d)# aov() is used to summarize the analysis of variance model.  > summary(av1)# returns the results of basic statistical calculations  **Df Sum Sq Mean Sq F value Pr(>F)**  **ind 3 44361 14787 2.149 0.123**  **Residuals 22 151351 6880** |

If given level of significance (α) is less than the p-value(0.123) then we accept the null hypothesis.

**Ex2 (Two way classification)**

A tea company appoints four salesmen A,B,C and D and observes their sales in three seasons – Summer, Winter and Monsoon. The figures (in lakhs) of sales are given in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Season/Salesman** | **A** | **B** | **C** | **D** |
| Summer | 36 | 36 | 21 | 35 |
| Winter | 28 | 29 | 31 | 32 |
| Monsoon | 26 | 28 | 29 | 29 |

1. Do the salesmen significantly differ in performance?
2. Is there significant difference between the seasons?

**Solution:**

**H01:**β1= β2= β3=……..= βk=0 i.e, Sales of different salesmen do not differ significantly

**H02:**α1=α2=……..=αk=0 i.e, Sales in different seasons do not differ significantly

|  |
| --- |
| > s1=c(rep(1:3,rep(4,3)))  > s1  [1] 1 1 1 1 2 2 2 2 3 3 3 3  > s2=rep(c("A","B","C","D"),3)  > s2  [1] "A" "B" "C" "D" "A" "B" "C" "D" "A" "B" "C" "D"  > obs=c(36,36,21,35,28,29,31,32,26,28,29,29)  > d=data.frame(s1,s2,obs*)# data.frame is used to display in a format as a table.*  > seasons=factor(s1*)# factor() is used to categorize the data or represent categorical data and store it on multiple levels.*  > salesmen=factor(s2)  > av2=aov(obs~seasons+salesmen*)#~is used to group* ***obs*** *with* ***seasons and salesmen***  > summary(av2)  **Df Sum Sq Mean Sq F value Pr(>F)**  **seasons 2 32 16.00 0.706 0.531**  **salesmen 3 42 14.00 0.618 0.629**  **Residuals 6 136 22.67** |

If given level of significance (α) is less than the p-value (given by Pr(>F)),then we accept corresponding null hypothesis. Here both hypotheses are accepted.

Hence seasons do not differ significantly, as well as salesmen do not differ significantly.