Numerical Examples - Probability & Statistics

Numerical examples are taken from the book – "Business Analytics, The Science of Data-Driven Decision Making" by U. Dinesh Kumar, Wiley Publication (unless stated otherwise).

TABLE OF CONTENTS

Probability Basics

Example 1. Few Real life experiments and sample space

Experiment: Outcome of a football match

Sample Space = S = {Win, Draw, Lose}

Experiment: Predicting customer churn at an individual customer level

Sample Space = S = {Churn, No Churn}

Experiment: Predicting percentage of customer churn

Sample Space = $S = \{X \mid X \in R, 0 \leq X \leq 100\}$, that is X is a real number that can take any value between 0 and 100 percentage.

Experiment: Life of a turbine blade used in an aircraft engine

Sample Space = $S = \{X \mid X \in R, 0 \leq X < \infty\}$, that is X is a real number that can take any value between 0 and ∞.

Example 2. Probability Estimation using Relative Frequency

A website displays 10 advertisements and the revenue generated by the website depends on the number of visitors to the site clicking on any of the advertisements displayed on the website. The data collected by the company has revealed that out of 2500 visitors, 30 visitors clicked on 1 advertisement, 15 clicked on 2 advertisements, and 5 clicked on 3 advertisements. Remaining did not click on any of the advertisements. Calculate

(a) The probability that a visitor to the website will click on an advertisement.

(b) The probability that the visitor will click on at least two advertisements.

(c) The probability that a visitor will not click on any advertisements.

Solution:

(a) Number of customers clicking an advertisement is 50 and the total number of visitors is 2500. Thus, the probability that a visitor to the website will click on an advertisement is

$$\frac{50}{2500} = 0.02$$

(b) Number of customers clicking on at least 2 advertisements is 20. Thus, the probability that a visitor will click on at least 2 advertisements is

$$\frac{20}{2500} = 0.008$$

(c) Probability that a visitor will not click on any advertisement is

$$\frac{2450}{2500} = 0.98$$

Conditional/Bayes

Example 1.

Black boxes used in aircrafts are manufactured by three companies A, B and C. 75% are manufactured by A, 15% by B, and 10% by C. The defect rates of black boxes manufactured by A, B, and C are 4%, 6%, and 8%, respectively. If a black box tested randomly is found to be defective, what is the probability that it is manufactured by company A?

Solution:

Let $P(A)$, $P(B)$, $P(C)$ be events corresponding to the black box being manufactured by companies A, B, and C, respectively, and $P(D)$ be the probability of defective black box. We are interested in calculating the probability $P(A|D)$.

$$P(A \mid D) = \frac{P(D \mid A) \times P(A)}{P(D)}$$

Now $P(D|A) = 0.04$ and $P(A) = 0.75$. Using Eq. (3.14):

$$P(D) = 0.75 \times 0.04 + 0.15 \times 0.06 + 0.10 \times 0.08 = 0.047$$

So

$$P(A \mid D) = \frac{0.04 \times 0.75}{0.047} = 0.6382$$

Example 2.

The proportion of people in a given community who have a certain disease is 0.005. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.99. If a person does not have the disease, the

probability that the test will produce a positive signal is 0.01. If a person tests positive, what is the probability that the person actually has the disease?

(Ref.: Statistics for Engineers and Scientists, Third Edition, William Navidi, McGraw-Hill)

Solution

Let D represent the event that the person actually has the disease, and let $+$ represent the event that the test gives a positive signal. We wish to find $P(D|+)$. We are given the following probabilities:

$$P(D) = 0.005 \qquad P(+|D) = 0.99 \qquad P(+|D^c) = 0.01$$

Using Bayes' rule (Equation 2.27),

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

$$= \frac{(0.99)(0.005)}{(0.99)(0.005) + (0.01)(0.995)}$$

$$= 0.332$$

In the Example, only about a third of the people who test positive for the disease actually have the disease. Note that the test is fairly accurate; it correctly classifies 99% of both diseased and non-diseased individuals. The reason that a large proportion of those who test positive are actually disease-free is that the disease is rare—only 0.5% of the population has it. Because many diseases are rare, it is the case for many medical tests that most positives are false positives, even when the test is fairly accurate. For this reason, when a test comes out positive, a second test is usually given before a firm diagnosis is made.

Joint Probability

Example 1.

At an e-commerce customer service center a total of 112 complaints were received. 78 customers complained about late delivery of the items and 40 complained about poor product quality. (a) Calculate the probability that a customer complaint will be about both late delivery and product quality. (b) What is the probability that a complaint is only about poor quality of the product?

Solution:

Let A = Late delivery and B = Poor quality of the product. Let $n(A)$ and $n(B)$ be the number of cases in favour of A and B. So $n(A) = 78$ and $n(B) = 40$. Since the total number of complaints is 112 (here complaints is treated as the sample space), hence

$$n(A \cap B) = 118 - 112 = 6$$

Probability of a complaint about both delivery and poor product quality is

$$P(A \cap B) = \frac{n(A \cap B)}{\text{Total number of complaints}} = \frac{6}{112} = 0.0535$$

$$\text{Probability that the complaint is only about poor quality} = 1 - P(A) = 1 - \frac{78}{112} = 0.3035$$

Example 2.

A system contains two components, A and B. Both components must function for the system to work. The probability that component A fails is 0.08, and the probability that component B fails is 0.05. Assume the two components function independently. What is the probability that the system functions?

(Ref.: Statistics for Engineers and Scientists, Third Edition, William Navidi, McGraw-Hill)

Solution

The probability that the system functions is the probability that both components function. Therefore:

P*(*system functions*)* = **P***(*A functions and B functions*)*

Since the components function independently,

$\mathbf{P}$(A functions and B functions) = $\mathbf{P}$(A functions)$\mathbf{P}$(B functions)

= [1 - $\mathbf{P}$(A fails)][1 - $\mathbf{P}$(B fails)]

= (1 - 0.08)(1 - 0.05)

= 0.874

Example 3.

Of the microprocessors manufactured by a certain process, 20% are defective. Five microprocessors are chosen at random. Assume they function independently. What is the probability that they all work? (Ref.: Statistics for Engineers and Scientists, Third Edition, William Navidi, McGraw-Hill)

Solution

For **i** = 1, . . . , 5, let **Ai** denote the event that the **i**th microprocessor works. Then

$\mathbf{P}$(all 5 work) = $\mathbf{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap \mathbf{A}_3 \cap \mathbf{A}_4 \cap \mathbf{A}_5)$

= $\mathbf{P}(\mathbf{A}_1)\mathbf{P}(\mathbf{A}_2)\mathbf{P}(\mathbf{A}_3)\mathbf{P}(\mathbf{A}_4)\mathbf{P}(\mathbf{A}_5)$

= $(1 - 0.20)^5$

= 0.328

Central Limit Theorem

Example 1.

It is believed that college students in Bangalore spend on average 80 minutes daily on texting using their mobile phones and the corresponding standard deviation is 25 minutes. Data from a sample of 100 students were collected for calculating the amount of time spent in texting. Calculate the probability that the average time spent by this sample of students will exceed 84 minutes.

Solution:

Using the central limit theorem, the mean of the sampling distribution is 80 and the corresponding standard deviation is $25/\sqrt{100} = 2.5$. The probability that the sample average is more than 84 minutes is given by

$$P\left(Z > \frac{84-80}{2.5}\right) = P(Z > 1.6) = 0.05479$$

Example 2.

The value of insurance claims received at an insurance company follows exponential distribution with mean INR 4200. If a sample of 100 claims is taken from the population, calculate the probability that the total claim will exceed INR 5,00,000.

Solution:

According to CLT, the summation of random variables follows a normal distribution with mean $n\mu$ and standard error $\sigma\sqrt{n}$. Note that for an exponential distribution mean and standard deviation are same.

The probability that the total claim will exceed INR 5,00,000 is

$$P\left(Z > \frac{5,00,000 - n\mu}{\sigma\sqrt{100}}\right)$$

In this case $n = 100$, $\mu = \sigma = 4200$, and Z is the standard normal variate. So

$$P(Z > 5,00,000) = P\left(Z > \frac{5,00,000 - 100 \times 4200}{4200 \times \sqrt{100}}\right) = P(Z > 1.90476) = 0.02841$$

That is, there is 2.8% chance that the total claim will exceed INR 5,00,000.

Estimating the size of the population

Example 1.

A hospital is interested in estimating the average time it takes to discharge a patient after the clearance (discharge note) by the doctor. Calculate the required sample size at a confidence of 95% and maximum error in estimation of 5 minutes. Assume that the population standard deviation is 30 minutes.

Solution:

We know that $\boldsymbol{D} = 5$, s = 30, α = 0.05, and $|\boldsymbol{Z}\alpha/2| = 1.96$ for α = 0.05. we get,

$$n = \left[\frac{Z_{\alpha/2} \times \sigma}{D}\right]^2 = \left[\frac{1.96 \times 30}{5}\right]^2 \approx 138$$

Binomial Distribution

Example 1.

Fashion Trends Online (FTO) is an e-commerce company that sells women apparel. It is observed that about 10% of their customers return the items purchased by them for many reasons (such as size, color, and material mismatch). On a particular day, 20 customers purchased items from FTO. Calculate: (a) Probability that exactly 5 customers will return

the items. (b) Probability that a maximum of 5 customers will return the items. (c) Probability that more than 5 customers will return the items purchased by them. (d) Average number of customers who are likely to return the items.

(e) The variance and the standard deviation of the number of returns.

Solution: In this case, the value of $n = 20$ and $p = 0.1$.

(a) Probability that exactly 5 customers will return the items purchased is

$$P(X=5)=\binom{20}{5}\times(0.1)^5\times(0.9)^{15}=0.03192$$

(b) Probability that a maximum of 5 customers will return the items purchased is

$$P(X\le 5)=\sum_{k=0}^{5}\binom{20}{k}\times(0.1)^k\times(0.9)^{20-k}=0.9887$$

(c) Probability that more than 5 customers will return the product is

$$P(X>5)=1-P(X\le 5)=1-\sum_{k=0}^{5}\binom{20}{k}\times(0.1)^k\times(0.9)^{20-k}=1-0.9887=0.0113$$

(d) The average number of customers who are likely to return the items is

$$E(X) = n \times p = 20 \times 0.1 = 2$$

(e) Variance of a binomial distribution is given by

$$\text{Var}(X) = n \times p \times (1 - p) = 20 \times 0.1 \times 0.9 = 1.8$$

and the corresponding standard deviation is 1.3416.

Example 2.

Die Another Day (DAD) hospital recruit's nurses frequently to manage high attrition among the nursing staff. Not all job offers from DAD hospital are accepted. Based on the past recruitment data, it was estimated that only 70% of offers rolled out by DAD

hospital are accepted. (a) If 10 offers are made, what is the probability that more than 5 and less than 8 candidates will accept the offer from DAD hospital? (b) During March 2017, DAD required 14 new nurses to manage attrition. What should be the number of offers made by DAD hospital so that the average number of nurses accepting the offer is 14?

Solution:

(a) Probability that the number of accepted offers will be greater than 5 and less than 8 out of 10 offers is given by

$$P(5<X<8)=P(X=6)+P(X=7)=\binom{10}{6}\times 0.7^6\times 0.3^4+\binom{10}{7}\times 0.7^7\times 0.3^3=0.4669$$

(b) For binomial distribution, $E(X) = n \times p$, and we have to find the number of offers such that on average 14 nurses accept the offer:

$$n\times p=14\Rightarrow n=\frac{14}{p}=\frac{14}{0.7}=20$$

That is, the hospital should make 20 offers to ensure that the expected number of accepted offers is 14.

Poisson Distribution

Example 1.

On average, 20 customers per day cancel their order placed at Fashion Trends Online. Calculate the probability that the number of cancellations on a day is exactly 20 and the probability that the maximum number of cancellations is 25.

Solution:

The probability that the number of cancellations is exactly 20 is given by

$$P(X=20)=\frac{e^{-20}20^{20}}{20!}=0.0888$$

Probability that the maximum number of cancellation will be 25 is given by

$$P(X\leq 25)=\sum_{k=0}^{25}\frac{e^{-20}20^{k}}{k!}=0.8878$$

Example 2.

The number of calls arriving at a call center follows a Poisson distribution at 10 per hour. Calculate the probability that the number of calls over a 3-hour period will exceed 30.

Solution:

Since the average calls per hour is 10 ($\lambda = 10$), and we are interested in finding the calls over 3 hours, the mean number of calls over 3 hours is $\lambda t = 30$. Probability that the number of calls will be more than 30 is given by

$$P(X>30)=1-P(X\leq 30)=1-\sum_{k=0}^{30}\frac{e^{-\lambda t}\times(\lambda t)^{k}}{k!}=1-\sum_{k=0}^{30}\frac{e^{-30}\times(30)^{k}}{k!}=0.4516$$

Geometric Distribution

Example 1.

Local Dhaniawala (LD) is an online grocery store and has an innovative feature which predicts whether the customer has forgotten to buy an item which is very common among customers of grocery items. The probability that a customer buys milk in each shopping visit is 0.2.

(a) Calculate the probability that the customer's first purchase of milk happens during the 5^{th} visit.
(b) Calculate the average time between purchases of milk.
(c) If a customer has not purchased milk during the past 3 shopping visits, what is the probability that the customer will not buy milk for another 2 visits?

Solution:

(a) Probability that the customer's first purchase of milk happens on 5^{th} trip is given by

$$P(X=5)=(1-0.2)^4\times 0.2=0.08192$$

(b) The average time between purchase of milk is

$$E(X)=\frac{1}{p}=\frac{1}{0.2}=5$$

(c) Given that a customer has not purchased milk during the past 3 shopping visits, the probability that the customer will not buy for another 2 visits is given by

$$P(X>3+2\,|\,X>3)=P(X>2)=(1-p)^2=(1-0.2)^2=0.64$$

Normal Distribution

Example 1.

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution. Note NORMSDIST(**Z**) in the solution, which is a function in

Microsoft Excel to calculate CDF value of a standard normal distribution. (a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily? (b) What proportion of customers are spending less than 20 minutes? (c) What proportion of customers are spending between 50 minutes and 100 minutes?

Solution:

It is given that $\mu = 68$ minutes and $\sigma = 12$ minutes.

(a) Proportion of customers spending more than 90 minutes is given by

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90)$$

The standard normal random variable value for $X = 90$ is given by

$$Z = \frac{x - \mu}{\sigma} = \frac{90 - 68}{12} = 1.8333$$

That is, $F(X = 90) = F(Z = 1.8333)$. From standard normal distribution table, we can get the value of $F(Z)$ for $Z = 1.8333$. The area under the standard normal distribution curve for $Z = 1.8333$ is 0.9666. Thus,

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90) = 1 - 0.9666 = 0.0334$$

Alternatively, using Excel, we get

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - \text{Normdist}(90, 68, 12, \text{true}) = 0.0334$$

(b) Proportion of customers spending less than 20 minutes is

$$P(X \leq 20) = F(20)$$

Using Excel function, we have Normdist(20, 68, 12, true) = 3.1671×10^{-5}

(c) Proportion of customers spending between 50 and 100 minutes is given by

$$\begin{aligned} P(50 \leq X \leq 100) &= F(100) - F(50) \\ &= \text{Normdist}(100, 68, 12, \text{true}) - \text{Normdist}(50, 68, 12, \text{true}) \\ &= 0.9293 \end{aligned}$$

Example 2.

At Die Another Day (DAD) hospital, nurses are given an additional bonus of INR 1,00,000 if they stay for more than 36 months with DAD hospital. The average stay of nurses follows a normal distribution with an average of 28 months and the corresponding

standard deviation is 4.8 months. Calculate (a) The expected number of nurses who will be given bonus and the value of bonus that will be given if 50 new nurses join DAD hospital in the current month, (b) What will be the additional amount paid if DAD hospital changes the policy that they will give bonus if the stay exceeds 24 months? What assumptions are made in this case?

Solution:

a) Expected number of nurses and the value of bonus:

Expected number of nurses who will be getting bonus = $50 \times P(X \geq 36)$

$$P(X \geq 36) = 1 - \text{Normdist}(36, 28, 4.8, \text{true}) = 0.04779$$

Expected number of nurses who will be getting bonus = $50 \times 0.04799 = 2.389518$

Expected value of bonus given = $50 \times P(X \geq 36) \times 100{,}000 = \text{INR } 238951.76$

(b) The additional bonus given is

$$50 \times 1{,}00{,}000 \times \left[\text{Normdist}(36, 28, 4.8, \text{true}) - \text{Normdist}(24, 28, 4.8, \text{true})\right] = 3749406$$

The major assumption here is that the policy change is unlikely to change the attrition behaviour of the nurses, which may not be true. Since the nurses now know that if they stay for 24 months, they will get the bonus, the distribution parameter values are likely to change.

Maximum Likelihood

Example 1.

A sample of ten new bike helmets manufactured by a certain company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not. Let p = P (flawed helmet), i.e., p is the proportion of all such helmets that are flawed. Define (Bernoulli) random variables $X_1, X_2, \ldots, X_{10}$ by

$$X_1 = \begin{cases} 1 \text{ if 1st helmet is flawed} \\ 0 \text{ if 1st helmet isn't flawed} \end{cases} \quad \cdots \quad X_{10} = \begin{cases} 1 \text{ if 10th helmet is flawed} \\ 0 \text{ if 10th helmet isn't flawed} \end{cases}$$

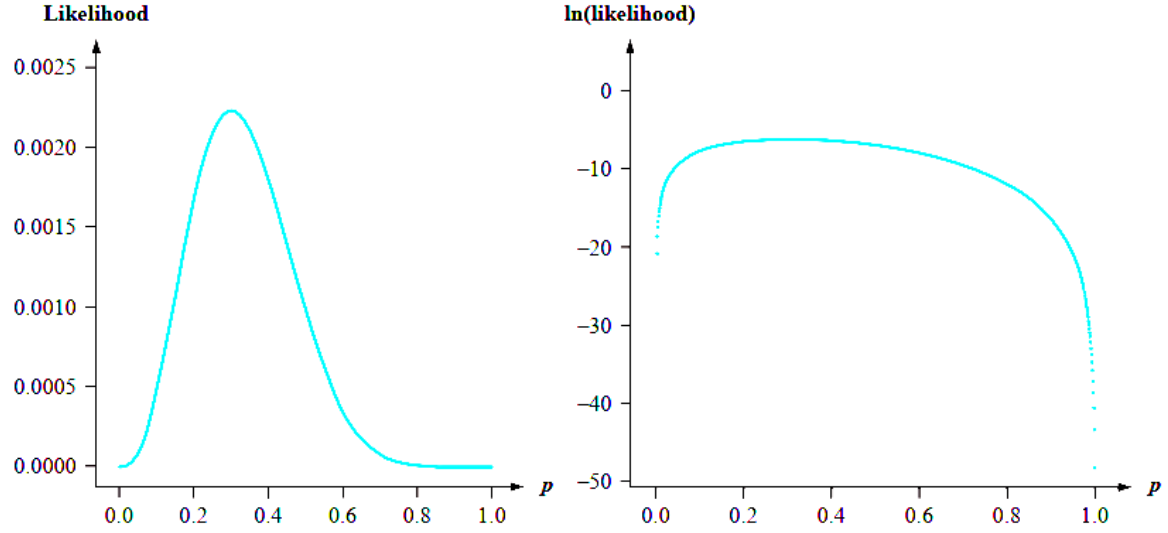
Then for the obtained sample, $X_1=X_3=X_{10}=1$ and the other seven X_i's are all zero. The probability mass function of any particular X_i is $p^{Xi}(1- p)^{1-Xi}$, which becomes p if $X_i=1$ and

1-p when X_i=0. Now suppose that the conditions of various helmets are independent of one another. This implies that the X_i's are independent, so their joint probability mass function is the product of the individual pmf's. Thus the joint pmf evaluated at the observed Xi's is:

$f(x_1, \ldots, x_{10}; p) = p\,(1-p)\,p \ldots p = p^3(1-p)^7$ ----------(1)

Suppose that p=0.25. Then the probability of observing the sample that we actually obtained is $(0.25)^3(0.75)^7$=0.002086. If instead p=0.50, then this probability is $(0.50)^3(.50)^7$=0.000977. For what value of p is the obtained sample most likely to have occurred? That is, for what value of p is the joint pmf (1) as large as it can be? What value of p maximizes (1)? Figure (a) given below, shows a graph of the likelihood

(1) as a function of p. It appears that the graph reaches its peak above p = 0.3 the proportion of flawed helmets in the sample. Figure (b) given below shows a graph of the natural logarithm of (1); since ln[g(u)] is a strictly increasing function of g(u), finding u to



maximize the function g(u) is the same as finding u to maximize ln[g(u)]

Figure (a) Graph of the likelihood (joint pmf) (1) from Example (b) Graph of the natural logarithm of the likelihood

We can verify our visual impression by using calculus to find the value of p that maximizes (1). Working with the natural log of the joint pmf is often easier than

working with the joint pmf itself, since the joint pmf is typically a product so its logarithm will be a sum. Here

$$\ln[f(x_1, \ldots, x_{10}; p)] = \ln[p^3(1-p)^7] = 3\ln(p) + 7\ln(1-p)$$

Thus

$$\frac{d}{dp}\{\ln[f(x_1, \ldots, x_{10}; p)]\} = \frac{d}{dp}\{3\ln(p) + 7\ln(1-p)\} = \frac{3}{p} + \frac{7}{1-p}(-1)$$

$$= \frac{3}{p} - \frac{7}{1-p}$$

[the (-1) comes from the chain rule in calculus]. Equating this derivative to 0 and solving for p gives 3(1-p) =7p, from which 3=10p and so p=3/10=0.3 as conjectured. That is, our point estimate is $\check{p}$=0.30. It is called the maximum likelihood estimate because it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample. In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from figure (b) above.

(Ref. of above example: Probability and statistics for Engineering and science, Eighth Edition, Jay L. Devore, Cengage Learning)

Z-test

Example 1. One-sample Z-test for mean

An agency based out of Bangalore claimed that the average monthly disposable income of families living in Bangalore is greater than INR 4200 with a standard deviation of INR 3200. From a random sample of 40,000 families, the average disposable income was estimated as INR 4250. Assume that the population standard deviation is INR 3200. Conduct an appropriate hypothesis test at 95% confidence level (α = 0.05) to check the validity of the claim by the agency.

Solution:

Note: In contexts such as this, we set alternative hypothesis as the statement that we would like to prove.

Claim: Average disposable income is more than INR 4200.

Let m and s denote the mean and standard deviation in the population. The corresponding null and alternative hypotheses are

$$H_0: \mu \leq 4200$$
$$H_A: \mu > 4200$$

Since we know the population standard deviation, we can use the *Z*-test. The corresponding *Z*-statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{4250 - 4200}{3200 / \sqrt{40000}} = 3.125$$

This is a right-tailed test. The corresponding **Z**-critical value at **α**= 0.05 for right-tailed test is approximately 1.64 [in Excel NOMRSINV (1 - **α**) that is NORMSINV (0.95) gives the critical value for the right-tailed test]. Since the calculated **Z**-statistic value is greater than the **Z**-critical value, we reject the null hypothesis. The corresponding p-*value* = 0.00088 [*p*-value in Excel is given by 1 - NORMSDIST (*Z*-statistic value), that is 1 – NORMSDIST (3.125) in this case]. The critical value, *Z*-statistic value, and the corresponding *p*-value are shown in Figure below.

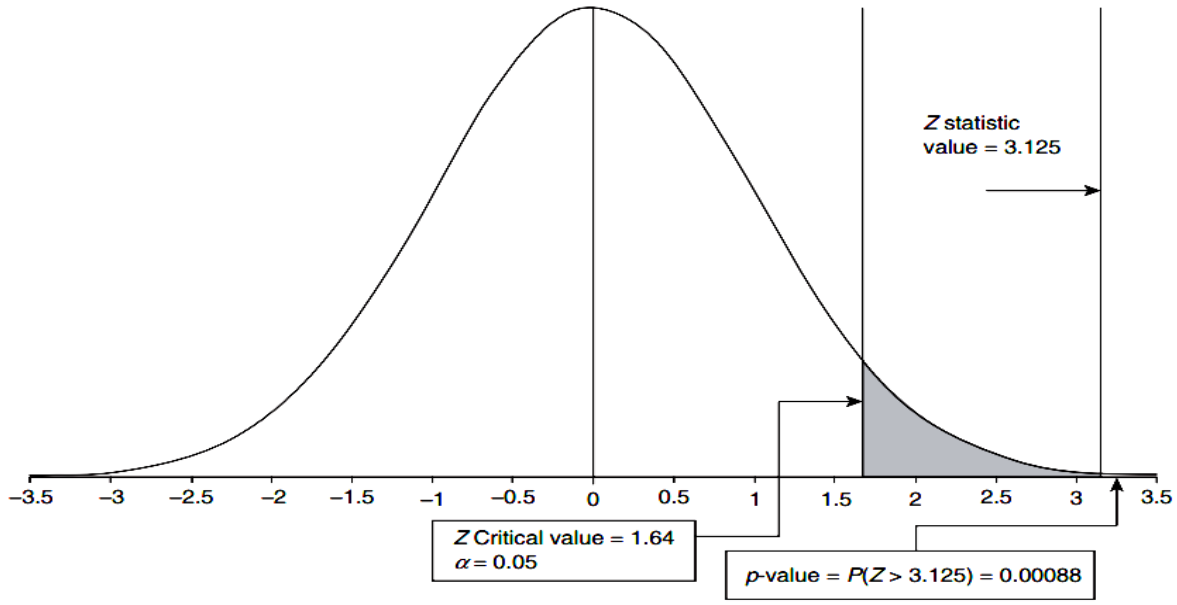


Figure: Critical value, Z-statistic value, and corresponding p-value.

IMPORTANT

Z-statistic measures the standardized difference between estimated value of mean and the hypothesis value of mean. Z = 3.125 implies that the sample mean is at 3.125 standard deviations away from the hypothesized population mean given that the null hypothesis is true

Example 2. One-sample Z-test for mean

A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents. Table below: shows processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level **α**= 0.05 to verify the claim made by the passport office.

Table: Passport processing time

16	16	30	37	25	22	19	35	27	32
34	28	24	35	24	21	32	29	24	35
28	29	18	31	28	33	32	24	25	22
21	27	41	23	23	16	24	38	26	28

Solution:

Null and alternative hypotheses in this case are given by

$$H_0: \mu \geq 30$$
$$H_A: \mu < 30$$

From the data in Table 6.6, the estimated sample mean is 27.05 days.

The standard deviation of the sampling distribution $\sigma / \sqrt{n} = 12.5 / \sqrt{40} = 1.9764$.

The value of Z-statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{27.05 - 30}{12.5 / \sqrt{40}} = -1.4926$$

The critical value of left-tailed test for **α** = 0.05 is –1.644. Since the critical value is less than the **Z**-statistic value, we fail to reject the null hypothesis. The **p**-value for **Z** = -1.4926 is 0.06777 which is greater than the value of **α**. That is, there is no strong evidence against null hypothesis so we retain the null hypothesis, which is **μ** ≥ 30. Figure shows the calculated **Z**-statistic value and the rejection region.

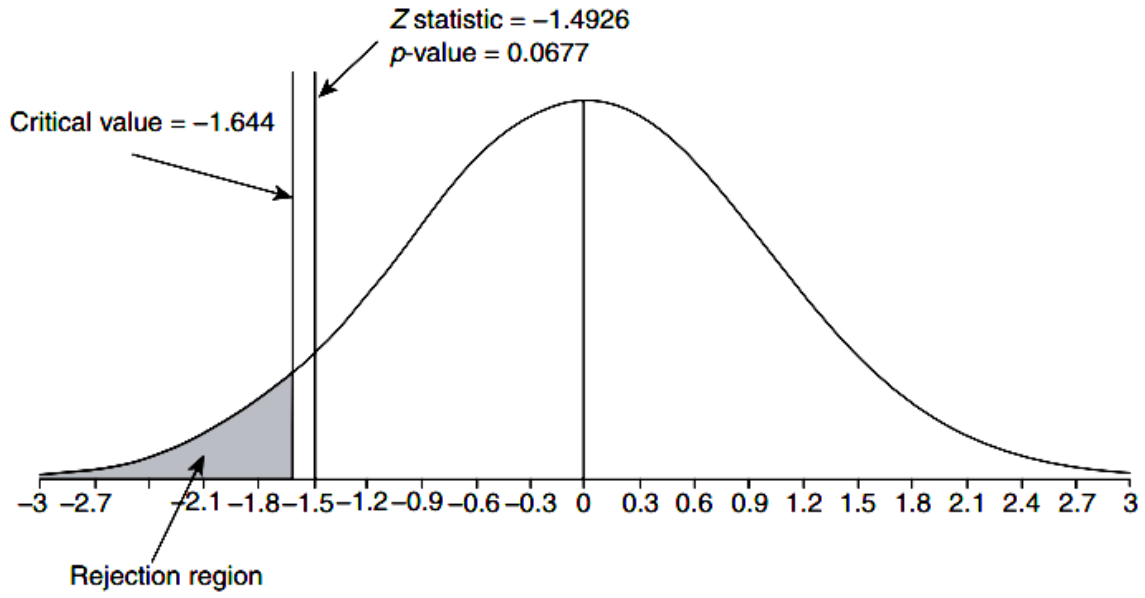


Figure: Left-tailed test for above Example

Example 3. One- sample Z-test for mean

According to the company IQ Research, the average Intelligence Quotient (IQ) of Indians is 82 derived based on a research carried out by Professor Richard Lynn, a British Professor of Psychology, using data collected from 2002 to 2006 (Source: IQ Research). The population standard deviation of IQ is estimated as 11.03. Based on a sample of 100 people from India, the sample IQ was estimated as 84. **(a)** Conduct an appropriate hypothesis test at **α** = 0.05 to validate the claim of IQ Research (that average IQ of Indians is 82). **(b)** Ministry of education believes that the IQ is more than 82. If the actual IQ (population mean) of Indians is 86, calculate the Type II error and the power of hypothesis test.

Solution:

(a) Hypothesis test: It is given that $\mu = 82$, $\sigma = 11.03$, $n = 100$, and $\bar{X} = 84$.

The null and alternative hypotheses in this case are:

$$H_0: \mu = 82$$

$$H_A: \mu \neq 82$$

Since the direction of alternative hypothesis is both ways, we have a two-tailed *t*-test. The test statistics is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{84 - 82}{11.03 / \sqrt{100}} = 1.8132$$

For a two-tailed test, the critical values at **α/2** = 0.025 are -1.96 and 1.96 [in Excel NORMSINV (0.025) = -1.96 and NORMSINV (1 - 0.025) = 1.96]. Since the calculated **Z**-statistic value is within the critical values, we fail to reject the null hypothesis (retain the null hypothesis). Figure 6.7 shows the rejection regions and the **Z**-statistic value in this case. Since the **Z**-statistic value is 1.8132 and falls on the right tail, we first calculate normal distribution beyond 1.8132 which is equal to 0.0348. Since this is a two-tailed

test, the **p**-value is twice the area to the right side of the **Z**-statistic value, which is = 0.0698, that is the **p**-value in this case is 0.0698.

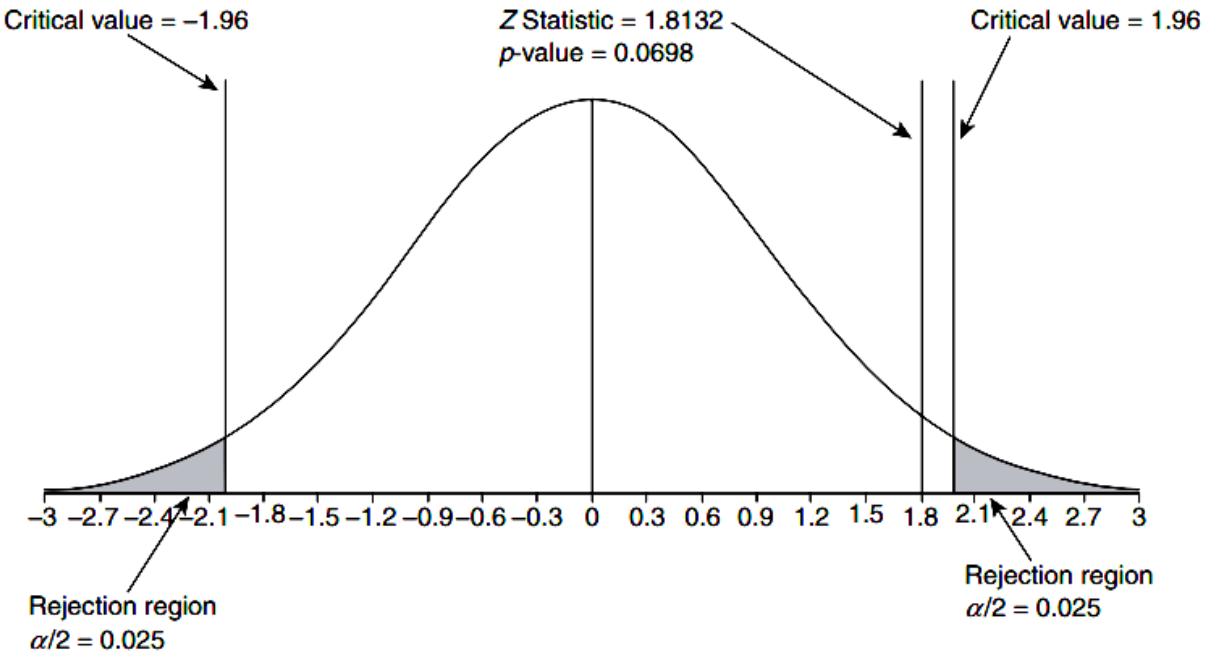


Figure: Z-statistic, critical values, and the rejection region for above Example

In a two-tailed test, the p-value is two times the tail area.

(a) Calculating Type II Error and Power of Test: In this case, the null and alternative hypotheses are

$$H_0: \mu \leq 82$$
$$H_A: \mu > 82$$

Note that ministry of education believes that the average IQ is 86 (thus we have to carry out a right-tailed test). Type II error is the conditional probability of retaining a null hypothesis when it is false, that is $P(\text{retaining } H_0 \mid H_0 \text{ is false})$.

The mean and standard deviation of Z-statistic in null hypothesis are 82 and 1.103, respectively. For the standard normal distribution the critical value for a right tailed test when $\alpha = 0.05$ is 1.644. The corresponding critical value for the normal distribution $N(82, 1.103)$ is

$$X_{\text{critical}} = \mu + Z_\alpha \times \sigma / \sqrt{n} = 82 + 1.644 \times 1.103 = 83.8133$$

That is, under normal distribution $N(82, 1.103)$, the region beyond 83.8133 is the rejection region (rejection of null hypothesis).

Now consider the normal distribution **N** (86, 1.103). Area under this normal distribution may take values below 83.8133 which is region of retaining the null hypothesis, although the actual mean in this case is 86. Thus, we will be retaining the null hypothesis when it is incorrect resulting in Type II error, **β** (Figure below). For the normal distribution **N** (86, 1.103), the probability of the variable taking value less than 83.8133 (the critical value) is given by

$$P(X \leq 83.8133) = P\left(Z \leq \frac{83.8133 - 86}{1.103}\right) = 0.0237$$

That is, the Type II error $\beta = 0.0237$

The power of test, $1 - \beta = 1 - 0.0237 = 0.9763$

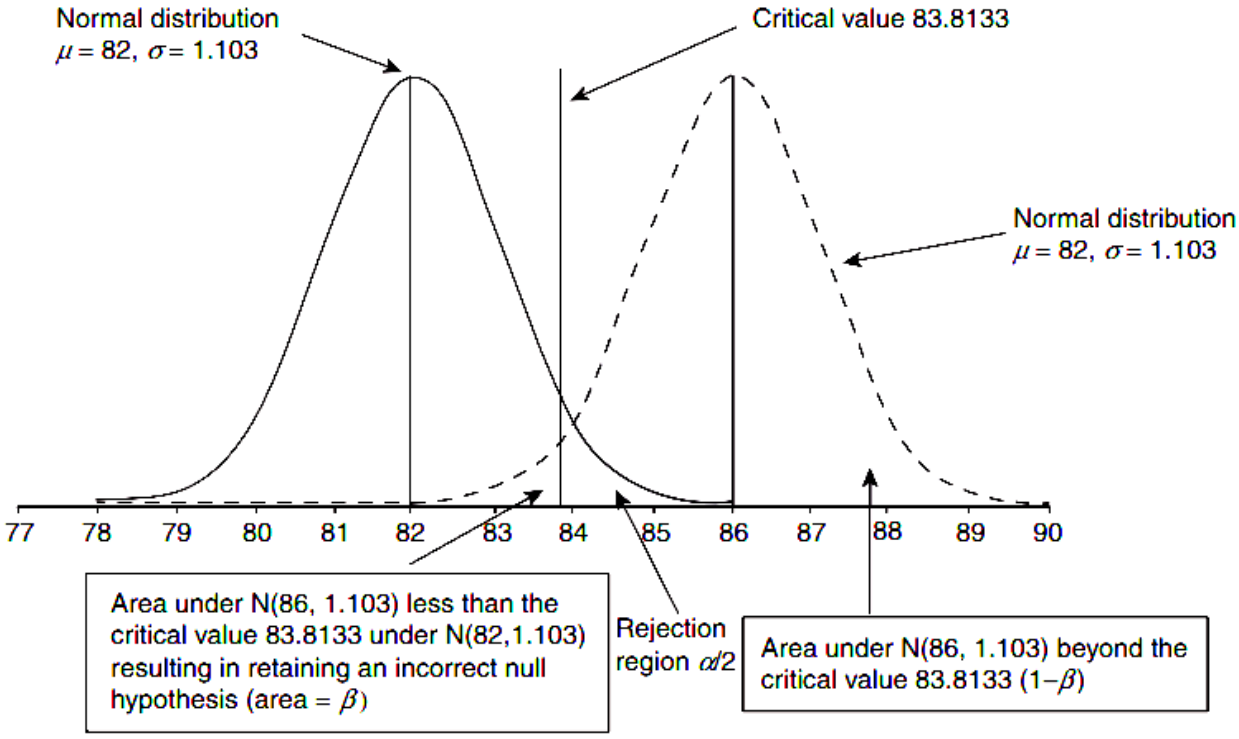


Figure: Type II error and power of hypothesis test

Example 4. Two-sample Z-test for mean

The Dean of St Peter School of Management Education (SPSME) believes that the graduating students with specialization in Marketing earn at least INR 5000 more per month than the students with specialization in Operations Management. To verify his belief, the Dean collected a sample data from his graduating students, given in Table Conduct an appropriate hypothesis test at $\alpha = 0.05$ to check whether the difference in monthly salary is at least 5000 more for students with marketing specialization compared

to operations specialization. Assume that the salary of students with marketing specialization and operations specialization follow normal distribution.

Table: Sample values on marketing and operations students

Specialization	Sample Size	Estimated Mean Salary (in Rupees) per Month	Population Standard Deviation
Marketing	120	67,500.00	7,200
Operations	45	58,950.00	4,600

Solution:

We have $n_1 = 120$, $n_2 = 45$, $\bar{X}_1 = 67,500$, $\bar{X}_2 = 58,950$, $\sigma_1 = 7,200$ and $\sigma_2 = 4,600$.

The null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 5000$$
$$H_A: \mu_1 - \mu_2 > 5000$$

The corresponding test statistic value is

$$Z = \frac{(67500 - 58950) - 5000}{\sqrt{\frac{7200^2}{120} + \frac{4600^2}{45}}} = \frac{3550}{949.85} = 3.7374$$

The critical value of Z at $\alpha = 0.05$ is 1.64 [= NORMSINV(1 − 0.05)]. Since the Z-statistic value is higher than the Z-critical value, we reject the null hypothesis. The corresponding p-value is 9.29×10^{-05}.

Example 5. One-sample Z-test for proportion

According to a study exactly 12% of gift cards purchased from e-commerce portals are never used. The manager of an e-commerce company wanted to test whether this claim is true. She collected data of 250 gift card purchases and found that 22 gift cards were not used till its expiry date.

(a) Conduct an appropriate hypothesis test at 5% significance to check whether the claim that exactly 12% gift cards are never used is true or not.

(b) Calculate the 95% confidence interval for the proportion of gift cards that are not used.

Solution:

(a) The estimated value of proportion of gift cards not used is

$$\hat{p} = \frac{22}{250} = 0.088$$

$$n \times \hat{p} \times (1 - \hat{p}) = 250 \times 0.088 \times (1 - 0.088) = 20.064 > 10$$

so we can use $\hat{p}$ to calculate the population standard deviation. The initial claim is that the percentage of unused gift cards is equal to 12%. The null and alternative hypotheses are

$$H_0 : p = 0.12$$

$$H_A : p \neq 0.12$$

The value of the Z-statistic is given by

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.088 - 0.12}{\sqrt{\frac{0.12 \times (1-0.12)}{250}}} = -1.557$$

Note that the critical values are −1.96 and 1.96 (two-tailed test). Since the calculated value of Z is not part of the rejection region (greater than −1.96 and less than 1.96), we retain the null hypothesis that p = 0.12, the corresponding p-value is 0.1195. This is a two-tailed test, and thus the p-value is 2 × 0.05973 ≈ 0.1195.

(b) The 95% confidence interval for the proportion is given by

$$\left[\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}\right] = \left[0.088 \pm 1.96 \times \sqrt{\frac{0.088 \times (1-0.088)}{20}}\right] = [0.0528,\ 0.1231]$$

From the confidence interval estimate for the proportion of unused gift cards we can infer that the proportion of unused gifts is likely to lie between 0.0528 and 0.1231 at 95% confidence level.

Example 6. Two-sample Z-test for proportion

The marketing manager of a company believes that the non-affluent customers are sensitive to discounts compared to affluent customers. To validate this hypothesis, discount coupons were sent to non-affluent and affluent customers and the data is provided in Table below. Use an appropriate hypothesis to check whether there is any difference in proportion of customers who use discount coupons at $\alpha = 0.05$.

Table: Data related to increase in heights from different groups

Group	Sample Size	Number of Customers using Discount Coupons	Estimated Proportion
Non-Affluent	500	145	0.29
Affluent	300	42	0.14

Solution:

The null and alternative hypothesis are

$$H_0 : p_1 = p_2$$

$$H_A : p_1 \neq p_2$$

From the above Table, the estimated values of proportions are $\hat{p}_1 = 0.29$ and $\hat{p}_2 = 0.14$.

The pooled proportion is (assuming null hypothesis is true)

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{500 \times 0.29 + 300 \times 0.14}{500 + 300} = 0.2338$$

The test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.29 - 0.14) - 0}{\sqrt{0.2338\left(\frac{1}{500} + \frac{1}{300}\right)}} = \frac{0.15}{0.03091} = 4.8528$$

The Z-critical at $\alpha = 0.05$ for two-tailed test is 1.96. Since the calculated value of Z is more than the Z-critical value, we reject the null hypothesis. That is, the non-affluent customers are sensitive to coupons, that is they use more coupons.

t-Test

Example 1. One-sample t-test for mean

Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing a Bollywood movie. The industry believes that the production house will require at least INR 500 million (50 crore) on average. It is assumed that the Bollywood movie production cost follows a normal distribution. Production cost of 40 Bollywood movies in millions of rupees are shown in Table below. Conduct an appropriate hypothesis test at $\alpha = 0.05$ to check whether the belief about average production cost is correct.

Table. Production cost of Bollywood movies

601	627	330	364	562	353	583	254	528	470
125	60	101	110	60	252	281	227	484	402
408	601	593	729	402	530	708	599	439	762
292	636	444	286	636	667	252	335	457	632

Solution:

It is given that the production cost of Bollywood movies follows a normal distribution; however, the standard deviation of the population is not known and we to estimate the

standard deviation value from the sample. Thus, we have to use the t-test for testing the hypothesis. From the sample data in Table above, we get the following values:

$$n = 40,\ \bar{X} = 429.55, \text{ and } S = 195.0337$$

The null and alternative hypotheses are

$$H_0: \mu \leq 500$$

$$H_A: \mu > 500$$

The corresponding test statistic is

$$t\text{-statistic} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{429.55 - 500}{195.0337/\sqrt{40}} = -2.2845$$

Note that this is a one-tailed test (right-tailed) and the critical **t**-value at **α** =0.05 under right-tailed test, **t** critical = 1.6848 [in Excel TINV (2**α**, **df**) will return right-tailed critical value at significance of **α**, in this example **α** = 0.05, the corresponding critical **t**-value using Excel function is TINV (0.1, 39) = 1.6848, that is the critical value is 1.6848]. Since **t**-statistic value is less than the critical **t**-value, we retain the null hypothesis. See the Figure below.

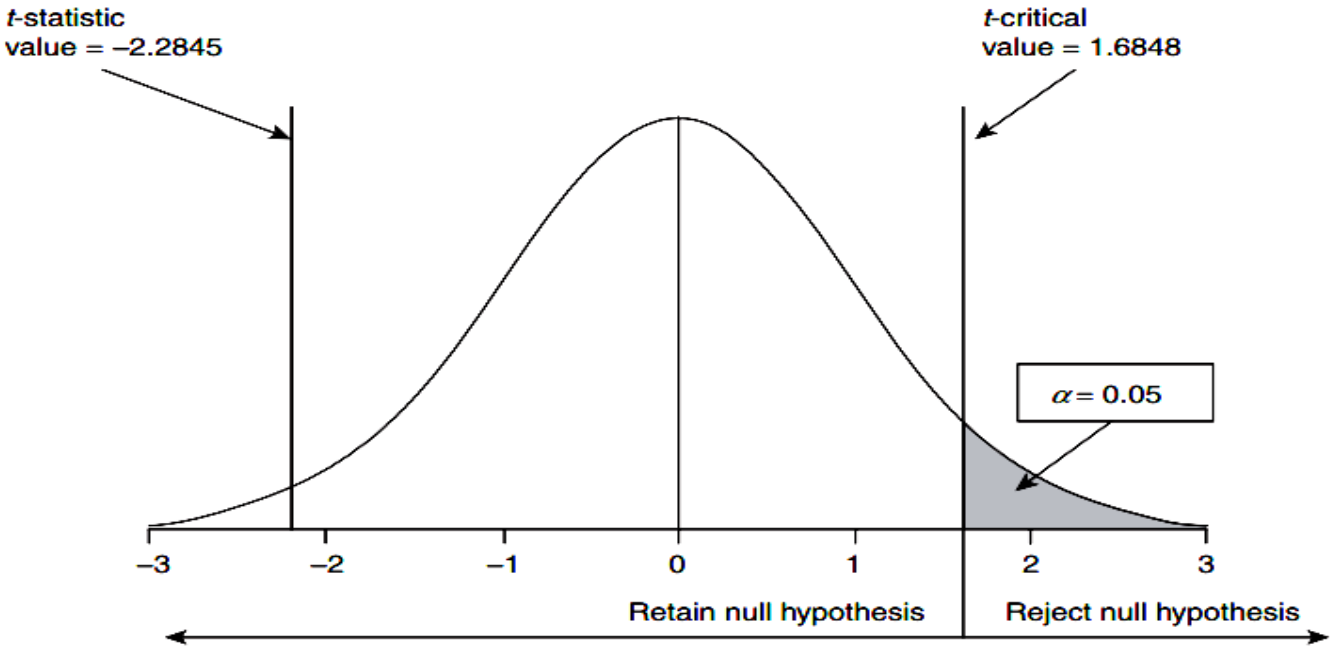


Figure: Critical value, t-statistic value for t-test in Example above.

Example 2. One-sample t-test for mean

According to statistics released by the Department of Civil Aviation, the average delay of flights is equal to 16.8 minutes, flight delays are assumed to follow a normal distribution. However, from a sample of 50 flights, the average delay was estimated to be 19.5 minutes and the sample standard deviation was 6.6 minutes. Conduct a hypothesis test to disprove the claim that the average delay is equal to 16.8 minutes at **α**= 0.01.

Solution:

Given $n = 50,\ \bar{X} = 19.5,\ S = 6.6$

Null and alternative hypotheses are

$$H_0: \mu = 16.8$$

$$H_A: \mu \neq 16.8$$

The corresponding *t*-statistic value is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{19.5 - 16.8}{6.6/\sqrt{50}} = 2.8927$$

The critical **t**-value for two-tailed **t**-test when **α** = 0.01 and degrees of freedom = 49 is 2.67 [in Excel, TINV (0.01, 49) = 2.68 or T.INV.2T (0.01, 49) = 2.68]. Since the calculated **t**-statistic value is greater than the **t**-critical value, we reject the null hypothesis. The corresponding **p**-value is 0.0057 [in excel T. DIST (**t**-statistic value, degrees of freedom, tails) returns the **p**-value, T.DIST.2T (2.8927, 49) = 0.0057]. The values of **t**-statistic, **t**-critical value, rejection and retention regions are shown in Figure below.

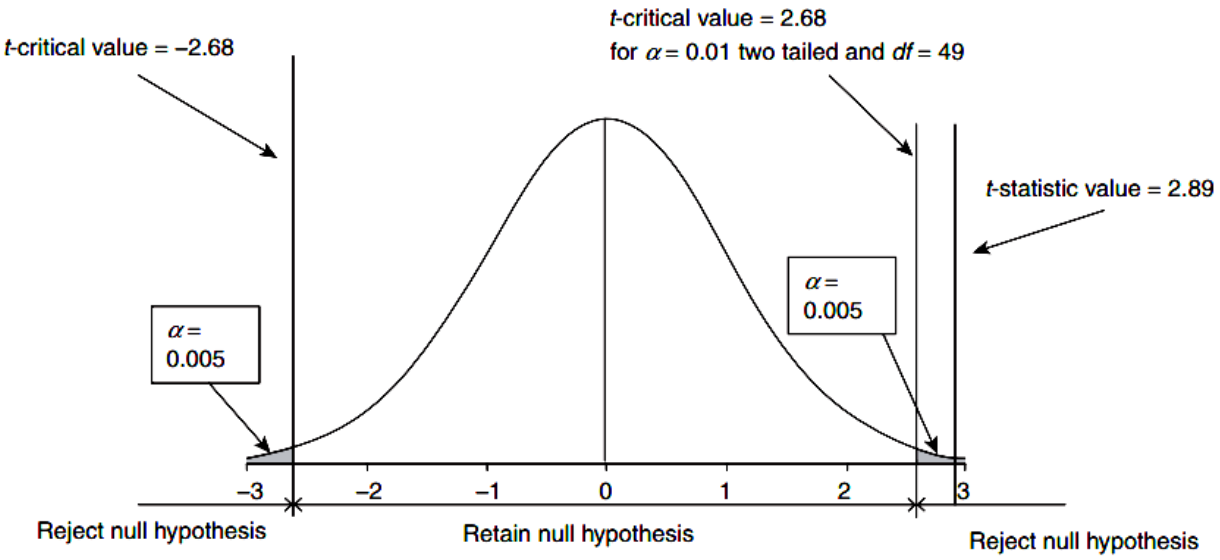


Figure: t-statistic, t-critical, rejection and acceptance regions for Example above.

Example 3. One-sample t-test for mean

A researcher believes that people drink more coffee on Mondays than other days of the week. Based on a sample of 50 coffee drinkers, the mean difference was estimated as 14 ml and the corresponding standard deviation was 8.5 ml. Conduct an appropriate hypothesis test at **α** = 0.1 to check the claim that people drink on average 10 ml more coffee on Mondays compared to other days of the week.

Solution:

We are given $n = 50, D = 14, S_d = 8.5$, and $\mu_d = 10$. The null and alternative hypotheses are

$$H_0: \mu_d \leq 10$$

$$H_A: \mu_d > 10$$

The test statistic is

$$t = \frac{D - \mu_d}{S_d / \sqrt{n}} = \frac{14 - 10}{8.5 / \sqrt{50}} = 3.3275$$

Critical value of **t** for **α** = 0.1 and **df** = 49 is 1.2990 [in Excel T.INV (1 - 0.1, 49) = 1.2990]. Since **t**-statistic value is greater than **t**-critical value, we reject the null

hypothesis. That is, there is evidence from the data that people drink at least 10 ml more coffee on Mondays than other days. The corresponding **p**-value is 0.000834 [in Excel, T. DIST.RT(3.3275, 49) = 0.000834].

Example 4. Two-sample t-test for mean

A company makes a claim that children (in the age group between 7 and 12) who drink their health drink will grow taller than the children who do not drink that health drink. Data in Table below shows average increase in height over a one-year period from two groups: one drinking the health drink and the other not drinking the health drink. At α = 0.05, test whether the increase in height for the children who drink the health drink is at least 1.2 cm.

Table: Data related in increase in heights from different groups

Group	Sample Size	Increase in Height (in cm) during the Test Period	Standard Deviation Estimated from Sample
Drink health drink	80	7.6 cm	1.1 cm
Do not drink health drink	80	6.3 cm	1.3 cm

Solution:

We have $n_1 = 80$, $n_2 = 80$, $\bar{X}_1 = 7.6$, $\bar{X}_2 = 6.3$, $\sigma_1 = 1.1$, and $\sigma_2 = 1.3$. (Note that the sample standard deviations need not be same for samples although the population standard deviations are same.)

The null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 1.2$$

$$H_A: \mu_1 - \mu_2 > 1.2$$

Pooled variance is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} = \frac{79 \times 1.1^2 + 79 \times 1.3^2}{80 + 80 - 2} = 1.45$$

The **t**-statistic is:

$$t = \frac{(7.6 - 6.3) - 1.2}{\sqrt{1.45\left(\frac{1}{80} + \frac{1}{80}\right)}} = 0.5252$$

The **t**-critical value for one-tailed **t**-test when **α**= 0.05 and degrees of freedom = 158 (80 + 80 – 2) is 1.6546. Since the calculated **t**-statistic value is less than **t**-critical value we retain the null hypothesis. That is, the difference between two groups is less than 1.2 and the corresponding right-tailed test has a **p** -value of 0.3.

Hypothesis Test for Equality of Population Variances

Example 1.

Preetha Dallal is an investment advisor and she believes that the variance of stock prices of manufacturing companies and information technology companies are the same. To verify the claim, variances of stock prices from these two sectors are collected and are shown in Table below. Conduct an appropriate test at **α**= 0.1 to check whether the variances of stock prices in two industry sectors are equal or not.

Table: Data related in variance in stock prices from different groups

Group	Sample Size	Variance Estimated from Sample
Manufacturing Firms	80	42
Information Technology Firms	52	36

Solution:

Given $S_1^2 = 42$ and $S_2^2 = 36$. The null and alternative hypotheses are

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

The corresponding F-statistic is given by

$$F_{(n_1-1,\, n_2-1)} = F_{(79,\, 51)} = \frac{42}{36} = 1.1666$$

The left and right critical values are given by 0.6635 and 1.5407, respectively. Since the calculated value is between the two critical values, we will retain the null hypothesis. In Excel, the right critical value of the **F**-test is given by F.INV. RT (**α**/2, **df-N**, **df-D**), where **df-N** is the degrees of freedom in the numerator and **df-D** is the degrees of freedom in the denominator. The left critical value in Excel is given by (1/FINV (**α**/2, **df-D**, **df-N**)). That is, we reverse the degrees of freedom and calculate FINV and take the inverse to get the left critical value under **F**-test.

Figure below shows the $F_{(79,\, 51)}$ -distribution with left and right critical values. and the calculated **F**-statistic value.

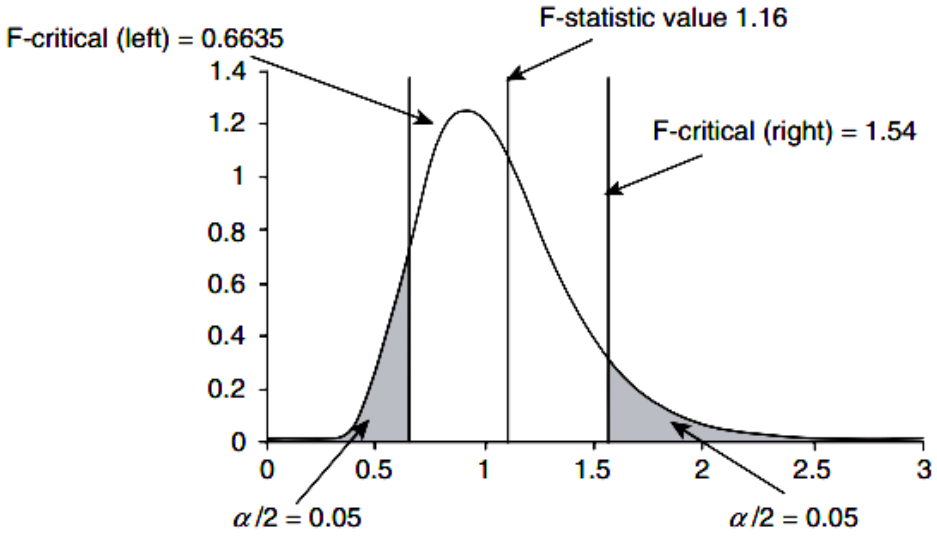


Figure: F distribution with left and right critical values.

An alternative and frequently used approach to test the variances of two groups is through right tailed **F** test to simplify the calculations. In such cases, the null and alternative hypotheses are set as

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_A: \sigma_1^2 > \sigma_2^2$$

While performing a right-tailed test, we choose sample with larger variance as S_1^2. In Example 6.13, we will designate $S_1^2 = 42$ and $S_2^2 = 36$. So the *F*-statistic for right-tailed test is

$$F_{(79,\,51)} = \frac{42}{36} = 1.1666$$

The critical **F**-value for the right-tailed test is 1.39 [in Excel, FINV (**α**, **df**1, **df**2) or F.INV.RT(**α**, **df**1, **df**2) functions can be used for getting the **F**-critical value, **df**1 and **df**2 are degrees for freedom for numerator and denominator, respectively]. Since the calculated **F**-statistic value is less than the **F**-critical value, we retain the null hypothesis.

Confidence Interval

Example 1.

A sample of 100 patients was chosen to estimate the length of stay (LoS) at a hospital. The sample mean was 4.5 days and the population standard deviation was known to be 1.2 days.

(a) Calculate the 95% confidence interval for the population mean.

(b) What is the probability that the population mean is greater than 4.73 days?

Solution:

(a) 95% confidence interval for population mean: We know that $\bar{X} = 4.5$ and $\sigma = 1.2$ and thus $\sigma / \sqrt{n} = 1.2 / \sqrt{100} = 0.12$.

The 95% confidence interval is given by

$$(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n},\ \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = (4.5 - 1.96 \times 0.12,\ 4.5 + 1.96 \times 0.12)$$
$$= (4.2648,\ 4.7352)$$

The Excel function CONFIDENCE(α, σ, n) [or CONFIDENCE.NORM(α, σ, n)], where α is the significance, σ is the population standard deviation, and n is the sample size, returns the value $Z_{\alpha/2} \times \sigma / \sqrt{n}$. For current problem CONFIDENCE(0.05, 1.2, 100) = 0.235196. The corresponding confidence interval is

$$(4.5 - 0.235196,\ 4.5 + 0.235196) = (4.2648,\ 4.7352)$$

(b) Note that 4.73 is the upper limit of the 95% confidence interval from part (a), thus the probability that the population mean is greater than 4.73 is approximately 0.025.

Example 2.

Amount of time (measured in hours) spent by 20 students on an online course is given in Table below. Assuming that the population of time spent follows a normal distribution and standard deviation is 3.1 hours, calculate the 90% confidence interval for the mean time spent by the students.

Table: Sample time spent by student on an online course

4.7	9.3	8	7.4	9.2	1.7	7.2	8.6	9	6.9
9.2	11.2	7.6	4.9	5.3	2.8	12.3	10.6	5.7	3.8

Solution: The estimate mean from the sample is $\bar{X} = 7.27$ and the sampling distribution's standard deviation is $\sigma / \sqrt{n} = 3.1 / \sqrt{20} = 0.6932$.

The 90% confidence interval is given by

$$(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n},\ \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = (7.27 - 1.64 \times 0.6932,\ 7.27 + 1.64 \times 0.6932)$$
$$= (6.1332,\ 8.4068)$$

Example 3.

A retail store was interested in finding the proportion of customers who pay through cash (as against credit or debit card) for the merchandize they buy at the store. From a sample of 100 customers, it was found that 70 customers paid by cash. Calculate the 95% confidence interval for proportion of customers who pay by cash.

Solution: In this case, $n = 100$, $\hat{p} = 70/100 = 0.7$ and $\hat{q} = 1 - \hat{p} = 0.3$. Since $n \times \hat{p} \times \hat{q} = 100 \times 0.7 \times 0.3 = 21 \geq 10$, we can use the confidence interval equation provided in Eq. (5.4).

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}}$$

$$\Rightarrow 0.7 - 1.96\sqrt{\frac{0.7 \times 0.3}{100}} \leq p \leq 0.7 + 1.96\sqrt{\frac{0.7 \times 0.3}{100}} = 0.6102 \leq p \leq 0.7898$$

That is, the 95% confidence interval for **p** is (0.6102, 0.7898). That is, we are 95% confident that the interval (0.6102, 0.7898) contains the true population proportion of the customers who pay by cash.

Example 4.

An online grocery store is interested in estimating the basket size (number of items ordered by the customer) of its customer order so that it can optimize its size of crates used for delivering the grocery items. From a sample of 70 customers, the average basket size was estimated as 24 and the standard deviation estimated from the sample was 3.8. Calculate the 95% confidence interval for the basket size of the customer order.

Solution: We know that $n = 70$, $\bar{X} = 24$, $S = 3.8$ and $t_{0.025,69} = 1.995$ [using TINV(0.05, 69) in Microsoft Excel].
The confidence interval for size of basket using Eq. (5.6) is given by

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 24 \pm 1.995\frac{3.8}{\sqrt{70}} = 24 \pm 0.9061$$

Thus the 95% confidence interval for the size of the basket is (23.09, 24.91).

Example 5.

Time taken to manufacture an aircraft door is a random variable due to several manual processes and assembly of more than 1000 parts to make the aircraft door. The sources of variability in door assembly include factors such as non-availability of parts, manpower, and machine tools. It is known that the time to assemble a door follows a normal distribution. The variance of the time taken to manufacture the door was estimated to be 324 hours based on a sample of 50 doors. Calculate a 95% confidence interval for the variance in manufacturing aircraft door.

Solution: We know that $n = 50$, $S^2 = 324$, $\chi^2_{0.025,49} = 70.22$, $\chi^2_{0.975,49} = 31.55$ [the value of χ^2 can be calculated using Microsoft Excel function CHIINV($\alpha/2$, df) or CHISQ. INV.RT($1 - \alpha/2$, df)].

The 95% confidence interval for variance is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = \left[\frac{49\times 324}{70.22}, \quad \frac{49\times 324}{31.55}\right] = [226.09, \ 503.20]$$

The 95% confidence interval for standard deviation is [15.04, 22.43].

Linear Regression

Example 1.

Using the Hooke's law data in Table below, compute the least-squares estimates of the spring constant and the unloaded length of the spring. Write the equation of the least-squares line.

(Ref.: Statistics for Engineers and Scientists, Third Edition, William Navidi, McGraw-Hill Publication)

Table. Measured lengths of a spring under various loads

Weight (lb) x	Measured Length (in.) y	Weight (lb) x	Measured Length (in.) y
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

Solution:

The estimate of the spring constant is $\hat{\beta}_1$, and the estimate of the unloaded length is $\hat{\beta}_0$. From Table above.

$$\bar{x} = 1.9000 \qquad \bar{y} = 5.3885$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 26.6000$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} = 5.4430$$

We compute:

$$\hat{\beta}_1 = \frac{5.4430}{26.6000} = 0.2046$$

$$\hat{\beta}_0 = 5.3885 - (0.2046)(1.9000) = 4.9997$$

The equation of the least-squares line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Substituting the computed values for $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$y = 4.9997 + 0.2046x$$

Example 2.

Table below provides the annual salary in rupees of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10. Develop a simple linear regression model by estimating the model parameters.

Table: Salary of MBA students versus their grade 10 marks

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62.00	270000	26	64.60	250000
2	76.33	200000	27	50.00	180000
3	72.00	240000	28	74.00	218000
4	60.00	250000	29	58.00	360000
5	61.00	180000	30	67.00	150000
6	55.00	300000	31	75.00	250000
7	70.00	260000	32	60.00	200000
8	68.00	235000	33	55.00	300000
9	82.80	425000	34	78.00	330000
10	59.00	240000	35	50.08	265000
11	58.00	250000	36	56.00	340000
12	60.00	180000	37	68.00	177600
13	66.00	428000	38	52.00	236000
14	83.00	450000	39	54.00	265000
15	68.00	300000	40	52.00	200000
16	37.33	240000	41	76.00	393000
17	79.00	252000	42	64.80	360000
18	68.40	280000	43	74.40	300000
19	70.00	231000	44	74.50	250000
20	59.00	224000	45	73.50	360000
21	63.00	120000	46	57.58	180000
22	50.00	260000	47	68.00	180000
23	69.00	300000	48	69.00	270000
24	52.00	120000	49	66.00	240000
25	49.00	120000	50	60.80	300000

the estimated values of β_0 and β_1 are given by

$$\hat{\beta}_0 = 61555.3553 \text{ and } \hat{\beta}_1 = 3076.1774$$

The corresponding regression equation is given by

$$\hat{Y}_i = 61555.3553 + 3076.1774 X_i$$

where $\hat{Y}_i$ is the predicted value of Y for a given value of X_i.

The equation can be interpreted as follows: for every one percentage increase in grade 10 marks, the salary of the MBA students will increase at the rate of 3076.1774 on an average. The notations $\hat{\beta}_0$ and $\hat{\beta}_1$ are used to denote that these are estimated values of the regression coefficients from the sample of 50 students. The Microsoft

Table: Excel output for SLR model

	Coefficients	Standard Error	*t*-stat	*p*-value
Intercept	61555.35534	66701.901	0.9228	0.3607
Percentage in grade 10	3076.177438	1031.5258	2.9821	0.0044

Correlation

Example 1.

The average share prices of two companies over the past 12 months are shown in Table below. Calculate the Pearson correlation coefficient.

Table: Share prices (monthly average) of two companies over last 12 month

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are $\bar{X} = 292.9717$ and $\bar{Y} = 229.8292$.

The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

The values are shown in the table below.

Table: Calculation of correlation coefficient

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	−18.39	−10.33	189.97	338.25	106.6917
287.96	242.92	−5.01	13.09	−65.61	25.12	171.3699
290.35	245.90	−2.62	16.07	−42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	−21.22	−7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	−33.17	−15.40	510.82	1100.36	237.1343
263.02	201.86	−29.95	−27.97	837.72	897.10	782.2743
286.03	204.23	−6.94	−25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98

From table we have

$$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$$

$$\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$$

$$\text{Correlation coefficient } r = \frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

In Microsoft Excel, CORREL(array 1, array 2) will give the Pearson product moment correlation value.

ANOVA

Example 1. One-way ANOVA

Ms Rachael Khanna the brand manager of ENZO detergent powder at the 'one stop' retail was interested in understanding whether the price discounts have any impact on the sales quantity of ENZO. To test whether the price discounts had any impact, price discounts of 0% (no discount), 10% and 20% were given on randomly selected days. The quantity (in kilograms) of ENZO sold in a day under different discount levels is shown in Table below. Conduct a one-way ANOVA to check whether discount had any significant impact on the average sales quantity at $\alpha = 0.05$.

Table: Sales of ENZO at different price discounts

No Discount (0% discount)									
39	32	25	25	37	28	26	26	40	29
37	34	28	36	38	38	34	31	39	36
34	25	33	26	33	26	26	27	32	40
10% Discount									
34	41	45	39	38	33	35	41	47	34
47	44	46	38	42	33	37	45	38	44
38	35	34	34	37	39	34	34	36	41
20% Discount									
42	43	44	46	41	52	43	42	50	41
41	47	55	55	47	48	41	42	45	48
40	50	52	43	47	55	49	46	55	42

Solution:

In this case, the number of groups

$$k = 3;\ n_1 = n_2 = 30;\ \mu_1 = 32,\ \mu_2 = 38.77,\ \mu_3 = 46.4;\ and\ \mu = 39.05.$$

The sum of squares of between groups variation (SSB) is given by

$$SSB = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2 = 30 \times [(32 - 39.05)^2 + (38.77 - 39.05)^2$$

$$+ \ (46.4 - 39.05)^2] \ = \ 3114.156$$

So

$$MSB \ = \ \frac{SSB}{k-1} \ = \ \frac{3114.156}{2} \ = \ 1557.078$$

The sum of squares of within the group variation is given by

$$SSW \ = \ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \ = \ \sum_{j=1}^{30} (Y_{1j} - 32)^2 \ + \ \sum_{j=1}^{30} (Y_{2j} - 38.77)^2$$

$$+ \ \sum_{j=1}^{30} (Y_{3j} - 46.4)^2 \ = \ 2056.567$$

$$MSW \ = \ \frac{SSW}{n-k} \ = \ \frac{2056.567}{90-3} \ = \ 23.63$$

The F-statistic value is

$$F_{2,\,87} \ = \ \frac{MSB}{MSW} \ = \ \frac{1557.078}{23.6387} \ = \ 65.86$$

The critical F-value with degrees of freedom (2, 87) for $\alpha = 0.05$ is 3.101 [Excel function FINV(0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The p-value for $F_{2,\,87} = 65.86$ is 3.82×10^{-18} [using Excel function FDIST(65.86, 2, 87) or F.DIST.RT(65.86, 2, 87)]. Since the calculated F-statistic is much higher than the critical F-value, we reject the null hypothesis and conclude that the mean sales quantity values under different discounts are different. The Excel output of ANOVA is shown in Table below.

Table: One-way ANOVA excel output

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
No Discount	30	960	32	27.17241		
10% Discount	30	1163	38.76667	20.46092		
20% Discount	30	1392	46.4	23.28276		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	3114.15556	2	1557.078	65.86986	3.82E-18	3.101296
Within Groups	2056.56667	87	23.6387			
Total	5170.72222	89				

Example 2. One-way ANOVA

Share Raja Khan (SRK) is a top stockbroker and believes that the average annual stock return depends on the industrial sector. To validate his belief, SRK collected annual return of shares from three different industrial sectors − consumer goods, services, and industrial goods. The annual return of shares in 2015−2016 for different sectors is shown in Table below.

Table: Annual return of stocks under different industrial sector

Annual return on 30 consumer goods stocks									
6.32%	14.73%	11.95%	12.36%	10.28%	3.81%	10.15%	11.06%	6.29%	5.15%
8.44%	14.28%	8.89%	5.98%	6.96%	11.62%	5.22%	5.34%	5.93%	7.10%
10.91%	8.20%	10.19%	9.04%	8.61%	9.39%	2.63%	2.77%	4.76%	9.60%
Annual return on 30 services stocks									
13.70%	3.58%	1.36%	17.41%	10.01%	10.88%	15.63%	–0.04%	10.32%	7.40%
11.48%	9.71%	11.19%	8.21%	1.64%	1.45%	10.12%	13.85%	–10.27%	5.26%
12.05%	4.47%	8.71%	5.59%	10.02%	7.65%	10.03%	7.87%	6.59%	13.60%
Annual return on 30 industrial goods stocks									
6.74%	7.11%	5.69%	2.48%	5.42%	8.00%	2.55%	8.34%	4.99%	3.39%
8.73%	13.85%	5.29%	9.06%	2.84%	5.82%	7.66%	4.12%	9.10%	8.76%
10.77%	1.48%	4.71%	10.66%	0.44%	2.94%	6.55%	2.84%	3.90%	7.28%

Solution:

In this case, the number of groups

$k = 3;\ n_1 = n_2 = 30;\ \mu_1 = 0.082,\ \mu_2 = 0.079,\ \mu_3 = 0.0605;\ and\ \mu = 0.0743$

.

The sum of squares of between groups variation (SSB) is given by

$$SSB = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2 = 30 \times [(0.082 - 0.0743)^2 + (0.079 - 0.0743)^2$$

$$+ (0.0605 - 0.0743)^2] = 0.0087$$

Therefore

$$MSB = \frac{SSB}{k-1} = \frac{0.0087}{2} = 0.0043$$

The sum of squares of within the group variation is given by

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{j=1}^{30} (Y_{1j} - 0.082)^2 + \sum_{j=1}^{30} (Y_{2j} - 0.079)^2$$

$$+ \sum_{j=1}^{30} (Y_{3j} - 0.0605)^2 = 0.1463$$

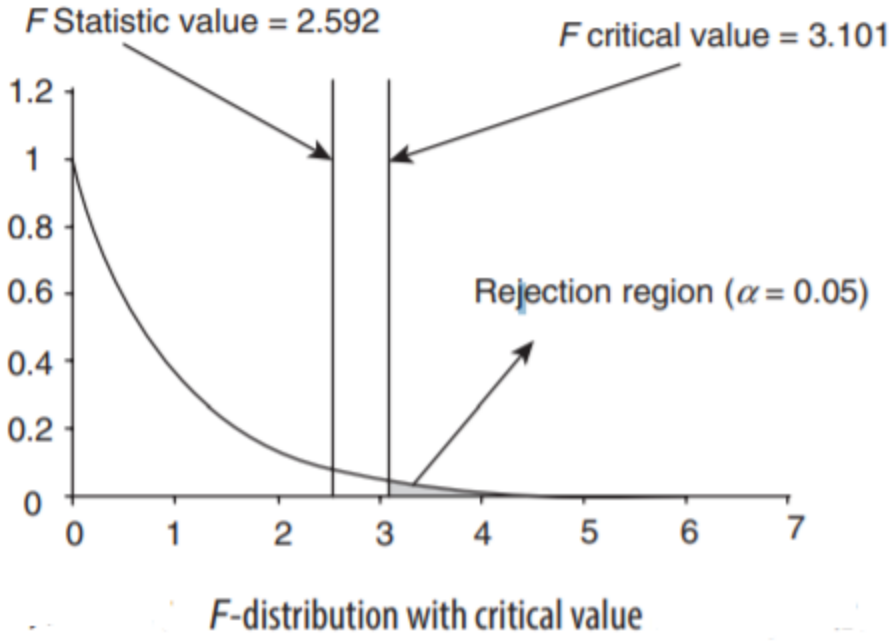
So

$$MSW = \frac{SSW}{n-k} = \frac{0.1463}{90-3} = 0.0016$$

The F-statistic value is

$$F_{2,\,87} = \frac{MSB}{MSW} = \frac{0.0043}{0.0016} = 2.592$$

The critical F-value with degrees of freedom (2, 87) for α = 0.05 is 3.101 [Excel function FINV(0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The P-value for $F_{2,\,87}$ = 2.592 is 0.0805 [using Excel function FDIST(2.592, 2, 87) or F.DIST.RT(2.592, 2, 87)]. Since the calculated F-statistic is less than the critical F-value, we retain the null hypothesis and conclude that the average annual returns under industrial sectors consumer goods, services, and industrial goods are not different (the below figure shows the F-critical value and F-statistic value for an F-distribution with degrees of freedom 2 and 87 for numerator and denominator, respectively).



The Excel output of ANOVA is shown in Table below.

Table: Microsoft excel ANOVA table

ANOVA: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Consumer Goods	30	2.4796	0.082653	0.00101		
Services	30	2.3947	0.079823	0.003073		
Industrial Goods	30	1.8151	0.060503	0.000963		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F critical*
Between Groups	0.008722	2	0.004361	2.59294	0.080572	3.101296
Within Groups	0.146317	87	0.001682			
Total	0.155039	89				

Example 3. Two-way ANOVA

Table below shows the sales quantity of detergents at different discount values and different locations collected over 20 days. Conduct a two-way ANOVA at $\alpha = 0.05$ to test the effects of discounts and location on the sales.

Table: Sales quantity at different locations under different discount rates

	Location 1		Location 2		
	Discount		Discount		
0%	10%	20%	0%	10%	20%
20	28	32	20	19	20
16	23	29	21	27	31
24	25	28	23	23	35
20	31	27	19	30	25
19	25	30	25	25	31
10	24	26	22	21	31
24	28	37	25	33	31
16	23	33	21	26	23
25	26	27	26	22	22
16	25	31	22	28	32
18	22	37	25	24	22
20	24	28	23	23	29
17	26	25	23	26	25
26	28	23	24	16	34
16	21	26	20	30	30
21	27	33	23	22	25
24	25	28	18	16	39
19	20	30	19	25	32
19	26	30	19	34	29
21	26	26	30	23	22

The two-way ANOVA with replication (since the data in Table above is repeated for locations) output from Microsoft Excel is shown in Table below.

Table: Two-way ANOVA with replication excel output

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Sample (Location)	7.008333	1	7.008333	0.443898	0.506593	3.92433
Columns (Discount)	1240.317	2	620.1583	39.27997	1.06E-13	3.075853
Interaction	84.81667	2	42.40833	2.686085	0.07246	3.075853
Within	1799.85	114	15.78816			
Total	3131.992	119				

In Table above, the sample stands for the row factor (which in this case is location), column stands for the column factor (discount in this case), and interaction stands for interaction effect (location × discount). The p-value for locations (data in rows) is 0.5065, thus it is not statistically significant (we retain the null hypothesis that the locations have no statistically significant influence on sales), whereas for discount rates (data in column) the p-value is 1.06×10^{-13}, so we reject the null hypothesis (that the discount rate has influence on sales). The p-value for the interaction effect is 0.0724 and is not significant. That is, only the factor discount is statistically significant at $\alpha = 0.05$.