**Testing of Hypothesis using R**

by

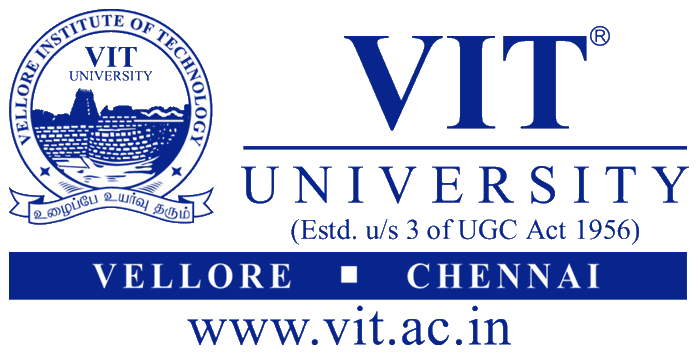
Marla Akhil Reddy 13BCE1078

Jagmohan Singh 13BCE1056

**MAT207 (Applied Probability, Statistics and Reliability)** Project report

Submitted to

**SCHOOL OF ADVANCED SCIENCES**



**VIT UNIVERSITY, CHENNAI**

**Vandalur – Kelambakkam Road**

**Chennai – 600127.**

**APRIL 2015**

Faculty In-charge: **Prof. Kaliyappan.M**

**ABSTRACT**

**R** is a [programming language](http://en.wikipedia.org/wiki/Programming_language) and software environment for statistical computing and graphics. The R language is widely used among [statisticians](http://en.wikipedia.org/wiki/Statistician) and [data miners](http://en.wikipedia.org/wiki/Data_mining) for developing [statistical software](http://en.wikipedia.org/wiki/Statistical_software) and data analysis. Polls, [surveys of data miners](http://en.wikipedia.org/wiki/Rexer's_Annual_Data_Miner_Survey), and studies of scholarly literature databases' provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible.

Here,In this Project we have applied different tools and functionalities avaliable from R to device different codes to implement in Testing of Hypothesis.We have explained different hypothesis testing taking two examples one for single and two var data.In the end we have explained the result form the output we got from appling the perticular values to code in R.

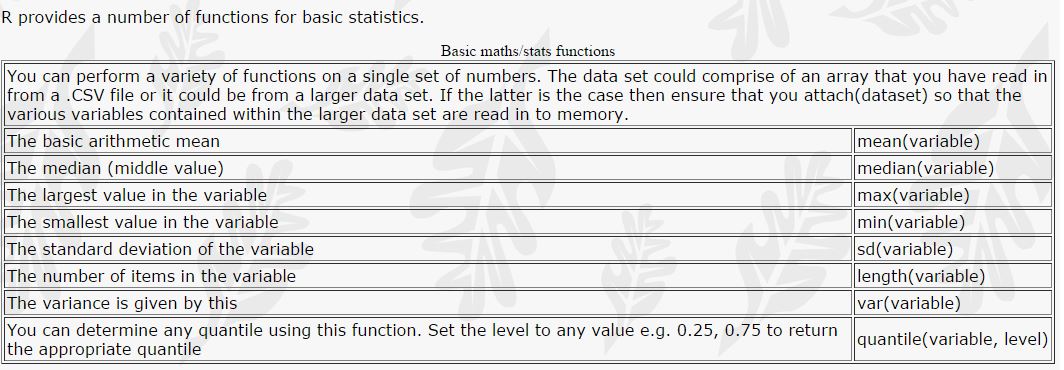
**TABLE OF CONTENTS**

* **Abstract**
* **What can we Analysis?**
* **Basic Functions used in Project**
* **Z-Test**
* **F-Test**
* **T-Test**
* **Chi-Square Test**
* **Conclusion**

**WHAT CAN WE ANALYSE?**

* **If we have a certain data in ‘.ODT’ format or in excel file. We can generate moment generating function, characteristic function and joint distribution function and analyse it in R.**
* **If we give data .We can code in R to get general summary like mean, variance, Highest and lowest values etc.. From that we can generate required function accordingly.**
* **Further we can get general distribution curves like Gamma, Binomial, Poisson etc.. From Data or Compare with general distributions like Normal etc..**

**BASIC FUNCTIONS USED IN PROJECT**

****

**Z-TEST**

**1.[Two sample](http://statistic-on-air.blogspot.in/2009/07/two-sample-z-test.html)** [**Z-test**](http://statistic-on-air.blogspot.in/2009/07/two-sample-z-test.html) **(Difference of Means)**

Comparison of the means of two independent groups of samples, taken from two populations with known variance.

**Problem:**

We are asked to compare the average heights of two groups. The first group (A) consists of individuals of Indian nationality (the variance of the Indian population is 5); the second group is taken from individuals of German nationality (the variance of German population variance is 8.5). The data are given below:

A: 175, 168, 168, 190, 156, 181, 182, 175, 174, 179  
B: 185, 169, 173, 173, 188, 186, 175, 174, 179, 180  
since we have the variance of the population, we must proceed with a two sample Z-test. .Here to test significance of difference b/w the means of the samples from given data ( Sample is from same or different population )

**Required R code:**

**Test of Normality:**

**kurtosis test and skewness test**

A more formal way of looking at the normality is by testing whether the kurtosis and skewness are significantly different from zero.

Take data and Keep it in

kurtosis.test() and skew.test() .If We see any significance difference between two functions then we can conclude that data don't follow normal distribution.

Both these tests are one-tailed, so you'll need to multiply the p-value by 2 to become two-tailed. If your p-value become larger than one you'll need to use 1-kurtosis.test() instead of kurtosis.test.

Code:

1. kurtosis.test <- function (x) {

m4 <- sum((x-mean(x))^4)/length(x)

s4 <- var(x)^2

kurt <- (m4/s4) - 3

sek <- sqrt(24/length(x))

totest <- kurt/sek

pvalue <- pt(totest,(length(x)-1))

pvalue

}

2. skew.test <- function (x) {

m3 <- sum((x-mean(x))^3)/length(x)

s3 <- sqrt(var(x))^3

skew <- m3/s3

ses <- sqrt(6/length(x))

totest <- skew/ses

pt(totest,(length(x)-1))

pval <- pt(totest,(length(x)-1))

pval

}

**Main code:**

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

z.test2sam = function(a, b, var.a, var.b){

n.a = length(a)

n.b = length(b)

zeta = (mean(a) - mean(b)) / (sqrt(var.a/n.a + var.b/n.b))

return(zeta)

}

**Solution:**

H0: µ1 = µ2 ( Samples has been taken from same Population)

H1: µ2 ≠ µ2 ( Samples have been taken from different Population)

The function z.test2sam provides in output the value of zeta, after receiving in input two vectors (a and b), the variance of the first population (var.a) and the variance of the second population (var.b).  
Using this function we obtain:   
  
  
a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)  
b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)  
  
z.test2sam(a, b, 5, 8.5)  
[1] -2.926254

The value of zeta is greater than the value of the critical value zeta tabulated for alpha equal to 0.05 (z-tabulated = 1.96 for a two-tailed test): then we reject the null hypothesis in favour of the alternative hypothesis. We conclude that the two means are significantly different.

**2.[One](http://statistic-on-air.blogspot.in/2009/07/one-sample-z-test.html)** [**sample Z-test**](http://statistic-on-air.blogspot.in/2009/07/one-sample-z-test.html) **( Single Mean )**

**Comparisons of the sample mean with know population mean and standard deviation.**

**Problem:**  
10 volunteers have done an intelligence test; here are the results obtained. The mean obtained at the same test, from the entire population is 75. You want to check if there is a statistically significant difference (with a significance level of 95%) between the means of the sample and the population, assuming that the sample variance is known and equal to 18.

65, 78, 88, 55, 48, 95, 66, 57, 79, 81

Here we test whether given sample of size n has been drawn from a population sample with mean ‘µ’.

**Required R code:**

a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)

z.test = function(a, mu, var){

zeta = (mean(a) - mu) / (sqrt(var / length(a)))

return(zeta)

}

**Solution:**

H0: sample mean = µ

H1: sample mean >/< µ

function z.test; it receives in input a vector of values (a), the mean of the population to perform the comparison (mu), and the population variance (var); it returns the value of zeta.

a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)  
  
z.test(a, 75, 18)  
[1] -2.832353  
  
The value of zeta is equal to -2.83, which is higher than the critical value Zcv = 1.96, withalpha = 0.05 (**2-tailed test**). We conclude therefore that the mean of our sample is significantly different from the mean of the population.

**3.Z-test of Single Proportion**

The null hypothesis of the two-tailed test about population proportion can be expressed as follows:

p = p0

where *p*0 is a hypothesized value of the true population proportion *p*.

Let us define the test statistic *z*in terms of the sample proportion and the sample size:

Then the null hypothesis of the two-tailed test is to be *rejected*if *z*≤−*zα∕*2 or *z*≥ *zα∕*2 , where *zα∕*2is the 100(1 − *α*) percentile of the standard normal distribution.

**Problem:**

Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?

**Required R code:**

h -> 12

t -> 20

p0 = .5 # hypothesized value

z.test = function(h,t,p0){

pbar = h/t # sample proportion

n = t # sample size

z = (pbar???p0)/sqrt(p0???(1???p0)/n)

return(z)

}

**Code with α significance:**

alpha = .05

z.testalpha = function(alpha){

z.half.alpha = qnorm(1???alpha/2)

return(c(???z.half.alpha, z.half.alpha))

}

**Solution:**

The null hypothesis (H0)is that *p*= 0*.*5.

We use z.test() function to calculate z value

z.test(12,20,0.5)

[1] 0.8944272

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

We use function z.testalpha(alpha) to calculate significance value of alpha

* z.testalpha(0.05)

[1] -1.959964 1.959964

**Answer**

The test statistic 0.89443 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do *not*reject the null hypothesis that the

**4. Parametric Z-test ( Difference of Proportions)**

**X1 and x2 be the no. of persons pocessing the given attribute in random samples of sizes n1 and n2 from two populations respectively.**

**The sample proportions given by**

**P1 = X1/n1 and P2 = X2/n2**

**Z = P1- P2 /**

**Problem:**

The owner of a betting company wants to verify whether a customer is cheating or not. To do this want to compare the number of successes of one player with the number of successes of one of his employees, of which he is certain that he is not cheating. In a month's time, the player performs 74 bets and wins 30; the player in the same period of time making 103 bets, wins 65. Your client is a cheat or not?

**Required R code:**

z.prop = function(x1,x2,n1,n2){

numerator = (x1/n1) - (x2/n2)

p.common = (x1+x2) / (n1+n2)

denominator = sqrt(p.common \* (1-p.common) \* (1/n1 + 1/n2))

z.prop.ris = numerator / denominator

return(z.prop.ris)

}

**Solution:**

We will calculate z value by z.prop() function

Z.prop function calculates the value of Z, receiving input the number of successes (x1 and x2), and the total number of games (n1 and n2). We apply the function just written with the data of our problem:  
  
  
z.prop(30, 65, 74, 103)  
[1] -2.969695  
  
We obtained a value of z greater than the value of z-tabulated (1.96), which leads us to conclude that the player that the director was looking at is actually a cheat, since its probability of success is higher than a non-cheat user.

**F-TEST**

Objective of this test is to determine whether two independent estimates of population variance differ significantly or whether two samples may be regarded as drawn from normal population having same variance.

Before proceeding with the *t-test*, it is necessary to evaluate the sample variances of the two groups, using a Fisher's F-test to verify the *homoscedasticity* (*homogeneity of variances*).

**Problem:**

A: 175, 168, 168, 190, 156, 181, 182, 175, 174, 179  
B: 185, 169, 173, 173, 188, 186, 175, 174, 179, 180

Calculate F test to compare two variances.

**Required R code:**

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

var.test(a,b)

qf(0.95, 9, 9)

**Solution:**

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)  
b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)  
  
var.test(a,b)  
  
F test to compare two variances  
  
data: a and b  
F = 2.1028, num df = 9, denom df = 9, p-value = 0.2834  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
0.5223017 8.4657950  
sample estimates:  
ratio of variances  
2.102784

We obtained p-value greater than 0.05, then we can assume that the two variances are homogeneous. Indeed we can compare the value of F obtained with the tabulated value of F for alpha = 0.05, degrees of freedom of numerator = 9, and degrees of freedom of denominator = 9, using the function qf(p, df.num, df.den):  
  
  
qf(0.95, 9, 9)  
[1] 3.178893  
  
Note that the value of F computed is less than the tabulated value of F, which leads us to accept the null hypothesis of homogeneity of variances.  
**NOTE:** The F distribution has only one tail, so with a confidence level of 95%,p = 0.95. Conversely, the *t-distribution* has two tails, and in the R's function(p, df) we insert a value p = 0975 when you're testing a two-tailed alternative hypothesis.

**T-TEST**

**1.[One sample Student's](http://statistic-on-air.blogspot.in/2009/07/one-sample-students-t-test.html)** [**t-test**](http://statistic-on-air.blogspot.in/2009/07/one-sample-students-t-test.html) **( Single Mean)**

**Comparison of the sample mean with a known value, when the variance of the population is not known.**

**Problem:**  
10 volunteers have done an intelligence test; here are the results obtained. The average result of the population which received the same test, is equal to 75. You want to check if the sample mean is significantly similar (when the significance level is 95%) to the average population, assuming that the variance of the population is not known.

65, 78, 88, 55, 48, 95, 66, 57, 79, 81

**Required R code:**

a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)

mu <- 75

t.test (a, mu=75)

**Solution:**

The **Student's t-test** for a **single sample** have a pre-set function in R we can apply immediately.  
It is the **t.test (a, mu)**

a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)  
  
t.test (a, mu=75)  
  
**One Sample t-test  
  
data: a  
t = -0.783, df = 9, p-value = 0.4537  
alternative hypothesis: true mean is not equal to 75  
95 percent confidence interval:  
60.22187 82.17813  
sample estimates:  
mean of x  
71.2**

The function t.test on one sample provides in output the value of t calculated; also gives us degrees of freedom, the confidence interval and the average (mean of x).  
In order to take your statistic decision, you can proceed in two ways. We can compare the value of t with the value of the tabulated student t with 9 degrees of freedom. If we do not have tables, we can calculate the value t-tabulated in the following way:  
  
  
qt(0.975, 9)  
[1] 2.262157  
  
  
The function qt (p, df) returns the value of t computed considering the significance level (we chose a significance level equal to 95%, which means that each tail is the 2.5% which corresponds to the value of *p = 1 - 0.025*), and the degrees of freedom. By comparing the value of t-tabulated with t-computed, t-computed appears smaller, which means that we accept the null hypothesis of equality of the averages: our sample mean is significantly similar to the mean of the population.  
  
Alternatively we could consider the p-value. With a significance level of 95%, If **p-value is greater than 0.05** then **we accept the null hypothesis H0**; if **p-value is less than 0.05** then **we reject the null hypothesis H0** in favour of the alternative hypothesis H1.

**2.Two sample Student’s t-test (Samples of equal variances)**

**t-Test to compare the means of two groups under the assumption that both samples are random, independent, and come from normally distributed population with unknown but equal variances  
  
Problem:**A: 175, 168, 168, 190, 156, 181, 182, 175, 174, 179  
B: 185, 169, 173, 173, 188, 186, 175, 174, 179, 180

**Required R code:**

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

t.test(a,b, var.equal=TRUE, paired=FALSE)

**Solution:**

Here actually before going to the Problem .We need to Test for Equity of Variance. This can be done by F-TEST.

We now use t.test for homogenous variances.

t.test( a, b, var.equal=TRUE, paired=FALSE)

Two Sample t-test  
  
data: a and b  
t = -0.9474, df = 18, p-value = 0.356  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-10.93994 4.13994  
sample estimates:  
mean of x mean of y  
174.8 178.2  
  
We obtained p-value greater than 0.05, then we can conclude that the averages of two groups are significantly similar. Indeed the value of t-computed is less than the tabulated t-value for 18 degrees of freedom, which in R we can calculate:  
  
  
qt(0.975, 18)  
[1] 2.100922  
  
This confirms that we can accept the null hypothesis H0 of equality of the means.

**3.[Two sample Student's](http://statistic-on-air.blogspot.in/2009/07/two-sample-students-t-test-2.html)** [**t-test**](http://statistic-on-air.blogspot.in/2009/07/two-sample-students-t-test-2.html)**(variance unknown)**

Comparison of the averages of two independent groups, extracted from two populations at variance unknown; sample variances are not homogeneous.

**Problem:**

A: 175, 168, 168, 190, 156, 181, 182, 175, 174, 179  
B: 120, 180, 125, 188, 130, 190, 110, 185, 112, 188

**Required R code:**

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(120, 180, 125, 188, 130, 190, 110, 185, 112, 188)

t.test(a,b, var.equal=FALSE, paired=FALSE)

qf(0.95, 9, 9)

**Solution :**

Now we need to check for non homogeneouity by F-test

On observing we get p value less than 0.05 . So we can say its Non-Homogenous.

To make the comparison between the two groups, we use the function t.test with not homogeneous variances (var.equal = FALSE, which can also be omitted, because the function works on non-homogeneous variance by default) and independent samples (paired = FALSE, which can also be omitted, because by default the function works on independent samples) in this way:  
  
  
t.test( a, b, var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test  
  
data: a and b  
t = 1.8827, df = 10.224, p-value = 0.08848  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-3.95955 47.95955  
sample estimates:  
mean of x mean of y  
174.8 152.8

We obtained p-value greater than 0.05, then we can conclude that the means of the two groups are significantly similar). Indeed the value of t is less than the tabulated t-value for 10,224 degrees of freedom, which in R we can calculate:  
  
  
qt(0.975, 10.224)  
[1] 2.221539  
  
We can accept the hypothesis H0 of equality of means.

**4.Paired Student’s t-test**

Comparison of the means of two sets of paired samples, taken from two populations with unknown variance.

**Problem:**  
A school athletics has taken a new instructor, and want to test the effectiveness of the new type of training proposed by comparing the average times of 10 runners in the 100 meters. Are below the time in seconds before and after training for each athlete.  
  
Before training: 12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3  
After training: 12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1

**Required R code:**

a = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)

b = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)

t.test(a,b, paired=TRUE)

qt(0.975, 9)

**Solution:**

In this case we have two sets of paired samples, since the measurements were made on the same athletes before and after the workout. To see if there was an improvement, deterioration, or if the means of times have remained substantially the same (hypothesis H0).

t.test(a,b, paired=TRUE)

Paired t-test  
  
data: a and b  
t = -0.2133, df = 9, p-value = 0.8358  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-0.5802549 0.4802549  
sample estimates:  
mean of the differences  
-0.05  
  
The p-value is greater than 0.05, then we can accept the hypothesis H0 of equality of the averages. In conclusion, the new training has not made any significant improvement (or deterioration) to the team of athletes.  
Similarly, we calculate the t-tabulated value:  
  
  
qt(0.975, 9)  
[1] 2.262157  
  
*t-computed < t-tabulated*, so we accept the null hypothesis H0.

**Problem:**

Suppose now that the manager of the team (given the results obtained) fired the coach who has not made any improvement, and take another, more promising. We report the times of athletes after the second training:  
  
Before training: 12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3  
After the second training: 12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0

**Required R code:**

**Less alternative:**

a = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)

b = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)

t.test(a,b, paired=TRUE, alt="less")

**More alternative:**

a = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)

b = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)

t.test(a,b, paired=TRUE, alt="greater")

**Solution:**

Now we check if there was actually an improvement, ie perform a t-test for paired data, specifying in R to test the alternative hypothesis H1 of improvement in times. To do this simply add the syntax alt = "less" when you call the t-test:

t.test(a,b, paired=TRUE, alt="less")

t.test(a,b, paired=TRUE, alt="less")  
  
Paired t-test  
data: a and b  
t = 5.2671, df = 9, p-value = 0.9997  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf 2.170325  
sample estimates:  
mean of the differences  
1.61

t.test(a,b, paired=TRUE, alt="greater")  
  
Paired t-test  
  
data: a and b  
t = 5.2671, df = 9, p-value = 0.0002579  
alternative hypothesis: true difference in means is greater than 0  
95 percent confidence interval:  
1.049675 Inf  
sample estimates:  
mean of the differences  
1.61

**CHI-SQUARED TEST**

**CHI-SQUARED TEST for Independent of Attributes**

**Problem:**

The owner of a betting company wants to verify whether a customer is cheating or not. To do this want to compare the number of successes of one player with the number of successes of one of his employees, of which he is certain that he is not cheating. In a month's time, the player performs 74 bets and wins 30; the player in the same period of time making 103 bets, wins 65. Your client is a cheat or not?

**Required R code:**

prop.test(x = c(30, 65), n = c(74, 103), correct = FALSE)

prop.test(x = c(30, 65), n = c(74, 103), correct=TRUE) # for Small Samples

qchisq(0.950, 1)

**Solution:**

We can solve it in Parametric Z-test or Non parametric chi-Squared Test

z.prop(x1,x2,n1,n2)

Or by Chi-Squared Test

Here we test whether two attributes are considered to be as independent or not

We solve the problem with the test of chi-square applied to a 2x2 contingency table. In R there is the function prop.test.  
  
  
prop. test(x = c(30, 65), n = c(74, 103), correct = FALSE)

2-sample test for equality of proportions without continuity correction  
  
data: c(30, 65) out of c(74, 103)  
X-squared = 8.8191, df = 1, p-value = 0.002981  
alternative hypothesis: two.sided  
95 percent confidence interval:  
-0.37125315 -0.08007196  
sample estimates:  
prop 1 prop 2  
0.4054054 0.6310680

Correct= False for large samples

In the case of small samples (low value of n), you must specify correct = TRUE, so as to change the computation of chi-square based on the continuity of Yates:  
  
  
prop.test(x = c(30, 65), n = c(74, 103), correct=TRUE)  
  
2-sample test for equality of proportions with continuity correction  
  
data: c(30, 65) out of c(74, 103)  
X-squared = 7.9349, df = 1, p-value = 0.004849  
alternative hypothesis: two.sided  
95 percent confidence interval:  
-0.38286428 -0.06846083  
sample estimates:  
prop 1 prop 2  
0.4054054 0.6310680

In both cases, we obtained p-value less than 0.05, which leads us to reject the hypothesis of equal probability. In conclusion, the customer is a cheat. For confirmation we compare the value chi-square-value calculated with the chi-square-tabulation, which we calculate in this way:  
  
  
qchisq(0.950, 1)  
[1] 3.841459  
  
qchisq function calculates the value of chi-square as a function of alpha and degrees of freedom. Since chi-square-calculated is greater than chi-square-tabulation, we conclude by rejecting the hypothesis H0

**CONCLUSION**

By this we can show that various complex and heavy data analysis in Probability can be easily solved in using mathematical sofwares like R.