Hypothesis testing using R

M P Gururajan, Hina A Gokhale and Dayadeep Monder

Indian Institute of Technology Bombay, Mumbai

In this session, we are going to learn some R commands for hypothesis testing. Specifically, we solve some of the problems from Ross using R.

1 Solved problems 8.3a to 8.3e

In 8.3a problem, it is given that a signal is received with a known standard deviation, namely, 2; that is, a random noise of N(0,4) is added to the signal. The average of 5 signals, sent independently is 9.5. We are asked to test the hypothesis that the actual signal sent is 8.

The hypothesis that the actual mean of the signals sent is 8 is called the null hypothesis H_0.

The first step in hypothesis testing is to identify the level of significance, α. Let us choose it to be 5% or 0.05.

We then calculate the test statistic:

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \tag{1}$$

In our case, $\mu_0 = 8$; $\bar{X} = 9.5$; $n = 5$ and $\sigma = 2$.

Hence, the test statistic is

```
Xbar = 9.5
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
```

We need to compare this test statistics with the critical value for accepting / rejecting the hypothesis. In this case, the critical value is decided by the standard normal distribution, $P(Z > z_{\alpha/2}) = \alpha/2$. Let us identify the $z_{\alpha/2}$

```
z <- qnorm(0.975)
```

By comparing the z value with the test statistic, we can either accept or reject the hypothesis. The following script is the complete version that solves the given problem:

```
Xbar = 9.5
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
```

```
alpha = 0.05
z <- qnorm(1-0.5*alpha)
TestStat

## [1] 1.677051

z

## [1] 1.959964

if(TestStat < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}

## [1] "Accept the null hypothesis"
```

As indicated in this problem, in the above script, if you change the α value to 10% (that is, 0.1), the hypothesis will be rejected.

In 8.3b, we consider the cases where $\bar{X} = 8.5$ and $\bar{X} = 11.5$. Consider $\bar{X} = 8.5$.

```
Xbar = 8.5
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
TestStat

## [1] 0.559017

z <- 2*pnorm(-TestStat)
print("The p-value associated with this test statistic is")

## [1] "The p-value associated with this test statistic is"

z

## [1] 0.5761501

Xbar = 11.5
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
TestStat

## [1] 3.913119

z <- 2*pnorm(-TestStat)
print("The p-value associated with this test statistic is")
```

```
## [1] "The p-value associated with this test statistic is"

z

## [1] 9.111162e-05
```

Large p-values mean that the null hypothesis will be accepted; and, very small p-values mean that the null hypothesis will be rejected.

In 8.3c, we are interested in calculating the probability of accepting the null hypothesis when the value received is 10. The following script solves this problem:

```
Xbar = 10
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
TestStat

## [1] 2.236068

alpha = 0.05
z <- qnorm(1-0.5*alpha)
prob <- pnorm(TestStat+z)-pnorm(TestStat-z)
prob

## [1] 0.3912205
```

In 8.3d, our interest is in finding the number of signals that should be sent to reject the null hypotheses of $\mu = 8$ with $\alpha = 0.05$ with 75% probability when the actual mean of the signals is 9.2. This is done using the formula

$$n \approx \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{(\mu_1 - \mu_0)^2}. \tag{2}$$

The following script calculates n:

```
alpha = 0.05
beta = 0.25
Var = 4
mu1 = 9.2
mu0 = 8
n <- (qnorm(0.5*alpha)+qnorm(beta))^2*Var/(mu1-mu0)^2
n

## [1] 19.27874
```

Let us check that with 20 signals, an average of 9.2 would indeed be rejected about 75% of the time. Note that we calculate the acceptance probability and obtain the rejection probability by subtracting it from unity.

```
Xbar = 9.2
mu0 = 8
n = 20
```

```
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
TestStat

## [1] 2.683282

alpha = 0.05
z <- qnorm(1-0.5*alpha)
prob <- pnorm(TestStat+z)-pnorm(TestStat-z)
1-prob

## [1] 0.7652593
```

Finally, in 8.3e, we assume that we know that the signal is at least as large as 8. Using this, we solve the problem 8.3a, namely, the acceptance or rejection of null hypotheses. The following R script accomplishes the task:

```
Xbar = 9.5
mu0 = 8
n = 5
sigma = 2
TestStat <- sqrt(n)*(Xbar-mu0)/sigma
alpha = 0.05
z <- qnorm(1-alpha)
TestStat

## [1] 1.677051

z

## [1] 1.644854

if(TestStat < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}

## [1] "Reject the null hypothesis"

print("The p-value is")

## [1] "The p-value is"

pnorm(-TestStat)

## [1] 0.04676626

print("That is, for any alpha above the given p-value the null hypothesis will be rejected")

## [1] "That is, for any alpha above the given p-value the null hypothesis will be rejected"
```

2 Solved problems 8.3h and 8.3j

In 8.3h, we have data of average daily use of water of 20 homes (in gallons). We are asked to test the null hypothesis that the daily average use is 350 gallons. Unlike the previous problems, we do not know the variance and hence we have to estimate it from the data. This also means using t-distribution to calculate the critical region. The following R script does the hypothesis testing:

```
WaterConsumption <- c(340,344,362,375,356,386,354,364,332,402,340,
                      355,362,322,372,324,318,360,338,370)
Xbar <- mean(WaterConsumption)
mu0 = 350
n = 20
Sigma <- sd(WaterConsumption)
TestStat <- sqrt(n)*(Xbar-mu0)/Sigma
alpha = 0.05
z <- qt(1-alpha,n-1)
TestStat
```

```
## [1] 0.7778411
```

```
z
```

```
## [1] 1.729133
```

```
if(TestStat < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}
```

```
## [1] "Accept the null hypothesis"
```

```
print("The p-value is")
```

```
## [1] "The p-value is"
```

```
2*pt(-TestStat,n-1)
```

```
## [1] 0.4462411
```

```
print("That is, for any alpha above the given p-value
      the null hypothesis will be rejected")
```

```
## [1] "That is, for any alpha above the given p-value \n      the null hypothesis will be rejected"
```

In the case of 8.3j, we test for a null hypothesis which is one sided and the variance is evaluated from the data and hence t-distribution is used for testing the hypothesis.

```
ServiceTime <- c(8.6,9.4,5.0,4.4,3.7,11.4,10.0,7.6,14.4,
                 12.2,11.0,14.4,9.3,10.5,10.3,7.7,8.3,6.4,
                 9.2,5.7,7.9,9.4,9.0,13.3,11.6,10.0,9.5,6.6)
```

```
Xbar <- mean(ServiceTime)
mu0 = 8
n = 28
Sigma <- sd(ServiceTime)
TestStat <- sqrt(n)*(Xbar-mu0)/Sigma
alpha = 0.05
z <- qt(1-alpha,n-1)
TestStat
```

```
## [1] 2.25753
```

```
z
```

```
## [1] 1.703288
```

```
if(TestStat < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

```
print("The p-value is")
```

```
## [1] "The p-value is"
```

```
pt(-TestStat,n-1)
```

```
## [1] 0.01613468
```

```
print("That is, for any alpha above the given p-value
      the null hypothesis will be rejected")
```

```
## [1] "That is, for any alpha above the given p-value \n      the null hypothesis will be rejected"
```

3 Solved problems 8.4a to 8.4c

In 8.4a, we are given the lives of tires in units of 100 kilometres for tires produced using two different methods. In method A, standard deviation is 40 and the data consists of 10 data points. In Method B, the standard deviation is 60 and the data consists of 8 data points. We are asked to test the hypothesis that the means in these two cases are the same with 5% level of significance. The following script gives the answer to the question:

```
n = 10
m = 8
VarA = 40*40
VarB = 60*60
dataA <- c(61.1,58.2,62.3,64,59.7,66.2,57.8,61.4,62.2,63.6)
```

```
dataB <- c(62.2,56.6,66.4,56.2,57.4,58.4,57.6,65.4)
Xbar <- mean(dataA)
Ybar <- mean(dataB)
TestStat <- (Xbar-Ybar)/sqrt((VarA/n) + (VarB/m))
alpha = 0.05
z <- qnorm(1-0.5*alpha)
TestStat
```

```
## [1] 0.06579433
```

```
z
```

```
## [1] 1.959964
```

```
if(TestStat < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}
```

```
## [1] "Accept the null hypothesis"
```

```
print("The p-value is")
```

```
## [1] "The p-value is"
```

```
pnorm(-TestStat)
```

```
## [1] 0.4737708
```

8.4b is similar to the previous problem but for the fact that the variances are not known and hence they have to be estimated and the t-distribution should be used for the analysis. The data given is one the consumption of vitamin C (10) and placebo (12) and we are asked to evaluate if vitamin C made any difference and identify the level of significance for acceptance or rejection of the hypothesis given the number of days it has taken for the patients to get cured of the cold. Here is the script that answers the question:

```
n = 10
m = 12
VitC <- c(5.5,6.0,7.0,6.0,7.5,6.0,7.5,5.5,7.0,6.5)
Placebo <- c(6.5,6.0,8.5,7.0,6.5,8.0,7.5,6.5,7.5,6.0,8.5,7.0)
Xbar <- mean(VitC)
Ybar <- mean(Placebo)
XVar <- var(VitC)
YVar <- var(Placebo)
Sp2 <- ((n-1)*XVar+(m-1)*YVar)/(n+m-2)
TestStat <- (Xbar-Ybar)/sqrt(Sp2*((1/n)+(1/m)))
alpha = 0.05
z <- qt(1-alpha,n+m-2)
TestStat
```

```
## [1] -1.898695
```

```
z
```

```
## [1] 1.724718
```

```
if(abs(TestStat) < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

Finally, in 8.4c, we test the same hypothesis as in 8.4a, assuming that the variances are equal and unknown.

```
n = 10
m = 8
dataA <- c(61.1,58.2,62.3,64,59.7,66.2,57.8,61.4,62.2,63.6)
dataB <- c(62.2,56.6,66.4,56.2,57.4,58.4,57.6,65.4)
Xbar <- mean(dataA)
Ybar <- mean(dataB)
XVar <- var(dataA)
YVar <- var(dataB)
Sp2 <- ((n-1)*XVar+(m-1)*YVar)/(n+m-2)
TestStat <- (Xbar-Ybar)/sqrt(Sp2*((1/n)+(1/m)))
alpha = 0.05
z <- qt(1-alpha,n+m-2)
TestStat
```

```
## [1] 1.028023
```

```
z
```

```
## [1] 1.745884
```

```
if(abs(TestStat) < z){
  print("Accept the null hypothesis")
} else {
    print("Reject the null hypothesis")
}
```

```
## [1] "Accept the null hypothesis"
```

```
2*pt(-TestStat,n+m-2)
```

```
## [1] 0.3192305
```

4 Home work

You have had access to several data sets so far. Some of them are amenable to this type of analysis. For example, taking the rain fall data, you can answer questions of the following type: is the annual average rain fall over Goa the same as Kerala? Is the average rain fall in the months of June, July, August, September and October in Vidharba region the same as North interior Karnataka? Specifically, with what level of confidence we can assert these statements? Similarly, taking the literacy data, we can compare the different states and union territories in terms of different parameters. Take these data and paly around with questions of this sort!!