Data Analytics
Mini Project - 4

Abhijeet Singh Panwar (ID : 201351005)

October 26, 2016

Instructor :

Prof. Bhargab Chattopadhyay

Indian Institute of Information Technology, Vadodara

0.1 Question 1

Consider the dataset stored in the file bp.xlsx. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods — a finger method and an arm method — from the same 200 patients.

Part A

Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer.

Solution

From boxplot, we can observe that spread(or variance) of both the distribution,

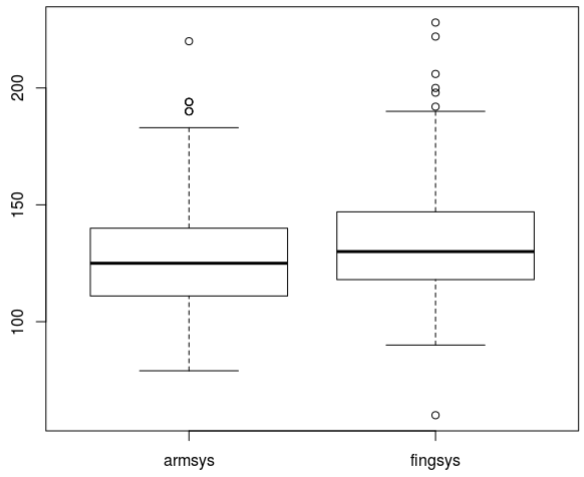

Figure 1: Box-plot of Fingsys & Armsys

are nearly equal.

Part B

Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

Solution

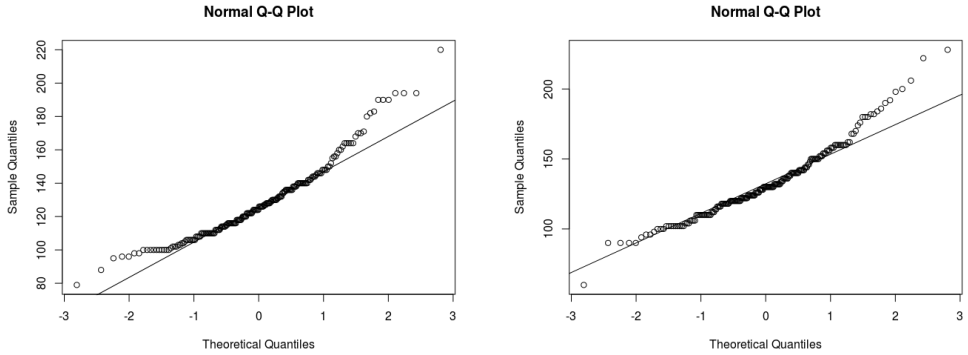

Figure 2: Q-Q plot of both Armsys & Fingsys

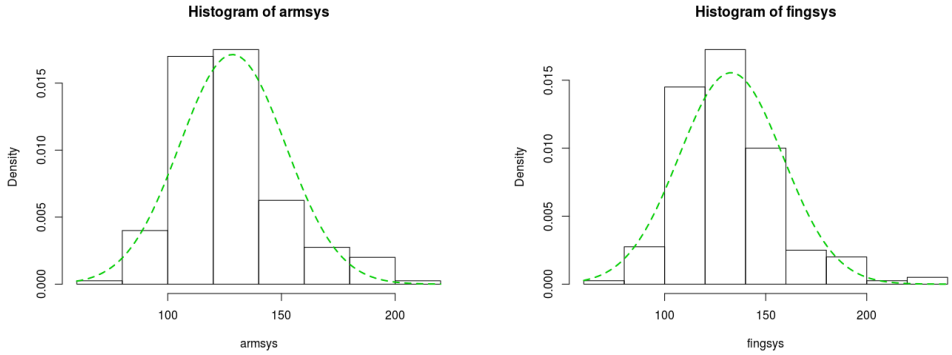

Figure 3: Q-Q plot of both Armsys & Fingsys

From seeing both, qqplot & histogram for **Armsys** & **Fingsys**, **it is evident that both the distributions are normal in nature**. Hence, in case of qqplot, more the points are towards y=x line, more the chances of given dataset to be normal in nature. Similarly, in case of histogram for both dataset, the normal curve nearly fits the histogram.

Part C

Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold?

Solution

Assumption

1. Variance of both distribution are equal. *Reffered to Part B*

The confidence interval for difference of mean of two distribution is:

$$[-72.19, 63.60]$$

Part D

Perform an appropriate 5% level test to see if there is any difference in the means of the two methods. Be sure to clearly set up the null and alternative hypotheses. State your conclusion. What assumptions, if any, did you make to construct the

interval? Do they seem to hold?

Solution

Null Hypothesis(H_o) : $(mean)_{Armsys} - (mean)_{Fingsys} = 0$

Alternative Hypothesis(H_a) : $(mean)_{Armsys} - (mean)_{Fingsys} \neq 0$

For, $\alpha = 0.05$

We will use t-test for hypothesis testing, as in this case population variace is unknown & assumed to be equal for both distribution.

Test statistic(t) = $(mean)_{Armsys} - (mean)_{Fingsys} \div S_p * \sqrt{1/n + 1/m}$

Here, S_p, is pooled sample variance.
n, is size of Armsys distribution.
m, is size of Fingsys distribution.

If $\mid t \mid >$ PValue, we reject Null Hypothesis,

Through calculation,

t, comes out to be -0.00715 & PValue comes out to be 1.965.

Therefore, null hypothesis is accepted, this shows that there is no difference in means of two methods.

Part E

Do the results from (c) and (d) seem consistent? Justify your answer.

Solution

0.2 Question 2

Suppose we are interested in testing the null hypothesis that the mean of a normal population is 10 against the alternative that it is greater than 10. A random sample of size 20 from this population gives 9.02 as the sample mean and 2.22 as the sample standard deviation.

Part A

Set up the null and alternative hypotheses.

Solution

The null hypothesis(H_o) will be : $\mu = 10$, where μ is population mean.

And, the alternative hypothesis(H_a) will be $\mu > 10$.

Part B

Which test would you use? What is the test statistic? What is the null distribution of the test statistic?

Solution

T-tests will be implemented, as population variance(σ) is unknown.

Therefore, the test-statistic(t) will be, $(\overline{X} - \mu_o) \div s/\sqrt{n}$.

Here, $\overline{X}$ *is sample mean,*
and n is sample size.

Student's T-distribution will be the null distribution for out test statistic.

Part C

Compute the observed value of the test statistic.

Solution

Value of test statistic based on above formula will be, **-1.974186**

Part D

Compute the p-value of the test using the usual way.

Solution

P-values for T-tests (F_ν is the cdf of T-distribution with the suitable number ν of degrees of freedom).

Therefore, using the formula $\rightarrow$ 1 - $F_a(t_{obs})$.

P-value = 0.03153941

Part E

Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare?

Solution

From Monte Carlo estimation, the p-value thus obtained is 0.0344. And, pvalue calculated from two different methods, comes out to be nearly equal.

Part F

State your conclusion at 5% level of significance.

Solution

As, $\alpha \rightarrow 0.05$, is greater than P_value, i.e. 0.03. Therefore, **null hypothesis is rejected**.

0.3 Question 3

According to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average credit limits of $2635 and $2887, respectively. Suppose that the sample standard deviations of these two samples were $365 and $412, respectively.

Part A

Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.

Solution

Given:
Mean Credit limit for January 2011:

- Sample Population $\rightarrow$ 400
- Sample Mean($\overline{X}$) $\rightarrow$ 2635$
- Sample standard deviation(s_1) $\rightarrow$ 365

Mean Credit limit for May 2011:

- Sample Population $\rightarrow$ 500
- Sample Mean($\overline{Y}$) $\rightarrow$ 2887$
- Sample standard deviation(s_2) $\rightarrow$ 412

Therefore, for a experiment having sample size ≥ 30 & unknown population variances, for,ula implemented for calculation of confidence interval for the difference in mean will be,

$$\overline{X} - \overline{Y} \pm \sqrt{\frac{{S_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

For this case, Confidence Interval for difference in mean credit limits will be,

$$[201.17, 302.82]$$

Part B

Perform an appropriate 5% level test to see if the mean credit limit of all credit

cards issued in May 2011 is greater than the same in January 2011. Be sure to specify the hypotheses you are testing, and justify the choice of your test. State your conclusion.

Solution

Null Hypothesis(H_o) : $(mean\ credit\ limit)_{May} - (mean\ credit\ limit)_{January} > 0$

Alternative Hypothesis(H_a) : $(mean\ credit\ limit)_{May} - (mean\ credit\ limit)_{January} \leq 0$

For $\alpha = 0.05$

As, population variance is unknown we will use T-tests to calculate test statistic.

$$t = (\overline{X} - \overline{Y})/\sqrt{({s_1}^2/n) + ({s_2}^2/m)}$$

Using this, value of **t** comes out to be: 9.71

Now, For a left-tail alternative,

- reject H_o, if t $\leq$ -t_α
- accept H_o, if t $> -t_\alpha$, where t_α is the critical value.

  From caluculation through R code -t_α comes out to be: 1.646.

  Therefore, it is evident that H_o is accpted, which in turn shows that mean credit limit of May 2011 is greater than mean credit limit for January 2011.

0.4 R Code

0.4.1 Question 1

```
#boxplot module

library(gdata)
library(xlsx)
#reading input file
data = read.xlsx("bp.xlsx", sheetName = "Sheet1", header = TRUE)

#converting the data into numberic foramt for frther calculation
armsys=as.numeric(data[,1])
fingsys = as.numeric(data[,2])

boxplot(armsys,fingsys, names=c("armSys", "fingSys"))

#PART B

#histogram of armsys and a normal curve with mean and
#standard deviation equal to that of sample

hist(armsys,prob=TRUE)
a=mean(armsys)
v=sd(armsys)
curve(dnorm(x,mean = a,sd=v), col=3, lty=2,lwd=2,add=TRUE)

#histogram of fingsys and a normal curve with mean and
#standard deviation equal to that of sample
hist(fingsys,prob=TRUE)
b = mean(fingsys)
c = sd(fingsys)
curve(dnorm(x,mean=b,sd=c),col=3,lty=2,lwd=2,add=TRUE)

#Q-Q plot for both datasets to identify whether their
#distribution is normal or not
qqnorm(armsys)
qqline(armsys)

qqnorm(fingsys)
qqline(fingsys)

#Part C

X = mean(armsys)
Y = mean(fingsys)
v1 = var(armsys)
v2 = var(fingsys)
alpha = 0.05

# Assuming that the two sample have equal variances
```

```
CI  = X - Y + c(-1, 1)*qnorm(1-(alpha/2))*sqrt(v1 + v2)

print(CI)

# Part D
Alpha = 0.05

# If |t| > PValue, we reject Null Hypothesis

# Null Hypothesis:: X - Y = 0
# Alternative Hypothesis:: X - Y != 0

# Preparing Test Statistics

l1 = length(armsys)
l2 = length(fingsys)

pooledSamVar = ((l1 - 1)*v1 + (l2 - 1)*v2)/(l1 + l2 - 2)

t = (X - Y)/pooledSamVar*sqrt(1/l1 + 1/l2)

PValue = qt(1-(Alpha/2), (l1 + l2 - 2))

if( PValue <= t &&  PValue >= -t  ) {
  print("Null Hypothesis Rejected")
} else { print("We can not Reject Null Hypothesis")
}
print(PValue)
```

0.4.2 Question 2

```
x = 9.02#sample mean
u = 10#population mean
s = 2.22#sample standared error
n = 20#sample size
#calculating test statistic
t= (x-u)/(s/sqrt(n))
t
#PART E
# mean of the whole population
X = 10
#Sample mean, standard deviation and sample size
X_bar = 9.02
sd = 2.22
n= 20

#test statistic t
t = (X_bar - X)/(sd/sqrt(n))
t
#to generate p-values given value of test statistic & degree of freedom
pt(t,df=n-1)
count <- 0
for(i in 1:10000){
```

```
t_sim <- rt(1,n-1)
if(t_sim < t){
count <- count + 1
}
}
p_val <- count/10000
p_val

#PART F
aplha=0.05
if (p_val<alpha)
{
  print("Reject Null Hypothesis")
}else{
  print("Accept Null Hypothesis")
}
```

0.4.3 Question 3

```
#standard error for January 2011
s2 = 365
#standard error for May 2011
s1 = 412
#mean credit limit for January 2011
y = 2635
#mean credit limit for May 2011
x = 2887
#number of samples for calculation of mean credit limit
#of January 2011
m = 400
#number of samples for calculation of mean credit limit
#of May 2011
n = 500
alpha = 0.05

#Part A
# Calculating Confidence Interval for difference in means
#of two samples
a=(s1^2/n)+(s2^2/m)
z = qnorm(1-(alpha)/2)
lower_confidence = x-y-z*sqrt(a)
upper_confidence = x-y+z*sqrt(a)
lower_confidence
upper_confidence

#Part B

a=(s1^2/n)+(s2^2/m)
#calculation of degree of freedom for calculating t_alpha
#using Satterthewaite Equation
v = (a)^2/(((s1^4)/(n^2*(n-1)))+((s2^4)/(m^2*(m-1))))
#calculating test statistic
t = (x-y)/sqrt(a)
```

```
#calculating critical value t_alpha
t_alpha = qt(alpha,v)

if(t > -t_alpha)
{
  print("H_0 is accepted, i.e. mean credit limit of cards issued
  in May 2011 is more than that issued in January 2011")
}
if(t <= -t_alpha)
{
  print("H_0 is rejected, i.e. mean credit limit of cards issued
  in May 2011 is not more than that issued in January 2011")
}
```