Data Analytics Mini Project 3

M.Tech- Group 9.2

Name of the Student: Vijay Deshpande
Student ID: 201761003
Role: Question 2

Name of the Student: Jayant Mahawar
Student ID: 201761004
Role: Question 1 and 3

Exercise-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.

(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.

(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.

(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?

(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?

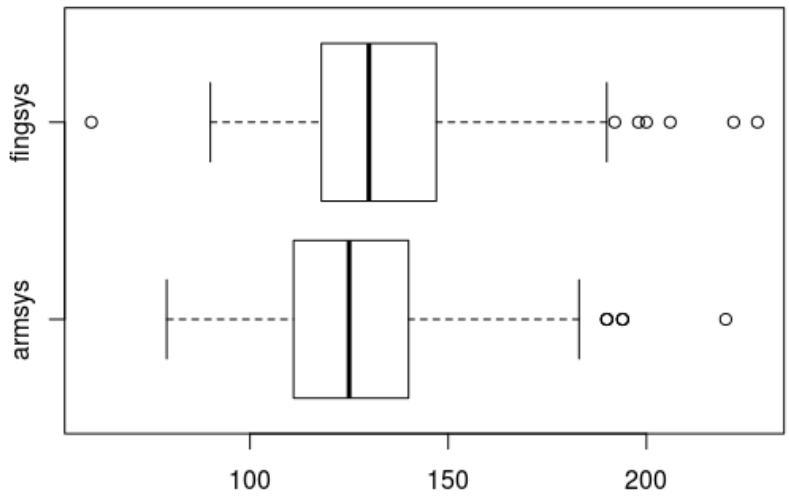
(e) Do the results from (c) and (d) seem consistent? Justify your answer.

Solution

a)
This is the summary of the given data

```
> summary(mydata)
      armsys          fingsys
 Min.   : 79.0   Min.   : 60.0
 1st Qu.:111.5   1st Qu.:118.0
 Median :125.0   Median :130.0
 Mean   :128.5   Mean   :132.8
 3rd Qu.:140.0   3rd Qu.:146.5
 Max.   :220.0   Max.   :228.0
```

Boxplot of the given data

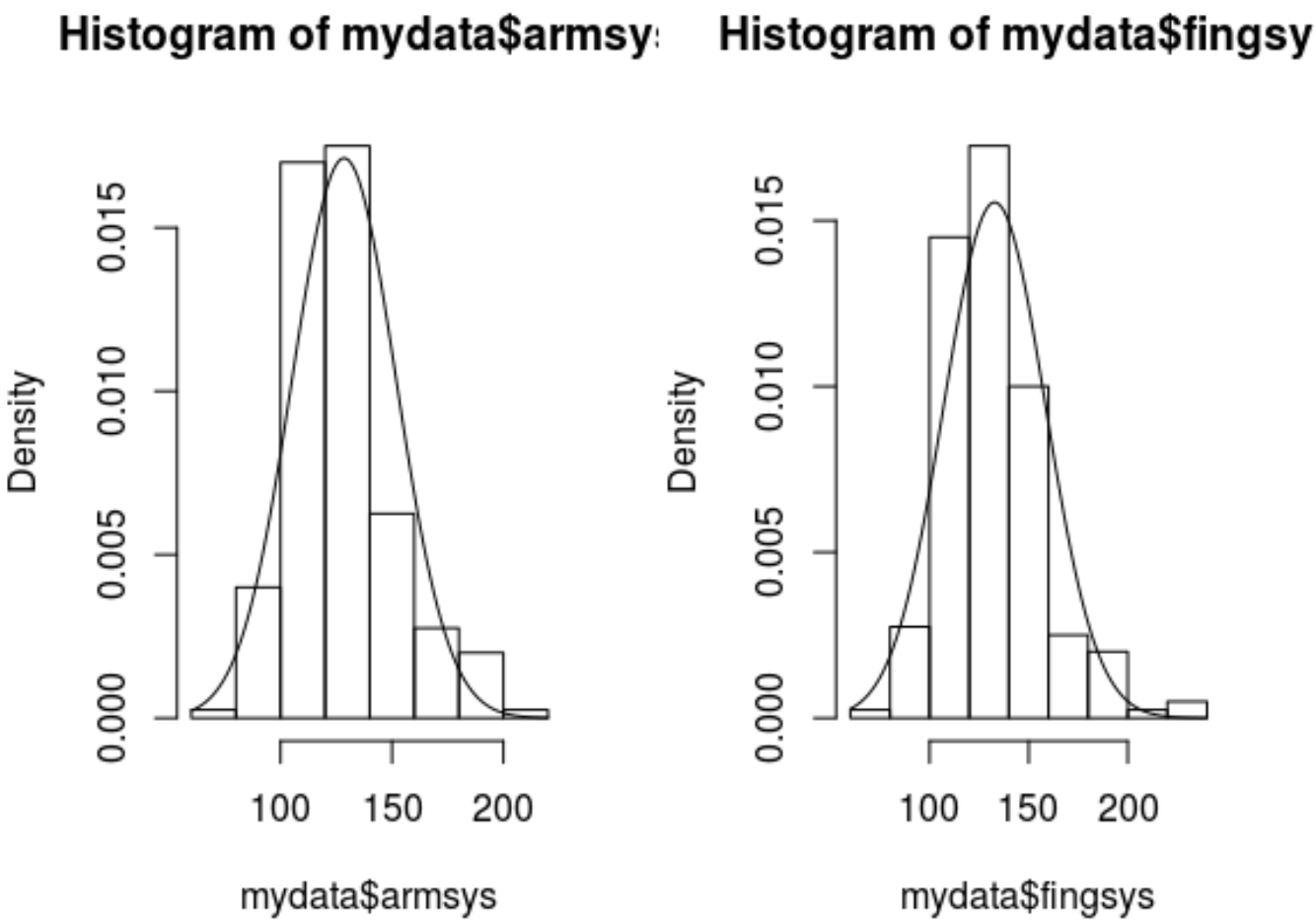


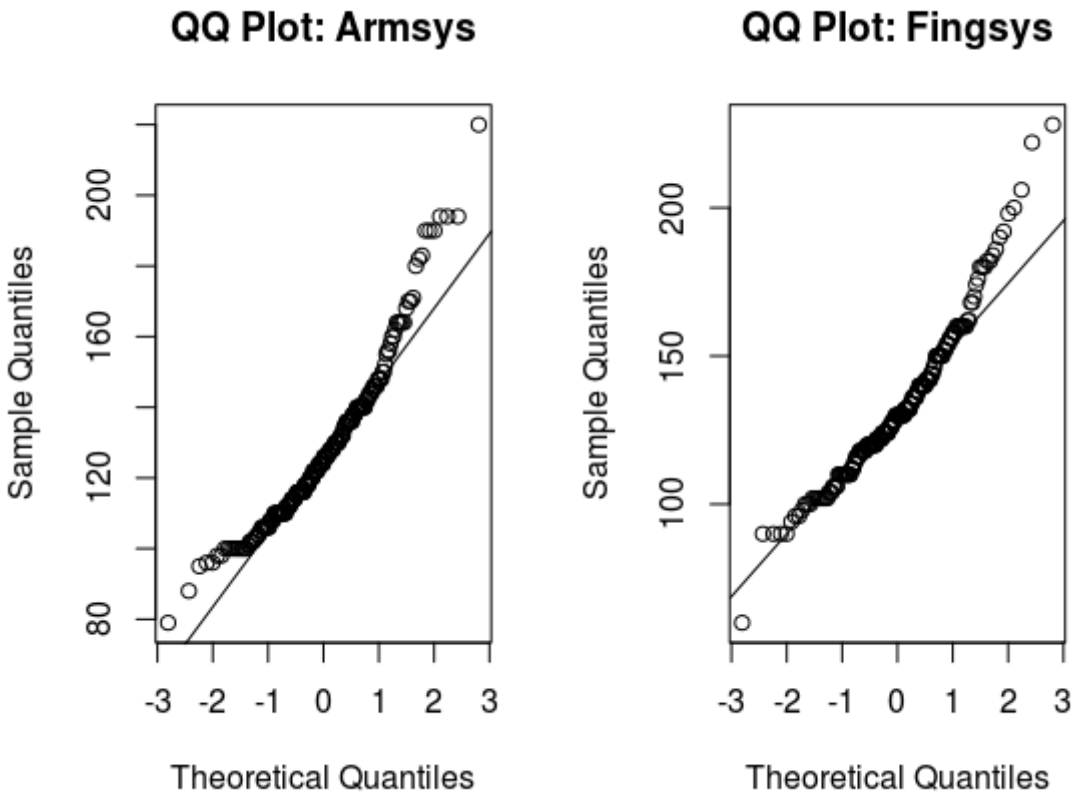
Finding interquartile range of two methods of measuremets.

```
> IQR(mydata$armsys)
[1] 28.5
> IQR(mydata$fingsys)
[1] 28.5
```

Since two datasets gives same interquartile range also from the above boxplot we can see that boxes overlaps with both median so no difference can be claimed in distributions.

b)





From the above histogram and qq plot the data seems to be normal . In qq plot the highly densile region lie above the line. The deviation at tail are due to presence of outliers in the data as seen from the boxplot above.

c)

95% confidence interval for difference of mean

Confidence interval for the difference of means(known standard deviations)

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

Here we have $\bar{X}$=128.5 , $\bar{Y}$=132.8 , n=m=200 , σ X= 23.287 , σ Y= 25.65, α=0.05
Here z0.05/2 = z0.025 = 1.96
We get, lower limit L = -9.1 ,Upper limit U= 0.5
95% confidence interval I=(-9.1,0.5)
the difference between means in the population is likely to be between -9.1 and 0.5, here zero belongs to

the confidence interval. So we can say two methods have identical mean.
In this process, we assume that our data follows normal distribution. Since our sample is large enough ,by
Central Limit Theorem our data will follow normal distribution.

(d)
If both sample sizes are 30 or larger the central limit theorem is in effect. The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

If the population variances are unknown, the sample variances are used.

Assumptions :

Two independent random samples
Each drawn from a normally distributed population (Q:1(b))

Hypotheses
H_0 : $\mu_1 = \mu_2$
H_1 : $\mu_1 \neq \mu_2$

Test statistic
This is a two sample z test.

Distribution of test statistic
If the assumptions are correct and H_0 is true, the test statistic is distributed as the normal distribution.

Decision rule
With $\alpha = .05$, the critical values of z are -1.96 and +1.96. We reject H_0 if $Z < -1.96$ or $Z > +1.96$.

Statistical decision
Accept H_0 because $|-1.753323| < 1.96$.

Conclusion
From these data, it can be concluded that the population means are equal. A 95% confidence interval would give the same conclusion.
p=2*pnorm(-abs(z))
[1] 0.07954652
p-Value is greater than the significance level 0.05.

(e)
Results in part (c) and part (d) seems consistent. In (c) we find confidence interval of difference of mean and zero belongs to interval which show both mean are identical. Also in (d) we get the same result by hypothesis testing.

Exercise-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.

(a) Set up the null and alternative hypotheses.
(b) Which test would you use? What is the test statistic? What is the null distribution of the test statistic?
(c) Compute the observed value of the test statistic.
(d) Compute the p-value of the test using the usual way.
(e) Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare? (8 points)
(f) State your conclusion at 5% level of significance.

Solution:

(a)
Null Hypothesis: H_0 :$\mu=10$
Alternative Hypothesis: H_1 :$\mu>10$

b) Since the given sample size is less than 30 so we will use T test

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Null distribution of test statistic is t-distribution using n-1 degree of freedom.

c) Observed value of test statistic is -1.974186.

d) The p-value for test is 0.9684606

e)We simulate the function to calculate p value 3000 times by creating random sample with given parameters of the distribution we saw that average p value of the simulation (i.e 0.964) is close to observed p value .

f)
Conclusion at 5% level
Critical value F is 1.729133

The test statistic -1.974186 is less than the critical value of 1.729133. Hence, at .05 significance level, we can accept null hypothesis of mean=10.

Exercise -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. 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.

Is your decision in Question 3 of mini project 2 different from this or same?

Solution:

μ_x=2887
μ_y=2635
σ_x=412*412
σ_y=365*365
n_x=500
n_y=400

Null Hypothesis : $\mu_x = \mu_y$
Alternative Hypothesis : $\mu_x < \mu_y$
Since number of observations are lagre(>30) we will perform Z-test.

$$Z = \frac{\mu_x - \mu_y - D}{\sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)}}$$

According to null hypothesis we assumed no differnce. So, D=0.

$Z_\alpha = 1.64$ for $\alpha = 0.05$

For right tailed distribution we reject hypothesis when $|Z| \geq Z_\alpha$.

So we reject the hypothesis i.e. there is a variability in two population.

Decision in mini project 2 is same as the decision in mini project 3.

R CODE

1)

```
summary(mydata)
boxplot(mydata,horizontal=TRUE)

IQR(mydata$armsys)
IQR(mydata$fingsys)
```

#Histogram

```
par(mfrow=c(1,2))
hist(mydata$armsys,probability = TRUE)
mean(mydata$armsys)
[1] 128.52
sd(mydata$fingsys)
[1] 23.28758
x=mydata$armsys
curve(dnorm(x,mean=128.52,sd=23.287),add=TRUE)
hist(mydata$fingsys,probability = TRUE)
mean(mydata$fingsys)
[1] 132.815
sd(mydata$fingsys)
[1] 25.6482
x=mydata$fingsys
curve(dnorm(x,mean=132.815,sd=25.6482),add=TRUE)
```

#QQPLOT

```
par(mfrow=c(1,2))
qqnorm(mydata$armsys,main="QQ Plot: Armsys")
qqline(mydata$armsys)
qqnorm(mydata$fingsys,main="QQ Plot: Fingsys")
qqline(mydata$fingsys)
```

d)

```
z = (mean(mydata$armsys) - mean(mydata$fingsys)) /
(sqrt(var(mydata$armsys)/length(mydata$armsys) + var(mydata$fingsys)/length(mydata$fingsys)))
z
[1] -1.753323
qnorm(0.025)
[1] -1.959964
p=2*pnorm(-abs(z))
[1] 0.07954652
```

2)

c)
#Test statistic
```
xbar=9.02 # sample mean
mu0=10 #hypothesized value
sd=2.22 #sample standard deviation
n=20 #sample size
t= (xbar-mu0)/(sd/sqrt(n))
t #test statistic
[1] -1.974186
```

d)
#p value of test statistic
```
pval =1-pt(t,19) #p-value
pval
[1] 0.9684606
```

e)
#Monte-Carlo Simulation
```
xbar=9.02 # sample mean
mu0=10 #hypothesized value
sd=2.2 #sample standard deviation
n=20 #sample size
nSims=10000
pval=0
p <-numeric(nSims)
sum=0
for( i in c(1:nSims))
{
a=rnorm(n, mean=xbar, sd=sd)
meana=mean(a)
t = (meana-mu0)/(sd(x)/sqrt(n))
pval= (1-pt(-t,n-1))
sum=sum+pval
}
sum/nSims
[1] 0.964
```

f)**#Critical value**
```
T=qt(1-0.05,df=20-1)
T
[1] 1.729133
```

3)
```
m1=2887
m2=2635
var1=365*365
var2=412*412
n1=400
n2=500
```

```
z = (m1 - m2) / (sqrt(var1/n1 + var2/n2))
z
[1] 9.717132
```