Data Analytics

Group-3

Mini Project 3

Date- 24/10/17

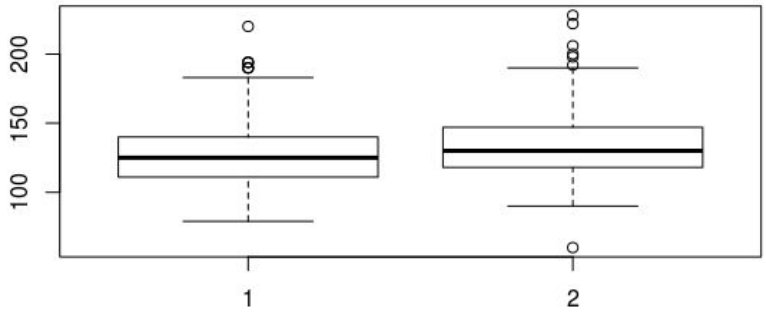
Name	Contribution(s)
Mayank Sharma (201452040)	Exercise-2(a)(b)(c)
Ravi Kiradoo (201451007)	Exercise-2(d)(e)(f)
Manish Singla (201452018)	Exercise-3
Rahul Chaurasia (201451039)	Exercise-1(a)
Chirag Garg (201451052)	Exercise-3
Shikar Dhing (201452021)	Exercise-1(c)
Himanshu Soni (201452046)	Exercise-1(b)
Sahil Luthra (201452043)	Exercise-1(e)
Shubham Solanki (201452001)	Exercise-1(d)

Exercise 1:

Question: 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.

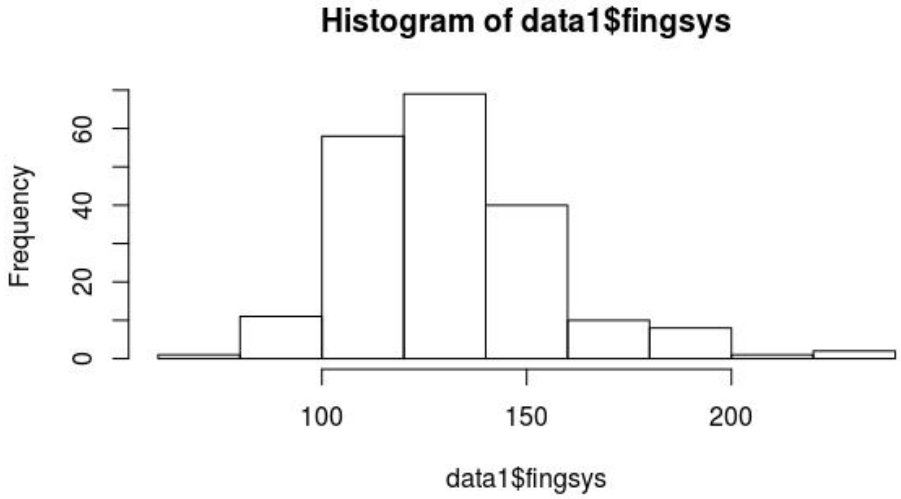
(a): Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer.

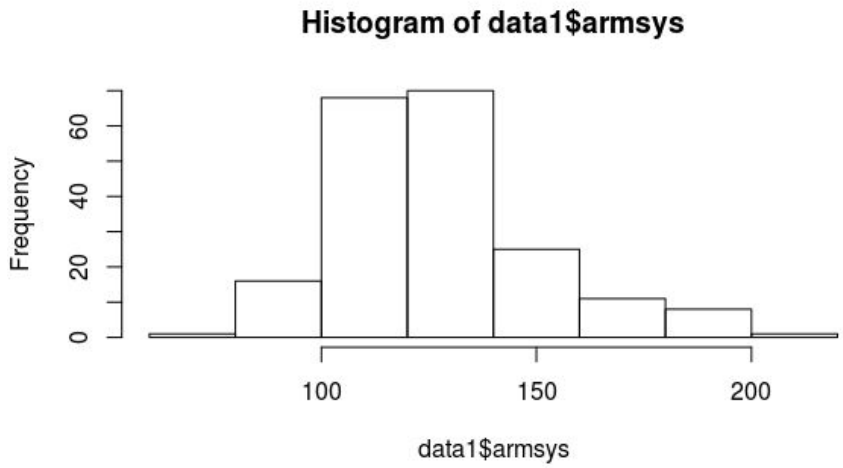
1→ Boxplot of arm systolic blood pressure
2→ Boxplot of finger systolic blood pressure



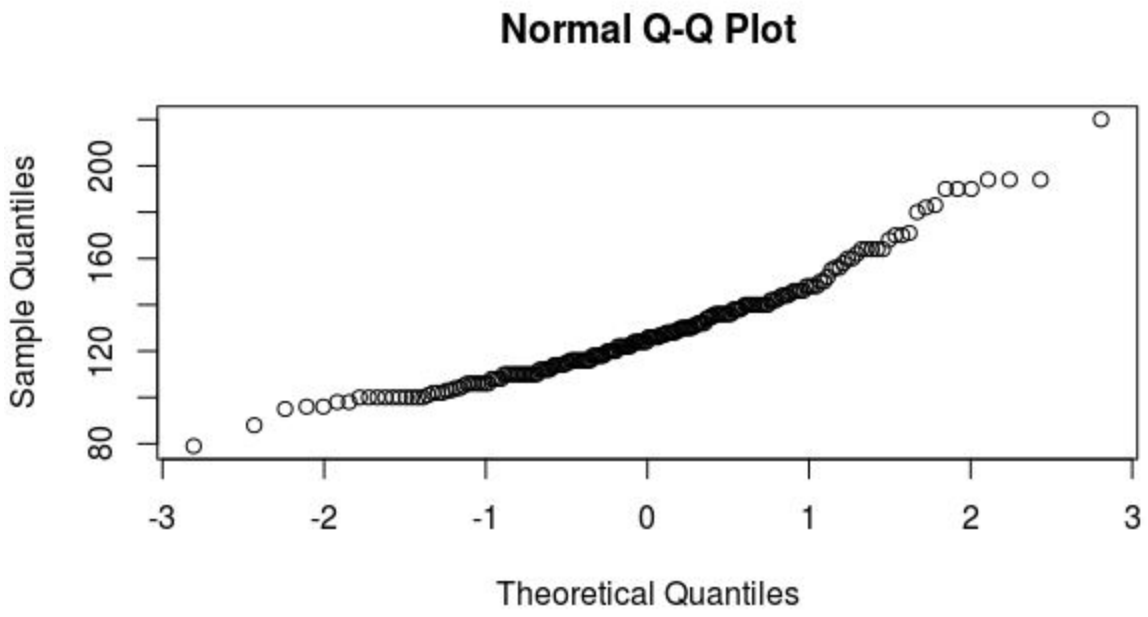
Justification: The two distribution obtained are similar as seen by boxplot. There are few more outliers in finger method as compared to arm method of measuring the blood pressure.

(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.

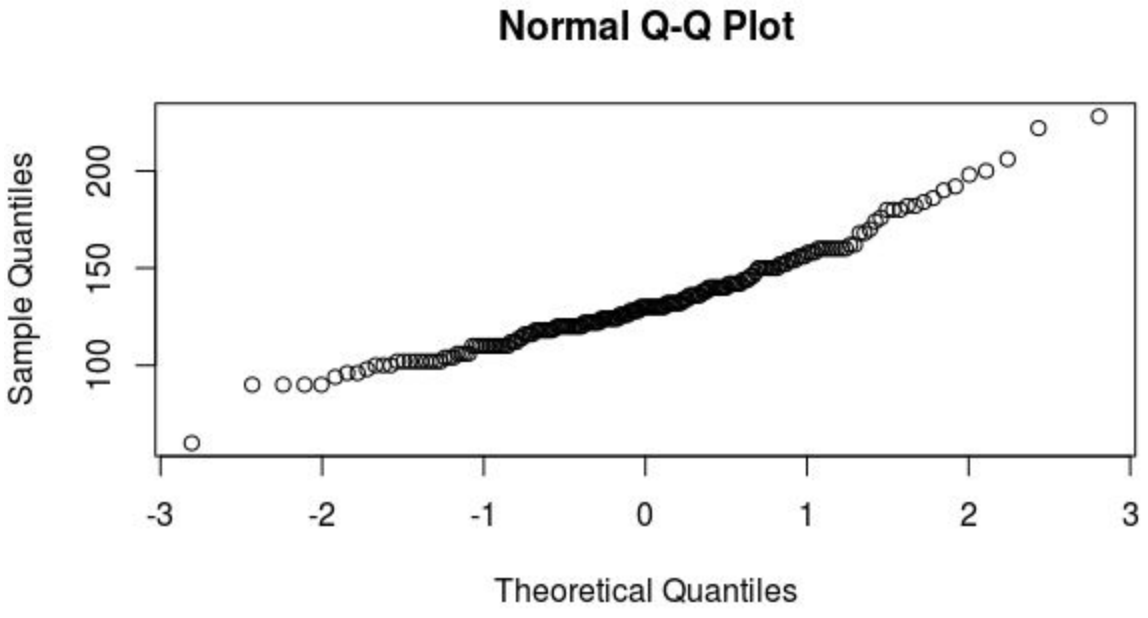




→ Normal Q-Q plot of data1$fingsys



→ Normal Q-Q plot of data1$armsys



Justification: From the histogram and QQ-plot it can be inferred that the data follows normal distribution. But in case of finger method the distribution is slightly more skewed with respect to the arm method.

(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?

```
        Welch Two Sample t-test

data:  data1$armsys and data1$fingsys
t = -1.7533, df = 394.35, p-value = 0.08032
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.1109747  0.5209747
sample estimates:
mean of x mean of y
  128.520   132.815
```

Justification: Since the p-value obtained from the z-test is 0.08(greater than 0.05), there is weak evidence against the null hypothesis, so we fail to reject the null hypothesis.

Which infers that there is no evidence that there is difference between the means of the two methods of measurement taken into consideration.

(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?

u1 = mean of arm method,
u2 = mean of finger method

H0: u1 - u2 = 0 (Null Hypothesis)
H1: u1 != u2 (Alternative Hypothesis)
alpha= 0.05
We will conduct two-tailed test to check if we can ignore null hypothesis

u11-> sample mean of arm method,
u22-> Sample mean of finger method
s1-> sample standard deviation for arm method
s2-> sample standard deviation for finger method

Z = (u11-u22)-(u1-u2)/SD
SD = sqrt(s1^2/n1 + s2^2/n2) # n1,n2 sample size
mean_diff<-mean(Data$armsys)-mean(Data$fingsys)
comb_sd<-var(Data$armsys)/sqrt(nrow(Data)) +
var(Data$fingsys)/sqrt(nrow(Data))
comb_sd<- sqrt(comb_sd)
Z_ob= (mean_diff - 0)/comb_sd
Z_alpha_cal<-qnorm(1-0.05/2)

We can reject H0 (Null hypothesis) if |Z_ob| > Z_alpha_cal

Z_ob =-0.4662346
Z_alpha_cal (+) = 1.96
Z_alpha_cal (-) = -1.96

since |Z_ob| < Z_alpha_cal
So we cannot reject Null hypothesis.

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

Consider the 95% confidence interval

$$-12.1 \leq \mu_1 - \mu_2 \leq -2.49$$

Since zero is not in the interval, the null hypothesis that $\mu_1 - \mu_2 = 0$ can be rejected at the 0.05 level. Moreover, since all the values in the interval are negative, the direction of the effect can be inferred: $\mu_1 < \mu_2$.

Whenever a significance test rejects the null hypothesis that a parameter is zero, the confidence interval on that parameter will not contain zero. Therefore either all the values in the interval will be positive or all the values in the interval will be negative. In either case, the direction of the effect is known.

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.
Sol^n :

H_0: $\mu = 10$, Null Hypothesis

H_1: $\mu > 10$, Alternative Hypothesis

(b) Which test would you use? What is the test statistic? What is the null distribution of the test statistic?
Sol^n :

We would use “t” test to get our statistics since population variance is unknown. Test statistics is “t” statistics. The null distribution of our statistics is “t” distribution because our sample size is very small(<30).

(c) Compute the observed value of the test statistic.
Sol^n :

$$\text{Test statistics } t = \frac{(\bar{x} - \mu)\sqrt{n}}{S}$$

Where $\bar{x}$ = sample mean

μ = population mean

n = sample size

S = sample variance

By substituting the values in above formula, we get:

$$t = \frac{(9.02 - 10)\sqrt{20}}{2.22}$$

t = -1.974186

(d) Compute the p-value of the test using the usual way.

Sol^n :

By calculating the p-value using r with t = -1.974186 and degree of freedom = 19, we get (the r code is in appendix):

p = 0.94 for alternative hypothesis

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

Sol^n :

For this, we run 100 simulations of following process:

We generate random t distribution values with a size (n) = 20 and degree of freedom = 19. Then, we calculate the probability of the value being greater than t = -1.974186.

After running 100 simulations, we get a sample of p-values with size 100 and then we calculate the mean of that sample which comes out to be 0.96 that is approximately equals to the calculated p-value. And because p-values are unbiased so the p-value of population would be equal to the expectation of sample p-value and hence we have used the sample mean to compare the two.

(f) State your conclusion at 5% level of significance.

Sol^n :

For the null hypothesis, we get p-value of 0.06 which is greater than α = 0.05 so we do not have strong evidence to reject the null hypothesis.

Same goes for the alternate hypothesis, we get p-value of 0.94 which is greater than α = 0.05 so we do not have strong evidence to reject the alternate hypothesis either.

Exercise 3:

Question: 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.

H0(Null Hypothesis): Difference between mean credit limit in May 2011 and Jan 2011 is less than or equal to zero.

H1 (Alternate Hypothesis): Difference between mean credit limit in May 2011 and Jan 2011 is greater than zero.
We conducted t-test to validate hypothesis.

Question: Is your decision in Question 3 of mini project 2 different from this or same?
Inference:
The p-value is 0 here so we can ignore the null hypothesis, The p-value plays an important role to reject the null hypothesis. We carried out t-test because population variance is not given. The sample variances are also not equal. Thus, the decision in question 3 is same for both Mini Project 2 and mini project 3. In min project2, we took confidence interval and concluded that the credit limit in may is statistically different from credit limit in January. Here, we also obtained the same result through hypothesis testing.

Appendix:

Code:

Exercise 1→

```
#Part (a)

data1=read.csv("bp.csv")#reading the data file
library("ggplot2")
boxplot(data1$armsys,data1$fingsys)# boxplot of the two variables given

#Part(b)
#histograms of the given data set
hist(data1$fingsys)
hist(data1$armsys)

#QQ-plot of the given data set
library("graphics")
qqnorm(data1$armsys)
qqnorm(data1$fingsys)

#Part(c)
#t-test for testing the difference in the means of the two methods with C.I of 95%
t.test(data1$armsys,data1$fingsys)
```

Exercise 2 →

```
#Part(d)
xbar=9.02
mu=10
s=2.22
n=20
t=((xbar-mu)/s)*sqrt(n) #the value of t by formula
p=(1 - 2*pt(-abs(t),df=n-1)) #the p-value by formula

#Part(e)
vec_p<- vector() #initializing empty vector of sample p-values
for( i in seq(from=1,to=100,by=1)){rd = rt(n,n-1)
```

```
p=(sum(rd>t))/n
vec_p<-c(vec_p,p)
}#runs 100 simulations and stores p-values in vec_p
ex_p <- mean(vec_p) #mean of vec_p
```

Exercise 3 →

```
#t-test
alpha <- 1-0.95
m1 <-2635 # Mean credit limit for January
s1 <- 365  # Standard Deviation for January
n1 <- 400  # Sample size for January
m2 <- 2887 # Mean credit limit for May
s2 <- 412  # Standard Deviation for May
n2 <- 500 # Sample size for May
```

For calculating p-value from given hypothesis.

```
#degree of freedom for different sample size
df <- (((s1^2 /n1) + (s2^2/n2))^2)/((((s1^2 /n1)^2)/(n1-1)) + (((s2^2/n2)^2)/(n2-1)))

#standard error
s12.error = sqrt(((s1^2 /n1) + (s2^2/n2)))

#t statistics
t = (m2-m1)/s12.error

cat("The p-value for Hypothesis test: ",(1-pt(t, df = df)))
```