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.

(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.

Report:

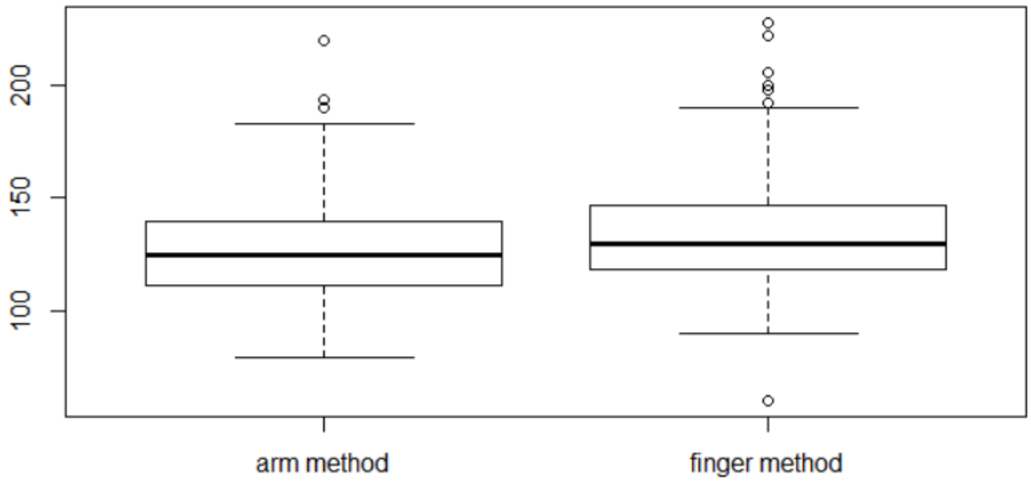
First I need to read data from the document. The document is .xlsx file. I did a change and save it into .csv file. And then I use the command to load data into R:

```
data=read.csv("F:\\6313 statistic for DS\\project\\4\\bp.csv",header=T,sep = ",")
armsys=data$armsys
fingsys=data$fingsys
```

After that, I use the command below to show the distribution of two samples of data.

```
boxplot(armsys,fingsys,names=c('arm method','finger method'))
```

The graph is shown below and we can see that those two samples have similar distribution of data, include IQR, skewness. Even though the data of finger method seems bigger than arm method’s in corresponding position, but we can directly say that those two distributions are same.



(b)

I will get histograms and QQplot of those two sample by using the commands below:

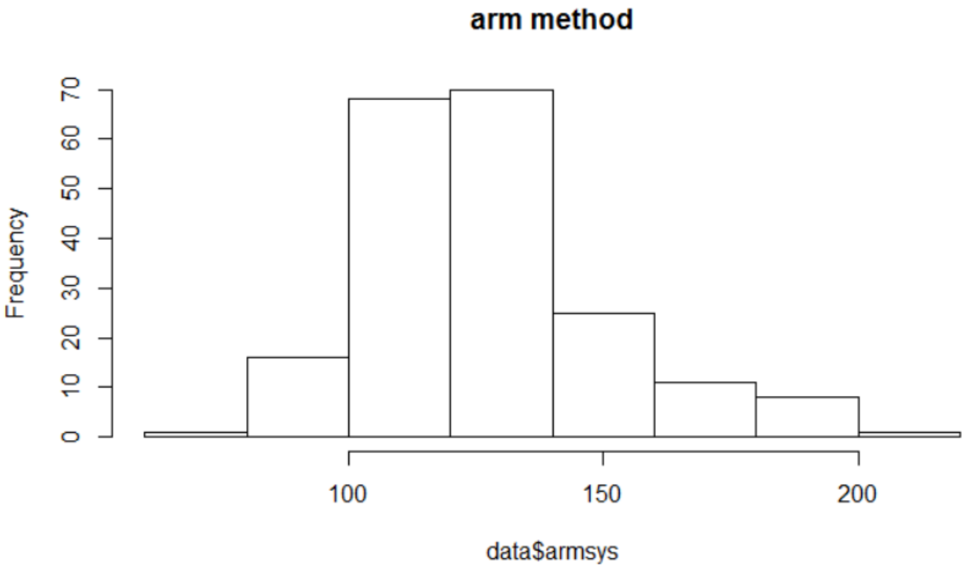
Histograms:

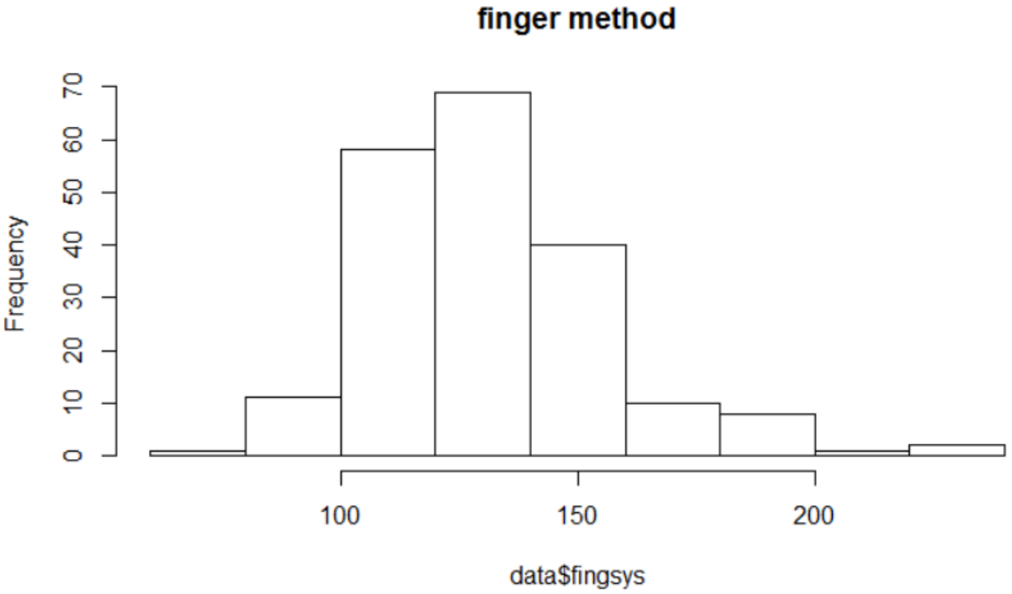
```
hist(armsys, main ='arm method')
hist(fingsys, main ='finger method')
QQplots:
qqnorm(armsys,main='arm method')
qqline(armsys)
qqnorm(fingsys,main='finger method')
qqline(fingsys)
```

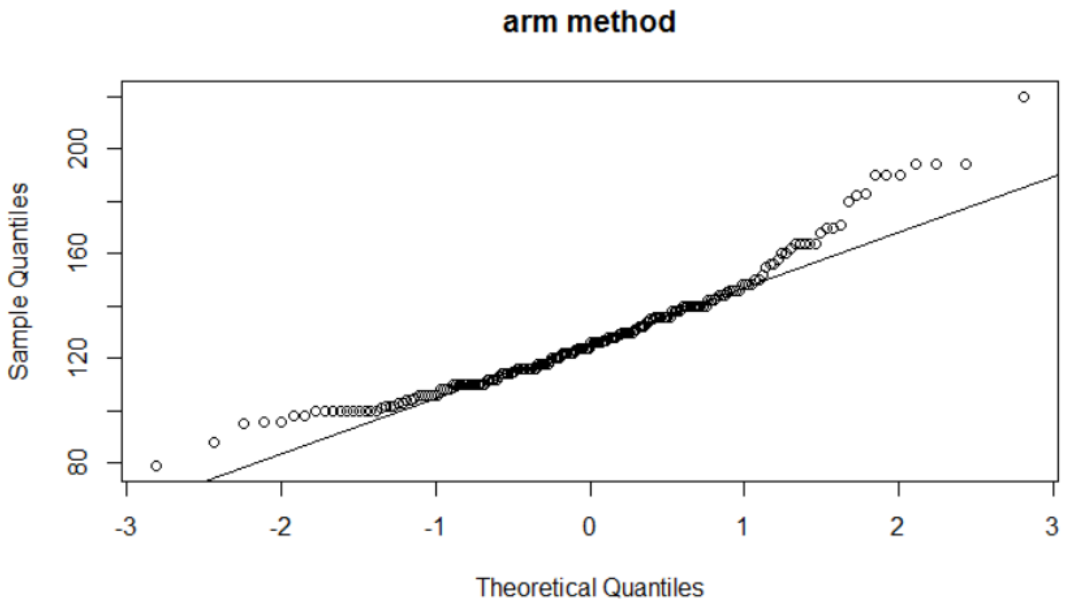
After that I get the graphs below.

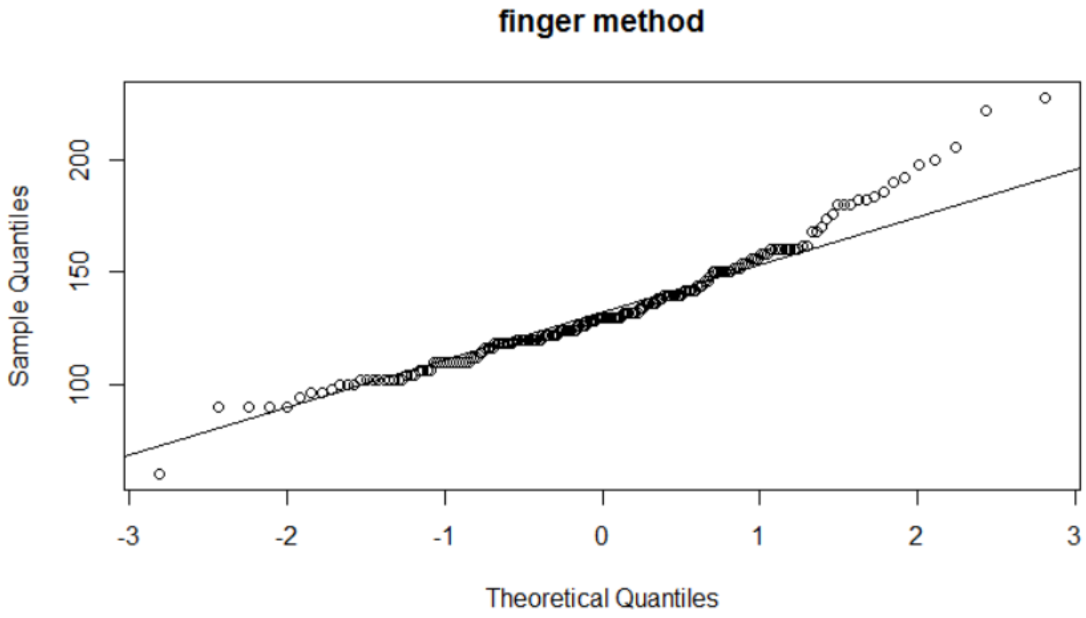
From two histograms I can see that both of the distributions are approximate normal distribution but skewed right.

From QQplots I can see that most of the data which close to central stay on the line, only some data which far from central not stay on the line. So that I can assume they are normal distribution. In another way, because of the large size of each sample (200>30), so that I can assume they are normal distribution.


arm method
Frequency
0
10
20
30
40
50
60
70
100
150
200
data$armsys


finger method
Frequency
0
10
20
30
40
50
60
70
100
150
200
data$fingsys





(c)

From the question, I know that there are two independent samples and unequal variance. I want to calculate CI for u_x-u_y with unknown standard deviation. Because the sample size is large(more than 30), so I choose the function of Z-distribution:

$$\hat{\theta} \pm z_{\alpha/2} \cdot s(\hat{\theta}).$$

The functions in R is shown below:

```
mean(armsys)
mean(fingsys)
n<-200
m<-200
alpha<-0.05
mean(armsys)-mean(fingsys)+c(-1,1)* qnorm(1-alpha/2)*sqrt(sd(armsys) ^2/200+sd(fingsys)
^2/200)
```

The answer is [1] -9.0961939 0.5061939

Because 0 is in the interval.

So we can say that the two methods have identical means.

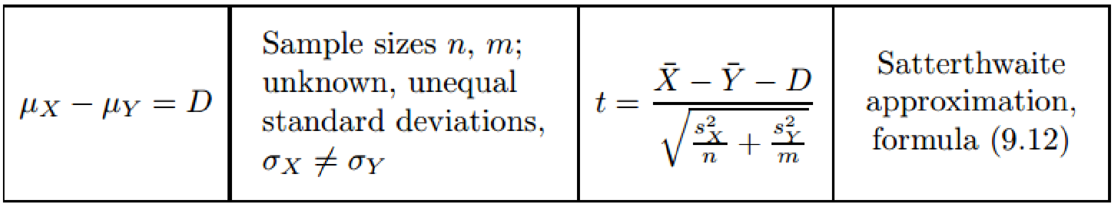
The assumptions I made is assume these two methods are both normal distribution.

The assumption seems hold.

(d)

H0 means $u_x=u_{y,}$, the value is equal between mean of two methods.

H1 means u_x!=u_y, the value is not equal between mean of two methods, this is a two-sided test. From the question we know they are two different samples and we don’t know the variance of population. So that I choose the function:



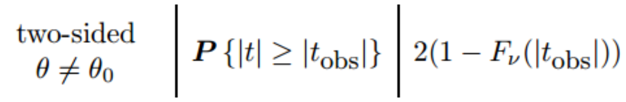
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown, unequal standard deviations, $\sigma_X \neq \sigma_Y$	$t = \dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Satterthwaite approximation, formula (9.12)

Code is shown below:

```
D<-0
T<-(mean(armsys)-mean(fingsys)-D)/sqrt(var(armsys)/n+var(fingsys)/m)
```

The value of T equals [1] -2.62779

So that the function of calculating P-value is:



two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\|t\| \geq \|t_{\text{obs}}\|\}$	$2(1 - F_\nu(\|t_{\text{obs}}\|))$

The function is shown in code as below:

```
Pvalue<-2*(1-pt(abs(T)，199))
```

The answer of Pvalue is [1] 0.009264109

Because the P-value is smaller than value of a, which is 0.05, so that we reject H0.

(e)

The result from (c) and (d) is consistent. The mean of the differences is -4.295 and the confidence interval is [-9.1023998, 0.5123998] at 95% confidence level. -4.295 is in the interval. So 95% chance that there is no difference between two means. The p-value shows that, given the sample, if the null hypothesis is true, then the probability is very low. The confidence interval of p-value calculation present complementary views to indicate that the null hypothesis is not valid.