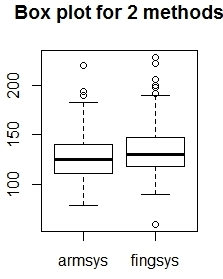
**Mini** **Project** **3**

**Exercise 1: Section 1**

We first convert the txt file to csv file and then use read.csv to read the blood pressure measurements from the same 200 patients using either the arm method (named as arm) or the finger method (named as finger).

(a). We first use the boxplot to examine the distribution of the measurements from the two datasets, which displays the Q1,Q2,Q3 and the range of [Q1‐1.5\*IQR, Q3+1.5\*IQR].

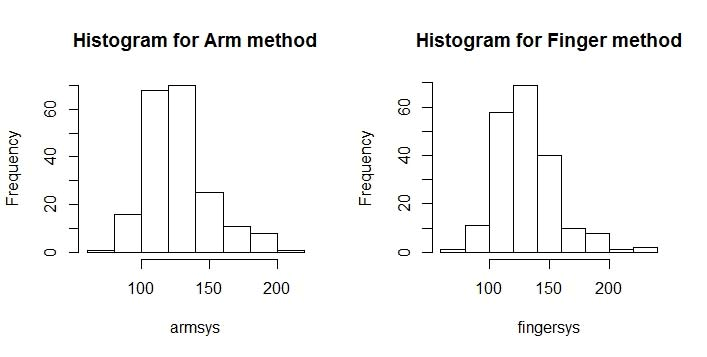


From this boxplot, we can observe that the medians and IQRs are fairly similar between the two sample sets, with the measurements from the Finger method having a slightly higher median value and also there are a number of outliers.

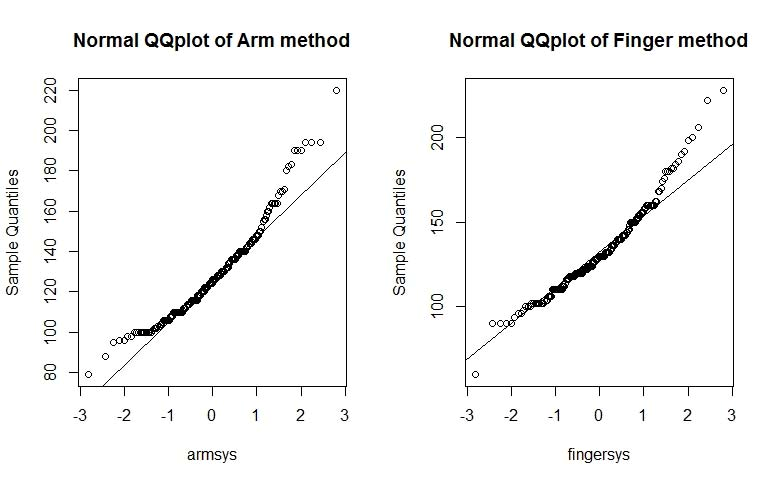
(b). We then constructed the histograms and QQ‐plot for these two datasets. The histogram for Arm is as:

And the histogram for Finger is as:

Both histograms appear to be right-skewed, which means that the mean is larger than the median. Thus, the assumption of normality for these two datasets may not be valid.



We further plotted the QQ‐plots, which is as:



The QQ-plot showed that the two datasets are closely linearly related.

(c). Since the measurements from the Arm and the Finger are from the same 200 patients. They will be considered as paired samples. Thus, we could just calculate the difference between these two methods for each patient, which is saved as dif.

We then calculated the sample mean, sample standard deviation and calculated the 95% confidence interval, which is [‐6.316529, ‐2.273471]. Since 0 is not within this interval, we concluded that the means from these two methods are not identical. In fact, the mean from the arm method appears to be smaller than that from the finger method.

In order to reach this conclusion, we assumed that mean of the differences could be approximated by a normal distribution. In this specific case, since N=200, and could be considered as sufficient large so that the central limit theorem could become valid.

**Section 2: The R code:**

#we read‐in the data set from the file

data = read.csv("bp.csv",header=TRUE, sep="\t")

#we construct the boxplot for the arm and finer data, using the 1.5\*IQR

boxplot(data,main="Box plot for 2 methods")

par(mfrow=c(1,2))

#we then plot the histograms for the arm and finger data

hist(x = data$armsys, main="Histogram for Arm method",xlab="armsys") hist(x=data$fingsys, main="Histogram for Finger method",xlab = "fingersys") par(mfrow=c(1,2))

#we then plot the QQ‐plots for the arm and the finger data

qqnorm(data$armsys, main = "Normal QQplot of Arm method", xlab = "armsys")

qqline(data$armsys)

qqnorm(data$fingsys, main = "Normal QQplot of Finger method", xlab = "fingersys")

qqline(data$fingsys)

#we then calculate the difference between arm and finger dif = data$armsys‐data$fingsys

# calculate the sample mean, standard deviation, and then calculate the 95% CI

alpha = 0.05

mu = mean(dif)

se = sd(dif)/sqrt(200)

interval = mu +c(‐1,1)\*qnorm(1‐alpha/2)\*se

**Exercise 2: Section 1:**

We constructed a function conf.int to randomly generate n random variables following the Bernoulli(p) distribution, and then use these samples to compute the 95% confidence interval. We also calculated the nominal 95% confidence interval for each p and n pairs. The results are summarized in the following table:

Table 1: simulated 95% CI

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | N=5 | N=10 | N=30 | N=50 | N=100 |
| P=0.05 | [0,0] | [0,0] | [0,0] | [‐0.01,0.09] | [0,0.09] |
| P=0.1 | [0,0] | [‐0.05,0.45] | [‐0.02.0.16] | [0.03,0.21] | [0.07,0.21] |
| P=0.25 | [‐0.15,0.55] | [0.02.0.58] | [0.03,0.30] | [0.14,0.38] | [0.16,0.33] |
| P=0.5 | [0.17,1.03] | [0.30,0.90] | [0.35,0.71] | [0.40,0.68] | [0.37,0.57] |
| P=0.9 | [1,1] | [1,1] | [0.70,0.97] | [0.82,0.98] | [0.85,0.97] |
| P=0.95 | [1,1] | [1,1] | [1,1] | [0.87,1.00] | [0.89,0.99] |

Table 2: nominal 95% CI

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | N=5 | N=10 | N=30 | N=50 | N=100 |
| P=0.05 | [-0.14,0.24] | [‐0.08,0.18] | [‐0.03,0.13] | [‐0.01,0.11] | [0.01,0.09] |
| P=0.1 | [‐0.16,0.36] | [‐0.08,0.28] | [‐0.01,0.21] | [0.02,0.18] | [0.04,0.16] |
| P=0.25 | [‐0.13,0.63] | [‐0.02,0.52] | [0.10,0.40] | [0.13,0.37] | [0.16,0.33] |
| P=0.5 | [0.06,0.94] | [0.19,0.81] | [0.32,0.68] | [0.36,0.64] | [0.40,0.60] |
| P=0.9 | [0.64,1.16] | [0.71,1.08] | [0.79,1.01] | [0.82,0.98] | [0.84,0.96] |
| P=0.95 | [0.76,1.14] | [0.81,1.08] | [0.87,1.03] | [0.89,1.01] | [0.91,0.99] |

From the above two tables, we could clearly observe that with increasing N, the simulated 95% CI matches more closely to the nominal 95% CI. It appears N=30 could yield fairly close CI compared to the nominal CI. Thus, we would recommend at least N>=30, and ideally N>=50.

This observation holds for all different p values, as the key mechanism here is the central limit theorem, which states that when N becomes larger, the mean of the sample more closely resemble a normal distribution, which is how the simulated CI calculation is based.

**Section 2: R code:**

#generate a function to generate random numbers following the bernoulli

#distribution using the Monte Carlo method conf.int <‐ function(n, p, alpha) {

x = runif(n)

y = (x<p)

z = as.numeric(y) #generate n random numbers following Bernoulli(p)

#we then calculate the sample mean, standard deviation and the confidence interval;

mu = mean(z)

sigma = sqrt(mu\*(1‐mu))

ci <‐ mu + c(‐1, 1) \* qnorm(1 ‐ (alpha/2)) \* sigma/sqrt(n)

return(ci)

}

#define p, aplha and n values;

p = 0.90 alpha = 0.05 n=5

conf.int(n, p, alpha)

#calculate the nominal confidence interval

nominal.ci = p + c(‐1,1) \* qnorm(1 ‐ (alpha/2)) \* sqrt(p\*(1‐p)/n)

nominal.ci