**Statistical Methods for Data Science (Fall 2018)**

**Mini Project 5 (Solution)**

# Haoran Lou

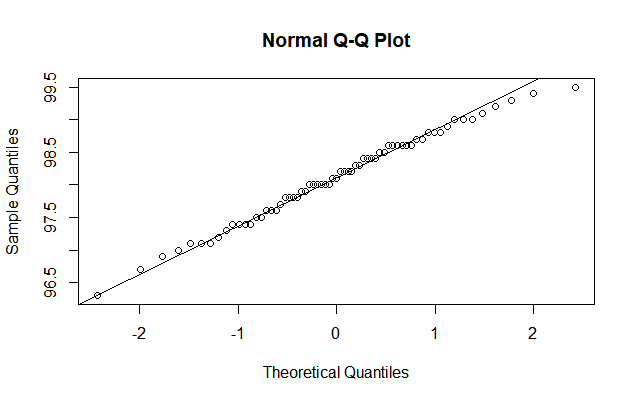
# hxl180012

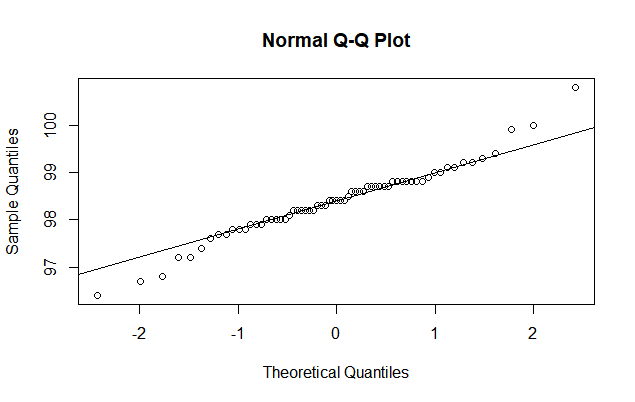
# Bo Yang

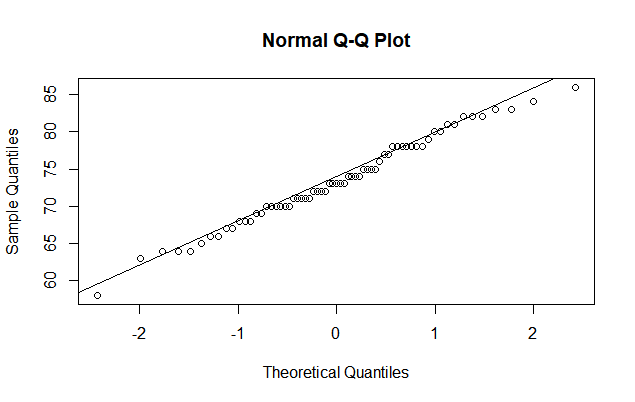
# bxy180002

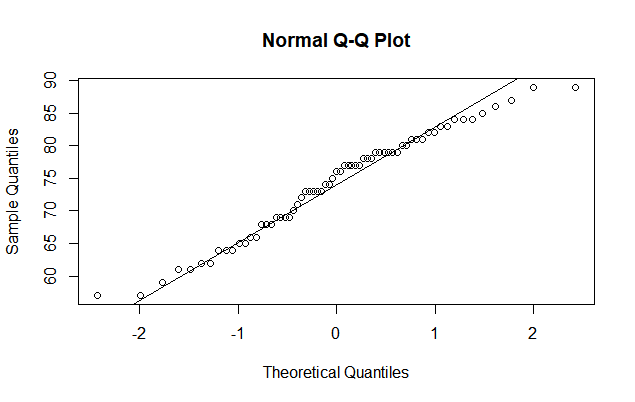
**1.**

First we generate 4 qqplots of male temperature, female temperature, male heart rate, female heart rate.





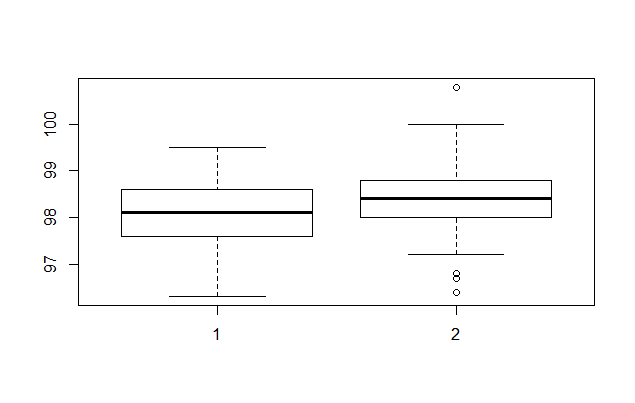




We can know from the plots that the data is approximately in normal distribution.

**(a)**

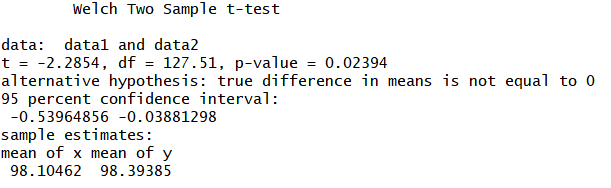
We have side-by-side boxplot of temperature of male and female below:



Then hypothesis:

H0: Males and females same in mean body temperature.

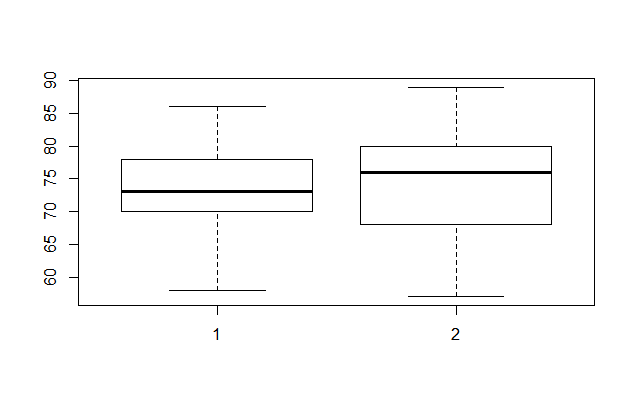
H1: Males and females differ in mean body temperature.



Based on the results above, use t-test(normal) to get the p-value=0.0239<0.05. So, we reject the null hypothesis. Males and females differ in mean body temperature.

**(b)**

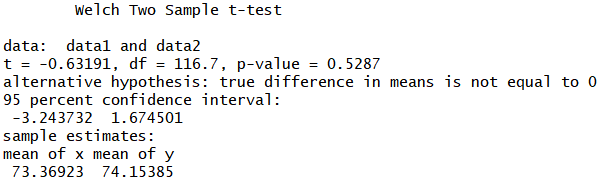
We have side-by-side boxplot of heart rate of male and female below:



Then hypothesis:

H0: Males and females same in mean heart rate.

H1: Males and females differ in mean heart rate.

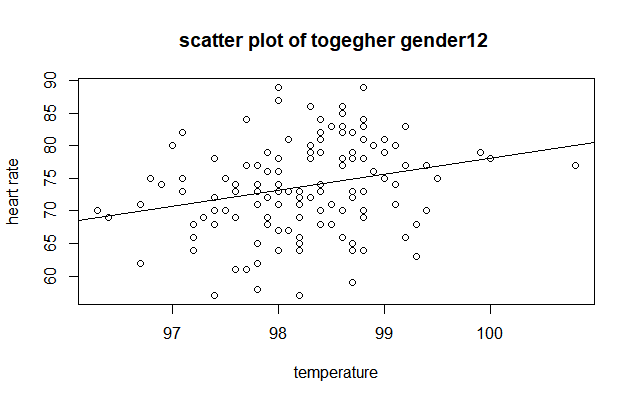


Based on the results above, use t-test(normal) to get the p-value =0.5287>0.05. So, we can’t reject the null hypothesis. Males and females same in mean heart rate.

**(c)**

**(i) Is there a linear relationship between body temperature and heart rate?**

The scatter plot of temperature and heart rate is blow:

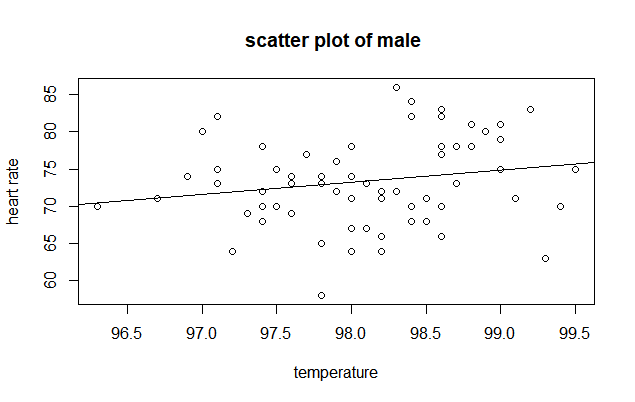


From the scatter plot we can see that there isn’t a linear relationship between body temperature and heart rate. And the correlation = 0.253654.

(P.S. p-value of correlation-test is 0.003591. Can we reject null hypothesis and consider it has correlation?......)

**(ii) Is there a linear relationship between body temperature and heart rate of male?**

The scatter plot of temperature and heart rate is blow:

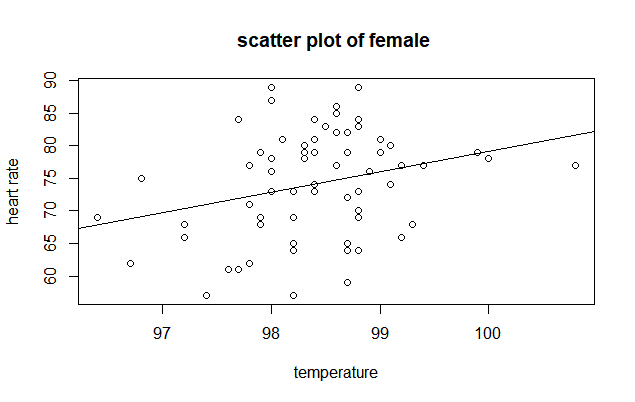


From the scatter plot we can see that there isn’t a linear relationship between body temperature and heart rate. And the correlation = 0.1953894.

(p-value of correlation-test is 0.1184252).

**(iii) Is there a linear relationship between body temperature and heart rate of male?**

The scatter plot of temperature and heart rate is blow:



From the scatter plot we can see that there isn’t a linear relationship between body temperature and heart rate. And the correlation = 0.2869312.

(P.S. p-value of correlation-test is 0.02048, can we think that it has correlation?....).

**2.**

**(a)**

Use n=30 and lamda=0.1

I got estimates of coverage probability of large sample interval is 0.9260, estimates of coverage probability of bootstrap percentile method interval is 0.9390.

**(b)**

Estimates of coverage probability of large sample interval:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| lamda  n | 0.01 | 0.1 | 1 | 10 |
| 5 | **0.3756** | **0.3830** | **0.3704** | **0.3680** |
| 10 | **0.8550** | **0.8582** | **0.8666** | **0,8662** |
| 30 | **0.9222** | **0.9260** | **0.9286** | **0.9290** |
| 100 | **0.9432** | **0.9460** | **0.9484** | **0.9454** |

Estimates of coverage probability of bootstrap percentile method interval:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| lamda  n | 0.01 | 0.1 | 1 | 10 |
| 5 | **0.9010** | **0.8966** | **0.8994** | **0.8954** |
| 10 | **0.9152** | **0.9210** | **0.9216** | **0.9164** |
| 30 | **0.9366** | **0.9390** | **0.9430** | **0.9396** |
| 100 | **0.9462** | **0.9478** | **0.9516** | **0.9502** |

**(c)**

(1) In case of the large-sample interval, when n is about 100 it can get a 94 percent probability accuracy.

In case of the bootstrap interval, when n is about 30 is can be 94 percent accurate for the interval.

From the data above, we can see that the probabilities mostly won’t be affected by λ.

I would recommend bootstrap percentile method, since the bootstrap

**(d)**

# Rcode：

1. setwd("C:/Users/lou/Desktop/R/R exercises")
2. #############################quetion 1##############################
3. data\_th=read.csv("bodytemp-heartrate.csv")
4. ##4 qqplots
5. ##temperature qqplots
6. body\_t<-data\_th**$body\_temperature**
7. male\_b<-subset(body\_t,data\_th**$gender**==1)
8. female\_b<-subset(body\_t,data\_th**$gender**==2)
9. ##male temperature
10. qqnorm(male\_b)
11. qqline(male\_b)
12. ##female temperature
13. qqnorm(female\_b)
14. qqline(female\_b)
15. #heart rate qqplots
16. body\_hr<-data\_th**$heart\_rate**
17. male\_hr<-subset(body\_hr,data\_th**$gender**==1)
18. female\_hr<-subset(body\_hr,data\_th**$gender**==2)
19. ##male heart rate
20. qqnorm(male\_hr)
21. qqline(male\_hr)
22. ##female heart rate
23. qqnorm(female\_hr)
24. qqline(female\_hr)
26. #(a)
27. data1=data\_th**$body\_temperature**[which(data\_th**$gender**==1)]
28. data2=data\_th**$body\_temperature**[which(data\_th**$gender**==2)]
30. boxplot(data1,data2)
32. tresult1=t.test(data1,data2,alternative="two.sided",mu=0)
34. #p-value=0.02394<0.05 reject null hypothesis
36. #(b)
37. data1=data\_th**$heart\_rate**[which(data\_th**$gender**==1)]
38. data2=data\_th**$heart\_rate**[which(data\_th**$gender**==2)]
40. boxplot(data1,data2)
42. tresult2=t.test(data1,data2,alternative="two.sided",mu=0)
44. #p-value=0.5287>0.05 can't reject null hypothesis
46. #(c)
47. #########(I)##############
48. #scatter plot xlab is temperature, ylab is heart rate.
49. plot(data\_th**$body\_temperature**,data\_th**$heart\_rate**,
50. xlab = "temperature",
51. ylab = "heart rate",
52. main = "scatter plot of togegher gender12 ")
53. abline(lm(data\_th**$heart\_rate** ~ data\_th**$body\_temperature**))
54. #summary correlation
55. corresult=cor.test(data\_th**$body\_temperature**,data\_th**$heart\_rate**,alternative="two.sided")
56. #cor=0.253654,p-value=0.003591 reject null hypothesis,rho>0,correlation exist.
58. #########(II)#############
59. #scatter plot xlab is temperature, ylab is heart rate.
60. plot(data\_th**$body\_temperature**[which(data\_th**$gender**==1)],data\_th**$heart\_rate**[which(data\_th**$gender**==1)],
61. xlab = "temperature",
62. ylab = "heart rate",
63. main = "scatter plot of male ")
64. abline(lm(data\_th**$heart\_rate**[which(data\_th**$gender**==1)] ~ data\_th**$body\_temperature**[which(data\_th**$gender**==1)]))
65. #summary correlation
66. corresult1=cor.test(data\_th**$body\_temperature**[which(data\_th**$gender**==1)],data\_th**$heart\_rate**[which(data\_th**$gender**==1)],alternative="two.sided")
67. #cor=0.1953894,p-value= can't reject null hypothesis, accept rho=0.
69. ###########(III)###########
70. #scatter plot xlab is temperature, ylab is heart rate.
71. plot(data\_th**$body\_temperature**[which(data\_th**$gender**==2)],data\_th**$heart\_rate**[which(data\_th**$gender**==2)],
72. xlab = "temperature",
73. ylab = "heart rate",
74. main = "scatter plot of female ")
75. abline(lm(data\_th**$heart\_rate**[which(data\_th**$gender**==2)] ~ data\_th**$body\_temperature**[which(data\_th**$gender**==2)]))
76. #summary correlation
77. corresult2=cor.test(data\_th**$body\_temperature**[which(data\_th**$gender**==2)],data\_th**$heart\_rate**[which(data\_th**$gender**==2)],alternative="two.sided")
78. #cor=0.2869312,p-value=0.02048 reject null hypothesis,rho>0.

81. ############################question 2######################################
83. #bootstrap function
84. #function of using bootstrap to compute CI
85. boot.conf.interval = function(x,n){
86. lamda.star = 1/mean(x)
87. mean.boot.dist = replicate(1000,mean(rexp(n,lamda.star)))
88. ci = sort(mean.boot.dist)[c(25,975)]
89. **return**(ci)
90. }
92. #(a)
93. #pareameter
94. n=30
95. lamda=0.1
97. #simulation
98. numprocess=5000
99. #use varible sum to count how many 1/lamda fall in the intv.
100. c1\_sum=0
101. c2\_sum=0
103. **for** (i **in** 1:numprocess){
105. x=rexp(n,lamda)
107. #confidence interval 1
108. t1=t.test(x,alternative="two.sided")
109. c1\_low=1/t1**$conf**.int[2]
110. c1\_high=1/t1**$conf**.int[1]
112. **if** ((lamda<c1\_high)&(lamda>c1\_low)){
113. c1\_sum=c1\_sum+1
114. }**else**{
115. c1\_sum=c1\_sum
116. }
118. #confidence interval 2
119. c2 = boot.conf.interval(x,30)
120. **if** ((1/lamda<c2[2])&(1/lamda>c2[1])){
121. c2\_sum=c2\_sum+1
122. }**else**{
123. c2\_sum=c2\_sum
124. }
126. print(i)
128. }
130. Procover1=c1\_sum/numprocess
131. Procover2=c2\_sum/numprocess

134. #(b)
135. numprocess=5000
136. #use two list to store n and lamda.
137. n\_list=c(5,10,30,100)
138. lamda\_list=c(0.01,0.1,1,10)
139. ##use two matrix to store result of probabilities
140. Procovermatrix1=matrix(0,nrow=length(n\_list),ncol=length(lamda\_list))
141. Procovermatrix2=matrix(0,nrow=length(n\_list),ncol=length(lamda\_list))
143. **for** (k1 **in** 1:length(n\_list)){
145. **for** (k2 **in** 1:length(lamda\_list)){
147. n=n\_list[k1]
148. lamda=lamda\_list[k2]
150. c1\_sum=0
151. c2\_sum=0
152. # 5000 times process..
153. **for** (i **in** 1:numprocess){
155. data=rexp(n,lamda)
157. #confidence interval 1
158. t1=t.test(data,alternative="two.sided")
159. c1\_low=1/t1**$conf**.int[2]
160. c1\_high=1/t1**$conf**.int[1]
162. **if** ((lamda<c1\_high)&(lamda>c1\_low)){
163. c1\_sum=c1\_sum+1
164. }**else**{
165. c1\_sum=c1\_sum
166. }
168. #confidence interval 2
169. c2 = boot.conf.interval(data,n)
170. **if** ((1/lamda<c2[2])&(1/lamda>c2[1])){
171. c2\_sum=c2\_sum+1
172. }**else**{
173. c2\_sum=c2\_sum
174. }
175. print(i)
176. }
177. #confidence interval 1(large sample)
178. Procovermatrix1[k1,k2]=c1\_sum/numprocess
179. #confidence interval 2(boot)
180. Procovermatrix2[k1,k2]=c2\_sum/numprocess
182. }
183. }