

Statistics for Computer Science

Assignment 2

Kanitha Chim

501453

Field of Study Software System and Service Management

Faculty of Informatics
Masaryk University

May 24, 2020

Exercise 3

1. Write down the formula for likelihood function of Poisson distribution.

The formula for the Poisson probability mass function is:

$$P(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

In Poisson distribution, the parameter of interest is λ . Having sequence of X_n , the probability of observing the sequence X_n will be the product of probabilities of each of them.

Therefore, the kernel of likelihood function of Poisson distribution is:

$$L(\lambda|X) = \prod_{i=1}^N \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

2. Write down the formula for log-likelihood function of Poisson distribution.

The formula for log-likelihood function of Poisson distribution is obtained by using natural logarithm on the likelihood function of Poisson distribution.

Therefore, the kernel of log-likelihood function of Poisson distribution is:

$$l(\lambda|X) = \ln \left(\prod_{i=1}^N \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \right)$$

$$l(\lambda|X) = \sum_{i=1}^N X_i \ln \lambda - N\lambda$$

3. Write down the likelihood equation and work out the exact formula for $\hat{\lambda}$.

$$L(\lambda) = \prod_{i=1}^N \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = e^{-N\lambda} \frac{\lambda^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!}$$

$$\ln L(\lambda) = -N\lambda + \sum_{i=1}^N x_i \ln(\lambda) - \ln \left(\prod_{i=1}^N x_i! \right)$$

$$\frac{d \ln L(\lambda)}{d\lambda} = -N + \sum_{i=1}^N x_i \frac{1}{\lambda}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^N x_i}{N}$$

4. Create your own R-function for calculating the value of log-likelihood function of Poisson distribution.

```

1 #set up sequence of x according to exercise description
2 x <- c(117, 109, 109, 89, 120, 88, 99, 103, 109, 91, 107, 101, 109, 117,
        96, 95, 129, 96, 105, 98)
3
4 #n number of observation
```

```

5 n <- 20
6
7 #the function will take 2 parameters lambda and x
8 #n is number of observations
9 #lambda is mean
10 #x is the sequence of observed values
11 #finally it will return the value of log-likelihood of Poisson
    distribution
12 poi.log.likelihood <- function(lambda, x){
13   n <- length(x)
14   log.like <- sum(x) * log(lambda) - n * lambda
15   return(-log.like)
16 }
17
18 #store value of poi.log.likelihood in variable ans.poi.log.like
19 ans.poi.log.like <- poi.log.likelihood(mean(x), x)

```

The value of log-likelihood function of Poisson distribution is -7612.856.

5. Using function optimize() find $\hat{\lambda}$. Compare it to the estimate you get from the exact formula.

```

20 #this function will take 2 parameters x and n
21 #x is the sequence of observed values
22 #n is number of observation
23 #finally it will return lambda hat which is the mean
24 lambda.hat <- function(x, n){
25   return(sum(x)/n)
26 }
27
28 #store value of estimate lambda in ans.lambda.hat
29 ans.lambda.hat <- lambda.hat(x, n)
30
31 #using optimize function to obtain lambda hat
32 #the optimize function take the log-likelihood of Poisson distribution
    to optimize with the given interval
33 #optimize function will return maximum and objective value
34 #in this case we interest in the value of maximum
35 lambda.hat.est <- optimize(f = poi.log.likelihood, interval = c(88, 129)
    , maximum = T, x = x)$maximum

```

The exact value of $\hat{\lambda}$ is 104.35 and the estimate value of $\hat{\lambda}$ is 128.9999.

Using function optimize the value of $\hat{\lambda}$ is greater than using the exact function.

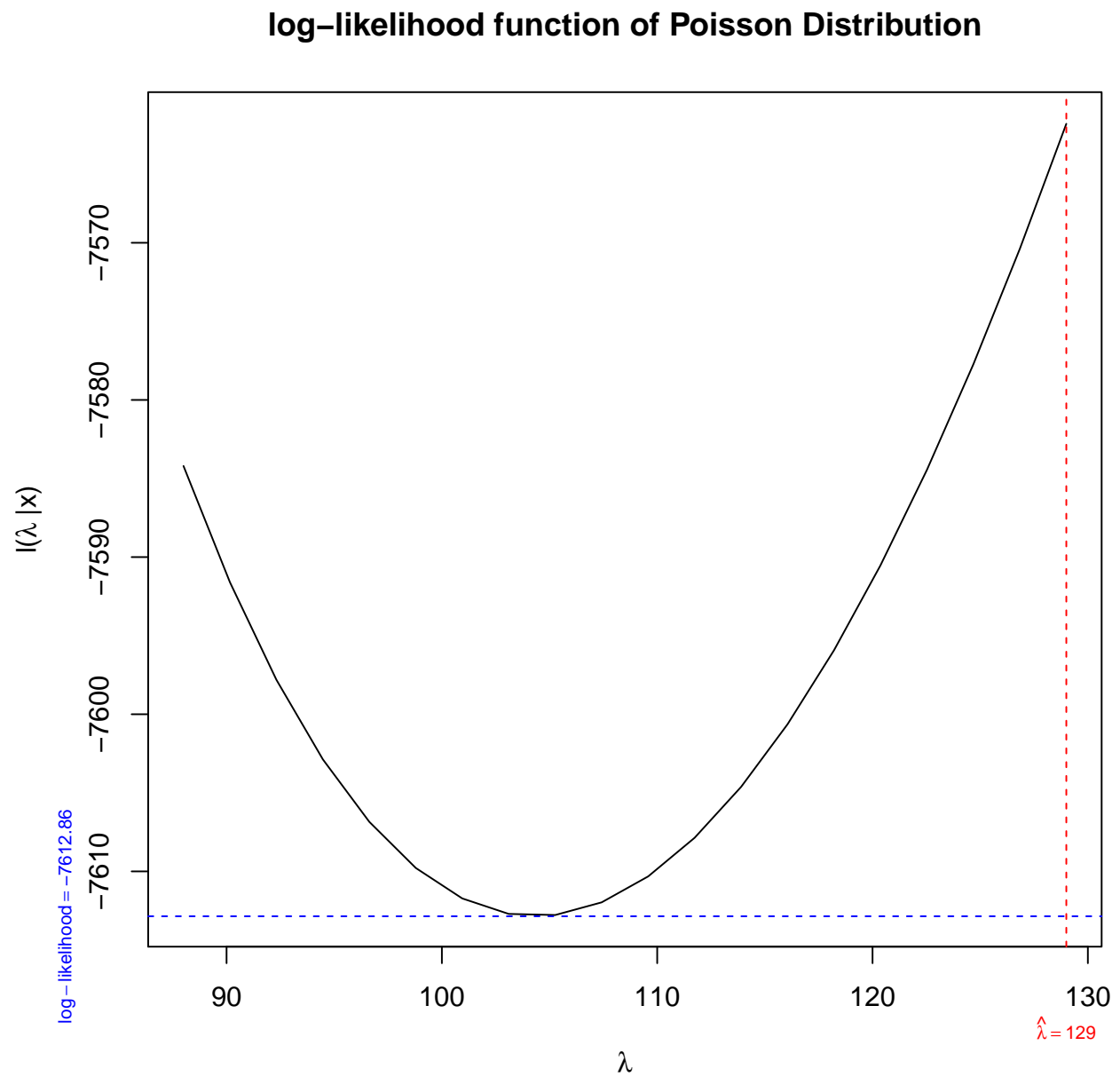
6. Plot the log-likelihood function, highlight the maximum and denote the maximum likelihood estimate in plot margin.

```

36 #set sequence of lambda for x-axis
37 lambda.seq <- seq(from=min(x), to=max(x), length=20)
38
39 #prepare value for y-axis

```

```
40 l.lambda <- apply(X = as.matrix(lambda.seq), MARGIN = 1, FUN = poi.log.
    likelihood, x=x)
41
42 #generate plot for log-likelihood function of poisson distribution
43 plot(lambda.seq, l.lambda, type = 'l', main="log-likelihood function of
    Poisson Distribution", xlab = bquote(lambda), ylab =bquote(paste('l('
    , lambda, ' |x)'))))
44
45 #set value of lambda estimation on vertical line
46 abline(v = lambda.hat.est, col = 'red', lty =2)
47
48 #set value of log-likelihood on horizontal line
49 abline(h = ans.poi.log.like, col = 'blue', lty = 2)
50
51 #use mtext to set symbol and value of lambda estimation on x-axis
52 mtext(bquote(hat(lambda) == .(round(lambda.hat.est, 2))), side = 1, line
    = 2,
53       at = lambda.hat.est, cex = 0.7, col = 'red')
54
55 #use mtext to set symbod and value of log-likelihood on y-axis
56 mtext(bquote(log-likelihood == .(round(ans.poi.log.like, 2))), side = 2,
    line = 2,
57       at = ans.poi.log.like, cex = 0.7, col = 'blue')
```



Exercise 4

1. Write down the null and the alternative hypotheses in mathematical form.

Null hypothesis— $H_0 : \rho = \rho_0$

Alternative hypothesis— $H_1 : \rho \neq \rho_0$,

where $\rho_0 = 0$

2. Calculate the value of test statistic.

```

58 #set working directory
59 setwd(getwd())
60
61 #read data from file txt
62 body <- read.table(file = 'body-measurements.txt', header = T)
63
64 #get only body height of female
65 body.f.height <- body[body$sex == 'f', 'body.H']
66
67 #get only neck of female
68 body.f.neck <- body[body$sex == 'f', 'neck.C']
69
70 #na.omit will clean data that have NA value
71 body.f.height <- na.omit(body.f.height)
72 body.f.neck <- na.omit(body.f.neck)
73
74 #create dataframe for body height and neck
75 df <- data.frame(body.f.height, body.f.neck)
76
77 #get length of the data
78 n.x <- length(body.f.height)
79 n.y <- length(body.f.neck)
80 n <- n.x <- n.y
81
82 #value of rho from null hypothesis
83 rho0 <- 0
84
85 #significant level
86 alpha <- 0.05
87
88 #calculate the estimate of rho using function cor()
89 rho.est <- cor(body.f.height, body.f.neck, method = c("pearson", "
    kendall", "spearman"))
90
91 #calculate ZR using Fisher Z -variable
92 ZR <- 1/2 * log((1+ rho.est)/(1- rho.est))
93
94 #calculate the value of xi
95 xi <- 1/2 * log((1+ rho0)/(1-rho0))
96
97 #calculate the value of test statistics
98 zW <- sqrt(n -3)*(ZR - xi)
99 # round(zW, 5) => 1.39395

```

Result: $zW = 1.39395$

3. Calculate the critical region and make your decision.

```

100 #calculate critical value using qnorm for lower bound
101 z.CR.l <-qnorm(alpha/2, lower.tail = T)
102 # round(z.CR.l, 5) => -1.95996
103
104 #calculate critical value using qnorm for upper bound
105 z.CR.u <-qnorm(alpha/2, lower.tail = F)
106 # round(z.CR.u, 5) => 1.95996

```

Result: $W = (-\infty, -1.95996) \cup (1.95996, \infty)$

The test statistic $z_W = 1.39395$ does not belong to the critical region. Therefore we do not reject H_0 .

4. Calculate the p-value and make your decision.

```

107 #calculate the p-value using test statistic value and pnorm()
108 p.val.zW <- 2 * (1 - pnorm(abs(zW)))
109 # round(p.val.zW, 5) => 0.16333

```

Result: $p\text{-value} = 0.16333$

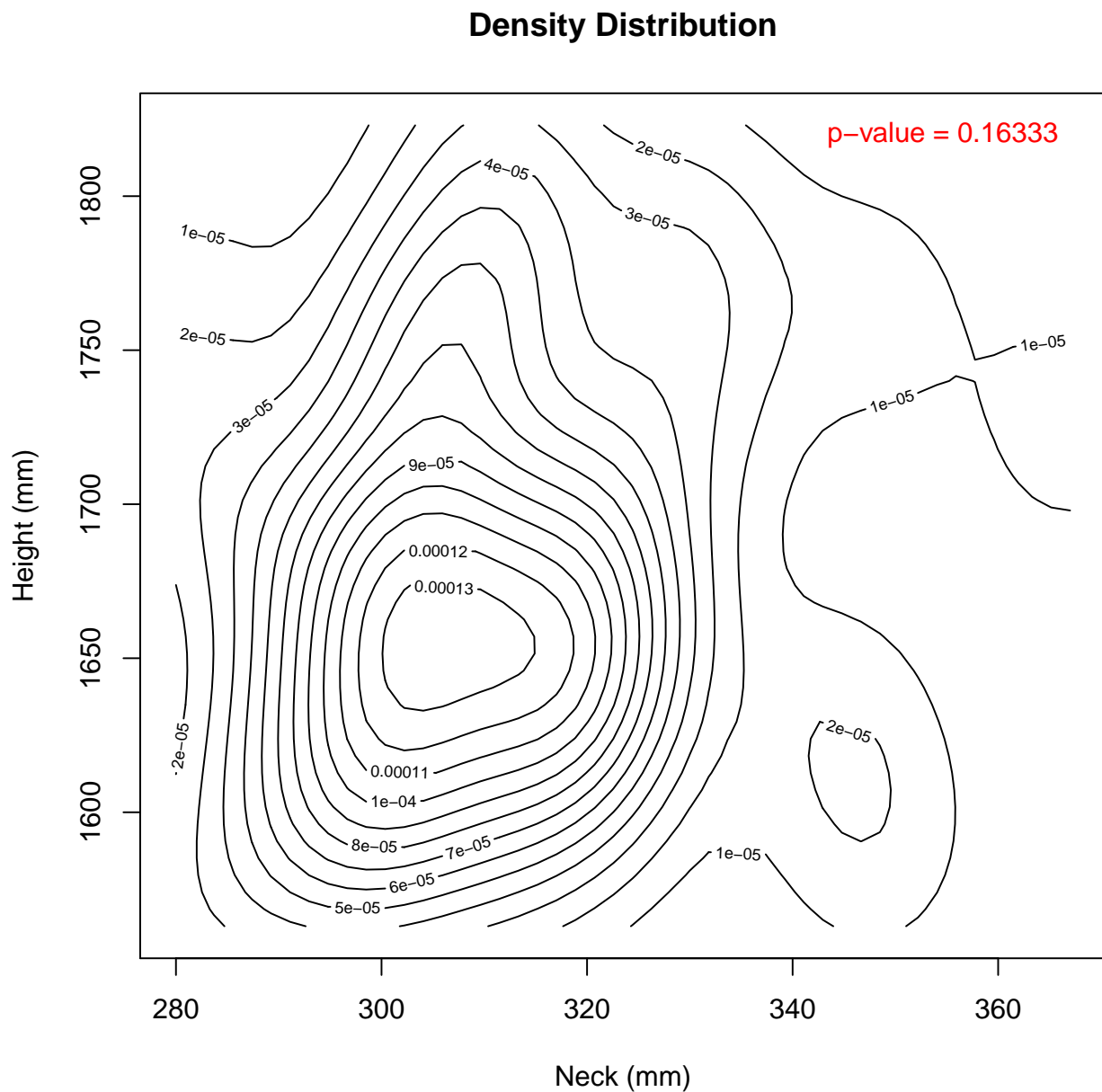
The p-value is 0.16333 which is greater than $\alpha = 0.05$. Therefore we do not reject H_0 .

5. Plot the density of the distribution, that the test statistic follows, and visualise the p-value.

```

110 #create data frame for body neck and height
111 df <- data.frame(body.f.neck, body.f.height)
112
113 #load library called MASS for function kde2d
114 library(MASS)
115
116 #call function kde2d to generate density of body neck and height
117 body_density <- kde2d(df$body.f.neck, df$body.f.height, n=n)
118
119 #plot bivariate distribution
120 plot1 <- contour(body_density, xlab='Neck (mm)', ylab= 'Height (mm)',
121                 main='Density Distribution')
122
123 #add text showing p-value on top right of the plot
124 text(x = 355, y = 1820, labels = "p-value = 0.16333 ", xpd = NA, col = '
red')

```



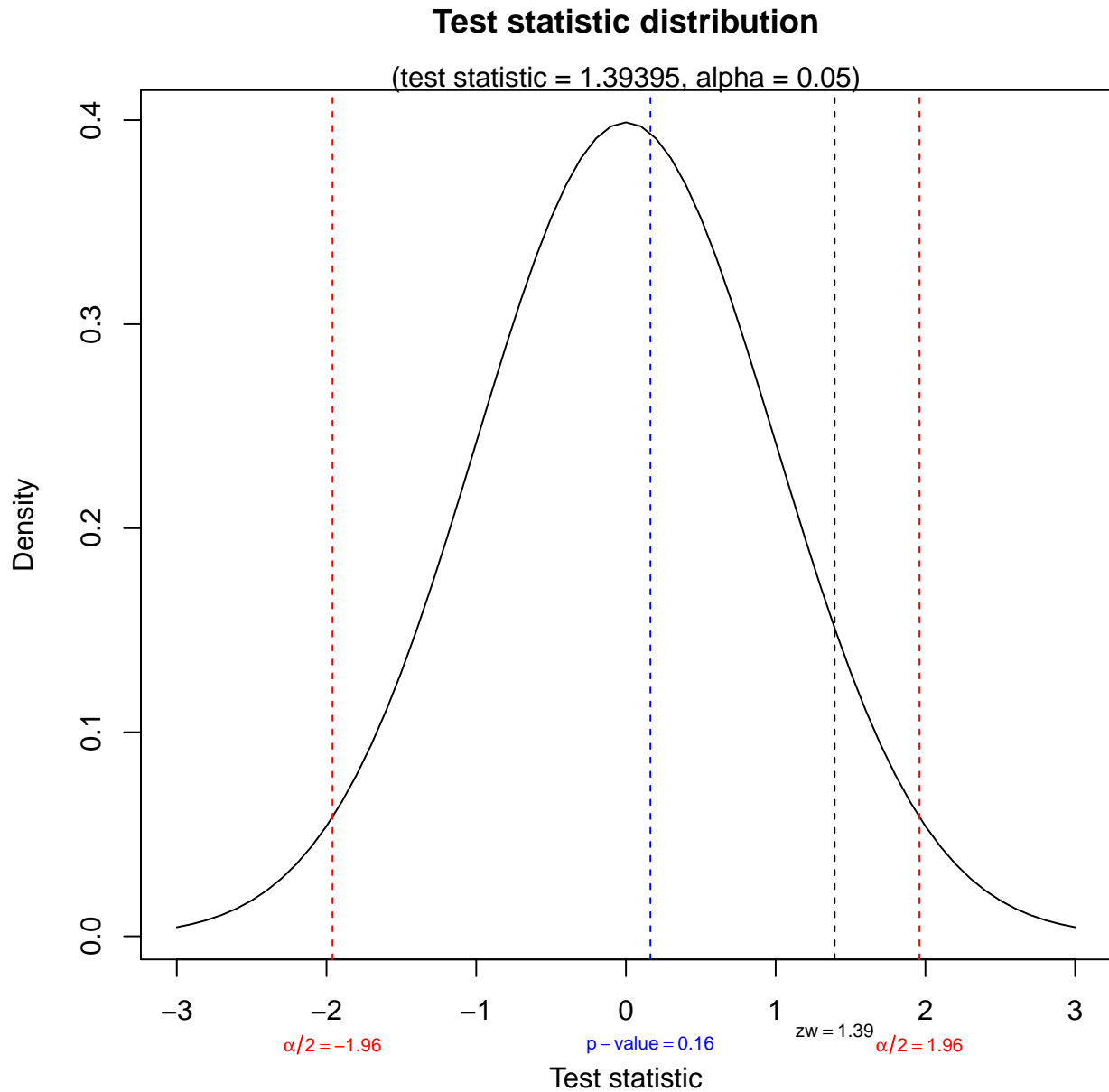
```

124 #set sequence of test statistic value
125 zw.seq <- seq(-3,3, by = .1)
126
127 #using dnorm for density distribution
128 dvalues <- dnorm(zw.seq)
129
130 #plot the test statistic
131 plot2 <- plot(zw.seq, dvalues, type = 'l', xlab='Test statistic', ylab =
      'Density', main = 'Test statistic distribution')
132
133 #add text under main title
134 text(x = 0, y = 0.42, labels = '(test statistic = 1.39395, alpha = 0.05)')

```



```
      ', col = 'black', xpd= NA)
135
136 #add vertical line for test statistic value and text
137 abline(v = zW, col = 'black', lty =2)
138 mtext(bquote(zw == .(round(zW, 2))), side = 1, line = 1.5,
139       at = zW, cex = 0.7, col = 'black')
140
141 #add vertical line for lower bound critical region and text
142 abline(v = z.CR.l, col = 'red', lty =2)
143 mtext(bquote(alpha/2 == .(round(z.CR.l, 2))), side = 1, line = 2,
144       at = z.CR.l, cex = 0.7, col = 'red')
145
146 #add vertical line for upper bound critical region and text
147 abline(v = z.CR.u, col = 'red', lty =2)
148 mtext(bquote(alpha/2 == .(round(z.CR.u, 2))), side = 1, line = 2,
149       at = z.CR.u, cex = 0.7, col = 'red')
150
151 #add vertical line for p-value and text
152 abline(v = p.val.zW, col = 'blue', lty =2)
153 mtext(bquote(p-value == .(round(p.val.zW, 2))), side = 1, line = 2,
154       at = p.val.zW, cex = 0.7, col = 'blue')
```



6. Calculate the confidence interval for ρ and make your decision.

```

155 #calculate the confidence interval Uaplha/2 = 1.95996 for lower bound
156 CI.zW.l <- tanh(ZR - 1.95996/sqrt(n-3))
157 # round(CI.zW.l, 5) => -0.08418
158
159 #calculate the confidence interval Uaplha/2 = 1.95996 for upper bound
160 CI.zW.u <- tanh(ZR + 1.95996/sqrt(n-3))
161 # round(CI.zW.u, 5) => 0.46209

```

Result: $CI : (-0.08418, 0.46209)$

$\rho_0 = 0$ belongs to the value of $CI : (-0.08418, 0.46209)$. Therefore we do not reject H_0 .

7. Interpret your conclusion.

Statistical conclusion:

H_0 is not rejected on a significant level $\alpha = 0.05$, because (1) test statistics does not belong to critical region, (2) $\rho_0 = 0$ belongs to the confidence interval, and (3) p-value is greater than 0.05.

Verbal conclusion:

We are not rejecting null hypothesis that is stated there is no correlation between body height and neck circumference of females.

Exercise 5

1. Write down the null and the alternative hypotheses in mathematical form.

Null hypothesis— $H_0 : p_1 - p_2 = p_0$

Alternative hypothesis— $H_1 : p_1 - p_2 \neq p_0$,

where $p_0 = 0$

2. Calculate the value of test statistic.

```

162 #first group data n1: number of product from group 1, x1: number of
    faulty product from group 1
163 n1 <- 200
164 x1 <- 32
165
166 #calculate the estimate probability of group 1
167 p1.est <- x1 / n1
168
169 #second group data n2: number of product from group 2, x2: number of
    faulty product from group 2
170 n2 <- 230
171 x2 <- 21
172
173 #calculate the estimate probability of group 2
174 p2.est <- x2 / n2
175
176 #calculate the estimate standard deviation
177 sd.est <- sqrt((p1.est*(1-p1.est))/n1 + (p2.est*(1-p2.est))/n2)
178
179 #value of p from null hypothesis

```

```

180 p0 <- 0
181
182 #significant level
183 alpha <- 0.05
184
185 #calculate the value of test statistics using Wald Test
186 zW.obs <- (p1.est - p2.est - p0)/sd.est
187 # round(zW.obs, 5) => 2.13765

```

Result: $zw = 2.13765$

3. Calculate the critical region and make your decision.

```

188 #calculate the critical region for lower bound
189 z.CR.l <- qnorm(alpha/2, lower.tail = T)
190 # round(z.CR.l, 5) => -1.95996
191
192 #calculate the critical region for upper bound
193 z.CR.u <- qnorm(alpha/2, lower.tail = F)
194 # round(z.CR.u, 5) => 1.95996

```

Result: $W = (-\infty, -1.95996) \cup (1.95996, \infty)$

The test statistic $zw = 2.13765$ belongs to the upper bound critical region. Therefore we reject H_0 .

4. Calculate the p-value and make your decision.

```

195 #calculate the p-value using value of test statistic and pnorm()
196 p.val.zW <- 2 * (1 - pnorm(abs(zW.obs)))
197 # round(p.val.zW, 5) => 0.03255

```

Result: $p\text{-value} = 0.03255$

The p-value is 0.03255 which is smaller than $\alpha = 0.05$. Therefore we reject H_0 .

5. Calculate the confidence interval and make your decision.

```

198 #calculate the confidence interval for lower bound Ualpha/2 = 1.95996
199 CI.zW.l <- p1.est - p2.est - (1.95996 * sd.est)
200 # round(CI.zW.l, 5) => 0.00571
201
202 #calculate the confidence interval for upper bound Ualpha/2 = 1.95996
203 CI.zW.u <- p1.est - p2.est + (1.95996 * sd.est)
204 # round(CI.zW.u, 5) => 0.13168

```

Result: $CI : (0.00571, 0.13168)$

$p_0 = 0$ does not belong to the value of $CI : (0.00571, 0.13168)$. Therefore we reject H_0 .

6. Interpret your conclusion.

Statistical conclusion:

H_0 is rejected on a significant level $\alpha = 0.05$, because (1) test statistics belongs to critical region, (2) $p_0 = 0$ does not belong to the confidence interval, and (3) p-value is less than 0.05.

Verbal conclusion:

We are rejecting null hypothesis that the probability of getting a faulty product are the same in both factories.