qbs120_ps5_gibran

Gibran Erlangga

10/14/2021

Question 1

```
# BiocManager::install("multtest")
library(multtest)
```

```
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
data(golub)
gene1.values = golub[1,]
gene1.values[1:5]
```

```
## [1] -1.45769 -1.39420 -1.42779 -1.40715 -1.42668
```

(a) What is the MLE of (^mle)?

```
mle_mean <- mean(gene1.values)
mle_mean
```

```
## [1] -1.129013
```

(b) Is your estimate from part a) unbiased? Justify.
Yes, as this is the output of MLE.

(c) Is your estimate from part a) consistent? Justify.
Yes, it is consistent because if follows the Method of Moments formula.

(d) If you cannot assume the data are normally distributed, is your estimate from a) for E[X] = still valid?
Yes!

(e) What is the MLE of 2 (^2)?

```
mle_var <- sum((gene1.values-mle_mean)^2 / length(gene1.values))
mle_var
```

```
## [1] 0.3364397
```

(f) Is your estimate from part e) unbiased? Justify.
Yes, because if follows the Method of Moments formula.

(g) Is your estimate from part e) consistent? Justify.
No, because it needs to follow the distribution of its parent.

(h) If you cannot assume the data are normally distributed, is your estimate from e) for V ar(X) = 2 still valid?
No, because the variance will be different for every distribution.

(i) What is the distribution of ^mle? If you cannot assume the data are normally distributed, does this sampling distribution still hold?
It is a normal distribution, and yes, it still holds.

(j) What is the MSE of the ^mle estimate?

```
mse <- sum(((gene1.values-mle_mean) / length(gene1.values))^2)
mse
```

```
## [1] 0.008853675
```

(k) Use the distribution of ^mle to compute a 95% CI for . Assume that the 2 = ^2.

```r
n <- length(gene1.values)
abc <- qt(0.05/2, df=n-1)
s <- sd(gene1.values)

confidence_interval <- c(mle_mean - ((s*abc)/sqrt(n)), mle_mean + ((s*abc)/sqrt(n)))
confidence_interval
```

```
## [1] -0.9358015 -1.3222249
```

Question 2

(Based on Rice 8.3) One of the earliest applications of the Poisson distribution was made by Student (1907) in studying errors made in counting yeast cells or blood corpuscles with a haemacytometer. In this study, yeast cells were killed and mixed with water and gelatin; the mixture was then spread on a glass and allowed to cool. Four different concentrations were used. Counts were made on 400 squares, and the data are summarized in the data.frame below. In this data.frame, each of the "concen.*" columns records the number of squares associated with that concentration for which the number of counted cells equals the value in the "cells" column.

```r
yeast.counts = data.frame(cells=0:12,
                          concen.1 = c(213,128,37,18,3,1,0,0,0,0,0,0,0),
                          concen.2 = c(103,143,98,42,8,4,2,0,0,0,0,0,0),
                          concen.3 = c(75,103,121,54,30,13,2,1,0,1,0,0,0),
                          concen.4 = c(0,20,43,53,86,70,54,37,18,10,5,2,2))
```

a) Compute the MLE estimate of the parameter λ for each of the four sets of data.

The MLE for λ in Poisson RV X is defined by the sample mean:

$$\hat{\lambda} = \bar{X}$$

Here's the implementation of lambda calculation in R:

```r
n = 400
cells = 0:12

get_lambda <- function(cells, concen, n) {
  return(sum(cells*concen)/n)
}

lambda_concen <- c()

lambda_concen[1] <- get_lambda(cells, yeast.counts$concen.1, n)
lambda_concen[2] <- get_lambda(cells, yeast.counts$concen.2, n)
lambda_concen[3] <- get_lambda(cells, yeast.counts$concen.3, n)
lambda_concen[4] <- get_lambda(cells, yeast.counts$concen.4, n)

cat(paste("lambda value for concentration 1 is", lambda_concen[1], "\n",
          "lambda value for concentration 2 is", lambda_concen[2], "\n",
          "lambda value for concentration 3 is", lambda_concen[3], "\n",
          "lambda value for concentration 4 is", lambda_concen[4], "\n"
            ))
```

```
## lambda value for concentration 1 is 0.6825
##  lambda value for concentration 2 is 1.3225
##  lambda value for concentration 3 is 1.8
##  lambda value for concentration 4 is 4.68
```

b) Approximate the theoretical standard error of the $\hat{\lambda}$ values computed for Problem 1 part a). Do not use simulation.
Formula is $\frac{\sqrt{\hat{\lambda}}}{400}$. Therefore,

```
std_error <- c()
std_error[1] <- sqrt(lambda_concen[1] / 400)
std_error[2] <- sqrt(lambda_concen[2] / 400)
std_error[3] <- sqrt(lambda_concen[3] / 400)
std_error[4] <- sqrt(lambda_concen[4] / 400)

cat(paste("standard error for concentration 1 is", std_error[1], "\n",
          "standard error for concentration 2 is", std_error[2], "\n",
          "standard error for concentration 3 is", std_error[3], "\n",
          "standard error for concentration 4 is", std_error[4], "\n"
            ))
```

```
## standard error for concentration 1 is 0.0413067791046458
##  standard error for concentration 2 is 0.0575
##  standard error for concentration 3 is 0.0670820393249937
##  standard error for concentration 4 is 0.10816653826392
```

c) For the $\hat{\lambda}$ values compute for Problem 1 part a), estimate the standard error using the parametric bootstrap. How do these values compare to the approximate theoretical values? Do these results match your expectations?

```
set.seed(20)
n_sample = 10000
N = 400

for (i in 1:n_sample) {
  sim_1 <- rpois(n_sample, lambda_concen[1])
  sim_2 <- rpois(n_sample, lambda_concen[2])
  sim_3 <- rpois(n_sample, lambda_concen[3])
  sim_4 <- rpois(n_sample, lambda_concen[4])
}

mean_sim_1 = c()
mean_sim_2 = c()
mean_sim_3 = c()
mean_sim_4 = c()

for (i in 1:N) {
  s_1 <- mean(sample(sim_1, N))
  mean_sim_1 <- append(mean_sim_1, s_1)
  s_2 <- mean(sample(sim_2, N))
  mean_sim_2 <- append(mean_sim_2, s_2)
  s_3 <- mean(sample(sim_3, N))
```

```
  mean_sim_3 <- append(mean_sim_3, s_3)
  s_4 <- mean(sample(sim_4, N))
  mean_sim_4 <- append(mean_sim_4, s_4)
}

cat(paste("estimated standard error for concentration 1 is", sd(mean_sim_1), "\n",
          "estimated standard error for concentration 2 is", sd(mean_sim_2), "\n",
          "estimated standard error for concentration 3 is", sd(mean_sim_3), "\n",
          "estimated standard error for concentration 4 is", sd(mean_sim_4), "\n"
            ))
```

```
## estimated standard error for concentration 1 is 0.0394596962736985
##  estimated standard error for concentration 2 is 0.0564280265510954
##  estimated standard error for concentration 3 is 0.0650488345567619
##  estimated standard error for concentration 4 is 0.10269273630093
```

d) Find an approximate 95% confidence interval for each estimate.

Formula for getting an approximation of confidence interval for λ of Poisson distribution is denoted as:

$$\left[\hat{\lambda} - z\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z\sqrt{\frac{\hat{\lambda}}{n}}\right]$$

with:
- $\hat{\lambda}$ is sample mean, - n = 400, - z is equal to $\alpha/2$ (area under the density curve of a standard normal distribution)

We want 95% confidence interval, so we know that $\alpha = 0.05$.

We know from the reference table that for 95% CI, the value for z is 1.96.

Plugging in all the numbers into the above equation, we get:

```
get_ci_upper <- function(lambda, z, n) {
  return(lambda + z*sqrt(lambda/n))
}

get_ci_lower <- function(lambda, z, n) {
  return(lambda - z*sqrt(lambda/n))
}

concen_ci_upper <- c()
concen_ci_lower <- c()

# concen 1
concen_ci_lower[1] <- round(get_ci_lower(lambda_concen[1], 1.96, n), digit=3)
concen_ci_upper[1] <- round(get_ci_upper(lambda_concen[1], 1.96, n), digit=3)

# concen 2
concen_ci_lower[2] <- round(get_ci_lower(lambda_concen[2], 1.96, n), digit=3)
concen_ci_upper[2] <- round(get_ci_upper(lambda_concen[2], 1.96, n), digit=3)

# concen 3
```

```r
concen_ci_lower[3] <- round(get_ci_lower(lambda_concen[3], 1.96, n), digit=3)
concen_ci_upper[3] <- round(get_ci_upper(lambda_concen[3], 1.96, n), digit=3)

# concen 4
concen_ci_lower[4] <- round(get_ci_lower(lambda_concen[4], 1.96, n), digit=3)
concen_ci_upper[4] <- round(get_ci_upper(lambda_concen[4], 1.96, n), digit=3)


cat(paste("95% confidence interval for lambda value in concentration 1 are",
          concen_ci_lower[1], "and", concen_ci_upper[1], "\n",
          "95% confidence interval for lambda value in concentration 2 are",
          concen_ci_lower[2], "and", concen_ci_upper[2], "\n",
          "95% confidence interval for lambda value in concentration 3 are",
          concen_ci_lower[3], "and", concen_ci_upper[3], "\n",
          "95% confidence interval for lambda value in concentration 4 are",
          concen_ci_lower[4], "and", concen_ci_upper[4], "\n"
          ))
```

```
## 95% confidence interval for lambda value in concentration 1 are 0.602 and 0.763
##  95% confidence interval for lambda value in concentration 2 are 1.21 and 1.435
##  95% confidence interval for lambda value in concentration 3 are 1.669 and 1.931
##  95% confidence interval for lambda value in concentration 4 are 4.468 and 4.892
```

e) Compare observed and expected counts. The PDF of Poisson RV X with $\lambda > 0$ is defined as:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Observation would be:

```r
get_expected_count <- function(lambda, k, n) {
  return(round((lambda^k / factorial(k)) * exp(-lambda) * n, digit=2))
}

get_exp_1 <- c()
get_exp_2 <- c()
get_exp_3 <- c()
get_exp_4 <- c()

for (i in cells) {
  get_exp_1 <- append(get_exp_1, get_expected_count(lambda_concen[1], i, n))
  get_exp_2 <- append(get_exp_2, get_expected_count(lambda_concen[2], i, n))
  get_exp_3 <- append(get_exp_3, get_expected_count(lambda_concen[3], i, n))
  get_exp_4 <- append(get_exp_4, get_expected_count(lambda_concen[4], i, n))
}

concen_1_observed <- yeast.counts$concen.1
concen_1_expected <- get_exp_1
concen_2_observed <- yeast.counts$concen.2
concen_2_expected <- get_exp_2
concen_3_observed <- yeast.counts$concen.3
concen_3_expected <- get_exp_3
```

```r
concen_4_observed <- yeast.counts$concen.4
concen_4_expected <- get_exp_4

data.frame(concen_1_observed,
           concen_1_expected,
           concen_2_observed,
           concen_2_expected,
           concen_3_observed,
           concen_3_expected,
           concen_4_observed,
           concen_4_expected)
```

```
##    concen_1_observed concen_1_expected concen_2_observed concen_2_expected
## 1                213            202.14               103            106.59
## 2                128            137.96               143            140.96
## 3                 37             47.08                98             93.21
## 4                 18             10.71                42             41.09
## 5                  3              1.83                 8             13.59
## 6                  1              0.25                 4              3.59
## 7                  0              0.03                 2              0.79
## 8                  0              0.00                 0              0.15
## 9                  0              0.00                 0              0.02
## 10                 0              0.00                 0              0.00
## 11                 0              0.00                 0              0.00
## 12                 0              0.00                 0              0.00
## 13                 0              0.00                 0              0.00
##    concen_3_observed concen_3_expected concen_4_observed concen_4_expected
## 1                 75             66.12                 0              3.71
## 2                103            119.02                20             17.37
## 3                121            107.11                43             40.65
## 4                 54             64.27                53             63.41
## 5                 30             28.92                86             74.19
## 6                 13             10.41                70             69.44
## 7                  2              3.12                54             54.16
## 8                  1              0.80                37             36.21
## 9                  0              0.18                18             21.18
## 10                 1              0.04                10             11.02
## 11                 0              0.01                 5              5.16
## 12                 0              0.00                 2              2.19
## 13                 0              0.00                 2              0.86
```

Question 3

(Based on Rice 8.9) How would you respond to the following argument? This talk of sampling distributions is ridiculous! Consider Example A of Section 8.4. The experimenter found the mean of the number of fibers to be 24.9. How can this be a "random variable" with an associated "probability distribution" when it's just a number? The author of this book is guilty of deliberate mystification!

In this particular case, the sample we use is 23 fibers. The sample mean of it is a random variable, as this is calculated from the sample of fibers we collected from the factory. If we were to repeat the sampling process again, there is a very small chance that the value for sample mean to be 24.9, as this is drawn from a random variable. We can never guarantee that the sample mean value to be 24.9 before we start sampling the data.

Question 4

(Based on Rice 8.13) In Example D of Section 8.4, the MOM estimate was found to be $\hat{\alpha} = 3\bar{X}$. In this problem, you will consider the sampling distribution of $\hat{\alpha}$.

(a) Show that $E[\hat{\alpha}] = \alpha$, i.e., the estimate is unbiased. By the linearity of expectation and the fact that $E(X_i) = \frac{\alpha}{3}$ for i {1, 2, 3, ..., n}, we have:

$$\begin{aligned} E(\hat{\alpha}) &= E(3 \cdot \bar{X}) \\ &= 3 \cdot E(\bar{X}) \\ &= \frac{3}{n} \sum_{i=1}^{n} E(X_i) \\ &= \frac{3}{n} \sum_{i=1}^{n} \frac{\alpha}{3} \\ &= \alpha \end{aligned}$$

Therefore, the estimate $\hat{\alpha}$ is unbiased for α.

(b) Show that $Var(\hat{\alpha}) = (3 - \alpha^2)/n$. Hint: What is $Var(\bar{X})$?

We know that:

$$Var(X_1) = E(X_1^2) - E(X_1)^2$$

with density function of X_1 as:

$$f(x) = \frac{1 + \alpha x}{2}, x \in [-1, 1]$$

for $x \in [-1, 1]$:

$$\begin{aligned} E(X_1^2) &= \int_{-1}^{1} x^2 \cdot f(x)\ dx \\ &= \int_{-1}^{1} x^2 \cdot \frac{1 + \alpha x}{2}\ dx \\ &= \frac{1}{2} \int_{-1}^{1} x^2\ dx + \frac{\alpha}{2} \int_{-1}^{1} x^3\ dx \\ &= \frac{1}{2}(\frac{x^3}{3})|_{-1}^{1} + \frac{1}{2}(\frac{x^4}{4})|_{-1}^{1} \\ &= \frac{1}{3} \end{aligned}$$

Therefore, $Var(X_1$ is:

$$\begin{aligned} Var(X_1) &= \frac{1}{3} - (\frac{\alpha}{3})^2 \\ &= \frac{3 - \alpha^2}{9} \end{aligned}$$

The fact that X_i are independent, then:

$$\begin{aligned} Var(\hat{\alpha}) &= Var(3 \cdot \bar{X}) = 9 \cdot Var(\bar{X}) = \frac{9}{n^2} \sum_{i=1}^{n} Var(X_i) \\ &= \frac{9}{n^2} \sum_{i=1}^{n} \frac{3-\alpha^2}{9} \\ &= \frac{9}{n^2} \cdot n \cdot \frac{3-\alpha^2}{9} \\ &= \frac{3-\alpha^2}{n} \end{aligned}$$

(c) Use the CLT to deduce that a normal approximation to the sampling distribution of $\hat{\alpha}$. According to this approximation, if n = 20 and $\alpha = 1$, what is the $P(\hat{\alpha} > 0.5)$? Define in terms of $\Phi()$, the CDF for the standard normal. CLT says that for a large value of n,

$$\frac{\hat{\alpha} - E(\hat{\alpha})}{\sqrt{Var(\hat{\alpha})}} \overset{D}{\approx} N(0,1)$$

Plugging the results of mean and variance of $\hat{\alpha}$ from previous questions, and the fact that it is approximately distributed on standard normal RV Z, we have:

$$\hat{\alpha} \overset{D}{\approx} N(\alpha, \frac{3-\alpha^2}{n})$$

Therefore,

$$\begin{aligned} P(|\hat{\alpha}| > 0.5) &= 1 - P(|\hat{\alpha}| \leq 0.5) \\ &= 1 - P(-0.5 \leq \hat{\alpha} - \alpha \leq 0.5) \\ &= 1 - P(\frac{-0.5}{\sqrt{\frac{3-\alpha^2}{n}}} \leq \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{3-\alpha^2}{n}}} \leq \frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}}) \\ &\approx 1 - [\Phi(\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}}) - \Phi(-\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}})] \\ &= 1 - [\Phi(\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}}) - 1 + \Phi(-\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}})] \\ &= 2 - 2\Phi(\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}}) \end{aligned}$$

Plug $n = 20$ and $\alpha = 1$ to the formula,

$$\begin{aligned} P(|\hat{\alpha}| > 0.5) &= 2 - 2\Phi(\frac{0.5}{\sqrt{\frac{3-\alpha^2}{n}}}) \\ &= 2 - 2\Phi(1.58) \\ &= 2 - 2\Phi(1.58) \\ &= 2 - 2 \cdot 0.9429 \\ &= 0.1142 \end{aligned}$$

Question 5

(Based on Rice 8.58) For a population in Hardy-Weinberg equilibrium, alleles occur with the following frequencies:

$$AA : (1-\theta)^2$$
$$Aa : 2\theta(1-\theta)$$
$$aa : \theta^2 2$$

For a specific sample of 190 people, the haptoglobin types occurred as follows:

$$X1 : Hp1-1 : 10$$
$$X2 : Hp1-2 : 68$$
$$X3 : Hp2-2 : 112$$

Assume the haptoglobin genotype for this population is in Hardy-Weinberg equilibrium.

(a) Find the mle of θ.

We know that n = 190.

Likelihood function:

$$\begin{aligned} likelihood(\theta) &= P(X=1|\theta)^{10} P(X=2|\theta)^{68} P(X=3|\theta)^{112} \\ &= (1-\theta)^{20}(2\theta\ (1-\theta))^{68}(\theta)^{224} \\ &= (1-\theta)^{88} 2^{68} \theta^{292} \end{aligned}$$

apply natural logarithm on both ends, we get:

$$\begin{aligned} l(\theta) &= ln(lik(\theta)) \\ &= 68\ ln(2) + 292\ ln(\theta) + 88\ ln(1-\theta) \end{aligned}$$

Derivative of l = 0 to get the max value:

$$\begin{aligned} l'(\theta) &= 0 \\ \frac{292}{\theta} - \frac{88}{1-\theta} &= 0 \\ 292 - 380\theta &= 0 \\ \theta &= 0.768 \end{aligned}$$

second derivative of l:

$$l''(\theta) = -\frac{292}{\theta^2} - \frac{88}{(1-\theta)^2} < 0$$

which value is always negative, hence l is a concave function.

Therefore, $\tilde{\theta} = 0.7684$.

(b) Find the asymptotic variance of the mle. Estimating variance of $\tilde{\theta}$:

$$Var(\tilde{x}) \approx \frac{1}{n \ . \ I(\theta)}$$

with $I(\theta)$ as:

$$I(\theta) = E([\frac{\partial}{\partial\theta} ln \ f(X|\theta)]^2)$$

Therefore,

$$\begin{aligned} Var(\tilde{\theta}) &= \frac{1}{E([l'(\theta)]^2)} \\ &= \frac{1}{E(l''(\theta))} \\ &= \frac{1}{E(-\frac{292}{\theta^2} - \frac{88}{(1-\theta)^2})} \\ Var(\tilde{\theta}) &\approx \frac{1}{\frac{292}{\theta^2} + \frac{88}{(1-\theta)^2}} \end{aligned}$$

(c) Find an approximate 99% confidence interval for θ.

Based on the asymptotic normality of MLE, CI of θ is:

$$\left[\tilde{\theta} - z \ .s_{\tilde{\theta}}, \tilde{\theta} + z \ .s_{\tilde{\theta}}\right]$$

with $\tilde{\theta}$ as MLE of θ and $s_{\tilde{\theta}}$ as the estimated standard error of $\tilde{\theta}$.

For 99% CI, $\alpha = 0.01$. From the reference table, we know that the value for z is 2.58.

Then, we get the value for estimated variance of $s_{\tilde{\theta}}$ using the result we get in b). Plugging the numbers in, we get:

$$\begin{aligned} Var(\tilde{\theta}) &= \frac{1}{E(-\frac{292}{0.7684^2} - \frac{88}{(1-0.7684)^2})} \\ s^2_{\tilde{\theta}} &= \frac{1}{2135.6} \\ &= 0.000468 \end{aligned}$$

Therefore, the estimated standard error of $\tilde{\theta}$ is:

$$s_{\tilde{\theta}} = \sqrt{0.000468} = 0.0216$$

Lastly, plug the numbers into the CI formula, we get:

$$[0.7684 - 2.58 \cdot 0.0216, 0.7684 + 2.58 \cdot 0.0216] = [0.7127, 0.8241]$$

(d) Use the parametric bootstrap to estimate the sampling distribution of the MLE of θ. Plot this distribution along with the asymptotic distribution. How does the shape of the bootstrap sampling distribution compare to the asymptotic distribution?

(e) Use the bootstrap sampling distribution to estimate the variance of the MLE of θ. How does the bootstrap variance compare with the asymptotic variance?

(f) Compute the 99% CI for θ using the bootstrap percentile approach. How does this compare with the CI computed in part c)?