Probability Theory and Statistics 3, Assignment 5

Jiyeon Kim *, Younjoo Mo †

November 27, 2020

Exercise 1

a.

part 1) Showing that Q is a pivotal quantity

$$X_i \sim EXP(\theta) \tag{1}$$

$$\Rightarrow \text{MGF of } X: \ M_X(t) = \frac{1}{1-\theta t} \tag{2}$$

$$Y_n := X_1 + \ldots + X_n \tag{3}$$

$$\Rightarrow M_Y(t) \overset{\text{indep.}}{=} \left[M_X(t)\right]^n = \left(\frac{1}{1-\theta t}\right)^n \tag{4}$$

$$\Rightarrow Y_n \sim GAM(\theta, n) \tag{5}$$

*Student number: 12470236
†Studentnumber: 12475440

$$Q = \frac{2}{\theta}\sum_{i=1}^{n} X_i = \frac{2}{\theta} Y_n \tag{6}$$

$$\Rightarrow M_Q(t) = E\left[e^{t \cdot \frac{2}{\theta} Y}\right] = M_Y\left(t \cdot \frac{2}{\theta}\right) = \left(\frac{1}{1-2t}\right)^n \tag{7}$$

$$\Rightarrow Q \sim GAM(2, n) \tag{8}$$

The distribution of Q does not depend on n, while Q is a random variable that is a function only of $X_1, \ldots, X_n$ and θ. Therefore, Q is a pivotal quantity.

part 2) Deriving a $100\gamma\%$ lower confidence limit for θ based on Q

$$Q \sim GAM(2, n) = \chi^2(2n) \tag{9}$$

$$\gamma = P\left[Q < \chi^2_\gamma(2n)\right] \tag{10}$$

$$= P\left[\frac{2}{\theta}\sum_{i=1}^{n} X_i < \chi^2_\gamma(2n)\right] \tag{11}$$

$$= P\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_\gamma(2n)} < \theta\right] \tag{12}$$

Therefore, the one-sided $100\gamma\%$ lower confidence limit is as following

$$\ell_n^Q(X_1, \ldots, X_n) = \frac{2\sum_{i=1}^{n} X_i}{\chi^2_\gamma(2n)} \tag{13}$$

b.

$$Y := X_{1:n} \tag{14}$$

$$\Rightarrow \text{(PTS2 Theorem 6.15) } F_Y(y) = 1 - [1 - F_X(y)]^n = 1 - e^{-\frac{ny}{\theta}}\ , \quad (y > 0) \tag{15}$$

$$\Rightarrow f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{n}{\theta} \cdot e^{-\frac{ny}{\theta}}\ , \quad (y > 0) \tag{16}$$

$$\Rightarrow Y \sim EXP\left(\frac{\theta}{n}\right) \tag{17}$$

In order to apply theorem 11.4.2, we get derivative of $F_y(y)$ with respect to θ

$$\frac{d}{d\theta} F_Y(y) = \frac{-ny}{\theta^2} e^{\frac{-ny}{\theta}} \tag{18}$$

Since $\frac{d}{d\theta} F_y(y)$ is a decreasing function of θ and we are deriving lower confidence limit, we use the first equation in theorem 11.4.2

$$\gamma = F_Y(y) = F_{X_{1:n}}(x_{1:n}) = 1 - e^{\frac{-ny}{\theta}} \tag{19}$$

$$\Rightarrow y = x_{1:n} = -\frac{\theta \ln(1-\gamma)}{n} \tag{20}$$

If we solve the equation w.r.t θ so that other variables go on the right side (21)

$$\theta = -\frac{ny}{\ln(1-\gamma)} \quad = -\frac{nx_{1:n}}{\ln(1-\gamma)} \tag{22}$$

Therefore, the one-sided $100\gamma\%$ lower confidence limit for θ for $Y = X_{1:n}$ is as following

$$\ell_n^Y(X_1, \ldots, X_n) \ = \ -\frac{nX_{1:n}}{\ln(1-\gamma)} \quad ,(0 < \gamma < 1) \tag{23}$$

c.

part 1 $\big)$ $E\big[$the confidence limit from (a)$\big]$ for $\gamma = 0.95$

$$E\left[\ell_n^Q\right] = E\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_\gamma(2n)}\right] = \frac{2}{\chi^2_{0.95}(2n)} \cdot n \cdot E[X] = \frac{2n\theta}{\chi^2_{0.95}(2n)} \overset{\text{PTS2 App. B4}}{=} \frac{2n\theta}{\chi^2_{\alpha=0.05}(2n)} \tag{24}$$

i) $n = 1$

$$E\left[\ell_1^Q\right] = \frac{2\theta}{\chi^2_{\alpha=0.05}(2)} \overset{\text{PTS2 App. B4}}{=} \frac{2\theta}{5.991} \approx 0.33\theta \tag{25}$$

ii) $n = 2$

$$E\left[\ell_2^Q\right] = \frac{4\theta}{\chi^2_{\alpha=0.05}(4)} \overset{\text{PTS2 App. B4}}{=} \frac{4\theta}{9.488} \approx 0.42\theta \tag{26}$$

iii) $n = 50$

$$E\left[\ell_{50}^Q\right] = \frac{100\theta}{\chi^2_{\alpha=0.05}(100)} \overset{\text{PTS2 App. B4}}{=} \frac{100\theta}{124.342} \approx 0.80\theta \tag{27}$$

part 2 $\big)$ $E\big[$the confidence limit from (b)$\big]$ for $\gamma = 0.95$

$$E\left[\ell_n^Y\right] = E\left[-\frac{nX_{1:n}}{\ln(1-\gamma)}\right] = -\frac{n}{\ln 0.05} \cdot E\left[X_{1:n}\right] = -\frac{\theta}{\ln 0.05} \tag{28}$$

$$E\left[\ell_1^Y\right] \approx 0.33\theta \tag{29}$$

$$E\left[\ell_2^Y\right] \approx 0.33\theta \tag{30}$$

$$E\left[\ell_{50}^Y\right] \approx 0.33\theta \tag{31}$$

d.

For n=1,the $E\left[\text{the confidence limit from (a)}\right]$ is 0.33θ , 0.42θ when n=2 and reaches 0.80θ when sample size is 50. From this result we can see that$E\left[\text{the confidence limit from (a)}\right]$ gets closer toθ.

However, $E\left[\text{the confidence limit from (b)}\right]$ is free from n, remaining 0.33θ for all possible n.

As sample size increases in the case of $\sum_n^{i=1} x_i$, the expected confidence limit gets closer to θ.It is because $\sum_n^{i=1} x_i$ is sufficient for θ, meaning it doesn't depend on θ(Definition 10.2.1,Theorem 10.2.1).

Therefore we think the confidence limit of Q is better since we can interpret this result as $\sum_n^{i=1} x_i$ containing most information about θ.

The outcome has already been expected. In general, we have more power to estimate the population as we have the larger size of sample. The result in (a) uses the whole sample data, thus it uses the advantage of the large number. On the other hand, the result in (b) uses only the smallest observation to estimate θ, no matter how large n is, or regardless of the rest of the data.

Exercise 2

a.

If $\hat{e}$ is an unbiased estimator for e, then $E[\hat{e}] = e$ holds. So we can try to choose c by using the equation $E[\hat{e}] = e$. As the first step, we should get $E[\hat{e}]$.

$$E[\hat{e}] = E\left[1 - c \cdot \frac{\hat{p}_X}{\hat{p}_Y}\right] = E\left[1 - \frac{cX}{Y}\right] = 1 - cE[X] \cdot E\left[\frac{1}{Y}\right] \tag{32}$$

We need $E\left[\frac{1}{Y}\right]$ for further calculations.

$$f_Y(y) = \binom{n}{y}(p_Y)^y(1-p_Y)^{n-y}\ , \quad (y = 0, 1, 2, \ldots, n) \tag{33}$$

$$\Rightarrow E\left[\frac{1}{Y}\right] = \sum_{y=0}^{n}\left[\frac{1}{y}\cdot\binom{n}{y}(p_Y)^y(1-p_Y)^{n-y}\right] = \text{undefined} \tag{34}$$

$E\left[\frac{1}{Y}\right]$ is undefined because there is 0 in a denominator. Therefore, $E[\hat{e}]$ is undefined as well, and we cannot say $E[\hat{e}] = e$. In conclusion, it is not possible to choose a c such that $\hat{e}$ is unbiased for e, in this case.

b.

In general, we would have solved c by $E[\hat{e}] = E[e]$. But since it is not possible as shown in (a), we use definition 7.5.1 X_n has an asymptotically normal distribution and theorem 7.6.1, proving that $\hat{p_X} \xrightarrow{\mathsf{P}} p_X$. And we can apply theorem 7.7.4.3 when we deal with $\frac{\hat{p_X}}{\hat{p_Y}}$. Therefore, $c = 1$ is better than c becomes the other number.

c.

In order to estimate the efficiency, we substitute $n = 15000,\ \hat{p_X} = \frac{5}{15000},\ \hat{p_Y} = \frac{90}{15000}$ to $\hat{e}$.

$$\hat{e} = 1 - \frac{\hat{p_X}}{\hat{p_Y}} \tag{35}$$

$$= 1 - \frac{5/15000}{90/15000} \tag{36}$$

$$= 1 - \frac{1}{18} \tag{37}$$

$$= \frac{17}{18} \tag{38}$$

d.

Objective

1. An approximate 99% lower confidence limit ℓ, such that $P[\ell < \hat{e}] = 0.99$
2. Getting a numerical value for ℓ

To find the lower confidence limit, we will apply CLT to $\hat{p_X}$. So we will explain about it first, with introducing a new random variabe J.

$$J \sim BIN(1, p_X) \tag{39}$$

$$X = J_1 + \ldots + J_n \sim BIN(n, p_X) \tag{40}$$

$$\hat{p_X} = \frac{X}{n} = \frac{J_1 + \ldots + J_n}{n} \tag{41}$$

J has a finite mean $E[J] = p_X$ and a finite non-zero variance $Var[J] = p_X(1 - p_X)$. Thus by CLT,

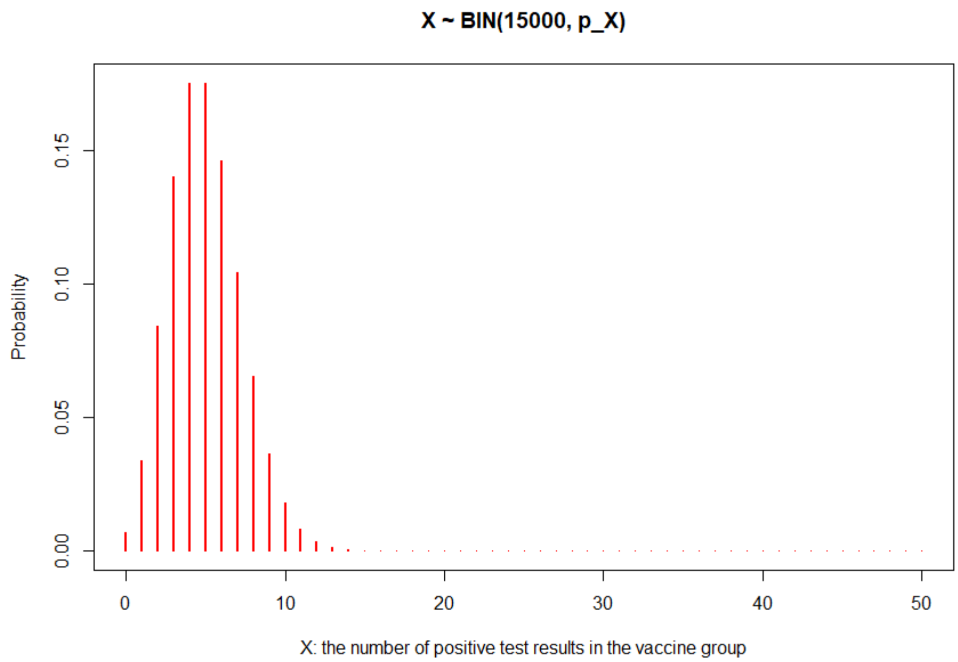
$$\sqrt{n} \cdot \frac{\hat{p_X} - p_X}{\sqrt{p_X(1 - p_X)}} \xrightarrow{\text{d}} Z \sim N(0, 1) \tag{42}$$

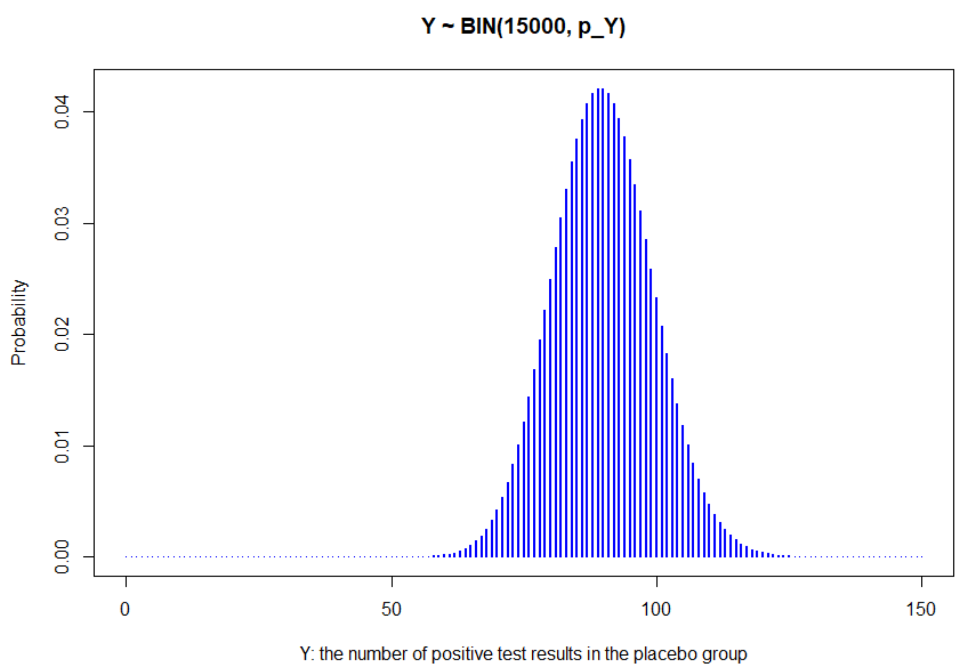
$$
\begin{aligned}
0.99 &= P[\ell < \hat{e}] && (43)\\
&= P\left[\ell < 1 - \frac{\hat{p_X}}{\hat{p_Y}}\right] && (44)\\
&= P\left[1 - \ell > \frac{\hat{p_X}}{\hat{p_Y}}\right] && (45)\\
&\text{Replace } \hat{p_Y} \text{ with } p_Y \quad (\because \hat{p_Y} \xrightarrow{\mathsf{P}} p_Y) && (46)\\
&\approx P\left[p_Y(1-\ell) > \hat{p_X}\right] && (47)\\
&= P\left[\sqrt{n}\cdot\frac{p_Y(1-\ell)-p_X}{\sqrt{p_X(1-p_X)}} > \sqrt{n}\cdot\frac{\hat{p_X}-p_X}{\sqrt{p_X(1-p_X)}}\right] && (48)\\
&\text{By (43),} && (49)\\
&\approx P\left[\sqrt{n}\cdot\frac{p_Y(1-\ell)-p_X}{\sqrt{p_X(1-p_X)}} > Z\right] \overset{\text{App. C3}}{=} P[2.326 > Z] && (50)
\end{aligned}
$$

$$
\begin{aligned}
&\text{From (51),} && (51)\\
&\sqrt{n}\cdot\frac{p_Y(1-\ell)-p_X}{\sqrt{p_X(1-p_X)}} \;=\; 2.326 && (52)\\
&\Rightarrow \ell = 1 - \frac{1}{p_Y} \times \left[\frac{2.326\sqrt{p_X(1-p_X)}}{\sqrt{n}} + p_X\right] && (53)
\end{aligned}
$$

If we substitute $n = 15000,\ p_X = \frac{5}{15000},\ p_Y = \frac{90}{15000}$ to (54), then approximately 0.8866641 is the numeric value for the 99% lower confidence limit ℓ.

e.


X ~ BIN(15000, p_X)
Probability
0.00
0.05
0.10
0.15
0
10
20
30
40
50
X: the number of positive test results in the vaccine group


Y ~ BIN(15000, p_Y)
Probability
0.00
0.01
0.02
0.03
0.04
0
50
100
150
Y: the number of positive test results in the placebo group

i) The first graph is drawn for $X \in [0, 50]$ and the second graph is drawn for $Y \in [0, 150]$, instead of $[0, 15000]$. It is to make the graph to be more visible.

ii) Approximately 0.873 is the 99% lower confidence limit for the efficiency, gotten by the simulation. The R code is following.

```
# The given Moderna data

n = 15000
x = 5
y = 90


# p_X and p_Y which are replaced by its MLE, based on the Moderna data

p_X = x / n
p_Y = y / n


# Graphs of the distribution of X and Y, respectively

X <- dbinom(0:50, size=15000, prob=p_X)
plot(0:50, X, main=c("X ~ BIN(15000, p_X)"),
     xlab="X: the number of positive test results in the vaccine group",
     ylab="Probability",
     type='h', lwd=2, col="red")
Y <- dbinom(0:150, size=15000, prob=p_Y)
plot(0:150, Y, main=c("Y ~ BIN(15000, p_Y)"),
```

```
    xlab="Y:␣the␣number␣of␣positive␣test␣results␣in␣the␣placebo␣group",
    ylab="Probability",
    type='h', lwd=2, col="blue")


# Ramdomly generated p_X hats and p_Y hats

p_X.hat <- rbinom(10^6, size=15000, prob=p_X) / n
p_Y.hat <- rbinom(10^6, size=15000, prob=p_Y) / n


# 10^6 e hats

e.hat = 1 - (p_X.hat / p_Y.hat)


# The 1st percentile of e hats
# = the 99% lower confidence limit for the efficiency

quantile(e.hat, prob=0.01)
```