STAT5703 HW1

Huang Chao, Wancheng Chen, Chengchao Jin

Exercise 1.

Question 1.

To calculate p^{th}, we need to find $Q_D(p)$ such that $P(D \leq Q_D(p)) = p$. Then, the previous function can be transformed to

$$\int_0^{Q_D(p)} \lambda e^{-\lambda D} dD = 1 - e^{-\lambda Q_D(p)} = p$$

So, from this equation, $Q_D(p)$ can be expressed by

$$Q_D(p) = -\frac{1}{\lambda} \ln(1-p)$$

Question 2.

From question(a), we already obtain the equation for $Q_D(p)$ which is $Q_D(p) = -\frac{1}{\lambda}\ln(1-p)$. Then, to find the MLE of $Q_D(p)$, we can find the MLE of λ first and then replace λ with its Maximum Likelihood Estimator $\hat{\lambda}^{MLE}$. $D_1, ..., D_n$ are i.i.d. Exponential random variables with parameter λ, the log-likelihood function is

$$\ell(\lambda; D_1, ..., D_n) = n \ln \lambda - \sum_{i=1}^{n} \lambda D_i$$

The MLE $\hat{\lambda}^{MLE}$ is

$$\hat{\lambda}^{MLE} = \frac{1}{\bar{D_n}}$$

and

$$Q_D(p)^{MLE} = -\frac{1}{\hat{\lambda}^{MLE}} \ln(1-p) = -\bar{D_n} \ln(1-p)$$

Question 3.

$D_1, ..., D_n \overset{i.i.d.}{\sim} Exp(\lambda)$. Then the CLT tells us that

$$\sqrt{n}(\bar{D_n} - \mu) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

Hence, by Delta Method we can get,

$$\sqrt{n}(Q_D(p) + \frac{\ln(1-p)}{\lambda}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \frac{(\ln(1-p))^2}{\lambda^2})$$

Then, for *approximate* $(1-\alpha)$-confidence interval,

$$L(D) = -\bar{D_n} \ln(1-p) - \frac{z_{1-\alpha/2} \times \ln(1-p)}{\lambda\sqrt{n}}$$

$$R(D) = -\bar{D_n}\ln(1-p) + \frac{z_{1-\alpha/2} \times \ln(1-p)}{\lambda\sqrt{n}}$$

So, the *approximate* $(1-\alpha)$-confidence interval for $Q_D(p)$ is $[-\bar{D_n}\ln(1-p) - \frac{z_{1-\alpha/2}\times\ln(1-p)}{\lambda\sqrt{n}}, -\bar{D_n}\ln(1-p) + \frac{z_{1-\alpha/2}\times\ln(1-p)}{\lambda\sqrt{n}}]$

Question 4.

We know that if $D_1, ..., D_n$ are independent exponential random variables with parameter λ, then

$$\lambda\bar{D_n} \sim \Gamma(n, \frac{1}{n})$$

So, $\lambda\bar{D_n}$ is independent of the parameter λ, which means it is an exact pivot. To construct an exact confidence interval of the median, we can first transform $\lambda\bar{D_n}$ to χ^2 distribution. Then,

$$2n\lambda\bar{D_n} \sim \chi^2_{2n}$$

Hence, for any $\alpha \in (0,1)$,

$$P(\chi^2_{1-\alpha/2,2n} < 2n\lambda\bar{D_n} < \chi^2_{\alpha/2,2n}) = 1-\alpha$$

Since $Q_D(0.5) = -\bar{D_n}\ln 0.5$, then

$$P(-\frac{\ln 0.5\chi^2_{\alpha/2,2n}}{2n\lambda} < Q_D(0.5) < -\frac{\ln 0.5\chi^2_{1-\alpha/2,2n}}{2n\lambda}) = 1-\alpha$$

Hence, the $(1-\alpha)$ exact confidence interval of the median is,

$$Q_D(0.5) \in [-\frac{\ln 0.5\chi^2_{\alpha/2,2n}}{2n\lambda}, -\frac{\ln 0.5\chi^2_{1-\alpha/2,2n}}{2n\lambda}]$$

Exercise 3.

Question 1.

As $R_1 - \mu$ has a zero mean distribution, all moments with odd orders are zero. Therefore, we have,

$$\begin{aligned}\gamma = \mathbf{E}[R_1^3] &= E[(R_1-\mu+\mu)^3] \\ &= \mathbf{E}[(R_1-\mu)^3 + 3(R_1-\mu)^2\mu + 3(R_1-\mu)\mu^2 + \mu^3] \\ &= 3\mu\mathbf{E}[(R_1-\mu)^2] + \mu^3 \\ &= 3\mu Var[R_1-\mu] + \mu^3 \\ &= \mu^3 + 3\mu\sigma^2\end{aligned}$$

Question 2.

(a) Since $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ has the distribution of $\mathbf{N}(\mu, \sigma^2/n)$, similarly to Q1, we can derive $\mathbf{E}[\bar{R}^3] = \mu^3 + 3\mu\frac{\sigma^2}{n}$. So the bias is $\mathbf{E}[\hat{\gamma} - \gamma] = -\frac{n-1}{n}\mu\sigma^3$.

(b) $\hat{\gamma}$ is not consistent. Since $\bar{R} \sim N(\mu, \frac{\sigma^2}{n})$, we have,

$$\begin{aligned}\Pr[|\bar{R}^3 - (\mu^3 + 3\mu\frac{\sigma^2}{n})| \geq \epsilon] &= 1 - \Pr[|\bar{R}^3 - (\mu^3 + 3\mu\frac{\sigma^2}{n})| \leq \epsilon] \\ &= 1 - \Phi(\sqrt{n}\frac{(\mu^3 + 3\mu\frac{\sigma^2}{n} + \epsilon)^{\frac{1}{3}} - \mu}{\sigma^2}) \\ &\quad + \Phi(\sqrt{n}\frac{(\mu^3 + 3\mu\frac{\sigma^2}{n} - \epsilon)^{\frac{1}{3}} - \mu}{\sigma^2}) \\ &\to 1 - \Phi(\sqrt{n}\frac{(\mu^3 + \epsilon)^{\frac{1}{3}} - \mu}{\sigma^2}) + \Phi(\sqrt{n}\frac{(\mu^3 - \epsilon)^{\frac{1}{3}} - \mu}{\sigma^2}) \\ &\to 1 - \Phi(\infty) + \Phi(-\infty) \\ &= 1 - 1 + 0 = 0, \textit{as } n \to \infty \textit{ with fixed } \epsilon\end{aligned}$$

So $\hat{\gamma}$ converges to $\mu^3 + 3\mu\frac{\sigma^2}{n} \to \mu^3$, so it is not consistent to the estimated parameter $\gamma = \mu^3 + 3\mu\sigma^3$.

Question 3.

Since we have $\mathbf{E}[R_1R_2R_3] = \mu^3$ and $\mathbf{E}[\hat{\gamma}] = \mu^3 + \frac{3\mu\sigma^2}{n}$, we have $3\mu\sigma^2/n = \mathbf{E}[\hat{\gamma}] - \mathbf{E}[R_1R_2R_3]$.Therefore, we can choose $n\hat{\gamma} - (n-1)R_1R_2R_3$ as the unbias estimator, whose mean is exactly μ^3.

Question 4.

(a) Since $\mathbf{E}[\tilde{\gamma} - \gamma] = n * \frac{1}{n}\mathbf{E}[R_1^3] - \gamma = 0$, the bias is 0.
(b) $\tilde{\gamma}$ is consistent. Using LLT, $\tilde{\gamma} \overset{p}{\sim} \mathbf{E}[R_1^3] = \gamma$. So it's consistent.

Question 5.

Since the minimal sufficient statistics for normal distributions are $\bar{R} = \frac{1}{n}\sum\limits_{i=1}^{n} R_i$ and $\overline{R^2} = \frac{1}{n}\sum\limits_{i=1}^{n} R_i^2$. And they are also complete statistics. According to the Rao-Blackwell, we only need to find the conditional expection of an unbiased estimator by setting the two statistics as the condition. Therefore $\gamma_{UVME} = \mathbf{E}[\tilde{\gamma}|\bar{R}, \overline{R^2}]$. In the following, we use T to denote the condition. We have,

$$\begin{aligned}\mathbf{E}[\tilde{\gamma}|T] &= \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} R_i^3|T] \\ &= \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R} + \bar{R})^3|T] \\ &= \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}[(R_i - \bar{R})^3 + 3(R_i - \bar{R})^2\bar{R} \\ &\quad + 3(R_i - \bar{R})\bar{R}^2 + \bar{R}^3]|T]\end{aligned}$$

By using symmmetry of the conditional distribution, one can prove that all (conditional) moments of $R_i - R$

which have odd orders are zero. Therefore, we have,

$$\begin{aligned}
\mathbf{E}[\tilde{\gamma}|T] &= \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}[3(R_i-\bar{R})^2\bar{R}+\bar{R}^3]|T] \\
&= \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}[3R_i^2\bar{R}-6R_i\bar{R}^2+3\bar{R}^3+\bar{R}^3]|T] \\
&= \mathbf{E}[\frac{3}{n}\bar{R}\sum_{i=1}^{n}R_i^2-2\bar{R}^3|T] \\
&= 3\bar{R}\overline{R^2}-2(\bar{R})^3
\end{aligned}$$

Exercise 4.

Load dataset

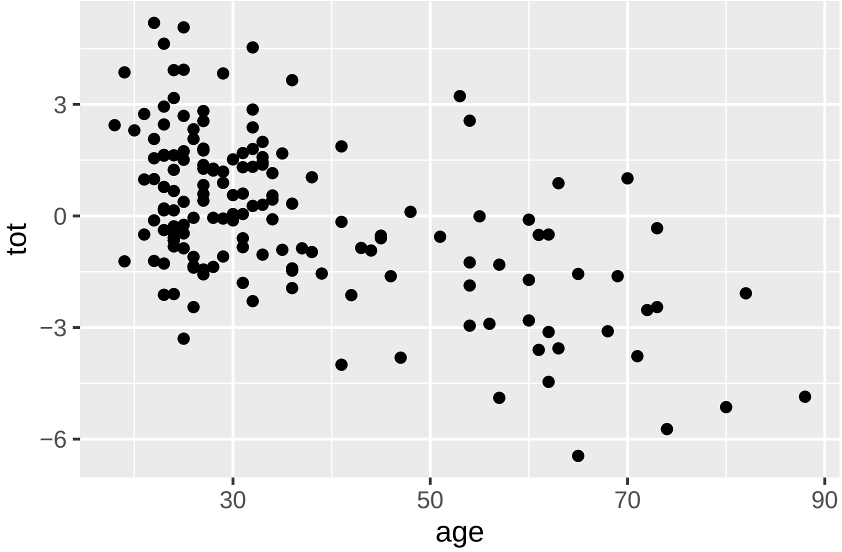
```
lines <- readLines("kidney.txt")
```

```
numbers_vec <- lapply(lines,
  function (line)  stringr::str_extract_all(line, "[-]?\\d+[\\.]?\\d*")) %>%
  unlist(recursive = FALSE) %>%
  Filter(f = function(x) length(x) == 3) %>%
  Map(f = function(x) lapply(x, as.numeric))
```

```
df <- do.call(rbind.data.frame, numbers_vec)
colnames(df) <- c("id", "age", "tot")
rownames(df) <- df$id
```

Question 1.

```
library(ggplot2)
scatterPlot <- ggplot(df, mapping = aes(x=age, y=tot)) + geom_point()
scatterPlot
```



The scatter plot shows that "age" and "tot" have a negative relationship and it could be fitted with a linear model.

Question 2.

I would like to use tot as the response variable because it is more intuitively reasonable to say that the tot function is affected by age.

Question 3.

```
Corr = cor(df$age, df$tot)
Corr
```

```
## [1] -0.5718387
```

Negative sign, since the correlation is negative. Without any calculation, I also expect the intercept to be positive and the slope to be negative. First, as age variable increases, the overall function of kidney, tot, tends to be lower and therefore the slope should be negative. Then, around age=20, the values of tot scatter around tot=0. So clearly, since the slope is negative, the value of tot should be larger than 0 at age=0. Intuitively, the overall function of kidney for a baby should be positive. So the intercept should be positive.

Question 4.

α denotes the expected value when the input is 0, while β denotes how much the response will change if the input is increased or decreased by 1.

Question 5.

```
linearModel <- lm(tot ~ age, data = df)
summary(linearModel)
```

```
##
## Call:
## lm(formula = tot ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2018 -1.3451  0.0765  1.0719  4.5252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.860027   0.359565   7.954 3.53e-13 ***
## age         -0.078588   0.009056  -8.678 5.18e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 155 degrees of freedom
## Multiple R-squared:  0.327,  Adjusted R-squared:  0.3227
## F-statistic: 75.31 on 1 and 155 DF,  p-value: 5.182e-15
```

From the model, we can see that α=2.860027 and β=-0.078588. So, when age=0, the tot is estimated to be 2.860027. Also, for each year increase of age, the tot is estimated to decrease by 0.078588. Both parameters have p-value much smaller than 0.05 so they are both statistically significant.

Question 6.

I would use the geometry to intepret these two parameters. In linear algebra, the estimates provide the optimal group of parameters to project a high-dimension vector to a two-dimension plane with minimal loss.

Question 7.

```
#prediction <- function(age) linearModel$coefficients * age + linearModel$
beta <- as.numeric(linearModel$coefficients["age"])
alpha <- as.numeric(linearModel$coefficients["(Intercept)"])
predict <- function (age) alpha + beta * age
```

```
predict(100)
```

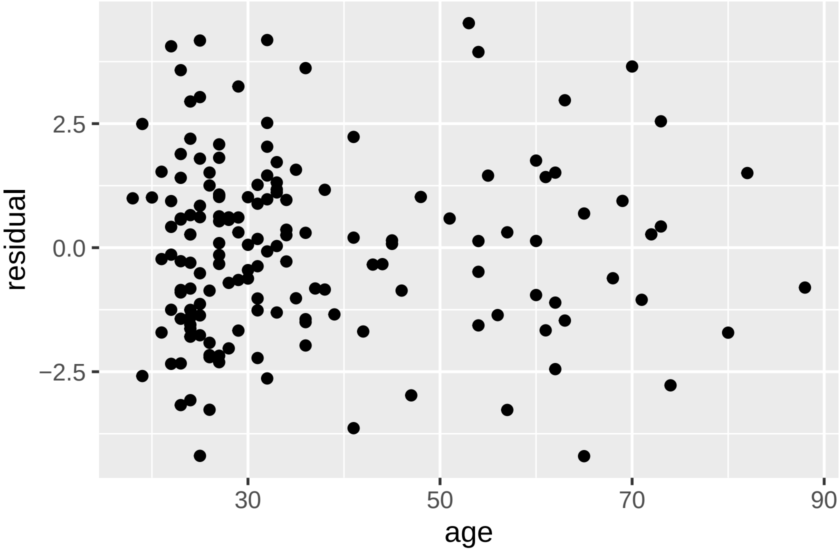
```
## [1] -4.998815
```

The prediction seems reasonable.

Question 8.

```
res_df <- df %>%
  dplyr::mutate(prediction = predict(df$age)) %>%
  dplyr::mutate(residual = tot - prediction)
```

```
ggplot(res_df) + geom_point(aes(x=age, y=residual))
```



The plot shows that the residuals are randomly distributed around 0, so it seems reasonable to assume that errors ϵ_i are i.i.d..

Question 9.

```
minus <- function(x, y) max(y,x) - min(y,x)
betaIntNormal <- Reduce(minus, confint(linearModel)[2,])
betaIntAsym <- Reduce(minus, confint.default(linearModel)[2, ])
```

The asymptotic one seems to have a shorter interval.

Question 10.

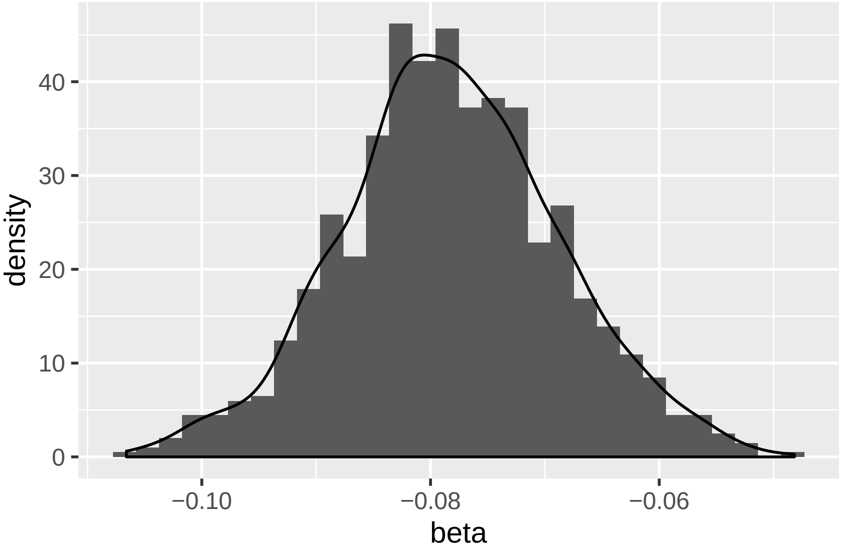
```
boot.stat <- function(data, indices){
data <- data[indices, ] # select cases in bootstrap sample
mod <- lm(tot ~ age, data=data) # refit model
coef(mod)["age"] # return coefficient vector
}
```

```
set.seed(12345) # for reproducibility
df.boot <- boot::boot(data=df, statistic=boot.stat, R=1000)
```

```
bootResult <- as.data.frame(df.boot$t) %>%
  dplyr::rename(beta=V1)
```

```
ggplot(bootResult) +
  geom_histogram(aes(x=beta,y=..density..)) +
  geom_density(aes(x=beta, y=..density..))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
confInts <- boot::boot.ci(df.boot)
```

```
## Warning in boot::boot.ci(df.boot): bootstrap variances needed for
## studentized intervals
```

```
basicInt <- Reduce(minus, confInts$basic[4:5])
percentInt <- Reduce(minus, confInts$percent[4:5])
```

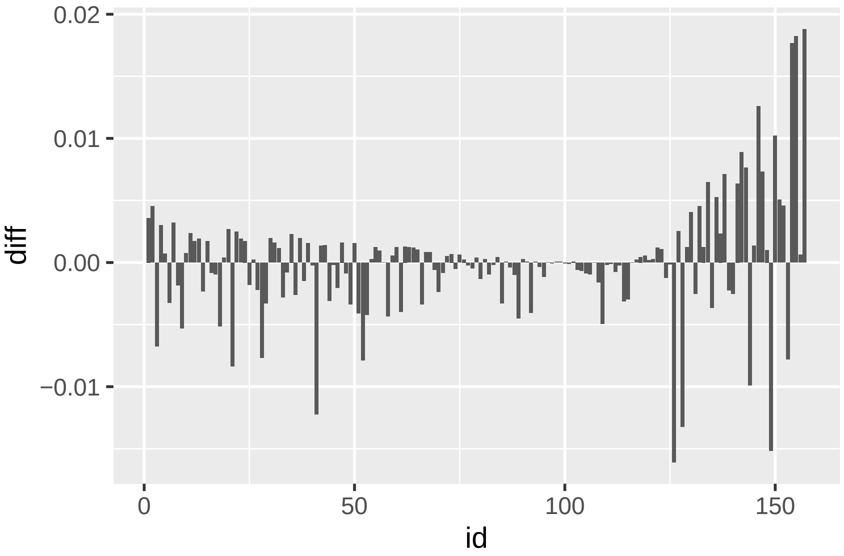
The bootstrap interval is larger.

Question 11.

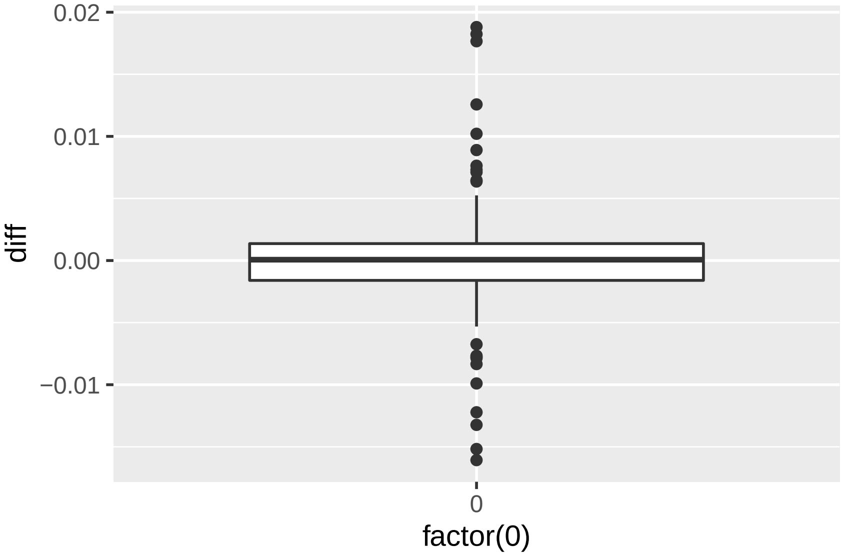
```
LeaveOneOutCorr <- function (idx)  {
  df_tmp <- df %>% dplyr::filter(id != idx)
  cor(df_tmp$age, df_tmp$tot) - Corr
}
```

```
corr_diff <- unlist(Map(LeaveOneOutCorr, seq.int(1, nrow(df))))
df_lou <- df %>% dplyr::mutate(diff=corr_diff)
```

```
ggplot(df_lou) + geom_col(aes(x=id, y=diff))
```

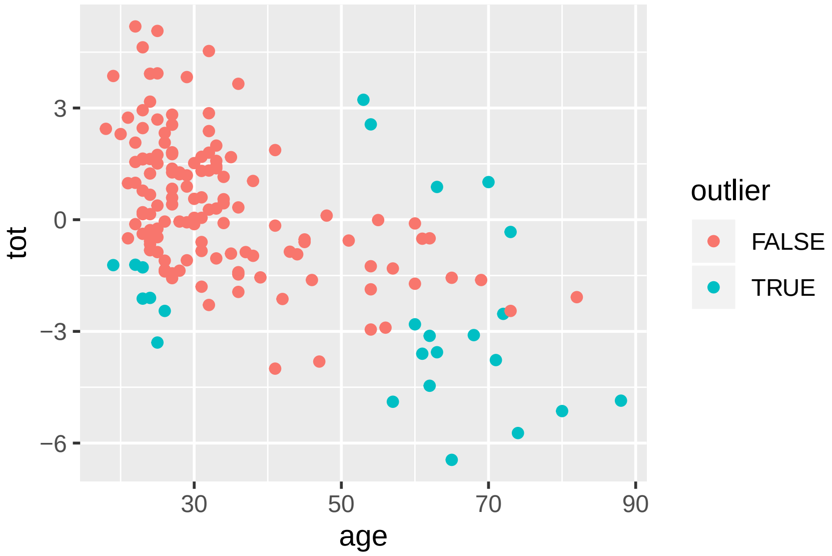


```
ggplot(df_lou, aes(x=factor(0), y=diff)) + geom_boxplot()
```



```
outlierDetect <- function(corrVal) {
  ifelse (abs(corrVal) > 0.005, TRUE, FALSE)
}
df_outlier <- df_lou %>% dplyr::mutate(outlier=outlierDetect(diff))
```

```
ggplot(df_outlier) + geom_point(aes(x=age, y=tot, color=outlier))
```



There are some data points which are more influential than others. In the above plot, they are marked as "outlier"s with special leave-one-out differences.

Bonus Question.

Question 1.

The author want to answer the question whether techno-scientific findings are inevitable or not by fitting the findings dataset using a Poisson model. The optimal parameter chosen after experiments tends to show that techno-scientific discoveries are not inevitable and highly depends on luck. I think for me, the choice of Poisson distribution seems reasonable, since techno-scientific findings are odd and can happen at a low probability. And Poisson distribution is quite suitable for modeling the probability of rare events happening.

Question 2.

Since there are no data for singleton and no-findings in the dataset, so using a truncated model will not give weird expected values for $k = 0$ or $k = 1$.

Question 3.

Suppose $X \sim Poisson(\mu)$, then we can derive the expectation and variance of Y using $\mathbf{E}[X]$ and $Var[X]$. We have,

$$\begin{aligned}\mathbf{E}[X] &= \mu \\ &= \sum_{k=0}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} \\ &= \mu e^{-\mu} + \sum_{k=2}^{\infty} k \frac{e^{-\mu}\mu^k}{k!}\end{aligned}$$

If we denote $C = \frac{1}{1-e^{-\mu}-\mu e^{-\mu}}$, we have,

$$\begin{aligned}\mathbf{E}[Y] &= C \sum_{k=2}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} \\ &= C(\mu - \mu e^{-\mu})\end{aligned}$$

Similarly, we can derive $Var[Y]$ with the help of $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$. We have,

$$Var[Y] = C\mu(2\mu - e^{-\mu}) - C^2\mu^2(1 - e^{-\mu})^2$$

#### Question 4. The data can be saved as a dataframe as below,

```
tbl1 <- data.frame(
  k = seq.int(2, 9),
  count = c(179, 51, 17, 6, 8, 1, 0, 2)
)
tbl1
```

```
##   k count
## 1 2   179
## 2 3    51
## 3 4    17
## 4 5     6
## 5 6     8
## 6 7     1
## 7 8     0
## 8 9     2
```

Then the log likelihood function can be derived as,

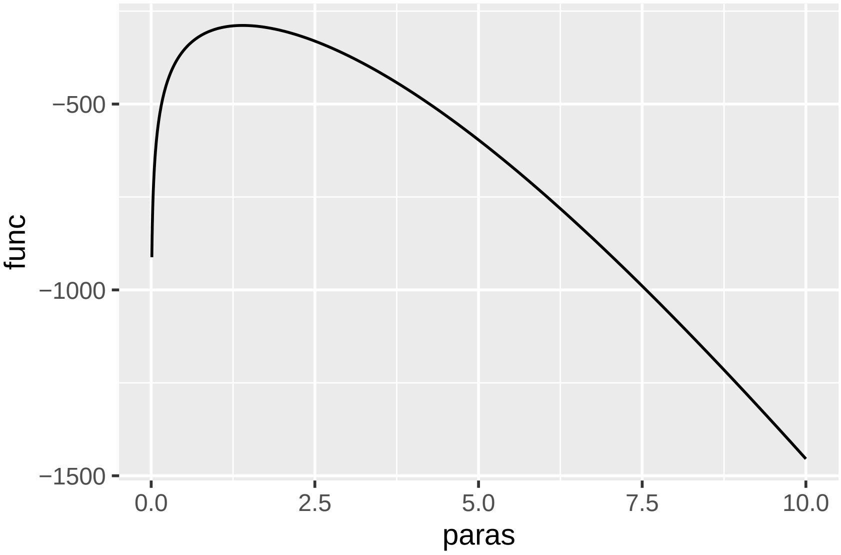
$$\begin{aligned}\log L &= \log(\prod_{k=2}^{9}(C\frac{e^{-\mu}\mu^k}{k!})^{COUNT_k}) \\ &= \sum_{k=2}^{9} COUNT_k \log(\frac{e^{-\mu}\mu^k}{k!})\end{aligned}$$

```
logL <- function (mu) {
  prob_dist <- function(x) {
    exp(-mu) * mu^x / factorial(x) /
      (1 - exp(-mu) - mu * exp(-mu))
  }
  data_ <- tbl1 %>%
    dplyr::mutate(prob=prob_dist(k)) %>% # likelihood
    dplyr::mutate(likelihood=count*log(prob)) # log-likelihood
  sum(data_$likelihood) # sum them up
}
```

And the log likelihood can be plotted as below,

```
plot_curve <- function(pars, f) {
 df_curve <- data.frame(
  paras = pars,
  func = unlist(lapply(pars, f))
 )
 ggplot(df_curve, aes(x=paras, y=func)) + geom_line()
}

plot_curve(10^seq.int(-2, 1, 0.01), logL)
```



Question 5.

The algorithm I choose is “BFGS”, implemented in “optimx:optimx” function. Since it’s a convex and nonlinear optimization problem as plotted above, this algorithm will converge shortly. The results and code are shown below,

```
opt <- optim(as.vector(c(1)), method = "BFGS", fn=function(x) {-logL(x)}, gr = NULL)
opt$par
```

```
## [1] 1.398391
```

Question 6.

Since the given distribution follows the regularity conditions, the asymptotic dsitribution would be a normal distribution,

$$\sqrt{n}(\hat{\mu}^{MLE} - \mu_0) \rightarrow N(0, I(\mu_0)^{-1})$$

where the fisher information can be calculated as,

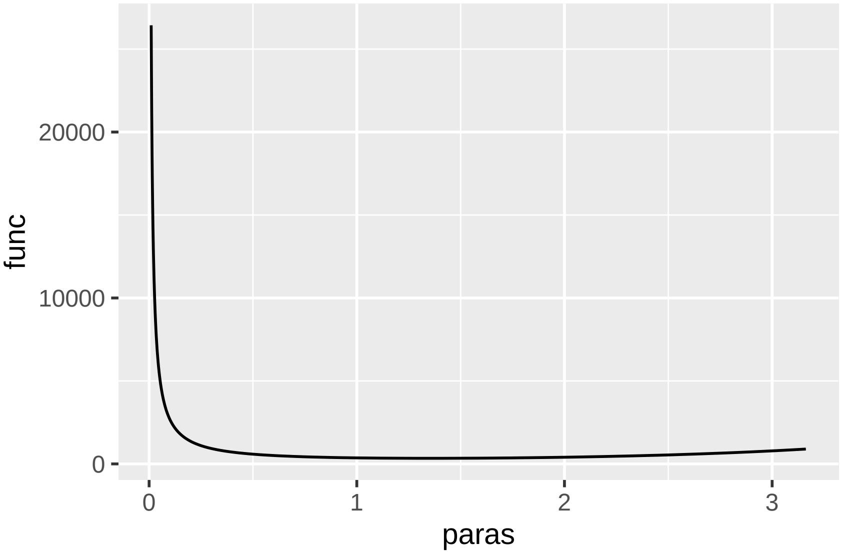
$$I(\mu_0) = -\mathbf{E}[\frac{\partial^2 \log(L)}{\partial \mu^2}] = \frac{n}{\mu} + \frac{n(\mu + e^{-\mu} - 1)}{e^{-\mu}(e^{\mu} - 1 - \mu)^2}$$

```
fisher <- function(mu) {
  n <- 264
  n / mu + n*(mu+exp(-mu)-1)/exp(-mu)/(exp(-mu)-1-mu)^2
}
optimize(fisher, lower=0, upper=10)
```

```
## $minimum
## [1] 1.35379
##
## $objective
## [1] 337.4855
```

Also, from the curve below, we can notice that the curve of fisher information around the MLE or optimal μ is quite flat. Therefore, we use MLE to calculate fisher information, which is 337.8257851.

```r
plot_curve(10^seq.int(-2, 0.5, 0.01), fisher)
```



Question 7.

Given the asymptotic distribution given by Q6, we have the confidence interval as,

$$\mu \in [\mu_{ML} - 1.96\frac{I(\mu_{ML})^{-1}}{\sqrt{n}}, \mu_{ML} + 1.96\frac{I(\mu_{ML})^{-1}}{\sqrt{n}}] = [1.39803, 1.39875]$$

Question 8.

It seems like a reasonable choice since different groups of majors have quite different value of μ as mentioned in the paper. But it would be hard to evaluate the mathematical properties of this estimator.

Question 9.

Our ML estimator is 1.3983907, which is quite similar to the result ($\mu = 1.4$) given by the paper. Both of them can show evidence that the techno-scientific findings are not inevitable.