Exam 1

Claudio Robelo

2022-08-30

Question 1

##loading data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- -read.delim("coalcold.tab", header = TRUE, sep = "\t", quote = "")
glimpse(data)
```

```
## Rows: 313
## Columns: 49
## $ DURAT    <dbl> -3, -7, -20, -6, -7, -2, -17, -27, -49, -4, -29, -49, -6, -23~
## $ BELGIUM  <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -~
## $ CANADA   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ DENMARK  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ FINLAND  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ FRANCE   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ICELAND  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ IRELAND  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ISRAEL   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ITALY    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NETHER   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NORWAY   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ PORTUG   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SPAIN    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SWEDEN   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ UK       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ T        <int> -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -~
## $ POPINFL  <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -~
## $ OPPINFL  <int> -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -3, -~
## $ ROPINFL  <int> -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -~
## $ PROX     <int> -45, -45, -45, -45, -45, -45, -45, -45, -45, -45, -45, -45, -~
## $ IDENT    <int> -1, -1, -1, -1, -1, -2, -2, -2, -2, -2, -2, -1, -1, -1, -1, -~
```

```
## $ VOLAT    <int> -93, -93, -93, -93, -93, -62, -62, -62, -62, -62, -62, -106, ~
## $ RESPONSE <int> -4, -4, -5, -5, -5, -5, -5, -5, -3, -5, -3, -1, -1, -3, -1, -~
## $ INVEST   <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -~
## $ POLAR    <int> -11, -11, -11, -11, -6, -3, -3, -3, -2, -1, -1, -5, -11, -11,~
## $ FRACT    <int> -656, -656, -656, -656, -634, -599, -599, -599, -620, -592, -~
## $ NUMST2   <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, 0, -1, -1, -1, -1, -1, -1~
## $ PARLBAS  <int> -54, -54, -79, -79, -63, -50, -50, -50, -52, -49, -59, -84, -~
## $ CRISIS   <int> -10, -24, -7, -7, -45, -51, -4, -6, -10, -23, -2, -29, -65, -~
## $ FORMAT   <int> -2, -5, -2, -3, -4, -1, -3, -1, -1, -2, -1, -1, -4, -5, -5, -~
## $ OPPCONC  <dbl> 0.86792, 0.86792, 0.14103, 0.14103, 0.58065, 0.79592, 0.79592~
## $ ELTIME2  <int> 0, 0, 0, 0, -1, -1, 0, 0, -1, -1, 0, -1, -1, 0, -1, -1, 0, -1~
## $ ELTIMEB  <int> 0, 0, 0, -1, 0, 0, 0, -1, 0, 0, -1, 0, 0, -1, 0, 0, -1, 0, 0,~
## $ ELTIMEN  <int> -1, -1, -1, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1,~
## $ CARETK2  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, ~
## $ NONPART  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ELECTC   <int> 0, 0, 0, -1, 0, 0, 0, -1, -1, 0, -1, -1, 0, 0, 0, 0, 0, 0, 0,~
## $ ELECTM   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SINGPAR  <int> 0, 0, 0, 0, 0, -1, -1, -1, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ EXTSUP   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SALIEN   <dbl> -190.4412, -190.4412, -190.4412, -190.4412, -190.4412, -261.2~
## $ OPPCON2  <dbl> -1.00000, -1.00000, -0.52381, -0.52381, -0.97297, -0.78000, -~
## $ OPPCOND  <int> -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ NP       <dbl> -2.90698, -2.90698, -2.90698, -2.90698, -2.73224, -2.49377, -~
## $ DECPROX  <int> -33, -33, -33, -33, -33, -50, -50, -50, -50, -50, -50, -75, -~
## $ CIEP12   <int> -1, -1, -1, -1, -1, -1, -1, -1, 0, -1, -1, 0, -1, -1, 0, -1, ~
## $ CIEP24   <int> -1, -1, -1, -1, -1, -1, -1, 0, 0, -1, 0, 0, -1, -1, 0, -1, -1~
## $ CIEPTW   <int> -1, -1, -1, -1, -1, -1, -1, -1, 0, -1, -1, 0, -1, -1, 0, -1, ~
```

Estimate* the rate and mean (reciprocal of the rate) for the exponential distribution. Use the parametric bootstrap to obtain a 95% confidence interval for each quantity.

mean = 1/lambda

```
#lambda with optim
lambda_fn <- function(lambda, x) {
log_lik <- length(x)*log(lambda) - lambda*sum(x)
}
est <- optim(par = 1, lower = 0, upper = 100, fn = lambda_fn,
             control = list(fnscale = -1), x = data$DURAT , method = 'Brent')
#rate
mlest<- est$par
#mean
1/est$par
```

```
## [1] 0.01
```

```
#parametric bootstrap
##rate
bs <- 2000
bs_est <- numeric(bs) # a container for the estimates
for (i in 1:bs) {
bs_y <- rexp(data$DURAT, mlest)
bs_est[i] <- mean(bs_y)
}

print(quantile(bs_est, probs = c(0.025, 0.975)), digits = 2)
```

```
##   2.5%  97.5%
## 0.0089 0.0111
```

```
##mean
bs <- 2000
bs_est <- numeric(bs) # a container for the estimates
for (i in 1:bs) {
bs_y <- rexp(data$DURAT, mlest)
bs_est[i] <- 1/mean(bs_y)
}

print(quantile(bs_est, probs = c(0.025, 0.975)), digits = 2)
```

```
##  2.5% 97.5%
##    90   112
```

#2 Simulation

```
#fake datasets to compare
fake_list <- list()
for(i in 1:5){
fake_list[[i]] <- rexp(nrow(data), rate = bs_est)
}

bind_cols(fake_list, data$DURAT) %>%
janitor::clean_names() %>%
rename('sim1'=x1,
'sim2'=x2,
'sim3'=x3,
'sim4'=x4,
'sim5'=x5,
'data'=x6) -> fake_data
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
```

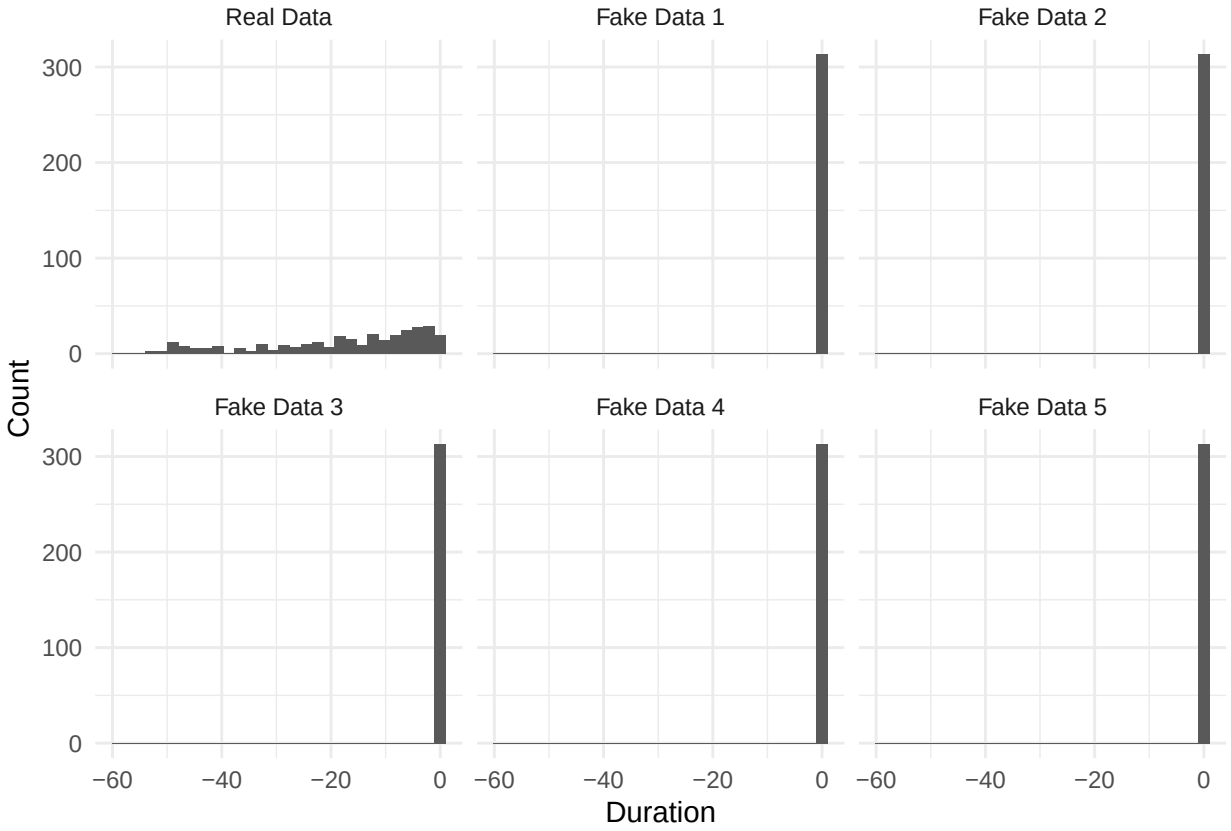
```
fake_data %>%
pivot_longer(cols = c(1:6)) -> fake_data
kableExtra::kable(fake_data %>%
group_by(name) %>%
summarize(mean = mean(value)),format = 'latex',digits = 3)
```

name	mean
data	-18.495
sim1	0.010
sim2	0.010
sim3	0.011
sim4	0.011
sim5	0.009

```
fake_data %>%
ggplot(aes(value))+
```

```
geom_histogram()+
facet_wrap(vars(name), labeller = as_labeller(c('data' = 'Real Data',
                                                'sim1'= 'Fake Data 1',
                                                'sim2'= 'Fake Data 2',
                                                'sim3'= 'Fake Data 3',
                                                'sim4'= 'Fake Data 4',
                                                'sim5'= 'Fake Data 5')))+
labs(y='Count', x='Duration')+
theme_minimal()
```

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



The observed data presents a much more skewed histogram, where the mean and range of values is small when compared with the predicted distribution. As such this is not the ideal distribution when compared with the actual data.