CHL5226 Assignment 3

Zhengyang Fei

1. **[10 marks]** The public health department of a city regularly samples lake and river water to ensure it is safe for swimming. Suppose that the bacterial count in unit water samples follows a Poisson distribution with mean μ.

Suppose that the exact count cannot be determined, rather it can only be determined if there are any bacteria in a sample of volume v. In this case, n independent volume v samples are analyzed and are either positive or negative for the bacteria. Suppose y of the n samples test negative. Determine the MLE $\hat{\mu}$ of μ, the mean bacterial count in the water.

Problem 1

Let X be the bacterial count in a sample of unit water with volume v. Since the count follows a Poisson distribution with mean μ per unit volume, the mean count in a sample of volume v is μv. That is,

$$x_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu v)$$

For a sample with no bacteria, the probability

$$P(X = 0) = \frac{(v\mu)^0 e^{-v\mu}}{0!} = e^{-v\mu}$$

is the likelihood that a given sample will test negative for bacteria.

Now, let Y be the total number of negative samples (no bacteria samples) out of the n samples. Then $Y \sim \text{binomial}(n, e^{-\mu v})$. So our likelihood kernel

$$L(\mu) = \left(e^{-v\mu}\right)^y \left(1 - e^{-v\mu}\right)^{n-y}$$

and the log-likelihood kernel

$$l(\mu) = y \ln\left(e^{-v\mu}\right) + (n-y) \ln\left(1 - e^{-v\mu}\right) = -yv\mu + (n-y) \ln\left(1 - e^{-v\mu}\right)$$

Differentiating with respect to μ and setting to zero, we have

$$\frac{d\ell(\mu)}{d\mu} = -yv + (n-y) \frac{ve^{-v\mu}}{1 - e^{-v\mu}} = 0$$

Lastly, solve for $\hat{\mu}$, we have

$$\hat{\mu} = -\frac{1}{v} \ln\left(\frac{y}{n}\right)$$

2. **[10 marks]** Below are 10 observations corresponding to wait times (to the nearest day) for a specialist referral letter.

1 20 13 25 4 7 5 26 36 32

Wait times are believed to come from an exponential distribution with mean θ.

$$f(x) = \frac{1}{\theta}e^{-x/\theta}$$

(a) Determine the MLE $\hat{\theta}$ from the approximate likelihood based on $f(x)$.

(b) If we consider each observation x_i to correspond to an interval $[x_i - 0.5, x_i + 0.5]$ determine the MLE $\hat{\theta}$ using the exact likelihood.

(c) Plot the relative log likelihood functions for the two approaches and determine approximate 95% confidence intervals (using $r(\theta)$) for the two approaches (approximate and exact).

(d) Obtain an approximate 95% confidence interval using the normal approximation based on the likelihood from part (a).

(e) Historical data suggest the mean is 14 days. Test if the data are consistent with this value.

(f) You have now obtained three 95% confidence intervals. They would all be expected to have 95% coverage probabilities. Determine if this is so. You may use simulation. Which intervals are also likelihood intervals?

Problem 2

(a) The log-likelihood for the exponential distribution is

$$\ell(\theta) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

Hence, we find the score function and set it to 0 to solve for $\hat{\theta}$

$$\frac{d\ell(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0 \implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Using our data, our MLE estimate

$$\hat{\theta} = \frac{1 + 20 + \cdots + 32}{10} = 16.9$$

(b) We treat each x_i as a continuous observation with an interval $[x_i - 0.5, x_i + 0.5]$. Then the probability for each x_i is,

$$P(x_i - 0.5 \leq X \leq x_i + 0.5) = \int_{x_i - 0.5}^{x_i + 0.5} \frac{1}{\theta}e^{-x/\theta}\,dx$$

$$= e^{-(x_i - 0.5)/\theta} - e^{-(x_i + 0.5)/\theta}.$$

So our likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \int_{x_i-0.5}^{x_i+0.5} \frac{1}{\theta} e^{-x/\theta}\, dx$$

$$= \prod_{i=1}^{n} \left(e^{-(x_i-0.5)/\theta} - e^{-(x_i+0.5)/\theta} \right)$$

$$= \prod_{i=1}^{n} e^{-x_i/\theta} \left(e^{-0.5/\theta} - e^{0.5/\theta} \right)$$

And our log-likelihood function is

$$\ell(\theta) = -\frac{\sum_{i=1}^{n} x_i}{\theta} + n \ln\left(e^{0.5/\theta} - e^{-0.5/\theta} \right)$$

Taking derivative and setting to 0 and solve for $\hat{\theta}$,

$$\frac{d\ell(\theta)}{d\theta} = -\frac{\sum_{i=1}^{n} x_i}{\theta^2} + n \cdot \frac{-\frac{0.5}{\theta^2} e^{0.5/\theta} + \frac{0.5}{\theta^2} e^{-0.5/\theta}}{e^{0.5/\theta} - e^{-0.5/\theta}}$$

$$\hat{\theta} = \left(\ln\left(\frac{\bar{x}+0.5}{\bar{x}-0.5} \right) \right)^{-1}$$

Using $\bar{x} = 16.9$, we get the exact MLE

$$\hat{\theta} = 16.8950$$

(c) Recall the relative log-likelihood function $r(\theta) = \ell(\theta) - \ell(\hat{\theta})$.

- Recall from part (a):

$$\ell(\theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} x_i \quad \text{and}$$

$$\hat{\theta}_{\text{approx}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  Hence

$$r_{\text{approx}}(\theta) = \ell_{\text{approx}}(\theta) - \ell_{\text{approx}}(\hat{\theta}_{\text{approx}})$$

$$= -n \ln(\theta) - \frac{\sum_{i=1}^{n} x_i}{\theta} + n \ln\left(\frac{\sum_{i=1}^{n} x_i}{n} \right) + n$$

$$= -n \ln(\theta) - \frac{n\bar{x}}{\theta} + n \ln\left(\frac{\bar{x}}{n} \right) + n$$

$$= -10 \ln(\theta) - \frac{169}{\theta} + 10 \ln(16.9) + 10$$

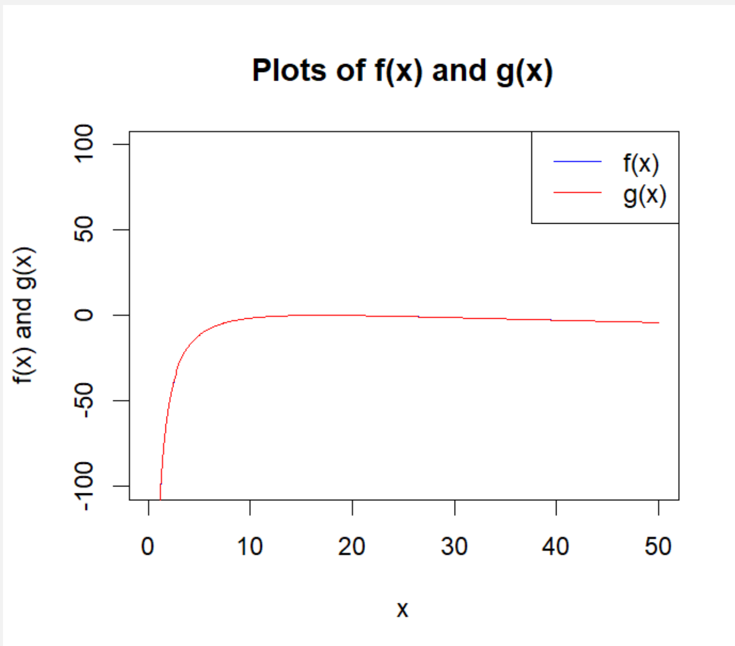
– Recall from part (b):

$$\hat{\theta}_{\text{exact}} = \left(\ln\left(\frac{\bar{x}+0.5}{\bar{x}-0.5}\right)\right)^{-1}$$

Hence

$$\begin{aligned} r_{\text{exact}}(\theta) &= \ell_{\text{exact}}(\theta) - \ell_{\text{exact}}(\hat{\theta}_{\text{exact}}) \\ &= -\frac{\sum_{i=1}^{n} x_i}{\theta} + n\ln\left(e^{0.5/\theta} - e^{-0.5/\theta}\right) + \frac{\sum_{i=1}^{n} x_i}{\hat{\theta}} - n\ln\left(e^{0.5/\hat{\theta}} - e^{-0.5/\hat{\theta}}\right) \\ &= -\frac{169}{\theta} + 10\ln\left(e^{0.5/\theta} - e^{-0.5/\theta}\right) + \frac{169}{16.8950} - 10\ln\left(e^{0.5/16.8950} - e^{-0.5/16.8950}\right) \end{aligned}$$

Now, using R we can plot the RLFs (code at bottom). Here f and g are the approximate and exact approaches respectively.



To obtain the 95% confidence intervals, we set $r_{\text{approx}}(\theta) \geq \ln(0.147)$ and solve that the 95% CI is $(9.6359, 33.6969)$. Similarly, we set $r_{\text{exact}}(\theta) \geq \ln(0.147)$ and obtain the 95% CI to be $(9.6398, 33.7043)$ (code at bottom).

(d) Now we use 95% confidence interval with normal approximation based on the likelihood from part (a). Note that the information function

$$S(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3}\sum x_i$$

For $\theta = 16.9$, we get $S(\theta) = 0.035$. Hence, our CI

$$16.9 \pm 1.96 \times \sqrt{0.035} = (6.4233, 27.3766)$$

(e) Conducting a likelihood ratio test (see code) with $H_0 : \theta = 14$, the result is summarized below

Case	**D**	**p-value**
Approximate	0.377731	0.538820
Exact	0.376398	0.539537

In both cases, we see data provides very little evidence that $\theta \neq 14$.

(f) Using simulations, from the table of summarized results (code below) we see that this is indeed true.

# Runs	nobs	LI for Exact	LI for Approximate	Normal CI
1000	10	0.955	0.955	0.902

The interval from part (c) are likelihood intervals, since we used the relative log-likelihood to estimate them at 14.7%.

3. [**10 marks**] Write an R function to perform Newton's method and use it to obtain the MLEs from the previous question, both exact and approximate log likelihoods.

Problem 3

We summarize the result first in a table.

Initial Theta	# Iterations	Theta Hat
20.97108	6	16.9
20.97108	24	16.89507

```
# Score function for approximate case
expS <- function(theta, x, n) {
  -(n/theta) + (1/theta^2) * sum(x)
}

# Information function
expI <- function(theta, x, n) {
  (n / theta^2) - (2 / theta^3) * sum(x)
}

# The data
x <- c(1, 20, 13, 25, 4, 7, 5, 26, 36, 32)
n <- length(x)  # Number of observations
max_iter <- 100  # To avoid infinite loop in case of non-convergence
iter <- 0 # Counter
set.seed(1030)

th0=runif(1,1,24)
cat("initial theta:", th0)
```

```
initial theta: 20.97108
```

```
eps = 1e-7 # threshold
for (i in 1:max_iter){
  th.i <- th0 - (expS(th0, x, n) / expI(th0, x, n))
  iter <- iter + 1
  if ((abs(th.i - th0) < eps || iter >= max_iter)) {
    th0 = th.i
    break
  }
  th0 <- th.i

}
cat("\n#iteration:", iter, "\ntheta_hat:", th0)
```

```
#iteration: 6
theta_hat: 16.9
```

```r
# Score function for exact case
expS.exact <- function(theta, x, n) {
  -n/theta^2 * (exp(-1/theta)/(1-exp(-1/theta))) + sum(x-0.5)/theta^2
}

# Information function for exact case
expI.exact  <- function(theta, x, n) {
  t1 <- -2*n*exp(-1/theta)/(theta^3*(1-exp(-1/theta)))
  t2 <- +n/theta^4 * (exp(1/theta)/(exp(1/theta)-1)^2)
  t3 <- + 2*sum(x-0.5)/theta^3
}

# The data
x <- c(1, 20, 13, 25, 4, 7, 5, 26, 36, 32)
n <- length(x)  # Number of observations
max_iter <- 100  # To avoid infinite loop in case of non-convergence
iter <- 0 # Counter
set.seed(1030)
th0.exact = runif(1, 1, 24)
cat("initial theta:", th0.exact)
```

```
initial theta: 20.97108
```

```r
eps = 1e-7
for (i in 1:max_iter){
  th.i <- th0.exact + expS.exact(th0.exact, x, n) / expI.exact(th0.exact, x, n)
  iter <- iter + 1
  if (abs(th.i - th0.exact) < eps || iter >= max_iter) {
    th0.exact = th.i
    break
  }
  th0.exact = th.i

}
cat("\n#iteration:", iter, "\ntheta_hat:", th0.exact)
```

```
#iteration: 24
theta_hat: 16.89507
```

4. **[10 marks]** Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent exponential variables. The X_i's have mean θ and the Y_i's have mean $\lambda\theta$ where λ and θ are positive unknown parameters.

(a) Derive expressions for $\hat{\lambda}$ and $\hat{\theta}$. **Note:** Use the density functions to form the likelihood rather than attempting the exact likelihood approach.

(b) Suppose the data below are survival times for patients on two different treatments. The survival times are assumed to be exponential with a mean θ for treatment A and mean $\lambda\theta$ for treatment B.

Treatment A:	9	186	25	6	44	115
Treatment B:	1	18	6	25	14	45

Find $\hat{\lambda}$ and $\hat{\theta}$.

(c) Plot and therefore determine the 10% likelihood region for λ and θ. **Note:** It is expected that you will use the computer for this.

(d) Determine the 10% likelihood interval for λ using $R_{\max}(\lambda)$ or $r_{\max}(\lambda)$.

(e) Test the hypothesis $H : \lambda = 1$. Based on your results is there convincing evidence that the groups differ in survival and if so, which group seems to be better?

Problem 4

(a) We first find $\hat{\theta}$ and then $\hat{\lambda}$

- We proceed as usual: finding log-likelihood, score function by taking derivative, and finally setting score function equal to zero and solve for MLE.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(X_i;\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-X_i/\theta} \\
&= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} X_i/\theta}, \\
\ell(\theta) &= \ln L(\theta) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} X_i, \\
\frac{d\ell(\theta)}{d\theta} &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i = 0, \\
\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} X_i.
\end{aligned}
$$

– Similarly,

$$
\begin{aligned}
L(\theta, \lambda) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-X_i/\theta} \times \prod_{j=1}^{n} \frac{1}{\lambda\theta} e^{-Y_j/(\lambda\theta)}, \\
\ell(\theta, \lambda) &= \ln L(\theta, \lambda) = -n\ln(\theta) - \frac{\sum_{i=1}^{n} X_i}{\theta} - n\ln(\lambda) - n\ln(\theta) - \frac{1}{\theta}\sum_{j=1}^{n}\frac{Y_j}{\lambda}, \\
\frac{\partial \ell(\theta, \lambda)}{\partial \lambda} &= -\frac{n}{\lambda} + \frac{1}{\lambda^2\theta}\sum_{j=1}^{n} Y_j = 0, \\
\hat{\lambda} &= \frac{\sum_{j=1}^{n} Y_j}{n\theta}.
\end{aligned}
$$

Using $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we can substitute and simplify,

$$
\begin{aligned}
\hat{\lambda} &= \frac{\sum_{j=1}^{n} Y_j}{n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}, \\
\hat{\lambda} &= \frac{\sum_{j=1}^{n} Y_j}{\sum_{i=1}^{n} X_i}.
\end{aligned}
$$

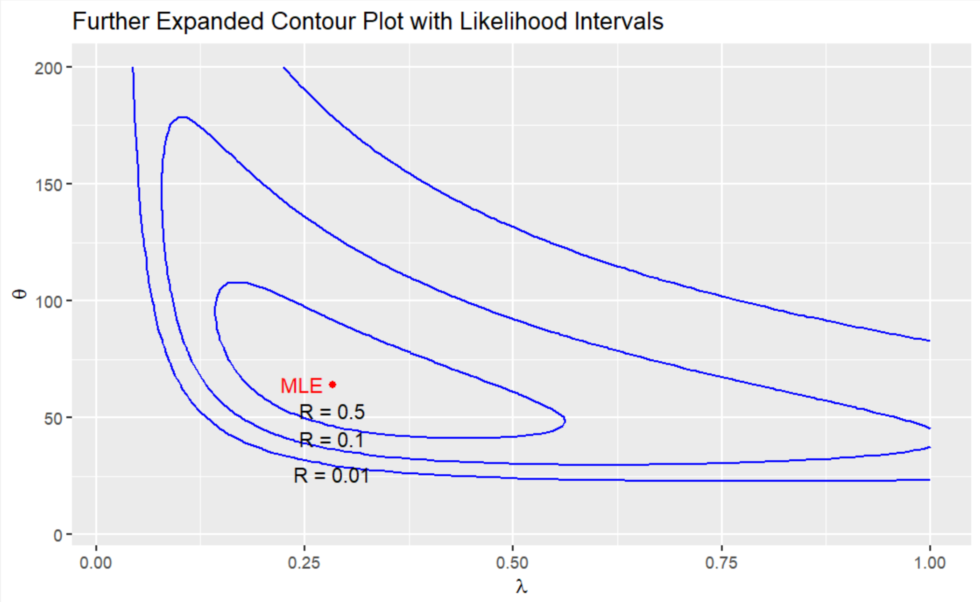
Hence, we get that

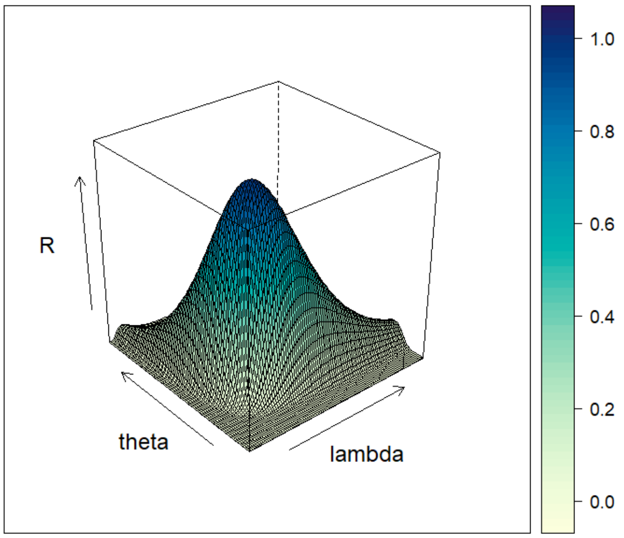
$$
\begin{aligned}
\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\hat{\lambda} &= \frac{\sum_{j=1}^{n} Y_j}{\sum_{i=1}^{n} X_i}.
\end{aligned}
$$

(b) From R, we use the MLE estimate formulas we have from part (a) and the data to get that

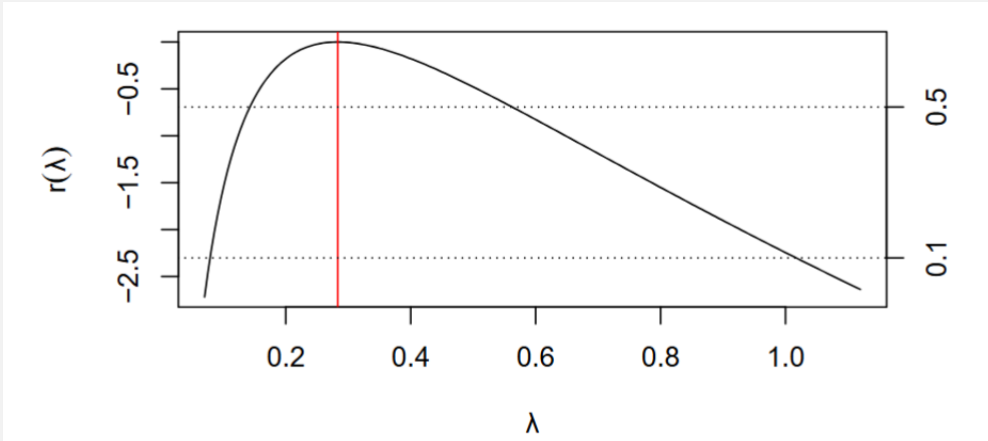
$$\hat{\theta} = 64.1667 \quad \text{and} \quad \hat{\lambda} = 0.2831$$

(c) We show the contour plot and 3-dimensional plot (refer to code: Part c and Part c (again))





(d) We can first plot the relative likelihood



Furthermore, we can precisely determine the 10% likelihood interval

$$(0.0787, 1.01723)$$

which supports the graph.

(e) Using the likelihood ratio test for hypothesis testing for $H_0 : \lambda = 1$, we get a test statistic $D = 4.490225$ which gives us p-val $= 0.034089$. Since p-val < 0.05, this means that the data provide very little evidence that $\lambda \neq 1$.