18.650 Problem Set 3 Spring 2017
Statistics for Applications
Due Date: Fri 3/3/2017, prior to 4:00pm
Where: Electronically to Stellar website (preferred)
or Problem Set Box (outside 4-174)

Problems from John A. Rice, Third Edition. [*Chapter.Section.Problem*]

1. Problem 8.10.21. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. with density function

$$f(x \mid \theta) = \begin{cases} e^{-(x-\theta)}, & if \quad x \geq \theta \\ 0, & otherwise \end{cases}$$

(a). (3 points) Find the method of moments estimate of θ.

The first moment of X is

$$\begin{aligned} \mu_1 &= E[X] = \int_\theta^\infty x e^{-(x-\theta)} dx \\ &= \theta + \int_0^\infty y e^{-y} dy \\ &= \theta + [(y)(-e^y)]|_{y=0}^{y=\infty} + \int_0^\infty e^{-y} dy \\ &= \theta + 1 \end{aligned}$$

(The second line follows by tranforming to $y = x - \theta$; the third line follows from integration-by-parts.)

Equating the sample first moment to the population first moment:

$$\begin{aligned} \mu_1 &= \hat{\mu}_1 \\ \theta + 1 &= \frac{1}{n}\sum_1^n X_i = \overline{X} \\ \Longrightarrow \quad \hat{\theta} &= \overline{X} - 1 \end{aligned}$$

(b). (4 Points) Find the mle of θ. The likelihood of the data is

$$\begin{aligned} lik(\theta) &= f(X_1, \ldots, X_n \mid \theta) \\ &= \prod_{i=1}^n f(X_i \mid \theta) \\ &= \prod_{i=1}^n [e^{-(X_i-\theta)} \mathbf{1}_{[\theta,\infty)}(X_i)] \\ &= [e^{-\sum_{i=1}^n (X_i-\theta)}] \prod_{i=1}^n [\mathbf{1}_{[0,X_i]}(\theta)] \\ &= [e^{-\sum_{i=1}^n X_i} e^{n\theta}][\mathbf{1}_{[0,min(X_1,\ldots,X_n)]}(\theta) \end{aligned}$$

$lik(\theta)$ is maximized by maximizing θ subject to $\theta \leq X_i$,

for all $i = 1, \ldots, n$

i.e., $\hat{\theta}_{MLE} = min(X_1, \ldots, X_n)$

(c). (5 Points) Find a sufficient statistic for θ. Consider

$$T(X_1, \ldots, X_n) = min(X_1, \ldots, X_n)$$

The distribution function of T, $F_T(t)$ satisfies

$$\begin{aligned}
[1 - F_T(t)] &= P(T > t) \\
&= P(X_1 > t, X_2 > t, \ldots X_n > t) \\
&= \prod_{i=1}^n P(X_i > t) \\
&= \prod_{i=1}^n [e^{-(t-\theta)}] \\
&= [e^{-n(t-\theta)}]
\end{aligned}$$

for values $t \geq \theta$.

The density of T is simply the derivative:

$$f_T(t \mid \theta) = ne^{-n(t-\theta)}, \;\; t \geq \theta.$$

The conditional density of the sample given $T = t$ is

$$\begin{aligned}
f(X_1, \ldots, X_n \mid T, \theta) &= \frac{f(X_1, \ldots, X_n \mid \theta)}{f_T(t \mid \theta)} \\
&= \frac{[e^{-\sum_{i=1}^n X_i} e^{n\theta}][\mathbf{1}_{[0, min(X_1, \ldots, X_n)]}(\theta)]}{[e^{-n(t-\theta)}]\mathbf{1}_{[0,t]}(\theta)} \\
&= [e^{-\sum_{i=1}^n (X_i - t)}][\prod_{i=1}^n \mathbf{1}_{[t,\infty)}(X_i)]
\end{aligned}$$

The density function does not depend on θ, so

$T = min(X_1, \ldots, X_n)$ is sufficient for θ.

2. Problem 8.10.45. **A Random walk Model for Chromatin**

The html in Rproject3.zip "*Rproject3//Rproject3_rmd_rayleigh_theory.html*" details estimation theory for a sample from a Rayleigh distribution.

Note: Parts (h) and (i) are for extra credit.

(a). (3 Points) MLE of θ :

Data consisting of:

$$R_1, R_2, \ldots, R_n$$

are i.i.d. $Rayleigh(\theta)$ random variables. The likelihood function is

$$\begin{aligned}
lik(\theta) &= f(r_1, \ldots, r_n \mid \theta) = \prod_{i=1}^n f(r_i \mid \theta) \\
&= \prod_{i=1}^n \left[\frac{r_i}{\theta^2} exp\left(\frac{-r_i^2}{2\theta^2}\right) \right]
\end{aligned}$$

The log-likelihood function is

$$\begin{aligned}
\ell(\theta) &= \log[lik(\theta)] \\
&= [\textstyle\sum_1^n log(r_i)] - 2nlog(\theta) - \frac{1}{\theta^2}\sum_1^n [r_i^2/2]
\end{aligned}$$

The mle solves $\frac{d}{d\theta}\ell(\theta) = 0$:

$$\begin{aligned}
0 &= \tfrac{d}{d\theta}(\ell(\theta)) \\
&= -2n(\tfrac{1}{\theta}) + 2(\tfrac{1}{\theta^3})\textstyle\sum_1^n [r_i^2/2] \\
\implies \hat{\theta}_{MLE} &= (\tfrac{1}{n}\textstyle\sum_1^n [r_i^2/2])^{1/2} \\
&= \left(\tfrac{1}{2n}\textstyle\sum_1^n r_i^2\right)^{\frac{1}{2}}
\end{aligned}$$

(b). (3 Points) Method of moments estimate:

The first moment of the $Rayleigh(\theta)$ distribution is

$$\begin{aligned}
\mu_1 &= E[R \mid \theta] = \int_0^\infty r f(r \mid \theta) dr \\
&= \int_0^\infty r \frac{r}{\theta^2} exp(\frac{-r^2}{2\theta^2}) dr \\
&= \frac{1}{\theta^2} \int_0^\infty r^2 exp(\frac{-r^2}{2\theta^2}) dr \\
&= \frac{1}{\theta^2} \int_0^\infty v \cdot exp(\frac{-v}{2\theta^2})[\frac{dv}{2\sqrt{v}}] \text{ (change of variables: } v = r^2) \\
&= \frac{1}{2\theta^2} \int_0^\infty v^{\frac{3}{2}-1} \cdot exp(\frac{-v}{2\theta^2}) dv \\
&= \frac{1}{2\theta^2} \Gamma(\frac{3}{2})(2\theta)^{\frac{3}{2}} \\
&= \sqrt{2}\theta\Gamma(\frac{3}{2}) = \sqrt{2}\theta \times (\frac{1}{2})\Gamma(\frac{1}{2}) \\
&= \theta \times \frac{\sqrt{\pi}}{\sqrt{2}}
\end{aligned}$$

(using the facts that $\Gamma(n+1) = n\Gamma(n)$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$)

The MOM estimate solves:

$$\begin{aligned}
\mu_1 &= \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n R_i = \overline{R} \\
\theta \times \frac{\sqrt{\pi}}{\sqrt{2}} &= \overline{R} \\
\Longrightarrow \quad \hat{\theta}_{MOM} &= \overline{R} \times \frac{\sqrt{2}}{\sqrt{\pi}}
\end{aligned}$$

(c). (4 Points) Approximate Variance of the MLE and method of moments estimate.

The approximate variance of the MLE is $Var(\hat{\theta}_{MLE}) \approx \frac{1}{nI(\theta)}$

where

$$\begin{aligned}
I(\theta) &= E[-\frac{d^2}{d\theta^2}(log(f(x \mid \theta)))] \\
&= E[-\frac{d^2}{d\theta^2}[log(\frac{x}{\theta^2} exp(-\frac{x^2}{2\theta^2}))]] \\
&= E[-\frac{d}{d\theta}[-2(\frac{1}{\theta}) - (\frac{x^2}{2})(-2)\theta^{-3}]] \\
&= E[-[(\frac{2}{\theta^2}) + (x^2))(-3)\theta^{-4}]] \\
&= 3\theta^{-4}E[x^2] - (\frac{2}{\theta^2}) = 3\theta^{-4}(2\theta^2) - (\frac{2}{\theta^2}) \\
&= \frac{4}{\theta^2}
\end{aligned}$$

So, $Var(\hat{\theta}_{MLE}) \approx \frac{\theta^2}{4n}$

Variance of the MOM estimate of Rayleigh Distribution Parameter:

The MOM estimate

$\hat{\theta}_{MOM} = \overline{R} \times \frac{\sqrt{2}}{\sqrt{\pi}}$

has variance:

$Var(\hat{\theta}_{MOM}) = (\frac{\sqrt{2}}{\sqrt{\pi}})^2 Var(\overline{R}) = (\frac{2}{\pi})\frac{Var(R)}{n}$

$$\begin{aligned} Var(R) &= E[R^2] - (E[R])^2 \\ &= 2\theta^2 - (\sqrt{\tfrac{\pi}{2}}\theta)^2 \\ &= \theta^2(2 - \tfrac{\pi}{2}) \end{aligned}$$

So, $Var(\hat{\theta}_{MOM}) = \theta^2(2 - \frac{\pi}{2})(\frac{2}{\pi})(\frac{1}{n}) = \theta^2(\frac{4}{\pi} - 1)(\frac{1}{n}) \approx \frac{\theta^2}{n} \times 0.2732$

This exceeds the approximate $Var(\hat{\theta}_{MLE}) \approx \frac{\theta^2}{n} \times 0.25$

For parts (d), (e), (f), (g), See the R script file: (3 points for each of parts (d), (e), (f), and (g) *Rproject3_script4_Chromatin_solution.r*

(h) and (i), (5 Points each)

(h) For this part we need to define an R function that will generate random deviates from a $Rayleigh(\theta)$ distribution. Following the hint in the problem:

Suppose X follows a Rayleigh distribution with $\theta = 1$. We show first that $Y = \theta X$ follows a Rayleigh distribution with parameter θ.

The cdf of X is

$$F_X(c) = 1 - P(X \geq c) = 1 - e^{-x^2/2}.$$

Note that the density $f(x \mid \theta)$ is the derivative of the cdf.

If $Y = \theta X$, then the cdf of Y is:

$$\begin{aligned} F_Y(c) = P(Y \leq c) &= P(\theta X \leq c) \\ &= P(X \leq c/\theta) \\ &= 1 - exp(-(c/\theta)^2/2) \end{aligned}$$

The derivative of $F_Y(c)$ gives the density of the Rayleigh(θ) distribution.

To generate random values of X, we use Proposition D of Section 2.3:

Let U be a uniform random variable on $[0,1]$ and let $X = F^{-1}(U)$. Then the cdf of X is F. This is true by the proof:

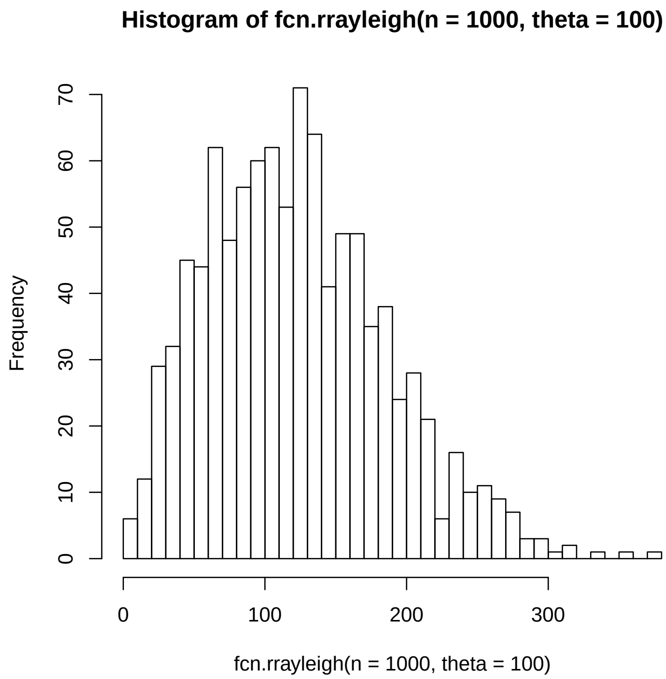
$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

So, we can generate random values of the Rayleigh($\theta = 1$) distribution by generating a random $U \sim Uniform[0,1]$ and computing

$$X = F_X^{-1}(U) = \sqrt{2 * (-log(1-U))}$$

The following function will generate random deviates from a $Rayleigh(\theta)$ distribution

```
> fcn.rrayleigh<-function(n=1, theta=1){
+   vec.U=runif(n)
+   vec.X = sqrt(-2.*log(1-vec.U))
+   vec.Y = theta*vec.X
+   return(vec.Y)
+ }
> # Test out the function by generating a large sample
> #
> hist(fcn.rrayleigh(n=1000,theta=100), breaks=50)
```



This function is used for parts (h) and (i) in R script file:

$Rproject3\_script4\_Chromatin\_solution.r$

3. Problem 8.10.51 **Double Exponential (Laplace) Distribution**

   (6 Points)

   The double exponential distribution is

   $$f(x \mid \theta) = \frac{1}{2}e^{|x-\theta|}, \ -\infty < x < \infty.$$

   For an iid sample of size $n = 2m+1$, show that the mle of θ is the median of the sample.

   Let $X_1, \ldots, X_n$ denote the sample random variables with outcomes $x_1, \ldots, x_n$. The likelihood function of the data is

   $$\begin{array}{rcl} lik(\theta) & = & \prod_{i=1}^n f(x_i \mid \theta) = \prod_{i=1}^n [\frac{1}{2}e^{-|x_i-\theta|}] \\ & = & (\frac{1}{2})^n e^{-\sum_{i=1}^n |x_i-\theta|} \end{array}$$

   This is maximized by minimizing the sum in the exponent:

   $$g(\theta) = \sum_{i=1}^n |x_i - \theta|$$

   Note that $g(\theta)$ is a continuous function of θ and its derivative exists at all points θ that are not equal to any x_i

$$
\begin{aligned}
g'(\theta) &= \tfrac{d}{d\theta} g(\theta) = \textstyle\sum_{i=1}^{n} [-1 \times \mathbf{1}(x_i > \theta) + (+1) \times \mathbf{1}(x_i < \theta)] \\
&= (-1) \times [\textstyle\sum_{i=1}^{n} \mathbf{1}(x_i > \theta)] + (+1) \times \textstyle\sum_{i=1}^{n} \mathbf{1}(x_i < \theta)] \\
&= \begin{cases} positive & if \quad \theta > median(x_i) \\ negative & if \quad \theta < median(x_i) \end{cases}
\end{aligned}
$$

It follows that $g(\theta)$ is minimized at $\theta = median(x_i)$. A graph of $g(\theta)$ is piecewise linear with slope changes at each of the x_i values; the slope at any given θ (not equal to an x_j) is

$$count(x_i < \theta) - count(x_i > \theta).$$

4. Problem 8.10.58 **Gene Frequencies of Haptoglobin Type**

(3 points for each part)

Gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2. Plato et al. published the following data on Haptoglobin Type in a sample of 190 people

Haptoglobin Type		
Hp1-1	Hp1-2	Hp2-2
10	68	112

This is precisely the same problem as Example 8.5.1.A of the text and class notes which corresponds to count data: $(X_1, X_2, X_3) \sim Multinomial(n = 3, p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$ distribution.

(a). Find the mle of θ

- $(X_1, X_2, X_3) \sim Multinomial(n, p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2))$

- Log Likelihood for θ

$$
\begin{aligned}
\ell(\theta) &= log(f(x_1, x_2, x_3 \mid p_1(\theta), p_2(\theta), p_3(\theta))) \\
&= log(\tfrac{n!}{x_1!x_2!x_3!} p_1(\theta)^{x_1} p_2(\theta)^{x_2} p_3(\theta)^{x_3}) \\
&= x_1 log((1-\theta)^2) + x_2 log(2\theta(1-\theta)) \\
&\qquad + x_3 log(\theta^2) + (\text{non-}\theta \; terms) \\
&= (2x_1 + x_2) log(1-\theta) + (2x_3 + x_2) log(\theta) + (\text{non-}\theta \; terms)
\end{aligned}
$$

- First Differential of log likelihood:

$$\ell'(\theta) = -\frac{(2x_1 + x_2)}{1-\theta} + \frac{(2x_3 + x_2)}{\theta}$$

$$\implies \hat{\theta} = \frac{2x_3 + x_2}{2x_1 + 2x_2 + 2x_3} = \frac{2x_3 + x_2}{2n} = \frac{2(112) + 68}{2(190)} = 0.76842$$

(b). Find the asymptotic variance of the mle.

- $Var(\hat{\theta}) \longrightarrow \dfrac{1}{E[-\ell''(\theta)]}$
- Second Differential of log likelihood:

$$\begin{aligned}\ell''(\theta) &= \frac{d}{d\theta}[-\frac{(2x_1+x_2)}{1-\theta}+\frac{(2x_3+x_2)}{\theta}] \\ &= -\frac{(2x_1+x_2)}{(1-\theta)^2}-\frac{(2x_3+x_2)}{\theta^2}\end{aligned}$$

- Each of the X_i are $Binomial(n, p_i(\theta))$ so

$$\begin{aligned}E[X_1] &= np_1(\theta) = n(1-\theta)^2 \\ E[X_2] &= np_2(\theta) = n2\theta(1-\theta) \\ E[X_3] &= np_3(\theta) = n\theta^2\end{aligned}$$

- $E[-\ell''(\theta)] = \dfrac{2n}{\theta(1-\theta)}$
- $\hat{\sigma}^2_{\hat{\theta}} = \dfrac{\hat{\theta}(1-\hat{\theta})}{2n} = \dfrac{0.76842(1-0.76842)}{2\times 190} = 0.0004682898 = (.02164)^2$

Parts (c), (d), and (e): see the R script

Rproject3_script1_multinomial_simulation_Problem_8_57.r