Problem Set 3: Written Part

Your Name Goes Here

Details

Due Date

This assignment is due at 4:00 PM on Friday, Feb 22.

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 20% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited in the space I have provided at the end of this problem set.

Problem I: Bird Hops (Adapted from Rice problem 8.8)

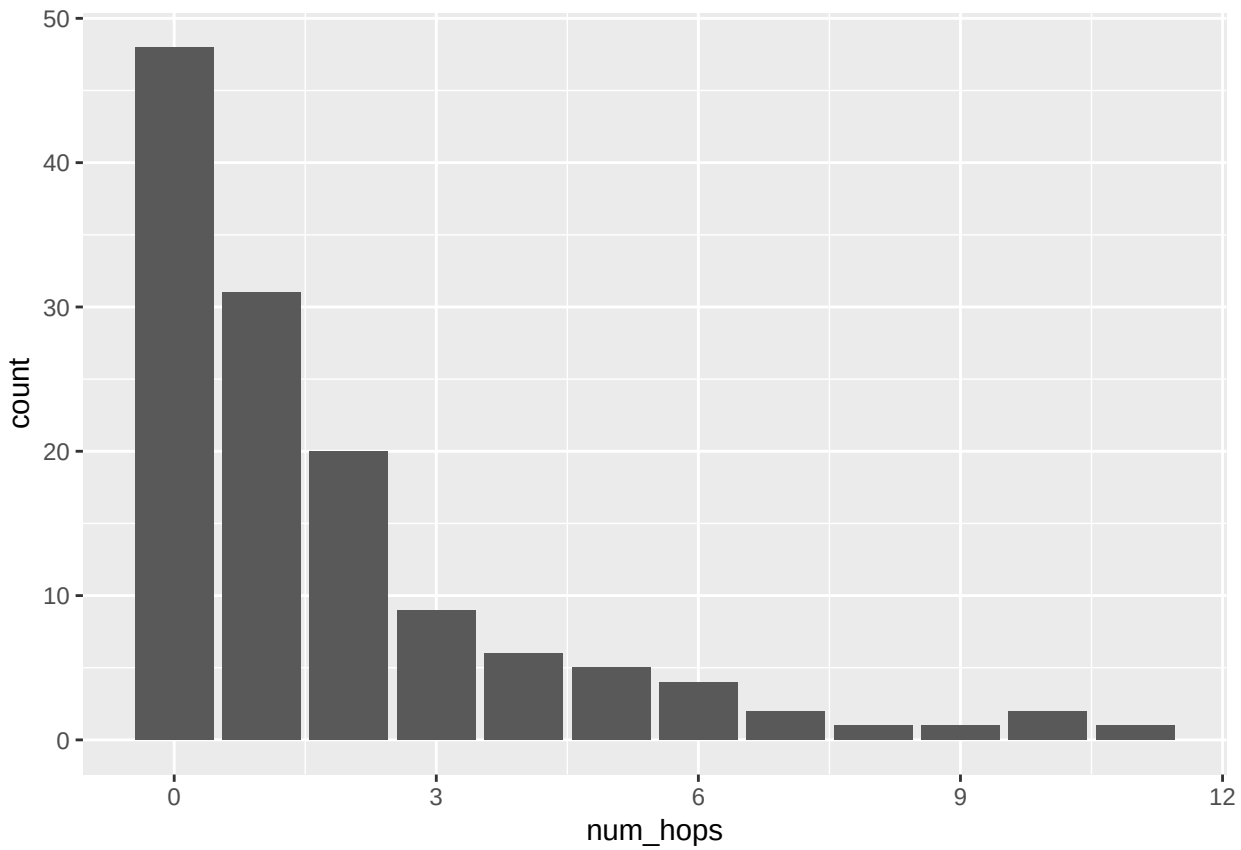
In an ecological study of the feeding behavior of birds, researchers counted the number of hops each of several birds took between flights. The R code below reads in the data and makes an initial plot:

```
bird_hops <- read_csv("http://www.evanlray.com/data/rice/Chapter%208/hops.csv") %>%
  transmute(
    num_hops = num_hops - 1
  )
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   num_hops = col_double()
## )
```

```
ggplot(data = bird_hops, mapping = aes(x = num_hops)) +
  geom_bar()
```




```
bird_hops %>%
  count(num_hops)
```

```
## # A tibble: 12 x 2
##    num_hops     n
##       <dbl> <int>
##  1        0    48
##  2        1    31
##  3        2    20
##  4        3     9
##  5        4     6
##  6        5     5
##  7        6     4
##  8        7     2
##  9        8     1
## 10        9     1
## 11       10     2
## 12       11     1
```

After subtracting 1 from the originally reported counts (as is done in the code above), the number of hops is 0 if the bird took off directly (0 hops before successful flight), 1 if the bird took one hop before taking off, and so on. Let's model these data with a Geometric(p) distribution. In the written part of this assignment you will find the maximum likelihood estimator for the parameters of a geometric distribution, and in the R part of the assignment you will fit the model to this data set.

Note that there are multiple parameterizations of the geometric distribution in common use. For this problem, please use the parameterization given in the "probability distributions" handout.

(1) In a sentence or two, explain why the geometric distribution is a reasonable model for these data. What does the parameter p represent?

(2) Find the maximum likelihood estimator of p.

(3) Is your result from part (2) a random variable or a number? If it is a random variable, explain why it is random. If it is a number, explain why it is not random.

Problem II: Spatial Organization of Chromosome (Rice Problem 8.45)

A human chromosome is a very large molecule, about 2 or 3 centimeters long, containing 100 million base pairs (Mbp). The cell nucleus, where the chromosome is contained, is in contrast only about a thousandth of a centimeter in diameter. The chromosome is packed in a series of coils, called chromatin, in association with special proteins (histones), forming a string of microscopic beads. It is a mixture of DNA and protein. In the G0/G1 phase of the cell cycle, between mitosis and the onset of DNA replication, the mitotic chromosomes diffuse into the interphase nucleus. At this stage, a number of important processes related to chromosome function take place. For example, DNA is made accessible for transcription and is duplicated, and repairs are made of DNA strand breaks. By the time of the next mitosis, the chromosomes have been duplicated. The complexity of these and other processes raises many questions about the large-scale spatial organization of chromosomes and how this organization relates to cell function. Fundamentally, it is puzzling how these processes can unfold in such a spatially restricted environment.

In the 1990's, a series of experiments (Sachs et al., 1995; Yokota et al., 1995) were conducted to learn more about the spatial organization of chromosomes. Pairs of small DNA sequences (size about 40 kbp) at specified locations on human chromosome 4 were flourescently labeled in a large number of cells. The distances between the members of these pairs were then determined by flourescence microscopy. (The distances measured were actually two-dimensional distances between the projections of the paired locations onto a plane.) The empirical distribution of these distances provides information about the nature of large-scale organization.

There has long been a tradition in chemistry of modeling the configurations of polymers by the theory of random walks. As a consequence of such a model, the two-dimensional distance should follow a Rayleigh distribution. If $X \sim \text{Rayleigh}(\theta)$ (with parameter $\theta > 0$), then the probability density function is given by

$$f(x|\theta) = \frac{x}{\theta} \exp\left(\frac{-x^2}{2\theta}\right)$$

for positive values of x (and the probability density function is 0 for non-positive values of x). Note that I have made a slight modification to the standard form of the probability density function here. This modification will simplify the algebra you have to do for this problem, and does not change anything fundamental about the distribution.

In the problems below, you may find the following to be helpful:

$$\begin{aligned} E(X) &= \left(\frac{\theta\pi}{2}\right)^{1/2} \\ E(X^2) &= 2\theta \\ Var(X) &= 2\theta - \frac{\theta\pi}{2} \end{aligned}$$

In this problem, you will fit the Rayleigh distribution to some of the experimental results. You will find the maximum likelihood estimator in the written part, and fit it to the actual data set in the R part. The entire data set comprises 36 experiments in which the separation between the pairs of flourescently tagged locations ranged from 10 Mbp to 192 Mbp. In each such experimental condition, about 100–200 measurements of two-dimensional distances were determined. This exercise will be concerned just with the data from one of these experiments. The R code below reads the data in, displays the first few rows of the data and the dimensions of the data set, and makes an initial plot.

```
chromatin <- read_csv("http://www.evanlray.com/data/rice/Chapter%208/Chromatin/data05.txt",
  col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double()
## )
```

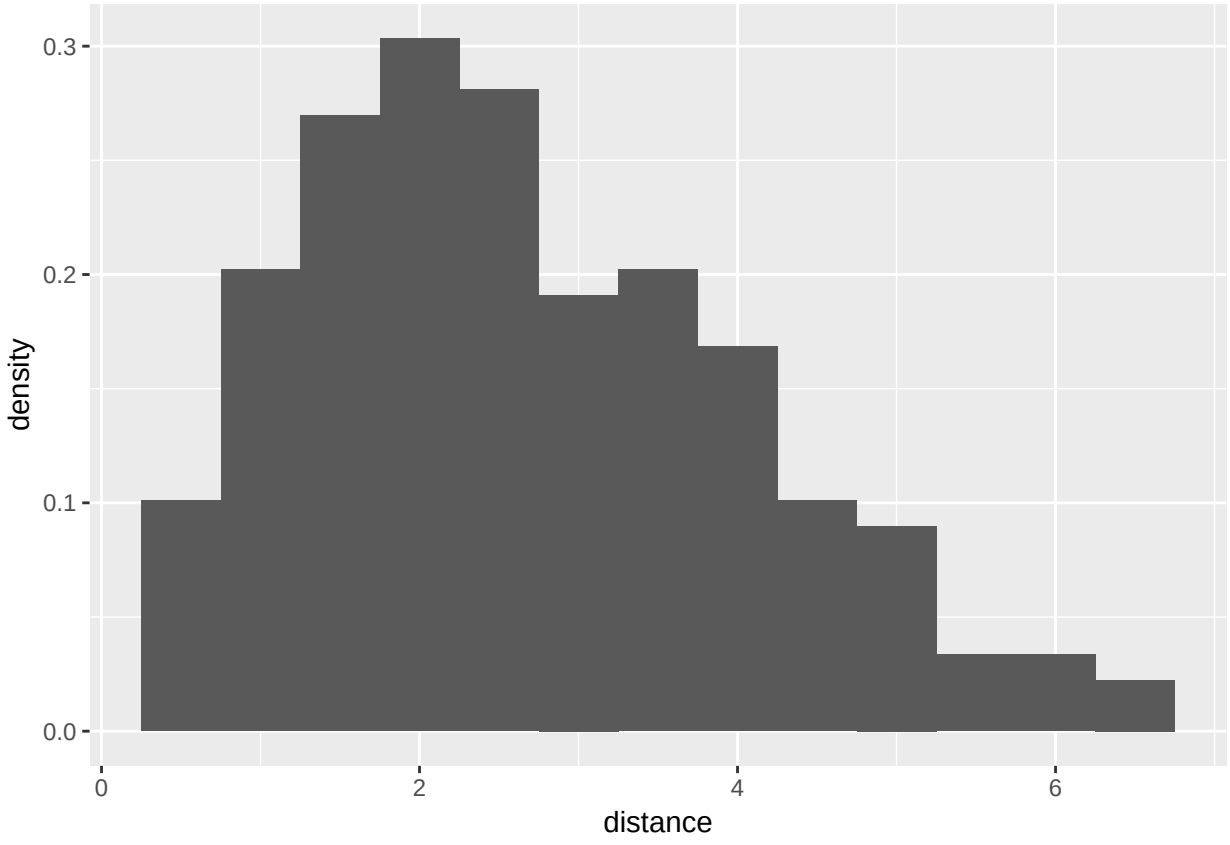
```
colnames(chromatin) <- "distance"
chromatin <- mutate(chromatin, distance_squared = distance^2)

head(chromatin)
```

```
## # A tibble: 6 x 2
##   distance distance_squared
##      <dbl>            <dbl>
## 1     4.26            18.1
## 2     3.81            14.5
## 3     1.83             3.35
## 4     5.2             27.0
## 5     1.62             2.62
## 6     1.94             3.76
```

```
dim(chromatin)
```

```
## [1] 178   2
```

```
ggplot(data = chromatin, mapping = aes(x = distance, y = ..density..)) +
  geom_histogram(binwidth = 0.5)
```



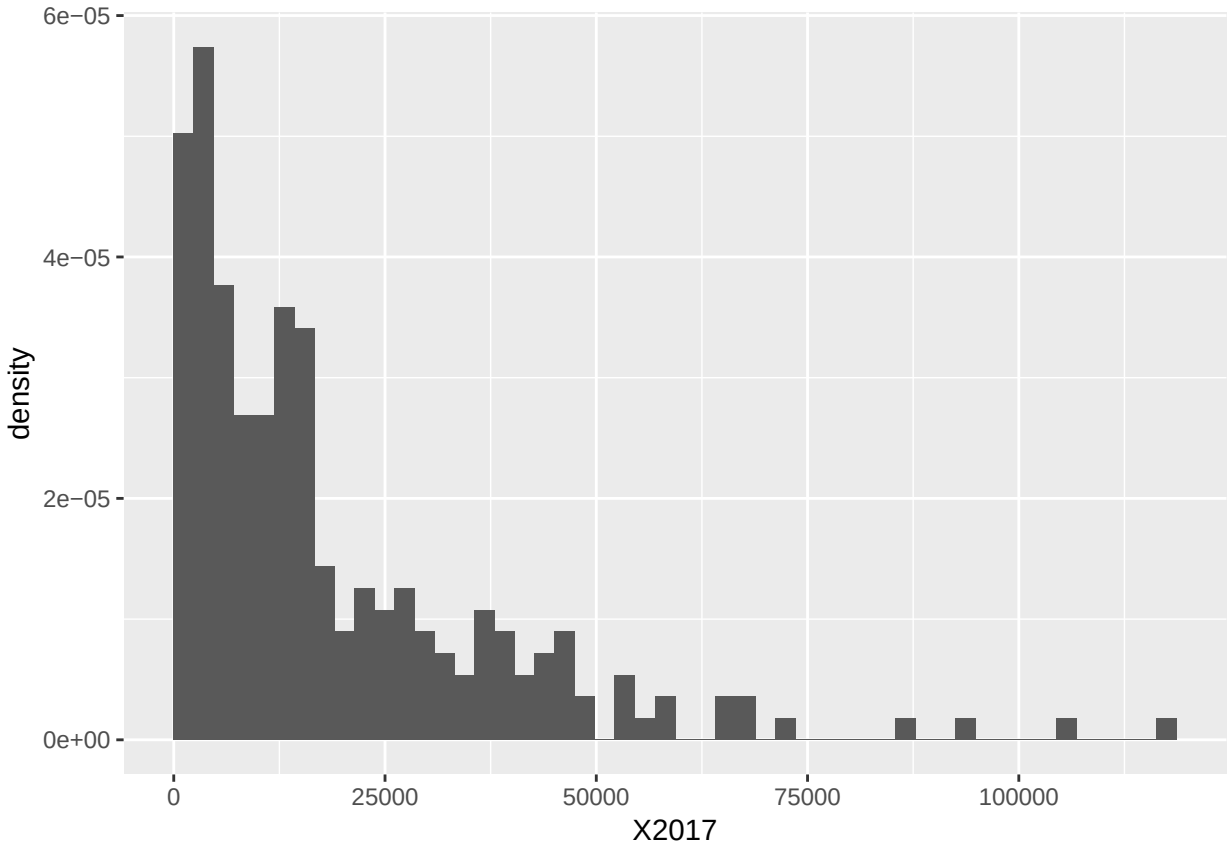
(1) Find an expression for the maximum likelihood estimator of θ based on a sample $X_1, \ldots, X_n \sim$ Rayleigh(θ).

(2) Is the maximum likelihood estimator that you found in part (1) an unbiased estimator of θ?

Problem III: Per Capita GDP

The code below reads in and plots a data set with measurements of per capita GDP at purchasing power parity as of 2017 for 235 countries, measured in inflation-adjusted 2011 international dollars; these data are from the World Bank, here: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD. Per capita GDP can be roughly interpreted as the amount of income generated in a country in one year divided by the number of people living in that country. The purchasing power parity adjustment attempts to adjust GDP to account for differences in cost of living in different countries.

```
gdp <- read.csv("http://www.evanlray.com/data/worldbank/worldbank_percap_gdp_ppp.csv")

gdp <- gdp %>%
  filter(!is.na(X2017))

ggplot(data = gdp, mapping = aes(x = X2017)) +
  geom_histogram(mapping = aes(y = ..density..), boundary = 0, bins = 50)
```



A lognormal distribution is often used to model non-negative variables that are skewed right, like incomes. In the written part of this assignment you will find the maximum likelihood estimator for the parameters of a lognormal distribution, and in the R part of the assignment you will fit the model to this data set.

For the purpose of this assignment, let's assume that the per capita GDP of different countries in a given year can be modelled as independent, identically distributed random variables (this is not actually reasonable, but may be good enough if our goal is to describe the distribution of values for per capita GDP across different countries).

Let's adopt the model $X_i \overset{\text{i.i.d.}}{\sim} \text{lognormal}(\mu, \sigma)$, $i = 1, \ldots, n$.

The pdf of a lognormal distribution is given by $f(x|\mu,\sigma) = x^{-1}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\frac{\{\log(x)-\mu\}^2}{\sigma^2}\right]$

(1) Find the maximum likelihood estimators of μ and σ. For this problem, you do not have to check second-order conditions to verify that you have found a global maximum of the log-likelihood function.