

# Homework 1

Amir Voloshin

**Due on Sunday, January 22 at 11:59 pm**

**Note:** You may work with a partner but you must writeup and submit your own assignment. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gradescope in a zip file. Include any addition files (e.g. scanned handwritten solutions) in zip file with the pdf.

## Text Analysis of JK Rowling's Harry Potter Series

### Question 1

You are interested in studying the writing style and tone used by JK Rowling (JKR for short), the author of the popular Harry Potter series. You select a random sample of chapters of size  $n$  from all of JKR's books. You are interested in the rate at which JKR uses the word *fire* in her writing, so you count how many times the word *fire* appears in each chapter in your sample,  $(y_1, \dots, y_n)$ . In this set-up,  $y_i$  is the number of times the word *fire* appeared in the  $i$ -th randomly sampled chapter. In this context, the population of interest is all chapters written by JKR and the population quantity of interest (the estimand) is the rate at which JKR uses the word *fire*. The sampling units are individual chapters. Note: this assignment is partially based on text analysis package known as [tidytext](#). You can read more about tidytext [here](#).

#### 1a.

Model: let  $Y_i$  denote the quantity that captures the number of times the word *fire* appears in the  $i$ -th chapter. As a first approximation, it is reasonable to model the number of times *fire* appears in a given chapter using a Poisson distribution. *Reminder:* Poisson distributions are for integer outcomes and useful for events that occur independently and at a constant rate. Let's assume that the quantities  $Y_1, \dots, Y_n$  are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter  $\lambda$ ,

$$p(Y_i = y_i \mid \lambda) = \text{Poisson}(y_i \mid \lambda) \quad \text{for } i = 1, \dots, n.$$

Write the likelihood  $L(\lambda)$  for a generic sample of  $n$  chapters,  $(y_1, \dots, y_n)$ . Simplify as much as possible (i.e. get rid of any multiplicative constants)

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \propto \prod_{i=1}^n e^{-\lambda} \lambda^{y_i} \\ &= e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i} \end{aligned}$$

#### 1b.

Write the log-likelihood  $\ell(\lambda)$  for a generic sample of  $n$  articles,  $(y_1, \dots, y_n)$ . Simplify as much as possible. Use this to compute the maximum likelihood estimate for the rate parameter of the Poisson distribution.

log-likelihood

$$\begin{aligned}l(\lambda) &= \log(L(\lambda)) = \ln\left(\prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}\right) \\&= \sum_{i=1}^n \ln\left(\frac{e^{-\lambda} \lambda^{y_i}}{y_i!}\right) \\&= \sum_{i=1}^n [\ln(e^{-\lambda}) - \ln(y_i!) + \ln(\lambda^{y_i})] \\&= \sum_{i=1}^n [-\lambda - \ln(y_i!) + \ln(\lambda^{y_i})] \\&= -\lambda n - \sum_{i=1}^n \ln(y_i!) + \ln(\lambda) \sum_{i=1}^n y_i\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}\frac{d}{d\lambda} l(\lambda) &= \frac{d}{d\lambda} [-\lambda n - \sum_{i=1}^n \ln(y_i!) + \ln(\lambda) \sum_{i=1}^n y_i] \\&= -n + \frac{1}{\lambda} \sum_{i=1}^n y_i = 0 \\&\Rightarrow n = \frac{1}{\lambda} \sum_{i=1}^n y_i \\&\Rightarrow \lambda = \frac{\sum_{i=1}^n y_i}{n}\end{aligned}$$

From now on, we'll focus on JKR's writing style in the Harry Potter book, *The Goblet of Fire*. This book has 37 chapters. Below is the code for counting the number of times *fire* appears in each chapter of *The Goblet of Fire*. We use the `tidytext` R package which includes functions that parse large text files into word counts. The code below creates a vector of length 37 which has the number of times the word *fire* was used in that chapter (see [https://uc-r.github.io/tidy\\_text](https://uc-r.github.io/tidy_text) for more on parsing text with `tidytext`)

```
library(tidyverse)      # data manipulation & plotting
library(stringr)        # text cleaning and regular expressions
library(tidytext)       # provides additional text mining functions
library(harrypotter)    # text for the seven novels of the Harry Potter series

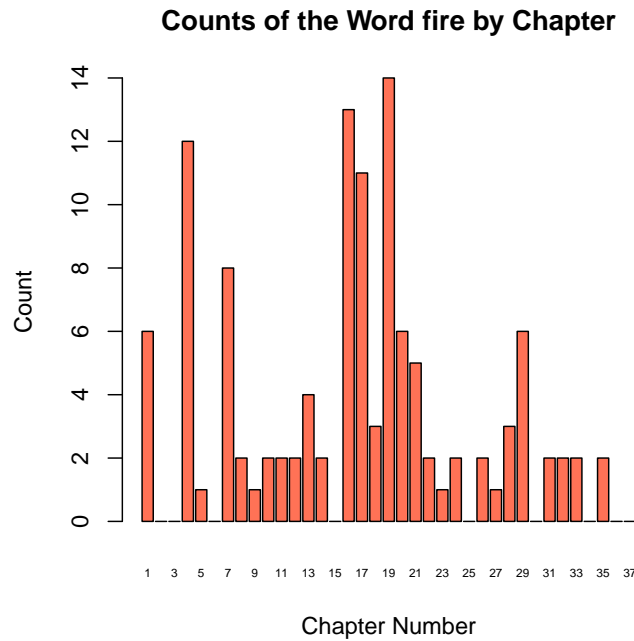
text_tb <- tibble(chapter = seq_along(goblet_of_fire),
                  text = goblet_of_fire)
tokens <- text_tb %>% unnest_tokens(word, text)
word_counts <- tokens %>% group_by(chapter) %>%
  count(word, sort = TRUE) %>% ungroup
word_counts_mat <- word_counts %>% spread(key=word, value=n, fill=0)

fire_counts <- word_counts_mat$fire
```

1c.

Make a bar plot where the heights are the counts of the word *fire* and the x-axis is the chapter.

```
# Barplot of fire
chap <- seq(1, 37, by = 1)
barplot(fire_counts, main = 'Counts of the Word fire by Chapter', xlab = 'Chapter Number', ylab = 'Count')
```



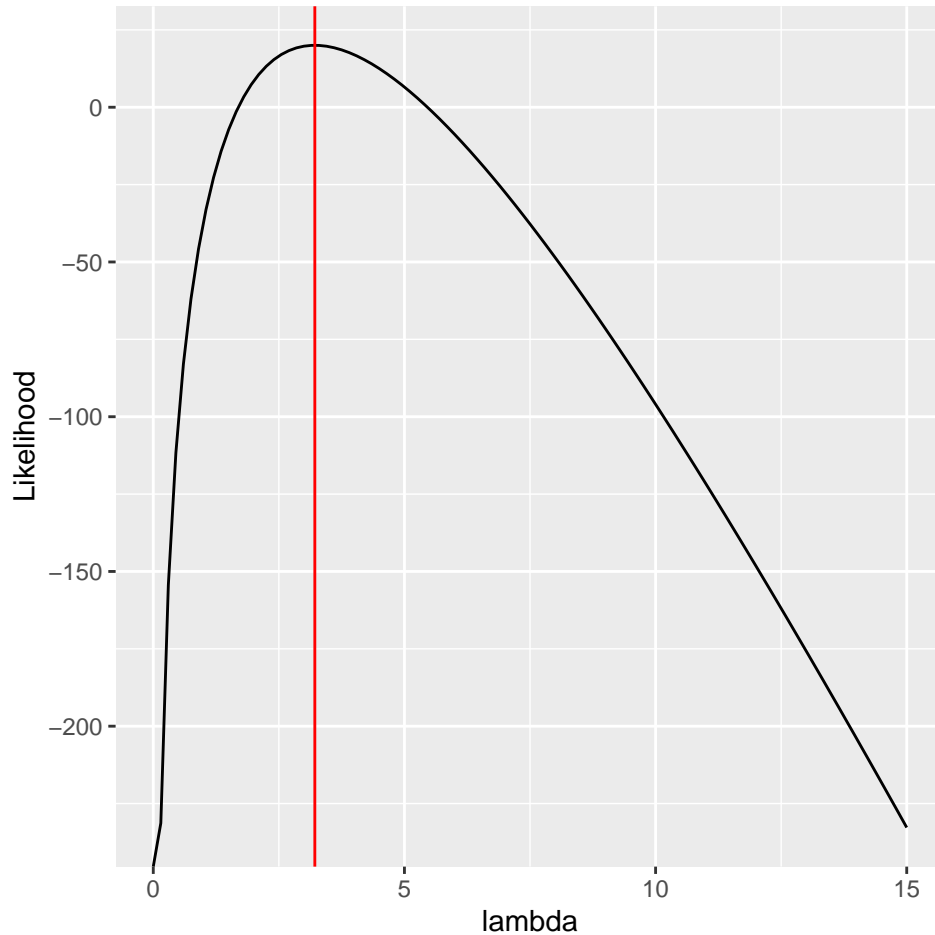
1d.

Plot the log-likelihood of the Poisson rate of *fire* usage in R using the data in `fire_counts`. Then use `fire_counts` to compute the maximum likelihood estimate of the rate of the usage of the word *fire* in The Goblet of Fire. Mark this maximum on the log-likelihood plot with a vertical line (use `abline` if you make the plot in base R or `geom_vline` if you prefer `ggplot`).

```
# log-likelihood function without multiplicative constants
y <- fire_counts
L1 <- function(lambda){
  -(lambda*length(y)) + log(lambda)*sum(y)
}

# MLE
MLE1 <- sum(fire_counts)/length(fire_counts)

# Plot
ggplot(data = data.frame(lambda = 0), mapping = aes(x=lambda)) + stat_function(fun = L1) +
  xlim(0,15) + scale_y_continuous(name = "Likelihood") + geom_vline(xintercept = MLE1, col = 'red')
```



## Question 2

For the previous problem, when computing the rate of *fire* usage, we were implicitly assuming each chapter had the same length. Remember that for  $Y_i \sim \text{Poisson}(\lambda)$ ,  $E[Y_i] = \lambda$  for each chapter, that is, the average number of occurrences of *fire* is the same in each chapter. Obviously this isn't a great assumption, since the lengths of the chapters vary; longer chapters should be more likely to have more occurrences of the word. We can augment the model by considering properties of the Poisson distribution. The Poisson is often used to express the probability of a given number of events occurring for a fixed “exposure”. As a useful example of the role of the exposure term, when counting the number of events that happen in a set length of time, we need to account for the total time that we are observing events. For this text example, the exposure is not time, but rather corresponds to the total length of the chapter.

We will again let  $(y_1, \dots, y_n)$  represent counts of the word *fire*. In addition, we now count the total number of words in each chapter  $(\nu_1, \dots, \nu_n)$  and use this as our exposure. Let  $Y_i$  denote the random variable for the counts of the word *fire* in a chapter with  $\nu_i$  words. Let's assume that the quantities  $Y_1, \dots, Y_n$  are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter  $\lambda \cdot \frac{\nu_i}{1000}$ ,

$$p(Y_i = y_i \mid \nu_i, 1000) = \text{Poisson}(y_i \mid \lambda \cdot \frac{\nu_i}{1000}) \quad \text{for } i = 1, \dots, n.$$

In the code below, `chapter_lengths` is a vector storing the length of each chapter in words.

```
chapter_lengths <- word_counts %>% group_by(chapter) %>%
  summarize(chapter_length = sum(n)) %>%
  ungroup %>% select(chapter_length) %>% unlist %>% as.numeric
```

## 2a.

What is the interpretation of the quantity  $\frac{\nu_i}{1000}$  in this model? What is the interpretation of  $\lambda$  in this model? State the units for these quantities in both of your answers.

The quantity  $\frac{\nu_i}{1000}$  is the representation of the number of words in the chapter and is used to scale the effective  $\lambda$  in the poisson distribution expression.

In this model,  $\lambda$  is the rate of the word fire per 1000 words.

## 2b.

List the known and unknown variables and constants, as described in lecture 2. Make sure your include  $Y_1, \dots, Y_n, y_1, \dots, y_n, n, \lambda$ , and  $\nu_i$ .

Observed/Known & Var > 0:

$Y_1, \dots, Y_n$

Observed/Known & Var = 0:

$y_1, \dots, y_n, n, \nu_i$

Unobserved/Unknown & Var > 0:

NA

Unobserved/Unknown & Var = 0:

$\lambda$

## 2c.

Write down the likelihood in this new model. Use this to calculate maximum likelihood estimator for  $\lambda$ . Your answer should include the  $\nu_i$ 's.

Likelihood Function

$$L(\lambda) = \frac{\prod_{i=1}^n \exp\left(\frac{-\lambda\nu_i}{1000}\right) \frac{\lambda\nu_i}{1000}^{y_i}}{y_i!} \\ \propto \prod_{i=1}^n \exp\left(\frac{-\lambda\nu_i}{1000}\right) \frac{\lambda\nu_i}{1000}^{y_i}$$

Log-Likelihood Function

$$l(\lambda) = \ln(L(\lambda)) = \ln\left(\frac{\prod_{i=1}^n \exp\left(\frac{-\lambda\nu_i}{1000}\right) \frac{\lambda\nu_i}{1000}^{y_i}}{y_i!}\right) \\ = \sum_{i=1}^n \ln\left(\frac{\exp\left(\frac{-\lambda\nu_i}{1000}\right) \frac{\lambda\nu_i}{1000}^{y_i}}{y_i!}\right) \\ = \sum_{i=1}^n \ln\left(\exp\left(\frac{-\lambda\nu_i}{1000}\right)\right) - \ln(y_i!) + \ln\left[\frac{\lambda\nu_i}{1000}^{y_i}\right] \\ = \sum_{i=1}^n \frac{-\lambda\nu_i}{1000} - \sum_{i=1}^n \ln(y_i!) + \sum_{i=1}^n \ln\left[\frac{\lambda\nu_i}{1000}^{y_i}\right] \\ = \frac{-\lambda \sum_{i=1}^n \nu_i}{1000} - \sum_{i=1}^n \ln(y_i!) + \sum_{i=1}^n \ln\left(\frac{\lambda\nu_i}{1000}\right) y_i$$

## Maximum-Likelihood Function

$$\begin{aligned}
 \frac{d}{d\lambda} l(\lambda) &= \frac{d}{d\lambda} \left[ -\frac{\lambda \sum_{i=1}^n \nu_i}{1000} - \sum_{i=1}^n \ln(y_i!) + \sum_{i=1}^n \ln\left(\frac{\lambda \nu_i}{1000}\right) y_i \right] \\
 &= \frac{-\sum_{i=1}^n \nu_i}{1000} + \sum_{i=1}^n \lambda \nu_i^{-1} \frac{\nu_i}{1000} y_i \\
 &= \frac{-\sum_{i=1}^n \nu_i}{1000} + \sum_{i=1}^n \frac{y_i}{\lambda} = 0 \\
 \Rightarrow \sum_{i=1}^n \frac{y_i}{\lambda} &= \frac{\sum_{i=1}^n \nu_i}{1000} \\
 \Rightarrow \lambda \sum_{i=1}^n \nu_i &= 1000 \sum_{i=1}^n y_i \\
 \lambda &= \frac{1000 \sum_{i=1}^n y_i}{\sum_{i=1}^n \nu_i}
 \end{aligned}$$

## 2d.

Compute the maximum likelihood estimate and save it in the variable `lambda_mle`. In 1-2 sentences interpret its meaning (make sure you include units in your answers!).

```
lambda_mle <- (1000*sum(fire_counts)) / sum(chapter_lengths)
```

```
. = ottr::check("tests/q2d.R")
```

My MLE computed above is 0.6201728. What this MLE means is that for my data above, the number appearances of the word *fire* by chapter is distributed as  $Poisson(\lambda \frac{\nu_i}{1000})$ . In other words, we get a different Poisson distribution for each chapter, where the common part for all the distributions is  $\lambda$  which is the rate of the word *fire* per thousand words.

## 2e.

Plot the log-likelihood from the previous question in R using the data from on the frequency of *fire* and the chapter lengths. Add a vertical line at the value of `lambda_mle` to indicate the maximum likelihood.

```
y <- fire_counts
v <- chapter_lengths
```

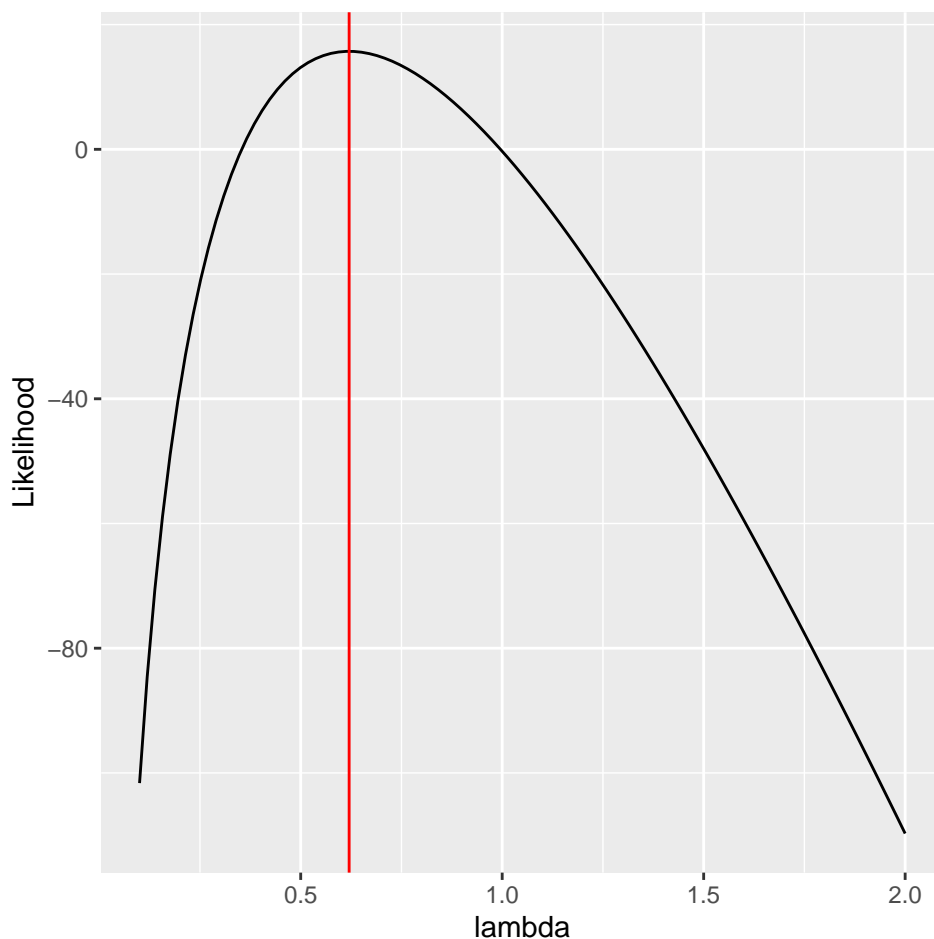
```
# Defining log likelihood function
```

```
L2_help <- function(z) {
  log(z*v/1000)*y
}
```

```
L2 <- function(lambda) {
```

```
  -(lambda/1000)*sum(v) + colSums(sapply(lambda, L2_help)) # Using colSums() and sapply() because using
}
```

```
ggplot(data = data.frame(lambda = 0.000001), mapping = aes(x=lambda)) + stat_function(fun = L2) +
  xlim(0.1, 2) + scale_y_continuous(name = "Likelihood") + geom_vline(xintercept = lambda_mle, col =
```



### Question 3

Correcting for chapter lengths is clearly an improvement, but we're still assuming that JKR uses the word *fire* at the same rate in all chapters. In this problem we'll explore this assumption in more detail.

#### 3a.

Why might it be unreasonable to assume that the rate of *fire* usage is the same in all chapters? Comment in a few sentences.

It would be unreasonable to assume the the rate of fire usage is the same in all chapters because usage of the word fire will vary based on the subject of the chapter. In a chapter where the subject is spells, disasters, problems, etc. there will likely be more appearances of the word fire. While in other chapters the word fire could be totally irrelevant and wont appear.

#### 3b.

We can use simulation to check our Poisson model, and in particular the assumption that the rate of *fire* usage is the same in all chapters. Generate simulated counts of the word *fire* by sampling counts from a Poisson distribution with the rate  $(\hat{\lambda}_{MLE}\nu_i)/1000$  for each chapter  $i$ .  $\hat{\lambda}_{MLE}$  is the maximum likelihood estimate computing in 2d. Store the vector of these values for each chapter in a variable of length 37 called `lambda_chapter`. Make a side by side plot of the observed counts and simulated counts and note any similarities or differences (we've already created the observed histogram for you). Are there any outliers in the observed data that don't seem to be reflected in the data simulated under our model?

```

observed_histogram <- ggplot(word_counts_mat) + geom_histogram(aes(x=fire), col = "lightblue", fill = 'lightblue',
  xlim(c(0, 25)) + ylim(c(0,10)) + ggtitle("Observed")

lambda_chapter <- (lambda_mle*chapter_lengths)/1000

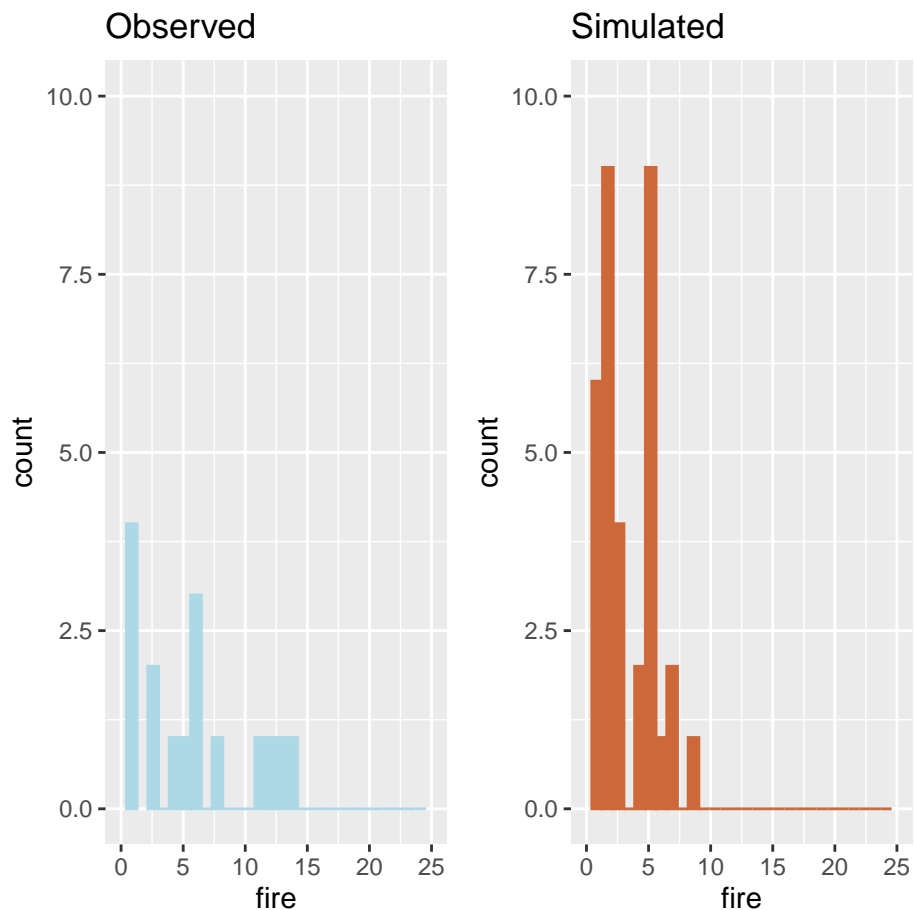
simulated_counts <- tibble(fire = rpois(37, lambda_chapter))

simulated_histogram <- ggplot(simulated_counts) + geom_histogram(aes(x=fire), col = 'sienna3', fill = 'sienna3',
  xlim(c(0, 25)) + ylim(c(0,10)) + ggtitle("Simulated")

## This uses the patchwork library to put the two plots side by side
observed_histogram + simulated_histogram

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

```



```

. = ottr::check("tests/q3b.R")

```

```

##
## All tests passed!

```

The two histograms are similar in the sense that they both have a lot of observations in chapters 1-10, but different in the later chapters. The observed data has a few counts in chapters 11-14 while the simulated counts have no counts after chapter 8. So the counts in chapters 11-14 of the observed data are outliers in our model.



3c. Assume the word usage rate varies by chapter, that is,

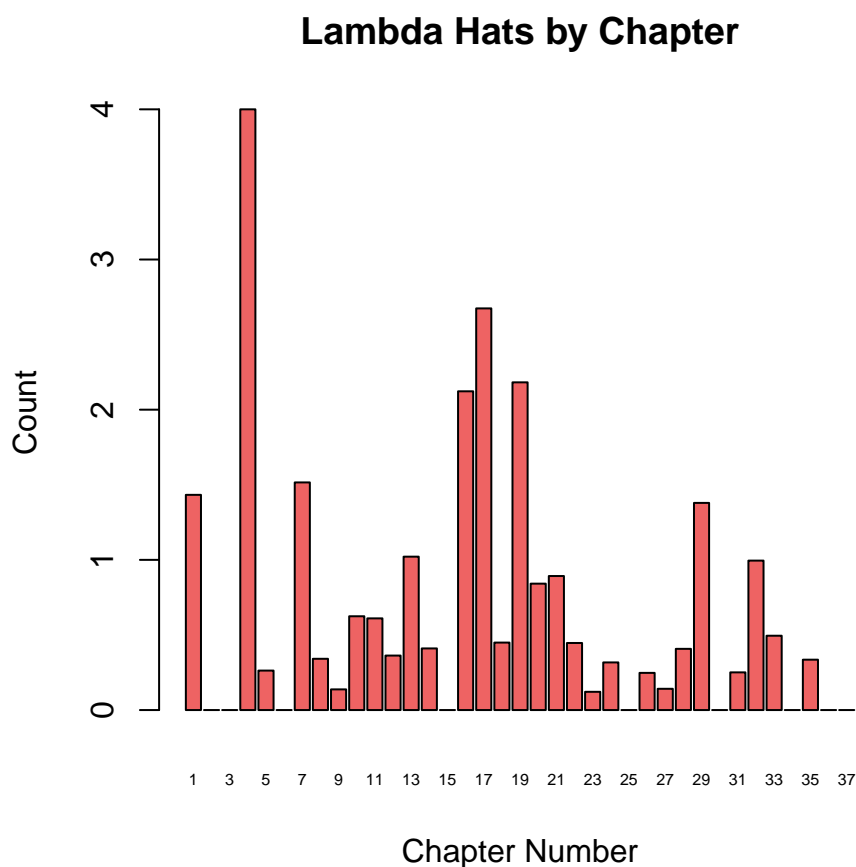
$$p(Y_i = y_i \mid \lambda, \nu_i, 1000) = \text{Poisson}(y_i \mid \lambda_i \cdot \frac{\nu_i}{1000}) \quad \text{for } i = 1, \dots, n.$$

Compute a separate maximum likelihood estimate of the rate of *fire* usage (per 1000 words) in each chapter,  $\hat{\lambda}_i$ . Make a bar plot of  $\hat{\lambda}_i$  by chapter. Save the chapter-specific MLE in a vector of length 37 called `lambda_hats`. Which chapter has the highest rate of usage of the word *fire*? Save the chapter number in a variable called `most_fiery_chapter`.

```
# Maximum likelihood estimate
lambda_hats <- 1000*fire_counts / chapter_lengths

# Most fiery chapter
most_fiery_chapter <- which.max(lambda_hats)

# Make a bar plot of the MLEs, lambda_hats
barplot(lambda_hats, main = 'Lambda Hats by Chapter', xlab = 'Chapter Number', ylab = 'Count', col = 'red')
```



```
. = ottr::check("tests/q3c.R")
```

```
##
## All tests passed!
```

## Question 4

Let's go back to our original model for usage rates of the word *fire*. You collect a random sample of book chapters penned by JKR and count how many times she uses the word *fire* in each of the chapter in your

sample,  $(y_1, \dots, y_n)$ . In this set-up,  $y_i$  is the number of times the word *fire* appeared in the  $i$ -th chapter, as before. However, we will no longer assume that the rate of use of the word *fire* is the same in every chapter. Rather, we'll assume JKR uses the word *fire* at different rates  $\lambda_i$  in each chapter. Naturally, this makes sense, since different chapters have different themes and tone. To do this, we'll further assume that the rate of word usage  $\lambda_i$  itself, is distributed according to a  $\text{Gamma}(\alpha, \beta)$  with known parameters  $\alpha$  and  $\beta$ ,

$$f(\Lambda = \lambda_i \mid \alpha, \beta) = \text{Gamma}(\lambda_i \mid \alpha, \beta).$$

and that  $Y_i \sim \text{Pois}(\lambda_i)$  as in problem 1. For now we will ignore any exposure parameters,  $\nu_i$ . Note: this is a “warm up” to Bayesian inference, where it is standard to treat parameters as random variables and specify distributions for those parameters.

#### 4a.

Write out the the data generating process for the above model.

```
for (i in 1:37)
- Generate  $\lambda_i$  from  $\text{Gamma}(\alpha, \beta)$ 
return  $\lambda = (\lambda_1, \dots, \lambda_{37})$ 
```

#### 4b.

In R simulate 1000 values from the above data generating process, assume  $\alpha = 10$  (shape parameter of `rgamma`) and  $\beta = 1$  (rate parameter of `rgamma`). Store the value in a vector of length 1000 called `counts`. Compute the empirical mean and variance of values you generated. For a Poisson distribution, the mean and the variance are the same. In the following distribution is the variance greater than the mean (called `overdispersed`) or is the variance less than the mean (underdispersed)? Intuitively, why does this make sense?

```
## Store simulated data in a vector of length 1000
counts <- rep(NA, 1000)
```

```
for (i in 1:1000){
  lambda_i <- rgamma(1, 10, 1)
  counts[i] <- lambda_i
}
```

```
print(mean(counts))
```

```
## [1] 10.12978
```

```
print(var(counts))
```

```
## [1] 10.42634
```

The mean and variance are very similar to each other, since I am using `rgamma`, which randomly simulates the observations, the mean and variance aren't always the same. Some of the time the mean is greater than variance, and other times vice-versa. But they are always very close to each other, and this is very intuitive because we are observing a rate. In theory, if something is occurring at a consistent rate, the mean and variance should be the same, hence a poisson distribution. In other words, the mean is how often the event occurs, and the variance is the distance between those events(which should be how often it occurs), so it makes a lot of sense that the mean and variance are very close to each other.

```
. = ottr::check("tests/q4b.R")
```

```
##
```

```
## All tests passed!
```

**4c.**

List the known and unknown variables and constants as described in lecture 2. Make sure your table includes  $Y_1, \dots, Y_n$ ,  $y_1, \dots, y_n$ ,  $n$ ,  $\alpha$ , and  $\beta$ .

Observed/Known & Var  $> 0$ :

$Y_1, \dots, Y_n$ ,  $\lambda$

Observed/Known & Var  $= 0$ :

$y_1, \dots, y_n$ ,  $n$ ,  $\alpha$ ,  $\beta$

Unobserved/Unknown & Var  $> 0$ :

$\lambda$

Unobserved/Unknown & Var  $= 0$ :

NA