Homework 1

PSTAT 115, Fall 2020

Due on Sunday, October 18, 2020 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. 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 *dark* in her writing, so you count how many times the word *dark* appears in each chapter in your sample, $(y_1, ..., y_n)$. In this set-up, y_i is the number of times the word *dark* appeared in the i-th randomly sampled chapter. In this context, the population of interest is all chapters written by JRK and the population quantity of interest (the estimand) is the rate at which JKR uses the word *dark*. 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 *dark* appears in the i-th chapter. As a first approximation, it is reasonable to model the number of times *dark* 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 $Y_1, ...Y_n$ are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter λ,

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

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

$$\mathcal{L}(\lambda) = \mathbb{P}(Y_i = y_i \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

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

1b.

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

$$l(\lambda) = \log(\mathcal{L}(\lambda)) = -\lambda n + \sum_{i=1}^{n} y_i \log(\lambda)$$

By taking the partial derivative in respect to λ and equate it to 0:

$$\frac{\mathrm{d}l(\lambda)}{\mathrm{d}\lambda} = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} = 0$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

So the maximum likelihood estimate for the parameter is $\bar{y}$.

From now on, we'll focus on JKR's writing style in the last Harry Potter book, *The Deathly Hallows*. This book has 37 chapters. Below is the code for counting the number of times *dark* appears in each chapter of *The Deathly Hallows* . 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 *dark* was used in that chapter (see https://uc-r.github.io/tidy_text for more on parsing text with `tidytext`)

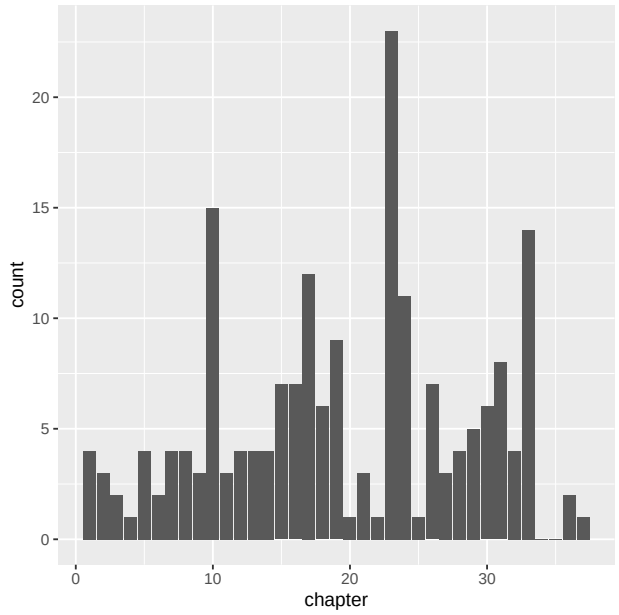
```
library(tidyverse)      # data manipulation & plotting
library(stringr)        # text cleaning and regular expressions
library(tidytext)       # provides additional text mining functions
library(harrypotter)    # text for the seven novels of the Harry Potter series

text_tb <- tibble(chapter = seq_along(deathly_hallows),
                  text = deathly_hallows)
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)

dark_counts <- word_counts_mat$dark

text_tb <- tibble(chapter = seq_along(deathly_hallows),
                  text = deathly_hallows)
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)
```

1c.

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

```
df_darkCounts = data.frame(count = dark_counts,
                           chapter = text_tb$chapter)


df_darkCounts %>%
  ggplot(aes(x = chapter, y = count)) +
  geom_histogram(stat = 'identity', binwidth = 0.5)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



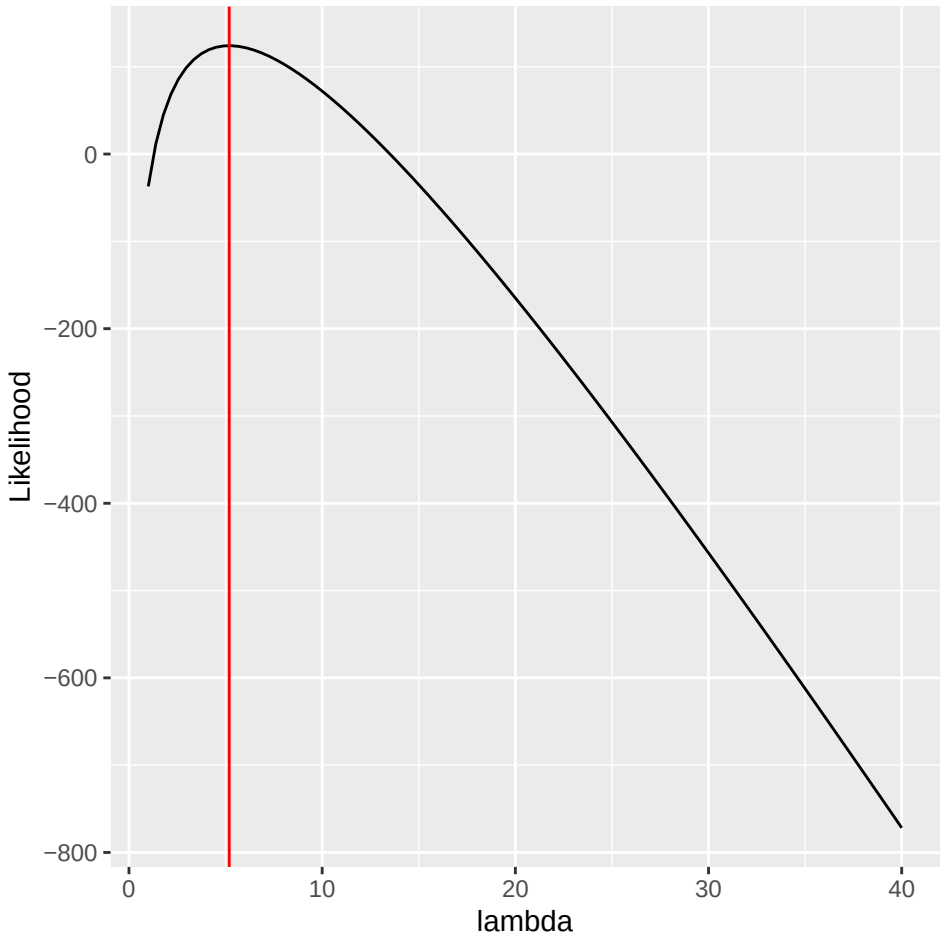
1d.

Plot the log-likelihood of the Poisson rate of *dark* usage in R using the data in `dark_counts`. Then use `dark_counts` to compute the maximum likelihood estimate of the rate of the usage of the word *dark* in The Deathly Hallows. 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`).

```
n = 37 # the lenth of charpter

L <- function(lambda){ #Log-Likelihood Function
  -lambda*n+sum(dark_counts)*log(lambda)
  #+rep(sum(log(factorial(dark_counts))))
}

lambda_mle = mean(dark_counts)
lambda_mle
```

```
## [1] 5.189189
```

```
df_logLikelihood = data.frame(lambda = mean(dark_counts))

ggplot(data = df_logLikelihood,mapping = aes(x=lambda)) +
  stat_function(fun = L)+
  xlim(1,40)+scale_y_continuous(name = "Likelihood")+
  geom_vline(xintercept = lambda_mle, color = 'red')
```



Question 2

For the previous problem, when computing the rate of *dark* 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 *dark* 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 then number of events that happen in a set length of time, we to 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 $(y_1, ..., y_n)$ represent counts of the word *dark*. In addition, we now count the total number of words in each each chapter $(\nu_1, ..., \nu_n)$ and use this as our exposure. Let Y_i denote the random variable for the counts of the word *dark* in a chapter with ν_i words. Let's assume that the quantities $Y_1, ...Y_n$ are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter $\lambda \cdot \frac{\nu_i}{1000}$,

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

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
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

2a.

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

ν_i means the total numbers of words in each chapter, $\frac{\nu_i}{1000}$ meansper 1000 words.

λ is the average rate of the word *dark* appears in a given chapter, and the unit for λ is *dark* per chapter

2b.

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

Known, Var > 0: $Y_1, ..., Y_n$

Known, Var $= 0$: $y_1, ..., y_n$, n, ν_i

Unknown, Var > 0:

Unknown, Var $= 0$:, λ

2c.

Write down the likelihood in this new model. Use this to calculate maximum likelihood estimator for λ. Your answer should include the ν_i's.

Suppose $\beta_i = \frac{\lambda v_i}{1000}$, then:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \frac{e^{-\beta_i}\beta_i^{y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^{n}\beta_i}\prod_{i=1}^{n}\beta_i^{\sum_{i=1}^{n}y_i}}{\prod_{i=1}^{n}y_i!} \propto e^{-\sum_{i=1}^{n}\beta_i}\prod_{i=1}^{n}\beta_i^{\sum_{i=1}^{n}y_i}$$

Substitue back in $\frac{\lambda v_i}{1000}$ for β_i:

$$\mathcal{L}(\lambda) \propto e^{-\sum_{i=1}^{n}\frac{\lambda v_i}{1000}}\prod_{i=1}^{n}\frac{\lambda v_i}{1000}^{\sum_{i=1}^{n}y_i}$$

By taking the partial derivative in respect to λ and equate it to 0:

$$\frac{\mathrm{d}l(\lambda)}{\mathrm{d}\lambda} = -\sum_{i=1}^{n}\frac{v_i}{1000} + \frac{\sum_{i=1}^{n}y_i}{\lambda} = 0$$

$$\hat{\lambda} = 1000\frac{\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}v_i}$$

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!).

```
# YOUR CODE HERE
lambda_mle = sum(dark_counts)/sum(chapter_lengths)*1000
lambda_mle
```

```
## [1] 0.9652801
```

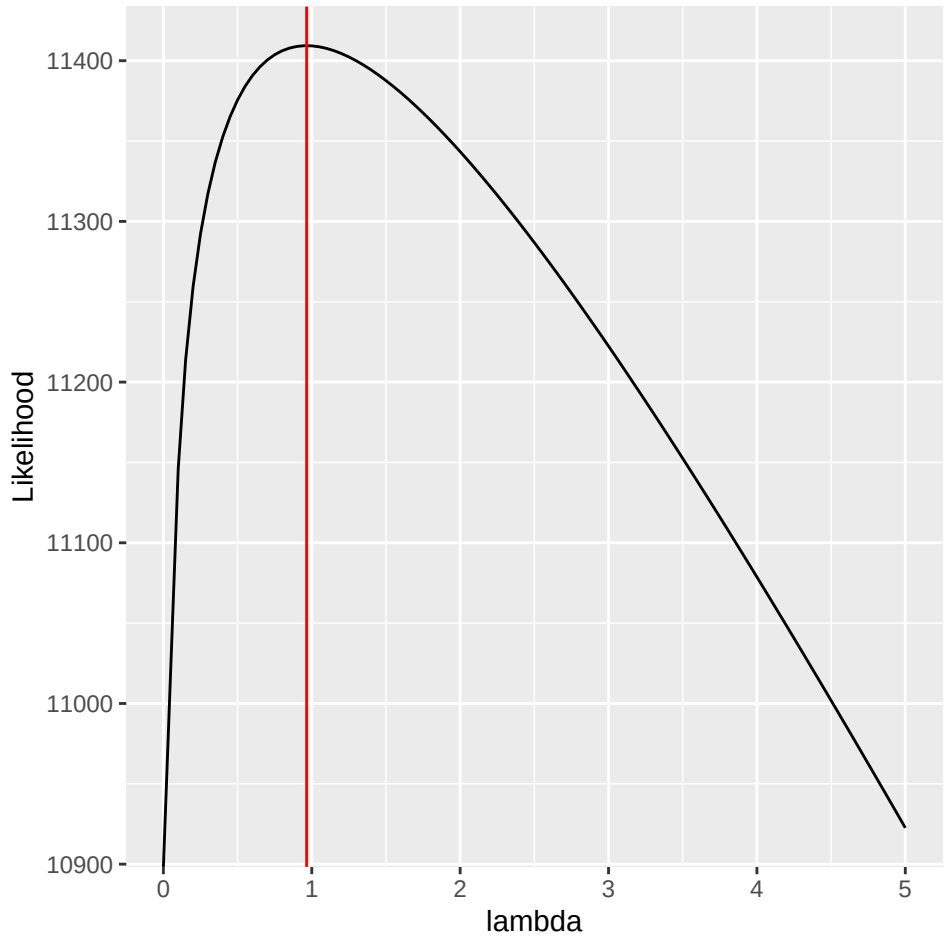
```
. = ottr::check("tests/q2d.R")
```

The maximum likelihood estimate $\hat{\lambda}$ 0.9652801, in this scenario can be interpreted as the best estimate for the rate at which the word *dark* occurs in each chapter every 1000 words is once.

2e.

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

```
y=dark_counts
vi=chapter_lengths
L = function(lambda){#Log-Likelihood Function
  lambda * -sum((vi)/1000)+
  sum(y)*log(lambda*prod(vi/1000))
}


df_logLikelihood2 = data.frame(lambda = lambda_mle)

ggplot(data = df_logLikelihood2, mapping = aes(x=lambda))+
  stat_function(fun = L)+
  xlim(0,5)+
  scale_y_continuous(name = "Likelihood")+
  geom_vline(xintercept = lambda_mle, color = 'red')
```



Question 3

Correcting for chapter lengths is clearly an improvement, but we're still assuming that JKR uses the word *dark* 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 *dark* usage is the same in all chapters? Comment in a few sentences.

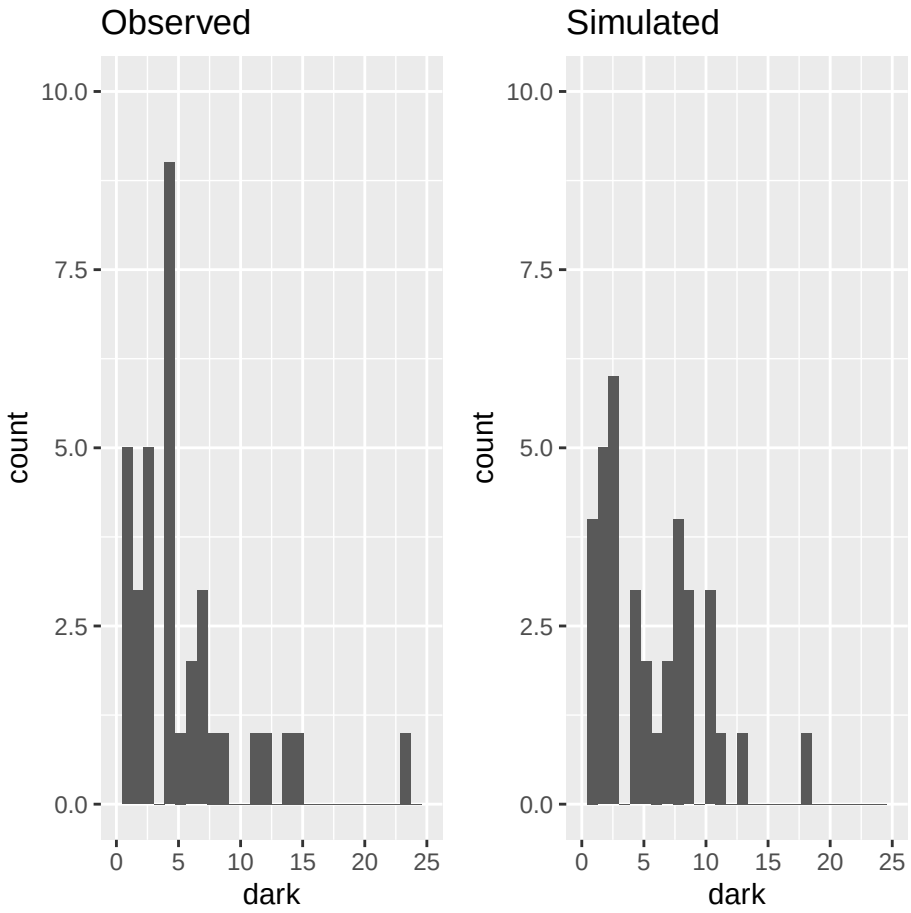
This is probably not a good assumption since different chapters contain different elements that make up the overall plot, for example in some postive chapter may not use *dard* at all, but some chapter scenes relavant with dark, then the rate of *dark* will appear more often.

3b.

We can use simulation to check our Poisson model, and in particular the assumption that the rate of *dark* usage is the same in all chapters. Generate simulated counts of the word *dark* by sampling counts from a Poisson distribution with the rate $(\hat{\lambda}_{\text{MLE}}\nu_i)/1000$ for each chapter i. $\hat{\lambda}_{\text{MLE}}$ 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=dark)) +xlim(c(0, 25)) + ylim(c(0,


lambda_chapter <- rpois(37, (lambda_mle*chapter_lengths/1000))

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

simulated_histogram <- ggplot(simulated_counts) + geom_histogram(aes(x=dark)) +xlim(c(0, 25)) + ylim(c(

## 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`.
## Warning: Removed 2 rows containing missing values (geom_bar).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 2 rows containing missing values (geom_bar).
```




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

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

In observed histogram, there are few outliers, while most of data are clustered in simulated data, and the

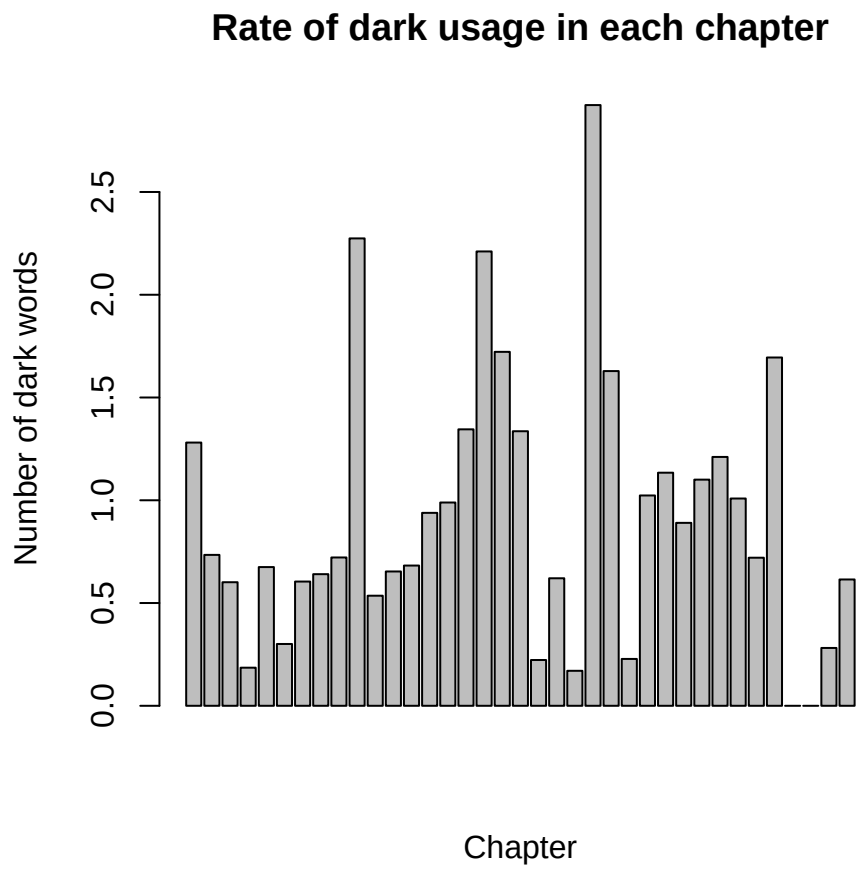
overall range of observed data is larger than simulated data.

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

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

Compute a separate maximum likelihood estimate of the rate of *dark* 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 dark? Save the chapter number in a variable called `darkest_chapter`.

```
# Maximum likelihood estimate
lambda_hats <- (1000*dark_counts)/(chapter_lengths)

darkest_chapter <- 23

# Make a bar plot of the MLEs, lambda_hats
barplot(lambda_hats,
        main = "Rate of dark usage in each chapter",
        ylab = "Number of dark words",
        xlab = "Chapter")
```




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

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

Question 4

Let's go back to our original model for usage rates of the word *dark*. You collect a random sample of book chapters penned by JKR and count how many times she uses the word *dark* in each of the chapter in your sample, $(y_1, ..., y_n)$. In this set-up, y_i is the number of times the word *dark* appeared in the i-th chapter, as before. However, we will no longer assume that the rate of use of the word *dark* is the same in every chapter. Rather, we'll assume JKR uses the word *dark* at different rates λ_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 λ_i itself, is distributed according to a Gamma(α, β) with known parameters α and β,

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

and that $Y_i \sim \text{Pois}(\lambda_i)$ as in problem 1. For now we will ignore any exposure parameters, ν_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.

The rate of word usage λ_i is distributed according to Gamma(α, β), so based on the Gamma distrbution, we can generate λ_i. y_i is the number of times the word dark appeared in the i-th chapter, it is distributed according to Poisson(λ_i). We can use the λ_i from Gamma(α, β) to get Y_i.

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 `overdispsersed''`) `or is the variance less than the mean` (underdispersed'')? Intuitively, why does this make sense?

```
## Store simulated data in a vector of length 1000
lambda_dist <- rgamma(1000, shape=10,scale=1)
counts <- rpois(1000, lambda_dist)

print(mean(counts))
```

```
## [1] 10.197
```

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

```
## [1] 21.82001
```

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

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

The variance greater than the mean, so it is overdispsersed. Since λ is a random variable in this case, so the variance will be greater.

4c.

List the known and unknown variables and constants as described in lecture 2. Make sure your table includes $Y_1, ..., Y_n$, $y_1, ..., y_n$, n, λ, α, and β.

Known, Var > 0: $Y_1, ..., Y_n$

Known, Var $= 0$: $y_1, ..., y_n$, α, and β

Unknown, Var > 0: λ

Unknown, Var $= 0$:

Extra Credit.

Compute $p(Y_i \mid \alpha, \beta) = \int p(Y_i, \lambda_i \mid \alpha, \beta) d\lambda_i$. *Hint:* The gamma function is defined as $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$.

$$
\begin{aligned}
p(Y_i \mid \alpha, \beta) &= \int_0^\infty p(Y_i, \lambda_i \mid \alpha, \beta) d\lambda_i \\
&= \int_0^\infty p(Y_i \mid \lambda_i) p(\lambda_i \mid \alpha, \beta) d\lambda_i \\
&= \int_0^\infty \frac{-e^{-\lambda_i} \lambda_i^{y_i}}{y_i} \frac{\lambda_i^{\alpha-1} \beta^\alpha e^{-\beta\lambda_i}}{\Gamma(\alpha)} d\lambda_i \\
&= \int_0^\infty \frac{-\lambda_i^{y_i+\alpha-1} e^{-(1+\beta)\lambda_i}}{y_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} d\lambda_i \\
&= \int_0^\infty \frac{-\lambda_i^{y_i+\alpha-1} e^{-(1+\beta)\lambda_i} (\beta+1)^{(a+y_i)}}{\Gamma(y_i+\alpha)} d\lambda_i \frac{\Gamma(y_i+a)}{y_i!(\beta+1)^{(a+y_i)}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{\Gamma(y_i+\alpha)}{y_i!(\beta+1)^{a+y_i}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{\Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)} \cdot \frac{\beta^\alpha}{(\beta+1)^(a+y_i)} \\
&= \binom{\alpha+y_i-1}{y_i} (\frac{\beta}{\beta+1})^\alpha (\frac{1}{\beta+1})^{y_i}, i = 1, 2, 3, ..., n
\end{aligned}
$$

Hence is it is

$$NB(\alpha, \frac{\beta}{1+\beta})$$

Fill in the blank.

You just showed that a Gamma mixture of Poisson distributions is a ___ negative binomial distribution_$(\alpha, \frac{\beta}{1+\beta})$_____.