Homework 1

Your name here

PSTAT 115 (F25)

Assignment due on Sunday, October 12 at 11:59 pm"

Note: You may work with a partner but you must writeup and submit your own assignment. Submit your Quarto file (.qmd) and the compiled pdf on Gradescope in a zip file.

Text Analysis of JRR Tolkien's "The Lord of the Rings' ' '

Question 1

You are interested in studying the writing style and tone used by JRR Tolkein, the author of the the Lord of the Rings books. You select a random sample of chapters of size n from all of Tolkien's books. You are interested in the rate at which Tolkien uses a particular word in his writing, so you count how many times the word appears in each chapter in your sample, $(y_1, ..., y_n)$. As an example, consider the word "fire' '. 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 Tolkien and the population quantity of interest (the estimand) is the rate at which Tolkien 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 $Y_1, ...Y_n$ 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)

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

We'll focus on Tolkien's writing style in the first novel, the Lord of the Rings (LOTR). The novel is split into two "books" with a total of 23 chapters between them. Below is the code for counting the number of times *fire* appears in each chapter of LOTR . 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 23 which has the number of times the word *fire* 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

LOTR <- readRDS("LOTR.RDS") ## read in the text of

stop_words <- tidytext::stop_words
tokens <- LOTR %>% unnest_tokens(word, text)
tokens <- tokens %>% dplyr::filter(!word %in% stop_words$word)

all_word_counts <- tokens %>% group_by(book, chapter) %>%
  count(word, sort = TRUE) %>% ungroup
all_word_counts_mat <- all_word_counts %>% spread(key=word, value=n, fill=0)

word_counts <- all_word_counts_mat[["fire"]]
```

1c.

Change the word "fire" in the code above to another word of your choosing, so that `word_counts` represents the counts of that word in each chapter. The word you choose must appear at least 50 times total in LOTR. Make a bar plot where the heights are the counts of the word you choose and the x-axis is the chapter.

```
tibble(chapter=1:length(word_counts), word=word_counts) %>%  ggplot() + geom_col(aes(x=chapte
```

```
. = ottr::check("tests/q1c.R")
```

1d.

Plot the log-likelihood of the Poisson rate of usage of the word in `word_counts`. Then use `word_counts` to compute the maximum likelihood estimate of the rate of the usage of the word in the LOTR. 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`).

Type your answer here, replacing this text.

```
# YOUR CODE HERE
```

Question 2

For the previous problem, when computing the rate of word 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 your chosen word 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 your chosen word. In addition, we now count the total number of all 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 your chosen word 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 $\lambda \cdot \frac{\nu_i}{1000}$,

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

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

```
chapter_lengths <- all_word_counts %>% group_by(book, 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 λ in this model? State the units for these quantities in both of your answers.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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 <- NULL # YOUR CODE HERE
```

Type your answer here, replacing this text.

2e.

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

```
# YOUR CODE HERE
```

Question 3

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

Type your answer here, replacing this text.

3b.

We can use simulation to check our Poisson model, and in particular the assumption that the rate of the word usage is the same in all chapters. Generate simulated counts of your word 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 23 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 counts in the observed data that seem very different from the data simulated under our model? Note: simulated data from the proposed model is a useful way of doing model checking (called "posterior predictive model checking") that we'll discuss later in the quarter.

```
observed_histogram <- tibble(word=word_counts, chapter=1:length(word_counts)) %>% ggplot() +
  xlim(c(0, max(word_counts)+1)) + ggtitle("Observed")


lambda_chapter <- NULL # YOUR CODE HERE
```

```
simulated_counts <- tibble(word = rpois(length(word_counts),
                           lambda=lambda_chapter))

simulated_histogram <- NULL # YOUR CODE HERE

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

Type your answer here, replacing this text.

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

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 word usage (per 1000 words) in each chapter, $\widehat{\lambda}_i$. Make a bar plot of $\widehat{\lambda}_i$ by chapter. Save the chapter-specific MLE in a vector of length 23 called `lambda_hats`. Which chapter has the highest rate of usage of the word you chose? Save the chapter number in a variable called `highest_rate_chapter`.

```
# Maximum likelihood estimate
lambda_hats <- NULL # YOUR CODE HERE

highest_rate_chapter <- NULL # YOUR CODE HERE

# Make a bar plot of the MLEs, lambda_hats
# YOUR CODE HERE
```

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

Question 4

Let's go back to our original model for usage rates of the word. You collect a random sample of book chapters penned by Tolkien and count how many times he uses the word 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 appeared in the i-th chapter, as before. However, we will no longer assume that the rate of use of the word is the same in every chapter. Rather, we'll assume Tolkien uses the word 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.

Type your answer here, replacing this text.

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

# YOUR CODE HERE

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

```
. = ottr::check("tests/q4bp.R")
```

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.

Type your answer here, replacing this text.

Question 5

Please reflect on your use of AI tools for this assignment. If you used an AI tool (e.g., ChatGPT, Claude, Gemini etc), please specify which one and describe how you used it. Then, reflect on how it affected your ability to complete the assignment and your understanding of the material, including both the positive and negative impacts. What content might you need to revisit in order to gain a deper understanding? This is an exercise in transparency: you will get full credit if you response appears thoughtful and honest. Your answer should be about one paragraph long.

Type your answer here, replacing this text.