The Secerts of Mordern Detective Story

Gender distribution of victims, proportion of testimonial clues before and after 1905, average age that authors made their first publications.

Group YEL: Eric Li, Lisa Cheng, Yinuo Zhao

December 3, 2021

Introduction

Include here a few sentences to introduce the problem and provide context. You might want to briefly summarize the data in words (what is the data and what is it used for). You can present the questions you are investigating here.

Data Summary

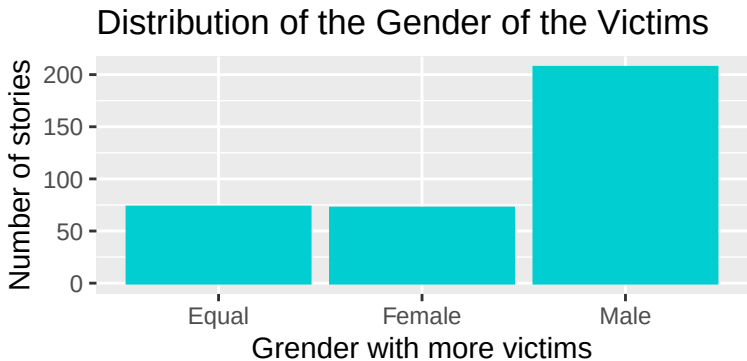
Each of our research questions require specific data wrangling in order for them to be carry out as a statistical research. However, there are several steps of which they all share in common. This includes the selection of required variables (define a new data set that contains only the few variables needed for the question), the removal of missing (NA) values, as well as the creation of all new variables that are essential for the research question.

Research Question 1 - Statement of the Question

The research question is: **Is the proportion of male victims equal to the proportion of female victims from the early 1800s to the early 1900s?**

Motivation: Unlike the current society that put general equality at a critical position for social morality, the society in the 1800s $\sim$ 1900s haven't brought this topic to people's attention. This lead to a reasonable prediction that there might be a trend of authors prefer using male characters over female characters. Instead of going down the usual direction that explores the difference between the number of male and female main character (in this case, the detective), I would like to investigate the phenomenon of gender inequality by comparing the difference in proportion of male and female victims. This allows us to further explore this topic from a different perspective other than the common way of thinking.

Research Question 1 - Relevant Visualization



Research Question 1 - Set-up of Statistical Model/Method

Data Wrangling

I first divided the data into three categories:
- The number of stories with an equal number of male and female victims
- The number of stories with more male victims
- The number of stories with more female victims

I created a new variable called "diff" and saved each row by the number of male victims equals the number of female victims as "Equal", "Male", or "Female." Since the number of male victims equals the number of female victims in the first category, I did not use that in the one-sample p test.

Statistical Method

I used the one-sample p test to capture the difference between the proportion of stories with more male victims and stories with more female victims.

Research Question 1 - Results and Interpretation

- For this test, my null hypothesis test is that the proportion of books with more male victims is equal to the proportion of books with more female victims. On the other hand, my alternative hypothesis is that the proportion of books with more male victims is not equal to that of books with more female victims.

$$H_0 : p_{\textit{stories with more male victims}} = p_{\textit{stories with more female victims}}$$

$$H_1 : p_{\textit{stories with more male victims}} \neq p_{\textit{stories with more female victims}}$$

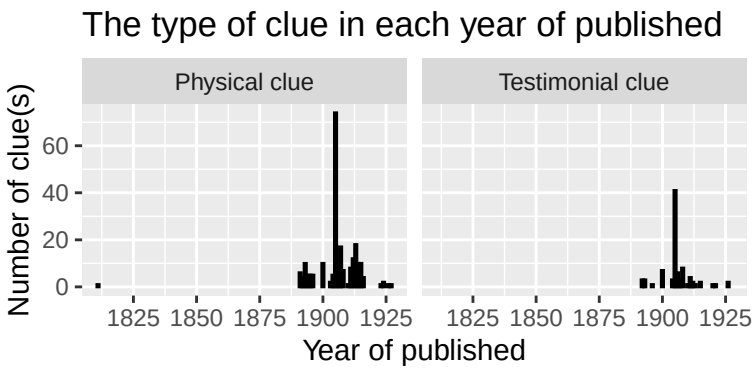
- With the p-test, I obtained the p-value(0.0012). Then by using a significant level of 0.01. We can reject the null hypothesis since the p-value is smaller than the significant level. Finally, conclude that there is a significant difference between the proportion of male victims and female victims.

Research Question 2 - Statement of the Question

The research question is: **Is the proportion of testimonial clues in detective stories published before 1905 is the same as the proportion after 1905?**

The authors did not start using essential clues in the detective stories until around 1890, so the use of clue types changed rapidly. This research will allow us to investigate if the author changed the type of important clue over time or with the change of time, how the mainstream of detective authors writing about clues has changed.

Research Question 2 - Relevant Visualization



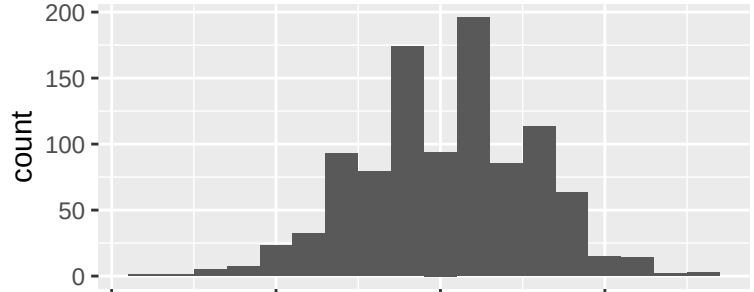
The above figure is a bar chart which clearly show the number of each clues for detective books in each year of published. We can see that physical clues range from 0 to 75, and the testimonial clue range from 0 to about 43. According to the figure, it shows that author tend to use physical clue more than testimonial clue.

Research Question 2 - Set-up of Statistical Model/Method

$$H_0 : p_{testimonial\_before\_1905} - p_{testimonial\_after\_1905} = 0$$

$$H_A : median_{testimonial\_before\_1905} - median_{testimonial\_after\_1905} \neq 0$$

```
## [1] 0.07687229
```



Research Question 2 - Results and Interpretation

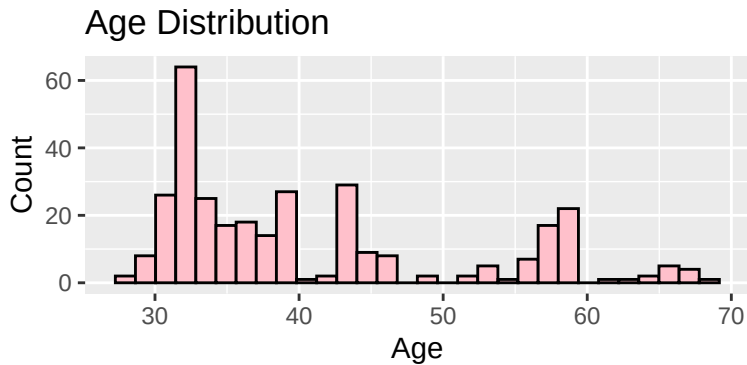
- From the result of hypothesis test, the p-value is 0.934 which is larger than 0.10 so that there is no evidence against the null hypothesis.
- This means that the difference between the proportion of testimonial clue before 1905 and the proportion of testimonial clue after 1905 is almost the same, it did not change a lot.
- Some of the reason for the result is that at about 1900s, the structure and system of the detective stories has just developed, most of the people simply imitated the writing styles of others or famous people. At that time, some mainstream writing styles had not been formed, so the type of essential clue did not change much.

Research Question 3 - Statement of the Question

The research question is: **What is the range of plausible values for the average age of authors when the detective story is first published?**

Motivation: The years between 1800 to 1900 is the era of the birth of modern detective stories, and there were no such things that existed before this period of time. Therefore, writing detective stories can be seen as a complete innovation to the wide range of existing English literature. This research question focuses on discovering who the people who took this important step to pioneer modern detective stories are. Furthermore, since age is generally an adequate predictor of education levels and the amount of life experience, we can also find some characteristics that modern detective story writers share in common.

Research Question 3 - Relevant Visualization



The above figure is a histogram that shows the distribution of the authors' ages when the story is first published. We can see that all age ranges from around 27 to around 68. It is more common for authors to publish detective stories in their 30s to 40s than later in their life. No story was pub

Research Question 3 - Set-up of Statistical Model/Method

Data Wrangling

- Selected to include the only two variables needed for this specific research question: age of the author and date of the publication of the story.
- Removed all observations that are missing any of these two variables.
- Created a new variable that represents the age of the author when the story is published. This is done by subtract these two variables.
- There exist several unreasonable values in this new variable we created(negative ages, and extremely small ages), we exclude these observations from the data since this is probably due to the mistakes during data collection.

Statistical Method

- The research question is completed using the bootstrap method. This method will give us the range of plausible values for the parameter we are trying to find, which is this case is the average age of authors when the detective story is first published.

Research Question 3 - Results and Interpretation

- From the result of bootstrapping, the 2.5% percentile is 39.59 and the 97.5% percentile is 41.91.
- This means that we are 95% confident that the average age of authors when the detective story is first published is between 39.59 to 41.91.
- If we repeated this procedure many times, 95% of those confidence intervals will include the true average age of authors when the detective story is first published for all detective stories from the early 1800s to the early 1900s.

Limitations

Even though the process of data collection was conducted with consideration to ensure that all data is mostly accurate, this accuracy is still somewhat not guaranteed based on our findings. In the third research question, we found that some detective stories were published when the author was less than one year old, or even at negative ages(these detective stories were removed for the reliability of the research). Since this is impossible, it is an evidence that there exists incorrect data in the data set. While it is possible to exclude these incorrect data for this specific research question, we cannot conclude how much data is erroneous in other variables, leading to erroneous results for other research questions.

Overall Conclusions

Brief Summary of the Results Research Question 1: Research Question 2: Research Question 3:

Overall Conclusions

Further more sdaslkdjalksjdlkansdnkjxkjjxjxjxjcjcjoji-
jjxjxxjxjxjxjxjxjjxjxjxjxjjxjxjxjxjxjxjjxjxjxjxjxjxjxjxjxjxjxj

References and Acknowledgements (optional)

The authors would like to thank "TA name" for their helpful suggestions and comments that improved the presentation of this poster.