The Creation 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

The three questions that our group tends to investigate are “Is the proportion of books with more male victims equal to the proportion of books with more female victims?”, “Is the proportion of testimonial clues in detective stories published before 1905 same as the proportion after 1905?”, and “What is the range of plausible values for the average age of authors when the detective story was first published?” The data that we used for the project is a sample for over 300 short detective stories from the early 1800s to the early 1900s. As this was an important time period for the development of detective stories, we are interested in exploring how did detective stories change over the 100 years and how were the features(victims, essential clues, and the author) of detective stories influenced by the society at that time.

Data Summary

Each of our research question used different variables in the dataset and required specific data wrangling in order for them to be carry out as statistical researches. However, there are several steps of which they all share in common. These includes:

- The selection of required variables (define a new data set that contains only the few variables needed for the question),
- The removal of observations that have missing (NA) values for the required variables,
- The creation of new variables needed for the specific research question.

The detailed data summary will be within each research question below.

Research Question 1 - Statement of the Question

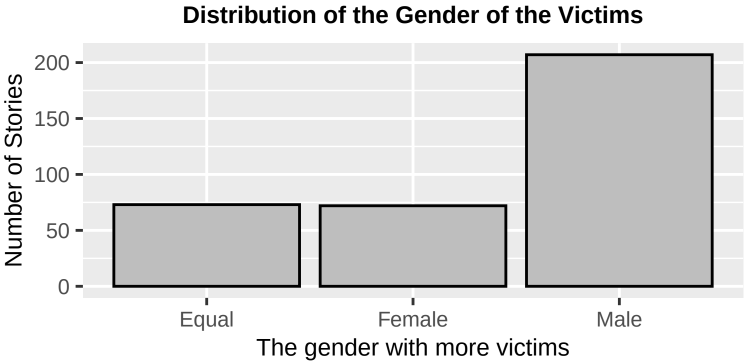
The research question is:

- Is the proportion of books with more male victims equal to the proportion of books with more female victims?

Motivation and Significance:

- Unlike the current society that puts gender 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 leads 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), we 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



The above barplot is a distribution of the dominant gender of victims for all books. We can see that while the number of books with equal number of male and female victims is similar to the number of books with more female victims, the number of books with more male victims triples the other two categories.

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

Variable Used

The number of victims of gender male, and The number of victims of gender female

Data Wrangling

- Created a new variable that divides all detective stories into three categories based on the above two variable:
    - Story with More male victims,
    - Story with more female victims, and
    - Story with equal number of male and female victims. (This is removed from the data since in this category, stories have the same number of male and female victims, which is irrelevant to our research question)

Statistical Method

This research question uses the one proportional hypothesis test. It creates a large number of simulations under our hypothesis(the null hypothesis), and checks how many of those simulations are as extreme as the value we obtained from the existing data(test statistic). It allows us to conclude whether the null hypothesis is true since if the test statistic is very rare among all simulations, the null hypothesis is likely not true. Oppositely, if most simulations result in a value close to the test statistic, then it is likely true.

Research Question 1 - Results and Interpretation

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

$$H_0 : p_{stories\ with\ more\ male\ victims} - p_{stories\ with\ more\ female\ victims} = 0$$

$$H_A : p_{stories\ with\ more\ male\ victims} - p_{stories\ with\ more\ female\ victims} \neq 0$$

- The test statistic (difference between the proportion of books with more male victims and proportion of books with more female victims, calculated using the available data) is : 0.484
- 10000 simulations were made under the assumption that our null hypothesis is true. Of all simulations, none of them were as extreme as the test statistic (p-value is 0).
- This indicates that if the null hypothesis is true, the situation as the test statistic is extremely rare, thus we have strong evidence against the null hypothesis.
- Since the null hypothesis is incorrect, we can conclude that there is a significant different between the proportion of books with more male victims and proportion of books with more female victims.

Research Question 2 - Statement of the Question

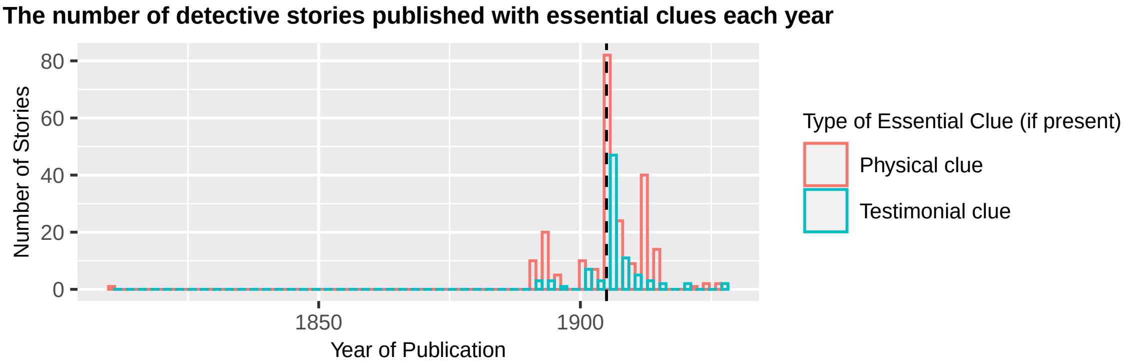
The research question is:

- Is the proportion of testimonial clues in detective stories published before 1905 same as the proportion after 1905?

Motivation and Significance:

- Based on the information given in this data set, essential clues are categorized as physical clues and testimonial clues. The authors of modern detective stories did not start giving essential clues to the readers until around 1890, therefore it was a complete new feature introduced to detective stories. Given that detective story itself was also newly introduced to the field of literature in 1800s, there were no rules or common pattern that authors had to follow. This leads to an interesting idea that how did modern detective story change overtime as it was developing? This research will allow us to investigate whether authors changed the type of essential clue before and after 1905. Alternatively, 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 histogram that compares the number of physical clues and testimonial clues, with respect to time. The x-axis is the year of publication and the y-axis is the number of stories. The doted line is the year 1905, which is also the line that separates the two groups of this research question: stories published before and after 1905. From the graph, we can see that there are always more physical clues than testimonial clues, which leads to a reasonable prediction that the proportion of testimonial clue is the same before and after 1905.

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

Variable Used

The type of essential clue, The year of publication

Data Wrangling

- Mutate a new variable that divides all detective stories into two categories based on the year of publication:
  - Stories published before 1905
  - Stories published after 1905

Statistical Method

The method used for this question is a two proportions hypothesis test (compares the difference between two proportions from two groups). In this case, the two groups are detective stories that were published before 1905 and after 1905. The null hypothesis H_0 is that the proportion of testimonial clues in stories that were published before 1905 is equal to the proportion of testimonial clues in stories that were published after 1905. On the contrary, the alternative hypothesis is that the two proportions are not equal.

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

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

Research Question 2 - Results and Interpretation

- The test statistic (difference between the proportion of testimonial clues in detective stories published before 1905 and the proportion after 1905, calculated using the available data) is : 0.0769
- 1000 simulations were made under the assumption that our null hypothesis is true. Of all simulations, 94.3% of them were as extreme as the test statistic (p-value is 0.934).
- This indicates that if the null hypothesis is true, the situation as the test statistic is very common, thus we have no evidence against the null hypothesis.
- Since the null hypothesis is correct, we can conclude that there is no difference between the proportion of testimonial clues in detective stories published before 1905 and the proportion after 1905, for all detective stories between early 1800s to early 1900s.

Research Question 3 - Statement of the Question

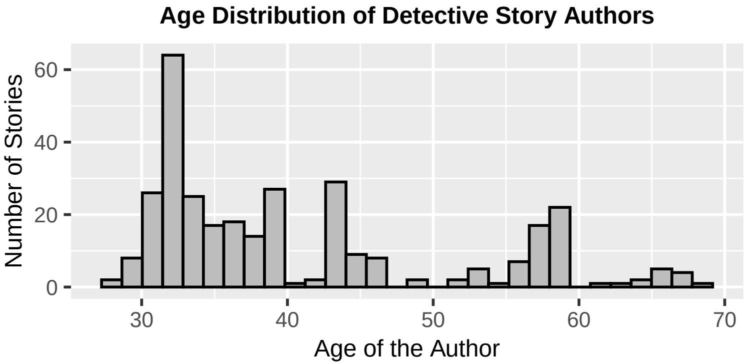
The research question is:

What is the range of plausible values for the average age of author when the detective story was 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 are the people that took this important step and became the pioneers of modern detective stories. 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. The x-axis is the age of the author and the y-axis is the number of stories 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 author published detective stories before 25 years old and after 70 years old.

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

Variable Used

Birthday of the authors, date of the first publication of the stories.

Data Wrangling

- Selected to include the only two variables needed for this specific research question.
- 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

- 1000 simulations were made by re-sample the original data we have. This means to randomly take out a sample that has the same number of observations as our data, but with replacement (some observations might be picked multiple times while others were never selected). For each of the 1000 sample, we calculate the average age of all authors when the detective story is first published.
- From the 1000 results of average ages, the 2.5% percentile is 39.59 and the 97.5% percentile is 41.91. Therefore the middle 95% of all results is between 39.59 to 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(confidence interval).
- 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

Accuracy of the Data

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.

Small Dataset

Besides, the data set contains information of 352 detectives stories between in 1800s to 1900s, which is only a small proportion of all detective stories written in this period of time. Insufficient amount of data may result in lacking variety and thus a higher probability of the data being biased. This may effect the reliability of the research questions negatively since the result could also be biased.

Overall Conclusions

Research Question 1:

The result suggests that the proportion of male victims is significantly higher than the proportion of female victims for detective stories from the early 1800s to the early 1900s. Recall our previous discussion about characters and gender equality, the result agrees with our initial guess that authors may prefer male characters over female. However, unlike the detectives that demonstrate professional skills and intelligence, victims are the people who suffer in the story. Without knowing the details of the story, it is hard to reasonably conclude whether having more male victims represents gender discrimination toward female or male. Besides, the first efforts to achieve equality for women also occurred in the 1800s, thus if we collect more data for detective stories written afterward (not limited to just 100 years), we might observe that the proportion of male and female characters were becoming more balanced. Therefore, we suggest two topics for future analyses: exploring gender equality from the character's experience in the story, and how does the proportion of female character changes along with the waves of feminism.

Research Question 2:

We concluded that the proportion of testimonial clues in detective stories published before and after 1905 is the same (the difference between the two proportions is zero). One of the possible reasons is that people imitated the writing styles of other published detective stories. In the early 1800s, the structure of detective stories had just started to develop, mainstream writing styles had not been formed, thus there wasn't any rule to refer to.

(Continuing on the next page)

Overall Conclusions

Research Question 2:

This can be seen as a form of conformity from a psychological perspective. It is possible that after the first author provided an essential clue, others follow the pattern. The type of essential clue also remained similar, and therefore the proportion of testimonial clues did not change over time. More analyses can be done to investigate how other features of detective stories changed over time and therefore explore how detective story was developed.

Research Question 3:

We found that the range of plausible value of the authors' average age when the detective stories are first published is between 39.59 to 41.91 years old. Therefore, middle age people are the central force contributing to the development of modern detective stories in the early 1800s to early 1900s. As previously discussed in the motivation of this research question, we are interested in discovering the characteristics of people who started to write detective stories. From the visualization, we see that all authors were aged above 25. Together, this leads to the idea that younger people were less likely to write detective stories. This might be a result of insufficient knowledge and life experience, since detective story generally requires rigorous logic and sophisticated writing skills for the readers to understand the story. We can enhance this research by collecting data of the authors educational background and social-economical status to further conclude on their characteristics. However, although most authors in this dataset have passed away, the risk of re-identification may still exist to their families, and we have to consider the ethics of doing this research.

References and Acknowledgements

The authors would like to thank TA Uzair for the helpful suggestions and comments.

- ggplot2 title: main, axis and legend titles: http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles
- Beamer theme gallery: https://deic-web.uab.cat/~iblanes/beamer_gallery/individual/CambridgeUS-dolphin-default.html
- ggplot2 histogram plot: Quick start guide - R software and data visualization: http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization
- ggplot title, subtitle and caption https://www.datanovia.com/en/blog/ggplot-title-subtitle-and-caption/
- British Detective Fiction in the 19th and Early 20th Centuries: https://oxfordre.com/literature/oso/viewentry/10.1093$002facrefore$002f9780190201098.001.0001$002facrefore-9780190201098-e-240
- Gender Discrimination: https://law.jrank.org/pages/22615/Gender-Discrimination-History.html
- Conformity: https://www.psychologytoday.com/ca/basics/conformity