Detective Novels In A Nutshell

Group 55: Statistically Yours

Colin Sihan Yang, Xinxiang Gao, Ming Hon Yeung, Qian Yi

December 3, 2021

Overall Introduction

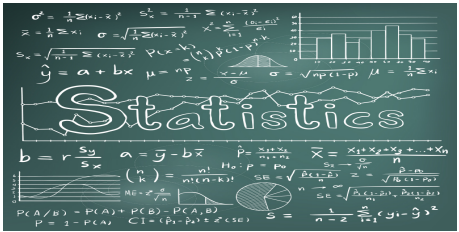


Photo onPennState Eberly College of Science

Goals of The Project

In this project, to help these two professors on their research, we want to obtain the detailed information of the detective stories and get to know some potential association between the characteristics of the stories from that period through statistical methods.

Data Summary

Variables

- total number victims: the total number of male and female victims
- total number culprits: the total number of male, female, non-binary and unknown culprits.
- sex: The gender of detective story's author
- text words: The total number of words in story altogether
- book length: A binary categorical variable with "short book"(word_text $<$7500) and "long book"(word_text $>=$ 7500)

Data_Wrangling

First, we selected the variable we would like to discover.
Second, we filtered out the observations with NA values on the variable above.
Third, we created three more variables:
total number of victims by adding them together,
total number of culprits by adding them together as well,
and book length, dividing the novels into two groups by the number of text words.

Research Question 1: One Group Hypothesis Test

Objective

As a group, we are interested to find out if the proportion of female authors in the detective stories is the same as the proportion of females on a national level.

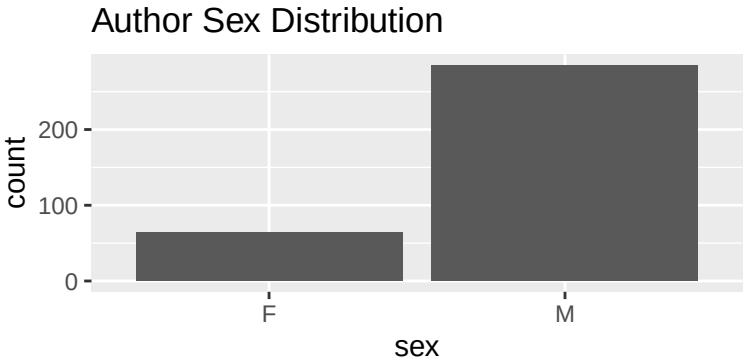
Question Description

Is the proportion of female authors in the detective novels the same as the proportion of females across the country?

Variable Used

The variable we will be using for this specific question is sex

Relevant Visualization



Sample

The sample for this question is all the authors in the 352 detective stories.

Population

All detective stories authors from early 1800s to early 1900s.

Visualization Description

From the barplot, we see that the distribution of sex for male is greater than female at a pretty significant level. We approximate that the number of female authors is 60, and the number of male authors is around 280. The number of male authors is approximately four times the number of female authors.

Exploration

- Method: one-proportion hypothesis test
- Parameter: p_f is the proportion of female authors
- Null hypothesis

$$H_0 : p_f = 0.51$$

  The proportion of female authors in the detective stories is 0.51, which is the national female ratio.
- Alternative hypothesis

$$H_A : p_f \neq 0.51$$

  The proportion of female authors in the detective stories is not 0.51.

Conclusion and Limitation

Conclusion and Interpretation

With a p-value equal to 0, there's a very strong evidence against the null hypothesis that the proportion of female authors is 0.51, which is 51%. In other words, the proportion of female authors of the detective stories from early 1800 to early 1900 is not equal to the proportion of females on a national level. The possible reason could be during that period, male were more engaged in creating literature like detective novels.

Limitations

With the p-value that we get, our concern would be we might make a type 1 error in which we reject the null hypothesis, but it turns out the null hypothesis is true.

Research Question 2: Multiple Linear Regression

Objective

In this question, we want to use linear regression model to predict the total number of culprits based on the total number of victims and the length of books.

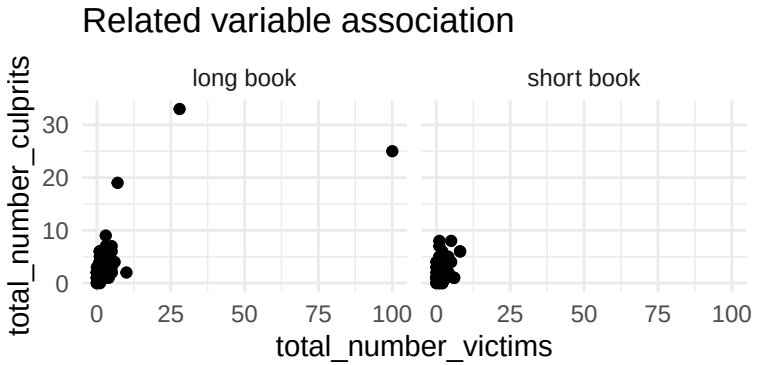
Question Description

Is there an association between total number of culprits and victims as well as book length?

Variable Used

- total number victims
- total number culprits
- book length

Relevant Visualization



Population

Population is all the murder cases in the detective stories from early 1800s to early 1900s.

Visualization

According to the scatter plot, we could see a moderate positive linear association between the total number of victims and the total number of culprits. The strength of the association is similar between short books and long books.

Exploration

- Method: Multiple Linear Regression Model
- Response variable: number of culprits
- Predictor variable 1: length of the book
- Predictor variable 2: number of victims

Conclusion and Limitations

Conclusion and Interpretation

Because the p-values are relatively large, there's no association between number of culprits, victims, or the length of the book. Perhaps, there might be a much larger number of variables that could affect the number of culprits in the stories, and indicating all of those variables would involve huge work.

Limitations

The sample size of the data may be too small to predict the total culprits, which could make the result imprecise for verification. Moreover, since we didn't test multiple models, there might be a better model that predict the number of culprits based on other predictor variables or other combinations.

Research Question 3: Bootstrap confidence interval

Objective

As a group, we want to find out the plausible range of the mean value of the number of words for the detective stories.

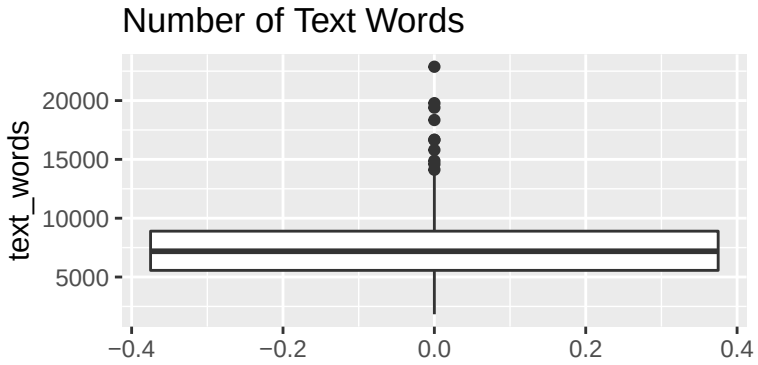
Question Description

What is the range of plausible values for the average text words in the detective stories from early 1800s to early 1900s?

Variabled Used

- text words

Relevant Visualization



Exploration and Visualization(Continuing)

- Method: bootstrap confidence interval
- Original data: detectives_full_clean with number of text words in the sample detective novels.
- Bootstrap sample: boot_samp, which selects observations whose number is equal to original data with replacement.
- Confidence interval: (7224.76, 7829.33)

Population

The population for this specific research question is all the detective stories in the early 1800s to early 1900s.

Visualization Description

According to the boxplot, the text word number ranges from 2000 to 22500. The median is about 7000. The distribution is right-skewed, so the mean would be greater than median.

Conclusion and Limitations

Conclusion and Interpretation

According to the bootstrap sampling, we are 95% confident that the true mean text number of all detective novels/stories is between 7224.76 and 7829.33. This implies that authors around that period preferred to write relatively short novels and finished off early. Since we are selecting the mean, it is also possible that some of the authors tended to write long novels, while others write shorter ones.

Limitations

Although we repeat 1000 trials in our bootstrap sampling method, there is still a probability that the confidence interval does not capture the true mean text numbers of the detective stories from early 1800s to early 1900s due to our relatively small sample size.

Final Conclusion

To sum up, we investigated three different research questions and provide three conclusions as well as interpretations respectively.
As we set up our analysis, some of the results turns out to be surprising.
We hope that our project could provide a unique insight to the audience as this research keeps producing some interesting fact from the perspective of detective novels in early 1800s to early 1900s.

Reference

The dataset, detectives_full, is from detective_data.csv.
Thanks to Prof. Bolton and Caetano as well as Prof. Hammond and Stern for letting us encounter the very special opportunity
to work with this program.
Thanks to TAs for answering our problems.