# Hardy Boys Analyze Detective Novels

Ashvat Ranadive, Ege Mert Akin, Malhar Padir, Manit Gosalia

December 3, 2021

# Introduction

Through this project, we were tasked to assist Professors Adam Hammond (English) and Simon Stern (Law) on their project, "The Birth of the Modern Detective Story". We were provided a dataset which contained a collection of variables referring to the names of the detective stories, their authors, plot summary and other relevant details, published in the period between 1800 and 1900, this served as our project's population. We analysed the different variables and tried to come up with possible research questions that could relate them and lead to interesting conclusions. The various statistical analysis methods we used in our project are data wrangling, hypothesis testing, bootstrapping, and regression. Our project targets those who have a very limited background knowledge about the statistical methods we have used in our analysis, Professors Adam Hammond (English) and Simon Stern (Law) for example.

## Project Goals

- To find if there is a difference between the median number of words (how long the novel is) for novels that provide sufficient clues to guess the outcome and novels that don't.
- To determine with confidence what is the plausible range of words in the novel before the plot is revealed that would result in that novel having a high satisfaction rating (4 and above).
- To examine the relation between the total number of culprits and the satisfaction rating.
  Why did we chose this question?

## Research Question 1

**Is the median number of words the same, for stories that provide sufficient clues in details for the reader to guess the solution and the stories that don't?**

Data Summary: The variables used in the research question are the number of words in a story and whether the story provided sufficient clues in details for the reader to guess the solution ("Yes" or "No" indicated whether it did or not). There were missing values in both of the variables, hence we cleaned our dataset by filtering those to get a more accurate result. The population for the first research question includes the detective stories from early 1800s to 1900s that includes the number of words data and whether the story provided sufficient clues in details for the reader to guess the solution data.

Statistical Method being used to explore the research Question:

- Hypothesis Testing

The null hypothesis in hypothesis testing is the claim that we assume that it is true. Our null hypothesis ($H_0$) is that the median number of words for the stories that provide sufficient clues for the reader to guess the solution and the ones that don't is the same. The alternative hypothesis is the opposite of the null hypothesis. The alternative hypothesis ($H_1$) is that the median number of words for the stories that provide sufficient clues for the reader to guess the solution and the ones that don't, is different.

The null hypothesis:

- $H_0 : \tilde{x}_{sufficient\_clues} = \tilde{x}_{not\_sufficient\_clues}$

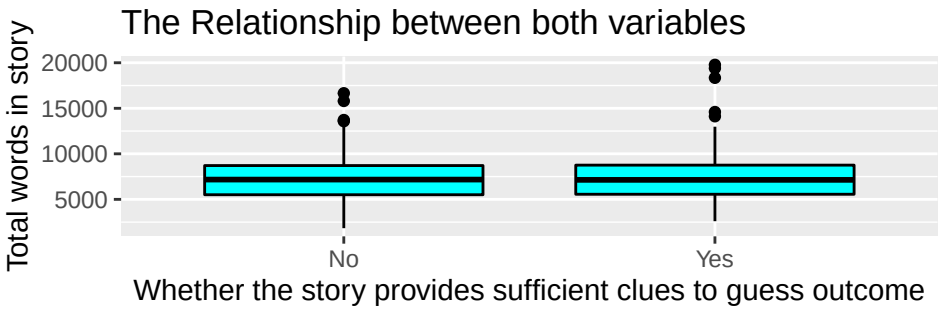
The alternative hypothesis:

- $H_1 : \tilde{x}_{sufficient\_clues} \neq \tilde{x}_{not\_sufficient\_clues}$

$\tilde{x}$ sign is used for representing the median.

- We used hypothesis testing in order to understand whether our claim (null hypothesis is true. The hypothesis test helped us to calculate the p-value for our hypothesis test which demonstrated whether our claim was true or not.
- We used a boxplot in order to visualize the relation between the number of words and the stories that provide sufficient clues in details and the ones that don't.
- We visualized two boxplots next to each other in order to compare the number of words by categorizing the stories that provide sufficient clues as "Yes" and the stories that don't provide sufficient clues as "No" on the x - axis of the graph. The y-axis represents the number of words in a story.

## Visualization




According to the graph, the median for stories that provide clues and the ones that don't is about 7000 which is almost the same for both. This indicates that our null hypothesis was meaningful. This motivated us to do a hypothesis test to make sure about our claim and check the p-value to understand whether our null hypothesis is true. The inter-quartile range (the range in the box where 50% of values are) is almost the same for both cases as well. Both of the boxplots are right-skewed since both the total number of words in a story decreases after the median.

An important thing from the graph that should be noted is the outliers for the stories that provide sufficient clues. There are 3 obvious outliers for that which are higher than 15000 words. These outliers are probably due to fact that some authors wrote more in order to include more clues and details in their story.

```
## [1] 0.917
```

Interpretation:

Our p-value is 0.917. The p-value is quite high which indicates that there is no evidence against our null hypothesis since 0.917 is higher than 0.1. This means that our assumption about medians number of words being the same for stories that provide sufficient clues in details for the reader to guess the solution and the ones that don't, was successful. On the other, we should note that there were some limitations which might have cuased our p-value this high.

From this result we can conclude that including more sufficient clues in details for the reader to guess the solution does not mean that an author should write more in order to accomplish this. They can still achieve this by including sufficient details with concise descriptions which would not affect the number of words.

# Research Question 2

**What is the range of plausible "before_reveal_words" for "satisfaction_rating" to be high where a high satisfaction rating refers to one which is greater than 4**

Data Summary:

First we made sure to remove all null values of data for the satisfaction score and before reveal words using the filter function. Then we mutated a new Boolean variable (True or False) to check if the satisfaction rating is greater than or equal to 4. The reason we chose 4 to be the threshold of high is becuase that would be the equivalent of what most surveys consider a "good" score with 5 being "very good" and 3 being "average". The population for the first research question includes a sample from the detective stories from early 1800s to 1900s that only includes stories with satisfcation ratings over 4 and 5.

## Statistical Method being used:

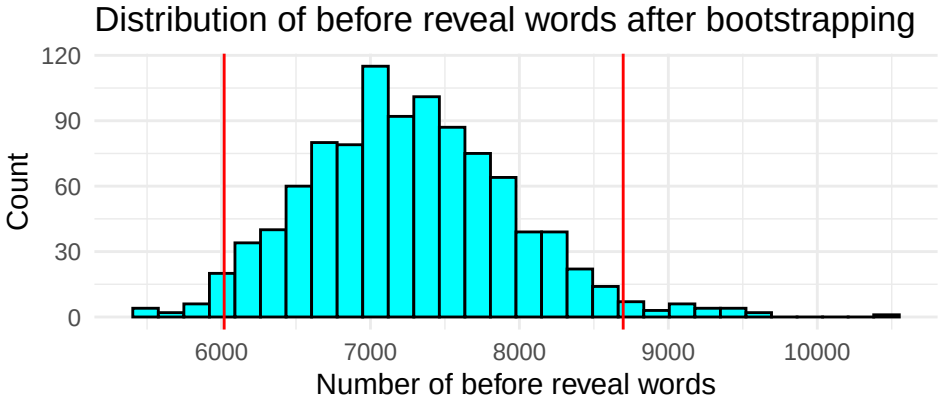
- Bootstrapping

The reason we choose bootstrapping is because using confidence intervals after bootstrapping the sample data, we are able to derive a range of words with 95 percent confidence that would increase the chance of gaining a satisfaction rating higher than or equal to 4.

The statistical method we used to analyse this research question is bootstrapping, a model in which a sample of a certain size is extracted from the population and used for analyzing our research question and deriving an appropriate conclusion. From this sample, we branch out various bootstrap sample of the same size and calculate the statistic, the estimate of the required parameter which is the median of number of words before the reveal in this case, for each bootstrap sample. The confidence interval implemented in our bootstrap analysis is 95% i.e., excluding the first and last 2.5% of the distribution, all the other values can be used to accurately find the median of before reveal words.

Visualization:

The graph below represents a histogram showing the bootstrapped data of the mean before reveal words for each repetition of the simulation for data that only includes stories with a satisfaction score greater than or equal to 4. The histogram also includes red vertical lines that show the confidence intervals.



```
##      2.5%    97.5%
## 6017.825 8696.939
```

Interpretation:

The plausible range of number of words before the reveal for the satisfaction rating to be high came out to be between 6017 and 8696 words. This shows that the detective stories in which the number of before reveal words is between 6017 and 8696 have a high satisfaction rating i.e., greater than 4. From this we can conclude that authors can aim to write number of words within this range before the reveal to publish stories that can satisfy the readers. From our histogram we can see the number of detective stories with the corresponding number of before reveal words and 2 vertical lines which represents the confidence intervals, 6017 and 8696 specifically.

## Research Question 3

**Is there an association between the satisfaction rating of stories and total number of culprits?**

Data Summary: The total number of culprits is calculated as the sum of the variables, for the number of male, female, non-binary, and unknown gender culprits. We created a new variable for the total culprits, which accounts for all the culprits in a story, known and unknown. The data was then cleaned by removing missing values for the variable. In addition, since there was an outlier where total culprits was 2000 in a story which massively skewed the results, out only those observations with total culprits lesser than 2000 were retained and the rest were filtered out. The population is detective stories from 1800s to early 1900s that include data about satisfaction rate and the number of culprits

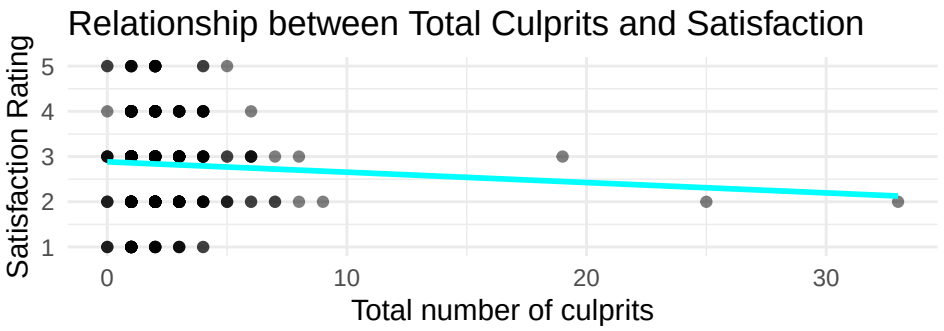
Statistical Method used to explore the research question:

- Simple Linear Regression

We used linear regression to understand the relationship between the number of total culprits in a story and the satisfaction rating. We chose this method to understand whether higher number of total culprits resulted in higher satifaction or lower satisfaction rates. We fitted a line that is close to as many points as possible on the scatterplot so that it could show the relationship more accurately.

Visualization:

We used a scatterplot to visualize our linear regression model. It contained the categorical variable, satisfaction rating, on the y axis and, numerical variable, the number of total culprits, on the x axis. The line we fitted on the graph shows the relationship between these variables and it is fitted in a way that it is close to as many points as possible at the same time. From the line, it can be seen that the satisfaction rating decreases slightly as the total number of culprits increases.



```
##                   Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)     2.88312721 0.07141576 40.371019 1.486808e-133
## total_culprits -0.02290147 0.02063802 -1.109674  2.679036e-01
```

Interpretation:

The slope of the fitted line (-0.023) from the linear regression model plotted above has a negative slope. Slope shows how a change in y-axis for a 1-unit change in x-axis. This means as the total number of culprits in a detective novel increase, the reader's satisfaction rating decreases. However, since the slope parameter is close to zero using total number of culprits to predict satisfaction rating is not that useful, because if it is closer to zero it means that the relationship between the two variables is less significant.

```
## [1] -0.05929496
```

The correlation value was also found as -0.06. This shows the correlation of the points around our fitted line. Since it is close to zero we can conclude that the linear association between the number of culprits and satisfaction rating was not that significant.

From this authors can learn that including less culprits in a story may result in more satisfied readers.

## Limitations

Limitations to the Hypothesis Testing: Since our population only includes detective stories from early 1800s to early 1900s, there were many other detective stories that were not included from 1900 to 2000. This might have caused less accurate results. Therefore, doing hypothesis test with more stories would result in more accurate conclusions.

Limitations to the Bootstrapping Method: This statistical method will not work accurately when the sample size is small because it can only use the limited information it has from that original small sample. If the original sample is not large enough to represent the whole population, then the bootstrap analysis will result in misleading results. This method of analysis is very restricted in the type of information it can give, because it only uses the original sample and does not refer to the whole population.

Limitations to the Linear Regression: Even though satisfaction rating was a numerical variable, it looked like it was a categorical variable on the graph since the satisfactions ratings were as integers rather than decimals. This made it harder to fit a best fit line as satisfaction ratings were not that accurate since it couldn't take values between integers such as 3.4. If it did the best fit line would summarize the the relation between the number of culprits and satisfaction rating better.

## Conclusion

- In relation to Research Question 1, we can conclude that author's do not need to write more to provide sufficient clues and details for the readers. They can still manage to do this by being concise. In our opinion, including details with concise explanations would result in more satisfied readers.
- In relation to Research Question 2, we aimed to find the plausible range of words for the satisfaction rating to be high. We were able to find this range with certain confidence however, as mentioned the method has its limitations. However, we do suggest, if Professor Adam Hammond decides to write his own set of mystery novels one day, he should try to make sure the number of before reveal words is within the range of 6000 and 8700.
- In relation to Research Question 3, even though there was not a significant linear relationship between the number of total culprits and satisfaction rating, the satisfaction ratings tend to decrease as the total number of culprits increase. This helps us conclude that if authors and our audience decide to write a detective story, they should not have too many culprits in their stories in order to satisfy the readers.

## Acknowledgements

*The authors would like to thank Cassidy Wang and Uzair for their helpful suggestions and comments that improved the presentation of this poster.*

*The authors would also like to thank Professor Bolton and Professor Caetano for their continuous guidance and support.*