Exploring Possible Indicators Leading to a Satisfying Detective Novel

Min Gi Kwon, Aidan Li, Youssef Soliman, Sarah Xu (82. Poutine Pouteam)

December 3, 2021

Introduction

Our team is helping Professors Adam Hammond and Simon Stern on their project, "The Birth of the Modern Detective Story". We're going to use data to look into detective stories from the early 1800s to the early 1900s and see what makes them enjoyable.

Our data is a random sample of over 300 short detective stories from the time period. In this project, we will use statistical inference to make certain conclusions about all the detective stories from this time period based on what we can learn from our data sample.

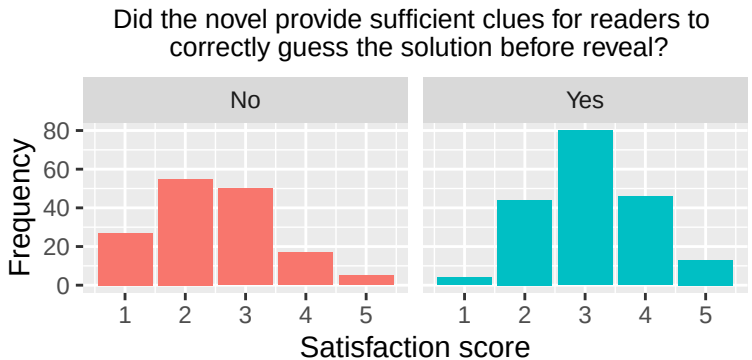
We will be investigating three questions in this project and making statistical inferences from our results to learn more about detective novels from the early 1800s to the early 1900s.

Question 1: Question Statement & Data Wrangling

Question: Are detective novels that provide sufficient clues in sufficient detail to allow readers to correctly guess the solution before the reveal as satisfying as detective novels that do not?

- We believe that people like detective novels partly because it is fun to try to solve the mystery for themselves as the story progresses.
- We renamed `does_the_story_provide_sufficient_clues_in_sufficient_detail_to_allow_a` to `sufficient_clues_to_guess_correctly` (categorical variable) and `how_satisfying_is_this_story_as_a_piece_of_detective_fiction` to `satisfaction_score` (discrete numerical variable)
- We selected only the `sufficient_clues_to_guess_correctly` **(indicates presence of sufficient clues)** and `satisfaction_score` **(indicates satisfaction rating)** variables, and filtered out observations with any missing values in these columns. *There are 341 observations remaining for our analysis.*

Question 1: Data Visualization



These two barplots compare the frequency distribution of satisfaction levels for novels in our given dataset that provided sufficient clues to allow readers to correctly guess the solution, and those that did not. There were 154 novels for 'no' (did not provide sufficient clues), and 187 novels for 'yes' (provided sufficient clues).

Question 1: Statistical Methods

- To answer this question, we carried out a two-group hypothesis test at the 5% significance level to test whether the mean satisfaction rating of detective novels that provided sufficient clues for the reader to correctly guess the solution was different to the mean satisfaction rating of novels that did not provide sufficient clues.
- The null hypothesis states that the mean satisfaction rating of detective novels with sufficient clues is the same as the novels without sufficient clues, while the alternate hypothesis states there is a difference in mean satisfaction rating between the groups.

In a hypothesis test, we assume the null hypothesis is true and see if there is enough evidence based on simulated data to reject the null hypothesis and claim the alternate hypothesis is true instead.

We have the following equations:

$$H_0 : \mu_{sufficient} - \mu_{insufficient} = 0$$

$$H_A : \mu_{sufficient} - \mu_{insufficient} \neq 0$$

where $\mu_{sufficient}$ is the mean satisfaction rating of the novels that provided sufficient clues, and $\mu_{insufficient}$ is the mean rating of the insufficient clues group.

Question 1: Results

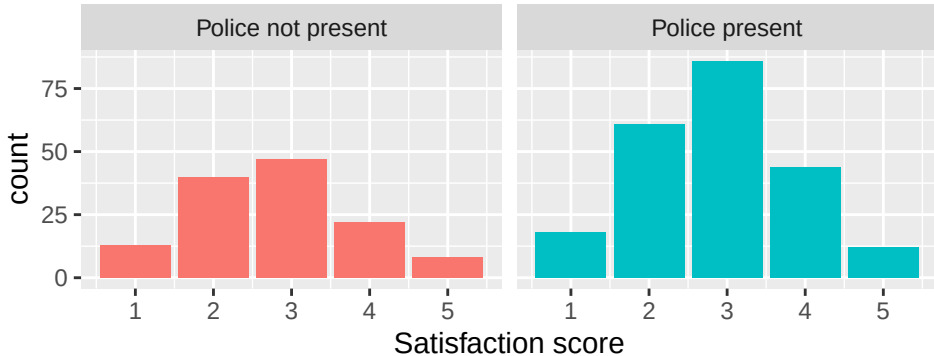
- Our hypothesis test returned a p-value of 0. The smaller the p-value, the greater the evidence against the null hypothesis; a p-value of 0 indicates there is extremely strong evidence against the null hypothesis. So, at the 5% significance level, we can reject the null hypothesis.
- This allows us to make a very strong claim that there is a difference in mean satisfaction score between detective novels (written in the 1800/1900s) that provided sufficient clues in sufficient detail to allow readers to correctly guess the solution before the reveal, and detective novels that did not.
- Overall, the presence of clues in detective novels from the 1800s-1900s appears to have a noticeable effect on the reader's satisfaction.

Question 2: Question Statement & Data Wrangling

Question: Does police involvement in a detective story impact how satisfying the story is as a piece of detective fiction?

- Police involvement often causes more harm than good by interfering with the protagonist's investigation. The trope might be cliche, to the reader's dismay.
- We created a new variable `police_involvement` that tells us whether police were present in a novel's investigation based on its corresponding value for the variable `what_is_the_role_of_the_police_force_in_solving_the_crime`.
- We filtered out all missing values for the new variable `police_involvement`. In total, we had 351 observations for our statistical analysis.
- We selected only the `police_involvement` **(identifies police involvement in a novel, categorical variable)** and `how_satisfying_is_this_story_as_a_piece_of_detective_fiction` **(measures satisfaction rating, discrete numerical variable)** variables.

Question 2: Data Visualization



These two barplots compare the frequency distribution of satisfaction levels for novels in our given dataset that had police involvement and novels that did not have police involvement. There were 130 novels for 'Police not present', and 221 novels for 'Police present'.

Question 2: Statistical Methods

- We carried out a two-group hypothesis test at the 5% significance level to test whether the mean satisfaction rating of detective novels that had police involved in the investigation process was different to the mean satisfaction rating of novels that did not have police involvement.

$$H_0 : \mu_{police\_present} - \mu_{police\_not\_present} = 0$$

$$H_A : \mu_{police\_present} - \mu_{police\_not\_present} \neq 0$$

- The null hypothesis states the mean satisfaction rating of detective novels with police involvement is the same as the group without police involvement, while the alternate hypothesis states the mean satisfaction rating differs between the groups.
- We support or reject the null hypothesis by observing the presence of simulated data that is rare enough to argue against it.

Question 2: Results

- Our hypothesis test returned a p-value of 0.485. This p-value indicates there is practically no evidence against the null hypothesis. So, based on this hypothesis test, we fail to reject the null hypothesis at the 5% significance level, and we cannot claim with any evidence that there is a difference in the mean satisfaction rating between detective novels with and without police involvement.
- Based off this statistical analysis, police involvement in detective stories from the 1800s/1900s does not seem to affect how satisfying the story is.

Question 3: Question Statement & Data Wrangling

Question: Is there an association between the readability of a story and its satisfaction score?

- This question stems from a fairly common belief among academics that if a text is easier to read, it is more understandable and as such the reader is more satisfied after reading it.
- In order to answer this question, we first had to extract the full text version of as many stories as we could.
- This was achieved by using the `story_url` variable associated with the stories and visiting each link programatically through a process called web scraping.
- After de-duplication and text parsing was complete, there were 239 unique observations for the analysis to take place.

Question 3: Data Wrangling (cont.)

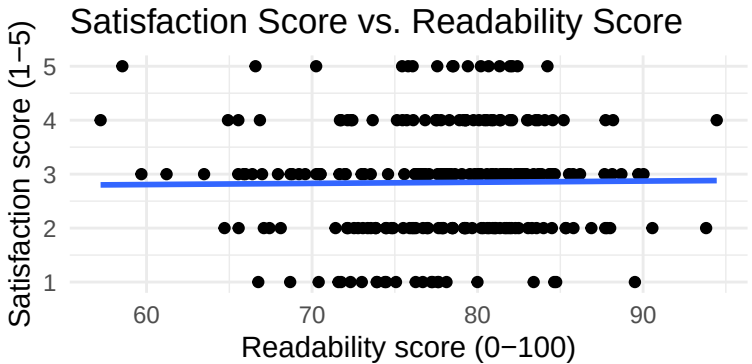
- We used the Flesch-Kincaid Reading Ease score to assign a score between 0-100 for each piece of text (note that a higher score implies the text is less complex and easier to read).

The Flesch-Kincaid score was calculated using the following formula:

$$FK_{RE} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- The number of sentences were extracted based off of the position of certain punctuation within the text and the words were determined through a tokenization algorithm, which splits the text up into its individual word components.
- The syllables within our text were extracted using a hyphenation algorithm used to split up words into their individual components.

Question 3: Data Visualization



This scatterplot visualizes the distribution and nature of the reading scores when compared to the satisfaction scores with a fitted line. Each point represents an individual novel. Clearly, there seems to be no association between the two variables based off of this visualization alone.

Question 3: Statistical Methods

- The equation modeling the potential linear relationship between our readability scores (continuous numerical variable) and satisfaction scores (discrete numerical variable) is represented as such:

$$\hat{y}_{\text{satisfaction score}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{readability score}}$$

- In order to statistically determine if there is a relationship between our two variables, we perform a hypothesis test with the hypothesis at the 5% significance level as follows:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

- If our null hypothesis (H_0) is not rejected, then there is no evidence of a linear association between our readability score and satisfaction score.

Question 3: Results

- After performing the hypothesis test, we determine a p-value of ~0.84 which indicates that there is no statistical evidence for a linear association between readability and satisfaction scores. Subsequently, we fail to reject the null hypothesis at the 5% significance level due to a lack of evidence against it.
- Based off of this statistical analysis, there does not seem to be any linear association between the readability of a detective novel from the 1800s-1900s and its satisfaction score.

Limitations

- In our hypothesis tests, there is always a small chance we could draw the wrong conclusions. We could be guilty of either a Type I error in Q1 or a Type II error in Q2. Simply put, this means that we reject the null hypothesis or fail to reject the null hypothesis respectively although we should have done the opposite.
- We cannot necessarily say correlation implies causation. Many elements in a story could act as confounding variables affecting how satisfying the story is. The author's name (bias based on reputation) could be a confounder, for example.
- For the third question, a good portion of the dataset was lost due to incorrect URLs being used in the dataset or webpages that could not be parsed. Furthermore, the method used to extract text introduces minor artifacts into the final texts due to the unstructured nature of the webpages' construction. This could introduce incorrect reading scores and skew the results.
- Technically, satisfaction score is a categorical variable but we have treated it as a numerical variable for our analyses.

Conclusion

Our statistical analyses suggest the following:

- The presence of clues in detective novels from the 1800s-1900s appears to have a noticeable effect on the reader's satisfaction.
- Police involvement in detective stories from the 1800s-1900s does not seem to affect how satisfying the story is.
- There does not seem to be any linear association between the readability of a detective novel from the 1800s-1900s and its satisfaction score.
- Overall, it is difficult to pinpoint what exactly makes a story satisfying. Two reasonable assumptions were inconclusive in reality after investigation. Not really "elementary, my dear Watson".

References and Acknowledgements (optional)

We would like to thank the creators of these libraries that were essential for our project.

- https://stringr.tidyverse.org/ **stringr** String manipulation library for R
- https://cran.r-project.org/web/packages/urltools/vignettes/urltools.html **urltools** URL parsing library
- https://rvest.tidyverse.org/ **rvest** Web scraping library for R
- https://cran.r-project.org/web/packages/httr/index.html **httr** HTTP library, allowing to set request timeouts
- https://purrr.tidyverse.org/ **purrr** Used for the `map` function
- https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.html **koRpus.lang.en** R Text analysis library for Flesch-Kincaid score