Homework4

Surafel Geleta ssg2775 https://github.com/surafelgeleta/SDS315_HW4

Problem 1

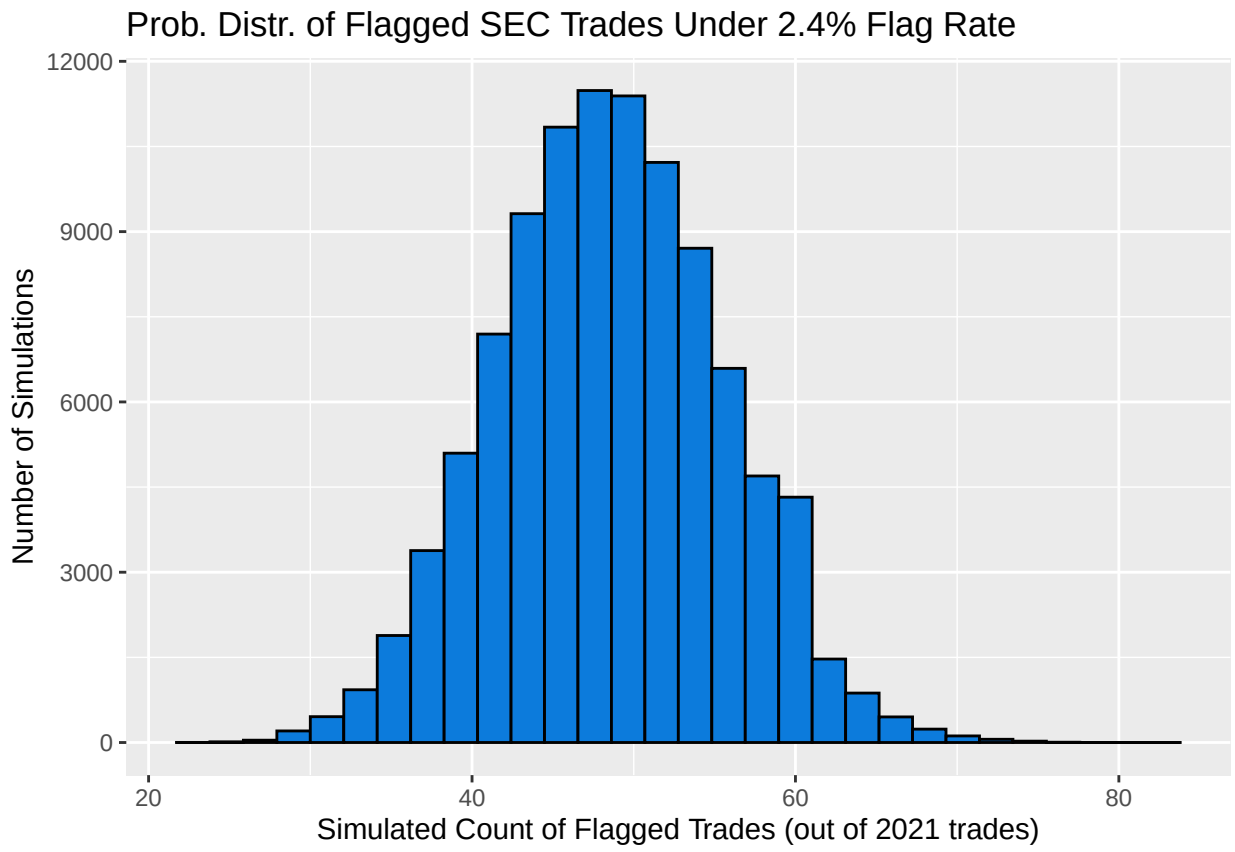
Null Hypothesis

The null hypothesis is that across time, the proportion of Iron Bank securities trades that are flagged by the SEC's detection algorithm is 0.024.

Test Statistic

The test statistic is the number of trades made from the Iron Bank that were flagged by the SEC's algorithm, which was 70 flagged trades out of 2021. This comes out to a proportion of flagged trades of 0.0346363.

Probability Distribution



The graph above displays the probability distribution of the test statistic assuming that Iron Bank securities trades are flagged at a rate of 2.4%.

P-Value

The p-value or probability of the Iron Bank observing a count of at least 70 flagged trades out of 2021 by chance with a flag rate of 2.4% is 0.00205.

Conclusion

Using an threshold of 0.05 and considering that $0.00205 < 0.05$, the p-value appears to be significant at the 95% level and may lend enough evidence to reject the null hypothesis that Iron Bank is flagged at a rate of 2.4% over time, but it may help to collect flagging data at different points of time to make a more conclusive decision about the null hypothesis's plausibility.

Problem 2

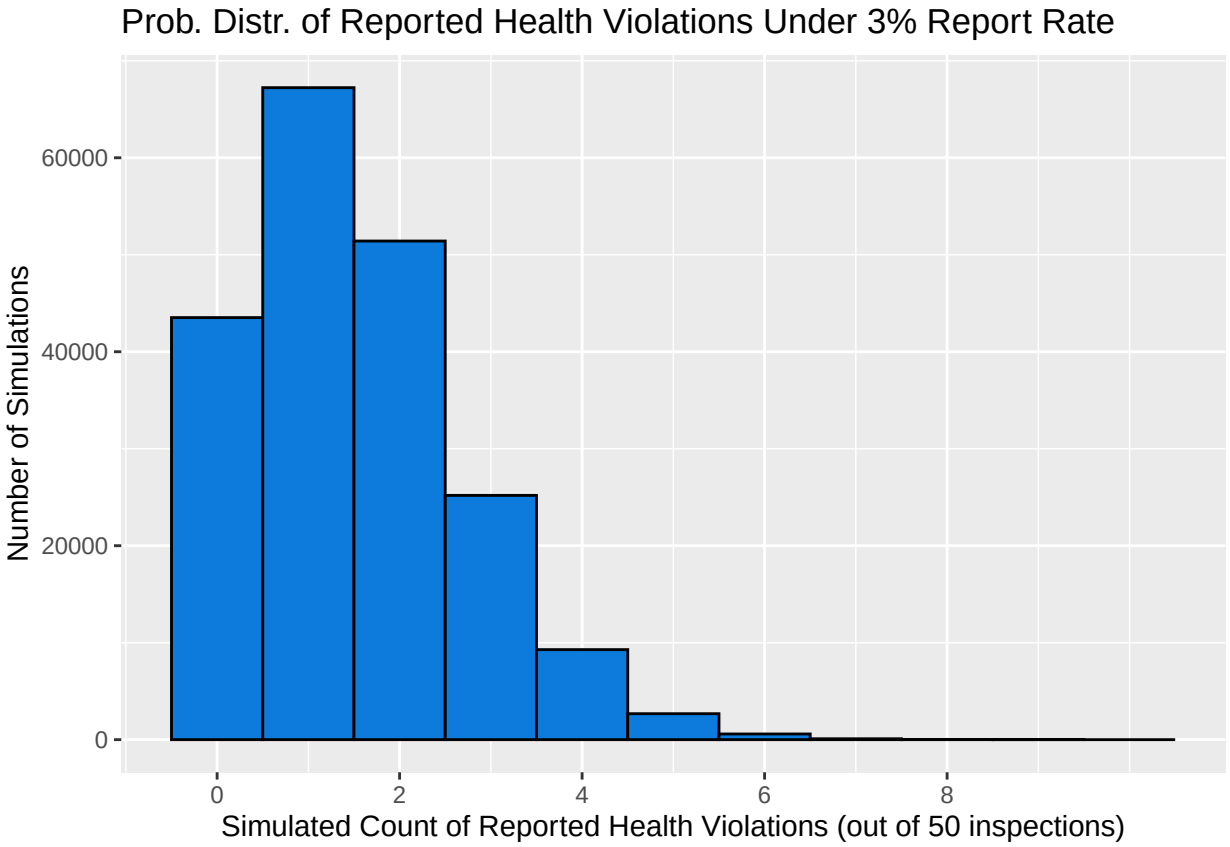
Null Hypothesis

The null hypothesis is that Gourmet Bites's health code violation report rate is consistent over time as the average citywide report rate of 3%.

Test Statistic

The test statistic is the count of health violation reports that Gourmet Bites had over the last years, totaling 8 reported violations out of 50 inspections and amounting to a 16% health violation report rate.

Probability Distribution



The graph above displays the probability distribution of counts of reported health violations among 50 inspections if the citywide average report rate is 3%.

P-Value

The p-value or probability of a restaurant in the city observing a count of 8 or more reported health code violations out of 50 inspections by chance with a citywide report rate of 3% is 0.00012.

Conclusion

Using a threshold of 0.05 and considering that the p-value of $0.00012 < 0.05$, there is evidence to reject the null hypothesis that Gourmet Bites's health code violation report rate is consistent over time as the average citywide report rate of 3%, and it may warrant further investigation into Gourmet Bites particularly because actual health code violations may be unevaluated because the test statistic only includes those reported.

Problem 3

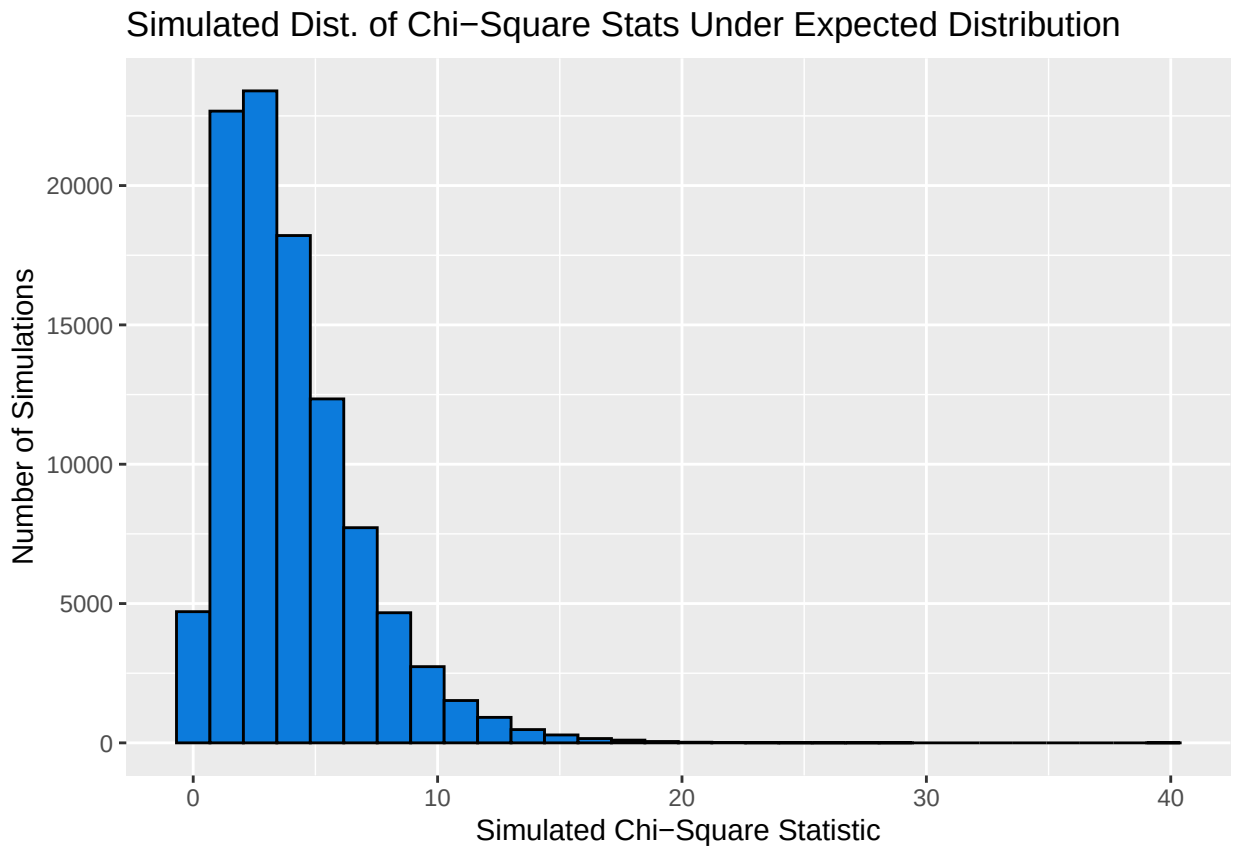
Null Hypothesis

The null hypothesis is that the distribution of group counts among the impaneled jurors in the 20 trials is consistent with the multinomial distribution of groups in the county's eligible jury pool.

Test Statistic

A chi-square test statistic is calculated as the test statistic. The expected counts of juror group distributions is calculated by multiplying the countywide eligible jury pool by the total number of impaneled jurors that were involved in the 20 trials overseen by the same judge. Using the chi-square goodness-of-fit test, the chi-square test statistic is **12.42639**.

Probability Distribution



The graph above displays the distribution of 100000 simulated chi-square statistics under the assumption that the null hypothesis is true. What this means is that for each simulated chi-square statistic, the "observed" group counts were calculated using the multinomial expected distribution of the jury groups, as well as the size of the observed counts in the test statistic (240 jurors).

P-Value

The probability of observing a jury group distribution at least as extreme as the one seen among the 240 impaneled jurors by chance, assuming that the distribution of group counts are consistent with the expected multinomial jury group distribution, is **0.01425**.

Conclusion

Using a threshold of 0.05, since $P(T \mid H_o) = 0.01425 < 0.05$, the data do support the idea that the observed group distribution of the jurors overseen by that particular judge are not consistent with, and significantly

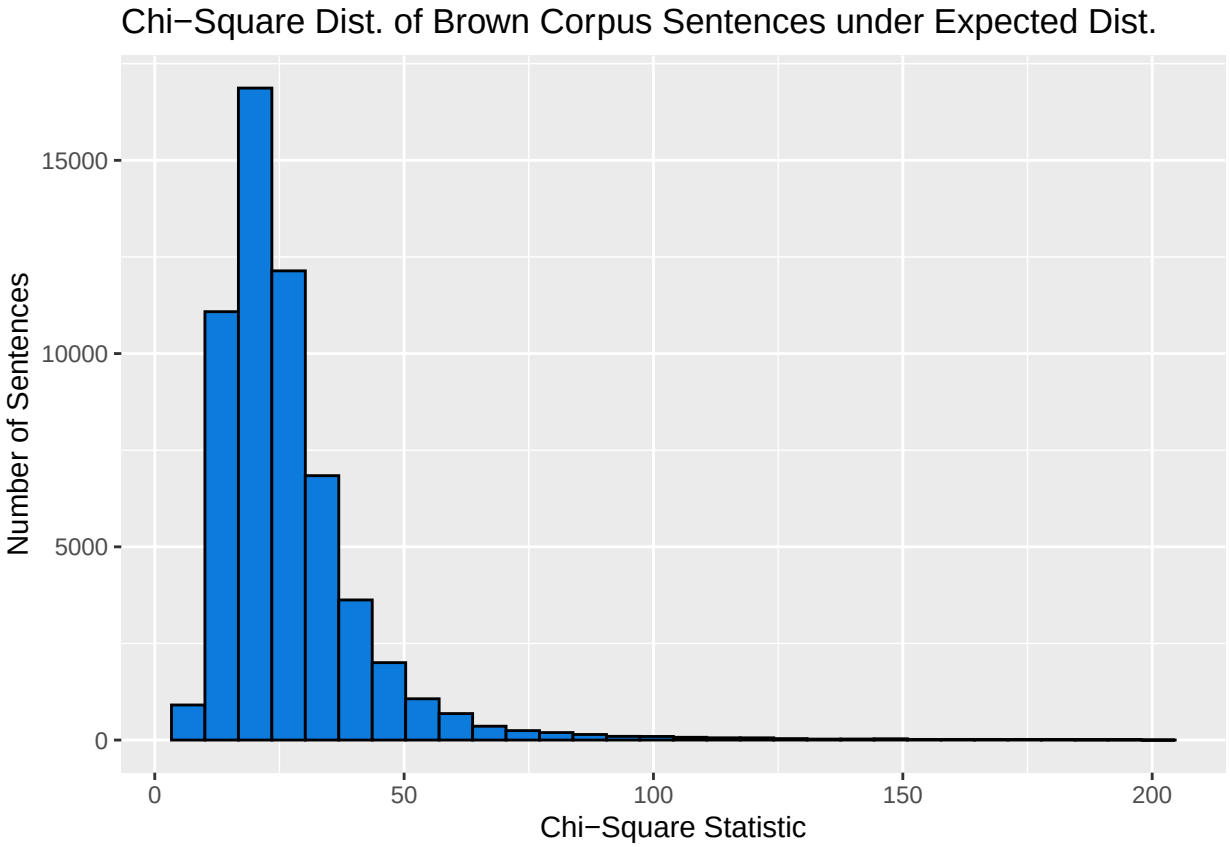
different from, the countywide eligible jury pool group distribution. While this probably warrants further investigation into the judge, it by no means establishes that the judge is engaging in systematic bias against or in favor of particular groups. Some explanations apart from bias in favor of and/or against groups for the difference in the judge's jury group distribution and the countywide jury pool group distribution include:

- The judge's working hours, which may result in the judge being more likely to select jurors that are available for jury duty for the particular trials the judge precedes. It may be the case that certain groups are disproportionately more or less likely to be available at certain times.
- The group distribution of the eligible jury pool may not be representative of the people who actually participate in jury duty in general. Eligible potential jurors often have certain reasons for not wanting to participate in jury duty (work, school, familial and homemaking responsibilities), and these reasons may vary from group to group.
- Potential jurors from certain groups may be more likely to exhibit "clear bias", thus resulting in them getting removed from jury duty "for cause". Moreover, it may also be the case that this particular judge has a lower bar for "clear bias", resulting in a juror group distribution notably different from the juror group distributions selected by other judges.

In order to investigate this matter further, it might be helpful to look at differences among jurors selected and not selected by the judge when the jurors had similar backgrounds including availability hours and clear bias or lack thereof, similar to a blocking technique. Then, it may be easier to evaluate potential systematic bias on the judge's part when looking between jurors who would have no reason to be removed or not considered other than for availability (largely falling under "automatic exemption) or bias (falling under"for cause").

Problem 4

Part A



Above is the distribution of chi-square statistics across a collection of 56745 sentences and phrases from the Brown Corpus, with comparisons made between observed letter counts in Brown Corpus sentences and expected letter counts calculated using the number of letters in a sentences multiplied by the Project Gutenberg English letter frequency distribution. This distribution indicates the range of expected chi-squared statistics for standard, human-written English sentences according to Gutenberg's letter frequency distribution.

Part B

Table 1: P-Values of Chi-Square Test Statistics of Ten Sentences

Sentence	P-Value
1	0.879
2	0.513
3	0.981
4	0.866
5	0.981
6	0.132
7	0.964
8	0.981

9	0.981
10	0.513

The sentence generated and watermarked by an LLM was most likely sentence 6. This is because it has the lowest p-value of all ten sentences, indicating the probability of observing the sentence's letter distribution or a letter distribution more extreme than that by chance assuming that all of the above sentences were human-written and not generated by an LLM was 0.132. This sentence or any sentence with a letter distribution more extreme being the least likely to have occurred by chance suggests that it was LLM watermarked and generated, given that we know that one of the ten sentences was LLM generated and watermarked.