Homework 5

Lucas Chiang

2024-02-24

Github Link:

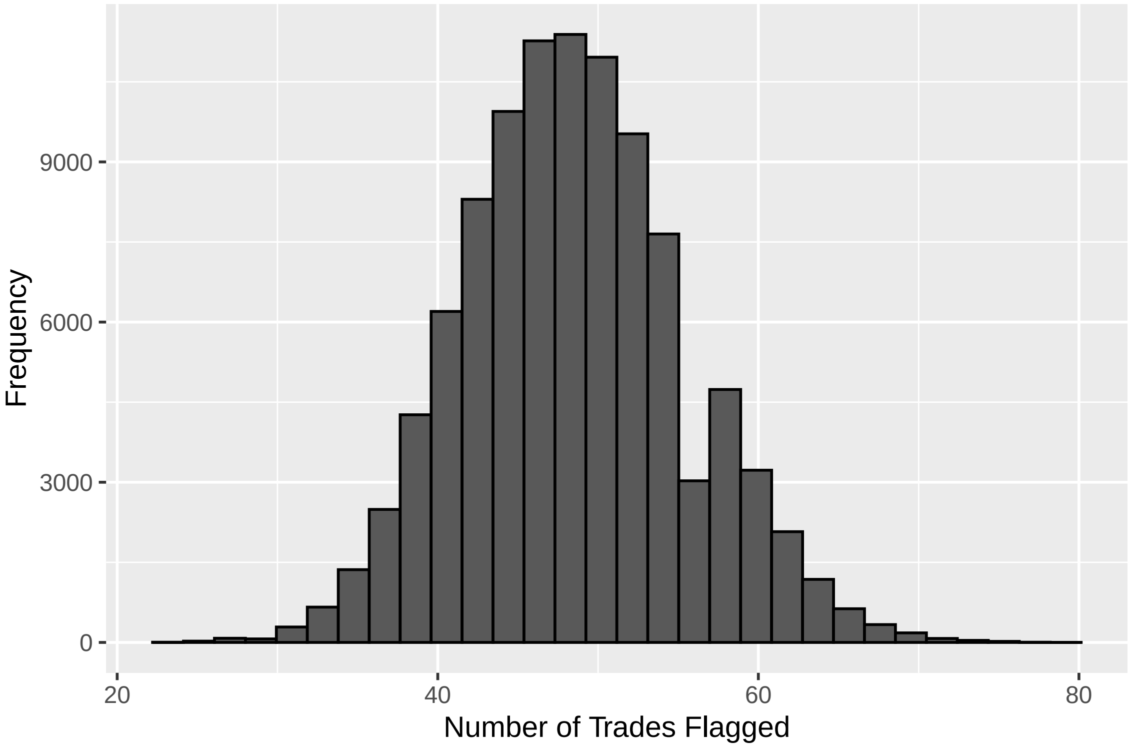
Problem 1 - Iron Bank

Description: Here we are investigating if 70 out of 2021 (3.45%) of trades being flagged by the SEC in the Iron Bank, is consistent with the random variability of trading patterns with a 2.4% baseline rate of trades being flagged by the SEC algorithm.

Null Hypothesis: The securities trades from Iron Bank are flagged at a baseline rate of 2.4% in the long run.

Test Statistic: The test statistic I used here is a number of trades flagged over the course of 2021 trades to weigh evidence against the null hypothesis with the baseline rate being flagged being 2.4%.

100000 Simulated Runs of Flagged Trades over The Course of 2021 Trades



P-Value: The p-value here expresses the probability of obtaining 70 or more flags over the course of 2021 trades under a baseline probabilty of 2.4%. The p-value here is 0.00213.

Conclusion: Since the p-value is less than .05, it is statistically significant, so we can reject the null hypothesis by convention and say that the Iron Bank does not have a baseline rate of 2.4% of flagged trades and should be investigated; as a matter of personal opinion to which the null hypothesis is plausible, it is highly unlikely because the p-value is .00213 a very small number below the .05 threshold.

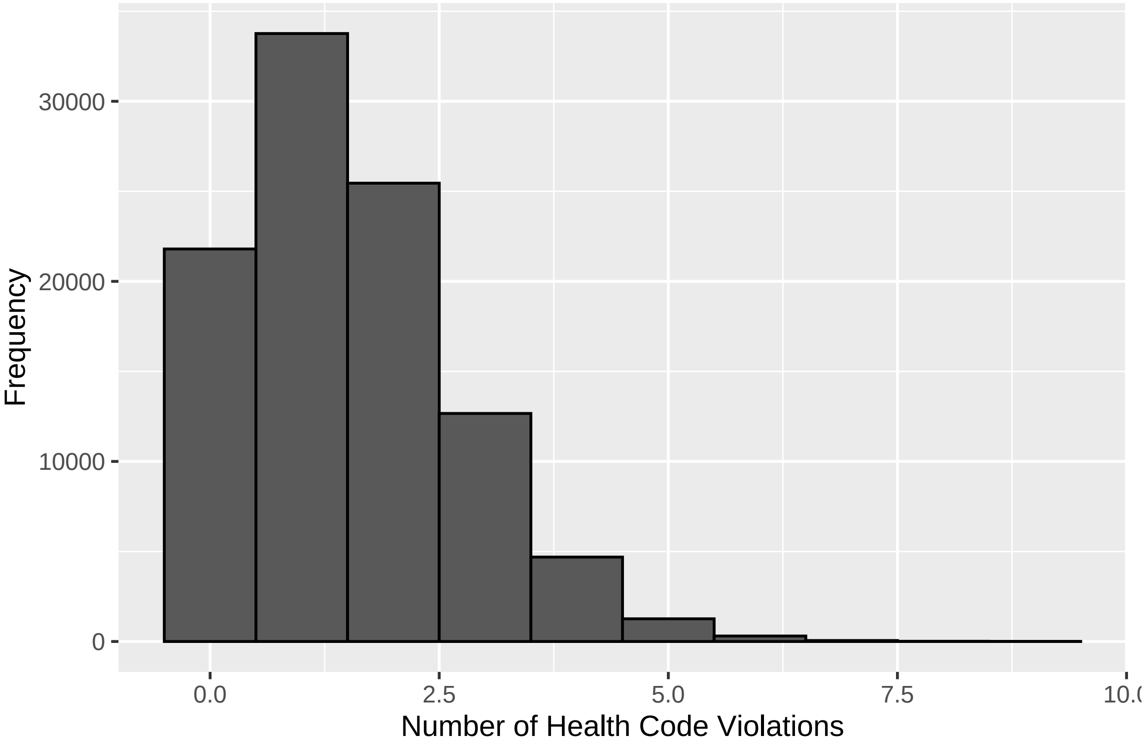
Problem 2 - Health Inspections

Description: Here we are investigating if 8 out of 50 inspections of health code violations in the restaurant chain Gourmet Bites, is consistent with the random variability of health code violations at a 3% baseline rate on average for the restaurants in the city. If not, then there is solid evidence for the Health Department to take action.

Null Hypothesis: The baseline rate of health code violations for the local restaurant chain Gourmet Bites, is consistent with the citywide average rate of 3%.

Test Statistic: The test statistic used here is the number of health code violations that is brought up under 50 inspections in a baseline rate of 3%.

100,000 Simulated Runs of Health Code Violations Over the Course of 50 Inspections



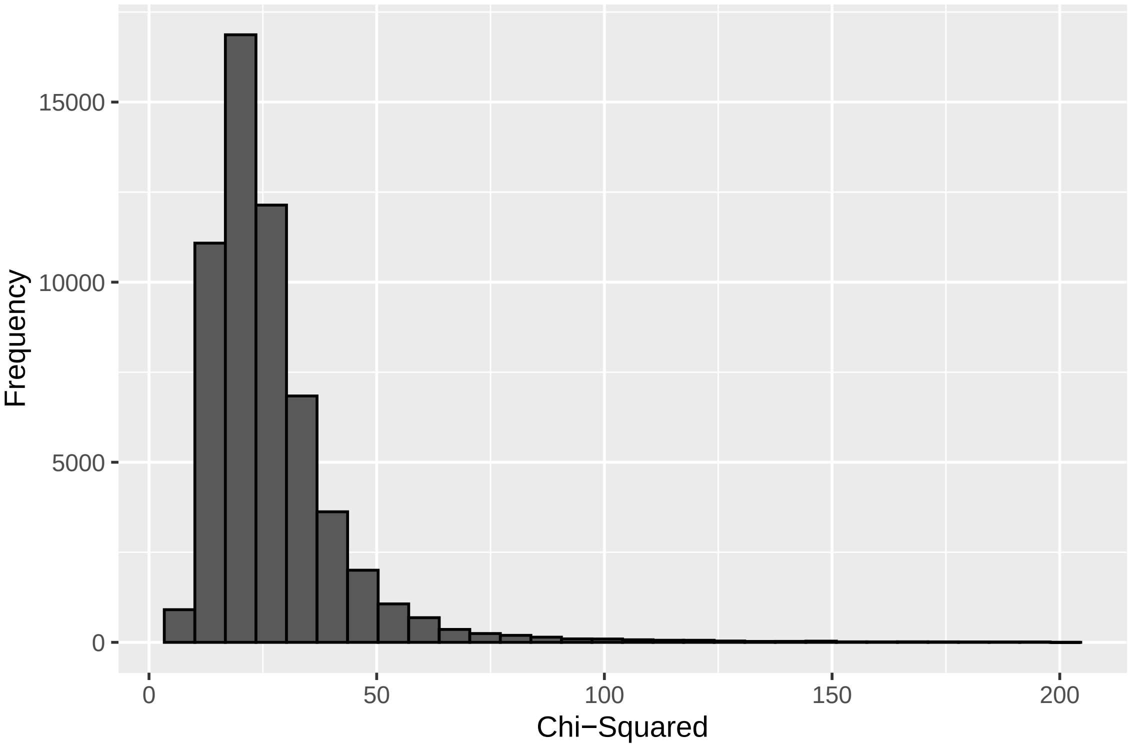
P-Value: The p-value here expresses the probability of getting 8 or more health code violations over the course of 50 inspections with an average rate of 3%. The p-value here is .00014.

Conclusion: Since the p-value is well below the conventional threshold of .05, this value is statistically significant, and we can reject the null hypothesis and say that there is solid evidence for the Health Department to take action against Gourmet Bites; this p-value is extremely close to 0, so it is very unlikely that the null hypothesis is plausible.

Problem 3: LLM Watermarking

Part A: The Null or Reference Distribution

Null Distribution of Chi-Squared Values from English Sentences Extracted by Brown Corpus



Part B: Checking for a Watermark

Here, we test 10 sentences, one of which is produced by a large language model. We will figure which one it is by comparing their chi-squared values to the typical English letter distribution, and see where that chi-square value lies in the null distribution in order to calculated p-values. Doing so will allow us to see which sentence is the most likely to have been created by a large language model.

Sentence	Chi-Squared Values for Test Sentences
1	22.931
2	13.051
3	46.286
4	23.546
5	23.676
6	96.453
7	28.271
8	9.635
9	44.929
10	49.961

Sentence	P-Values for Test Sentences
1	0.513
2	0.926
3	0.076
4	0.489
5	0.484
6	0.009
7	0.328
8	0.988
9	0.084
10	0.059

The sentence that has been produced by a large language model is sentence 6 with a p-value of .009. We know this because in this case, the p-value is an indication of how likely it is that a human-written sentence has a chi-squared deviation statistic that is at least 96.453 or greater when compared to the null distribution. Since the p-value is extremely low and is below .05, we can reject the idea or the null hypothesis that this particular sentence resulted from chance and had a baseline distribution of typical English texts. Although sentences 3, 9, and 10, might warrant investigation because they are near the .05 threshold, sentence 6 has a much lower p-value out of all the sentences, and since we know there is only one that was produced by the LLM model, this one is the most likely to have been. Therefore, we know that sentence 6 was likely produced by a large language model which manipulated the frequency of the letters deliberately to create a watermark.