HW4

Genavieve Middaugh

2025-02-14

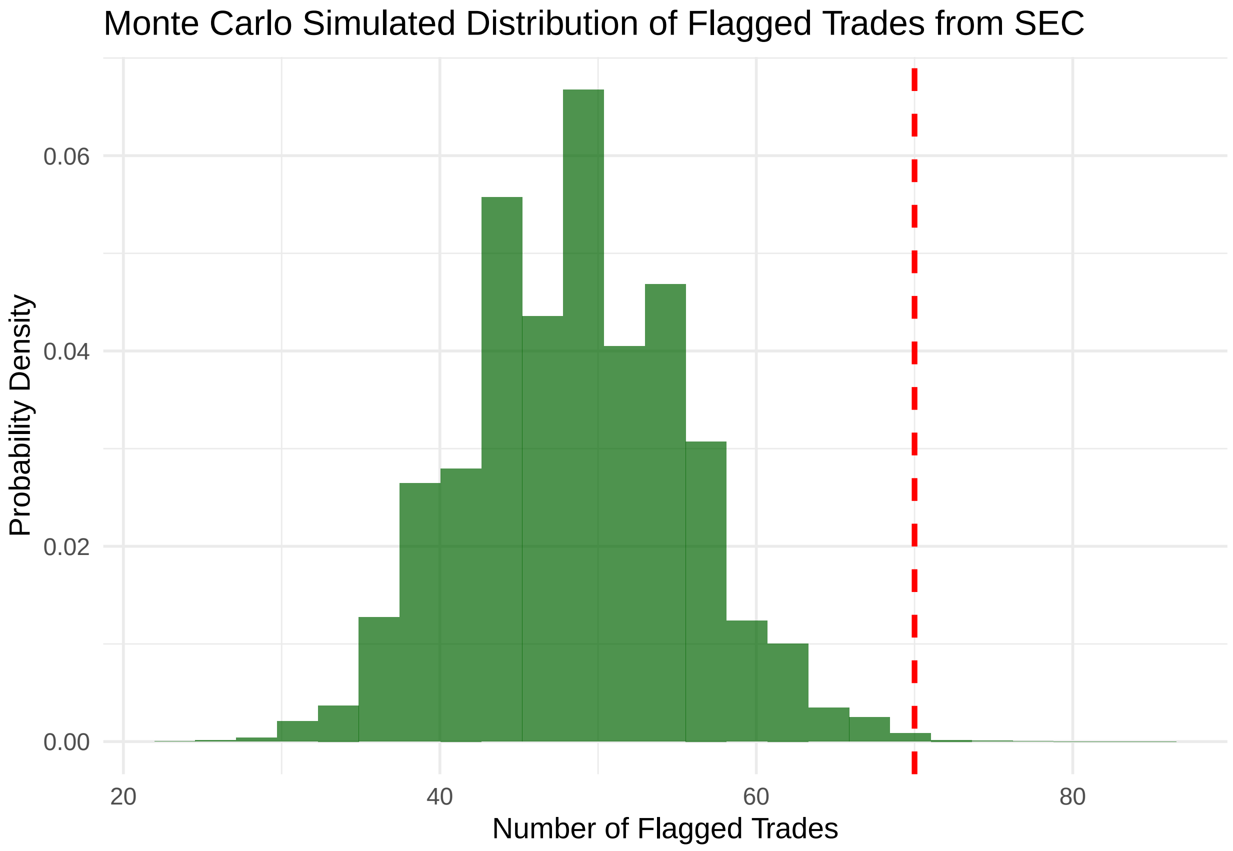
Problem 1 - Iron Bank

In this problem we are testing if the observed data is consistent with the Securities and Exchange Commissions, null hypothesis that over time security trades from Iron Bank are flagged at the same 2.4% rate as other traders.

Null Hypothesis: 2.4 will be our null hypothesis, this number comes from the Iron Bank employees trades that were flagged at this baseline rate of 0.024 compared to other traders.

Test Statistic: The number of flagged trades out of 2021 total trades which is 70/2021 = 0.03465

Monte Carlo Simulation and P-Value: Simulate 100000 trials where we assume each trade is 2.4% chance of being flagged. So after each trial, it will count how many trades get flagged. Below is the simulation.



```
## [1] 0.00185
```

The graph is bell shaped indicating little variation or skew, but there is a bit of inconsistency in the middle, with seemingly drastic leaps down and then upp again. The p-value we get is 0.00185, so under the null hypothesis there is only a 0.185% chance of observing 70 or more flagged trades by random chance. One would argue that this is relatively low number, it is below 0.05 and when you put it in context of the simulation, this means that the idea of being flagged 70 times is very odd, and doesn't or shouldn't happen often under a 2.4% flag rate.

Conclusion: Since the p value is much smaller than 0.05 we have strong evidence to reject the null hypothesis. This value suggests that the flagged trade rate for Iron Bank employees is significantly higher than expected in the regular trading scene.

Problem 2 - Health Inspections

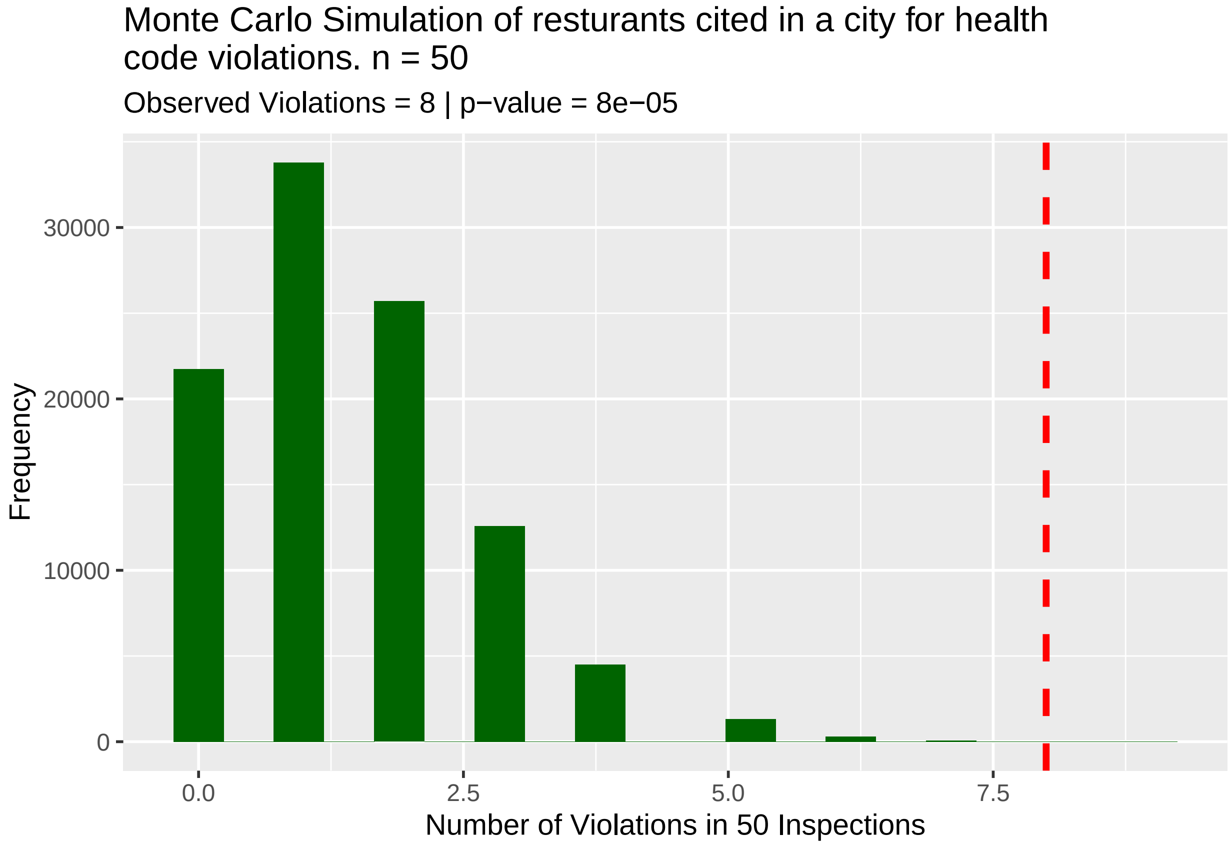
The Health Department wants to ensure that any action taken is based on solid evidence that Gourmet Bites' rate of health code violations is significantly higher than the citywide average of 3%. In Problem 2 we are looking at a local health department that was investigating gourmet Bites. We know that Gourmet Bites was inspected 50 times out of the 1500 times that inspections were conducted, and we know that 8 of these 50 inspections led to citations.

Question: Are the observed data for Gourmet Bites consistent with the Health Department's null hypothesis that, on average, restaurants in the city are cited for health code violations at the same 3% baseline rate?

Hypothesis null hypothesis is 0.03 restaurants in city are cited for health code violations at same 3% baseline rate

Test Statistic the number of health code violations in 50 inspections

Use a Monte Carlo simulation (with at least 100,000 simulations) to calculate a p-value under this null hypothesis.



```
## [1] 8e-05
```

Our graph here is showing the Monte Carlo Simulation ran to simulate 50 of these Gourmet Bites restaurants being inspected and checking if they are cited at a 3% rate. When we compare our observed value of 8 citations out of the 50 inspections the p-value which is the mean of the results from this simulation as long as they are greater than our observed value of 8. That mean comes out to be 8×10^{-5}. This value is significantly lower than the basic standard of 0.05, this is suggesting that the 8 citations is out of the 3% range, and suggests that this is strong evidence against the null hypothesis that inspections resulting in citation is 3%.

Problem 3 - Evaluating Jury Selection for Bias

In this problem we are using information known on how jury selection works and information given by the county which anonymized racial and ethnic categories into 5 groups.

breakdown of county eligible jury pool

Group 1 - 30 % Group 2 - 25% Group 3 - 20% Group 4 - 15% Group 5 - 10%

Corresponding group counts for em-paneled Jurors in 20 Trials seen by judge (each jury has 12 jurors)

Group 1 - 85 Group 2 - 56 Group 3 - 59 Group 4 - 27 Group 5 - 13

We want to determine if the distribution of jurors is significantly different from county population proportions. If so does this suggest systematic bias in jury selection? What other explanations might exist, and how could you investigate further?

Do determine this we will be using a Chi-Squared goodness of fit test to find if the distribution of em-paneled jurors is significantly deviant from the expected county proportions.

Null Hypothesis this is the distribution of em-paneled jurors that match the expected distribution from the county jury pool.

Alt Hypothesis The distribution of em-paneled jurors that do not match the expectations from the county jury pool.

```
## [1] 72 60 48 36 24
```

```
## [1] 12.42639
```

```
## [1] 0.01506
```

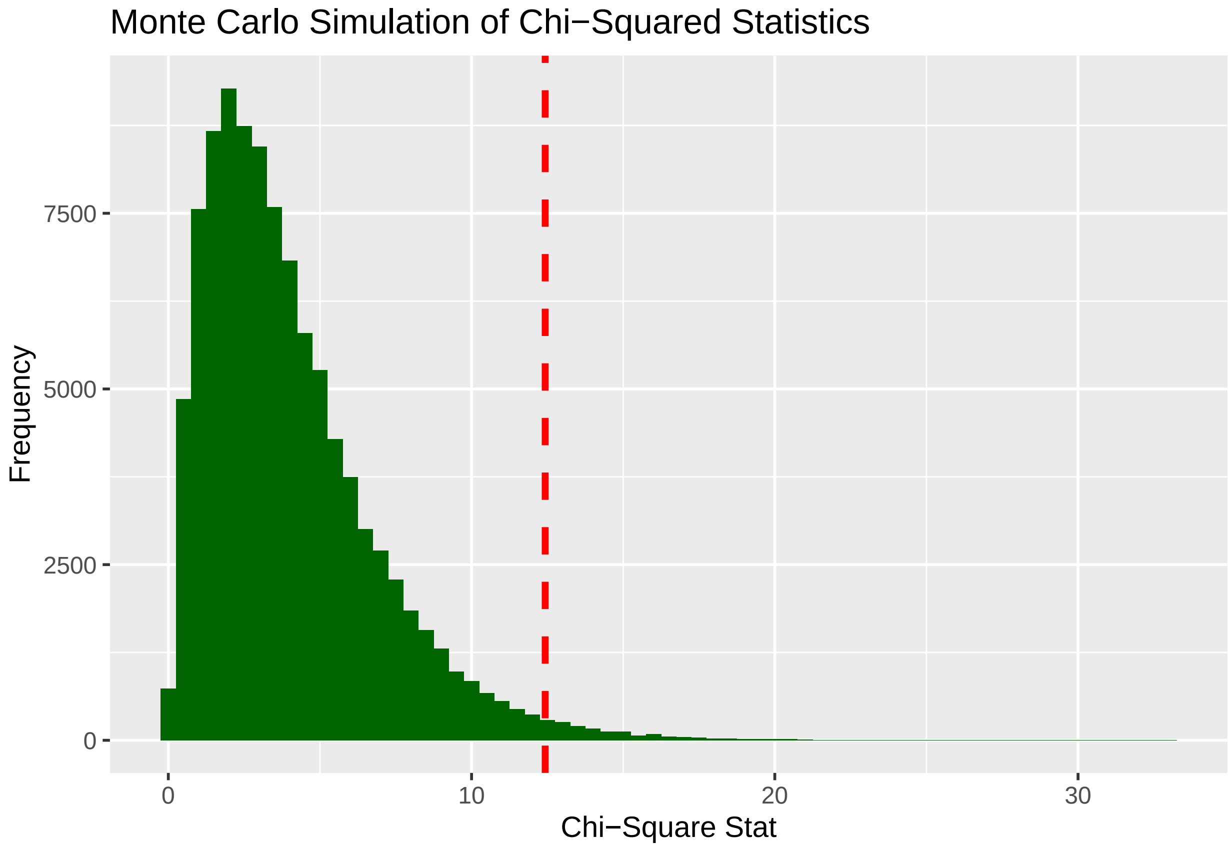


Table 1: Summary of chi Squared Statistics

	Min	Q1	Median	Mean	Q3	Max
25%	4.689436	17.64296	23.25522	26.95035	31.3806	199.2418

Problem 4: LLM watermarking

In problem 4 we are focusing on the watermarking of AI generated text.We are testing 10 sentences and testing if there is a watermark or not. 9 of them are normal sentences completely human made, and 1 was generated by a Large Language Model with a watermark. Which sentence is it?

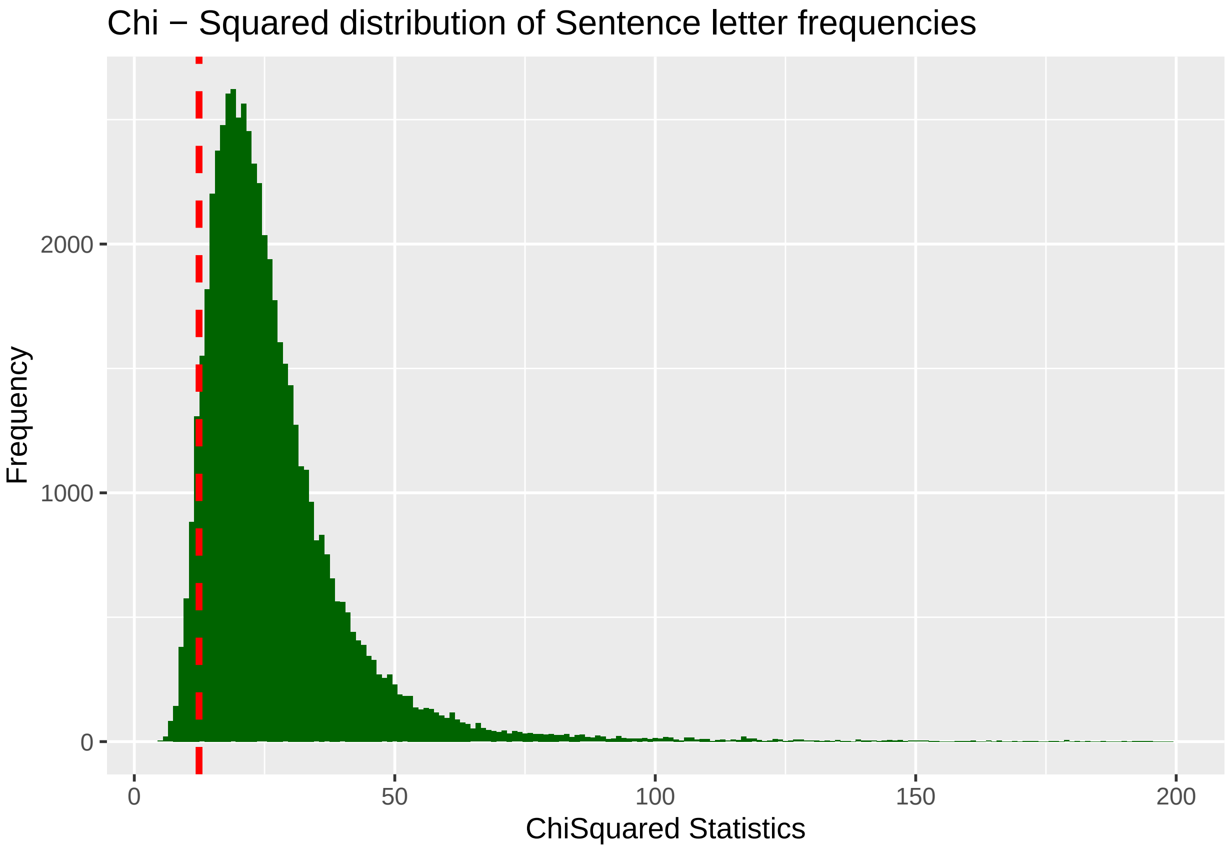
Part A: the null or reference distribution

Using data in the brown_sentences.txt file that contains collectionof english sentences fromt he Brown Corpus (well known and widely used text corpus in linguistics and natural language processing). Here we want to calculate a null distribution of hte chi-squared test statistic based on letter frequencies.

We want to know "what does the chi-squared statistic look like across lots of normal English sentences not generated by an LLM?"

Above we are given the minimum chi-statistic is 4.6 ranging all the way to roughly 200.

Then Below is the distribution that this summary comes from.



This graph suggests that the differences between the observed counts of each letter in each sentence and the expected counts given the data are large. This then can be defined as the range of Chi-Squared values you could expect to see in normal English sentences based on the predefined letter frequency. We will use this for part B.

Part B: checking for watermark