Homework 4

Ryann Liu

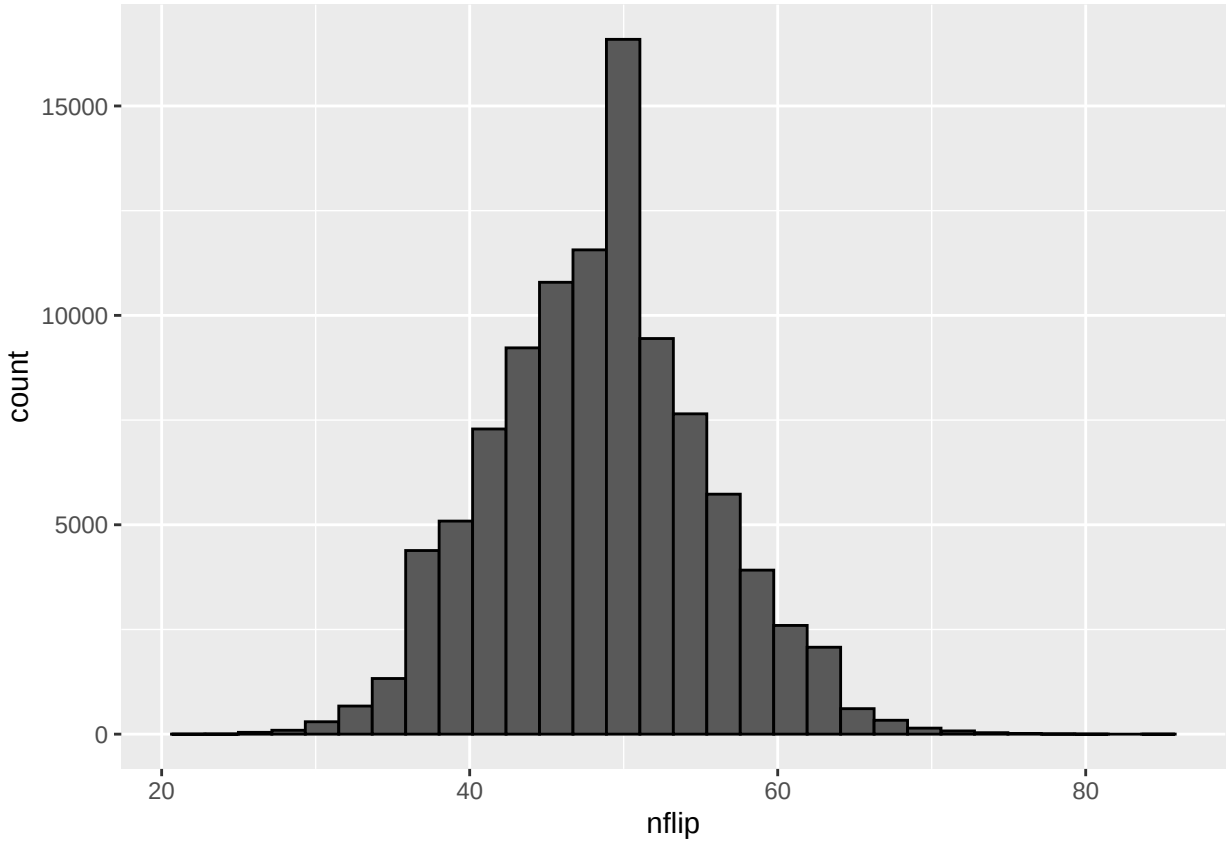
2025-02-16

Link to my Github repo: https://github.com/ryannrliu/HW4_Scott

Problem 1: Iron bank

1) Null Hypothesis: Over the long run, securities trades from the Iron Bank are flagged at the same 2.4% baseline rate as that of other traders.
2) Test Statistic: Number of flagged trades out of all 2021 trades.
3) P-Value: 0.002
4) Conclusion: Because the p-value is small (0.002), the null hypothesis is unsupported by the data and we can conclude that the SEC should probably investigate the flagged trades.

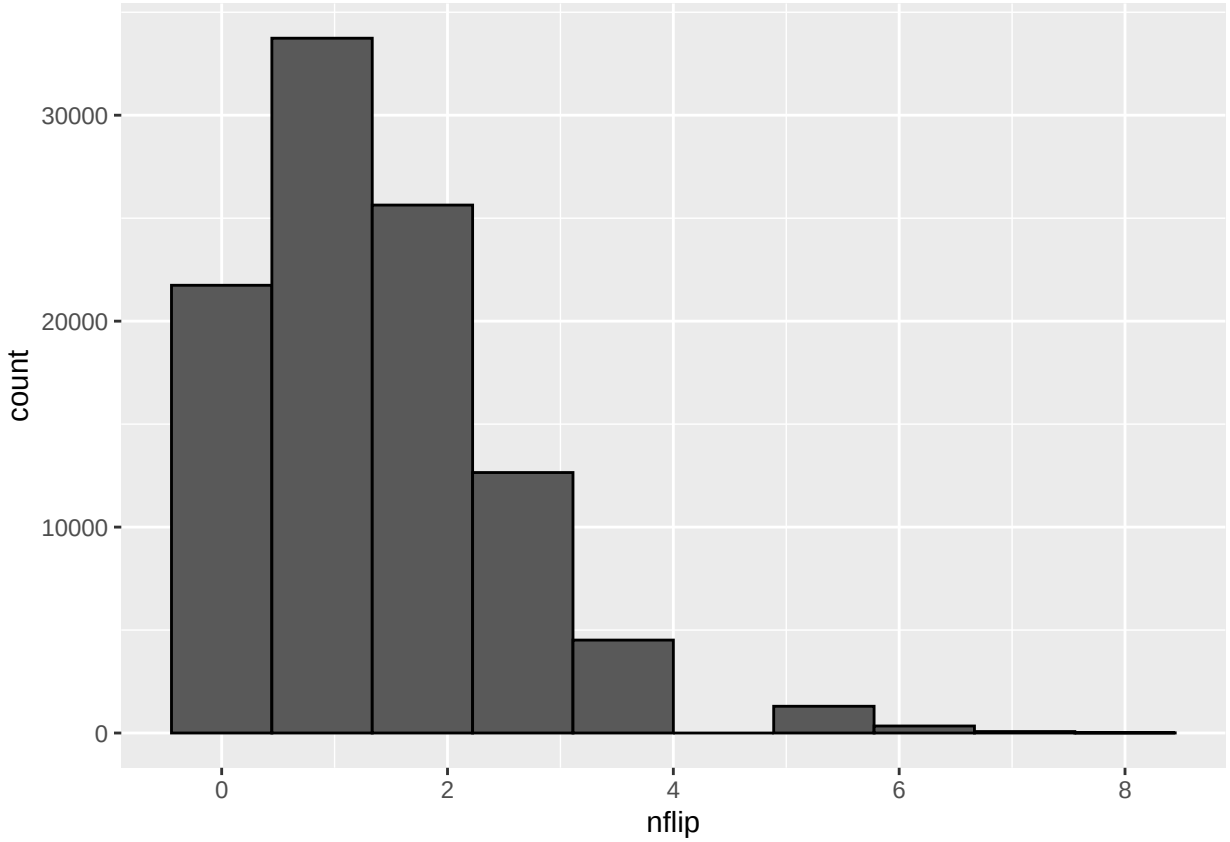
```
## [1] 0.00202
```



Problem 2: Health Inspections

1) Null Hypothesis: On average, restaurants in the city are cited for health code violations at the same 3% baseline rate.
2) Test Statistic: Number of health code violations for Gourmet Bites.
3) P-Value: 0.0001
4) Conclusion: Because the p-value is very low, we can conclude that there is evidence within the data that Gourmet Bites has a significantly higher average of health code violations.

```
## [1] 0.00011
```



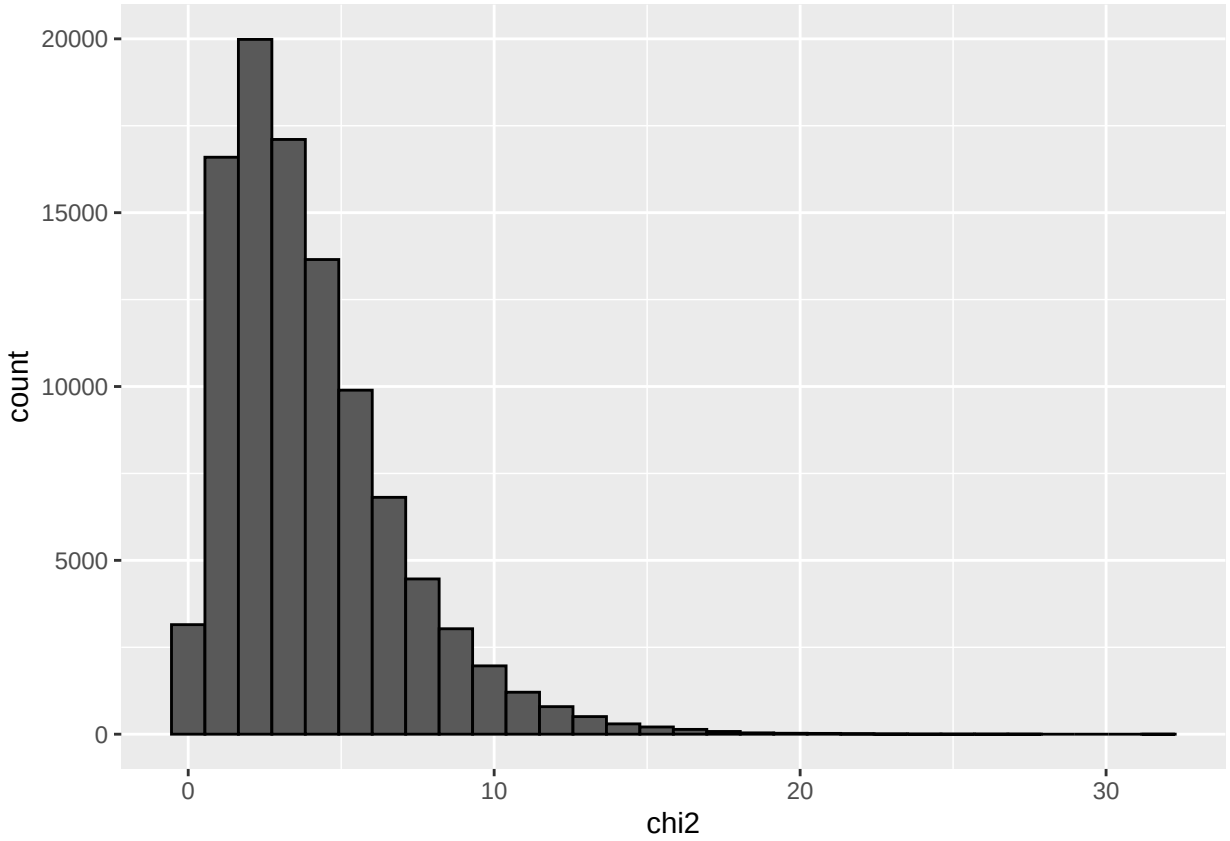
Problem 3: Evaluating Jury Selection for Bias

To determine whether the distribution of jurors empaneled by this judge is significantly different from the county's population proportions, I calculated a chi squared statistic that compared the expected demographic breakdown in percentage to the observed breakdown of groups for the empaneled jurors in 20 trials. My null hypothesis is: The distribution of jurors empaneled by this judge is similar to the county's population proportions. My test statistic is: The expected number of members in each group.

I used 100,000 Monte Carlo simulations to create an expected chi squared statistic, and compared it to the observed to calculate a p-value of 0.0141. Below is a table of the expected number members based on the given percentages, as well as a distribution of the simulated chi square values. Because the p-value is relatively low, I will conclude that there is a significant difference, which could possible suggest systematic bias. Also, the table of observed vs. expected shows significantly differing numbers, which also aligns with

this conclusion. This could also depend on the demographic of the specific area, as if the population has significantly less people in one group, then this could affect the judge's ability to have sufficient representation of that group. One could investigate further by also taking the overall populations demographic into account.

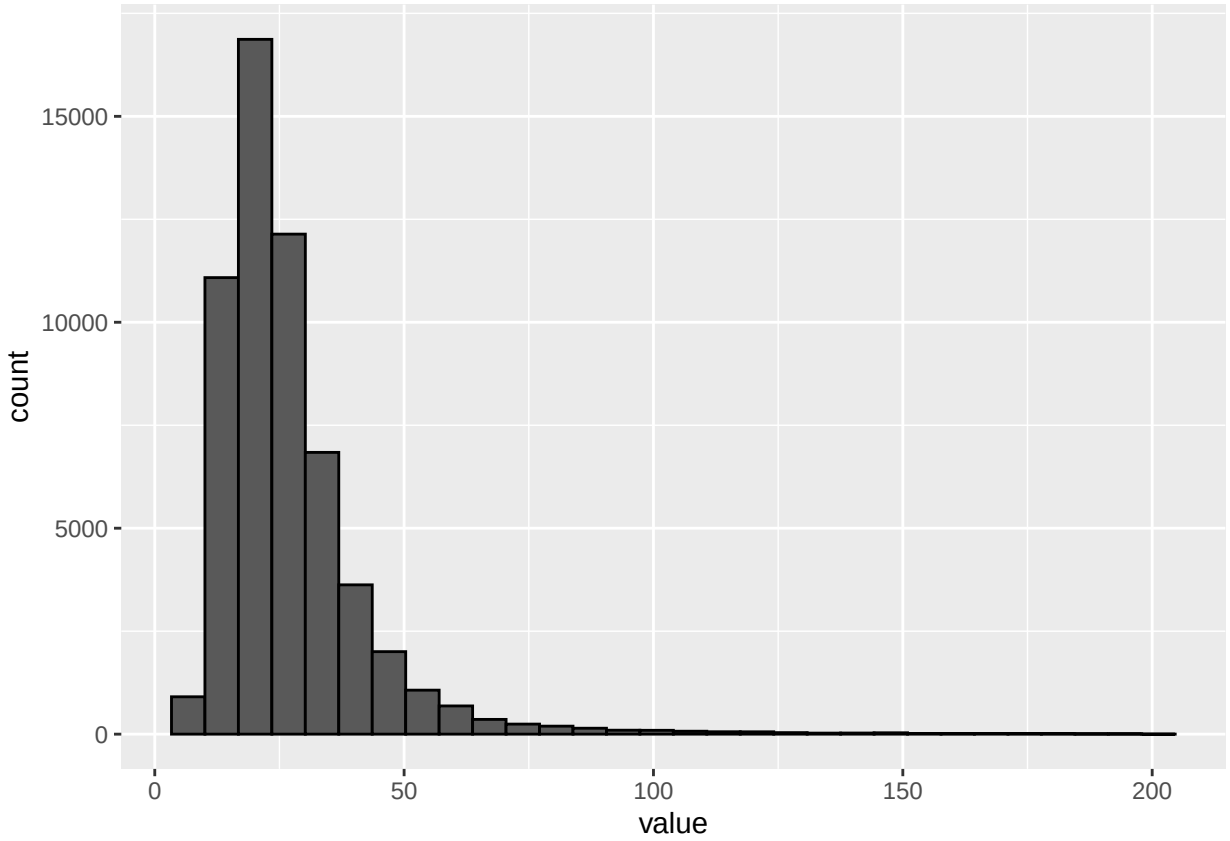
```
## # A tibble: 5 x 2
##   observed expected
##      <dbl>    <dbl>
## 1       85       72
## 2       56       60
## 3       59       48
## 4       27       36
## 5       13       24
```




```
## [1] 0.01443
```

Problem 4: LLM Watermarking

Below is the null distribution of the chi-squared test statistic based on letter frequency of English sentences from the Brown Corpus.



Here, I calculated the chi-squared statistic for 10 sentences. I then compared the values I got to the previous null chi-squared values, and calculated the following p-values. Based on the given p-values, sentence 6 was most likely created by an LLM due to the significantly lower p-value.

```
## # A tibble: 10 x 2
##    index value
##    <int> <dbl>
##  1     1 0.513
##  2     2 0.926
##  3     3 0.076
##  4     4 0.489
##  5     5 0.484
##  6     6 0.009
##  7     7 0.328
##  8     8 0.988
##  9     9 0.084
## 10    10 0.059
```