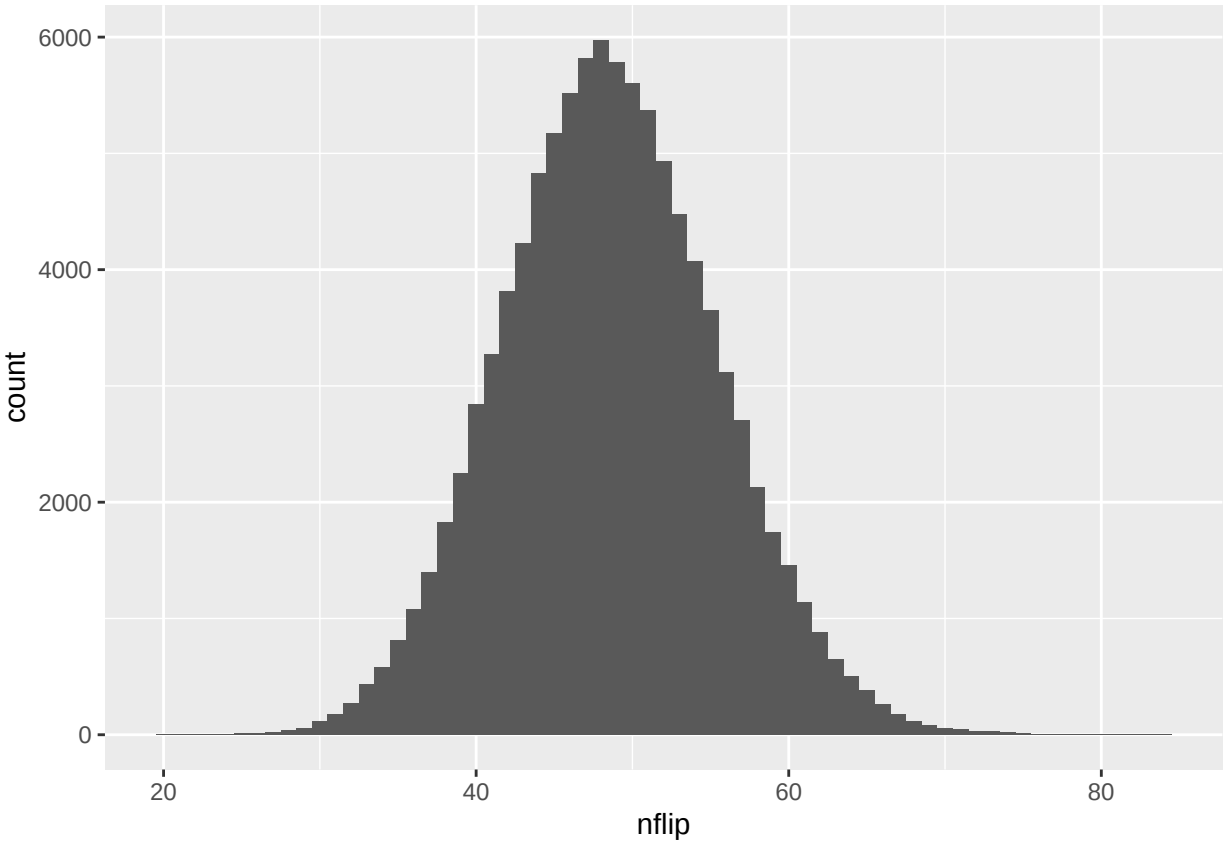
Homework4

Daniel Wu (EID: djw3627)

2025-02-12

To access my GitHub repository, click here: https://github.com/DanielWu3627/SDS315/tree/main. Please check the file named **Homework4.Rmd**.

Problem 1



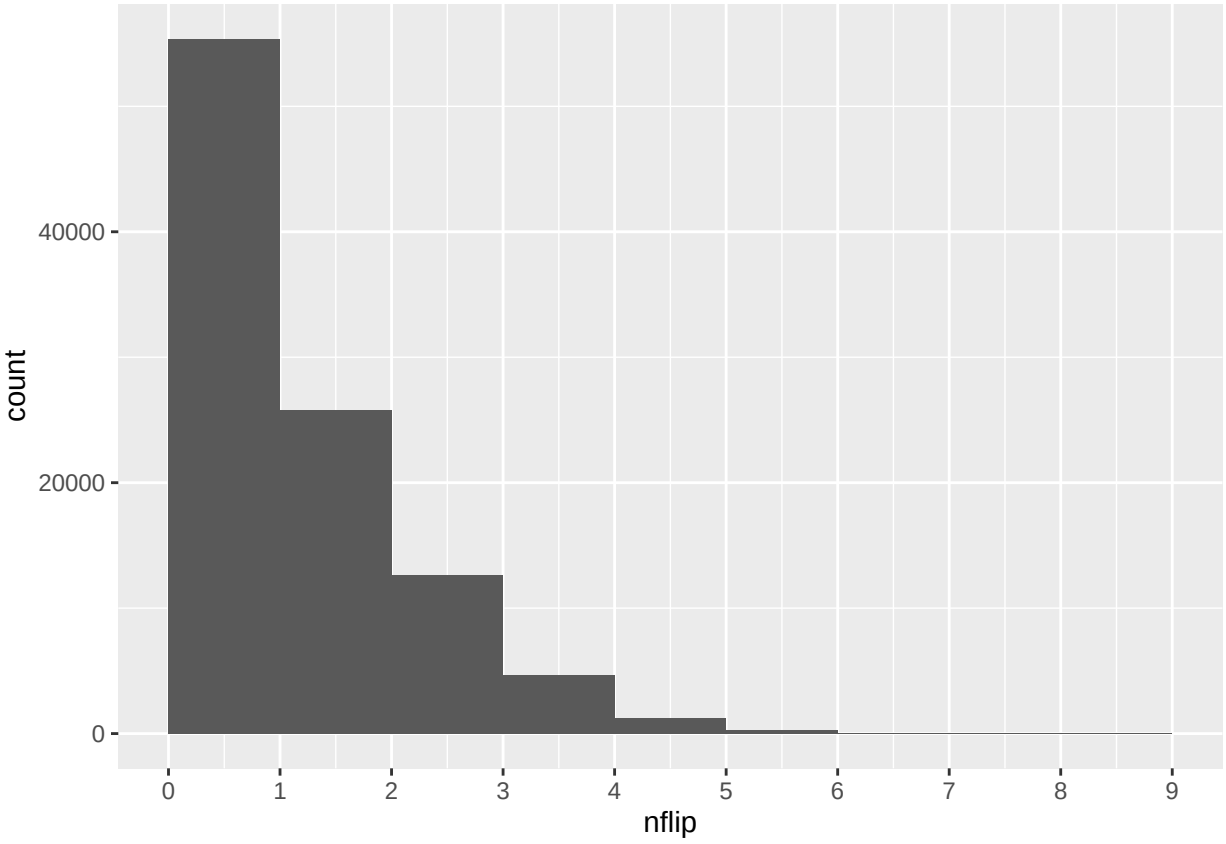

```
## [1] "The p-value is 0.002"
```

The histogram shows the distribution of the number of flagged trades out of the 2021 trades, assuming the null hypothesis is true.

- The null hypothesis I am testing is that the security trades from the Iron Bank being flagged is at the baseline rate of 2.4%.
- The test statistic I used to measure the evidence against the null hypothesis is the number of flagged trades. In the observed data, there are 70 flagged trades out of 2021.

- According to the histogram showing the probability distribution of the test statistic, there is a very small amount of chance of having 70 or more flagged trades.
- The p-value is about 0.002, meaning that assuming the null hypothesis is true, there is only a 0.2% chance that the number of trades flagged by chance will be 70 or higher. Therefore, the null hypothesis does NOT look plausible in light of the data. There are too many flagged trades from the Iron Bank. The SEC should conduct further investigation!!

Problem 2




```
## [1] "The p-value is 1e-04"
```

The histogram shows the distribution of the number of health code violation out of 50 inspections, assuming the null hypothesis is true.

- The null hypothesis I am testing is that the rate at which Gourmet Bites is being flagged for health code violations is 3%, just like all restaurant inspections.
- The test statistic I used to measure the evidence against the null hypothesis is the number of health code violations. In this data, 8 health code violations are reported for Gourmet Bites out of 50 inspections.
- According to the histogram showing the probability distribution of the test statistic, there is a very small chance that the number of health cde violations is at or beyond 8.
- The p-value is about 0.00011, meaning that assuming the null hypothesis is true, there is only a 0.01% chance that the number of violations reported for Gourmet Bites by chance will be 3% or higher. Therefore, the null hypothesis does NOT look plausible in light of the data.

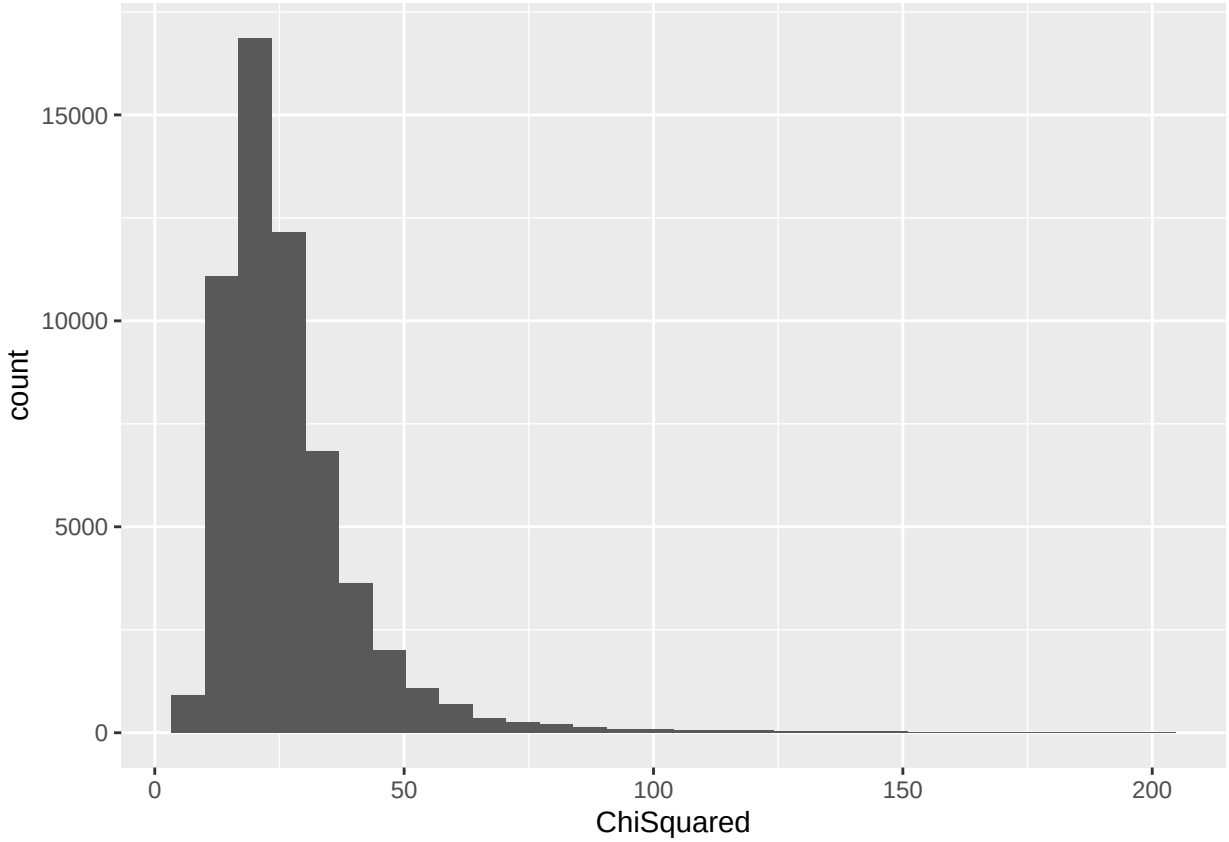
Problem 3

```
## [1] "The p-value is 0.01388"
```

- The null hypothesis I am testing is that the distribution of juries empaneled by the judgew is now significantly different from the county's population proportions.
- The test statistic is the chi-squared statistic. In the observed data, the chi-squared statistic is calculated using the group counts from 20 trials.
- The null distribution is obtained through 100,000 Monte Carlo simulations of group counts of jurors. Then, chi-squared statistics is calculated for each simulation.
- Given that the p-value is about 0.01, the null hypothesis is rejected. there is a significant difference in the distribution of jurors empaneled by the judge from the county's population proportions. This indeed suggests systematic bias in jury selection. However, there is a non-random pool composition. For instance the pool of available jurors might not reflect the expected demographic distribution due to socioeconomic, geographic, or other factors. For example, if certain groups are less likely to register for jury duty or more likely to be excused, this could affect representation. To investigate further, we can examine selection procedures to see if any steps disproportionately exclude certain groups. Furthermore, look at multiple jury pools empaneled by different judges to determine whether the bias is persistent or if this was an anomaly.

Problem 4

Part A



The histogram shows the null distribution of the chi-squared statistics from all sentences in the Brown Corpus.

Part B

```
##    Sentence P-Value
## 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
```

The sentence that is watermarked is sentence 6. Given the p-values, sentence 6 has a p-value of 0.009, which is the smallest p-value, indicating its letter frequency is most significantly different from normal English Letter frequency.