Analysis of Emergency Service Claims by Income Category

Introduction

Healthcare utilization varies significantly across socioeconomic groups, due to disparities in healthcare access, overall health shaped by quality of life, and financial ability to afford care. Access to emergency healthcare services is particularly critical, as it can be a matter of life or death. However, disparities in socioeconomic status often dictate an individual's ability to obtain such care.

This study explores the relationship between income levels and the frequency of emergency service claims. Using the 2022 dataset “Medicare Physician & Other Practitioners - by Provider and Service” from the Centers for Medicare and Medicaid Services, combined with ZIP code-level Adjusted Gross Income (AGI) data from the IRS, the following question is answered: how does the frequency of emergency service claims differ between high-poverty and affluent ZIP codes? By comparing claim frequencies across these socioeconomic groups, this analysis aims to identify disparities in access to emergency services and patterns of healthcare utilization.

Statistical Approach

Since the dataset did not explicitly define affluent and high-poverty categories, the first step after cleaning and preparing the data for emergency services analysis was to create these classifications. The IRS data included AGI stubs ranging from 1 to 6, where 1 represents the lowest income bracket (under $25,000) and 6 represents the highest income bracket (over $200,000). ZIP codes with the majority of tax returns falling within AGI stubs 1 and 2 were categorized as “high-poverty”, while ZIP codes with the majority of returns in stubs 5 and 6 were categorized as “affluent”, leaving stubs 3 and 4 for the “middle” category. This grouping allowed for 659 observations in the affluent category and 6,846 observations in the high-poverty category. I then sampled 659 out of the 6,846 to allow for more valid one-to-one comparisons between the two categories. This also preserves the focus of the research question, which is to examine differences in claim behavior between low-income and high-income populations.

Next, I performed exploratory data analysis to identify any major issues in the data. I first calculated aggregate descriptive statistics of each group, specifically central tendency measure calculations, in order to understand the spread and averages of the data. I also created visualizations of each group's distribution, which revealed that the total claims for both the high-poverty and affluent groups were heavily right-skewed. I applied a log transformation to each group, which brought the data closer to normality.

With the data prepared, I performed diagnostic checks to assess the assumptions of linearity, homoskedasticity, and normality, in order to ensure the suitability of the linear model. The log transformation effectively addressed the issue of skewed claim data, and there were no significant violations of linearity or homoskedasticity detected. Although the normality

assumption showed some deviations in the tails, the large sample size of the dataset makes it reasonable to assume that normality holds, due to the Central Limit Theorem. With these checks successfully completed, I performed a simple regression t-test to compare the two income groups, and utilized heteroskedasticity-consistent standard errors to enhance the robustness of the analysis.

Results

The goal of the Two-Sample T-Test was to determine whether the difference in the frequency of claims between the high-poverty and affluent groups was statistically significant. Here, the null hypothesis states that there is no difference in the mean log-transformed total claims between the affluent and high-poverty groups, while the alternative hypothesis states there is a statistically significant difference.

The p-value for the test was 0.01572, which is less than the significance level. This provides evidence to reject the null hypothesis and conclude that the true difference in means between the two groups is not 0. The test results reported that the mean of the log-transformed total claims for the affluent group is 6.011561, corresponding to approximately 408 claims. Similarly, the mean of the log-transformed total claims for the high-poverty group is 6.179731, translating to an average of 483 claims.

The confidence interval results indicate that the high-poverty group has 18.31% more claims than the affluent group. Specifically, the 95% confidence interval suggests that the high-poverty group has between 3.1% and 35.7% more claims than the affluent group.

Conclusion

This analysis demonstrates a statistically significant difference in the frequency of emergency service claims between high-poverty and affluent ZIP codes, with individuals in high-poverty areas making approximately 18% more claims than those in affluent areas. However, certain assumptions and limitations must be considered when interpreting these results.

One key assumption is the classification of ZIP codes into the high-poverty and affluent categories. While groups were constructed based on IRS AGI data to facilitate precise comparisons, they are ultimately arbitrary groupings. Additionally, while the statistical analysis focused on per-group averages, the sheer difference in the original count of observations — ten times higher in high-poverty ZIP codes — may suggest broader disparities. This could reflect a greater prevalence of health issues in lower-income populations, potentially driving a higher demand for emergency services compared to higher-income groups.

A limitation of this study is the sample size used for the t-test, which included 659 observations per group. While this is sufficient to achieve reliable statistical results, a larger sample size per group would improve the accuracy of the results, providing estimates that more closely represent the true population values.

These findings highlight a disparity in emergency healthcare utilization that aligns with known socioeconomic inequities. Future research could expand on this by exploring underlying factors contributing to these differences, such as access to preventive care, provider availability, and population health challenges. Improving healthcare access and education about healthcare in underserved areas might address the disparities observed in this analysis.

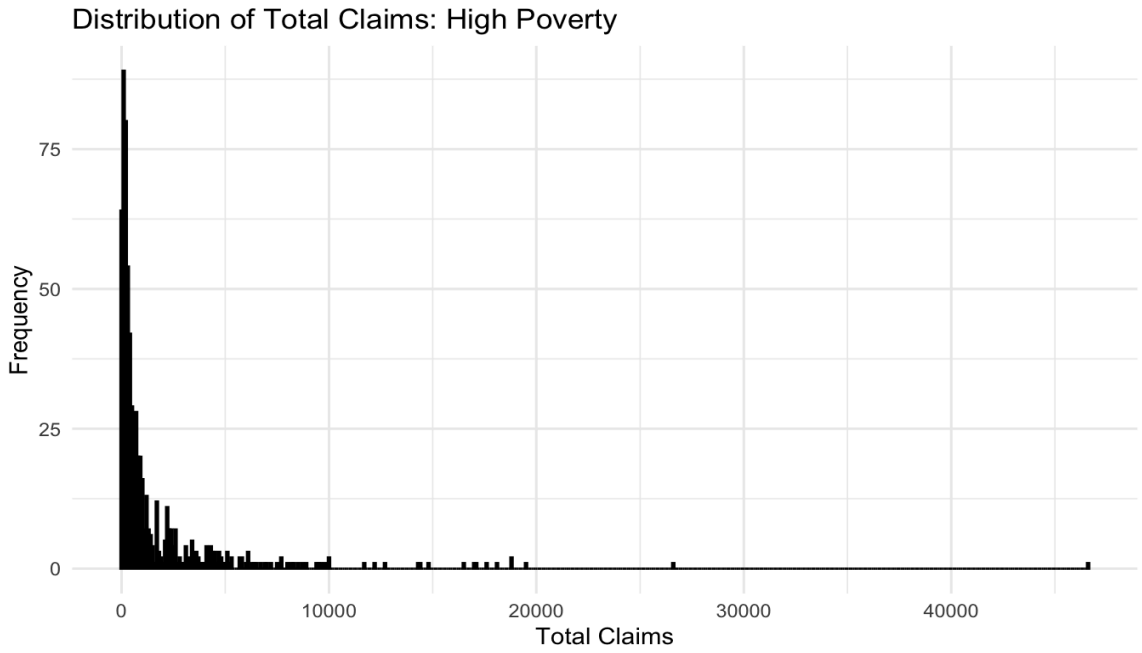
Appendix

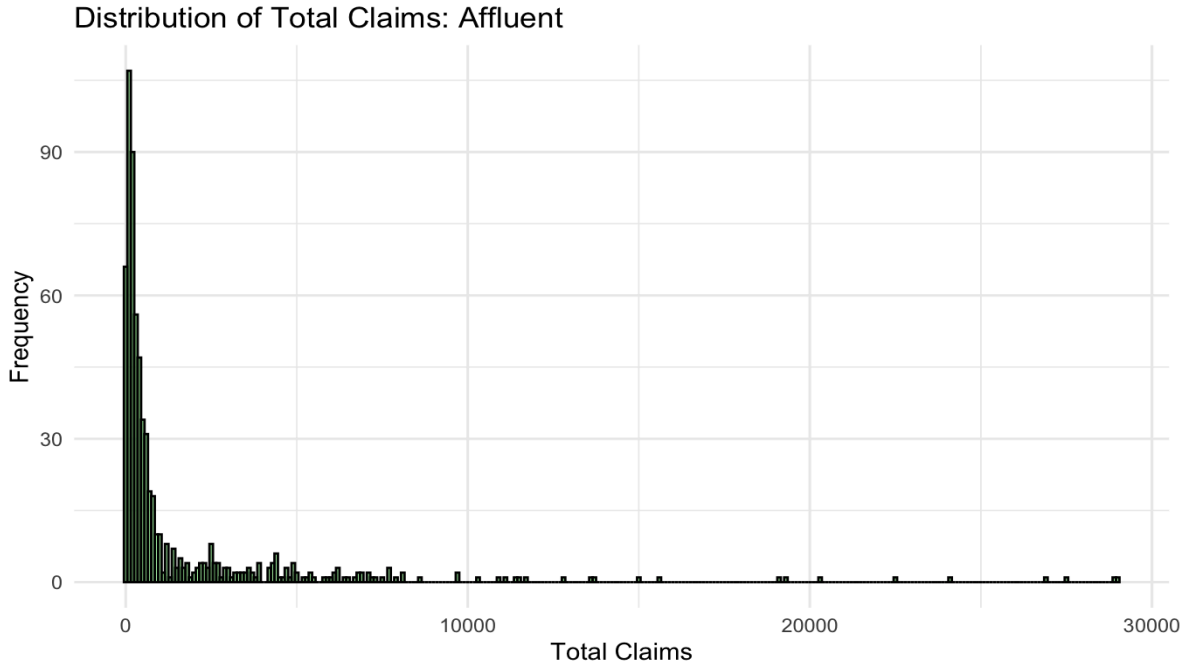
GitHub Link: https://github.com/snigdhapakala/506_final_project

Summary Statistics Table

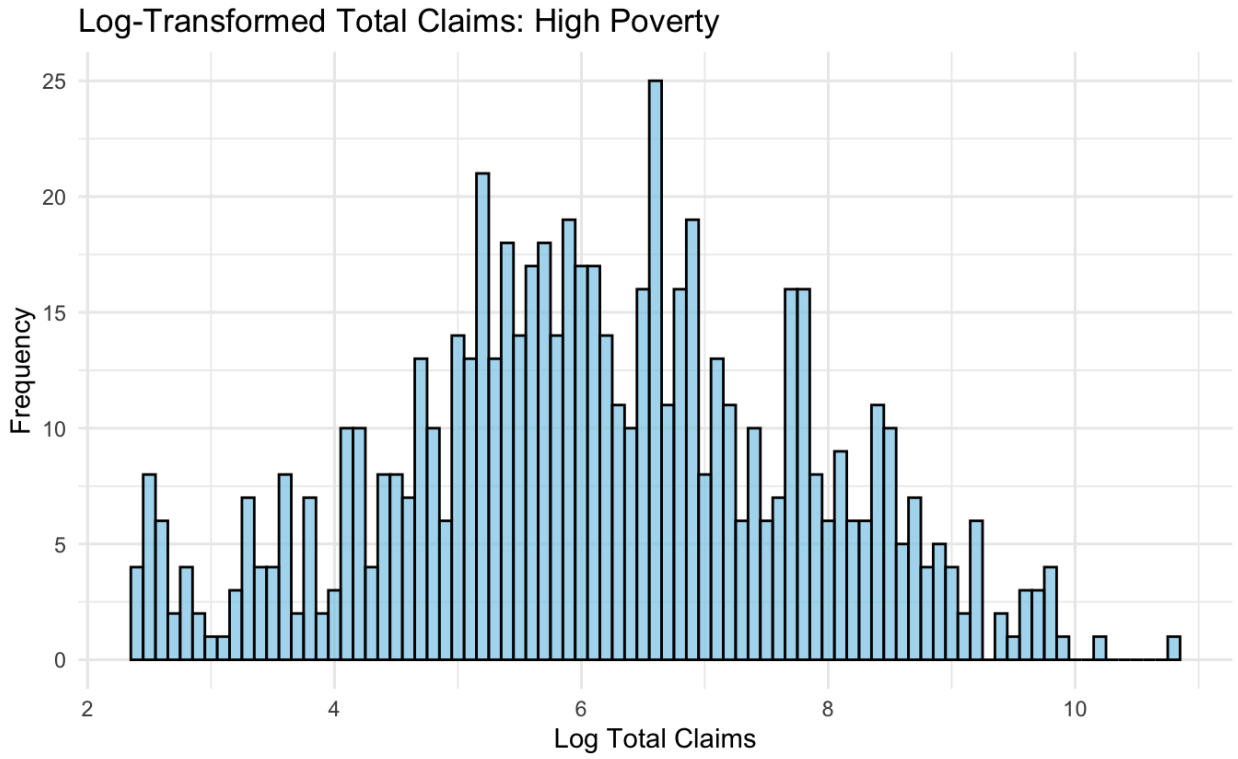
Income Category	Number of ZIPs	Total Claims	Claims/Beneficiary	Mean Claims/ZIP	Median Claims/ZIP	SD Claims
affluent	659	1064216	1.03	1614.90	368.0	3536.35
high_poverty	6846	12631448	1.04	1845.08	461.5	4021.98

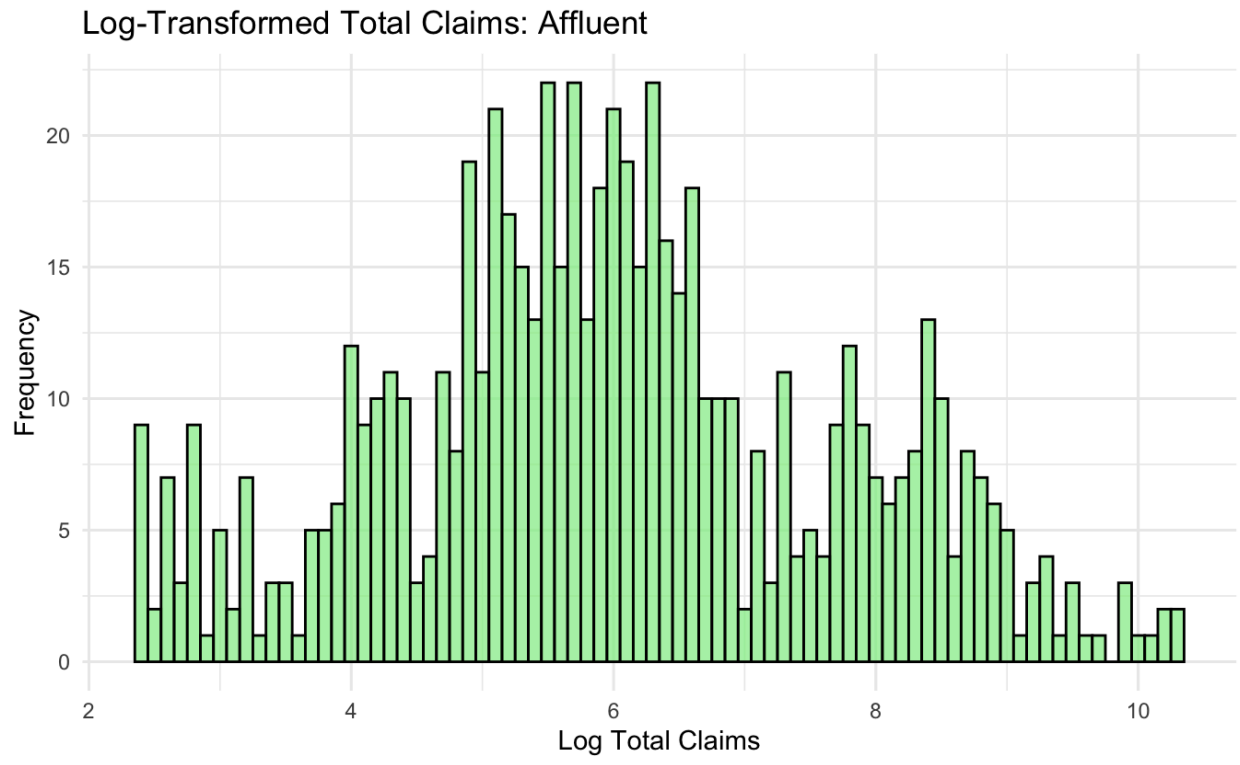
Total Claims Distributions Per Income Category



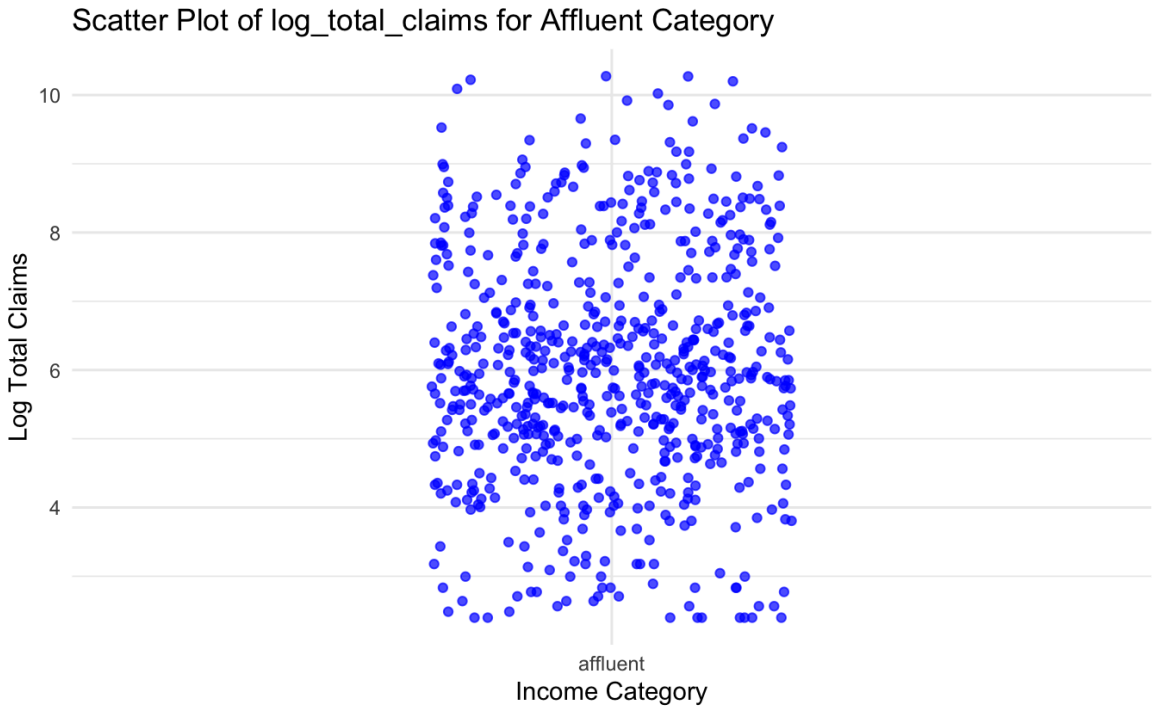


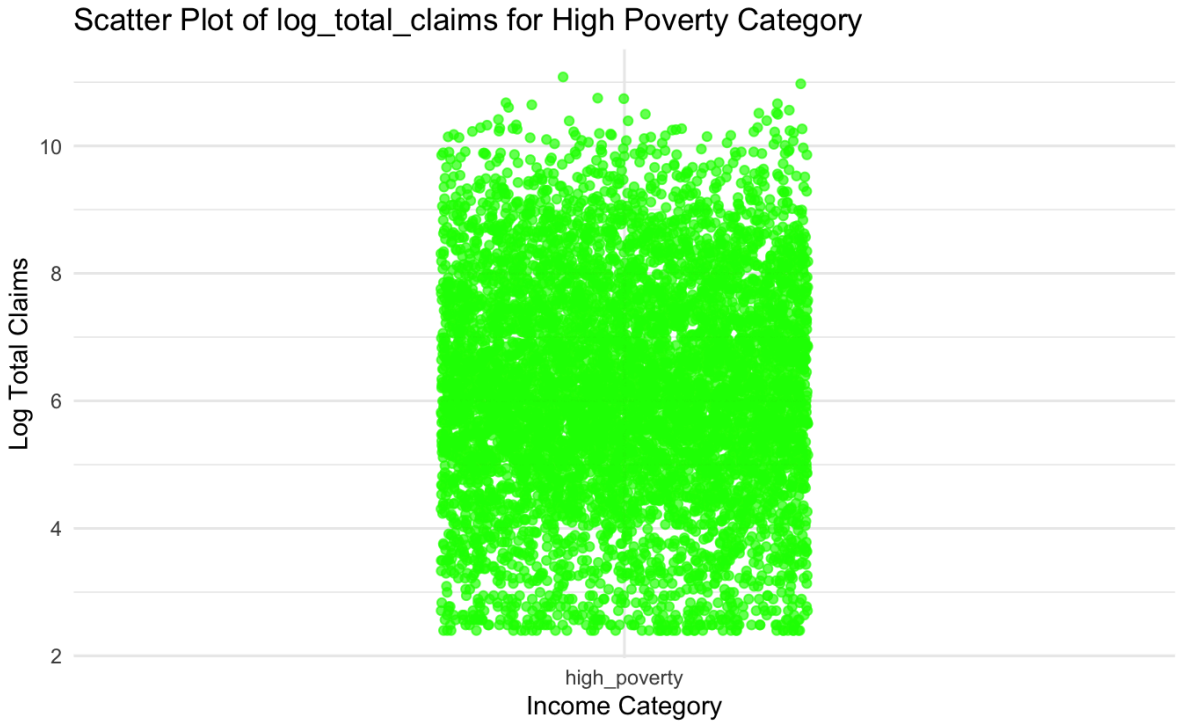
Log-Transformed Claims Per Income Category



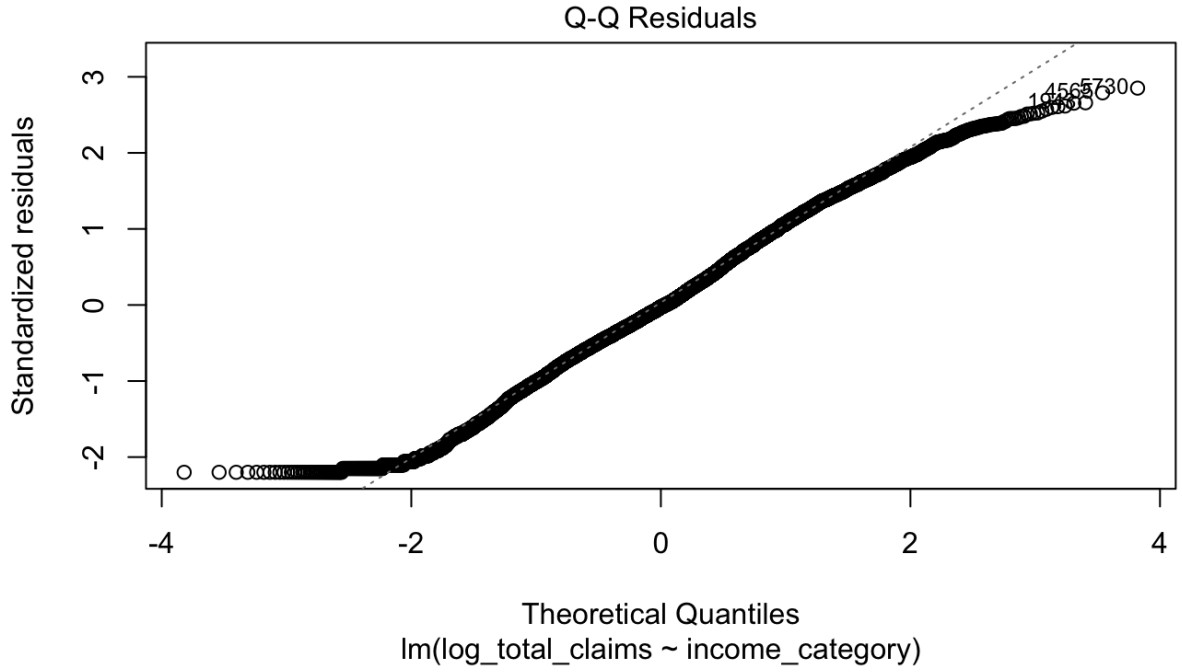


Linearity Checks Per Income Category





Normality Check



Output of Welch Two Sample t-test

```
        Welch Two Sample t-test

data:  log_total_claims by income_category
t = -2.4205, df = 793.17, p-value = 0.01572
alternative hypothesis: true difference in means between group affluent and group high_poverty
is not equal to 0
95 percent confidence interval:
 -0.30455484 -0.03178608
sample estimates:
    mean in group affluent mean in group high_poverty
                  6.011561                   6.179731
```