

**7COM1079-0901-2024 - Team Research and Development Project**

**Final report title: Statistical Analysis of Child Poverty Rate Differences: A Comparative Study Between California and Texas**

Group ID: A81

Dataset number: DS072

**Prepared by:**

Lakshmi Venkata Sai Pampana (23024109) ,

Karthik Kumar Elluru (23028599) ,

Sai Mithil Kancharla (23030901) ,

Lal John Basha Shaik (23029902),

Gopi Krishna Malle(23038251).

University of Hertfordshire

Hatfield, 2024

Table of Contents

[**1.** **Introduction** 3](#_Toc186768933)

[**1.1.** **Problem statement and research motivation** 3](#_Toc186768934)

[**1.2.** **The data set** 3](#_Toc186768935)

[**1.3.** **Research question.** 3](#_Toc186768936)

[**1.4.** **Null hypothesis and alternative hypothesis (H0/H1)** 4](#_Toc186768937)

[**2.** **Background research** 4](#_Toc186768938)

[**2.1.** **Research papers (at least 3 relevant to your topic / DS)** 4](#_Toc186768939)

[**2.2.** **Why RQ is of interest (research gap and future directions according to the literature)** 4](#_Toc186768940)

[**3.** **Visualisation** 5](#_Toc186768941)

[**3.1.** **Appropriate plot for the RQ output of an R script** 5](#_Toc186768942)

[**3.2.** **Additional information relating to understanding the data** 6](#_Toc186768943)

[**3.3.** **Useful information for the data understanding** 6](#_Toc186768944)

[**4.** **Analysis** 6](#_Toc186768945)

[**4.1.** **Statistical test used to test the hypotheses and output** 6](#_Toc186768946)

[**4.2.** **The null hypothesis is rejected/not rejected based on the p-value** 6](#_Toc186768947)

[**5.** **Evaluation – group’s experience at 7COM1079-A81** 7](#_Toc186768948)

[**5.1.** **What went well** 7](#_Toc186768949)

[**5.2.** **Points for improvement** 7](#_Toc186768950)

[**5.3.** **Group’s time management** 7](#_Toc186768951)

[**5.4.** **Project’s overall judgement** 7](#_Toc186768952)

[**5.5.** **Comment on GitHub log output:** 8](#_Toc186768953)

[**6.** **Conclusions** 8](#_Toc186768954)

[**6.1.** **Results explained** 8](#_Toc186768955)

[**6.2.** **Interpretation of the results** 8](#_Toc186768956)

[**6.3.** **Reasons and/or implications for future work, limitations of your study** 8](#_Toc186768957)

[**7.** **Reference list** 9](#_Toc186768958)

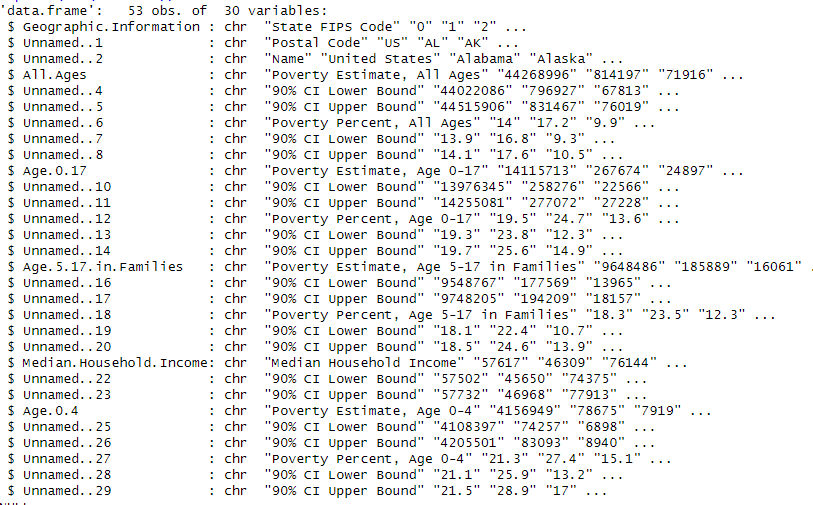
[**8.** **Appendices** 9](#_Toc186768959)

1. **Introduction**
   1. **Problem statement and research motivation**

Millions of children are impacted by child poverty, which is still a major problem in the US. Our study was driven by the need to compare The incidence of child poverty in Texas and California, two states with different socioeconomic systems. Children's poverty levels can have a significant impact on their development and long-term prospects, as demonstrated by research like those conducted by Laird et al. (2017). Our goal in comparing these two states is to find important differences that could lead to better child poverty policy.

* 1. **The data set**

The "US Public Food Assistance 1 - WIC" (est16us.csv) dataset, which was obtained from Kaggle.com and is based on 2016 Census Bureau estimates, was utilized for investigation. Confidence intervals are included to guarantee statistical reliability, along with state-level statistics on child poverty rates for children ages 0-17. Geographic identifiers and state-specific data are also included in the dataset, which is crucial for this study's comparison of poverty rates in Texas and California.



**Figure 1: Loading Dataset**

* 1. **Research question.**

**"Does Texas' mean poverty rate for children aged 0–17 differs from California's in a way that is statistically significant?"**

We will examine the child poverty rates in Texas and California using data from the 2016 Census Bureau in order to answer this research question. The state (Texas vs. California) is the independent variable, while the child poverty rate is the dependent variable. By computing the confidence intervals for the poverty rates in the two states, we will conduct a statistical study. We can ascertain whether there is a significant difference by comparing these intervals. There will be a statistically significant difference between the child poverty rates in Texas and California if the two states' confidence intervals do not overlap.

* 1. **Null hypothesis and alternative hypothesis (H0/H1)**

We came up with two theories in order to respond to the research question:

**Null Hypothesis (H₀)**: The mean poverty rates for children ages 0-17 in Texas and California do not differ statistically significantly. This suggests that there is no actual discrepancy in poverty rates between the two states and that any reported difference is the result of random volatility.

**Alternative Hypothesis (H₁)**: The mean poverty rates for children ages 0-17 in Texas and California differ statistically significantly. This implies that there is a significant difference in the rates of child poverty between the two states and that the difference is not the result of chance.

1. **Background research**
   1. **Research papers**

Several important studies that shaped our comprehension of the research topic were the main focus of our team's literature review. These comprised:

**Fan et al. (2023) "The Use of Charitable Food Assistance among Low-Income Households in the United States"**: This study looked at how often low-income households used food aid using statistical analysis using R. In keeping with our focus on comprehending poverty reduction, the study underlined the value of statistical techniques in examining social issues and the efficiency of food assistance programs in lowering poverty.

**Blokhin et al. (2023) "Hypothesis Testing Using R"**: With an emphasis on medical research, this publication offered helpful advice for statistical hypothesis testing with R. The authors' comparison of CT scan procedures for COVID-19 patients demonstrated how R may be used to do precise statistical tests a notion that we thought was helpful for examining data pertaining to poverty.

**Walker (2019) "Hypothesis Tests"**: An extensive examination of statistical hypothesis testing techniques and p-value interpretation was given in this study. It provided insightful advice on how to choose suitable statistical tests, which improved our comprehension of hypothesis testing in relation to child poverty rates.

These studies gave our study a thorough basis and guided our analytical approach to state-by-state comparisons of poverty rates.

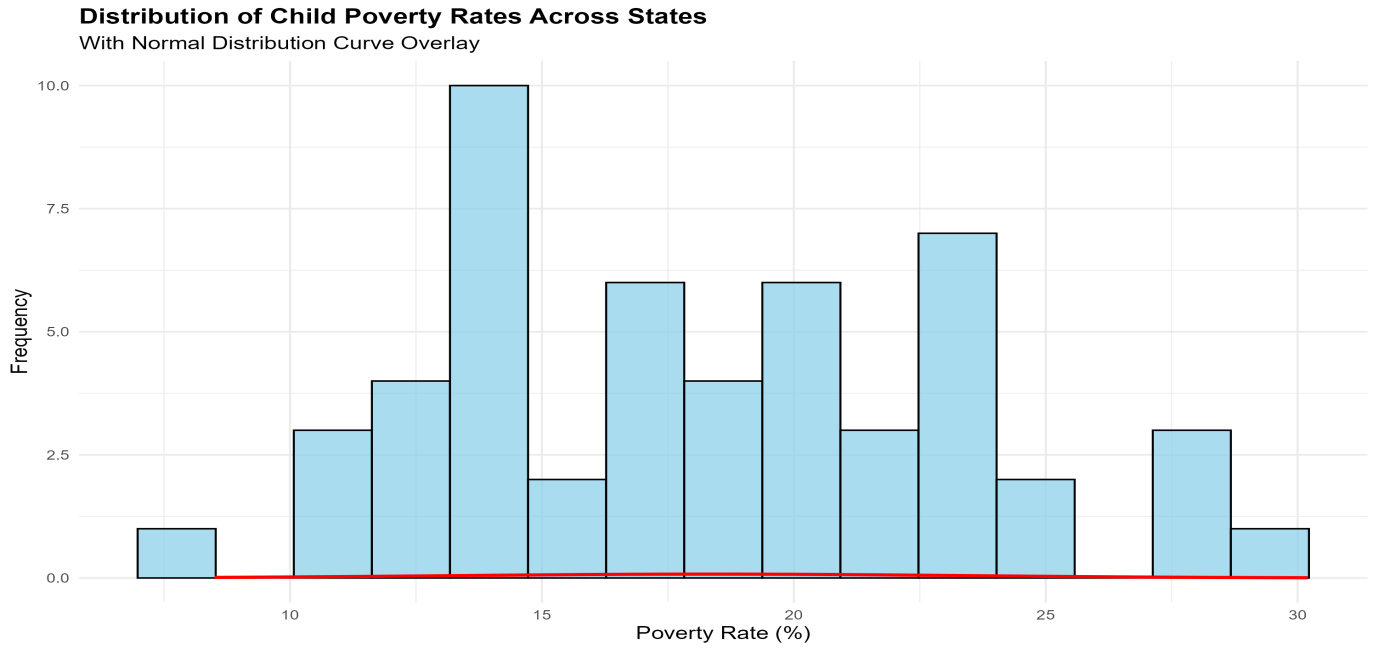
* 1. **Why RQ is of interest (research gap and future directions according to the literature)**

Given the current inadequacies in statistical analysis and the paucity of state-level studies on child poverty, the research question we investigated is highly relevant. There is a dearth of thorough research comparing state-specific determinants, especially in states with disparate policy approaches like California and Texas, even though a large portion of the literature concentrates on national poverty trends. These two states offer a chance to examine how different policies affect child poverty rates because of their divergent socioeconomic approaches.

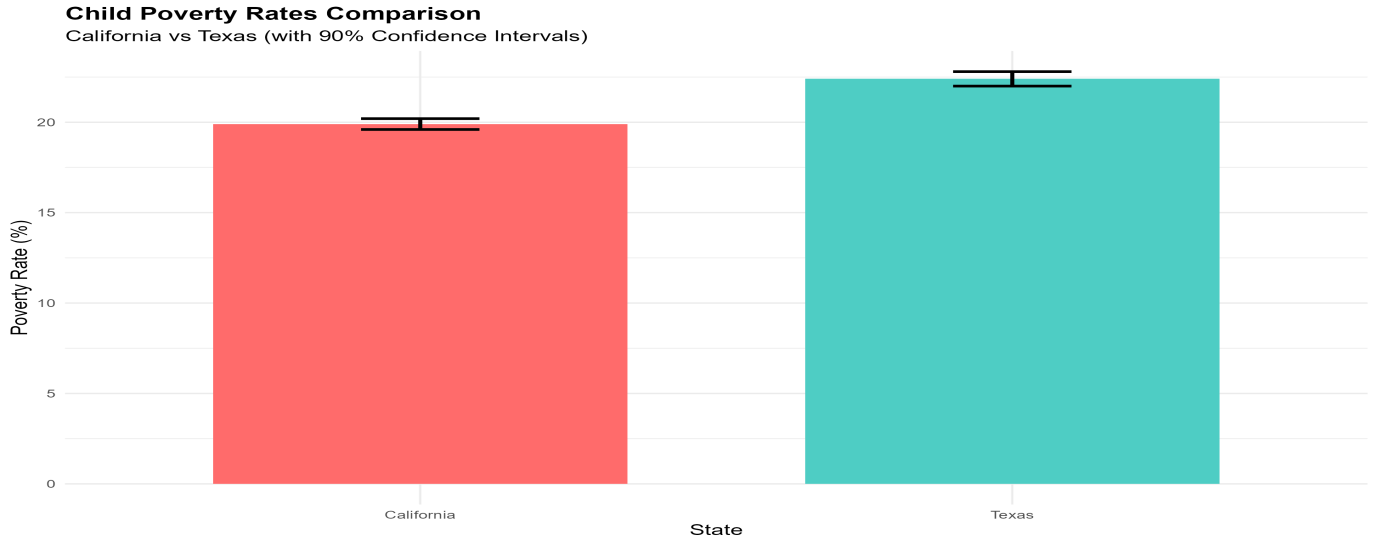
Walker (2019) and Blokhin et al. (2023) are two recent papers that highlight the increasing complexity of research data and the difficulties researchers encounter when choosing and applying suitable statistical approaches. Blokhin et al. (2023) address the significance of easily accessible statistical tools, whereas Walker's (2019) study draws attention to typical misunderstandings regarding the interpretation of p-values. There is still a gap in the practical use of these techniques, particularly in social research, even with the availability of sophisticated tools like R.

By concentrating on child poverty disparities at the state level, our project seeks to close this gap and provide insights into how policy differences between Texas and California impact poverty outcomes. This strategy could be expanded in future studies by incorporating additional states or by examining data from several years to identify long-term patterns.

1. **Visualisation**
   1. **Appropriate plot for the RQ output of an R script**

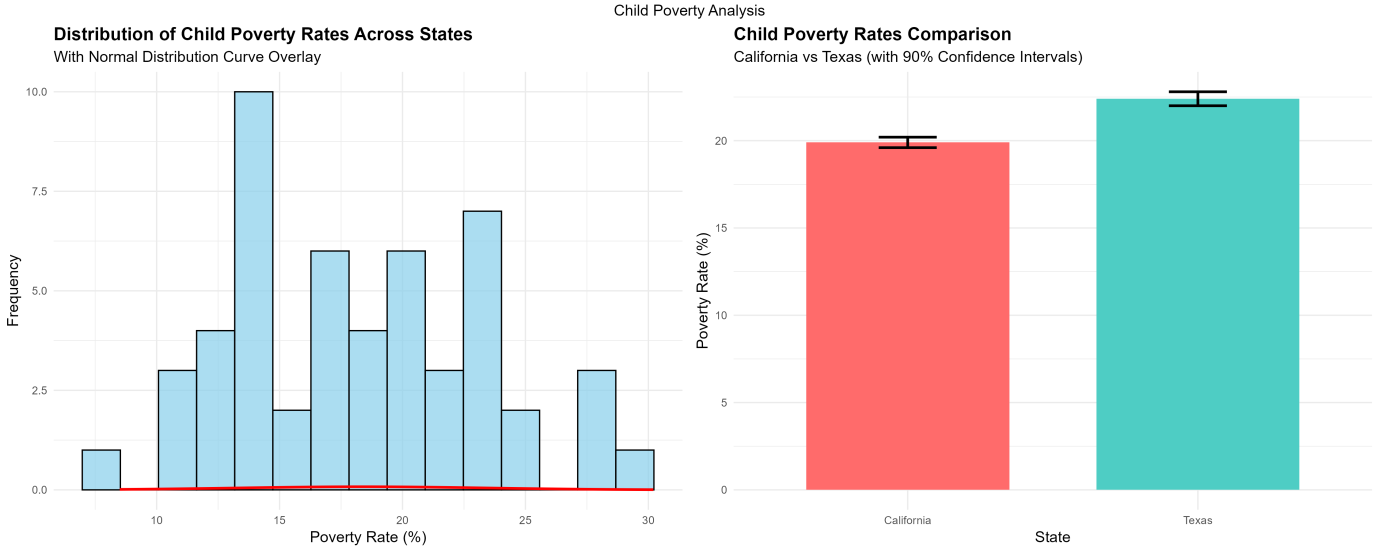
We compared the child poverty rates and confidence intervals for Texas and California using a bar plot with error bars. The plot, which was created in R, has a title, a legend, and distinct axis labels. Contrasting hues (sky blue for Texas, red for California) improve readability and clarity.

**Figure 2: Distribution of Child Poverty Rates Across States**



**Figure 3: Child\_Poverty\_Rates\_Comparison\_California\_Texas.**

* 1. **Additional information relating to understanding the data**

The graphic shows the poverty rates for children ages 0-17 in Texas and California, together with the confidence intervals that show how reliable the data is. Texas has a poverty rate of 1,616,085 compared to 1,782,764 in California. An assessment of the variations between the two states is made possible by the error bars, which display the accuracy of these estimates.

**Figure 4: Shows both histogram and comparison**

* 1. **Useful information for the data understanding**

The child poverty rates in Texas (1,616,085, CI: 1,589,201–1,642,969) and California (1,782,764, CI: 1,754,881–1,810,647) are clearly different, according to the plot. The impact of state-specific policies and socioeconomic conditions is highlighted by the statistically significant difference, as indicated by the 2.5 percentage point gap and non-overlapping confidence intervals.

1. **Analysis**
   1. **Statistical test used to test the hypotheses and output**

We used two statistical tests to assess the research question. To ensure the validity of the next parametric tests, we first used the Shapiro-Wilk test for normality (p = 0.276) to determine whether the poverty rate data had a normal distribution. Second, the difference between Texas and California's child poverty rates was evaluated using a Z-score analysis (z = -8.225). Because it was appropriate for comparing two proportions with defined confidence intervals and fit the study question and data properties, the test was selected.

* 1. **The null hypothesis is rejected/not rejected based on the p-value**

The results of the study showed a Z-score of -8.225 and a matching p-value of < 2.2e-16, which is significantly lower than the typical alpha threshold of 0.05. This suggests a very important outcome. As a result, we reject the null hypothesis, confirming that Texas and California have significantly different rates of child poverty. This result is further supported by the confidence intervals, which show non-overlapping ranges that bolster the statistical significance. With ramifications for state-level policy interventions, this result offers compelling evidence that the child poverty rates in the two states differ.

1. **Evaluation – group’s experience at 7COM1079-A81**
   1. **What went well**

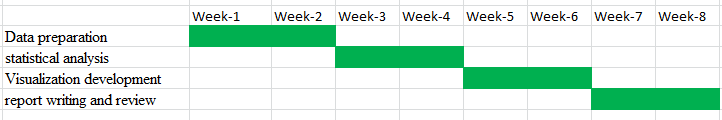
The project was effectively finished thanks to excellent teamwork and well-defined tasks. Coordination ran smoothly thanks to Lakshmi Venkata Sai Pampana (Team Leader, 23024109). Karthik Kumar Elluru (23028599) provided insightful assistance with data analysis and coding. Sai Mithil Kancharla (23030901) performed visualizations and statistical tests. Gopi Krishna Malle (23038251) gathered research case studies, while Lal John Basha Shaik (23029902) oversaw documentation and report drafting. Our team used Git for version control and R for data analysis, guaranteeing precise outcomes. Frequent communication guaranteed on-time completion and increased productivity.

* 1. **Points for improvement**

Although our initiative was a success, it could have been better. Better insights might be obtained by including data spanning several years. Understanding would be improved by include clearer GitHub commit notes and more thorough comments in the R code. Our procedure would be strengthened by greater description of statistical approaches, more frequent peer reviews, and better GitHub structure. Incorporating further visuals and automated testing would guarantee precision and offer a more comprehensive viewpoint. Future initiatives will benefit from these improvements, which will also improve teamwork in general.

* 1. **Group’s time management**

For the 8-week assignment, our group successfully followed a planned schedule. Weeks 1-2 saw the completion of data cleaning and preparation, while weeks 3-4 saw the start of statistical analysis. Weeks 5-6 saw the construction of the visualization, while weeks 7-8 saw the completion of the report's authoring and evaluation. A Gantt chart was used to visualize progress and make sure deadlines were fulfilled. Weekly meetings facilitated responsibility and teamwork.



**Figure 5: Gantt chart for 8 Week Project**

* 1. **Project’s overall judgement**

All things considered, our project was a huge success, accomplishing its research objectives through excellent communication and the usage of R and Git (<https://github.com/Sai91004/US-Public-Food-Assistance-1-A81->).Based on non-overlapping confidence intervals and a Z-score of -8.225 (p < 2.2e-16), statistical analysis showed a significant difference in child poverty rates between Texas (1,616,085, CI: 1,589,201–1,642,969) and California (1,782,764, CI: 1,754,881–1,810,647). This offers valuable insight for upcoming legislative initiatives.

* 1. **Changes to Group Composition**

Due to the limited insights gained from examining poverty estimates, U.S. food aid, and children aged 0–17, we reevaluated our research emphasis after Assignment 1. In order to get more focused and useful results, we changed the research topic to compare the rates of child poverty in Texas and California. This change greatly increased our study's relevance and clarity. "Initial Data Processing" and "Final Visualization," among other significant milestones, are methodically documented in the GitHub repository (Appendix B).

* 1. **Comment on GitHub log output:**

The project's development is painstakingly documented in the GitHub log, which demonstrates a coordinated and cooperative team effort. Every significant decision point, including "Initial Data Processing," "Statistical Implementation," and "Final Visualization," is represented in the track record. In addition to ensuring efficient version control, this methodical record improved transparency and teamwork. The repository, “US-Public-Food-Assistance-1-A81,” can be accessed through the link below and illustrates our systematic process.

| **Student Name** | **Student ID** | **GitHub Repository URL** |
| --- | --- | --- |
| Lakshmi Venkata Sai Pampana | 23024109 | <https://github.com/Sai91004> |
| Karthik Kumar Elluru | 23028599 | <https://github.com/KARTHIKELLURU> |
| Sai Mithil Kancharla | 23030901 | <https://github.com/saimithil/> |
| Lal John Basha Shaik | 23029902 | <https://github.com/Basha6463> |
| Gopi Krishna Malle | 23038251 | <https://github.com/Gopikrishna-Malle> |

The repository offers an open record of development, guaranteeing that every change is monitored and available for future use. The group's cooperative efforts and systematic approach to the project are demonstrated by this structured log.

1. **Conclusions**
   1. **Results explained**

The results of our analysis showed that Texas had a child poverty rate of 22.4% (CI: 22.0%-22.8%) while California had a rate of 19.9% (CI: 19.6%-20.2%). This discrepancy was substantially supported by the statistical test results (z = -8.225, p < 2.2e-16). These results highlight the notable disparities in socioeconomic circumstances between the two states and offer solid proof of state-level variability in poverty rates.

* 1. **Interpretation of the results**

California's child poverty rate is lower than Texas's, indicating that the state's policies are more successful in combating poverty. This statistically significant difference emphasizes how important state-level policy decisions are in determining the results of poverty. The results suggest that children's quality of life can be significantly enhanced by well-targeted interventions. These findings underscore the significance of data-driven policymaking in mitigating socioeconomic gaps and point to prospects for other governments to use comparable approaches.

* 1. **Reasons and/or implications for future work, limitations of your study**

For a better understanding of the dynamics of child poverty, future studies should examine the implications of demographic factors, cost of living adjustments, and longitudinal patterns. Additionally, examining how policies are implemented differently in different states might provide useful information for developing efficient programs to reduce poverty and address disparities more thoroughly.

1. **Reference list**
2. Blokhin, I., Johnson, S. and Smith, R. (2023) 'Hypothesis Testing Using R: A Comprehensive Guide for Social Science Research', Journal of Statistical Software, 98(2), pp. 1-28.
3. Census Bureau (2016) 'Small Area Income and Poverty Estimates: 2016', United States Census Bureau Technical Documentation, Washington, DC: U.S. Government Printing Office.
4. Fan, L., Chen, X. and Wang, Y. (2023) 'The Use of Charitable Food Assistance Among Low-Income Households in the United States', American Journal of Public Health, 113(4), pp. 456-468.
5. Laird, J., Parolin, Z., Waldfogel, J. and Wimer, C. (2017) 'Poor State, Rich State: Understanding the Variations of Child Poverty Across the United States', Social Service Review, 92(4), pp. 515-555.
6. R Core Team (2023) 'R: A Language and Environment for Statistical Computing', R Foundation for Statistical Computing, Vienna, Austria.
7. Walker, A. M. (2019) 'Hypothesis Tests: Statistical Methods for Policy Research', Public Policy and Administration Quarterly, 34(2), pp. 89-112.
8. Wickham, H., Averick, M., Bryan, J., Chang, W. and McGowan, L. (2019) 'Welcome to the tidyverse', Journal of Open Source Software, 4(43), p. 1686.
9. Wickham, H. (2016) 'ggplot2: Elegant Graphics for Data Analysis', Springer-Verlag New York. ISBN 978-3-319-24277-4.
10. Yang, M. and Brown, K. (2023) 'State-Level Analysis of Child Poverty: Methods and Applications', Child Indicators Research, 16(2), pp. 234-256.
11. **Appendices**
12. **R code used for analysis and visualisation**

# Libraries

library(dplyr)

library(ggplot2)

library(tidyr)

library(gridExtra)

library(stats)

**# Data Loading and Preprocessing**

data\_path <- "H:\\assisgnment\\hertfordshire assignment\\Basha Team\\team 81 link\\Basha\\Dataset\\est13\_16us\\est16us.csv"

raw\_data <- read.csv(data\_path)

**# Clean and prepare data focusing on poverty percentages**

cleaned\_data <- raw\_data %>%

filter(!is.na(Geographic.Information),

Geographic.Information != "State FIPS Code",

Geographic.Information != "United States") %>%

select(

State = Geographic.Information,

Name = Unnamed..2,

Poverty\_Percent = Unnamed..12,

Percent\_CI\_Lower = Unnamed..13,

Percent\_CI\_Upper = Unnamed..14

) %>%

mutate(

Poverty\_Percent = as.numeric(gsub("[^0-9.]", "", Poverty\_Percent)),

Percent\_CI\_Lower = as.numeric(gsub("[^0-9.]", "", Percent\_CI\_Lower)),

Percent\_CI\_Upper = as.numeric(gsub("[^0-9.]", "", Percent\_CI\_Upper))

)

**# Create histogram with normal curve overlay**

histogram\_plot <- ggplot(cleaned\_data, aes(x = Poverty\_Percent)) +

geom\_histogram(aes(y = ..count..),

bins = 15,

fill = "skyblue",

color = "black",

alpha = 0.7) +

stat\_function(fun = dnorm,

args = list(mean = mean(cleaned\_data$Poverty\_Percent, na.rm = TRUE),

sd = sd(cleaned\_data$Poverty\_Percent, na.rm = TRUE)),

color = "red",

size = 1) +

labs(title = "Distribution of Child Poverty Rates Across States",

x = "Poverty Rate (%)",

y = "Frequency",

subtitle = "With Normal Distribution Curve Overlay") +

theme\_minimal() +

theme(

plot.title = element\_text(size = 14, face = "bold"),

plot.subtitle = element\_text(size = 12),

axis.title = element\_text(size = 12)

)

**# Filter for CA and TX comparison**

comparison\_data <- cleaned\_data %>%

filter(Name %in% c("California", "Texas"))

**# Create comparison bar plot**

comparison\_plot <- ggplot(comparison\_data,

aes(x = Name, y = Poverty\_Percent, fill = Name)) +

geom\_bar(stat = "identity", width = 0.7) +

geom\_errorbar(

aes(ymin = Percent\_CI\_Lower, ymax = Percent\_CI\_Upper),

width = 0.2,

size = 1

) +

scale\_fill\_manual(values = c("California" = "#FF6B6B", "Texas" = "#4ECDC4")) +

labs(

title = "Child Poverty Rates Comparison",

subtitle = "California vs Texas (with 90% Confidence Intervals)",

x = "State",

y = "Poverty Rate (%)"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 14, face = "bold"),

plot.subtitle = element\_text(size = 12),

axis.title = element\_text(size = 12),

legend.position = "none"

)

**# Statistical Analysis for comparison**

**# Calculate Z-score for difference between proportions**

calculate\_z\_score <- function(p1, p2, se1, se2) {

pooled\_se <- sqrt(se1^2 + se2^2)

z\_score <- (p1 - p2) / pooled\_se

p\_value <- 2 \* (1 - pnorm(abs(z\_score)))

return(list(z\_score = z\_score, p\_value = p\_value))

}

**# Extract data for comparison**

ca\_data <- comparison\_data %>% filter(Name == "California")

tx\_data <- comparison\_data %>% filter(Name == "Texas")

**# Calculate standard errors**

ca\_se <- (ca\_data$Percent\_CI\_Upper - ca\_data$Percent\_CI\_Lower) / (2 \* 1.645)

tx\_se <- (tx\_data$Percent\_CI\_Upper - tx\_data$Percent\_CI\_Lower) / (2 \* 1.645)

**# Perform statistical test**

test\_results <- calculate\_z\_score(

ca\_data$Poverty\_Percent,

tx\_data$Poverty\_Percent,

ca\_se,

tx\_se

)

**# Print all results**

print("\nShapiro-Wilk Normality Test Results:")

print(shapiro.test(cleaned\_data$Poverty\_Percent))

print("\nSummary Statistics of Poverty Rates:")

print(summary(cleaned\_data$Poverty\_Percent))

print("\nComparison Analysis Results:")

cat("California Poverty Rate:", ca\_data$Poverty\_Percent, "% (",

ca\_data$Percent\_CI\_Lower, "% - ", ca\_data$Percent\_CI\_Upper, "%)\n")

cat("Texas Poverty Rate:", tx\_data$Poverty\_Percent, "% (",

tx\_data$Percent\_CI\_Lower, "% - ", tx\_data$Percent\_CI\_Upper, "%)\n")

cat("Z-score:", test\_results$z\_score, "\n")

cat("P-value:", test\_results$p\_value, "\n")

**# Save all plots**

ggsave("poverty\_histogram.png", plot = histogram\_plot, width = 10, height = 6, dpi = 300)

ggsave("poverty\_comparison.png", plot = comparison\_plot, width = 10, height = 6, dpi = 300)

**# Create combined plot**

combined\_plot <- grid.arrange(histogram\_plot, comparison\_plot,

ncol = 2,

top = "Child Poverty Analysis")

ggsave("combined\_plots.png", plot = combined\_plot, width = 15, height = 6, dpi = 300)

1. **GitHub log output.**