Estimating National Hourly Wages: A Bootstrap Approach

Lowell Monis, Adit Parmar, Atticus Gondoly, Claire Medema

May 2, 2025

Contents

Motivation

Hourly wage is a core economic indicator, reflecting the financial well-being of workers, the cost of labor to employers, and the economic vitality of different U.S. regions. Given the complexity of modeling wage data across the country, this project seeks to estimate the national average hourly wage in 2024 using state-level data. We employ nonparametric bootstrap methods due to the limitations of parametric modeling on a small and right-skewed dataset.

Data

In the first in-class project workday, we were able to find a dataset that we all agreed had potential to yield interesting and pertinent results. The dataset included several different components to it, containing the average hourly wage within the United States over the span of about 50 years, as well as the average hourly wage for each state in each year. Since the average hourly wage over time is not independent, we decided to focus on estimating the average wage for the country as a whole during a specific year using the data from each state, operating under the assumption that the wages from each state are independent of one another. For our analysis, we isolate data from the year 2024.

Population: The entire working population of the United States.

Sample: Median hourly wages for all 50 states (excluding D.C. for reasons discussed below).

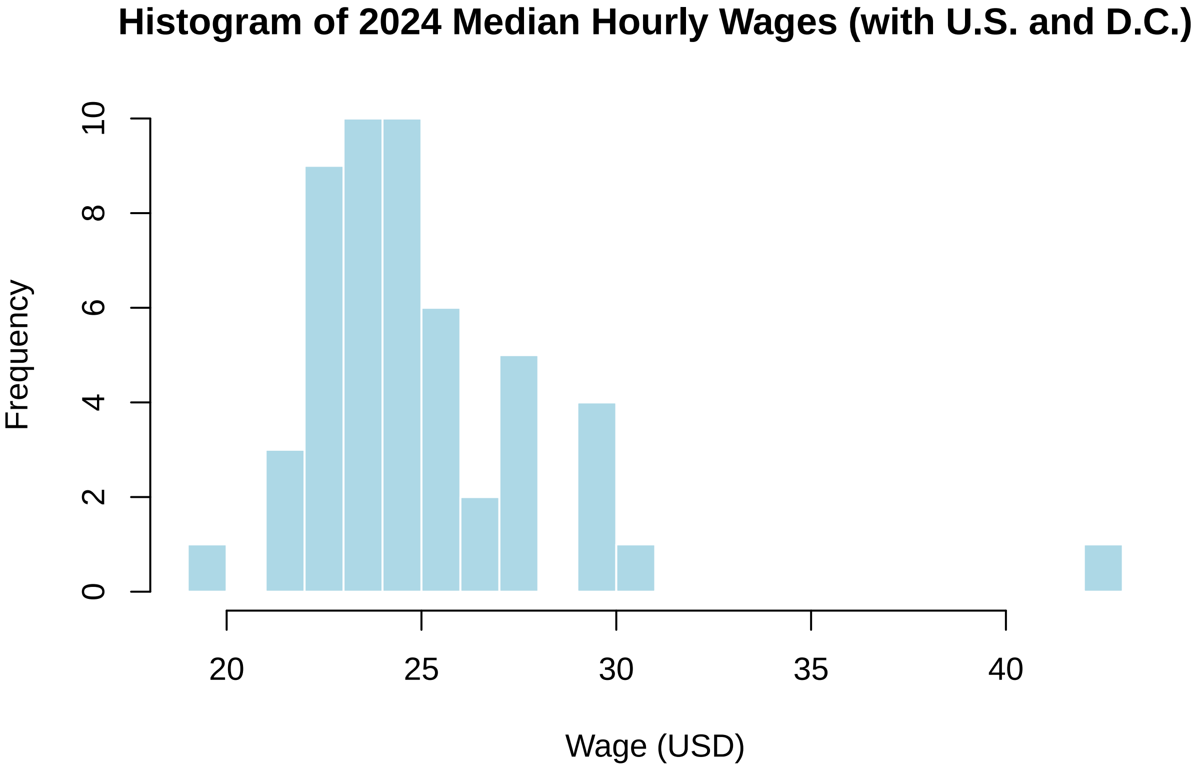
Variable of interest: State-level median hourly wage in 2024.

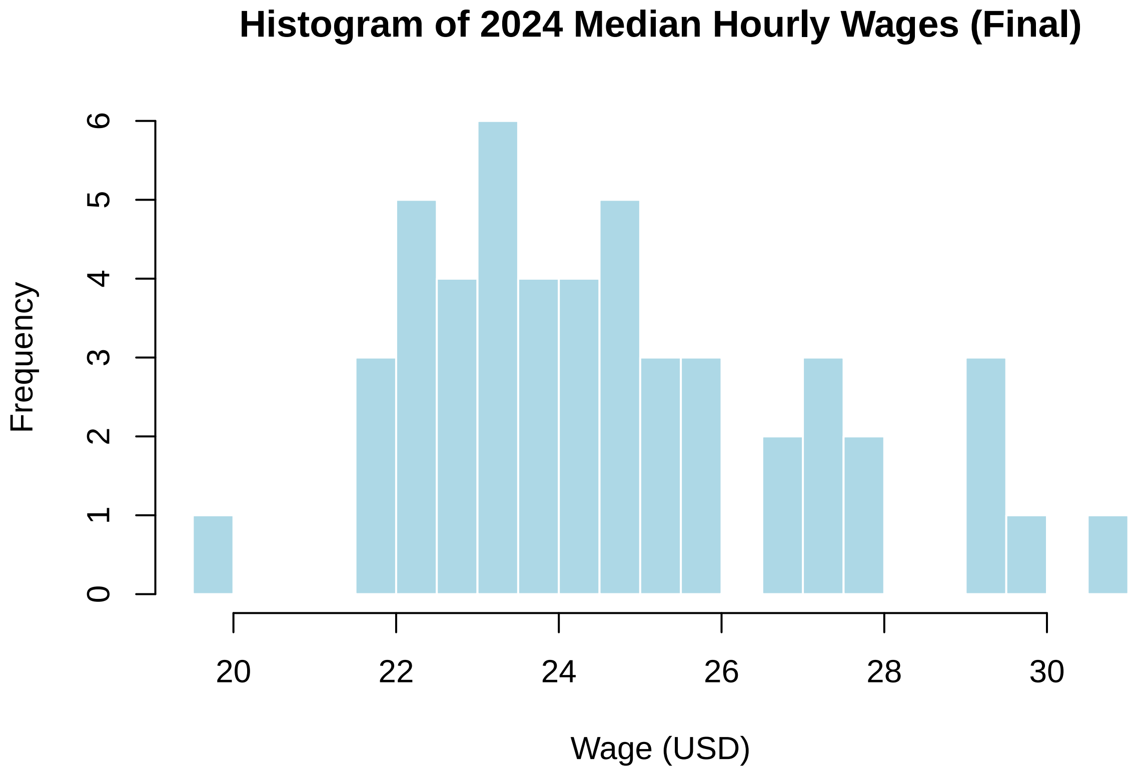
The median is used in this analysis because it is widely recognized as a robust measure of central tendency, particularly in the presence of skewed data or outliers. Since our dataset consists of the median hourly wages of individual states—themselves already measures of central tendency—it is conceptually appropriate to apply a robust estimator to this second layer of aggregation. While the sample mean is sensitive to extreme values, the median resists distortion from disproportionately high or low state wages, making it a more reliable summary of the overall distribution.

However, it is important to acknowledge that while the median is robust, it is not a sufficient statistic for many standard distributions, meaning it may not capture all the relevant information contained in the sample for estimating population parameters. Despite this, its resilience to non-normality and its interpretability in skewed settings make it a pragmatic and defensible choice for our nonparametric approach to estimating the central wage tendency in 2024.

Assumptions and Distributional Models

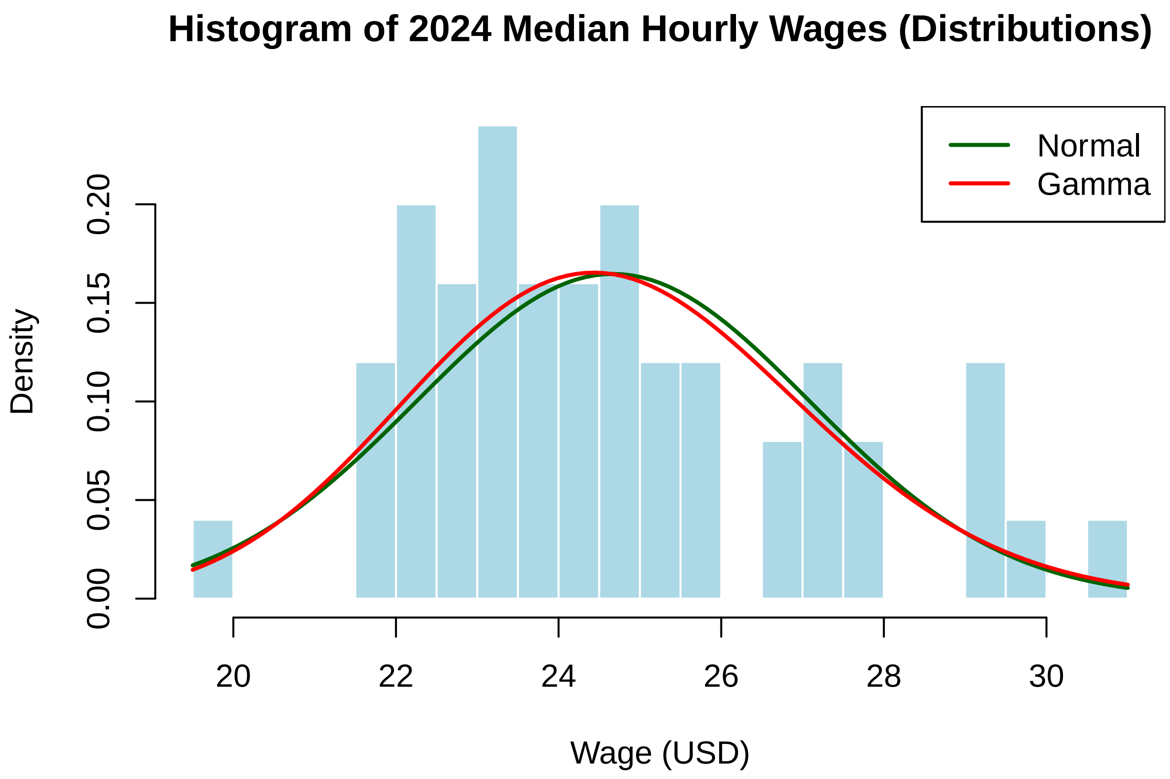
To study median hourly wages across the U.S. in 2024, we begin by analyzing the distribution of our chosen variable: state-level median wages. Initially, we consider whether this data can reasonably be modeled by one of the distributions discussed in class, such as the Normal or Gamma distributions.





At first glance, the Normal distribution may seem like a viable model, as the variable is continuous, positive, and generally central around a mean. Median wages also tend to follow smooth economic trends across states and demographic groups, particularly when segmented by education levels. This makes the assumption of approximate symmetry and bell-shaped behavior tempting.

We also considered the Gamma distribution, which is frequently used for modeling positive skewed data like income or waiting times. Given that the data is a non-negative, continuous random variable, the Gamma model seems structurally appropriate.



However, a closer inspection of the histogram reveals moderate right skewness and a major outlier in Washington D.C. While we ultimately exclude D.C. from the analysis for being a geographically and demographically incomparable unit, the overall distribution remains asymmetric even after its removal. This persistent skewness violates the assumptions of normality.

Further, the sample size is limited to 50 states, which restricts our ability to reliably estimate parameters of any parametric distribution, especially one like the Gamma that is sensitive to variation and tail behavior in small samples.

Additionally, we do not have access to a probability density function (pdf) or sufficient prior information about the population distribution of wages across states to justify likelihood-based inference.

For these reasons, we conclude that neither the Normal nor Gamma distributions offer a statistically valid foundation for inference in this context.

Given these limitations and the nature of our data, we adopt a nonparametric bootstrap approach. This method:

- Requires no assumptions about the underlying population distribution.
- Works well with small samples.
- Respects the structure of observed data, including its skewness and potential outliers.
- Allows us to empirically estimate the sampling distributions of statistics like the sample mean and sample median.

The bootstrap technique is especially appropriate here because we are not sampling from an infinite population but rather analyzing a full census of states, and we are interested in understanding variation as if this sample were drawn repeatedly from a broader population model.

In summary, bootstrap sampling provides a flexible, assumption-free framework that fits the empirical characteristics of our data and supports robust inference. The only assumption we do make is that the wages from each state are independent of one another.

Research Question

In this analysis, we tried to address:

> How can we estimate the average hourly wage in the United States based on data from each state in 2024? Which method is better?

This question seeks to estimate a central tendency—such as the mean or median—of hourly wages across the U.S. by using state-level data. While the dataset gives us 50 discrete values (one per state), our true objective is to infer a population-level characteristic: the typical hourly wage that represents the U.S. labor market as a whole. In other words, we are trying to move beyond a simple description of the sample to make a meaningful statement about the overall population of U.S. workers.

The characteristic we are estimating is the central tendency of hourly wages—either the mean or median wage—for the entire U.S. workforce in 2024. This is an inherently population-level question because it goes beyond the 50 state medians we observe and instead aims to describe how wages are distributed at the national level. The assumption is that the state-level medians offer a representative snapshot of national wage patterns, though they may not capture population-weighted nuances without adjustment.

This research question is of significant practical interest to a wide range of audiences:

- Policy makers and economists may use such national wage estimates to guide decisions about minimum wage adjustments, tax policy, or federal aid distribution.
- Employers and HR departments might use this data to benchmark salaries and ensure wage competitiveness across different regions.
- Workers and labor unions are directly impacted, as understanding the national wage landscape informs negotiations, job mobility decisions, and advocacy for fair compensation.
- Researchers and journalists rely on such estimates when reporting on wage inequality, labor market trends, and the health of the economy.

By identifying an accurate national wage estimate from the state-level data, this project contributes to a clearer understanding of wage conditions across the United States. It enables comparisons over time, across states, and across education or industry sectors—ultimately supporting more equitable and evidence-based decision-making.

Estimators

To address our research question about the central tendency of wages in the United States, we propose two estimators for the underlying population characteristic: the **sample mean** and the **sample median** of state-level median hourly wages.

The sample mean is a commonly used estimator of central tendency. It is defined as the arithmetic average of all observed state median wages and is computed as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In the context of this project, it provides an estimate of what a “typical” state earns in terms of hourly wage. The mean is useful for summarizing the overall wage level and is particularly helpful when comparing wages over time or across regions. However, the sample mean is sensitive to outliers—such as Washington D.C., which has a much higher wage than all other states. This makes the mean less reliable in skewed distributions, where outliers can unduly influence the result. It is however a sufficient estimator since it covers the entire data.

The sample median is defined as the value that separates the higher half of the data from the lower half. It is calculated by ordering the state median wages and selecting the middle value (or the average of the two middle values if the number of states is even). The median is robust to outliers and less affected by skewed distributions, making it particularly useful in our case. It is however, less sufficient.

Our histogram reveals moderate right-skewness, even after removing outliers such as D.C. As such, the sample median serves as a compelling estimator of central tendency, offering a clearer picture of what most states earn without being distorted by a few extremely high values.

Both the sample mean and median are appropriate estimators because they aim to quantify the same underlying characteristic: the center of the national wage distribution. However, they offer different statistical and practical perspectives on what constitutes a “typical” wage:

- The mean reflects the arithmetic average across all states, but may be skewed by high-wage regions.
- The median better reflects what a randomly chosen state might experience, especially when the distribution is not symmetric.

While the sample mean is a sufficient statistic for the normal distribution and widely used in parametric analysis, it performs poorly in skewed or heavy-tailed contexts. The sample median, while not sufficient, provides greater robustness in the presence of non-normality and outliers.

Given the skewness and small sample size, the median may offer a more stable and representative estimate. However, by comparing both estimators using bootstrap resampling, we can empirically evaluate their sampling variability, bias, and overall performance—helping us determine which estimator is better suited for our nonparametric framework.

Calculation and Evaluation of Estimators

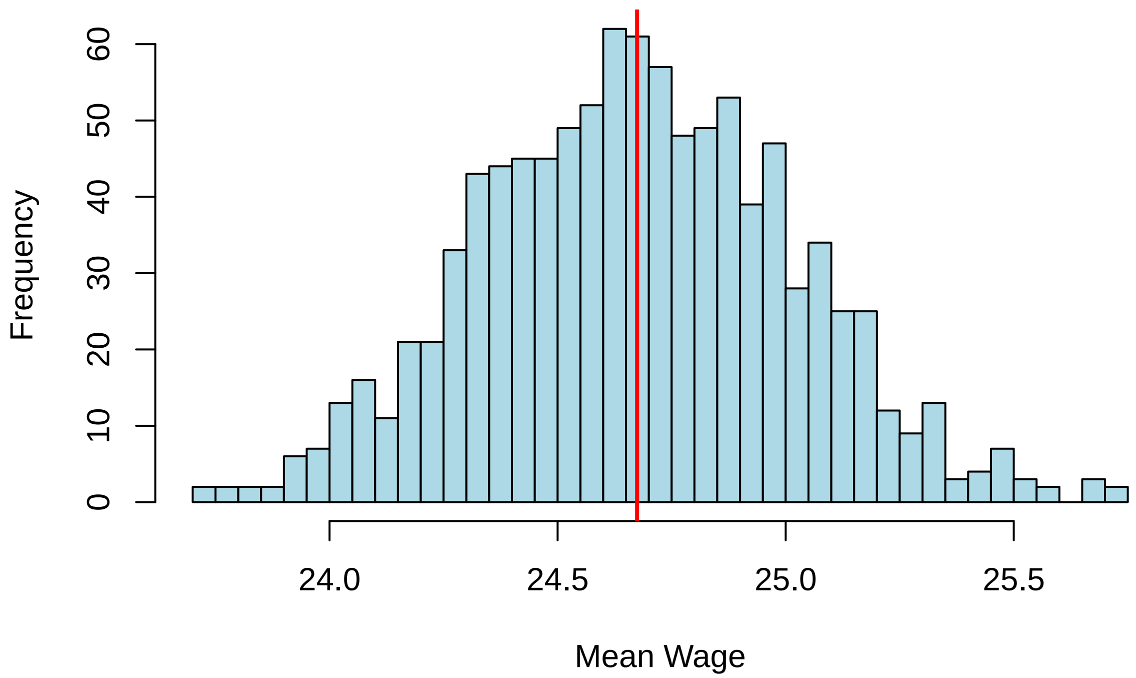
To estimate the central tendency of state-level hourly wages in the United States for 2024, we applied our two proposed estimators: the sample mean and the sample median. Based on the raw data (excluding Washington D.C. and the national average), we obtained the following:

Sample Mean: USD 24.67

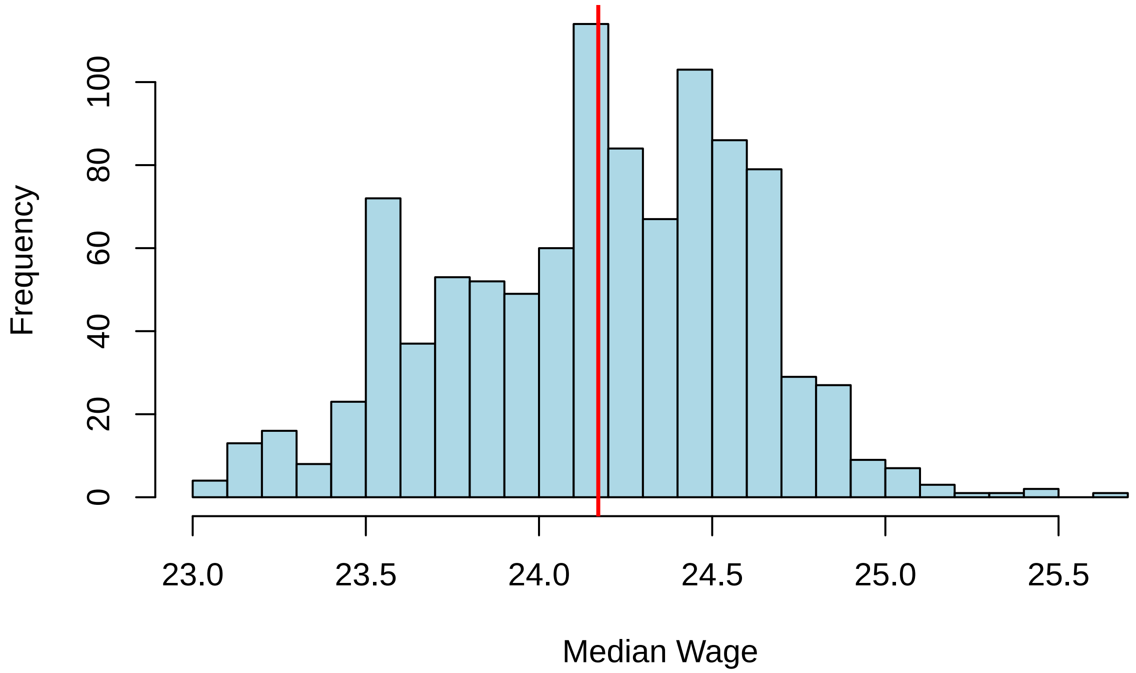
Sample Median: USD 24.24

Because the distribution is moderately right-skewed, we expected some minor inflation in the mean due to higher-wage states. To account for sample variability and better approximate the sampling distribution in order to evaluate it, we applied nonparametric bootstrap resampling with 1,000 iterations.

Bootstrap Sampling Distribution of Mean Wage (2024)



Bootstrap Sampling Distribution of Median Wage (2024)



Bootstrap Mean: USD 24.67

Bootstrap Median: USD 24.17

The sample estimator and bootstrap estimator are close, indicating very little to no bias in both estimators for this sample.

Despite the greater bias in the median, it may still be the superior estimator. Thus, we judge by its mean squared error (MSE):

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

This formula shows that even a biased estimator can outperform an unbiased one if it achieves a sufficiently lower variance.

In our case, the variance of the bootstrap means is 0.1174356, and that of the bootstrap medians is 0.1920202.

Thus, from the above formula, the MSE for the bootstrap means is 0.1174366 and that of the bootstrap medians is 0.1943434.

In this case, the mean outperforms the median under the MSE criterion, due to its significantly lower variance and zero bias. Although the median is more robust to outliers and skew, the actual data distribution in this project—while moderately skewed—does not distort the mean enough to outweigh its lower variability. The median also could not fit into the bias-variance tradeoff well.

The mean is thus sufficient (covers the data), efficient (low variance), and has low MSE. These are reasons to believe that we can better estimate the central tendency of the data for 2024 using the sample mean.

Conclusion

Although the median is robust and resistant to outliers, our analysis using bootstrap sampling shows that the mean is a better estimator

This suggests that—for this dataset—the mean provides a more efficient estimate of the true central tendency, despite the slight right-skewness observed. The median, while still informative, introduces more variability and a small negative bias that increases its total estimation error.

1. The mean is unbiased and had lower variance than the median.
2. The median was slightly biased and had a higher MSE, making it less favorable for this particular dataset.
3. The mean is a sufficient estimator.
4. The mean is the more efficient estimator here.

Although robustness is an important property, efficiency outweighs it when the data distribution does not strongly violate assumptions.

Thus, the mean is the preferred estimator in this analysis. It offers a precise, unbiased, and statistically efficient estimate of the national hourly wage across states, grounded in sound theoretical evaluation through the bootstrap method. With approximations done using these methods, we can continue to track trends in wages over different years, education levels, and even age or gender, and contribute to the knowledge of those in the work force, ensuring that they know what a fair wage looks like. As for the value of the average hourly wage, it looks like the sample mean says it is $24.67.

References

Economic Policy Institute, State of Working America Data Library, “Hourly wage, average - Average real hourly wage (2024$),” 2025.

Appendix

Data Used

The mean hourly wages are organized by alphabetical order of state.

```
##  [1] 22.20945 27.21600 24.62100 21.96600 26.60216 29.29000 27.96009 24.56000
##  [9] 23.55880 23.28776 23.65438 22.92333 25.48111 23.06200 23.17933 22.98840
## [17] 21.71000 21.58964 24.43400 29.18600 30.80706 24.19560 25.66400 19.98114
## [25] 23.01333 23.89412 23.53000 24.01000 27.78152 29.65752 22.20030 26.60247
## [33] 22.27600 24.87600 24.83800 22.24800 25.68600 25.30163 27.14698 22.57036
## [41] 23.00800 24.29200 23.19200 24.67127 25.90253 27.44733 29.13000 22.17300
## [49] 25.01200 22.93600
```

R Script

```r
# Package imports
library(tidyverse)

# Reading the data and initial cleaning
wages <- read.csv("data/median_hourly_wage_state_epi.csv")
wages_2024 <- subset(wages, format(as.Date(date), "%Y") == "2024") %>%
  select(-date) %>%
  pivot_longer(cols = everything(),
               names_to = "State", values_to = "Hourly.Wage.2024")

# Visualizing via histogram
hist(wages_2024$Hourly.Wage.2024, xlab="Wage (USD)", breaks=30,
     main="Histogram of 2024 Median Hourly Wages (with U.S. and D.C.)",
     col = "lightblue",
     border = "white")

# Cleaning out DC and US, revisualizing
wages_2024 <- wages_2024 %>%
  filter(!(State %in% c("District.of.Columbia", "United.States")))
hist(wages_2024$Hourly.Wage.2024, xlab="Wage (USD)", breaks=30,
     main="Histogram of 2024 Median Hourly Wages (Final)",
     col = "lightblue",
     border = "white")

# Visualizing to gauge distributions
hist(wages_2024$Hourly.Wage.2024,
     xlab = "Wage (USD)",
     breaks = 30,
     main = "Histogram of 2024 Median Hourly Wages (Distributions)",
     freq = FALSE,
     col = "lightblue",
     border = "white")

wage_mean <- mean(wages_2024$Hourly.Wage.2024, na.rm = TRUE)
wage_sd <- sd(wages_2024$Hourly.Wage.2024, na.rm = TRUE)
```

```r
curve(dnorm(x, mean = wage_mean, sd = wage_sd),
      col = "darkgreen", lwd = 2, add = TRUE)

wage_mean <- mean(wages_2024$Hourly.Wage.2024, na.rm = TRUE)
wage_var <- var(wages_2024$Hourly.Wage.2024, na.rm = TRUE)

gamma_shape <- wage_mean^2 / wage_var
gamma_rate <- wage_mean / wage_var

curve(dgamma(x, shape = gamma_shape, rate = gamma_rate),
      col = "red", lwd = 2, add = TRUE)

legend("topright", legend = c("Normal", "Gamma"),
       col = c("darkgreen", "red"), lwd = 2)

# Set seed and parameters
set.seed(442)
B <- 1000

# Extract the numeric wage vector
wage_vec <- wages_2024$Hourly.Wage.2024

# Bootstrap sample means and medians
boot_means <- replicate(B, mean(sample(wage_vec, replace = TRUE)))
boot_medians <- replicate(B, median(sample(wage_vec, replace = TRUE)))

# Calculate bootstrap point estimates
boot_mean <- mean(boot_means)
boot_median <- mean(boot_medians)

# Bootstrap histogram for mean
hist(boot_means,
     main = "Bootstrap Sampling Distribution of Mean Wage (2024)",
     xlab = "Mean Wage",
     col = "lightblue", breaks = 30)
abline(v = boot_mean, col = "red", lwd = 2)

# Bootstrap histogram for median
hist(boot_medians,
     main = "Bootstrap Sampling Distribution of Median Wage (2024)",
     xlab = "Median Wage",
     col = "lightblue", breaks = 30)
abline(v = boot_median, col = "red", lwd = 2)
```