U.S. Salary Disaparities Across Various Demographics and Industries

S&DS 230 Final Project

Shane Lee

1) Problem Context

It is well known that salaries vary significantly across individuals, often influenced by demographic factors such as gender, race, and age. While one might envision a future in which compensation is determined solely by internal attributes—such as interpersonal skills or domain-specific expertise—that ideal remains distant. In the present, demographic characteristics continue to play a meaningful role in shaping earnings. This paper investigates how wages differ across gender, age, and occupational industry, with the aim of quantifying these disparities and understanding the extent to which they persist.

2) Data Background

The American Community Survey (ACS) Public Use Microdata Sample (PUMS) is a nationally representative dataset that provides detailed information on individuals and housing units, with data available at the state level. It includes a wide range of demographic variables, particularly those related to employment. These include annual wage income, class of worker (e.g. private sector, government, self-employed), average hours worked per week, and more. The following is a list of the variables used in this analysis:

- WAGP - Wages/salary (continuous)
- OCCP - Occupation (categorical)
- COW - Class of worker (categorical)
- ESR - Employment Status (categorical)
- WKHP - Hours worked per week (continuous)
- AGEP - Age (continuous)
- SEX - Sex (categorical)

The exact details of these variables can be found in the 2023 ACS PUMS Data Dictionary. For this paper, we take the data for individuals living in Massachusetts.

3) Data Cleaning

There are two initial issues with using every entry in the dataset as-is.

The first is that not all individuals are employed full-time. To address this, the dataset is filtered based on several variables. For the COW variable, only individuals employed in the private sector or government are retained; those who are self-employed or unemployed are excluded. For the ESR variable, individuals who are unemployed or in the armed forces are also removed. Finally, using the WKHP variable, only those who report working an average of 30 hours per week or more are kept.

The second issue is the presence of outliers in reported wages. Some individuals in the dataset are marked as working full-time but earning less than $20,000 annually—below what would be expected under Massachusetts minimum wage laws. Assuming these cases reflect reporting discrepancies or data quality issues, they are excluded from the analysis. Additionally, the dataset applies top-coding to wage values: individuals earning more than $715,000 are all recorded as earning exactly $715,000 to protect their anonymity. Because their true incomes are unknown, these entries are also excluded from the final dataset.

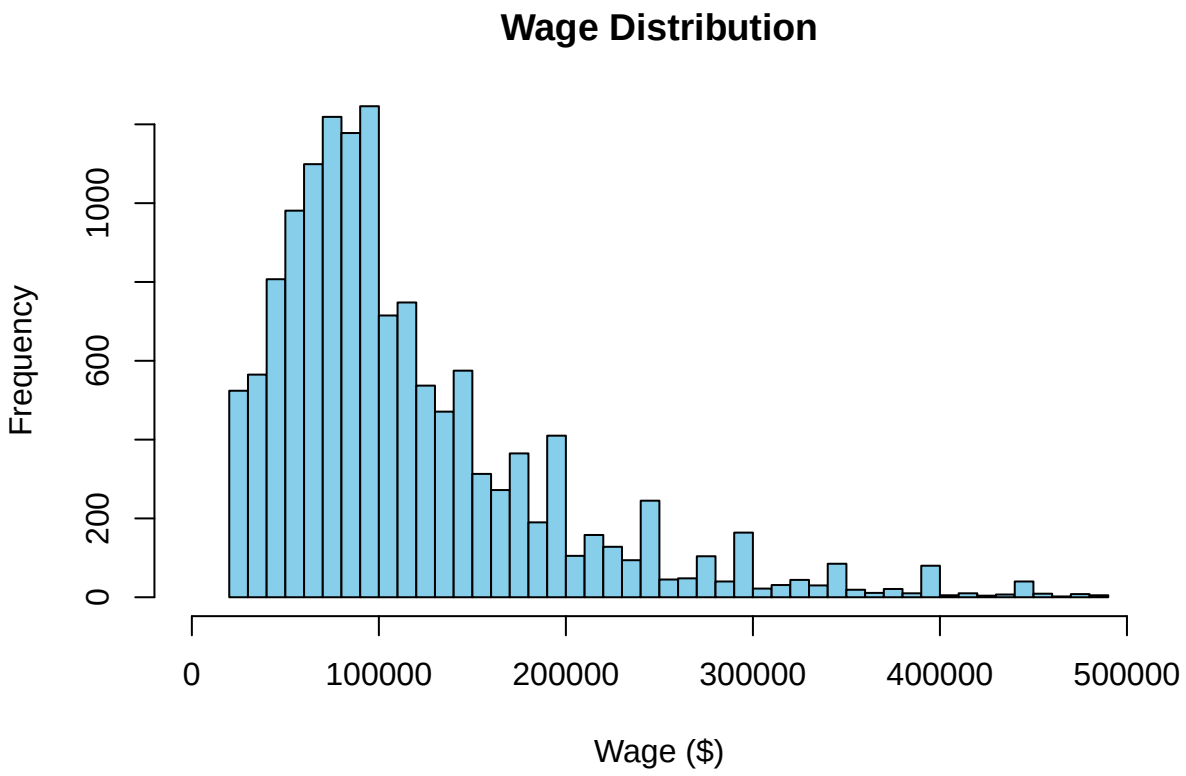
The OCCP variable includes hundreds of unique codes, each corresponding to a specific occupation, such as "Photographer" or "Retail Salesperson." These occupations can be grouped into broader industry categories—referred to in this paper as occupation types—such as "Engineering" or "Healthcare." The full list of occupational codes and their associated categories is available on the ACS Occupation Codes website.

While all occupation types are included in the original dataset, this analysis focuses on a selected subset to maintain clarity and statistical power. The following occupation types are used in this paper, with representative examples in parentheses:

- Computer and Mathematical (e.g. Software Engineer)
- Education (e.g. Teacher)
- Finance (e.g. Financial Analyst)
- Healthcare (e.g. Dentist)
- Legal (e.g. Paralegal)
- Management (e.g. Human Resources Manager)
- Science (e.g. Astronomer)

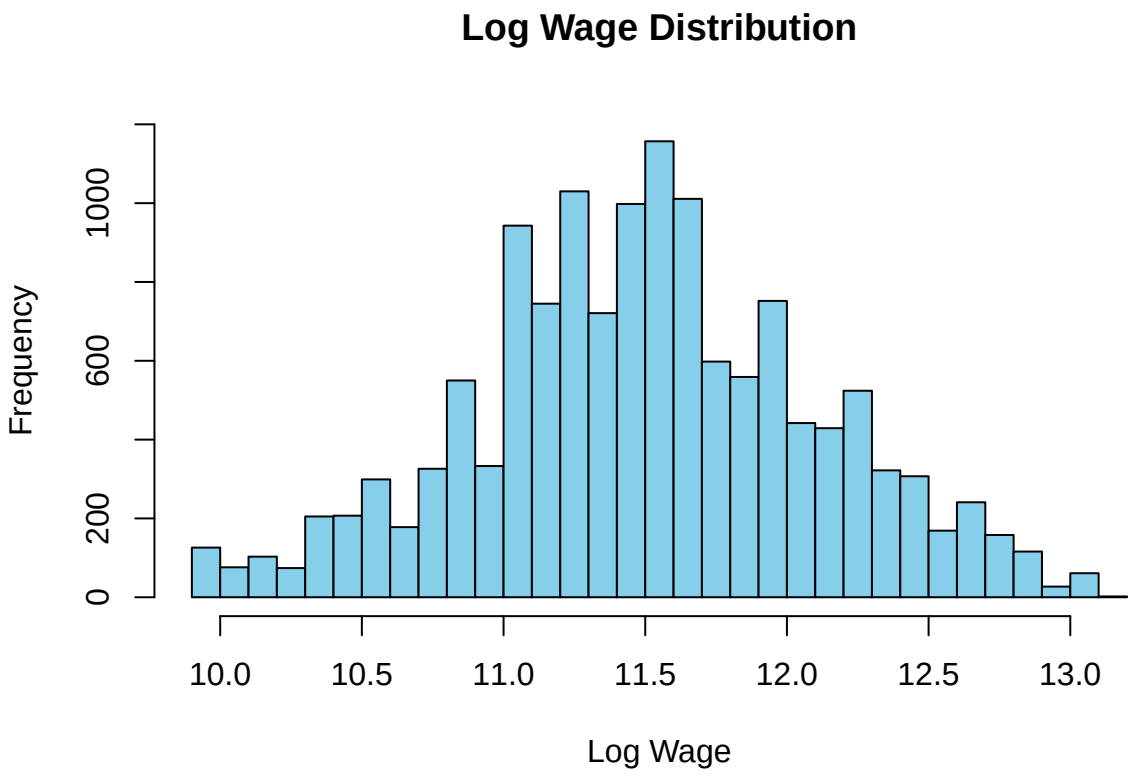
4) Initial Data Plots and Summaries

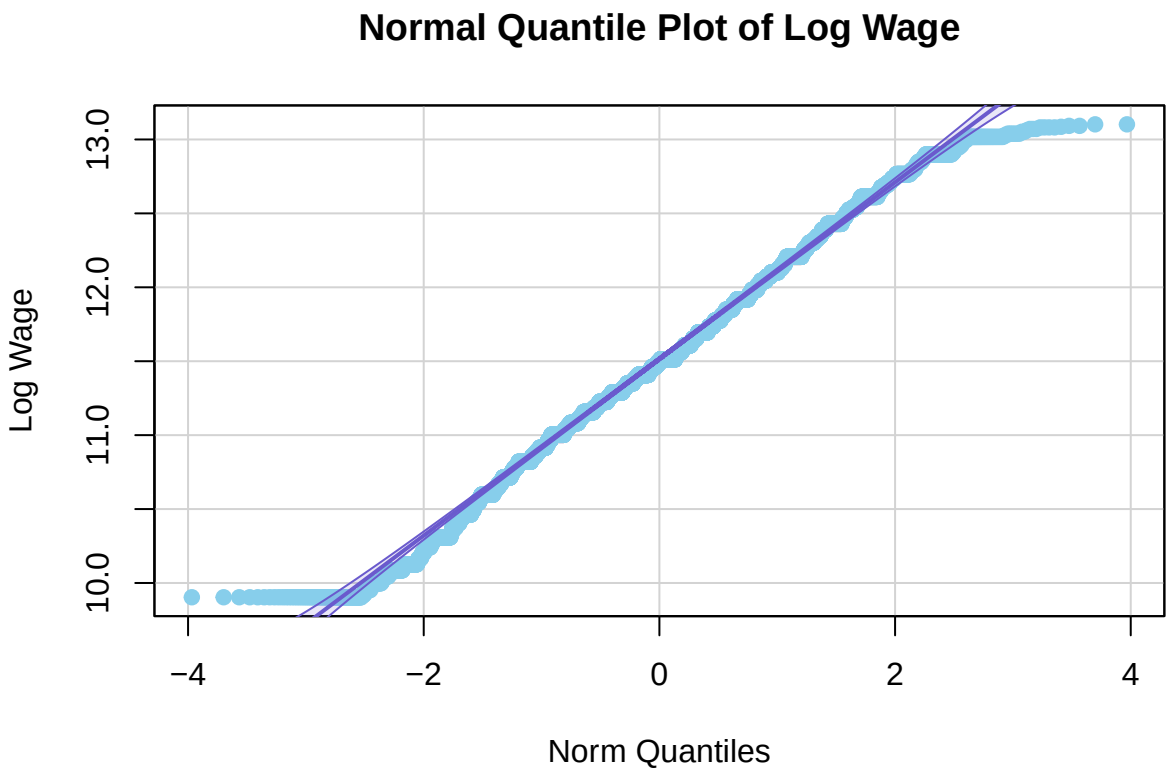
Here is a histogram of wages of individuals in the dataset.



It's clearly not very normally distributed, which is a vital assumption for some of the statistical tests done later in the analysis section of this paper. Here is a normal quantile plot of the wages to further demonstrate its non-normal distribution.

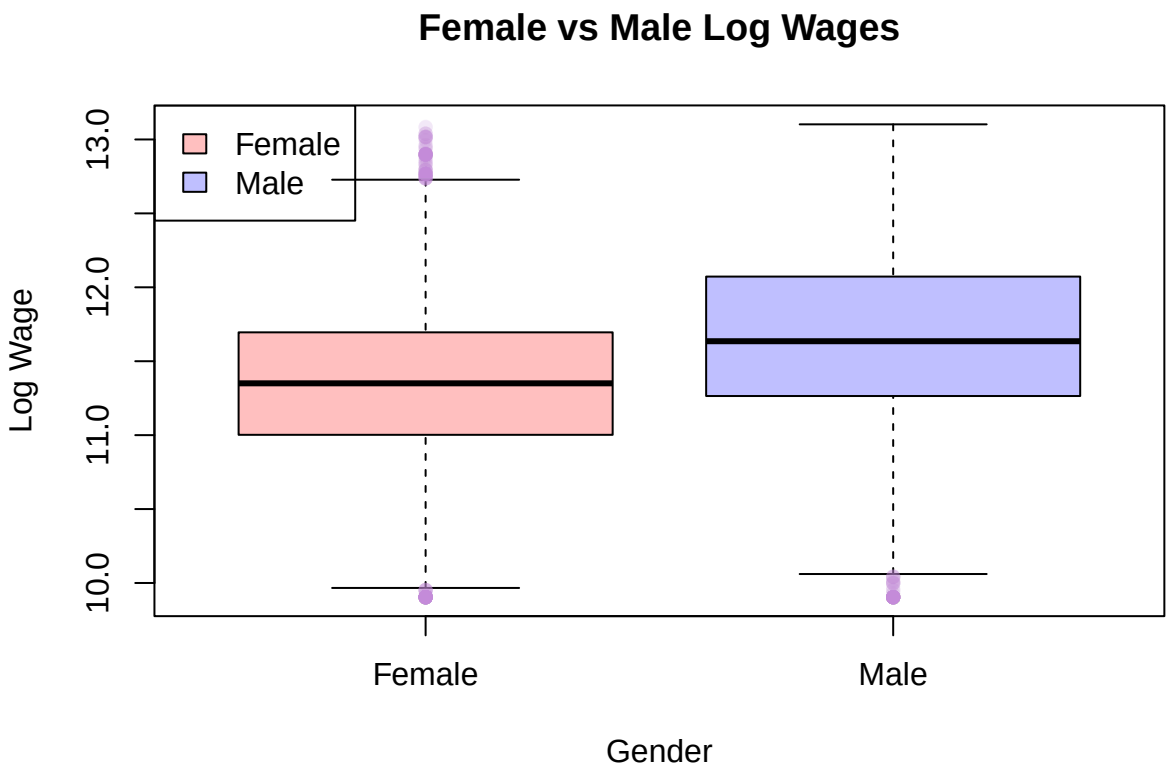
For the purposes of achieving normality of data, the natural logarithm of the wage is used for these tests instead.

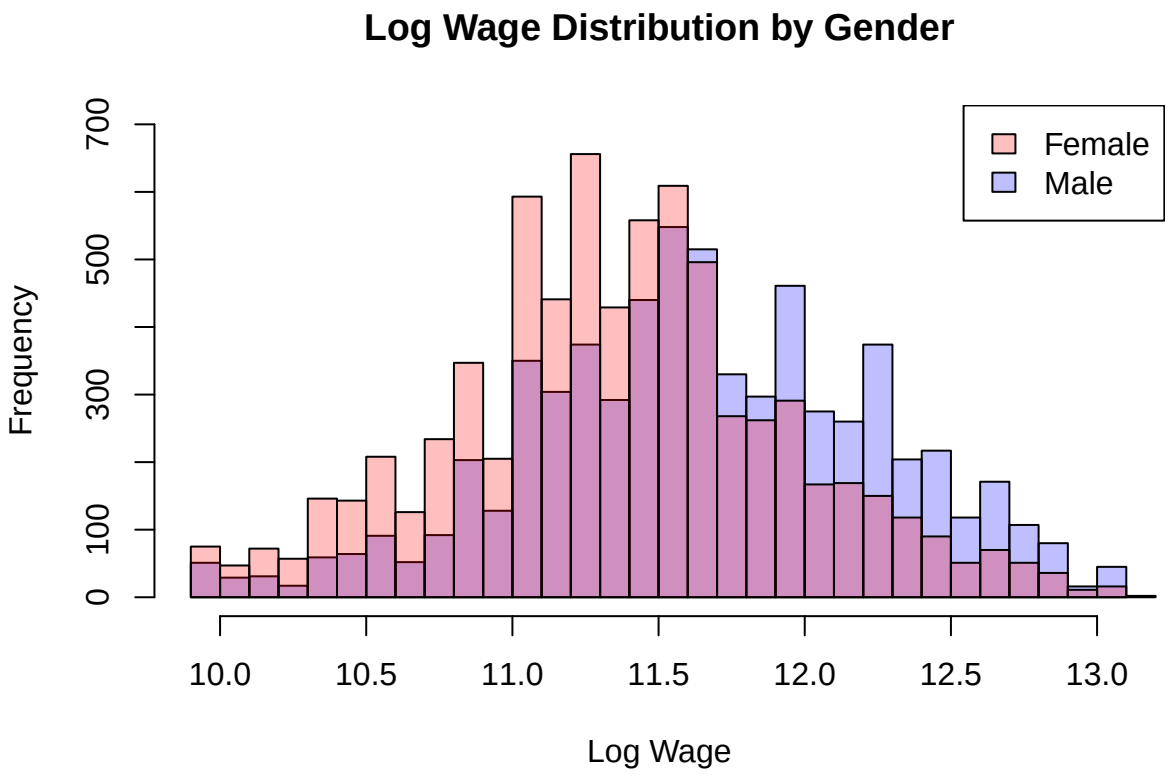

Log Wage Distribution
Frequency
0
200
600
1000
10.0
10.5
11.0
11.5
12.0
12.5
13.0
Log Wage



The log wages are approximately normally distributed as demonstrated by the histogram and the normal quantile plot. The normal quantile plot does have some deviations near the ends caused by the wages being bounded by $20,000 and $715,000, but this should not cause any issues with the statistical tests.

Now for the wage distribution comparisons between different demographics. First is gender:


Female vs Male Log Wages
Female
Male
13.0
12.0
11.0
10.0
Log Wage
Female
Male
Gender

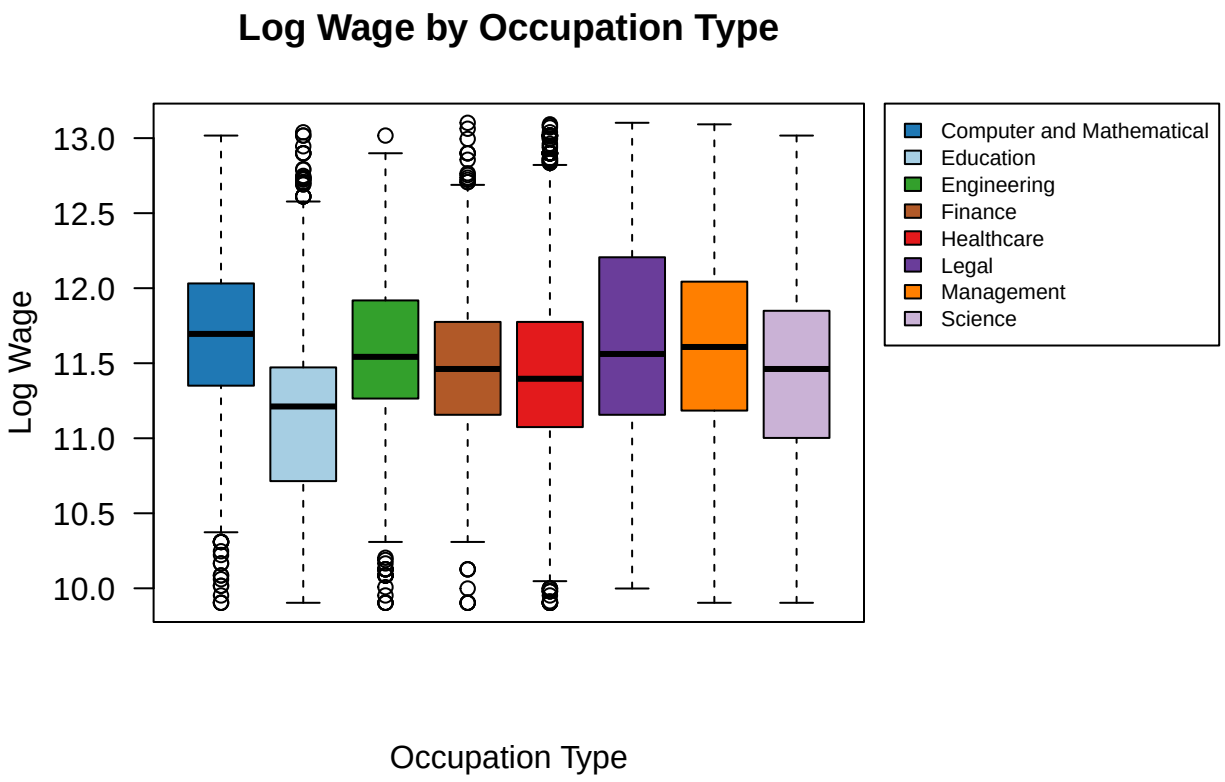


From an initial glance, it appears that salaries are higher for male individuals than for female individuals, which should not be particularly surprising.

Now for age:

Based on the density of the individual points, there appears to be a very slight positive correlation between age and log wage. This should also make sense; older individuals will likely have more years of experience and thus are more valuable as employees than, say, new hires.

And lastly, salaries across different occupation types:



It seems that occupation types that fall under "Computer and Mathematical" have the highest median salaries, which makes sense considering the types of jobs in this category (e.g. software engineers, data scientists, statisticians, actuaries, etc.). Also not surprising is that Education jobs have the lowest median salaries.

5) Analysis

5.1) Differences in Log Wage by Gender

Since the standard two-sample t-test assumes equal variance, the standard deviations of log wages for female and male individuals are compared to ensure their ratio is less than 2.

```
## [1] "Std. Dev. of Log Wage for Female: 0.577"
```

```
## [1] "Std. Dev. of Log Wage for Male: 0.605"
```

```
## [1] "Ratio: 1.048"
```

The two-sample t-test is therefore a valid test to use here.

```
##
##  Two Sample t-test
##
```

```
## data:  df2$LogWage[df2$Gender == "Female"] and df2$LogWage[df2$Gender == "Male"]
## t = -28.33, df = 13787, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3050141 -0.2655378
## sample estimates:
## mean of x mean of y
##  11.35891  11.64418
```

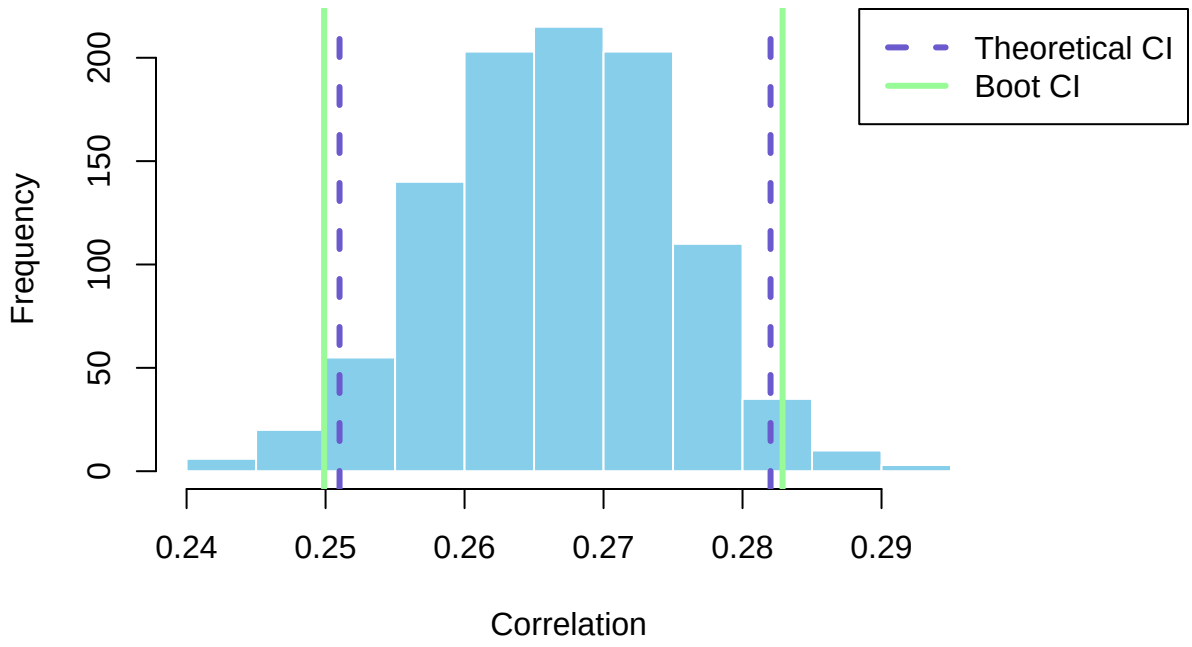
```r
n_boot <- 1000
boot_corrs <- numeric(n_boot)

for (i in 1:n_boot) {
  sample_indices <- sample(1:nrow(df2), replace = TRUE)
  sample_data <- df2[sample_indices, ]
  boot_corrs[i] <- cor(sample_data$LogWage, sample_data$Age)
}

par(mar = c(5, 5, 5, 6))

hist(boot_corrs,
     main = "Bootstrap Distribution of Correlation (LogWage vs Age)",
     xlab = "Correlation",
     col = "skyblue",
     border = "white")

ci <- quantile(boot_corrs, c(0.025, 0.975))
abline(v = ci, lwd = 3, col = "palegreen")
abline(v = cor.test(df2$LogWage, df2$Age)$conf.int, lwd = 3, col = "slateblue", lty = 2)
legend("topright",
       c("Theoretical CI","Boot CI"),
       lwd = 3,
       xpd = TRUE,
       inset = c(-0.25, 0),
       col = c("slateblue","palegreen"), lty = c(2, 1))
```




```r
n_perm <- 1000
perm_diffs <- numeric(n_perm)

obs_diff <- with(df2, mean(LogWage[Gender == "Female"]) - mean(LogWage[Gender == "Male"]))

# Permutation test
for (i in 1:n_perm) {
  perm_labels <- sample(df2$Gender)
  perm_diffs[i] <- with(df2, mean(LogWage[perm_labels == "Female"]) -
                          mean(LogWage[perm_labels == "Male"]))
}

# Plot histogram of permuted differences
hist(perm_diffs,
     main = "Permuted Mean Log Wage Differences by Gender",
     xlab = "Difference in Mean Log Wages (Female - Male)",
     col = "skyblue", border = "white",
     xlim = c(-0.3, 0.04))
abline(v = obs_diff, col = "palegreen", lwd = 2)

# Two-tailed p-value
p_value <- mean(abs(perm_diffs) >= abs(obs_diff))
cat("Observed difference:", round(obs_diff, 2), "\n")
```

```
## Observed difference: -0.29
```

```r
cat("Permutation p-value:", round(p_value, 4), "\n")
```

```
## Permutation p-value: 0
```

```r
text(x = obs_diff - 0.01,
     y = 10,
     labels = "Observed Difference = -0.29",
     col = "darkgrey",
     srt = 90,  # rotate vertically
     adj = 0)
```