Gender Pay Gap Testing in R

Lindsay Williamson

CSU Global

Gender Pay Gap Testing in R

Decades after pay discrimination became illegal, a pay gap continues to exist between men and women. It is important to fix this issue, so organizations are turning to data analytics and statistical analysis to easily find instances of intentional and unintentional discrimination. While the unadjusted pay gap is calculated as the difference between the mean male salary and the mean female salary, the adjusted pay gap is harder to calculate, yet it provides a more accurate picture (Ziv, 2019). One way to understand the data is to perform hypothesis testing and regression analysis. By having tools in place to make gender pay easy to analyze, organizations can make significant progress toward pay equity.

To begin the analysis, the dataset was saved in CSV format and imported into RStudio. While this dataset was already cleaned up, it is worthwhile to group similar job titles and departments so there is a large mixture of male and female in each group to provide enough data (Chamberlain, 2017).

A picture containing computer, filled, full, laptop

Description automatically generated

*Figure 1*. Dataset showing salary, gender and other variables. Taken from: Chamberlain, A.

(2017, April). How to analyze your gender pay gap: an employer’s guide. Retrieved from Glassdoor:https://www.glassdoor.com/research/app/uploads/sites/2/2019/03/GD\_Report\_AnalyzingGenderPayGap\_v2-2.pdf

Summary statistics were performed on the data to provide the initial overview of the data. The median and mean base pay were also calculated after grouping by gender. The preliminary results show that female pay is under the mean by over $3,000. It also shows that the average male pay is higher than the median by about $5,000 and higher that average female pay by over $8,000.

A picture containing table

Description automatically generated

*Figure 2*. Summary statistics in R.

A screenshot of a cell phone

Description automatically generated

*Figure 3*. Mean and median pay by gender.

The question to be addressed by hypothesis testing is, based on the dataset, are women paid less than men? The expectation is that men are paid higher than women.

H0 (null hypothesis): there is no difference between average salary of the male and female population, and

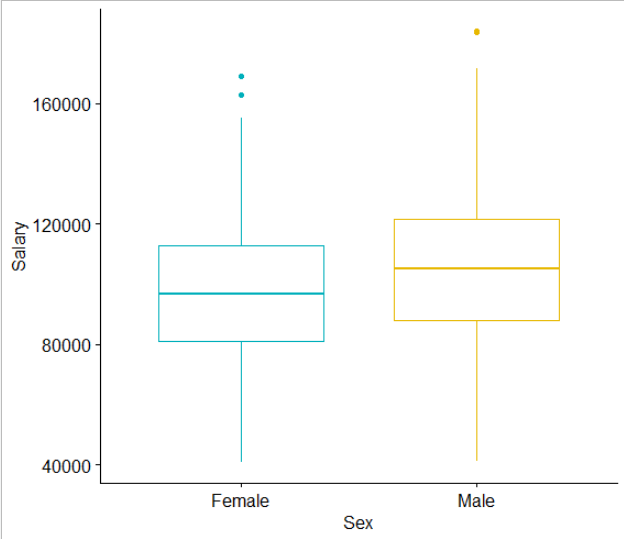
Ha (alternative hypothesis): there is a difference between the average salary of the male and female population.

The boxplot below clearly shows a gender pay gap and is a useful visualization to get an understanding of what the hypothesis will reveal.

> ggboxplot(genderpay, x = "gender", y = "totalpay",

+ color = c("#00AFBB", "#E7B800"),

+ ylab = "Salary", xlab = "Sex")



*Figure 4*. Boxplot showing differences between male and female salary.

Using a 95% confidence interval, the result of the t-test is that the average of male salaries is about $8,500 higher than the average of female salaries.

A p-value lower than 0.05 means there is strong evidence to reject the null hypothesis, so the alternative is accepted (Unpaired two-sample t-test in R, n.d.). A p-value higher than 0.05 indicates that there is not enough evidence to reject the null hypothesis (Unpaired two-sample t-test in R, n.d.). The p-value is lower than 0.05, so the null hypothesis is rejected.

A picture containing bird

Description automatically generated

*Figure 5*. Welch two sample t-test hypothesis testing code and results in R.

To analyze this further, a multivariable regression analysis can be used to provide clarity on a gender pay audit based on certain employee characteristics like education level, performance evaluation rating, seniority and job title (Chamberlain, 2017). This will show whether there is an adjusted pay gap. A dummy variable was created equal to 1 for males and 0 for females (Frye, 2019). If there is no pay gap according to the regression analysis, there will be no salary advantage for being male. The models below show the regression models with an estimate of the approximate pay gap percentage. Model 1 is a simple regression that shows the unadjusted pay gap, which is just the pay gap between male and female. Model 2 is a multivariable regression that compares the variable of education and performance evaluation along with gender. Model 3 compares education, department, seniority, job title, age and performance evaluation. The results also show the outliers (Frye, 2019). This is where there may be a gender pay equity issue, and it requires further review. For model 1, the coefficient of .095 means there is approximately a 9.5% pay gap. Model 2 shows 10.01% and model 3 shows a gap of 1.1%. Once all variables are considered, there is only a 1.1% gender pay gap (Chamberlain, 2017). The reason for the unadjusted pay gap at this company isn’t because women in the same role with similar education, performance rating, seniority and department are paid less; it’s because there are more males represented in higher paid positions like engineering and management (Chamberlain, 2017).

> model1 <- lm(log\_base ~ male, data = data)

> # Add controls for age, education and performance evaluations.

> model2 <- lm(log\_base ~ male + perfEval + age\_bin + edu, data = data)

> # Add all controls. ("adjusted" pay gap.)

> model3 <- lm(log\_base ~ male + perfEval + age\_bin + edu + dept + seniority + jobTitle, data = data)

A picture containing bird

Description automatically generated

A picture containing bird

Description automatically generated

A screenshot of a cell phone

Description automatically generated

*Figure 6*. Code and model results in R.

One interesting relationship that was exposed through this analysis was that, even though there did not appear to be any bias regarding pay where the job characteristics are similar between men and women, the unadjusted pay gap was still high. This could point to an issue with hiring and promotions since it was clear from the hypothesis testing and simple regression that men on average make more than women. Performing hypothesis testing and regression analysis for a gender pay audit is an important step to ensuring equitable treatment (Chamberlain, 2017). Data analytics can now provide the insight into these issues where datasets are considerable. While many human resources departments do not have data analysts on staff to perform statistical analysis, learning how to accomplish complex data analysis is becoming increasingly important (Ziv, 2019). The gender pay gap is a crucial problem to address and the tools to do it are available and easy to use.

References

Chamberlain, A. (2017, April). *How to analyze your gender pay gap: an employer's guide*. Retrieved from Glassdoor: https://www.glassdoor.com/research/app/uploads/sites/2/2019/03/GD\_Report\_AnalyzingGenderPayGap\_v2-2.pdf

Frye, C. (2019, November 12). *Tableau and R for analytics projects*. Retrieved from LinkedIn Learning: https://www.linkedin.com/learning/tableau-and-r-for-analytics-projects/create-a-logistic-regression-model-in-r?u=2245842

*Unpaired two-sample t-test in R*. (n.d.). Retrieved from STHDA: http://www.sthda.com/english/wiki/wiki.php?id\_contents=7600

Ziv, S. (2019, July 11). *3 Ways The Gender Pay Gap Is Even Bigger Than You Think* . Retrieved from Forbes: https://www.forbes.com/sites/shaharziv/2019/07/11/gender-pay-gap-bigger-than-you-thnk/#4e313fd87d8a