Hypothesis Test: Is the Average Income Differ Due to Gender?

The topic of the Gender Pay Gap is a very relevant and controversial topic currently. This topic is very important and relevant to me as I am a woman. “According to Bureau of Labor Statistics data, in 2020, women’s annual earnings were 82.3% of men’s, and the gap is even wider for many women of color (5 facts).” This means that for every dollar earned by a male, a female earns only 82 cents. Seeing this fact about the Gender Pay Gap led me to the thought that the average salary of men and women has to be different.

Research Question: I will be assessing if the average income among males and females is the same or not.

Parameters:

* mean income of males
* mean income of females

Hypothesis: There is not a difference in the average income between males and females, i.e. .

Hypothesis Test:

Vs.

i.e.

There is not a difference in the average income between males and females.

There is a difference in the average income between males and females.

We will be using a significance level of .

Analysis/Methods:

* Imported and cleaned up the data.
* I used the head and tail functions to ensure the data imported correctly.
  + The data consists of 5 different variables: Gender, Current Salary, Years of Education, Minority Classification, Date of Birth.
* I used the summary function on the Current Salary Variable to see values such as the minimum salary, maximum salary, and the average salary.
* The mean function was then used on the Current Salary variable to obtain the average salary.
* Then I did a summary of the data to see information like that there are 474 different peoples salaries in this data along with a summary of each variable of the data.
* A frequency table of Gender (Male & Female) was created.
* A frequency table of Years of education and Gender was created.
* A Standard Mean Table of the Current Salary categorized by Gender (Male & Female) was created.
* A Standard Deviation Table of the Current Salary categorized by Gender (Male & Female) was created.
* A One-way ANOVA Table was created modeling the Current Salary variable as a function of the Gender variable.
* Post Hoc Analysis (Simultaneous Tests for General Linear Hypothesis & Simultaneous Confidence Intervals) for the One-way ANOVA table was created.
* Assessed the equality of variances by performing Levene’s Test for Homogeneity of variance.
* Tested whether samples are originated from the same distribution by performing the Kruskal-Wallis Rank Sum Test for the Current Salary variable as a function of the Gender Variable.
* Pairwise Comparisons using Wilcoxon Rank Sum Test with Continuity Correction.
* Calculated pairwise comparisons between group levels with corrections for multiple testing using the Wilcoxon Rank sum test for the five different fats of interest and the two subtypes of interest
  + The adjustment methods used was the Bonferroni Correction
* Created a Box plot of the mean with the Current Salaries as the y values and the Gender for the x values
* Tested each level of the subtypes for normality using the Shapiro Test

Results:

Gender Frequency Table:

|  |  |
| --- | --- |
| Male: | Female: |
| 216 | 258 |

From the Gender Frequency Table we can see that there are 42 females than males in this study.

Years of Education and Gender Table:

|  |  |  |
| --- | --- | --- |
| Years of Education: | Female: | Male: |
| 8 | 30 | 23 |
| 12 | 128 | 62 |
| 14 | 0 | 6 |
| 15 | 33 | 83 |
| 16 | 24 | 35 |
| 17 | 1 | 10 |
| 18 | 0 | 9 |
| 19 | 0 | 27 |
| 20 | 0 | 2 |
| 21 | 0 | 1 |

When observing the Frequency table of years of Education and Gender Table we can see that the highest number of years of education Females have is 17 while the highest number of years of education Males have is 21 years.

The Mean Table of the Current Salary categorized by Gender:

|  |  |
| --- | --- |
| Gender: | Current Salary: |
| Female: | $26,031.92 |
| Male: | $41,441.78 |

We know that the mean is the average of the given data. When observing the Mean Table of the Current Salary categorized by Gender we can see that the average salary for Males is $15,409.86 higher than the average current salary for females. Shown above is the average salary between male and female. Notice that we can see a difference in the average salary of males and females. The salary of the men being higher by about 15,000.

The Standard Deviation Table of the Current Salary categorized by Gender:

|  |  |
| --- | --- |
| Gender: | Current Salary: |
| Female: | 7,558.021 |
| Male: | 19,499.214 |

Shown above is the Current Salary categorized by Gender Standard Deviation Table. We know that standard deviation is a measure of how dispersed the data is in relation to the mean, i.e. the standard deviation is heavily dependent on the mean. As expected since the Male category had the largest average, it also has the highest standard deviation.

One-way ANOVA Table Modeling Current Salary as a function of Gender:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Degrees of Freedom: | Sum Square: | Mean Squared Error: | F-Statistic: | P-Value: |
| Gender: | 1 | 2.792e+10 | 2.792e+10 | 119.8 | <2e-16 |
| Residuals: | 472 | 1.100e+11 | 2.330 |  |  |

Notice that the degrees of freedom is one, which is what we would expect as there are two categories of the Gender variable, male & female. “The sum of squares measures the deviation of data points away from the mean value. A higher sum-of-squares result indicates a large degree of variability within the data set, while a lower result indicates that the data does not vary considerably from the mean value (Kenton).” Notice that the Sum of Squares obtained is very large at a value of 2.792e+10. Thus, there is a lot of variability in the data set as the Sum of Squared is a very large value. “The mean squared (MSE) tells you how close a regression line is to a set of points (Mean).” The mean squared error is obtained by obtaining each distance from each point to the regression line and then squaring the values (Mean). “The smaller the mean squared error, the closer you are to finding the line of best fit (Mean).” Note that the mean squared error is a large value at 2.792e+10, the same value as the Sum of Squares. “The larger the F-statistic, the greater the variation between sample means relative to the variation within the samples (Zach).” When observing the ANOVA table we can see that we have a large F-Statistic at a value of 119.8. “An F statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different (F Statistic).” Thus, since we have a large F-Statistic we know that the means of the Current Salary of Males & Females is statistically significant. We can use the P-value to decide if the difference between group means is statistically significant or not (Zach). The P-Value obtained from the one-way ANOVA table is <2e-16 which is smaller than the chosen alpha value of 0.05. Thus, we will reject the null hypothesis that there is not a difference in the average income between males and females. Hence, the difference between the mean of the Current Salary of Males & the Current Salary of Females is statistically significant.

Post Hoc Analysis (Simultaneous Test for General Linear Hypotheses):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Linear Hypothesis: | Estimate: | Standard Error: | T Value: | P-Value: |
| Male – Female = 0 | 15,410 | 1,408 | 10.95 | <2e-16 |

Post Hoc Analysis tests were performed with using the Tukey Test. A Tukey Test can be used to determine exactly which groups means are different (Stephanie). When observing the Post Hoc Analysis (Simultaneous Test for General Linear Hypothesis), we can see that the obtained P-Value is <2e-16, which is smaller than the chosen alpha value of 0.05. This implies that the difference between the mean of the Current Salary of Males & the Current Salary of Females is statistically different.

Post Hoc Analysis (Simultaneous Confidence Intervals):

|  |  |  |  |
| --- | --- | --- | --- |
| Linear Hypothesis: | Estimate: | Lower Bound: | Upper Bound: |
| Male – Female = 0 | 15,409.8616 | 12,643.3219 | 18,176.4014 |

When looking at the Post Hoc Analysis (Simultaneous Confidence Intervals), we know that the difference between the mean of the current salary of Males and the current salary of Females is considered statistically significant since zero is not contained in the associated confidence interval (Frost). We know that if zero would have been contained in the confidence interval that the means of the current salary of Males and the current salary of Females would have been the same (Frost). Recall that the difference between the mean of the current salary of Males and the current salary of Females is also considered statistically significant when looking at the obtained p-value for the Simultaneous Test for General Linear Hypothesis.

Recall that the width of the confidence interval determines how precise the estimate is (Frost).

Levene’s Test:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Degrees of Freedom: | F Value: | P-Value: |
| Group: | 1 | 51.602 | 2.666e-12 |
|  | 472 |  |  |

“Levene’s test is used to check that variances are equal for all samples when your data comes from a non normal distribution (Stephanie Levene).” When observing the obtained Levene Test we can see that the P-Value is 2.666e-12 which a lot smaller than the selected alpha value of 0.05. Thus, we can conclude that the variances are not the same.

Kruskal-Wallis Rank Sum Test:

|  |  |
| --- | --- |
| Kruskal-Wallis Chi-Squared: | 150.96 |
| Degrees of Freedom: | 1 |
| P-value: | < 2.2e-16 |

“The Kruskal-Wallis Test (sometimes called the one-way ANOVA on ranks) is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups if an independent variable on a continuous or ordinal dependent variable (Kruskal-Wallis).” The Kruskal-Wallis Test is considered to be the nonparametric alternative to a one-way ANOVA table (Kruskal-Wallis). Note that the Kruskal-Wallis Test cannot tell us exactly which groups are statistically significant, it will only tell us that at least two groups are different (Kruskal-Wallis). When observing the Kruskal-Wallis Rank Sum Test we can see that the obtained P-Value (<2.2e-16) is smaller that the alpha value of 0.05. This implies that the difference in the means of two or more groups is statistically significant. Since we are only observing the difference in the means of two groups, we have more evidence that the difference in the means of the current salary of Men and the current salary of Women is statistically significant.

Pairwise Comparisons using Wilcoxon Rank Sum Test with Continuity Correction:

|  |  |
| --- | --- |
|  | Female: |
| Male: | <2e-16 |

The Wilcoxon Rank Sum Test is a nonparametric test serving as an alternative test to the independent sample t test (Ellis). “It tests whether the average sum of the ranks (and thus the medians) of the two samples differ significantly from each other (Ellis).” When observing the results of the Pairwise Comparisons using Wilcoxon Rank Sum Test with Continuity Correction we can see that when the Male groups is compared to the Female group a P-Value of <2e-16 is obtained. Thus, we know that the groups differ significantly from each other as the obtained P-Value is below the alpha level of 0.05.

Box Plot:

Chart, box and whisker chart

Description automatically generated

|  |
| --- |
| f = Female |
| m = Male |

When observing the box plot, we can see that green dot represents the mean of each group, and the boxes represent the middle 50%. First, notice that the mean of the Current Salary for Males is significantly higher than the mean of the Current Salary for Females. The Male group has a higher spread compared to the Female group as it has a lot of outliers at the top.

Conclusion:

Overall, from the obtained results we have statistically significant evidence that the mean of the Current Salary of Men is larger than the mean of the Current Salary of Females from our specific data.

MATH 5320 Specific Part: