# Statistical Analysis Report

## 1. EDA

Fig 1. Mean salary by age group

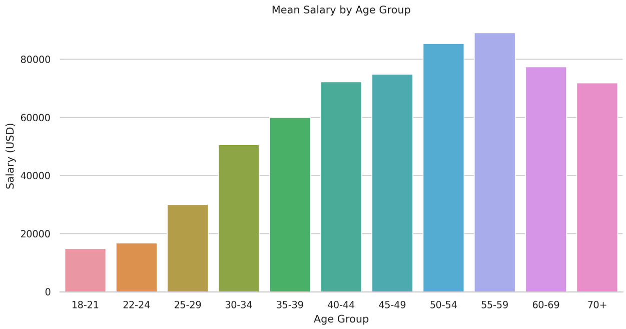
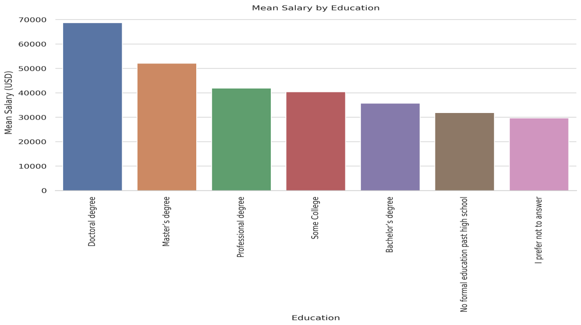


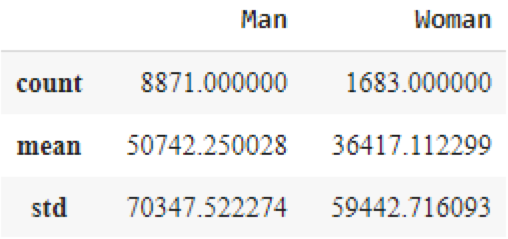
Fig 2. Mean salary by education level



In figure 1, we can see that the average salary increases as age increases, peaks around age 50-60, then slowly decreases. From figure 2 we can see that participants with higher education levels such as Doctoral degree and Master's degree reported much higher salaries on average. Another plot also indicates that Switzerland, USA, and Israel are the top three countries with the highest mean salary levels in this dataset (see “Countries with Highest Salaries on Average” plot in notebook).

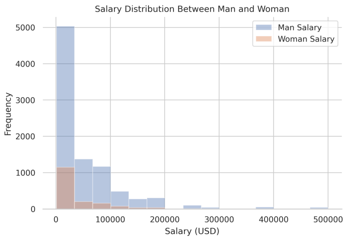
## 2. Estimating difference between average salary of men vs. women

Table 1. Descriptive statistics



| | Man | Woman |
|---|---|---|
| count | 8871.000000 | 1683.000000 |
| mean | 50742.250028 | 36417.112299 |
| std | 70347.522274 | 59442.716093 |

Fig 3. Salary distributions for men vs. women



Based on figure 3 and the results from normality test and Levene’s test for homoscedasticity (see notebook), the original data is right-skewed and does not satisfy the equal variances assumption. Thus, the original data is not suitable for a two-sample t-test. To compare the difference in means of the two gender groups, bootstrapping is used to generate the sample distributions in the figures below. Distributions of both bootstrapped samples and the difference in means all appear normal.

Fig 4. Bootstrapped salary distributions for men vs. women

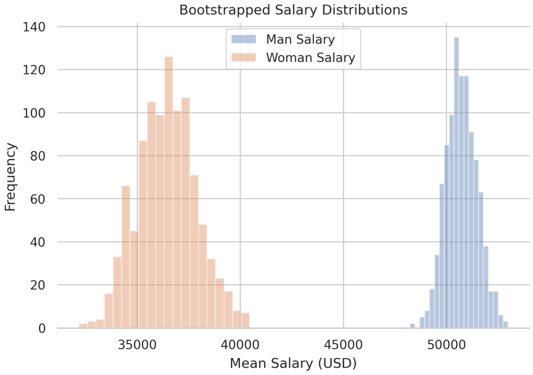
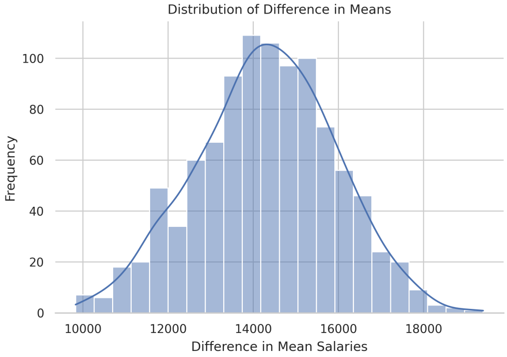


Fig 5. Distribution of difference in means

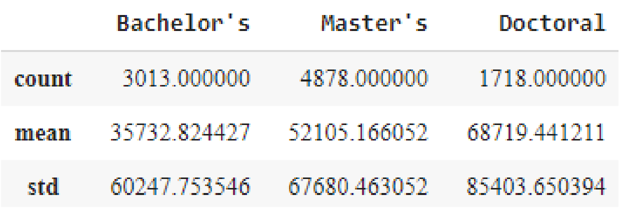


After bootstrapping, the normality test and the figures above suggest normality, which can also be confirmed with Central Limit Theorem (CLT). Since the Levene's test states that the two groups have different variances (see notebook), Welch's t-test is performed. With a p-value of 0.0 which is smaller than the 0.05 threshold, it is safe to determine that the difference between mean salaries for men and women is statistically significant. Along with the descriptive statistics and figure 4, we can conclude that the average salary for men is significantly higher than that for women.

Two-sample t-test is computationally light-weight, and easy to interpret. However, the independence, normality, and homoscedasticity assumptions need to be satisfied beforehand. Sampling method is also critical in this case, since convenience sampling or non-random allocation would render the test meaningless. The t-test should not be used for comparing more than two groups, otherwise Type I error would increase drastically.

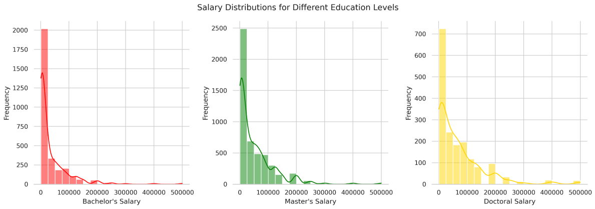
### 3. Estimating differences between mean salaries of  Bachelor’s vs. Master’s vs. Doctoral degree

Table 2. Descriptive statistics



| | Bachelor's | Master's | Doctoral |
|---|---|---|---|
| **count** | 3013.000000 | 4878.000000 | 1718.000000 |
| **mean** | 35732.824427 | 52105.166052 | 68719.441211 |
| **std** | 60247.753546 | 67680.463052 | 85403.650394 |

Fig 6. Salary distributions for different education levels



Based on figure 6 and the results from normality test and Levene’s test for homoscedasticity (see notebook), the original three groups of data are right-skewed and do not satisfy the homoscedasticity assumption. Thus, ANOVA is not recommended here. To compare the differences in means of the three education groups, bootstrapping is used to generate the sample distributions in the figures below. Distributions of the three bootstrapped samples and the between-group differences appear normal.

Fig 7. Bootstrapped salary distributions

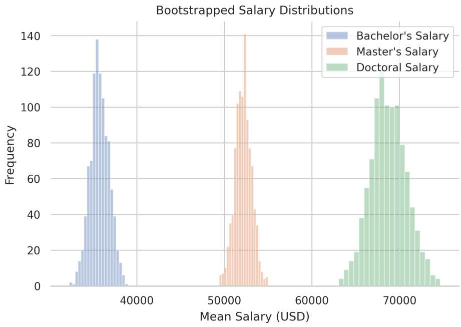
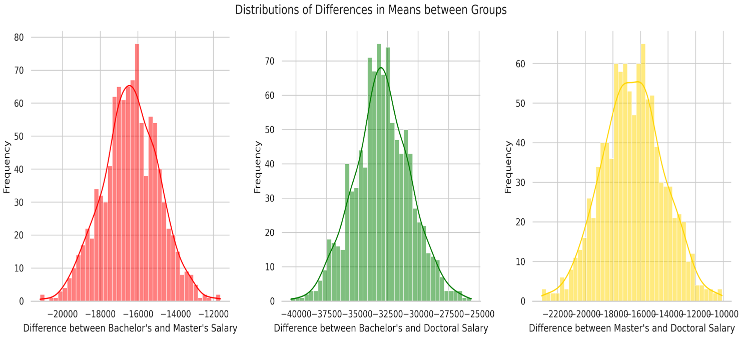


Fig 8. Distributions of between-group differences



After bootstrapping, the above plots agree with normality test results and CLT such that the resulting distributions are normal. The between-group mean differences also follow normal distributions. On average, Bachelor's salary is about $17,000 less than Master's and about $33,000 less than Doctoral’s, while Master's salary is about $17,000 less than Doctoral’s. Although Levene's test says that the three groups have different variances, since the standard deviations for the three bootstrapped mean distributions are actually not far apart (1071.8, 941.7, and 2043.4), we can proceed with the parametric ANOVA test. The ANOVA result (p-value=0) indicates that there is a significant difference between the average salaries of Bachelor's, Master's, and Doctoral participants.

Although ANOVA has a robust design and can increase statistical power, it only proves the existence of between-group differences without specification, which would require post-hoc tests or planned comparisons.