Rosalie Day – Statistical Data Analysis Capstone 1
Prediction of Income from Education and Sector Status
December, 2019

This study explores whether a person's income is less or greater than $50,000. The focus is on the education and work class variables. Descriptive statistics, distribution and count plots for this sample, included education levels, ages, work class (employment sector and status), hours worked per week and income. This study does not explore race, gender and marital information, although these will be used for classification of income outcomes in the predictor set.

Counts for education levels, zero grades completed through a doctorate degree versus income outcomes and work-class. High school graduates comprised the largest level in almost all work class, employment sector and status, categories. The "some college" was second largest followed by bachelor's degrees.

The inferential statistical analysis includes hypothesis testing:

1. Outcomes, greater than $50k income, are more likely with bachelor's degrees than any education level short of a bachelor's degree;
2. Work class is associated with education level; and
3. Education is correlated with hours worked.

The first tests the difference in income outcomes between having specifically a bachelor's degree (BA), and no graduate degrees, with not having a bachelor's degree (noBA). Appropriate filters were set to select only the rows of interest: the noBA had a sample size of 36372; and the BA, 8025. This hypothesis test uses a two-sample t-test from the scipy stats library with parameters set for equal variances.

The null hypothesis posits that there is no differential effect on the income outcome with respect to having a bachelor's degree; it is the same without a BA and with only a BA. The results were a significant t-statistic = -52.408681619739234 and a p-value = 0.0. From this comparison of income outcomes, the probability is 0.0 that the 2 samples, BA and noBA, are alike, therefore rejecting the null hypothesis. The alternative hypothesis that a bachelor's degree impacts the income outcome holds.

Second, testing the association of the categorical variables, education and workclass, will use the chi-squared test. The null hypothesis is that education and workclass are independent variables. A cross-tabulation of education and workclass was created from crosstab function in pandas. The chi-squared statistic = 3906.1080264021853 with degrees of freedom = 120 and p-value = 0.0 Therefore, the null hypothesis that these two variables are independent can be rejected. Workclass, which is a combination of work-sector and -status, and level of education are correlated.

Third, if the hours worked were correlated positively, that relationship could affect the income outcome. Testing for correlation of education level and hours worked per week is implemented with correlation pairs of the researchpy library. The null hypothesis, that educ_num and hours-per-week are independent variables, cannot be rejected with the very small r value of 0.1437. There appears to be no significant correlation between education levels and hours worked per week.