Allison Schlarmann

Suzette Beltran

**Statistical Methods and Computing**

**Analysis of Income Data**

The dataset (income.dat) used for this project was obtained from <http://homepage.divms.uiowa.edu/~kcowles/Datasets/income.info>

**Data Info:**

Personal income and demographic data from the March, 2011 supplement to

the Current Population Survey. Data on all 80,976 respondents aged 25

to 64 years who were currently in the labor force and who listed their

race as Asian, black, or white. This is a random sample from all such

residents of the United States.

variables and coding:

Sex 1=male, 2=female

Income Total personal income, dollars

Race Person's race, 1=white, 2=black, 4=Asian

Age Person's age in years

Educ Educational attainment,

 1=less than high school

 2=some high school but no diploma

 3=high school graduate

 4=some college but less than bachelor's degree

 5=bachelor's degree

 6=master's, professional, or doctoral degree

*Note: Alpha Value used for this dataset = 0.05*

**Research question**

 If there is a relationship among each individual variable and income, which category has the greatest mean income within each variable (education, race, age, gender)?

**Method of Analysis**

1. Determining the distribution of data:

After running the data through proc univariate, it is apparent the distribution is skewed right, with P-Value <.0001, and a large number of outliers. The distribution is not normal because the P-Value is not greater than the alpha value 0.05 or 5%.



The proceeding data was collected under a transformed distribution to normal since this distribution was skewed right.



Though there are some outliers, the distribution is somewhat normal.

1. Comparing income of genders:

 (\* Sex 1=Male, 2=female)

 ***H*0 :** μ1 = μ2

***H*a** : μ1 ≠ μ2

 We ran a two sample ttest on income.dat. The t value for pooled method was 31.57 and 31.56 for the Satterthwaite method. Both P-Values were <.0001, this indicates the null hypothesis (that the mean income for both genders is equal) is rejected and that the difference of income among genders is extremely statistically significant. Men receive the largest income.



The P-value at 0.0001 is statistically significant so we reject the null hypothesis that the mean income is equal for the two gender.



The distributions of wealth in terms of gender appear to be similar.





The mean log for males is $10.40, and $9.99 for females. We can infer the income for males is greater than females.

1. Comparing income of race:

One question we had about the data was how the mean income related in regards to race. In other words, we compared the mean incomes of the following races: white, black, and Asian. To do this we ran an ANOVA test in order to see which means, if any, were different from the others. Our hypothesis test was as follows:

**H0** : μw = μb = μa

**Ha** : μw ≠ μb and/or μw ≠ μa and/or μb ≠ μa

The following is the F test we computed to determine if we could or could not reject the null hypothesis:



The P-value at 0.0001 is statistically significant so we reject the null that the mean income for each race is equal.

Since P = 0.0001 < α = 0.05, we could reject the null and conclude that at least one mean income was different from the others.

The following is the t test we computed to determine which mean(s) was different:



From the above table, we can infer that the only mean incomes that do not differ significantly at α = 0.05 are whites and Asians.

From this, we could conclude that the mean incomes were not significantly different for whites compared to Asians, but they were significantly different for whites compared to blacks and Asians compared to blacks.

The following is the mean incomes in regards to race:



From this, we can infer that whites may have the greatest mean income.

1. Comparing income of education:

Another question we had about the data was how the mean income related in regards to education. In other words, we compared the mean incomes of the following education levels: (1) less than high school, (2) some high school, (3) high school graduate, (4) less than bachelor’s degree, (5) bachelor’s degree, and (6) master’s degree or above. To do this we ran an ANOVA test in order to see which means, if any, were different from the others. Our hypothesis test was as follows:

H0: μ1 = μ2 = μ3 = μ4 = μ5 = μ6

HA: μ1 ≠ μ2 and/or μ1 ≠ μ3 and/or μ1 ≠ μ4 and/or μ1 ≠ μ5 and/or μ1 ≠ μ6

and/or μ2 ≠ μ3 and/or μ2 ≠ μ4 and/or μ2 ≠ μ5 and/or μ2 ≠ μ6

and/or μ3 ≠ μ4 and/or μ3 ≠ μ5 and/or μ3 ≠ μ6

and/or μ4 ≠ μ5 and/or μ4 ≠ μ5

and/or μ5 ≠ μ6

The following is the F test we computed to determine if we could or could not reject the null hypothesis:



The P-value at 0.0001 is statistically significant so we reject the null hypothesis that the mean incomes for each education level are equal.

Since P = 0.0001 < α = 0.05, we could reject the null and conclude that at least one mean income was different from the others.

The following is the t test we computed to determine which mean(s) were different:



From the above table, we can infer that the only mean incomes that do not differ significantly at α = 0.05 are some high school and less than high school.

From this, we can conclude that the only mean incomes that are not significantly different are those of workers with less than a high school education and those with some high school. The mean incomes in regards to all other levels of education are significantly different.

The following is a comparison of the mean incomes in regards to education:

 

From this, we can infer that mean income increases as education level increase.

1. Relationship of age on income:

***H*0 :** β =0

***H*a :** β ≠ 0

Regression of income on age resulted in $R^{2}$ = 0.010 meaning there is not really a relationship of age on income. There is no significant age that has the largest income mean. P-Value is <.0001, the null hypothesis is rejected.



The regression line is y = 0.017x + 0.027 and R^2 = 0.010.

**SAS Code**

\*Create formats;

**proc** **format**;

value genderfmt **1** = 'male' **2** = 'female';

value racefmt **1** = 'white' **2** = 'black' **4** = 'Asian';

value educfmt **1** = 'less than high school'

 **2** = 'some high school'

 **3** = 'high school graduate'

 **4** = 'less than bachelor'

 **5** = 'bachelor'

 **6** = 'master or above';

**run**;

\*Read in data;

**data** income;

input gender wages race age educ;

format gender genderfmt.;

format race racefmt.;

format educ educfmt.;

logwages = log(wages+1);

datalines;

\*Copy and paste data here;

;

**run**;

\*Plot data to check if normal;

**proc** **univariate** plot data = income;

var logwages;

**run**;

\*Run a two sample t-test on the income means in regards to gender;

**proc** **ttest** data = income;

class gender;

var logwages;

**run**;

\*Check the means of each income in regards to gender;

**proc** **univariate** data = income;

class gender;

var logwages;

**run**;

\*Run ANOVA on mean incomes in regards to education level and check which means are different;

**proc** **anova** data = income;

class educ;

model logwages = educ;

means educ / bon alpha = **0.05**;

**run**;

\*Calculate mean incomes in regards to education;

**proc** **means** data = income;

class educ;

var logwages;

**run**;

\*Calculate the regression line of mean incomes in terms of age;

**proc** **reg** data = income;

model logwages = age;

**run**;

\*Run ANOVA on mean incomes in regards to race and check which means are different;

**proc** **anova** data = income;

class race;

model logwages = race;

means race / bon alpha = **0.05**;

**run**;

\*Calculate mean incomes in regards to race;

**proc** **means** data = income;

class race;

var logwages;

**run**;

**Contributions**

Suzette Beltran: Wrote summaries/interpretations of data, compiled data into one document, ran code, reviewed all project material

Allison Schlarmann: Wrote summaries/interpretations of data, ran code, reviewed all project material