



For this project, you should access Pew Research Center's 2014 dataset on Higher education, gender, and income ([GenderEduIncome.csv](#)  ).

Your job is to investigate the effect of sex and education on income. For example, you might hypothesize that sex and education each have a main effect on a person's income, but that the two interact, such that men with more education earn significantly more than women with more education.

Run a full-factorial two-way ANOVA. Report your findings.

Interpret the data. Are your hypotheses supported? What is the effect size(s)? Are there any limitations to your data analysis/findings? Create a bar graph that displays mean income by sex and education (include 95% confidence intervals as error bars). You may use any statistic software of your preference to create the bar graph.

Robert Bruffey

### INST314 Project 3: Factorial ANOVA

#### Research Question:

Does sex and the degree of education have an effect on income?

#### Hypothesis:

2\*3 factorial design

##### Main Effect: Sex

$H_0$ : Male and Females have equal mean incomes

$$\mu_{\text{male}} = \mu_{\text{female}}$$

$H_1$ : Male and Females mean incomes are not equal

$$\mu_{\text{male}} \neq \mu_{\text{female}}$$

##### Main Effect: Education

$H_0$ : Each level of education has the same mean incomes

$$\mu_{\text{college}} = \mu_{\text{highschoolgrad}} = \mu_{\text{lesshighschool}}$$

$H_1$ : There is at least one level of education that does not have the same mean income.

##### Interaction: Sex X Education

$H_0$ : differences between means of males and females are equal between education levels.

$H_1$ : there is at least one difference among means of males and females between education levels that is not equal.

## ANOVA Factorial Test

```
library(readr)
Income <- read_csv("~/Downloads/GenderEduIncome.csv")
View(income)
b = aov(appxincome~sex*educ,data=income)
summary(b)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	1.365e+10	1.365e+10	1.376	0.24092
educ	2	1.031e+11	5.156e+10	5.197	0.00563 **
sex:educ	2	8.169e+10	4.085e+10	4.117	0.01646 *
Residuals	1582	1.569e+13	9.921e+09		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

414 observations deleted due to missingness

## Effect Sizes

Overall  $R^2 = SS_{\text{between}}/SS_{\text{total}} = (1.365 \times 10^{10} + 1.031 \times 10^{11} + 8.169 \times 10^{10}) / (1.365 \times 10^{10} + 1.031 \times 10^{11} + 8.169 \times 10^{10} + 1.569 \times 10^{13}) = 1.9844 \times 10^{11} / 1.5888 \times 10^{13} = .01245$

The effect size value of .01245 tells us that there is a 1.245% of variance in the DV(Sex/Educ) explained by IV(ApproxIncome).

Sex  $\eta^2 = SS_{\text{between}}/SS_{\text{total}} = (1.365 \times 10^{10}) / (1.5888 \times 10^{13}) = 8.59 \times 10^{-4}$

The effect size value of  $8.59 \times 10^{-4}$  tells us that there is a .08% of variance in the DV(Sex) explained by IV(ApproxIncome).

Educ  $\eta^2 = SS_{\text{between}}/SS_{\text{total}} = (1.031 \times 10^{11}) / (1.5888 \times 10^{13}) = .006489$

The effect size value of .006489 tells us that there is a .6% of variance in the DV(Educ) explained by IV(ApproxIncome).

Sex:Educ  $= (8.169 \times 10^{10}) / (1.5888 \times 10^{13}) = .00514$

The effect size value of .00514 tells us that there is a .514% of variance in the DV(Sex and Educ) explained by IV(ApproxIncome).

## Observations:

Based off our results from the ANOVA Factorial test, we can see that the for the first main effect of sex has a p-value of 0.24092. Since, .24092 is greater than our significance level of .05, we therefore fail to reject our null hypothesis for the sex main effect. Therefore, we can still say that males and females have equal mean incomes throughout our data. For our next main effect of education, we found a p-value of 0.00563. Since, 0.00563 is less than .05 we can

reject our null hypothesis for our education main effect. By doing this we can conclude that there is at least one education level income mean that is not equal to the others. Finally, for our interaction between age and education, we got a p-value of 0.01646. Again, since 0.01646 is less than .05 we can therefore reject our null hypothesis for our interaction. By doing this, we can conclude that there is at least one difference in means among males and females between education levels that is not equal.

### **Limitations:**

Limitations from our data could be represented in that some of the data was found to be incomplete and reported an approximate income that was not applicable(N/A), this could mean that the person could have been unemployed. If the person was unemployed then perhaps they could have been recorded as having an income of \$0, but that would have caused the data to be much more skewed left. Another limitation of the data could be that we are not aware of the age of the data participants. This could mean that the person could have been older and have worked enough in a career to have a higher income than someone who just got out of college. Also, for the p-values that we found that were less than our significance level and we rejected the null hypothesis, we are not sure which category of the data that has a different mean income than the others. For example, we concluded that there is at least one educational mean income that is not equal to the others, but we are unsure as to which level of education was different than the others or if there was more than one that was different.

### **Subsets:**

- male = subset(income,sex == "Male") #male incomes  
m = subset(male, appxincome >0)
- female = subset(income, sex == "Female") #female incomes  
f = subset(female,appxincome > 0)
- college = subset(income, educ == "College") #college incomes  
c = subset(college,appxincome >0)
- high = subset(income, educ == "High School Grad") #high school grad incomes  
h = subset(high,appxincome>0)
- less = subset(income, educ == "Less High School") #less high school incomes  
l = subset(less,appxincome>0)
- mc = subset(m, educ == "College") #males who went to college incomes
- mh = subset(m,educ == "High School Grad") #males who graduated high school incomes
- ml = subset(m, educ == "Less High School") #males who did not graduate high school incomes
- fc = subset(f,educ == "College") #females who went to college incomes
- fh = subset(f, educ == "High School Grad") #females who graduated high school incomes
- fl = subset(f, educ == "Less High School") #females who did not graduate high school incomes

### Means

1.  $\mu_{\text{males}} = 89290.78$
2.  $\mu_{\text{females}} = 83626.01$
3.  $\mu_{\text{college}} = 92225.38$
4.  $\mu_{\text{highschoolgrad}} = 77184.58$
5.  $\mu_{\text{lesshighschool}} = 70384.62$
6.  $\mu_{\text{male/college}} = 91071.43$
7.  $\mu_{\text{male/highschoolgrad}} = 90390.95$
8.  $\mu_{\text{male/lesshighschool}} = 71791.67$
9.  $\mu_{\text{female/college}} = 93428.43$
10.  $\mu_{\text{female/highschoolgrad}} = 59837.84$
11.  $\mu_{\text{female/lesshighschool}} = 68465.91$

### 95% Confidence Intervals

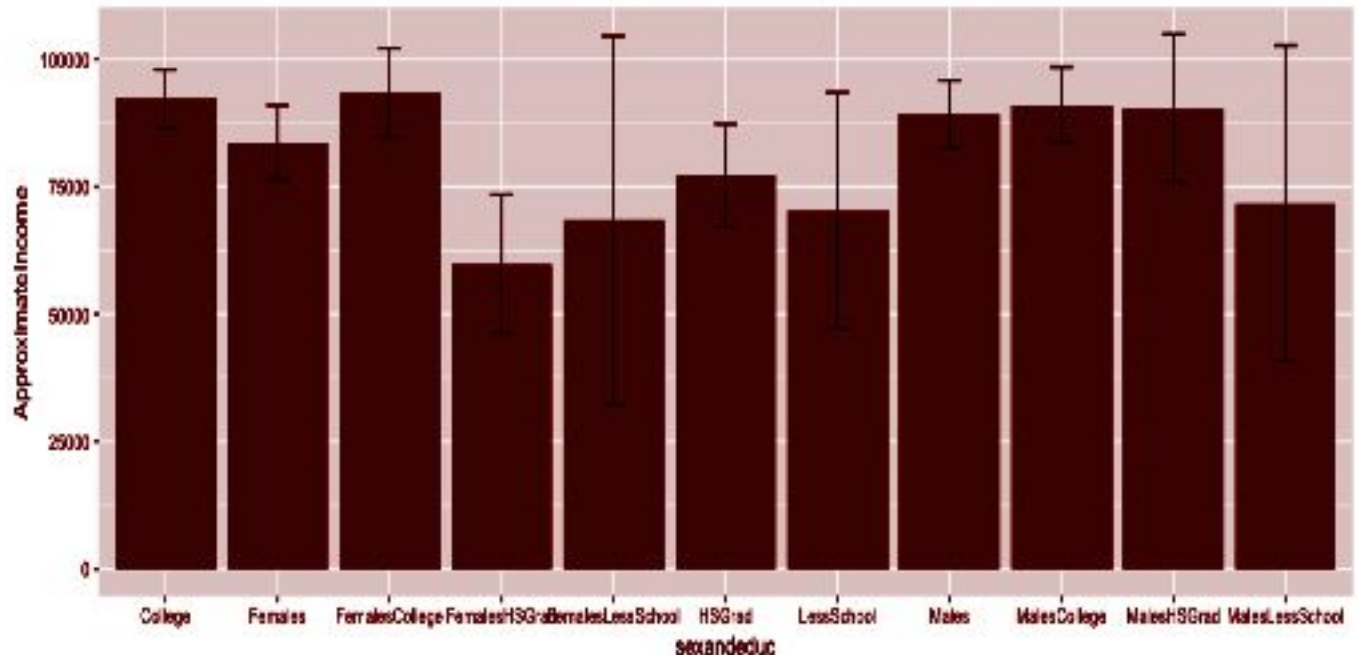
1. Male mean error = 6622.448, lower limit = 82668.33, upper limit = 95913.23
2. Female mean error = 7332.549, lower limit = 76293.46, upper limit = 90958.55
3. College mean error = 5722.325, lower limit = 86503.05, upper limit = 97947.7
4. High School Grad mean error = 10145.65 , lower limit = 67038.93, upper limit = 87330.23
5. Less High School mean error = 23078.1 , lower limit = 47306.52, upper limit = 93462.71
6. Male/College mean error = 7380.19, lower limit = 83691.24, upper limit = 98451.62
7. Male/High School Grad mean error = 14460.85, lower limit = 75930.1, upper limit = 104851.8
8. Male/Less High School mean error = 30885.76 , lower limit = 40905.9, upper limit = 102677.4
9. Female/College mean error = 8822.592, lower limit = 84605.84, upper limit = 102251
10. Female/High School Grad mean error = 13534.84, lower limit = 46303, upper limit = 73372.68
11. Female/Less high school mean error = 36116.26, lower limit = 32348.65, upper limit = 104582.2

### Bar Graph

```
> d.ci = data.frame(sexandeduc = c("Males", "Females", "College", "HSGrad", "LessSchool",  
"MalesCollege",  
"MalesHSGrad", "MalesLessSchool", "FemalesCollege", "FemalesHSGrad", "FemalesLessSchool"  
), ApproximateIncome =  
c(mean(m$appxincome), mean(f$appxincome), mean(c$appxincome), mean(h$appxincome), mean(l$appxincome), mean(mc$appxincome), mean(mh$appxincome), mean(ml$appxincome), mean(fl$appxincome), mean(fmc$appxincome), mean(fmh$appxincome), mean(fml$appxincome)))
```

```
fc$appxincome),mean(fh$appxincome),mean(fl$appxincome)), ci.upper =
c(95913.23,90958.55,97947.7,87330.23,93462.71,98451.62,104851.8,102677.4,102251.7,3372.
68,104582.2),ci.lower =
c(82668.33,76293.46,86503.05,67038.93,47306.52,83691.24,75930.1,40905.9,84605.84,46303
,32348.65))
```

```
> ggplot(d.ci,aes(x=sexandeduc,y =
ApproximateIncome))+geom_bar(stat="identity")+geom_errorbar(aes(ymin=ci.lower,ymax=ci.up
per),width=.2)
```



From the bar graph you can see that most of the data means are relatively in the same income range. As for outliers, females that are high school grads have the lowest mean incomes. Another interesting note is that females with less high school have the second lowest mean income, but then they also have the highest dispersion of their error bar for the 95% confidence interval. This means the data for females with less high school has the widest range of incomes, but based on the mean, most of them tend to be on the lower end of the spectrum.