HW4 Solutions

October 10, 2017
1 Introduction

A random sample of $n = 2584$ Americans with paying jobs in 2005 were selected from the 1979 National Longitudinal Survey of Youth (NLSY). They were asked about their income and the number of years of education in 2006. Education was simplified into 5 distinct levels: less than 12 years, 12 years (high school diploma), 13-15 years (some college), 16 years (bachelors degree) or more than 16 years. The question of interest is whether there is an association between income and educational level.

2 Statistical Methods Used

The data are summarized by means and SDs in Table 1. Side-by-side boxplots were used to compare the incomes of individuals based on their education level in Figure 1. Due to a few individuals with large incomes, the incomes are severely right-skewed with variability increasing as the median and mean income increase. Most individuals have incomes less than $100,000. Figure 1 makes it hard to see the spread in these majority of “common” incomes. The log$_{10}$-transformed incomes are displayed in a boxplot in Figure 2. As suggested by Display 3.8 in the textbook, the latter figure shows that on the log scale, the data are much more symmetric and also have about the same spread. Also, the spread in the majority of incomes that are less than $100,000 = 10^5$ are better represented.

Performing the ANOVA on the log-transformed incomes would appear to satisfy the normality and constant variance assumption of the ANOVA (that will wait until HW5). However, as instructed, the untransformed incomes were analyzed by ANOVA instead. The resulting ANOVA table is presented in Table 2. Because we have a large random sample ($n \geq 30$) from each group, we need not assume normality of the incomes for each group. However, Table 1 and Figure 1 suggest that the constant variance assumption is NOT met.

The incomes were also analyzed by a permutation test using the variance of the means as the test statistic to test whether there was any difference in mean incomes across the 5 educational levels. The approximate permutation distribution of the variance of the means test statistic over $10^5$ simulations is shown in the left pane of Figure 3. The variance in means for the actual data set is $4.45 \times 10^8$. There was not a single simulated data set that attained or exceeded this value; hence, $p$-value $\approx 0$.

In the right pane is the approximate permutation distribution of the $F$ test statistic used by ANOVA. The actual $F$ statistic for the data was $F = 89.7$. There was not a single simulated data set that attained or exceeded this value; hence, we still arrive at the conclusion that $p$-value $\approx 0$.

The ANOVA and permutation tests require that the data are from a RS from each group and that the groups are independent. These appear to be satisfied.
3 Summary of Statistical Findings

The data suggest that there is a difference in the mean income depending on educational level ($p$-value < 0.00005 via either ANOVA or permutation test). It appears that income increases, on the average, as the number of years of education increase. We will assess this statement with statistical significance in HW5.

4 Scope of Inference

Because these data are from a random sample from Americans who took the NLSY in 1979, then these results suggest that for all Americans who took the NLSY in 1979, mean income is associated with education. These data, from an observational study, do not suggest that education caused the difference in mean income.

5 Appendix

5.1 Tables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;12</td>
<td>28301.45</td>
<td>21021.90</td>
<td>136.00</td>
</tr>
<tr>
<td>12</td>
<td>36864.90</td>
<td>29369.73</td>
<td>1020.00</td>
</tr>
<tr>
<td>13-15</td>
<td>44875.96</td>
<td>33913.54</td>
<td>648.00</td>
</tr>
<tr>
<td>16</td>
<td>69996.97</td>
<td>64256.80</td>
<td>406.00</td>
</tr>
<tr>
<td>&gt;16</td>
<td>76855.46</td>
<td>65428.29</td>
<td>374.00</td>
</tr>
</tbody>
</table>

Table 1: Summary of incomes by 5 education levels

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>4</td>
<td>688235137515.90</td>
<td>172058784378.98</td>
<td>89.61</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residuals</td>
<td>2579</td>
<td>4951742721103.09</td>
<td>1920024319.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results of ANOVA

5.2 Figures
Figure 1: Incomes by Education level
Figure 2: Incomes by Education level
par(mfrow=c(1,2))

hist(var.mean, prob=T, main="Hist. of Var(means)",
     xlab="Var(means)") # a density histogram
abline(v=var.test.stat, lwd=3) # puts a vertical line at the observed variance

# Graph the approximate permutation distribution of the F-stat
hist(F, prob=T, main="Hist. of F", xlab="F = EMS/MSF") # a density histogram
abline(v=F.stat, lwd=3) # puts a vertical line at the observed variance
curve(df(x, 4, 24), add=TRUE, col=2, lwd=3)

5.3 R-code

# Import the data
library(Sleuth3)
d = ex0525
summary(d)

## Subject Educ Income2005
## Min. : 2 12 :1020 Min. : 63
## 1st Qu.:1586 13-15: 648 1st Qu.: 23000
## Median :3108 16 : 406 Median : 38231
## Mean : 3494 <12 : 136 Mean : 49417
## 3rd Qu.:4636 >16 : 374 3rd Qu.: 61000
## Max. :12140 Max. :703637

# Condition the data
d$Educ = factor(as.character(d$Educ), levels = c("<12", "12", "13-15", "16",">16"))
Mean = tapply(d$Income2005, d$Educ, mean)
SD = tapply(d$Income2005, d$Educ, sd)
n = tapply(d$Income2005, d$Educ, length)
sum.tab = cbind(Mean, SD, n)

# Make some plots of the data
boxplot(Income2005 ~ Educ, xlab="Education level in years",
       ylab="Income in dollars", data=d)
boxplot(log10(Income2005) ~ Educ, xlab="Education level in years",
       ylab="log10-transformed income in dollars", data=d)

# Conduct ANOVA
Figure 3: Simulation results: an approximation to the permutation distribution for two test statistics. The left pane shows a histogram of the variance of the means. The right pane shows a histogram of the F statistic.
m.aov = lm(Income2005 ~ Educ, data=d)
sum.aov = anova(m.aov)

var.test.stat = var(tapply(d$Income2005, d$Educ, mean))
var.test.stat

## [1] 445300306

F.stat = 89.613  # from the ANOVA

gmean = mean(d$Income2005)

num_sim = 10  # Draw num_sim randomizations
var.mean = numeric(num_sim)  # storage vector
EMS = numeric(num_sim)  # storage vector
MSF = numeric(num_sim)  # storage vector
gmean = mean(d$Income2005)

# generate num_sim random assignments and calculate the variance in means
for (i in 1:num_sim)
{
  grp <- sample(d$Educ, 2584, replace=FALSE)
  n.sim = tapply(d$Income2005, grp, length)
  Mean.sim = tapply(d$Income2005, grp, mean)
  var.mean[i] <- var(Mean.sim)
  EMS[i] <- sum(n.sim*(Mean.sim - gmean)^2)/4  # Extra credit
  MSF[i] <- sum((n-1)*SD^2)/sum(n-1)  # Extra credit
}

F = EMS/MSF

# Graph the approximate permutation distribution of the variance in means
par(mfrow=c(1,2))
hist(var.mean, prob=T, main="Hist. of Var(means)",
     xlab="Var(means)")  # a density histogram
abline(v=var.test.stat, lwd=3)  # puts a vertical line at the observed variance

# Graph the approximate permutation distribution of the F-stat
hist(F, prob=T, main="Hist. of F", xlab = "F = EMS/MSF")  # a density histogram
abline(v=F.stat, lwd=3)  # puts a vertical line at the observed variance
curve(df(x,4,24), add=TRUE, col=2, lwd=3)
# Get the p-value wrt var.test.stat
sum(var.mean>=var.test.stat)/num_sim

## [1] 0

# Get the p-value wrt F.test.stat
sum(F>=F.stat)/num_sim

## [1] 0