Econ 400 Problem Set 5 Question 1

Travis Cao

March 15, 2021

You have collected 14,925 observations from the Current Population Survey. There are 6,285 females in the sample, and 8,640 males. The females report a mean of average hourly earnings of \$16.50 with a standard deviation of \$9.06. The males have an average of \$20.09 and a standard deviation of \$10.85. The overall mean average hourly earnings is \$18.58.

(a) Using the t-statistic for testing differences between two means (section 3.4 of your textbook), decide whether or not there is sufficient evidence to reject the null hypothesis that females and males have identical average hourly earnings.

Our testing hypotheses are the following:

$$\begin{aligned} H_0 &: \overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}} = 0 \\ H_1 &: \overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}} \neq 0 \end{aligned}$$

To perform such test, we need the standard error of estimator $\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}$. By default, we should actually consider the case that the two population variances don't naturally equal to one another. This means that our standard error will use the unequal population variance formula:

$$\begin{aligned} se(\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}) &= \sqrt{\frac{s^2_{\text{female}}}{n_{\text{female}}} + \frac{s^2_{\text{male}}}{n_{\text{male}}}} \\ &= \sqrt{\frac{9.06^2}{6285} + \frac{10.85^2}{8640}} \\ &= 0.1634 \end{aligned}$$

so that our t-statistic under unequal population variance is constructed as the following:

$$\begin{aligned} t &= \frac{(\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}) - (\mu_{\text{female},H_0} - \mu_{\text{male},H_0})}{\sqrt{\frac{s^2_{\text{female}}}{n_{\text{female}}} + \frac{s^2_{\text{male}}}{n_{\text{male}}}}} \\ &= \frac{(16.50 - 20.09) - 0}{0.1634} \\ &= -21.9706 \end{aligned}$$

To find out the critical value, we need the degree of freedom for our t-distribution, which is calculated as the following under unequal population variance:

$$\begin{aligned}\frac{(s^2_{\text{female}}/n_{\text{female}} + s^2_{\text{male}}/n_{\text{male}})^2}{\frac{(s^2_{\text{female}}/n_{\text{female}})^2}{n_{\text{female}}-1} + \frac{(s^2_{\text{male}}/n_{\text{male}})^2}{n_{\text{male}}-1}} &= \frac{(9.06^2/6285 + 10.85^2/8640)^2}{\frac{(9.06^2/6285)^2}{6285-1} + \frac{(10.85^2/8640)^2}{8640-1}} \\ &= \frac{(0.0131 + 0.0136)^2}{\frac{0.0131^2}{6284} + \frac{0.0136^2}{8639}} \\ &= 14368\end{aligned}$$

Intuitively, the sample size is very large (14925 observations), so it makes sense that the degree of freedom for t-distribution is also very large. And under a large sample size, t-distribution collapses onto a standard normal z-distribution.

Under 5% size, the cutoff value for a two-sided z-distribution is 1.96. And since $|t| = 21.9706 > 1.96$, we reject the null hypothesis and accepts the alternative under 5% significance level.

(b) You decide to run two regressions: first, you simply regress average hourly earnings on an intercept only. Next, you repeat this regression, but only for the 6,285 females in the sample. What will the regression coefficients be in each of the two regressions?

The first model regress average hourly earnings on an intercept only:

$$AHE_i = \beta_0 + u_i$$

In this model, the intercept represents the mean level of AHE across the whole population, since

$$\begin{aligned}&E[AHE_i] = E[\beta_0 + u_i] = E[\beta_0] + E[u_i] = \beta_0 + 0 = \beta_0 \\ \Rightarrow \quad &\hat{\beta}_0 = \overline{AHE}\end{aligned}$$

so in this model, $\hat{\beta}_0 = 18.58$, the overall sample AHE mean.

The second model also regress average hourly earnings on an intercept only, but this time the sample size is restricted to only females. In this case, the intercept represents the mean level of AHE among females, since

$$\begin{aligned}E[AHE_i|\text{female} = 1] &= E[\beta_0 + u_i|\text{female} = 1] \\ &= E[\beta_0|\text{female} = 1] + E[u_i|\text{female} = 1] = \beta_0 \\ \Rightarrow \quad \hat{\beta}_0 = \overline{AHE}_{\text{female}}&\end{aligned}$$

so in this model, $\hat{\beta}_0 = 16.50$, the sample AHE mean among females.

(c) Finally you run a regression over the entire sample of average hourly earnings on an intercept and a binary variable DFemme where this variable takes on a value of 1 if the individual is a female, and is 0 otherwise. What will be the value of the intercept? What will be the value of the coefficient of the binary variable?

In this model:

$$AHE_i = \beta_0 + \beta_1 \text{DFemme}_i + u_i$$

Recall that

$$\begin{aligned} E[AHE_i|\text{female} = 0] &= E[\beta_0 + \beta_1 \text{DFemme}_i + u_i|\text{female} = 0] = \beta_0 \\ E[AHE_i|\text{female} = 1] &= E[\beta_0 + \beta_1 \text{DFemme}_i + u_i|\text{female} = 1] = \beta_0 + \beta_1 \end{aligned}$$

so β_1 represents the differences in mean AHE between male and female groups:

$$E[AHE_i|\text{female} = 1] - E[AHE_i|\text{female} = 0] = \beta_1$$

When estimating these βs, the estimated values are thus the sample analog of population parameters:

$$\begin{aligned} \hat{\beta}_0 &= \overline{AHE}_{\text{male}} = 20.09 \\ \hat{\beta}_1 &= \overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}} = 16.50 - 20.09 = -3.59 \end{aligned}$$

(d) What is the standard error on the slope coefficient? What is the t-statistic?

From (c), we know that the slope coefficient is recording the difference between female and male average AHE levels. Based on same calculation in (a) (assuming unequal population variance between male and female groups):

$$\begin{aligned} se(\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}) &= \sqrt{\frac{s^2_{\text{female}}}{n_{\text{female}}} + \frac{s^2_{\text{male}}}{n_{\text{male}}}} \\ &= \sqrt{\frac{9.06^2}{6285} + \frac{10.85^2}{8640}} \\ &= 0.1634 \end{aligned}$$

and

$$\begin{aligned} t &= \frac{(\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}) - (\mu_{\text{female},H_0} - \mu_{\text{male},H_0})}{\sqrt{\frac{s^2_{\text{female}}}{n_{\text{female}}} + \frac{s^2_{\text{male}}}{n_{\text{male}}}}} \\ &= \frac{(16.50 - 20.09) - 0}{0.1634} \\ &= -21.9706 \end{aligned}$$

Same conclusion as (a): we reject the null hypothesis that $\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}} = 0$ at 5% significance level.

(e) Had you used the homoskedasticity-only standard error in (d) and calculated the t-statistic, how would you have had to change the test-statistic in (a) to get the identical result?

If homoskedasticity-only error is assumed, then the two groups should be equal variance, since homoskedasticity states that

$$Var(AHE|\text{DFemme}) = Var(\beta_0 + \beta_1\text{DFemme} + u|\text{DFemme}) = Var(u|\text{DFemme}) = \sigma^2$$

Thus, $Var(AHE|\text{DFemme} = 0) = Var(AHE|\text{DFemme} = 1) = \sigma^2$, meaning that male and female group in population has the same variance.

The only adjustment then needed in (a) is to use the equal variance formula from 310:

$$t = \frac{(\overline{AHE}_{\text{female}} - \overline{AHE}_{\text{male}}) - (\mu_{\text{female},H_0} - \mu_{\text{male},H_0})}{\sqrt{\frac{s^2_{\text{pooled}}}{n_{\text{female}}} + \frac{s^2_{\text{pooled}}}{n_{\text{male}}}}} \quad \sim \quad t_{n_{\text{female}}+n_{\text{male}}-2}$$

where

$$s^2_{\text{pooled}} = \frac{\sum_i^{n_{\text{female}}}(AHE_{\text{female},i} - \overline{AHE}_{\text{female}})^2 + \sum_i^{n_{\text{male}}}(AHE_{\text{male},i} - \overline{AHE}_{\text{female}})^2}{n_{\text{female}} + n_{\text{male}} - 2}$$