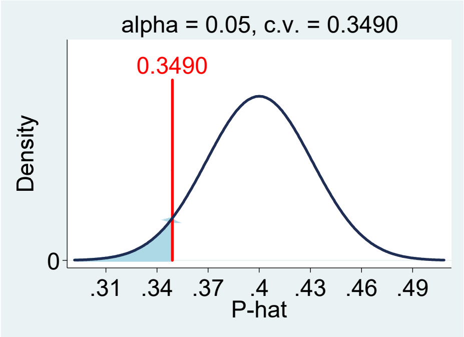
ECO220Y1Y, Test #4, Prof. Murdock: SOLUTIONS

March 8, 2019, 9:10 – 11:00 am

NOTE: The parts of the solutions [in brackets] are extra explanations and are *not* required parts of your answer.

(1)



$H_0: p = 0.40$

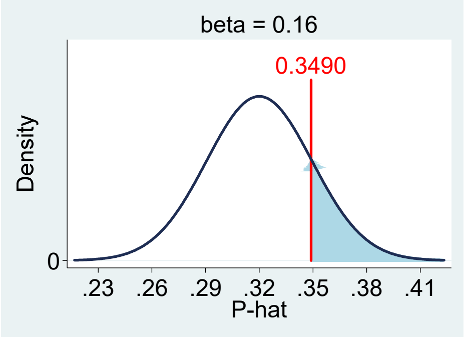
$H_1: p < 0.40$

Find standardized rejection region: $P(Z < -z_\alpha) = \alpha;\ P(Z < -1.645) = 0.05$

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \Rightarrow -1.645 = \frac{\hat{P} - 0.40}{\sqrt{\frac{0.40(1-0.40)}{250}}} = \frac{\hat{P} - 0.40}{0.030984} \Rightarrow c.v. = 0.3490$$

The unstandardized rejection region is $(-\infty,\ 0.3490)$ or $(0, 0.3490)$.

[In a random sample of 250 students, we need less than 34.9% supporting the changes to prove at a 5% significance level that less than 40% of the population supports the changes.]



Beta is the probability of NOT being in the rejection region (i.e. failing to reject the false null).

$$\beta = P(\hat{P} > 0.3490 \mid p = 0.32, n = 250) = ?$$

$$\beta = P(\hat{P} > 0.3490) = P\left(Z > \frac{0.3490 - 0.32}{\sqrt{\frac{0.32(1-0.32)}{250}}}\right) = P\left(Z > \frac{0.029}{0.029503}\right) =$$

$$P(Z > 0.98) = 0.5 - 0.3365 = 0.16$$

(2) (a) The ME is 0.048657491 $\left(= \frac{5.29445751 - 5.197142527}{2}\right)$. The relevant formula is: $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$. The degrees of freedom is very large, $\nu = 12{,}778$, so use $t_{\alpha/2} = 1.960$ for $\alpha = 0.05$. Hence, the margin of error (ME) is $1.960 * \frac{s}{\sqrt{12{,}779}}$.

Solving for s in:

$$0.048657491 = 1.960 * \frac{s}{\sqrt{12{,}779}}$$

yields $s = 2.81$.

(b) We need to make an inference about the difference in means using *independent* samples.

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

Depending on whether or not you assume equal variances, the test statistic is $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ or $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$.

[Either of the above two answers is acceptable in this case.]

(c) We need to make an inference about the difference in proportions (using *independent* samples).

$H_0: p_2 - p_1 = 0$

$H_1: p_2 - p_1 \neq 0$

The test statistic is: $z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}}$

(d) In China during the period from 2006-2017, on average GDP per capita grew by approximately 7.7 percent per year, which is very rapid growth.

(e) In China during the period from 2006-2017, in years where GDP per capita was 10 percent higher, on average the mean happiness (measured on a scale from 0 to 10) was 0.09 higher, which is quite small.

[To see that it is small, note that the average happiness (on a 10-point scale) in China is more that 2 below Canada: China is 5.246 versus Canada at 7.328. A 0.09 increase does not put much of a dent in the huge gap.]

(f) The histogram of the residuals for **Regression #1** must be Histogram <u>**D**</u>.

[Histogram A is way too spread out: none of the observations are above the line by 0.5 or more. Histogram B clearly has a mean above zero and the mean of the residuals must be zero. Histogram C clearly has a mean below zero. Histogram E is far too spread out: none of the observations are below the line by 0.8 or more. Histogram F is not nearly spread out enough: for example, for year 2009 the residual is below -0.2 but Histogram F says there are no residuals below -0.08.]

(3) (a)

$V[B_M - B_E] = 21.5^2 = 462.25$

$V[B_M - B_E] = V[B_M] + V[B_E] - 2 * CORR[B_M, B_E] * SD[B_M] * SD[B_E]$

$462.25 = 19.0^2 + 20.9^2 - 2 * CORR[B_M, B_E] * 19.0 * 20.9$

$CORR[B_M, B_E] = 0.42$

(b) This requires making an inference about the difference in means using *paired* data (<u>*not*</u> independent samples).

$H_0: \mu_d = 0$

$H_1: \mu_d \neq 0$

$t = \frac{\bar{d} - \Delta_0}{s_d/\sqrt{n}} = \frac{0.71 - 0}{\frac{19.5}{\sqrt{5{,}268}}} = \frac{0.71}{0.2687} = 2.64$ with degrees of freedom $\nu = n - 1 = 5{,}268 - 1 = 5{,}267$

$P - value \approx 2 * (0.5 - 0.4959) = 0.0082$

Yes, the difference is highly statistically significant, including at the 1% significance level.

(c) No. The difference between math scores, which average 44.9, and English scores, which average 44.2, is only 0.71, which is not even one point out of 100 points and is tiny. There is no meaningful difference in average performance in these two subjects.

[This tiny difference is statistically significant is because of the huge sample size.]

(d) From the aid sheets: $s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i-0)^2}{n-2}}$. The STATA output reports SSE of 1418634.27.

Plugging in: $s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1418634.27}{5{,}256-2}} = 16.432.$

(e) From the aid sheets: $b_1 \pm t_{\alpha/2} s.e.(b_1)$.

Plugging in: $0.3615425 \pm 1.960 * 0.0137301$, which is 0.3615425 ± 0.0269. This yields a LCL of 0.3346 and a UCL of 0.3884. Use 1.960 because of the very large degrees of freedom ($\nu = 5{,}254$).

[Note that we are not accurate to the 4th decimal place because we used rounded value of 1.960. See output below.]

```
      Source |       SS           df       MS      Number of obs   =     5,256
-------------+----------------------------------   F(1, 5254)      =    693.38
       Model |  187220.818         1  187220.818   Prob > F        =    0.0000
    Residual |  1418634.27     5,254   270.01033   R-squared       =    0.1166
-------------+----------------------------------   Adj R-squared   =    0.1164
       Total |  1605855.09     5,255  305.586126   Root MSE        =    16.432

------------------------------------------------------------------------------
     overall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   b_overall |   .3615425   .0137301    26.33   0.000     .3346259    .3884592
       _cons |   24.18992   .8856223    27.31   0.000     22.45373    25.92611
------------------------------------------------------------------------------
```

(4) (a) $10{,}000 * \frac{3{,}337}{407{,}846} = 10{,}000 * 0.008182 = 81.8$

[*Why* rate per 10,000 rather than proportions? For rare events, like ADHD, the proportions are so tiny that people have trouble interpreting them. Researchers often report rates per # people, where # is higher for rarer events. For example, the murder rate per 100,000 people.]

(b) To answer requires making an inference about the difference in population proportions using *hypothesis testing*.

Define Group 2 to be the *youngest*: $\hat{P}_2 = \frac{320+309}{36{,}577+36{,}319} = \frac{629}{72{,}896} = 0.00862873$

Define Group 1 to be the *oldest*: $\hat{P}_1 = \frac{225+240}{35{,}353+34{,}405} = \frac{465}{69{,}758} = 0.00666590$

To test if it is higher for younger kids requires a one-tailed test:

$H_0: (p_2 - p_1) = 0$

$H_1: (p_2 - p_1) > 0$

$z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}}$ where $\bar{P} = \frac{X_1+X_2}{n_1+n_2}$

$$\bar{P} = \frac{X_1+X_2}{n_1+n_2} = \frac{629+465}{72{,}896+69{,}758} = \frac{1{,}094}{142{,}654} = 0.007668905$$

$$z = \frac{(\hat{P}_2 - \hat{P}_1) - 0}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}} = \frac{0.00862873 - 0.00666590}{\sqrt{0.007668905(1-0.007668905)\left(\frac{1}{72{,}896} + \frac{1}{69{,}758}\right)}} = \frac{0.00196283}{\sqrt{0.00761009(0.00002805)}} = \frac{0.00196283}{0.00046202} = 4.25$$

The P-value is ≈ 0. The evidence is *extremely strong* to support the inference that the rate of ADHD diagnoses is higher for the younger children. We easily meet a 1% significance level and we even easily meet a 0.1% significance level.