Homework 17: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) Call μ_0 the population mean donated amongst those that would give money and are offered no match. Call μ_1 the population mean donated amongst those that would give money and are offered a 1:1 match. Similarly μ_2 and μ_3 correspond to a 2:1 and 3:1 match. The question asks three different things that would require three hypothesis tests: $H_0: \mu_1 - \mu_0 = 0$ versus $H_1: \mu_1 - \mu_0 > 0$; $H_0: \mu_2 - \mu_0 = 0$ versus $H_1: \mu_2 - \mu_0 > 0$; $H_0: \mu_3 - \mu_0 = 0$ versus $H_1: \mu_3 - \mu_0 > 0$. There is no point in doing these tests because we clearly have NO EVIDENCE in favor of our research hypotheses. Why? Because in all three cases the sample average amount donated with a match is LESS THAN the average amount donated with no match. If we did the formal hypothesis tests, our P-values would be above 0.5 (i.e. huge): we have no evidence that offering a match increases the mean amount donated among those donating.

(b) One sentence is potentially misleading: "We find that the match offer increases both the revenue per solicitation and the response rate." The seems to imply two effects when there is only one effect. The *proportion giving* money does increase when you offer a match compared to no match (i.e. the match causes an increase in the response rate). However, that is the only effect. The ONLY REASON the mean revenue per solicitation goes up is because a higher proportion of people give and so you are averaging in fewer zeros. [NOTICE: This paper is a good example of an analysis that involves both comparisons of proportions and comparisons of means: it illustrates concepts from Chapters 11 - 14.]

(c) In this case, because we know the amount donated cannot be less than 0, the mean and s.d. ($79.99 and $627.06) alone show the obvious presence of an outlier (or outliers). If we made the mistake of going ahead with the analysis and compared the amount donated for a 3:1 match versus a 2:1 match:

$H_0: \mu_3 - \mu_2 = 0$ versus $H_1: \mu_3 - \mu_2 \neq 0$ (Note: the question said difference, not increase)

$$S.E.(\bar{X}_3 - \bar{X}_2) = \sqrt{\frac{393201.5}{253} + \frac{1871.691}{252}} = 39.5$$

$$t = \frac{79.98696 - 45.3373}{39.5} = 0.88$$

$$\nu = \frac{(s_1^2 + s_2^2)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = 254.4 \approx 254$$

From our Student t table, obtain the critical value for 250 degrees of freedom (a good approximation) and a significance level of 0.05 for this two-tailed test: 1.969. The rejection region would be (-∞, -1.969) and (1.969, ∞). The test statistic is not in the rejection region so we fail to reject the null hypothesis. In other words, we do not have a statistically significant difference at a 5% level. We can approximate the P-value with the Student t table to be greater than 0.20 (remember it is a two-tailed test). Hence this result is not statistically significant at any reasonable significance level. [Note: You may be surprised that this large positive outlier that pulled up the mean so much did not cause a statistically significant difference. The outlier also made the s.d. very large, which the formulas interpret as raising sampling error.] The difference between an average donation of $79.99 and $45.34 would certainly be economically significant even if it is not statistically significant. However, because a single data point single-handedly caused this large this difference, we would not say that we have an economically significant result.

(2) (a) These are paired data and must be analyzed as such. First, you must remember Section 9.3 and Lecture 8 regarding the variance of linear combinations of variables. Applying that here for comparing <u>December with January</u>:

$$V[dec12 - jan13] = V[dec12] + V[jan13] - 2COV[dec12, jan13]$$

But we've been given correlations so use the other version of the formula:

$V[dec12 - jan13] = V[dec12] + V[jan13] - 2 * r * SD[dec12] * SD[jan13]$

$V[dec12 - jan13] = 415541.3 + 319237.4 - 2 * 0.5298 * 644.6249 * 565.011 = 348836.6$

$SD[dec12 - jan13] = \sqrt{348836.6} = 590.6239$

We also know that: $MEAN[dec12 - jan13] = MEAN[dec12] - MEAN[jan13] = 163.1968$

(Note: Your answers may differ very slightly because of rounding: the above numbers come directly from STATA.)

$H_0: \mu_d = 0$ versus $H_1: \mu_d \neq 0$

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{590.6239}{\sqrt{1200}} = 17.0498$$

$$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} = \frac{163.1968 - 0}{17.0498} = 9.57$$

$\nu = n - 1 = 1{,}199$

Here the critical value can be obtained from the Standard Normal table because the degrees of freedom are huge. However, we need not bother because the t test statistic is enormous and falls deep within any rejection region with a P-value = 0. Hence there is definitely a statistically significant difference in the mean credit card spending comparing the month of December with January. Further, the point estimate of the difference is $163.20, which is also economically significant.

Using the same approach to compare January and February we obtain:

$V[feb13 - jan13] = 317313.2 + 319237.4 - 2 * 0.4843 * 563.3056 * 565.011 = 328270.54$

$SD[feb13 - jan13] = \sqrt{328270.54} = 572.94898$

We also know that: $MEAN[feb13 - jan13] = MEAN[feb13] - MEAN[jan13] = 565.605 - 564.9852 = 0.6198$

$H_0: \mu_d = 0$ versus $H_1: \mu_d \neq 0$

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{572.94898}{\sqrt{1200}} = 16.539612$$

$$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} = \frac{0.6198 - 0}{16.539612} = 0.04$$

$\nu = n - 1 = 1{,}199$

Here the critical value can be obtained from the Standard Normal table because the degrees of freedom are huge. However, we need not bother because the t test statistic is tiny and will not fall within any rejection. Using the Normal table the P-value = 0.968 (remember this is a two-tailed test). Hence there is definitely not a statistically significant difference in the mean credit card spending comparing the month of January with February. Further, the point estimate of the difference is $0.62, which is also not economically significant.

Using the same approach to compare January and March we obtain:

$$V[mar13 - jan13] = 362509.8 + 319237.4 - 2 * 0.4555 * 602.0878 * 565.011 = 371837.54$$

$$SD[mar13 - jan13] = \sqrt{371837.54} = 609.78483$$

$$MEAN[mar13 - jan13] = MEAN[mar13] - MEAN[jan13] = 602.761 - 564.9852 = 37.7758$$

$H_0: \mu_d = 0$ versus $H_1: \mu_d \neq 0$

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{609.78483}{\sqrt{1200}} = 17.602972$$

$$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} = \frac{37.7758 - 0}{17.602972} = 2.15$$

$$\nu = n - 1 = 1{,}199$$

Here the critical value can be obtained from the Standard Normal table because the degrees of freedom are huge. At a 5% significance level the rejection region is (-∞, -1.96) and (1.96, ∞). Our test statistic falls in the rejection region so we reject the null and infer that there is a statistically significant difference in mean credit card spending comparing January and March. Using the Normal table the P-value = 0.0316 (remember this is a two-tailed test). Hence there is a statistically significant difference in the mean credit card spending comparing the month of January with February at the 5% level (but not at a 1% significance level). The point estimate of the difference is $37.78, which may or may not be economically significant. Given that average monthly spending is consistently over $500, a $37.78 difference is not very big (at least from the perspective of an individual customer).

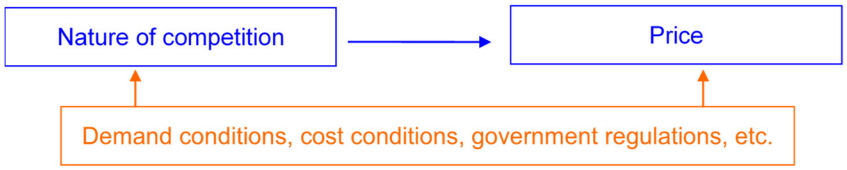
(b) The 95% CI for mean difference from December to January is (-196.6, -129.7). You should show your work. The 95% CI for mean difference from January to February is (-31.8, 33.1). The 95% CI for mean difference from January to March is (3.2, 72.3). The first interval is interpreted as we are 95% confident that the mean credit card spending of ALL of our customers (not just the 1200 in the sample) decreased between $127.7 and $196.6. You should interpret the other two cases.

(c) The hypothesis tests would be less powerful and the confidence interval estimates would be wider if we failed to recognize that these data are paired.

(3) (a) Hypothesis testing for Q1 and confidence interval estimation for Q2.

(b) These data are cross-sectional: different markets at the same point in time. These data are also observational. The presence of a monopolist or competition in a particular market has neither been randomly assigned by a researcher (as in experimental data) nor randomly assigned by other external forces (as in a natural experiment). In fact, firms *choose* whether or not to enter a market and compete, which means that the presence of a monopolist or competition in a particular market is not randomly set. This is the defining feature of observational data.

(c) This diagram illustrates the research question (in blue) and the confounding effects (in orange).



(d) The effect that we are interesting in is the blue arrow: how does the nature of competition (monopoly or competition) affect the price in a particular market. The confounding effects are that across different geographic markets the demand conditions, cost conditions and government regulations will vary. These differences would not be a problem if they only affected price, but we know that they will also affect the nature of competition because firms choose whether or not to enter a market based on demand conditions, cost conditions, government regulations, etc. It is the presence of the first orange arrow that makes these data observational and makes the nature of competition an endogenous variable. Unfortunately, this means that bias will creep into our inference about the magnitude of the blue arrow, which represents our research question. If we attribute all of the differences in price across markets to the presence or absence of a monopolist, we will have a biased estimate. The reason is that the other things (orange box) are systematically different among monopolized and competitive markets and part of the differences in price is attributable to these factors. Hence attributing all of the differences in price to the nature of competition would be wrong (i.e. would suffer an endogeneity bias).

(e) Suppose that isolated rural areas tend to be monopolized and have high input costs (expensive to ship gasoline) or that cities tend to have many competitors but high prices due to high taxes and expensive land. This illustrates how locations may not be otherwise comparable and how the confounding factors will affect price and the nature of competition. This leaves us with the troubling question: If two locations are really comparable, then why is one monopolized while other has competition? Without using more advanced techniques (you would learn in a 300-level statistics/econometrics course) it will be impossible to isolate the effect we are interested in: the confounding effects will be tangled up and cause bias.

(f) Define μ_M as the average price in all monopolized retail gasoline markets (population mean). Define μ_C as the average price in all competitive retail gasoline markets (population mean).

$H_0: (\mu_M - \mu_C) = 0$

$H_1: (\mu_M - \mu_C) > 0$

$$t = \frac{(\bar{X}_M - \bar{X}_C) - (\mu_M - \mu_C)}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_C^2}{n_C}}} = \frac{(1.87 - 1.80) - (0)}{\sqrt{\frac{0.0258199^2}{4} + \frac{0.0907115^2}{8}}} = \frac{0.07}{0.03457223} = 2.02$$

$$\nu = \frac{\left(\frac{s_M^2}{n_M} + \frac{s_C^2}{n_C}\right)^2}{\frac{1}{n_M - 1}\left(\frac{s_M^2}{n_M}\right)^2 + \frac{1}{n_C - 1}\left(\frac{s_C^2}{n_C}\right)^2} = \frac{\left(\frac{0.0258199^2}{4} + \frac{0.0907115^2}{8}\right)^2}{\frac{1}{3}\left(\frac{0.0258199^2}{4}\right)^2 + \frac{1}{7}\left(\frac{0.0907115^2}{8}\right)^2} \approx 9$$

Rejection region at a 5% significance level is (1.83, ∞). Because the test statistic of 2.02 falls in the rejection region, reject the null and conclude that prices are higher, in a statistically significant way, in markets that are monopolized compared to competitive. [Note: You may have also answered using the P-value approach and found that the P-value given the test statistic of 2.02 is between 0.05 and 0.025.] This is NOT the same question as Q1. Q1 is the *causal research question*. The question asked for part (f) is simply a *descriptive question*. There is a HUGE conceptual difference between asking whether prices are statistically different and asking about what caused that difference.

(g) No. No, we cannot conclude that monopolies cause higher prices. We have observational data and we believe that our control variable (the nature of competition) is endogenous. Hence, our sample means will be systematically different from each other not only because of the nature of competition but also because of other systematic differences across markets (cost structure, demand structure, etc.). Our statistical analysis above does not control for these other differences. It simply compares the raw means: the average in the monopolized markets and the average in the competitive markets. Further it attributes ALL differences in these means to either sampling noise or to differences in the nature of competition. But we know that other things cause a difference in the mean prices and differences in the

nature of competition. Hence, our analysis is biased. Despite a small P-value (which would be great if we did not have an endogenous control variable) we cannot conclude that monopolies cause higher prices. All we can say is that monopolized markets tend to have higher prices than competitive markets but that could be due to not only the nature of competition but also to other unobserved factors like demand structure and costs.

(If the other factors (cost structure, demand structure, etc.) did not cause differences in firms' choices about entering market and hence the nature of competition, then we would NOT have a problem and we could conclude causality. It is OK if these other factors affect price, but it is not OK that they also affect the nature of competition. Unfortunately it is entirely implausible to suggest that the nature of competition is exogenous and hence we cannot infer causality in this example.)

(4) (a) H_0: $\mu_{wage,real} - \mu_{wage,fake} = 0$ and H_1: $\mu_{wage,real} - \mu_{wage,fake} \neq 0$ (Note: You may have specified a directional hypothesis but most often researchers are referring to a two-tailed tests.) Of all the methods in Chapters 11, 12, 13, and 14 the method you would use in this case is hypothesis testing to make an inference about the difference between population means (mean wages) for independent samples (Section 14.2).

(b) H_0: $p_{unemployed,real} - p_{unemployed,fake} = 0$ and H_1: $p_{unemployed,real} - p_{unemployed,fake} < 0$ Of all the methods in Chapters 11, 12, 13, and 14 the method you would use in this case is hypothesis testing to make an inference about the difference between population proportions (proportion unemployed) (Section 12.8).

(c) For Part (a), the point estimate would be difference in the sample mean wages from the data used in the study: $\left(\bar{X}_{wage,real} - \bar{X}_{wage,fake}\right)$. For Part (b), the point estimate would be difference in the sample proportions that are unemployed from the data used in the study: $\left(\hat{P}_{unemployed,real} - \hat{P}_{unemployed,fake}\right)$.

(d) If a researcher says that the difference in wages is not significant, this means that the difference is either not statistically significant, not economically significant, or is both not statistically significant and not economically significant. Any difference in wages may have been small (in terms of dollars) and/or any difference may simply be because of sampling error.

(e) No. While this research and these results are interesting they are not conclusive. The authors of the paper present many cautions and caveats (one of which is visible in the included excerpt). Wages are also self-reported so maybe the people who lied and claimed a fake degree are also lying about their wages (saying that they are higher than they really are). Also, certainly the study is only looking at wages: most people would agree that part of the "value" of being educated (to both you and society) is separate from wages.

(5) (a) Panel B breaks out the original sample of 24,646 lottery participants (i.e. the winners and losers) into three different subgroups: those that had "no visits" to the ED BEFORE the lottery (i.e. the healthy people), those with "one visit" BEFORE the lottery (i.e. sometimes not healthy people), and those with two or more visits to the ED BEFORE the lottery (i.e. the chronically unhealthy people). Notice the description written in the title of Panel B and how 16,930 + 3,881 + 3,835 = 24,646. Within each subgroup (e.g. the healthy subgroup) some people won the lottery (got insurance) and others lost (no insurance). Hence what the Panel B results show is that regardless of whether we look at people who were healthier or sicker PRIOR to the lottery, we see POSITIVE and economically significant increases in the ED use with coverage (the opposite of what politicians had predicted). However, not all the results are statistically significant.

(b) Define p_C to be the proportion of all people in the control group (i.e. no Medicaid, lost the lottery) who did not have any visits to the ED in the pre-randomization period (i.e. before the lottery) that did visit the ED after the lottery. Define p_T to be the proportion of all people in the treatment group (i.e. got Medicaid, won the lottery) who did not have any visits to the ED in the pre-randomization period (i.e. before the lottery) that did visit the ED after the lottery.

H_0: $(p_T - p_C) = 0$

H_1: $(p_T - p_C) \neq 0$

$$z = \frac{\hat{P}_T - \hat{P}_C}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_T} + \frac{\bar{P}(1-\bar{P})}{n_C}}}$$

$$\bar{P} = \frac{X_T + X_C}{n_T + n_C}$$

The table tells us that $n_T + n_C = 16{,}903$. However, it does not report how many people are in the control group and treatment group. But it does tell us that $\hat{P}_T - \hat{P}_C = 0.067$ and that $\sqrt{\frac{\bar{P}(1-\bar{P})}{n_T} + \frac{\bar{P}(1-\bar{P})}{n_C}} = 0.029$ (and also that $\hat{P}_C = 0.225$ which implies that $\hat{P}_T = 0.292$). Hence we can find the z value:

$$z = \frac{\hat{P}_T - \hat{P}_C}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_T} + \frac{\bar{P}(1-\bar{P})}{n_C}}} = \frac{0.067}{0.029} = 2.31$$

$P - value = P(z < -2.31) + P(z > 2.31) = 2 * 0.0104 = 0.0208$. The table reports a P-value of 0.019: ours is off a tiny bit because we used rounded numbers in our calculations (and we used the Normal table instead of software).

(c) Define μ_C to be the mean number of ED visits post-randomization (i.e. after the lottery) of all people in the control group (i.e. no Medicaid, lost the lottery) who did not have any visits to the ED in the pre-randomization period (before the lottery). Define μ_T to the mean number of ED visits post-randomization (i.e. after the lottery) of all people in the treatment group (i.e. got Medicaid, won the lottery) who did not have any visits to the ED in the pre-randomization period (before the lottery).

$H_0: (\mu_T - \mu_C) = 0$

$H_1: (\mu_T - \mu_C) \neq 0$

While it is clear that these are independent samples (not paired data), it is unclear if they assumed equal variances or not. Even though your textbook cautions against it, researchers often use the equal variances assumption.

$$t = \frac{(\bar{X}_T - \bar{X}_C) - \Delta_0}{\sqrt{\frac{s_p^2}{n_T} + \frac{s_p^2}{n_C}}}$$

From the table $(\bar{X}_T - \bar{X}_C) = 0.261$ and $\Delta_0 = 0$ (and also that $\bar{X}_C = 0.418$, which implies that $\bar{X}_T = 0.679$). Again, the table does not give us enough to compute s_p^2 and n_1 and n_2. But, it does tell us that $\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = 0.084$. Hence we can find the t value: $t = \frac{0.261}{0.084} = 3.11$. Give the large degrees of freedom, we can use the Normal table very accurately approximate the P-value. $P - value = P(t < -3.11) + P(t > 3.11) \approx 2 * P(Z > 3.11) = 2 * 0.0009 = 0.0018$. The table reports a P-value of 0.002 and our calculations round to exactly that.

(d) There are two reasons. One is the bigger difference in mean visits comparing the treatment and control groups (0.652 versus 0.380): other things equal, this leads to a bigger test statistic and smaller P-value. (It is easier to reject a null hypothesis of no difference in mean visits comparing the two groups if we see a big difference in the sample.) However, if you look at the standard error for that estimate, it is also smaller than the next row (0.254 versus 0.648). Remember the formula for the standard error of the difference between two sample means, independent samples: $SE[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$, which is an estimate of $SD[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$. In addition to sample sizes, it is also a function of the variance of ED visits among those in the treatment group and those in the control group. These must be smaller to explain the smaller standard error. It makes sense the s.d. would be smaller for the group that had one ED visit compared to the group that had two+ visits as the latter group likely includes some rather unwell people who had lots of visits: i.e. a long right tail.