SOLUTIONS

Part 1:

(1) (a)

$H_0{:}\,(p_2 - p_1) = 0$
$H_1{:}\,(p_2 - p_1) \neq 0$

$$\bar{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{0.34 * 1007 + 0.39 * 1005}{1007 + 1005} = 0.365$$

$$z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}} = \frac{0.39 - 0.34}{\sqrt{\frac{0.365(1-0.365)}{1007} + \frac{0.365(1-0.365)}{1005}}} = 2.33$$

$$P - value = P(Z < -2.33) + P(Z > 2.33) = 0.0099 + 0.0099 = 0.0198$$

Hence the 5 percentage point change in positive public opinions on immigration is statistically significant at a 5% significance level but not at a 1% significance level. (Note to markers: The above treated population 1 as 2010 and population 2 as 2012 but students could have reversed it and obtained a test statistic of -2.33. The P-value is identical as is the interpretation and conclusion.)

(b) $CI\ estimator{:}\ \hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = \hat{P} \pm ME$

The sample would include about 291 18 – 34 year olds: 291 = 0.29*1005. Hence assuming a 5% significance level (i.e. 95% confidence level) the ME of for 18 – 34 year olds is: $ME_{18-34} = z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = 1.96\sqrt{\frac{0.48(1-0.48)}{291}} = 0.057$

For all adults it would be: $ME_{18+} = z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = 1.96\sqrt{\frac{0.39(1-0.39)}{1005}} = 0.030$

There are two reasons that the margin of error (ME) is bigger for the young adults. The biggest reason is that we are only looking at a sub-sample and hence have a much smaller sample size (291 versus 1005). Another reason is that the ME is bigger the closer the proportion is to 0.5: hence even if the sample sizes were equal the ME for young adults (18 – 34 years old) would be larger.

(c)

$H_0{:}\,p = 0.5$
$H_1{:}\,p > 0.5$

The sub-sample of those aged 35+ will be about 714 (=0.71*1005).

$$z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.58 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{714}}} = 4.28$$

The P-value is about 0 (< 0.001), which means we have overwhelming statistical evidence in favor of the research hypothesis: we can conclude at any reasonable significance level that in 2012 a majority of middle-aged to older Canadians are of the opinion that illegal immigrants who are working in Canada should be deported.

(2) (a)

$H_0: (\mu_1 - \mu_2) = 0$
$H_1: (\mu_1 - \mu_2) \neq 0$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(84.25 - 84.09) - 0}{\sqrt{\frac{4.644^2}{340} + \frac{4.609^2}{322}}} = \frac{0.163}{0.3597} = 0.453$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{4.644^2}{340} + \frac{4.609^2}{322}\right)^2}{\frac{1}{339}\left(\frac{4.644^2}{340}\right)^2 + \frac{1}{321}\left(\frac{4.609^2}{322}\right)^2} = \frac{0.016745203}{0.000025427} \approx 658$$

Given the large degrees of freedom (>400) we can use the Standard Normal table as an excellent approximation to the Student t table. $P - value \approx P(Z < -0.453) + P(Z > 0.453) = 0.33 + 0.33 = 0.66$

Note to markers: If students assumed equal variances it makes no difference because the sample variances are very similar to each other. Assuming equal variances is also an acceptable approach in this particular example.

$$s_p^2 = \frac{(340 - 1)4.644^2 + (322 - 1)4.609^2}{340 + 322 - 2} = 21.409$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(0.163) - 0}{\sqrt{\frac{21.409}{340} + \frac{21.409}{322}}} = 0.453$$

$$\nu = n_1 + n_2 - 2 = 340 + 322 - 2 = 660$$

(b) Students graduating high school in five years are on average 0.862 years older than those graduating in four years and this difference is highly statistically significant. (This is not surprising.) There is no significant difference in the high school average across these groups: the difference is tiny – much less and one half of one percent – and is not statistically significant. High school grades differ very little between those finishing in four versus five years. However, the GPA in first year of university is significantly *lower* for those who graduated high school in 4 years: the difference is highly statistically significant and the point estimate is 0.237 out of 4.0, which is a large difference.

(3) (a)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15}$
$H_1: Not\ all\ the\ slopes\ are\ zero$

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} = \frac{0.316/15}{(1 - 0.316)/(662 - 15 - 1)} = 19.9$$

This test statistic is deep into the rejection region even at a 1% significance level (the table does not give the exact critical value for this case but it would be around 2), which means that this model is highly statistically significant overall. (The P-value of this test is near zero.) (Note to markers: Students may have missed the note in the table regarding the five indicator variables for each parent's level of education (total of 10 assuming two parents) and hence incorrectly thought that $k = 5$. However, if everything else is correct do NOT deduct a point for this.]

(b) After controlling for high school average, age, sex, immigrant status, and parent's level of education those students who graduated high school in only four years have grades in Intro to MGT that are 8 percentage points lower and first year university GPAs that are 0.6 lower on a four-point scale. Both differences are highly significant: both statistically and in the very large size of the differences in grades in that course and in first-year overall.

(c) After controlling for graduating in only four years, age, sex, immigrant status, and parent's level of education those students who have a high school average that is one percentage point higher on average have an Intro to MGT grade that is 1.03 percentage points higher and a GPA that is 0.09 points higher out of 4.00.

(d) *No* absolutely not. The coefficients must be interpreted holding the other included x variables fixed. The negative and highly statistically significant coefficient on age means that after controlling for whether a student graduated in four or five years, high school average, sex, immigrant status, and parent's level of education those students who are older tend to earn lower marks. If we wished to see how age is associated with performance generally, we would run a simple regression (i.e. with only one x variable: age). I suspect that there is a POSITIVE and not negative association between university performance and age: look back at question (2): students that are one year older (five years of high school) perform better. Do not make the mistake of interpreting a multiple regression coefficient as the association between two variables: it is not.

(e) In the regression where the grade in *Intro to MGT* is the y variable, the coefficient on female is statistically significant and negative $\left(t = \frac{-2.421}{0.717} = -3.38\right)$. After controlling for graduating in only four years, high school average, age, immigrant status, and parent's level of education, females on average earn marks that are 2.4 percentage points lower than males in Intro to MGT. However, in the regression where *GPA* is the y variable, the coefficient on female is not statistically significant $\left(t = \frac{-0.047}{0.057} = -0.82\right)$: we cannot rule out the possibility that there is no difference between males and females in first year university grades overall (i.e. GPA) after controlling for graduating in only four years, high school average, age, immigrant status, and parent's level of education.

(f) This hypothesis CANNOT be tested using the reported regression results. However, we CAN test it WITHOUT collecting more data. We could use the same data but we would need to run two new regressions that are like the existing two regressions but that also include an interaction term: *Female*Four-year graduate*. If the hypothesis is true we would expect to see a positive and statistically significant coefficient on that interaction term, which means that the negative effect of finishing high school in four years is less for females than males after controlling for other included x variables.

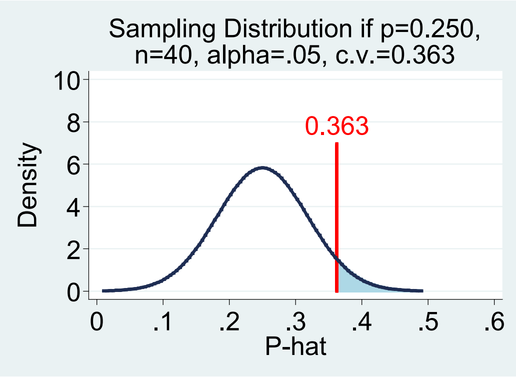
(4) The *power* of the statistical test is 0.35.

$H_0: p = 0.25$
$H_1: p > 0.25$

$$\sigma_{\hat{P}} = SD[\hat{P}] = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.25(1-0.25)}{40}} = 0.0685$$

$$P(Z > 1.645) = 0.05$$

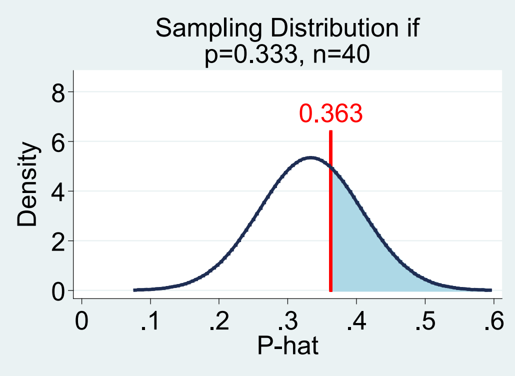
$$Critical\ Value = c.v. = 1.645 * 0.0685 + 0.25 = 0.363$$



$$Power = P\left(\hat{P} > 0.363 \mid p = \frac{1}{3}, n = 40\right)$$

$$\sigma_{\hat{P}} = SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1/3(1-1/3)}{40}} = 0.0745$$

$$Power = P\left(Z > \frac{0.363-1/3}{0.0745}\right) = P(Z > 0.40) = 0.5 - 0.1554 = 0.3446 \approx 0.35$$



(Note: With software the exact value is 0.347 but you cannot get an exact answer with just the provided statistical table.)

Part 2:

(1) If the y variable were population size (i.e. not logged), the coefficient on *time* would be ___. **(A)**

(2) On average, population grew by ___ in Edmonton between 2001 and 2013. **(B)**

(3) If the variable *time* were measured as the year (i.e. 2001, 2002, …, 2013) what would change? **(C)**

(4) If we had monthly data instead of annual data we would expect an R^2 that is ___. **(A)**

(5) Provided that the sample size is sufficiently large, what does the Central Limit Theorem say? **(B)**

(6) If you randomly select a person in Canada receiving EI benefits in Jan. 2000, what is the chance that person lives in Ontario and is in "Management occupations"? **(B)**

(7) For "Trades, transport and equipment operators and related occupations" in Ontario in Jan. 2013, which is true? **(B)**

(8) If you randomly select three EI beneficiaries in Jan. 2013, what is the chance two of them live in Ontario? **(E)**

(9) Looking at the eleven rows of EI numbers by occupational category for Canada, a correlation matrix of the EI numbers for every possible pair of years is greater than 0.98 in all cases. Looking back at the table, clearly Jan. 2010 stands out relative to the others. How can it be that Jan. 2010 is extremely highly correlated with the other years? **(A)**

(10) Regarding the distribution of the variable *Item list price ($)*, which is CORRECT? **(D)**

(11) Regarding the distribution of the variable *Purchase rate (purchases/views)*, which is CORRECT? **(B)**

(12) Which of the four variables shows signs of being negatively skewed (i.e. left skewed)? **(B)**

(13) Consider the variable *Average viewer distance (kilometers)*. Using its mean and standard deviation reported in the table, if it were perfectly Normally distributed what should the 10th percentile be? **(C)**

(14) For Model 1 (top three graphs), which is the FUNDAMENTAL problem? **(C)**

(15) For Model 2 (bottom three graphs) an OLS regression yields a coefficient of 1.3 on $\ln(Salary_i)$. It is highly statistically significant. How should you interpret it? **(D)**

(16) The exact OLS results for Model 2 are $ln(\widehat{Sales})_{\imath} = 0.7628 + 1.3165 * ln(Salary_i)$ with R^2 = 0.29. Consider data for one additional company where the CEO has a salary of $1 (virtually works for free). Which value of *Sales* for this extra company would result in the biggest *increase* in the R^2 if the regression were re-run including it (i.e. n = 177)? **(A)**

(17) A 2012 paper "The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases" uses several experiments. In one experiment, participants make a series of choices between $1 with certainty or taking a bet. Specifically, in each round the participant can get $1 (for sure) or call heads or tails as the experimenter flips a coin in plain view. If the participant is correct s/he gets \$2.50 but otherwise gets nothing. Each participant plays 15 rounds. Suppose you are a participant and you decide to bet in 10 of the 15 rounds. Your expected *total payoff* is \$17.50. What is the standard deviation of your *total payoff*? **(D)**

(18) Looking at the results in Column (1), is the coefficient on *Return* statistically significant (i.e. $\neq 0$)? **(A)**

(19) Looking at the results in Column (2), is the coefficient on *Return*NEG* statistically significant (i.e. $\neq 0$)? **(A)**

(20) Looking at the results in Column (3), is the coefficient on *Return* statistically significant (i.e. $\neq 0$)? **(B)**

(21) Looking at the regression results in Column (1), how should you interpret the coefficient on *Return*? After controlling for university fixed effects and a time trend, on average when a university obtains a return on its endowment that is **___. (D)**

(22) Looking at regression (4), in which of these cases should you use the 0.81 coefficient (and *not* the 0.14 coefficient) to help predict the change in the dependent variable (y-variable)? **(C)**

(23) The inclusion of *University fixed effects* in all regressions enables the models to control for the fact that ___. **(A)**

(24) How does the number of x variables (right-hand-side variables) differ between Regressions (1) and (4)? **(C)**

(25) In testing the slope of a simple regression, which would be a Type I error? **(B)**

(26) To address the research question, how should the hypotheses be set up? **(D)**

(27) Which conclusion should be drawn? There is ___ with a pay-for-performance compensation scheme. **(E)**

(28) The exact same experiment is repeated with employees at a different firm. It obtains a P-value less than 0.0001 for the same hypothesis test. Which conclusion should be drawn? **(C)**