ECO220Y1Y, Test #3, Prof. Murdock: SOLUTIONS

January 18, 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)

$H_0: p = 0.75$

$H_1: p < 0.75$

$\hat{P} = 0.739$

$$z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.739 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{1{,}000}}} = \frac{-0.011}{\sqrt{0.0001875}} = \frac{-0.011}{0.0137} = -0.803$$

$$P - value = P(Z < -0.803) \approx 0.5 - 0.2881 = 0.21$$

Given this large P-value, we do *not* have sufficient evidence to prove that in 2018 less than 75% of Canadians either oppose or somewhat oppose allowing Stats Canada to access personal financial data. While the sample proportion is below 75%, given the sample size of 1,000, we simply haven't proven (at any reasonable burden of proof) that the population proportion – i.e. for *all* adult Canadians – is below 75%.

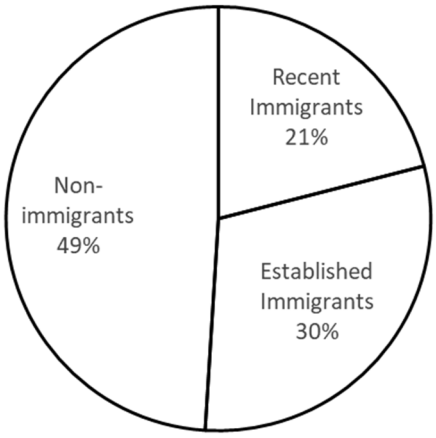
[Remember that we *cannot* say that we "accept the null" or that we have "proven the null" or the "evidence supports the null." While we cannot rule out that 75% of all Canadians either oppose or somewhat oppose the access, there are many other null hypotheses that we could also not rule out (e.g. $H_0: p = 0.74$). That does not make them all true.]

(2) True. While the income distribution is extremely positively skewed, the *percentiles* will be Uniformly spread out and the chance of any equal width interval is the same. For *income*, there is much more density at low income values than high ones – again, because of positive skew – so that the same width of interval in dollars does NOT have the same probability: it will be much higher in the lower income area.

(3) (a) **Toronto's Segregated Immigrant Population, 1981**

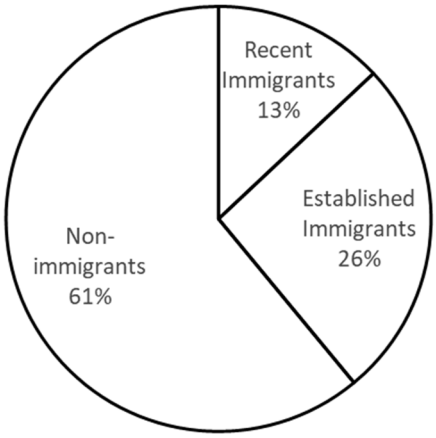
Low Income Neighbourhoods

531,300 people



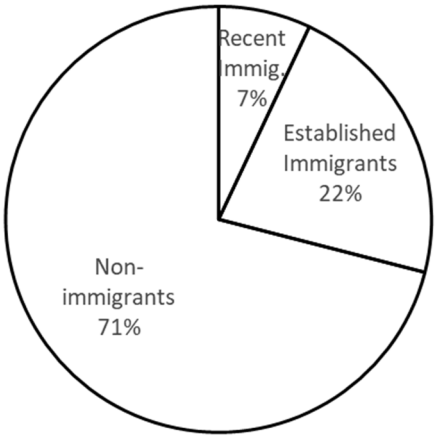
Middle Income Neighbourhoods

1,329,300 people



High Income Neighbourhoods

239,400 people



Work on next page >>>>>

For three blanks, find marginal probabilities – $P(L), P(M), P(H)$ – and multiply by population in 1981.

Low Income Neighbourhoods: $531{,}300 = 2{,}100{,}000 * (0.124 + 0.053 + 0.076) = 2{,}100{,}000 * 0.253$

Middle Income Neighbourhoods: $1{,}329{,}300 = 2{,}100{,}000 * (0.387 + 0.081 + 0.165) = 2{,}100{,}000 * 0.633$

High Income Neighbourhoods: $239{,}400 = 2{,}100{,}000 * (0.081 + 0.008 + 0.025) = 2{,}100{,}000 * 0.114$

To fill in the pie charts, compute the nine relevant conditional probabilities and multiply by 100 to get percentages.

Low Inc.: $P(N \mid L) = \frac{P(N\&L)}{P(L)} = \frac{0.124}{0.253} = 0.49$, $P(R \mid L) = \frac{P(R\&L)}{P(L)} = \frac{0.053}{0.253} = 0.21$, $P(E \mid L) = \frac{P(E\&L)}{P(L)} = \frac{0.076}{0.253} = 0.30$

Mid. Inc.: $P(N \mid M) = \frac{P(N\&M)}{P(M)} = \frac{0.387}{0.633} = 0.61$, $P(R \mid M) = \frac{P(R\&M)}{P(M)} = \frac{0.081}{0.633} = 0.13$, $P(E \mid M) = \frac{P(E\&M)}{P(M)} = \frac{0.165}{0.633} = 0.26$

High Inc.: $P(N \mid H) = \frac{P(N\&H)}{P(H)} = \frac{0.081}{0.114} = 0.71$, $P(R \mid M) = \frac{P(R\&H)}{P(H)} = \frac{0.008}{0.114} = 0.07$, $P(E \mid M) = \frac{P(E\&H)}{P(H)} = \frac{0.025}{0.114} = 0.22$

(b)

Joint Probability Table: Toronto's <u>NOT</u> Segregated Immigrant Population, <u>2016</u>

	L	**M**	**H**
N	0.269	0.149	0.112
R	0.067	0.037	0.028
E	0.172	0.095	0.071

Absent segregation, the income level of the neighborhood and immigration status would be *independent*. For independent events $P(A \,\&\, B) = P(A) * P(B)$. Hence, answering requires finding the correct marginal probabilities for 2016 and then multiplying them to compute what the joint probabilities would be without segregation.

[For further explanation, recall Question (1) on Test #2, Nov. 2018. Also, Professor Hulchanski defines exactly what he means by segregation in the figure titled "Toronto's Segregated Immigrant Population, 2016." Showing the segregation issue is the *purpose* of that figure. Also, he uses the word segregation in a mainstream manner. Comparing the differences across the three pie charts shows the segregation. Again, see Question (1) on Test #2, Nov. 2018 for how the pie charts would look if there were no segregation.]

$P(L) = 0.508 = \frac{1{,}368{,}000}{1{,}368{,}000+757{,}000+568{,}000}$ $\quad$ $P(N) = \frac{1{,}425{,}700}{1{,}425{,}700+910{,}300+355{,}700} = 0.530$

$P(M) = 0.281 = \frac{757{,}000}{1{,}368{,}000+757{,}000+568{,}000}$ $\quad$ $P(E) = \frac{910{,}300}{1{,}425{,}700+910{,}300+355{,}700} = 0.338$

$P(H) = 0.211 = \frac{568{,}000}{1{,}368{,}000+757{,}000+568{,}000}$ $\quad$ $P(R) = \frac{355{,}700}{1{,}425{,}700+910{,}300+355{,}700} = 0.132$

$P(N\&L) = P(N) * P(L) = 0.530 * 0.508;\ P(N\&M) = 0.530 * 0.281;\ P(N\&H) = 0.530 * 0.211$

$P(R\&L) = P(R) * P(L) = 0.132 * 0.508;\ P(R\&M) = 0.132 * 0.281;\ P(R\&H) = 0.132 * 0.211$

$P(E\&L) = P(E) * P(L) = 0.338 * 0.508;\ P(E\&M) = 0.338 * 0.281;\ P(E\&H) = 0.338 * 0.211$

(4) $P(X < 2{,}400) + P(X > 4{,}200) = P\left(Z < \frac{2{,}400-3{,}147}{492}\right) + P\left(Z > \frac{4{,}200-3{,}147}{492}\right) = P(Z < -1.52) + P(Z > 2.14) = (0.5 - 0.4357) + (0.5 - 0.4838) = 0.08$

Eight percent of newborns either weigh less than 2,400 grams or weigh more than 4,200 grams.

(5)

$E\left[\sum_{i=1}^{17} P_i + \sum_{i=1}^{4} L_i + \sum_{i=1}^{2} M_i + \sum_{i=1}^{3} H_i\right] = 17 * 2{,}000 + 4 * 4{,}500 + 2 * 8{,}700 + 3 * 16{,}000 = 117{,}400\ kg$

$V\left[\sum_{i=1}^{17} P_i + \sum_{i=1}^{4} L_i + \sum_{i=1}^{2} M_i + \sum_{i=1}^{3} H_i\right] = 17 * 800^2 + 4 * 1{,}100^2 + 2 * 1{,}900^2 + 3 * 2{,}500^2 = 41{,}690{,}000\ kg^2$

$SD = \sqrt{41{,}690{,}000} = 6{,}457\ kg$

[The mean (i.e. expected) *total* weight is 117,400kg with a standard deviation of 6,457kg. We can add variances – $V\left[\sum_{i=1}^{17} P_i + \sum_{i=1}^{4} L_i + \sum_{i=1}^{2} M_i + \sum_{i=1}^{3} H_i\right] = \sum_{i=1}^{17} V[P_i] + \sum_{i=1}^{4} V[L_i] + \sum_{i=1}^{2} V[M_i] + \sum_{i=1}^{3} V[H_i]$] – because an independence assumption *is* reasonable in this context (i.e. that the weights are unrelated across vehicles).]

(6) (a) We need to make an inference about the difference in population proportions.

Define Group 2 to be the *youngest*: $\hat{P}_2 = \frac{317+320+309}{34{,}415+36{,}577+36{,}319} = \frac{946}{107{,}311} = 0.0088155$

Define Group 1 to be the *oldest*: $\hat{P}_1 = \frac{225+240+232}{35{,}353+34{,}405+31{,}285} = \frac{697}{101{,}043} = 0.0068981$

The point estimate of the difference is $\left(\hat{P}_2 - \hat{P}_1\right) = 0.0019174$. Next, obtain the CI estimate of the difference:

$$(\hat{P}_2 - \hat{P}_1) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$$

$$0.0019174 \pm 1.96\sqrt{\frac{\frac{946}{107{,}311}\left(1-\frac{946}{107{,}311}\right)}{107{,}311} + \frac{\frac{697}{101{,}043}\left(1-\frac{697}{101{,}043}\right)}{101{,}043}}$$

$$0.0019174 \pm 1.96 * 0.0003863$$

$$0.0019174 \pm 0.0007571$$

The lower confidence limit (LCL) is 0.00116 and the upper confidence limit (UCL) is 0.00267.

We are 95% confident that among all kindergarteners the insurance claims-based rate of ADHD diagnoses per 10,000 children is between 11.6 to 26.7 *higher* for the youngest kindergarteners born in June, July or August compared to the oldest kindergarteners born in September, October or November. The rate per 10,000 children is 88.2 for the youngest versus 69.0 for the oldest, which is a whopping 28 percent higher. Even the lower confidence limit is that it is 17 percent higher. Hence, it seems that youth is inappropriately causing higher diagnosis rates of ADHD. The causal interpretation can be justified on the grounds that birth month is pretty much random and unlikely to be correlated with any lurking/unobserved/confounding/omitted variables. [In fact, the authors check for any possible confounding variables and do not find any.] [**NOTE:** It is valid to do the entire analysis switching the definition of groups 1 and 2, so long as you are consistent: e.g. the interpretation would be how much *lower* the diagnosis rate is for older children.]

(b) Unlike the comparison in Part (a) where we compared the oldest versus the youngest quartiles, we cannot be confident that there is any difference between the slightly older and slightly younger kindergarteners (the middle two quartiles) because the CI estimate of the difference spans both positive and negative values. We are 90% confident that ADHD diagnoses per 10,000 children is between 2.1 *lower* to 11.4 *higher* when comparing children born in March through May (slightly younger) with those born in December through February (slightly older). This is not surprising as there is not much age difference between kindergarteners in the second and third quartiles of the age distribution whereas there is a big age difference between the first and fourth quartiles of the age distribution. Also, even though we use only a 90% confidence level here, which gives a narrower interval, we still cannot be sure that being a little younger increases the ADHD diagnosis rate compared to being a little older.

(7) (a) In a random sample of 10 adult Canadians in 2016, there is about a 2/10,000 chance of getting a sample mean age as small as 29.7 years, which means such a young sample mean is extremely unlikely.

In formal notation: $P(\bar{X} \leq 29.7 \mid \mu = 48.520,\ \sigma = 18.369,\ n = 10) \approx \frac{2}{10{,}000}$.

(b) The only reasonable choice is Histogram #3. A sample should be similar to the population, aside from sampling error, and a sample size of 100 is pretty large. A Normal shape would be highly unlikely given the shape of the population. Further, Histogram #2 shows far too little variation. Histograms #1 and #2 confuse the distribution of $\bar{X}$ with the distribution of X.

[Note the question clearly asks about the distribution of a sample, *not* the sampling distribution of the sample mean. To reinforce that, all three histogram choices had the x-axis labeled as "Age (in years)." If we were thinking about a sampling distribution of the sample mean then the x-axis would have to be labelled something like $\bar{X}$ or "X-bar (mean age in years)." In contrast, the STATA summary is about a sampling distribution and the top label is clear: "X-bar."]