Stat 322 (F17) - Assignment #3

(Due ~~Wed. Nov. 15th at 4:00~~ Fri. Nov. 17th at 12:00 pm (noon) in appropriate STAT 322 slot in assignment box #15 outside the Math Tutorial Centre (MC 4066/4067). Electronic submissions or in-class submissions will not be accepted.

SOLUTIONS (/45)

1) Read the Waterloo News (March 5/2014) article *Energy drinks linked to teen health risks* posted in the articles on LEARN.

   a) Based on the article, define the following:

      i) **[2]** The target population.

      The target population is the population being referred to when drawing conclusions. Based on the article, the target population is (Canadian) 'teens' or 'teenagers' (could also be interpreted as (Canadian) high school students).

      ii) **[2]** The study population (i.e. the sampling frame).

      The study population is the population from which the sample was selected. Based on the last line of the article, the study population here is "junior and senior high school students from three provinces in Atlantic Canada."

      iii) **[2]** The sample.

      The 8210 high school students surveyed.

   b) **[4]** Give two population attributes, and the values of their estimates obtained from the sample, referred to in the article.

   The only two attributes and their estimates referred to in the article are:

   - π_1 = proportion of population that consumed energy drinks at least once in the past year ($\hat{\pi}_1 \approx .65$ ("nearly two thirds"))
   - π_{21} = proportion of population that consume energy drinks at least once per month ($\hat{\pi}_2 \approx .20$ ("more than 20 percent"))

c) Describe each of the following errors in the context of the study, and for each error, discuss whether you feel the error would be negligible or considerable.

i) **[3]** Frame error (or Study error)

This is the difference in attributes between the target population and the study population – in this case, the difference in consumption rates (proportions) between all (Canadian) teenagers (or high school students) and junior and senior high school students from three provinces in Atlantic Canada.

This error might be significant if, for example, energy drink consumption rates for teens in the three Atlantic provinces were higher/lower than for teens from the rest of Canada.

ii) **[3]** Sample error

This is the difference in attributes between the study population and the sample – in this case, the difference in consumption rates (proportions) between teens (or high school students) from three provinces in Atlantic Canada and that of the 8210 students surveyed.

This error might be significant if, for example, consumption rates for students who were absent from school (or not registered in school) at the time of the survey, or who would choose not to respond to the survey if selected, were higher/lower than consumption rates of all high school students from the three Atlantic Provinces.

iii) **[3]** Measurement error

In this study, measurement error is the difference in the consumption rates (proportions) based on the students' responses and the actual consumption rates of the students surveyed.

This error might be significant if, for example, students tended to underrepresent (or overrepresent) their energy drink consumption, so that the sample proportions based on the reported values would be lower/higher than the sample proportions based on the actual values.

2) Open the associated journal article through the 'published in Preventative Medicine' link in the news article (also posted on LEARN), and read about the methodology in the Methods section (primarily, the *Participants* section) to answer the following questions:
(Note: for consistency, use the sample size provided in the Waterloo News article for calculation purposes)

a) **[2]** Update your definition of the study population from question 1).

The study population is more specifically defined in the journal article: students "in grades 7, 9, 10 and 12 attending public schools in the three Atlantic Provinces Nova Scotia, New Brunswick, and Newfoundland & Labrador…"

b) **[5]** Based on the description of the sample design, clearly describe how stratified sampling, cluster sampling, and simple random sampling were used in this survey. Be sure to clearly define the strata and the clusters.

According to the journal article:

"The sample design was a **two stage stratified cluster sample** of **randomly selected classes containing at least 20 students in each of the four surveyed grades within each health region** in the three participating provinces."

Exactly how stratified and cluster sampling was employed is a bit unclear, but this is one reasonable protocol inferred from the article:

- Stratified sampling: The three Atlantic provinces were separated into health regions. Each health region would then constitute a stratum.

  An additional stratification level was likely grades – with each of grades 7, 9, 10, 12 being the strata.

- Cluster sampling: Classes containing at least 20 students constituted the clusters.

- Simple random sampling: The clusters (classes) were randomly selected using simple random sampling.

c) **[3]** According to the webpage, https://www.populationpyramid.net/canada/2012/, approx. 5% of Canadians are of the age from which the sample was taken. If we assume that this proportion is consistent with the provinces involved in this survey, calculate the finite population correction factor associated with the standard deviation of estimates associated with this survey (note that the total population of the relevant provinces is given).

N = size of the study population = number of students in grades 7, 9, 10 and 12 in Nova Scotia, New Brunswick, and Newfoundland & Labrador. Based on the article, the **total** population of these three provinces = 2,187,434. Assuming 5% of this population is from the age demographic associated with study population, then

$$N = 2187434(.05) = 109372$$

$$fpc = 1 - \frac{n}{N} = 1 - \frac{8210}{109372} = 0.925$$

d) Based on your results in 2c) and on the information in the Results section, calculate an approx. 95% confidence interval for:

i) **[3]** The proportion of Canadian teens who used energy drinks at least once in the past year.

$$\hat{\pi} \pm 1.96\sqrt{fpc}SE(\hat{\pi})$$

$$\hat{\pi} \pm 1.96\sqrt{fpc}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$$.622 \pm 1.96\sqrt{.925}\sqrt{\frac{.622(.378)}{8210}}$$

$$.622 \pm .010$$

$$= (.612,.632)$$

ii) **[3]** The proportion of Canadian teens who used energy drinks at least once a month in the past year.

$$195 \pm 1.96\sqrt{.925}\sqrt{\frac{.195(.805)}{8210}}$$

$$.195 \pm .008$$

$$= (.187,.203)$$

iii) **[3]** The average age of Canadians from the age demographic in the frame (Grades 7, 9, 10, 12).

$$\hat{\mu} \pm 1.96\sqrt{fpc}SE(\hat{\mu})$$

$$\hat{\mu} \pm 1.96\sqrt{fpc}\frac{s}{\sqrt{n}}$$

$$15.2 \pm 1.96\sqrt{.925}\frac{0.06*}{\sqrt{8210}}$$

$$15.2 \pm .001$$

$$= (15.201,15.199)$$

*note that the value of sample standard deviation, s, given in the article, 0.06 years or approx. 3 weeks, is far too low, and is likely a typo. It is likely 0.6 years, or approx. 7 months, which is much more realistic. These types of statistical errors and typos are common in journal articles.

If we consider the reported value of 0.06 to be the standard error of the sample mean, $SE(\hat{\mu}) = \frac{s}{\sqrt{n}} = 0.06$, then the resulting confidence interval would be:

$$15.2 \pm 1.96\sqrt{.925}(.06)$$

$$15.2 \pm 0.11$$

$$= (15.31,15.19)$$

e) **[3]** Which of the intervals calculated for <u>the proportions</u> in d) is wider? Why?

The interval for the proportion in d) i) is wider, since the standard error of the estimator, $SE(\hat{\pi})=\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, is larger for values of $\hat{\pi}$ closer to 0.5.

f) **[4]** For the proportion in d i), calculate the sample size required for the results to be 'accurate to within 1 percentage point, 99 times out of 100'.

'99 times out of 100' $\Rightarrow z = 2.575$

$$1.96\sqrt{1-\frac{n}{N}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = .01$$

$$\Rightarrow n = \frac{1}{\frac{1}{N}+\frac{.01^2}{2.575^2\,\hat{\pi}(1-\hat{\pi})}}$$

$$= \frac{1}{\frac{1}{109372}+\frac{.01^2}{2.575^2(.622)(.378)}}$$

$$= 13645$$

You would need a sample size of approx. 13645 for the estimate of this proportion to be accurate to within .01, with 99% confidence.