Stat 332 – Assignment 1
Prof. Samuel Wong – Winter 2019
Due: Friday, January 25 at 10:30am on Crowdmark

General instructions: You may submit your work using one or more of the following ways:

- Type out work, for example using 'Latex', 'R Markdown', or Word.
- Present scans/photographs of handwritten work. If you choose this option ensure your work is legible. Illegible work will receive no credit.

For data analysis problems: When you are using R (which we strongly encourage), you must clearly present your final answers in addition to the commands you used.

1. A researcher wishes to study the prevalence of drug use among homeless people in the US. From a list of Health Care for the Homeless clinics across the country, she randomly selects 25 clinics. She spends a day at each of the 25 clinics and asks questions to all the patients that visit the clinic that day.

   Identify the target population, sampled population, sampling frame, sampling unit, and observation unit. Briefly describe a possible source of selection bias.

2. A headline claims that "45% of computer users encountered a significant malfunction in the last year". Reading the article, you find that the data were obtained from subscribers to a tech magazine who were asked to fill out an online survey. The survey asked respondents to answer YES/NO for 123 types of computer problems that they may have encountered in the last year. Survey respondents were also entered into a draw to win a new Macbook Pro.

   (a) What was the target population and frame for this survey?

   (b) Think about potential sources of error and bias resulting from (i) how the frame was constructed; (ii) how the sample was obtained; (iii) how the responses were measured. Provide one answer for each of (i), (ii), and (iii).

3. A survey is conducted using the following sampling design: for each unit $1, ..., N$ in the population, we flip a fair coin (i.e., with 0.5 probability of heads) to decide whether this person will be included in the survey or not. Let S be the random variable denoting the set of persons that are included in the survey and y_i the response variable recorded for each person. *You must show your work for all parts of this problem.*

   (a) Show that, under this design, $\tilde{\tau} = 2\sum_{i \in S} y_i$ is an unbiased estimator for the population total τ.

   (b) Find the variance of $\tilde{\tau}$ under this design.

   (c) Find an unbiased estimator for $\text{Var}(\tilde{\tau})$ under this design and show that your estimator is unbiased.

4. *Analyzing a SRS.* Download the dataset `bookssn.csv` from Learn. It contains a SRS of 91 Florida residents from a Pew Research Survey. The two variables listed in the file are:

   - *books*: The number of books read by the respondent in the past year.
   - *sn*: Whether the respondent accesses Facebook and/or Twitter daily (1 = yes).

   (a) Estimate the average number of books read by a Florida resident in the past year. Report the estimate, SE of the estimate, and a 95% CI.

(b) Estimate the proportion of Florida residents who use Facebook/Twitter daily. Report the estimate, SE of the estimate, and a 95% CI.

(c) You feel these estimates are not precise enough (the 95% CIs are quite wide, after all). How many respondents would you need, to ensure that the total width of the 95% CI for the population mean of *books* will be less than 0.1? Show your work.