# Assignment #1

## Stat 332: Sampling and Experimental Design (Spring 2024)

Professor: Christian Boudreau — Due at 11:59 pm on Sunday Jun 2, 2024

Instructions:

- Teamwork is allowed and encouraged; however, everyone must hand in their own assignment showing that they understood what they wrote. You are not allowed to collaborate with anyone who is not a STAT 332 classmate.
- Justify all your answers.
- One solution per question; if you provide multiple solutions, you will receive 1/2 the points of the inferior solution.
- Write legibly; if the TA or I have to guess at your answers, it's unlikely to be in your favour.
- Make sure you include your `R` code and relevant output for Q#4.
- Crowdmark:
  - You should soon received an automated email from Crowdmark saying that you can now submit/upload your solution. If you have not received the email by noon on Monday, please start by checking your spam or junk mail folder. If the email is not there, email me.
  - To ensure smooth and efficient grading, please make sure to upload your solution separately for each question.
  - Crowdmark allows you to resubmit anytime before the due date; once the due date has passed, your submission is locked in and you will thus not be able to resubmit.
  - Late submissions will be accepted, but are subject to a 5% per hour penalty*; click here for more information on Crowdmark and late submissions.
  - Technical difficulties with Crowdmark:
    - Consult Crowdmark Help
    - Watch this short video about submitting an assignment on Crowdmark
    - E-mail me
    - If tragedy strikes and you cannot upload your assignment to Crowdmark, email it to me, so I have proof of when you completed your assignment (otherwise, the above mentioned lateness penalty will apply).
- Recall that each of you is allowed a one time 1 or 2-day extension with a reduced lateness penalty. That reduced penalty is 5% for an additional 24 hours and 10% for an additional 48 hours. Please know that you have to email me (at least) 12 hours before the original due date to take advantage of this reduced penalty.
- See course outline for more information on assignments, and how much each of them contributes to your overall grade.

*This is the default penalty, and is applied when there are no extenuating circumstances.

**Question #1** (10 points)**:**

Recall that some sampling designs $p(S)$ have a fixed sample size $n$ (e.g., SRS), whereas others have a random sample size $n_S$ (e.g., Bernoulli sampling). Prove that

$$V(\hat{t}_\pi) = -\frac{1}{2}\sum_{k\in U}\sum_{\ell\in U}(\pi_{k,\ell} - \pi_k\pi_\ell)\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2$$

is equivalent to

$$V(\hat{t}_\pi) = \sum_{k\in U}\sum_{\ell\in U}(\pi_{k,\ell} - \pi_k\pi_\ell)\,\frac{y_k}{\pi_k}\,\frac{y_\ell}{\pi_\ell}$$

for fixed size sampling designs. Note that you are <u>not</u> allowed to assume any given sampling design (this includes SRS).

**Question #2** (10 points)**:**

This question is concerned with political polls. Those polls usually consists of 1000 adults, which are traditionally selected using random digit dialing (RDD) or stratified RDD, and interviewed by phone. Since RDD is very similar to SRS, let's simply assume for this question that those 1000 individuals are selected using SRS.

(a) According to Statistics Canada, at the time of the last federal elections in 2015, there were 28,422,880 eligible voters in Canada. Since you didn't get called by any survey firm during the last federal election, you're now wondering what were your odds of being called, and ask yourself the following questions:

   i) What is my probability of being selected to take part in a given such political poll of 1000 adults?

   ii) Now, 181 political polls were conducted during the last federal election in 2015. Assuming that selection is done independently from one poll to another, what was the probability that I would *not* get selected in any of the 181 polls? In other words, what was the probability that I would *not* get called by any of these 181 polls?

   iii) How many such pools must be conducted so I have a probability of at least 25% of being in one or more of them?

(b) After investigating your odds of being called to take part in a political poll, you started to wonder why those polls usually consists of 1000 adults/eligible voters. Show that such a sample size will guarantee, to the survey firm, that they will be able estimate the percentage of indviduals that will vote for any given candidate with a minimum precision of $\pm 3.1\%$, when using a 95% CI.

(c) On October 16, 2015, Nanos Research conducted a poll for the upcoming federal election. They found that 37.3% of the 2000 indviduals they selected via SRS and then interviewed by phone were planning to vote for the Liberal party. Provide a 95% CI for that point estimate.

**Question #3** (5 points):

Post 9 on our Piazza discussion forum asked about why the population variance was defined as

$$S_U^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y}_U)^2$$

instead of

$$\tilde{S}_U^2 = \frac{1}{N}\sum_{k\in U}(y_k - \bar{y}_U)^2$$

I answered that, amongst other things, this was done to ensure that $S_S^2$ is an unbiased estimator of $S_U^2$ under SRS.

Now, assume an SRS sampling design and that the population variance is defined as $\tilde{S}_U^2$. Consequently,

$$V_{\text{SRS}}(\hat{t}_\pi) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\tilde{S}_U^2$$

and $S_S^2$ is now a biased estimator of $\tilde{S}_U^2$. Compute the bias of $S_S^2$?

To be clear

$$S_S^2 = \frac{1}{n-1}\sum_{k\in 2}(y_k - \bar{y}_S)^2$$

Note: you can use what was proven in class by simply stating it. In other words, though you can, you do not have to compute $E(S_S^2)$ from scratch.

**Question #4** (10 points):

An instructor at the University of Florida conducted a small survey of the students in his introductory to statistics class. To this end, he selected a simple random sample of 57 students from the 350 students in his class. File `classurv.txt`, which contains the dataset, is available on the STAT 332 LEARN website. The file contains 11 variables, but this question is concerned with the 3 4 variables given in the table below.

| Variable | Description |
|---|---|
| `gender` | gender (1 = male, 2 = female) |
| ~~`age`~~ | ~~age (-9 = missing)~~ |
| `GPA` | GPA (-9 = missing) |
| `year` | year of study (1 = freshman, 2 = sophomore, 3 = junior, 4 = senior, 5 = other) |

Using the `survey package`, give a point estimate and corresponding 95% CI for each of the following:

(a) Estimate the average GPA of all students.

(b) Estimate the average GPA of male students

(c) Estimate the average GPA of female students.

(d) Estimate the proportions of male/female students.

(e) Estimate the proportions of students that are freshman/sophomore/junior/senior/other.

(f) Amongst freshmen, what are the proportions of male/female students.

(g) Repeat (f) but for seniors.

(h) Amongst the top performers (i.e., GPA $\geq 3.2$), what are the proportions of male/female students.

As mentioned in class, make sure to read the documentation of the `survey package`; in particular, the `svyby` function of the package can help answering parts (b) and (c).