1. A student wants to know the class average on a midterm exam she recently wrote. At the 8:30am lecture following the midterm, she is a little early and asks 10 of the people in the room what they got on the test. She calculates her estimate of the class average on what these 10 students say they scored on the test.

   (a) [**3 marks**] What sampling protocol did the person use? Explain.
   Solution:
   Convenience sampling because she collected the responses from students she were able to reach easily. Some may argue that the sampling protocol is judgment sampling since the student would have considered the students who arrive earlier are representative of the target population. We will accept both answers although the earlier argument is more convincing.

   (b) [**3 marks**] Will her sample be representative of the population as a whole? Explain.
   Solution:
   No, because students who arrived earlier may be more diligent with attending morning classes, or may be more consistent with their class attendance. We will accept other valid argument that the sample is not representative of the population.

2. [**2 marks**] Suppose the National Rifle Association in the U.S. sends a survey to 1000 of its members, asking "Do you believe that all illegal immigrants should be deported back to their country of origin?". 912 members return the survey, 59% responding "yes". This survey is then used as evidence that 59% of Americans believe that illegal immigrants should be deported. *These survey results given in this question are fictitious.*

   Which of the following statements are true? (Circle all that apply.)

   (a) Since only 912 out of 1000 people returned the survey, the survey gives absolutely no evidence about the feelings of NRA members toward illegal immigrants.

   (b) The biggest problem with this survey is that it is a mail-in survey, which can cause a delay of several weeks.

   (c) The sample size is far too small to be meaningful.

   (d) **The sample is not representative of the population of interest.**

3. Seat belt marks (bruising) on the body are sometimes used as evidence that a person was wearing a seat belt during a car crash. A study investigated seat belt marks in victims of fatal car crashes in Sydney Australia. In 74 fatalities (obtained through simple random sampling) in which the victim was wearing a seat belt, 27 victims showed seat belt marks.

(a) [**3 marks**] What is the estimated proportion of seatbelt-wearing victims of fatal car accidents that show seat belt marks?
Solution:
Let

$$y_i = \begin{cases} 1, & \text{if i-th victim showed seat belt marks,} \\ 0, & \text{otherwise.} \end{cases}$$

Let π denotes the proportion of seatbelt-wearing victims of fatal car accidents that show seat belt marks. The estimate,

$$\begin{aligned} \hat{\pi} &= \frac{1}{74}\sum_{i=1}^{74} y_i \\ &= \frac{27}{74} \\ &= 0.3649. \end{aligned}$$

(b) [**5 marks**] Find a 95% confidence interval for the proportion of seatbelt-wearing victims of fatal car accidents that show seat belt marks.
Solution:
From class, $Var(\tilde{\pi}) = \frac{\sigma^2}{n}$ as we ignored the finite population corrections. Hence $\hat{Var}(\tilde{\pi}) = \frac{\hat{\sigma}^2}{n}$, where $\hat{\sigma}^2$ can be simplified as $\frac{n}{n-1}\hat{\pi}(1-\hat{\pi})$. Hence,the standard error is given by

$$SE(\pi) = \sqrt{\frac{0.3649 \times (1 - 0.3649)}{74 - 1}} = 0.0563$$

and the 95% confidence interval for π is

$$0.3649 \pm 1.96 \times 0.0563.$$

(c) [**5 marks**] Ignoring the finite population corrections and using $\hat{\sigma}^2$ from this study, what sample size would be necessary to estimate the proportion with 95% confidence and margin of error 0.001?
Solution:
By ignoring the finite population corrections, we will solve for the sample size using

$$MOE = z_{\alpha/2} \times \sqrt{\frac{1}{n-1}\pi(1-\pi)}.$$

After some algebraic work, $n = 1 + z_{\alpha/2}^2 \times \frac{\pi(1-\pi)}{MOE^2}$. Using $\hat{\pi} = 0.3649$ from earlier, $n = 1 + 1.96^2 \times \frac{0.3649(1-0.3649)}{0.001^2} = 890284.1$. The sample size required is 890,285. *If used* $\hat{\pi} = \frac{27}{74}$ *instead of 0.6349, then the answer should be* $890,247.6 \approx 890,248$.

4. Suppose that we divide the population U into H mutually exclusive strata, $U_1, \ldots, U_H$ with sizes $N_1, \ldots, N_H$ such that $N = N_1 + \ldots N_H$. Define

- h-stratum mean: $\mu_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$
- population mean: $\mu = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}$

(a) [**3 marks**] Show that

$$\mu = \sum_{h=1}^{H} W_h \mu_h, \quad \text{where } W_h = \frac{N_h}{N}.$$

Solution:

$$
\begin{aligned}
\mu &= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} \\
&= \sum_{h=1}^{H}\frac{1}{N}\sum_{j=1}^{N_h} y_{hj} \\
&= \sum_{h=1}^{H}\frac{1}{N}\times N_h \times \underbrace{\frac{1}{N_h}\sum_{j=1}^{N_h} y_{hj}}_{\mu_h} \\
&= \sum_{h=1}^{H}\frac{N_h}{N}\mu_h \\
&= \sum_{h=1}^{H} W_h\mu_h \quad \text{, where } W_h = \frac{N_h}{N} \text{ is the stratum weight.}
\end{aligned}
$$

(b) **[3 marks]** Show that $\tilde{\mu}_{st} = \sum_{h=1}^{H} W_h\tilde{\mu}_h$. is an unbiased estimator of μ. *You may use results from SRS.*
Solution:

$$
\begin{aligned}
E(\tilde{\mu}_{st}) &= E\left(\sum_{h=1}^{H} W_h\tilde{\mu}_h\right) \\
&= \sum_{h=1}^{H} W_h E(\tilde{\mu}_h) \\
&= \sum_{h=1}^{H} W_h\mu_h, \quad \text{since each stratum is a SRS} \\
&= \mu.
\end{aligned}
$$