SM-4331 Exercise 2

1. In a population of size N, let μ_c be the mean of the N_c elements of the population included in the the frame population, and $\mu_{\neg c}$ be the mean of the $N_{\neg c}$ elements of the population not included in the frame population. Note that $N = N_c + N_{\neg c}$, and the population mean is given by

$$\mu = \frac{N_c}{N}\mu_c + \frac{N_{\neg c}}{N}\mu_{\neg c}$$

Let $\mathbf{y}_c = \{y_1, \ldots, y_n\}$ be a sample taken from the N_c elements in the frame population $\mathbf{Y}_c = \{y_1, \ldots, y_{N_c}\}$. Using this sample, the sample mean is

$$\bar{y}_c = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad y_i \in \mathbf{y}_c$$

(a) Find $\mathrm{E}(\bar{y}_c)$.

Solution: Since each y_i has expectation μ_c (they are obtained from the frame population), then

$$\begin{aligned}\mathrm{E}(\bar{y}_c) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(y_i) = \mu_c.\end{aligned}$$

(b) Show that if we use $\bar{y}_c$ as an estimator for the population mean μ, the bias due to undercoverage is

$$\mu_c - \mu = \frac{N_{\neg c}}{N}(\mu_c - \mu_{\neg c}).$$

Solution:

$$\begin{aligned}\mathrm{E}(\bar{y}_c - \mu) &= \mu_c - \mu \\ &= \mu_c - \frac{N_c}{N}\mu_c - \frac{N_{\neg c}}{N}\mu_{\neg c} \\ &= \frac{N_{\neg c}}{N}\mu_c - \frac{N_{\neg c}}{N}\mu_{\neg c} \\ &= \frac{N_{\neg c}}{N}(\mu_c - \mu_{\neg c})\end{aligned}$$

(c) Explain what happens to the bias when

i. the undercoverage is small, i.e. $N_{\neg c}/N$ is small.

Solution: If undercoverage is small, i.e. $N_{\neg c} \to 0$, then the bias tends to zero as well, and we get an unbiased estimator for the mean.

ii. the covererd and not-covered populations are similar, i.e. $\mu_c \approx \mu_{\neg c}$.

Solution: If $\mu_c \approx \mu_{\neg c}$, then $\mu_c - \mu_{\neg c} \approx 0$, so the estimator is unbiased.

2. Consider deploying a survey by landline telephone at home, in which the target population are adults aged 18 or over in the country.

   (a) Who would be included in the survey population? Comment on the coverage error of the target vs. survey population in different counties (e.g. developed vs developing world) and in different time periods (e.g. what is the non-coverage as a country becomes more and more developed?).

   Solution: The survey population are those adults aged 18 or over who has a registered telephone landline at their residence. Generally speaking, as a country transitions from an underdeveloped state to a developed state, the proportion of population who has access to landlines will increase. However, there comes a critical point in time where they transition again to mobile phones, so may not require to have a landline anymore.

   (b) Discuss how you would construct a sampling frame for such a telephone survey. What are the potential issues with such sampling frames?

   Solution: We would need access to a list of the entire population and their telephone numbers. We could obtain this from the phone book. Issues that could arise include:

   - Phone book is not kept up-to-date (undercoverage).
   - Duplicates (some residences might have two landlines).
   - Included non-residential addresses e.g. commercial businesses.

   (c) Repeat your answers to parts (a) and (b) above in the case of on-line surveys.

   Solution: Survey population would be individuals who have access to the internet. This depends on your target population. For example, if wanted to do a survey for all members of an organisation, it is likely that the organisation has a list of everyone's e-mail addresses, which can be used as a reliable sampling frame. However, if the population is the entire country, such a list does not exist. The issues that could arise are 1) duplication of e-mails; and 2) one e-mail for multiple individuals. Even so, coverage error varies by country (developed vs developing) and by time. Further, internet coverage

may also be related to certain characteristics (age, education, socioeconomic status, attitudes, etc.)

3. Suppose that we are interested in the mean μ of a variable y in a population of size $N = 1000$ with four (mutually exclusive and completely exhaustive) groups, A, B, C & D. Suppose further that in this population, the sizes and means of the groups are as follows:

	A	B	C	D
Size N_h	600	50	200	150
Mean μ_h	40	10	90	60

(a) Calculate the overall population mean μ.

Solution: $\mu = (600\times 40+50\times 10+200\times 90+150\times 60)/1000 = 51500/1000 = 51.5$.

(b) Suppose that we use stratified sampling with 50 elements sampled from each group. Determine the inclusion probabilities for each group.

Solution: The inclusions probabilities for each group A, B, C, and D would be $50/600 = 0.08$, $50/50 = 1$, $50/200 = 0.25$, and $50/150 = 0.33$, respectively.

(c) Suppose now that the stratified sample of 50 elements in each group happens to give the following sample means: $\bar{y}_A = 40$, $\bar{y}_B = 10$, $\bar{y}_C = 90$, $\bar{y}_D = 60$. Calculate the overall sample mean $\bar{y}$. Is it unbiased?

Solution:

$$\bar{y} = \frac{50 \times 40 + 50 \times 10 + 50 \times 90 + 50 \times 60}{50 \times 4} = 50$$

It is not unbiased.

(d) Define the **design weights** (a.k.a. sampling weights or base weights) to be the inverse of the inclusion probabilities, i.e.

$$w_h = 1/\,\mathrm{P}(\text{inclusion in group } h).$$

Define also the (sample) **weighted mean** to be

$$\bar{y}_w = \frac{\sum_{h\in\{A,B,C,D\}} w_h \bar{y}_h}{\sum_{h\in\{A,B,C,D\}} w_h}.$$

Calculate the sample weighted mean and show that it is unbiased.

Solution:

$$\bar{y} = \frac{(50/600) \times 40 + (50/50) \times 10 + (50/200) \times 90 + (50/150) \times 60}{50/600 + 50/50 + 50/200 + 50/150} = 51.5$$

It is unbiased since $\mu = 51.5$.

(e) Discuss briefly the logic of design weights, and the role that it plays to produce an unbiased estimator of the population mean.

Solution: Design weights will downweight the values obtain from a strata with high inclusion probabilities, so that the final estimator will better reflect and represent the entire population.

4. Consider the values of the data y concerning an artificial population shown below, which is stratified into 3 strata and clustered into 5 clusters.

	Cluster A	Cluter B	Cluster C	Cluster D	Cluster E
Stratum 1	10	9	8	8	10
Stratum 2	3	4	3	3	4
Stratum 3	7	7	6	6	6

Three kinds of sampling techniques were performed, with the following samples (each $n = 6$) realised:

- SRS: $\{10, 9, 8, 4, 4, 6\}$.
- Cluster sample: All data from Clusters A and E.
- Stratified sample: $\{10, 8\}$ from Stratum 1, $\{4, 3\}$ from Stratum 2, and $\{7, 6\}$ from Stratum 3.

Compare and contrast the three kinds of estimators for the population mean, and the variance of the respective estimators. Which sampling technique yields the least sampling error?

Solution: Firstly, we can calculate the true mean and variances of this population:

$$\mu = 6.20 \qquad S^2 = 6.17$$
$$(\mu_A, \mu_B, \mu_C, \mu_D, \mu_E)^\top = (6.7, 6.7, 5.7, 5.7, 6.3)^\top$$
$$(S_A^2, S_B^2, S_C^2, S_D^2, S_E^2)^\top = (12.3, 6.3, 6.3, 6.3, 10.3)^\top$$
$$(\mu_1, \mu_2, \mu_3)^\top = (9.0, 3.4, 6.2)^\top$$
$$(S_1^2, S_2^2, S_3^2)^\top = (1.0, 0.3, 0.7)^\top$$
$$S_{cl}^2 = 0.26$$

The sample mean from each kind of sampling teachnique is as follows:

- SRS: $\bar{y} = \frac{10+9+8+4+4+6}{6} = 6.83$.
- Cluster: $\bar{y}_{cl} = \frac{10+3+7+10+4+6}{6} = 6.50$.
- Stratified: $\bar{y}_{st} = \frac{10+8+4+3+7+6}{6} = 6.33$

and each estimator has variance as follows:

- SRS: $\text{Var}(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{15-6}{15\times 6} \times 6.17 = 0.62$.
- Cluster: $\text{Var}(\bar{y}_{cl}) = \frac{M-m}{Mm} S_{cl}^2 = \frac{5-2}{5\times 2} \times 0.26 = 0.08$.
- Stratified: $\text{Var}(\bar{y}_{st}) = \sum_{h=1}^{3} \frac{N_h - n_h}{N_h n_h} S_h^2 = \frac{5-2}{5\times 2}(1.0 + 0.3 + 0.7) = 0.60$

Stratified sampling yields the least error estimate $|6.33 - 6.20| = 0.13$. The largest sampling error is SRS, since the variance of the estimate is 0.62 (although stratified sampling yields a variance of 0.60, not too far off). Interestingly the variance from cluster sampling is the lowest (0.08), which indicates that a "high confidence" in the estimate of 6.50. We can see why the variance is low: it is a function of the variance of the cluster means, and since the clusters are homogenous, there is not much variation in there.

5. The Fish and Game department of a particular state was concerned about the direction of its future hunting programs. To provide for a greater potential for future hunting, the department wanted to determine the proportion of hunters seeking any type of game bird. A simple random sample of $n = 1000$ of the $N = 99000$ licensed hunters was obtained. Suppose 430 indicated they hunted game birds.

   (a) Estimate p, the proportion of licensed hunters seeking game birds, giving a 95% confidence interval for p.

Solution: $\hat{p} = 430/1000 = 0.43$. This estimator has variance

$$\text{Var}(\hat{p}) = \frac{N-n}{Nn} S^2 = \frac{N-1}{N-1} \cdot \frac{N-n}{Nn} S^2 = \frac{N-n}{(N-1)n} \sigma^2$$

where $\sigma^2 = p(1-p)$, which can be estimated by

$$\widehat{\text{Var}}(\hat{p}) = \frac{N-n}{(N-1)n} \hat{p}(1-\hat{p}) = 0.9899 \times \frac{0.2451}{1000} = 2.43 \times 10^{-4}$$

Assuming asymptotic normality, $\hat{p} \approx \text{N}\left(p, \widehat{\text{Var}}(\hat{p})\right)$. Thus, the 95% interval is

$$\hat{p} \pm 1.96 \times \sqrt{2.43 \times 10^{-4}} = (0.399, 0.461)$$

(b) Determine the sample size the department must obtain to estimate the proportion of game bird hunters, such that the interval width calculated in (a) does not exceed 0.02.

Solution: From the lecture slides we get

$$n > \frac{4z^2S^2/B^2}{1 + 4z^2S^2/(NB^2)}$$

Substitute in values $z = 1.96$, $S^2 = \hat{p}(1-\hat{p})N/(N-1) = 0.2451$, and $N = 99000$ to get

$$\begin{aligned} n &> \frac{4z^2S^2/B^2}{1 + 4z^2S^2/(NB^2)} \\ &= \frac{9415.7616}{1 + 0.0951} \\ &= 8598 \end{aligned}$$

(c) State any assumptions you used in your calculations in (a) and (b).

Solution: Finite population N. Note that if we did not use the population correction factor, we would still get similar results. Except here the assumption would be that there is a potentially infinite population (or N unknown), or we can say that $1 - n/N \approx 1$.

6. The European Social Survey partly aims to measure the trust in police effectiveness. In order to measure this concept, several questions relating to this topic was asked in a survey, among them this question: "How successful do you think the police are at preventing crimes where violence is used or threatened?". Respondents answered on a 5-point Likert scale (from 1 = Not at all successful to 5 = Completely successful). From the $n = 2{,}350$ respondents from the UK selected using simple random sampling, the following statistics were obtained:

$$\sum_{i=1}^{n} y_i = 7{,}815 \qquad \sum_{i=1}^{n} y_i^2 = 29{,}655$$

(a) Calculate the sample variance s^2.

Solution:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$
$$= \frac{1}{2349}\left(29655 - 2350\left(\frac{7815}{2350}\right)^2\right)$$
$$= 1.56$$

(b) The population of the UK is roughly 66 million people when this survey was conducted. Calculate the sample mean and the variance of this estimator.

Solution: The sample mean is $\bar{y} = 7815/2350 = 3.33$. It has variance

$$\mathrm{Var}(\bar{y}) = \frac{N-n}{Nn}S^2$$

which is estimated by

$$\widehat{\mathrm{Var}}(\bar{y}) = (1 - n/N)s^2/n = 0.999 \times 1.56/2350 = 6.64 \times 10^{-4}$$

(c) The European Social Survey aimed to obtain an estimator whose variance does not exceed $B = 0.0005$. Did they achieve this target, and if not, what sample size would have helped achieve this target?

Solution: They did not achieve this target, since the estimated variance is $0.00064 > 0.0005$. In order to achieve this target, they would need to obtain the following sample size:

$$n > \frac{s^2/B}{1 + s^2/(NB)}$$
$$= \frac{3120}{1 + 4.727 \times 10^{-5}}$$
$$= 3119.8$$

so the minimum sample size should be 3120.