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

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

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

   (c) Explain what happens to the bias when

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

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

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?).

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

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

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.

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

(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?

(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.

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

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: $\{10, 8, 3, 4, 4, 6, 5\}$.
- 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?

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.

(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.

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

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.

(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.

(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?