Survey Sampling
Statistics 4234/5234 — Fall 2018

Homework 1

Solutions:

1. For each of the following surveys: describe the target population, sampling frame, sampling unit, and observation unit; and discuss any possible sources of selection bias or inaccuracy of responses.

   (a) *A student wants to estimate the percentage of mutual funds whose shares went up in price last week. She selects every tenth fund listed in the Mutual Fund pages, and calculates the percentage of those in which the share price increased.*

   Target population is *all* mutual funds, sampling frame is only those listed in the Mutual Fund pages, so undercoverage is an issue. Specifically, the smaller funds are more likely to be excluded from those listings, and if smaller funds are systematically different from the larger ones (which seems plausible), then this is a source of bias.

   Sampling unit and observation unit are mutual fund.

   Sampling every tenth fund in the listing — a systematic sample — is not the best idea (I assume the funds are listed alphabetically?), but I don't see it introducing any huge bias.

   (b) *A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.*

   Target population is all cows in the region, the sampling frame is the list of farms. As long as the listing is complete (there aren't a whole bunch of small farms excluded from the list), then undercoverage should not be an issue.

   The sampling unit is the farm and the observation unit is the cow — this is an example of cluster sampling. While not the most statistically efficient approach, cluster sampling is an unbiased sampling method, as every cow in the region has the same change of being included in the sample.

   (c) *To study nutrient content of menus in boarding homes for the elderly in Washington State, researchers mailed surveys to all 184 licensed homes in Washington State, directed to the administrator and food service namager. Of those, 43 were returned by the deadline and included menus.*

   Target population is the 184 licensed boarding homes for the elderly in Washington State, sampling frame is the same. So no undercoverage.

   Sampling unit and observation unit are the same, boarding home.

   Voluntary response sampling is a biased sampling method, an approach that almost guarantees the sample will be unrepresentative of the population. In this case, it seems reasonable to think that the respondents to the survey are the better run (and better funded) boarding homes, with (i) the resources to allow them to take the time to complete the survey, and (ii) a belief in the value of this research, and hence in the importance of nutrition.

   Another source of bias is measurement bias, as respondents are likely to exaggerate the nutritional value of their menus.

2. Let the discrete random vector (X, Y) have a joint probability mass function $p(x, y)$ given by the following table:

		y			
		1	2	3	4
x	1	0.12	0.21	0.24	0.03
	2	0.06	0.06	0.12	0.06
	3	0.02	0.03	0.04	0.01

(a) The marginal distributions of X and Y are given, respectively, by

x	1	2	3
$p_1(x)$	0.60	0.30	0.10

and

y	1	2	3	4
$p_2(y)$	0.20	0.30	0.40	0.10

(b) $P(X = 1, Y = 2) = 0.21 \neq (0.60)(0.30) = P(X = 1)P(Y = 2)$ so X and Y are not independent.

(c)

$$E(X) = \sum xP(X = x) = 1(0.60) + 2(0.30) + 3(0.10) = 1.5$$

$$E(Y) = \sum yP(Y = y) = 1(0.20) + 2(0.30) + 3(0.40) + 4(0.10) = 2.4$$

$$E(3X - 2Y) = 3E(X) - 2E(Y) = 3(1.5) - 2(2.4) = -0.3$$

Also

$$E(X^2) = \sum x^2P(X = x) = 1(.60) + 4(.30) + 9(.10) = 2.7$$

$$E(Y^2) = \sum y^2P(Y = y) = 1(.20) + 4(.30) + 9(.40) + 16(.10) = 6.6$$

so

$$V(X) = E(X^2) - (EX)^2 = 2.7 - 1.5^2 = 0.45$$

$$V(Y) = E(Y^2) - (EY)^2 = 6.6 - 2.4^2 = 0.84$$

(d)

$$\begin{aligned} E(XY) &= \sum\sum xyP(X = x, Y = y) \\ &= 1(0.12) + 2(0.21) + 3(0.24) + 4(0.03) \\ &\quad + 2(0.06) + 4(0.06) + 6(0.12) + 8(0.06) \\ &\quad + 3(0.02) + 6(0.03) + 9(0.04) + 12(0.01) \\ &= 3.66 \end{aligned}$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 3.66 - (1.5)(2.4) = 0.06$$

$$V(3X - 2Y) = 9V(X) + 4V(Y) - 12\text{Cov}(X, Y) = 9(0.45) + 4(0.84) - 12(0.06) = 6.69$$