CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 14

Survey Design and Analysis

Preface. The material for Lecture 14 is adapted from the third edition of *Sampling Techniques* by William G. Cochran (Wiley, 1977). The relevant pages of *Sampling Techniques* are indicated in brackets, although Lecture 14 is meant to be self-contained. Also, at the risk of stating the obvious, Lecture 14 is not a substitute for a course like CPH 631!

Notation for simple random sampling. [page 20] The population consists of N individuals (or other units such as households) with scores $y_1, y_2, \ldots, y_N$ on some numeric variable of interest such as the amount of money spent out-of-pocket on health care during 2008. The sample consists of n ($\leq N$) individuals. For convenience we represent their scores as $y_1, y_2, \ldots, y_n$ even though they may not be the first n members of the population.

The population mean $N^{-1}\sum_{i=1}^{N} y_i = N^{-1}(y_1 + \cdots + y_N)$ is denoted $\bar{Y}$, and the sample mean $n^{-1}\sum_{i=1}^{n} y_i = n^{-1}(y_1 + \cdots + y_n)$ is denoted $\bar{y}$. These conventions for capital and lower case letters differ from those more widely employed in statistical inference, where a capital letter denotes a random variable and a lower case letter denotes a generic numerical value. Here, a capital letter denotes a population quantity while a lower case letter denotes a sample quantity.

Likewise, we define capital S^2 to be the population variance $(N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and lower case s^2 to be the sample variance $(n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Simple random sampling. [page 18] We have a simple random sample of size n if every group of n individuals in the population has the same probability of being selected for the sample.

For instance, suppose that there are $N = 4$ students named "A", "B", "C", and "D" in an advanced public health class. Suppose that we wish to draw a sample of size $n = 2$.

One option is to flip a fair coin, choosing "A" and "B" if the coin comes up heads but choosing "C" and "D" if the coin comes up tails. Note that each individual has a 50% chance of being included in the sample. Yet, this does not yield a simple random sample. There is no chance for "A" and "C" to be selected (together), for "A" and "D" to be selected, for "B" and "C" to be selected, or for "B" and "D" to be selected.

Simple random sampling requires a 1/6 probability for "A" and "B" to be selected, a 1/6 probability for "A" and "C", a 1/6 probability for "A" and "D", a 1/6 probability for "B" and "C", a 1/6 probability for "B" and "D", and a 1/6 probability for "C" and "D". This can be achieved by rolling a fair six-sided die. Note that, here too, each individual has a 50% chance of being included in the sample.

Estimating a population mean by simple random sampling. [pages 22-24 and 27] Let $f := n/N$ denote the fraction of individuals in the population that are included in the sample. A natural estimate of the population mean $\bar{Y}$ is the sample mean $\bar{y}$, which has expected value $\bar{Y}$ and variance $S^2(1 - f)/n$. Thus, the sample mean is "unbiased" in that it has no systematic tendency (across repeated simple random samples) to underestimate or overestimate the population mean.

We refer to $1 - f$ as the "finite population correction factor". The finite population correction factor is not mentioned in introductory statistics courses, where we generally assume that the population is so large as to be effectively infinite. However, if $f > 0.05$, and especially if $f > 0.10$, taking the finite population correction factor into account is worthwhile because it allows us to validly shorten our confidence intervals.

Rather than use

$$\bar{y} \pm z_{1-\alpha/2} s/\sqrt{n} \quad \text{or} \quad \bar{y} \pm t_{n-1,1-\alpha/2} s/\sqrt{n},$$

as we do in an introductory statistics course, we can use

$$\bar{y} \pm z_{1-\alpha/2} s\sqrt{1-f}/\sqrt{n} \quad \text{or} \quad \bar{y} \pm t_{n-1,1-\alpha/2} s\sqrt{1-f}/\sqrt{n}$$

as a $100(1-\alpha)\%$ confidence interval for the population mean.

Discussion Questions. Consider the extreme case that $n = N$. What does the finite population correction factor do to the confidence interval? Can you provide an intuitive explanation for this?

Notation for stratified random sampling. [page 90] The original population consisting of N individuals is partitioned into L subpopulations with $N_1, \ldots, N_L$ individuals such that $N_1 + \cdots + N_L = N$. The scores for the individuals in subpopulation h ($1 \le h \le L$) are denoted $y_{h1}, \ldots, y_{hN_h}$. For each h ($1 \le h \le L$), a sample of size n_h ($\le N_h$) is drawn from subpopulation h. For convenience we represent their scores as $y_{h1}, \ldots, y_{hn_h}$ even though they may not be the first n_h members of subpopulation h.

The subpopulation mean $N_h^{-1} \sum_{i=1}^{N_h} y_{hi} = N_h^{-1}(y_{h1} + \cdots + y_{hN_h})$ is denoted $\bar{Y}_h$, and the corresponding sample mean $n_h^{-1} \sum_{i=1}^{n_h} y_{hi} = n_h^{-1}(y_{h1} + \cdots + y_{hn_h})$

is denoted $\bar{y}_h$.

Likewise, we define capital S_h^2 to be the subpopulation variance $(N_h - 1)^{-1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$ and lower case s_h^2 to be the corresponding sample variance $(n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$.

Finally, we let f_h denote the sampling fraction n_h/N_h within subpopulation h and W_h the proportion of the full population N_h/N belonging to subpopulation h.

Stratified random sampling. [page 89] We have a stratified random sample of total size $n = n_1 + \cdots + n_L$ if we draw a simple random sample of size n_h from subpopulation h for each h $(1 \leq h \leq L)$ and if these L simple random samples are drawn independently.

One reason to perform stratified random sampling is that we want to explicitly estimate the mean within each subpopulation, not just for the population as a whole. Here the subpopulations may be defined in terms of demographic characteristics such as gender, race, and income or based on health characteristics such as whether and how much a person smokes.

A second reason to pursue stratified random sampling is convenience in administering a survey. Here the subpopulations may be defined geographically, according to the locations of offices from which the survey efforts are coordinated.

A third reason to employ stratified random sampling is to obtain a more precise estimate of the population mean than may be possible with a simple random sample. We discuss this idea further below.

Estimating a population mean by stratified random sampling. [pages 91-92 and 95-96] Obviously $\bar{y}_h$ is a natural estimate for $\bar{Y}_h$. Thus, since

$$\bar{Y} = N^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{L} (N_h/N) N_h^{-1} \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{L} W_h \bar{Y}_h,$$

we can adopt

$$\bar{y}_{st} := \sum_{h=1}^{L} W_h \bar{y}_h$$

to estimate the population mean $\bar{Y}$.

Note that $\bar{y}_{st}$ is not generally equal to the sample mean $\bar{y}$, which has the representation

$$\bar{y} = \sum_{h=1}^{L} (n_h/n) \bar{y}_h,$$

unless $n_h/n = N_h/N$ for all h $(1 \leq h \leq L)$. As we will see later, such "proportional allocation" is not necessarily advantageous.

The expected value of $\bar{y}_{st}$ is $\bar{Y}$, while the variance is $\sum_{h=1}^{L} W_h^2 S_h^2 (1-f_h)/n_h$. Hence, a $100(1-\alpha)\%$ confidence interval for the population mean is

$$\bar{y}_{st} \pm z_{1-\alpha/2} \sqrt{\sum_{h=1}^{L} W_h^2 s_h^2 (1-f_h)/n_h} \quad \text{or} \quad \bar{y}_{st} \pm t_{df,1-\alpha/2} \sqrt{\sum_{h=1}^{L} W_h^2 s_h^2 (1-f_h)/n_h},$$

where

$$df := \frac{\left(\sum_{h=1}^{L} g_h s_h^2\right)^2}{\sum_{h=1}^{L} (g_h^2 s_h^4)/(n_h - 1)} \quad \text{and} \quad g_h := N_h (N_h - n_h)/n_h.$$

Discussion Questions. Consider the extreme case that $S_h^2 = 0$ (and, hence, $s_h^2 = 0$) for all h. What does this imply about the manner in which the subpopulations have been determined? What do you conclude about the circumstances under which stratified random sampling may yield a much more precise estimate of the population mean than simple random sampling?

Optimal allocation in stratified random sampling. [pages 96-99] Consider the following scenario. We intend to survey a total of $n = 1200$ individuals from three subpopulations with $N_1/N = 0.60$, $N_2/N = 0.20$, $N_3/N = 0.20$, $S_1^2 = 100$, $S_2^2 = 100$, and $S_3^2 = 400$. Assuming that n/N is negligibly small, so that each f_h can be treated as if it were 0, how shall we choose n_1, n_2, and n_3 to make the variance of $\bar{y}_{st}$ as small as possible?

Strategy #1: Equal allocation. If we take $n_1 = n_2 = n_3 = 400$, then the variance of $\bar{y}_{st}$ is

$$0.6^2(100)/400 + 0.2^2(100)/400 + 0.2^2(400)/400 = 0.1400.$$

Strategy #2: Proportional allocation. If we take $n_1 = 720 = 0.60(1200) = (N_1/N)n$, $n_2 = 240 = 0.20(1200) = (N_2/N)n$, and $n_3 = 240 = 0.20(1200) = (N_3/N)n$, then the variance of $\bar{y}_{st}$ is

$$0.6^2(100)/720 + 0.2^2(100)/240 + 0.2^2(400)/240 = 0.1333.$$

Strategy #3: Neyman allocation. If we take $n_1 \propto (N_1/N)S_1$, $n_2 \propto (N_2/N)S_2$, and $n_3 \propto (N_3/N)S_3$, then $n_1 + n_2 + n_3 = 1200$ requires that $n_1 = 600$, $n_2 = 200$, and $n_3 = 400$. The variance of $\bar{y}_{st}$ is

$$0.6^2(100)/600 + 0.2^2(100)/200 + 0.2^2(400)/400 = 0.1200.$$

Clearly, the Neyman allocation is superior to equal allocation and proportional allocation. In fact, the Neyman allocation is superior to any other scheme for choosing n_1, n_2, and n_3. Moreover, the superiority of the Neyman allocation is not just an artifact of this example but rather a general phenomenon.

The reasoning behind the Neyman allocation is as follows.

If two subpopulations have the same variance, then the subpopulation making up a larger fraction of the full population should receive a greater allocation of the total sample size, as the individuals in this subpopulation play a larger role in determining the mean of the full population.

If two subpopulations make up the same fraction of the full population, then the subpopulation with the greater variance should receive a greater allocation of the total sample size, as the mean of this subpopulation is more difficult to estimate.

Cluster sampling. [page 233] Although time limitations prevent me from giving a mathematical treatment of cluster sampling in Lecture 14, I want to make a few remarks about it.

The motivation for cluster sampling is that simple random sampling may be prohibitively expensive. For instance, suppose that our population consists of Lexington households and that we want to survey 3000 households. One problem we face is that there may not be a complete, current, and accurate list of all Lexington households. Even if there were such a list, locating and traveling to 3000 households spread throughout the city could be very time-consuming.

A cluster sample entails assigning population members to groups called "clusters" and then taking a simple random sample of the clusters — with the intention of surveying all population members in each selected cluster. A simple random sample of 100 Lexington city blocks will yield a cluster sample containing approximately 3000 households if there are an average of 30 households per city block.

Discussion Question. Having homogeneous subpopulations can help precision in stratified random sampling, but having homogeneous clusters can hurt precision in cluster sampling. Can you give an intuitive explanation for this?