14.0 Surveys

- Answer Questions
- Random Samples
- Bias
- Variance
- Major Issues
- Problems

14.1 Random Samples

In a simple random sample of n units from a population,

- each unit is equally likely to be chosen,
- each pair of units is equally likely to be chosen,
- and so forth, until
- each n-tuplet is equally likely to be chosen.

Sampling is done without replacement.

Survey sampling is done in order to learn about a population parameter, e.g.:

- a population proportion
- a population mean
- a population standard deviation (more rarely).

One wants to get both:

- an estimate of the unknown parameter, and
- a numerical description of the uncertainty in the estimate.

The uncertainty (or total survey error in the estimate has two parts:

- bias, which is the systematic deviation of the estimate from the true value, and
- variance, which is the deviation due to chance error.

The standard deviation of an unbiased estimator is called the standard error of the estimate. From our previous work,

- the standard error of an average is $sd/\sqrt{n}$
- the standard error of a sum is $sd * \sqrt{n}$.

14.2 Bias

Survey researchers have identified many kinds of bias:

- Response Bias. The phrasing of the query can affect the response. For example: "Do you feel that liberal tax-and-spend policies and irresponsible Washington insiders have hurt your long-term financial security?"

- Nonresponse Bias. People who are missed in a survey often tend to be systematically different from those who are found, and this systematic difference can be reflected in their opinions. For example: People who are not at home to answer survey phone calls may be working three jobs to make ends meet, and their view of Bush's economic policies might differ from those with more leisure time.

- Household Bias. Often pollsters only talk to one person per home, so large households are underrepresented. Consider what would happen at a fraternity, for example.
- Interviewer Bias. The interviewer may not follow the complex survey instructions. (These instructions can be difficult, and perhaps the interviewer feels pressure to have many short interviews rather than one lengthy one.)
- Nonrespondent Bias. People who refuse to participate in the survey may be different from those who do. For example, consider a survey of sexual behavior in Utah.
- Liars. Some people misrepresent themselves. For example, consider the poll results in the David Duke campaign.

Pollsters and survey experts try to correct for known biases. Techniques include:

- upweighting responses from large households;
- upweighting data from poor people or those with little formal education;
- using a model, such as the continuum of nonresponse theory.

The continuum of nonresponse theory says that people who are hard to catch or unwilling to respond can have their responses estimated as a function (perhaps a linear regression function) of the difficulty in finding people or the cost of convincing them to respond.

For example, the people who are never home to answer their phone may have opinions more like those whom the pollsters had to call 15 times before making contact, and less like those of the people who answered on the first call.

14.3 Variance

The variance portion of the total survey error is due to sampling.

All sampling begins (explicity or implicitly) with a frame. A frame is a list of all eligible respondents.

If one has a simple random sample from the frame, then variance is small. For large populations, we can ignore the distinction between sampling with replacement and sampling without replacement.

But simple random sampling is expensive. For this reason, people have tried alternatives:

- Quota Sampling. Here interviewers are told to find a fixed number of people in each of various demographic segments (e.g., rich white middle-aged men, rich black middle-aged men, etc.). But this introduces bias.
- Multistage Cluster Sampling. Here one divides the frame up into groups that are relatively homogeneous and cheap to survey. Then one draws the groups at random and sends interviewers to survey at random within those groups.

Many of these fixes involve weighting the sample, but this can be problematic. Different weights can give dramatically different answers.

14.4 Major Issues in Surveys

The three major problems faced by survey firms are:

- bias reduction strategies
- cognitive design of questionnaires
- falling response rates.

Question/survey design requires piloting and planning. Unscrupulous pollsters like to ask loaded questions, but even pollsters with integrity can have trouble finding neutral wording. Even the order of the options can have a big effect.

To increase response rates, consider:

- CATI and CAPI
- incentives
- mail $<$ web $<$ phone $<$ personal interviews
- interviewer flexibility

Other ways to improve survey quality include using mathematical models of non-response, or creating stables of respondents.

14.5 Problems

Problem 1: You have a simple random sample of 25 Duke students and record their GPA. Your sample mean is 3.1, and the sample standard deviation is .3. What is the proability that the true mean (the EV of the box) is less than 3?

Note that in order to solve this, we have to assume that the standard deviation of the sample is equal to the *sd* of the box. In practice, there is a very easy way to handle this, but we will not talk about that until later in the course.

$$\begin{aligned}
\mathbf{P}[EV < 3] &= \mathbf{P}[-EV > -3] \\
&= \mathbf{P}[\bar{X} - EV > \bar{X} - 3] \\
&= \mathbf{P}[\frac{\bar{X} - EV}{sd/\sqrt{n}} > \frac{\bar{X} - 3}{sd/\sqrt{n}}] \\
&\doteq \mathbf{P}[Z > \frac{\bar{X} - 3}{sd/\sqrt{n}}] \\
&= \mathbf{P}[Z > (3.1 - 3)/(.3/\sqrt{25})] \\
&= \mathbf{P}[Z > 1.66]
\end{aligned}$$

From the standard normal table, we know this has chance (1/2) (100 - 90.11) $= 4.945\%$, so the probability that the mean Duke GPA is below a B is .04945.

Problem 2: You are an auditor, and wish to estimate the total debt of a company. The company has 900 outstanding bills. You draw a simple random sample of size 16, and see that the sample mean is $500 and the sample standard deviation is $100.

In particular, you want to estimate the probability that the company owes more than $455,000.

As before, we must assume that the sample standard deviation is equal to the standard deviation of the 1000 bills.

Describe the box model for this situation. How does the EV relate to the question?

The probability of excessive debt is

$$
\begin{aligned}
\mathbf{P}[\ \textbf{debt}\ > 455,000] &= \mathbf{P}[900 * \bar{X} - \ \textbf{debt}\ < 900 * \bar{X} - 455,000] \\
&= \mathbf{P}[\frac{900 * \bar{X} - \ \textbf{debt}}{\sqrt{900} sd} < \frac{900 * \bar{X} - 455,000}{\sqrt{900} sd}] \\
&\doteq \mathbf{P}[Z < \frac{900 * \bar{X} - 455,000}{\sqrt{900} sd}] \\
&= \mathbf{P}[Z < [450,000 - 455,000]/(30 * 100)] \\
&= \mathbf{P}[Z < -1.67].
\end{aligned}
$$

From before, we know the probability of this is .04945.