Chapter 19. Sample Surveys

- A sample is part of a population. A parameter is a numerical fact about a population. Usually a parameter cannot be determined exactly, but can only be estimated. A statistic can be computed from a sample, and used to estimate a parameter. A statistic is what the investigator knows. A parameter is what the investigator wants to know.

- The method used to draw the sample matters.

- Some methods are terrible. (Literary Digest poll’s choice of Landon [section 2]: i. selection bias—only people listed on telephone books and club membership lists were polled, ii. non-response bias—only 2.4 million people bothered to reply, out of the 10 million people who got the questionnaire.) (R5)

- Handpicking the sample to get a representative cross-section (quota sampling) tends not to work very well. (Gallup poll’s “election” of Dewey [section 3]: unintentional bias on the part of the interviewers.) (Another example:

<table>
<thead>
<tr>
<th>Math 1680-003</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrats</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Republicans</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Independents</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Suppose the percentages of male and female are known, but the percentages of party affiliation are unknown. In a quota sample of ten persons, you will pick 4 men and 6 women. But it is still up to the interviewer who to pick, as long as he gets 4 men and 6 women. So he may subconsciously choose too many Republican, say 3 MR, 1 MD, 3 FR, 2 FD, 1 FI. In this case, you get a perfect cross-section in terms of gender, but you get a wrong estimate of the percentage of Republican.

- Haphazard selection may be even worse. (R6: convenience samples)

- The best methods for drawing a sample involve the planned introduction of chance—probability methods:
  - the interviewers have no discretion at all as to whom they interview;
  - there is a definite procedure for selecting the sample, and it involves the planned use of chance.

- Simple random sampling means drawing at random without replacement. (A2)

- Real sample surveys, of course, use methods much more complicated than simple random sampling — multistage cluster sampling.

- Multistage cluster sampling is a probability method, but quota sampling is not—it fails both features of probability methods.

Chapter 20. Chance Errors in Sampling

- The sample percentage will differ from the population percentage:

  \[
  \text{sample percentage} = \text{population percentage} + \text{chance error.}
  \]

- With a simple random sample, the expected value for the sample percentage equals the population percentage.

- To compute the SE for a percentage, first get the SE for the corresponding number; then convert to percent, relative to the size of the sample. (A2,5,7)

  \[
  \text{SE for percentage} = \frac{\text{SE for number}}{\text{size of sample}} \times 100\%.
  \]

- If the probability histogram for the number follows the normal curve, so does the probability histogram for the percentage because conversion to percent is only a change of scale. (figure 3) (R5)

- When estimating percentages, accuracy depends mainly on the absolute size of the sample, rather than size relative to the population. To justify this, i) consider drawing the same amount with replacement from two boxes with different sizes but same proportion of 1’s; ii) consider drawing without replacement, but the number of draws is just a tiny fraction of the size of the box, which is the usual case. (R6)
What if the number of draws is not a tiny fraction of the size of the box? Indeed, when drawing without replacement, the box does get a bit smaller, reducing the variability slightly. So the SE for drawing without replacement is a little less than the SE for drawing with replacement.

\[
\text{SE when drawing WITHOUT replacement} = \text{correction factor} \times \text{SE when drawing WITH replacement}
\]

where the correction factor is

\[
\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - 1}}
\]

Chapter 21. The Accuracy of Percentages

• How accurate is an estimated percentage likely to be? Answer: (i) accuracy is determined by the SE; (ii) the estimate is likely to be about right, but off by an SE or so.

• The bootstrap method: a procedure for estimating the standard error from the sample—substitution of estimates for parameters in the formula. The estimate is good when the sample is reasonably large. (Example 1 on page 378, A1, A9)

• Confidence intervals for the parameter (eg. population percentage)

1. A confidence interval is used when estimating an unknown parameter from sample data. The interval gives a range for the parameter—and a confidence level that the range covers the true value. (C1, C2, C6, B4) For instance:

   (a) The interval “sample percentage ± 1 SE” is a 68%-confidence interval for the population percentage.
   (b) The interval “sample percentage ± 2 SEs” is a 95%-confidence interval for the population percentage.
   (c) The interval “sample percentage ± 3 SEs” is a 99.7%-confidence interval for the population percentage.

2. The chance variability is in the sampling process, not in the parameter. For example, it seems natural to say “There is a 95% chance that the population percentage is between 75% and 83%.” But there is a problem here. In the frequency theory, a chance represents the percentage of the time that something will happen. No matter how many times you take a sample, the population percentage will not change. Either this percentage was between 75% and 83%, or not. There really is no way to define the chance that the parameter will be in the interval from 75% to 83%. That’s why statisticians have to turn the problem around slightly. They realize that the chances are in the sampling procedure, not in the parameter. And they use the new word “confidence” to remind you of this.

3. Probabilities are used when you reason forward, from the box to the draws; confidence levels are used when reasoning backward, from the draws to the box.

4. A sample percentage will be off the population percentage, due to chance error. The SE tells you the likely size of the amount off. Confidence levels were introduced to make this idea more quantitative.

• Warning: The formulas for simple random samples may not apply to other kinds of samples.

Chapter 23. The Accuracy of Averages

• The average of the draws will be around the average of the box, give or take an SE or so.

\[
\text{SE for average of draws} = \frac{\text{SE for sum}}{\text{number of draws}}
\]

• With a simple random sample, the SE of the average is estimated by substituting the SD of the sample for the unknown SD of the box. The estimate is good when the sample is large. Then, confidence intervals are obtained by going the right number of SEs either way from the average of the sample. (C4)

• The formulas in this chapter are for simple random sampling, i.e. draws from a box, and should not be applied mechanically to other kinds of samples. (D4)

• The calculations for confidence levels depend on the normal approximation. (D3)