Class Worksheet

April 7 and 9

Math 125 *Kovitz* 2025

Sampling (Chapter 19)

Sample Surveys

The statistician infers the values of population parameters from the values of statistics, which are obtained from the sample.

When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a large scale.

Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.

Some samples are really bad. To find whether a sample is any good, ask how it was chosen. Was there selection bias? non-response bias? You may not be able to answer these questions just by looking at the data.

In quota sampling, the sample is hand-picked to resemble the population with respect to some key characteristics. The method seems reasonable, but does not work very well. The reason is unintentional bias on the part of the interviewers.

Simple random sampling means drawing at random without replacement.

To minimize bias, an impartial and objective probability method should be used to choose the sample.

In general, use the following formula:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Chance Errors in Sampling (Chapter 20)

For a simple random sample from a given population, the following equation holds for the percentage (in a specified category)

percentage in sample = percentage in population + 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 a percent, relative to the size of the sample. As a cold mathematical formula.

$$\textbf{SE for percentage} = \frac{\textbf{SE for number}}{\textbf{size of sample}} \times 100\%.$$

Multiplying the size of a sample by some factor divides the SE for a percentage not by the whole factor—but by its square root. (For instance, multiplying the size of the sample by 4 divides the SE for the percentage by $\sqrt{4} = 2$.) This is exact when drawing without replacement, provided the number of draws is small by comparison with the number of tickets in the box.

If you wish, you may compute the SE for a percentage directly from the SD of the box by the formula

$$\text{SE for percentage} = (\text{SD of box}/\sqrt{\text{number of draws}}) \times 100\%.$$

It is a fact that the SD of a 0–1 counting box is always .5 or less. This means that the SE for a percentage is always less than $\dfrac{50\%}{\sqrt{\text{number of draws}}}$, no matter what the percentage of the box.

Warning: if the problem involves classifying and counting to get a percent, put 0's and 1's in the box.

Using the normal curve.

When drawing at random from a box of 0's and 1's, the percentage of 1's among the draws is likely to be around ______ , give or take ______ or so. The expected value for the percentage of 1's among the draws fills in the first blank. The SE for the percentage of 1's among the draws fills in the second blank.

The Correction Factor

When estimating percentages by drawing a sample without replacement, it is the absolute size of the sample which determines accuracy, not the size relative to the population. This is true when the sample is only a small part of the population, which is the usual case.

The SE for drawing without replacement is always a little less than the SE for drawing with replacement. There is a mathematical formula to deal with this:

$$\begin{matrix}\text{SE when drawing} \\ \text{WITHOUT Replacement}\end{matrix} = \begin{matrix}\text{correction} \\ \text{factor}\end{matrix} \times \begin{matrix}\text{SE when drawing} \\ \text{WITH replacement}\end{matrix}$$

The formula for this factor looks complicated:

$$\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - \text{one}}}$$

When the number of tickets in the box is large relative to the number of draws, the correction factor is nearly 1 and can be ignored. Then it is the absolute size of the sample which determines accuracy, through the SE for drawing with replacement.

Problems to think about

A researcher wishes to determine the average number of times all adult males living in the United States have been married. Discuss a method of taking a sample of size 2500. Would it be a good idea to select a city at random and randomly choose 2500 names from the telephone directory? If not, what other method might be used?

A fair coin is tossed 64 times. Find the expected value and the SE for the number of heads and for the percent of heads.

A box contains 16 red marbles and 9 blue ones. Four hundred draws will be made at random with replacement from this box. The percentage of red marbles among the draws will be around ________%, give or take ________% or so.

A certain city has 300,000 residents age 18 and over, and the average gross income is \$27,000 with an SD of \$13,000. Of them, 64% are married, and 8% have incomes over \$80,000.

A simple random sample of 900 people will be drawn from this population.

To estimate the chance that between 6% and 10% of the people in the sample have incomes over \$80,000, a box model is needed.

(a) Should the number of tickets in the box be 900, or 300,000?

(b) Each ticket in the box shows
a zero or a one   a gross income

(c) True or false: the SD of the box is \$13,000.

(d) True or false: the number of draws is 900.

(e) Find the chance (approximately) that between 6% and 10% of the people in the sample have incomes over \$80,000.

(f) With the information given, can you find the chance (approximately) that between 5% and 7% of the people in the sample have incomes over \$100,000? Either find the chance, or explain why you need more information.

To find the chance (approximately) that the total gross income of the people in the sample is over \$24,000,000, a box model is needed. Work through parts (a) through (d); then find the chance or explain why you need more information.

You have hired a polling organization to take a simple random sample from a box of 1,200,000 tickets, and estimate the percentage of 1's in the box. Unknown to them, the box contains 50% 0's and 50% 1's. How far off should you expect them to be if they draw 100,000 tickets?

A survey organization wants to take a simple random sample in order to estimate the percentage of people who will vote for a certain candidate. To keep the costs down, they want to take as small a sample as possible. But their client will only tolerate chance errors of one half of a percentage point or so in the estimate. Should they use a sample of size 100, 2,500, 10,000, or 40,000? You may assume the population to be very large.