MAS 170 Elementary Statistics Spring 2020
Sampling, Ch 16--21,23
(revised 4/16/2020)

A sequence of random draws from a box model (see notes on Probability) is called a ***sample*** from the box. A sequence of random draws taken without replacement is called a ***simple random sample***. The ***sample sum*** is the sum of the draws. The ***sample average*** is the average of the draws. For a box whose tickets are labeled with 0's and 1's, the percentage of 1's drawn in a sample is called the ***sample percentage of 1's***.

Example:

Box contents:
1,2,3

All possible samples of size 2, with replacement:
(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)

All possible sample sums:
2,3,4,3,4,5,4,5,6

All possible sample averages:
1,1.5,2,1.5,2,2.5,2,2.5,3

The sample sum, the sample average, and the sample percentage of 1's are examples of ***statistics***. A ***statistic*** is any number that is computed from a sample. Another important statistic is the ***sample SD*** (the SD of the list of draws). By contrast, any number that can be computed from the numbers in the box is called a ***parameter***. For example, the average and the SD of the numbers in the box are parameters. Sampling theory is the art of exploiting the relationships between parameters and statistics to do these two kinds of tasks.

Prediction: If you know what is in a given box, calculate the probability that a given statistic will fall in a given range of values.

Inference: If you do not know all the details of the contents of a box, but you have access to one or more samples from the box, make educated guesses about the parameters of the box.

Sampling theory relies on the key formulas (1) and (2) below that express the fundamental relationships between the contents of a box model and samples coming out of the box. First we have to introduce the quantities involved. The average of the list of all possible sample sums is called the ***expected sum***, which we will denote by "expected(sum)". The SD of the list of all sample sums is called the ***SE for the sum***, which we will denote by "SE(sum)". Similarly, the average of the list of all sample averages is called the ***expected average*** (denoted "expected(ave)") and the SD of the list of all sample averages is called the ***SE for the average*** (denoted "SE(ave)"). For a box containing only 0's and 1's, the average of all the sample percentages of 1's is called the ***expected percentage of 1's*** (denoted "expected(%1's)") and the SD of the list of all sample percentages of 1's is called the ***SE for the percentage of 1's*** (denoted "SE(%1s)").

For the example above, we have the following.

expected(sum) = (2+3+3+4+4+4+5+5+6)/9 = 4
SE(sum) = sqrt((-2)^2 + 2*(-1)^2 + 3*(0^2) + 2*(1)^2 + 2^2)/9)
= (approx) 1.15
expected (average) = 2
SE(average) = (approx) 0.58

For samples with replacement, we have the following formulas. Equation (2) is called the ***Square Root Law***.

(1) expected(sum) = (average of the box) * (number of draws)
expected(ave) = average of the box
expected(%1's) = percent 1's in the box

(2) SE(sum) = (SD of the box) * sqrt(number of draws)
SE(ave) = SE(sum) / sqrt(number of draws)
SE(%1's) = SE(sum) / sqrt(number of draws) * 100%

For samples without replacement, there is no change in formula (1), but each of the equations in (2) is replaced by

(2') SE(without replacement) = SE(with replacement) * (correction factor)

where the ***correction factor*** is given by the following.

correction factor = sqrt((N – n) / (N – 1)), where
N = the number of tickets in the box
n = the number of draws in the sample

The reason we say that (1) and (2) are just two equations and not six is because of the simple relationships between sum, average, and percentage of 1's. The average of a list of numbers is the sum of that list of numbers divided by the number of items in the list; the percentage of 1's (in a list of 0's and 1's) is the average of that same list of 0's and 1's times 100%. Applying these simple relationships (summarized in equation (0), below) to the first of the three equations in each of (1) and (2) produces the rest of the equations in (1) and (2).

(0) average(list) = sum(list) / (#items in the list)
(percent 1's in a list of 0's and 1's) = (average of the list) * 100%

Sampling theory derives its power from a fact called the ***Central Limit Theorem***. The Central Limit Theorem says that, no matter what numbers are in a box model, the histogram of all possible (standardized) sample sums (which is the same as the standardized sample average or the standardized sample percentage of 1's) is close to the standard normal distribution, as long as the number of draws is sufficiently large. [Our textbook is vague about how large is "sufficiently large". For a box with a uniform distribution (that is, a histogram for which all the blocks are close to even in height), 25 draws or more is sufficiently large. For a box that is very uneven (that is, a histogram with blocks of very different heights), sufficiently large might be 100 or more draws.]

Basic prediction problem
==========

Armed with these facts and formulas, here is the skeleton of our basic prediction problem.

Given: a description of a chance process
Question: find the probability that a sample (with a sufficiently large number of draws) from a box model for the chance process produces a sum or average in a given range

The procedure to solve the problem has these steps:

(first) Construct the box model that matches the question. How many tickets are in the box? What numbers are on the tickets? How many draws? Are the draws taken with or without replacement? Is the question about the sum or the average of the draws?

(second) Calculations using (1), (2) or (2')
ave(box)
SD(box)
expected(sum,ave,%1's) (depending on the question)
SE(sum,ave,%1's)

(third) Sketch a normal curve (justified by the Central Limit Theorem) and shade an area that answers the question. The center of the horizontal axis is expected(sum,ave,%1's)) (depending on the question). Find the z values needed for your shaded area using one of the formulas

$$z = \frac{(\text{sum,ave,\%1's in the question}) - (\text{expected(sum,ave,\%1's)})}{\text{SE(sum,ave,\%1's)}}$$

and finally use the normal distribution table to find the area that answers the given question.

Basic Inference Problem
==========

The basic inferential statistics problem in our text goes like this: you are given information about a sample from a box with unknown contents. You are asked to estimate the average of the box, and to make an educated statement about how far off that estimate might be from the actual value.

Let's write ave(observed) and SD(observed) to denote the average and SD of the list of numbers in the given sample. By virtue of eqn. (1), it is intuitively reasonable to use ave(observed) to estimate ave(box). It is not so clear, but it turns out to be okay for large enough samples, to use SD(observed) to estimate SD(box). By (2), it makes sense that our actual ave(observed) will be off from the ave(box) by SE(ave), which we can estimate by

(3) SE(ave) = (estimated) SD(observed) / sqrt(number of draws)

(times the correction factor, if sampling without replacement). Equation (3) is called the ***bootstrap*** estimate.

It makes intuitive sense to report our estimate of ave(box) as follows.

ave(observed) plus/minus (estimated SE(ave))

For a positive number z, let's write Area(z) to denote the area under the normal curve in the range −z to +z. The interval

(4) ave(observed) plus/minus z*(estimated SE(ave))

is called an ***Area(z) confidence interval for ave(box)***. For example, a 95% confidence interval for the average of the box is the observed sample average plus or minus 2 times the estimated SE for the sample average.

By virtue of the Central Limit Theorem, we can make a precise probability statement about the interval (4). We can say that approximately Area(z) (percent) of all Area(z) confidence intervals (imagine many samples of the given sample size) will contain the average of the box. For example, approximately 95% of all 95% confidence intervals (calculated from many samples) will contain the box average.