DISCOVERY **#30 & 31: Sampling, EV, SE**

April 3, 2019 and April 5th, 2019

Karle Flanagan and Wade Fagen-Ulmschneider

In our discussion of random variables, we started with games of chance because they easily translate into probability models. We know all of the outcomes and their probabilities. In the next section, we are going to see how random variables relate to gathering information about large populations from small samples.

We take a sample to find out about a larger population. We usually don't have the resources to gather information on everyone in the whole population so instead, we select a small sample and use it to make inferences about the larger population. ***Obviously, the closer the sample resembles the population, the more accurate our inferences will be.***

Terminology

- **Population:** the whole class of individuals about whom the investigator wants to generalize.
- **Sample:** the part of the population the investigator examines.
- **Inferences:** generalizations about the population that come from the sample.
- **Parameters:** numerical facts about the population.
- **Statistics:** estimates of the parameters computed from the sample.

Main Idea- Sample should be as representative of the population as possible.

Consider the following methods of sample selection. Which is the **best method** for drawing a sample as much like the population as possible?

- **Quota Sampling:** The researcher hand picks the sample to resemble the population on all relevant characteristics. For example, if the population is 50% female and 50% Democratic, he makes sure his sample is as well.

- **Self-Selected Sample:** The researcher publicly posts the survey (on TV, newspapers, Internet, etc.) and allows anyone to respond.

- **Probability Sample:** The sample is *randomly* selected using a planned introduction of chance. The simplest probability method is *simple random sampling* where everyone in the population has an equal chance of being chosen.

Random Samples are best for surveys for exactly the same 2 reasons that randomized controls are best for experiments:

- **Blind chance works better than human judgment. Human judgment introduces bias. Random selection is most likely to make the sample as like the population as possible because it eliminates selection bias**. With enough subjects, **random differences average out**, not only on the characteristics that the researcher has identified as relevant but on *all* characteristics, including hidden ones that the researcher might not realize are important.

- **Random samples can be translated into probability models so that we can use statistics to measure the accuracy of our estimates.** Random samples are based on ***chance procedures***. **Without a chance process, we can't use statistics to analyze the accuracy of the survey**. We could make an estimate about the population from a non-random sample but it would be meaningless because we would have no way of knowing how far off it is.

Probability Methods eliminate selection bias but there are still other possible sources of bias. The 2 main ones are:

Non-response bias- Random selection eliminates bias in who is chosen for the survey. But not everyone who is chosen responds. The people who respond may be systematically different from those who don't. This introduces non-response bias. Researchers reduce this bias by weighting the responses of people according to the difficulty of obtaining their response. Those responses that were hard to get are weighted more.

Response bias- Phrasing of questions influences responses. For example, responses to questions about abortion are different if phrased as "a women's right to choose" as opposed to "killing an unborn baby".

Chance Error and Bias

There are many sources of bias. But even if they were all eliminated our estimate would still be likely to be off due to **chance error.**

Estimate = Parameter + Bias + Chance Error

Expected Value and Standard Error

For means (averages):

Expected Value of the Sample Mean	Standard Error of the Sample Mean
$EV_{avg} = E(\overline{X}) = \mu$	$SD_{avg} = SD / \sqrt{n}$

For percents: The formulas are the same as the formulas for averages, just remember to multiply by 100 to get your answers in percent form.

Expected Value of the Sample Percent	Standard Error of the Sample Percent
EV%=population percent=p	$SE\% = (SD / \sqrt{n}) * 100\%$

Also, to calculate the SD of a population with yes's and no's (1s and 0s) where p is the proportion of 1's in the population, and 1-p is the proportion of 0's in the population, use this formula: $\sigma = \sqrt{p(1-p)}$

Next, we will talk about *inference*: drawing conclusions about the population from what's known about the sample.

So far with gambling games, we've known exactly what's in the population and used it to draw conclusions about possible samples. Now we'll go in the opposite direction (which is much more common). We know the composition of one sample, and we use it to estimate the composition of the population.

Example 1: Percents
In February of 2019, a CNN Poll of 1,011 adults nationwide asked the following question: "Do you think the government should provide a national health insurance program for all Americans, even if this would require higher taxes?" 54% answered 'Yes'. The 1,011 adults were chosen as a *simple random sample*.

a) Estimate the percentage of all American adults who would favor a national health insurance program.

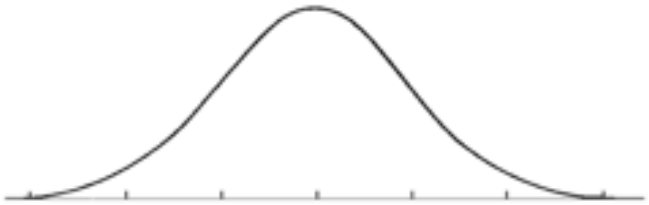
b) Obviously, our best estimate of the percentage of people in the general population who favor a national health insurance program will be the sample percent. But give or take what amount? What is the SE of the sample percentage?

Confidence Intervals

In Example 1 we got a sample percentage of 54% favoring a national health insurance program. The SE was estimated to be about _____%.
So we estimate that about 54% of all Americans give or take about _____% favor a national health insurance program even if it means raising taxes.

***How confident are we that our estimate is right?** How sure are we that if we really polled all US adults we'd get 54% ± _____% saying they favor a national health insurance program?

We can use the normal curve to answer that question because we know the probability histogram for the sample percent follows the normal curve (Central Limit Theorem).



We know that if hundreds of pollsters all took random samples of 1,011 people asking the same question, about 68% of them would get sample percents within 1 SE of the true population percent, and about 95% of them would get sample percents within 2 SE's of the true population percent.

So we can be about 68% sure that our sample percent is within 1 SE of the true population percent and about 95% sure that our sample percent is within 2 SE's of the true population percent.

Example 1 continued (Feb. 2019 random poll of 1,011 adults found 54% in favor of a national health insurance program.)
c) Suppose we wanted to be about 80% sure that the true population % lies in our confidence interval, how many SE's do we need to attach to our estimate of 54% favoring a national health insurance program?

This means:

d) Find the following confidence intervals for the % of all US adults who favor a national health insurance program.

68% confidence interval= 54% ± _____* _____% = (_____% - _____%)

90% confidence interval= 54% ± _____* _____% = (_____% - _____%)

95% confidence interval= 54% ± _____* _____% = (_____% - _____%)

99% confidence interval= 54% ± _____* _____% = (_____% - _____%)

e) Which of the following statements is true?

i) Our 95% confidence interval from part d can be applied to all adults worldwide.

ii) Our 95% confidence interval from part d can be applied to all adults nationwide.

iii) Our 95% confidence interval from part d can be applied to all US females.

iv) Our 95% confidence interval from part d can be applied to all US college students.

*****A sample is ONLY representative of the population it was drawn from. (no subgroups or wider populations)**

Sometimes we want to look at surveys that ask questions with numerical answers, like: "How many minutes of lecture did you sleep through?" "How much money did you make last year?" or "How much do you weigh?"

In these surveys, the relevant statistic is the average, so now we're interested in the EV and SE of the sample average.

Basically, we're doing exactly what we did in example 1, except we're dealing with the sample average instead of the sample percent. The population mean (μ) is unknown, however, we do know the sample mean $(\overline{X})$.

Example 2: Money! Thinking of your own situation, how much money per year would you need to make in order to consider yourself rich. A random sample of 1,572 adults nationwide was taken and their average was $150,000 and the sample SD was $158,600.

a) Estimate the average amount of money that all American adults would need to make to consider themselves rich.

b) Calculate the standard error of the sample average.

c) What is an 80% confidence interval for the average amount of money all American adults would need to consider themselves rich?

Choosing How Many People to Poll

Obviously the more people we poll the more accurate our sample statistics will be in estimating the true population parameters.

How many people do we need to poll to achieve a 95% confidence interval with a margin of error of 2%? How about 4%?

Example 1

In a pre-election poll in a close race you may want a 95% confidence interval with a small margin of error, say only 2%.

a) Estimate how many people you'd need to poll to get a 95% confidence interval with only a 2% margin of error?

b) If you were willing to accept a 4% margin of error, you'd only need to sample how many people?

c) In calculating how many people to poll we assumed the SD of the population to be 0.5 which is the biggest the SD can be in a 0-1 population. What if the SD of the population was really less, say only 0.4, then how many people would we need to poll to get a 4% margin of error?

A quick formula for choosing how many people to poll for margins of errors for Confidence Intervals: $\mathbf{n=(100*z*SD/M_{of}E)^2}$