The Accuracy of Percentages

Suppose we have the results of a sample. What can we say about their accuracy? What can we conclude about the population?

Where are we going?

Review of SE of a percentage.

Population $\Longrightarrow$ sample

Statistical inference: sample $\Longrightarrow$ population

The concept of a confidence interval: mechanics and interpretation.

Review: a 0-1 Box

- Box average = fraction of tickets which equal 1
- Box SD = $\sqrt{\text{(fraction of 0's) x (fraction of 1's)}}$

Population: 50,000 under age of 18 and 350,000 over age of 18. Take sample of 1000

- How many under 18?
- What % under 18?

Percentage = 100 x number÷1000

Sampling: like 1000 draws from a 0-1 box with 50,000÷400,000 = 12.5% ones.

EV of percentage = 100 x EV of number÷1000
= 12.5%

SE of percentage = 100 x SE of number÷1000

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

$$\text{SE of percentage} = \frac{\text{SE of number}}{\text{sample size}} \times 100\%$$

$$\text{SE of percentage} = \frac{\text{SE of number}}{n} \times 100\%$$

$$= \frac{\sqrt{n}\ \text{SD of box}}{n} \times 100\%$$

$$= \frac{\text{SD of box}}{\sqrt{n}} \times 100\%$$

Formula is exact for sampling with replacement and approximate without replacement.

Example

Population size = 500,000, percent unemployed = 20%, Sample size = 400

SD of Box = $\sqrt{.2 \times .8} = .4$

SE of sample percentage =

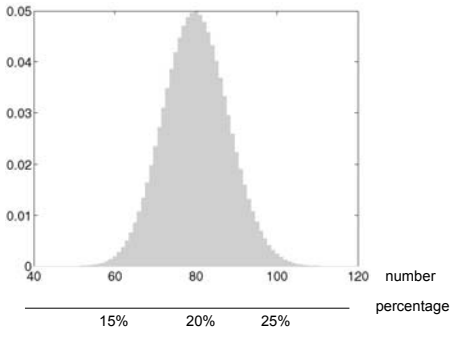
$$100 \times \frac{.4}{\sqrt{400}} = 2\%$$

Review: where did this come from?

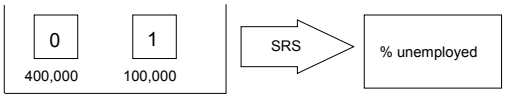
The chance of, say, 75 ones in 400 draws with replacement from a box with 20% ones is given by the

_____ formula

Binomial Probability Histogram: 400 draws with chance of success = 20%



We can do this problem:



Population Sample of 400

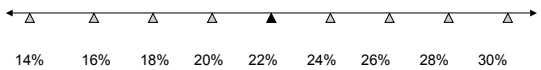
From a SRS we expect about ___%
unemployed give or take ___ or so.

This problem? Suppose you take a SRS of size 400 from a population of size 500,000 and you find 22% unemployed in the sample. What can you say about the population?

Population Sample

We know how to work from the population to the sample. How can we work *backwards* from the sample to the population? This is a problem of "**statistical inference**."

What if you knew that the SE was 2%?



Chance that sample % is less than 2 SE away from population percent is about ___ %

Chance that population percent is less than 2 SE away from sample % is about ___%.

Sample % plus or minus 2 SE is called a ___ % **confidence interval**.

Interpretation

Chance that sample % is less than 2 SE away from population percent is about 95%.

Chance that population percent is less than 2 SE away from sample % is about 95%.

What is the random object in these statements? The population % or the sample %?

Does this sentence make any sense: "The chance that the population % is within the interval 18% to 26% is 95%."

How can we get the SE?

We need the box SD and we don't know the box.

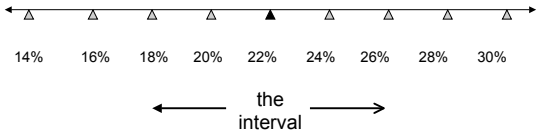
Box SD = $\sqrt{\text{(fraction 1's) x (fraction 0's)}}$

Bootstrap: Use the fraction of 1's and 0's in the sample.

Estimated Box SD = $\sqrt{.22 \times .78}$ = .41

The estimated SE of the sample percentage then turns out to be .41/20 = 2%.

A 95% confidence interval is 22% plus or minus 4%.



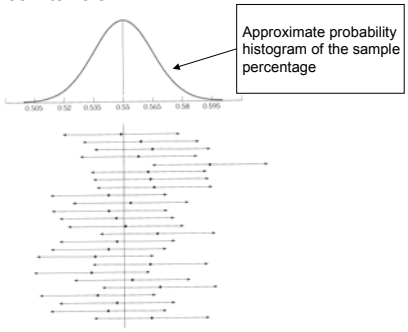
Different Sized Confidence Intervals

22% plus or minus 1 SE (2%) is a ____% confidence interval.

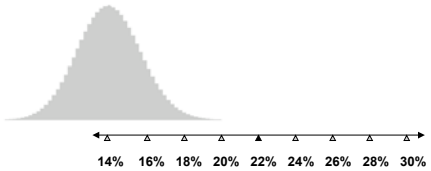
How could we find a 90% confidence interval? A 99% confidence interval?

The wider the interval the _____ the level of confidence. (Answer = “higher” or “lower”)

Example: 25 samples, each of size 1000 taken from a population with percentage = .55. The 25 resulting 95% confidence intervals.

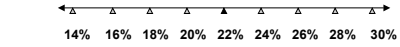


Another view of confidence intervals: For what values of the population percentage is the sample percentage likely or unlikely?



Would the 22% be likely if population percent was 14%?

16%?



18%?

19

20%?

22%?

20

24%?

26%?

21

A confidence interval consists of a collection of population percentages for which the sample percentage would be not too unlikely.

An Imaginary Conversation on the Meaning of a Confidence Interval

Statistician: From my simple random sample of 400, I estimate the population unemployment rate to be 22%.

Unemployed Person: I'm sure that I'm unemployed. Are you sure that 22% of the population is unemployed.

Statistician: No, I'm not sure.

Unemployed Person: What's the point of your doing a sample then?

Statistician: Well, I'm not sure, but I'm 95% confident that the percent unemployed is between 18% and 26%.

Unemployed Person: I'm 100% confident that I'm unemployed. What do you mean that you are 95% confident?

Statistician: I mean that the chance that an interval constructed in this way contains the population percentage is 95%.

Unemployed Person: Oh, I get it! There is a 95% chance that the population percentage is between 18% and 26%.

Statistician: No, that's not what I mean.

Unemployed Person: Say what?

Statistician: Think about it this way: The population percentage is a number, which we don't know. It is either between 18% and 26% or it isn't. There is no chance here, nothing random.

Unemployed Person: So what do you mean by "95% confident" then?

Statistician: I mean that in the long run, 95% of the intervals I construct in this way will contain the population percent. The intervals are random, but the population percent isn't.

Unemployed Person: In the "long run," we are all dead. You mean that the interval 18% to 26% is random?

Statistician: It's not random now, but it was before I took the sample. Like if I toss a coin: before I toss it, the outcome is random. But then once I've done it and get heads, say, there's nothing random anymore.

Unemployed Person: Oh, I get it. Like I'm in your sample, but before you took it I was like totally random. Now that you have taken the sample, I'm not random anymore. That really makes me feel a lot more secure, but I'm still unemployed. Would you mind taking the sample again?

Another Imaginary Conversation

Psychologist: Hey, you want to see some cool stats? I had 100 student volunteers and 21% of them could learn to wiggle their eyebrows in 3-4 time and simultaneously snap their fingers in 2-4 time. That gives me a confidence interval of 13% to 29%.

Statistician: What kind of a confidence interval is that?

Psychologist: You know, a 95% confidence interval. The plus or minus 2 SE kind.

Statistician: I get the algebra, but I don't get what it is a confidence interval for.

Psychologist: For the percent of people who can learn to wiggle their eyebrows in 3-4 time and simultaneously snap their fingers in 2-4 time.

Statistician: I don't know what people you are talking about. You didn't take a random sample from any population.

Psychologist: I think you're getting hostile. Look, statistics and probability are about chance, right?

Statistician: That's right.

Psychologist: And if I were to do my study again, I'd get different volunteers and the results wouldn't be exactly the same, right?

Statistician: I guess so.

Psychologist: So there is chance involved in my experiment and I'm using stats just like I learned from that expensive book. Stats are about chances, right?

Statistician: Right. But to make confidence intervals meaningful, you need a model for the chance error. Like the 100 people who you trained were a random sample from some big population.

Psychologist: Well, they are not really, but couldn't I pretend that they are a sample from all the people who could possibly have volunteered? Students, I mean.

Statistician: You can pretend anything you like.

Psychologist: OK. I think I'll call it a model for the chance error in my experiment.

Summary

- With a simple random sample, the sample percentage is used to estimate the population percentage.
- The sample percentage will be off by some amount---it's chance error.
- The SE of the sample percentage measures how large the chance error is likely to be

- Since we don't know the population percentage, we can use the sample percentages to estimate the SD of "the box" and the SE of the sample percentage.
- A confidence interval is the sample percentage plus and minus a certain number of SE's. The confidence level is found from the normal curve.
- A confidence statement is not a statement of probability.

? A simple random sample of 400 dwelling units is taken from a large population of units. In the sample, 22% of the units had a single occupant. What is a 90% confidence interval for the percentage of all the dwelling units which have a single occupant?

Approximate SE:

? In a simple random sample of 100 businesses in a city, 20% offered free parking to their employees. Of the 900 employees of these businesses, 45% commuted to work by private automobiles. If possible, find (1) a 90% confidence interval for the percentage of businesses in the city offering free parking to their employees and also (2) a 90% confidence interval for the percentage of workers in the city who commute to work by private automobile. If the methods in the text do not apply to either one or both of these, explain why in a single sentence.