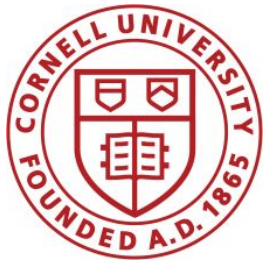


DSFA

Spring 2020

Lecture 21

Designing Experiments

Questions from the Past Week

- How can we quantify natural concepts like “center” and “variability”?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

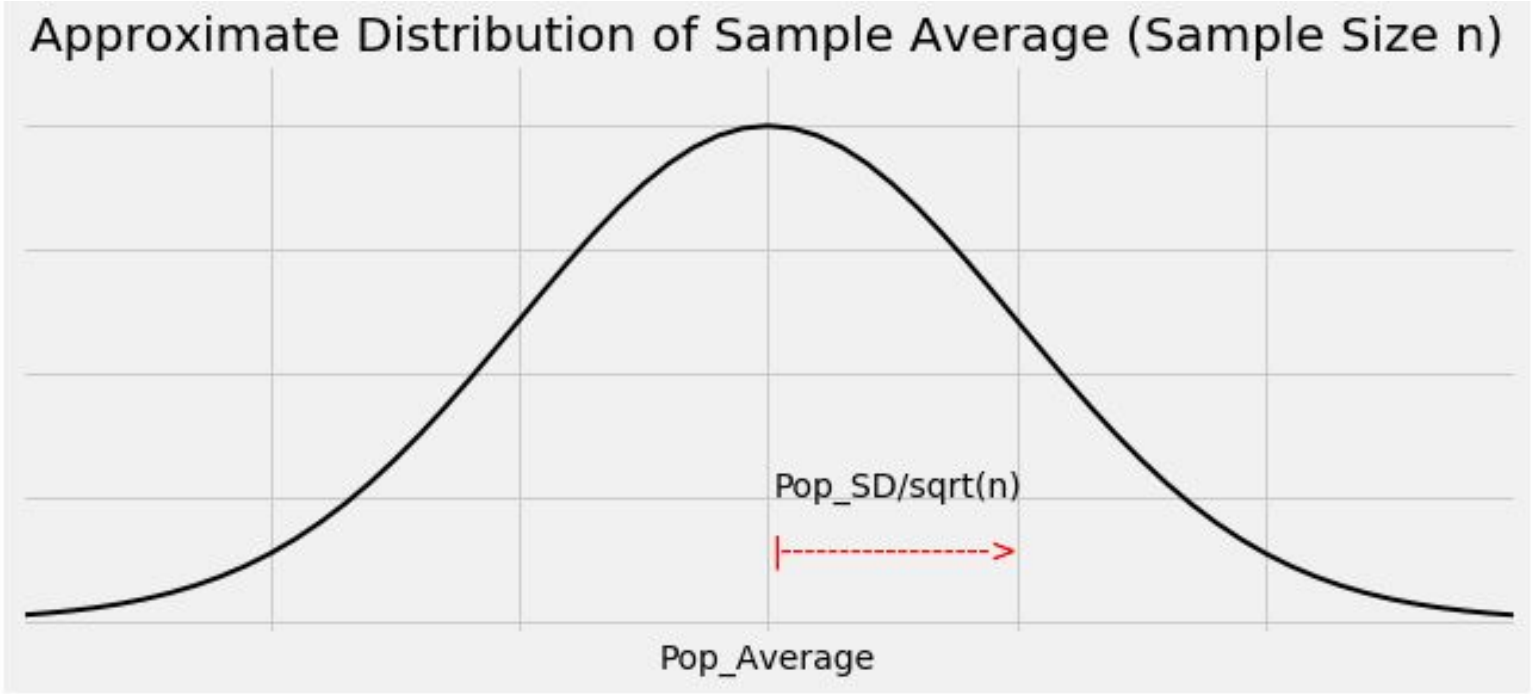
Distribution of the Sample Average

- Fix a large sample size.
- Draw all possible random samples of that size.
- Compute the average of each sample.
- You'll end up with a lot of averages.
- The distribution of those is called the *distribution of the sample average.*
- It's roughly normal, centered at the population average.
- SD = (population SD) / $\sqrt{\text{sample size}}$

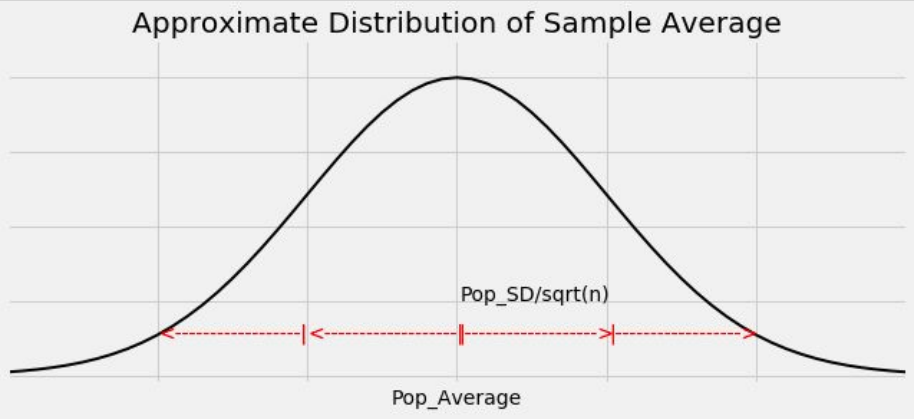
(Demo)

Confidence Intervals

Graph of the Distribution



The Key to 95% Confidence



- For about 95% of all samples, the sample average and population average are within **2 SD**s of each other.
- **SD** = SD of sample average

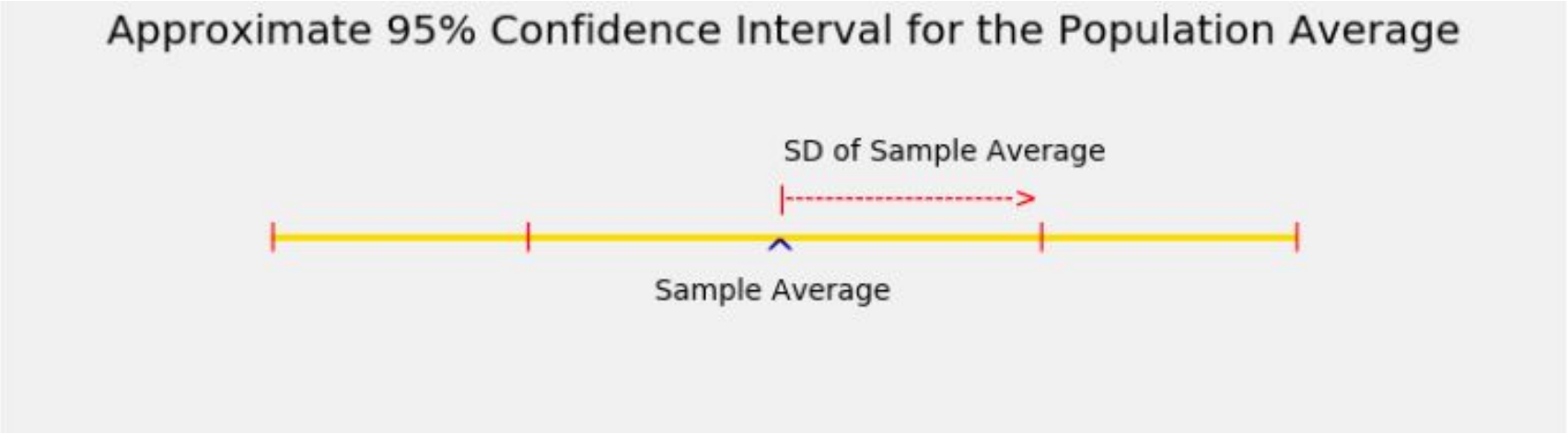
  = (population SD) / $\sqrt{\text{sample size}}$

Constructing the Interval

For 95% of all samples,

- If you stand at the population average and look two **SD**s on both sides, you will find the sample average.
- Distance is symmetric.
- So if you stand at the sample average and look two **SD**s on both sides, you will capture the population average.

The Interval



Width of the Interval

Total width of a 95% confidence interval for the population average

= 4 * SD of the sample average

= 4 * (population SD) / $\sqrt{\text{sample size}}$

Sample Proportions

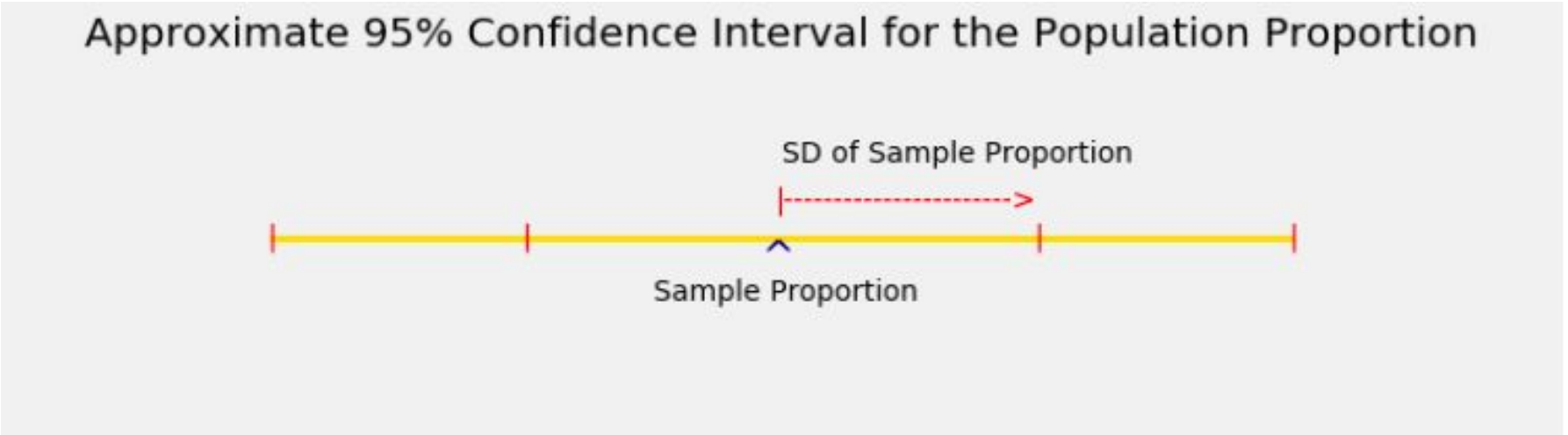
Proportions are Averages

- Data: 0 1 0 0 1 0 1 1 0 0 (10 entries)
- Sum = 4 = number of 1's
- Average = 4/10 = 0.4 = proportion of 1's

If the population consists of 1's and 0's (yes/no answers to a question), then:

- the population average is the proportion of 1's in the population
- the sample average is the proportion of 1's in the sample

Confidence Interval



Controlling the Width

- Total width of an approximate 95% confidence interval for a population proportion

  $= 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$

- The narrower the interval, the more accurate your estimate.
- Suppose you want the total width of the interval to be no more than 3%. How should you choose the sample size?

The Sample Size for a Given Width

$$0.03 = 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$$

- Left hand side is 3%, the maximum total width that you will accept
- Right hand side is the formula for the total width

$$\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.03$$

(Demo)

“Worst Case” Population SD

- $\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.03$
- SD of 0/1 population is at most 0.5
- $\sqrt{\text{sample size}} \geq 4 * 0.5 / 0.03$
- $\text{sample size} \geq (4 * 0.5 / 0.03) ** 2 = 4444.44$
- The sample size should be 4445 or more

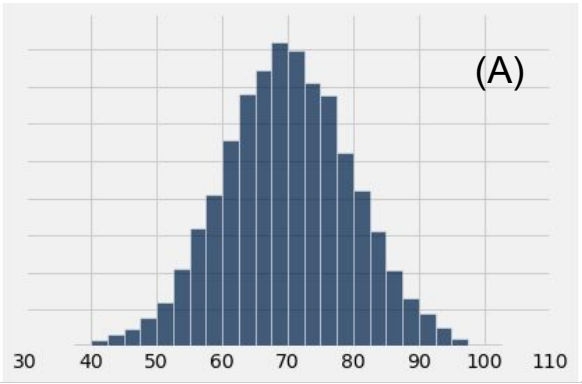
Discussion Question

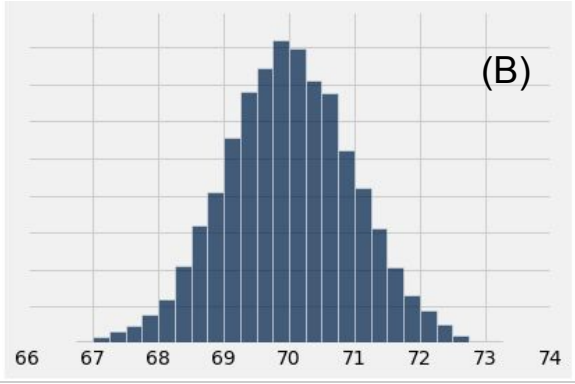
A city has 200,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. The distribution of the incomes [pick one and explain]:

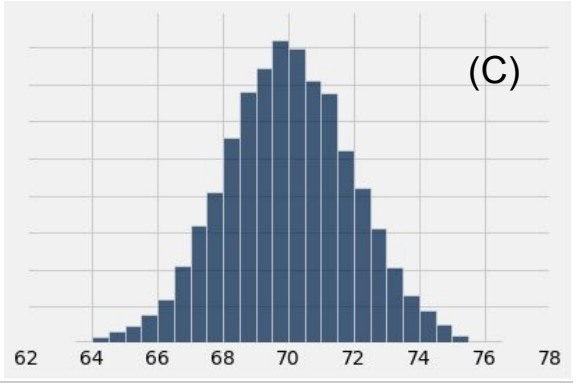
(a) is roughly normal because the number of households is large.
(b) is not close to normal.
(c) may be close to normal, or not; we can't tell from the information given.

Discussion Question

A population has average 70 and SD 10. One of the histograms below is the empirical distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one?







Discussion Question

- I am going to use a 68% confidence interval to estimate a population proportion.
- I want the total width of my interval to be no more than 2.5%.
- How large must my sample be?