190F Foundations of Data Science

Fall 2018

Lecture 26

Designing Experiments

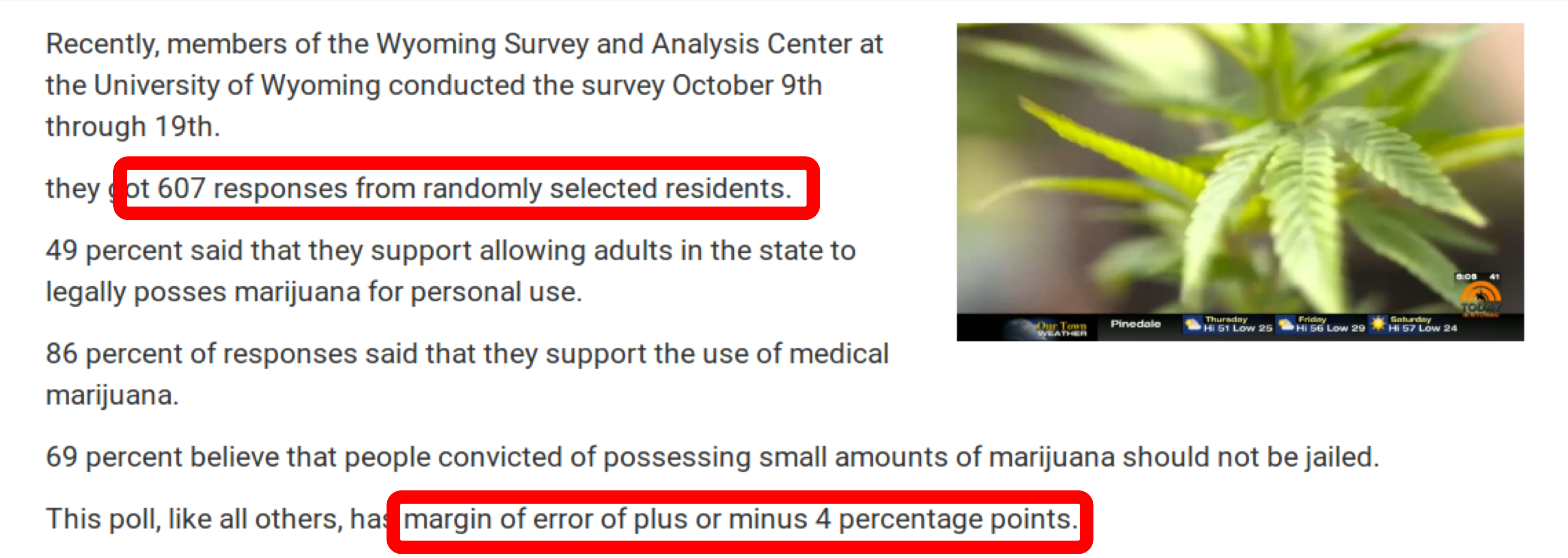
Questions we are studying

- 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?

How do they know?



Recently, members of the Wyoming Survey and Analysis Center at the University of Wyoming conducted the survey October 9th through 19th.

they got 607 responses from randomly selected residents.

49 percent said that they support allowing adults in the state to legally posses marijuana for personal use.

86 percent of responses said that they support the use of medical marijuana.

69 percent believe that people convicted of possessing small amounts of marijuana should not be jailed.

This poll, like all others, has margin of error of plus or minus 4 percentage points.

Variability of the Sample Average

- The distribution of all possible sample averages of a given size is called the *distribution of the sample average.*
- We approximate it by an empirical distribution.
- By the CLT, it's roughly normal:
  - Center = the population average
  - SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question

A city has 500,000 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 city has 500,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. A random sample of 900 households is taken.

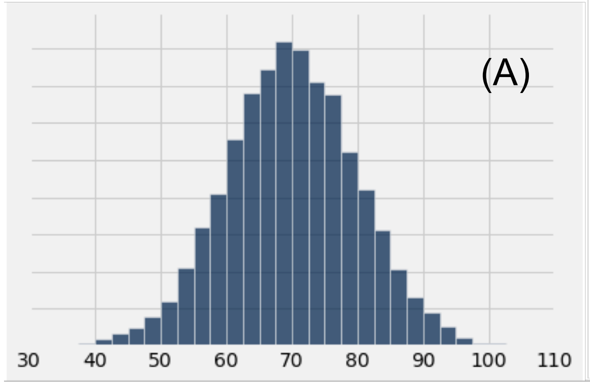
Fill in the blanks and explain:

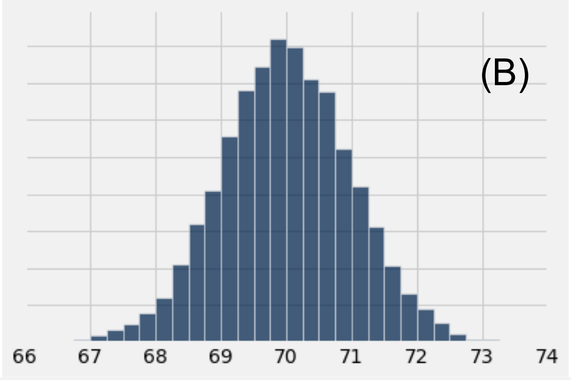
There is about a 68% chance that the average annual income of the sampled households is in the range

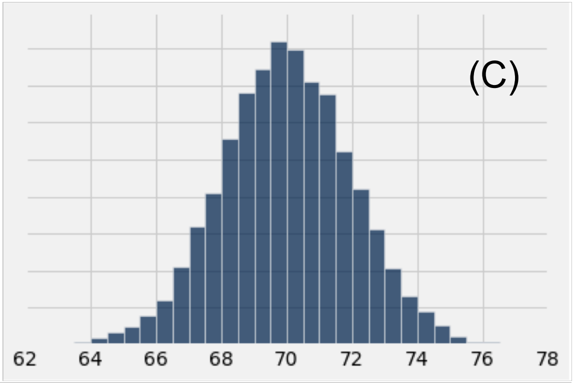
$___________ plus or minus $___________

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?

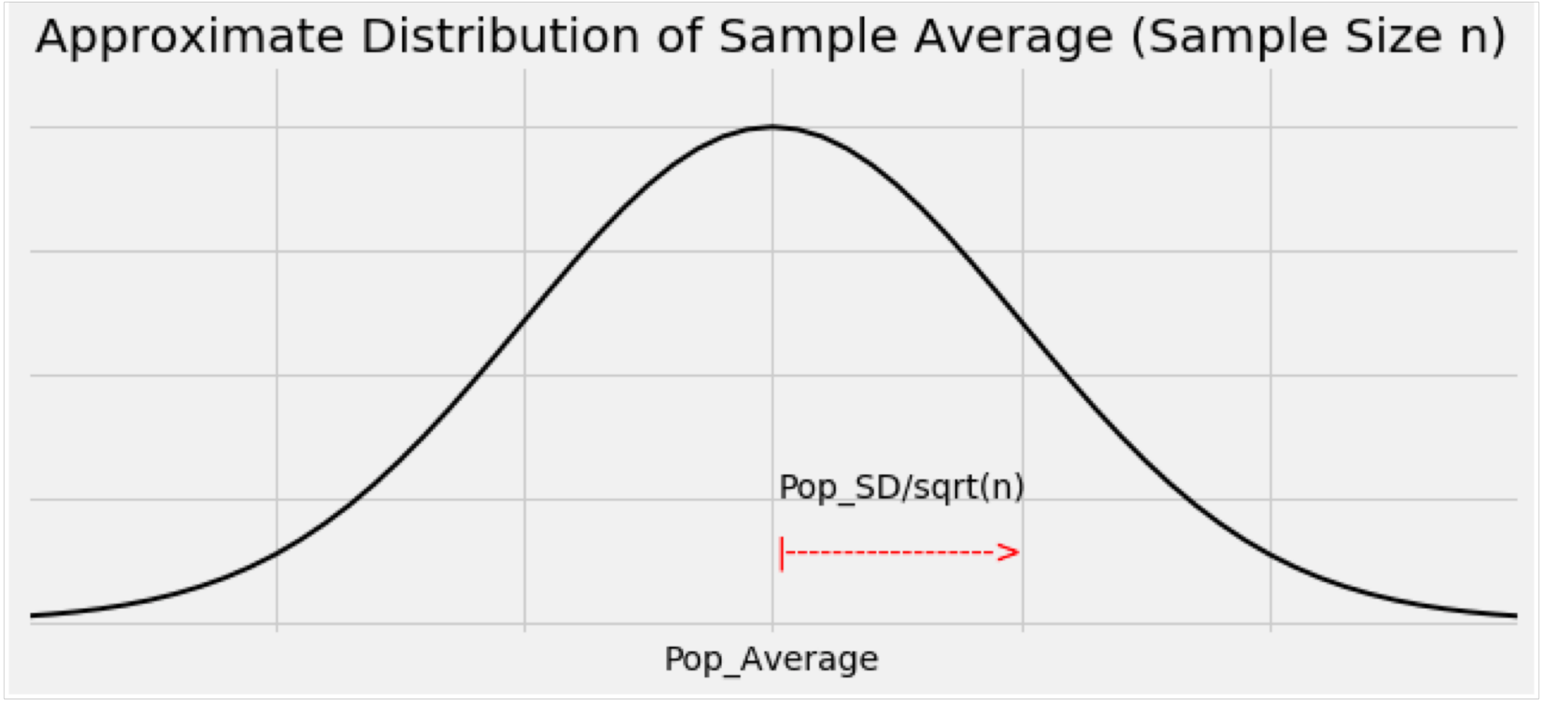




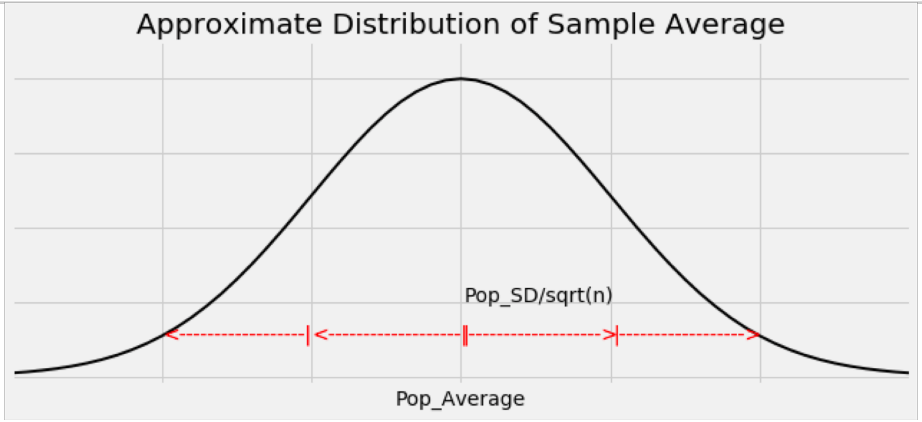


Confidence Intervals

Graph of the Distribution



The Key to 95% Confidence



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

- **SD** = SD of sample average

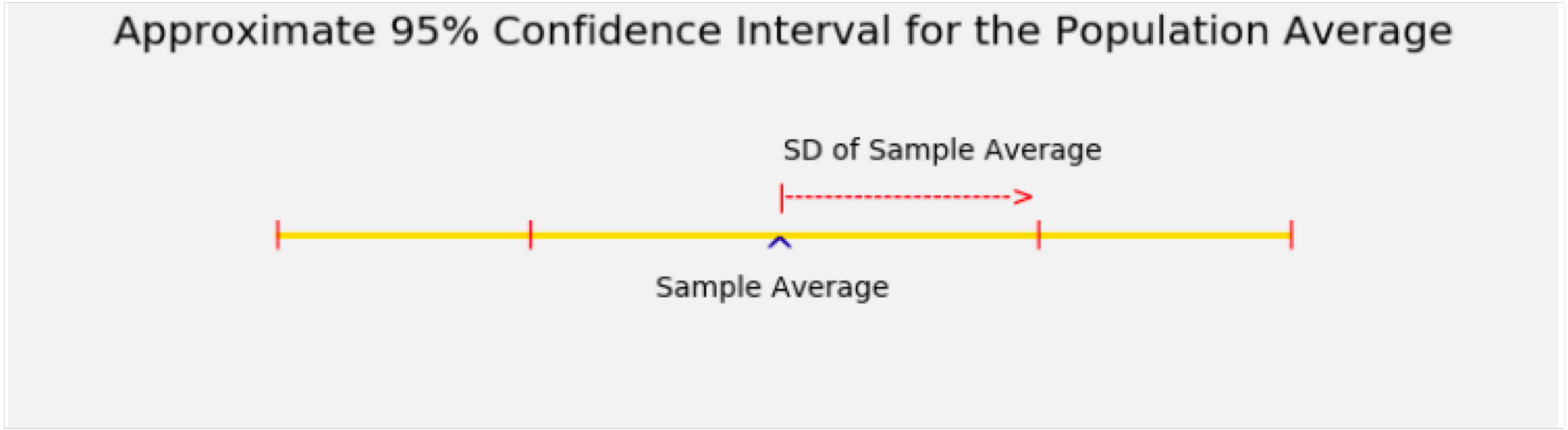
  = (population SD) / $\sqrt{\text{sample size}}$

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


Approximate 95% Confidence Interval for the Population Average
SD of Sample Average
Sample Average

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

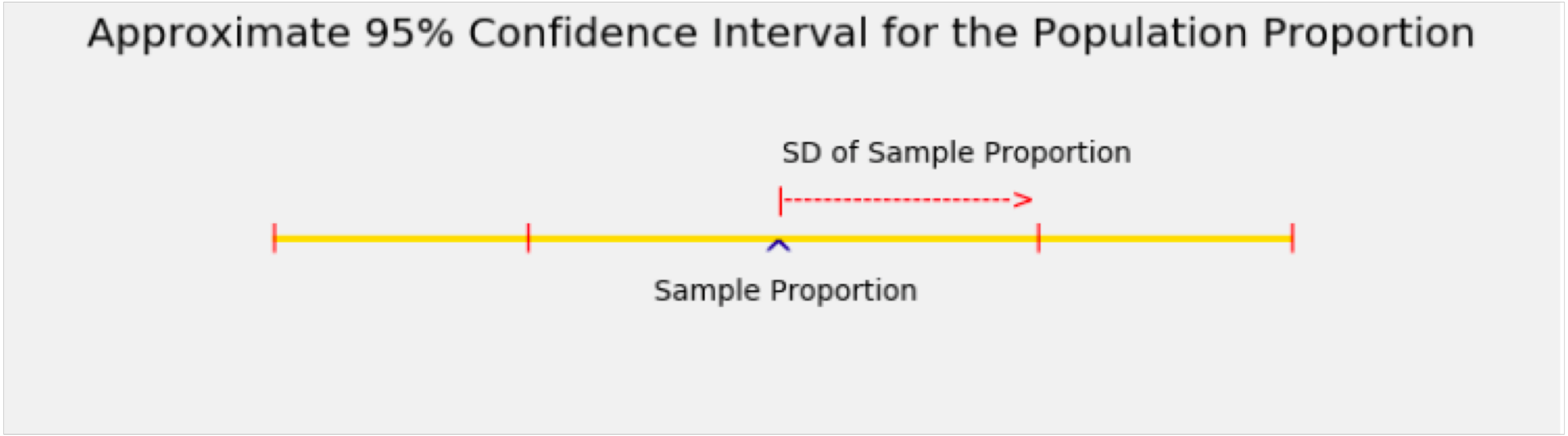
Proportions are Averages

- 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



Approximate 95% Confidence Interval for the Population Proportion

Controlling the Width

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

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

Discussion Question

- A researcher is estimating a population proportion based on a random sample of size 10,000.

Fill in the blank with a decimal:

- With chance at least 95%, the estimate will be correct to within ____________________.

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 random sample be?

$2 * (0.5) / \mathrm{sqrt}(\text{sample size}) = 0.025$