

YData: An Introduction to Data Science

Lecture 28: Designing Experiments

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org



Announcements

- Homework assignment 07 grades were posted
- Project 2 checkpoint is due tonight
- Homework assignment 08 is due Monday, 4/12
- Homework assignment 09 will be published shortly – due next Thursday, 4/15

Questions for This Week

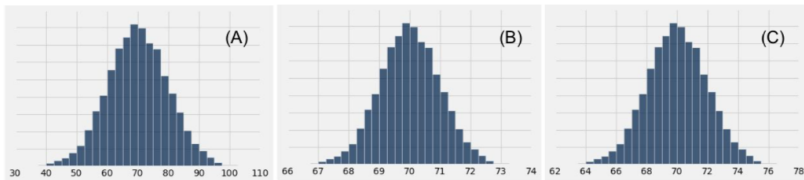
- How can we quantify natural concepts like “center” and “variability”?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

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

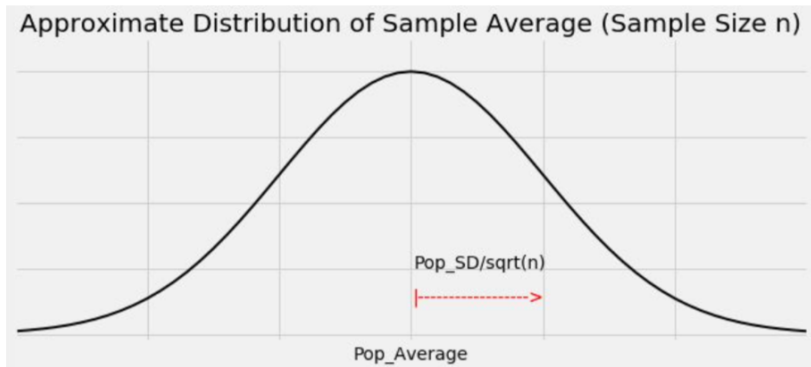
Discussion Question

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?

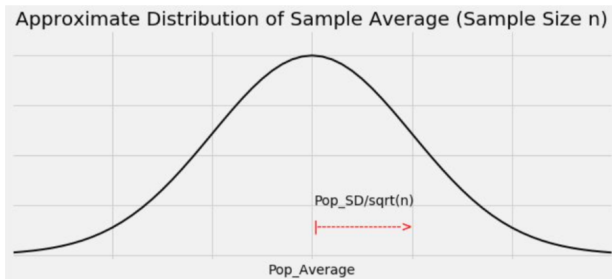


Confidence Intervals

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

The Interval

Approximate 95% Confidence Interval for the Population Average



Width of the Interval

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

$$= 4 * \text{SD of the sample average}$$

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

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
$$= 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$$
- The narrower the interval, the more accurate your estimate.
- Suppose you want the total width of the interval to be no more than 0.03. 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 0.03, 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 researcher is estimating a population proportion based on a random sample of size 10,000.

Fill in the blank with a decimal:

- With about 95% confidence, 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 0.025.
- How large must my random sample be?