DSFA

Spring 2018

Lecture 27

Designing Experiments

Announcements

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

- The distribution of those is called the *distribution of the sample average.*

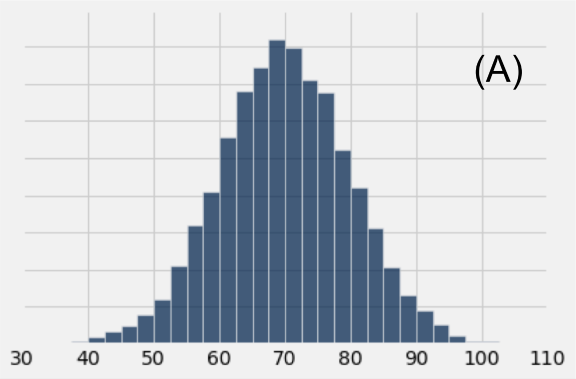
- It’s roughly normal, centered at the population average.

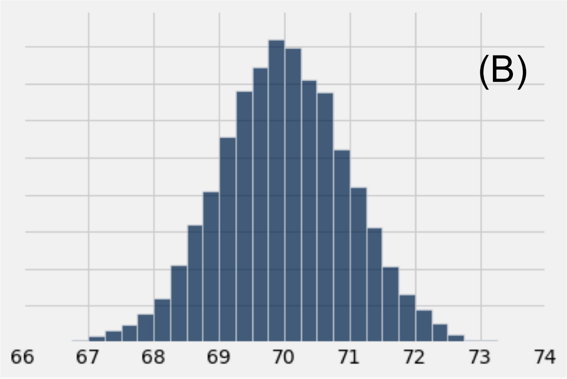
- SD of the sample average = (population SD) $/\sqrt{\text{sample size}}$

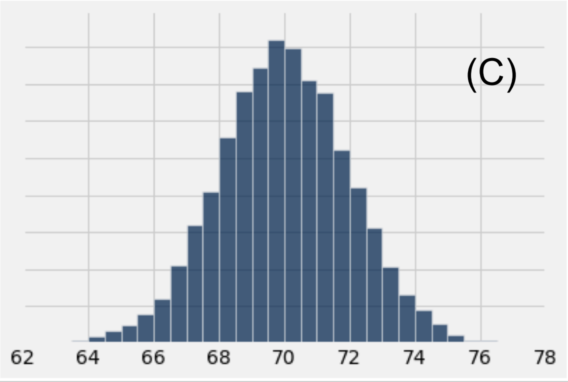
  -- this is often called the sample standard error

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?



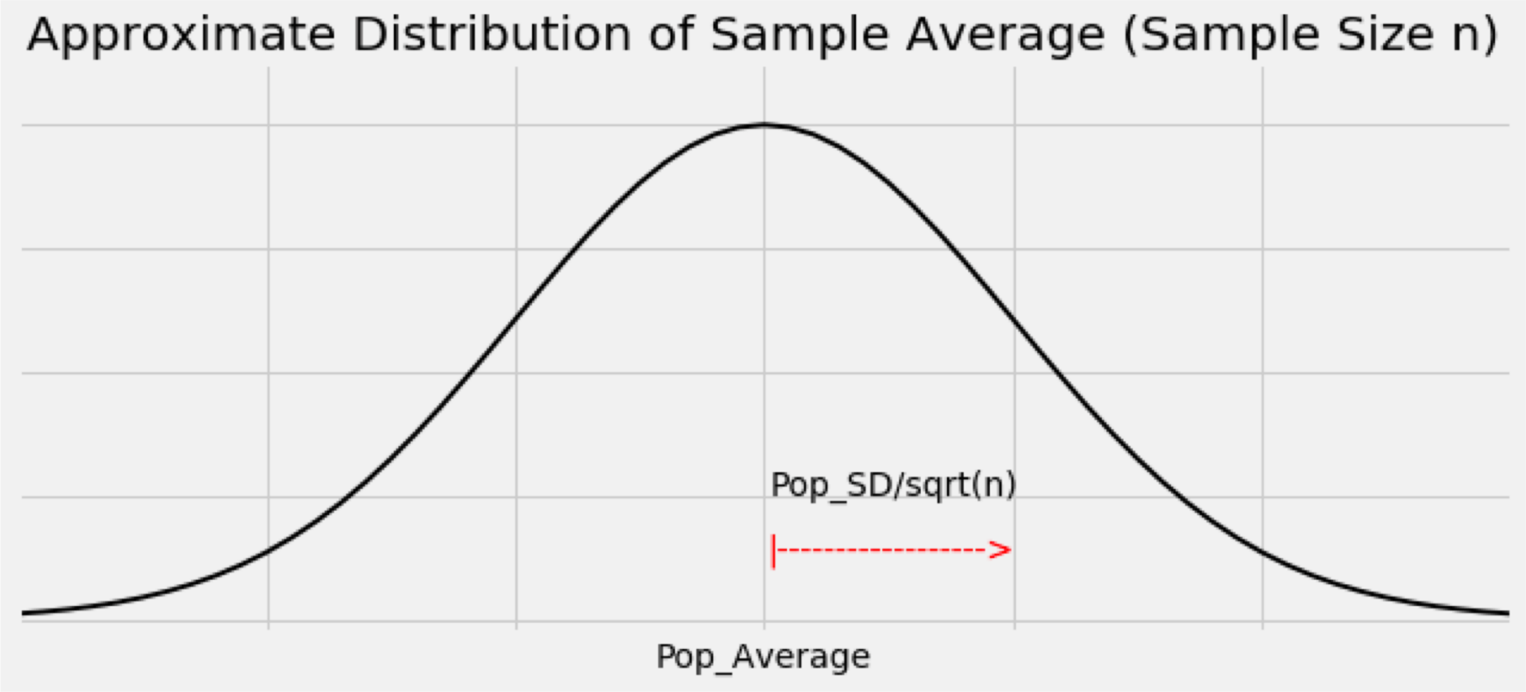




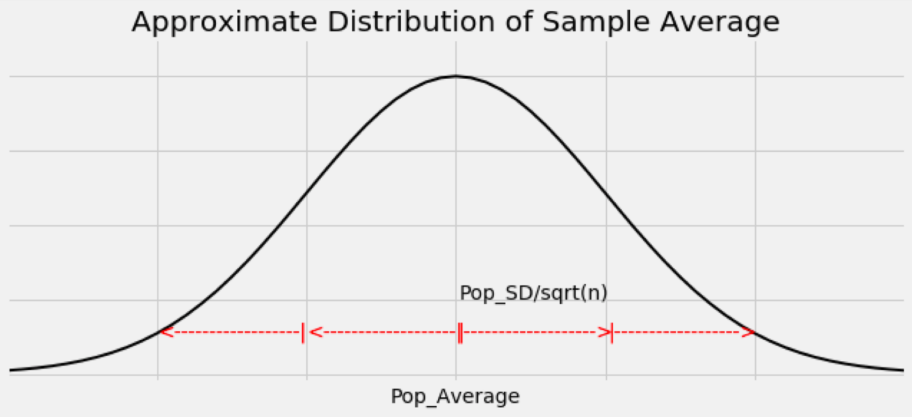
(Demo)

Confidence Intervals
(take 2)

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) / $\sqrt{\text{sample size}}$

The Key Idea for a 95% Confidence Interval for a given sample of size (*n*)

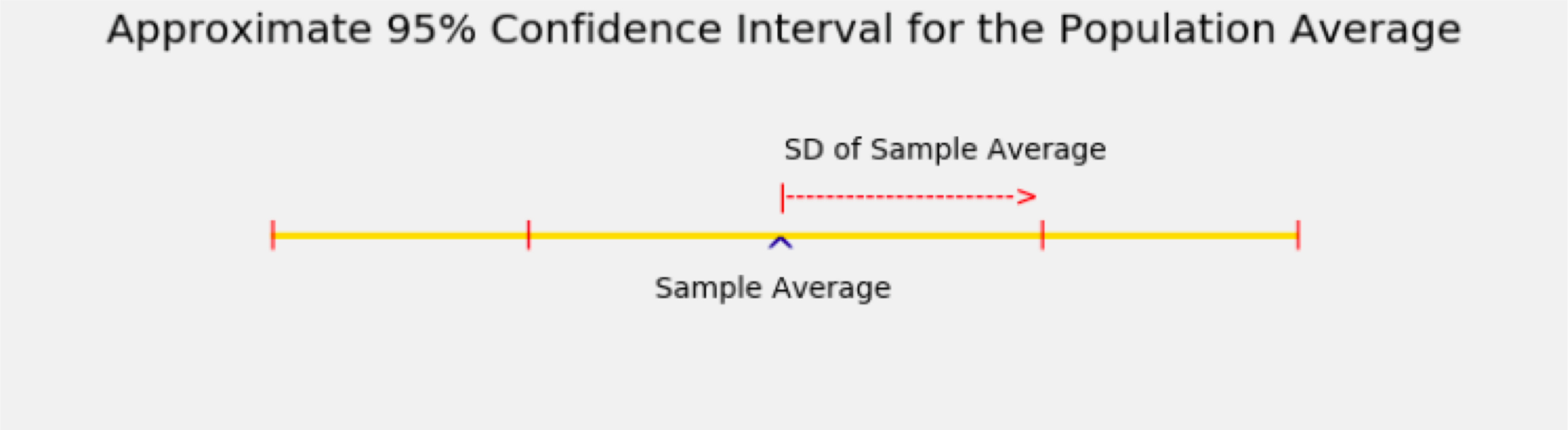
$$Prob\left(-2 \leq \frac{Sample\ Average - Population\ Average}{\frac{Pop\ SD}{\sqrt{n}}} \leq 2\right) = .95$$

$$Prob\left(-2 * \frac{Pop\ SD}{\sqrt{n}} \leq Sample\ Average - Population\ Average\ \ \leq 2 * \frac{Pop\ SD}{\sqrt{n}}\right) = .95$$

$$Prob\left(Sample\ Average - 2 * \frac{Pop\ SD}{\sqrt{n}} \leq\ Population\ Average\ \ \leq Sample\ Average + 2 * \frac{Pop\ SD}{\sqrt{n}}\right) = .95$$

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

The Interval


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

Sample Average ± 2*(SD of 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) $/\sqrt{\text{sample size}}$

Width of the Interval and Margin of Error

The width of a 95% confidence interval for the population average is about 4 * (Population SD) /$\sqrt{\text{sample size}}$.

Suppose we fix the total width of the interval to be 2*DELTA (DELTA is called the margin of error). Then

$$\text{DELTA} = 2*(\text{Population SD}) / \sqrt{\text{sample size}}$$

➡ sample size = 4*(Population SD)^2 / DELTA^2 (Demo)

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

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

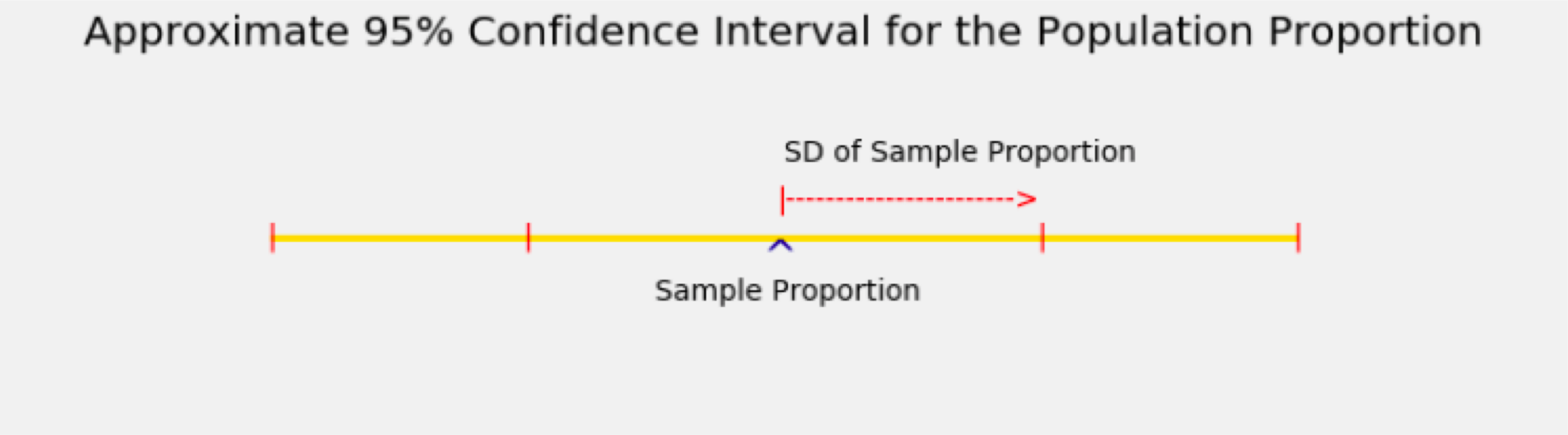
If the population consists of 1’s and 0’s (yes/no answers to a question), then:

- the population SD is the

$$\sqrt{(\text{proportion of 1's}) * (1 - \text{proportion of 1's})}$$

(Demo)

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) /$\sqrt{\text{sample size}}$

$$4 * \sqrt{(proportion\ of\ 1's) * (1 - proportion\ of\ 1's)} / \sqrt{sample\ \ size}$$

- The narrower the interval, the more accurate your estimate.

- Suppose you want the total width of the interval to be no more than 3% (a margin of error of 1.5%). 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% (a margin of error of 1.5%), the maximum total width that you will accept

- Right hand side is the formula for the total width

$$\text{Sample size} = [4 * (\text{SD of 0/1 population}) / 0.03]^2$$

$$\text{Sample size} = [2 * (\text{SD of 0/1 population}) / 0.015]^2$$

“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

“Better Case” Population SD

- Suppose from prior studies the proportion is around 25%.
- SD of 0/1 population is then
- $\sqrt{\text{proportion} * (1 - \text{proportion})} = \sqrt{.25 * (.75)}$ = 0.433
- sample size = $(4*(0.433) / 0.03)^2$ = 3333.33
- The sample size should be 3334 or more

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?