Data 8

Summer 2018

Lecture 25

Sample Means and Designing Experiments

Contributions by Vinitra Swamy (vinitra@berkeley.edu) and Fahad Kamran (fhdkmrn@berkeley.edu)
Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)

Announcements

Questions for This Lecture

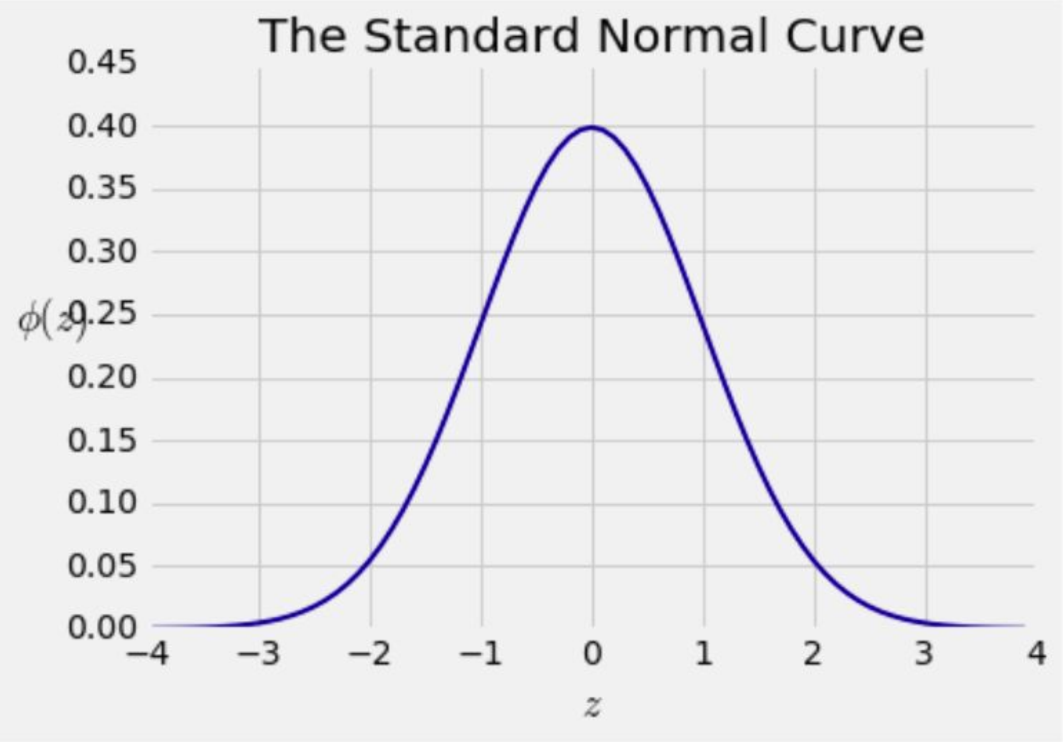
- How can we quantify natural concepts like “center” and “variability”?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

The Standard Normal Curve

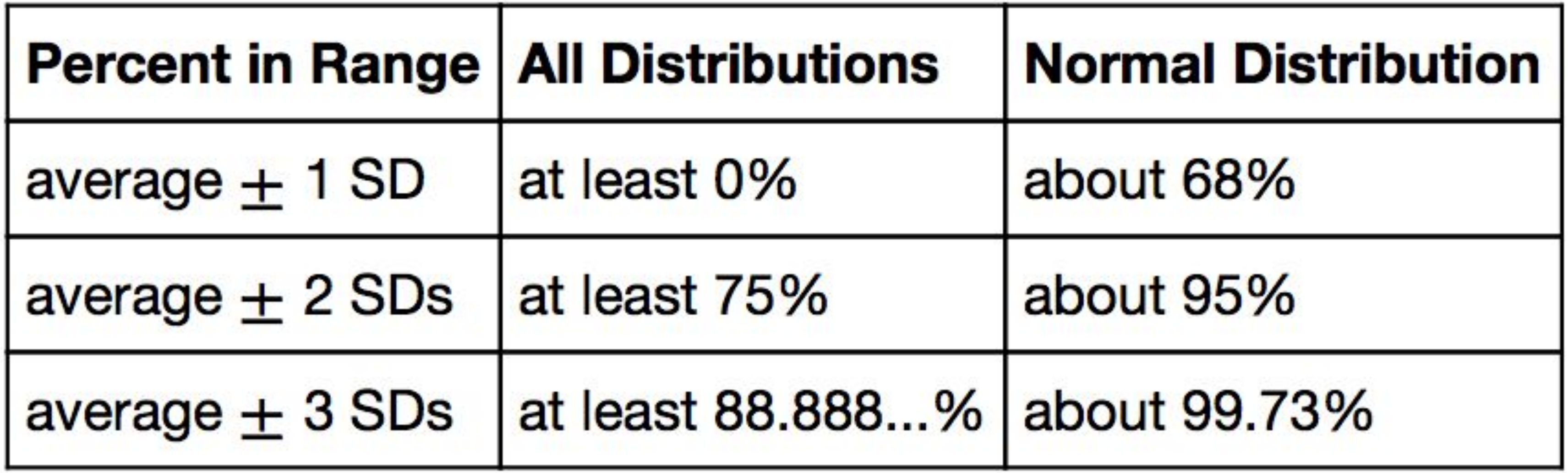
A beautiful formula that **we won’t use at all**:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$\phi(z)$
−4
−3
−2
−1
0
1
2
3
4
z

Bounds and Normal Approximations



Percent in Range	All Distributions	Normal Distribution
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

Sample Averages

- Central Limit Theorem
  - Describes bell-shaped distributions (normal curve) in the context of random sampling
- Many distributions we observed were **not bell-shaped**.
  - Empirical distributions of their sample averages are!
- Why do we care about sample averages?
  - They estimate population averages.

Distribution of the Sample Average

Why is There a Distribution?

- You have only one random sample, and it has only one average.
- But **the sample could have come out differently**.
- And then the sample average might have been different.
- So there are many possible sample averages.

Distribution of the Sample Average

- Imagine all possible random samples of the same size as yours. There are lots of them.

- Each of these samples has an average.

- The **distribution of the sample average** is the distribution of the averages of all the possible samples.

Shape of the Distribution

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the distribution of the sample sum
(or of the sample average) is roughly bell-shaped

(Demo)

Specifying the Distribution

Suppose the random sample is large.

- We have seen that the distribution of the sample average is roughly bell shaped.

- Important questions remain:
  - Where is the center of that bell curve?
  - How wide is that bell curve?

Center of the Distribution

The Population Average

The distribution of the sample average is roughly a bell curve centered at the population average.

Variability of the Sample Average

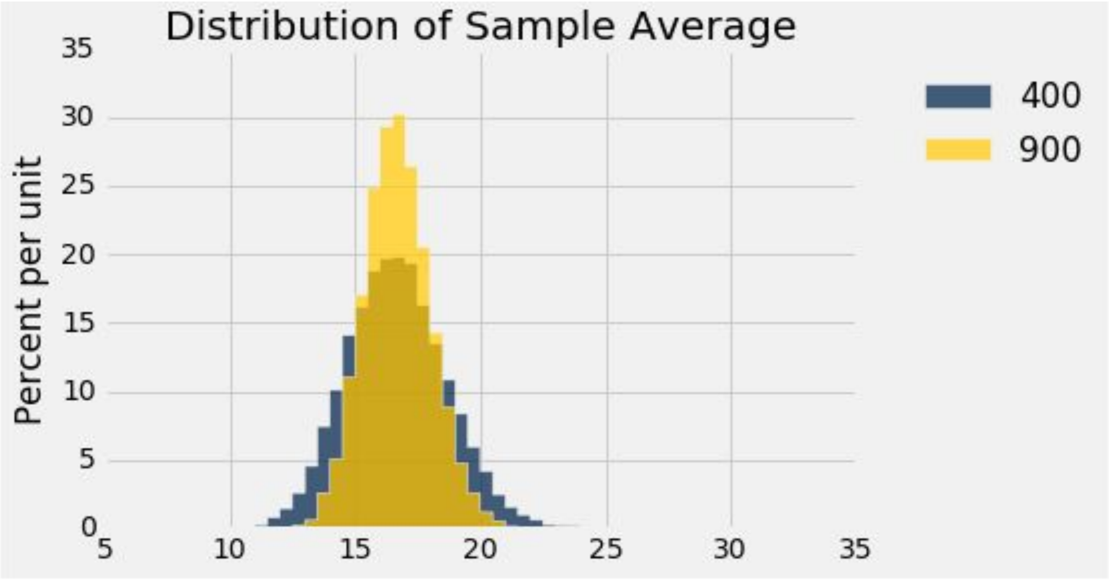
Why Is This Important?

- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding the variability of the sample mean helps us work out how large our sample has to be. (Demo)

Discussion Question

The gold histogram shows the distribution of ___________ values, each of which is ___________________________.

(a) 900
(b) 10,000
(c) a randomly sampled flight delay
(d) an average of flight delays



The Two Histograms

- The gold histogram shows the distribution of 10,000 values, each of which is an average of 900 randomly sampled flight delays.
- The blue histogram shows the distribution of 10,000 values, each of which is an average of 400 randomly sampled flight delays.
- Both are roughly bell shaped.
- The larger the sample size, the narrower the bell.

(Demo)

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 $____________

Break

Revisiting Questions

- 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?

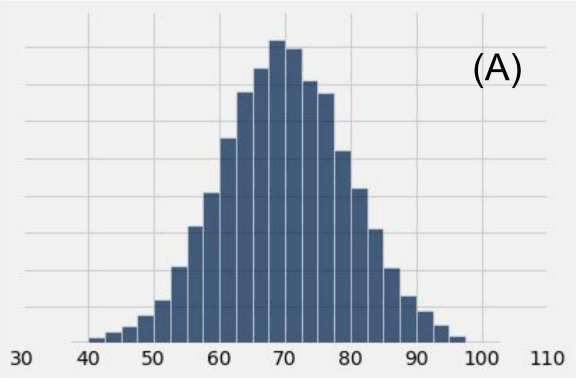
Distribution of the Sample Average

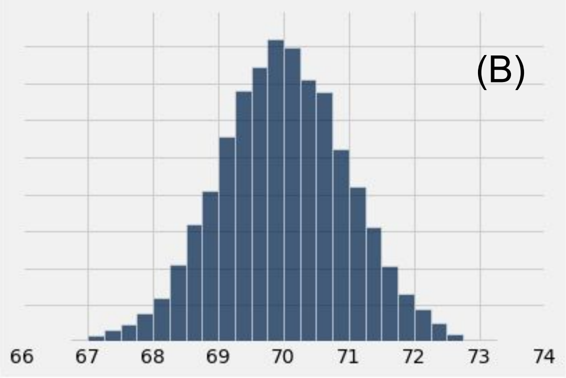
- Fix a large sample size.
- Draw all possible random samples of that size.
- Compute the average of each sample.
- You'll end up with a lot of averages.
- The distribution of those is called the *distribution of the sample average.*
- It's roughly normal, centered at the population average.
- SD = (population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

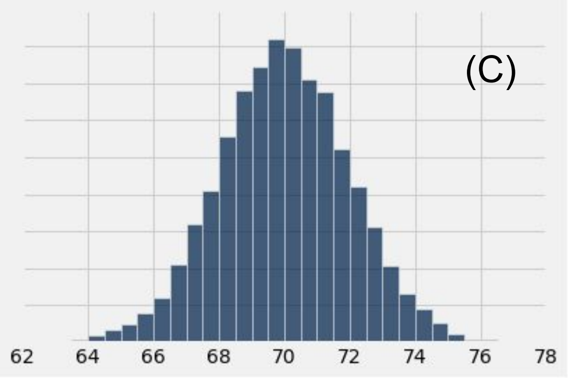
(Demo)

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?

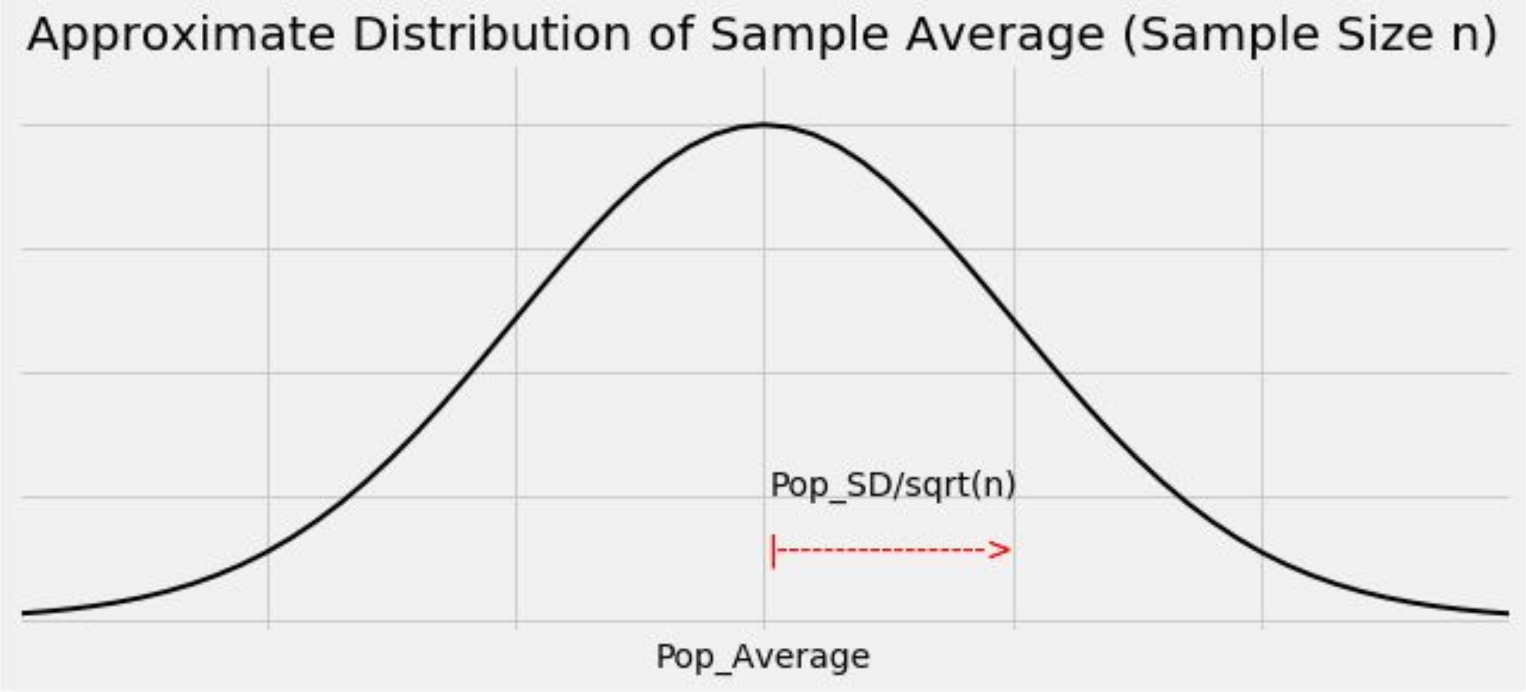




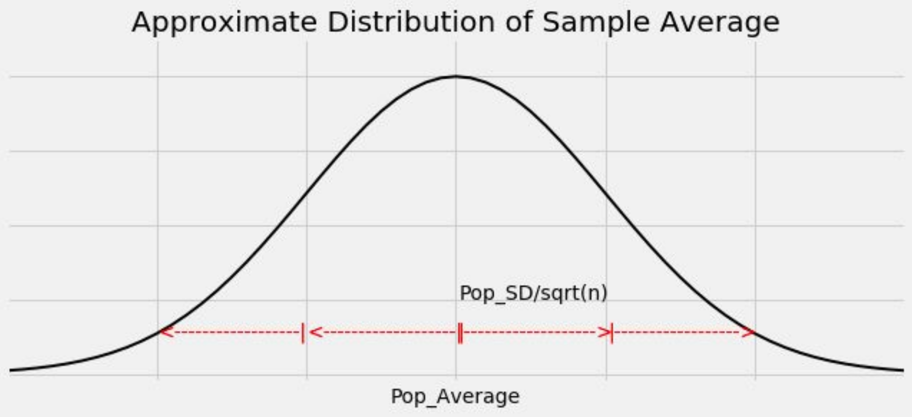


Confidence Intervals

Graph of the Distribution



The Key to 95% Confidence



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

- **Sample SD** = SD of sample average

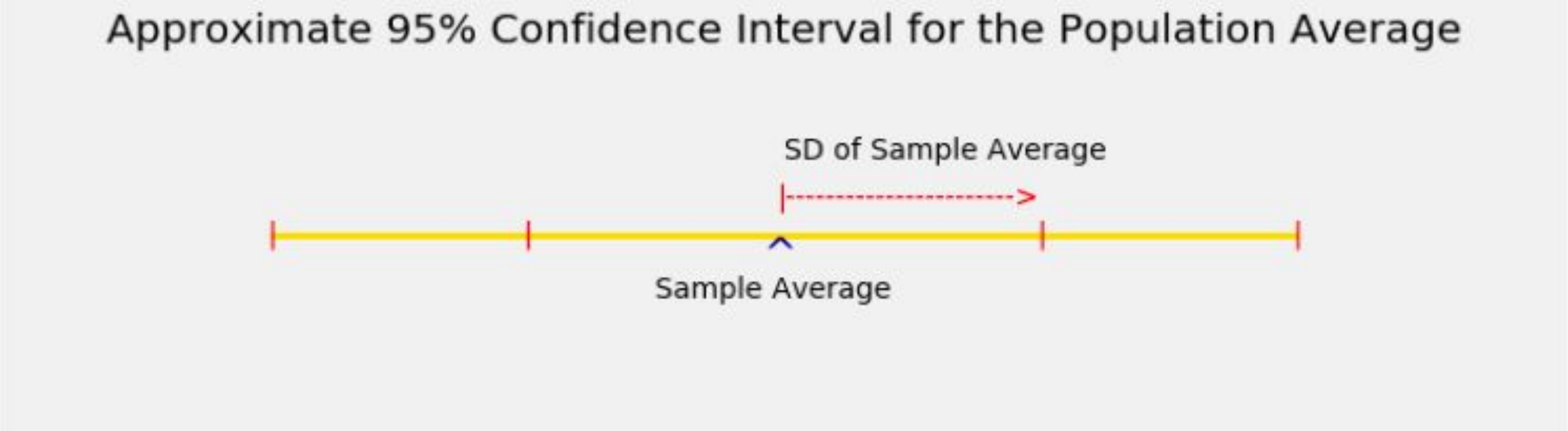
  = (population SD) / $\sqrt{\text{sample size}}$

Constructing the Interval

For 95% of all samples,

- If you stand at the population average and look two **Sample SDs** on both sides, you will find the sample average.
- Distance is symmetric.
- If you stand at the sample average and look two **Sample SDs** 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

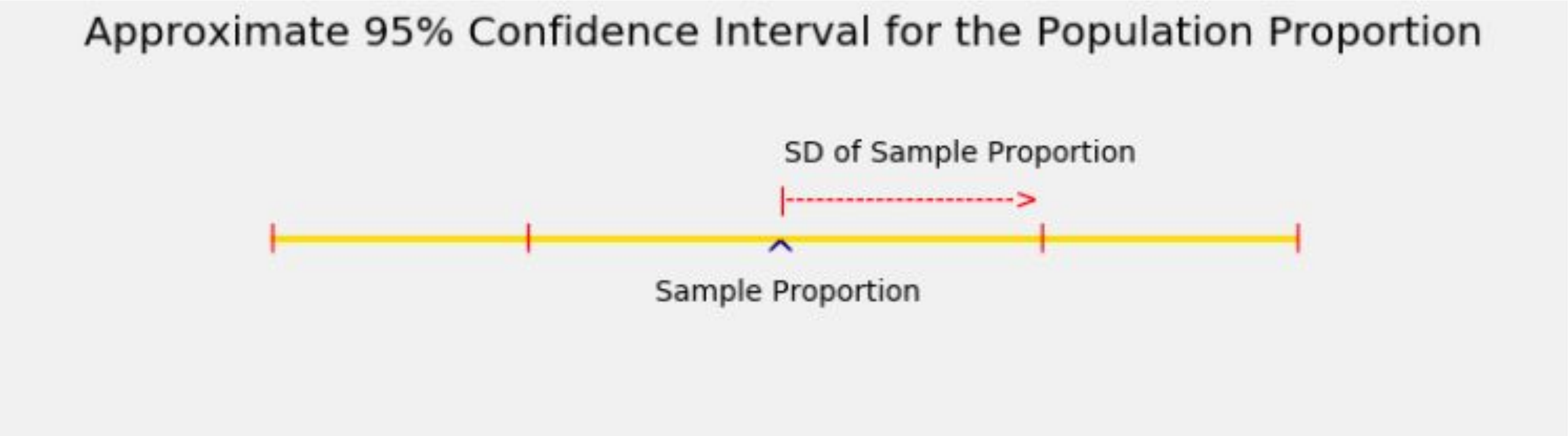
Proportions are Averages

- 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
SD of Sample Proportion
Sample Proportion

Controlling the Width

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

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

- 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

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

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

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

(Demo)

“Worst Case” Population SD

- $\sqrt{\text{sample size}}$ = 4 * (population SD) / 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: Why 10,000?

- 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

- Vinitra is going to use a 68% confidence interval to estimate a population proportion.
- Vinitra wants the total width of her interval to be no more than 2.5%.
- How large must her random sample be?