MODULE 8

Why the mean matters

CENTER AND SPREAD

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?

AVERAGE

THE AVERAGE OR MEAN

Data: 2, 3, 3, 9 Average = (2+3+3+9)/4 = 4.25

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- **Smoothing operator**: collect all the contributions in one big pot, then split evenly

(Demo – notebook 8.1,

Average(Mean))

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

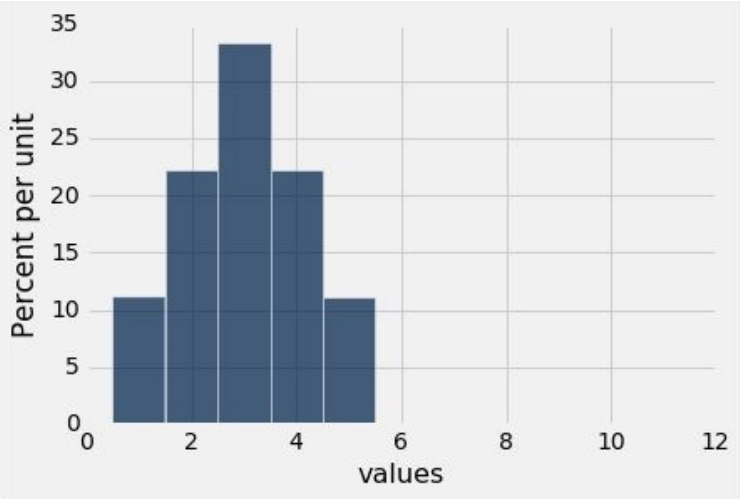
If the population consists of 1's and 0's (yes/no answers to a question), then:

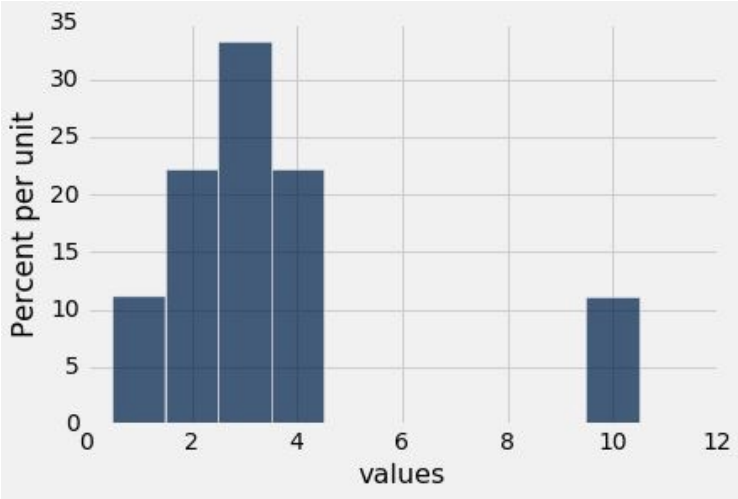
- the population average is the proportion of 1's in the population
- the sample average is the proportion of 1's in the sample

(Demo – notebook 8.1, Average(Mean))

DISCUSSION QUESTION

Are the medians of these two distributions the same or different? Are the means the same or different? If you say “different,” then say which one is bigger.





COMPARING MEAN AND MEDIAN

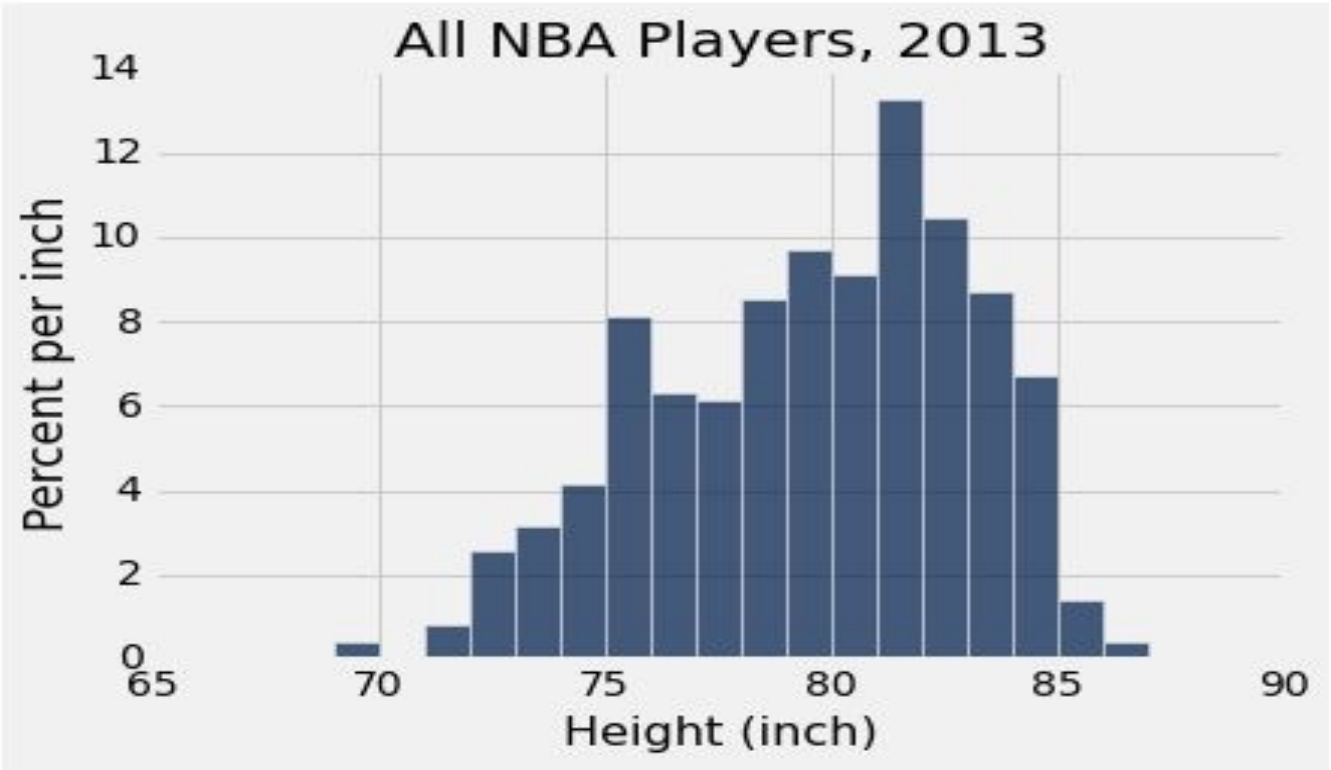
- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is **symmetric about a value**, then that value is both the average and the median.
- If the histogram is **skewed**, then the mean is pulled away from the median in the direction of the tail.

DISCUSSION QUESTION

Which is bigger?

(a) mean

(b) median



(Demo – Notebook 8.1, Discussion Question)

STANDARD DEVIATION

DEFINING VARIABILITY

Plan A: "biggest value - smallest value"

- Doesn't tell us much about the shape of the distribution

Plan B:

- Measure variability around the mean
- Need to figure out a way to quantify this

(Demo – Notebook 8.1, Standard Deviation)

HOW FAR FROM THE AVERAGE?

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average

  5 4 3 2 1

- SD has the same units as the data

WHY USE THE SD?

There are two main reasons.

- **The first reason:**

No matter what the shape of the distribution,
the bulk of the data are in the range "average ± a few SDs"

- **The second reason:**

We can quantify how much data is represented within a few SDs, i.e., Chevyshev's Bounds.

CHEBYSHEV'S INEQUALITY

HOW BIG ARE MOST OF THE VALUES?

No matter what the shape of the distribution,

the bulk of the data are in the range "average ± a few SDs"

Chebyshev's Inequality

No matter what the shape of the distribution,

the proportion of values in the range "average ± z SDs" is

at least $1 - 1/z^2$

CHEBYSHEV'S BOUNDS

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888…%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

No matter what the distribution looks like

(Demo – Notebook 8.2, Chebyshev's Bounds)

STANDARD UNITS

STANDARD UNITS

- How many SDs above average?
- **z = (value - average)/SD**
  - Negative z: value below average
  - Positive z: value above average
  - z = 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Chebyshev: At least 96% of the values of z are between -5 and 5

(Demo – Notebook 8.2, Standard Units)

DISCUSSION QUESTION

Find whole numbers that are close to:

(a) the average age

(b) the SD of the ages

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

... (1164 rows omitted)

THE SD AND THE HISTOGRAM

- Usually, it's not easy to estimate the SD by looking at a histogram.

- But if the histogram has a bell shape, then you can.

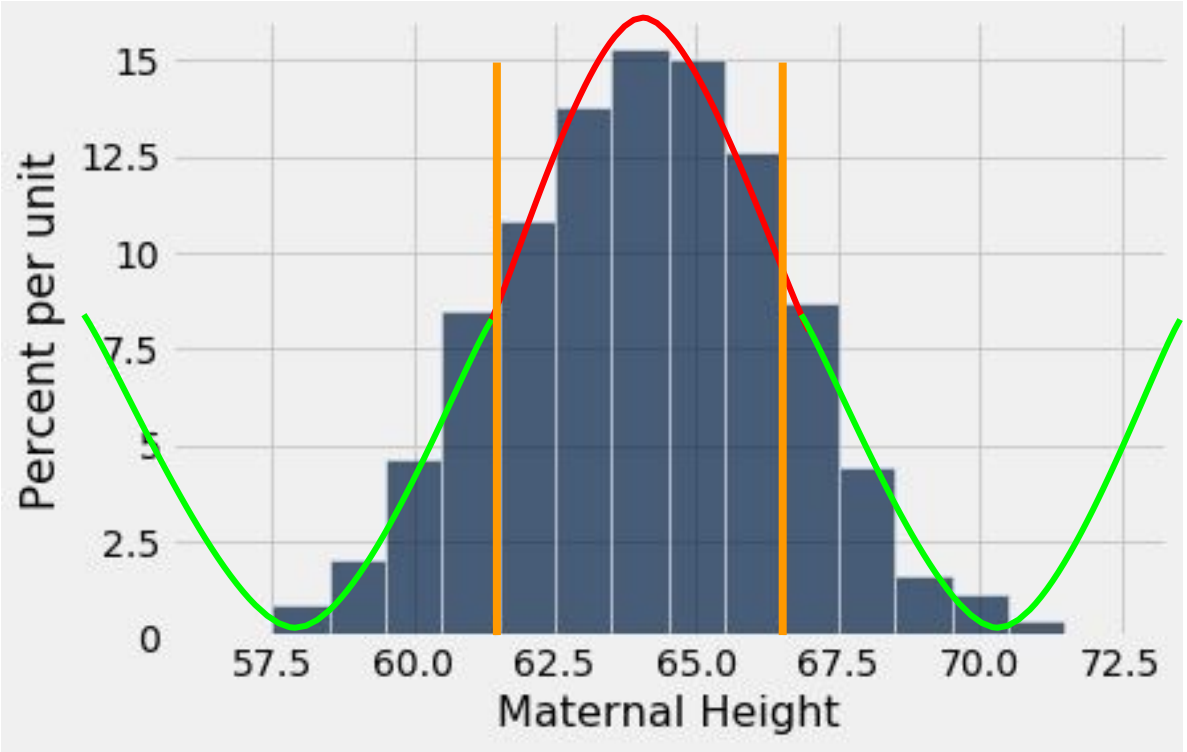
THE SD AND BELL-SHAPED CURVES

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

(Demo – Notebook 8.2, The SD and Bell-Shaped Curves)

POINT OF INFLECTION



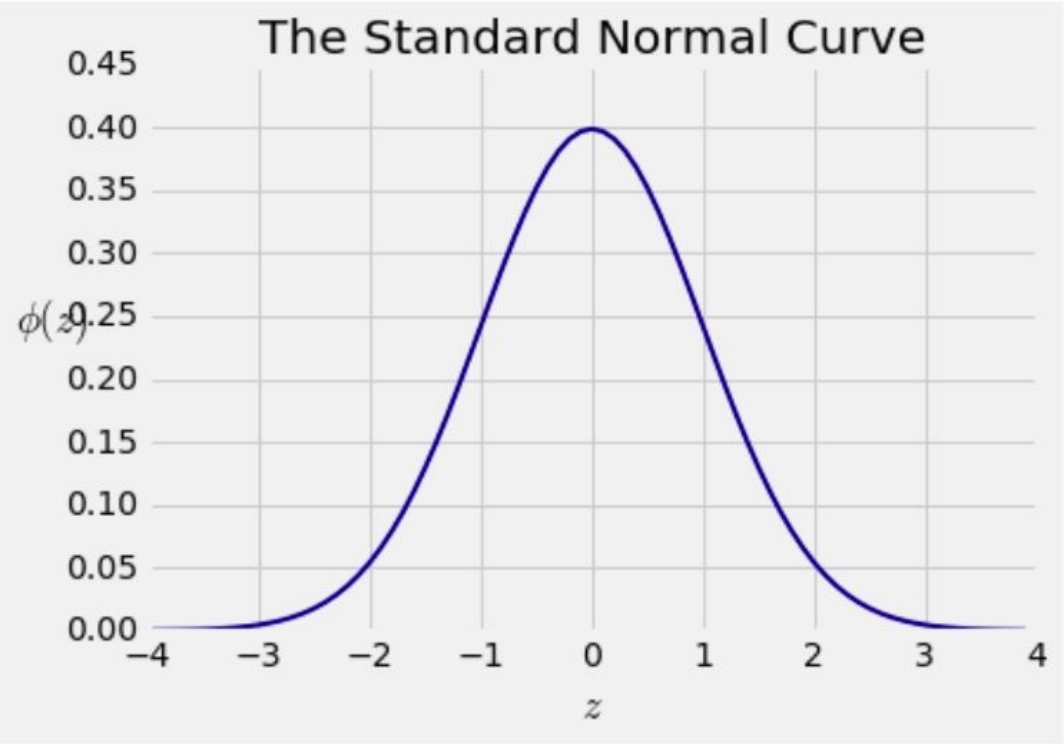
THE NORMAL DISTRIBUTION

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

NORMAL PROPORTIONS

HOW BIG ARE MOST OF THE VALUES?

No matter what the shape of the distribution,

the bulk of the data are in the range "average ± a few SDs"

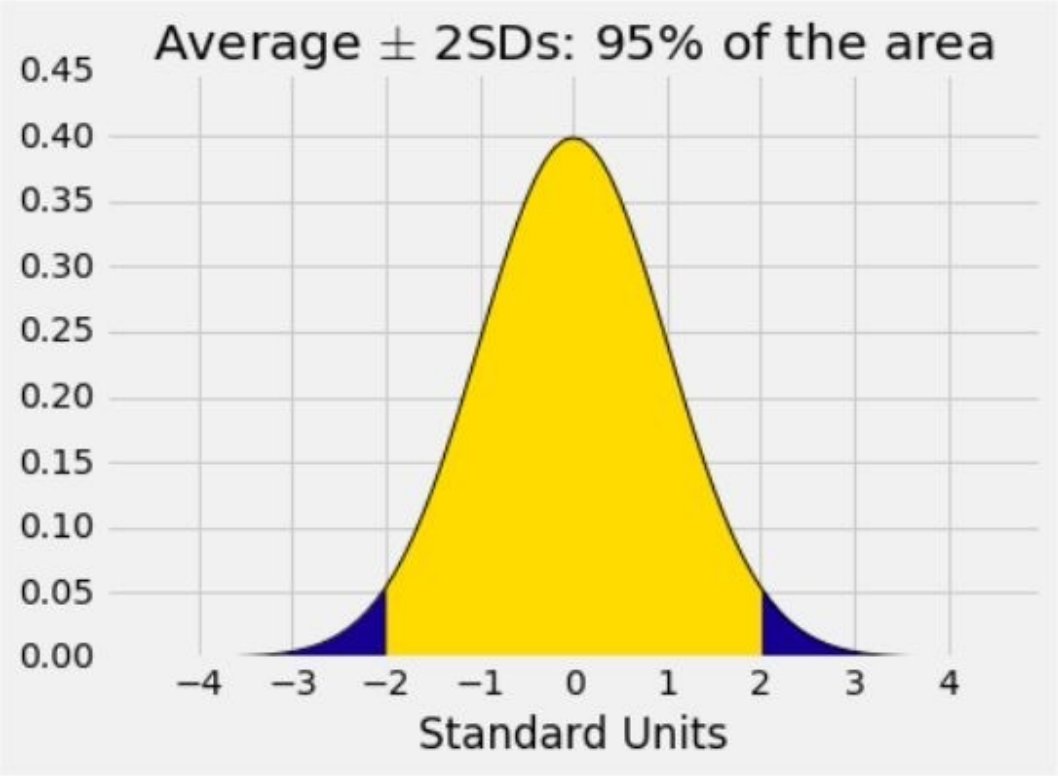
If a histogram is bell-shaped, then

- Almost all of the data are in the range "average ± 3 SDs"

BOUNDS AND NORMAL APPROXIMATIONS

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

A "CENTRAL" AREA



CENTRAL LIMIT THEOREM

SAMPLE AVERAGES

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.
- We care about sample averages because they estimate population averages.

CENTRAL LIMIT THEOREM

If the sample is

- large, and
- drawn at random with replacement,

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

the probability distribution of the sample sum (or the sample average) is roughly normal

(Demo)

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.

(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.

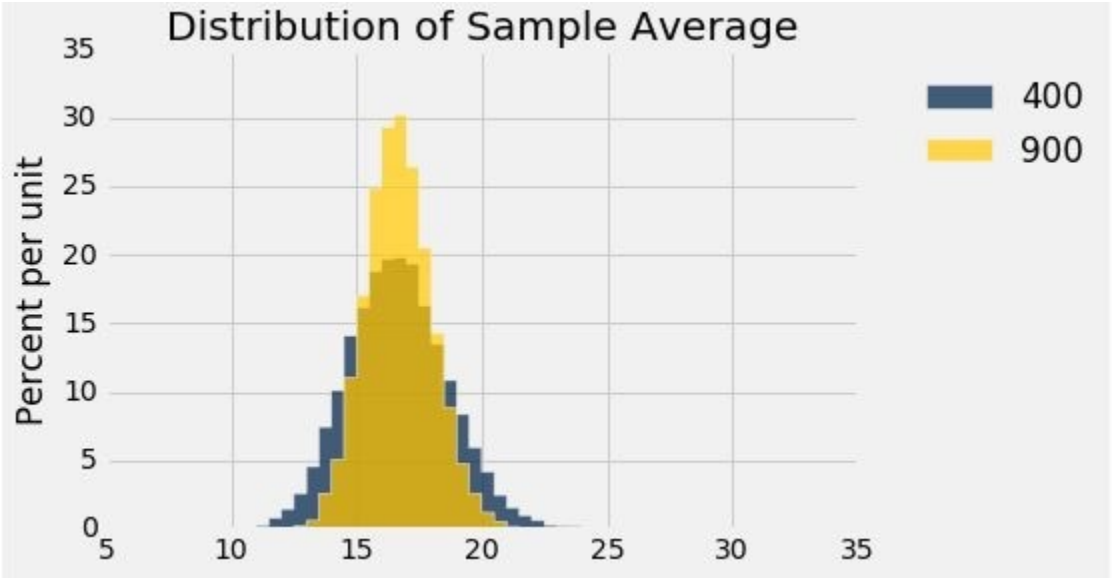
- If we want a specified level of accuracy, understanding the variability of the sample average helps us work out how large our sample has to be

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

(Demo)

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]:

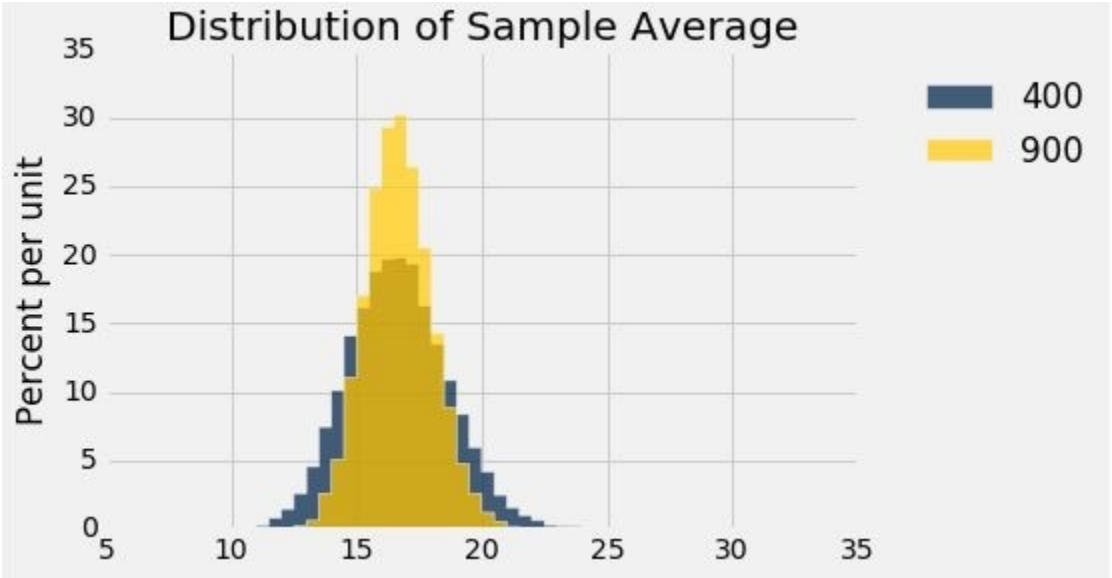
(a) is roughly normal because the number of households is large.
(b) is not close to normal.
(c) may be close to normal, or not; we can't tell from the information given.

AVERAGES OF LARGE SAMPLES

THE EFFECT OF SAMPLE SIZE

CLT: If the sample size is large, the distribution of a random sample average is roughly normal.

The bigger the sample, the smaller the spread of the distribution.



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

(Demo)

CENTRAL LIMIT THEOREM

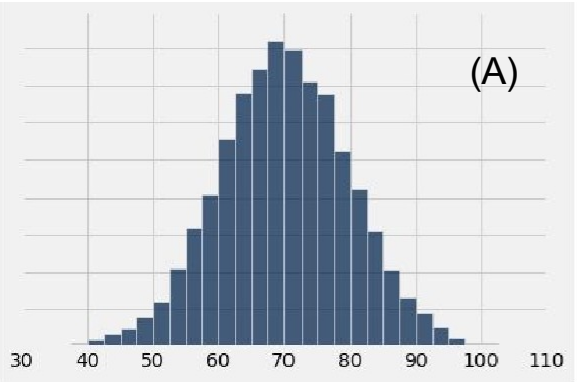
If the sample is large and drawn at random with replacement,

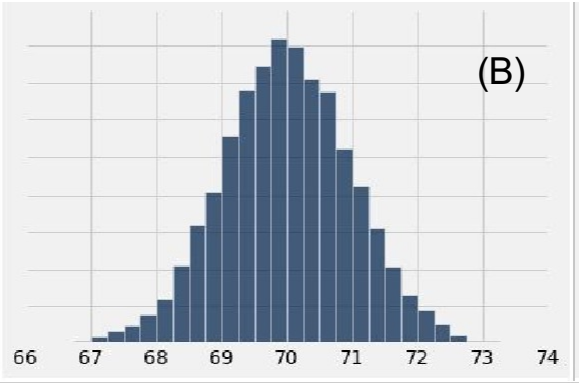
Then, *regardless of the distribution of the population,*

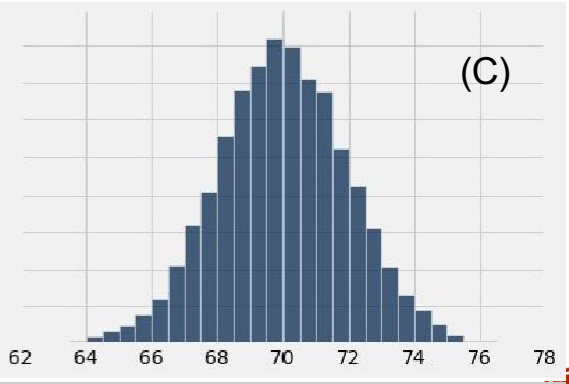
- **the probability distribution of the sample average:**
  - is roughly normal
  - mean = population mean
  - SD = (population SD) / √sample size

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?







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.

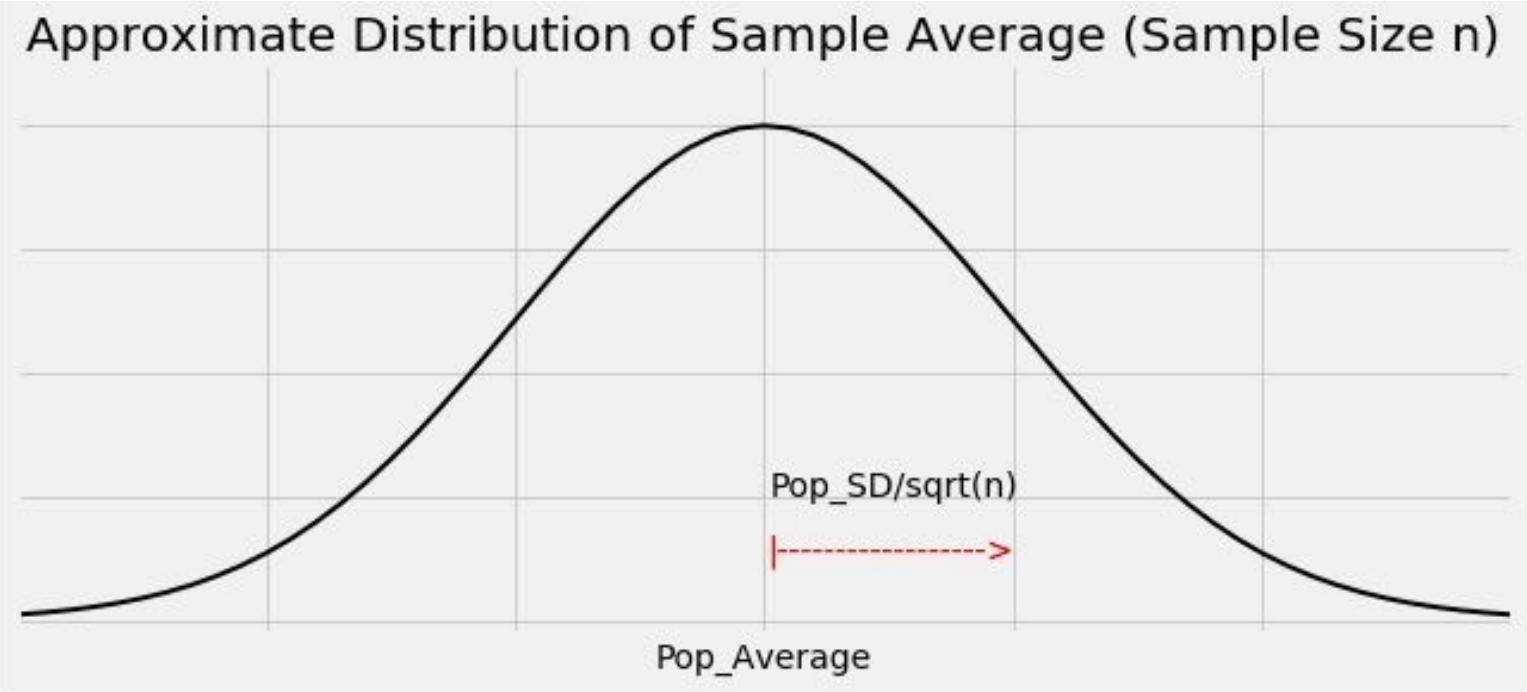
Fill in the blanks and explain:

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

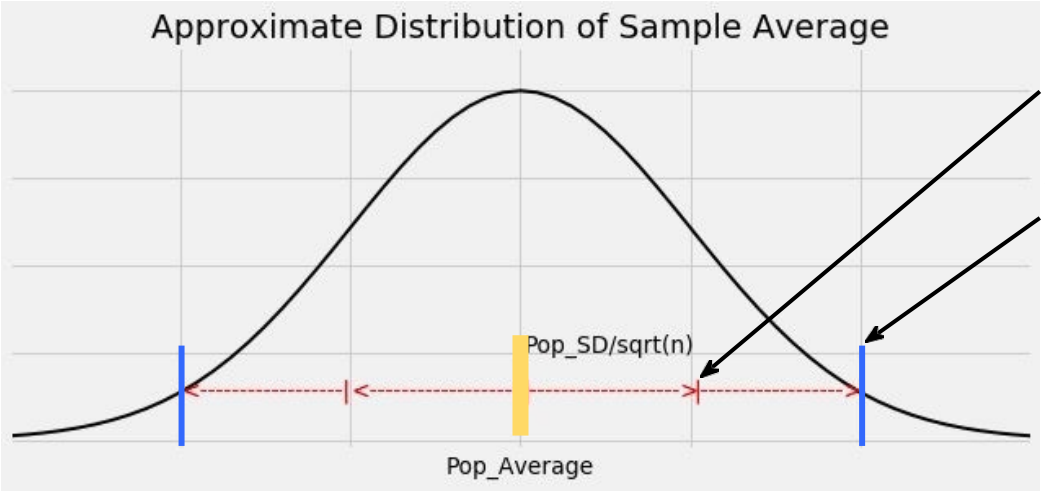
$_____plus or minus $_____

CONFIDENCE INTERVALS

GRAPH OF THE DISTRIBUTION



THE KEY TO 95% CONFIDENCE



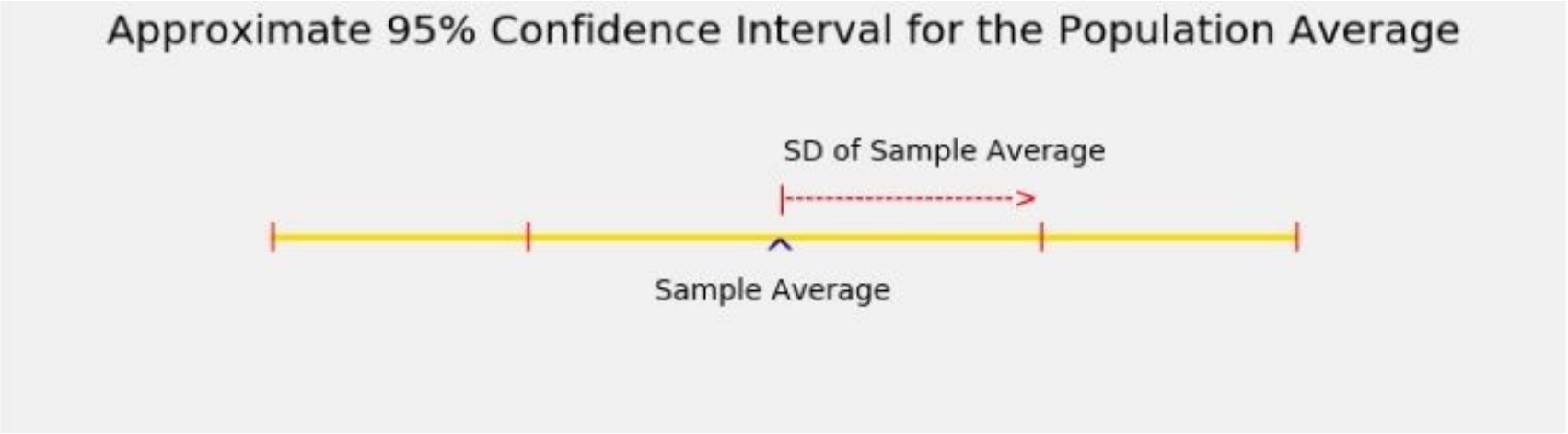
- For about 95% of all samples, the sample average and population average are within **2 SD**s of each other.
- **SD** = SD of sample average
  = (population SD) / $\sqrt{\text{sample size}}$

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



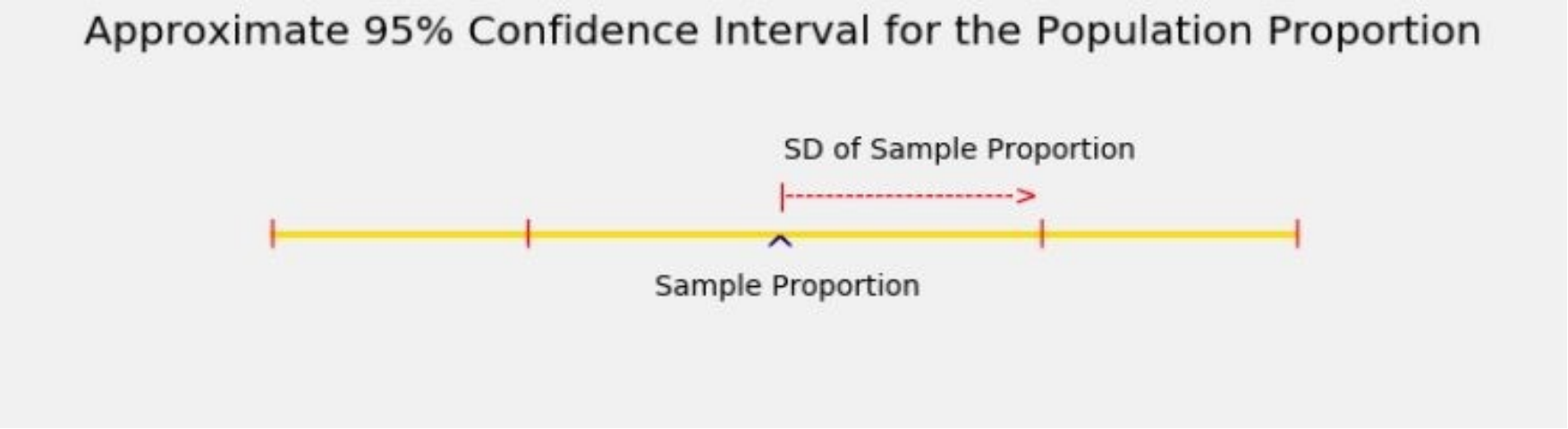
WIDTH OF THE INTERVAL

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

= 4 * SD of the sample average

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

CONFIDENCE INTERVAL



CONTROLLING THE WIDTH

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

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

- The narrower the interval, the more precise your estimate.
- Suppose you want the total width of the interval to be no more than 1%. How should you choose the sample size?

THE SAMPLE SIZE FOR A GIVEN WIDTH

$$0.01 = 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$$

- Left side: 1%, the max total width that you'll accept
- Right side: formula for the total width

$$\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.01$$

(Demo)

"WORST CASE" POPULATION SD

- $\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.01$
- SD of 0/1 population is at most 0.5
- $\sqrt{\text{sample size}} \geq 4 * 0.5 / 0.01$
- $\text{sample size} \geq (4 * 0.5 / 0.01) ** 2 = 40000$
- The sample size should be 40,000 or more

DISCUSSION QUESTION

https://www.scientificamerican.com/article/howcan-a-poll-of-only-100/

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?

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

QUESTIONS?