

COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 19: Central Limit Theorem

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

Sept 30, 2025

Copyright © 2026 Barnard College

April 8, 2026



Final Project Proposals Due Friday, 4/17

Today's Lecture

- Bootstrapping
- Confidence Intervals
- Standard Deviation
- Normal Distribution

Last Class: Normal Distributions

Standard Deviation (SD)

Standard deviation measures the variability around the mean

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

- No matter the shape of the distribution, the bulk of the data is in the range “average plus or minus a few standard deviations”

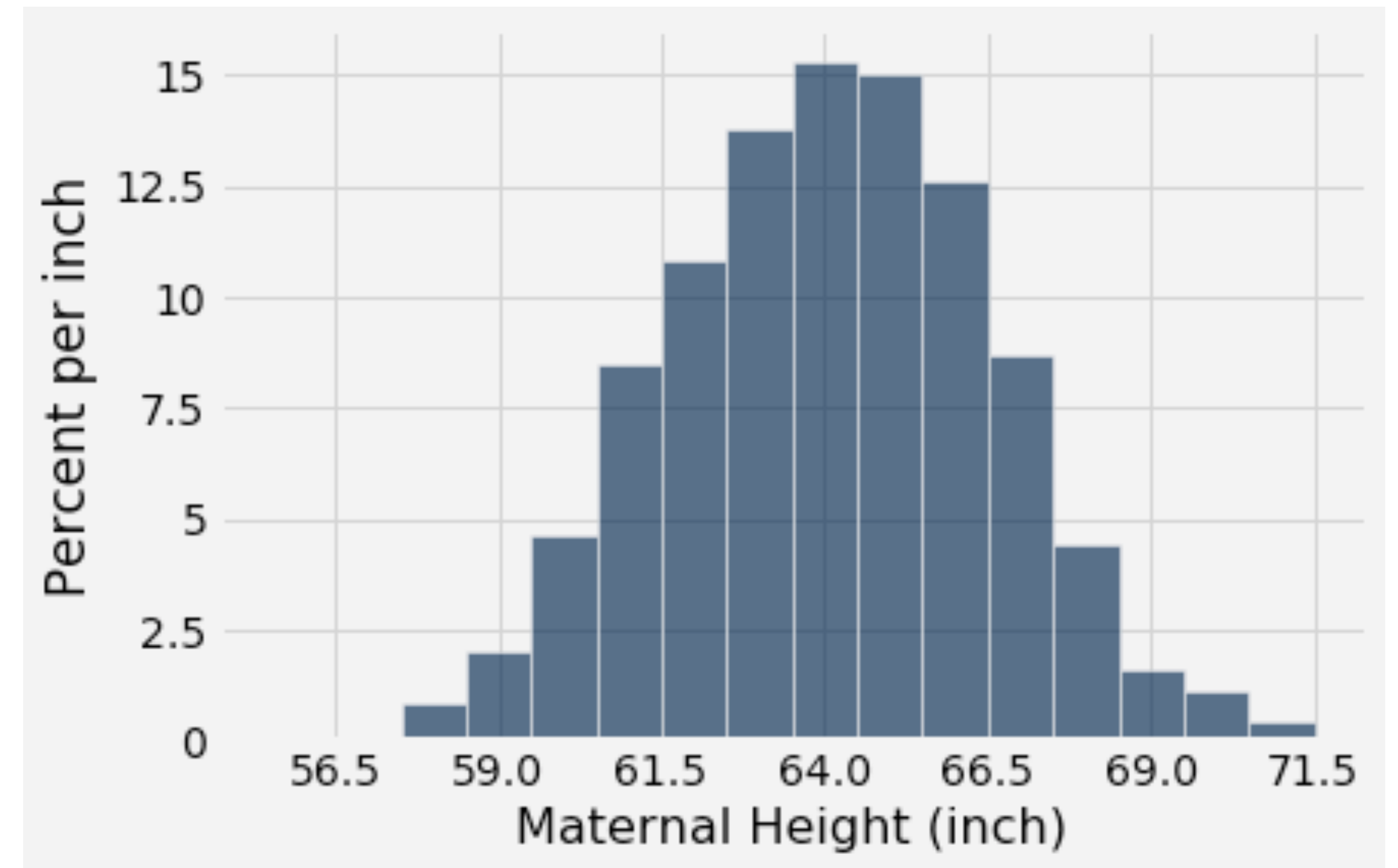
Standard Units

- The quantity z (from “average $\pm z$ SDs” in Chebychev’s inequality) measures **standard units**
- **Standard units** is the number of standard deviations away from the average
- To convert a value (v) to standard units, compare the deviation from the average (μ) with the standard deviation (SD):

$$z = \frac{v - \mu}{\text{SD}}$$

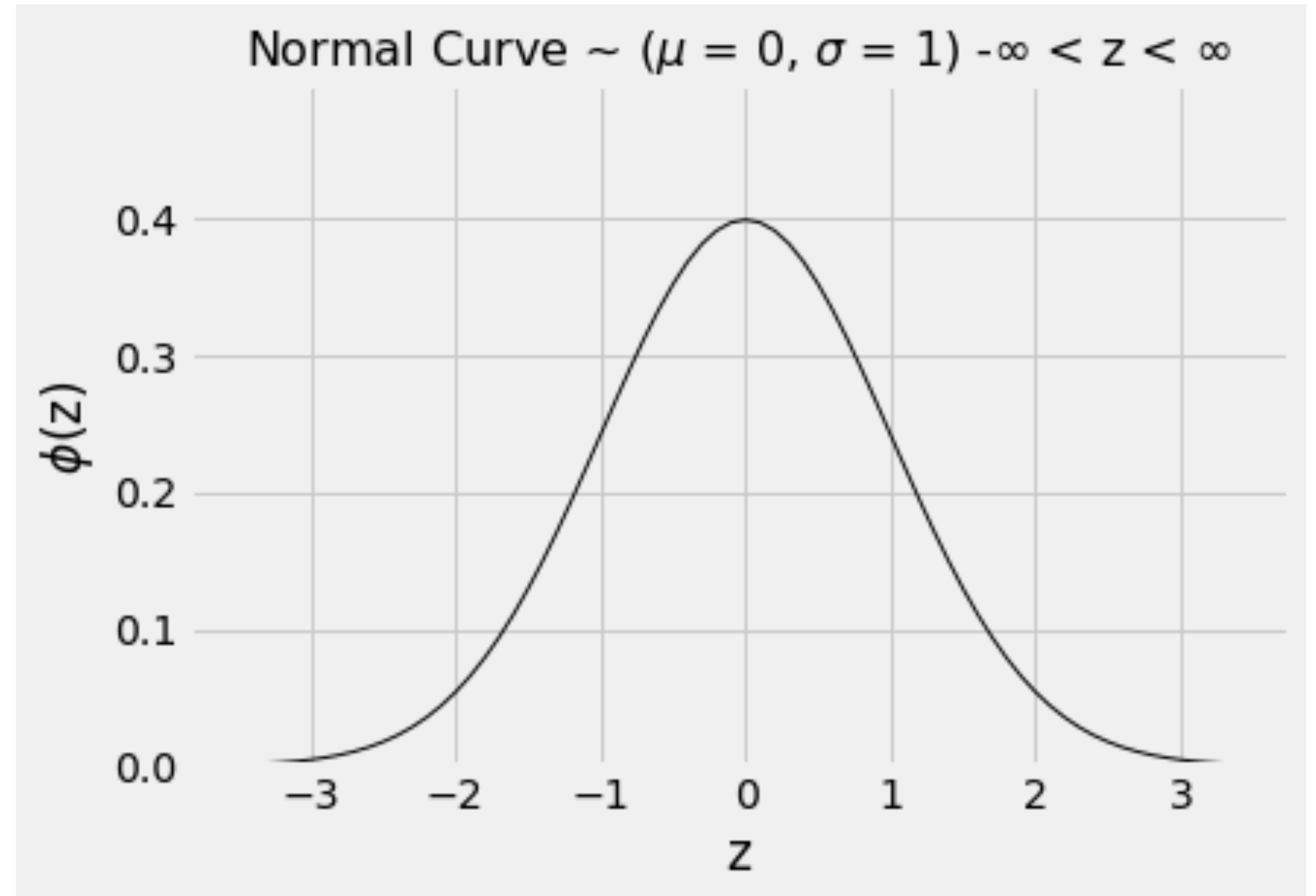
Bell Shaped Curves

- The normal curve / bell-curve is a very common distribution
- For bell-shaped (aka Gaussian distribution):
 - Average is at the center
 - SD is the distance between the average and the points of inflection on either side



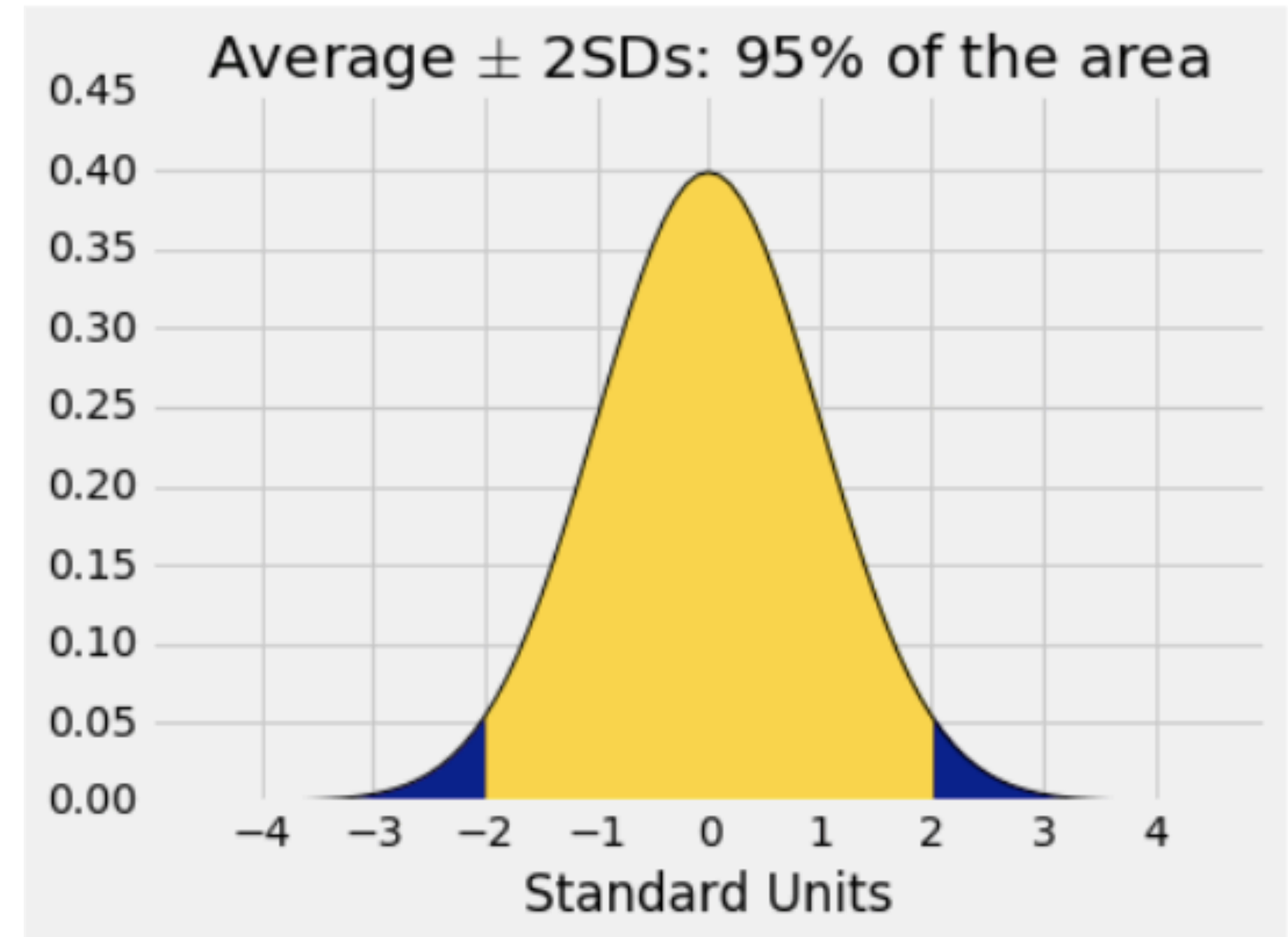
Normal Distribution

- On a standard normal curve, x-axis units are standard units
- Total area of the curve is 1
- Curve is symmetric around 0 (mean and median are both 0)
- Points of inflection are -1 and 1
- Standard deviation is 1



Application to Normal Distributions

- If a histogram is bell-shaped (normal), then 95% of the data is in the range average + 2 SDs
- Note this is much higher than Chebychev's bound of 75%
- 75% is a lower bound that applies to *all* distributions



Normal vs All Distributions

Range	All Distributions (Chebyshev's)	Normal Distribution
mean \pm 1 SDs	At least 0%	At least 68%
mean \pm 2 SDs	At least 75%	At least 95%
mean \pm 3 SDs	At least 89%	At least 99%

Central Limit Theorem

Central Limit Theorem

- Describes how a normal distribution is connected to random sample averages (the average of a sample we collect)
- We calculate sample averages because they can help us estimate population averages

Central Limit Theorem

Definition:

Is a sample is large and drawn at random with replacement

Then regardless of the distribution,

The **probability distribution of the sample average** is roughly normal

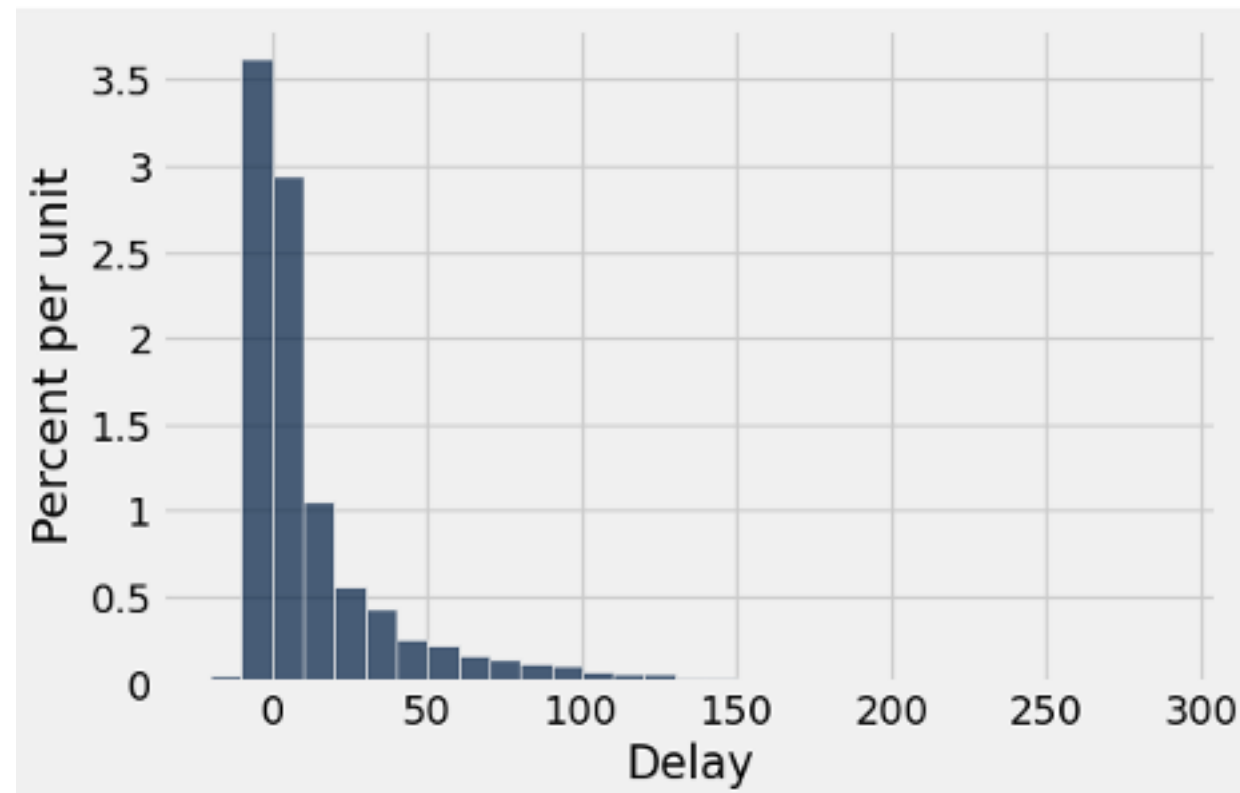
Distribution of Sample Averages

- If you have one random sample, you can take the average of values in that sample (*one sample average*)
- But that sample average could be different if you took different sample
- Distribution of sample averages is the distribution of *possible sample averages* if you were to draw different samples

Notebook Demo: Central Limit Theorem

Example: Delays in United Flights

Population



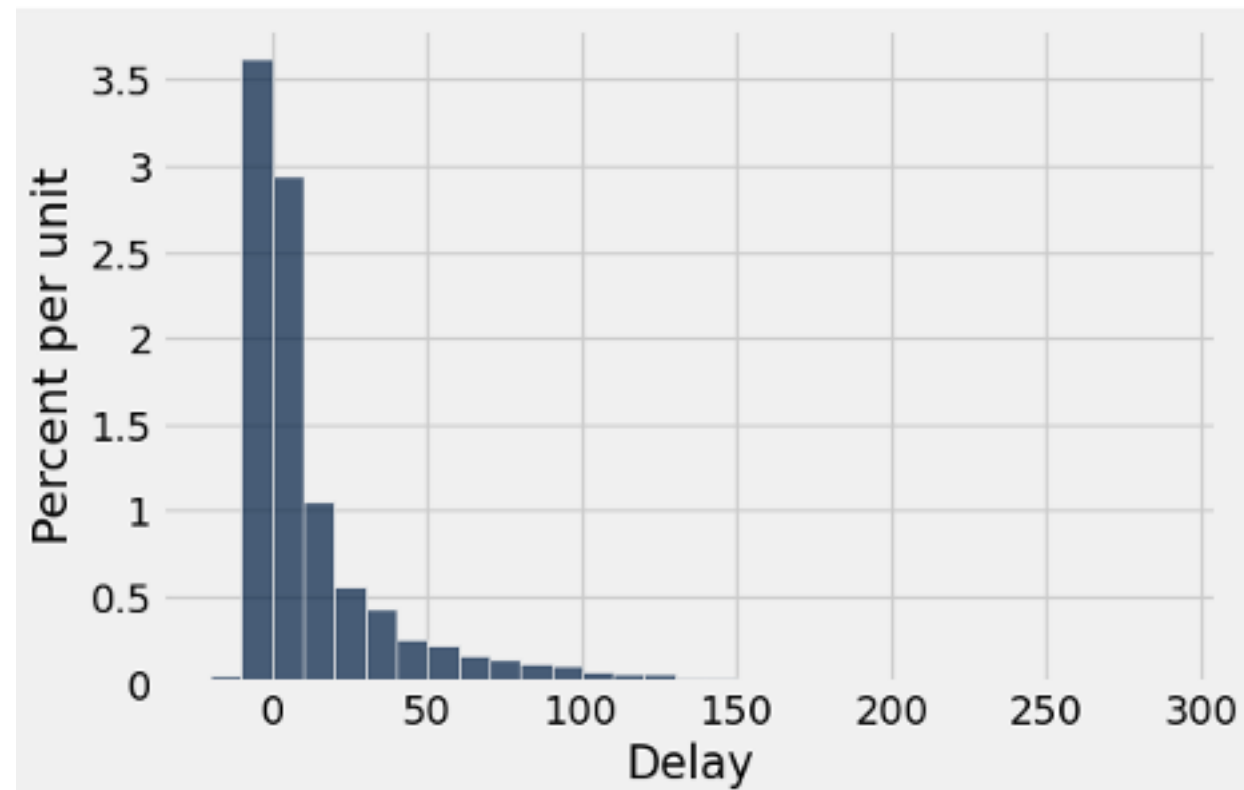
Population mean = 16.65

SD = 39.5

Example: Delays in United Flights

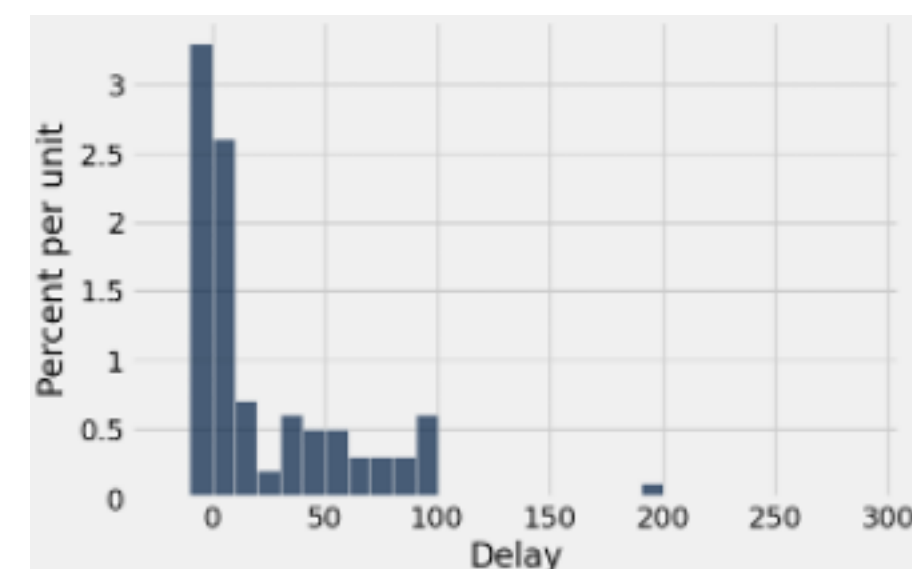
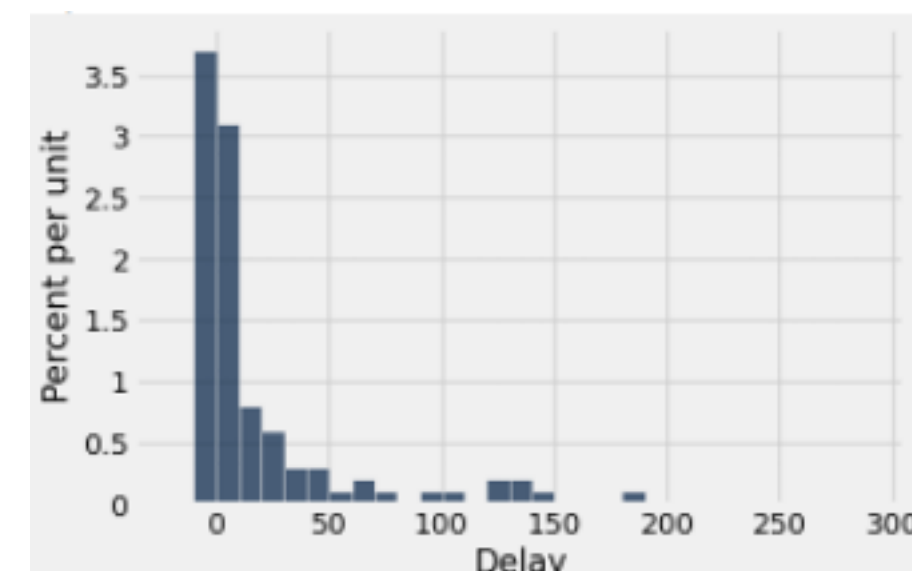
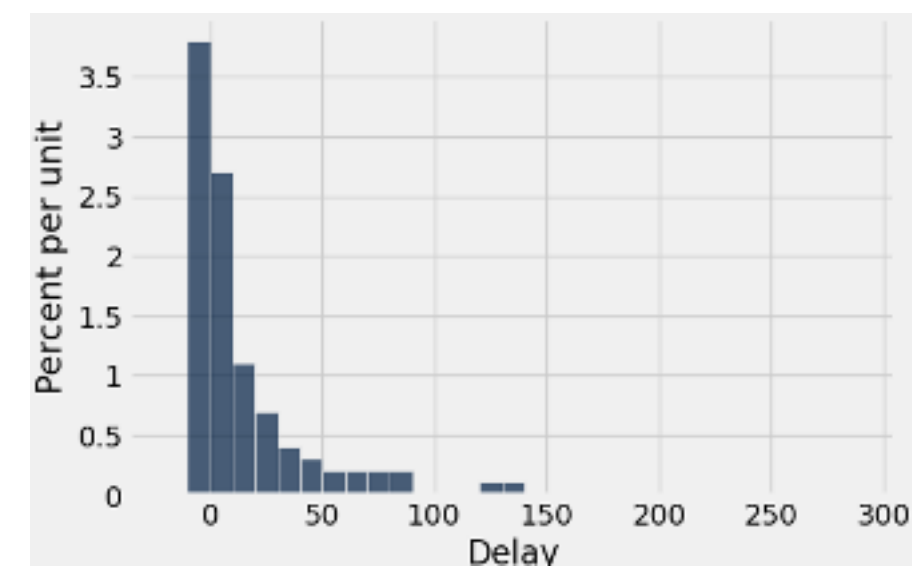
**Sample
Size = 100**

Population



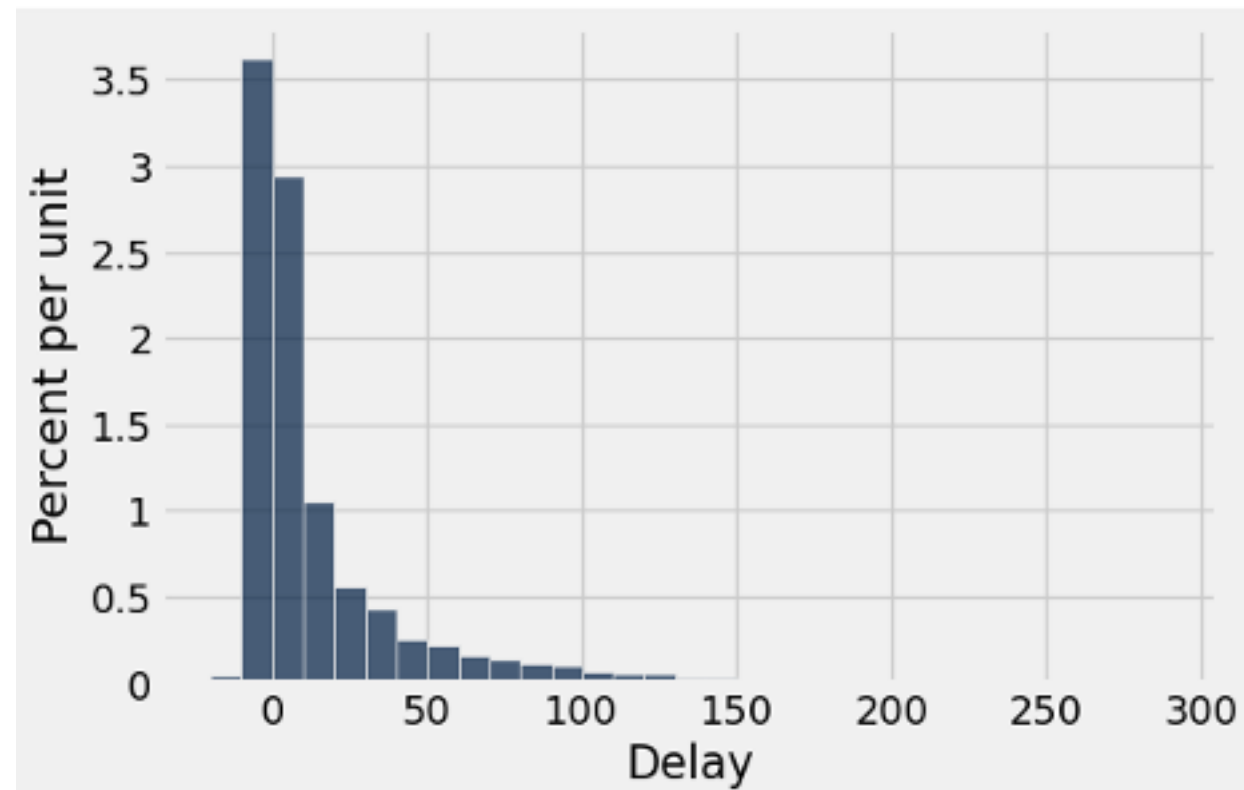
Population mean = 16.65

SD = 39.5



Example: Delays in United Flights

Population

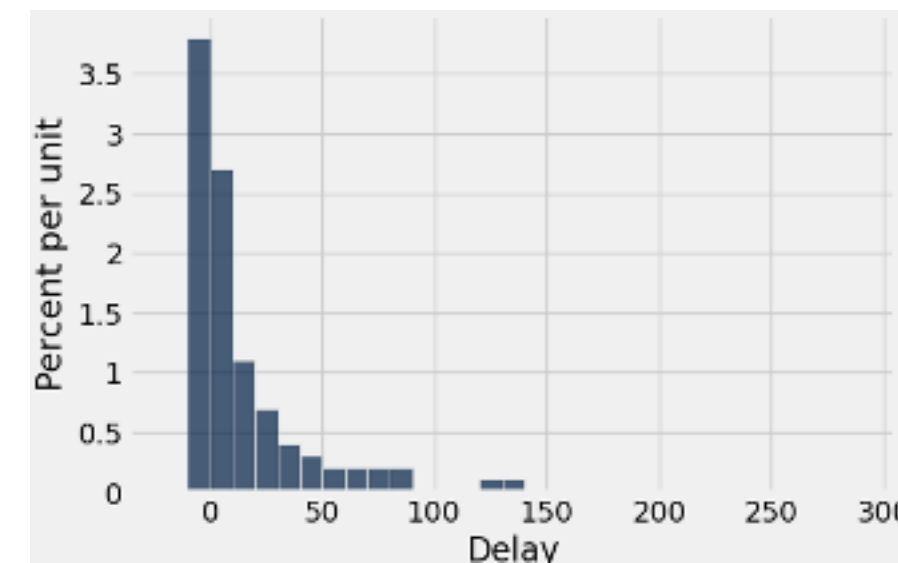


Population mean = 16.65

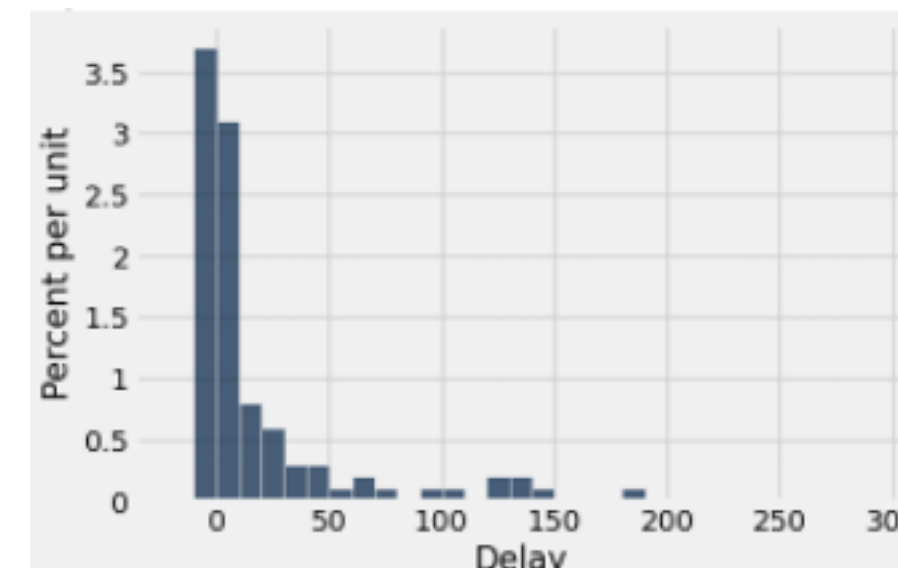
SD = 39.5

Sample
Size = 100

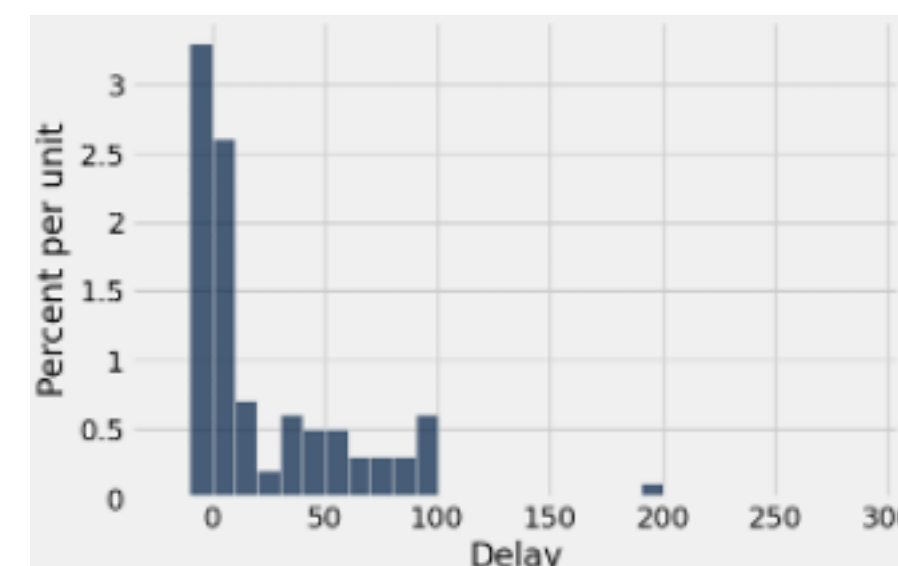
Statistic
Sample mean



$$\mu_1 = 13.1$$



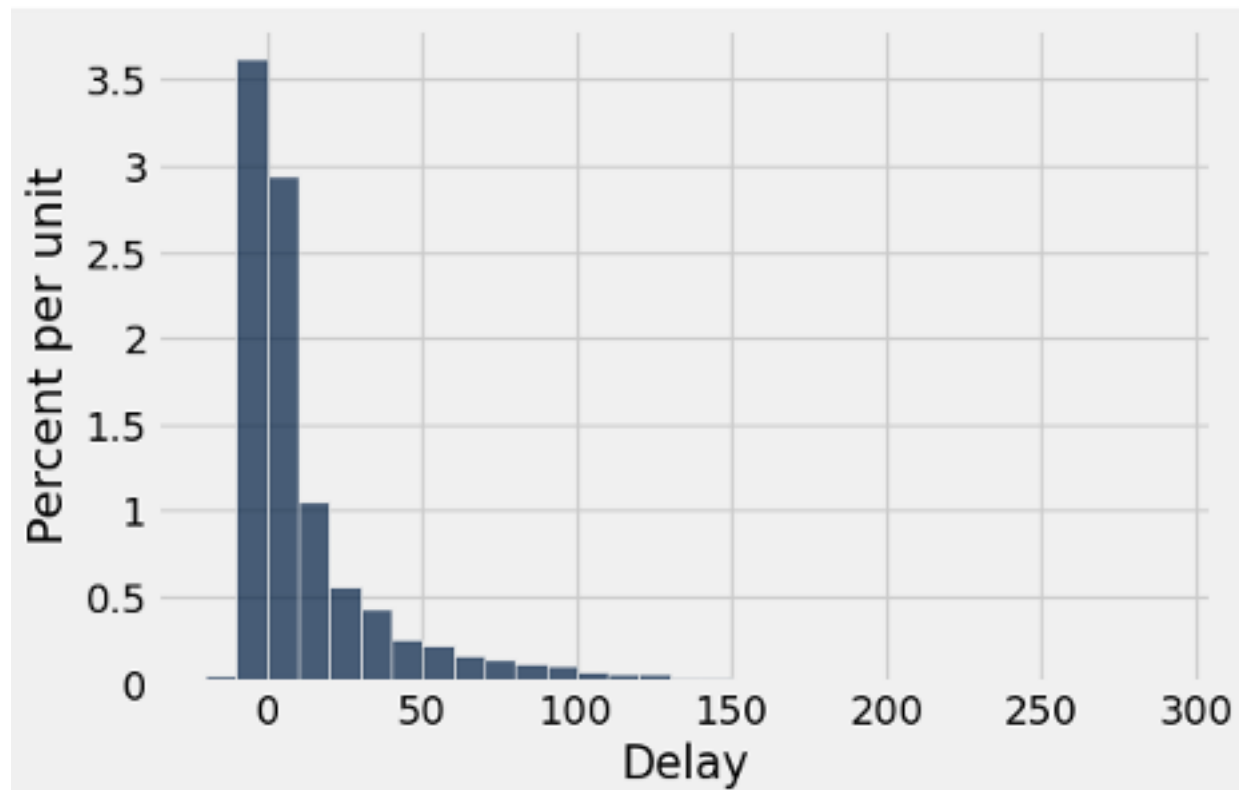
$$\mu_2 = 17.55$$



$$\mu_n = 21.95$$

Example: Delays in United Flights

Population



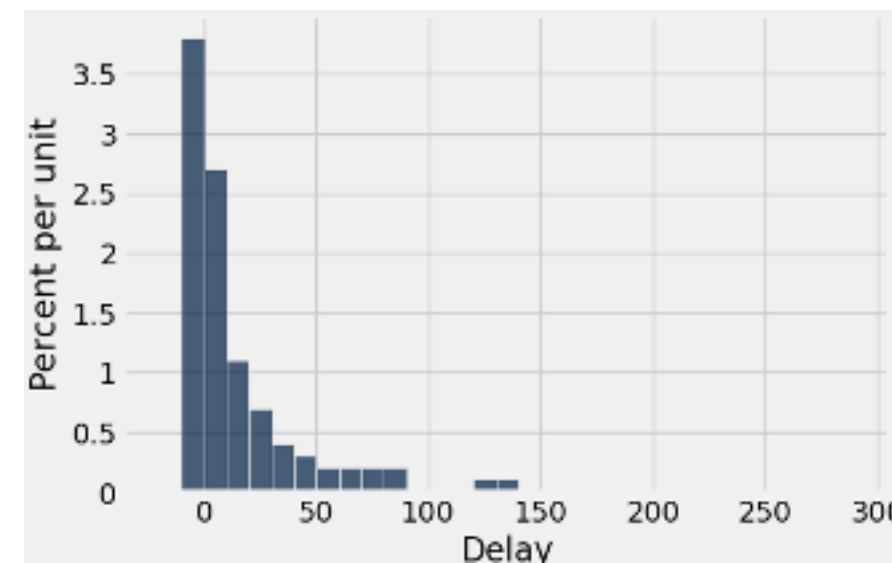
Population mean = 16.65

SD = 39.5

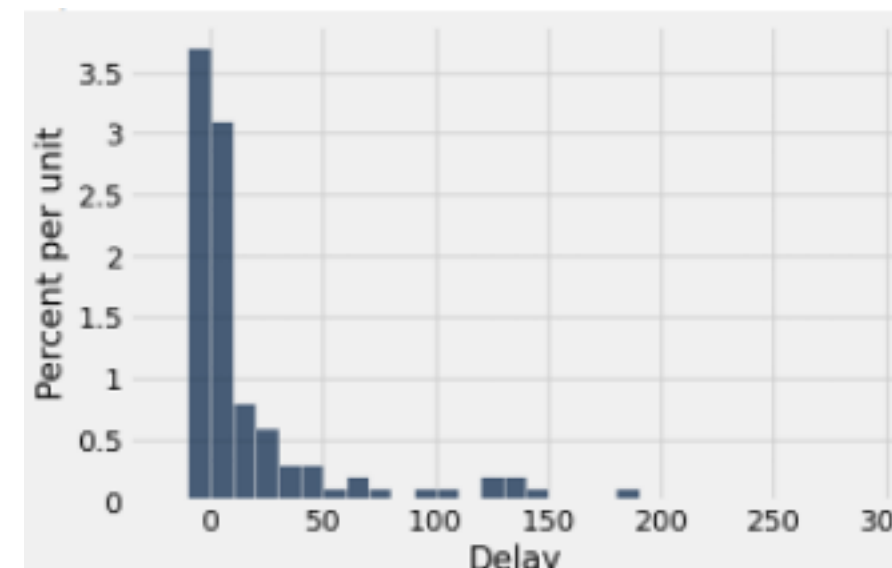
Sample
Size = 100

Statistic
Sample mean

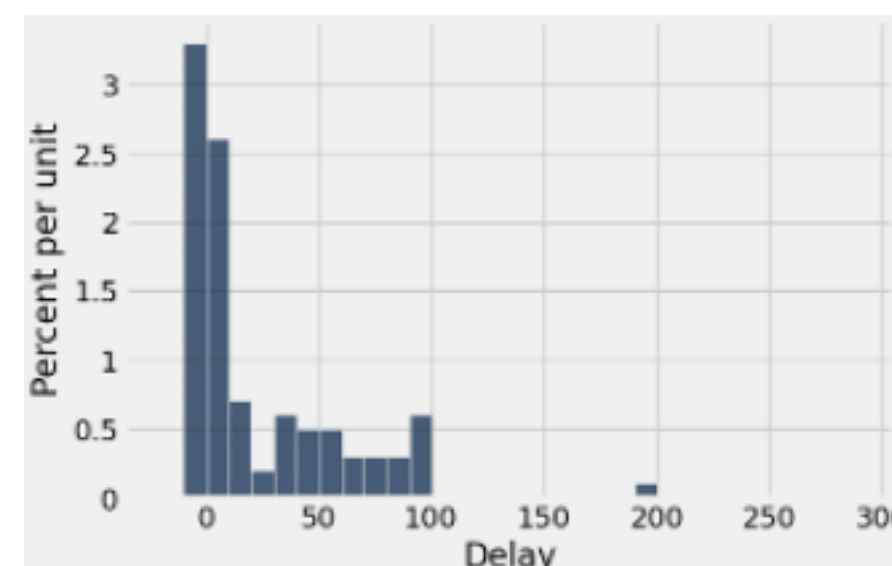
Distribution of
sample means



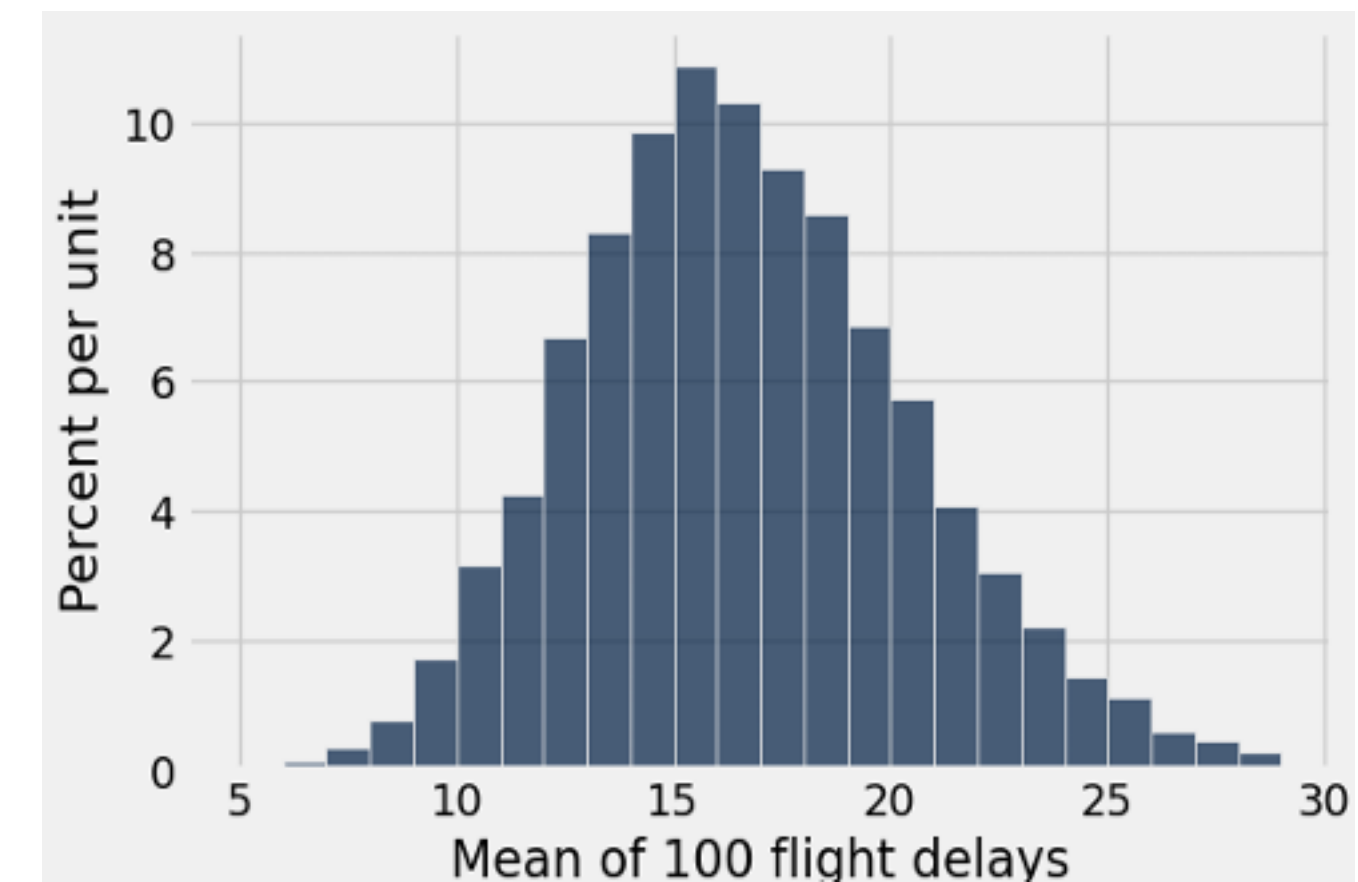
$$\mu_1 = 13.1$$



$$\mu_2 = 17.55$$



$$\mu_n = 21.95$$



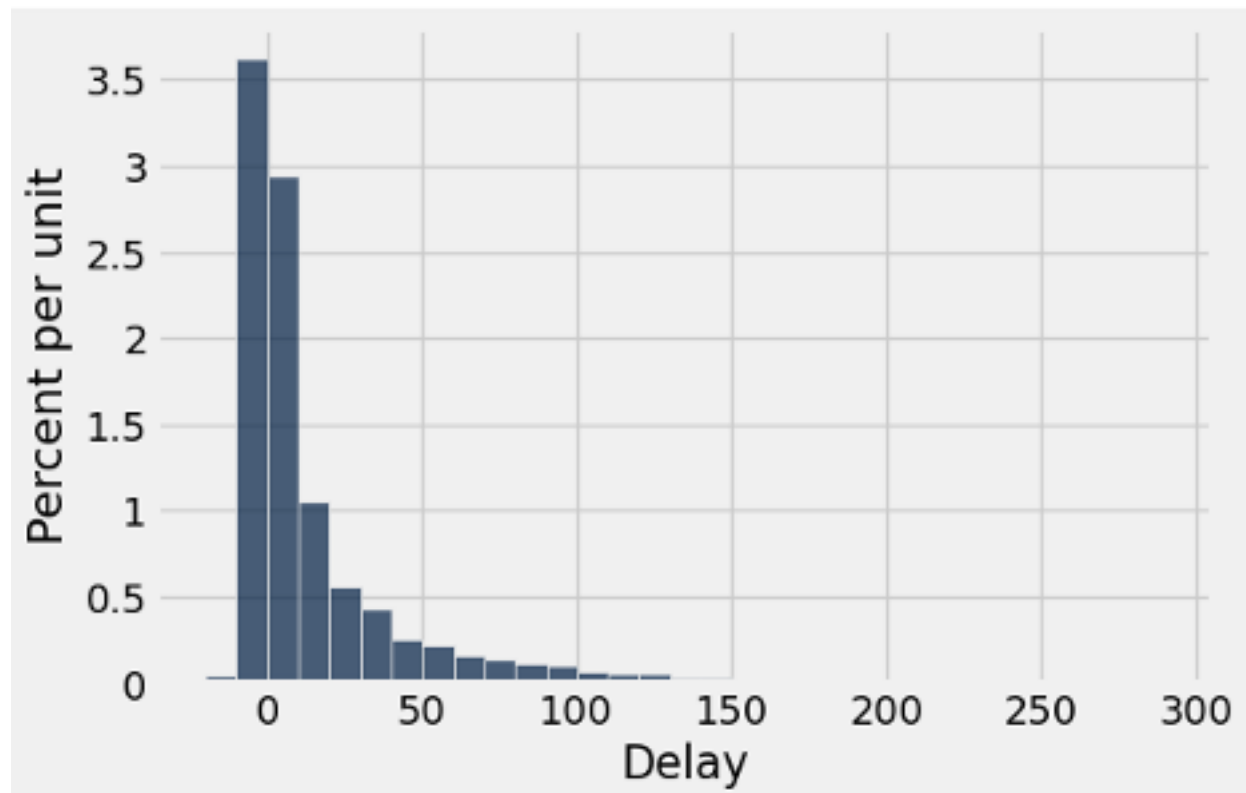
Sample average mean = 16.63

SD = 3.9

Example: Delays in United Flights

How might things change if we change the size of the sample of means?

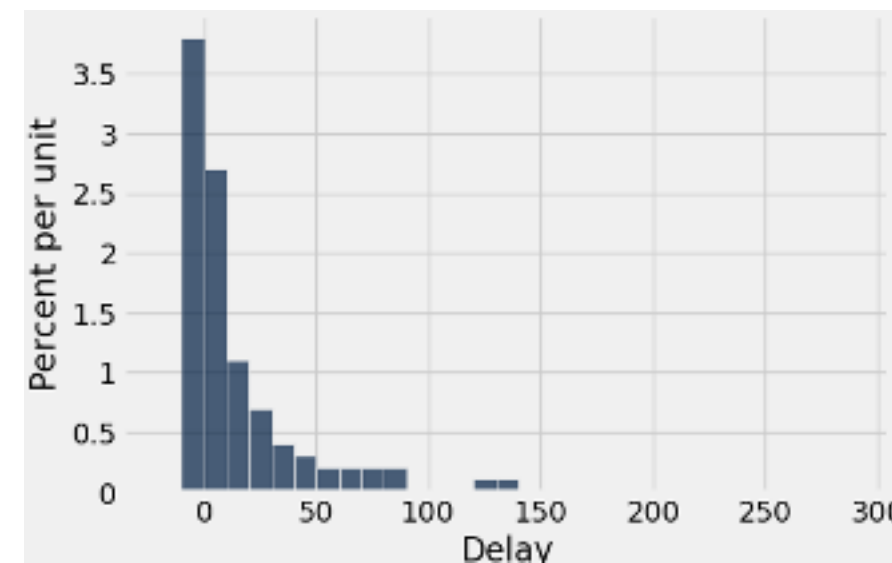
Population



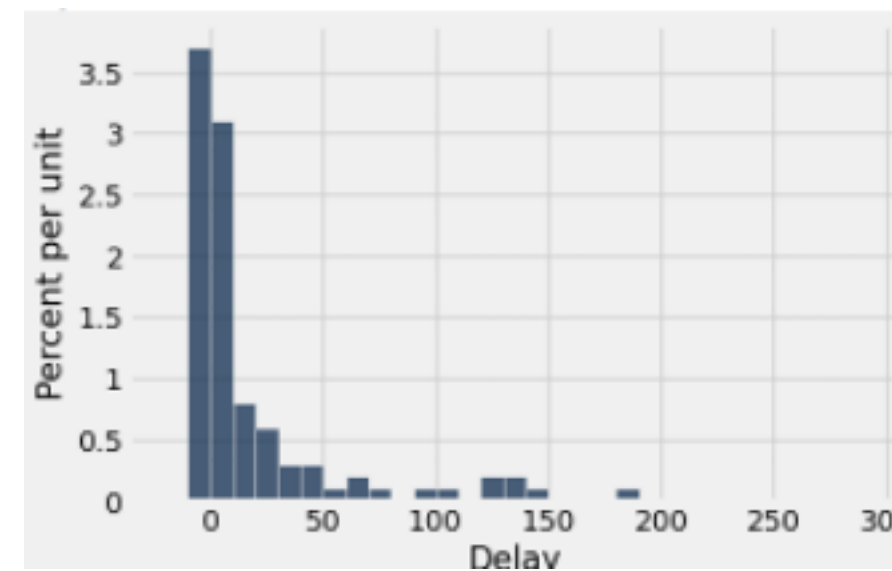
Population mean = 16.65

SD = 39.5

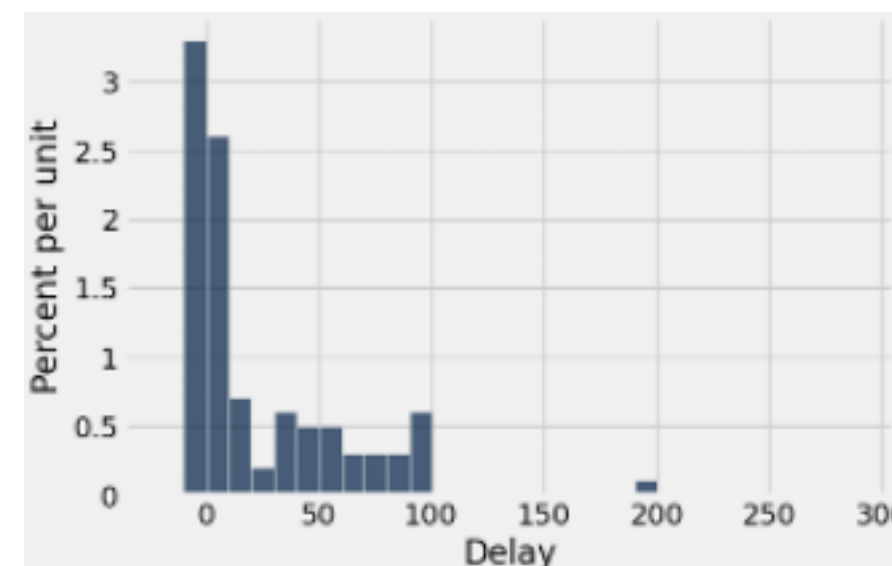
Sample Size = 100



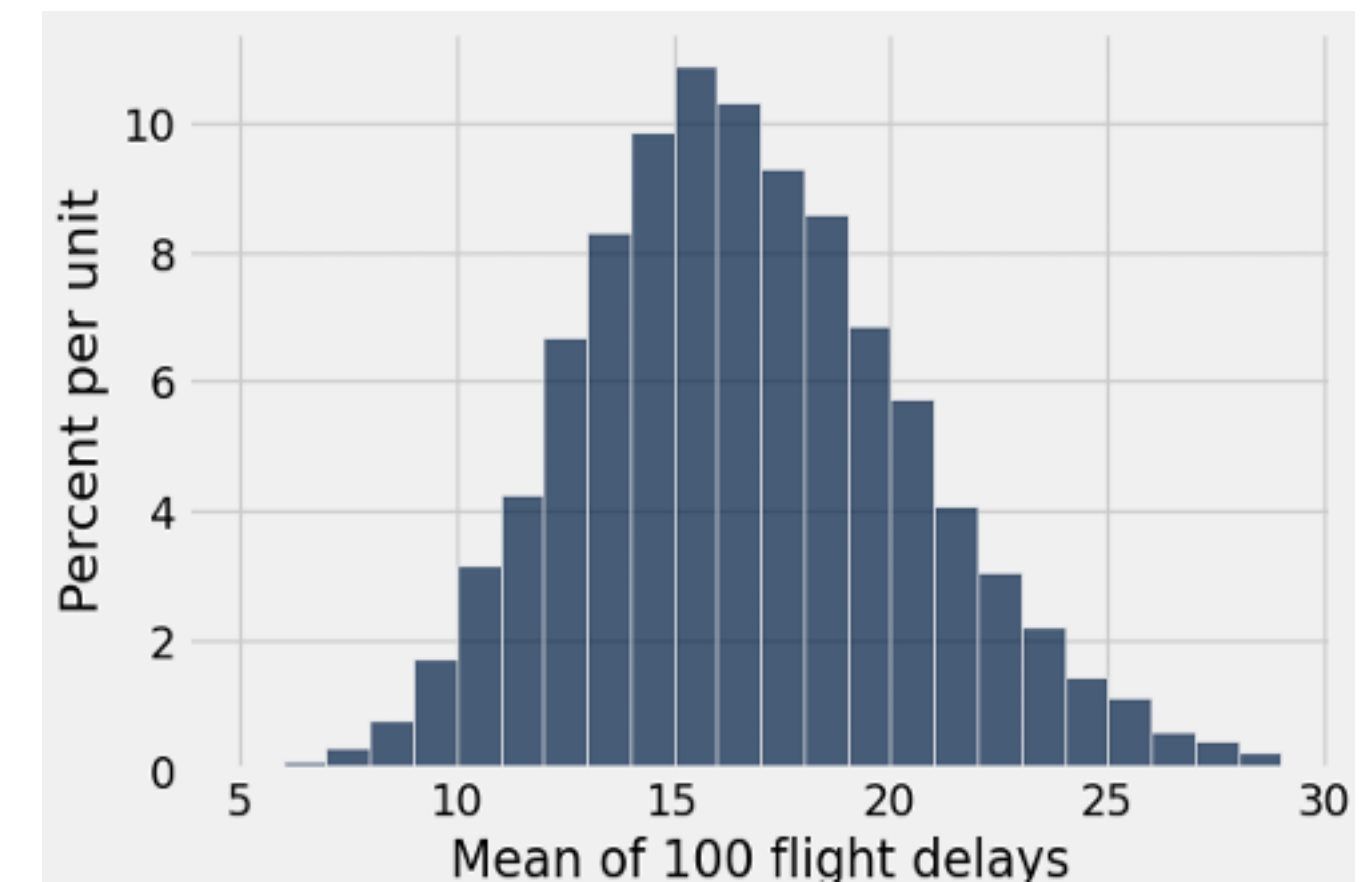
$$\mu_1 = 13.1$$



$$\mu_2 = 17.55$$



$$\mu_n = 21.95$$



Sample average mean = 16.63

SD = 3.9

Central Limit Theorem for Sample Mean

Definition:

If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population,

the **probability distribution of the sample mean** is roughly normal,

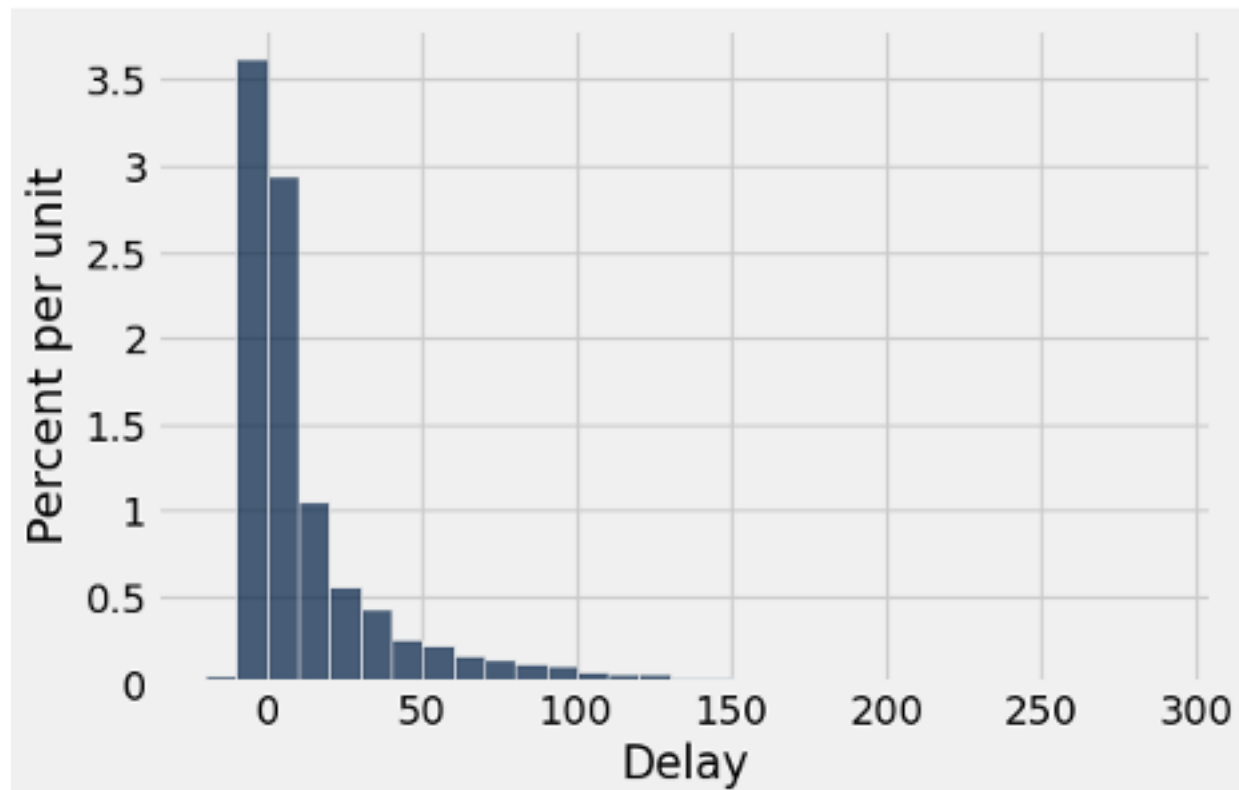
centered at the population mean (mean = population mean), with

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Notebook Demo: Standard Deviation + Sample Size

Example: Delays in United Flights

Population



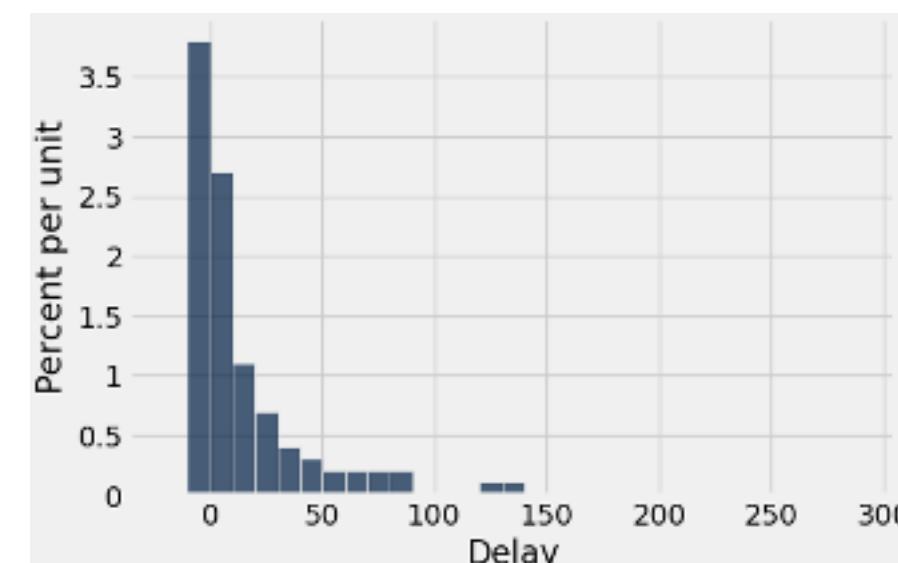
Population mean = 16.65

SD = 39.5

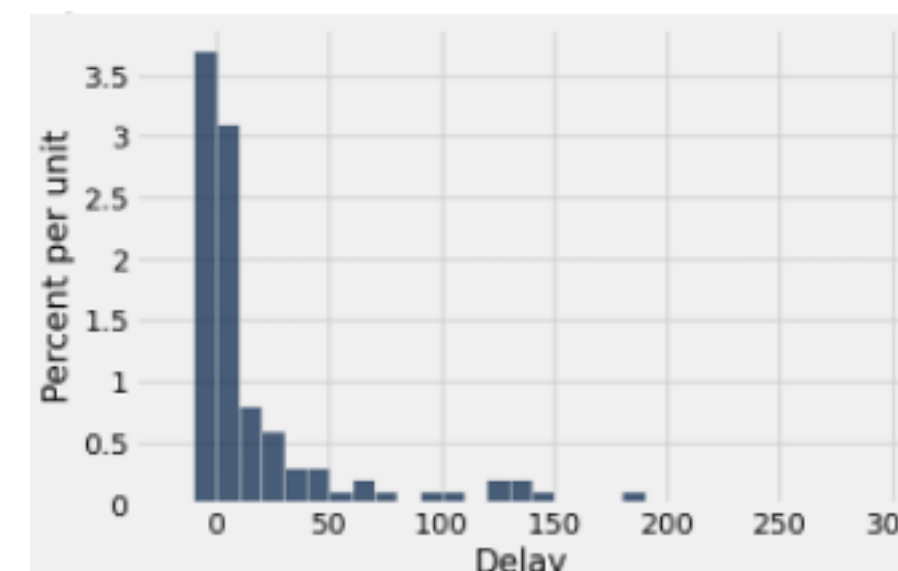
Sample
Size = 100

Statistic
Sample mean

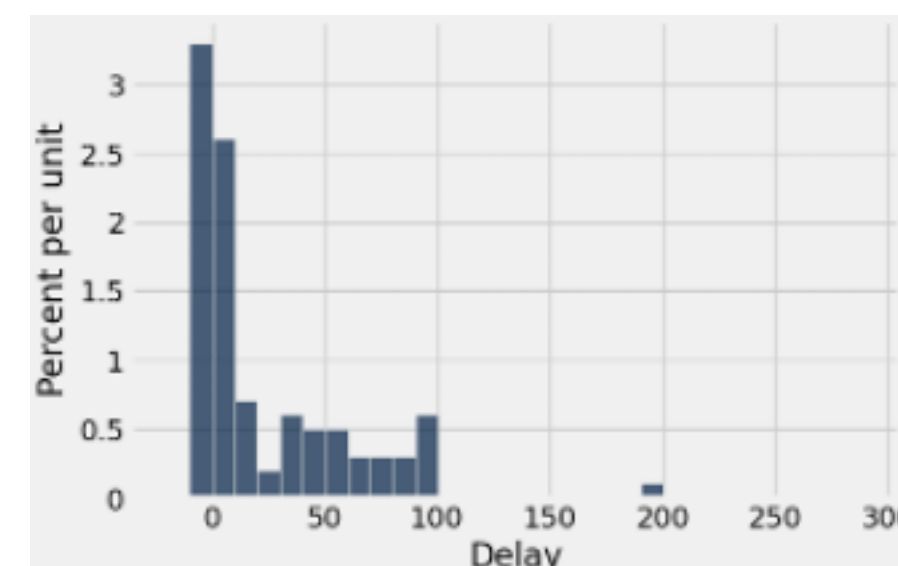
Distribution of
sample means



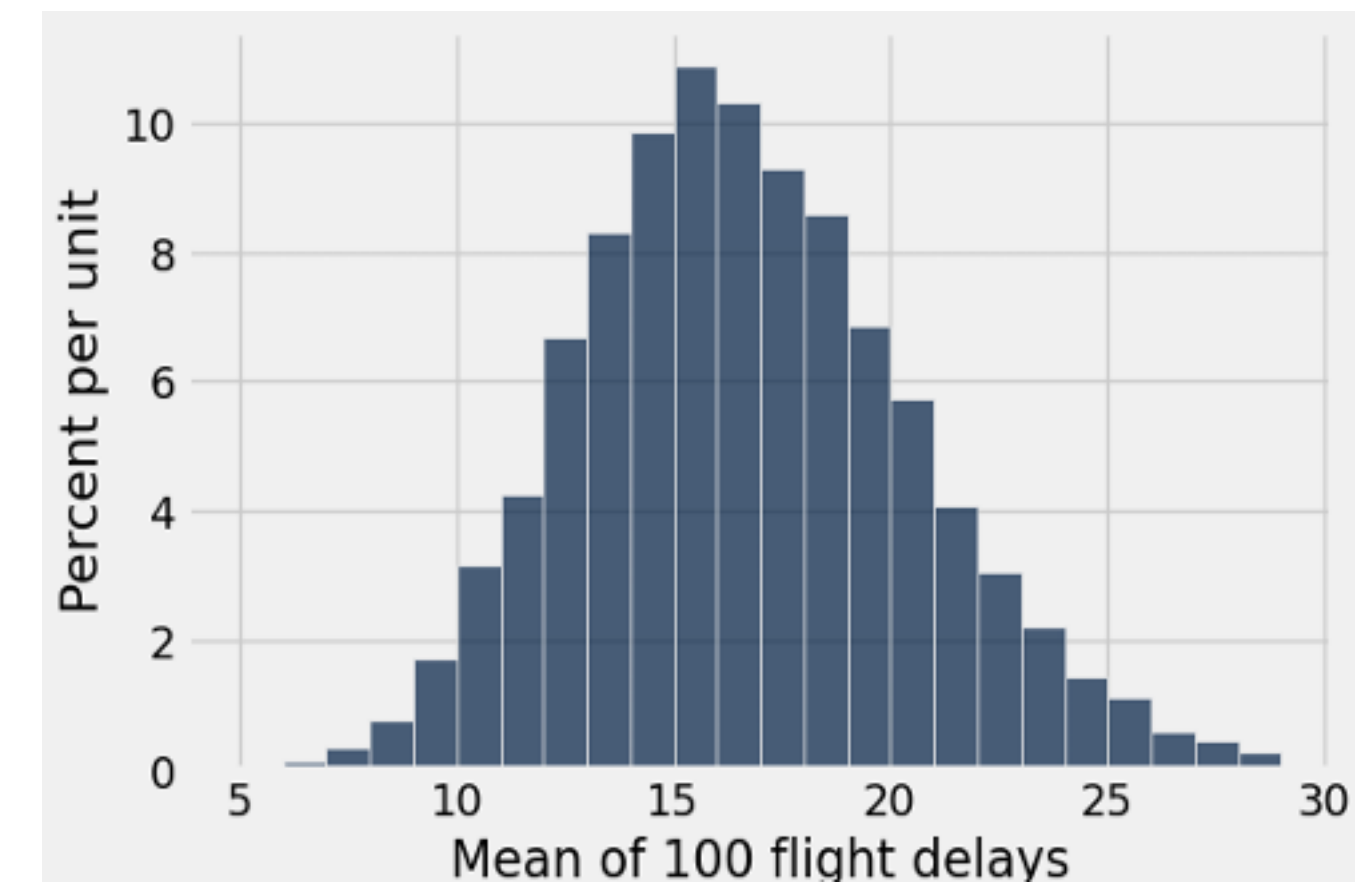
$$\mu_1 = 13.1$$



$$\mu_2 = 17.55$$



$$\mu_n = 21.95$$

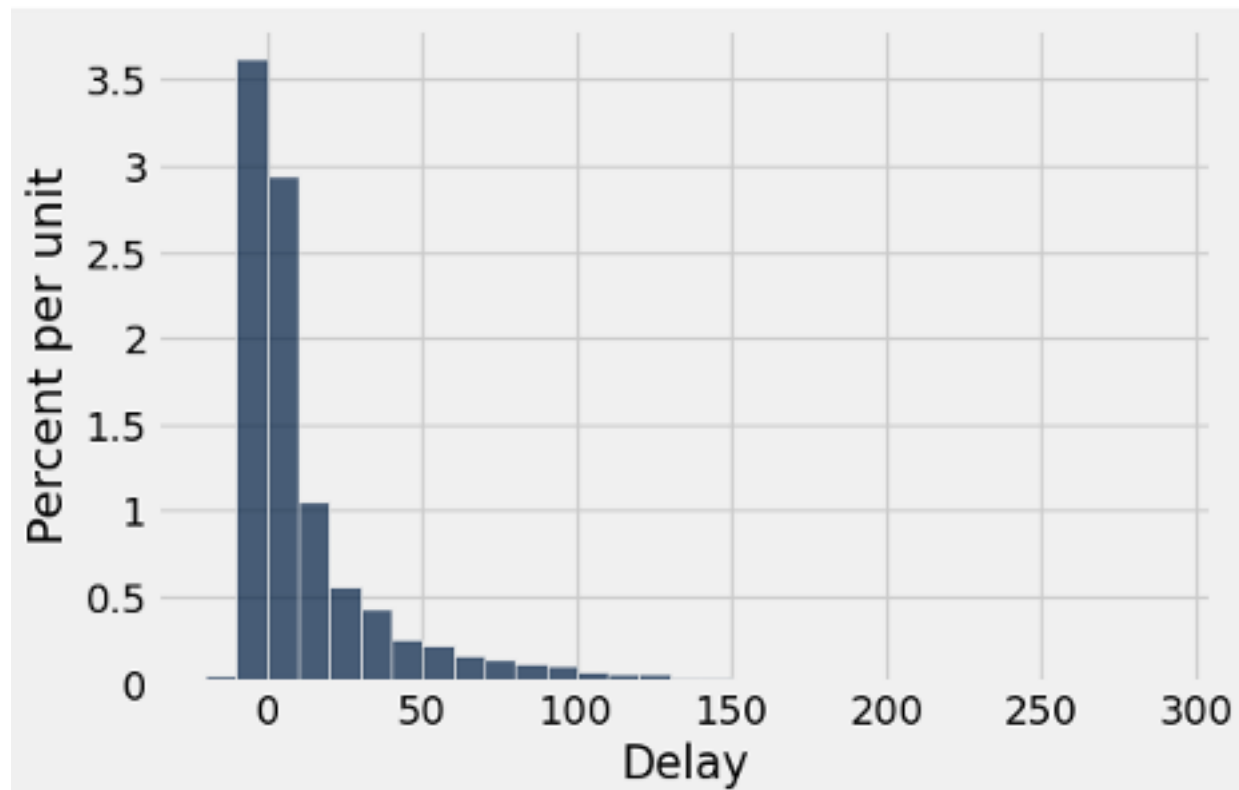


Sample average mean = 16.63

SD = 3.9

Example: Delays in United Flights

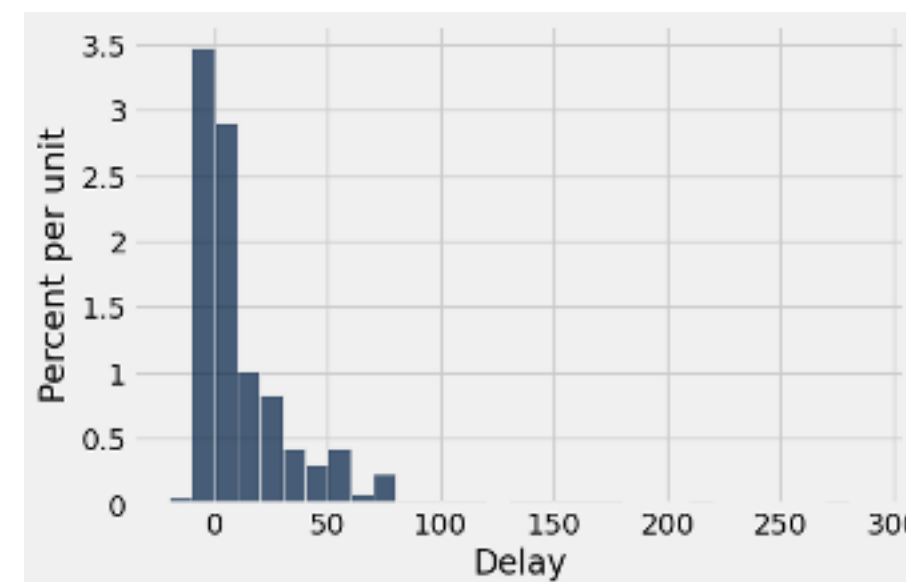
Population



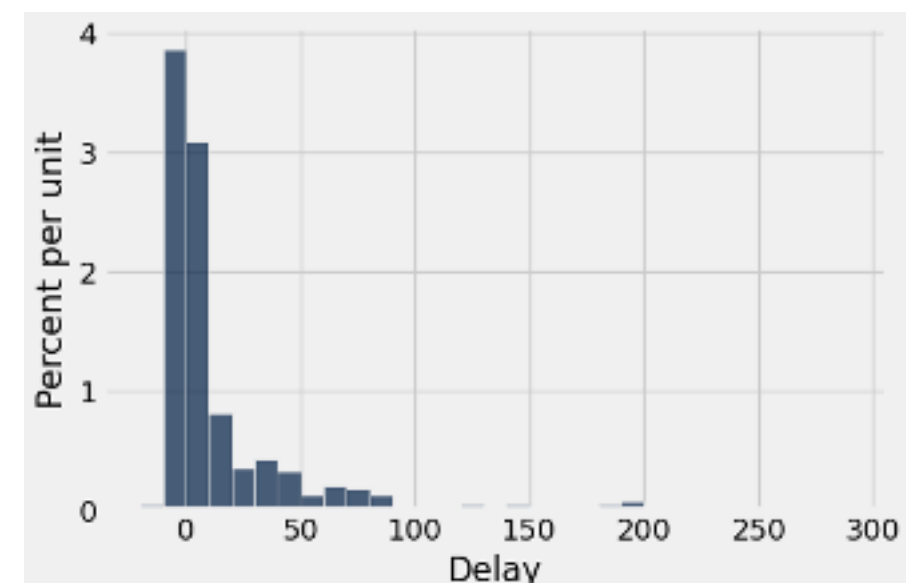
Population mean = 16.65

SD = 39.5

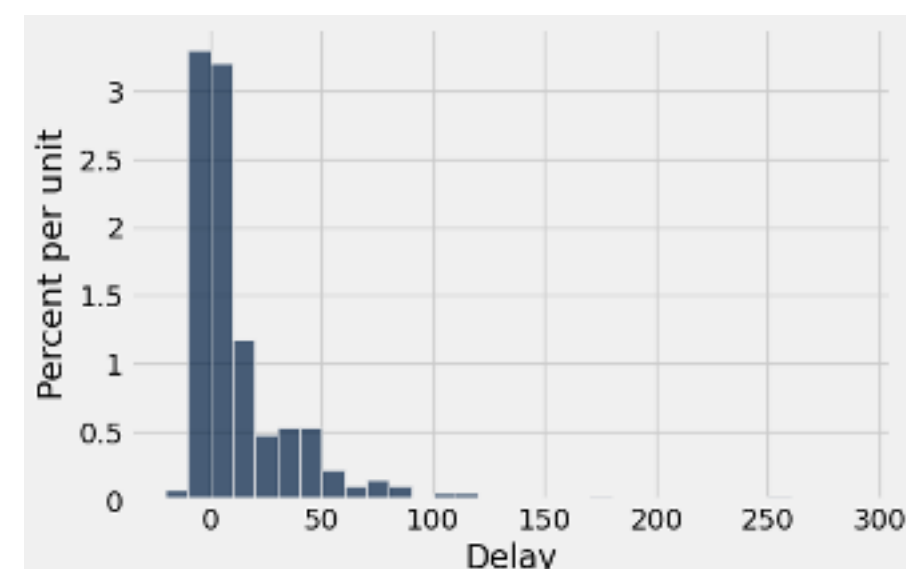
Sample
Size = 400



$$\mu_1 = 15.33$$



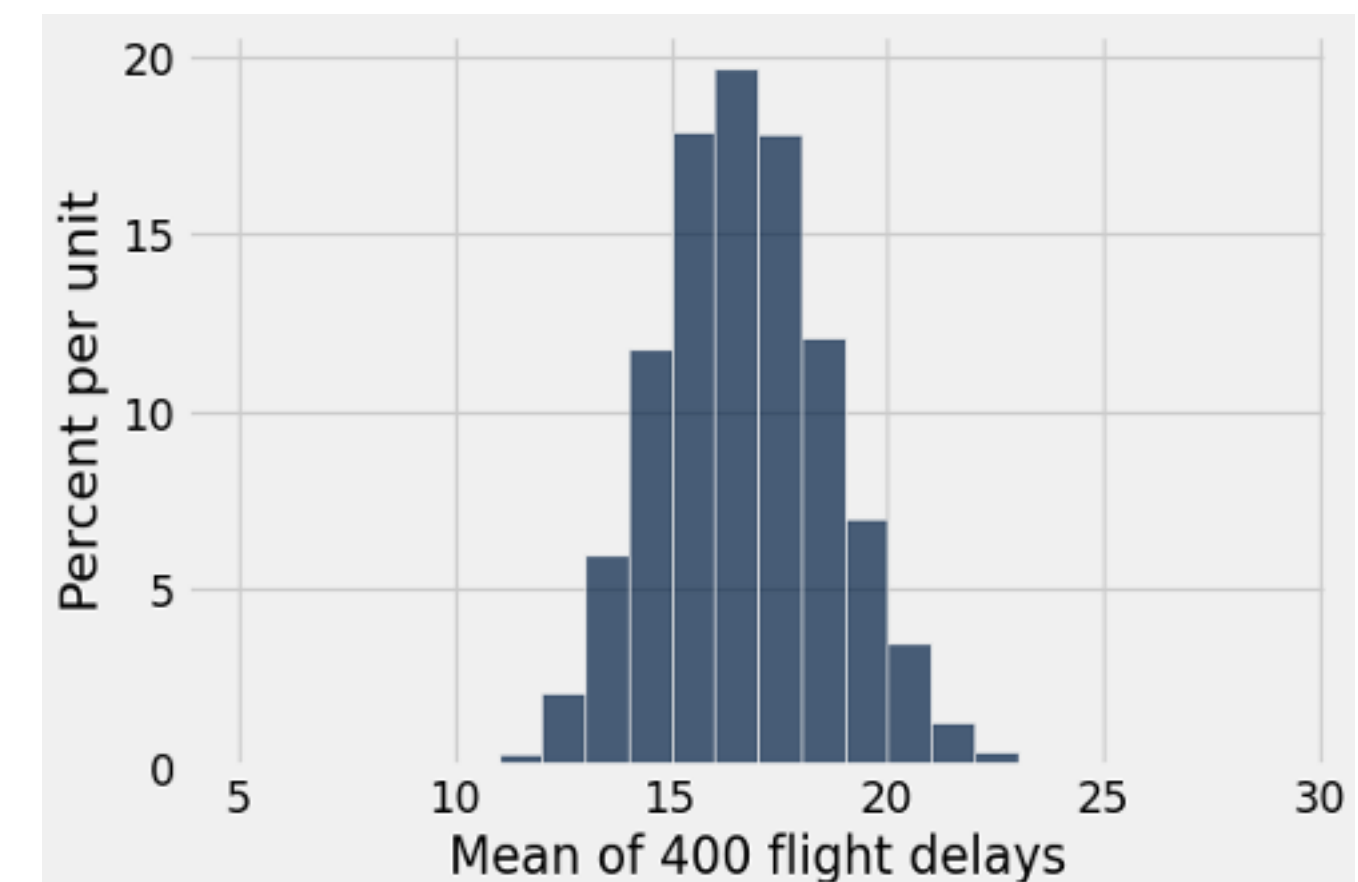
$$\mu_2 = 16.2$$



$$\mu_n = 18.95$$

Statistic
Sample mean

Distribution of
sample means



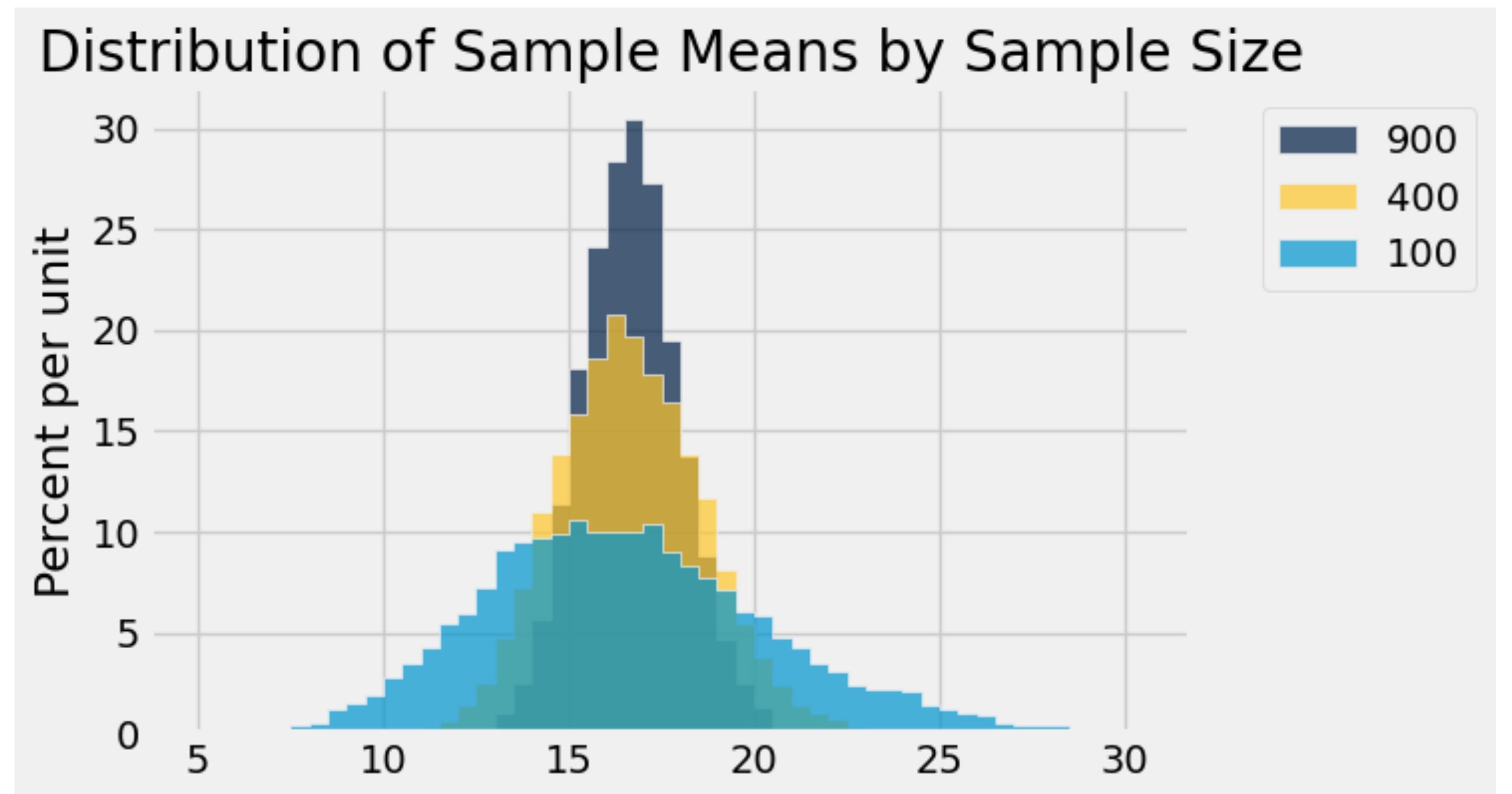
Sample average mean = 16.67

SD = 1.98

Distribution of Sample Means & Sample Size

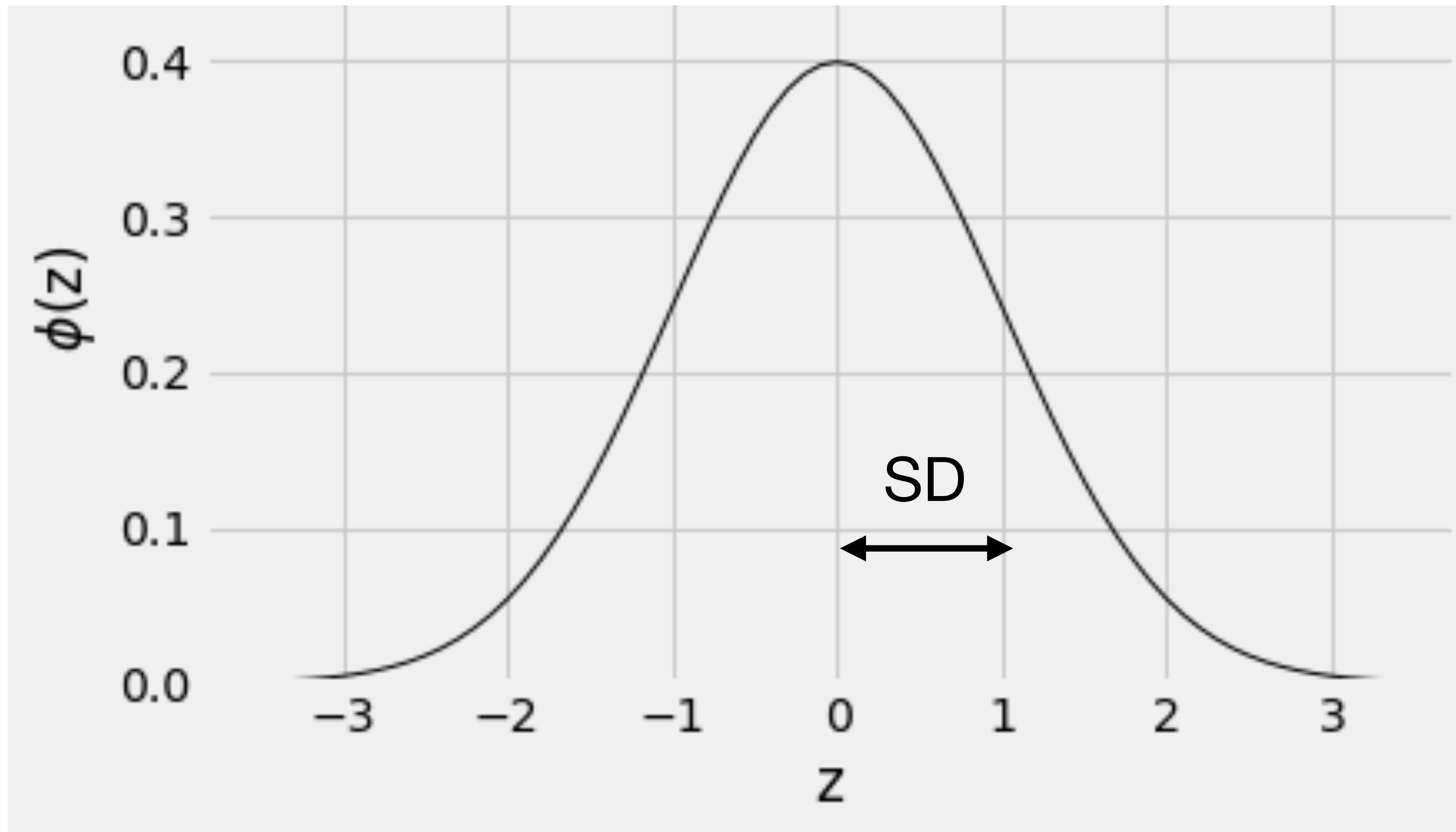
When increasing sample size,

- Distributions get narrower (closer to the true mean)
- Also get taller (higher probability around true mean)



Connecting to confidence intervals

Distribution of Sample Averages (for Sample Size n)



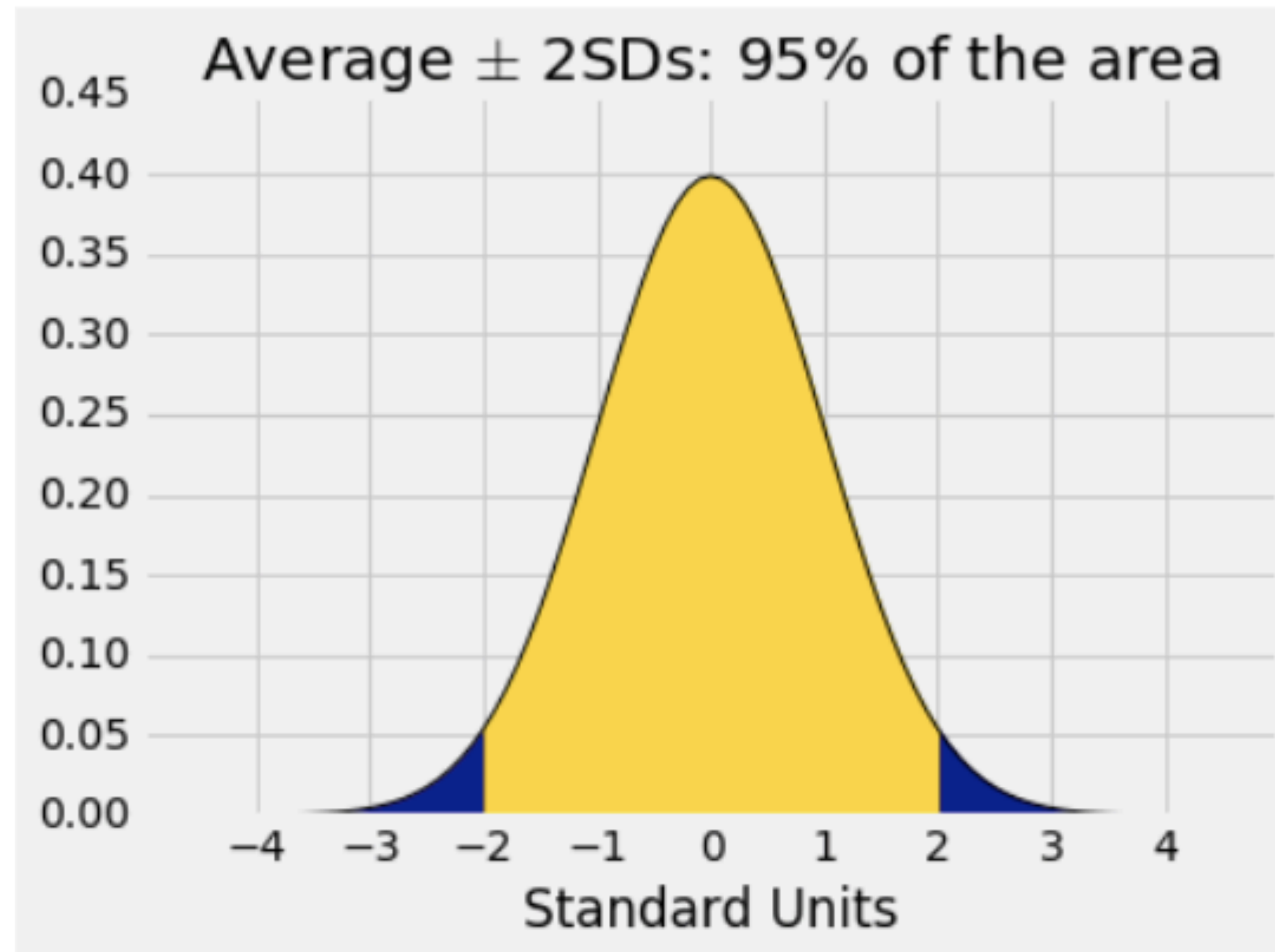
$$SD = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Normal vs All Distributions (Chebyshev's)

Range	All Distributions (Chebyshev's)	Normal Distribution
mean \pm 1 SDs	At least 0%	At least 68%
mean \pm 2 SDs	At least 75%	At least 95%
mean \pm 3 SDs	At least 89%	At least 99%

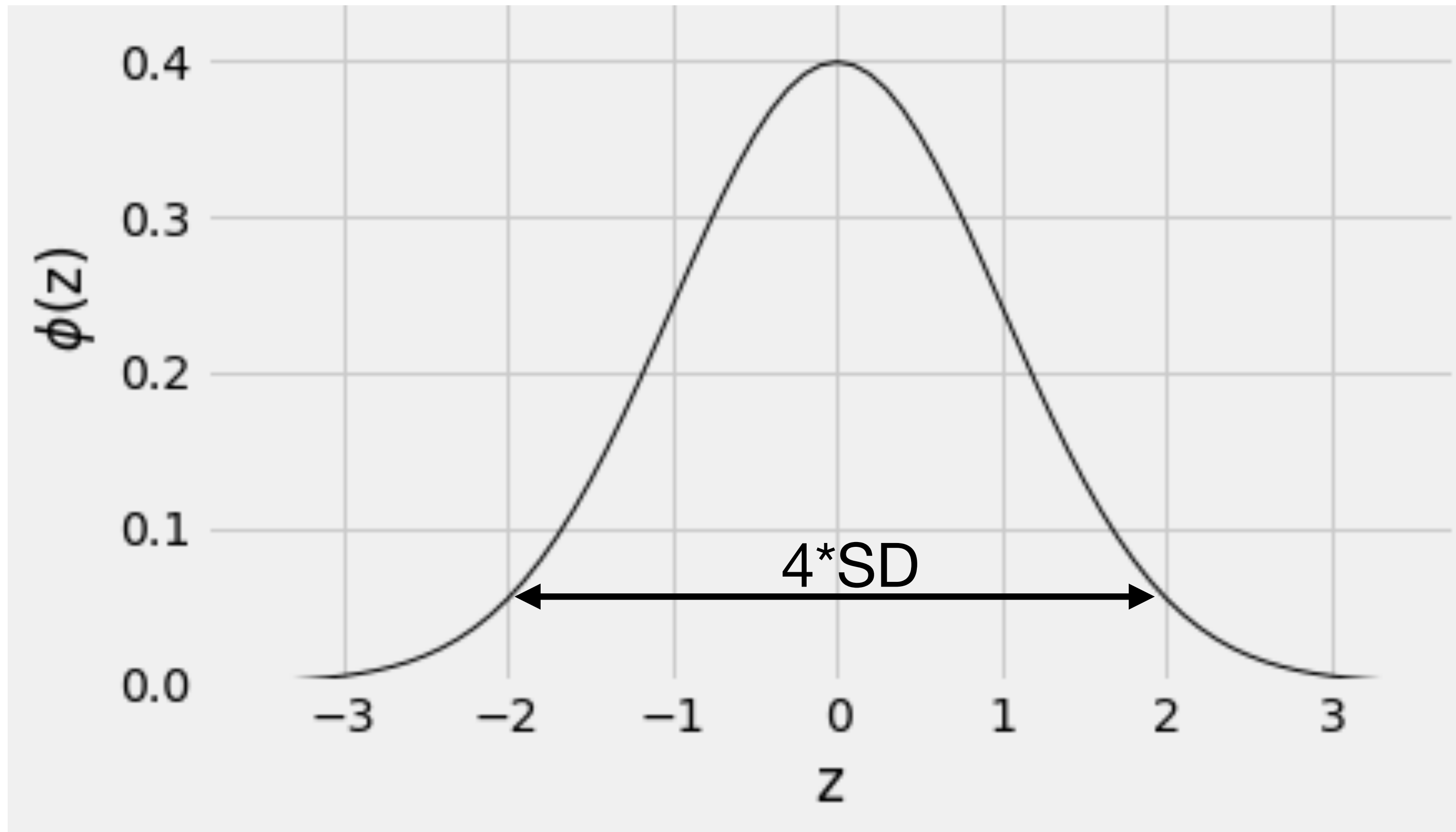
95% Confidence in Normal Distributions

- For a normal distribution, 95% of the data is in the range average \pm 2SDs



Connecting to confidence intervals

95% of the data is in the range average + 2SDs



width of the 95%
confidence interval =

$$4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Controlling the width

Total width of the 95% confidence interval

$$= 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- The more narrow the interval, the more precise your estimate

Suppose you want to bound the total width of an interval (say, at most 1%)

- How should you choose the sample size?

How large of a sample do we need?

- Recall for normal distributions: Width of 95 % confidence interval = 4 * SD
- Central Limit Theorem:

$$SD = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

How large of a sample do we need?

- Recall for normal distributions: Width of 95 % confidence interval = 4 * SD
- Central Limit Theorem:

$$SD = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$
$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

How large of a sample do we need?

- Recall for normal distributions: Width of 95 % confidence interval = 4 * SD
- Central Limit Theorem:

$$SD = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$
$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

Sample Proportions

- There's interesting data which is just 0 or 1
- Examples:
 - Polling for Candidate A vs Candidate B
 - Cancerous cells vs no cancerous cells
 - Disease vs no disease

Sample Proportions

Suppose we polled 10 voters on if they voted for Candidate A (0) or Candidate B (1)

Voter	1	2	3	4	5	6	7	8	9	10
	0	0	0	0	0	0	0	0	0	1

- What is the proportion who voted for B?

$$\frac{1}{10} = 0.1$$

Sample Proportions

Suppose we polled 10 voters on if they voted for Candidate A (0) or Candidate B (1)

Voter	1	2	3	4	5	6	7	8	9	10
	0	0	0	0	0	0	0	0	0	1

- What is the proportion who voted for B? $1/10 = 0.1$

Voter	1	2	3	4	5	6	7	8	9	10
	0	0	0	0	0	1	0	0	0	1

- What is the proportion who voted for B? $2/10 = 0.2$

Proportions are Averages

If the population consists of 1s and 0s (or yes/no), then the population average is the proportion of 1s in the population

- Likewise the sample average is the proportion of 1s in the sample

Voter	1	2	3	4	5	6	7	8	9	10
	0	0	0	0	0	1	0	0	0	1

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	
1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

$$\sigma = \sqrt{\frac{(1 - 0.1)^2 + 9 * (0. - 0.1)^2}{10}}$$

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

$$\begin{aligned} \sigma &= \sqrt{\frac{(1 - 0.1)^2 + 9 * (0. - 0.1)^2}{10}} \\ &= \sqrt{\frac{0.81 + 9 * 0.01}{10}} \end{aligned}$$

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

$$\begin{aligned} \sigma &= \sqrt{\frac{(1 - 0.1)^2 + 9 * (0. - 0.1)^2}{10}} \\ &= \sqrt{\frac{0.81 + 9 * 0.01}{10}} \\ &= \sqrt{\frac{0.9}{10}} \end{aligned}$$

Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	0.3
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

$$\begin{aligned} \sigma &= \sqrt{\frac{(1 - 0.1)^2 + 9 * (0. - 0.1)^2}{10}} \\ &= \sqrt{\frac{0.81 + 9 * 0.01}{10}} \\ &= \sqrt{\frac{0.9}{10}} \\ &= 0.3 \end{aligned}$$

Proportion and SDs

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	0.3
2	0.2	0.4
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1	

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

$$\begin{aligned} \sigma &= \sqrt{\frac{2 * (1 - 0.2)^2 + 8 * (0. - 0.2)^2}{10}} \\ &= \sqrt{\frac{2 * 0.64 + 8 * 0.04}{10}} \\ &= \sqrt{\frac{1.6}{10}} \\ &= 0.4 \end{aligned}$$

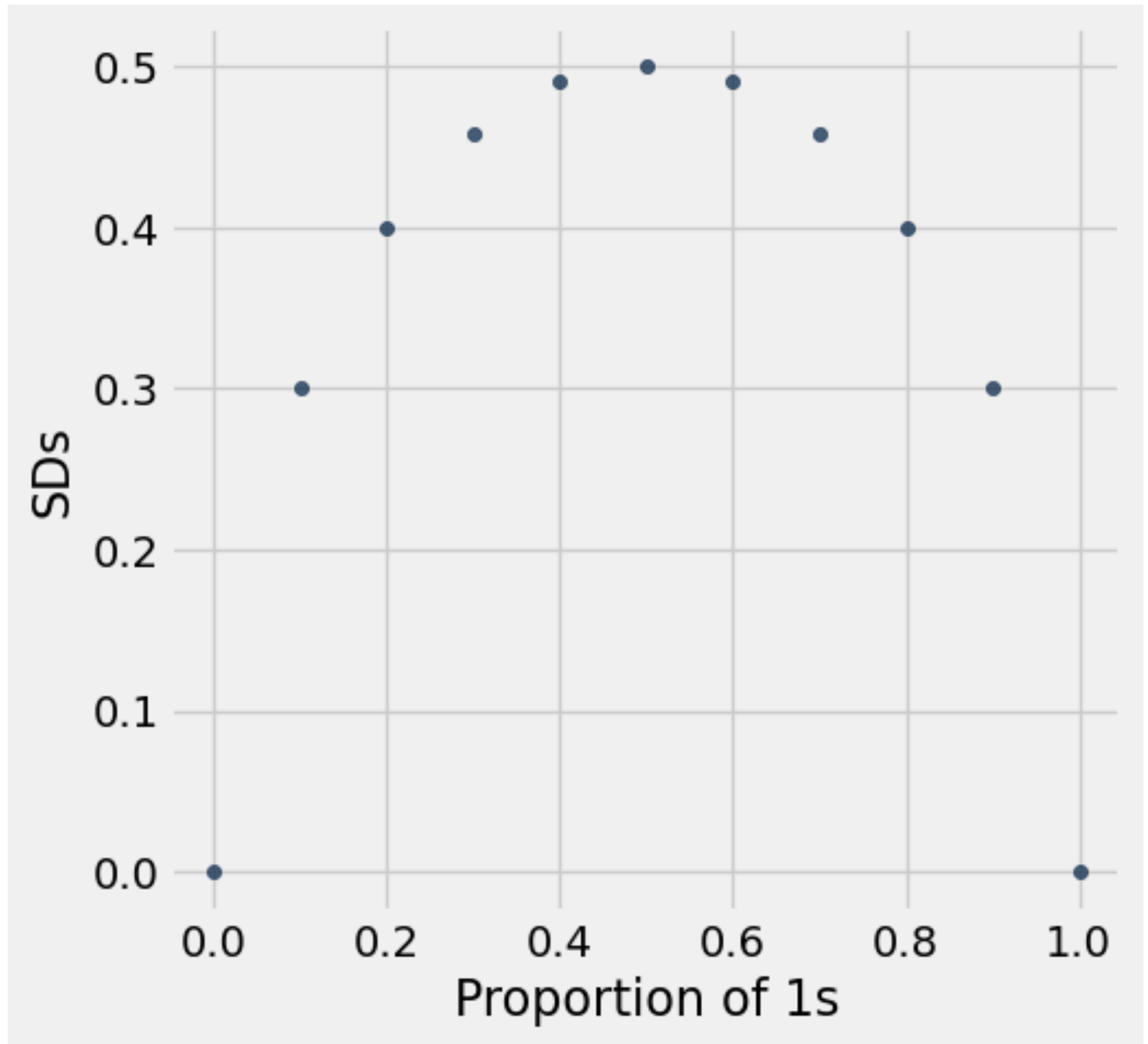
Proportion and SDs

$$\sigma = \sqrt{\text{avg} \left((v - \mu)^2 \text{ for } v \in \vec{V} \right)}$$

Number of 1s	Proportion of 1s	SDs
0	0	0
1	0.1	0.3
2	0.2	0.4
3	0.3	0.458258
4	0.4	0.489898
5	0.5	0.5
6	0.6	0.489898
7	0.7	0.458258
8	0.8	0.4
9	0.9	0.3
10	1	0

Population SD for Situations with 2 Outcomes

- For situations with only 2 outcomes, the SD ranges from 0 to 0.5, with a max value of 0.5
- Thus, to estimate worst case scenario (most conservative sample size needed), you can use the maximum $SD=0.5$



Example: Polling Sample Size

- Two candidates are up for election: Candidate A and Candidate B
- Candidate A wants to estimate with a 95% confidence interval what % of voters will vote for her
- How large of a sample should the candidate poll if they want to make this estimate with a desired accuracy of no wider than 1%?
 - Example: 95% confidence interval of (44%, 44.5%) is ok, but (44%, 46%) would be too inaccurate

Example: Polling Sample Size

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

Example: Polling Sample Size

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

Example: Polling Sample Size

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

Example: Polling Sample Size

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

$$\text{sample size} \geq 200^2$$

Example: Polling Sample Size

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

$$\text{sample size} \geq 200^2$$

$$\text{sample size} \geq 40,000$$

Next time

- Prediction!!
- Correlation
- Linear regression