Sampling Distributions

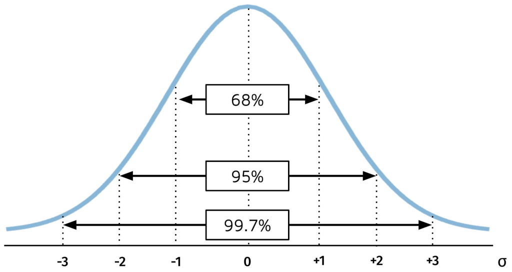
Lecture 6

Emma Ning, M.A.

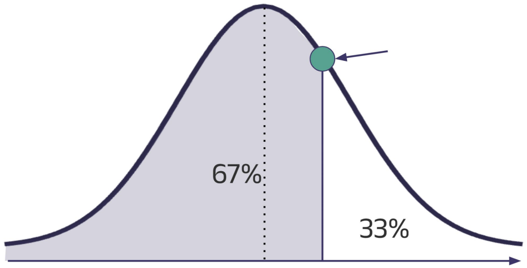
From our last lecture...

- Calculated z-scores $z = \dfrac{x - \mu}{\sigma}$

- Empirical Rule



- Probability



Throughout this, have you ever wondered...
Why are we using population parameters?

Didn't we say we almost never know the population?

Yes... You are right!

TODAY'S PLAN

01 Revisiting Sampling Error

02 Sampling Distribution

03 Distribution of Sample Means & Standard Error

04 Central Limit Theorem

05 Wrap Up

Learning objectives

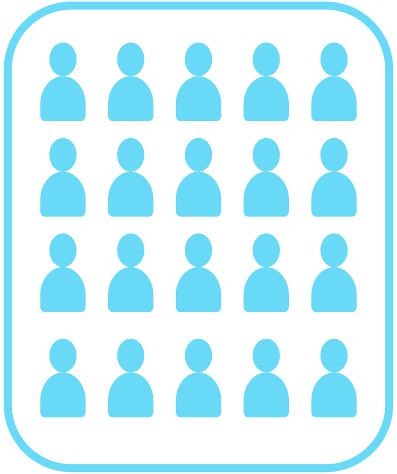
- Explain and differentiate between a **sample** vs. **sampling distribution**
- Explain and differentiate **sampling error** and **standard error**
- Describe the **distribution of sample means** (a sampling distribution)
- Calculate **standard error of the mean**, explain what it measures, and describe how it changes if the population standard deviation or sample size changes (numerator and denominator in its calculation)
- Describe the **central limit theorem** and explain why it is important in statistics

01

Revisiting Sampling Error

Sampling

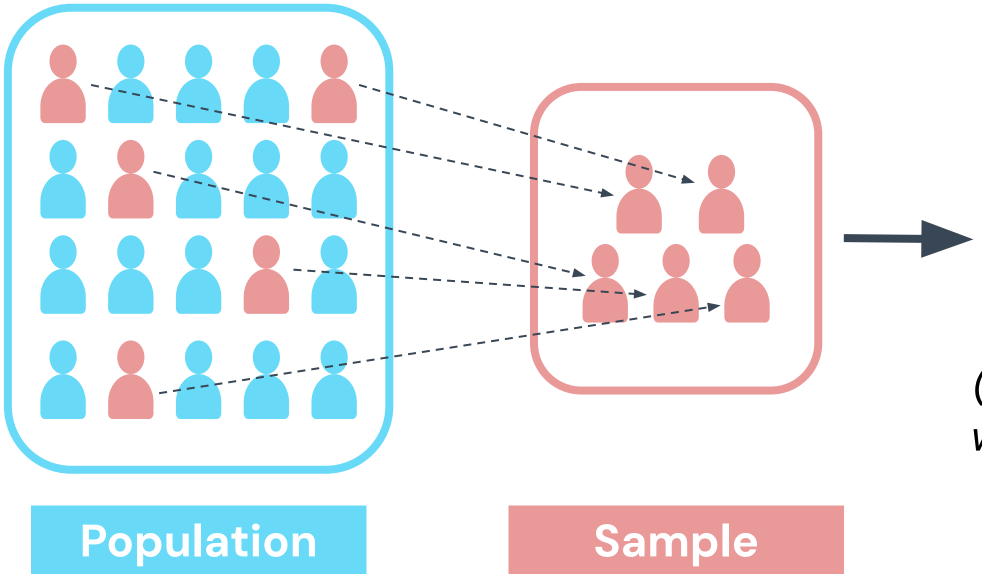
the process of selecting of a subset of the population for the purposes of a research study



Population

Sampling

the process of selecting of a subset of the population for the purposes of a research study



randomly selected

sampling error occurs

*(The whole population will **never** be perfectly represented by a sample)*

Sampling

In other words…

Every spoonful (**sample**) will have a slightly different mix of celery and carrots, and rarely match the exact average in the whole pot (**population**).

This mismatch <u>between sample statistics and population statistics</u> is **sampling error**. It is also referred to as **chance error**.

One more recap... Why do we care about samples?

We don't care about the sample for its own sake — we care about the **population**. However, the population is too hard/large/expensive to get. So we hope the sample would give us a snapshot of the population.

Just like the soup pot — we care about how the entire pot tastes. To make sure it's good for guests, we stir it well and try a spoonful, rather than tasting or drinking the whole thing.

This is called making an **inference** about the population.

Statistics

Descriptive Statistics

used to **summarize** and **describe** data

Mean, Median, SD...

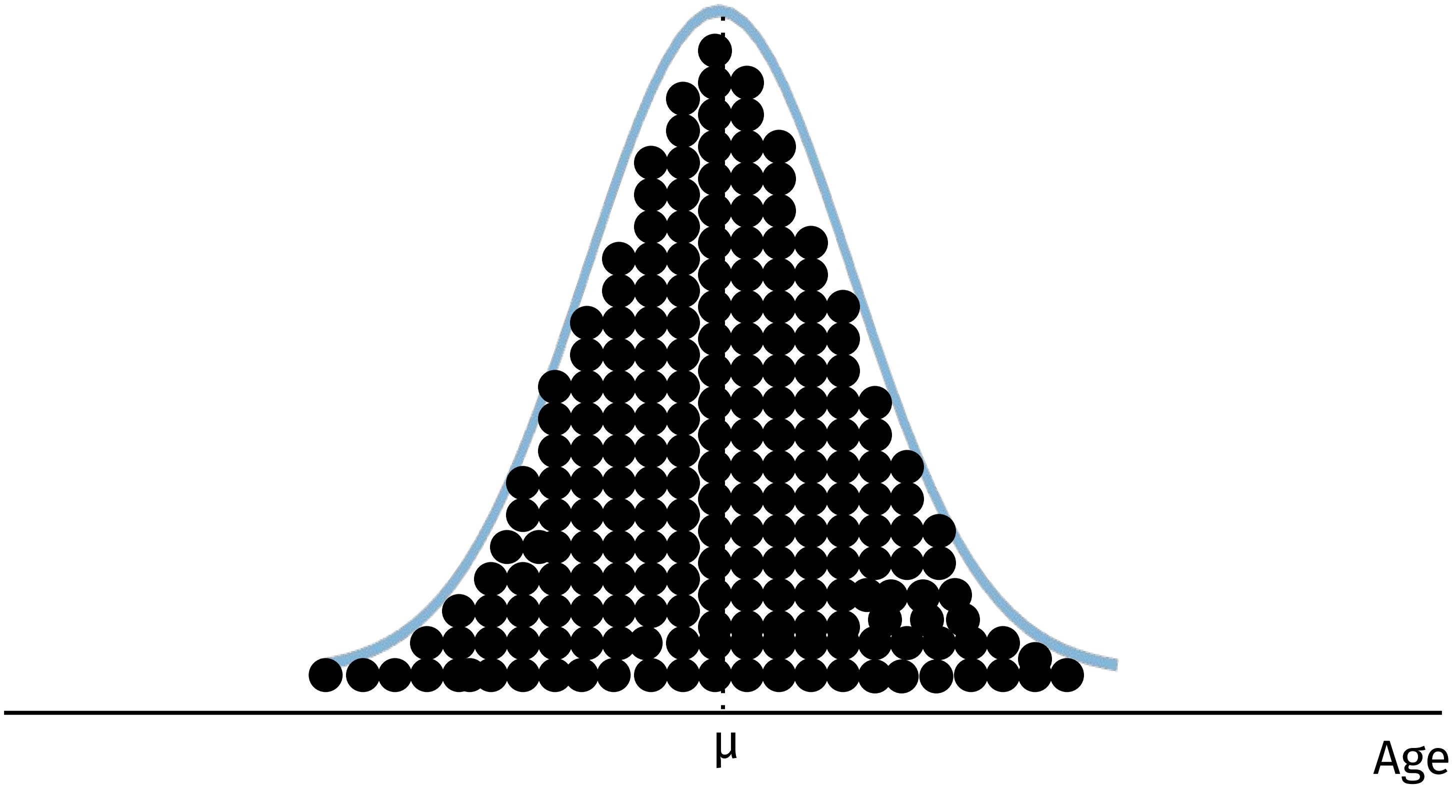
Inferential Statistics

techniques used to **make generalizations** about samples and apply them to populations

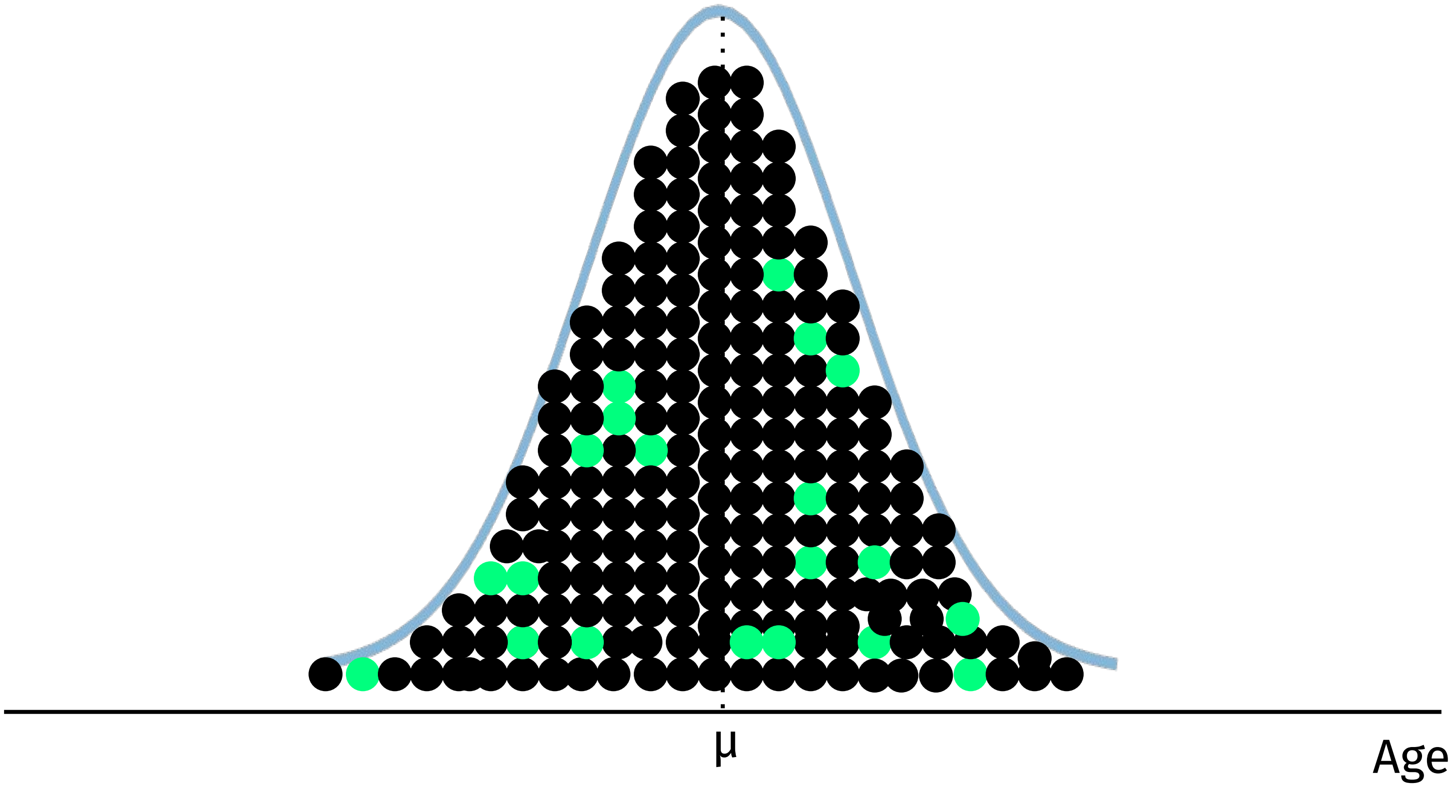
02

Sampling Distribution

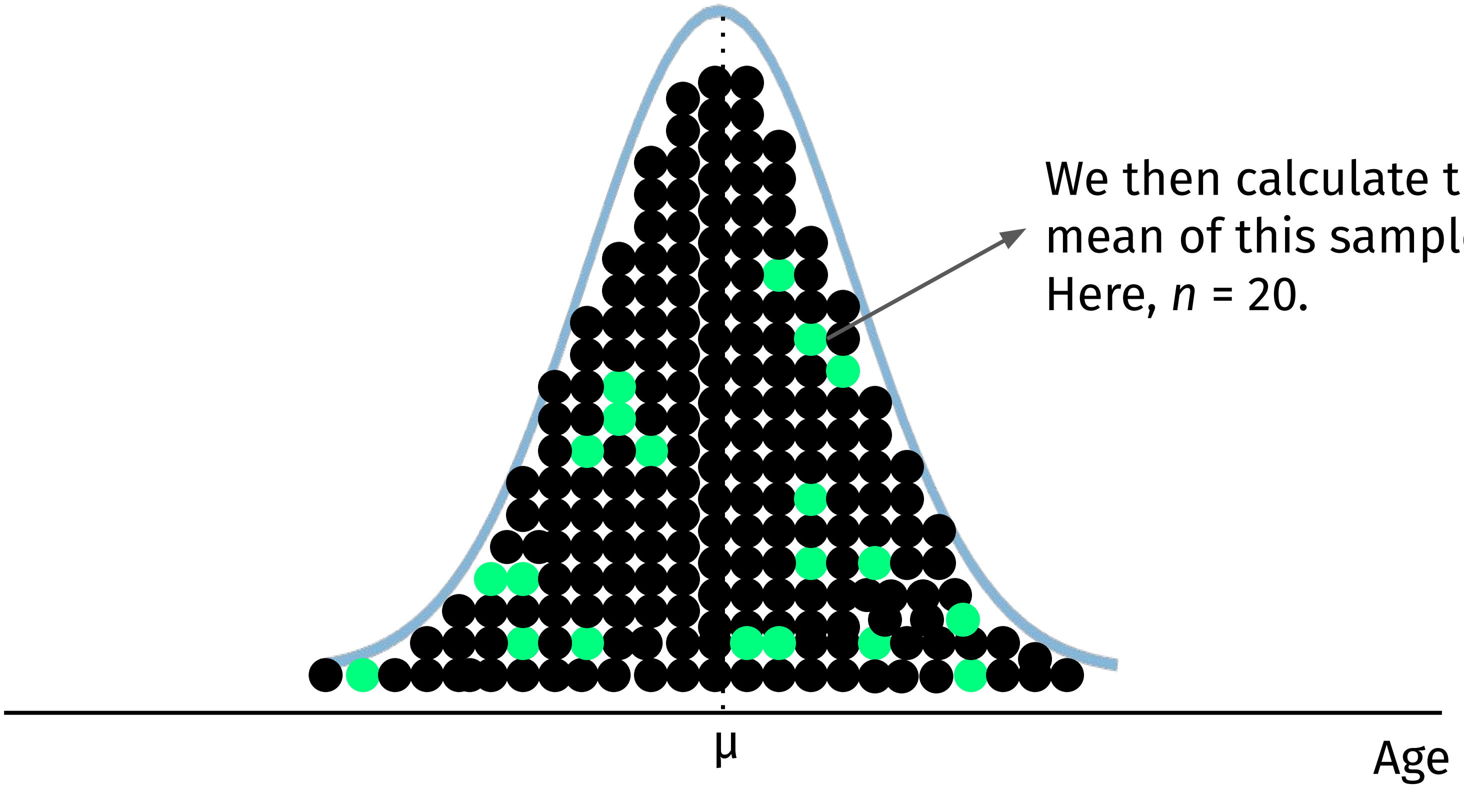
Our population of student ages



Step 1: Sample from our population



Step 2: Calculate our sample mean



We then calculate the mean of this sample. Here, *n* = 20.

Step 3: Keep a tally of our sample means

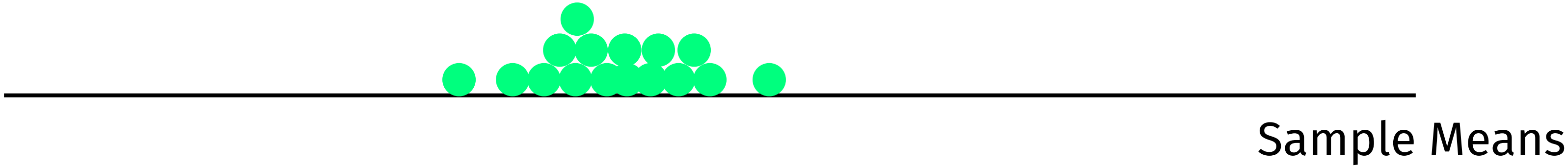
To get our distribution of sample means, we repeat Steps 1-3 many many times. Think hundreds and thousands of times.

Say this is the sample mean from our *n* = 20.

Sample Means

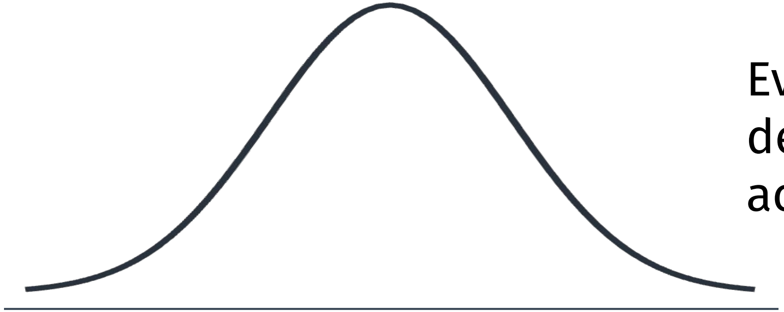
Step 3: Keep a tally of our sample means

Here I repeated this process 16 times, because there are 16 dots.
Each dot represents the mean of a sample.
Remember, for each sample, n = 20.



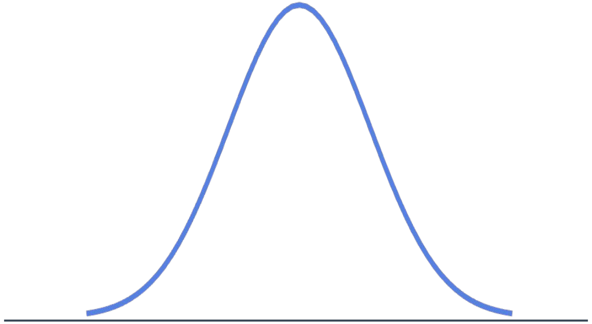
What do we get after repeating this process?

Population Distribution



Every dot/data point denotes **one person**'s actual score (e.g., age).

Distribution of **Sample means**



Every dot/data point denotes the **mean of one sample** (with many people in it).

What do we get after repeating this process?

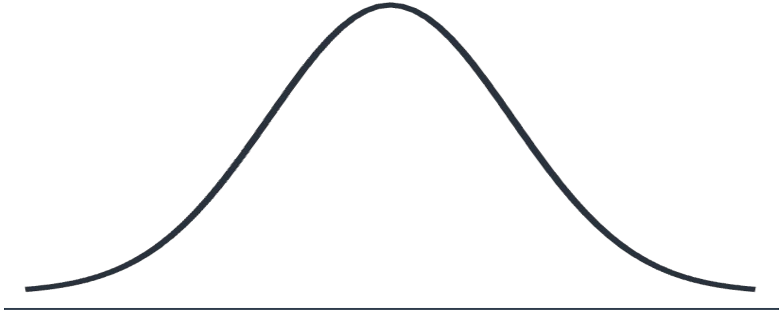
This repeated process of drawing samples of a fixed size from a population, calculating the mean for each sample, and keeping track of those means across many samples builds what we call a "sampling distribution".

Distribution of Sample means

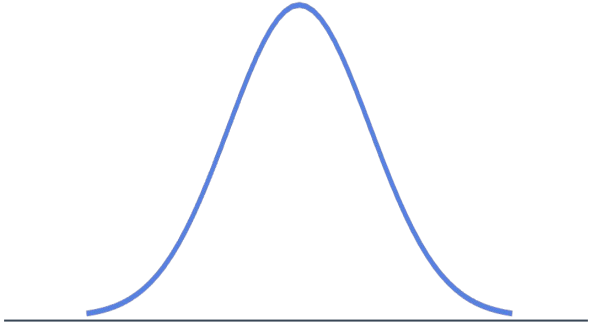
Every dot/data point denotes the **mean of one sample** (with many people in it).

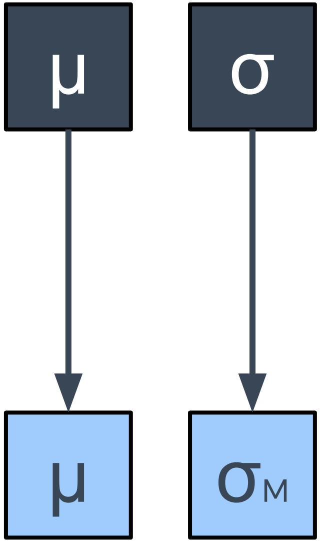
What do we get after repeating this process?

Population
Distribution



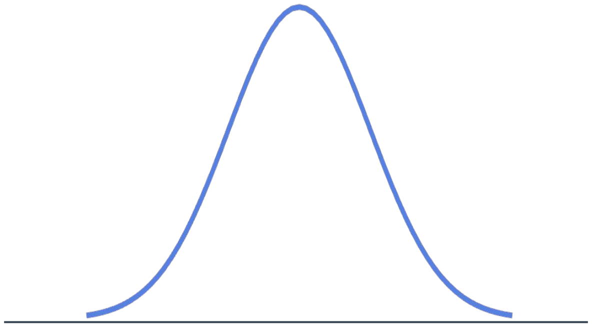
Distribution of
Sample means
(aka the sampling distribution of the mean)





What does the distribution of sample means mean?

Distribution of
Sample means
(aka the sampling distribution of the mean)



μ

σ_M

Every dot/data point denotes the **mean of one sample** (with many people in it).

The sampling distribution shows the variability of sample means from sample to sample, if we repeatedly drew samples of the same size from the population.

Why do we care about sampling distribution?

Like we said before, we usually only get to work with samples.
The sampling distribution shows us **how much** sample statistics, like the mean, **vary** from sample to sample.

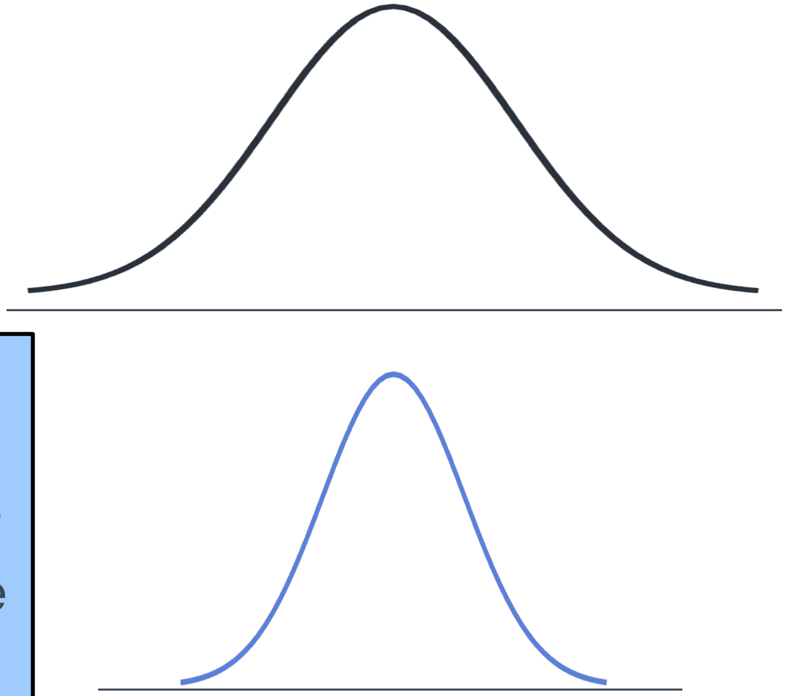
This variability matters because in real life, we almost always have just one sample. So we want to know: **how much can we trust that one snapshot**?

The more variable the sampling distribution, the less confident we are.
Vice versa.
Therefore, we are really interested in the **standard deviation of the sampling distribution.**

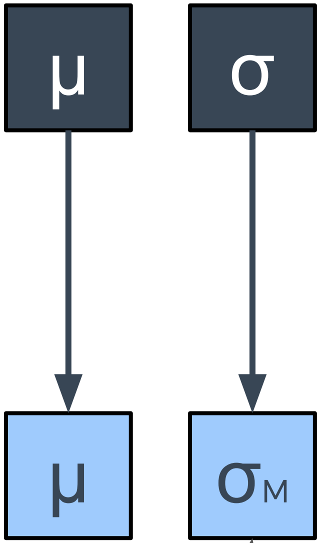
03

Distribution of Sample Means & Standard Error

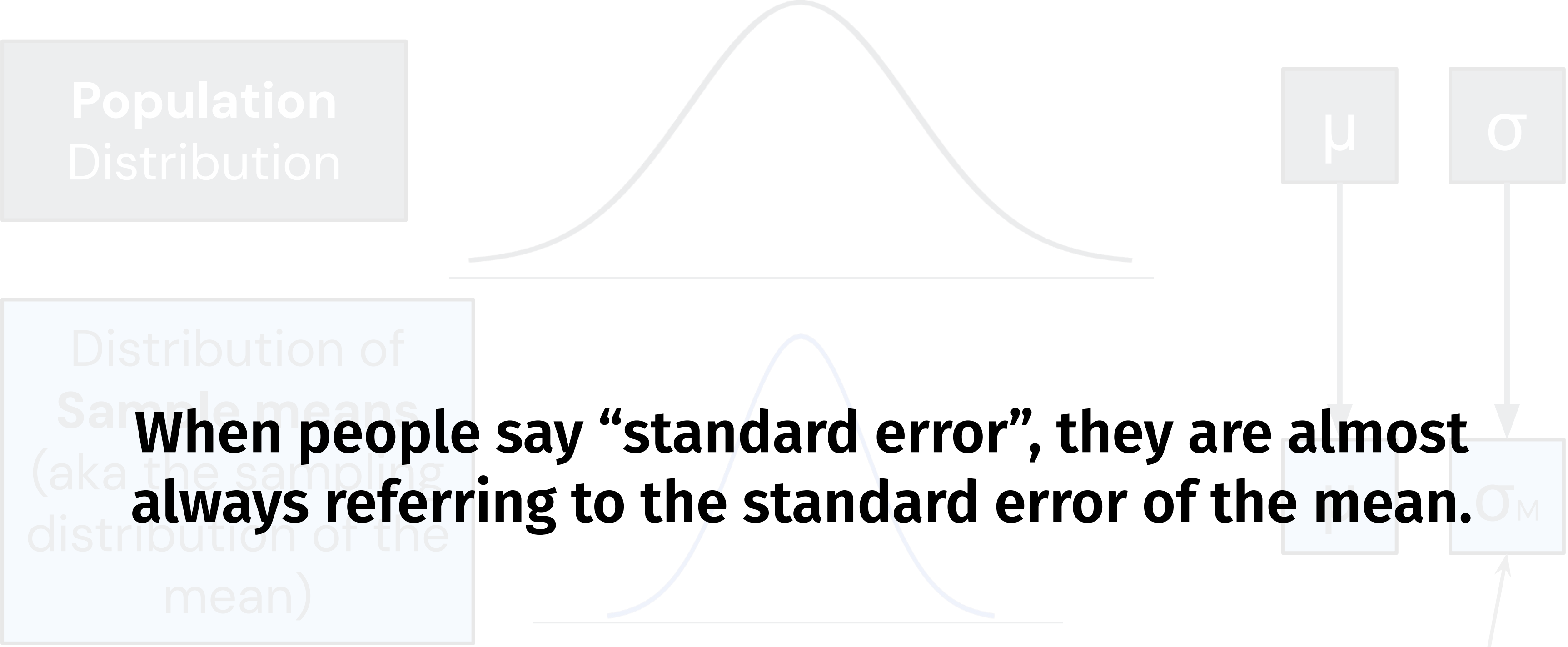
Population Distribution



Distribution of **Sample means** (aka the sampling distribution of the mean)



Standard error is the standard deviation of the sampling distribution. Since we are looking at the sampling distribution of the mean, the standard deviation of the lower curve, is called the **standard error of the mean (most commonly referred to as SE).**

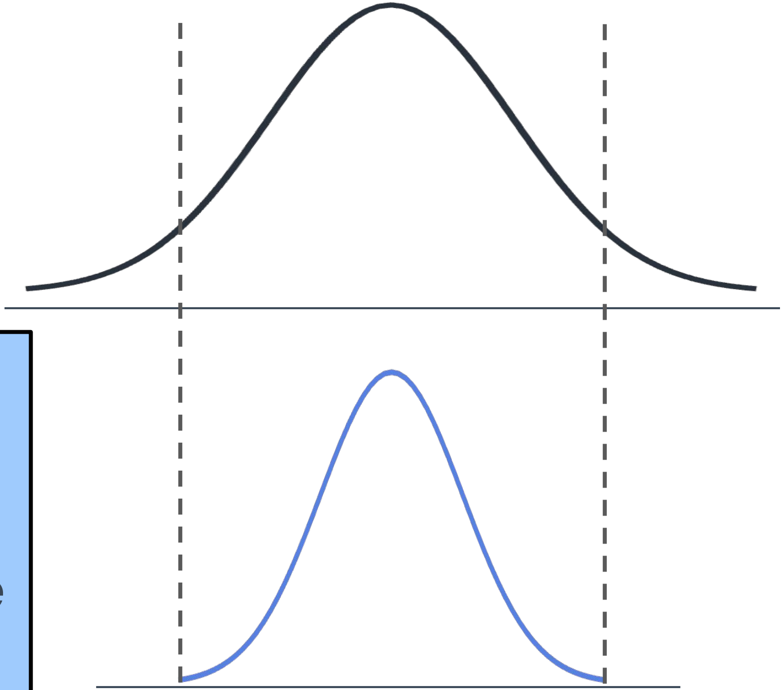


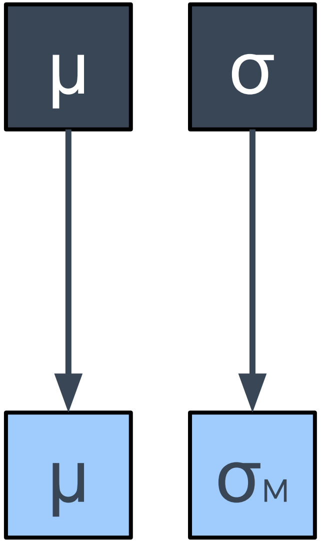
When people say “standard error”, they are almost always referring to the standard error of the mean.

Standard error is the standard deviation of the sampling distribution. Since we are looking at the sampling distribution of the mean, the standard deviation of the lower curve, is called the **standard error of the mean (most commonly referred to as SE).**

Population Distribution

Distribution of **Sample means** (aka the sampling distribution of the mean)





Wait, why is $\sigma_M < \sigma$?

Calculating Standard Error

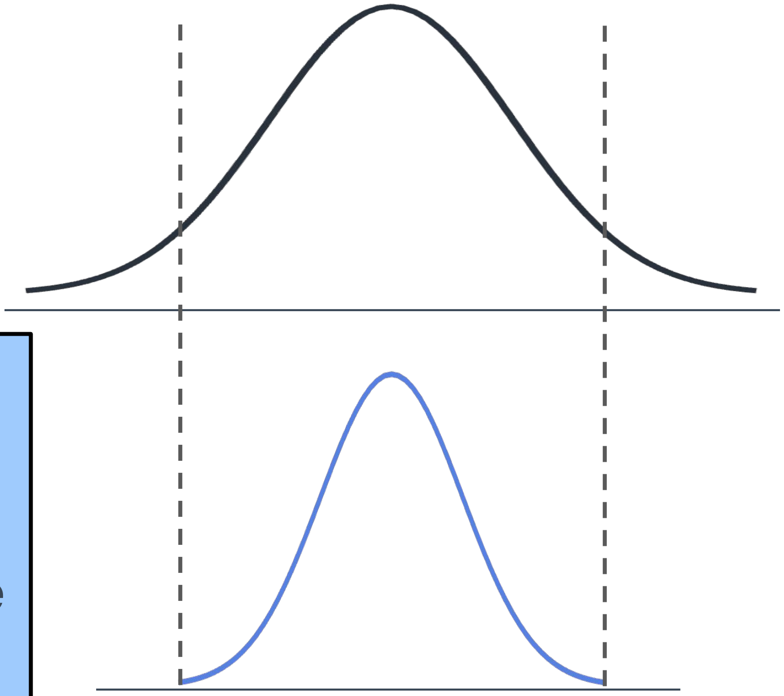
Population
standard deviation

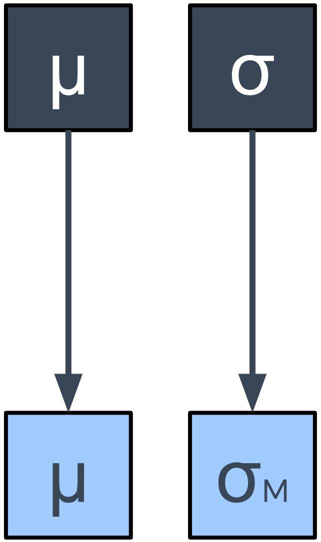
$$SE = \frac{\sigma}{\sqrt{n}}$$

Square root of
the sample size

Population Distribution

Distribution of **Sample means** (aka the sampling distribution of the mean)





$$SE = \frac{\sigma}{\sqrt{n}}$$

From this equation, we see that *SE* = σ if and only if n = 1. And that never happens. So σ_M is always smaller than σ. (SE is σ_M, they are interchangeable).

What else does the SE equation tell us?

$$SE = \frac{\sigma}{\sqrt{n}}$$

What happens to standard error as the **sample size increases**?

If we want to reduce the variability in our sampling distribution of the mean (aka SE), we should always aim for a <u>large sample size</u>.

Remember, we said earlier "This variability matters because in real life, we almost always have just one sample. So we want to know: how much can we trust that one snapshot?"

Having more people in your sample makes you trust your snapshot more. It represents the population more accurately.

SD of sampling distribution (aka Standard Error)

Think of the population as a pot of soup. Some soups—like tomato bisque—are really uniform. No matter where you sample, it'll taste the same. Others—like chicken noodle—are chunkier and more varied. One spoon might be more "carroty", another spoon might be more "chickeny".

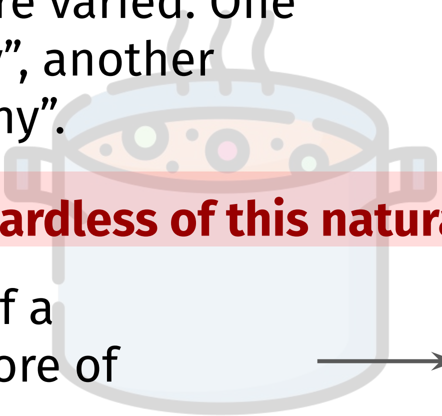
Sample statistics—like the sample mean—naturally bounce around from sample to sample.
How much they bounce around depends on the population's variability.

$$\downarrow SE = \frac{\sigma}{\sqrt{n}} \downarrow$$

SD of sampling distribution (aka Standard Error)

Think of the population as a pot of soup. Some soups—like tomato bisque—are really uniform. No matter where you sample, it'll taste the same. Others—like chicken noodle—are chunkier and more varied. One spoon might be more “carroty”, another spoon might be more “chickeny”.

Sample statistics—like the sample mean—naturally bounce around from sample to sample. How much they bounce around depends on the population's variability.



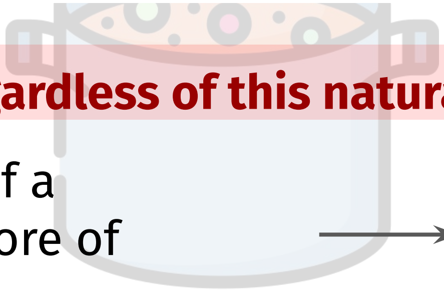
However, regardless of this natural variability:

If we took a ladleful instead of a spoonful, we'd average out more of the chunks—it'd taste more consistent.

Larger samples = Less variability

SD of sampling distribution (aka Standard Error)

$$SE = \frac{\sigma}{\sqrt{n}}$$



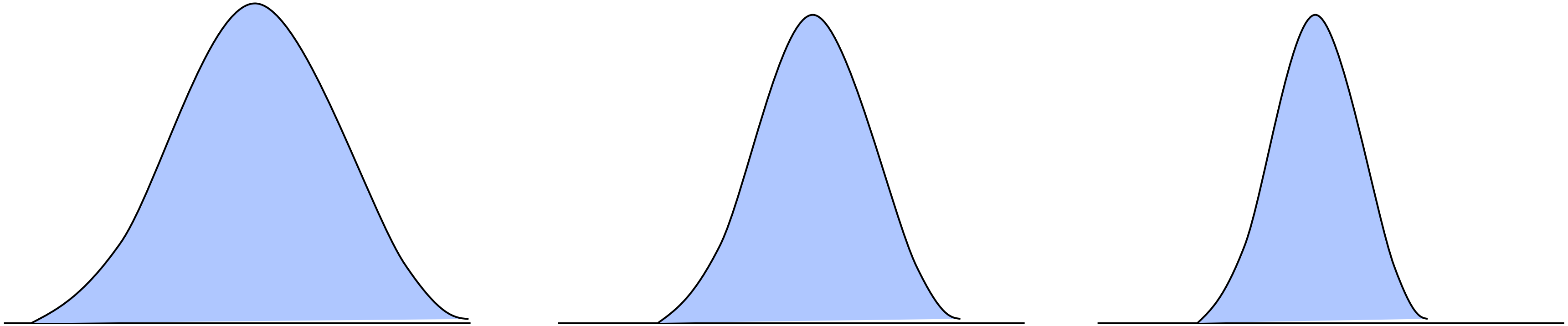
However, regardless of this natural variability:

If we took a ladleful instead of a spoonful, we'd average out more of the chunks—it'd taste more consistent.

→ Larger samples = Less variability

$$SE = \frac{\sigma}{\sqrt{n}}$$

$\sigma = 8$



n = 5 | n = 10 | n = 30

$$SE = \frac{8}{\sqrt{5}} \qquad SE = \frac{8}{\sqrt{10}} \qquad SE = \frac{8}{\sqrt{30}}$$

SE = **3.58** | SE = **2.53** | SE = **1.46**

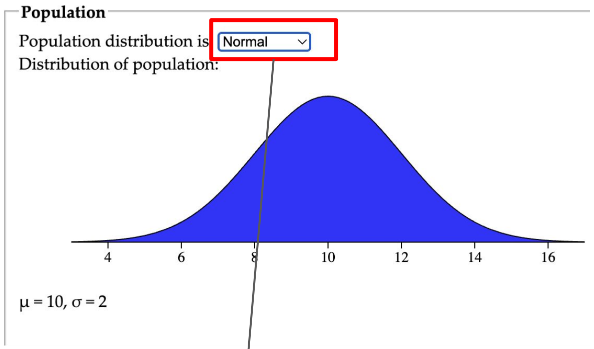
True or False?

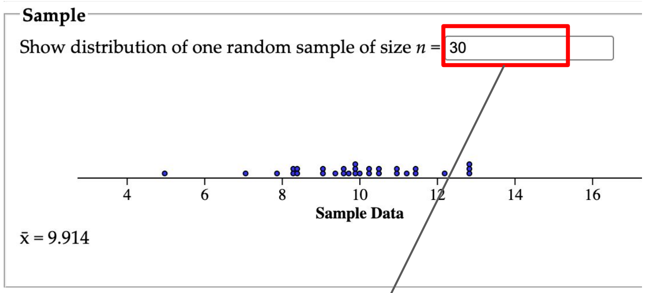
Discuss each statement with your table.

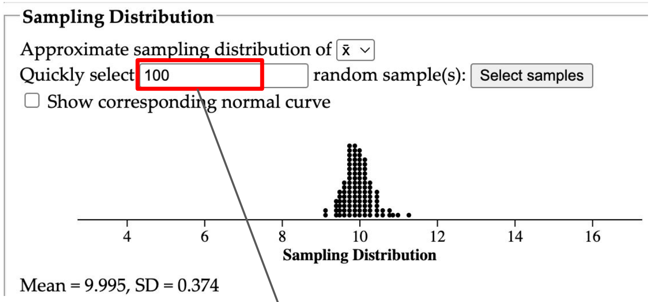
1	The standard error provides a method for defining and measuring sampling error.	T
2	As sample size (n) increases, the size of the standard error increases.	F
3	Standard deviation and standard error are the exact same thing.	F
4	When the sample consists of a single score ($n = 1$), the standard error is the same as the standard deviation ($\sigma_M = \sigma$).	T

ICA6: Let’s pause, and play around with this:

https://www.stapplet.com/sampdist.html







Let’s set it to normal for now.

Play around with this number. Notice the number of dots & how they stack up.

Play around with this number. Notice the number of dots & how they stack up.

Make sure you can explain the differences between sample vs. sampling distribution.

04

Central Limit Theorem

Central Limit Theorem

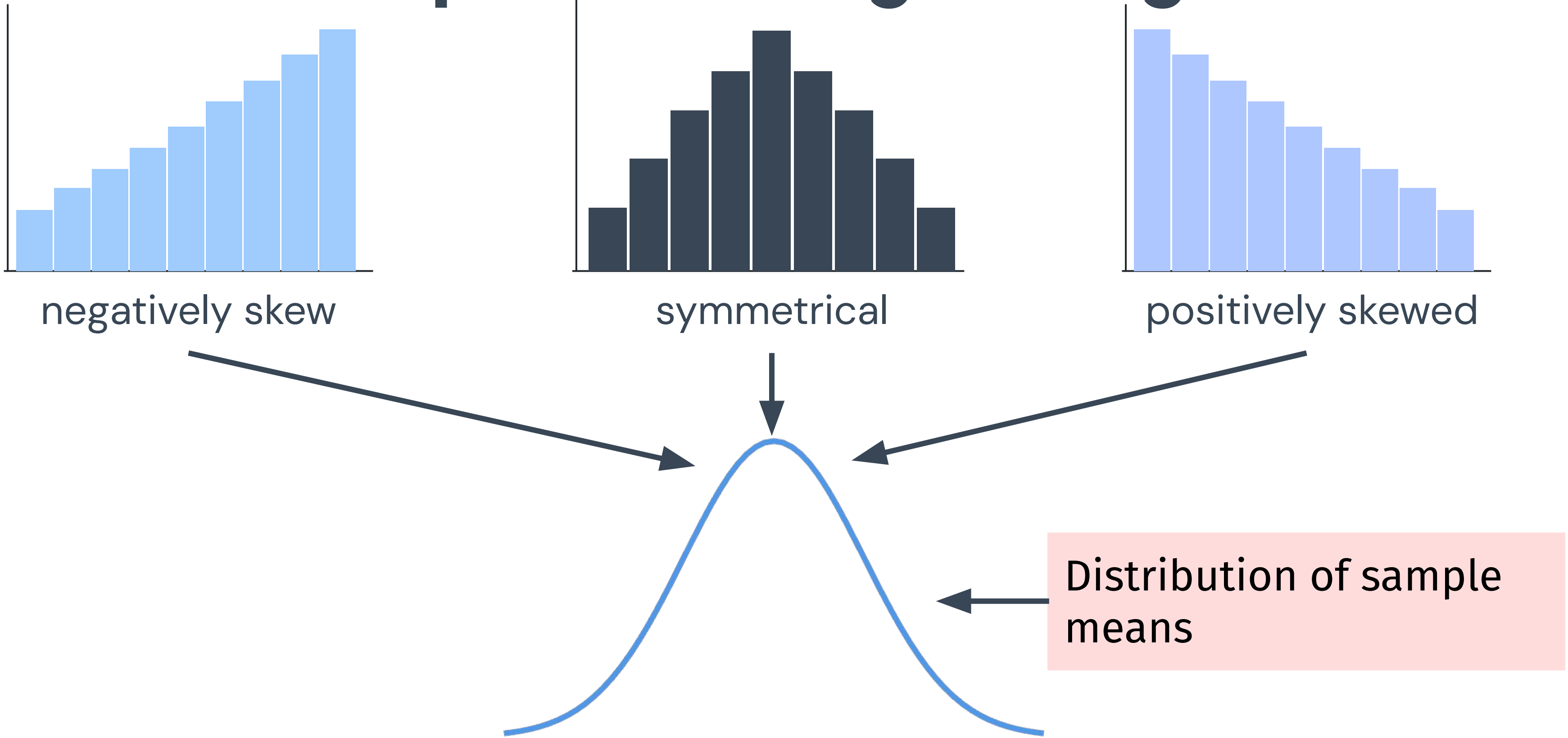
Formal Definition

For any population with a mean and standard deviation, the **distribution of sample means will approximate a normal curve** with sufficiently large samples, usually of **30** of greater. The shape of the distribution of samples means will get closer to the shape of the normal curve as the sample size increases.

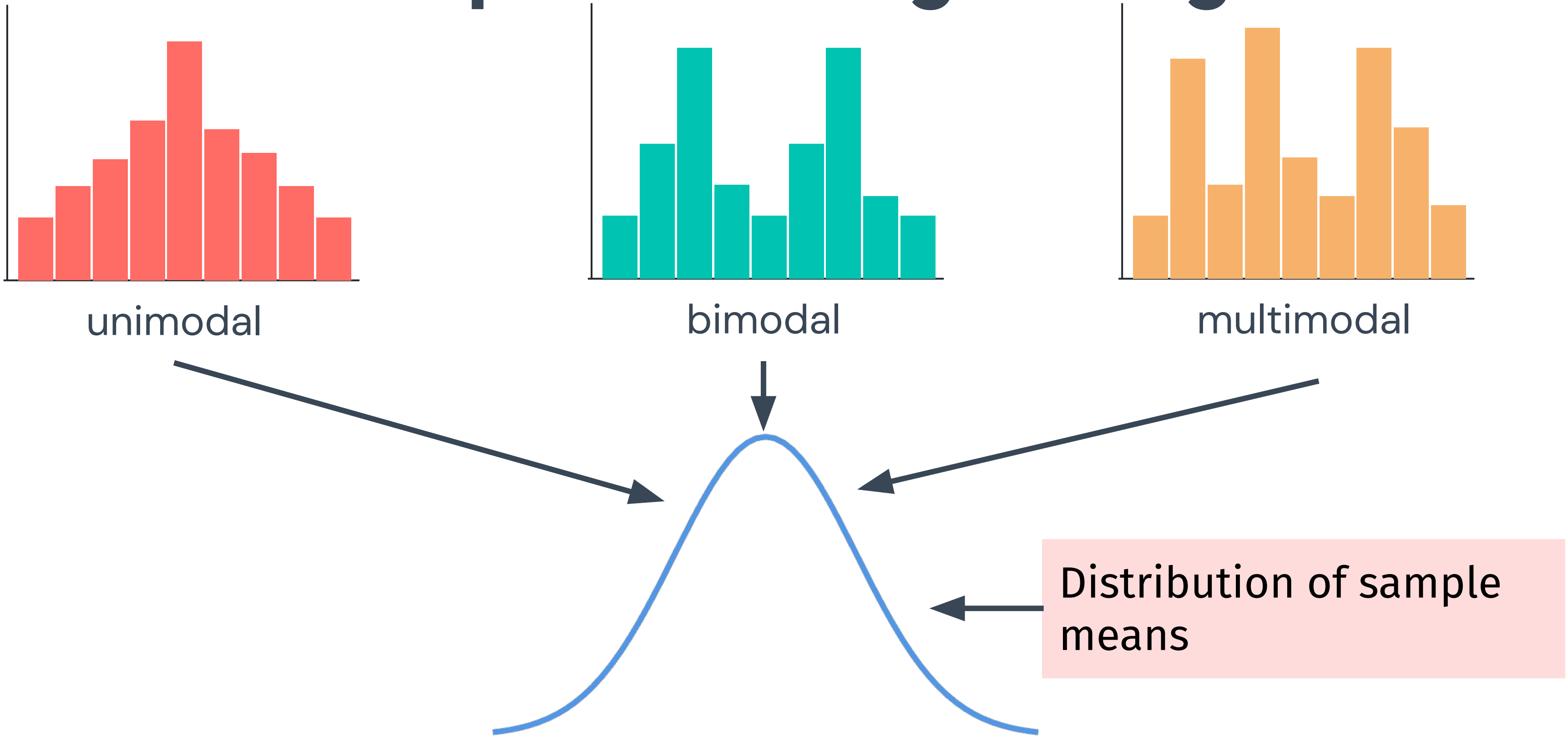
Simplified Definition

All **distributions of sample** means are roughly **normally distributed** as long as our **samples are large enough** (usually n > 30).

Shape does not matter as long as sample size is large enough:



Modality does not matter as long as sample size is large enough:



What does CLT tell us about study design?

30

is a **rule of thumb** for our ***minimum sample size*** for central limit theorem to apply.

In other words, we usually want at least 30 people in our sample. The **more skewed** the population, the **more people** we need in our sample.

<30

If a sample has less than 30 people, we can also apply the central limit theorem ***if the population is normally distributed*** on that variable.

What does CLT tell us about study design?

30

is a **rule of thumb** for our ***minimum sample size*** for central limit theorem to apply.

In other words, we usually want at least 30 people in our sample. The **more skewed** the population, the **more people** we need in our sample.

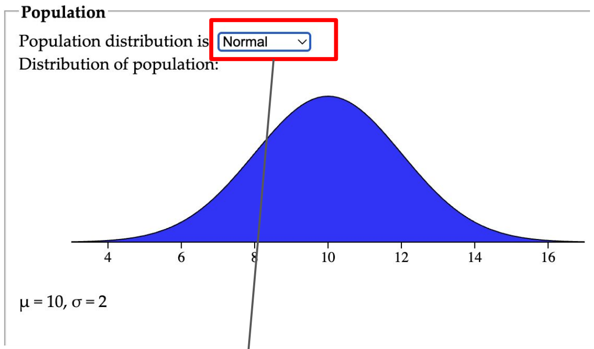
<30

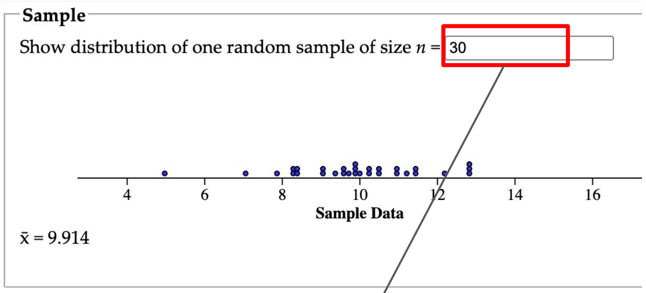
If a sample has less than 30 people, we can also apply the central limit theorem ***if the population is normally distributed*** on that variable.

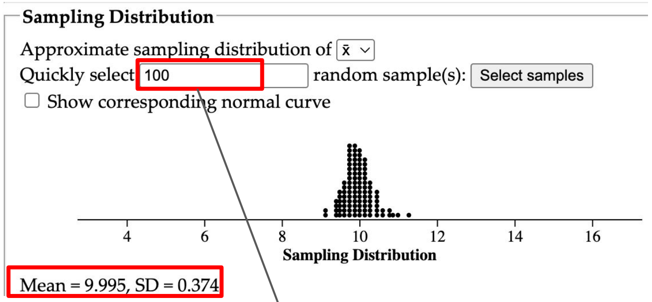
Both the Central Limit Theorem and the Standard Error calculation are telling you, the more people you have in your sample, the better!

ICA6: Let's come back to this website!

https://www.stapplet.com/sampdist.html







Now, change this to whichever option your heart desires.

CLT tells us this should be >30. Change it to below 30, run next step. Then change this to a massive number, run next step. Notice how the **shape & the Mean and SD** of the sampling distribution changes.

Let's keep this fixed to a large number. Let's say we repeatedly draw 10,000 samples.

Summary:

The larger the sample size, the lower the standard error.

The sampling distribution's mean equals the population mean, no matter the sample size.

As sample size increases, the sampling distribution becomes approximately normal—even if the population itself is not normal.

For each population distributions below, which would produce a sampling distribution that is normally distributed?

A Positively skewed distribution; sample size of 20

B Normal population distribution; sample size of 12

C Negatively skewed distribution; sample size of 50

D Bimodal distribution; sample size of 10

E Normal population distribution; sample size of 40

05

Wrap Up

Summary of today's lecture

- **Introduced & defined standard error (SE)**

Standard deviation of the sampling distribution.

- **Differentiated sampling distribution from the samples**
- **Learned about the factors that increase/decrease SE**

Sample size, population standard deviation

Summary of today's lecture

- **Central Limit Theorem, and how it relates to sampling distributions**

No matter what shape & modality of the population distribution, the sampling distribution of the mean will approximate a normal distribution as the sample size becomes large. The more skewed or irregular the population is, the larger the sample size needed for this approximation to hold.

- **Understand that SE tells us how trustworthy our sample is in giving us an idea about the overall population.**

The standard error tells us how much our sample statistic is expected to vary from sample to sample. Smaller SE means our sample statistic is a more precise estimate of the population value—so we can be more confident that our sample reflects the population.