SAMPLING & CONFIDENCE INTERVALS

RECAP

- Population
  - The larger group of individuals that a researcher wants to draw conclusions about

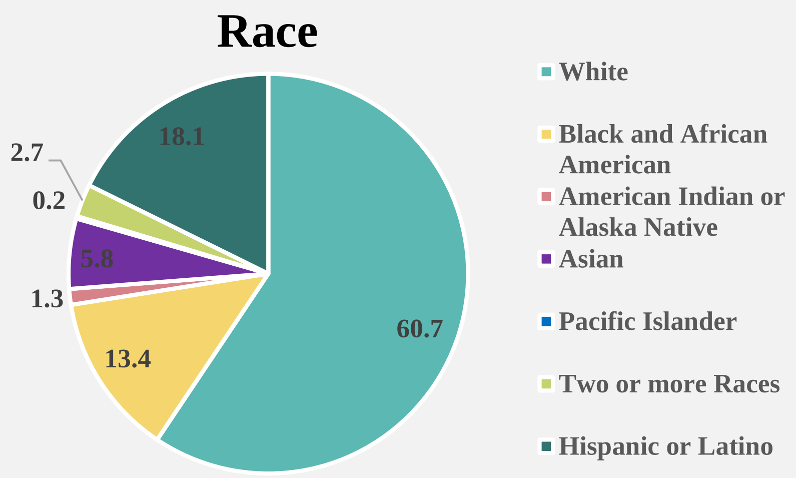
- Sample
  - A group of cases selected from a larger population

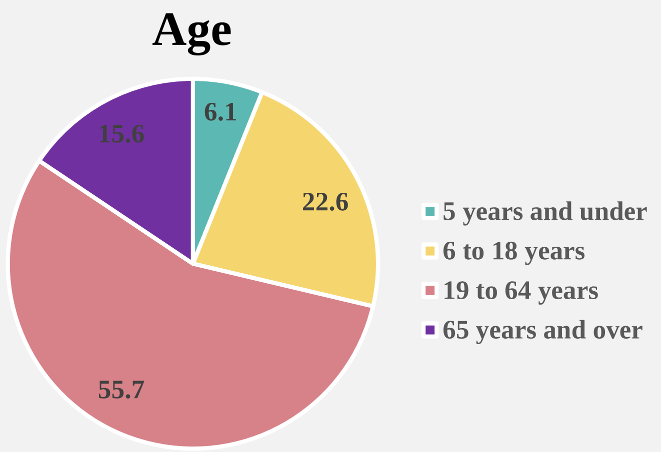
- Due to practical limitations, researchers often collect data from samples instead of populations.
  - The researcher's **sampling method**, or the way a researcher collects their sample, has important implications for **generalizing** the results to the larger population

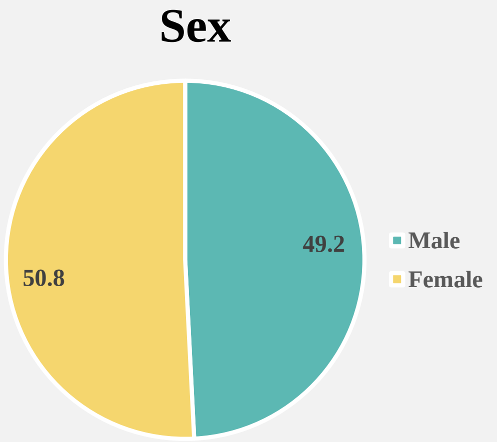
TYPES OF SAMPLES

Representative Sample: the participants have the same attributes as those that exist in the population and in approximately the same proportions, making *generalization* possible

Ex: US Population







TYPES OF SAMPLES

Convenience Sample: the participants are chosen based on how easily they can be obtained

- **Example:** Recruiting participants for a study by handing out fliers at the EMU around lunchtime

- Convenience samples are unlikely to be representative
  - If the sample is not representative, the results cannot be said to *generalize* to the larger population of interest.

COLLECTING A REPRESENTATIVE SAMPLE

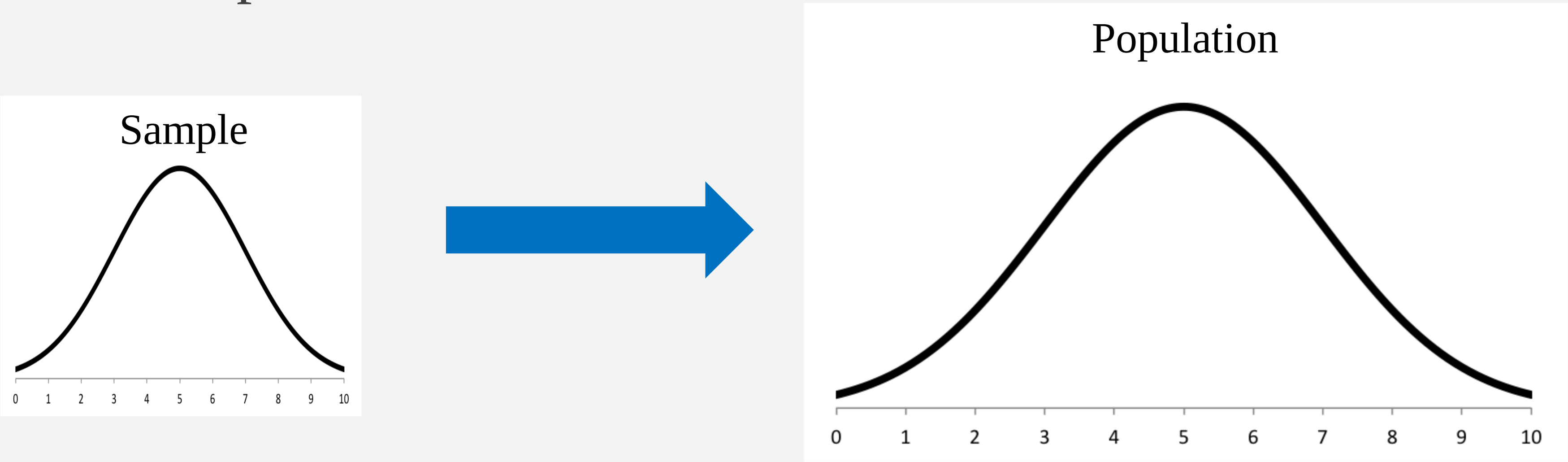
Methods for obtaining a representative sample:

1. **Random Sample**: recruiting participants from the population completely randomly; all cases have an equal chance of being selected for the sample
    - Example: Using a random number generator
  - Random sampling by itself does not guarantee a representative sample.

2. **Large sample size**
    - *The law of large numbers*: a larger sample is more likely to represent the population

PRACTICAL ISSUES WITH SAMPLING

- A researcher desires to be able to collect sample data and still make inferences about the population the sample came from.



- However, <u>even with a random sample</u>, there are potential issues that arise from the sampling process.

PRACTICAL ISSUES WITH SAMPLING

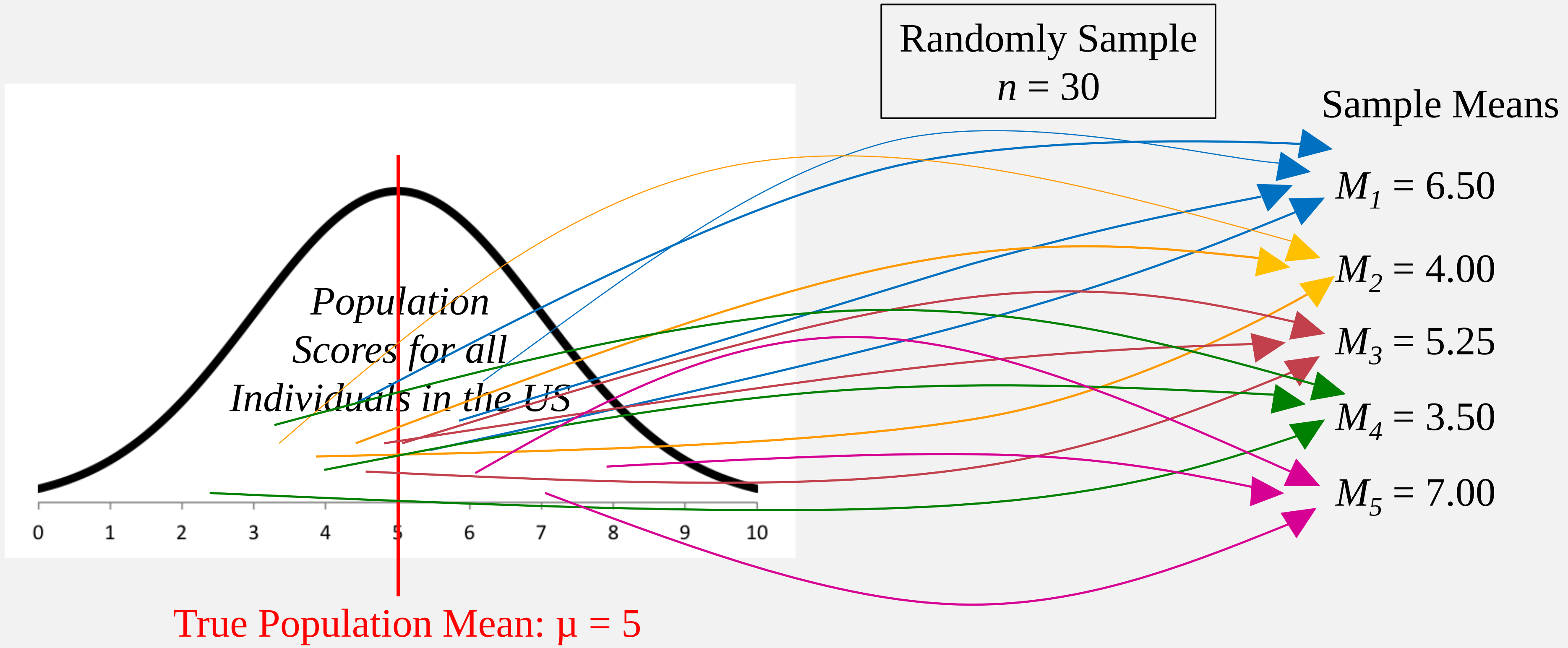
1. **Self-selection bias:** occurs when not everyone who is asked to participate in a study agrees to do so
   - The sample will be non-representative because only the characteristics of individuals who self-select into the study will be represented.
     - Example: Recruiting participants for a study on cheating behaviors.

2. **Consent rate:** the percentage of people who agree to participate; consent rate can be an indicator of whether self-selection bias has occurred
   - Rule of thumb: When consent rate is above 70%, self-selection bias is not a problem

3. **Sampling Error:** the discrepancy, due to random factors, between a sample statistic and the population parameter the sample is attempting to estimate.
   - Example: The difference between a sample mean and the true mean of the population that the sample came from.

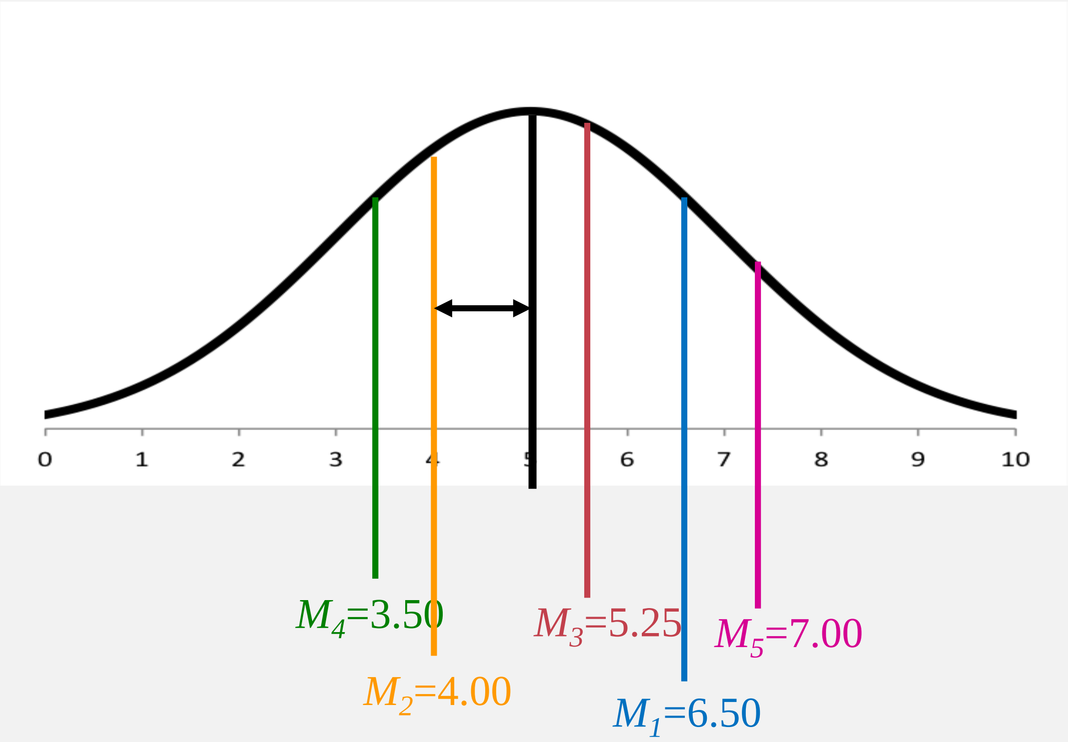
SAMPLING ERROR

- <u>Example</u>: Say a researcher is interested in assessing the attitudes of people in the US towards eliminating all student loans on a scale from 0 (extremely negative) to 10 (extremely positive). The researcher only has enough resources to randomly sample 30 individuals from the population.



SAMPLING ERROR

- Sampling error is due to completely random factors.
  - In other words, obtaining a sample mean that doesn’t perfectly match the true population mean occurs by *chance*.
- There is nothing systematic about sampling error.
  - You are equally likely to obtain a sample mean above or below the actual population mean.
  - This means that the sample mean is an *unbiased* estimator of the true mean of the population.



SAMPLING DISTRIBUTIONS

- A **sampling distribution** is a frequency distribution that's created by calculating all of the *possible* means (or another type of statistic, such as the median) that could be obtained by randomly sampling from a given population.

All Possible Sample Means with a Sample Size of n = 30

$M_1 = 6.50$

$M_2 = 4.00$

$M_3 = 5.25$

$M_4 = 3.50$

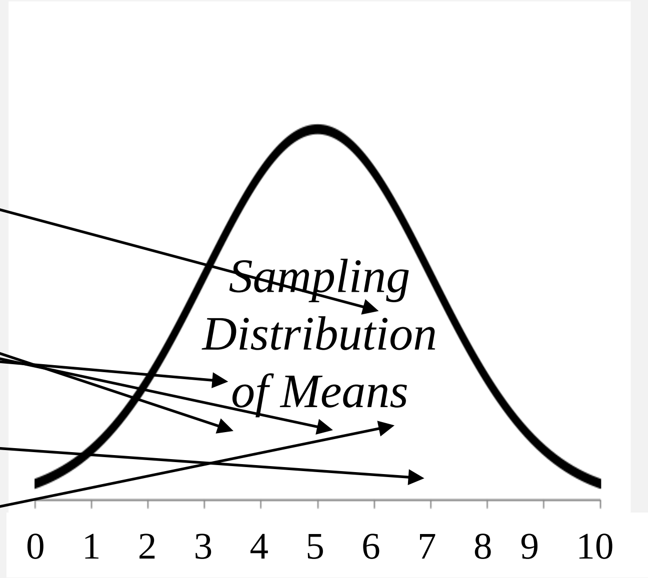
$M_5 = 7.00$

.

.

.

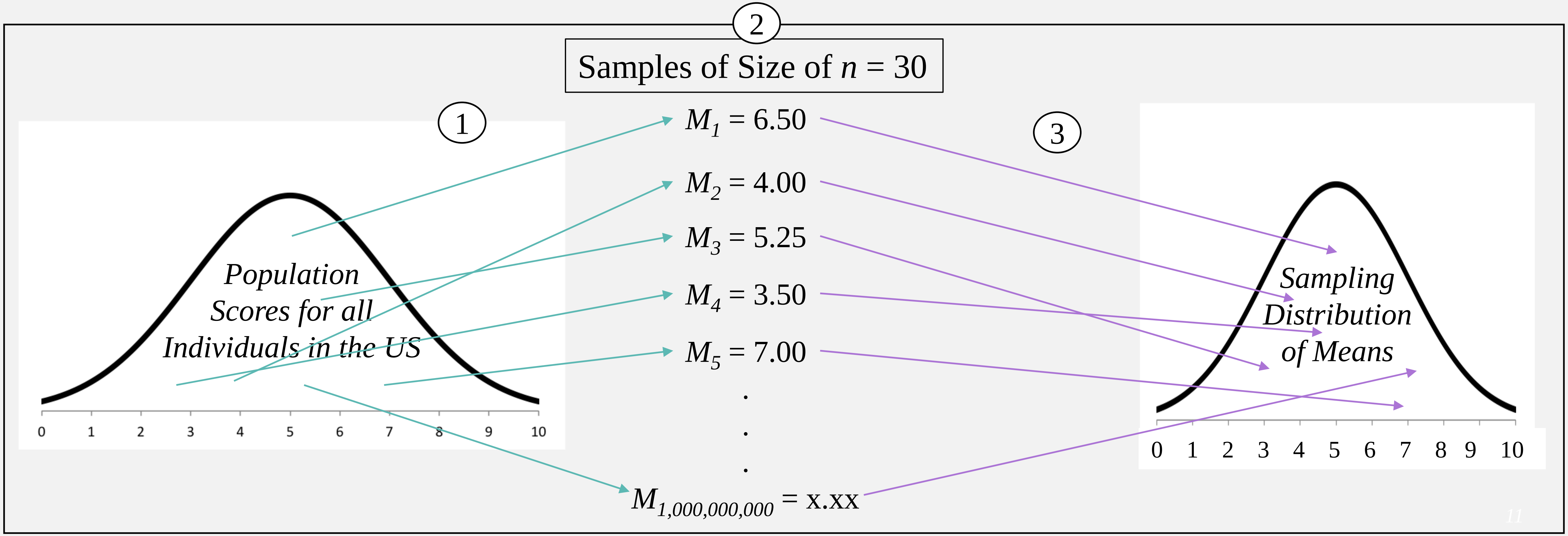
$M_{1,000,000,000} = \text{x.xx}$



SAMPLING DISTRIBUTION OF MEANS

To generate a sampling distribution:

1. Take repeated, random samples of a specified size from a population.
2. Calculate the mean for each sample.
3. Plot all of the sample means in a frequency distribution.



CENTRAL LIMIT THEOREM

Instead of having to construct a sampling distribution of means from scratch…

Central Limit Theorem: describes the characteristics of a sampling distribution of means when the sample size is large and every possible sample is obtained

If n is large ($n \geq 30$), then…

(1) The mean of the sampling distribution of means is equal to the population mean

$$\mu_M = \mu$$

(2) The **standard error of the mean** is equal to the standard deviation of the population divided by the square root of the sample size

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

(3) The sampling distribution of the mean will be normally distributed, no matter what the shape of the population is.

STANDARD ERROR OF THE MEAN

- **Standard error of the mean (σ_M)** : the measure of standard deviation for a sampling distribution of means
  - Describes the average amount of deviation of sample means from the mean of the sampling distribution of means

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

  - When σ is not known, use the estimate of the population standard deviation (*s*) to calculate the standard error, s_M

$$s_M = \frac{s}{\sqrt{n}}$$

- Q: What happens as you increase the size of the samples being taken from the population?

PRACTICE PROBLEMS

You just finished reading a Pew research report that surveyed everyone in the US and found that Americans spend an average of 180 minutes ($\mu = 180$) on their phones per day. The report does not provide the standard deviation of the population. You recently conducted a study in which you randomly selected 80 people and asked them the amount of time they spend on their phones each day. The mean of your sample is 164.69 minutes ($M = 164.69$). The standard deviation of your sample is 86.54 minutes ($s = 86.54$).

Based on the information given, construct a sampling distribution of means that represents all possible sample means you could obtain from the US population if you were to repeatedly sample 80 people from the population and calculate the average amount of time people spend on their phones each time.

Specify the following about your sampling distribution of means:

- The sampling distribution's shape
- The sampling distribution's mean
- The sampling distribution's standard error

PRACTICE PROBLEMS

1. What is the mean of the sampling distribution of means you constructed?

a) 86.54　　b) 164.69　　c) 180　　d) 1

2. What is the standard error of the sampling distribution of means you constructed?

a) 0　　b) 86.54　　c) 164.69　　d) 9.68

3. What is the shape of the sampling distribution of means you constructed?

a) Positively Skewed　　b) Negatively Skewed　　c) Normal　　d) Bimodal

PRACTICE PROBLEMS

Solutions

1. What is the mean of the sampling distribution of means you constructed?

a) 86.54 b) 164.69 c) 180 d) 1

2. What is the standard error of the sampling distribution of means you constructed?

a) 0 b) 86.54 c) 164.69 d) 9.68

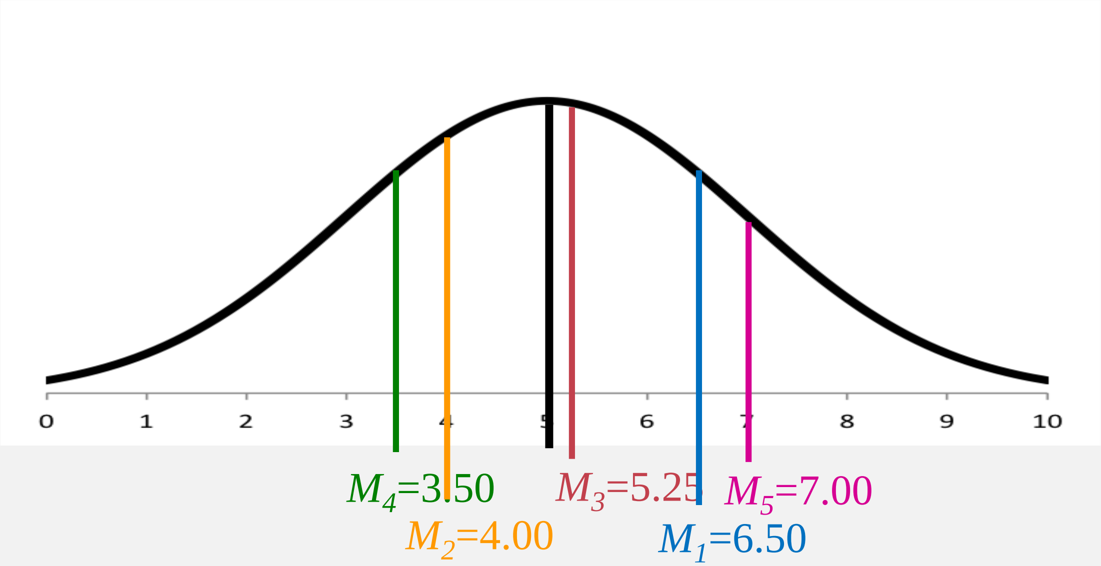
3. What is the shape of the sampling distribution of means you constructed?

a) Positively Skewed b) Negatively Skewed c) Normal d) Bimodal

POINT ESTIMATES

- **Point Estimate:** a single value that is used to estimate a population value
  - **Example:** A sample mean (*M*) is used to estimate the population mean (μ)

- Due to sampling error, a point estimate is unlikely to be exactly equal to the population parameter it's estimating**.**



INTERVAL ESTIMATES

- **Interval Estimate:** a range of values around a point estimate within which a population parameter is likely to exist
  - More likely to contain the true population parameter than a point estimate

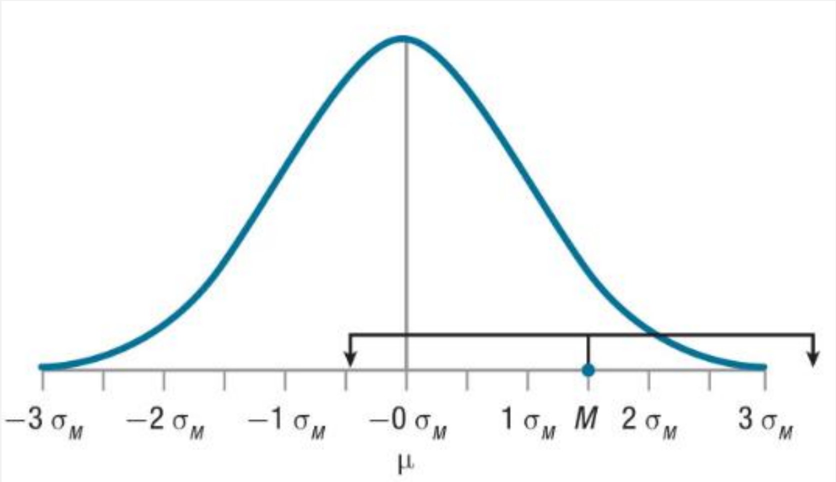
- A commonly reported interval estimate in psychological research is the **95% confidence interval**

95% CONFIDENCE INTERVAL

95% Confidence Interval

$$95\%\text{CI} = M \pm (1.96 \text{ x } \sigma_M)$$

- M, the sample mean
- σ_M, the standard error of the mean
    - When σ_M is unknown, use s_M
- ± 1.96, these are the standardized values that cover a range of 95% of means around the sample mean



95% CONFIDENCE INTERVAL

- **Example:** Let's return to our example in which we collected a sample of 80 individuals and asked them to report the number of minutes they spend on their phones each day. The mean of the sample was 164.69 (M = 164.69), and the standard deviation of the sample was 86.54 (s = 86.54).
- Calculate a 95% confidence interval around our point estimate of the population mean (around M = 164.69).

95% CI = $M \pm (1.96 \text{ x } \sigma_M)$

= 164.69 ± (1.96 *(86.54/√80))

= 164.69 ± (1.96 *9.68)

= 164.69 ± (18.97)

95% CI = ranges from 145.72 to 183.66

Q: What does this 95% CI mean?

Reporting in APA Style: 95%CI [145.72, 183.66]

95% CONFIDENCE INTERVAL

- *Officially*, a 95% confidence interval tells you that if you were to repeat the process of sampling and calculating 95% CI's for each sample mean, 95% of the time, the CI will contain the true population mean

- If you repeatedly sampled and calculated 95%CI's:
  - 95% of the time, the CI captures the true population mean, μ
  - 5% of the time, the CI does not capture the true population mean, μ

PRACTICE PROBLEMS

Joven wants to estimate the average number of books people read in a year. He randomly selects 65 people from the US population and asks them to report how many books they read the previous year. The sample mean is 7.50 (M = 7.50), and the standard deviation is 4.40 (s = 4.40).

Calculate the 95% CI around this point estimate for average number of books people in the US read per year. Report in APA style.

a) 95%CI [6.43, 8.57]

b) 95%CI [-1.12, 16.12]

c) 95%CI [6.95, 8.05]

d) 95%CI [6.00, 9.00]

PRACTICE PROBLEMS

Solution

$$95\%\ CI = 7.50 \pm (1.96 * (4.40/\sqrt{65}))$$
$$= 7.50 \pm (1.96 * 0.5458)$$
$$= 7.50 \pm 1.07$$

95% CI = ranges from 6.43 to 8.57

a) 95%CI [6.43, 8.57]
b) 95%CI [-1.12, 16.12]
c) 95%CI [6.95, 8.05]
d) 95%CI [6.00, 9.00]

WEEK 2 REMINDERS

- Homework 2 is due July 4th at 11:59pm

- Jamovi HW2 is due July 8th at 11:59pm