CH. 6 – Sampling Distributions

PY *221* Research Methods and Statistics I

Dr. Valenti

Outline for Ch. 6

1. Review of key facts
2. Sampling distribution of sample means
3. Central Limit Theorem
   - Standard error (SE)
4. Determining the probability of drawing a sample with particular characteristics, from a particular population (*return of the z-score!*)

Practice your understanding of what we've covered thus far, by indicating True or False, and correcting any false statements.

1. Not every individual score for a variable in a given sample will be equal to the mean score for that variable in that sample.
2. We can measure how far the typical score is from the mean score for a variable by calculating something called the *standard deviation*.
3. The location of an individual's raw score within a sample distribution of scores can be represented using a z-score.
4. We can use a person's z-score to determine the proportion of scores in that sample that fall below & above that person's score.
5. Even with a large, random sample, our sample statistics (e.g., x-bar) will always differ from the actual values of the parameters (e.g., μ).
6. If we pull several random samples from the same population, each sample will give us slightly different sample statistics (e.g., means).

All are TRUE.

#6 -- If we pull several random samples from the same population, each sample will give us slightly different sample statistics (e.g., means).

- How do we know if a sample mean is *slightly* different (and follows this rule) vs. *very* different (*too* different; and breaks this rule)?
- If the sample mean is TOO different from the other samples' means, what does that tell us about that sample?

Outline for Ch. 6

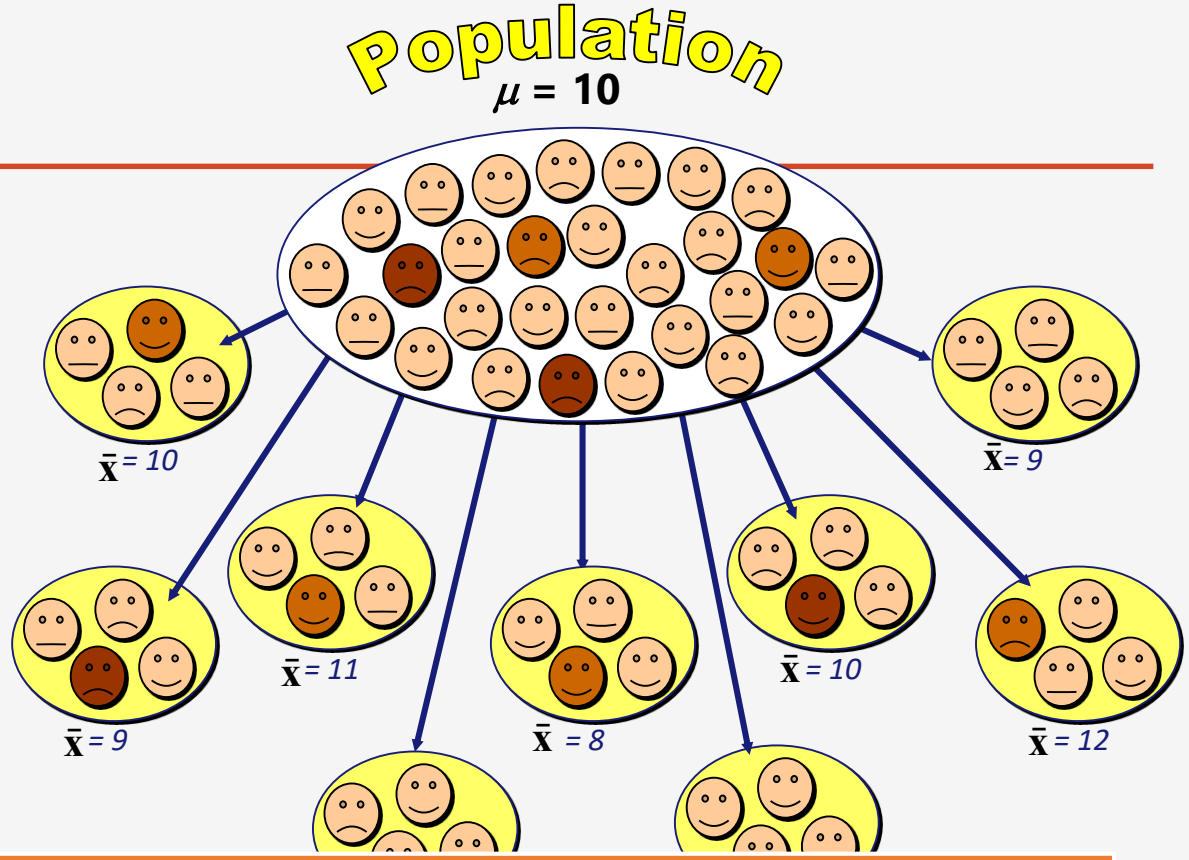
1. Review of key facts
2. **Sampling distribution of sample means**
3. Central Limit Theorem
   - Standard error (SE)
4. Determining the probability of drawing a sample with particular characteristics, from a particular population (*return of the z-score!*)

Thought experiment….

- Suppose we're interested in the **population of all professors at BSC,** and how **happy** they are.
- We decide to pull random samples of 4 professors at a time and calculate the mean happiness level for those 4 profs…

Pull random samples of 4 from the population of all BSC professors.

- *Then,* plot the sample means as a frequency distribution . . .
  - How many *samples* had a mean of **8**? (i.e., what's the frequency?)
  - How many *samples* had a mean of **9**?
  - Etc.

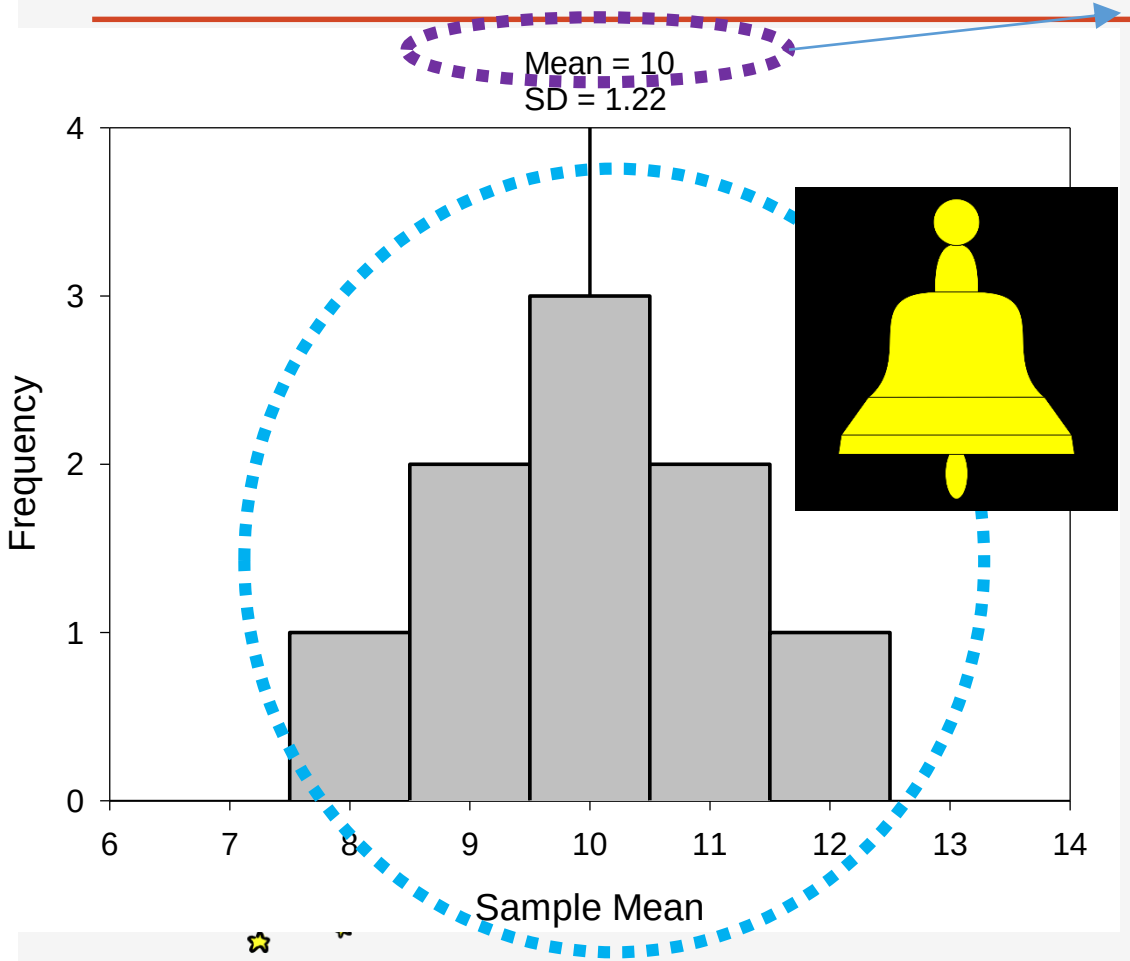


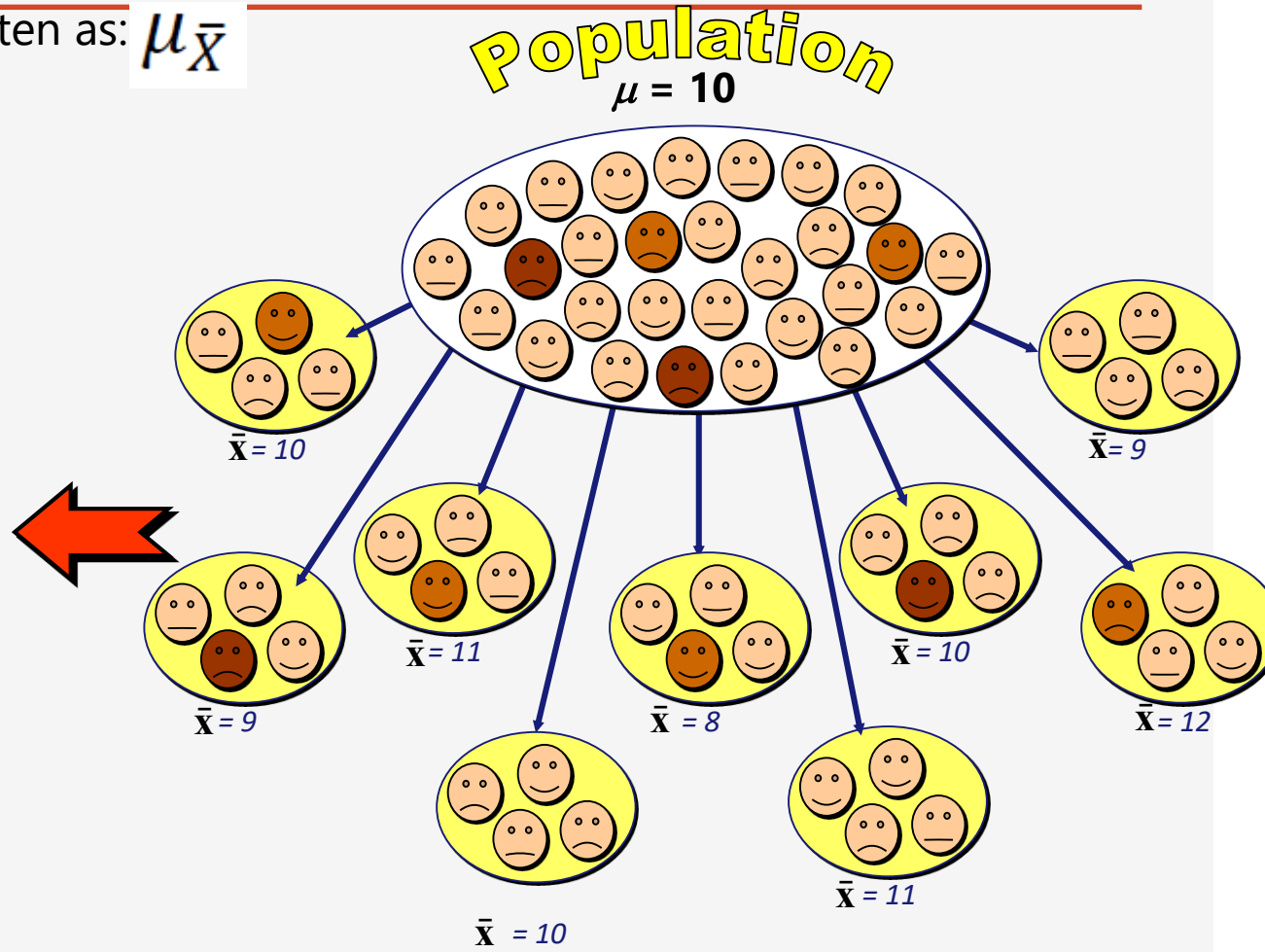
If we did this, what would go on the x axis and what would go on the y axis of our frequency distribution?

On X-axis, the sample means, i.e., x-bars (8, 9, 10, etc.)
On Y-axis, frequency of samples that had each mean

We've created the **sampling distribution** of sample means *aka* the ***sampling distribution***

Also sometimes written as: $\mu_{\bar{X}}$





Recap of creating a sampling distribution of sample means

1. Every sample is drawn randomly from a specified population
2. The sample size (N) is the same for all of these samples
3. The number of samples is very very large
4. The mean ($\overline{X}$) is calculated for each sample
5. Those sample means are arranged into a frequency distribution

   (aka the sampling distribution of sample means,

   aka the sampling distribution)

Outline for Ch. 6

1. Review of key facts
2. Sampling distribution of sample means
3. **Central Limit Theorem**
   - Standard error
4. Determining the probability of drawing a sample with particular characteristics, from a particular population (*return of the z-score!*)

CLT - Central Limit Theorem (derived by statisticians)

Assuming the sampling distribution of sample means is created from large (*N* = 30+) samples, the following is likely to be true:

1. the sampling distribution is **normal** (bell-shaped, symmetrical)
2. the **mean** of the sampling distribution ($\mu_{\bar{X}}$) = the mean of the population (μ)
3. the **standard deviation** of the sampling distribution ($\sigma_{\bar{X}}$) …. (*stay tuned!)*

Practice your understanding. Indicate *True* or *False*. Correct false statements.

1. A **sampling distribution** contains the number of people in the population on the y axis, and the possible values of the sample means on the x axis.
2. A **sampling distribution** contains the number of people in the sample on the y axis, and those people's scores on the x axis.
3. When the **CLT** refers to a "large" sample, "large" means at least 30 Ps.
4. When you draw many large samples randomly from the same population, each of their sample means will be identical.
5. According to the **CLT**, when you draw many large samples randomly from the same population, the average of all of their sample means will be equal to the population mean.

1. False
2. False
3. True
4. False
5. True

Outline for Ch. 6

1. Review of key facts
2. Sampling distribution of sample means
3. **Central Limit Theorem**
   - **Standard error**
4. Determining the probability of drawing a sample with particular characteristics, from a particular population (*return of the z-score!*)

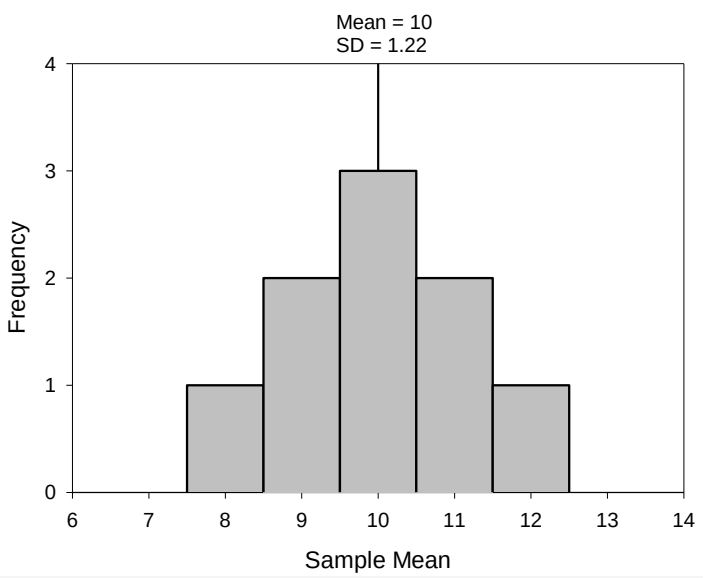
Standard deviation *of the* ***sampling distribution*** *of sample means*

Conceptually, what's meant by the "standard deviation of sample means"?

- the typical (i.e., standard/average) distance between individual *samples' means* and the *mean of all samples' means*, *OR*
- whether the individual samples' means are clustered closely or widely scattered around their own average

standard error (*SE*)

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$



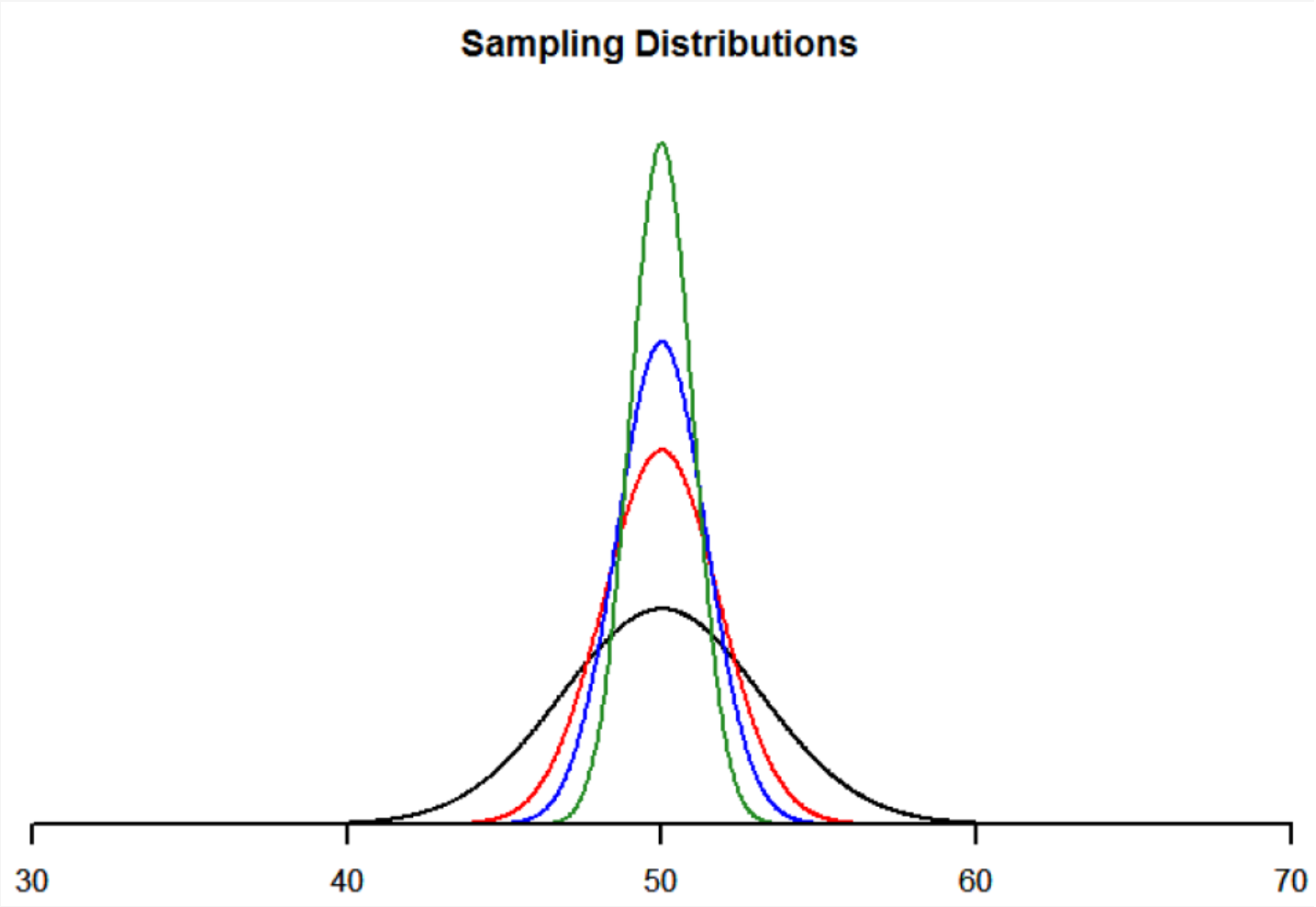
Standard error (*SE*), continued.

1. Which of these four sampling distributions (which color) has the largest/highest *SE*?

BLACK

2. Which of these has the smallest/lowest *SE*?

GREEN



Standard error (*SE*), continued

- Recall from CLT: mean of **sampling distribution** = the mean of the **population**
- *SE* describes whether sample means are clustered closely or widely scattered around <u>their own mean</u>
  - **Low** *SE* indicates:
    - sample means are clustered closely around the population mean → *most* sample means are similar to the population mean → *our one sample's* mean is probably similar to the population mean
  - **High** *SE* indicates:
    - sample means are widely scattered from the population mean → there's great variability between the means of different samples → there's a decent chance that our one sample's mean will <u>*not*</u> be similar to the population mean

Do researchers want their standard error to be low or high, assuming they're hoping to get an accurate estimate of the population mean from their one sample?

low (small)

Let's calculate the **standard error** for the following samples

$$SE = \sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

- Sample 1

  $\bar{x} = 9, s = 2, N = 36$ $SE = 0.333$

- Sample 2

  $\bar{x} = 9, s = 2, N = 100$ $SE = 0.200$

- Sample 3

  $\bar{x} = 9, s = 2, N = 900$ $SE = 0.067$

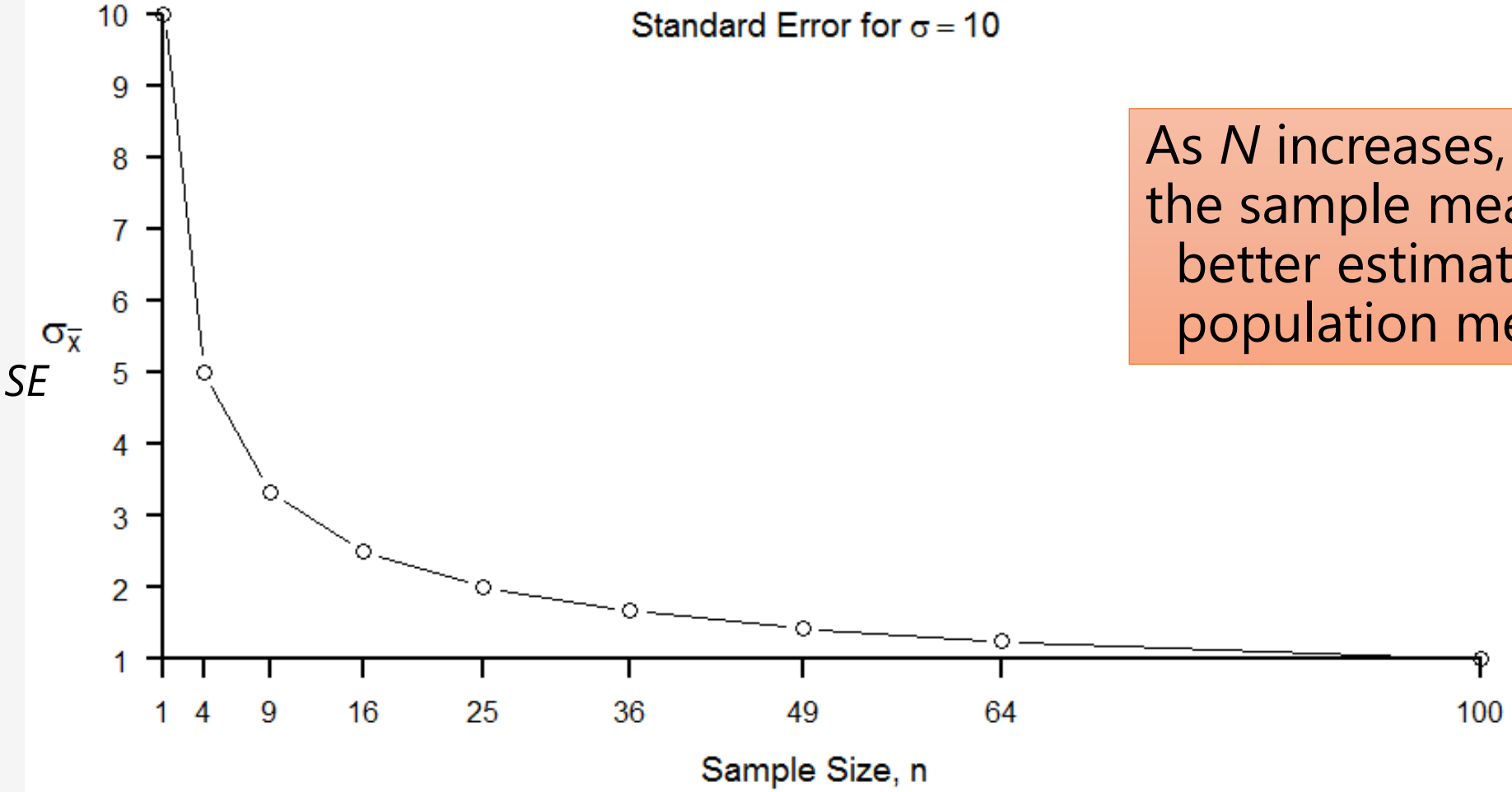
Bigger is better (in the case of samples)!

As *N* increases → *SE* decreases.

When *N* is larger, holding all else constant → the sample mean is a better estimate of the population mean.

Relationship between **sample size** and **standard error**

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$



As *N* increases, *SE* decreases → the sample mean becomes a better estimate of the population mean

Let's calculate the **standard error** for the following samples

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

- Sample 1
  $\overline{x} = 9, s = 1, N = 100$ $SE = 0.10$
- Sample 2
  $\overline{x} = 9, s = 10, N = 100$ $SE = 1.00$
- Sample 3
  $\overline{x} = 9, s = 20, N = 100$ $SE = 2.00$

As *s* **DE**creases → *SE* decreases.

As *s* **DE**creases → the sample mean becomes a better estimate of the population mean.

CLT - Central Limit Theorem (derived by statisticians)

Assuming the sampling distribution of sample means is created from samples that are large (*N* = 30+), the following is likely to be true:

1. The sampling distribution is **normal**
2. the **mean** of the sampling distribution ($\mu_{\bar{X}}$) = the **mean** of the population (μ)
3. The **standard deviation** of the sampling distribution (aka the *standard error*) can be computed using the equation:

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

Practice your understanding. Indicate *True* or *False*, and correct any false statements.

1. *Standard error* is also known as "the standard deviation of sample means."
2. Having a low standard error indicates that sample means are clustered fairly closely around the population mean.
3. Having a high standard error for your sample indicates that your sample's mean is not necessarily a good estimate of the actual population mean.
4. Having a smaller sample size will typically lead to a smaller standard error than having a larger sample size (holding all other values constant).
5. The "standard error" can be defined as the typical distance between an individual sample's mean, and the standard deviation of the population.
6. One way to think of "standard error" is that standard error captures how much we'd expect our means to differ from sample to sample, when these samples are randomly drawn from the same population.

1. True
2. True
3. True
4. False
5. False
6. True

Work For Tuesday . . .

- Make a plan for what blocks of time each week you will devote to PY 221.
- Read Ch. 6 on sampling distributions if you haven't already. Consider re-reading it if it was challenging the first time.
- Complete self-graded HW on Ch. 6.
- Read Ch. 7 to prepare for Tuesday's class.
- Work on exam corrections assignment.
- Complete Tuesday's quiz on Ch. 6.