Keeping up with work for this course…

- read Ch. 6 (10 pages) on sampling distributions.
- read Ch. 7 (20 pages) plus Ch 8 (pp. 154-end).
- watch the four videos on Moodle on Ch 6-7 in Misc Resources section.
- complete self-graded HW on Ch. 6.
- make some progress on Exam Autopsy.

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. What’s the probability of drawing a sample with a particular mean (or an even lower/higher mean) from a particular population?

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. **For a given variable, we can measure how far the typical score is from the mean score by calculating something called the *standard deviation*.**
3. **For normally distributed variables, the location of an individual's raw score within a sample's distribution of scores can be represented using a z-score.**
4. **We can use a person's z-score to determine (using a table) the proportion of scores in that sample that fall below & above that person's raw score.**
5. **Even with a large, random sample, our sample statistics (e.g., x-bar, the sample mean) will always differ from the population parameter values (e.g., μ, the pop mean), simply due to sampling error (natural variation).**
6. **If we pull several, large random samples from the same population, each sample will give us slightly different sample statistics (e.g., slightly different x-bars).**

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. For a given variable, we can measure how far the typical score is from the mean score by calculating something called the *standard deviation.*
3. For normally distributed variables, the location of an individual's raw score within a sample's distribution of scores can be represented using a z-score.
4. We can use a person's z-score to determine (using a table) the proportion of scores in that sample that fall below & above that person's raw score.
5. Even with a large, random sample, our sample statistics (e.g., x-bar, the sample mean) will always differ from the population parameter values (e.g., μ, the pop mean), simply due to sampling error (natural variation).
6. If we pull several, large random samples from the same population, each sample will give us slightly different sample statistics (e.g., slightly different x-bars).

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., x-bars).

Assume this is our population (imagine it's much larger, though).

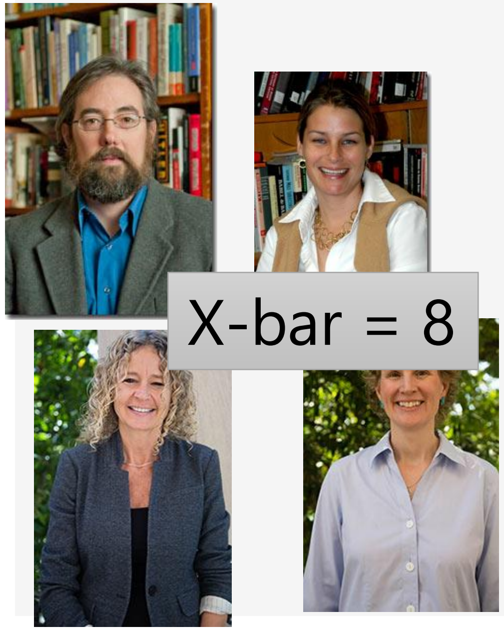
Assume these are large, randomly drawn samples from that population. We measure *age*.

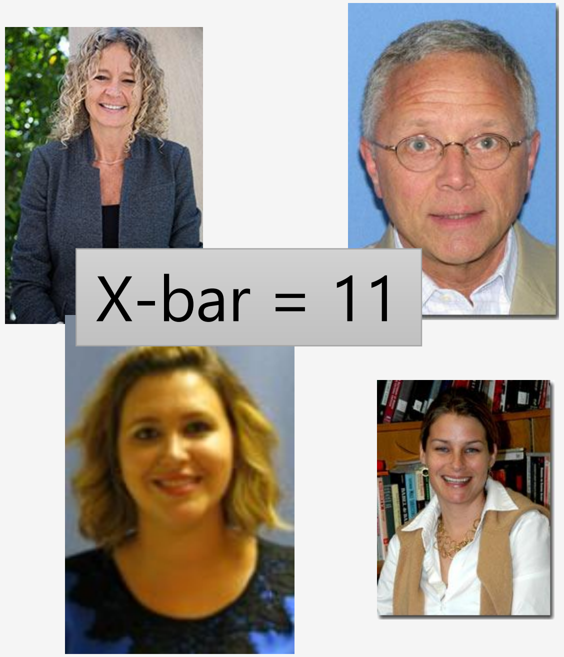
Outline for Ch. 6

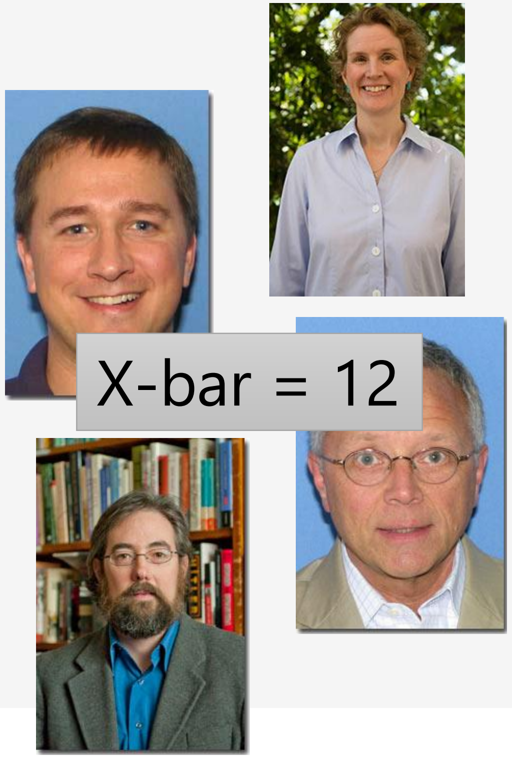
1. Review of key facts
2. **Sampling distribution of sample means**
3. Central Limit Theorem
   - Standard error (SE)
4. What's the probability of drawing a sample with a particular mean (or an even lower/higher mean) from a particular population?

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…

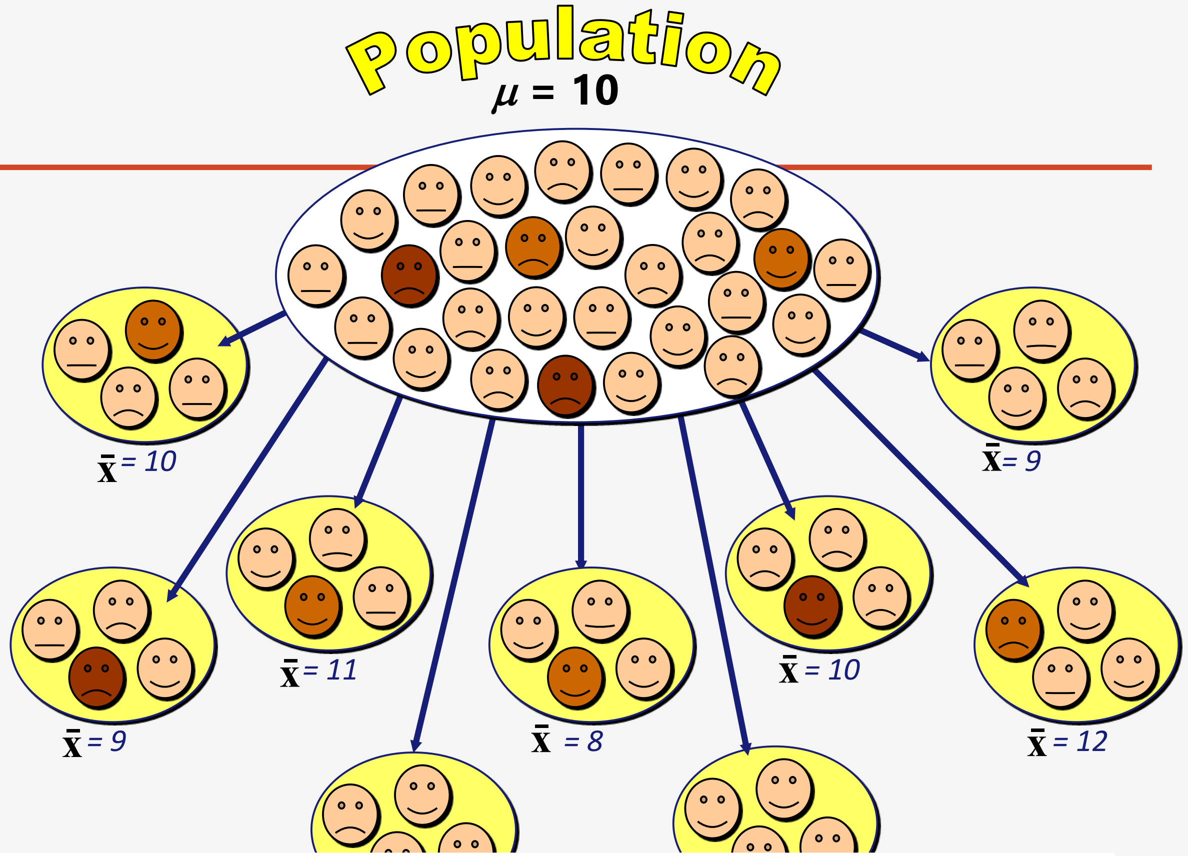






Pull random samples of 4 from the population of all BSC professors.

- *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.



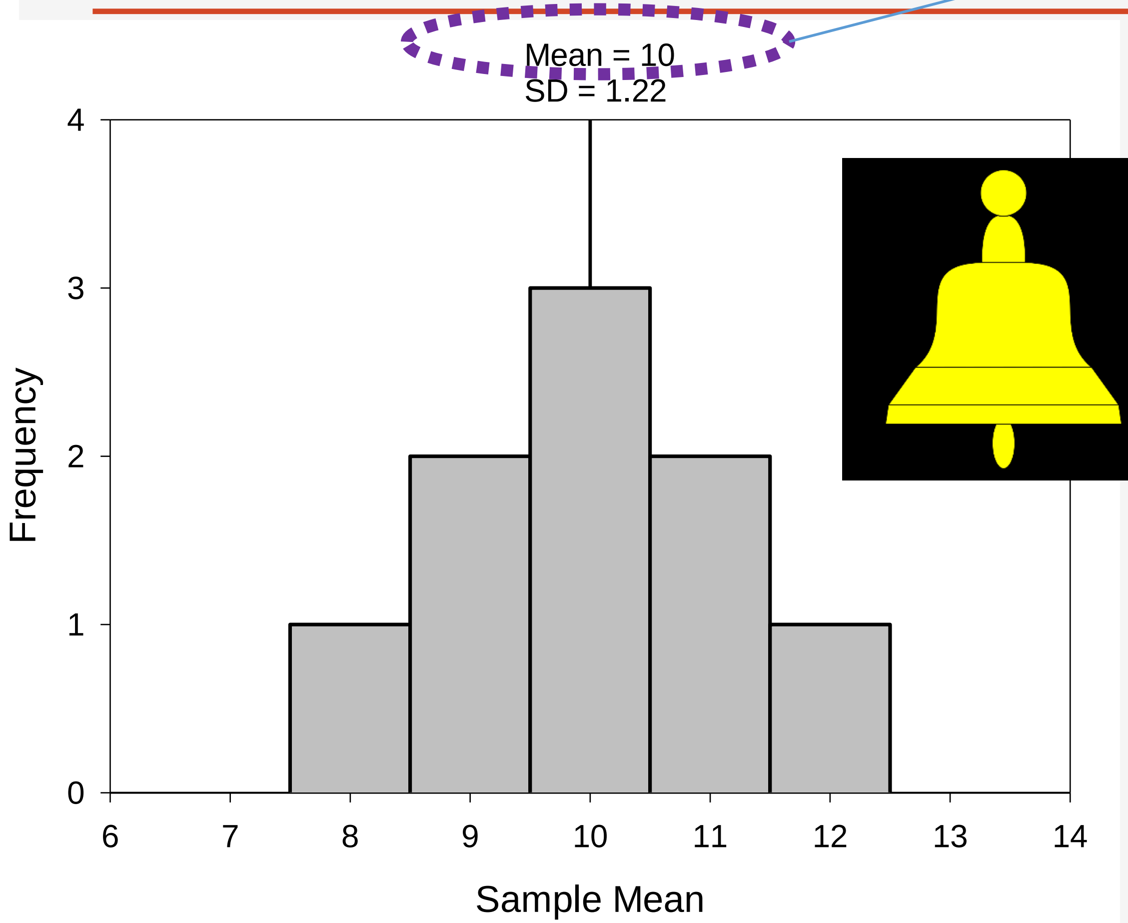
If we did this, how would we label the x axis and how would we label the y axis of our frequency distribution?

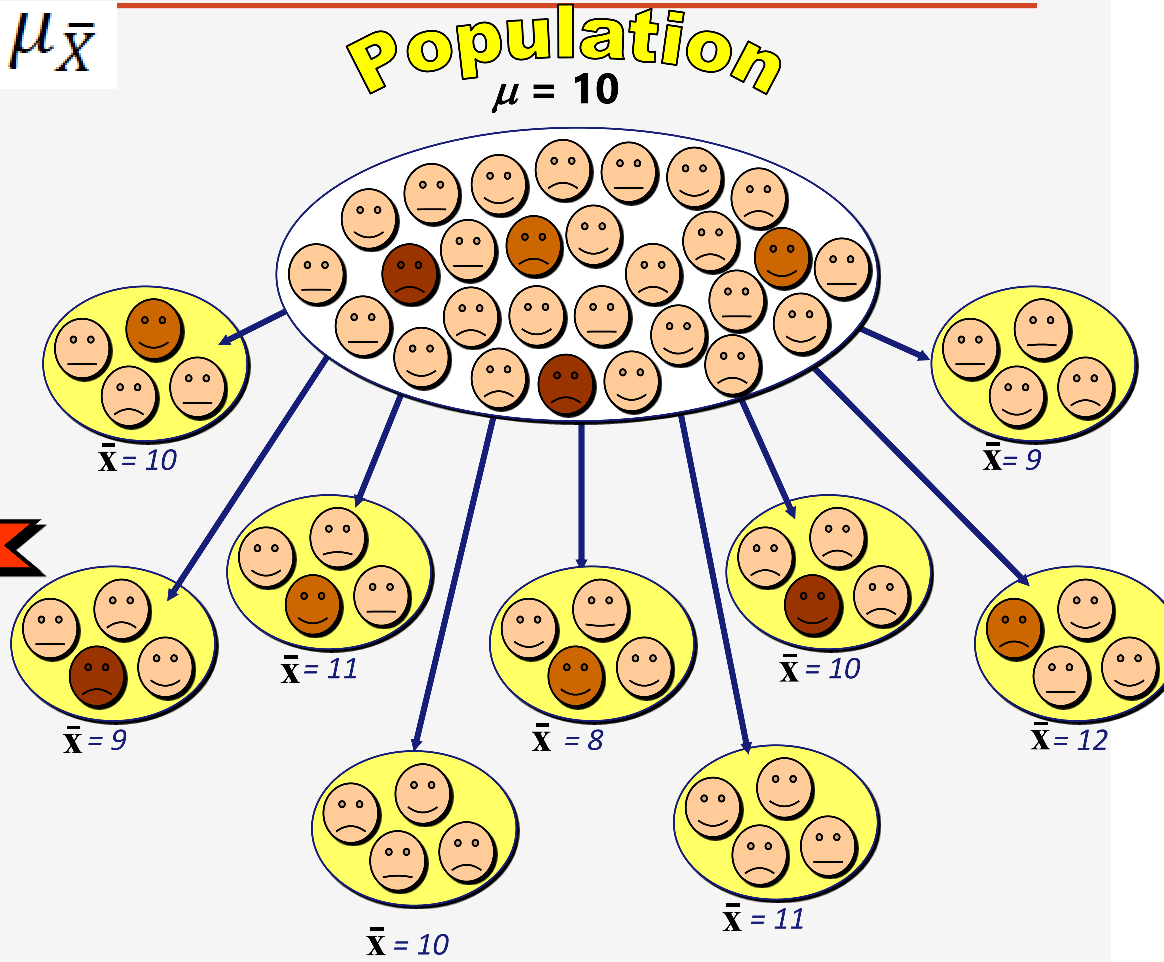
On X-axis, the sample means, i.e., x-bars (8, 9, 10, etc.)
On Y-axis, the frequency of samples that had each mean

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

It's the mean of all of the means!

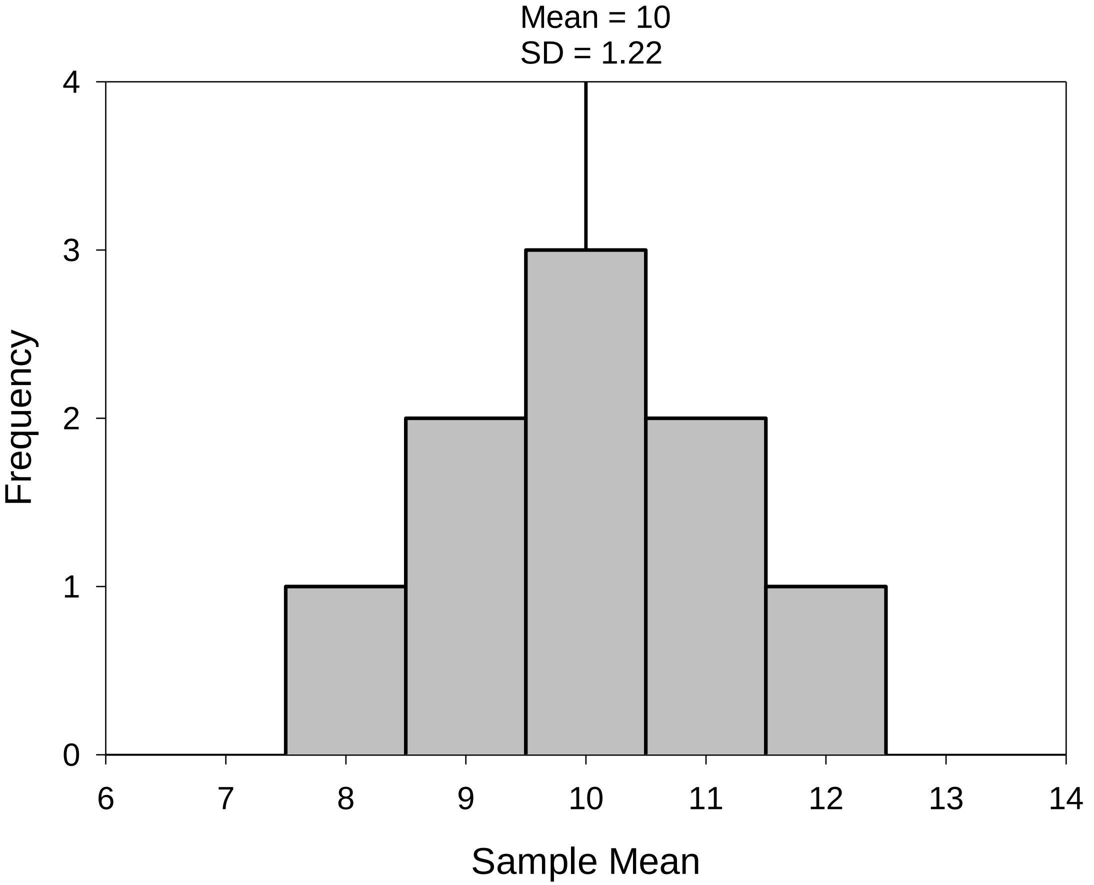
Also sometimes written as: $\mu_{\bar{X}}$





DRAW THIS HISTOGRAM IN YOUR NOTEBOOK. DON'T WORRY ABOUT THE NUMBERS BUT DO LABEL THE AXES & WRITE "SAMPLING DISTRIBUTION OF SAMPLE MEANS" ABOVE IT.

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



Our true/false statements from earlier . . .

5. Even with a large, random sample, our sample statistics (e.g., x-bar, the sample mean) will always differ from the population parameter values (e.g., μ, the pop mean), simply due to sampling error (natural variation).

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

REVIEW:
How does one create a <u>sampling distribution</u> of sample means?

1. Start by <u>randomly</u> drawing samples from a given population
   1. The sample size (*N*) is the same for all of these samples
   2. The number of samples is very, very large
2. Calculate the mean ($\bar{X}$) for each sample
3. Arrange those sample means ($\bar{X}$) into a frequency distribution

   (This distribution is known as the <u>sampling distribution of sample means</u>; for short, call it the <u>sampling distribution.</u>)

Practice your understanding. Indicate *True* or *False*. And please correct false statements.

Answers: Both are false. To correct both...
A sampling distribution contains the possible values of the sample means on the x axis, and the number of samples with those means on the y axis.

1. A **sampling distribution** contains the possible values of the sample means on the x axis and the number of people in the population whose scores match those means on the y axis.

2. A **sampling distribution** contains people's possible scores on the x axis, and the number of people in the sample who have each score on the y axis.

REVIEW:
How does one create a sampling distribution of sample means?

1. Start by randomly drawing samples from a given population
   i. The sample size (*N*) is the same for all of these samples
   ii. The number of samples is very, very large
2. Calculate the mean ($\bar{X}$) for each sample
3. Arrange those sample means ($\bar{X}$) into a frequency distribution

   (This distribution is known as the sampling distribution of sample means; for short, call it the sampling distribution.)

Who cares?!
Why do we need to know this?

Outline for Ch. 6

1. Review of key facts
2. Sampling distribution of sample means
3. **Central Limit Theorem**
   - Standard error
4. What's the probability of drawing a sample with a particular mean (or an even lower/higher mean) from a particular population?

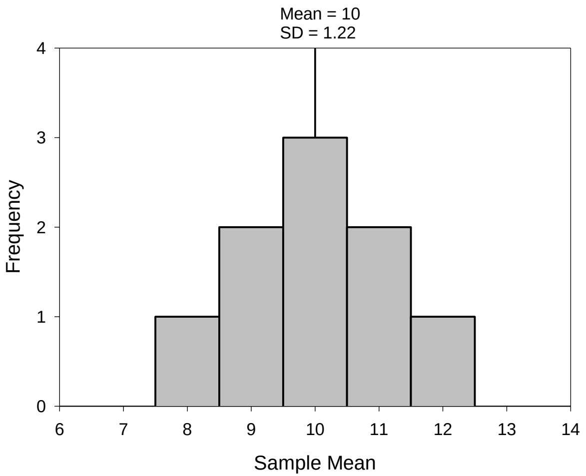
CLT - Central Limit Theorem (derived by statisticians)

Assuming the sampling distribution of sample means is created from large (N = 30+ scores) **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 ($\sigma_{\bar{X}}$) ….

Slides will be on Moodle. No need to copy.

Sampling distribution of sample means



Outline for Ch. 6

1. Review of key facts
2. Sampling distribution of sample means
3. **Central Limit Theorem**
   - **<u>Standard error</u>**
4. What's the probability of drawing a sample with a particular mean (or an even lower/higher mean) from a particular population?

Review

- Conceptually, what is meant by the **standard deviation** (in a regular distribution of scores for a given variable)?
  - the typical (or average or standard) distance between **individual participants' scores** in the sample and the **mean of all scores in that sample**.

EX: For our class distribution of Exam 1 scores, the typical distance between your nine individual exam scores and the mean score of 74 was **20 points**.

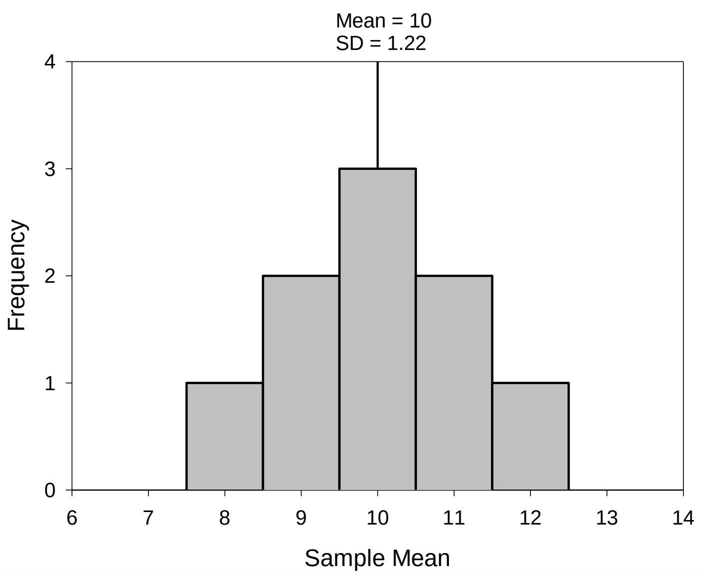
On average, each student's score is ~20 points away from the mean of 74

- Exam 1 Mean = 74%
- Exam 1 Standard deviation = 20 points

The standard deviation of sample means, aka **the standard error (SE)**

Conceptually, what's meant by the **standard error**?

- the typical distance between **individual *samples' means*** and the ***mean of all samples' means***, OR
- more simply, the extent to which the individual samples' means are clustered closely or widely scattered around their own average



The standard deviation of sample means, aka **the standard error (SE)**

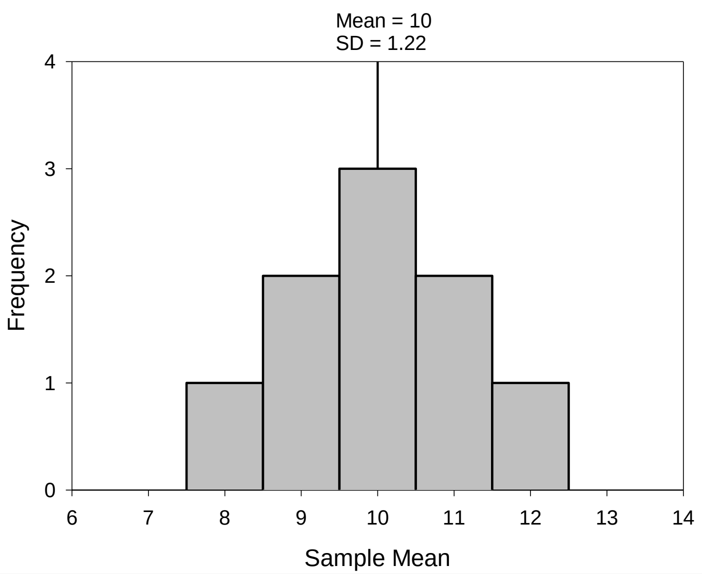
Conceptually, what's meant by the **standard error**?

- the typical distance between **individual *samples' means*** and the ***mean of all samples' means***, OR
- more simply, the extent to which the individual samples' means are clustered closely or widely scattered around their own average

standard error (*SE*)

$$\boxed{\sigma_{\overline{X}} = \frac{s}{\sqrt{N}}}$$

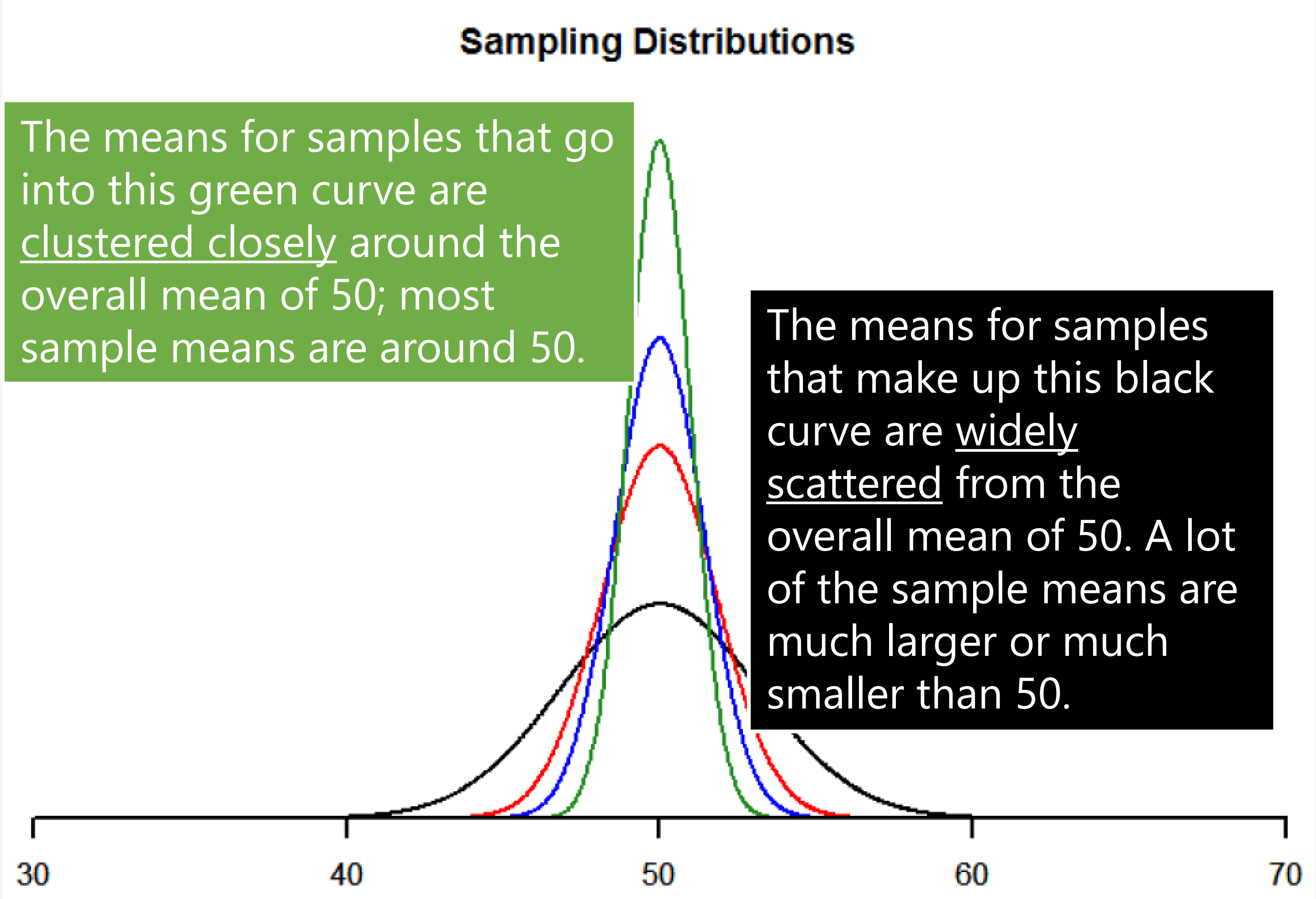
s = standard deviation calculated using your one sample's data
N = number of Ps in your one sample



Here are 4 different sampling distributions on the same set of axes

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

2. Which of these has the smallest *SE*? **GREEN**



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/small value for the standard error indicates that sample means are clustered fairly closely around the population mean.
3. The "standard error" can be described as the typical distance between an individual sample's mean, and the standard deviation of the population.

1. True
2. True
3. False

Corrected Statements from Prior Slide – two ways to correct #3

#3, *corrected*. The "standard error" can be described as the typical distance between an individual sample's mean, and the **MEAN OF ALL SAMPLE MEANS, WHEN THOSE SAMPLES ARE DRAWN RANDOMLY FROM THE SAME POPULATION**.

OR

#3, *corrected*. The "standard error" can be described as the typical distance between an individual sample's mean, and the **MEAN** of the population.

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 (i.e., the mean of the sampling distribution</u>)
  - **Low** *SE* indicates:
    - sample means are clustered closely around the population mean → *most* sample means in that sampling distribution 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 in that sampling distribution → 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/small* or *high/large*, assuming they are hoping to get an accurate estimate of the population mean from their one sample?

low (small)

Let's calculate the **standard error** for samples 2 and 3

- Sample 1

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

- Sample 2

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

- Sample 3

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

Bigger is better (in the case of samples)!

(standard error, or SE)

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

As *N* increases → *SE* decreases.

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