EDP308:
STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020

RAZ: Rebecca A. Zárate, MA

Overview

- Different Distributions
  - Population Distribution μ, σ
  - Data Distribution $\bar{x}, s$
  - Sampling Distribution $\mu_{\bar{x}}, \sigma_{\bar{x}}$
- Sampling Distribution
  - Distribution of your Sample Statistics
    - Notation
- The Central Limit Theorem
  - Normally Distributed Regardless of How Weird
  - Effects of Sample Size on Normality and Variation
- Standard Error
  - Calculating Standard Error
  - Effect of Sample Size on Standard Error
  - Effect of Variation on Standard Error

Different Distributions

Population Distribution: μ, σ

Data Distribution: $\bar{x}$, s

Sampling Distribution: $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$

Uncertainty

- We usually have to take a SAMPLE when we are trying to estimate a population PARAMETER (ex. mean, standard deviation)

What are some issues with taking a sample?

If we went out and took another sample, would our estimates be exactly the same?

Why or why not?

Uncertainty

If we went out and took another sample, would our estimates be exactly the same?

Why or why not?

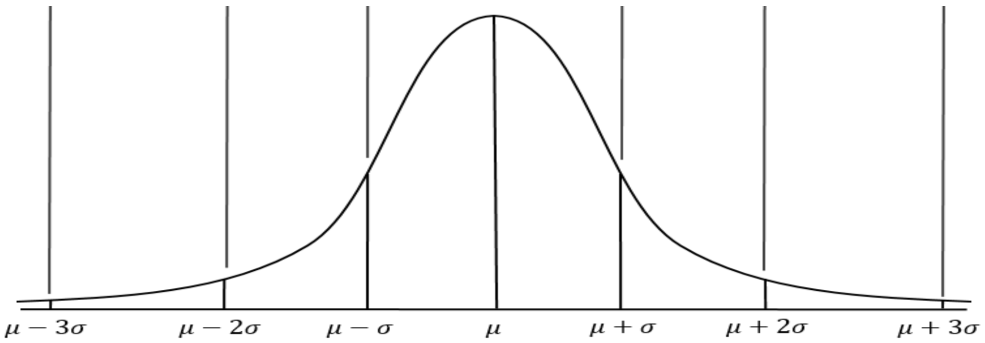
No… Most samples, even good ones, will be a little different from one another because of random chance associated with taking a sample.

Different Distributions

So far we have worked with both of these.

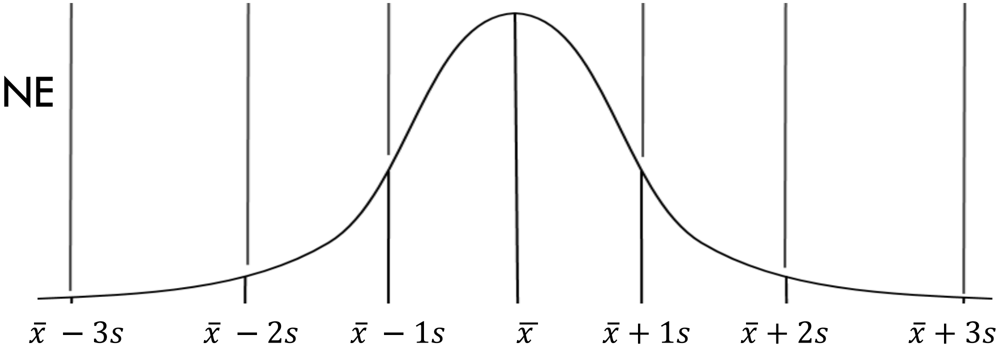
Population Distribution: μ, σ

Distribution for ENTIRE population



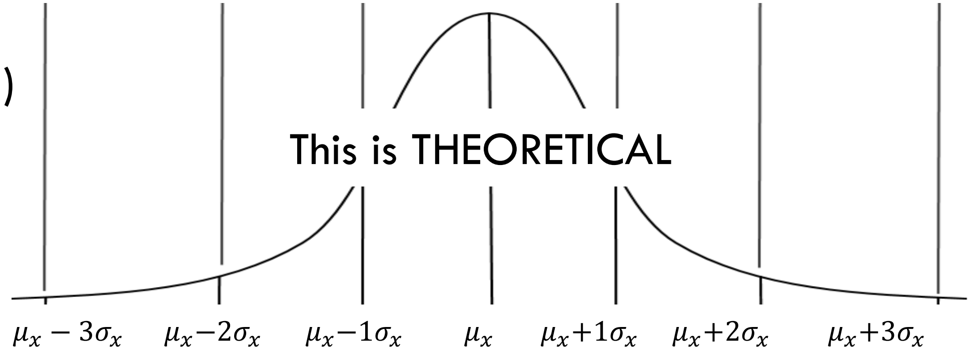
Data Distribution: $\bar{x}, \ s$

Distribution of the data from a ONE sample (a study) taken from the population



Sampling Distribution: $\mu_{\bar{x}}, \sigma_{\bar{x}}$

Distribution possible statistics ($\bar{x}, \ s$) of many samples



This is THEORETICAL

Take Note…

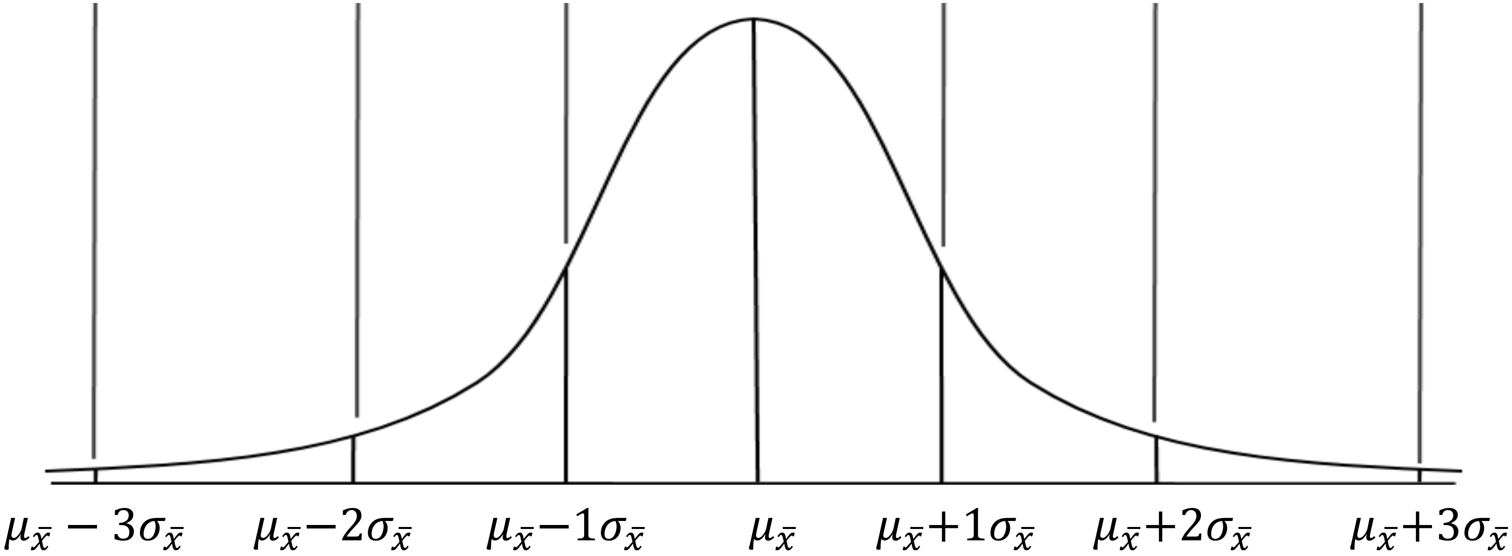
$$\text{SAMPL}\underline{\text{E}} \neq \text{SAMPL}\underline{\text{ING}}$$

(noun)

(not a verb in this case, rather an adjective)

Unfortunately, statisticians are terrible at naming things…
A “sample” is one single set of data,
while “sampling” refers to “sampling distribution”, a
theoretical distribution of all the sample sets.

Sampling Distribution


$\mu_{\bar{x}} - 3\sigma_{\bar{x}}$
$\mu_{\bar{x}}-2\sigma_{\bar{x}}$
$\mu_{\bar{x}}-1\sigma_{\bar{x}}$
$\mu_{\bar{x}}$
$\mu_{\bar{x}}+1\sigma_{\bar{x}}$
$\mu_{\bar{x}}+2\sigma_{\bar{x}}$
$\mu_{\bar{x}}+3\sigma_{\bar{x}}$

The sampling distribution is a theoretical distribution that shows us the distribution of a bunch of sample means ($\bar{x}$).

It's a SAMPLING distribution of STATISTICS (like the mean) we calculate from a SAMPLE.

Resampling

- If you took another sample from the population, do you think your statistics (i.e. mean and variance) would be the exact same?
  - Most likely no.

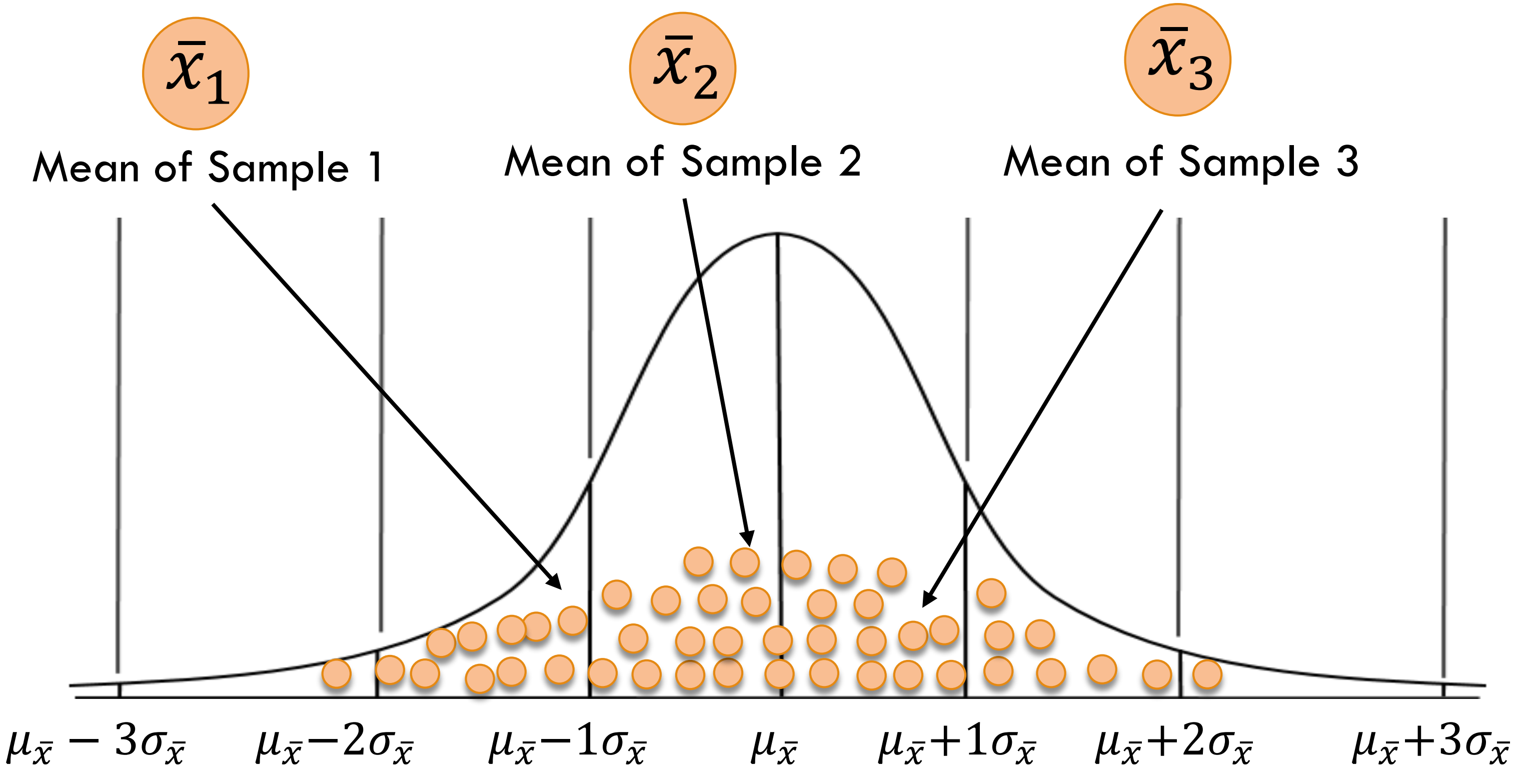
So how do we know (how can we gauge) how much each sample is going to vary from the other samples (i.e. another sample we take later)?

How much variation is there from *sample to sample*?

Distribution of Statistics

There are MEANS, not PEOPLE!

- Imagine each one of you in this class went out and took a sample of 100 people and asked how many hours a week they work on school outside of class. You calculated the average for *your* sample. Your classmates do the same. Now we have about 50 averages each calculated from a different sample. Then we can plot all these sample means on a curve to see how they distribute. This is the “Sampling Distribution”.



$\bar{x}_1$ Mean of Sample 1

$\bar{x}_2$ Mean of Sample 2

$\bar{x}_3$ Mean of Sample 3

$\mu_{\bar{x}} - 3\sigma_{\bar{x}}$ $\mu_{\bar{x}}-2\sigma_{\bar{x}}$ $\mu_{\bar{x}}-1\sigma_{\bar{x}}$ $\mu_{\bar{x}}$ $\mu_{\bar{x}}+1\sigma_{\bar{x}}$ $\mu_{\bar{x}}+2\sigma_{\bar{x}}$ $\mu_{\bar{x}}+3\sigma_{\bar{x}}$

Distribution of Statistics

Think back to what you know about variation.

In this case, do you think we want a lot of variation between the 50 calculated sample means? Or very little variation?

Do we want the data points to be clustered or spread out far away from each other?

Distribution of Statistics

Thinking back to what you know about variation, in this case, do you think we want a lot of variation between the 50 calculated sample means? Or very little variation?

Do we want the data points to be clustered for spread out far away from each other?

It would be great if the different sample statistics were very *similar* to each other.

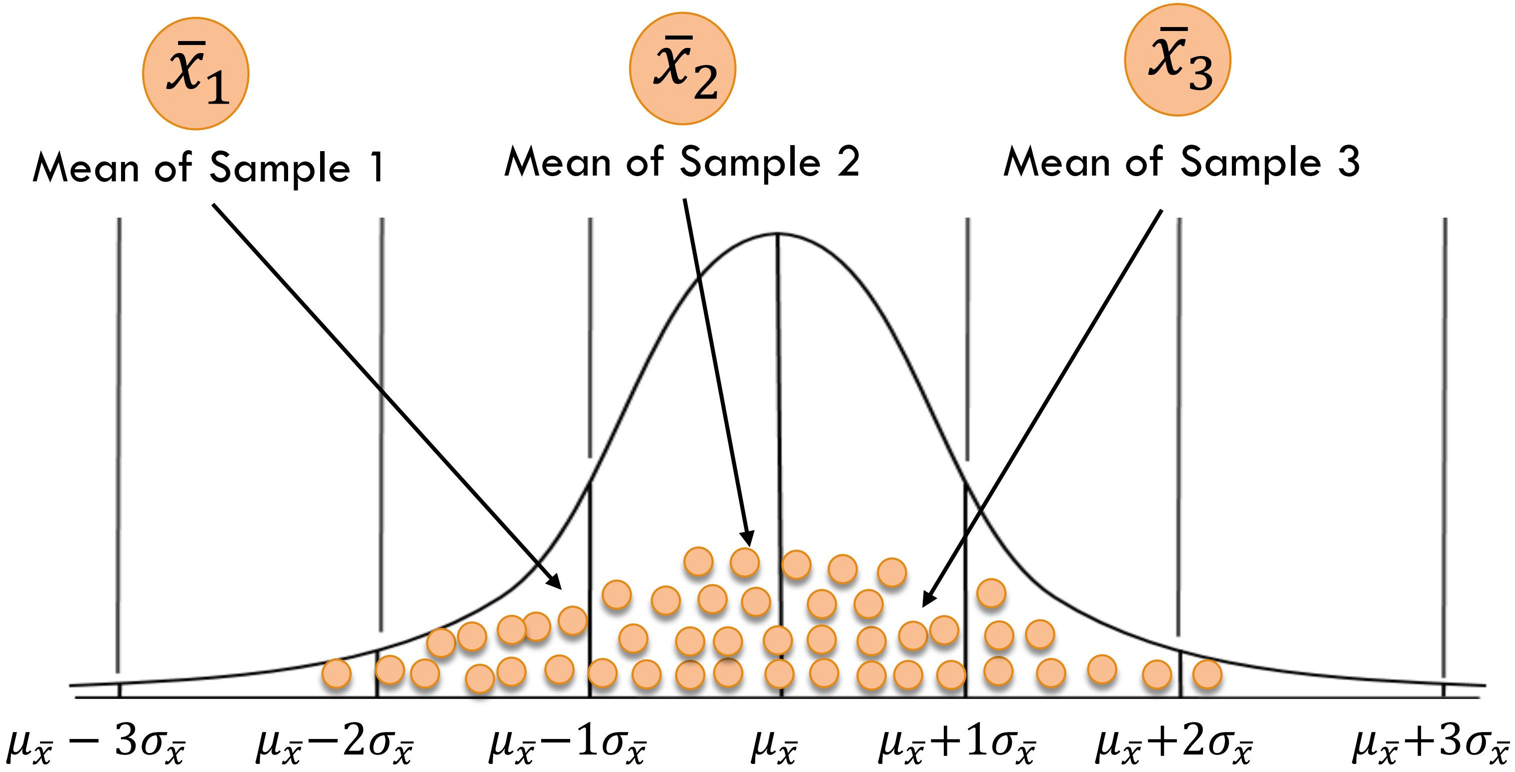
For example, if each of you went out to ask 100 students how many hours a week they work on school stuff outside of class, ideally, all of the averages you calculate should be fairly similar– assuming you are sampling correctly.

Would not be great if one person calculated an average of 2 hours and another person calculated an average of 55 hours. Lots of variation…

Distribution of Statistics

There are MEANS, not PEOPLE!

- □ Now, each one of these data points represents a statistic (ex. mean) calculated from one sample.



$\bar{x}_1$ Mean of Sample 1

$\bar{x}_2$ Mean of Sample 2

$\bar{x}_3$ Mean of Sample 3

$\mu_{\bar{x}} - 3\sigma_{\bar{x}}$
$\mu_{\bar{x}}-2\sigma_{\bar{x}}$
$\mu_{\bar{x}}-1\sigma_{\bar{x}}$
$\mu_{\bar{x}}$
$\mu_{\bar{x}}+1\sigma_{\bar{x}}$
$\mu_{\bar{x}}+2\sigma_{\bar{x}}$
$\mu_{\bar{x}}+3\sigma_{\bar{x}}$

Notations

So, if we are talking about a sample, where we would use $(\bar{x}, s)$, why then are we using $\mu_{\bar{x}}$ now? Isn't mu (μ) only for population data???

The Central Limit Theorem (CLT)

Central Limit Theorem states…

- As you take more samples (especially large n ones with replacement) and calculate statistics ($\bar{x}$, s), the distribution of those statistics (i.e. the sampl*__ing__* distribution) will look like a normal distribution
  - Even if the population *isn't* normally distributed
- And

$$\mu = \mu_{\bar{x}}$$

  - Now considered an "unbiased estimate"

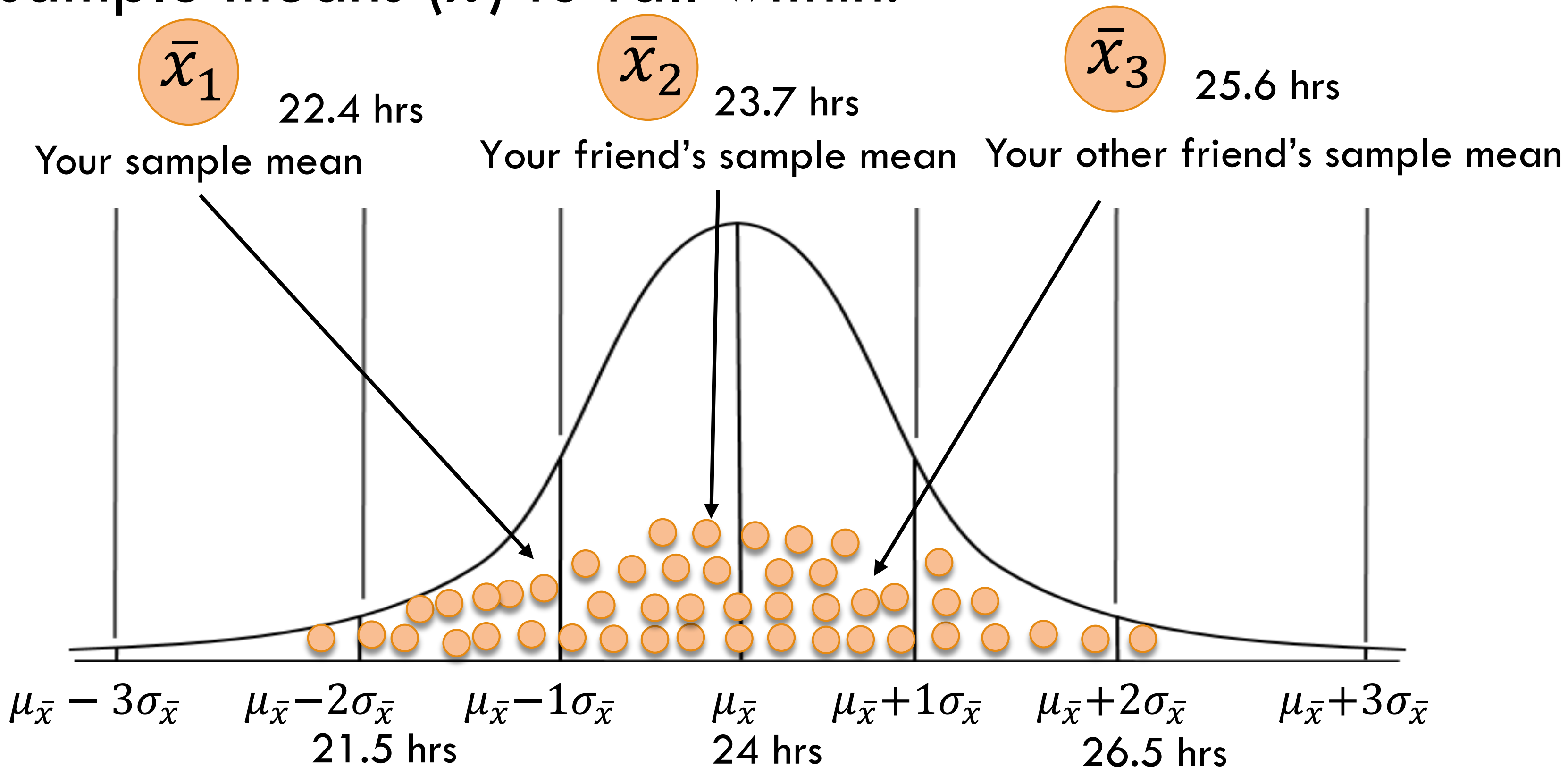
CLT… Huh?

- Again, imagine you and all of your classmates (about 50 of you) asked 100 people how much they work on school stuff outside of class. All of you comes together with your calculated means ($\bar{x}$), and we observe how they distribute. All of those means will be distributed normally because for each person that obtained a sample mean ($\bar{x}$) that was a little higher than the true mean (μ), there will be another student with a sample mean ($\bar{x}$) that is slightly lower than the true mean (μ). The average of all those sample averages is the sampling mean ($\mu_{\bar{x}}$) which will be equal to the true population mean (μ), the thing we are really interested in!

Distribution of Your Means

There are MEANS, not PEOPLE!

- □ Taken all together we can get a sampling mean ($\mu_{\bar{x}}$) and even a range that we would expect most of the sample means ($\bar{x}$) to fall within.



$\bar{x}_1$ 22.4 hrs

Your sample mean

$\bar{x}_2$ 23.7 hrs

Your friend's sample mean

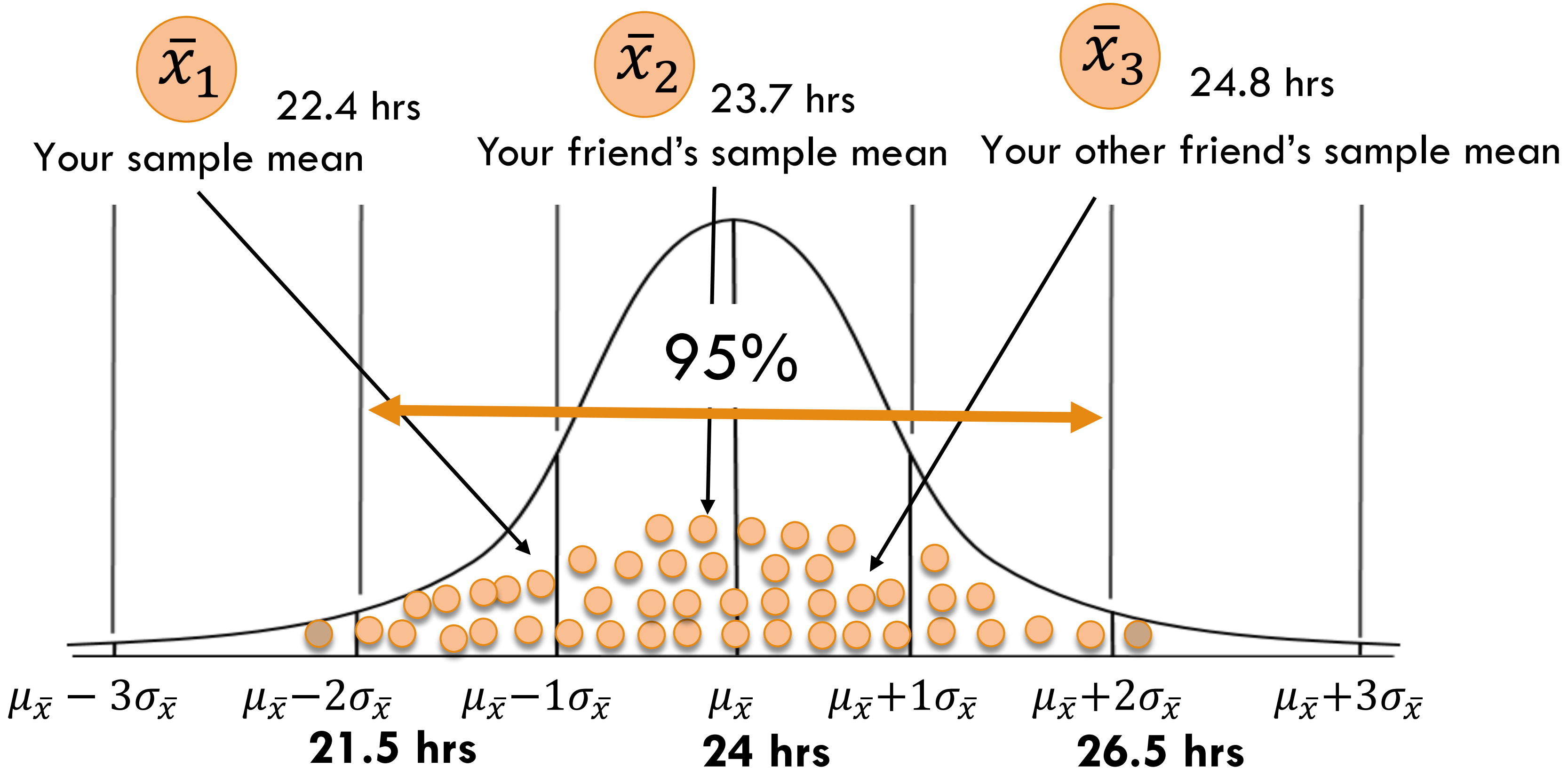
$\bar{x}_3$ 25.6 hrs

Your other friend's sample mean

$\mu_{\bar{x}} - 3\sigma_{\bar{x}}$
$\mu_{\bar{x}}-2\sigma_{\bar{x}}$
21.5 hrs
$\mu_{\bar{x}}-1\sigma_{\bar{x}}$
$\mu_{\bar{x}}$
24 hrs
$\mu_{\bar{x}}+1\sigma_{\bar{x}}$
$\mu_{\bar{x}}+2\sigma_{\bar{x}}$
26.5 hrs
$\mu_{\bar{x}}+3\sigma_{\bar{x}}$

Distribution of Your Means

There are MEANS, not PEOPLE!


$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
95%
$\mu_{\bar{x}} - 3\sigma_{\bar{x}}$
$\mu_{\bar{x}}-2\sigma_{\bar{x}}$
$\mu_{\bar{x}}-1\sigma_{\bar{x}}$
$\mu_{\bar{x}}$
$\mu_{\bar{x}}+1\sigma_{\bar{x}}$
$\mu_{\bar{x}}+2\sigma_{\bar{x}}$
$\mu_{\bar{x}}+3\sigma_{\bar{x}}$

22.4 hrs

Your sample mean

23.7 hrs

Your friend's sample mean

24.8 hrs

Your other friend's sample mean

21.5 hrs

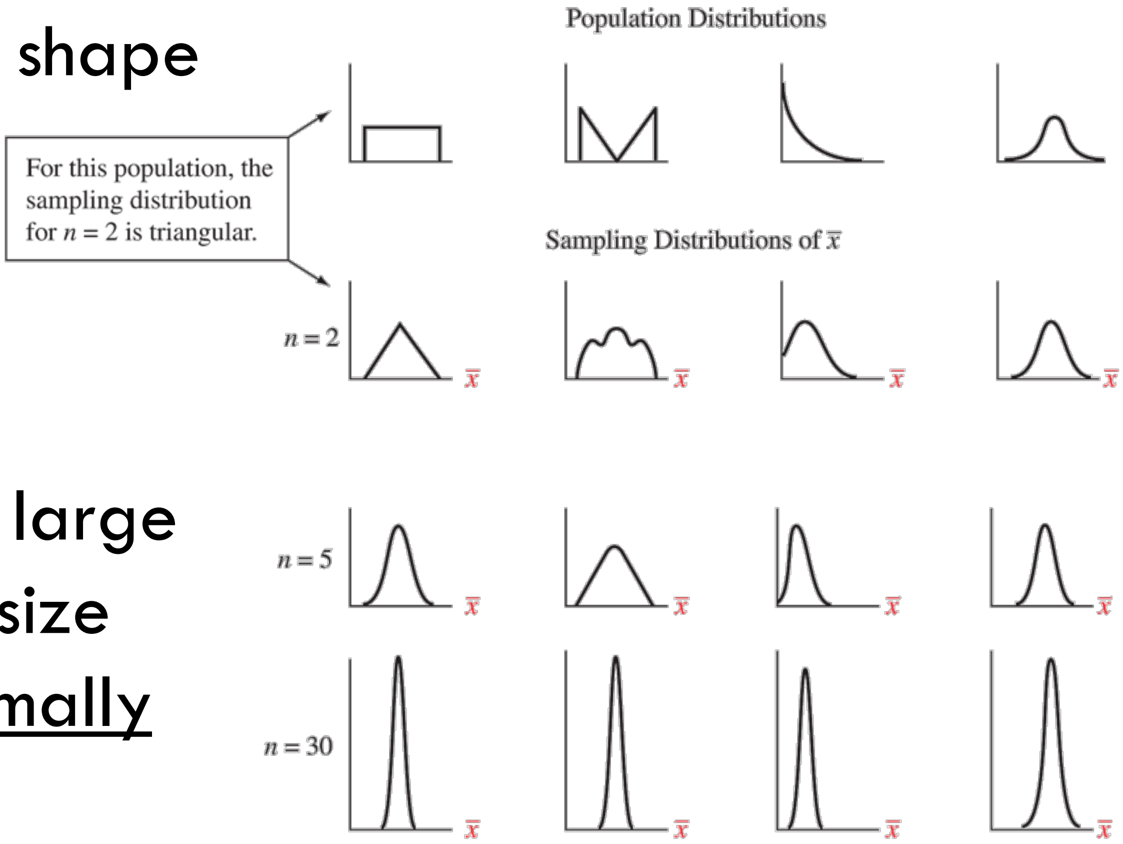
24 hrs

26.5 hrs

In this (made up) example, the true population mean (μ) is 24 hours and 95% of all sample means will fall between 21.5 hours and 26.5 hours. Notice how almost all 50 observations are within 21.5 and 26.5

Approaching Normality n ≥30

- No matter what shape the population distribution is the SAMPLING distribution will, eventually, with large enough sample size (n ≥30), be normally distributed…


Population Distributions
For this population, the sampling distribution for $n = 2$ is triangular.
Sampling Distributions of $\bar{x}$
$n = 2$
$n = 5$
$n = 30$
$\bar{x}$

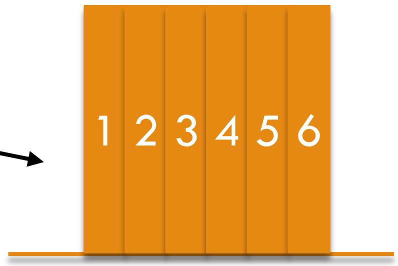
Normal Regardless of How Weird

- Let's see how this could be the case...
- What is the average of rolling a die X times?
  - Roll a die 10 times and take the average
    - Maybe the average ($\bar{x}$) was 4
  - Do that again (rolling 10x) but do that 100x
    - Take the average of all those averages
      - 2.5, 4.5, 3, 5.5, 2, etc.
    - Look at the distribution of those averages

Die Rolls have a Uniform Distribution

Each side has equal probability

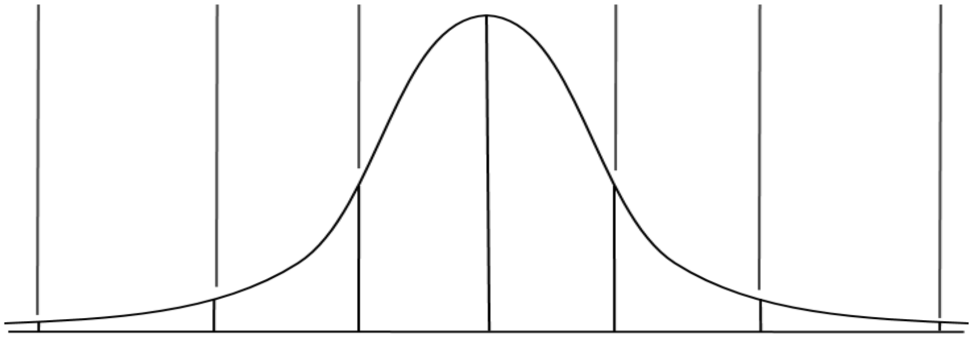
a 1 in 6 chance.



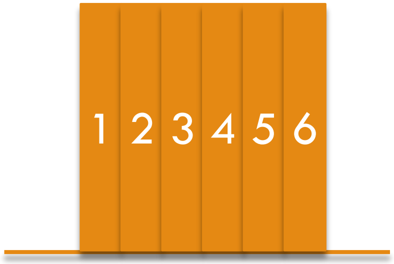
Normal Regardless of How Weird

- Let’s see how this could be the case… What is the average of rolling a die X times?
  - Roll a die 10 times and take the average
    - Maybe the average ($\bar{x}$) was 4
  - Do that again (rolling 10x) 100x
    - Take the average (μ_x) of all those averages
      - 2.5, 4.5, 3, 5.5, 2, etc.

    - Look at the distribution of those averages
      - It will be normally distributed and $\mu_{\bar{x}}$ will equal μ with it 3.5
    - Even though a die has a uniform distribution, the *sampling* distribution will be normally distributed.



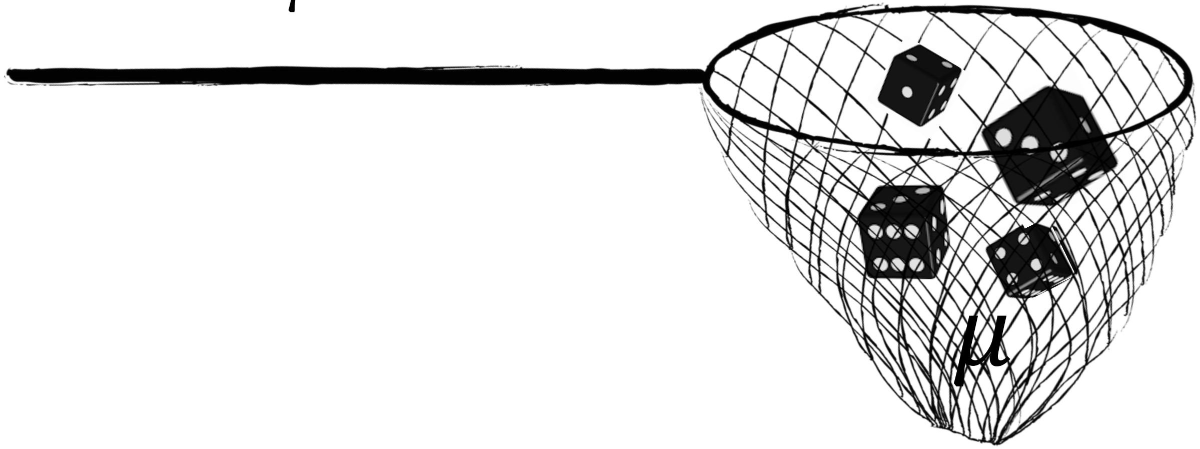
$\mu_{\bar{x}} = 3.5$



The Effects of Sample Size

The Bigger the Better

- In our die rolling example, we saw that the mean for any one sample was rarely if ever going to be exactly 3.5, the true population mean.
- With a small sample size, there will be more variation. Imagine only rolling a die 2x.
  - How accurate do you think that one average would be?
- Larger sample sizes (along with random sampling) are like casting a big net to capture as much "truth" as you can, so we can figure out true μ



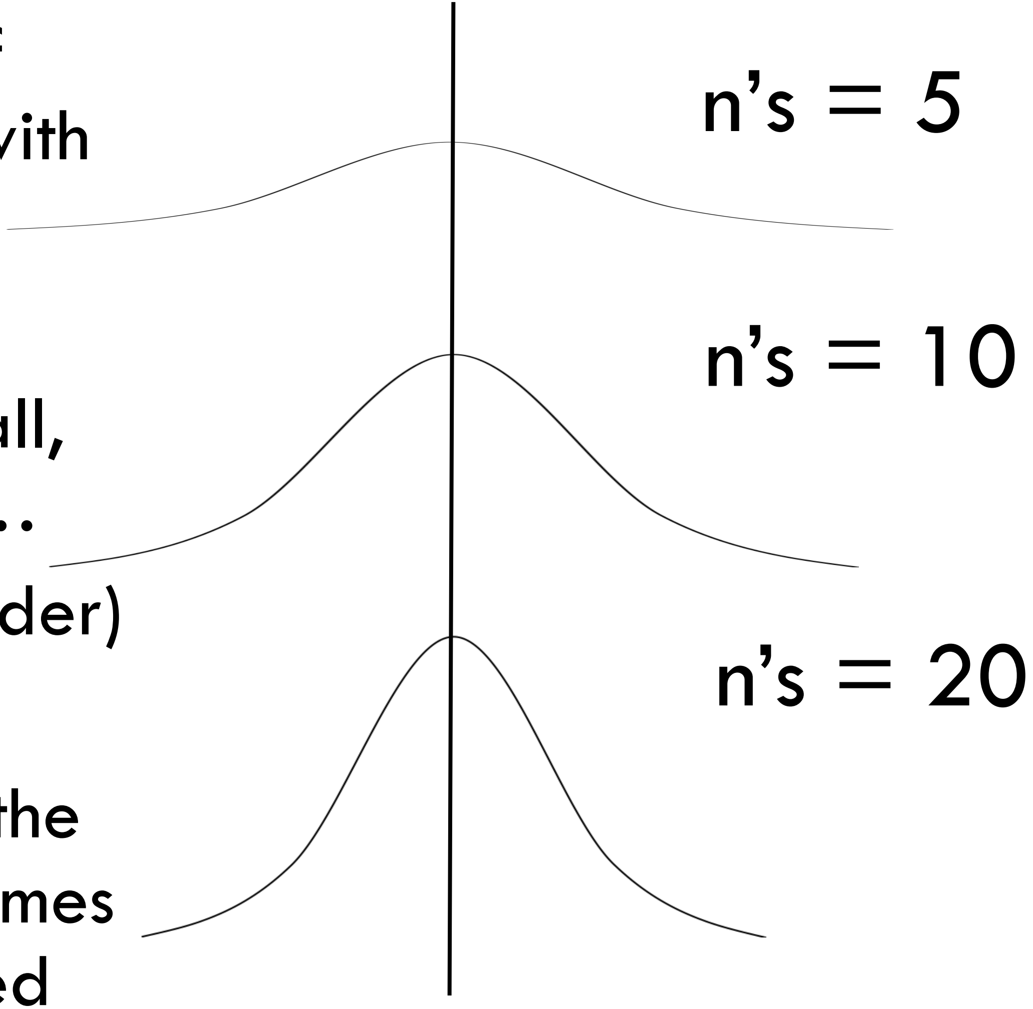
Compare Sample Sizes

These are examples of SAMPLING distributions with different sample sizes

When sample size is small, there is more variation…

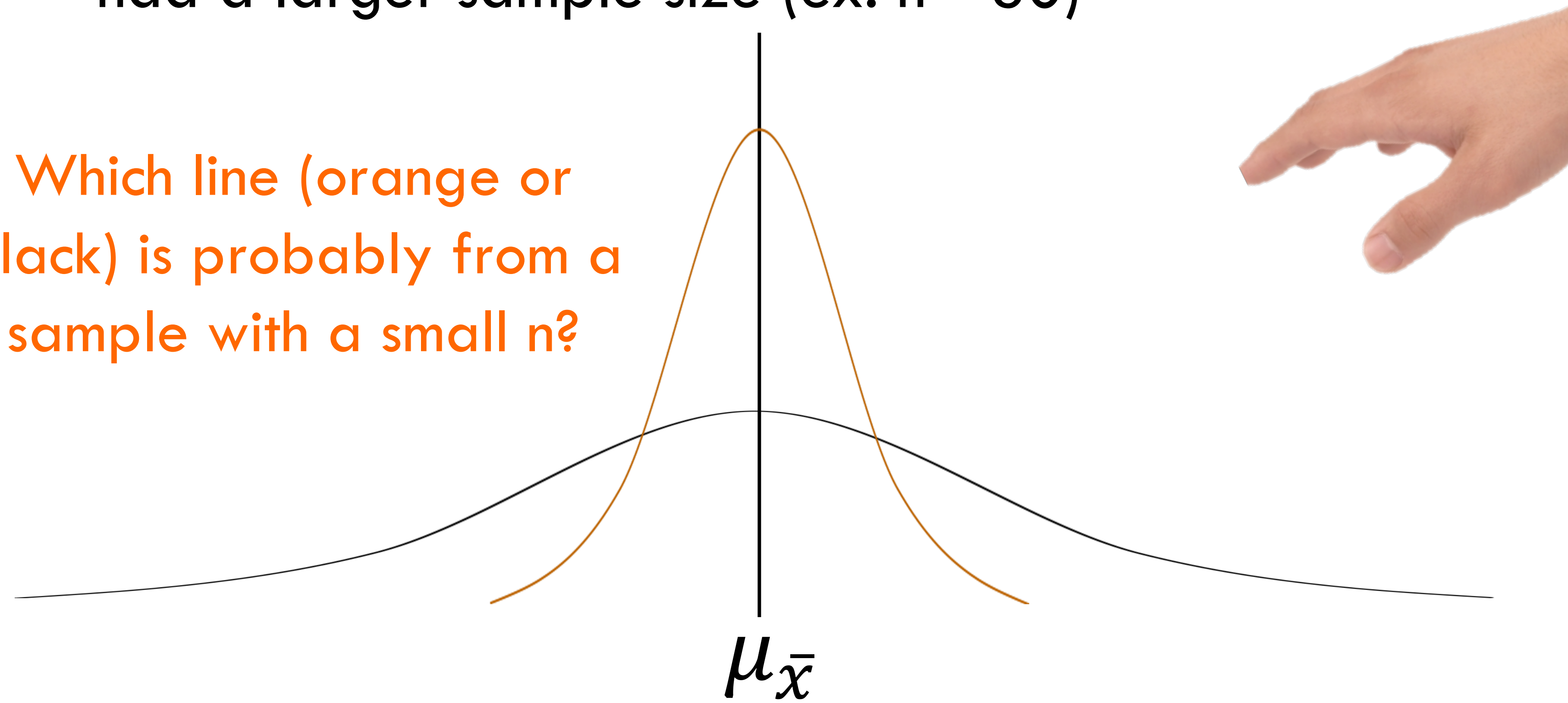
(the curve is flatter and wider)

As sample size increase, the sampling distribution becomes more normally distributed



Smaller n, more variability $\sigma_{\bar{x}}$

- □ When your sample is only n = 5, the average you calculate will vary more than it would if you had a larger sample size (ex. n =30)



Which line (orange or black) is probably from a sample with a small n?

Closer

● = $\bar{x}$ within 1 $\sigma_{\bar{x}}$ from true μ

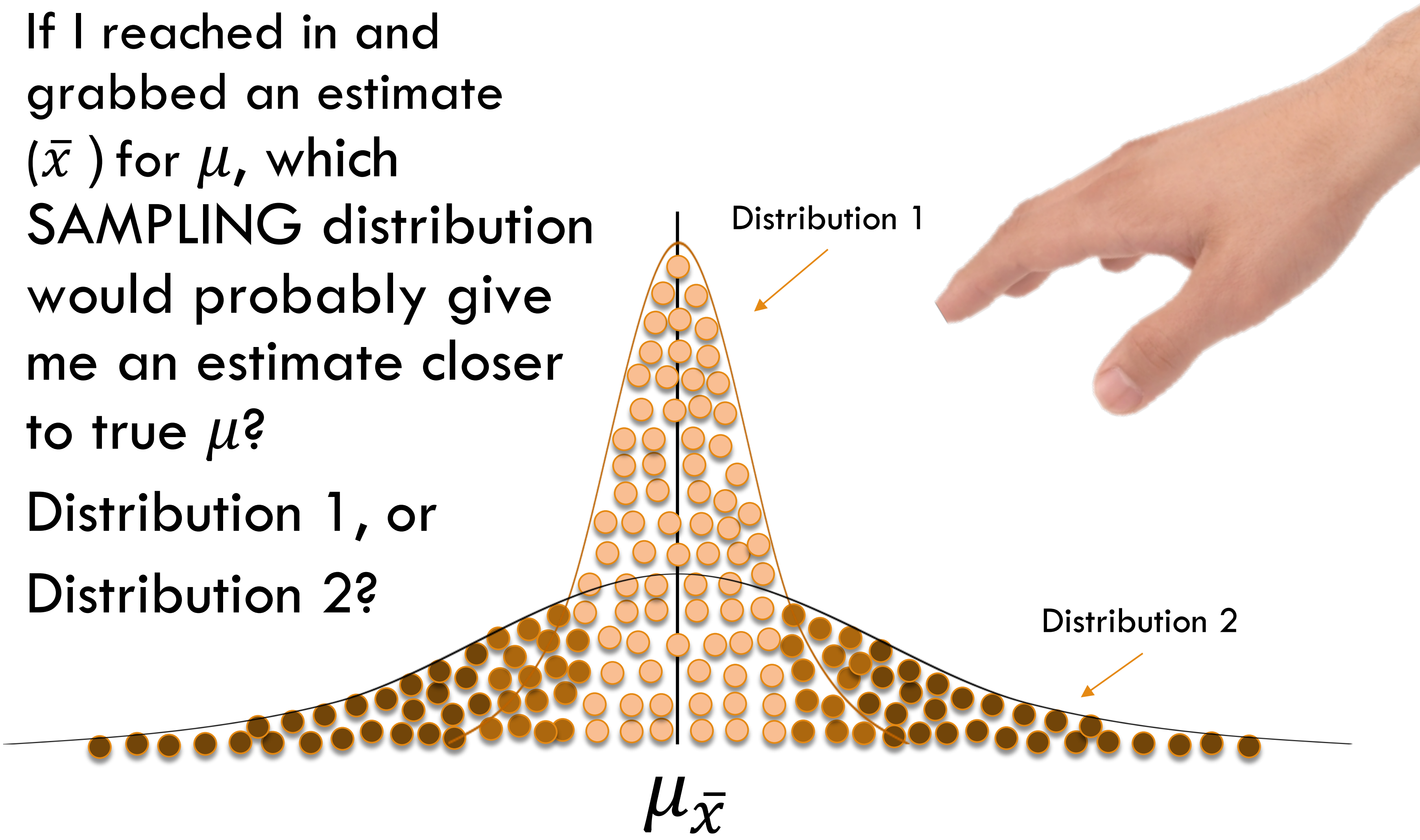
● = $\bar{x}$ within 2 $\sigma_{\bar{x}}$ from true μ

● = $\bar{x}$ within 3 $\sigma_{\bar{x}}$ from true μ

If I reached in and grabbed an estimate ($\bar{x}$) for μ, which SAMPLING distribution would probably give me an estimate closer to true μ?

Distribution 1, or

Distribution 2?



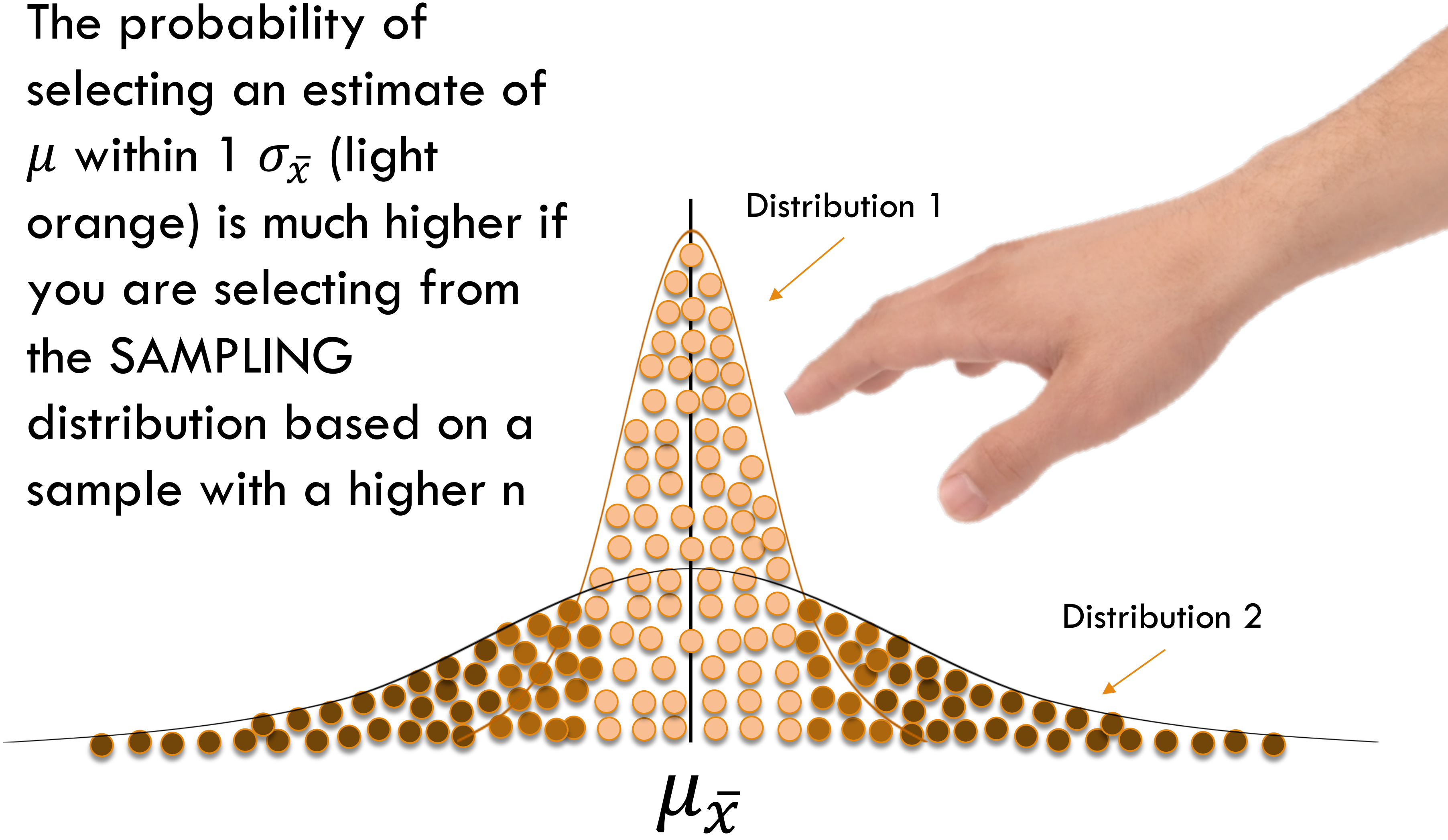
Closer

● = $\bar{x}$ within 1 $\sigma_{\bar{x}}$ from true μ

● = $\bar{x}$ within 2 $\sigma_{\bar{x}}$ from true μ

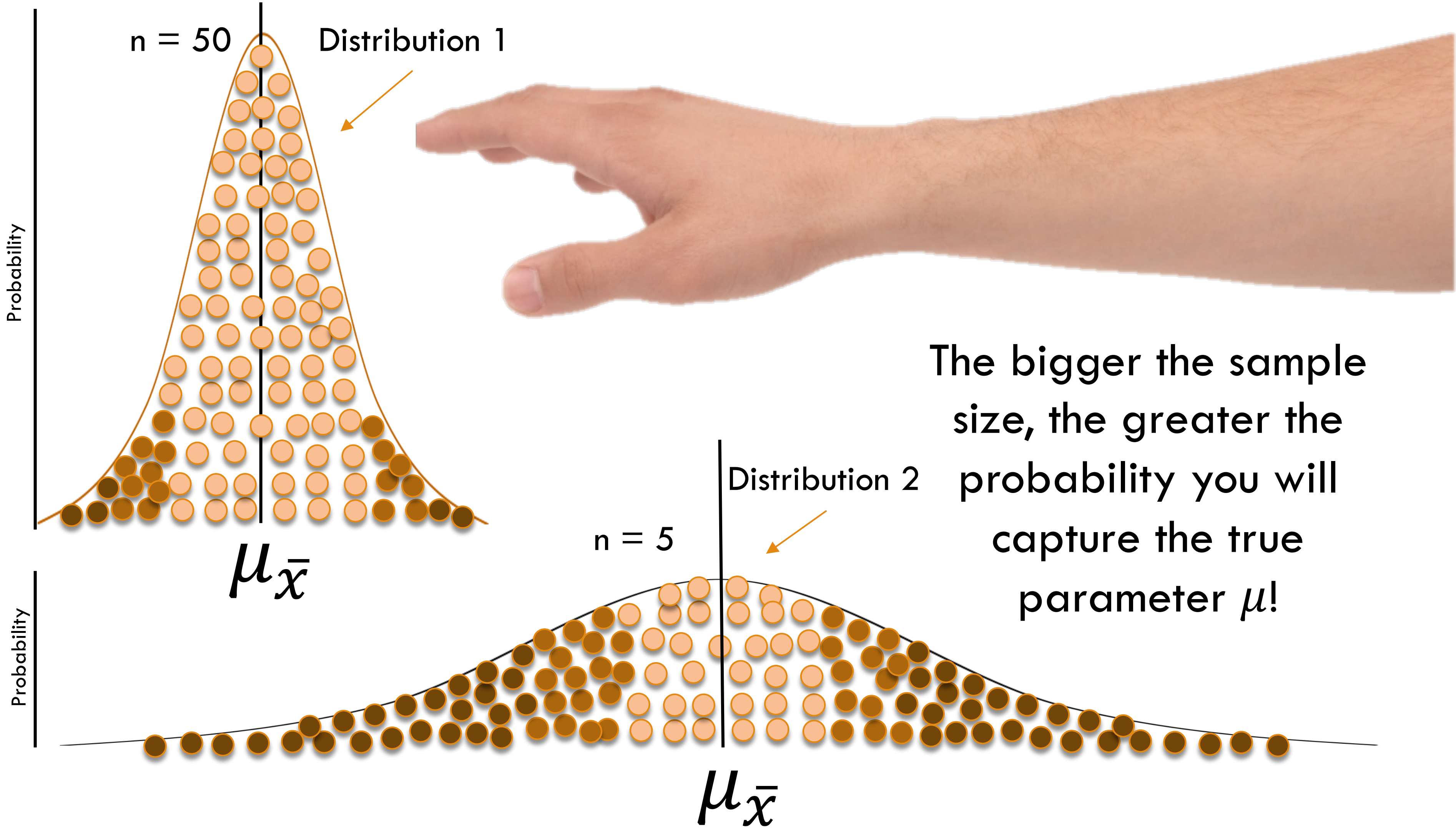
● = $\bar{x}$ within 3 $\sigma_{\bar{x}}$ from true μ

The probability of selecting an estimate of μ within 1 $\sigma_{\bar{x}}$ (light orange) is much higher if you are selecting from the SAMPLING distribution based on a sample with a higher n



Closer

● = $\bar{x}$ within 1 $\sigma_{\bar{x}}$ from true μ
● = $\bar{x}$ within 2 $\sigma_{\bar{x}}$ from true μ
● = $\bar{x}$ within 3 $\sigma_{\bar{x}}$ from true μ



The bigger the sample size, the greater the probability you will capture the true parameter μ!

Standard Error ($\sigma_{\bar{x}}$)

Confusing Naming… Again

- Unfortunately, there is a new name for standard deviation when we are talking about SAMPLING distributions…
- For Populations (σ)
  - “Sigma” (though most just say standard deviation)
- For Samples (s)
  - “Standard deviation”
- For Sampling Distributions
  - “Standard ERROR” ($\sigma_{\bar{x}}$)

Another measure of variability??

- You might have noticed this $\sigma_{\bar{x}}$ in the last few slides… Remember σ is standard deviation, or the average amount people differ from the mean. But now we are dealing with sample means rather than people, so ***$\boldsymbol{\sigma_{\bar{x}}}$ is the average amount the sample means differ from the true mean ($\boldsymbol{\mu}$).***
  - $\sigma_{\bar{x}}$ is called the "Standard Error"
- It is the measure of variability for a sampling distribution (rather than one sample), like the standard deviation of a sampling distribution, and it is determined by the standard deviation of the population and the size of the sample.

Calculating Standard Error

- When μ and σ are known, meaning they are population data, we can calculate standard error ($\sigma_{\bar{x}}$) for a given sample by taking the standard deviation of a population (σ) and divide it by the square root of the sample size for a given sample.

$$\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$$

- When μ and σ are unknown and you are using sample data, $\bar{x}$ and s, we have to *estimate* standard error ($\sigma_{\bar{x}}$) for the sample by taking the standard deviation (s) that you calculated from your sample and divide it by the square root of the sample size for your sample.

$$\frac{s}{\sqrt{n}} \underset{\uparrow}{\approx} \sigma_{\bar{x}}$$

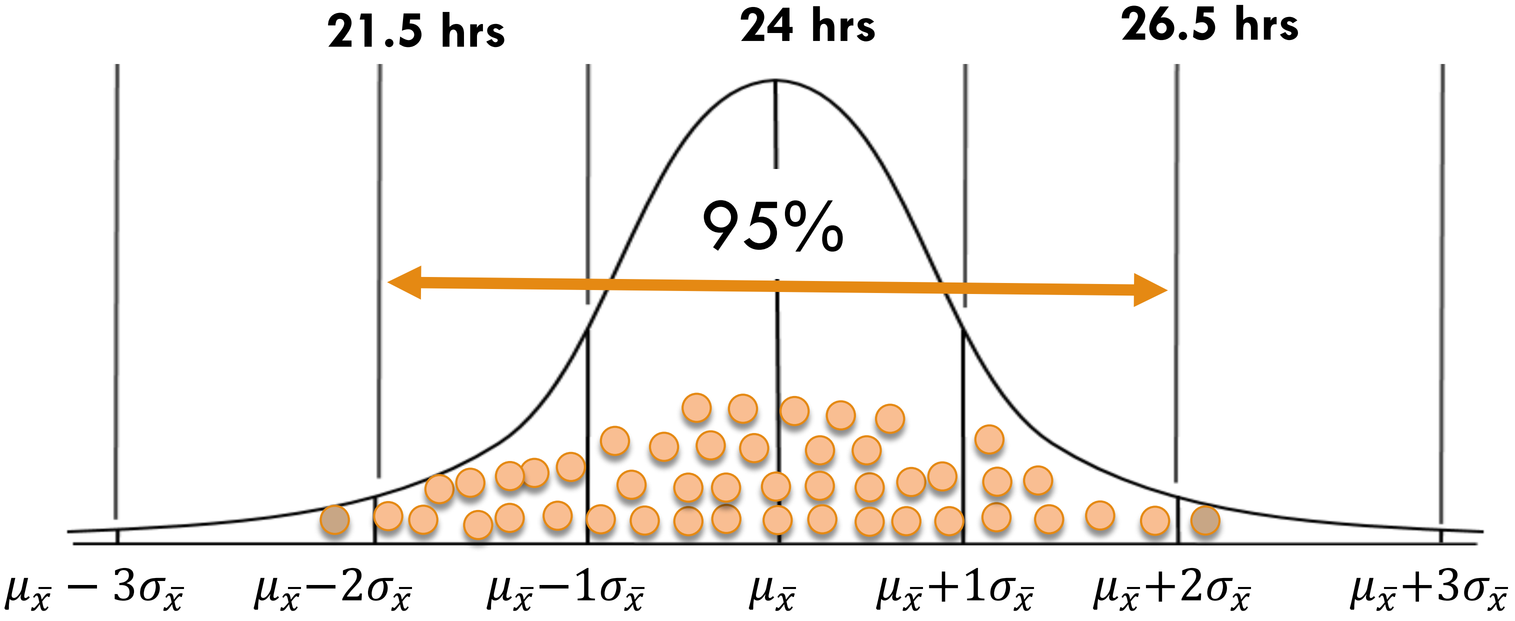
"approximately equal to"

Distribution of Your Means

There are MEANS, not PEOPLE!

- Back to our example, here the standard error ($\sigma_{\bar{x}}$) is 1.25. Based on the 68%, 95%, 99% rule, we know that 68% of the sample means will be between 22.75 and 25.25 hours, 95% will be between 21.5 and 26.5 hours, and 99% of the means will be between 20.25 and 27.75.

$$\mu_{\bar{x}} = 24$$
$$\sigma_{\bar{x}} = 1.25$$



Samples Size (n) and Standard Error (SE)

$\frac{s}{\sqrt{n}} \approx \sigma_{\bar{x}}$

$$\frac{\sigma}{\sqrt{n}} = \text{STANDARD ERROR} = \sigma_{\bar{x}}$$

What is the relationship between Standard Error (SE) and n?

Samples Size (n) and Standard Error ($\sigma_{\bar{x}}$)

- As sample size increases, the larger n can now, as the denominator, “eat up” some of that deviation
  - **Inversely related**

$$\frac{s}{\sqrt{n}} \approx \sigma_{\bar{x}}$$

$$\frac{\sigma}{\sqrt{n}} = \text{STANDARD ERROR} = \sigma_{\bar{x}}$$

↑ n then ↓ $\sigma_{\bar{x}}$

↓ n then ↑ $\sigma_{\bar{x}}$

Variation (σ) and Standard Error ($\sigma_{\bar{x}}$)

- □ As variation from your sample increases, larger standard deviation of your sample, the larger your standard error will be
  - **Directly related**

$$\frac{s}{\sqrt{n}} \approx \sigma_{\bar{x}}$$

$$\frac{\sigma}{\sqrt{n}} = \text{STANDARD ERROR} = \sigma_{\bar{x}}$$

↑ σ then ↑ $\sigma_{\bar{x}}$

Or s

↓ σ then ↓ $\sigma_{\bar{x}}$

Or s

Reality Check

- In reality we are not going to take numerous samples (because we are very busy and poorly funded researchers) and rarely do we know μ, σ, rather, we use what we know about distributions and the central limit theorem.
- We take ONE sample and try to uncover the truth with our one sample and, in some cases, use the sample to make inferences about the population.
  - That's the good stuff!

Up Next…

- Using what we know about sampling distributions to create an interval around a sample mean, and interval of confidence…

Confidence Intervals