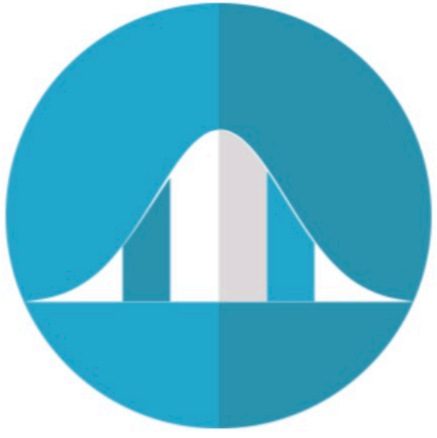
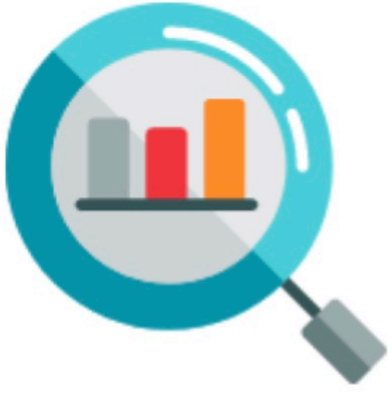


Statistics
Session-4



CLARUSWAY©
WAY TO REINVENT YOURSELF

Did you finish Statistics

(Central Limit Theorem and Confidence Intervals)

pre-class activity?

Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

Table of Contents

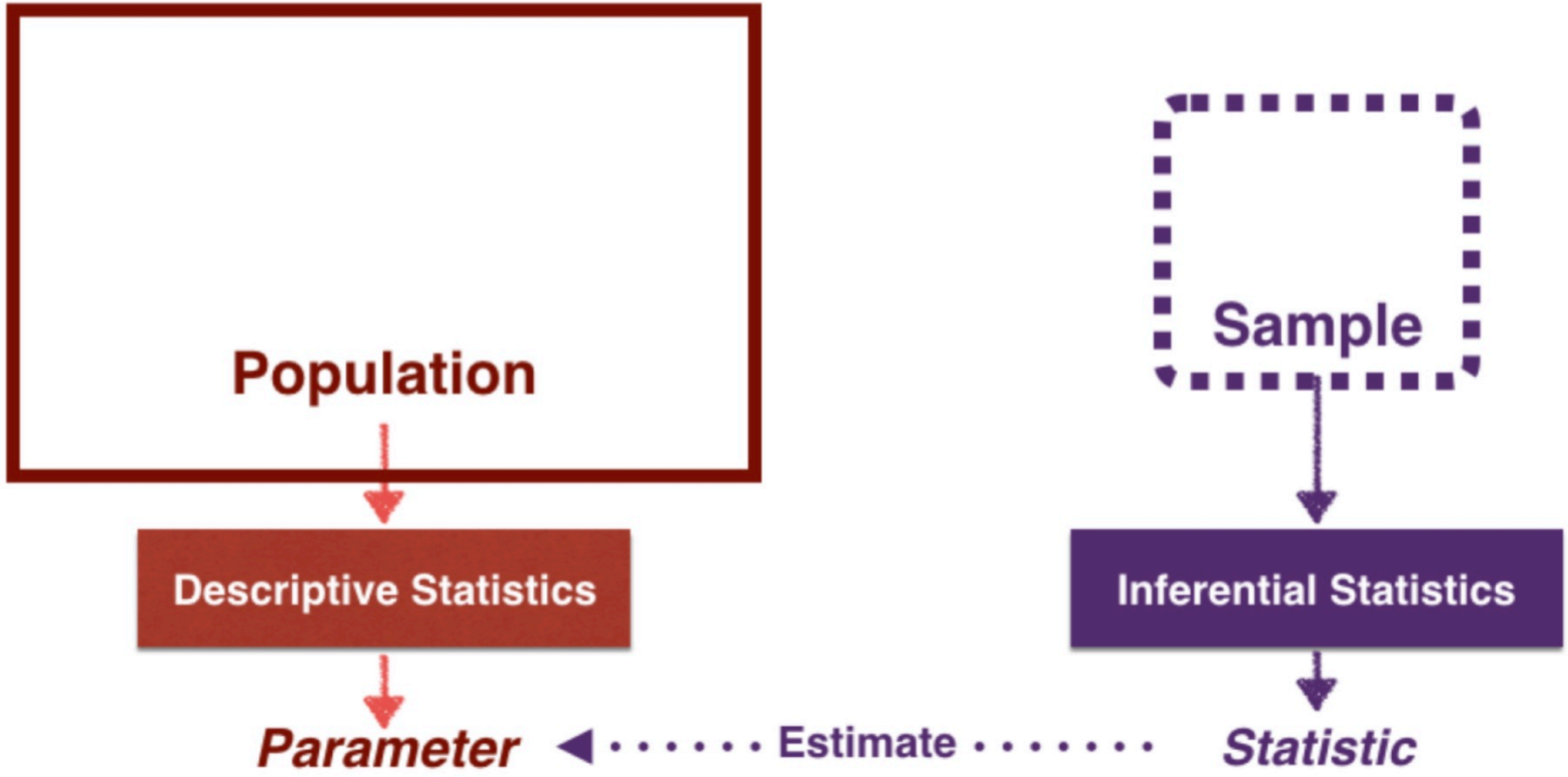
1 Sampling Distributions

Definition: Statistic

A statistic is a function of observable data and known constants

Generally, a statistic is used to estimate the value of a population parameter.

Sample Statistic



Definition: Sampling Distribution

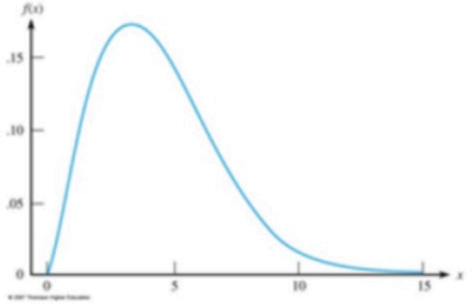
A sampling distribution is the probability distribution of a statistic (e.g. mean)

And the standard deviation of this statistic is called the standard error.

The concept of a sampling distribution is perhaps the most basic concept in inferential statistics.

Random Statistics

Consider drawing samples from a Weibull distribution



Six samples of size n=10 drawn from a Weibull distribution

Table 5.1 Samples from the Weibull Distribution of Example 5.19

Sample	1	2	3	4	5	6
1	6.1171	5.07611	3.46710	1.55601	3.12372	8.93795
2	4.1600	6.79279	2.71938	4.56941	6.09685	3.92487
3	3.1950	4.43259	5.88129	4.79870	3.41181	8.76202
4	0.6694	8.55752	5.14915	2.49759	1.65409	7.05569
5	1.8552	6.82487	4.99635	2.33267	2.29512	2.30932
6	5.2316	7.39958	5.86887	4.01295	2.12583	5.94195
7	2.7609	2.14755	6.05918	9.08845	3.20938	6.74166
8	10.2185	8.50628	1.80119	3.25728	3.23209	1.75468
9	5.2438	5.49510	4.21994	3.70132	6.84426	4.91827
10	4.5590	4.04525	2.12934	5.50134	4.20694	7.26081
$\bar{x}$	4.401	5.928	4.229	4.132	3.620	5.761
$\tilde{x}$	4.360	6.144	4.608	3.857	3.221	6.342
s	2.642	2.062	1.611	2.124	1.678	2.496

Note that the sample means, medians, and standard deviations are all different – randomness!

7

* Figure and table from *Probability and Statistics for Engineering and the Sciences*, 7th ed., Duxbury Press, 2008.

Sampling Distributions *Example*

- Let X_1 = # of spots on die #1.
- Let X_2 = # of spots on die #2.
- Then X_1, X_2 can be thought of as a system, where X_1 and X_2 have the same pmf or population distribution:

x	1	2	3	4	5	6
p(X)	1/6	1/6	1/6	1/6	1/6	1/6

Sampling Distributions *Example*

x	1	2	3	4	5	6
p(X)	1/6	1/6	1/6	1/6	1/6	1/6

$$\mu = \sum_x xp(x) = 1(1/6) + 2(1/6) + \ldots + 6(1/6) = 3.5$$

$$\sigma^2 = \sum_x (x-\mu)^2 p(x) = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots + (6-3.5)^2(1/6) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

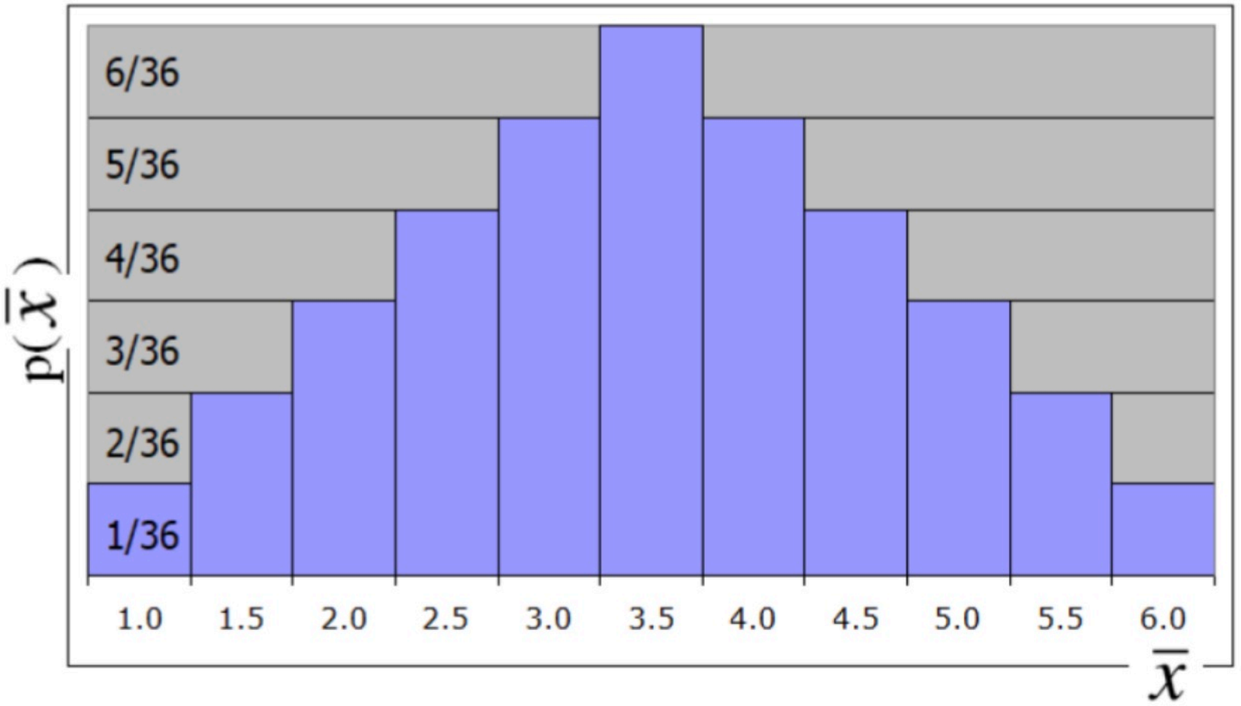
Sampling Distribution of $\bar{X}$

- Find the sampling distribution of $\bar{X} = \frac{X_1 + X_2}{2}$
- A sampling distribution is derived by looking at all samples of size n = 2 (i.e. two dice) and their means...

Sample		Sample		Sample	
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

Sampling Distribution of $\bar{X}$

$\bar{x}$	$p(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



$$\mu_{\bar{X}} = \sum_{\bar{x}} \bar{x} p(\bar{X} = \bar{x}) = 1(1/36) + 1.5(2/36) + \ldots + 6(1/36) = 3.5$$

$$\sigma^2_{\bar{x}} = \sum_{\bar{x}} (\bar{x} - \mu)^2 p(\bar{X} = \bar{x}) = (1-3.5)^2(1/36) + (1.5-3.5)^2(2/36) + \ldots + (6-3.5)^2(1/36) = 1.46$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.46} = 1.21$$

Sampling Distribution of the Sample Mean

Let $X_1,\ldots, X_n$ be a simple random sample from a distribution with mean value μ and standard deviation σ. Then

$$1.\ E\left(\overline{X}\right) = \mu_{\overline{X}} = \mu$$

$$2.\ V\left(\overline{X}\right) = \sigma^2_{\overline{X}} = \sigma^2 / n$$

$$3.\ \sigma_{\overline{X}} = \sigma\ / \sqrt{n}$$

} only because $X_1, \ldots, X_n$ are independent!

Standard Error of the Mean

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample's mean

n = the sample size

The sales of food and drink in Aunt Erma's Restaurant vary from day to day. Past records indicate that the daily sales follow a probability (population) distribution with a mean of **μ = $900** and a standard deviation of **σ = $300**.

How much variability would you expect in the weekly sample mean sales figures?
Find the **standard deviation of the sampling distribution of the sample mean**, and interpret this standard deviation.

Students, write your response!

Pear Deck Interactive Slide
Do not remove this bar

No Text Response
You didn't answer this question

SOLUTION

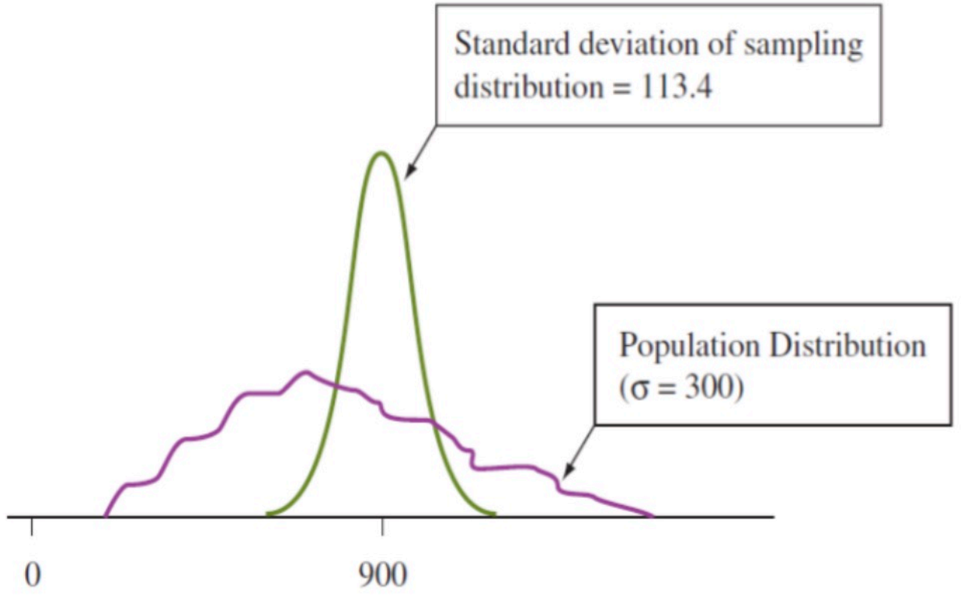
μ = $900

σ = $300

The sampling distribution of the sample mean for *n = 7* has *mean $900*.

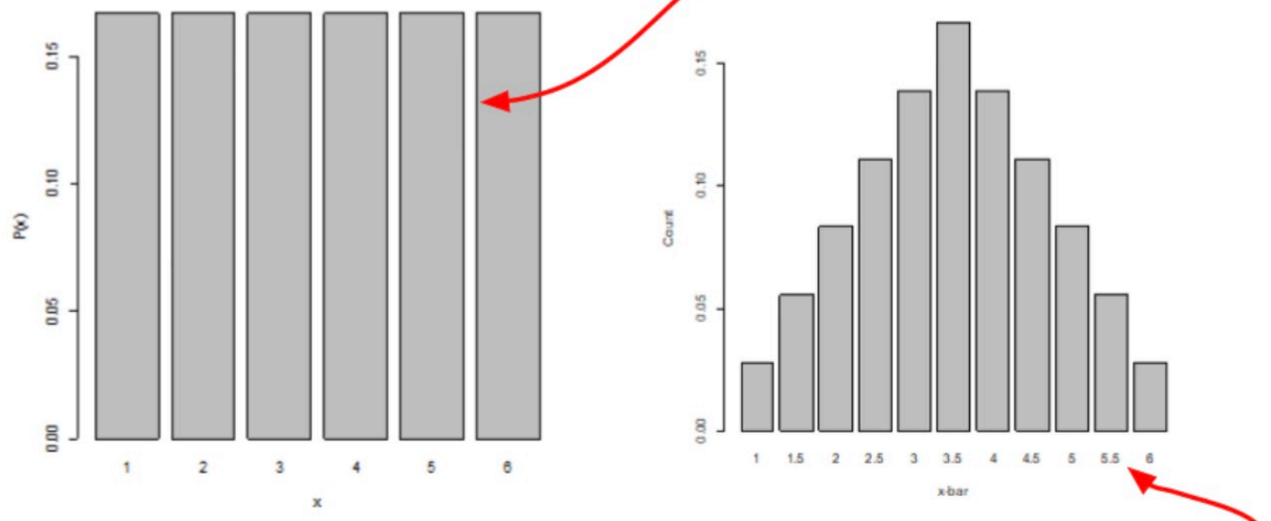
Its standard deviation equals

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{7}} = 113.4$$



Sampling Distribution of $\overline{X}$

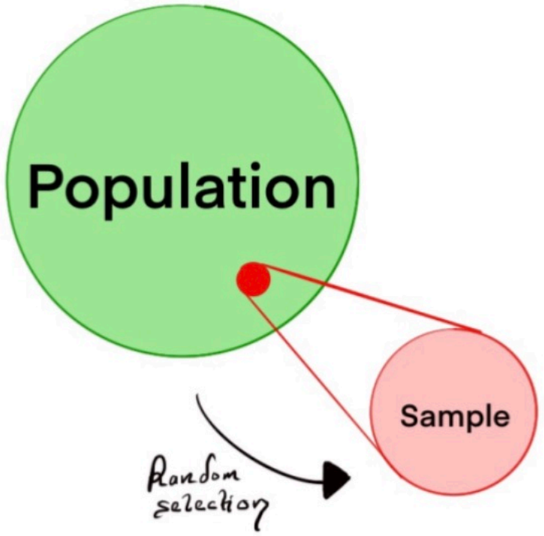
Compare the distribution of X...



...with the sampling distribution of $\overline{X}$

As well, note that mean of the sampling distribution is the same as the population mean, but that the sampling distribution is **less variable**.

What is the standard error of the mean?

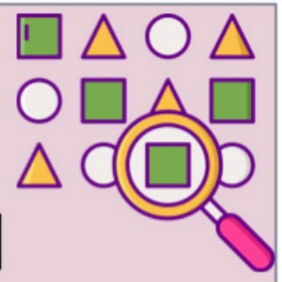


Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

Sampling Distributions and Inferential Statistics



- The examples specified a population
  - sampling distributions were determined from this population
- In practice, the process is the other way around
  - Estimate parameters of sampling distribution from sample data

Sampling Distributions

Keep in mind that all statistics have sampling distributions, not just the mean.

the sampling distribution of the proportion
the sampling distribution of the variance
the sampling distribution of the difference between means
the sampling distribution of Pearson's correlation
others....

Sampling Distribution of Sample Proportion

For a random sample of size n from a population with proportion p of outcomes in a particular category, the sampling distribution of the sample proportion in that category has

1. $\hat{p} = \frac{x}{n}$
2. $\mu_{\hat{p}} = p$
3. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

2 Central Limit Theorem

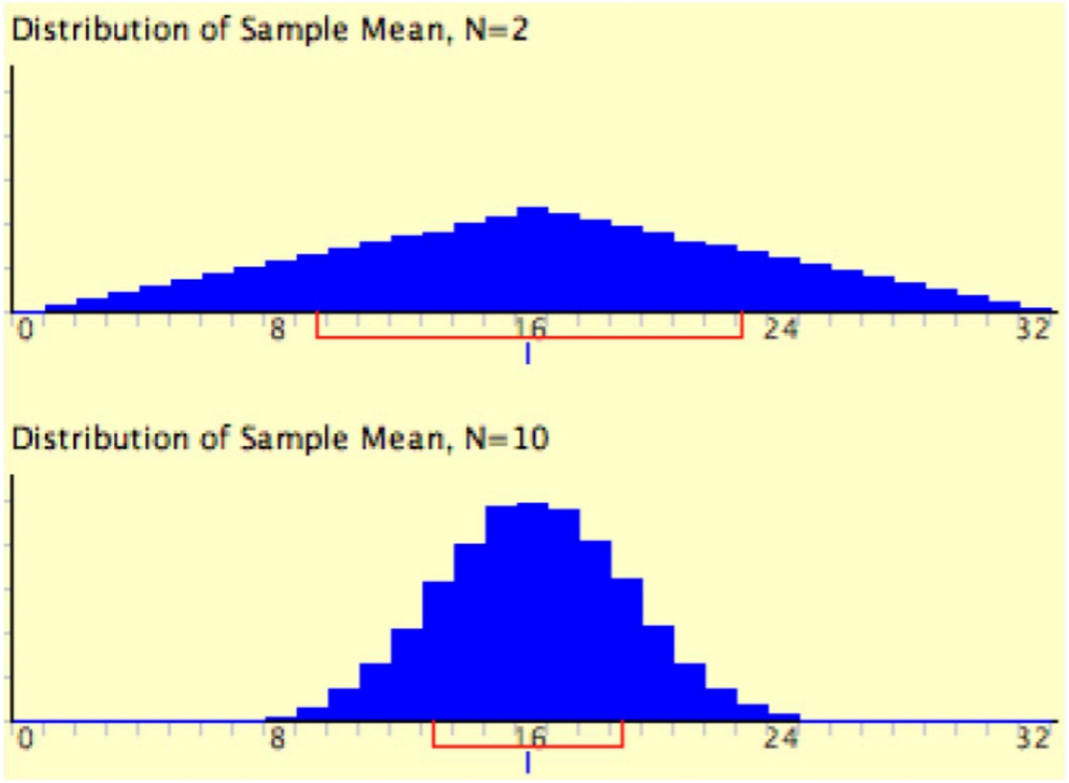
Central Limit Theorem

If we have a large enough sample size, the distribution of the sample mean is approximately normal, with a mean of the population mean and a standard deviation of the population standard deviation divided by the square root of n.

As the sample size increases, the sampling distribution of the sample mean has a more bell-shaped appearance.

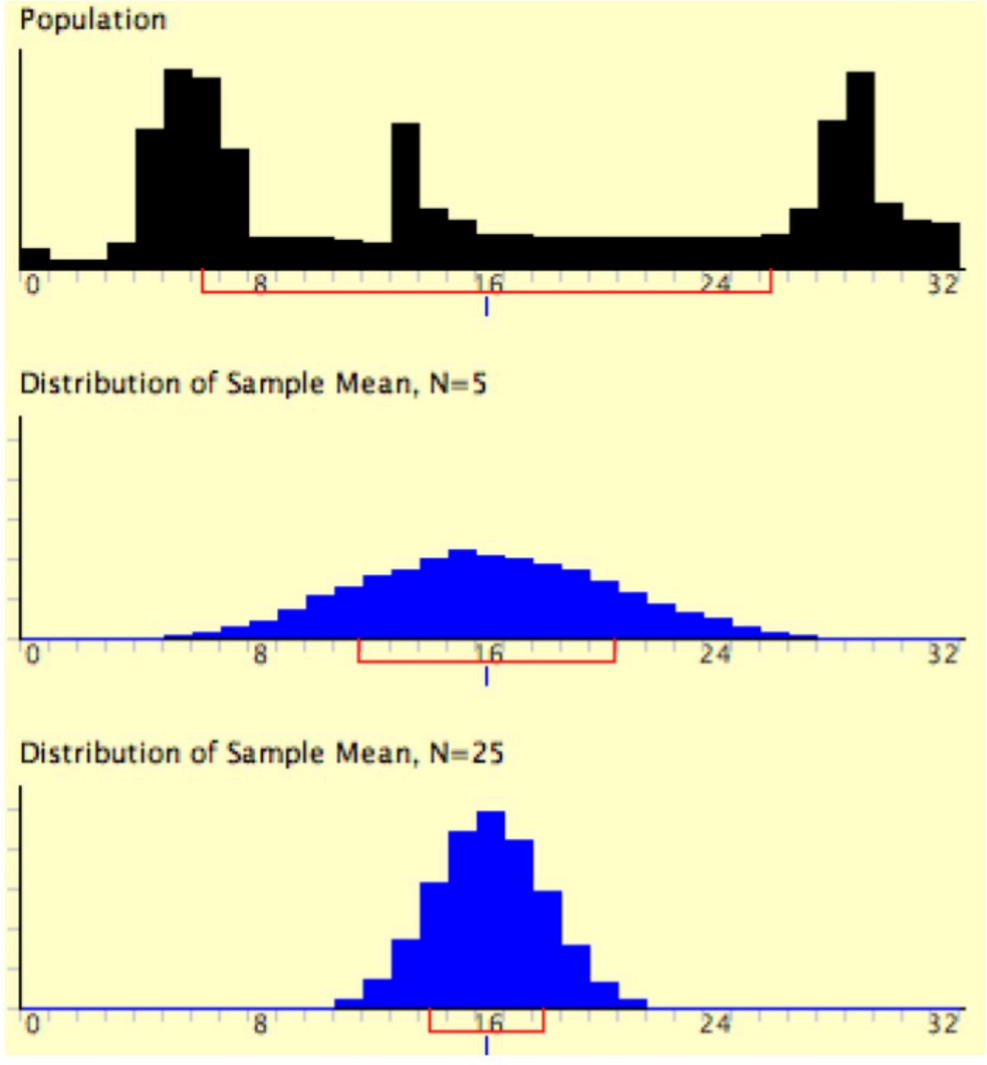
Central Limit Theorem

- Regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.

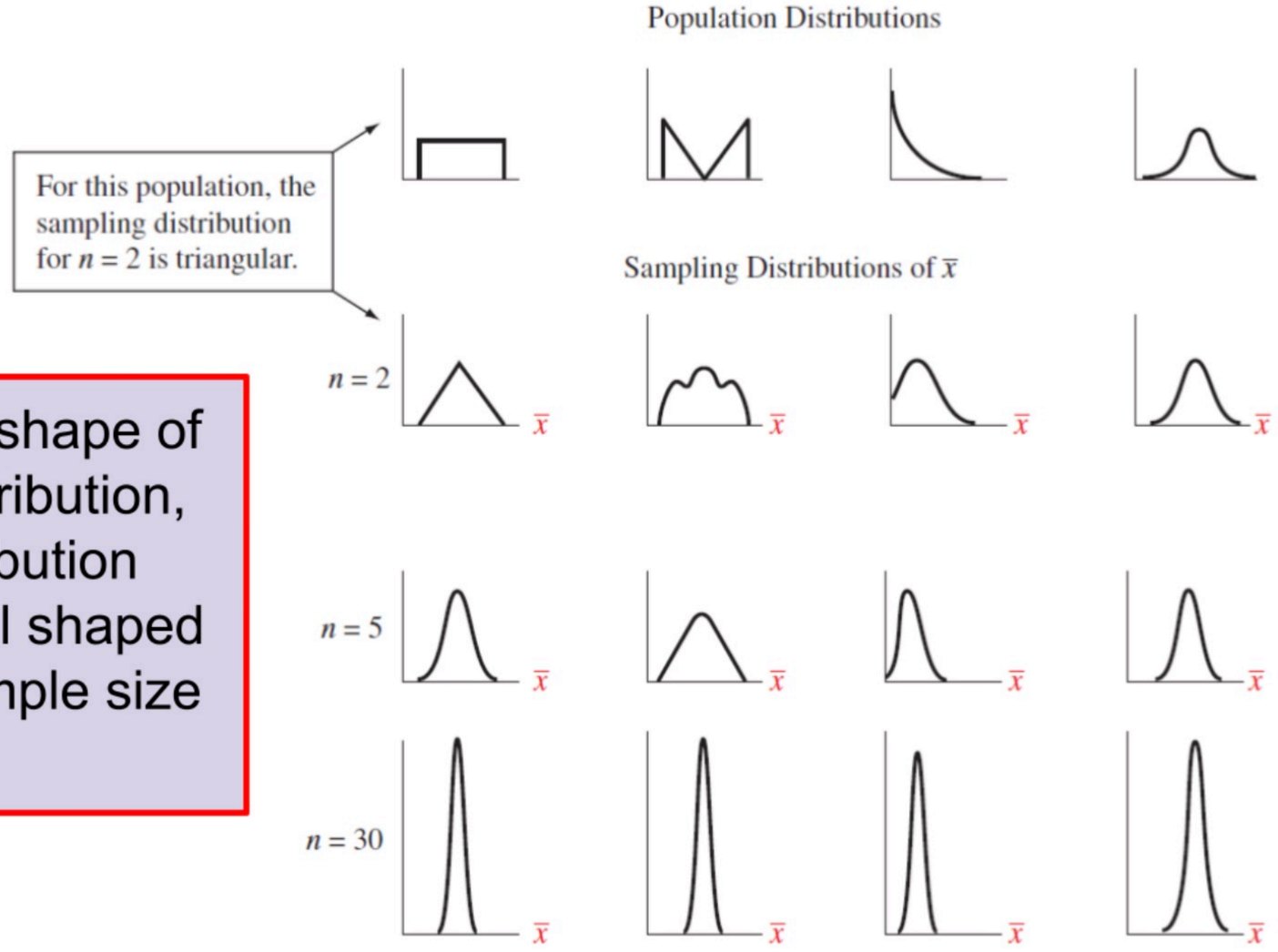


Central Limit Theorem

- Figure shows how closely the sampling distribution of the mean approximates a normal distribution even when the parent population is very non-normal.



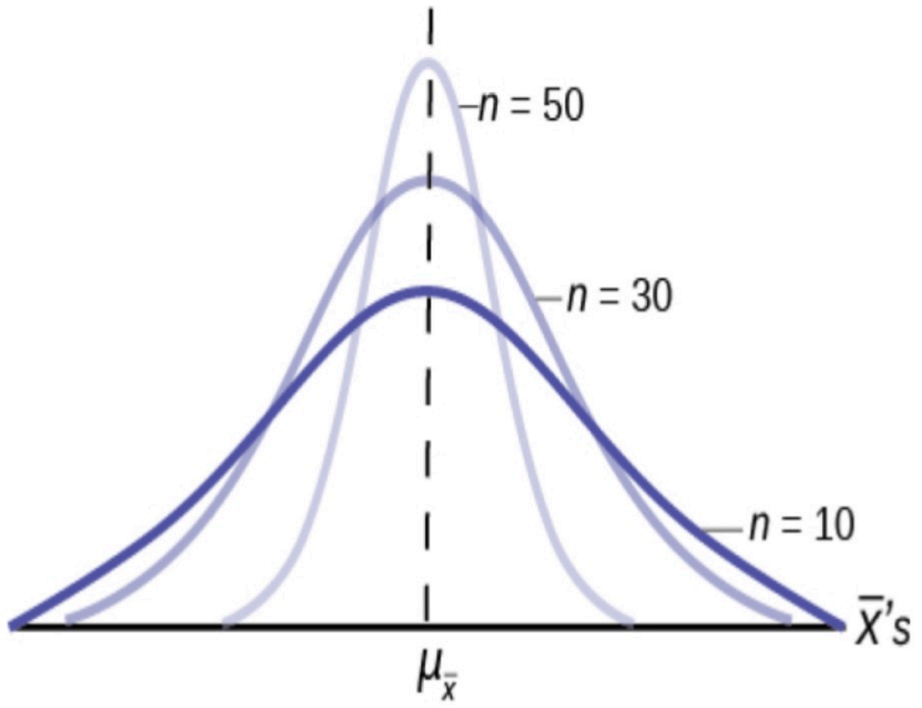
Central Limit Theorem



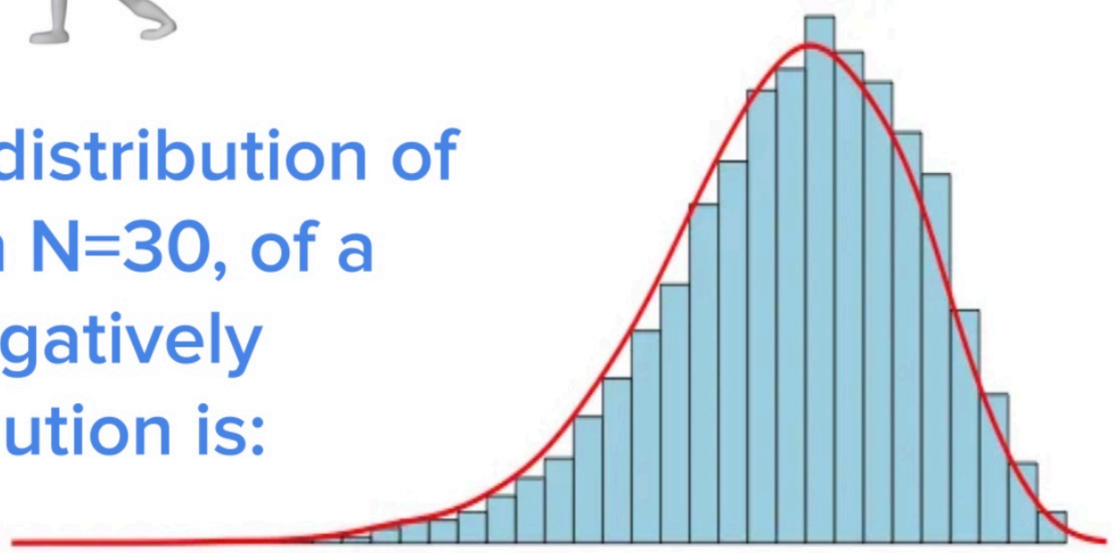
Regardless of the shape of the population distribution, the sampling distribution becomes more bell shaped as the random sample size n increases.

How Large is "Large Enough"?

- ★ In practice, some statisticians say that a sample size of 30 is large enough when the population distribution is roughly bell-shaped.
- ★ Others recommend a sample size of at least 40.



The sampling distribution of the mean, with N=30, of a moderately negatively skewed distribution is:



Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

London 4:33 am
New York 11:33 pm
Hong Kong 11:33 am
Options

The entire student body of 225 students took a test. These test scores have a mean of 75, a standard deviation of 10, and are slightly positively skewed.

If you randomly chose 25 of these test scores and calculated the mean over and over again, what could be the mean, standard deviation, and skew of this distribution?

Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

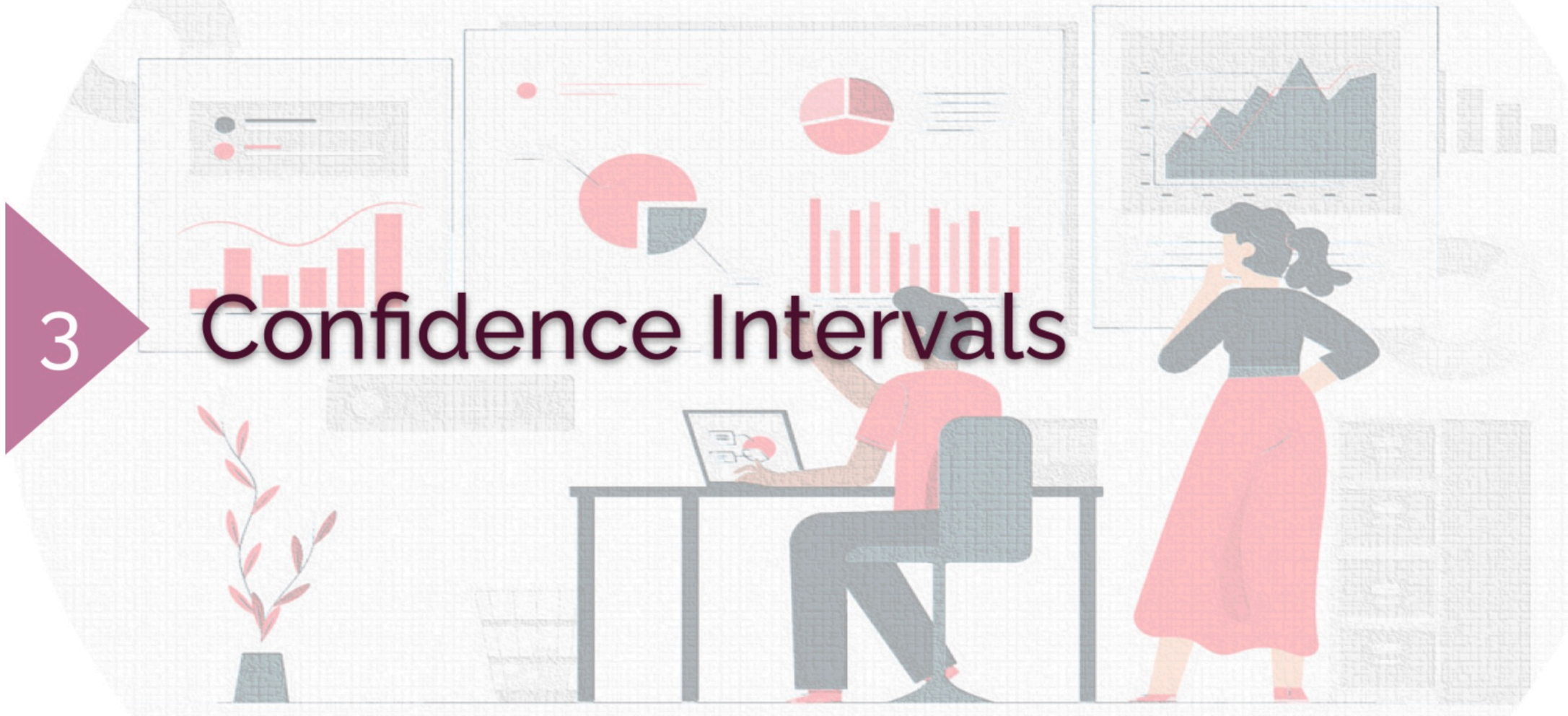
Example

The average male drinks 2 lt of water during an outdoor activity wit a std of 0.7 lt. You are planning a full day nature trip for random 50 men and you will bring 110 lt of water. What is the probability of that you will run out?

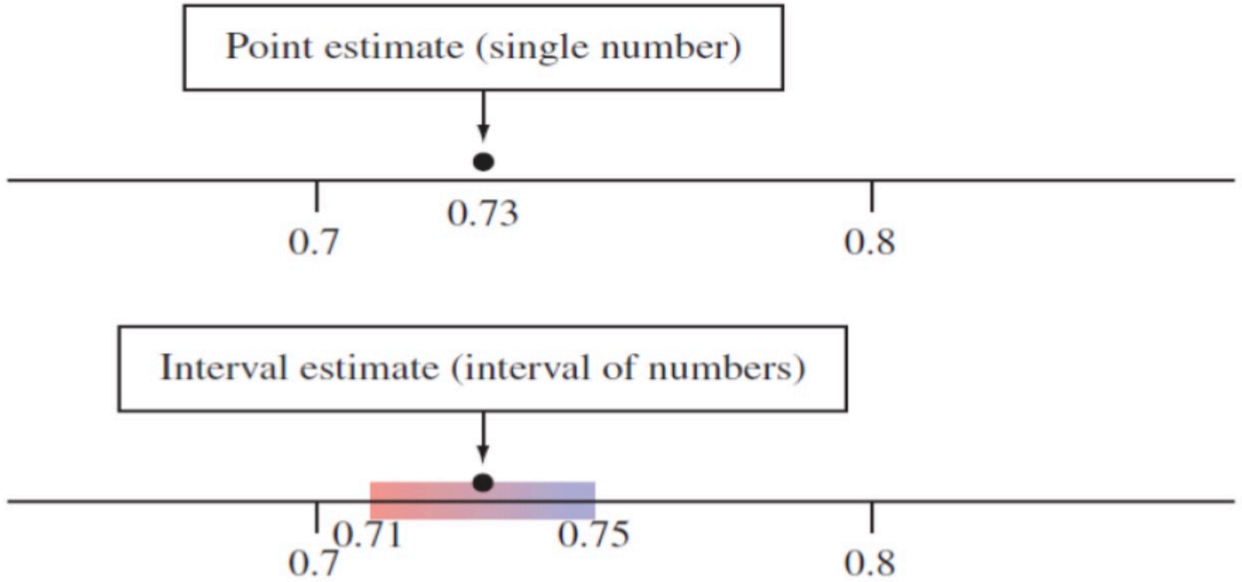
Answer

LINK to Answer 2

3 Confidence Intervals



Point and Interval Estimates



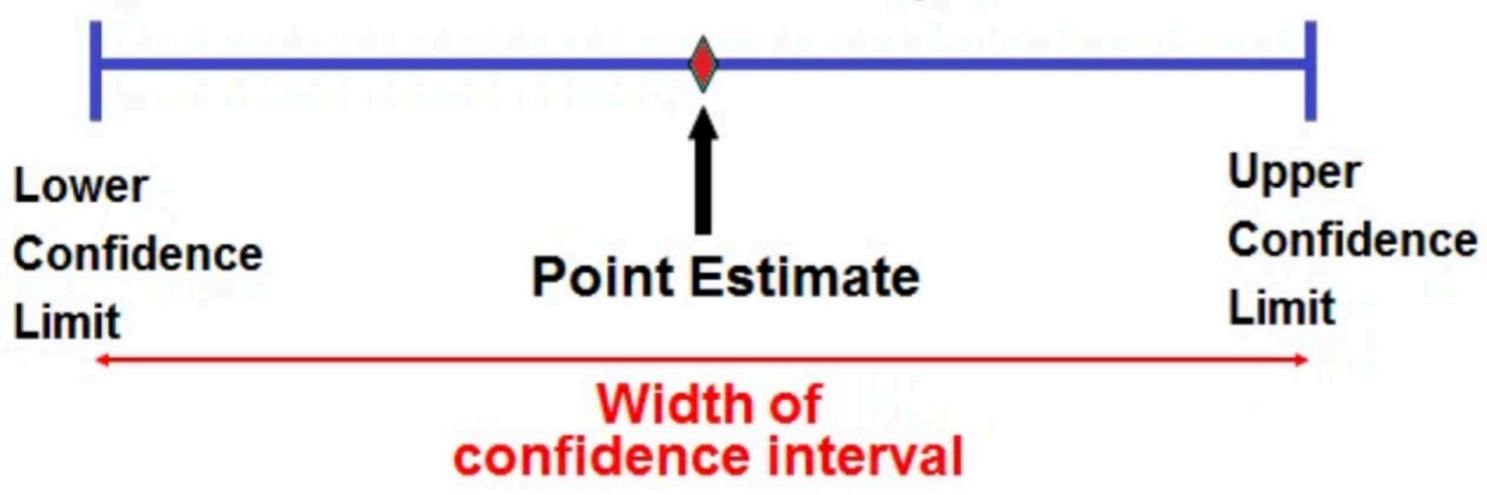
- ★ A **point estimate** is a *single number* that is our best guess for the parameter.
- ★ An **interval estimate** is an *interval of numbers* within which the parameter value is believed to fall.

Point and Interval Estimates

We can estimate a Population Parameter …		with a Sample Statistic (a Point Estimate)
Mean	μ	$\bar{x}$
Proportion	P	$\hat{p}$

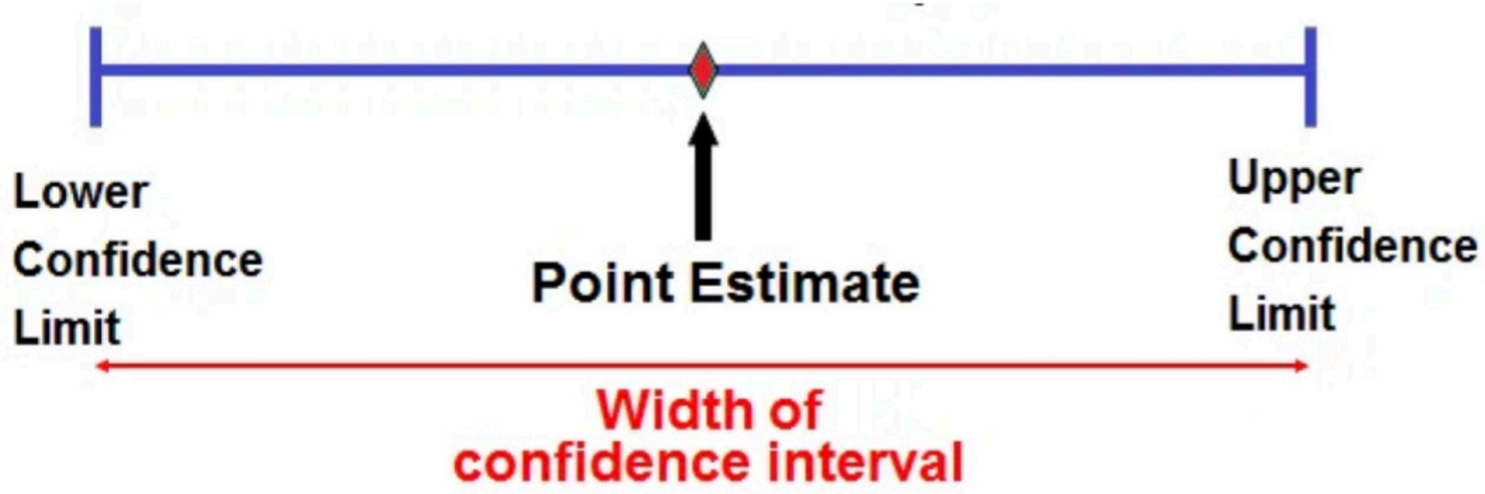
Interval Estimation

- ★ Instead of estimating a parameter with a single number, estimate it with an interval
- ★ Ideally, the interval will have two properties:
  - It will contain the target parameter with high probability
  - It will be relatively narrow



Interval Estimation

★ But, as we will see, interval endpoints are a function of the data,
- They will be variable
- So we cannot be sure the parameter will fall in the interval



Terminology

- ★ Interval estimators are commonly called confidence intervals (CIs)
- ★ Interval endpoints are called the upper and lower confidence limits
- ★ The probability the interval will enclose the parameter is called the confidence level
  - Notation: 1-α or 100(1-α)%
  - Usually referred to as "100(1-α)" percent CIs
- ★ Note α = probability the CI will miss!

Confidence Intervals

CONFIDENCE INTERVAL ESTIMATES

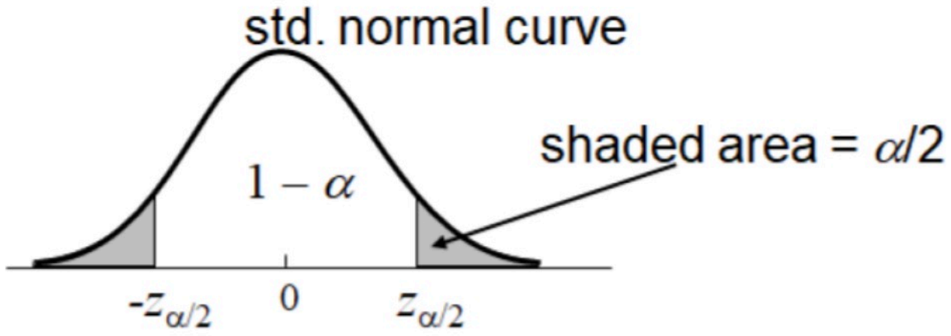
INTERVAL

$$\left[\text{Point estimate} - \text{reliability factor} * \text{standard error}\ ,\ \text{Point estimate} + \text{reliability factor} * \text{standard error}\right]$$

$$\left[\bar{x} - \text{reliability factor} * \frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + \text{reliability factor} * \frac{\sigma}{\sqrt{n}}\right]$$

Common Levels of Confidence

100(1 – α)% confidence interval = $\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



Level of Confidence (1-α)	α /2	$z_{\alpha/2}$
.90	.05	1.645
.95	.025	1.96
.99	.005	2.58

CI for on the Mean with σ known

- Sample from a normal distribution

$$2, 3, 5, 6, 9$$

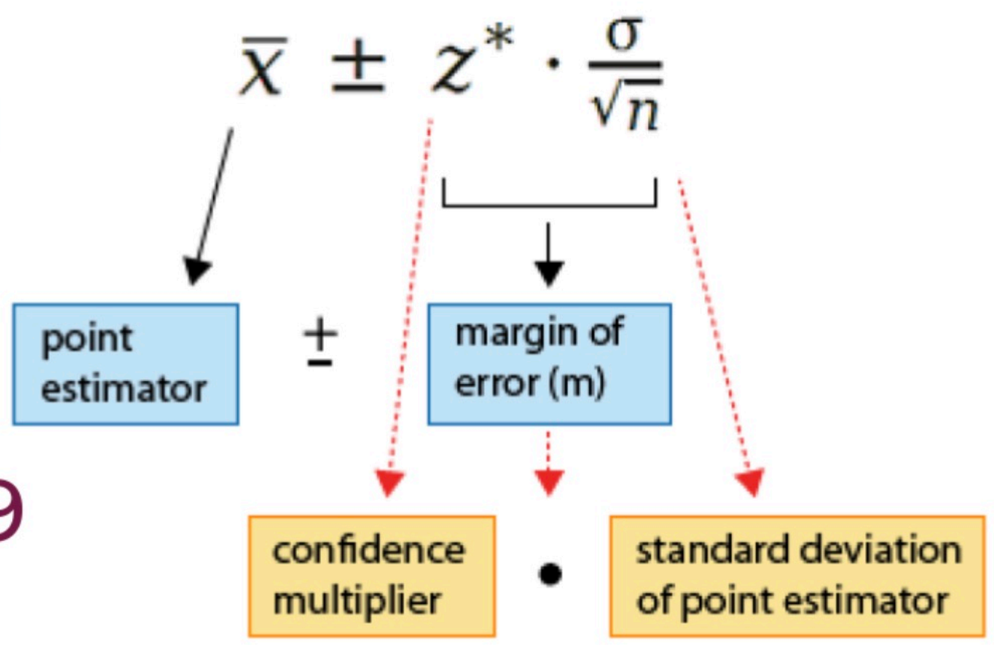
$$\bar{x} = \frac{2 + 3 + 5 + 6 + 9}{5} = 5$$

$$\sigma = 2.5$$

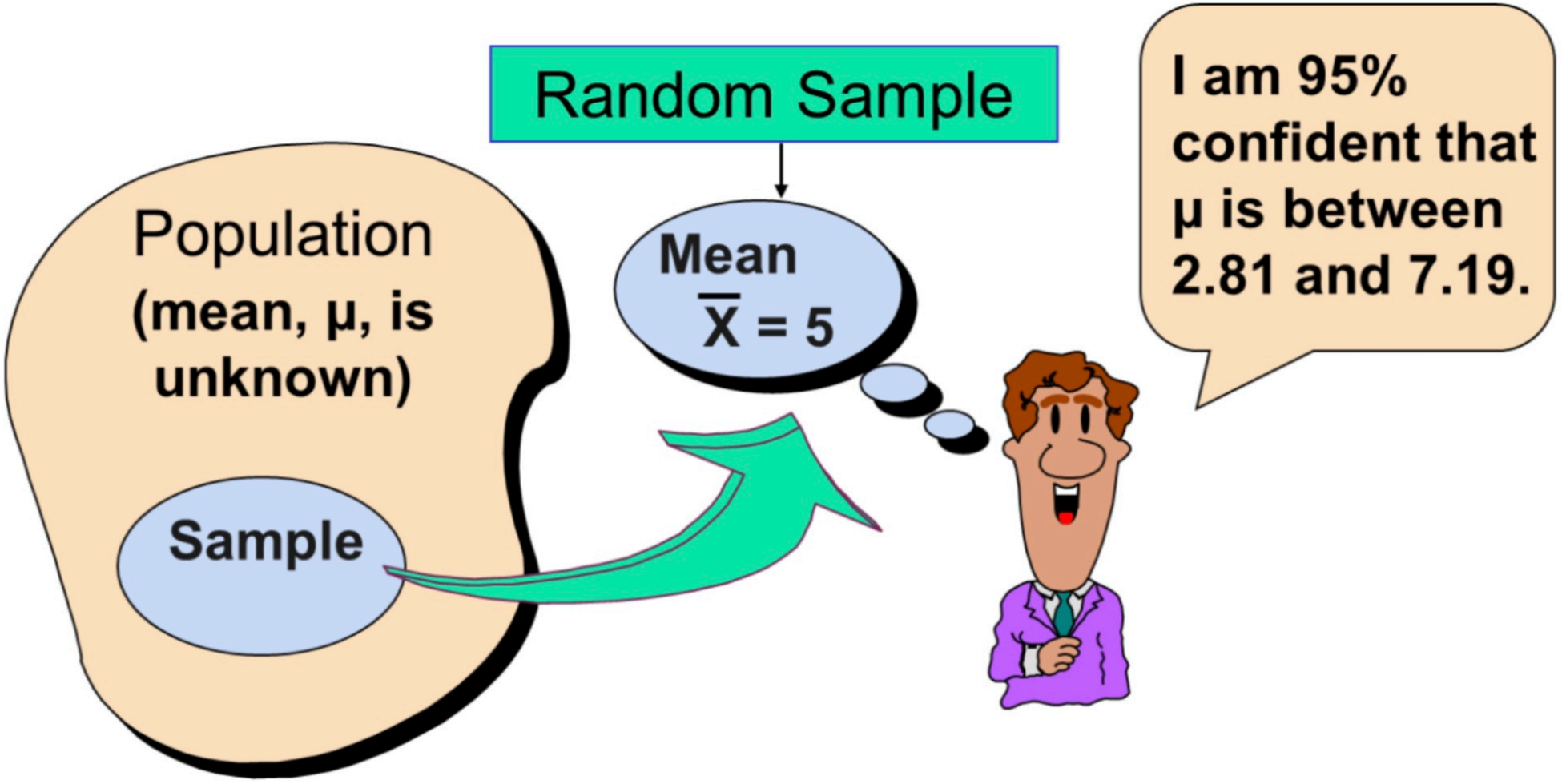
$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{5}} = 1.118$$

CI for on the Mean with σ known

- Lower limit
  - = 5 - (1.96)(1.118) = 2.81
- Upper limit
  - = 5 + (1.96)(1.118) = 7.19
- CI = [2.81, 7.19]



Estimation Process



Example

- A computer company samples demand over 25 time periods:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

- Its is known that the standard deviation of demand over time is 75 computers. We want to estimate the mean demand with 95% confidence in order to set inventory levels...

Example

- Sample size: n = 25 (given)
- Pop'n: $X_1, \cdots, X_{25} \overset{iid}{\sim} N(\mu = ?, \sigma = 75)$
- Statistic: $\bar{x} = 370.16$ (calculated from the sample)
- Statistic Critical Value: $z_{\frac{\alpha}{2}} = z_{\frac{.05}{2}} = z_{.025} = 1.96$
- Therefore, $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 370.16 \pm (1.96) \frac{75}{\sqrt{25}} = 370.16 \pm 29.40$

What is a Confidence Interval?

- ★ A confidence interval is a random interval
  - Random because it is a function of a random variable
- ★ Confidence level is the long-run percentage of intervals that will "cover" the population parameter
- ★ More .....
  - Confidence Interval for Variance
  - Confidence Interval for Proportion
  - Confidence Interval for the Difference Between Means

You take a sample (N = 25) of test scores from a population. The sample mean is 38, and the population standard deviation is 6.5.

What is the 95% confidence interval on the mean?

Students choose an option

No Multiple Choice Response
You didn't answer this question

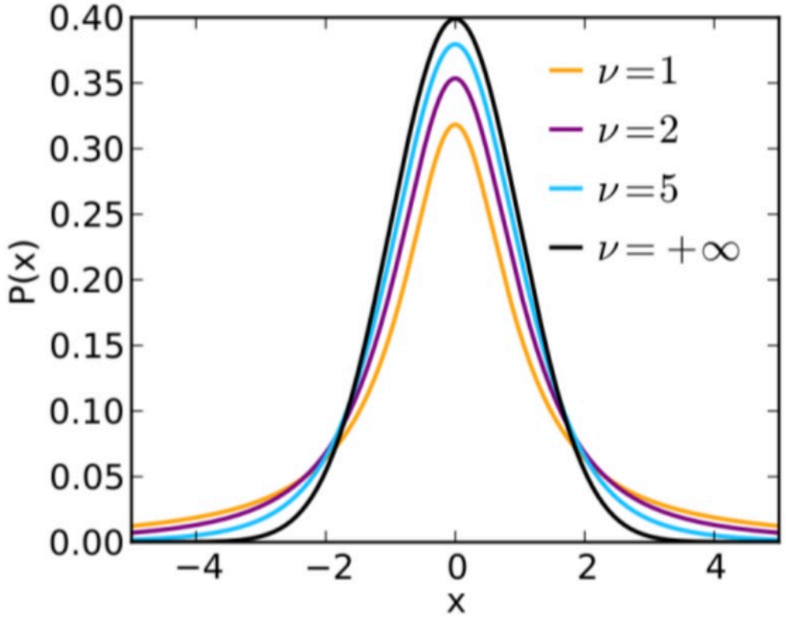
Need to know...

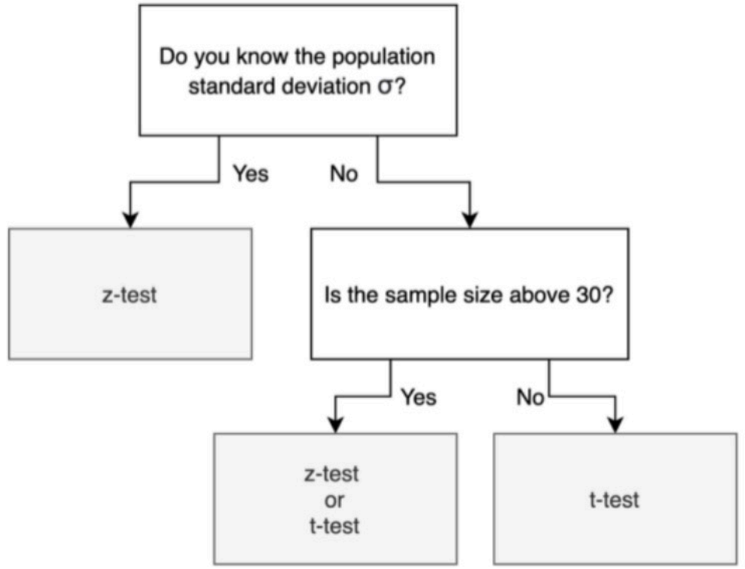
You should use the **t distribution** rather than the normal distribution when the variance is not known and has to be estimated from sample data.

t Distribution

t distribution (Student's t-distribution)

A probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.



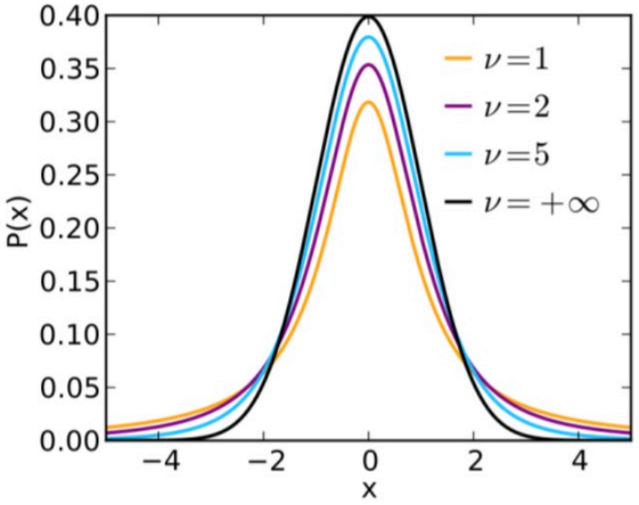


Degrees of Freedom

There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom. The degrees of freedom refers to the number of independent observations in a set of data.

df = sample size -1

The greater the degrees of freedom the more similar the t and normal distributions.



Students choose an option

No Multiple Choice Response

You didn't answer this question

t statistic

Sample sizes are sometimes small, and often we do not know the standard deviation of the population.

When either of these problems occur, statisticians rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

t Distribution *Example*

- Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days.
- If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

t Distribution *Example*

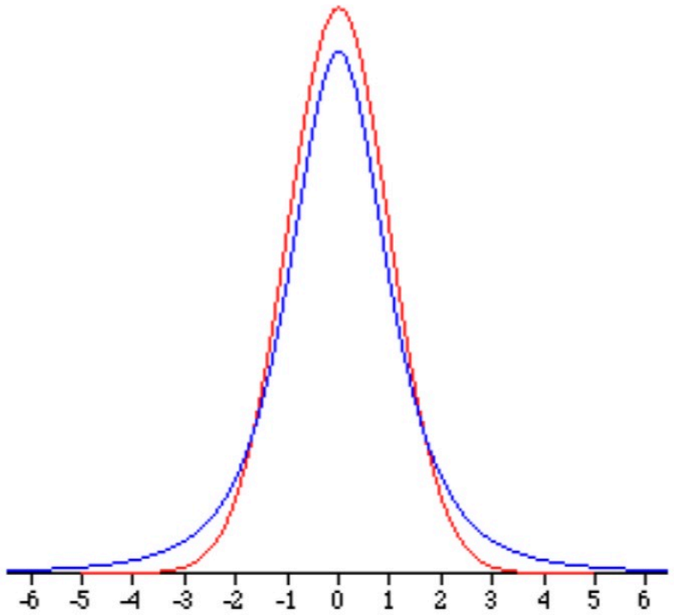
$$t = \frac{290-300}{\frac{50}{\sqrt{15}}}$$

```
stats.t.cdf(-0.774, 14)
```

```
0.2259202308781893
```

$$= \frac{-10}{12.909945} = -0.7745966$$

- ★ The degrees of freedom are equal to 15 – 1 = 14.
- ★ The population mean equals 300.
- ★ The sample mean equals 290.
- ★ The standard deviation of the sample is 50.
- ★ The cumulative probability: 0.226.



Which distribution is the t distribution?

Students choose an option

No Multiple Choice Response

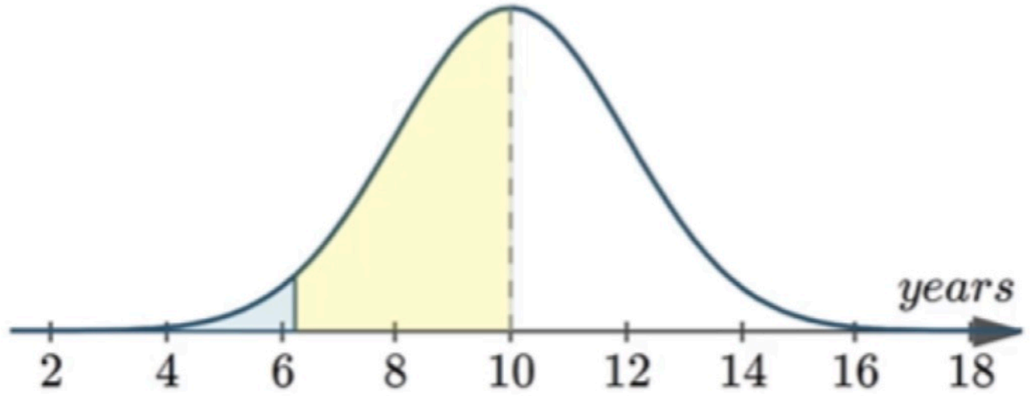
You didn't answer this question

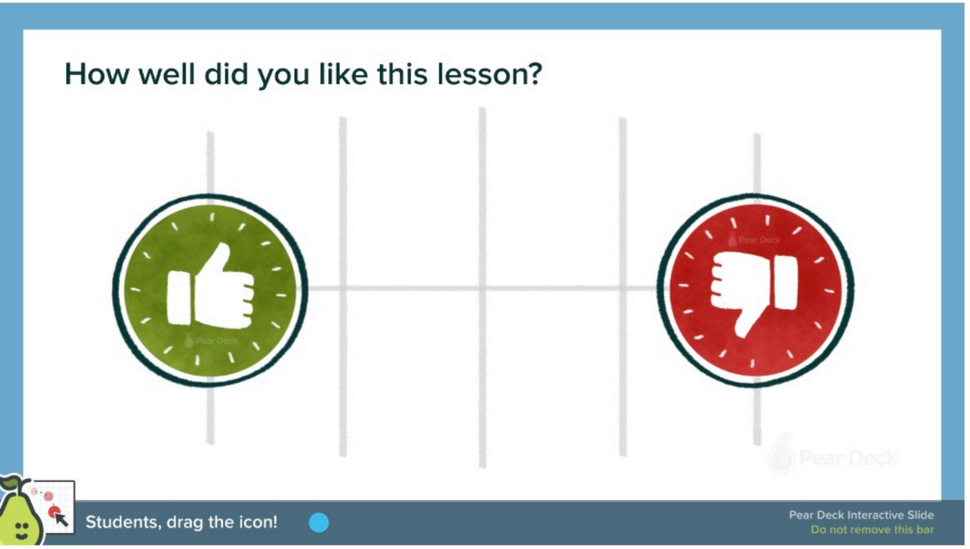
Interview Question (z Score)

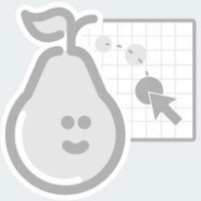
- The average life of a certain type of motor is 10 years, with a standard deviation of 2 years.
- If the manufacturer is willing to replace only 3% of the motors because of failures, how long a guarantee should she offer?
- Assume that the lives of the motors follow a normal distribution.

Answer

Solving this leads to x=6.24 , so the guarantee period should be 6.24 years.







No Draggable™ Response

You didn't answer this question

THANKS!

Any questions?

You can find me at:

- jason@clarusway.com