STA250 Probability and Statistics

Chapter 9 Notes

Estimation and Confidence Intervals

Asst. Prof. Abdullah YALÇINKAYA

Ankara University, Faculty of Science, Department of Statistics

ayalcinkaya@ankara.edu.tr

2023

STA250 Probability and Statistics

Reference Book

This lecture notes are prepared according to the contents of

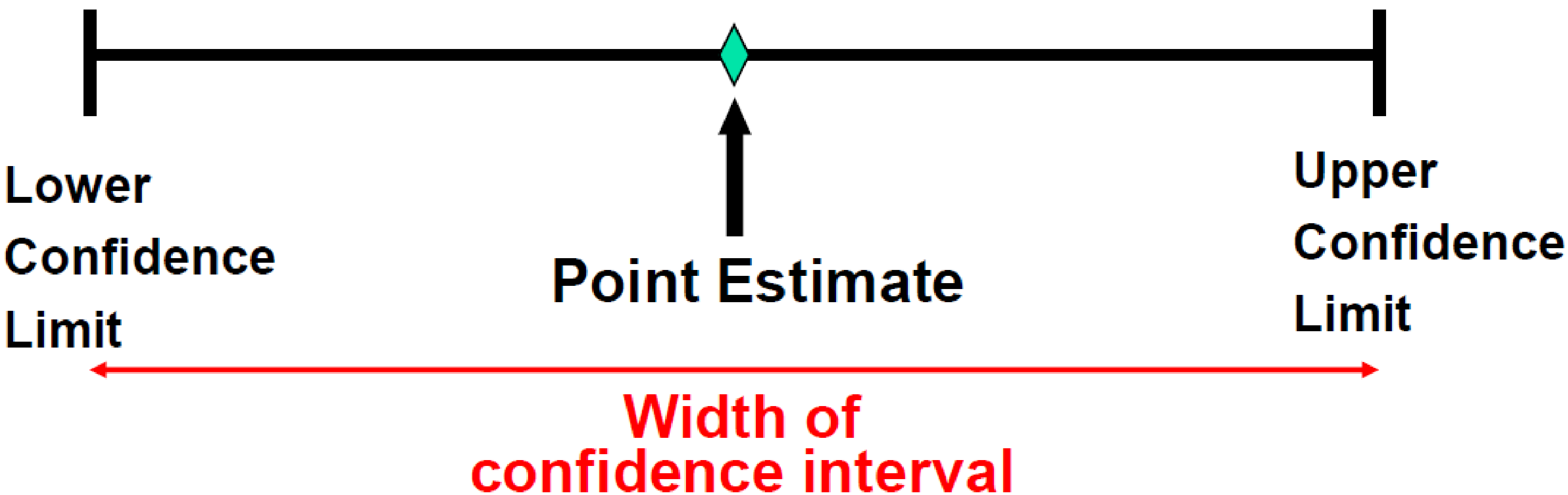
“PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS by Walpole, Myers, Myers and Ye”

Definitions

- **An estimator of a population parameter is**
  - a random variable that depends on sample information . . .
  - whose value provides an approximation to this unknown parameter
- **A specific value of that random variable is called an estimate**

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



What are the characteristics of a good estimator?

1. Unbiased Estimator

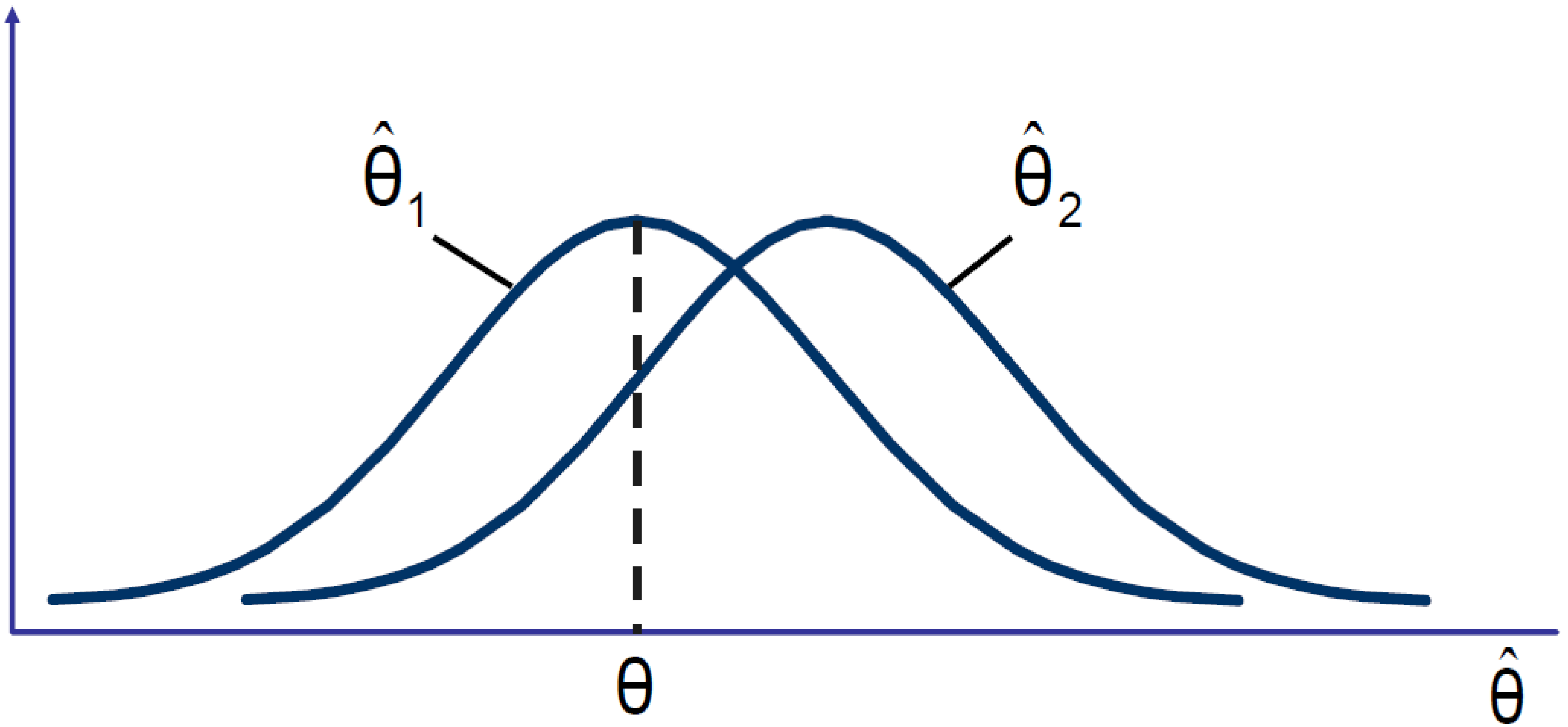
- A point estimator is said to be an unbiased estimator of the parameter θ if the expected value, or mean, of the sampling distribution of $\hat{\theta}$ is θ,

$$\boxed{E(\hat{\theta}) = \theta}$$

- Examples:
  - The sample mean $\bar{x}$ is an unbiased estimator of μ
  - The sample variance s^2 is an unbiased estimator of σ^2
  - The sample proportion $\hat{p}$ is an unbiased estimator of P

Unbiasedness

- $\hat{\theta}_1$ is an unbiased estimator, $\hat{\theta}_2$ is biased:



What are the characteristics of a good estimator?

2. Most efficient estimator

- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same population parameter θ, we want to choose the estimator whose sampling distribution has the smaller variance.

  - $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if $\boxed{\mathbf{Var}(\hat{\boldsymbol{\theta}}_1) < \mathbf{Var}(\hat{\boldsymbol{\theta}}_2)}$

  - The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is to the ratio of their variance:

$$\boxed{\boldsymbol{Relative\ efficiency} = \frac{\mathbf{Var}(\hat{\boldsymbol{\theta}}_2)}{\mathbf{Var}(\hat{\boldsymbol{\theta}}_1)}}$$

Confidence Intervals

- An interval estimate of a population parameter θ is an interval the form $\hat{\theta}_L < \theta < \hat{\theta}_U$.

- Such an interval is called an **interval estimate.**

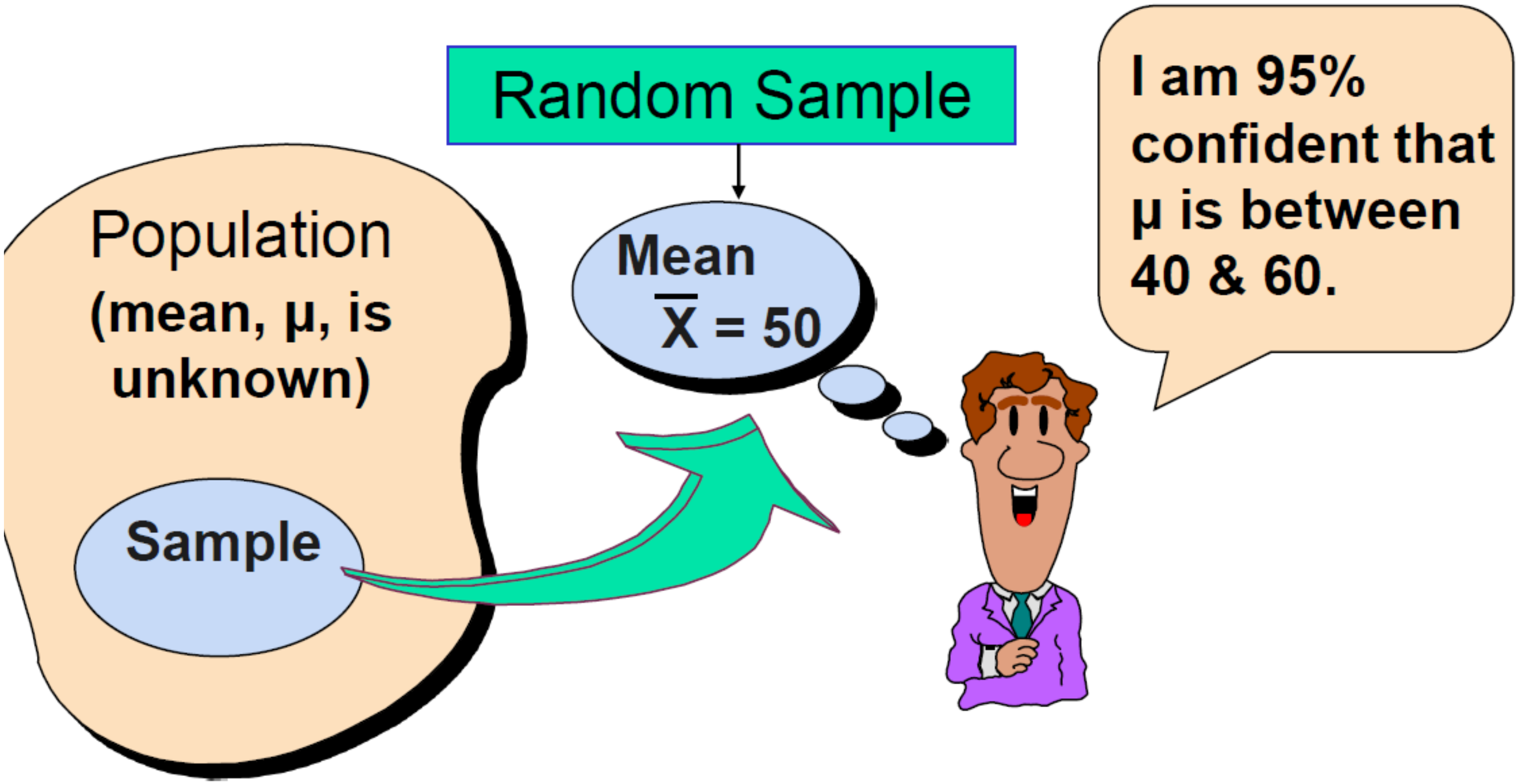
- **An interval estimate** provides more information about a population characteristic than does a point estimate

Confidence Interval Estimate

- **An integral gives a <u>range</u> of values:**

  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observation from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100% confident.

Estimation Process



Confidence Level, (1-α)

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter

Confidence Interval

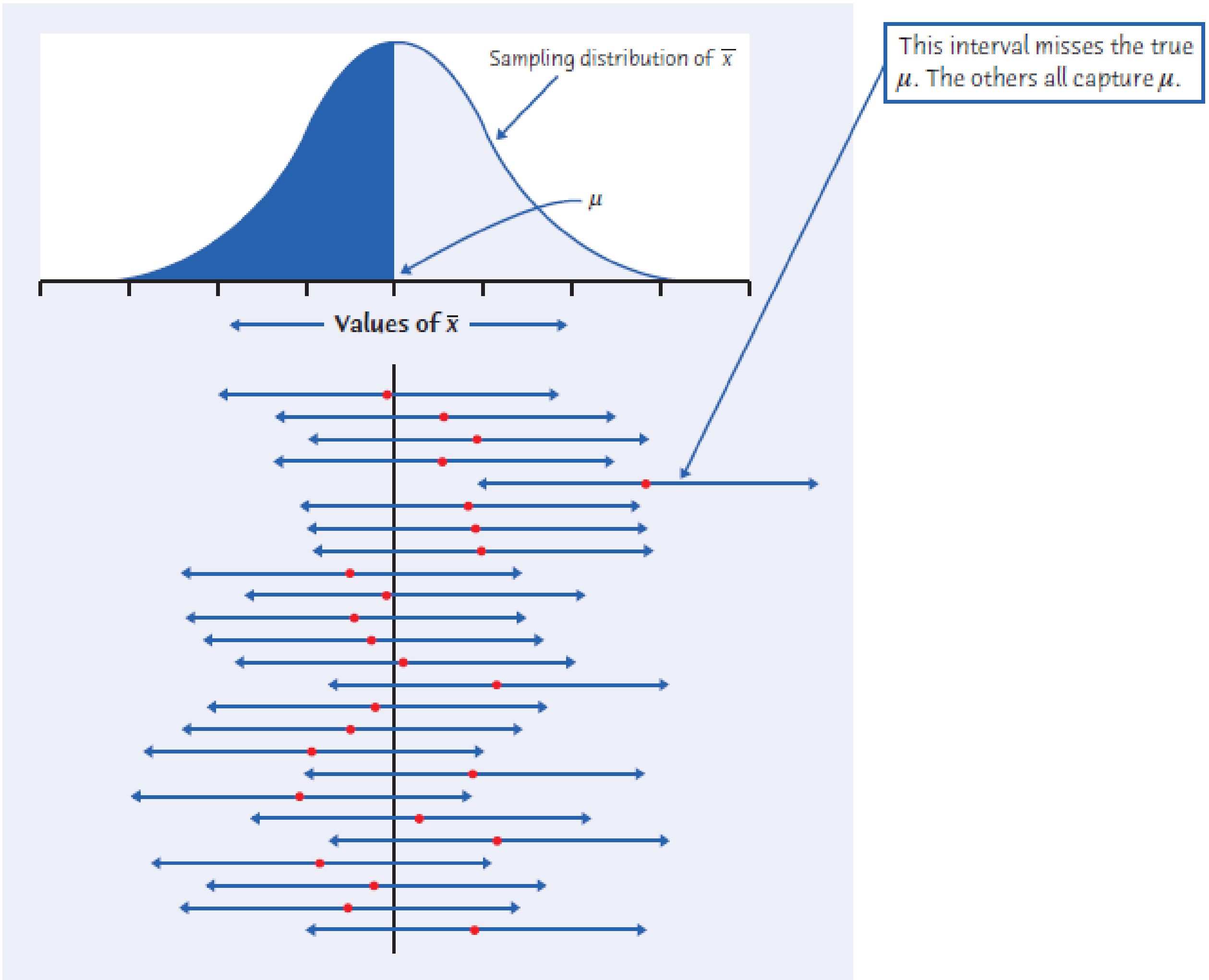
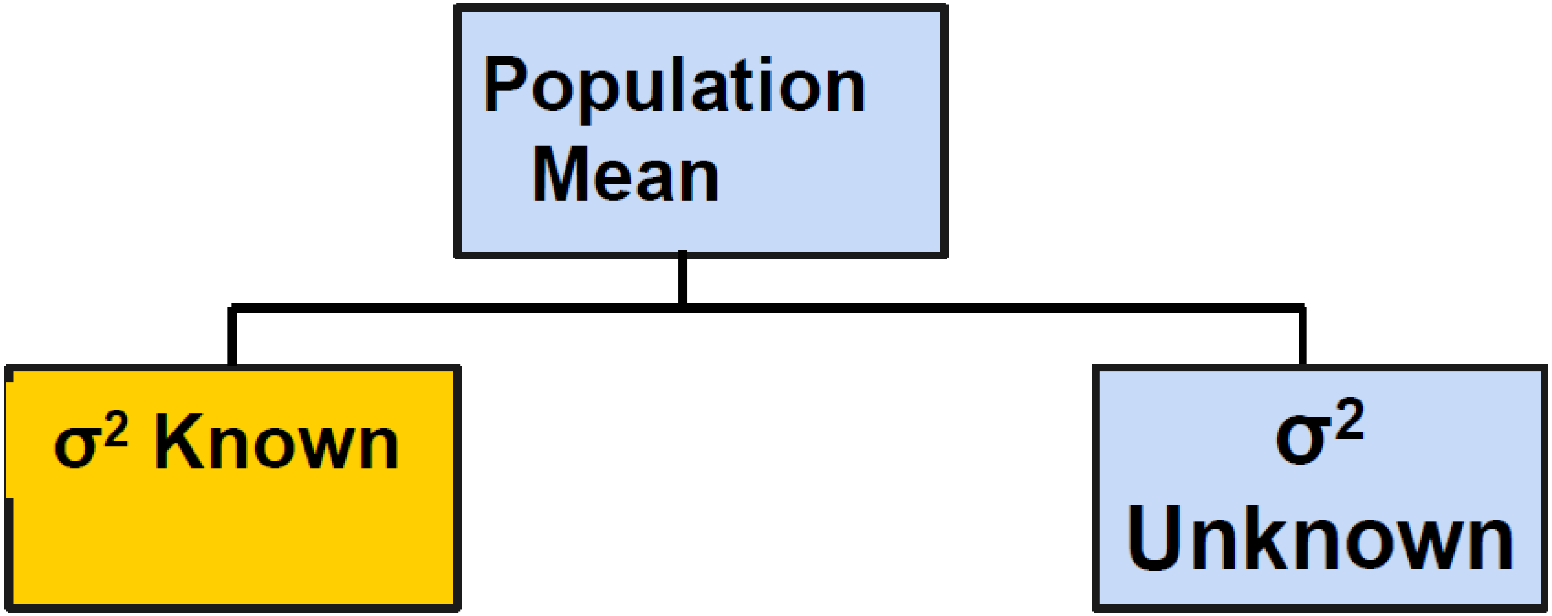


FIGURE 14.2

Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that contains the population mean μ.

Source: ***Mathematical Statistics with Applications,*** **Seventh Edition**
Dennis D. Wackerly, William Mendenhall III, Richard L. Scheaffer

Confidence Intervals



Confidence Intervals for μ (σ^2 Known)

- **When we use $\bar{x}$ to estimate μ, we don't expect the estimate to be exact. A confidence interval is a statement that we are 100(1-α)% confident that μ lies between two specified limits.**
- **If $\bar{x}$ is the mean of a random sample of size n from a <u>population with known variance σ^2</u>, then**

$$\boxed{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

is a 100(1-α)% confidence interval for μ.

- **Notes:**
  - Here $z_{\alpha/2}$ is the z value with area α/2 to the right.
  - For example, for a 95% confidence interval, α = .05, and $z_{.025} = 1.96$.
  - Population variance is σ^2 known
  - Population is normally distributed.
  - If population not normal, this is still okay if $n \geq 30$

Marjin of Error

- The confidence interval,

$$\boxed{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

- Can also be written as $\boxed{\bar{x} \pm ME}$
  where ME is called the margin of error

$$\boxed{ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

- The interval width, w, is equal to twice the margin of error

Reducing the Marjin of Error

$$\boxed{ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced ($\sigma\downarrow$)
- The sample size is increased ($n\uparrow$)
- The confidence level is decreased, $(1 - \alpha)\downarrow$

Comman Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Confidence Coefficient,* $1-\alpha$	$Z_{\alpha/2}$ *value*
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Example 9.2

The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

Solution 9.2

The point estimate of μ is $\bar{x} = 2.6$. The z-value leaving an area of 0.025 to the right, and therefore an area of 0.975 to the left, is $z_{0.025} = 1.96$ (Table A.3). Hence, the 95% confidence interval is

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

which reduces to $2.50 < \mu < 2.70$. To find a 99% confidence interval, we find the z-value leaving an area of 0.005 to the right and 0.995 to the left. From Table A.3 again, $z_{0.005} = 2.575$, and the 99% confidence interval is

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right),$$

or simply

$$2.47 < \mu < 2.73.$$

Example 9.3.

How large a sample is required if we want to be 95% confident that our estimate of μ in Example 9.2 is off by less than 0.05?

Solution: The population standard deviation is $\sigma = 0.3$. Then, by Theorem 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

Therefore, we can be 95% confident that a random sample of size 139 will provide an estimate $\bar{x}$ differing from μ by an amount less than 0.05.

Note:

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Example 9.4

A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

Determine a 95% confidence interval for the true mean resistance of the population.

Solution 9.4

A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is .35 ohms.

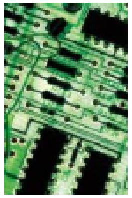
Solution:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\ (.35/\sqrt{11})$$

$$= 2.20 \pm .2068$$

$$\boxed{1.9932 < \mu < 2.4068}$$

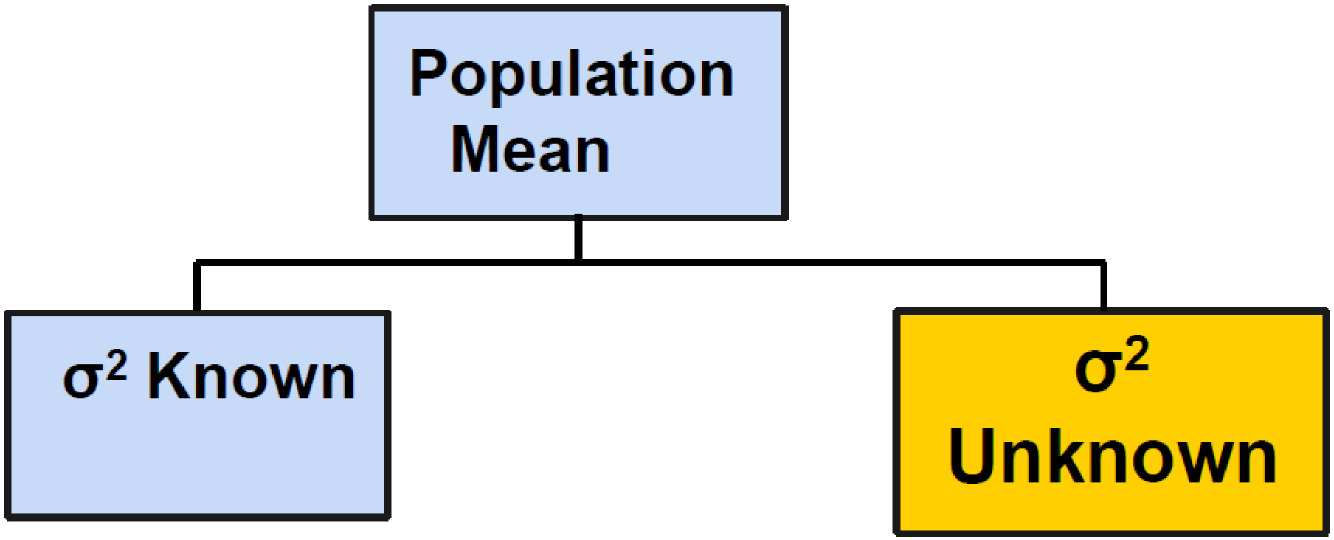


Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms

- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Confidence intervals



Student's t Distribution

- Consider a random sample of n observations
  - with mean $\bar{x}$ and standard deviation s
  - from a normally distributed population with mean μ

- Then the variable

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

follows the Student's t distribution with (n-1) degrees of freedom

Confidence Interval for μ (σ unknown)

- **Assumptions**
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\boxed{\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}}$$

where $t_{n-1,\,\alpha/2}$ is the critical value of the t distribution with $n-1$ d.f.

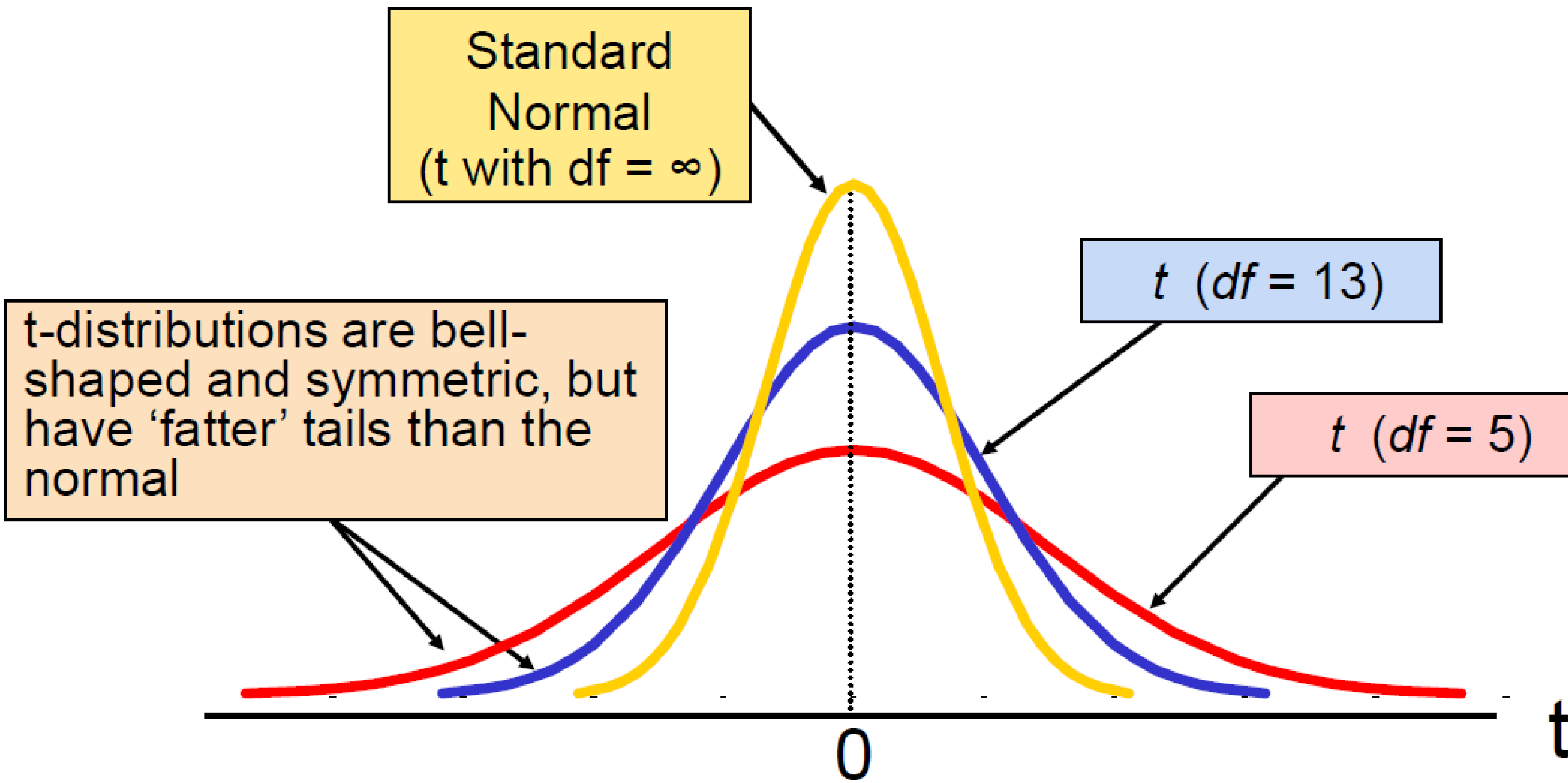
Student's t Distribution

- The t is a family of distribution
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

$$\boxed{\text{d.f.} = n - 1}$$

Student's t Distribution

Note: $t \longrightarrow Z$ as n increases

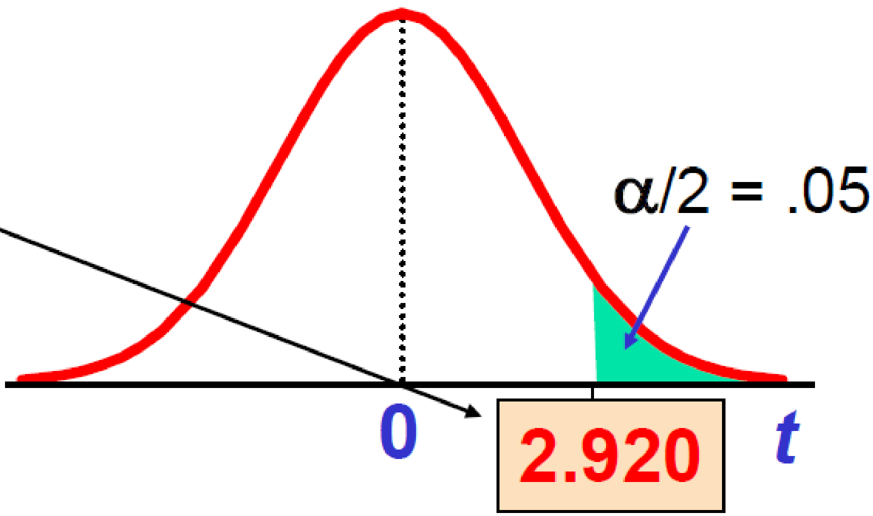


Student’s t Distribution

df	Upper Tail Area .10	.05	.025
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

The body of the table contains t values, not probabilities

Let: n = 3
df = n - 1 = 2
α = .10
$\alpha/2$ =.05



t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.282
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.576

Note: $t \longrightarrow Z$ as n increases

Example 1

The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

***Solution*:** The sample mean and standard deviation for the given data are

$$\bar{x} = 10.0 \qquad \text{and} \qquad s = 0.283.$$

Using Table A.4, we find $t_{0.025} = 2.447$ for $v = 6$ degrees of freedom. Hence, the

95% confidence interval for μ is

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

which reduces to $9.74 < \mu < 10.26$.

$$\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

Example 2

A do-it-yourself fencing company sells materials to and provides classes for homeowners wanting to erect privacy fences for their residences. Eight randomly selected invoices show sales for materials for fences of the following sizes (in meters)

50, 150, 225, 75, 200, 215, 195, 80

Find 95% Confidence Interval for the mean number of meters of material sold per customer. Assume the distribution to be normal.

Since X~ Normal → $\bar{X}$ ~ Normal but population standard deviation σ is unknown→ CI is [$\bar{X}$ $\pm$ $t_{n-1,\alpha/2}$ $(s/\sqrt{n})$]

Solution 2

$\bar{X}=\Sigma X_i/n$ = 1190/8= 148.75 meters is the expected sale for any random customer

$S=\sqrt{\dfrac{\mathbf{211900-8(148.75^2)}}{\mathbf{8-1}}}$ = 70.6 meters is the expected deviation from center

For $1\text{-}\alpha=0.95$ $\quad t_{n-1,\alpha/2} = 2.365$

→ CI is $[148.75 \pm (2.365)(70.6/\sqrt{8})]$

$[148.75 \pm 58.91$ meters] →[89.84 meters, 207.66 meters] is 95% CI estimate for population mean when σ unknown.

<u>Interpretation</u>: We are 95% confident that any random customer will be purchasing between 89.84 to 207.66 meters of fencing materials on the average.

Confidence Interval for Population Proportion, P

- An interval estimate for the population proportion (P) can be calculated by adding an allowance for uncertainty to the sample proportion ($\hat{p}$)

Confidence Interval for Population Proportion, P

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\boxed{\sigma_P = \sqrt{\frac{P(1-P)}{n}}}$$

- We will estimate this with sample data:

$$\boxed{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Confidence Interval for Population Proportion, P

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- where
- $z_{\alpha/2}$ is the standard normal value for the level of confidence desired
- $\hat{p}$ is the sample proportion
- n is the sample size

Example

A manufacturer of gunpowder has developed a new powder, which was tested in eight shells. The resulting muzzle velocities, in feet per second, were as follows:

3005	2925	2935	2965
2995	3005	2937	2905

Find a 95% confidence interval for the true average velocity μ for shells of this type. Assume that muzzle velocities are approximately normally distributed.

Solution If we assume that the velocities Y_i are normally distributed, the confidence interval for μ is

$$\overline{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right),$$

where $t_{\alpha/2}$ is determined for $n-1$ df. For the given data, $\overline{y} = 2959$ and $s = 39.1$. In this example, we have $n - 1 = 7$ df and, using Table 5, Appendix 3, $t_{\alpha/2} = t_{.025} = 2.365$. Thus, we obtain

$$2959 \pm 2.365\left(\frac{39.1}{\sqrt{8}}\right), \quad \text{or} \quad 2959 \pm 32.7,$$

as the observed confidence interval for μ. ∎

Example

Example 9.14: In a random sample of $n = 500$ families owning television sets in the city of Hamilton, Canada, it is found that $x = 340$ subscribe to HBO. Find a 95% confidence interval for the actual proportion of families with television sets in this city that subscribe to HBO.

Solution: The point estimate of p is $\hat{p} = 340/500 = 0.68$. Using Table A.3, we find that $z_{0.025} = 1.96$. Therefore, using method 1, the 95% confidence interval for p is

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}},$$

which simplifies to $0.6391 < p < 0.7209$.

Example

Example 9.15: How large a sample is required if we want to be 95% confident that our estimate of p in Example 9.14 is within 0.02 of the true value?

Solution: Let us treat the 500 families as a preliminary sample, providing an estimate $\hat{p} = 0.68$. Then, by Theorem 9.4,

$$n = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.8 \approx 2090.$$

Therefore, if we base our estimate of p on a random sample of size 2090, we can be 95% confident that our sample proportion will not differ from the true proportion by more than 0.02.

Next Lesson

- **Hypothesis Testing**

See you☺