W14 QMS 202 Lecture 1

Dr Boža Tasić

January 14, 2015

Confidence Interval Estimation - Learning Objectives

- In this Chapter we will learn
  - To construct and interpret confidence interval estimates for the population mean μ when the population standard deviation σ is known
  - To construct and interpret confidence interval estimates for the population mean μ when the population standard deviation σ is not known

The Standard Normal Distribution

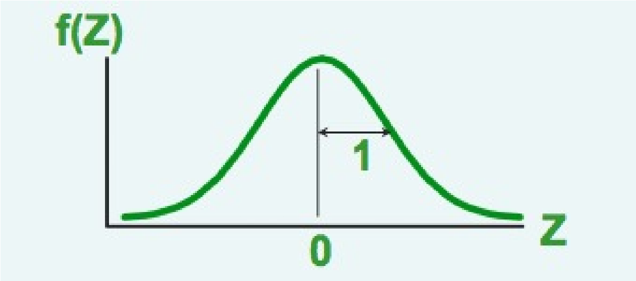
- Given a normal random variable X with the mean μ and the standard deviation σ we define the **Standardized Normal Random Variable** Z by

$$Z = \frac{X - \mu}{\sigma}$$

- The variable Z will have:
- The mean $\mu_{\mathrm{z}} = 0$ and
- The standard deviation $\sigma_{\mathrm{z}} = 1$.

The Standard Normal Distribution

- The standardized normal distribution Z will always have the mean 0 and the standard deviation 1.



- The values above the mean have **positive** Z-values. The values below the mean have **negative** Z-values.

Using CASIO Calculator to compute $P(Z \leq ?)$ or $P(Z < ?)$

- We will use $P(Z \leq ?) = 0.025$ as a model.

1. Select **STAT F5** (DIST) and then **F1**(NORM). Now select **F3**(InvN) and the following options:
2. Data: **F2** (Variable)
3. Tail: **F1** (Left)
4. Area: **0.025 EXE**
5. σ: **1 EXE**
6. μ: **0 EXE**
7. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Execute: key **EXE** or **F1**(Calc).

- You will get $xInv = -1.95996398$. This means $P(Z \leq -1.96) = 0.025$.

Using CASIO Calculator to compute $P(Z \geq ?)$ or $P(Z > ?)$

- We will use $P(Z \geq ?) = 0.025$ as a model.

1. Select **STAT F5** (DIST) and then **F1**(NORM). Now select **F3**(InvN) and the following options:
2. Data: **F2** (Variable)
3. Tail: **F1** (Right)
4. Area: **0.025 EXE**
5. σ: **1 EXE**
6. μ: **0 EXE**
7. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Execute: key **EXE** or **F1**(Calc).

- You will get *xInv* $= 1.95996398$. This means $P(Z \geq 1.96) = 0.025$.

Sampling Distributions

- Our main concern when making a statistical inference is drawing conclusions about a population, not about a sample.
- To estimate the values of population parameters we often use statistics calculated from samples.
- For example, we will use the sample mean (**a statistic**) to estimate the population mean (**a parameter**)
- A **sampling distribution** is a distribution of all of the possible values of a sample statistic for a given size sample selected from a population.

Sampling Distribution of the Mean

- The mean is the most widely used measure of central tendency. The sample mean is often used to estimate the population mean.
- Given $\{x_1, x_2, x_3, \ldots, x_N\}$, a population of size N we defined the **Arithmetic Mean** to be the number

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

  we defined the **Population Standard Deviation** to be the number

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

- The **sampling distribution of the mean** is the distribution of all possible sample means for a given size sample selected from a population.

The Standard Error of the Mean

- Different samples of the same size from the same population will give different sample means $\overline{X}$.
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean**.
- When sampling with replacement or sampling without replacement from large or infinite populations the standard error of the mean is defined by

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- Because we divide the standard deviation σ by $\sqrt{n}$, the standard error of the mean decreases as the sample size increases.

Sampling Distribution of the Mean for Normal Population

- We say that the population is normal if the values are normally distributed.
- If a population is normal with the mean μ and the standard deviation σ, the sampling distribution of $\overline{X}$ is also normally distributed with
- the mean

$$\mu_{\overline{X}} = \mu$$

  and
- the standard deviation

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

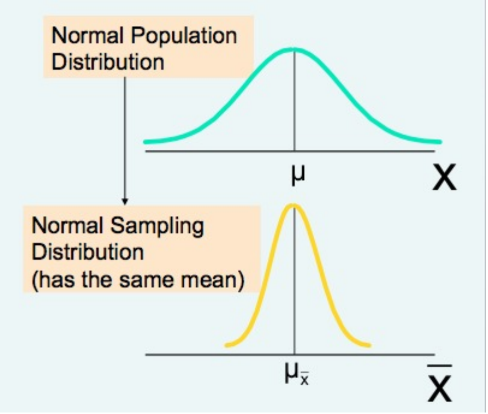
Z-value for Sampling Distribution of the Mean

- The same way we computed the Z-values for the variable X we can compute the Z-values for the sampling distribution of the mean $\overline{X}$.
- Z-value for the sampling distribution of the mean $\overline{X}$ is defined by

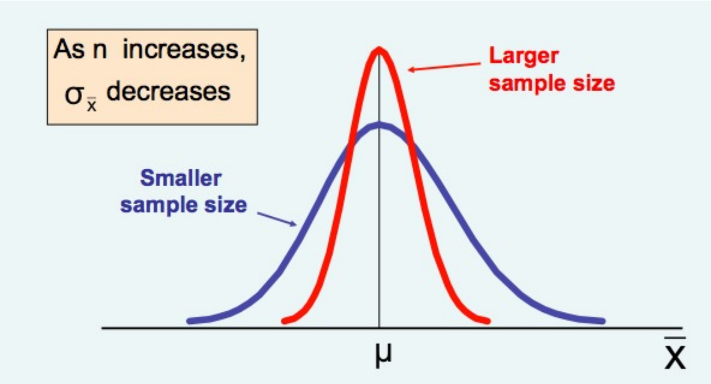
$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Where
  - $\overline{X}$ is the sample mean
  - μ is the population mean
  - σ is the population standard deviation
  - n is the sample size

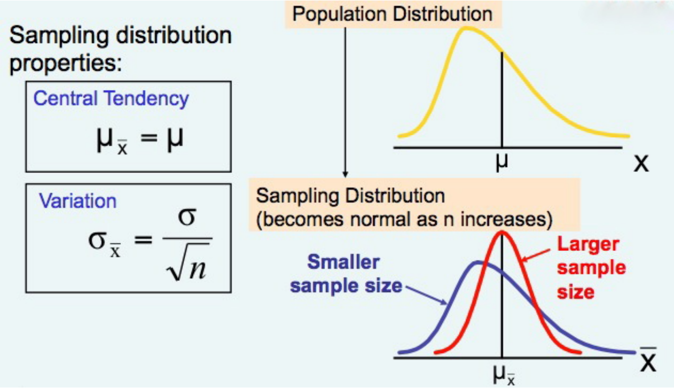
Sampling Distributions Properties

- Since $\mu_{\overline{X}} = \mu$, the sample mean $\overline{X}$ is unbiased.
- 

Sampling Distributions Properties

- As n increases, $\sigma_{\overline{X}}$ decreases
- 

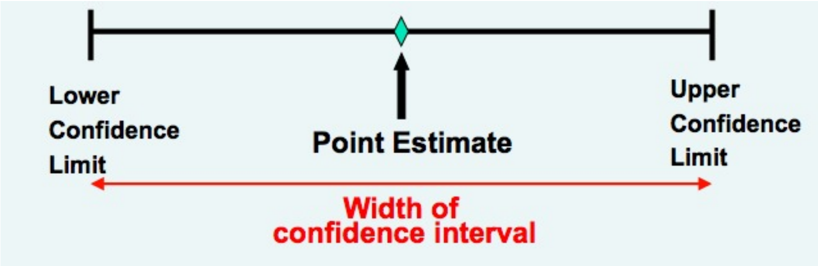
The Central Limit Theorem

- The **Central Limit Theorem** states that as the sample size (i.e., the number of values in each sample) gets *large enough*, the sampling distribution of the mean is approximately normally distributed. This is true regardless of the shape of the distribution of the individual values in the population.
- The Central Limit Theorem can be applied even if the population is not normal
- 

The Central Limit Theorem

- How large is large enough?
- For most distributions, $n \geq 30$ will give a sampling distribution that is nearly normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed.

Point and Interval Estimates

- A **point estimate** is the value of a single sample statistic,
- A **confidence interval estimate** provides additional information about the variability of the estimate
  - 

Confidence Interval Estimates

- An interval estimate provides more information about a population characteristic than does a point estimate
- It takes into consideration variation in sample statistics from sample to sample
- It is based on observations from 1 sample
- It is stated in terms of level of confidence, e.g.,
  - 99% confident, 95% confident
  - Can never be 100% confident

Confidence Interval Example

- Oxford Cereals fills thousands of boxes of cereal during an eight-hour shift. To be consistent with the labelling, boxes should contain a mean of 368 grams of cereal. The standard deviation of the cereal-filling process is 15 grams. Suppose that you randomly select a sample of 25 boxes without replacement during a shift.
- So, we have $\mu = 368$, $\sigma = 15$ and $n = 25$
- We can now find the interval that will contain 95% of the sample means based on sample of 25 boxes.
- If 95% of the means are in the interval, then 5% are outside the interval. We divide 5% into two equal parts of 2.5%. The Z value corresponding to an area of 0.025 is -1.96 (and hence the Z value corresponding to an area of 0.975 is 1.96).

Confidence Interval Example

- To get the interval we use the formula $\overline{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$
- To get the lower end point of the interval we use $Z = -1.96$. This gives us $\overline{X}_L = 368 - 1.96\frac{15}{\sqrt{25}} = 362.12$
- To get the upper end point of the interval we use $Z = 1.96$. This gives us $\overline{X}_U = 368 + 1.96\frac{15}{\sqrt{25}} = 373.88$
- So, the interval $(362.12, 373.88)$ contains 95% of the sample means.
- In general, by taking only the positive value for Z, the interval is determined by $(\mu - Z\frac{\sigma}{\sqrt{n}}, \mu + Z\frac{\sigma}{\sqrt{n}})$

Confidence Interval Example

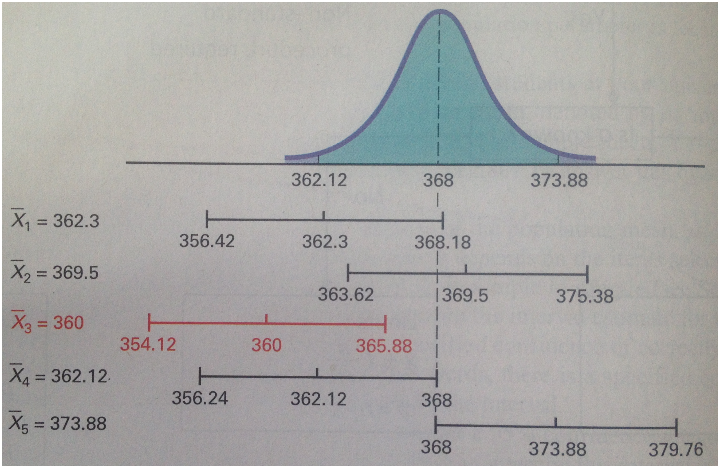
- When you do not know μ you use $\overline{X}$ to estimate μ.
- Suppose that we do not know μ but we have a sample of 25 with $\overline{X} = 362.3$
- So, in this case, the interval $(356.42, 368.18)$ contains 95% of the sample means.
- Since $356.42 \leq \mu \leq 368.18$ the interval based on this sample makes a correct statement about μ.
- What about the intervals from other samples of size 25?

Confidence Interval Example

- The following table sums up the results for 5 samples of size 25
-

Sample #	$\overline{X}$	Lower Limit	Upper Limit	Contain μ?
1	362.30	356.42	368.18	Yes
2	369.50	363.62	375.38	Yes
3	360.00	354.12	365.88	No
4	362.12	356.24	368.00	Yes
5	373.88	368.00	379.76	Yes

Confidence Interval Example

- 

Confidence Interval Example - Conclusion

- In practice you only take one sample of size n
- In practice you do not know μ and hence you do not know if the interval actually contains μ
- But, you do know that 95% of the intervals formed in this manner will contain μ
- Thus, based on the one sample you actually selected, you can be 95% confident your interval will contain μ. This is a 95% **confidence interval**.
- Note that 95% confidence is based on the fact that we used $Z = 1.96$.

Confidence Intervals

- The General Formula for Confidence Intervals is

$$\text{Point Estimate} \pm (\text{Critical Value})(\text{Standard Error})$$

- Where
  - **Point Estimate** is the sample statistic estimating the population parameter of interest
  - **Critical Value** is a table value based on the sampling distribution of the point estimate and the desired confidence level
  - **Standard Error** is the standard deviation of the point estimate

Confidence Level, $(1-\alpha)$

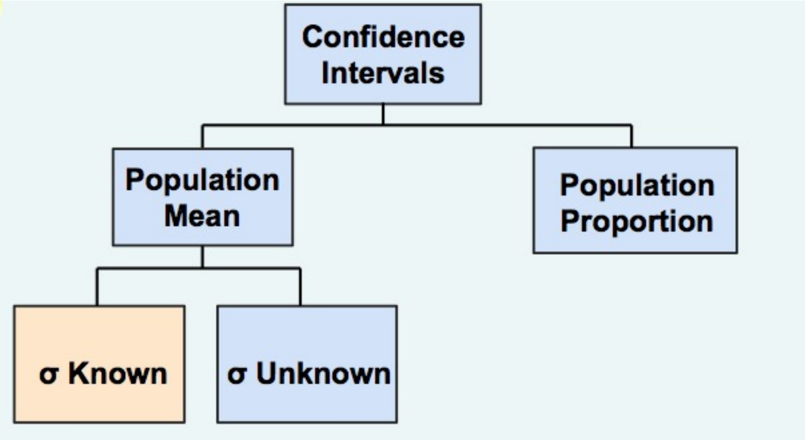
- The **Confidence Level** is given by the formula

$$(1-\alpha)\times 100\%$$

  where α is the proportion in the tails of the distribution that is outside the confidence interval.
- The proportion in the upper/ lower tail of the distribution is $\frac{\alpha}{2}$
- For example, if we want the confidence level 95% we will have $(1-\alpha) = 0.95$ that will imply $\alpha = 0.05$
- Interpretation: 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval will either contain or not contain the true parameter

Confidence Intervals

▶



Confidence Interval for μ (σ known)

- Assumptions:
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample
- The confidence interval for the mean μ when σ is known is determined by

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

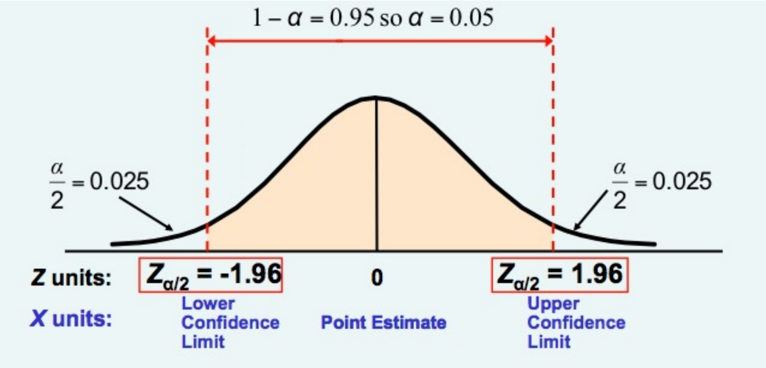
  or in the interval form as

- $$\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

  where
  - $\overline{X}$ is the point estimate
  - $Z_{\frac{\alpha}{2}}$ is the normal distribution critical value for a probability of $\frac{\alpha}{2}$ in each tail
  - $\frac{\sigma}{\sqrt{n}}$ is the standard error

Confidence Interval for μ (σ known)

- In case of 95% confidence interval we have:



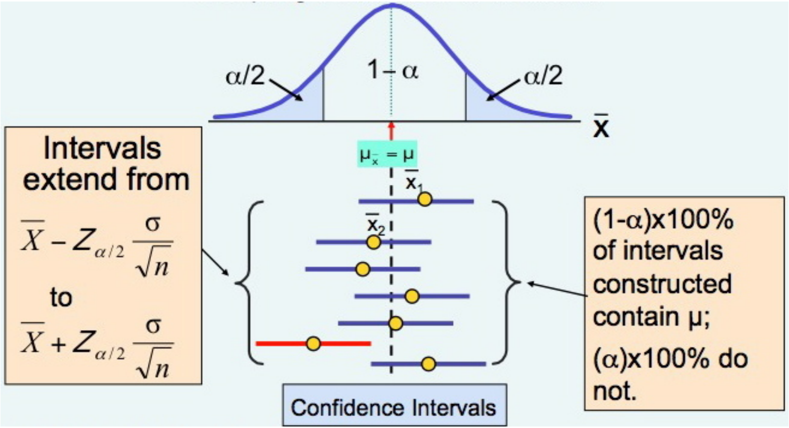
Confidence Interval for μ (σ known)

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

Confidence Level	Confidence Coefficient, $1-\alpha$	$Z_{\alpha/2}$ value
80%	0.80	1.28
90%	0.90	1.645
95%	0.95	1.96
98%	0.98	2.33
99%	0.99	2.58
99.8%	0.998	3.08
99.9%	0.999	3.27

Confidence Interval for μ (σ known)

- The summary of intervals and level of confidence



Confidence Interval for μ (σ known) - Example 1

- 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.
- We are given $\overline{X} = 2.2$, $\sigma = 0.35$ and $n = 11$.
- Using $1 - \alpha = 0.95$ we have $\alpha = 0.05$ and hence $Z_{\frac{\alpha}{2}} = 1.96$.
- Now we compute

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2.20 \pm 1.96 \frac{0.35}{\sqrt{11}} = 2.20 \pm 0.2068$$

- So, the interval is $1.9932 \leq \mu \leq 2.4068$
- **Conclusion**: 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.

Using CASIO Calculator -Example 1

1. Select **STAT**, **F4** (INTR), **F1**(Z) and then **F1**(1-S). Now in the **1-Sample ZInterval** menu select the following options:
2. Data: **F2** (Variable)
3. C-Level: **0.95 EXE**
4. σ: **0.35 EXE**
5. $\overline{x}$: **2.20 EXE**
6. n: **11 EXE**
7. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Press key **EXE**

- You will get Left $= 1.9932$ and Right $= 2.4068$, $\overline{x} = 2.20$ and $n = 11$. This means that the 95% confidence interval estimate is between Left $= 1.9932$ and Right $= 2.4068$ ohms.

Can You Ever Really Know σ ?

- Probably Not!
- In virtually all real world business situations, σ is not known.
- If there is a situation where σ is known, then you would also know μ. More precisely, you would be able to compute μ using the formula for σ. Try to express μ!
- If you truly know μ there would be no need to gather a sample to estimate it.

Confidence Interval for μ (σ unknown)

- Just as the mean of the population μ is usually unknown, you almost never know the standard deviation of the population, σ.
- If the population standard deviation σ is unknown, we can substitute the sample standard deviation s. This introduces extra uncertainty, since s is variable from sample to sample.
- So we use the **Student's t-distribution** instead of the normal distribution.
- **Student's t-distribution** was introduced by William Gosset while he was working for Guinness Breweries in Ireland. He wanted to make inferences about the mean when σ was unknown. It is commonly referred to as **t-distribution**.

Confidence Interval for μ (σ unknown)

- If the random variable X is normally distributed, then the following statistic has a **t-distribution** with $n-1$ **degrees of freedom**

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

- This expression has the same form as the Z statistic except that s is used to estimate the unknown σ.
- The t-distribution looks very similar to the standardized normal distribution Z. Both distributions are symmetrical and bell shaped with the mean 0. The t-distribution has more area in the tails and less in the centre than does the standardized normal distribution.

The concept of Degrees of Freedom

- Basically, the **degree of freedom** is a number of observations that are free to vary after a sample mean has been calculated.
- Suppose that a sample of five values has a mean 20. That is

$$\frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = 20$$

- How many values x_i do you need to know before you can determine the remainder of the values?
- Using $x_1 + x_2 + x_3 + x_4 + x_5 = 100$ we see that if we know four values the fifth value is not going to be free. For example if we know x_1, x_2, x_3 and x_4 then $x_5 = 100 - x_1 - x_2 - x_3 - x_4$. This means that you have 4-degrees of freedom.
- In general, if we know the sample mean $\overline{X}$ of n values, only $n-1$ of the sample values are free. This means that you have $n-1$ degrees of freedom.

Confidence Interval for μ (σ unknown)

- Assumptions:
  - Population standard deviation σ is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- The confidence interval for the mean μ when σ is unknown is determined by

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

  or in the interval form as

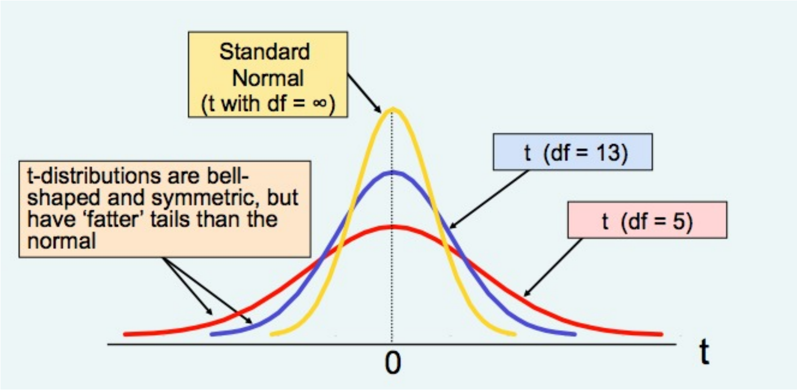
-

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

  where
  - $\overline{X}$ is the point estimate
  - $t_{\frac{\alpha}{2}}$ is the critical value of the t-distribution with $n-1$ degrees of freedom and an area of $\frac{\alpha}{2}$ in each tail.

Properties of t-Distribution

- Note that $t \to Z$ as n increases!



Confidence Interval for μ (σ unknown)

- This table illustrates the fact that $t \to Z$ as n increases numerically. The full table of $t_{\frac{\alpha}{2}}$ values for a sample of size n is given on the page 730, Table A.3.

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z (∞ d.f.)
0.80	1.372	1.325	1.310	1.28
0.90	1.812	1.725	1.697	1.645
0.95	2.228	2.086	2.042	1.96
0.99	3.169	2.845	2.750	2.58

Note: $t \longrightarrow Z$ as n increases

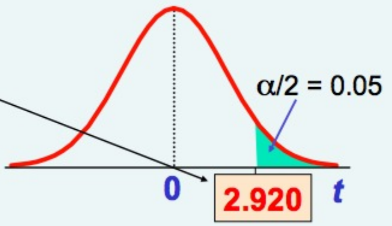
Confidence Interval for μ (σ unknown)

- $t_{\frac{\alpha}{2}}$ value depends on $n-1$ degrees of freedom and the upper tail area.

	Upper Tail Area		
df	.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 = 0.10
α/2 = 0.05



Using CASIO Calculator to determine the critical value $t_{\frac{\alpha}{2}}$

- Determine the critical value for an area of 0.025 in each tail with 99 degrees of freedom. We have $df = 99$ and $\frac{\alpha}{2} = 0.025$. We want the critical value $t_{\frac{\alpha}{2}}$

1. Select **STAT**, **F5** (DISTR), **F2**(t) and then **F3**(Invt). Now in the **Inverse Student-t** menu select the following options:
2. Data: **F2** (Variable)
3. Area: **0.025 EXE**
4. df: **99 EXE**
5. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
6. Press key **EXE**

- You will get xInv $= 1.98421695$. The calculator only gives you the upper tail area. The value for the lower tail area is -1.98421695 because the t-distribution is symmetric.

Confidence Interval for μ (σ unknown) - Example 2

- An accountant of Saxon Home Improvement wants to estimate the mean dollar amount listed on the sales invoices for the month. The accountant selects a sample of 100 sales invoices from the population of sales invoices during the month and and finds that the sample mean of the 100 sales invoices is \$110.27, with the standard deviation of \$28.95. Find the 95% confidence interval estimate for the sales invoices mean.
- We are given $\overline{X} = 110.27$, $s = 28.95$, $n = 100$.
- Using $1 - \alpha = 0.95$ we have $\alpha = 0.05$. Now either using your calculator or looking up into the table E.3 on page 812 (11th edition) we find that $t_{\frac{\alpha}{2}} = 1.9842$ for $df = 99$ and $\frac{\alpha}{2} = 0.025$.

Confidence Interval for μ (σ unknown) - Example 2

- Now we compute

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 110.27 \pm 1.9842 \frac{28.95}{\sqrt{100}} = 110.27 \pm 5.74$$

- So, the interval is $104.53 \leq \mu \leq 116.01$
- **Conclusion**: We are 95% confident that the mean amount of all the sales invoices is between $104.53 and $116.01.
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.

Using CASIO Calculator -Example 2

1. Select **STAT**, **F4** (INTR), **F2**(t) and then **F1**(1-S). Now in the **1-Sample tInterval** menu select the following options:
2. Data: **F2** (Variable)
3. C-Level: **0.95 EXE**
4. $\overline{x}$: **110.27 EXE**
5. sx: **28.95 EXE**
6. n: **100 EXE**
7. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Press key **EXE**

- You will get Left $= 104.53$ and Right $= 116.01$, $\overline{x} = 110.27$, $sx = 28.95$ and $n = 100$. This means that the 95% confidence interval estimate is between Left $= 104.53$ and Right $= 116.01$.

Confidence Interval for μ (σ unknown) - Example 3

- A manufacturing company produces electric insulators. If the insulators break when in use, a short circuit is likely. To test the strength of the insulators, you carry out destructive testing to determine how much force is required to break insulators. You measure force by observing how many pounds are applied to the insulatorbefore it breaks. The following table lists 30 values from this experiment.

1870	1728	1656	1610	1634	1784	1522	1696	1592	1662
1866	1764	1734	1662	1734	1774	1550	1756	1762	1866
1820	1744	1788	1688	1810	1752	1680	1810	1652	1736

- Construct a 95% confidence interval estimate for the population mean force required to break the insulator.

Using CASIO Calculator -Example 3

1. Select **STAT**, and enter the data into **List 1**. Then from the **STAT** mode select **F4** (INTR), **F2**(t) and then **F1**(1-S). Now in the **1-Sample tInterval** menu select the following options:
2. Data: **F1** (List)
3. C-Level: **0.95 EXE**
4. List: **List 1**
5. Freq: **1 EXE**
6. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Press key **EXE**

- You will get Left = 1689.96117 and Right = 1756.83883, $\overline{x} = 1723.4$, $sx = 89.5508332$ and $n = 30$. This means that the 95% confidence interval estimate is between Left = 1689.96117 and Right = 1756.83883 pounds .