CONFIDENCE INTERVAL

Important Terms :

1. Population vs. Sample

- Population: This is the *entire group* of individuals or items you are interested in studying. For example, if you want to study the average height of people in a country, the population is *everyone in the country*.
- Sample: Since studying the whole population can be difficult, you take a smaller group from the population to study. This smaller group is the *sample*. For example, you might study the heights of 1,000 randomly chosen people from the country.
  - Key Idea: The population is the "big picture," and the sample is a smaller "snapshot."

2. Parameter vs. Statistic

- Parameter: This is a number that describes the *entire population*. For example, the true average height of everyone in the country is a parameter.
- Statistic: This is a number that describes the *sample*. For example, the average height of the 1,000 people in your sample is a statistic.
  - Key Idea: Parameters are for populations, while statistics are for samples.

3. Inferential Statistics

- This is a method of using data from a *sample* to make predictions or draw conclusions about the *population*.
  - Example: If the average height in your sample is 5'7", inferential statistics helps you estimate that the average height in the entire country is *about* 5'7".
  - Key Idea: It's like making an educated guess about the whole population based on the sample.

4. Point Estimate

- A point estimate is a single value (number) that you use to estimate a population parameter.
  - Example: If the average height of your sample is 5'7", you use this as a point estimate for the average height of the entire population.
  - Key Idea: It’s a “best guess” for the population value based on your sample.

CONFIDENCE INTERVAL

Confidence Interval (CI):

A confidence interval is a range of values used to estimate a population parameter (like the population mean). Instead of giving one exact value (point estimate), we provide a range in which the true value is likely to lie.

- Example: "The average height of people is between 5'5" and 5'9" with 95% confidence."

Confidence Level :

The confidence level tells us how certain we are that the true population parameter falls within the confidence interval.

- Example: A 95% confidence level means that if we repeated the sampling 100 times, the confidence interval would contain the true population mean in 95 out of 100 samples.
- Common confidence levels: 90%, 95%, 99%.

Ways to Calculate Confidence Intervals :

1. Z-Procedure (When σ sigma is Known):

- Used when the population standard deviation (σ sigma) is known.
- Assumes data follows a normal distribution (or $n \geq 30$).
- Relies on the Z-distribution, which is fixed and symmetrical.

Key Idea:
We use the known sigma σ to measure how far the sample mean ($\bar{x}$ could vary from the true population mean, scaling this variation by the Z-value for the chosen confidence level.

2. T-Procedure (When sigma σ is Unknown):

- Used when the population standard deviation (σ sigma) is unknown.
- Uses the sample standard deviation (s) as an estimate of σ sigma.
- Relies on the T-distribution, which is wider and accounts for more uncertainty (especially in small samples).

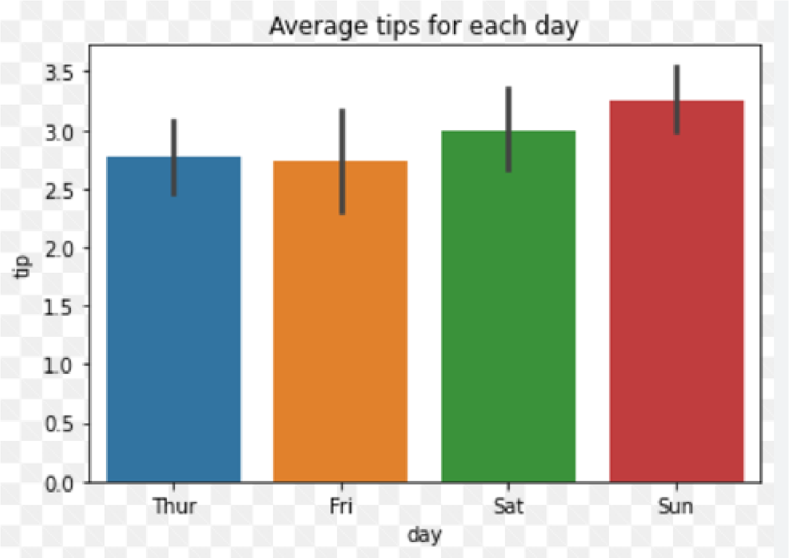
Key Idea:
Since σ sigma is unknown, the extra uncertainty in estimating it is incorporated using a T-value (which adjusts for smaller sample sizes with degrees of freedom).

Examples of Using Confidence Interval:

1. Bar Chart with Line on Top (Error Bars):

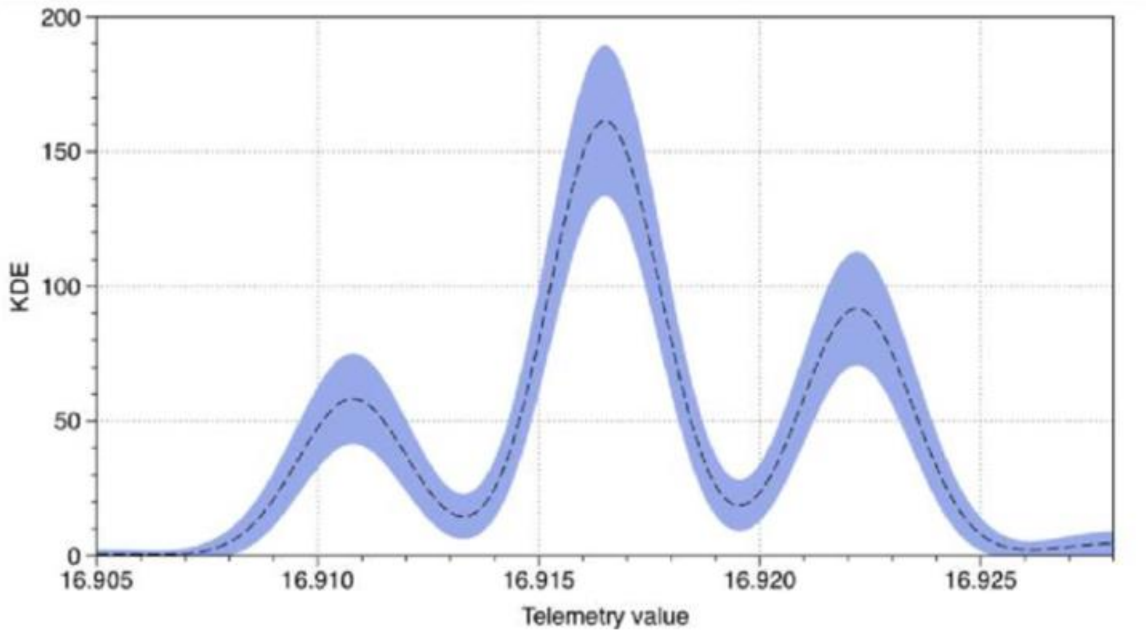
- Confidence intervals are shown as vertical lines (error bars) on top of bars.
- Purpose: Visualize the uncertainty or variability in the mean values of different groups.

CONFIDENCE INTERVAL



2. KDE Plot with Shadow (Shaded Area):

- **Confidence intervals are shown as a shaded region around the density curve.**
- **Purpose: Represent the range where the true density lies, adding uncertainty to the smooth distribution.**



Interpreting Confidence Interval:

A confidence interval is not saying there’s a 95% chance the true mean is in the interval. Instead, it means that if we repeated the experiment many times, 95% of the intervals we calculate would contain the true mean.

CONFIDENCE INTERVAL

Factors Affecting Margin of Error:

The margin of error determines the width of the confidence interval and depends on:

1. **Confidence Level (1−alpha):**
    - **Higher confidence levels (e.g., 99%) lead to wider confidence intervals because you want to be more certain.**
    - **Lower confidence levels (e.g., 90%) lead to narrower intervals but with less certainty.**
2. **Sample Size (n):**
    - **Larger samples reduce the margin of error and result in narrower intervals.**
    - **Smaller samples increase the margin of error and lead to wider intervals.**
3. **Population Standard Deviation (σ sigma):**
    - **Higher variability (larger σ sigma) leads to wider intervals.**
    - **Lower variability (smaller σ sigma) leads to narrower intervals.**

Calculating Confidence Interval Using Z-Procedure:

A confidence interval gives a range where the true population mean likely falls, based on sample data. The Z-procedure is used when the population standard deviation (σ sigma) is known.

Scenario Example

A company wants to estimate the average time employees spend commuting to work.

- **Sample mean (x̄): 40 minutes**
- **Population standard deviation (σ sigma): 5 minutes**
- **Sample size (n): 50**
- **Confidence level: 95% (Z=1.96Z)**

Formula for Confidence Interval:

$$CI = \bar{x} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

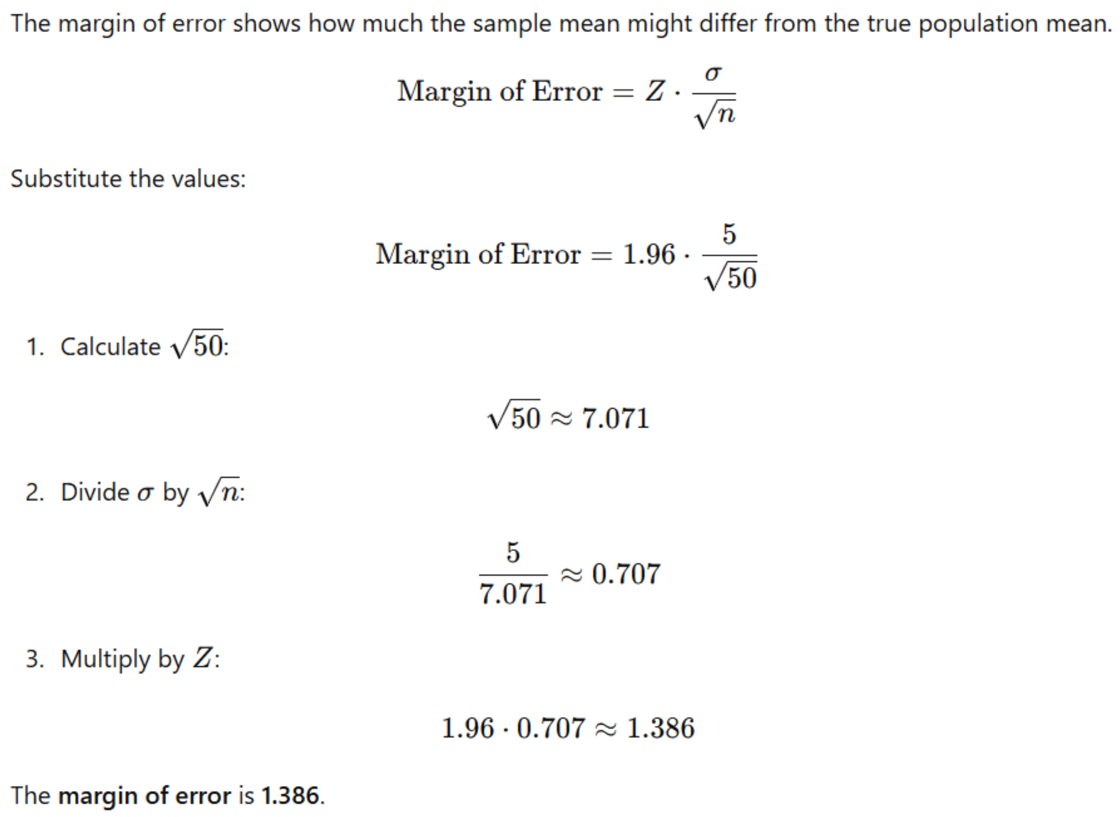
Where:

- **x̄: Sample mean**
- **Z: Z-score corresponding to the confidence level (e.g., 1.96 for 95%)**
- **sigma: Population standard deviation**

CONFIDENCE INTERVAL

- **n: Sample size**

Step 1: Calculate the Margin of Error



The margin of error shows how much the sample mean might differ from the true population mean.

$$\text{Margin of Error} = Z \cdot \frac{\sigma}{\sqrt{n}}$$

Substitute the values:

$$\text{Margin of Error} = 1.96 \cdot \frac{5}{\sqrt{50}}$$

1. Calculate $\sqrt{50}$:

$$\sqrt{50} \approx 7.071$$

2. Divide σ by $\sqrt{n}$:

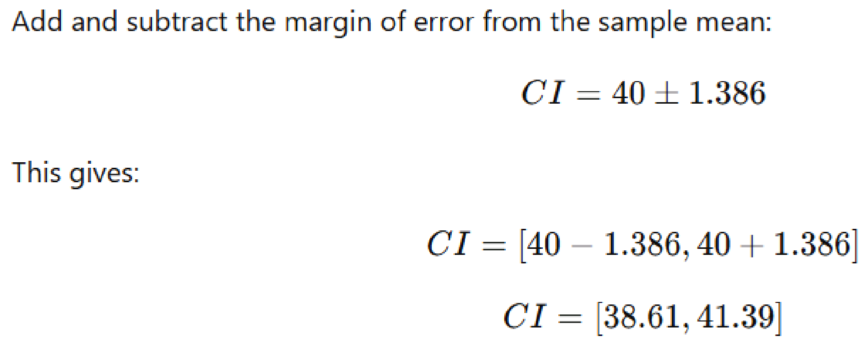
$$\frac{5}{7.071} \approx 0.707$$

3. Multiply by Z:

$$1.96 \cdot 0.707 \approx 1.386$$

The **margin of error** is **1.386**.

Step 2: Calculate the Confidence Interval



Add and subtract the margin of error from the sample mean:

$$CI = 40 \pm 1.386$$

This gives:

$$CI = [40 - 1.386, 40 + 1.386]$$

$$CI = [38.61, 41.39]$$

Interpretation

We are 95% confident that the true average commuting time for all employees lies between 38.61 minutes and 41.39 minutes.

CONFIDENCE INTERVAL

How to Find the Z-Value?

The Z-value determines how many standard deviations to include to cover the desired confidence level.

Steps to Find Z-Value:

1. **Understand Confidence Level:**
   - **For a 95% confidence level, 95% of the area is in the middle of the standard normal curve, and 5% is in the tails (split into 2.5% in each tail).**
2. **Find Tail Area:**
   - **Tail area = 1−Confidence Level=1−0.95=0.051**
   - **Each tail area = 0.05/2=0.025**
3. **Look Up Z-Value:**
   - **Use a Z-table or cumulative probability. For 95% confidence level:**
     - **Cumulative probability = 1−0.025=0.9751**
     - **Find 0.975 in the Z-table: Z=1.96Z =1.96.**

Calculating Confidence Interval Using the T-Procedure :

When we don't know the population standard deviation (σ sigma), we use the T-procedure to calculate the confidence interval. This method is used for small sample sizes (less than 30) or when we need to estimate the population standard deviation based on the sample.

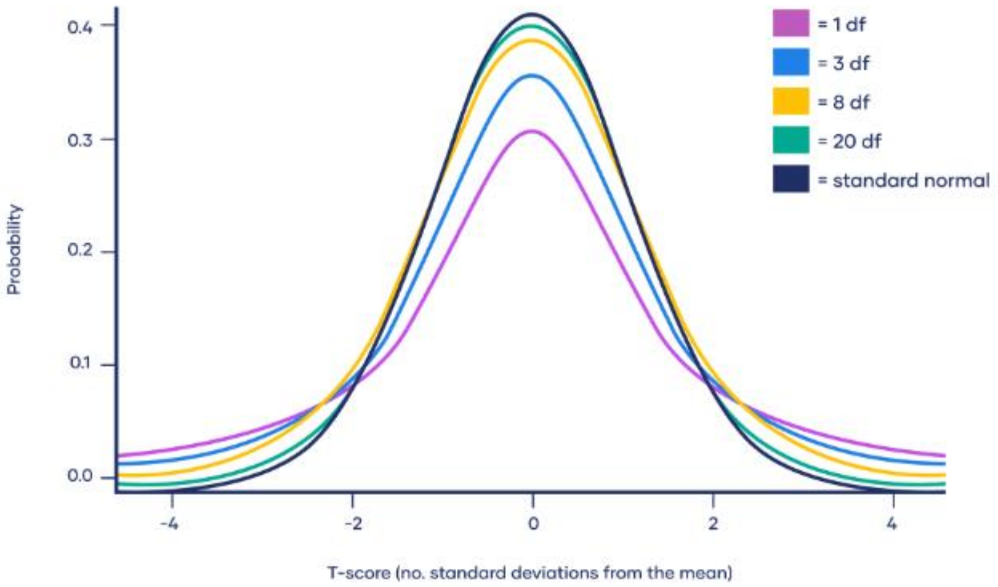
What is the Student's T-Distribution?

The T-distribution is like the normal distribution but has wider tails. This means it accounts for more variability in small samples.

- **The shape of the T-distribution depends on the sample size.**
- **Degrees of Freedom (df): The T-distribution is determined by df = n - 1, where nnn is the sample size.**

The larger the sample size, the closer the T-distribution is to the normal distribution. But for small samples, the T-distribution is more spread out, so we need to use it for more accurate results.

CONFIDENCE INTERVAL



How is the T-Distribution Used to Calculate Confidence Interval?

When we don't know σ sigma (the population standard deviation), we use the sample standard deviation (s). The T-value helps us adjust for the uncertainty in estimating σ.

Formula for Confidence Interval Using T-Procedure:

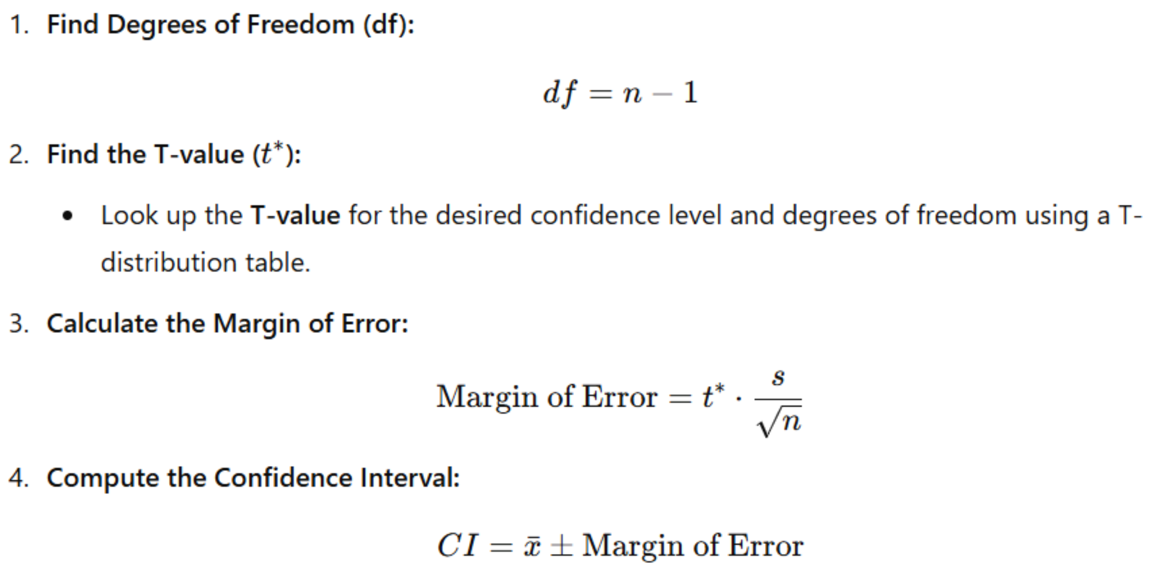
$$CI = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Where:

- **x¯: Sample mean**
- **t∗: T-value (based on the confidence level and degrees of freedom)**
- **s: Sample standard deviation**
- **n: Sample size**

CONFIDENCE INTERVAL

Steps to Calculate Confidence Interval Using T-Procedure:



1. **Find Degrees of Freedom (df):**

$$df = n - 1$$

2. **Find the T-value (t^*):**

- Look up the **T-value** for the desired confidence level and degrees of freedom using a T-distribution table.

3. **Calculate the Margin of Error:**

$$\text{Margin of Error} = t^* \cdot \frac{s}{\sqrt{n}}$$

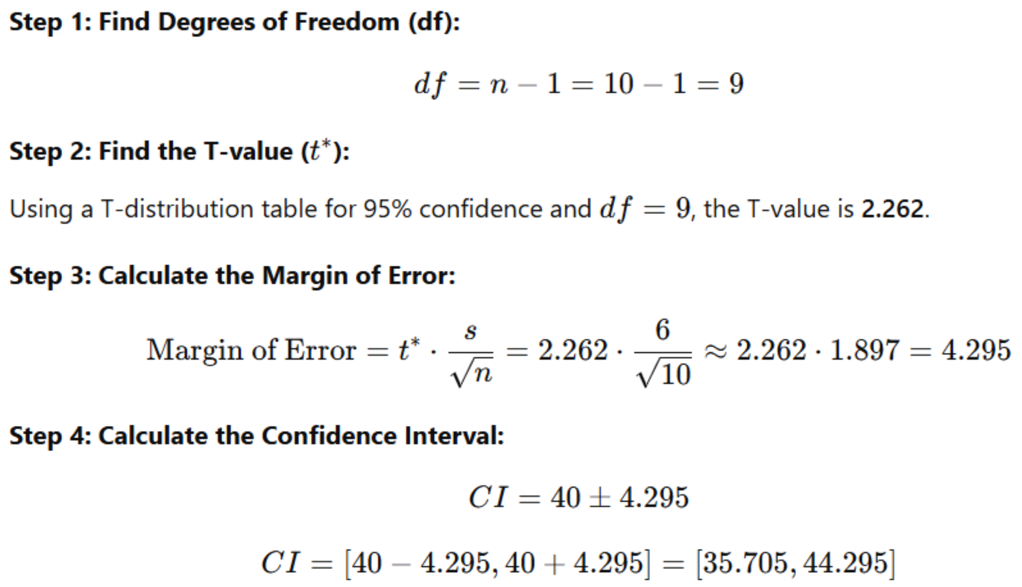
4. **Compute the Confidence Interval:**

$$CI = \bar{x} \pm \text{Margin of Error}$$

Example:

Imagine a small company wants to estimate the average hours its employees work per week.

- **Sample mean (x¯): 40 hours**
- **Sample standard deviation (s): 6 hours**
- **Sample size (n): 10**
- **Confidence level: 95%**



Step 1: Find Degrees of Freedom (df):

$$df = n - 1 = 10 - 1 = 9$$

Step 2: Find the T-value (t^*):

Using a T-distribution table for 95% confidence and $df = 9$, the T-value is **2.262**.

Step 3: Calculate the Margin of Error:

$$\text{Margin of Error} = t^* \cdot \frac{s}{\sqrt{n}} = 2.262 \cdot \frac{6}{\sqrt{10}} \approx 2.262 \cdot 1.897 = 4.295$$

Step 4: Calculate the Confidence Interval:

$$CI = 40 \pm 4.295$$

$$CI = [40 - 4.295, 40 + 4.295] = [35.705, 44.295]$$

CONFIDENCE INTERVAL

Interpretation:

We are 95% confident that the true average number of hours worked by all employees lies between 35.705 and 44.295 hours.

Why Use the T-Procedure?

The T-distribution is used when the sample size is small, and the population standard deviation is unknown. It accounts for the extra uncertainty in these situations by using wider tails to ensure a more accurate estimate. As the sample size increases, the T-distribution approaches the normal distribution, and both methods give similar results.