HCAI5DS02 – Data Analytics and Visualization.
Lecture – 06
Statistical Modeling
Confidence Interval and Statistical Inference.

Siman Giri

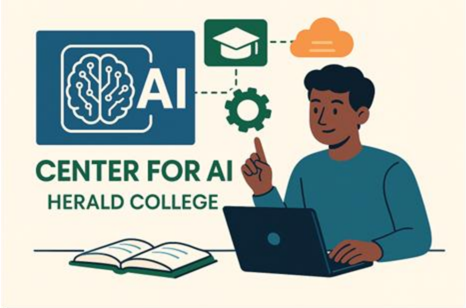


image generated via copilot.

HERALD COLLEGE | ing CENTER FOR AI.

3.3 How to make an Estimation?

Two Main Approaches to Estimation.

Approach.	**What it Does?**	**Example.**
Point Estimation ✓	Gives a **single best guess** for a parameter.	$\hat{p} = \frac{\text{clicks}}{\text{emails}} = 0.06$ ■
Interval Estimation	Gives a **range of plausible values** for the parameter (with confidence)	95% CI for p: [0.045, 0.075]

- A **point estimate** tells you what's **most likely.**
- An **interval** tells you **how sure** you are — and **what could go wrong**.

1.2 Parameters.

- A **parameter** is a fixed (but usually unknown) **numerical characteristic** of a **population or probability distribution.**
  - Think of it as describing the **true model** behind the data.
  - Parameters are often denoted using **Greek letters** **e.g.** $\boldsymbol{\mu, \sigma, \lambda, p}$.
  - Parameters are **not calculated** from data — they are **assumed** to exist **in the population.**
- Examples:

Parameter	Meaning	Example
μ	Population Mean	True average purchase amount.
σ^2	Population Variance	Variability in customer spending.
p	Population proportion	True click through rate.
λ	Rate parameter	Avg. visits.

"Why do we need parameters?"

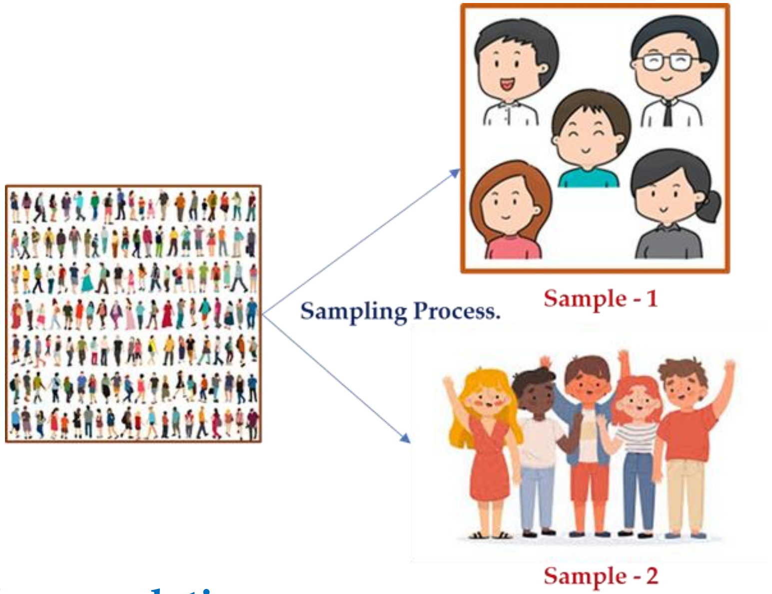
1.2.3 But There's a Catch …

- **Parameters** define the **underlying behavior** of a **process or population**.
  - Knowing the parameter helps us **describe, predict, and make decisions** using **data.**
- **But there's a Catch:**
  - Most **parameters** are **unknown** in the real world.
    - That's why we use **statistics** to **estimate** them using data.
  - **What are statistics?**

1.3 Statistic.

- A **statistic** is a **numerical summary** calculated from **sample data** used to **estimate** a **population parameter**.
  - **Statistics** are **observable** and **vary** from **sample to sample**.
  - Denoted by Roman letters (e.g. $\bar{x}, s^2, \hat{p}$) .
  - Statistics are our **best guesses** for unknown **parameters.**

Statistic Symbol	Meaning	Estimating Parameter
$\bar{x}$	Sample Mean	Estimates μ
s^2	Sample Variance	Estimates σ^2
$\hat{p}$	Sample proportion	Estimates p
Statistics (What we observe.)	Estimates →	Population (Truth)



- Parameters are to **populations** what statistics are to **samples**.
- We *use statistics to estimate parameters* — because **we rarely have full access** to **the population**.

2.1 Parameters Estimation: Introduction.

- **Key Idea:**
  - A **parameter** is a fixed but unknown quantity about the population
    - (e.g. the true average conversion rate.)
  - A **statistic** is a value we compute from a sample to estimate that parameter.
  - An **Estimation** is the process of using sample data to infer the value of an unknown population parameter.
  - It is a statistical method used when the true value is not directly observable.

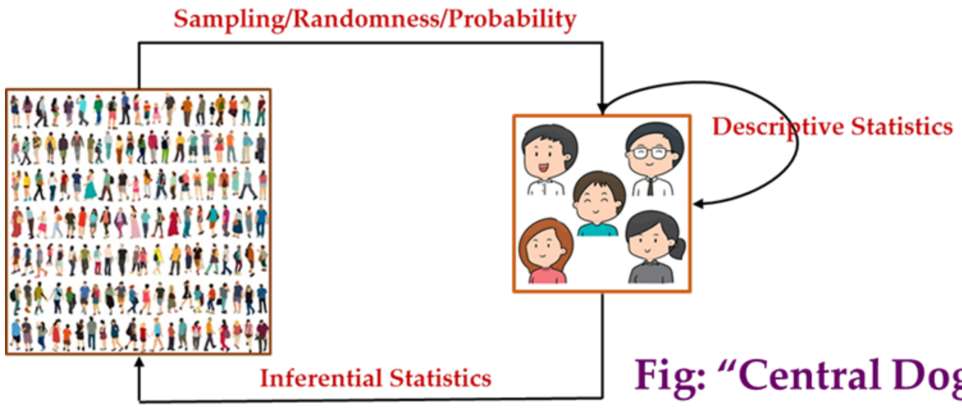
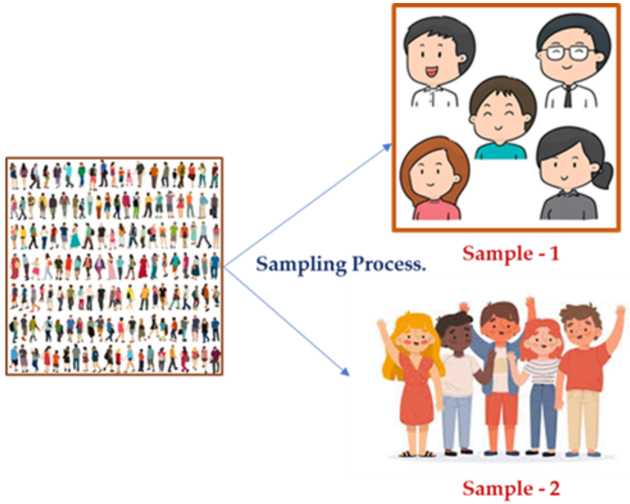


Fig: "Central Dogma" of Statistics.

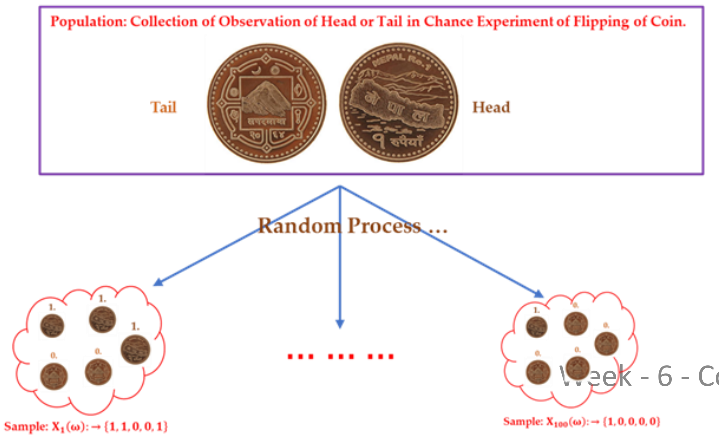


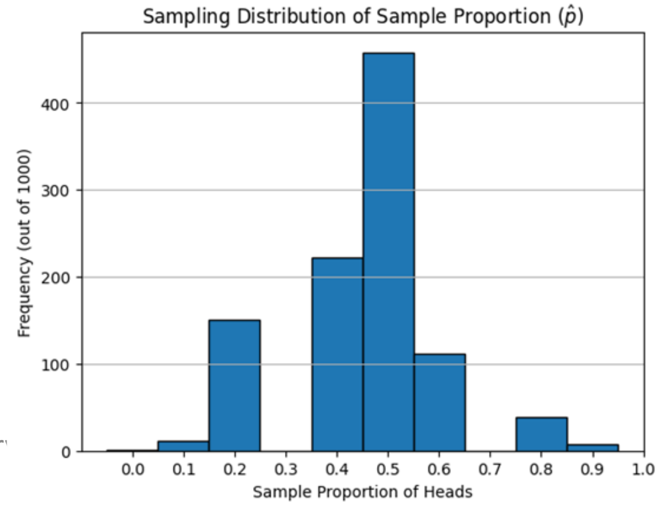
2.2 Estimator and Estimate.

- **Estimator:**
  - An **estimator** is a **mathematical rule or formula** applied to a **sample** to **compute an estimate** of an unknown population parameter.
  - The estimator itself is a random variable because it depends on which sample you get.
  - We usually denote estimators with a "hat":
    - $\hat{\theta} = \text{Estimator}(X_1, X_2, \ldots, X_n)$
  - **Common estimators:**
    - $\bar{X}$: estimates the population mean μ
    - $\hat{p}$: estimates the true proportion p
    - s^2: estimates population variance σ^2
- **Estimate:**
  - An **estimate** is the **numerical value** you get when you **apply the estimator to a specific data sample**.
  - It is not random but a fixed number (result) from the estimator applied to your collected sample.
  - Example: If $\hat{p} = \frac{\text{clicks}}{\text{emails}}$, and your sample has 50 clicks out of 1,000:
    - $\hat{p} = \frac{50}{1000} = 0.05 (\text{estimate})$

2.3.3 Towards Sampling Distributions.

- **Key Idea:**
  - A single estimate ($\hat{\mathbf{p}} = \mathbf{0.6}$), may not **match the true parameter ($\boldsymbol{p} = \mathbf{0.5}$)** ,
    - but the **average of many estimates** will be **close to the truth** (if the estimator is unbiased).
- **Why?**
  - **Law of Large Numbers:**
    - As the number of **sample increases**, the **mean of sampling distribution** of **converges to the true value**.
  - **Example:**
    - Suppose you **simulate 1,000 samples** of **size 5-coin flips** and **compute $\hat{\boldsymbol{p}}$** for each. The average of all those $\hat{p}$ values will be very close to 0.5. i.e.
      - $\mathbb{E}[\hat{\mathbf{p}}] = \mathbf{p} = \mathbf{0.5}$ **(if estimator is unbiased)**





2.4 Sampling Distributions.

- The **sampling distribution** of a statistic (or estimator) is the **probability distribution** of that statistic computed from all possible **random samples** of a **fixed size n** drawn from a given population.
  - It tells you **how the estimator (e.g., sample mean, proportion)** would vary **if you repeatedly sampled** from the same population.
- **Example:**
  - You have a population with a true parameter (e.g. **true mean μ**).
  - You **dray many samples** from **this population**:
    - $Sample\ 1 \rightarrow X_1^1, X_2^1, \dots, X_n^1$ → Compute $\bar{X}^1$
    - $Sample\ 2 \rightarrow X_1^2, X_2^2, \dots, X_n^2$ → Compute $\bar{X}^2$
    - … … …
    - $Sample\ m \rightarrow X_1^m, X_2^m, \dots, X_n^m$ → Compute $\bar{X}^m$
- **Now you have collection of means: $\bar{\mathrm{X}}^1, \bar{\mathrm{X}}^2, \dots, \bar{\mathrm{X}}^m$**
- These form the **sampling distribution of the estimator $\bar{\mathrm{X}}$**.

1. Before Confidence Interval.

{Sampling, Sampling Distribution and Sampling Error.}

1.1 Sampling Error.

- **Intuitions behind Sampling Error:**
  - We want to use a **sample to learn something about a population**, but **no sample is perfect!**
    - **Sampling error** is the **error resulting from using a sample to estimate a population characteristic**.
  - If we use a **sample mean $\bar{x}$** to **estimate μ**, chances are:
    - that $\bar{x} \neq \mu$ (sometimes they might be close, sometimes they might be not)
- These form the **sampling distribution of the estimator $\bar{X}$**.
- **We consider:**
  - How close **$\bar{x}$ to μ** ?
  - What if we took lots of samples and calculated $\bar{x}$ each time?
    - **Would these values cluster around μ ?**
  - What **would the shape of that distribution look like?**

1.1.1 Redefining: Sampling Distributions.

- The sampling distribution is **the distribution of a sample statistic like $\bar{x}$**
  - calculated **from all possible samples of a given size n**.
  - We in **general** focus on **the sampling distribution of the sample mean**.
- **In Layman terms:**
  - For a **variable X**, if we repeatedly take **sample size n** and compute $\bar{x}$ each time,
    - the **distribution of those sample means** is what **the sampling distribution of $\bar{x}$**.
- **Formal Definition of Sampling Error:**
  - Sampling error is the difference between
    - the **sample statistics like the sample mean**, (**denoted by $\bar{x}$**) and
    - the **true population parameter** (like population mean, **denoted by μ**).
  - Mathematically:
    - $\text{Sampling Error} = \bar{x} - \mu$
  - It arises because we only observe a subset of the population, not the whole.

1.2 Example: Average Delivery Time for a Food Delivery App.

- **Scenario:**
  - You are a data analyst at **QuickBite**, a food delivery company operating in a large metropolitan area.
  - Your team is interested in understanding the **average delivery time** (in minutes) across the city. The delivery times vary due to traffic, order volume, and distance. However, collecting the delivery time for **every order** is not feasible due to system constraints. So, you collect a **sample of size n = 2** from recent deliveries.
  - Assume, for learning purposes, that you know the entire **population of delivery times** for the last 5 orders from a specific zone:

Order ID	Delivery Time (min)
A	30
B	32
C	35
D	38
E	40

- **So, the population mean is:**
  - $\mu = \frac{30+32+35+38+40}{5} = 35$ **minutes.**

1.2 Example: Average Delivery Time for a Food Delivery App.

- **Now Suppose we take all samples of size $n = 2$.**
  - There are **10 possible combinations (ignoring order),**
    - and we can compute their **sample means** and **sample error**:

Sample	Sample Mean ($\bar{x}$)	Sampling Error $\bar{x} - \mu$
A,B	31	- 4.0
A,C	32.5	- 2.5
A,D	34	- 1.0
A,E	35	0.0
B,C	33.5	- 1.5
B,D	35	0.0
B,E	36	+ 1.0
C,D	36.5	+ 1.5
C,E	37.5	+ 2.5
D,E	39	+ 4.0

- The **sampling error** is defined for a single sample as:
  - **Sampling Error = $\bar{x} - \mu$**
  - Every sample will have its own sampling error
    - **some positive (overestimate error) and some negative (underestimate error).**
- **What happens over many Samples?**
  - If you take many samples, each with its own sampling error, then:
    - **Average sampling error across all possible samples = 0**
      - It is good, it **means our estimator is unbiased** i.e.
        - $\mathbb{E}[\bar{x}] = \mu$
  - The expected value or long run average of the sample mean $\bar{x}$ equals the true population mean μ.
- But just knowing the average error is zero, doesn't tell us **how big those errors tend to be,**
  - **or how it varies from sample to sample or how reliable our estimate is.**

1.3 Standard Error (SE).

- **Defintion:**
  - Standard Error is the **standard deviation** of a **sampling distribution**.
  - It measures the typical amount that a sample statistic like the sample mean
    - differs from the true population parameter due to random sampling.
  - For the sample mean:
    - Standard Error of the Mean $= \mathbf{SE_{\bar{x}}} = \frac{\sigma}{\sqrt{n}}$
  - **What if σ is unknown?**
    - In practice, we rarely know the population standard deviation.
    - So, we estimate it from the sample:
      - $\mathbf{SE_{\bar{x}}} \approx \frac{s}{\sqrt{n}}$
    - where **s** is the **sample standard deviation**.
- **Interpretation**
  - SE tells us **how much we expect the sample mean to vary** from **sample to sample**.
  - A **smaller SE** means the sample mean is **more stable** and likely closer to the population mean.
  - As **n increases**, **SE decreases** → **larger samples are more precise**.

2. Introduction to Confidence Interval.

2.1 Need for Confidence Interval.

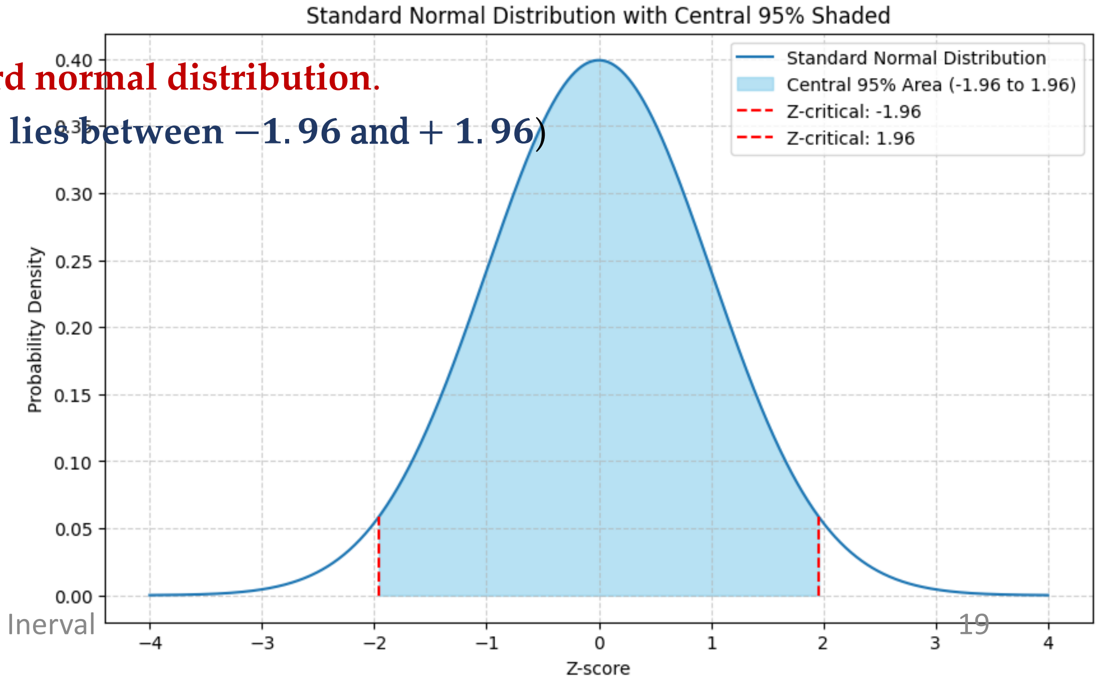
- **Recall the Point Estimate:**
  - A **point estimate** is a **single value estimate** of a **population parameter**.
  - We say that a **statistic is an unbiased estimator**
    - if the **mean of its distribution is equal to the population parameter**.
      - Otherwise, **it is a biased estimator**.
  - Point estimates are useful, but they only give us so much information.
  - The **variability of an estimate** is also important!!!
- **Why Confidence Intervals?**
  - When we take a sample from a population, we often compute a sample statistic (**like the sample mean $\bar{x}$**) to estimate the population parameter (**like the true mean μ**).
    - But because samples vary, **we want to capture the range** in **which μ** likely falls:
  - A **confidence interval (CI)** gives us such a range, using:
    - The point estimate: $\bar{x}$
    - Sampling variability: quantified using **standard error (S.E)**
    - Confidence Level: determines how wide the interval is **via a critical value**.

2.2 Confidence Interval.

- A confidence interval **gives a range of values**
  - that is **likely to contain** the **true population parameter (like the mean μ)**
    - with a **certain level of confidence (e.g. 95%).**
  - General Formula (**for the population mean μ**):
    - **Confidence Interval** $= \bar{x} \pm z^{*} \cdot SE$
    - Here:
      - $\bar{x}$: sample mean
      - z^{*}: critical value from the standard normal distribution(e. g. 1.96 for 95% confidence)
      - SE: standard error of the sample mean.

2.2.1 Critical Value

- **Formal Definition:**
  - A **critical value** is a value from a **probability distribution** (**usually the standard normal distribution**) such that:
    - The **area between the critical values** corresponds to the **confidence level e.g. 95%.**
    - The **area beyond the critical values** (**in the tails**) corresponds to **the significance level e.g. 5%** total or 2.5% in each tail.
    - In **confidence intervals**:
      - Suppose **we want a 95% confidence interval** for the **population mean**:
        - $\bar{\mathbf{x}} \pm \mathbf{z}^* \cdot \mathbf{S.E}$
        - Here: $\mathbf{z}^*$ **is the critical value** from the **standard normal distribution**.
        - For **95%,** $\mathbf{z}^* \approx \mathbf{1.96}$ (**because 95% of the area lies between** $-\mathbf{1.96}$ **and** $+\mathbf{1.96}$)



2.2.2 Margin of Error(ME).

- The **margin of error** is the **maximum likely difference** between the sample statistic and the population parameter at a **given confidence level**.
  - $\mathbf{Margin\ of\ Error = z^* \cdot SE}$
  - So:
    - $\mathbf{Confidence\ Interval = \bar{x} \pm Margin\ of\ Error}$
- **Example:**
  - Suppose:
    - $\mathbf{\bar{x} = 70\ \&\ SE = 2}$
    - For confidence level $\mathbf{95\% \rightarrow z^* = 1.96}$
  - Then:
    - $\mathbf{ME = 1.96 \cdot 2 = 3.92}$
    - $\mathbf{CI = 70\ \pm 3.92 = (66.08, 73.92)}$
  - Interpretation: **We are 95% confident that the true population mean μ lies between 66.08 and 73.92**.

2.3 Example Case Study – 1.

- **Scenario:**
  - **QuickNet**, a broadband internet service provider, is analyzing customer retention. The marketing team wants to assess how well current retention strategies are working, and to forecast future churn.
  - They conducted a survey of **800 customers** from last quarter and found that **68 customers canceled** their **subscription**.
  - The executive team asks:
    - **"What is the range in which the true churn rate likely lies?"**
    - **"How confident are we in this estimate?"**
- The **first Question we should be asking is**:
  - What distribution Models this scenario?
  - What is the Best Estimator?

2.3.1 What Distribution Models This Scenario?

- The **parameter** we are interested in is the **population churn rate**, denoted: **p=true proportion of customers who churn**
  - Which distribution models this scenario:
    - The **event churn is binary**: A customer **either churns (1) or does not churn (0).**
- **Model: Binomial Distribution:**
  - $\mathbf{X \sim Binomial(n = 800, p)}$
  - where:
    - $\mathbf{X}$: **number of churned customers**; $\mathbf{n = 800}$: **number of customers sampled**;
    - $\mathbf{p}$: **probability that a customer churns (what we want to estimate).**
- **Best Estimator:**
  - The **sample proportion** $\hat{p}$ is the **unbiased estimator of the population proportion p**:
    - $\hat{p} = \frac{x}{n} = \frac{68}{800} = 0.085$
    - So, $\mathbf{0.085}$ **or** $\mathbf{8.5\%}$ is **our point estimate of the churn rate**.
  - "If this **sample is representative**, then we estimate that around **8.5% of all QuickNet customers are likely to churn**."
  - This number is **not just a description** of the sample — it's a **statistical estimate** of the **overall churn rate** for the full customer base, based on the evidence you collected.
    - It's your **best guess for p.**
    - But it's still **an estimate** — so we add a **confidence interval** to express **uncertainty around it**.

2.3.2 Why Use a Normal Approximation?

- Working **directly with binomial probabilities (e.g., computing the confidence interval)** is difficult**, especially for large n.**
  - So, we approximate the **sampling distribution** of $\hat{p}$ with a **normal distribution**,
    - using the **Central Limit Theorem (CLT)**.
  - Conditions to **Use Normal Approximation**:
    - $n\,\hat{p} \geq 10$
    - $n(1 - \hat{p}) \geq 10$
  - Check:
    - $n\,\hat{p} = 800 \cdot 0.085 = 68 \geq 10$
    - $n(1 - \hat{p}) = 800 \cdot 0.915 = 732 \geq 10$
  - These conditions are satisfied, so we **can model the sampling distribution of $\hat{p}$** as:
    - $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$
  - Since **p is unknown**, we **approximate the standard error using $\hat{p}$**:
    - $SE_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

2.3.3 Step – by – Step Calculation.

- **Step 1: Standard Error (SE):**
  - $\mathbf{SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.085 \cdot 0.915}{800}} \approx 0.00986}$
- **Step 2: Critical Value:**
  - At **95% confidence**, the **critical z – value**:
    - $\mathbf{z^* = 1.96}$
- **Step 3: Margin of Error (ME):**
  - $\mathbf{ME = z^* \cdot SE = 1.96 \cdot 0.00986 \approx 0.0193}$
- **Step 4: Confidence Interval:**
  - $\mathbf{\hat{p} \pm ME = 0.085 \pm 0.0193 \Rightarrow (0.0657, 0.1043)}$
- **Business Interpretation:**
  - We are **95% confident** that the **true customer churn rate** lies **between 6.6% and 10.4%**.

2.4 Example Case Study – 2.

- **Context:**
  - **HelpPro Inc.** tracks customer service performance closely. One **key performance indicator (KPI)** is **average call duration**, because:
    - Longer calls → more cost per ticket
    - Shorter calls → more efficiency, but may hurt customer satisfaction
  - Their **internal target is 15 minutes per call**.
  - The **support manager** wants to **estimate the true average duration based on recent performance**, and see if **they're still meeting that target**.
- **Business Question:**
  - **"Is the average customer support call duration longer than our target of 15 minutes?"**
- **Data Collection:**
  - Sample size: **n = 100 customer support calls**
  - Sample mean: $\bar{x} = 16.2$ **minutes**
  - Known population standard deviation: $\sigma = 4.8$ **minutes**
  - Confidence level: **95%**

2.4.1 Step – by – Step Calculation.

- We will use the z – distribution, because the population standard deviation is known and n is large.
  - **Step 1: Compute the Standard Error:**
    - $\mathbf{SE = \frac{\sigma}{\sqrt{n}} = \frac{4.8}{\sqrt{100}} = \frac{4.8}{10} = 0.48}$
  - **Step 2: Determine the critical value:**
    - At 95% confidence, from the **standard normal distribution**,
      - $\mathbf{z^* = 1.96}$
  - **Step 3: Compute the Margin of Error:**
    - $\mathbf{ME = z^* \cdot SE = 1.96 \cdot 0.48 = 0.9408}$
  - **Step 4: Compute the Confidence Interval:**
    - $\mathbf{\bar{x} \pm ME = 16.2 \pm 0.9408 \Rightarrow (15.26, 17.14)}$
- **Business Interpretation:**
  - We are 95% confident that the **true average support call duration lies between 15.26 and 17.14 minutes**.

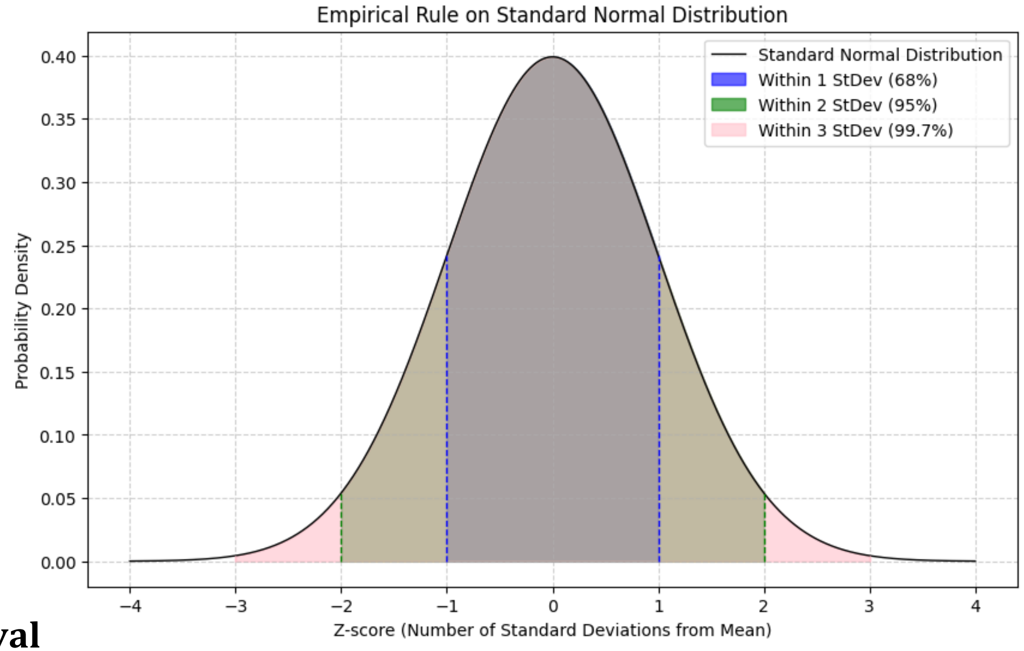
2.5 Why not Go for 99% Confidence?

- **Higher confidence sounds better … but there's a trade off:**
  - So, when estimating the **population mean μ**, the **confidence interval** becomes:
    - $\bar{x} \pm z^* \cdot SE$ **where** $SE = \frac{\sigma}{\sqrt{n}}$
  - The **z^* values** correspond **directly to the SD ranges:**

Table: Trade – Off : Confidence vs. Precision.

Confidence Level	Critical Value z^*	Approx SD Range	Margin of Error	CI Interval width
68%	1.00	±1 SE	Very Small	Very narrow
90%	1.645	±1.6 SE	Smaller	Narrower
95%	1.96	±2 SE	Medium	Moderate
99%	2.576	±2.6 SE	Big	Moderately Big
99.7%	3	±3 SE	Larger	Wider

As we increase the confidence level, the margin of error (ME) also increases.



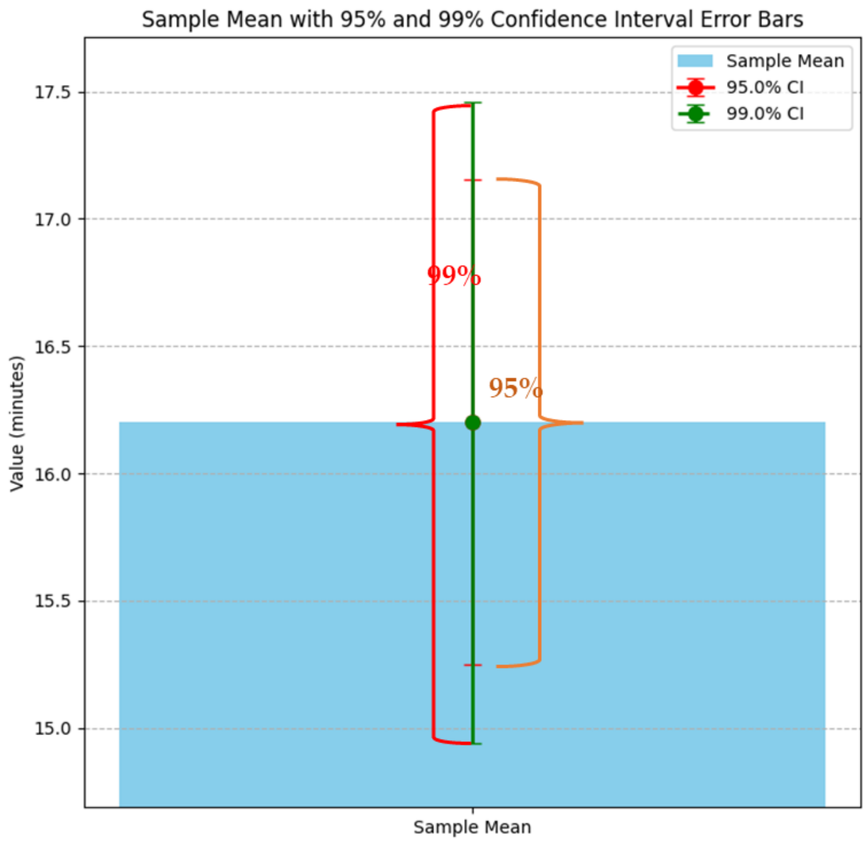
- **Summary:**
  - **Higher confidence → Higher margin of error → Wider interval**
  - **Lower confidence → Lower margin of error → Narrower interval**
- **The trade – off:**
  - **Do you want to be surer (95 to 99%) or do you want to be more precise (narrower range)?**

2.5.1 Why not Always Choose 99%?

- **More confidence = more uncertainty range**
- **Wider intervals** may be **less useful** for **decision-making**
- **Less actionable** in business **when precision matters (e.g., budgeting, planning)**
- May require **larger sample sizes** to keep ME small at higher confidence
  - Example: **HelpPro Call Duration** (**Sample Mean** $= \mathbf{16.2}, \boldsymbol{\sigma} = \mathbf{4.8}, \mathbf{n} = \mathbf{100}$)

Confidence	Margin of Error	Confidence Interval
95%	$\pm\mathbf{0.94}$	$(\mathbf{15.26}, \mathbf{17.14})$**minutes**
99%	$\pm\mathbf{1.24}$	$(\mathbf{14.96}, \mathbf{17.44})$**minutes**

  - **At 99%,** the interval **includes** the 15-min threshold
    - harder to make confident decisions about exceeding the target
- **Higher confidence isn't always better.**
  - **Balance confidence with the need for precision and actionability.**



3. When you do not Know Population σ?

3.1 Example Case Study.

- **Scenario: Estimating Average Delivery Time.**
  - A **logistics analyst** wants to estimate the **average delivery time for packages** sent last week.
  - They take a random sample of $\mathbf{n = 12}$ **delivery times**:
    - $\bar{\mathbf{x}} = \mathbf{42.5}$ **minutes**, $\mathbf{s = 5.4}$ **minutes**
  - Target: **Estimate the true average delivery time μ with 95% confidence**.
- **Let's Build a Confidence Interval:**
  - Recall the formula for CI when $\boldsymbol{\sigma}$ **is known:**
    - $\bar{\mathbf{x}} \pm \mathbf{z}^* \cdot \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$
  - **But do we know σ ?**
    - No, we only have the sample **standard deviation** $\mathbf{s = 5.4}$
    - Also, our sample is **small:** $\mathbf{n = 12}$
- This means we can not use the **z – distribution, why?**

3.1.1 Problem with Plugging s into a Z-Based CI.

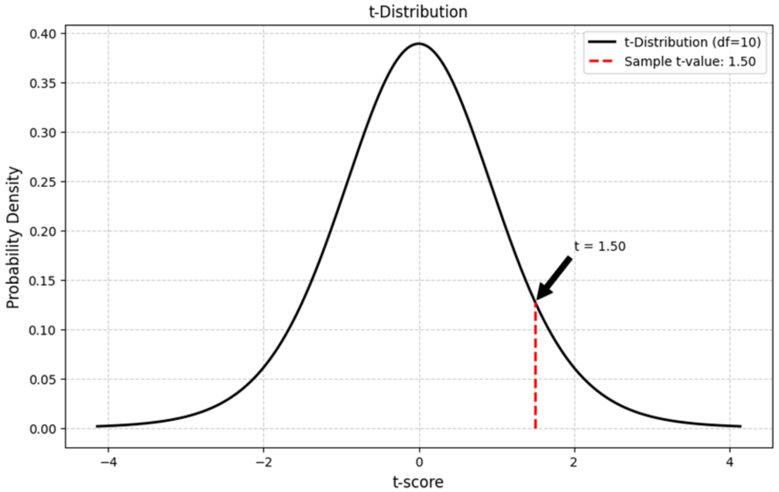
- If you plug **s** into the **z – distribution formula:**
  - $\bar{x} \pm z^* \cdot \frac{s}{\sqrt{n}}$
- You are:
  - Using a **noisy estimate of the standard deviation**.
    - This **estimation** use $n - 1$ in its calculation,
      - thus **reflects the noise and additional variability** (**Bessel's correction**).
  - If we apply the **z – based critical value**, which assumes **knowledge of true σ.**
- Thus, **if we do not adjust the critical value** to account **for that uncertainty**,
  - we may underestimate the total uncertainty.
- That's why we switch to the **t-distribution**, which adjusts **for this extra uncertainty.**

3.2 The t-Distribution.

- **aka t – student distribution:**
  - The **t-distribution** (also called **Student's t-distribution**) is a
    - **family of continuous probability distributions** used when **estimating population parameters**
  - When do we use it?
    - **We are estimating the population mean μ**
    - **The population standard deviation σ is unknown**
    - **You have to use the sample standard deviation s**
    - **And your sample size is small** $\mathbf{i.e.\ (n < 30)}$
  - The **t-distribution adjusts for the added uncertainty** introduced by **using s instead of σ**.
    - **t – statistic Formula:**
      - $\mathbf{t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}}$

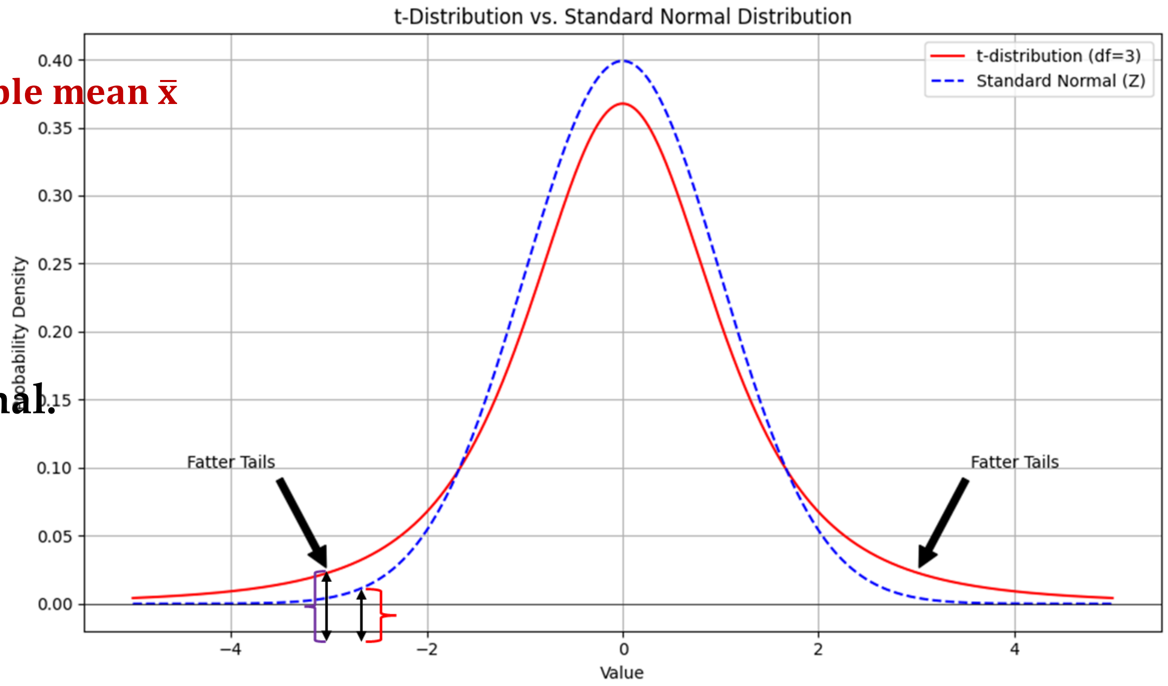
3.2.1 Characteristics of t – distribution.

- **Mean:**
  - The mean of the **t-distribution** is **0**.
  - This is analogous to the standard normal distribution (z distribution), which also has a mean of **0**.
- **Symmetry:**
  - The **t-distribution** is **symmetric around its mean**, similar to the **normal distribution**.
- **Variance:**
  - The variance of the **t-distribution** is **greater than 1** for small **sample sizes** but approaches **1** as the **sample size increases**.



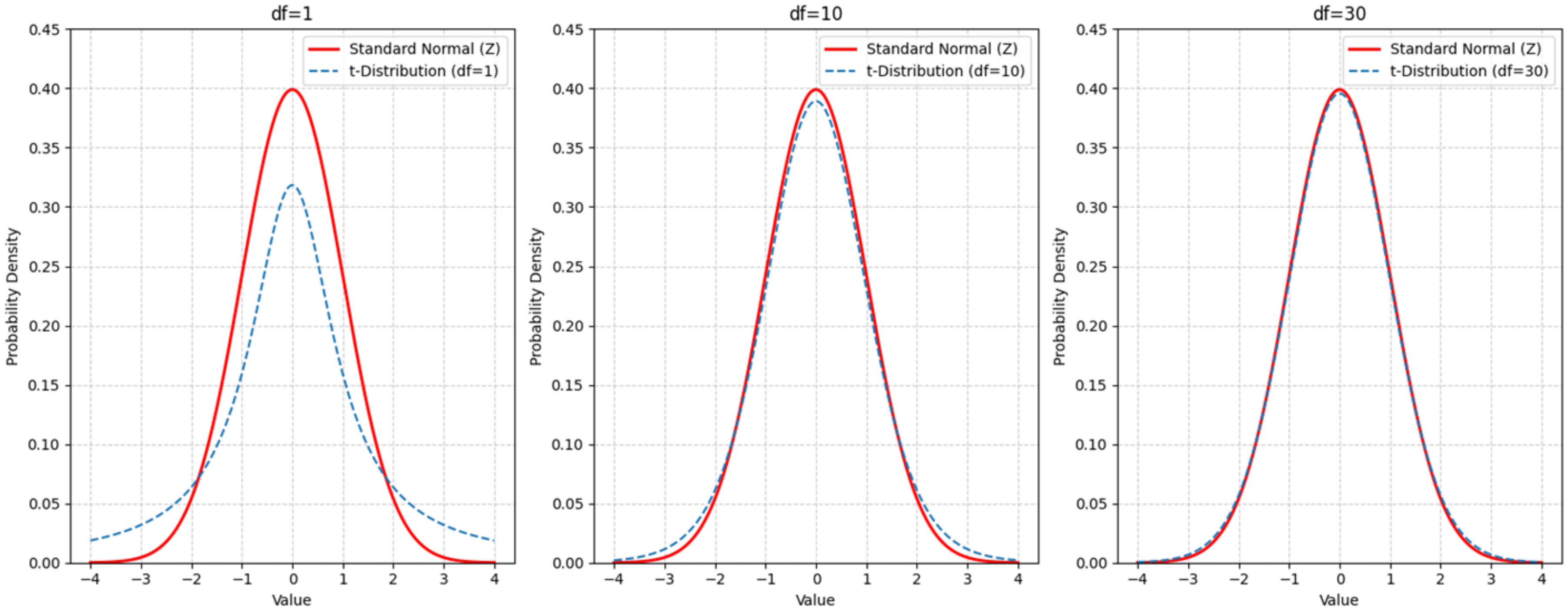
3.2.2 Characteristics of t – distribution.

- The **shape** of the t-distribution depends on the **degrees of freedom (df)**, which are typically calculated as:
  - **$df = n - 1$**
  - **where:**
    - **n = sample size**
    - **subtracting 1 accounts for the estimation of the sample mean $\bar{x}$**
- **Why it Matters?**
  - **Lower df(small samples) → heavier tails**
    - Reflects **greater uncertainty**
    - More probability in the extremes
  - **Higher df (larger samples)→ t curve approaches normal.**
    - More stable estimates
    - Less uncertainty



3.2.3 Characteristics of t – distribution.

- **Asymptotic Behavior**:
  - As the **degrees of freedom increase, the t-distribution approaches the standard normal distribution**.
  - This means that for **large sample sizes**, the **t-distribution and z-distribution** are **nearly indistinguishable**.



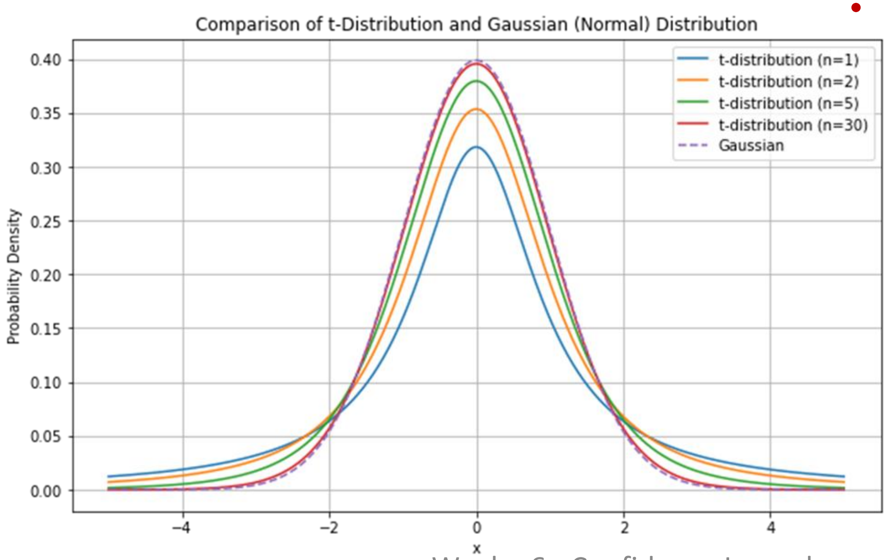
3.3 General Guideline: When to Use z vs. t Distribution.

Use the Z – distribution:

- When:
  - The population standard deviation σ is known.
  - Applies regardless of sample size n
  - **Formula:**
    - $\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$

Use the t –distribution:

- When:
  - The population standard deviation σ is unknown (which is common).
  - You use the sample standard deviation s instead
  - **Formula:**
    - $\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$ (with $df = n - 1$)



3.3.1 How sample size affects your choice.

Sample Size n	**Recommendation**	**Why?**
Small $n < 30$	Use t - distribution	More accurate, t – distribution has heavier tails to account for greater uncertainty.
Large $n \geq 30$	Z distribution is a reasonable approximation	t – distribution ≈ z – distribution as n increases by CLT.
	Still safer to use t - distribution	Technically more correct, even for large n and is always **more accurate**, because it properly accounts for the uncertainty in estimating variability using s.

Optional: Degree of Freedom – Intuition.

- Think of degrees of freedom as the number of values in a calculation that are free to vary.
- Example:
  - For any distribution with unknown population mean μ, and sample mean of 5, what could be the missing sample value below:

$\mathbf{i}$	$\mathbf{x_i}$	$\mathbf{x_i - \bar{x}}$
1	6	6-5 =2
2	4	4-5 = -1
3	?	

Constrained imposed by sample mean:
Sum of Deviations from the mean: By definition, the sum of the deviations of each observation from the mean must equal zero:

$$\sum(\mathbf{x_i} - \bar{\mathbf{x}}) = 0$$

Because of this constrained there is only one possible choice for our third observation i.e.

$$(\mathbf{x_1} - \bar{\mathbf{x}}) + (\mathbf{x_2} - \bar{\mathbf{x}}) + (\mathbf{x_3} - \bar{\mathbf{x}}) = 0$$

Rearrange the above algebraically:

$$\mathbf{x_1} + \mathbf{x_2} + \mathbf{x_3} - 3\bar{\mathbf{x}} = 0$$

$$\mathbf{x_1} + \mathbf{x_2} + \mathbf{x_3} = 3\,\bar{\mathbf{x}}$$

Let's find x_3:

$$6 + 4 + x_3 = 3\,\bar{x}$$

$$6 + 4 + x_3 = 3 \times 5$$

How many values of x_3 can satisfy the above equation?
Therefore in this condition third observation is not independent, once we know the mean and two observation. Thus for this example:

$$\mathbf{n = 3, we\ have\ (3 - 1) = 2\ degrees\ of\ freedom}$$

Thank You