# HCAI5DS02 – Data Analytics and Visualization.
# Lecture – 05
# Introduction to Statistical Modeling
## Quantifying Uncertainty with Parameter Estimations.

Siman Giri

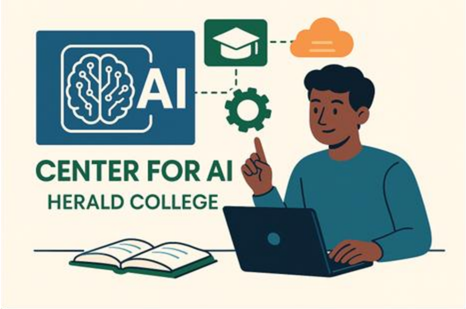



image generated via copilot.

HERALD COLLEGE | ing CENTER FOR AI.

# 1. What are Statistical Models?

## {Modelling the Randomness.}

# 1.0 What is Data Generating Process (DGP)?

- The **Data Generating Process (DGP)** refers to the **underlying mechanism**
  - often assumed to be probabilistic by which data is **produced** in the real world.
- It includes all the **random variables, relationships, and parameters** that define how observed data comes to be.
- Formal View:
  - A **DGP** is a **mathematical abstraction**: $\mathbf{X \sim f(x; \theta)}$
    - Where:
      - **X → observed data**
      - **f → probability distribution**
      - **θ → unknown parameters.**
- The **statistical model** is our attempt to represent this process, usually by specifying:
  1. The **form of the distribution** (e.g., Normal, Binomial, Poisson)
  2. The **assumptions** (e.g., independence, identically distributed)
  3. The **parameters** that govern the shape and behavior

# 1.1 What is Statistical Modeling?

- **Definition:**
  - A **statistical model** is **mathematical representation** of a **real-world process** **that describes how data is generated** using **random variables and unknown parameters** **governed** by **probability distributions.**
- **Key Components:**
  - **Random Variables:**
    - Quantities that vary from one observation to another e.g. number of clicks, time to purchase etc.
  - **Probability Distributions:**
    - Describes the likelihood of different outcomes e.g. Binomial, Normal, Poisson etc.
  - **Parameters:**
    - Unknown constants that shape the behavior of the distribution e.g. mean, variance, probability of success.
- Think of **statistical model** as a way to break down data into:
  - **Data = Structure (Model) + Randomness(Noise)**
    - **Structure (Model):** The expected or systematic part (e.g. average behavior, trend)
    - **Randomness (Noise):** The unpredictable variation around the expected value.

# 1.1.1 Real – World Examples:

- **Why it matters?**
  - **Statistical Models** help us:
    - Understand **patterns in data.**
    - Estimate unknown quantities.
    - Predict future outcomes.
    - Make informed **business decisions**.

| Scenario | Random Variable | Model | Parameter(s) |
|---|---|---|---|
| Email Click - through | $X = \#$ of clicks | Binomial | $p =$ Click – through rate |
| Website Traffic | $X =$ visits/hour | Poisson | $\lambda =$ Avg. visitis/hr |
| Time Until purchase | $T =$ waiting time | Exponential | $\lambda =$ Conv. rate |
| Purchase amount | $Y =$ purchase value | Normal | $\mu, \sigma$ |

*"Which scenario above would you expect more variability in? Why?"*

# Good to Know: Deterministic Model.

- **Definition:**
  - A **deterministic model**, is a **mathematical model** in which the **outcome** is completely **determined** by the **input values**, with **no randomness or uncertainty involved**.
  - The **same input** will always produce **the same output**.
- **Mathematical Form:**
  - $\mathbf{Y} = \mathbf{f}(\mathbf{X})$
  - Where:
    - $X =$ input(s)
    - $f =$ known function or rule
    - $Y =$ output, exact and predictable.
- **Q: Is Linear Regression Deterministic or Statistical Model?**

# Good to Know: Deterministic Model.

- **Linear regression has a deterministic structure, but it is a *statistical model***
  - because it **explicitly includes randomness (noise) in the outcome**.
- **Linear Regression Equation**:
  - $\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X} + \boldsymbol{\epsilon}$
    - $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X} \rightarrow$ Deterministic part – the systematic relationship.
    - $\boldsymbol{\epsilon} \rightarrow$ **Random error** – captures unexplained variability (noise).
- Thus:
  - The **model prediction**: $\hat{\mathbf{Y}} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X}$ is **deterministic.**
  - But the actual outcome $\mathbf{Y}$ is random due to $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$ making it ***statistical model***.

# 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 |
|---|---|---|
| $\mu$ | Population Mean | True average purchase amount. |
| $\sigma^2$ | Population Variance | Variability in customer spending. |
| p | Population proportion | True click through rate. |
| $\lambda$ | Rate parameter | Avg. visits. |

*"Why do we need parameters?"*

# 1.2.1 Why do we need Parameters?

- **Case Study: “The Email Campaign Dilemma”:**
  - **Context:**
    - You are a data analyst at **InsightX Marketing**, and your team is planning an **email campaign** to promote a new product. In your last campaign, you sent **10,000 emails** and recorded a **5% conversion rate**, where each successful conversion generated **Rs 1,200** in revenue. The marketing team asks you to help answer three key questions using **data-driven reasoning**:
  - **Questions:**
    1. **Forecast** the expected revenue from the current campaign using the historical conversion rate.
       - What is the **expected revenue**?
       - What is the **standard deviation (risk) of revenue** due to **conversion randomness**?
    2. **Simulate outcomes** from the campaign to show a range of possible revenues.
       - What are **the 5th percentile**, **median** and **95th percentile revenue scenarios**?
    3. The team is considering **three alternatives** to **increase revenue**:
       - Send 12, 000 emails (same content and conversion rate).
       - Improve the email design to increase the conversion rate to 6% with 10, 000 emails.
       - Target a more engaged list: 8,000 emails but with a 7% conversion rate.
         - **Which option would you recommend?**
         - **Compare expected revenue and variability.**

# 1.2.2 Solutions:

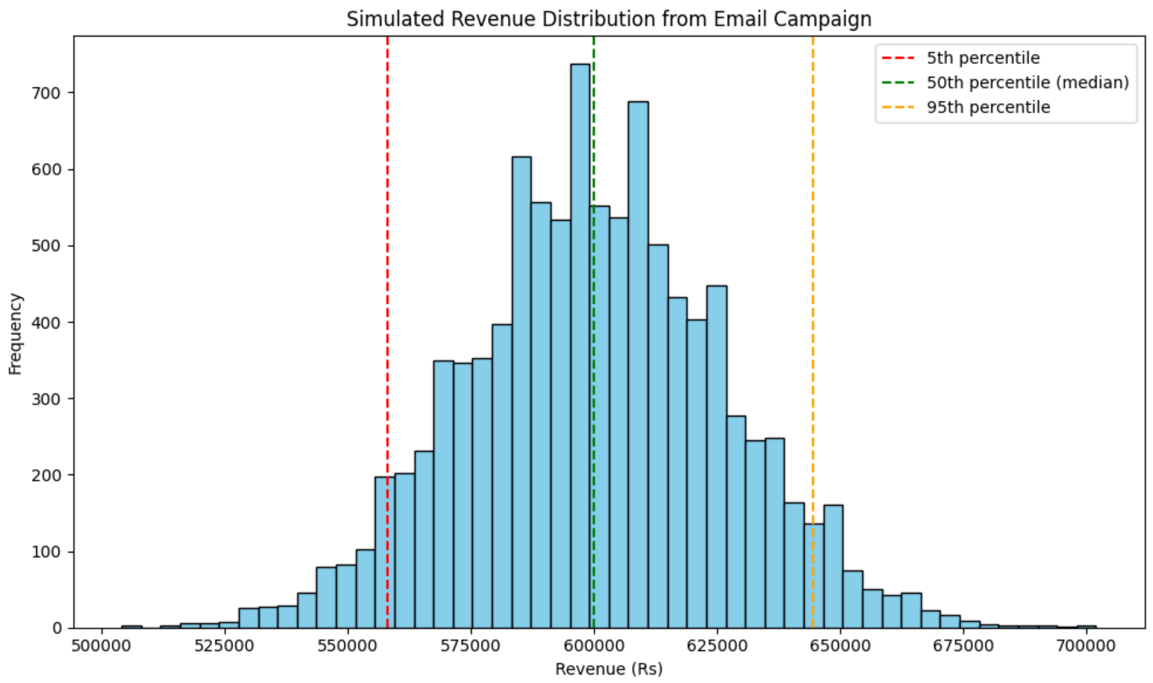
- Before we solve the case, let's first discuss **how we model this problem statistically.**
  - **What are we trying to model?**
    - We want to model the **number of people who convert (make a purchase)** after receiving an email.
  - **Selected Model: Binomial Distribution.**
    - **$X \sim \text{Binomial}(n, p)$**
      - **$X$: number of conversions (random variable)**
      - **$n$: number of emails sent (fixed, known)**
      - **$p$: probability of conversion per email (unknown parameter, estimated from past data)**
  - **Why Binomial?**
    - The **binomial model fits** because:

| Condition: | Real world match: |
|---|---|
| Fixed number of trials | We send a fixed number of emails $n = 10{,}000$ |
| Each trial is independent | One person's decision doesn't affect another's |
| Only two outcomes per trial | Either someone converts or doesn't |
| Constant probability of success $p$ | We assume a stable conversion rate from past data |

# 1.2.2 Solutions:

1. **Forecast Expected Revenue and Risk:**
   - **Expected Conversions:**
     - $= \mathbf{n \cdot p = 10,000 \times 0.05 = 500.}$
   - **Expected Revenue**
     - $= \mathbf{500 \times 1,200 = 600,000.}$
   - **Variance (Conversions)**
     - $= \mathbf{np(1-p) = \ 10,000 \cdot 0.05 \cdot 0.95 = 475.}$
   - **SD:**
     - $= \mathbf{\sqrt{475} \cdot 1,200 = 26,100.}$

2. **Simulate Revenue Outcomes:**
   - **Simulate $\mathbf{X \sim Binomial(10,000, 0.05)}$, compute:**
     - $\mathbf{Worst\ Case\ (5\%) \approx Rs, 549,000.}$
     - $\mathbf{Median\ (50\%) \approx 600,000.}$
     - $\mathbf{Best\ Case\ (95\%) \approx 651,000.}$



# 1.2.2 Solutions:

3. **Compare Strategic Alternatives:**

| Option | Emails | Conversion Rate | Expected Revenue | SD (Risk) |
|---|---|---|---|---|
| A | 12,000 | 5 | 720,000 | 28,540 |
| B | 10,00 | 6 | 720,000 | 26,832 |
| C | 8,000 | 7 | 672,000 | 23,904 |

- **Interpretation:**
  - **A and B give same revenue**, but **B has lower risk**, so **improving content** is more efficient.
  - **C gives lower revenue**, but may be **cheaper** if targeting a **smaller list** saves cost.

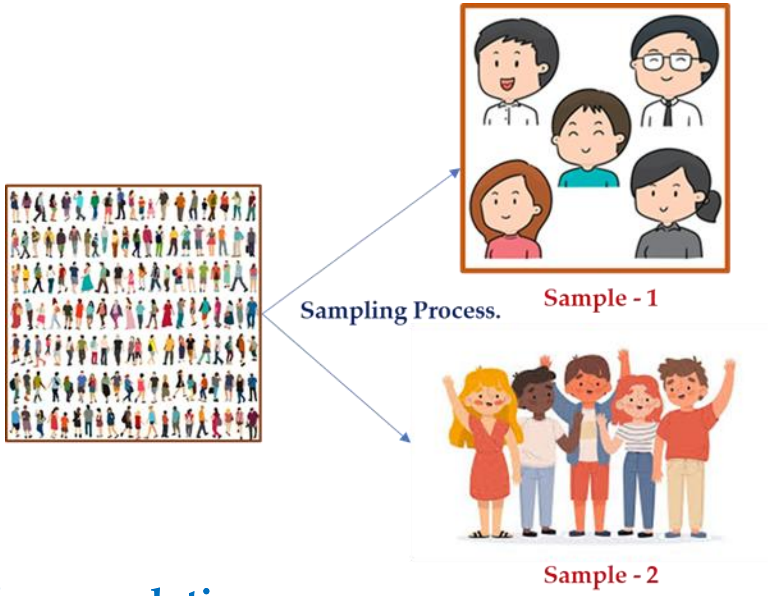
# 1.2.3 But There's a Catch …

- **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 $\mu$ |
| $s^2$ | Sample Variance | Estimates $\sigma^2$ |
| $\hat{p}$ | 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. Estimating Parameters from Data.

# 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.

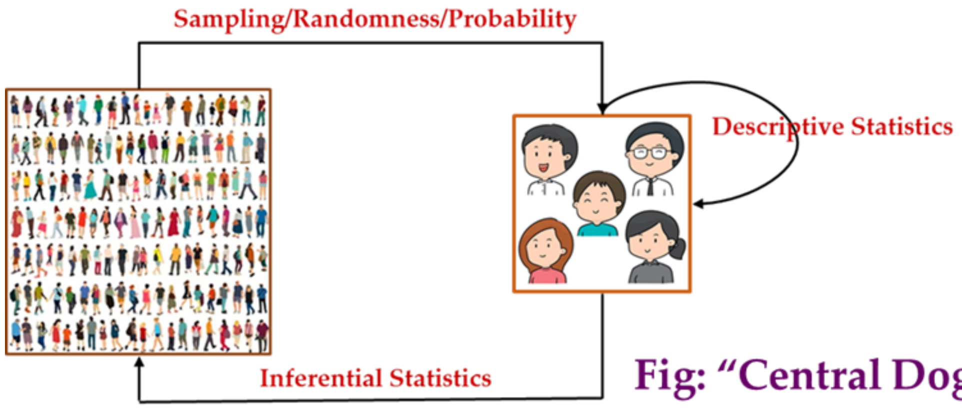
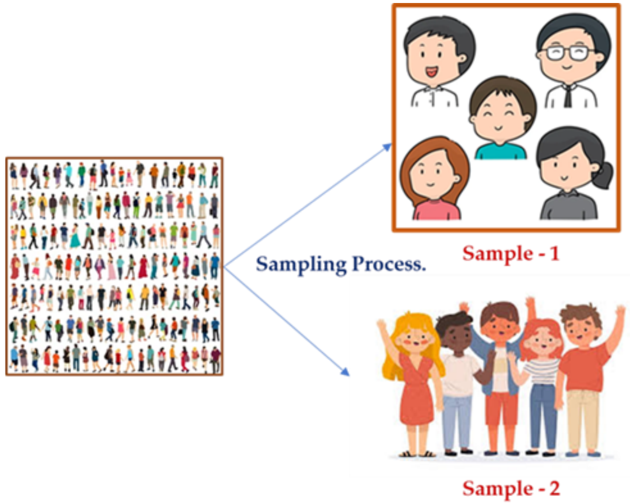



Fig: "Central Dogma" of Statistics.

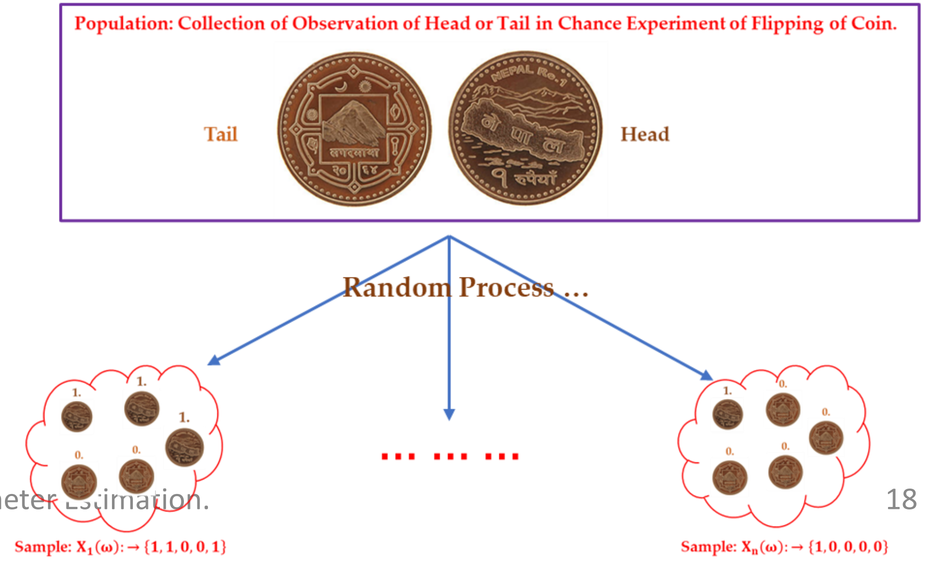


# 2.2 Estimator and Estimate.

- **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 $\mu$
    - $\hat{p}$: estimates the true proportion $p$
    - $s^2$: estimates population variance $\sigma^2$
- **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:
    - $\hat{p} = \frac{50}{1000} = 0.05(\text{estimate})$

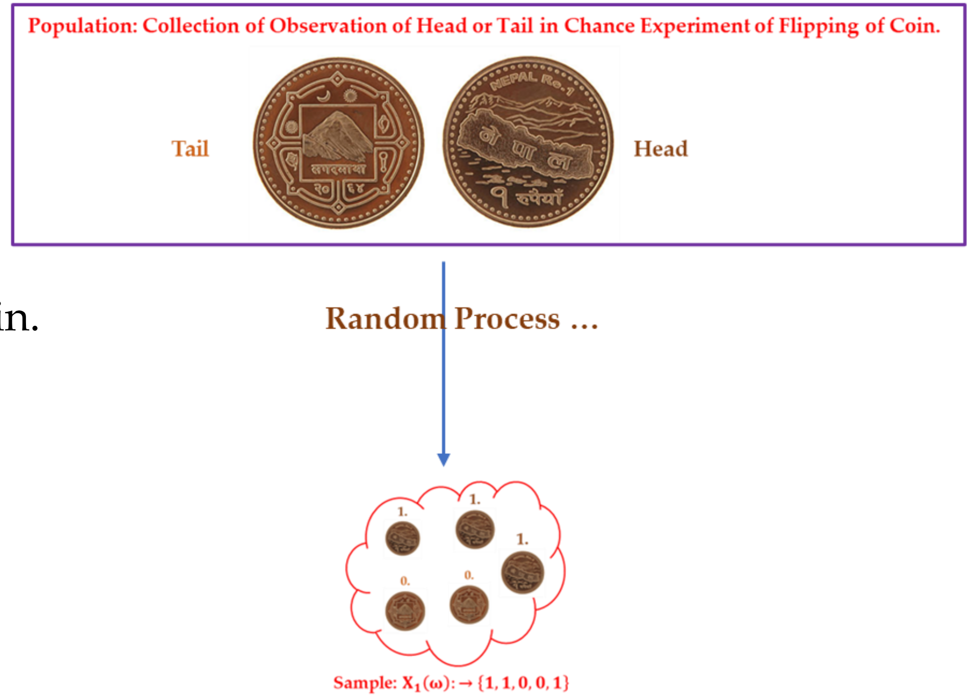
# 2.3.1 Estimator as a Random variable.

- **Core Idea:**
  - An **estimator** (e.g. ***sample mean*** $\overline{X}$) is a **function of sample data**, and **sample data** comes from **a random process.**
  - Therefore, the **estimator itself is random** – it varies **from sample to sample.**
- **Why it is a Random Variable?**
  - An **estimator** $\hat{\theta}$ is a function of the entire sample:
    - $\hat{\theta}: \Omega \to \mathbb{R}$ **where** $\hat{\theta}(\omega) = g\big(X_1(\omega), \dots, X_n(\omega)\big)$
    - **The estimator function g** could be the sample mean, proportion, or any summary, depending on the assumed underlying distribution.



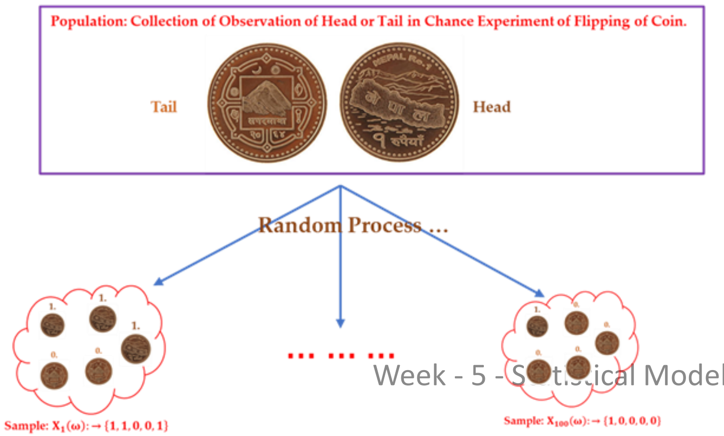
# 2.3.2 Back to Coin Flip Example.

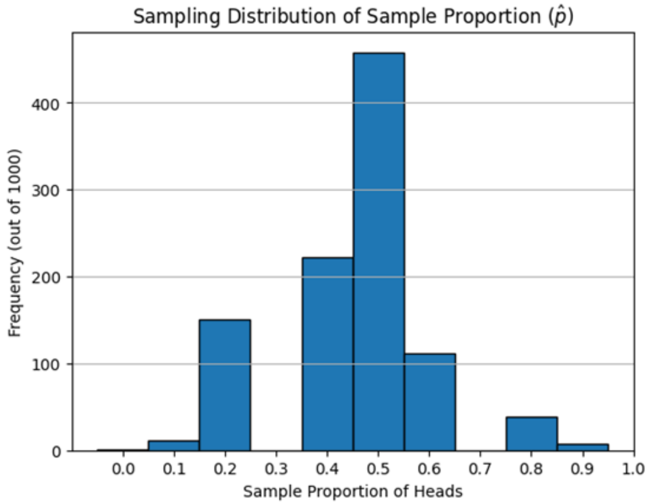
- Let $\mathbf{X_i} = \mathbf{1}$ **if heads**, **0 if tails**.
  - Estimator:
    - $\hat{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X_i}$ { assumed underlying distribution → Binomial.}
- Estimation from data:
  - Let's take samples of size $\mathbf{n = 5}$**-coin** flips.
- We compute the sample proportion of heads for sample:
  - $\mathbf{X_1\{1, 1, 0, 0, 1\} = \frac{1}{5}\{3\} = \frac{3}{5} = 0.6}$■
- Population parameter: fair coin → 0.5 subjective belief about flip of a coin.
  - **Is your estimate close to true value.**



# 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 $\hat{p}$** 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* $\rightarrow X_1^1, X_2^1, \ldots, X_n^1 \rightarrow$ Compute $\bar{X}^1$
    - *Sample 2* $\rightarrow X_1^2, X_2^2, \ldots, X_n^2 \rightarrow$ Compute $\bar{X}^2$
    - … … …
    - *Sample m* $\rightarrow X_1^m, X_2^m, \ldots, X_n^m \rightarrow$ Compute $\bar{X}^m$
- **Now you have collection of means:** $\bar{\mathbf{X}}^1, \bar{\mathbf{X}}^2, \ldots, \bar{\mathbf{X}}^m$
- These form the **sampling distribution of the estimator** $\bar{\mathbf{X}}$.

# 2.4.1 The Central Limit Theorem.

- We now turn our attention to one of the most fundamental results in statistics:
- The remarkable **The Central Limit Theorem.**

**The Central Limit Theorem**

Let $X_1, \dots, X_n$ be a **random sample** from a **distribution** with **finite mean $\mu$** and **finite variance $\sigma^2$**.
For $\bar{X}$ denoting **the sample mean**, if $n$ is sufficiently large then:

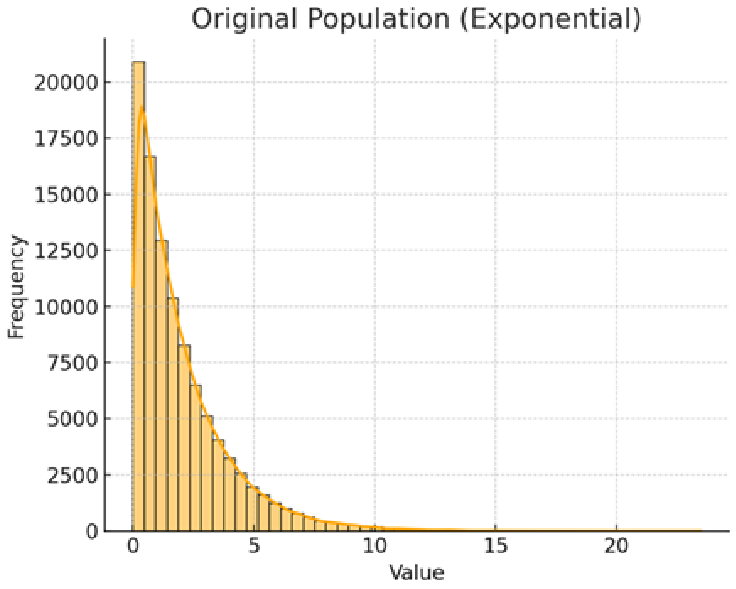
$$\bar{X} \ \widetilde{\text{approx}}.\ N\left(\mu, \frac{\sigma^2}{n}\right)$$

Where $\widetilde{\text{approx}}$ denotes "**approximately distributed as**".

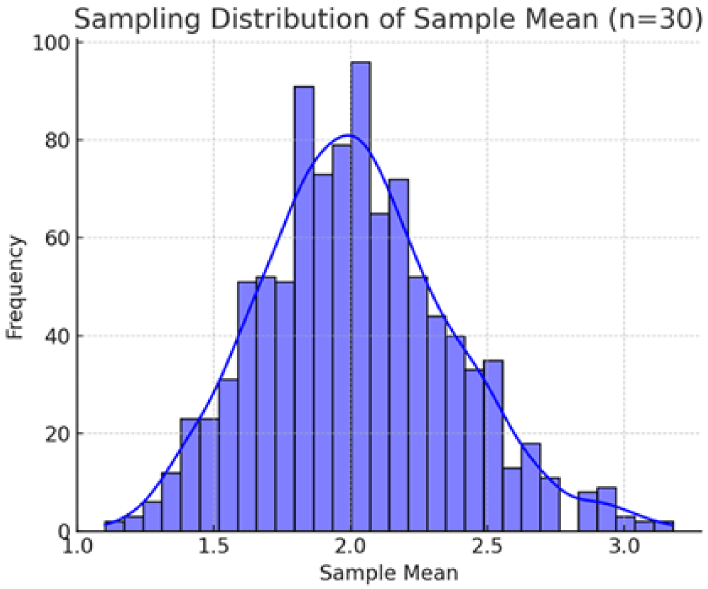
- Normally, a sample size of approximately $n = 30$ is considered to be **sufficiently large**.

# 2.4.1.1 CLT – Visual Intuition.

- Regardless of the **original population distribution** (it can be ***skewed, uniform, discrete***, etc.),
  - the **sampling distribution of the sample mean** will approach a *Normal distribution*
  - as the **sample size n** becomes large ***as long as*** **the population** has:
  - “a **finite mean (expectation)** and a **finite variance**.”




The original population follows an **Exponential distribution** highly skewed and non-Normal.

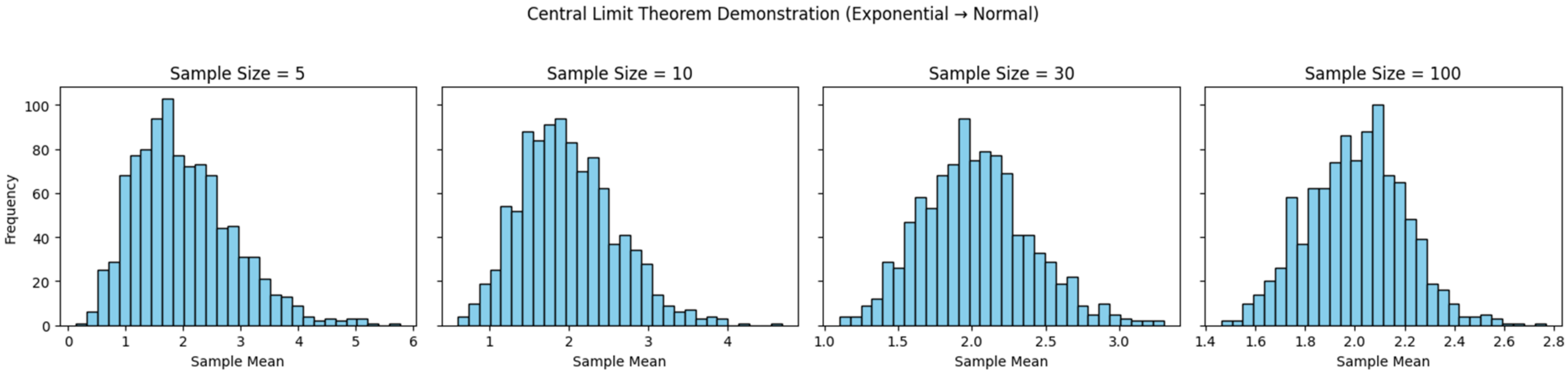



The distribution of **sample means** (each from 30 observations) is approximately **Normal**, centered around the true mean.

# 2.4.1.2 Misconceptions About the CLT.

1. **CLT says the population becomes Normal.**
   - **Wrong:** If I take enough samples, the population will become Normally distributes.
   - **Truth:** The Population distribution does not change.
   - CLT says the distribution of the sample mean becomes approximately Normal not the population itself.
2. **CLT applies no matter the sample size.**
   - **Wrong:** " CLT works even for very small samples."
   - **Truth:** The approximation to Normality improves with larger sample sizes.
   - For highly skewed or heavy tailed populations, you often need $\mathbf{n \geq 30}$ or more.
3. **CLT means each sample looks Normal.**
   - **Wrong:** "The data in each sample look Normal."
   - **Truth:** The sample mean (not the raw data) is approximately Normal.
   - Each individual sample may still look like the original skewed or non – Normal population.
4. **CLT only applies to the mean.**
   - **Wrong:** "CLT applies to any statistic."
   - **Truth:** CLT in its classic form applies to sums or averages of i.i.d variables.
   - Other statistics (like medians, variances) may not follow a Normal distribution unless special conditions hold.
5. **CLT doesn't need independence.**
   - **Wrong:** "CLT works even if the samples are dependent."
   - **Truth:** The classic CLT assumes i.i.d variables.
   - Dependence can violate CLT or require advanced versions (e.g. for time series).
6. **CLT implies exact Normality.**
   - **Wrong:** "CLT makes the sampling distribution exactly Normal."
   - **Truth:** CLT gives an approximation to Normality.
   - The distribution becomes closer to Normal as $n \rightarrow \infty$, but it's not perfect for finite n.

# 2.4.1.3 Convergence of Sample Means to Normality.




- **What It Demonstrates:**
  - You start **from a highly skewed population**.
  - For *small* $n(e.g.\,5)$, the **sampling distribution** is **still skewed**.
  - As **n increases**, the **distribution of the sample mean** becomes **more normal**.

# 3. Understanding the Quality of Estimators.

## {What makes an estimator “good”?}

# 3.1 What makes an estimator "good"?

- **Unbiasedness:**
  - As **estimator $\hat{\theta}$** is unbiased if:
    - $\mathbb{E}[\hat{\theta}] = \theta$
  - Example:
    - Sample mean $\bar{X}$ is **unbiased estimator** of population mean μ.
- **Consistency:**
  - As sample size increases, the estimator **converges in probability** to the true parameter:
    - $\hat{\theta}_n \xrightarrow{P} \theta$ **as** $n \to \infty$
  - **Larger samples → more accurate estimates**.
- **Efficiency:**
  - Among all **unbiased estimators**, the one with **lowest variance** is preferred.
  - Example:
    - The sample mean $\bar{X}$ is more efficient than the sample median for estimating $\mu$ in a Normal distribution.
- **Sufficiency:**
  - An estimator is sufficient if it captures all information about the parameter contained in the sample.
  - Formally, $T(X)$ is sufficient **for $\theta$** if:
    - $P(X|T(X), \theta) = P(X|T(X))$
  - Helps in reducing data without losing inferential power.

# 3.2 Bias and Variance of Estimator.

## Bias

- **Bias** measures how far the **average estimate** of a model is from the **true value** of the parameter.
  - $\mathbf{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$
- High bias → **systematic error**
  - (the estimator is consistently wrong).
- Low bias → estimator is centered around the true parameter.

## Variance

- Varaince measures **how much the estimator**
  - fluctuates **around** its **expected value**
    - across **different samples**.
    - $\mathbf{Var}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right]$
- **High variance estimator** is sensitive to **sample fluctuations.**
- **Low variance estimator** is stable **across different samples.**

**Remember Mean Squared Error (MSE):**

- MSE can be decomposed into:
  - $\mathbf{MSE}(\hat{\theta}) = \underbrace{\mathbf{Bias}(\hat{\theta})^2_{\text{systematic error}}} + \underbrace{\mathbf{Var}\,(\hat{\theta})_{\text{estimation error}}}$
- This is crucial for evaluating estimators in practice — especially in predictive modeling.

# 3.2.1 Systematic and Estimation Error.

## Systematic Error - Bias$^2$:

- **Systematic error** refers to
  - consistent, repeatable error that occurs because
  - the estimator or model is inherently misaligned with the true value.
- It reflects the bias of the estimator:
  - $\mathbf{Bias^2(\hat{\theta}) = (\mathbb{E}[\hat{\theta}] - \theta)^2}$.
- **Interpretation:**
  - The estimator is systematically off
    - target even if you had unlimited data,
    - it would still not center on the true parameter.

## Estimation Error (Variance):

- **Estimation error** reflects the **random variability**
  - in the estimator from sample to sample.
  - It's measured by the variance of the estimator:
    - $\mathbf{Var(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]}$
- **Interpretation:**
  - Even if the estimator is unbiased on average,
    - individual estimates can vary widely
    - depending on the sample drawn.

# 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**.

# 4.Methods of Point Estimation.

# 4.1 A Point Estimator: Introduction.

- A point estimator is a formula (or rule) used to calculate a single best guess of an unknown parameter based on sample data.
- **Formal Definition:**
  - Let **$\theta$** be a **population parameter** (like **mean $\mu$, proportion p**, etc.)
  - Then:
    - The **point estimator** is a **statistic $\hat{\theta}$**
    - The **point estimate** is the value computed **from your data** i.e.
      - $\hat{\theta} = \text{Estimator}(X_1, X_2, \ldots, X_n)$

| Parameter | Estimator - Statistic | Example Calculation |
|---|---|---|
| Mean revenue $\mu$ | $\bar{x}$ | Avg. of 100 transactions. |
| Proportion p | $\hat{p}$ | 30 clicks out of 500 emails |
| Variance $\sigma^2$ | $s^2$ | Sample variance from daily sales data. |

# 4.2 Techniques for Point Estimation.

- **What are we estimating?**
  - We want to estimate an unknown parameter $\theta$ (like mean, variance, or proportion) from data.
- Following are **common statistical methods** used to **derive point estimators**:

1. **Method of Moments (MoM):**
   - **Idea:**
     - Match sample moments (like the sample mean or variance) to the theoretical moments of the distribution.
   - **Process:**
     - Take the first k sample moments.
     - Set them equal to the first k population moments.
     - Solve for the unknown parameter(s).
   - **Example:** For a distribution **with mean μ,** use:
     - $\bar{x} = \mu \Rightarrow \hat{\mu} = \bar{x}$
   - **Use when:** Estimating parameters of **known distributions** like poison, exponential etc.

# 4.2 Techniques for Point Estimation.

2. **Maximum Likelihood Estimation (MLE):**
   - **Idea:**
     - Choose the **parameter value $\hat{\theta}$** that **maximizes** the **likelihood** of **observing your sample.**
   - **Process:**
     - Write the likelihood function:
       - $L(\theta) = P(\text{data}|\theta)$
     - Find $\hat{\theta} = \arg\max_{\theta} L(\theta)$
   - **Example:**
     - Suppose $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.
     - The MLE for $p$ is:
       - $\hat{p}_{MLE} = \frac{\sum X_i}{n}$
   - **Used when:**
     - You want **estimators** with nice mathematical properties
       - i.e. asymptotic normality, efficiency.

3. **Least Squares Estimation (LSE):**
   - **Idea:**
     - Minimize the **sum of squared errors**
     - between **observed and predicted values.**
   - **Used in:**
     - Regression models.
   - **Example:**
     - In **linear regression**:
       - $Y = \beta_0 + \beta_1 X + \epsilon$, LSE finds:
         - $\widehat{\beta_1} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

# 4.3 Maximum Likelihood Estimation.

- Let $X_1, X_2, \dots, X_n$ be a **random sample** from a population with **probability density (or mass) function** $f(x; \theta)$,
  - where:
    - $\theta \in \Theta$ is an **unknown parameter** (scalar or vector).
    - $\Theta$ is the parameter space.
- **Formal Definition:**
  - The **Maximum Likelihood Estimator (MLE)** of $\theta$, denoted $\hat{\theta}_{MLE}$, is the **value of $\theta$ that maximizes** the **likelihood function**, i.e. the **probability of observing the given data**:
    - $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta)$
      - where the likelihood function is:
    - $L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$
  - Alternatively, using the **log – likelihood for convenience**:
    - $\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta) \Rightarrow \hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell(\theta)$
    - The product assumes the $X_i$ are **iid.**
  - In probability: we fix $\theta$ and ask what is the chance of seeing this data?
  - In likelihood: we fix the data and ask: which $\theta$ makes this data most likely?

# 4.4 MLE for the parameter p of a Binomial Distribution.

1. **Binomial Distribution Overview:**
   - The Binomial distribution describes the number of successes $k$ in $n$ independent trials, each with success probability $p$.
   - Its **probability mass function (PMF)** is:
     - $\mathbf{P(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}}$
2. **Likelihood Function:**
   - Given **observed data k**, the **likelihood function L(p)** is the **probability** of **observing k** as a function **of p:**
     - $\mathbf{L(p) = \binom{n}{k} p^k (1 - p)^{n-k}}$
     - Since $\binom{n}{k} = \mathbf{C}$ does not **depend on p,** it is a **multiplicative constant**
     - **Thus:**
       - $\mathbf{L(p) = C \cdot p^k (1 - p)^{n-k}}$
     - Taking the **natural logarithm**:
       - $\mathbf{\log L(p) = \log\left[C \cdot p^k (1 - p)^{n-k}\right]}$

# 4.4 MLE for the parameter p of a Binomial Distribution.

3. **Solving logarithm equation:**
   - Recall the following logarithm rules:
     - **Product rule**: $\log(ab) = \log a + \log b$; **Power rule**: $\log(a^b) = b \log a$
   - Applying these rules:
     - $\log L(p) = \log C + \log\left(P^k (1-p)^{n-k}\right) \rightarrow$ (product rule)
     - $\log L(p) = \log C + \log p^k + \log(1-p)^{n-k} \rightarrow$ (Further product rule)
     - $\log L(p) = \log C + k \log p + (n-k)\log(1-p) \rightarrow$ (Power rule)
     - Since $\log C$ does **not depend on p**, it is treated as a constant when optimizing w.r.t $p$.
     - Thus, we write:
       - $\log L(p) = k \log p + (n-k)\log(1-p) + \text{constant}$
     - This is the **log – likelihood function** used for **MLE derivation**.
   - **We can Ignore the Constant, why?**
     - The MLE seeks **the value of p that maximizes $\log L(p)$.**
     - Since **$\log C$** does not change **with *p***, it has no effect on the location of the maximum.
   - Thus, we can drop it and simply work with:
     - $\boxed{\ell(p) = k \log p + (n-k)\log(1-p)}$

# 4.4 MLE for the parameter p of a Binomial Distribution.

4. **Maximizing the Log – likelihood:**
   - To find **MLE $\hat{p}$,** we maximize $\ell(p)$ with respect to $p$.
   - We take the **derivative of $\ell(p)$ w.r.t p** and **set it to zero**:

   i. **Computing the First derivative:**
      - $\frac{d\ell(p)}{d(p)} = \frac{d}{dx}[k \log p + (n-k)\log(1-p)]$
      - Compute the derivative term – by – term using the chain rule:
        - Derivative of $k \log p$:
          - $\frac{d(k \log p)}{dp} = \frac{k}{p}$
        - Derivative of $(n-k)\log(1-p)$:
          - $\frac{d}{dp}[(n-k)\log(1-p)] = (n-k) \cdot \frac{-1}{1-p} = -\frac{n-k}{1-p}$
      - Thus, the first derivative is:
        - $\frac{d\ell(p)}{dp} = \frac{k}{p} - \frac{n-k}{1-p}$

# 4.4 MLE for the parameter p of a Binomial Distribution.

ii. **Setting the First Derivative to Zero (Critical Point)**

- To **maximize $\ell(p)$**, we solve:
  - $\frac{k}{p} - \frac{n-k}{1-p} = 0$
  - $\frac{k}{p} = \frac{n-k}{1-p}$
  - $k(1-p) = (n-k)p$
  - $k - kp = np - kp$
  - $k = np$
- Solving for **p**:
  - $\hat{p} = \frac{k}{n}$

# 4.4 MLE for the parameter p of a Binomial Distribution.

5. **Final Result and Interpretation:**
   - The MLE $\hat{\mathbf{p}} = \frac{\mathbf{k}}{\mathbf{n}}$ is the **sample proportion of successes,**
     - which makes sense: the best estimate **for *p* is the observed success rate**.
   - **Final Answer:**
     - $\boxed{\hat{p} = \frac{k}{n}}$
     - The is the Maximum Likelihood Estimator (MLE) for the **parameter *p*** in a Binomial distribution.

# Home - Work

- **Case Study - Estimating Conversion Rate for a Marketing Campaign**

- **Scenario:**
  - You are a data analyst at a digital marketing firm. Your team just ran an **email campaign** targeting 5,000 customers.
  - The goal is to estimate the **conversion rate:**
    - the probability that a recipient makes a purchase after opening the email.
  - Out of the 5,000 recipients, **320 customers made a purchase**.
  - Your manager asks you to:
    - Estimate the **conversion rate**.
    - Evaluate how likely this observed data is under different values of p.
    - Explain why **MLE** is a natural method for this estimation.
- **Questions:**
  - What probability model would you use to model the number of conversions?
  - Write down the likelihood function for this model.
  - Derive the Maximum Likelihood Estimator (MLE) for the conversion probability p.
  - Compute the MLE using the given data.
  - Interpret your estimate in plain language.
    - What does this mean for the business team?

# Thank You