Inferential Statistics.
Point Estimation and Confidence Interval.

Prepared By: Siman Giri, Instructor: Ronit and Shiv for Herald Center for AI.

Summer, 2025

1 Learning Objectives.

- Explain the concept of point estimation and identify common estimators (mean, proportion, variance) in business contexts.
- Construct and interpret confidence intervals for population parameters using sample data.
- Evaluate the effect of sample size and confidence level on the width and reliability of a confidence interval.

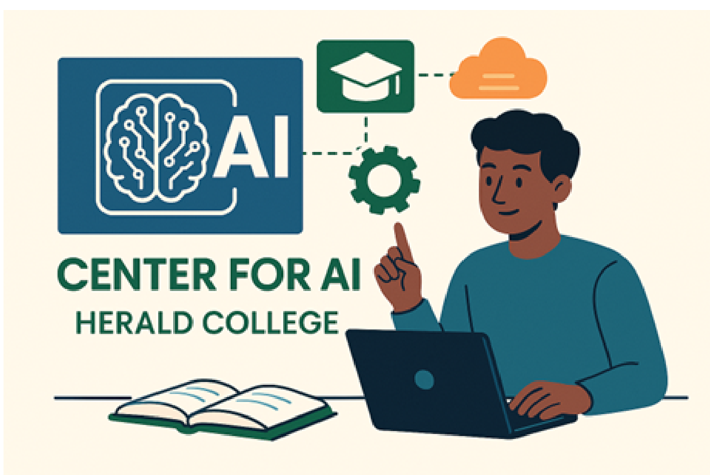

image generated via copilot.

2 Point Estimation:

2.1 Conceptual Understanding:

1. Define Maximum Likelihood Estimation. How does it differ from the Method of Moments?
2. Why is the logarithm of the likelihood function typically used when performing MLE? What are its advantages?
3. Suppose we observe a sample of customer inter-arrival times that follow an exponential distribution. What parameter would you estimate using MLE, and what would the likelihood function look like?

Also look at the section 5 for more detailed question on MLE.

3 Confidence Interval.

1. Find the t critical values for each combination of sample size and confidence level.

   (a) $\mathbf{95\%, n = 19}$

   (b) $\mathbf{90\%, n = 27}$

   (c) $\mathbf{80\%, n = 7}$

   Using Python and scipy.stats

   Compiling and Training the Model:

```
import scipy.stats as stats
def get_t_critical_value(confidence_level, n):
    df = n - 1 # Degrees of freedom
    alpha = 1 - confidence_level
    # Two-tailed critical value
    t_critical = stats.t.ppf(1 - alpha / 2, df)
    return t_critical
```

2. Find the sample size necessary in each of the following settings:

   (a) Suppose $\sigma = \mathbf{4}$ We wish to construct a $\mathbf{90}\%$ confidence interval with a maximum width of 10.

   (b) Suppose $\sigma = \mathbf{12}$ We wish to construct a $\mathbf{98}\%$ confidence interval with a maximum width of 15.

   (c) Suppose $\sigma = \mathbf{2}$ We wish to construct a $\mathbf{92}\%$ confidence interval with a maximum margin of error of 5.

   (d) Suppose $\sigma = \mathbf{10}$ We wish to construct a $\mathbf{99}\%$ confidence interval with a maximum margin of error of 3.

   Hint:

   Finding Required Sample Sizes for Confidence Intervals:

   To determine the minimum sample size ($\mathbf{n}$) needed for a confidence interval with a given width $\mathbf{W}$ or margin of error (ME), we use the formula:

$$\mathbf{n} = \left(\frac{\mathbf{z}_{\alpha/2} \cdot \sigma}{\text{ME}}\right)^{\mathbf{2}}$$

   Where:

   - $z_{\alpha/2}$ = critical value for the desired confidence level
   - σ = population standard deviation
   - ME = Margin of Error (half the width for symmetric CIs: ME $= \frac{W}{2}$)

3. **Balancing Statistical Priorities in Research Design:**

   - In statistical practice, researchers often face competing priorities:
     - Small margin of error (ME).
     - High Confidence level.
     - Small sample size.

   All of above contradicts with each other, Come up with a scenario:
     - where it will be especially important to have a small margin of error and a high level of confidence, and therefore worth spending a lot of resources on gathering a large sample size.
     - where it may be impractical to gather a large sample size. In your scenario, is it more important to prioritize interval width or confidence level?

   Explain your thought process.

4. **Marketing Campaign ROI:**

   - A sample of 40 marketing campaigns has an average ROI of 12% with a standard deviation of 3%. Construct a 95% confidence interval for the true average ROI.
   - **Solution:**
     Given:
     - `Sample Size:`$(\mathbf{n}) = \mathbf{40}$.
     - `Sample Mean:`$\mathbf{\bar{x}} = \mathbf{12}\%$.
     - `Sample Standard deviation:` $(\mathbf{s}) = \mathbf{3}\%$
     - `Confidence level:`$\mathbf{95}\%$

     Steps:
     - **Determine the critical valuez for** 95% **confidence:**

       $$\texttt{For a } \mathbf{95\%CI}\ \mathbf{z_{\alpha/2} = 1.96}.$$

     - **Calculate the Standard error (SE):**

       $$\mathbf{SE = \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{40}} \approx 0.474}.$$

     - **Compute the margin of error (M.E):**

       $$\mathbf{ME = z \times SE = 1.96 \times 0.474 \approx 0.929}.$$

     - **Construct the 95% CI:**

       $$\mathbf{CI = \bar{x} \pm ME = 12 \pm 0.929 = (11.07\%, 12.93\%)}.$$

5. **Fuel Efficiency of Delivery Trucks:**

   - A logistics company's fuel consumption was recorded for 36 delivery trucks. The average was 14 miles/gallon, with a standard deviation of 2 mpg. Construct a 90% confidence interval for the true mean fuel efficiency.

6. **Customer Satisfaction Survey:**
   - In a customer satisfaction survey, 240 out of 300 respondents said they were satisfied. Construct a 95% confidence interval for the proportion of satisfied customers.
   - **Hint for Solution:**
     Given:
     - `Sample Size:` $(\mathbf{n}) = \mathbf{300}$.
     - `Number of successes:` $\mathbf{x} = 240$.
     - `Sample proportion:` $\hat{\mathbf{p}} = \frac{\mathbf{240}}{\mathbf{300}} = \mathbf{0.8}$.
     - `Confidence level:` $= \mathbf{95}\%$.

     Use proportion to compute SE:

$$\mathbf{SE} = \sqrt{\frac{\hat{\mathbf{p}}(\mathbf{1} - \hat{\mathbf{p}})}{\mathbf{n}}}$$

7. **Software Subscriptions:**
   - A tech startup recorded monthly software subscriptions from a sample of 25 months. The sample mean was 3,400 subscriptions with a sample standard deviation of 600. Estimate the 99% confidence interval for the average subscriptions. {Hint: small sample - use t - distributions.}

8. Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7. Answer and Explain:
   (a) Should we use a z or a t critical value in this setting? Explain.
   (b) Find and interpret a 95% confidence interval for the mean number of eggs in a green sea turtle nest.
   (c) Find and interpret a 98% confidence interval for the mean number of eggs in a green sea turtle nest.
   (d) Suppose $\mu = \mathbf{105}$. Did the confidence intervals do a good job of estimating. Explain.

9. Some Herald students wanted to know how: long is an average commute to campus? A random sample of 32 students resulted in a mean of 31 minutes with standard deviation 18 minutes.
   (a) Should we use a z or a t critical value in this setting? Explain.
   (b) Find and interpret a 95% confidence interval for mean commute time.
   (c) Find and interpret a 90% confidence interval for mean commute time.

4 Confidence Interval - Coding Exercises.

Exercise 1 - Confidence Interval for Mean (Single Sample):

- **Scenario:** A company tracks the delivery time (in days) for their products. A random sample of 50 deliveries had an average time of 4.2 days with a sample standard deviation of 1.1 days.
  - **Task:**
    * Compute 95% confidence interval for the population mean delivery time using the t - distribution.

Exercise Code Template:

```python
import scipy.stats as stats
import numpy as np
# Given data
n = 50
sample_mean = 4.2
sample_std = 1.1
confidence = 0.95

# Your code here:
# Step 1: compute the standard error
# Step 2: get the critical t value
# Step 3: compute the margin of error
# Step 4: build the confidence interval
```

Exercise 2 - Confidence Interval for Proportion:

- **Scenario:** An e-commerce platform wants to estimate the proportion of customers who prefer express delivery. Out of 400 surveyed users, 128 preferred express delivery.
- **Tasks:**
  - Compute a 90% confidence interval for the true proportion of customers who prefer express delivery.

Exercise 2 Coding Template:

```python
# Given data
n = 400
x = 128
confidence = 0.90
# Your code here:
# Step 1: compute sample proportion
# Step 2: compute standard error for proportion
# Step 3: find the z critical value
# Step 4: compute confidence interval
```

Exercise 3 - Compare Two Means (Independent Samples):

- **Scenario:** A company wants to compare weekly sales (in $) from two different regions.
  - Region A: **n = 40**, **mean = 5200**, **SD = 610**.
  - Region B: **n = 35**, **mean = 4900**, **SD = 580**.
- **Tasks:**
  - Compute the 95% confidence interval for the difference in population means.

Exercise 3 code template.

```
# Given data
n1, mean1, std1 = 40, 5200, 610
n2, mean2, std2 = 35, 4900, 580
confidence = 0.95

# Your code here:
# Step 1: compute standard error for the difference
# Step 2: degrees of freedom (Welch's approximation if needed)
# Step 3: get critical t value
# Step 4: compute confidence interval for the difference
```

Exercise 4 - Visualizing Confidence Intervals:

- **Scenario:** You collected sample means from a simulation study of customer ratings (1 to 5 stars). You want to visualize how the confidence interval captures the true mean.
- **Tasks:**
  - Simulate 100 samples of size 30 from a population with mean = 3.6 and std = 0.8
  - For each sample, compute the 95% confidence interval
  - Plot the intervals and highlight how many contain the true mean (3.6)

Exercise 4 Coding Template:

```
import matplotlib.pyplot as plt
# Step 1: simulate 100 samples
# Step 2: compute CI for each
# Step 3: store lower and upper bounds
# Step 4: plot intervals, color based on whether they contain true mean
```

5 Optional: Point Estimation with MLE.

How to Derive a Maximum Likelihood Estimator (MLE)?

MLE is a method used to estimate parameters of a probability distribution by maximizing the likelihood that the observed data came from the assumed distribution.

Goal:

Given a probability distribution (e.g., Normal, Bernoulli, Binomial, Poisson), and observed data, find the parameter(s) that **maximize the likelihood** of the data.

Step - by - step Process:

Step 1: Define the Likelihood Function

Suppose we have a random sample:

$$\mathbf{X_1, X_2, \dots, X_n \sim f(x; \theta)}$$

Where:

- $\mathbf{f(x;\theta)}$: Probability Density Function (PDF) or Probability Mass Function (PMF)
- θ: The parameter(s) we want to estimate

The likelihood function is:

$$\mathbf{L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)}$$

Interpretation: `It's the probability (or density) of observing the specific sample values given the parameter` θ.

Step 2: Take the Log of the Likelihood (Log-Likelihood)

Because products are hard to work with, take the natural logarithm:

$$\ell(\theta) = \log \mathbf{L}(\theta) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \log \mathbf{f(X_i; \theta)}$$

Why? `Log turns products into sums and simplifies derivatives.`

Step 3: Differentiate the Log-Likelihood

Differentiate $\ell(\theta)$ with respect to θ:

$$\frac{\mathbf{d}}{\mathbf{d}\theta}\ell(\theta) \quad \text{(Score function)}$$

Step 4: Set Derivative to Zero and Solve

$$\frac{\mathbf{d}}{\mathbf{d}\theta}\ell(\theta) = \mathbf{0}$$

Solve this equation for θ. This gives the value $\hat{\theta}$, the MLE.

Step 5: Verify Maximum (Optional but Good Practice)

Check the second derivative:

$$\frac{\mathbf{d^2}}{\mathbf{d}\theta^{\mathbf{2}}}\ell(\theta) < \mathbf{0}$$

If this is negative, it confirms a maximum (not a minimum or saddle point).

5.1 Maximum Likelihood Estimation for Bernoulli Distribution

Problem Setup

We have data:

$$\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n} \sim \text{Bernoulli}(\mathbf{p})$$

where each $\mathbf{x_i} \in \{\mathbf{0}, \mathbf{1}\}$.
The probability mass function is:

$$\mathbf{P}(\mathbf{X} = \mathbf{x_i}) = \mathbf{p^{x_i}}(\mathbf{1} - \mathbf{p})^{\mathbf{1-x_i}}$$

Step 1: Likelihood Function

The likelihood function is:

$$\mathbf{L}(\mathbf{p}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{p^{x_i}}(\mathbf{1} - \mathbf{p})^{\mathbf{1-x_i}}$$

Step 2: Log-Likelihood Function

$$\ell(\mathbf{p}) = \log \mathbf{L}(\mathbf{p}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} [\mathbf{x_i} \log \mathbf{p} + (\mathbf{1} - \mathbf{x_i}) \log(\mathbf{1} - \mathbf{p})]$$

Let $\mathbf{S} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}$ (total successes), then:

$$\ell(\mathbf{p}) = \mathbf{S} \log \mathbf{p} + (\mathbf{n} - \mathbf{S}) \log(\mathbf{1} - \mathbf{p})$$

Step 3: Derivative of Log-Likelihood

$$\frac{\mathbf{d}\ell}{\mathbf{dp}} = \frac{\mathbf{S}}{\mathbf{p}} - \frac{\mathbf{n} - \mathbf{S}}{\mathbf{1} - \mathbf{p}}$$

Why these derivatives?

1. Derivative of $\log(\mathbf{p})$ with respect to $\mathbf{p}$

$$\frac{d}{d(p)} \log(p) = \frac{1}{p}$$

2. Derivative of $\log(1-\theta)$ with respect to θ:
   Using the chain rule:

$$\frac{d}{d(p)}\log(1-p) = \frac{1}{1-p}\cdot\frac{d}{d(p)}(1-p)$$
$$= \frac{1}{1-p}\cdot(-1)$$
$$= -\frac{1}{1-p}$$

Step 4: Find Maximum Likelihood Estimator

Set derivative equal to zero:

$$\frac{S}{p} - \frac{n-S}{1-p} = 0$$

Multiply through by $p(1-p)$:

$$S(1-p) = (n-S)p$$

Simplify:

$$S - Sp = np - Sp$$
$$S = np$$

Thus, the MLE is:

$$\hat{p}_{\text{MLE}} = \frac{S}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Interpretation

The maximum likelihood estimator for p is simply the sample proportion of successes.

5.2 Your Turn:

- **Derive MLE for the Mean of Normal Distribution with known Variance:**

- **Assume:**

$$\mathbf{X_1}, \ldots, \mathbf{X_n} \sim \mathcal{N}(\mu, \sigma^{\mathbf{2}}) \rightarrow \sigma^{\mathbf{2}}\ \texttt{known}$$

5.3 Case - Based MLE Problems:

1. Retail Conversion Rates – Bernoulli Distribution

A/B testing was conducted on two versions of a website. You observe that 120 out of 200 visitors on version A made a purchase.
Assume a Bernoulli distribution for purchases (success = 1).

(a) Derive the likelihood function for the probability of conversion θ.

(b) Find the Maximum Likelihood Estimate (MLE) of θ.

Sample Solution:
Let $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ be independent and identically distributed (i.i.d.) random variables where $\mathbf{X_i} \sim$ Bernoulli(θ).
The probability mass function is:

$$\mathbf{P}(\mathbf{X_i} = \mathbf{x_i}) = \theta^{\mathbf{x_i}}(\mathbf{1} - \theta)^{\mathbf{1}-\mathbf{x_i}}, \quad \mathbf{x_i} \in \{\mathbf{0}, \mathbf{1}\}$$

The likelihood function is:

$$\mathbf{L}(\theta) = \prod_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \theta^{\mathbf{x_i}}(\mathbf{1} - \theta)^{\mathbf{1}-\mathbf{x_i}}$$

Let $\mathbf{S} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{x_i}$, then:

$$\mathbf{L}(\theta) = \theta^{\mathbf{S}}(\mathbf{1} - \theta)^{\mathbf{n}-\mathbf{S}}$$

The log-likelihood function is:

$$\ell(\theta) = \log \mathbf{L}(\theta) = \mathbf{S} \log \theta + (\mathbf{n} - \mathbf{S}) \log(\mathbf{1} - \theta)$$

Differentiate with respect to θ:

$$\frac{\mathbf{d}\ell}{\mathbf{d}\theta} = \frac{\mathbf{S}}{\theta} - \frac{\mathbf{n} - \mathbf{S}}{\mathbf{1} - \theta}$$

Set derivative equal to zero:

$$\frac{\mathbf{S}}{\theta} - \frac{\mathbf{n} - \mathbf{S}}{\mathbf{1} - \theta} = \mathbf{0}$$

Solve for θ:

$$\frac{\mathbf{S}}{\theta} = \frac{\mathbf{n} - \mathbf{S}}{\mathbf{1} - \theta} \Rightarrow \mathbf{S}(\mathbf{1} - \theta) = (\mathbf{n} - \mathbf{S})\theta \Rightarrow \mathbf{S} = \mathbf{n}\theta \Rightarrow \hat{\theta} = \frac{\mathbf{S}}{\mathbf{n}}$$

For this data: $\mathbf{S} = \mathbf{120}, \mathbf{n} = \mathbf{200}$

$$\hat{\theta} = \frac{\mathbf{120}}{\mathbf{200}} = \mathbf{0.6}$$

Interpretation: The MLE for the conversion rate is 0.6, meaning 60% of visitors are estimated to make a purchase on version A.

2. Service Times – Exponential Distribution

A customer service center records the time (in minutes) taken to resolve tickets. The data is modeled using an exponential distribution:

$$\mathbf{f}(\mathbf{t}; \lambda) = \lambda \mathbf{e}^{-\lambda \mathbf{t}}$$

Given a sample: [5.2, 3.1, 4.7, 6.0, 2.8]

(a) Write the likelihood function.

(b) Derive the MLE for the parameter λ.

3. Customer Complaints – Poisson Distribution

The number of complaints received per day at a call center is believed to follow a Poisson distribution. Over 7 days, the observed complaints were: [2, 3, 4, 1, 0, 3, 2].

(a) Derive the MLE for the average rate λ of complaints per day.

(b) What does this estimate tell you from a business operations standpoint?

4. Normal Distribution – Revenue Per Transaction

The average revenue per transaction at a store is believed to follow a normal distribution. You collect the following sample (in dollars): [220, 250, 245, 230, 265, 240] Assume unknown μ and σ^2.

(a) Derive the MLEs for both parameters.

(b) Interpret the meaning of these estimates in the business context.

——————— The - End ———————-