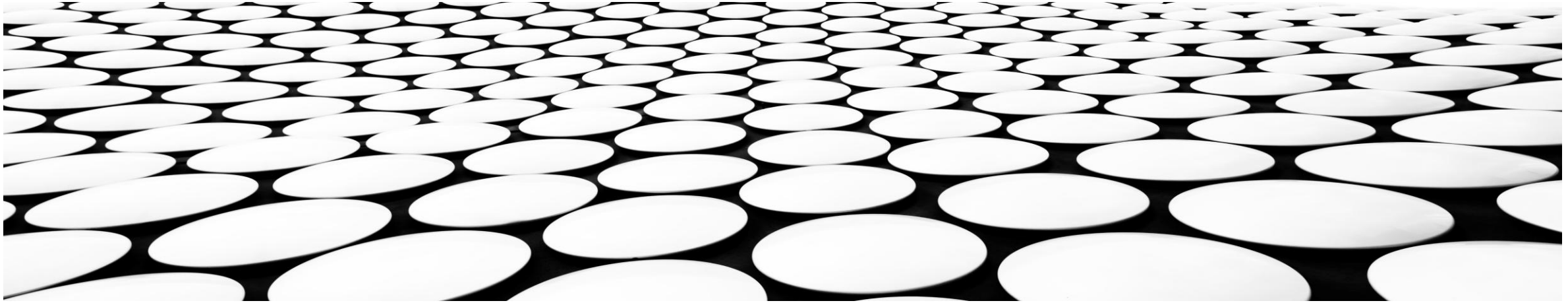
DATA MINING AND PREDICTIVE DATA ANALYTICS

CHAPTER-5

UNIVARIATE STATISTICAL ANALYSIS



STATISTICAL APPROACHES TO ESTIMATION AND PREDICTION

- In data analysis and statistics, we work with ***samples*** and try to make meaningful conclusions about the larger (entire) ***population***.
- This process of drawing conclusions from sample data is known as **Statistical Inference**.
- A sample is simply a subset of the population, preferably a representative subset.
- If the sample is not representative of the population, that is, if the sample characteristics deviate systematically from the population characteristics, statistical inference should not be applied

STATISTICAL APPROACHES TO ESTIMATION AND PREDICTION

- **Statistical Inference (Drawing population-level conclusions from sample)**
  - Statistical inference is a branch of statistics that **uses sample data to make generalizations, predictions, or decisions about a population**, while **quantifying the uncertainty** associated with these conclusions.
- It answers critical questions such as:
  - What can we say about the population based on the sample?
  - Are the observed patterns in the sample data genuine or due to random chance?
  - How confident (probability score) are we in the results?

STATISTICAL APPROACHES TO ESTIMATION AND PREDICTION

- **Role of Statistical Inference in Data Mining**
  - After completing Data understanding, Data preparation, Exploratory data analysis (EDA)… the next step is to **estimate** and **test** statements about the population using sample data.
  - This is done through **statistical inference**.
  - Statistical inference refers to the collection of methods used to:
    - **Estimate** unknown population parameters, and
    - **Test hypotheses** about those parameters using the information contained in a sample.
  - Thus, Statistical inference forms the core of predictive and inferential analytics in data mining workflows.

CHAPTER OVERVIEW

- Here, in this chapter, we examine univariate methods, statistical estimation, and prediction methods that analyze one variable at a time.
- These methods include **point estimation** and **confidence interval estimation** for **population means** and **proportions**.
- We discuss ways of reducing the margin of error of a confidence interval estimate.
- Then we turn to **hypothesis testing**, examining hypothesis tests for **population means** and **proportions**.

STATISTICAL INFERENCE

❑ Key Concepts

1. Population: Entire collection data of individuals or items of interest.
   - Example: All current and future customers of a cell phone company
2. Parameter: A numerical characteristic of a population
   - Examples: Population mean = μ

     Population proportion = π

     Population standard deviation = σ

   Parameters are usually unknown.

STATISTICAL INFERENCE

3. **Sample:** A subset of the population, ideally **representative**
   - Example: The *3333 customers* in the churn dataset
   - If the sample is **not representative**, inference becomes invalid.
4. **Statistic:** A numerical characteristic computed from a sample data
   - Examples:
     - Sample mean = $\bar{x}$
     - Sample proportion = p
     - Sample standard deviation = s
   - Statistics are used to **estimate** parameters.

STATISTICAL INFERENCE

- **Example**
  - Sample mean ($\bar{x}$) *customer service calls* → Used to estimate the unknown population mean (μ).
  - Sample *churn* proportion (p) → Used to estimate population *churn* proportion(π).
- **Example: Estimating the Mean Number of Customer Service Calls**
  - The true population mean number of customer service calls (μ) is unknown since complete data may not be collected or stored.
  - Data analysts use **estimation** to approximate μ.
  - If the sample mean is **1.563,** the estimated population mean number of customer service calls is **1.563**.

STATISTICAL INFERENCE

- Representativeness Matters
  - Inference is trustworthy only if the sample reflects the population.
  - Example: If the 3333 customers were mainly disgruntled customers, then:
    - Sample ≠ population
    - EDA results are not actionable
    - Estimates do not represent all customers

ESTIMATION

- **Estimation**
  - Estimation involves using sample data (statistics) to infer population parameters.
- **Point Estimation**
  - Provides a **single best estimate** of a population parameter.
  - Example: Sample mean ($\bar{x}$) estimates population mean (μ).
- **Interval Estimation**
  - Provides a **range of values** likely to contain the population parameter.
  - The most common is the **Confidence Interval (CI)**.
  - **Example:** 95% CI for mean = [50.2, 53.8]

    This means we are 95% confident the true mean lies within this range.

ESTIMATION

Point Estimation

- *Point estimation* refers to the use of a single known value of a statistic to estimate the associated population parameter.

Sample Statistic	Estimates Population Parameter
Mean $\bar{x}$	Mean μ
Standard deviation s	Standard deviation σ
Proportion p	Proportion π

Estimation Beyond Basic Parameters

- Any sample statistic can estimate its population counterpart.
- Examples: Sample maximum → estimate of population maximum
- Subgroup sample proportion → estimate of subgroup population proportion

ESTIMATION

- **How Confident Are We in Our Estimates?**
  - **Point estimates** such as the sample mean $\bar{x}$ are useful but incomplete.
  - They tell us the “best single guess,” but not **how accurate** that guess is.
  - Because a sample contains only part of the population information, the point estimate almost always differs from its corresponding population parameter.
- **Sampling Error**
  - Sampling error is defined as:

$$\text{Sampling Error} = \text{Point Estimate} - \text{Population Parameter}$$

  - For example, Sampling Error for a mean: $\left|\bar{x} - \mu\right|$
  - Since population parameters are unknown, the true sampling error is also unknown.

ESTIMATION

- **Why Point Estimates Are Insufficient**
  - Point estimates come with:
    - No measure of confidence in their accuracy
    - No probability statement associated with the estimate
    - No information about how close they are to the true value
- They can be thought of as point estimates directed toward an unknown true value—you can never be sure how close your estimate actually is.
- In contrast, an **interval estimate** has a measurable span, so it has a greater chance of covering the true value.
- This idea forms the basis of **confidence intervals**.

CONFIDENCE INTERVAL ESTIMATION OF THE MEAN

- A **confidence interval (CI)** provides:

  Point Estimate ± Margin of Error

  - Confidence interval estimation can be applied to any target parameter,
  - Commonly used for estimating the population mean and population proportion.

- **General Form of CI for a Population Mean**

$$\bar{x} \pm t_{\alpha/2,\, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

where:

- $\bar{x}$ = sample mean
- s = sample standard deviation
- $\frac{s}{\sqrt{n}}$ = standard error of the mean
- $t_{\alpha/2,n-1}$ = t-multiplier depending on confidence level and sample size

CONFIDENCE INTERVAL ESTIMATION OF THE MEAN

- **Margin of Error:** $E = t_{\alpha/2,\, n-1}\left(\frac{s}{\sqrt{n}}\right)$

What is $t_{\alpha/2,\, n-1}$?

- It is a **critical value** taken from the **t-distribution table**.
- It depends on the confidence level and the sample size.

Confidence Level	α	α/2
90%	0.10	0.05
95%	0.05	0.025
99%	0.01	0.005

- So for 95% CI, you look up: $t_{0.025,\, n-1}$
- Higher confidence → **wider interval** → **larger t value**.

CONFIDENCE INTERVAL ESTIMATION OF THE MEAN

- **Example 1:** Full Customer Base (Large Sample)
  - Find the 95% *t*-interval for the mean number of customer service calls for all customers
  - $\bar{x}$=1.563, *s*=1.315, *n*=3333, *CI*= 95% : $1.563 \pm t_{0.025,3332}\left(\frac{1.315}{\sqrt{3333}}\right)$
  - Result: (1.518, 1.608)
  - Margin of error: *E*=0.045
  - **Interpretation:**

    We are 95% confident that the true mean number of customer service calls is between 1.518 and 1.608.
  - Margin of error is relatively small due to large sample size

CONFIDENCE INTERVAL ESTIMATION OF THE MEAN

- **Example 2:** Subgroup Analysis (Small Sample)
  - Estimate the behavior of specific subsets of customers instead of the entire customer base,
  - Subgroup: Customers with *Intl_Plan* = Yes, *VMail_Plan* = Yes, *day_min* > 220.
  - *Given that, n* = 28, $\bar{x}$ =1.607, *s* = 1.403, *CI* = 95% :
  - Result: (0.873,2.341)
  - Margin of error: *E*=0.734
  - **Interpretation:**
    - We are 95% confident that the population mean number of customer service calls for all customers who have both plans and who have more than 220 minutes of day use falls between 0.873 and 2.341 calls.
    - Margin of error is much larger (wider interval) because the sample size is small.

HOW TO REDUCE THE MARGIN OF ERROR

- The three components affecting Margin of Error (*E*), are:
  - $t_{\alpha/2,\, n-1}$ — depends on confidence level
  - *s* — sample variability (cannot be changed)
  - *n* — sample size

$$E = t_{\alpha/2,\, n-1}\left(\frac{s}{\sqrt{n}}\right)$$

- There are two practical options:

1. Decrease the confidence level (**Not recommended)**

- Reduces $t_{\alpha/2,\, n-1}$ but weakens statistical reliability.

2. Increase sample size

- This is the **only recommended** method.

CONFIDENCE INTERVAL ESTIMATION OF THE PROPORTION

- Sample proportion: $p = x/n$
- ***Example:***
  - 483 of 3333 customers had churned,
  - So, an estimate of the *population proportion* π of all of the company's customers who churn is: x/n = 483/3333 = 0.1449
- **Confidence Interval for a Population Proportion is given by**: $p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- A Z-interval for the population proportion π is valid when: $np \geq 5, \quad n(1-p) \geq 5$
- Common Z-values: depends on the confidence level and for 90%, 95% and 99% confidence, the respective values are $Z_{0.10} = 1.645, \quad Z_{0.05} = 1.96, \quad Z_{0.01} = 2.575$

CONFIDENCE INTERVAL ESTIMATION OF THE PROPORTION

- **Example:** *Churn* Proportion
  - Find 95% confidence interval for the proportion of churners among the entire population of the company's customers
  - Given: x = 483, n = 3333, p = 483/3333 = 0.1449, CI = 95%
  - Result:(0.133, 0.157)
  - Margin of error: E=0.012
  - Interpretation:
    - We can estimate the true churn rate within 1.2% with 95% confidence.
    - The only way to reduce the margin of error at a constant confidence level is to increase sample size.

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

HYPOTHESIS TESTING FOR THE MEAN

- Hypothesis testing is a core component of **statistical inference**—the process of drawing conclusions about population parameters (like π or μ) based on sample data.
- A hypothesis test evaluates whether the data provide sufficient evidence to reject a preconceived claim about a population.
- We develop **two competing statements or hypotheses**:
  - **Null Hypothesis (H_0):** Represents the status quo or current belief about the population parameter.
    - Example: $H_0 : \mu = \mu_0$
  - **Alternative (Research) Hypothesis (H_1):** Represents the investigator's claim or suspected truth.
    - Examples: $H_1 : \mu > \mu_0, \quad \mu < \mu_0, \quad \mu \neq \mu_0$

HYPOTHESIS TESTING FOR THE MEAN

- A hypothesis test ends with one of two decisions:
  - **Reject H_0**
  - **Do not reject H_0**
- A useful analogy is a criminal trial:
  - **H_0:** Defendant is innocent
  - **H_1:** Defendant is guilty
  - Jury either **rejects H_0** (convicts) or **does not reject H_0** (acquits)

HYPOTHESIS TESTING FOR THE MEAN

- **Type I and Type II Errors**
  - Because decisions are based on sample data, errors are possible.
  - **Type I Error (α):** Rejecting H_0 when H_0 is true.

    → Convicting an innocent person.
  - **Type II Error (β):** Not rejecting H_0 when H_0 is false

    → Acquitting a guilty person.

Reality	Jury rejects H_0 (guilty)	Jury does not reject H_0 (not guilty)
H_0 true (innocent)	**Type I error**	Correct decision
H_0 false (guilty)	Correct decision	**Type II error**

- The probability of a Type I error is denoted α, while the probability of a Type II error is denoted β.

HYPOTHESIS TESTING FOR THE MEAN

- For a constant sample size, a decrease in α is associated with an increase in β, and vice versa.
- The probability of a Type I error (α) is called the **significance level**.
- Common values:
  - α = 0.05
  - α = 0.01
  - α = 0.10
- α is usually fixed *before* the test begins.

HYPOTHESIS TESTING FOR THE MEAN

- **Forms of Hypothesis Testing for the Mean (μ)**
  - We consider three standard hypothesis structures:

1. Left-tailed test: $H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$

2. Right-tailed test: $H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0$

3. Two-tailed test: $H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$

where, μ_0 represents a hypothesized value of μ .

HYPOTHESIS TESTING FOR THE MEAN

Test Statistic for the Mean

- When the population is normal or sample size is large (n > 30),

the **test statistic** $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- $\bar{x}$: sample mean
- s: sample standard deviation
- n: sample size
- t: measures how many standard errors $\bar{x}$ is from μ_0

- An extreme t-value means the data disagree strongly with H_0.
- How extreme is extreme? This is measured using the ***p-value***.

HYPOTHESIS TESTING FOR THE MEAN

- **Understanding the p-Value**
- The *p-value* is the probability of observing a sample statistic (such as mean or proportion) at least as extreme as the statistic actually observed, if we assume that H_0 is true.
- A small p-value will indicate conflict between the data and the null hypothesis. Thus, we will reject H_0 if the p-value is small.
- Interpretation:
  - **Small p-value → strong evidence against H_0 → reject H_0**
  - **Large p-value → insufficient evidence → do not reject H_0**
- p-value, being a probability value, always lies between 0 and 1

HYPOTHESIS TESTING FOR THE MEAN

How to compute p-value

Test Type	p-value
Left-tailed	$P(T \leq t)$
Right-tailed	$P(T \geq t)$
Two-tailed	$2P(T \leq t)$ if t < 0; $2P(T \geq t)$ if t > 0

- T → random variable following the theoretical t-distribution with degrees of freedom, df = n − 1 (used for probability calculations)
- Use t-distribution table or software to find the probability score

- We consider the *p*-value to be small if it is less than **α**.
- This leads us to the rejection rule:

Reject H_0 if *p*-value < α

HYPOTHESIS TESTING FOR THE MEAN

- **Example:**
  - Recall our subgroup of customers who have both the Intl_Plan and the Vmail_Plan and who have more than 220 Day_minutes.
  - Test - whether the mean (μ) number of customer service calls of all such customers differs from 2.4, and we set the level of significance α to be 0.05.
  - We would have a two-tailed hypothesis test:

$$\mathbf{H_0}: \mu = 2.4 \quad Vs \quad \mathbf{H_1}: \mu \neq 2.4$$

  - So, The null hypothesis will be rejected if the p-value is less than 0.05

HYPOTHESIS TESTING FOR THE MEAN

- **Example (contd..):**
  - Given $\bar{x} = 2.857, \quad s = 1.892, \quad n = 28, \quad \mu_0 = 2.4$
  - Test statistic: $t = \dfrac{2.857 - 2.4}{1.892/\sqrt{28}} = 2.16$
  - Since this is a two-tailed test: $\text{p-value} = 2P(T \geq 2.16) = 0.035$
  - Decision at α = 0.05: $0.035 < 0.05 \Rightarrow \text{Reject } H_0$
  - Interpretation: There is evidence at the 5% significance level that μ differs from 2.4.

HYPOTHESIS TESTING FOR THE MEAN

- **Assessing the Strength of Evidence Against the Null Hypothesis**
  - In classical hypothesis testing, a decision is made by comparing the **p-value** with a chosen **significance level α**:
    - If $\text{p-value} < \alpha$: Reject H_0
    - If $\text{p-value} \geq \alpha$: Do not reject H_0
  - Although widely used, this simple **yes/no** decision does not capture the full meaning of the p-value.
  - The p-value actually reflects **how strongly the sample data contradict the null hypothesis**, and this strength varies continuously.

HYPOTHESIS TESTING FOR THE MEAN

- **How Changing α Changes the Conclusion**
  - Suppose: p-value = **0.035** and Usual significance level α = **0.05**
    - Then: $0.035 < 0.05 \Rightarrow \text{Reject } H_0.$
  - But if we had chosen a stricter level: α = 0.01
    - Then: $0.035 > 0.01 \Rightarrow \text{Do not reject } H_0.$
  - The data and hypotheses did not change, yet the conclusion reverses.
  - **This highlights a major limitation:** the decision often depends more on the arbitrary choice of α than on the evidence itself.
  - Using the evidence-based approach, a p-value of 0.035 would be providing solid evidence against H_0 rather than a simplistic reject/not-reject decision

HYPOTHESIS TESTING FOR THE MEAN

- Strength of Evidence Based on P-Values
  - P-values offer graded, continuous evidence — not just a binary decision.
  - To interpret that, p-values can be grouped into meaningful categories.

Table: Strength of Evidence Against H_0 Based on P-Values

p-Value Range	Strength of Evidence Against H_0
$p \leq 0.001$	Extremely strong evidence
$0.001 < p \leq 0.01$	Very strong evidence
$0.01 < p \leq 0.05$	Solid evidence
$0.05 < p \leq 0.10$	Mild evidence
$0.10 < p \leq 0.20$	Slight evidence
$p > 0.20$	No evidence

HYPOTHESIS TESTING FOR THE PROPORTION

- Hypothesis testing can also be applied to the population proportion π, such as the proportion of customers who churn.
- When testing hypotheses about a proportion, the hypothesized population proportion is denoted by π_0
- This represents the benchmark or claimed value against which the sample evidence is evaluated.

- **Test Statistic**

For large samples, the test statistic for a proportion follows a standard normal (Z) distribution:

$$Z_{\text{obs}} = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

where:

- $\hat{p}$ = sample proportion
- π_0 = hypothesized population proportion
- n = sample size

HYPOTHESIS TESTING FOR THE PROPORTION

- **Hypotheses**
- The hypotheses take one of the following forms:
- Left-tailed test: $H_0 : \pi = \pi_0$ $H_1 : \pi < \pi_0$ $\text{p-value} = P(Z \leq z_{\text{obs}})$
- Right-tailed test: $H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$ $\text{p-value} = P(Z \geq z_{\text{obs}})$
- Two-tailed test: $H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$ $\text{If } z_{\text{obs}} > 0,\ \text{p-value} = 2P(Z \geq z_{\text{obs}})$

  $\text{If } z_{\text{obs}} < 0,\ \text{p-value} = 2P(Z \leq z_{\text{obs}})$

- The structure of the hypotheses and the rules for computing p-values are the same as those used for testing the mean.

HYPOTHESIS TESTING FOR THE PROPORTION

- **Example**
  - 483 of 3333 customers churned: $p = \frac{483}{3333} = 0.145$
  - We test whether the true population churn proportion differs from 0.15 using α=0.05 $H_0 : \pi = \pi_0 = 0.15$ $H_1 : \pi \neq 0.15$
  - Compute the Z statistic:

$$Z = \frac{0.145 - 0.15}{\sqrt{\frac{0.15(1-0.15)}{3333}}}$$

$$Z = -0.93 \text{ (approximately)}$$

  - Two-tailed p-value: $\text{p-value} = 2P(Z \leq -0.93) \approx 0.352$
  - Interpretation: $\text{p-value } (0.352) > \alpha(0.05),$
  - We do not reject H_0. There is no evidence that the true churn proportion differs from 15%.