Key Areas of Statistics

Statistics is broadly divided into:

1. Descriptive Statistics

2. Inferential Statistics

Both are essential in Data Science and ML.

---

3. Descriptive Statistics (Understanding Data)

Descriptive statistics is about describing the data. Data when collected is known as the ***raw data*** meaning unprocessed data. It must be processed to make some sense.

One of the initial steps in processing understanding the data is to describe the data. The tools applied to describe the data are: **Charts and Graphs (Graphical and Visual Techniques)** and **Numerical Methods**. Both of these tools are discussed in detail in separate chapters later.in

Some of the method used in Descriptive Statistics are discussed briefly here

(a) Measures of Central Tendency

These describe the "center" of data:

- Mean (average)
- Median (middle value)
- Mode (most frequent value)

Use:

- Feature understanding
- Detecting skewness

(b) Measures of Dispersion (Variability)

These show how spread out the data is:

- Variance
- Standard deviation
- Range
- Coefficient of variation (CV)
- Interquartile range (IQR)

The above allow us to measure and quantify variability in data. A knowledge of variability is critical in data analysis.

Example:

- High variance → more uncertainty in predictions (less reliable is data)

(c) Measures of Position

- Percentiles
- Quartiles
- Deciles

(d) Shape of Distribution

- Skewness (asymmetry)
- Kurtosis (tail heaviness)

These affect:

- Model assumptions
- Transformation decisions

(e) Data Visualization (Graphical Techniques or Data Visualization)

Key tools:

- Histograms
- Box plots
- Scatter plots and many others

Purpose:

- Identify patterns, trends, and anomalies

(f) Correlation Analysis (Studying the relationship between two variables)

It is important to understand how variables are related - Measures relationships between variables:

- Pearson correlation (linear relationship)
- Spearman correlation (rank-based)

Use:

- Feature selection
- Detect multicollinearity

All of the above tools are discussed in detail in two separate chapters under Descriptive Statistics:

Charts and Graphs with examples and computer instructions,

and Numerical Methods with computer instructions

Here we have provided an outline and brief description.

Inferential Statistics

The other broad area of statistics is Inferential Statistics. Here is an overview:

Inferential Statistics (Making Predictions & Decisions)

Inferential statistics allows us to **generalize from sample data to a population**. It is related to probability and uncertainty in our conclusion. The broad topics of importance area:

(a) Probability Theory (Core Foundation)- Probability concepts are critical to understand inferential statistics

---

Key concepts:

- **Random variables**
- **Probability distributions (Normal, Binomial, Poisson)**
- **Conditional probability**

Example:

- Likelihood of an event (e.g., fraud detection)

(b) Sampling Theory

- Random sampling
- Sampling distributions
- Central Limit Theorem (CLT)

Key idea: CLT- Even if data is not normal, sample means tend to be normal.

(c) Estimation

Estimating population parameters:

- Point estimates (single value)
- Confidence intervals (range of values)

Example:

- "Accuracy is 92% ± 2%"

(d) Hypothesis Testing

A fundamental tool in analytics and ML: Steps:

1. Define null hypothesis (H_0)
2. Define alternative hypothesis (H_1)
3. Collect the data and compute test statistic
4. Evaluate p-value or use other methods to test the hypothesis.

Common tests:

- t-test
- z-test
- testing proportions
- chi-square test
- ANOVA

Use cases:

- A/B testing
- Feature importance validation

(e) Regression Analysis

A cornerstone of ML: Predictive analytics tool

- Linear regression
- Logistic regression
- Variations and different types of regression

Statistical aspects:

- Coefficient estimation
- Significance testing
- Residual analysis

(f) Bayesian Statistics

Uses probability to update beliefs:

- Prior → initial belief
- Likelihood → data evidence
- Posterior → updated belief

Applications:

- Spam filtering
- Recommendation systems

(g) Nonparametric Methods

Used when assumptions (e.g., normality) do not hold:

- Rank tests
- Kernel density estimation
- Decision trees (ML connection)

---

5. Connection to Data Science & Machine Learning

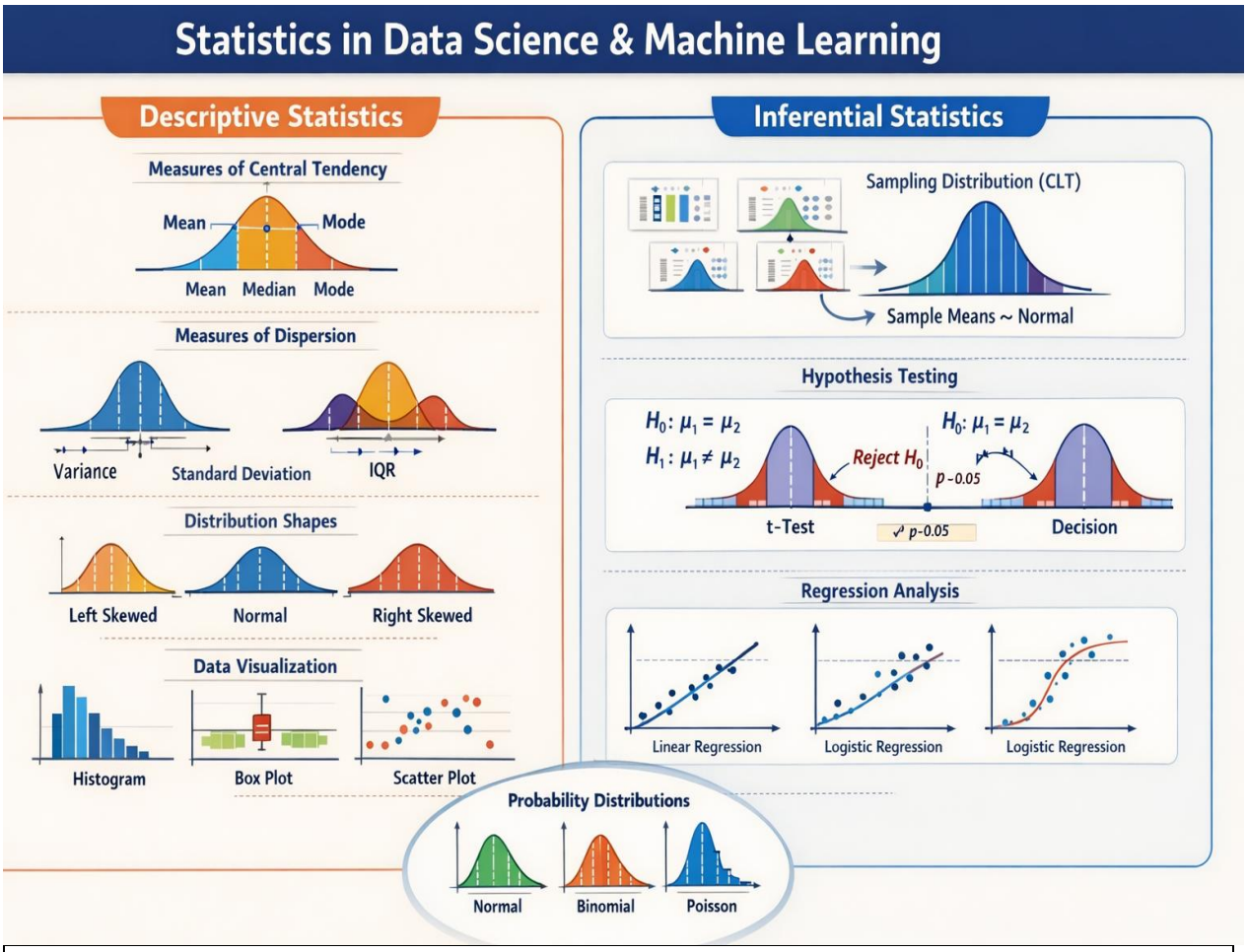
Area	Role of Statistics
Data Cleaning	Detect outliers, missing patterns
Feature Engineering	Correlation, distributions
Model Training	Optimization & estimation
Model Evaluation	Accuracy, precision, recall, ROC
Experimentation	A/B testing, hypothesis testing
Uncertainty Quantification	Confidence intervals, Bayesian methods

---

6. Key Takeaways

- Statistics provides the **language of data**.
- Descriptive statistics helps **understand data**.
- Inferential statistics enables **decision-making and prediction**.
- Machine learning is essentially **applied statistical modeling at scale**.

Visual diagrams (distribution plots, CLT illustration)



Here are a few examples provided. All the above topics are provided in separate chapters with concepts, details, computer applications and instructions, and several worked examples.

Some worked numerical problems

Worked numerical problems across the most important statistical topics used in Data Science, Analytics, and Machine Learning. I'll keep them structured and practical.

---

1. Descriptive Statistics Example

Problem:

A dataset of customer purchase amounts ($):

10, 12, 15, 20, 25, 30, 35

Step 1: Mean

$$\bar{x} = \frac{10 + 12 + 15 + 20 + 25 + 30 + 35}{7} = \frac{147}{7} = 21$$

Step 2: Median

Middle value = **20**

Step 3: Variance and Standard Deviation

Population Variance : $\sigma^2 = \frac{1}{n}\sum(x_i - \mu)^2$

Compute squared deviations:

Value	Deviation	Square
10	-11	121
12	-9	81
15	-6	36
20	-1	1
25	4	16
30	9	81
35	14	196

$$\sigma^2 = \frac{532}{7} \approx 76$$

Population Standard Deviation, $\sigma \approx \sqrt{76} \approx 8.72$

---

2. Probability Example

Problem:

A machine learning classifier correctly predicts 80% of cases.

What is the probability it predicts correctly 3 times in a row?

$$P = 0.8^3 = 0.512 \text{ (assuming independence)}$$

There is a **51.2% chance** of 3 consecutive correct predictions.

---

3. Central Limit Theorem (CLT)

Problem:

Population mean = 100, standard deviation = 20
Sample size = 25

Find the standard error.

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

👉 Interpretation:
Sample means will vary with standard deviation = **4**

---

4. Confidence Interval

Problem:

Sample mean = 50
Standard deviation = 10
Sample size = 100
Confidence level = 95% (z = 1.96)

$$CI = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$
$$= 50 \pm 1.96 \cdot \frac{10}{10} = 50 \pm 1.96$$
$$= (48.04, 51.96)$$

👉 Interpretation:
We are 95% confident the true mean lies in this interval.

---

5. Hypothesis Testing (t-test)

Problem:

A/B test results:

- Group A mean = 100
- Group B mean = 110
- Standard deviation = 15
- Sample size = 30 each

Test if difference is significant.

Step 1: Hypotheses

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

Step 2: Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

$$t = \frac{100 - 110}{\sqrt{\frac{225}{30} + \frac{225}{30}}} = \frac{-10}{\sqrt{7.5 + 7.5}} = \frac{-10}{\sqrt{15}} = \frac{-10}{3.87} \approx -2.58$$

Step 3: Decision

- Critical value ≈ ±2.04 (95% confidence)
- Since **$|t| > 2.04$ → Reject H_0**

👉 Conclusion:
The difference is **statistically significant**.

---

6. Linear Regression Example

Problem:

Fit a line to data:

X	Y
1	2

X	Y
2	4
3	5
4	4
5	5

Model:

$$y = mx + b$$

$$m$$

$$b$$

-10-8-6-4-2246810-10-5510y-interceptx-intercept

Using formulas:

- Slope (m) ≈ 0.6
- Intercept (b) ≈ 2.2

Final Model:

$$y = 0.6x + 2.2$$

👉 Interpretation:
For each unit increase in X, Y increases by **0.6**

---

7. Bayesian Update Example

Problem:

Spam detection:

- Prior probability of spam = 0.3
- Probability of word “free” in spam = 0.8
- Probability of word “free” in non-spam = 0.2

Find probability email is spam given word “free”.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$P(A)$$

8. Correlation Example

Problem:

Given paired data, correlation coefficient r = 0.85

Interpretation:

- Strong positive relationship
- Useful for feature selection in ML

Final Takeaways

These examples and problems demonstrate how statistics directly supports:

- 📊 Data understanding → Mean, variance
- 🎯 Prediction → Regression
- 📏 Decision making → Hypothesis testing
- 🔁 Learning from data → Bayesian inference
- 📈 Model reliability → Confidence intervals

To gain a better understanding of the above topics, please refer to individual chapters in subsequent sections