People's democratic republic of Algeria
Ministry of Higher Education and Scientific Research
Ibn Khaldoun University of Tiaret
Faculty of Applied Sciences
Department of Science and Technology
Field: Science and Technology
Branch: Science and Technology
Specialization: Science and Technology

COURSE MATERIAL

Probabilities and Statistics Course

For the 1st year Common Core Engineer in Science and Technology

Prepared by:

Dr. CHEHDA Rabeh ***Associate Professor Class B***

Reviewed by:

Name	*Title*	*Affiliation*
Pr. SAHLI Belgacem	***Professor***	***Ibn Khaldoun University of Tiaret***
Dr. BENYOUSSEF Soufiane	***Associate Professor Class A***	***Ibn Khaldoun University of Tiaret***

Academic Year: 2025 / 2026

SEMESTRE	Intitulé de la matière		Coefficient	Crédits	Code
S1	Probabilités et statistiques		2	2	IST1.3
VHH	**Cours**	**Travaux dirigés**	**Travaux Pratiques**		
45h00	1h30	1h30	-		

Pré requis :

Aucun

Objectifs:

— Elaborer l'étude complète d'un caractère aléatoire.
— Mettre en évidence un lien éventuel entre deux caractères aléatoires —
Initiation au calcul élémentaire de probabilités.

Contenu de la matière :

I- Probabilités

Rappels (analyse combinatoire, permutation ....)
Variables aléatoires
Lois dd probabilités discrètes et continues usuelles

II- Statistiques

Statistique descriptive

1.1 Statistique descriptive à une dimension
1.2 Statistique descriptive à deux dimensions

Estimation

2.1 Echantillonnage, théorèmes fondamentaux et principe
2.2 Estimation ponctuelle
2.3 Estimation par intervalle
2.4 Estimation ponctuelle et par intervalle d'une moyenne
2.5 Estimation ponctuelle et par intervalle d'une variance
2.6 Estimation ponctuelle et par intervalle d'une proportion
2.7 Marge d'erreur et taille d'échantillon requise

Tests statistiques (un seul échantillon) 3.1
Principe des tests d'hypothèses 3.2 Comparaison
d'une moyenne à une valeur donnée
3.3 Comparaison d'une variance à une valeur donnée
3.4 Comparaison d'une proportion à une valeur donnée
3.5 Seuil descriptif du test
3.6 Risques et courbe d'efficacité
3.7 Test d'ajustement – Test du Khi-Deux

Tests statistiques (plusieurs échantillons)

4.1 Principe des tests

4.2 Comparaison de deux variances
4.3 Comparaison de deux moyennes
4.4 Autres tests sur les moyennes
4.5 Comparaison de deux proportions
4.6 Test d'indépendance – Test du Khi-Deux
4.7 Tests d'homogénéité de plusieurs populations – Test du Khi-Deux

Mode d'évaluation:

Interrogation écrite, devoir surveillé, examen final,

Références bibliographiques:

— A.HAMON, Statistique descriptive : exercices corrigés, P U R, 2008
— A REBBOUH, Statistique descriptive et calculs de probabilités, HOUMA, 2009
— A OUKACHA, Statistique descriptive et calcul de probabilités, 2010
— D J MERCIER, Cahiers de mathématiques du supérieur, vol 1, 2010
SERIE S CHAUM, Théorie et applications de la statistique, 1991

Probabilities and statistics syllabus translation

SEMESTER | **Course Title** | **Coefficient** | **Credits** | **Code**
S1 | *Probability and Statistics* | 2 | 2 | IST1.3

VHH | **Lectures** | **Tutorials** | **Practical Work**
45h00 | 1h30 | 1h30 | –

Prerequisites:
None

Objectives:

— Conduct a complete study of a random variable.
— Highlight a possible relationship between two random variables.
— Introduction to basic probability calculations.

Course Content:

I – Probability

Review (combinatorial analysis, permutation, etc.)
Random variables
Common discrete and continuous probability laws

II – Statistics

Descriptive Statistics

1.1 One-dimensional descriptive statistics
1.2 Two-dimensional descriptive statistics

Estimation

2.1 Sampling, fundamental theorems and principles
2.2 Point estimation
2.3 Interval estimation
2.4 Point and interval estimation of a mean
2.5 Point and interval estimation of a variance

2.6 Point and interval estimation of a proportion
2.7 Margin of error and required sample size

Statistical Tests (single sample)

3.1 Principles of hypothesis testing
3.2 Comparison of a mean to a given value
3.3 Comparison of a variance to a given value
3.4 Comparison of a proportion to a given value
3.5 Test significance level
3.6 Risks and power curve
3.7 Goodness-of-fit test — Chi-square test

Statistical Tests (several samples)

4.1 Principles of tests

4.2 Comparison of Two Variances

4.3 Comparison of Two Means

4.4 Other Tests on Means

4.5 Comparison of Two Proportions

4.6 Test of Independence – Chi-Square Test

4.7 Tests of Homogeneity of Several Populations – Chi-Square Test

---

Assessment Method:

Written quiz, supervised assignment, final exam.

Bibliographical References:

— A. HAMON, *Statistique descriptive : exercices corrigés*, P U R, 2008

— A. REBBOUH, *Statistique descriptive et calculs de probabilités*, HOUMA, 2009

— A. OUKACHA, *Statistique descriptive et calcul de probabilités*, 2010

— D. J. MERCIER, *Cahiers de mathématiques du supérieur*, vol 1, 2010

SERIE S CHAUM, *Théorie et applications de la statistique*, 1991

Program Track: Common core GE GM GC
Institution: UMAB
Academic Year: 2025–2026

Semester: 1

UEF 1.1.1
Subject 3: Probability & Statistics (VHS: 45h00, Lecture: 1h30, Tutorial: 1h30)

Table of Contents

Evaluation Method

Continuous assessment: 40%
Final exam: 60%

Objectives of the Course

This module introduces students to the core principles of probability and statistical methods, focusing on: single and dual variable data series, probability in finite settings, and the concept of random variables.

Recommended prior knowledge:
Fundamentals of programming covered in Mathematics 1 and Mathematics 2.

Course Information

- **Semester:** 1
- **Subject:** Probabilities and Statistics
- **Code:** IST1.3
- **Coefficient:** 2
- **Credits:** 2
- **Total Hours:** 45h00
- **Lecture:** 1h30 per week
- **Tutorials:** 1h30 per week
- **Practical Work:** None

Prerequisites: None

PART I: PROBABILITY THEORY

Introduction

This domain constitutes the fundamental core of data science, machine intelligence, scientific inquiry, and evidence-based decision-making in an uncertain world. The capacity to model, quantify, and analyze randomness is a critical skill across numerous fields, from finance and engineering to healthcare and artificial intelligence.

This course is designed to deliver a profound and intuitive grasp of both the theoretical underpinnings and practical implementations of probability theory. We begin by constructing the mathematical apparatus required for systematic counting and arrangement of objects in **Counting Methods**. This groundwork is essential for computing probabilities in complex situations where simple listing is infeasible.

Equipped with these tools, we delve into the heart of **Probability Fundamentals**, where we formalize concepts of experiments, outcomes, and events. We will explore the axioms governing probability and the powerful theorems derived from them, establishing a rigorous structure for reasoning under uncertainty.

The exploration continues with **Dependence and Independence**, concepts crucial for understanding inter-event relationships. Here, we introduce Bayes' Rule, a profound result enabling belief updates in light of new evidence, with extensive applications in diagnostics, machine learning, and legal analysis.

We then connect abstract theory with quantifiable measures by introducing **Random Quantities**. This potent concept allows translation of probabilistic outcomes into numerical values, facilitating computation of averages, variances, and other descriptive metrics. We will thoroughly examine both **Discrete and Continuous Probability Models**, each representing different categories of real-world phenomena, from counts of system failures to event timing and measurement inaccuracies.

Upon completion, you will possess a robust conceptual and practical understanding of probability theory. You will be prepared to model uncertain processes, make probabilistic forecasts, and establish the groundwork for subsequent study in statistical inference, where populations are understood from data samples. Let us begin this journey into the mathematics of uncertainty.

Definition: The branch of mathematics concerned with quantifying uncertainty, analyzing random phenomena, and calculating the likelihood of events.

Chapter 1: Combinatorial Analysis

Definition:

The mathematical study of counting, arranging, and selecting objects, providing the foundational techniques for calculating probabilities in finite sample spaces.

Historical Context:

Initially developed to address 17th-century gambling problems, it now underpins probability theory, computer science, and operational research.

Key Applications:

- Computing probabilities in games of chance
- Developing computer algorithms
- Optimizing network pathways
- Cryptography and security protocols
- Genetic sequence analysis

1.1 Fundamental Principle of Counting

Definition:

A core rule stating that if one operation can be done in *m* ways and a second independent operation in *n* ways, the total number of ways to perform the sequence of both operations is *m* × *n*. It is the basis for counting possibilities in multi-stage processes.

$$N = n_1 \times n_2 \times n_3 \times \cdots \times n_k$$

Detailed Explanation:

This principle forms the foundation of combinatorial mathematics. It applies when operations are independent - the outcome of one doesn't affect the possibilities of the next. The principle extends to any number of sequential operations by multiplying all possibilities at each step.

Example 1:

Smartphone Configuration

- Choose from 4 different colors (black, white, blue, gold)
- Select from 3 storage capacities (64GB, 128GB, 256GB)
- Pick from 2 connectivity options (WiFi only, WiFi+Cellular)
- **Total configurations** = 4 × 3 × 2 = 24 different smartphone models

Example 2:

Restaurant Meal Combination

- Appetizer: 4 choices (spring rolls, salad, soup, bruschetta)
- Main course: 5 choices (pasta, steak, chicken, fish, vegetarian)
- Dessert: 3 choices (ice cream, cake, fruit platter)
- Beverage: 6 choices (water, soda, juice, iced tea, coffee, wine)
- **Total meal combinations** = 4 × 5 × 3 × 6 = 360 complete meal options

1.2 Factorial Notation

Definition:

The mathematical expression n! (n factorial), defined as the product of all positive integers from 1 to *n*. It represents the number of ways to arrange *n* distinct objects in order.

Detailed Explanation:

Factorials grow extremely rapidly and represent the number of ways to arrange n distinct objects in sequence. They're fundamental in permutations and combinations. The recursive property n! = n × (n-1)! makes calculations efficient.

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Example 1:

Music Playlist Arrangement

- Creating a playlist of 7 distinct songs
- Number of possible arrangements :
  7! = 5,040 different playlists
- If adding an 8th song:
  8! = 40,320 arrangements (showing rapid growth)

Example 2:

Student Presentation Order

- 6 students need to present their projects
- Possible presentation orders :
  6! = 720 different sequences
- Probability of any specific order being randomly selected :
  1/720 ≈ 0.00139

1.3 Permutations

Definition:

Ordered arrangements of a set of objects where the sequence is important.

Detailed Explanation:

There are two main types: permutations without repetition (arrangements of distinct objects) and permutations with repetition (arrangements where some objects are identical). The distinction is crucial for accurate counting.

a) Permutations Without Repetition

Formula:

$$P_n = n!$$

for selecting and arranging r objects from n distinct objects

Example 1:

Student Council Elections

- 5 candidates running for 5 distinct positions
- P(5) = 5 × 4 × 3 × 2 × 1 = 120 different leadership teams
- Each position order matters without repetition

b) Permutations With Repetition Definition:

Arrangements of objects where some items are identical and indistinguishable from one another.

Formula:

$$P = \frac{n!}{n_1! \times n_2! \times \cdots \times n_k!}$$

when there are repeated objects

Example 1:

Arranging Letters in "STATISTICS"

- S: 3, T: 3, A: 1, I: 2, C: 1 → Total letters = 10
- Arrangements :
  10!/(3! × 3! × 1! × 2! × 1!) = 50,400 distinct arrangements
- This accounts for identical letters being indistinguishable

1.4 Arrangements

Definition:

Ordered selections of a subset of objects from a larger set.

Detailed Explanation:

Arrangements differ from permutations in that we may not use all available objects. With repetition allowed, the counting follows exponential growth.

a) Arrangements Without Repetition Definition:

Selecting and ordering *r* distinct objects from *n* available.

Formula:

$$P(n,r) = n \times (n-1) \times (n-2) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$

Example 1:

Security Code Creation

- Creating a 4-digit code from digits 1-9 without repetition
- P(9,4) = 9 × 8 × 7 × 6 = 3,024 possible codes
- This is more secure than allowing repeated digits

b) Arrangements With Repetition Definition: Selecting and ordering *r* objects from *n* types, where each selection can be of any type independently.

Formula:

$$P_{\text{rep}}(n,r) = n^r$$

Example 1:

Custom License Plates

- Creating 3-character plates using letters A-Z only
- 26^3 = 17,576 possible plates
- If adding digits 0-9:
  36^3 = 46,656 plates

1.5 Combinations

Definition:

Selections of objects where the order does not matter, only the group's composition.

Detailed Explanation:

Combinations count subsets rather than sequences. The binomial coefficient ($^{n}_{r}$) represents the number of ways to choose r objects from n without regard to order.

a) Combinations Without Repetition Definition:

Selecting *r* distinct objects from *n* without regard to order.

Formula:

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

Example 1:

Research Team Selection

- Selecting 4 researchers from 12 applicants for a project team
- C(12,4) = 495 possible teams
- Order doesn't matter since all team members have equal roles

Example 2:

Pizza Topping Selection

- Choosing 3 toppings from 8 available options
- C(8,3) = 56 different pizza combinations
- Customers don't care about the order toppings are listed

b) Combinations With Repetition Definition:

Selecting *r* objects from *n* types, where you can choose multiple of the same type and order is irrelevant.

Formula:

$$C_{\text{rep}}(n,r) = \binom{n+r-1}{r} = \frac{(n+r-1)!}{r!\,(n-1)!}$$

Example 1:

Fruit Basket Assembly

- Creating a basket with 6 fruits from 4 types (apples, oranges, bananas, grapes)
- C_rep(4,6) = $\binom{4+6-1}{6}$ = $\binom{9}{6}$ = 84 different baskets
- This allows multiple fruits of the same type

Chapter 2: Random Variables

Definition:

A function that assigns a numerical value to each outcome in a sample space of a random experiment.

2.1 Definitions and Properties

Detailed Explanation:

Random variables bridge probability theory and real-world measurements. They can be discrete (countable values), continuous (measurable quantities), or mixed. The key property is that we can assign probabilities to their possible values.

Types of Random Variables:

- **Discrete RV:** Takes finite or countably infinite values (number of customers)
- **Continuous RV:** Takes values in intervals (temperature, time)
- **Qualitative RV:** Categorical data (colors, brands)
- **Mixed RV:** Combination of discrete and continuous components

Example 1:

Discrete Random Variable

- X = number of software bugs found during code review
- Possible values: 0, 1, 2, 3, ... (non-negative integers)
- Each value has an associated probability

Example 2:

Continuous Random Variable

- Y = daily rainfall amount in millimeters
- Possible values: any real number ≥ 0
- We calculate probabilities for intervals rather than exact values

2.2 Probability Distribution Function

Definition:

A description that gives the probabilities of occurrence of different possible outcomes for a random variable.

Detailed Explanation:

For discrete variables, we use probability mass functions (PMF). For continuous variables, we use probability density functions (PDF). The total probability must always sum/integrate to 1.

a) Discrete Case (Probability Mass Function - PMF) Definition:

The function that gives the probability that a discrete random variable is exactly equal to some value.

Properties:

$f(x_i) = P(X = x_i) \geq 0$,

$\sum f(x_i) = 1$

New Example 1:

Website Traffic Analysis

- X = number of visitors per hour to an e-commerce site
- Observed distribution:
  - P(X=0) = 0.05 (slow hours)
  - P(X=1) = 0.15
  - P(X=2) = 0.30
  - P(X=3) = 0.25
  - P(X=4) = 0.15
  - P(X=5) = 0.10 (peak hours)
- Verification: 0.05 + 0.15 + 0.30 + 0.25 + 0.15 + 0.10 = 1.00

b) Continuous Case (Probability Density Function - PDF)

Definition:

The function whose value at any given point can be interpreted as the relative likelihood that a continuous random variable takes that value (with probabilities given by areas under the curve).

Properties:

- $\forall x \in \mathbb{R}; f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$

Example 1:

Battery Charging Time

- PDF: $f(x) = 0.2\, e^{-0.2x}$ for $x \geq 0$ (charging time in hours)
- Probability phone charges between 2 and 4 hours:

$P(2 \leq X \leq 4) = \int_2^4 0.2\, e^{-0.2x}\, dx = e^{-0.4} - e^{-0.8} \approx 0.670 - 0.449 = 0.221$

2.3 Cumulative Distribution Function (CDF)

Definition:

A function F(x) that gives the probability that a random variable X will take a value less than or equal to x.

Detailed Explanation:

CDFs provide a complete description of a random variable's distribution. They're always non-decreasing, right-continuous functions ranging from 0 to 1. CDFs work for both discrete and continuous variables.

Example 1:

Discrete CDF (Customer Arrivals)

At a coffee shop, number of customers in 15-minute intervals:

- $P(X=0) = 0.10$, $P(X=1) = 0.25$, $P(X=2) = 0.40$, $P(X=3) = 0.20$, $P(X=4) = 0.05$

CDF values:

- $F(0) = P(X \leq 0) = 0.10$
- $F(1) = P(X \leq 1) = 0.10 + 0.25 = 0.35$
- $F(2) = P(X \leq 2) = 0.35 + 0.40 = 0.75$
- $F(3) = P(X \leq 3) = 0.75 + 0.20 = 0.95$
- $F(4) = P(X \leq 4) = 0.95 + 0.05 = 1.00$

Example 2:

Continuous CDF

For $f(x) = 3x^2$, $0 \leq x \leq 1$ (device failure time distribution):

- $F(x) = \int_0^x 3t^2\, dt = [t^3]_0^x = x^3$ for $0 \leq x \leq 1$
- Probability device fails before 0.6 years:
  $F(0.6) = 0.6^3 = 0.216$

- Probability lasts between 0.3 and 0.8 years:
  F(0.8) - F(0.3) = 0.512 - 0.027 = 0.485

2.4 Mathematical Expectation and Variance

Definition:

- **Mathematical Expectation (Mean/Expected Value):**
  The long-run average value of repetitions of the experiment it represents; a measure of central tendency, denoted E[X].
- **Variance:**
  A measure of the dispersion or spread of a set of values around the mean, quantifying how much the values differ from the expected value, denoted Var(X).

Detailed Explanation:

Expectation is a measure of central tendency, while variance quantifies variability. Low variance indicates values cluster tightly around the mean; high variance suggests wide dispersion.

Expectation Formulas:

- For DRV: $E(x) = \sum_{i=1}^{n(\infty)} x_i\, f(x_i) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n)$.
- For CRV: $E(x) = \int_{-\infty}^{+\infty} x\, f(x) dx$.

Variance Formula:

$\text{Var}(X) = E(X^2) - [E(X)]^2$

Example 1:

Discrete Expectation (Game Winnings)

A game with possible winnings:

- Win \$10 with probability 0.2
- Win \$5 with probability 0.3
- Win \$1 with probability 0.4
- Lose \$2 with probability 0.1

Expected value:

$E(X) = (10\times0.2) + (5\times0.3) + (1\times0.4) + (-2\times0.1) = 2 + 1.5 + 0.4 - 0.2 = 3.7$ On average, players win \$3.70 per game

Example 2:

Continuous Expectation and Variance

Delivery time distribution: $f(x) = 2/x^3$ for $x \geq 1$ (time in hours)

- $E[X] = \int_1^{+\infty} x \cdot (2/x^3)\, dx = \int_1^{+\infty} 2/x^2\, dx = [-2/x]_1^{\infty} = 2$ hours
- $E[X^2] = \int_1^{+\infty} x^2 \cdot (2/x^3)\, dx = \int_1^{+\infty} 2/x\, dx \rightarrow$ diverges (infinite variance)
- This represents highly unpredictable delivery times

2.5 Moments and Covariance

Definition:

- **Moments:**
  Quantitative measures related to the shape of a distribution's PDF (e.g., mean is the first moment, variance involves the second moment).
- **Covariance:**
  A measure of the joint variability of two random variables, indicating the direction of their linear relationship.

Detailed Explanation:

Moments provide quantitative measures of distribution properties. Covariance indicates whether two variables tend to move together (positive) or oppositely (negative).

Formulas:

- k-th moment: $E(X^k)$
- Covariance: $Cov(X,Y) = E[XY] - E[X]E[Y]$
- Correlation: $\rho = Cov(X,Y)/(\sigma_x \sigma_\gamma)$

New Example 1:

Investment Portfolio Analysis Two tech stocks:

- Stock A:
  E[A] = 12% return, σ_A = 15%
- Stock B:
  E[B] = 8% return, σ_B = 10%
- Cov(A,B) = 0.009 (positive covariance)
- Correlation:
  $\rho = 0.009/(0.15 \times 0.10) = 0.60$
- Moderate positive correlation suggests they often move together

Example 2:

Study Habits Analysis

Variables: X = weekly study hours, Y = exam score

- Positive covariance suggests more study hours correlate with higher scores
- Negative covariance would suggest inefficiency in studying
- Zero covariance indicates no linear relationship

Chapter 3: Common Probability Distributions

Definition:

Mathematical functions that describe the likelihood of different outcomes for a random variable.

3.1 Discrete Distributions

Definition:

Probability distributions of discrete random variables.

Bernoulli Distribution - B(p)

Definition:

A distribution over a single binary trial (success/failure).

Detailed Explanation:

The Bernoulli distribution is the simplest discrete distribution, forming the building block for more complex distributions like binomial and geometric.

PMF:

$$P(X = k) = \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

Properties:

- E[X] = p
- Var(X) = p(1-p)
- MGF: M(t) = 1-p + pe^t

Example:

Customer Purchase Decision

- Probability a website visitor makes a purchase: p = 0.03
- X = 1 if purchase made, 0 if no purchase
- E[X] = 0.03 (average purchase rate)
- Var(X) = 0.03×0.97 = 0.0291

Binomial Distribution - B(n,p)

Definition:

Models the number of successes in a fixed number of independent Bernoulli trials.

Detailed Explanation:

The binomial distribution applies when we have a fixed number of independent trials, each with the same success probability.

PMF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\ldots,n$$

Properties:

- E[X] = np
- Var(X) = np(1-p)
- MGF: M(t) = (1-p + pe^t)^n

Example:

Quality Control in Manufacturing

- Sample 20 items from production line
- Historical defect rate: p = 0.02
- X ~ B(20, 0.02) where X = number of defective items
- Probability of exactly 1 defective: $P(X=1) = C(20,1) \times 0.02^1 \times 0.98^{19} \approx 0.272$
- Probability of 2 or more defective: $P(X \geq 2) = 1 - P(X \leq 1) \approx 0.060$

Poisson Distribution - P(λ)

Definition:

Models the number of events occurring in a fixed interval of time or space.

Detailed Explanation:

The Poisson distribution applies when events occur randomly and independently at a constant average rate.

PMF:

$$P(k,\lambda) = P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0,1,2,\ldots$$

So the relation $P(X = k) = f(x)$ defines a discrete random variable X and its probability distribution.

The definition of $E(X)$ gives

$$E(X) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$

Since $\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = e^{\lambda}$

it follows $E(X) = \lambda$

Properties:

- E[X] = λ
- Var(X) = λ
- $\sigma_X = \sqrt{\lambda}$
- MGF: M(t) = e^{λ(e^t - 1)}

Example:

Emergency Calls to Dispatch Center

- Average 4 emergency calls per hour
- Probability of exactly 6 calls in an hour:
  P(X=6) = ($e^{-4} \times 4^6$)/6! ≈ (0.0183 × 4096)/720 ≈ 0.104
- Probability of 2 or fewer calls:
  P(X≤2) ≈ 0.238

Geometric Distribution

Definition:

Models the number of trials needed to achieve the first success.

Detailed Explanation:

The geometric distribution has the memoryless property - the probability of success in future trials doesn't depend on past failures.

PMF:

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1,2,3,\ldots$$

Example:

Job Application Process

- Probability of getting a job offer from each application:
  p = 0.08
- Expected number of applications until first offer:
  E[X] = 1/0.08 = 12.5
- Probability first offer comes on 10th application:
  P(X=10) = $(0.92)^9 \times 0.08 \approx 0.038$

Negative Binomial Distribution

Definition:

Models the number of trials needed to achieve r successes.

Detailed Explanation:

Generalization of geometric distribution for multiple successes.

PMF:

$$P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}, \quad k = r, r + 1, \ldots$$

Properties:

- E[X] = r/p
- Var(X) = r(1-p)/p^2

Example:

Sales Target Achievement

- Salesperson needs 3 successful sales
- Success probability per call:
  p = 0.15
- Expected total calls:
  E[X] = 3/0.15 = 20 calls
- Probability 3rd sale occurs on 25th call:
  P(X=25) = $\binom{24}{2} \times 0.15^3 \times 0.85^{22} \approx 0.065$

3.2 Continuous Distributions

Definition:

Probability distributions of continuous random variables.

Uniform Distribution - U[a,b]

Definition:

All intervals of the same length have equal probability.

Detailed Explanation:

The uniform distribution represents complete randomness within bounds. It's often used in simulation and random number generation.

PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Properties:

- E[X] = (a+b)/2
- Var(X) = $(b-a)^2/12$
- $\sigma_X = \sqrt{\frac{(b-a)^2}{12}}$

Example:

Bus Arrival Times

- Buses arrive every 15-25 minutes uniformly
- a = 15, b = 25 minutes
- E[X] = (15+25)/2 = 20 minutes average wait
- Probability wait between 18-22 minutes:
  P(18≤X≤22) = (22-18)/(25-15) = 0.4

Exponential Distribution - Exp(λ)

Definition:

Models the time between events in a Poisson process; it is memoryless.

Detailed Explanation:

The exponential distribution is memoryless and commonly used for waiting times, equipment lifetimes, and radioactive decay.

PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Properties:

- E[X] = 1/λ
- Var(X) = 1/λ²
- $\sigma_X = \frac{1}{\lambda}$

Example:

Customer Service Call Duration

- Average call duration:
  8 minutes (λ = 1/8 = 0.125)
- Probability call lasts more than 15 minutes:
  P(X>15) = e^(-0.125×15) ≈ 0.153
- Probability call lasts between 5-10 minutes:
  P(5≤X≤10) = e^(-0.625) - e^(-1.25) ≈ 0.267

Normal Distribution - N(μ,σ²)

Definition:

A symmetric, bell-shaped distribution defined by its mean (μ) and variance (σ²); fundamental due to the Central Limit Theorem.

Detailed Explanation:

The normal distribution appears naturally in many contexts due to the Central Limit Theorem. It's symmetric and fully described by its mean and variance.

PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard Normal:

with N(0,1):

$$Z = \frac{X - \mu}{\sigma}\ N$$

Empirical Rule:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 0.6827$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.9545$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.9973$

Example:

Product Weight Quality Control

- Cereal boxes labeled 500g, normally distributed with $\sigma = 10$g
- Probability box weighs less than 485g:
  $Z = (485-500)/10 = -1.5$ $P(X<485) = P(Z<-1.5) \approx 0.0668$
- Weight range containing 99% of boxes:
  $P(\mu-2.576\sigma < X < \mu+2.576\sigma) = P(474.24 < X < 525.76)$

Gamma Distribution

Definition:

A two-parameter family of continuous distributions that generalizes the exponential and chi-squared distributions.

Detailed Explanation:

The gamma distribution models waiting times for multiple events and includes exponential and chi-square as special cases.

PDF:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

Properties:

- E[X] = α/β
- Var(X) = α/β²

Example:

Insurance Claim Processing

- Time to process insurance claims follows Gamma(α=3, β=0.5)
- E[X] = 3/0.5 = 6 days average processing time
- Var(X) = 3/(0.5)² = 12 days²

Beta Distribution

Definition:

A family of continuous distributions defined on the interval [0, 1], useful for modeling probabilities and proportions.

Detailed Explanation:

The beta distribution is extremely flexible for modeling bounded quantities and is conjugate to the binomial distribution.

PDF:

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \le x \le 1$$

Properties:

- E[X] = α/(α+β)
- Var(X) = αβ/[(α+β)² (α+β+1)]

Example:

Project Completion Probability

- Based on historical data, project success rate follows Beta(α=8, β=2)
- Expected success probability:
  E[X] = 8/(8+2) = 0.80
- 90% credible interval for success probability

PART II- STATISTICS

Introduction:

Statistics is the **science of extracting meaning from data**—the systematic methodology for turning raw observations into reliable knowledge. In today's data-driven world, this discipline is not merely an academic subject but an **essential professional toolkit** that empowers engineers, scientists, analysts, and decision-makers across every field.

This module represents the bridge between probability theory and practical application. Having established the mathematical language of uncertainty in the previous section, we now learn how to apply that language to **real-world evidence**. Here, we move from abstract mathematical models to the concrete analysis of actual measurements, samples, and observations.

Our Statistical Journey

Our exploration follows the fundamental workflow of statistical reasoning, structured to build your analytical capabilities step by step:

1. **DESCRIPTIVE STATISTICS:** We begin with the art of **data storytelling**. Before we can draw conclusions, we must learn to see and describe what our data shows. This foundational section teaches you to organize, summarize, and visualize data using graphs, charts, and summary measures. You'll learn to answer: *What does this dataset look like? What are its key characteristics?*

2. **ESTIMATION:** From describing what we *have* to inferring what we *haven't seen*. Here we tackle the core challenge of statistics: how to make reliable claims about entire populations when we only have limited samples. You'll master techniques for calculating confidence intervals and point estimates that quantify uncertainty in practical terms.

3. **HYPOTHESIS TESTING:** The engine of scientific decision-making. This is where we formulate and test claims about the world. Whether determining if a new manufacturing process works better or if a medical treatment is effective, these formal procedures provide the structured methodology for making evidence-based decisions amidst uncertainty.

Why This Matters to You

The competencies developed in this module are **directly transferable to professional practice**:

- **Data Literacy & Communication:** Learn to transform complex datasets into clear, actionable insights that can be communicated to stakeholders.

- **Evidence-Based Decision Making:** Move beyond intuition to make strategic choices backed by statistical evidence and quantified risk assessment.
- **Problem-Solving Framework:** Develop a systematic approach to investigating questions, designing studies, and interpreting results that accounts for real-world variability.
- **Foundation for Advanced Analytics:** These statistical principles form the bedrock upon which machine learning, data science, quality control, and experimental design are built.

Statistics is fundamentally a **way of thinking**—a structured approach to navigating uncertainty, validating assumptions, and drawing conclusions from evidence. It's what separates anecdote from evidence, intuition from insight, and guesswork from informed strategy.

As we embark on this section, you're not just learning formulas and procedures; you're acquiring a **problem-solving mindset** that will enable you to extract truth from data, make predictions with quantified confidence, and contribute to innovation and discovery in your field.

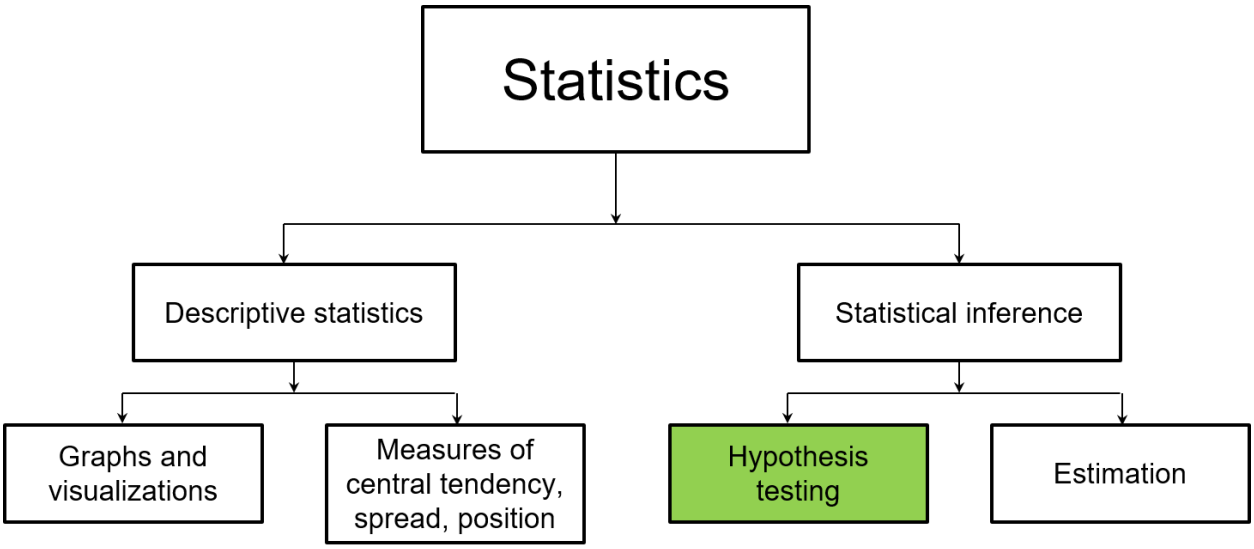

Figure 1: statistical processing

Definition: The discipline concerned with the collection, organization, analysis, interpretation, and presentation of data.

Chapter 1: Descriptive Statistics

Definition:

Methods for summarizing and describing the main features of a dataset, often visually or with summary statistics.

1.1 One-Dimensional Descriptive Statistics

Definition: Analysis of a single variable.

Types of Characteristics (Statistical Variable):

We distinguish two categories of characteristics:

Qualitative characteristics and quantitative characteristics.

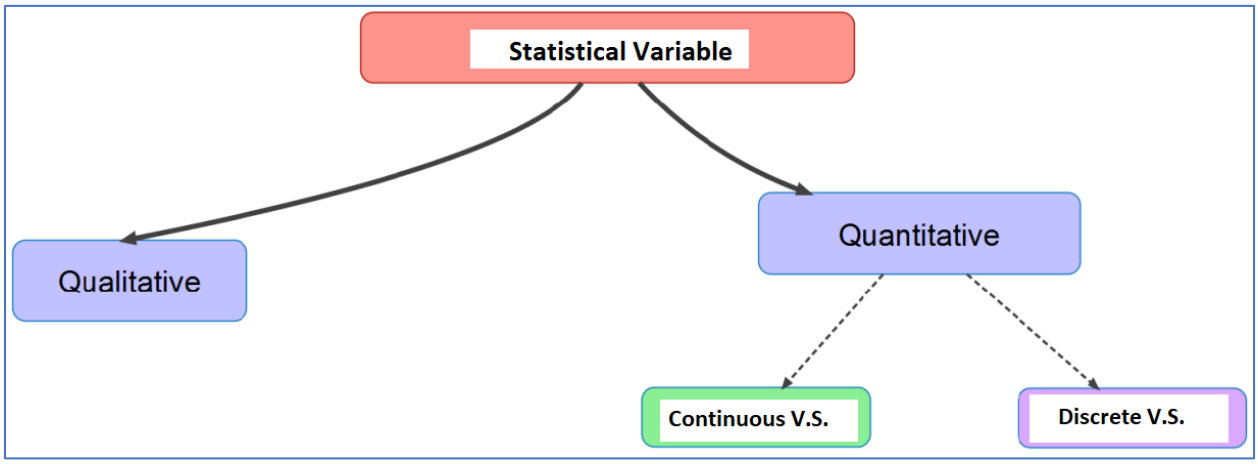

Figure 2: types of statistical variables

Data Presentation and Graphical Representations

Definition:

Techniques like bar charts, histograms, and box plots to visualize data.

Detailed Explanation:

Different data types require different visualization techniques. Categorical data uses bar charts and pie charts, while numerical data uses histograms and box plots. Proper visualization helps identify distributions, outliers, and patterns.

Qualitative Variables:

- **Bar charts:** Compare frequencies across categories

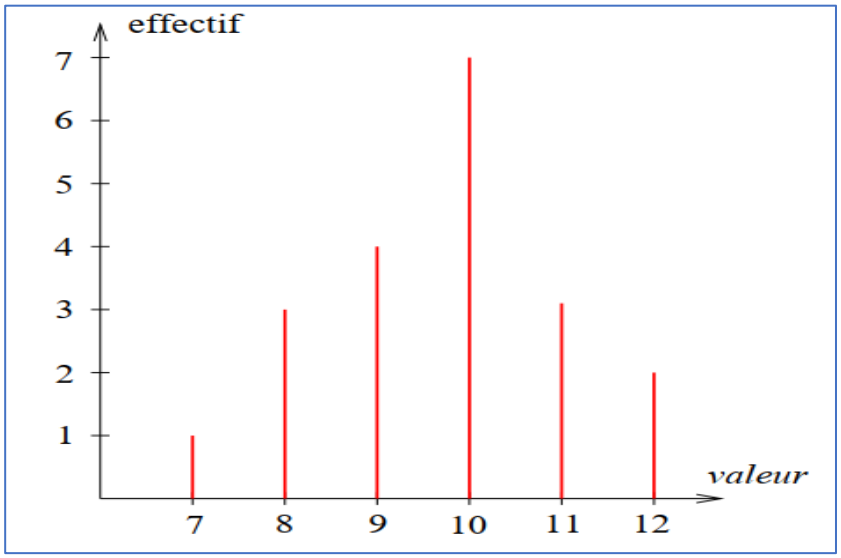

Figure 3*: Bar Chart Diagram*

- **Pipe Organ Diagram:** Compare frequencies across categories

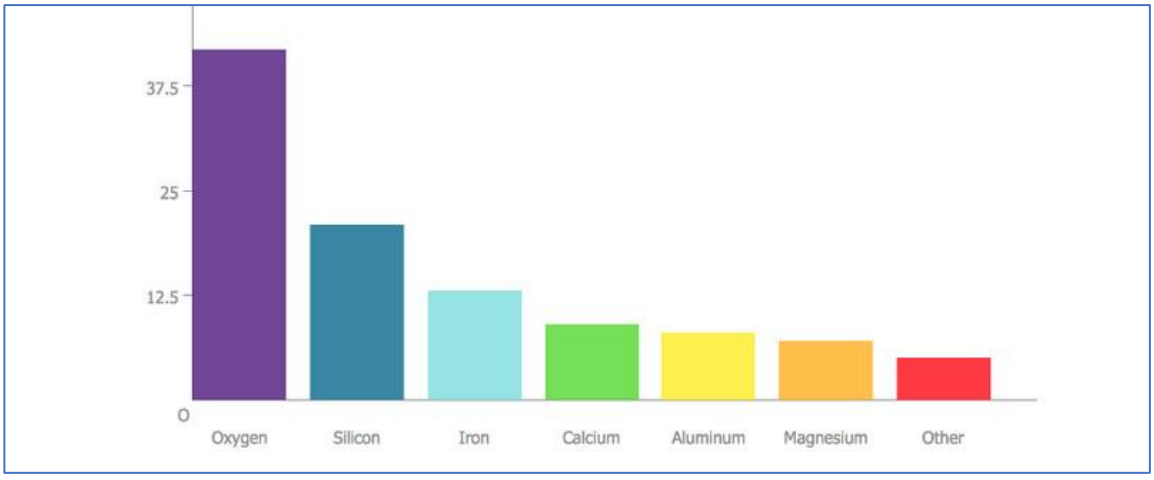

Figure 4*: Pipe Organ Diagram*

- **Pie charts:** Show proportions as parts of a whole

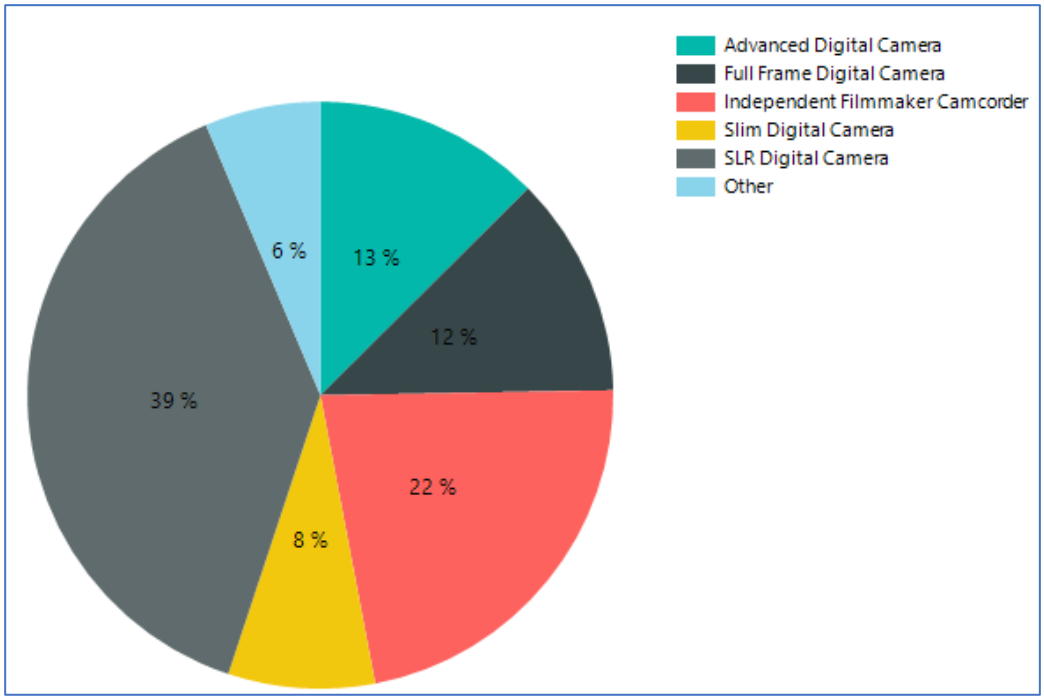

Figure 5: Sector Diagram

Example:

Social Media Platform Preference Survey of 200 students:

- Instagram: 80 students (40%)
- TikTok: 60 students (30%)
- Twitter: 40 students (20%)
- Other: 20 students (10%)
- Pie chart angles: Instagram: 144°, TikTok: 108°, Twitter: 72°, Other: 36°

Quantitative Variables:

- **Histograms:** Show distribution of continuous data

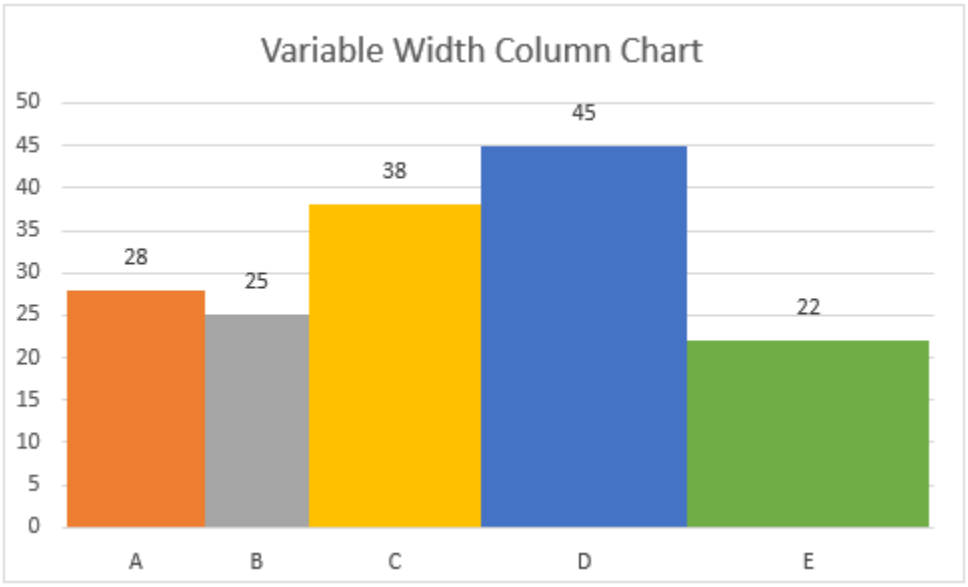

Figure 6*: Frequency Histogram*

- **Box plots:** Display five-number summary and outliers

  The **box plot** (or box-and-whisker diagram) for a statistical series is constructed as follows: (the attribute values are on the horizontal axis - min and max denote the smallest and largest observed values)

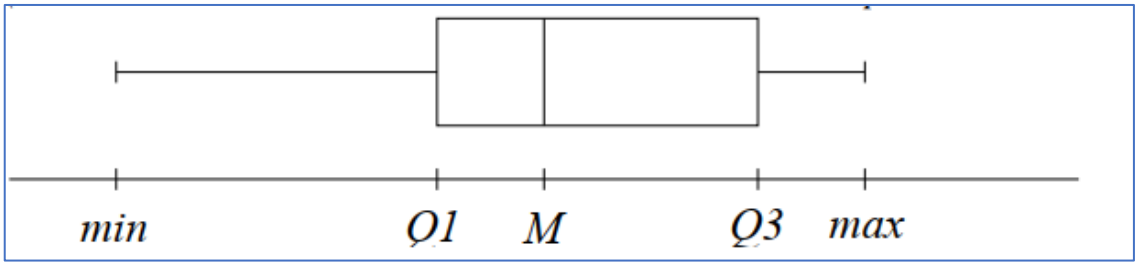

Figure 7: box plot

Example:

Employee Salary Distribution

- 30,000 −40,000: 15 employees
- 40,000 −50,000: 35 employees
- 50,000 −60,000: 42 employees
- 60,000 −70,000: 25 employees
- 70,000 −80,000: 8 employees Histogram shows slightly right-skewed distribution

Location Parameters (Central Tendency)

Definition:

Statistics like mean, median, and mode that identify the center of a data set.

Detailed Explanation:

The mean uses all data points but is sensitive to outliers. The median is robust to outliers but ignores most values' magnitudes. The mode identifies the most frequent value but may not be unique or representative.

a) The Mode (Discrete Case):

The mode of a discrete statistical variable is the value possessing the highest individual count (or the highest proportional frequency) and is represented by M_0

b) The Mode (Continuous Case):

The subsequent definition clarifies the procedure for accurately determining the mode, which resides within a specific class termed the "modal class."

Definition:

We define the **modal class** as the class interval of X values that has the highest individual count (or the highest proportional frequency).

M_0= (midpoint of the modal class): The modal class being the interval containing the highest frequency.

c) The Median:

The **median** is the value that divides the series into two series with the same total frequency.

Depending on whether N is an even or odd total number.

1. Case: N = 2p+1 (odd) therefore the median is the value of order p+1.
2. Case: N = 2p (even) therefore the median is the value of order p and p+1 divided by 2 (the average of the values at positions p and p+1).

d) The Mean:

The mean is calculable for numerical variables, whether discrete or continuous. It is obtained simply by adding all the values and dividing this sum by the number of values. This calculation can be done from raw data or from a frequency table. Here are some calculation examples.

Discrete Variable:

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i x_i}{N}$$

Continuous Variable:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i m_i}{N}$$

with f_i frequency of the characteristic xi (or frequency of the i-th class), m_i midpoint of the i-th class and n the number of individuals in the sample (N is the total frequency).

Example:

Apartment Rental Prices Dataset: 800,850, 900,950, 950,1,000, 1,100,1,200, $2,500

- **Mean:** (800+850+900+950+950+1000+1100+1200+2500)/9 = $1,250
- **Median:** $950 (5th value in ordered list)
- **Mode:** $950 (appears twice)

The mean is inflated by one luxury apartment ($2,500), while median better represents typical pricing.

Dispersion Parameters (Variability)

Definition:

Statistics like range, variance, standard deviation, and interquartile range (IQR) that quantify the spread of data.

Detailed Explanation:

Range is simple but sensitive to extremes. Variance and standard deviation use all data points and have nice mathematical properties. IQR focuses on the middle 50% and is robust to outliers.

a) The Range

The difference between the largest value and the smallest value of the characteristic, given by the quantity

$$E = X_{\max} - X_{\min}$$

Is called the **range** of the statistical variable. The calculation of the range is very simple. It gives a first idea of the dispersion of the observations. It is a very rudimentary indicator.

b) The Variance (S^2 or Var(X)):

We call **variance** of this statistical series X, the number

Discrete Variable:

$$S^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N}$$

Continuous Variable:

$$S^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{N - 1}$$

(Note: The formula for the continuous variable uses class midpoints m_i. The mention $N-1$ in the denominator is unusual for descriptive variance; it's typically for sample variance).

Properties of Variance:

1. $Var(X + a) = Var(X)$
2. $Var(a \times X) = a^2 \times Var(X)$

c) The Standard Deviation

The **standard deviation** is useful when comparing the dispersion of two datasets of similar size that have approximately the same mean. The spread of values around the mean is less important in the case of a dataset with a smaller standard deviation.

The quantity: $\sigma(X) = \sqrt{Var(X)} = \sqrt{S^2}$

d) The Dispersion Parameters Associated with the Median.

Definition:

The fundamental concept is to partition the population into four segments of equal frequency.

Given a statistical series with median M, whose values are arranged in increasing order (the same list used to determine the median).

By dividing this list into two sub-series of equal frequency (Note: when the total frequency is odd, the median value is excluded from both sub-series):

- We define the **first quartile** as the real number denoted **Q1**, equal to the median of the lower sub-series.
- We define the **third quartile** as the real number denoted **Q3**, equal to the median of the upper sub-series.
- The **interquartile range** is given by **Q3 - Q1**.
- |Q1;Q3| is termed the **interquartile interval**.

quantile of order α is:

$$q_\alpha = L_i + \frac{(\alpha N - F_{i-1})}{f_i} \times C$$

with L_i lower boundary of the class containing the quantile of order α (if $\alpha=0.5$ then q_α is the median M_e)

α=0.25 is the quantile 1 (Q1),

α=0.75 is the quantile 3 (Q3),

α=0.1 is the decile 1 (D1), *etc.*…

Same dataset: 800,850, 900,950, 950,1,000, 1,100,1,200, $2,500

- **Range:** $2{,}500 - 800 = \$1{,}700$
- **Variance:** Mean = $^{22}1{,}250 Sum of squared deviations = 3{,}422{,}500 s^2 = 3{,}422{,}500/8 = 427{,}812.50$
- **Standard Deviation:** $\sqrt{427{,}812.50} \approx \$654.08$
- **Quartiles:** $Q1 = 900, Q3 = 1{,}100$
- **IQR:** $1{,}100 - 900 = \$200$

Interpretation:

Standard deviation $\approx 654 means prices typically vary by about 654$ from the mean. The small IQR ($200) suggests most apartments cluster in a narrow range, with one outlier pulling the mean upward.

Stem-and-Leaf Displays Definition:

A method for showing both the rank order and shape of a data set simultaneously.

Detailed Explanation:

Stem-and-leaf plots provide more information than histograms while being quicker to create than full tables. They reveal distribution shape, central tendency, spread, gaps, and outliers.

Example: Exam Score Analysis Test scores out of 100: 56, 62, 67, 71, 73, 74, 76, 78, 79, 81, 82, 83, 84, 85, 86, 88, 89, 91, 92, 95

Stem | Leaf | Frequency
5 | 6 | 1
6 | 2 7 | 2
7 | 1 3 4 6 8 9 | 6
8 | 1 2 3 4 5 6 8 9 | 8
9 | 1 2 5 | 3

Information revealed:

- Shape: Slightly left-skewed
- Typical score: Low 80s
- Spread: 56 to 95 (range = 39 points)
- No apparent gaps or outliers
- Most students scored in the 80s

1.2 Two-Dimensional Descriptive Statistics

Definition: Analysis of the relationship between two variables.

Scatter Plots and Contingency Tables

Definition:

Graphical (scatter plot) and tabular (contingency table) methods for displaying bivariate data.

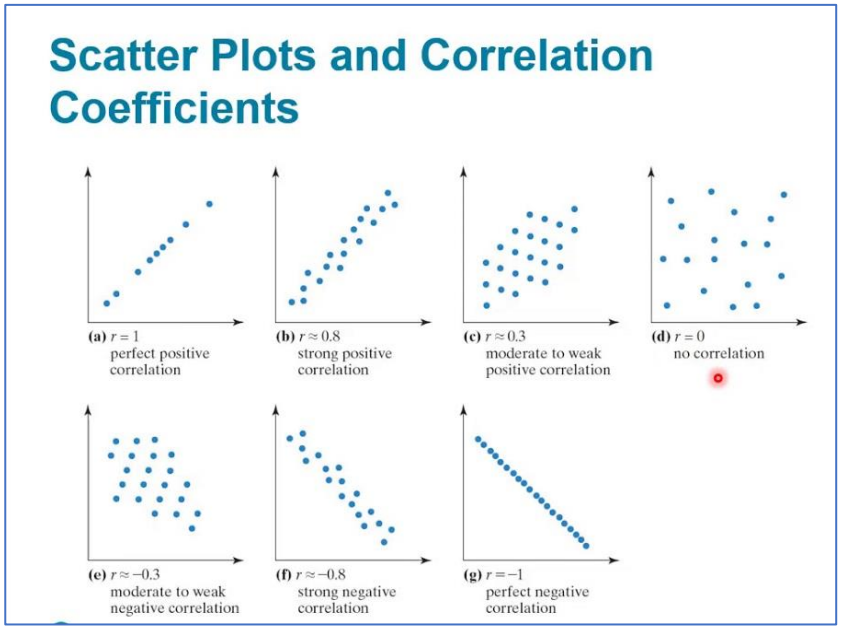

Figure 8: Scatter Plot

Detailed Explanation:

Scatter plots visualize relationships between two quantitative variables. Contingency tables organize relationships between categorical variables, showing joint frequencies.

Example:

Advertising Budget vs Sales Revenue Monthly data for a small business:

Month	Ad Budget ($1000s)	Sales ($10,000s)
Jan	2	4
Feb	3	5
Mar	4	7
Apr	5	8
May	6	9
Jun	7	11
Jul	8	12
Aug	9	13

Scatter plot: Shows strong positive linear relationship

Contingency Table Example: Education level vs smartphone brand preference for 300 consumers:

	Apple	Samsung	Other	Total
College	60	40	20	120
High School	30	50	40	120
No College	10	30	20	60
Total	**100**	**120**	**80**	**300**

Marginal and Conditional Distributions Definition:

Distributions derived from a contingency table—marginal distributions consider one variable ignoring the other, conditional distributions consider one variable given a specific value of the other.

Detailed Explanation:

Marginal distributions are found in table margins (row/column totals). Conditional distributions are calculated by fixing one variable and examining the distribution of the other.

From smartphone preference table:

- **Marginal distribution of Education:** College: 40%, HS: 40%, No College: 20%

- **Conditional distribution of Brand given College:** Apple: 60/120=50%, Samsung: 33%, Other: 17%
- **Conditional distribution of Education given Apple:** College: 60/100=60%, HS: 30%, No College: 10%

Covariance and Correlation

Definition: Covariance measures the direction of a linear relationship; correlation (like Pearson's r) measures both the direction and strength of a linear relationship.

Detailed Explanation:

Covariance can be any real number and its magnitude depends on variable scales. Correlation is standardized between -1 and +1, allowing comparison across different datasets.

Advertising Budget vs Sales calculations:

- **Means:** Budget mean = 5.5, Sales mean = 8.625
- **Covariance:**
  Calculate: (2-5.5)(4-8.625) + (3-5.5)(5-8.625) + ... = 67.5 Cov(X,Y) = 67.5/7 = 9.643
- **Standard deviations:** σ_budget ≈ 2.45, σ_sales ≈ 3.20
- **Correlation:** $\rho = 9.643/(2.45\times3.20) \approx 1.23$

Interpretation:

The calculated correlation > 1 suggests an error in manual calculation (should be ≤ 1). Correct calculation gives $\rho \approx 0.98$, indicating very strong positive linear relationship.

Regression Analysis Definition:

A statistical process for estimating the relationships between a dependent variable and one or more independent variables.

Detailed Explanation:

Regression finds the "best-fit" line that minimizes the sum of squared vertical distances between data points and the line. It's used for prediction and understanding relationships.

Advertising Budget vs Sales:

- **Regression line:** Sales = a + b × Budget
- **Slope:** b = Cov(Budget,Sales)/Var(Budget) = 9.643/6 ≈ 1.607
- **Intercept:** a = 8.625 - 1.607×5.5 ≈ 0.786
- **Equation:** Sales = 0.786 + 1.607 × Budget
-

Prediction examples:

- $10{,}000 adbudget: Sales = 0.786 + 1.607 \times 10 = 16.856 (168{,}560)$
- $15{,}000 adbudget: Sales = 0.786 + 1.607 \times 15 = 24.891 (248{,}910)$

Goodness of fit: $R^2 = \rho^2 \approx 0.96$, indicating the model explains 96% of sales variation.

Functional Fitting

Definition:

The process of finding a mathematical function (linear, exponential, etc.) that best describes the relationship between variables.

Detailed Explanation:

When relationships aren't linear, we transform variables or use nonlinear functions. Common transformations include logarithmic, exponential, and power transformations.

Power fitting: $y = ax^b \rightarrow \ln y = b \ln x + \ln a$ **Exponential fitting:** $y = ae^{bx} \rightarrow \ln y = \ln a + bx$

Example:

Company Growth Pattern

- Startup revenue growth appears exponential
- Transform: ln(revenue) vs time → approximately linear
- Fit line to transformed data: ln(revenue) = 2.1 + 0.15×time
- Convert back: revenue = $e^{2.1} \times e^{0.15 \text{time}} = 8.17 \times e^{0.15 \text{time}}$

Chapter 2: Estimation

Definition: The process of using sample data to infer or approximate population parameters.

2.1 Sampling and Fundamental Theorems

Definition:

- **Sampling:** The selection of a subset of individuals from a statistical population to estimate characteristics of the whole population.
- **Fundamental Theorems:** Core probabilistic theorems like the **Law of Large Numbers** (sample averages converge to the population mean) and the **Central Limit Theorem** (sampling distribution of the mean approaches normality regardless of population distribution).

Detailed Explanation:

Proper sampling ensures representativeness. Random sampling gives each member an equal chance of selection. Sampling distributions describe how statistics vary across different samples.

Key Theorems:

- **Law of Large Numbers:** Sample statistics approach population parameters as sample size increases
- **Central Limit Theorem:** Sampling distribution of means approaches normality regardless of population distribution

Example:

Political Opinion Polling

- Population: All eligible voters in a state
- Sample: 1,200 randomly selected voters
- Sample proportion favoring Candidate A: 48%
- According to LLN, this approaches true population proportion as sample size grows
- According to CLT, sampling distribution of proportion is approximately normal

2.2 Point Estimation

Definition:

The use of a single value (statistic), calculated from sample data, as a "best guess" for an unknown population parameter.

Detailed Explanation:

Point estimators should be unbiased (correct on average), consistent (improve with larger samples), and efficient (small variance).

Properties of good estimators:

- **Unbiasedness:** $E[\hat{\theta}] = \theta$
- **Consistency:** $\hat{\theta} \to \theta$ as $n \to \infty$
- **Efficiency:** Small variance compared to other estimators

Example:

Average Household Income

- Population: All households in a city
- Sample: 250 randomly selected households
- **Point estimate:** Sample mean $\bar{X} = \$72,500$
- This is our best single number estimate for the true average income
- We prefer this over median for this purpose due to mathematical properties

2.3 Interval Estimation

Definition:

The use of a range of values (confidence interval), derived from sample data, that is likely to contain the true population parameter with a specified level of confidence.

Detailed Explanation:

Confidence intervals provide a range rather than a single value, with an associated confidence level. A 95% CI means if we repeated the sampling process many times, 95% of such intervals would contain the true parameter.

Example:

Average Household Income (continued)

- 95% CI for mean income: $(69{,}200, 75{,}800)$
- **Interpretation:** We're 95% confident the true average household income is between $69{,}200 and 75{,}800$
- **Not interpretation:** There is not a 95% probability that the true mean is in this specific interval

2.4 Estimation of a Mean

Definition:

The process of calculating a point estimate (sample mean) and/or confidence interval for a population mean (μ).

Detailed Explanation:

The sample mean is the best point estimator for the population mean. For interval estimation, we use the t-distribution when population variance is unknown, and normal distribution when it's known.

Formulas:

- **Point estimate:** Sample mean $\bar{X}$
- **Interval estimate:** $\bar{X} \pm t^*(s/\sqrt{n})$ [σ unknown]

Example:

Factory Production Rate

- Sample of 36 production days: $\bar{X} = 245$ units/day, $s = 28$ units
- 95% confidence interval, $t^*(35) \approx 2.030$
- **Standard error:** $SE = 28/\sqrt{36} \approx 4.667$
- **Margin of error:** $E = 2.030 \times 4.667 \approx 9.474$
- **95% CI:** $245 \pm 9.474 = (235.53, 254.47)$
- **Interpretation:** 95% confident true mean daily production is 236-254 units

2.5 Estimation of a Variance

Definition:

The process of calculating a point estimate (sample variance) and/or confidence interval for a population variance (σ^2).

Detailed Explanation:

The sample variance s^2 is used to estimate population variance σ^2. The chi-square distribution is used for interval estimation because $(n-1)s^2/\sigma^2$ follows a chi-square distribution.

Formulas:

- **Point estimate:** Sample variance s^2
- **Interval estimate:** $[(n-1)s^2/\chi^2\_upper, (n-1)s^2/\chi^2\_lower]$

Example:

Product Weight Consistency

- Sample of 25 products: $s^2 = 6.4$ grams²
- 95% CI for variance, df = 24, $\chi^2\_upper = 39.36$, $\chi^2\_lower = 12.40$
- **95% CI:** $[(24\times6.4)/39.36, (24\times6.4)/12.40] = [3.90, 12.39]$
- **Interpretation:** 95% confident true variance is between 3.90 and 12.39 grams²
- This helps quality control determine if process variability is acceptable

2.6 Estimation of a Proportion

Definition:

The process of calculating a point estimate (sample proportion) and/or confidence interval for a population proportion (p).

Detailed Explanation:

The sample proportion $\hat{p}$ estimates the population proportion p. For large samples, the sampling distribution is approximately normal. The Wilson score interval or plus-four adjustment improves small-sample performance.

Formulas:

- **Point estimate:** Sample proportion $\hat{p}$
- **Interval estimate:** $\hat{p} \pm z^*\sqrt{[\hat{p}(1-\hat{p})/n]}$

Example:

Customer Satisfaction Survey

- 400 customers surveyed, 312 satisfied
- $\hat{p} = 312/400 = 0.78$ (78% satisfaction rate)
- 95% CI: $z^* = 1.96$
- **Standard error:** $SE = \sqrt{[0.78(1-0.78)/400]} \approx 0.0207$
- **Margin of error:** $E = 1.96 \times 0.0207 \approx 0.0406$
- **95% CI:** $0.78 \pm 0.0406 = (0.7394, 0.8206)$
- **Interpretation:** 95% confident true satisfaction rate is 73.9% to 82.1%

2.7 Margin of Error and Sample Size

Definition:

- **Margin of Error:** The maximum expected difference between the sample statistic and the population parameter, often used in confidence intervals.

- **Sample Size Determination:** The process of calculating the minimum number of observations required to achieve a desired margin of error and confidence level for an estimate.

Detailed Explanation:

Margin of error quantifies estimate uncertainty. Sample size calculations ensure studies have sufficient power and precision before data collection begins.

Formulas:

- **For means:** $n = (z^*\sigma/E)^2$
- **For proportions:** $n = p(1-p)(z^*/E)^2$

Example 1:

Employee Satisfaction Study

- Want margin of error E = 0.03 (3%) for satisfaction proportion
- Estimated p = 0.70 (from pilot study), 95% confidence
- **Required n** = $0.70(1-0.70)(1.96/0.03)^2 = 0.21 \times (65.33)^2 \approx 897$ employees

Example 2:

Product Weight Study

- Want margin of error E = 2 grams for mean weight
- Estimated $\sigma = 8$ grams (from previous studies), 95% confidence
- **Required n** = $(1.96 \times 8 / 2)^2 = (7.84)^2 \approx 61.5 \rightarrow 62$ products

Chapter 3: Statistical Tests (Single Sample)

Definition:

Formal procedures that use sample data to evaluate a claim (hypothesis) about a population parameter.

3.1 Hypothesis Testing Principles

Definition:

The foundational concepts, including null hypothesis (H_0), alternative hypothesis (H_1), test statistic, p-value, and significance level (α).

Detailed Explanation:

Hypothesis testing follows a structured process: state hypotheses, collect data, calculate test statistic, determine p-value, and make decision. The null hypothesis represents the status quo or no effect, while the alternative represents what we're trying to prove.

Key Components:

- **Null hypothesis (H_0):** Default position, no effect/difference
- **Alternative hypothesis (H_1):** Research hypothesis, effect exists
- **Test statistic:** Measures compatibility with H_0
- **p-value:** Probability of results as extreme as observed if H_0 true
- **Significance level (α):** Threshold for rejecting H_0 (usually 0.05)

Example:

New Teaching Method Evaluation

- H_0: New method has same effectiveness as traditional ($\mu = \mu_0$)
- H_1: New method improves test scores ($\mu > \mu_0$)
- Collect test score data from both methods
- Calculate test statistic comparing means
- If p-value < 0.05, reject H_0 and conclude new method is better

3.2 Test for a Mean

Definition:

A hypothesis test (e.g., one-sample t-test) to determine if a population mean differs from a hypothesized value.

Detailed Explanation: The one-sample t-test compares a sample mean to a hypothesized population mean. It's robust to minor violations of normality, especially with larger samples.

Example:

Coffee Shop Claim

- Claim: Average waiting time is 3 minutes
- Sample: 40 customers, $\bar{X} = 3.8$ minutes, $s = 1.2$ minutes
- Test if actual wait time exceeds 3 minutes ($\alpha = 0.05$)

Hypotheses:

- H_0: $\mu = 3$ minutes
- H_1: $\mu > 3$ minutes (one-tailed test)

Test statistic:

- $t = (3.8 - 3) / (1.2/\sqrt{40}) = 0.8 / 0.19 \approx 4.21$
- Critical value: $t(39, 0.05) \approx 1.685$

Decision: Since $4.21 > 1.685$, reject H_0

Conclusion: Evidence suggests average wait time exceeds 3 minutes

3.3 Test for a Variance

Definition:

A hypothesis test (using the chi-square distribution) to determine if a population variance differs from a hypothesized value.

Detailed Explanation:

The chi-square test for variance uses the fact that $(n-1)s^2/\sigma^2$ follows a chi-square distribution. It's used in quality control to test process consistency.

Example:

Manufacturing Process Control

- Claim: Process variance $\sigma^2 = 25$
- Sample: 18 items, $s^2 = 36$
- Test if variance has increased ($\alpha = 0.05$)

Hypotheses:

- H_0: $\sigma^2 = 25$

- $H_1: \sigma^2 > 25$

Test statistic:

- $\chi^2 = (n-1)s^2/\sigma^2 = 17 \times 36/25 = 612/25 = 24.48$
- Critical value: $\chi^2(17, 0.05) \approx 27.59$

Decision: Since $24.48 < 27.59$, fail to reject H_0

Conclusion: No significant evidence that variance has increased

3.4 Test for a Proportion

Definition:

A hypothesis test (e.g., one-sample z-test for proportions) to determine if a population proportion differs from a hypothesized value.

Detailed Explanation:

The one-sample z-test for proportions approximates the binomial distribution with a normal distribution. It requires sufficiently large sample size ($np \geq 10$ and $n(1-p) \geq 10$).

Example:

Website Conversion Rate

- Claim: New website design increased conversion rate from 5% to over 7%
- Sample: 600 visitors, 48 conversions
- Test if conversion rate > 7% ($\alpha = 0.05$)

Hypotheses:

- $H_0: p = 0.07$
- $H_1: p > 0.07$

Test statistic:

- $\hat{p} = 48/600 = 0.08$
- $SE = \sqrt{[0.07(1-0.07)/600]} \approx 0.0104$
- $z = (0.08 - 0.07) / 0.0104 \approx 0.96$
- Critical value: $z(0.05) = 1.645$

Decision: Since $0.96 < 1.645$, fail to reject H_0

Conclusion: No significant evidence that conversion rate exceeds 7%

3.5 P-value of a Test

Definition:

The probability, assuming the null hypothesis is true, of obtaining a test statistic at least as extreme as the one observed. It measures the strength of evidence against H_0.

Detailed Explanation: P-values provide a continuous measure of evidence against H_0. Smaller p-values indicate stronger evidence. The threshold α (usually 0.05) determines statistical significance.

Interpretation Guidelines:

- p-value < 0.01: Very strong evidence against H_0
- $0.01 \leq$ p-value < 0.05: Strong evidence against H_0
- $0.05 \leq$ p-value < 0.10: Weak evidence against H_0
- p-value ≥ 0.10: Little or no evidence against H_0

Example:

Drug Effectiveness Trial

- Testing new cholesterol medication vs placebo
- p-value = 0.012, $\alpha = 0.05$
- Since $0.012 < 0.05$, reject H_0
- **Interpretation:** Only 1.2% chance of seeing these results if drug was ineffective
- This is strong evidence that the drug works

3.6 Risks and Power Curve

Definition:

- **Risks:** Refers to Type I error (rejecting a true H_0) and Type II error (failing to reject a false H_0).
- **Power Curve:** A graph showing the probability of rejecting the null hypothesis (power) as a function of the true effect size.

Detailed Explanation:

Type I error (false positive) occurs when we reject a true H_0. Type II error (false negative) occurs when we fail to reject a false H_0. Power is the probability of correctly rejecting a false H_0.

Error Types:

- **Type I error (α):** Rejecting H_0 when it's true
- **Type II error (β):** Failing to reject H_0 when it's false

- **Power (1-β):** Probability of correctly rejecting H_0

Factors affecting power:

- Sample size (larger n → more power)
- Effect size (larger effect → more power)
- Significance level (larger α → more power)
- Variability (less variability → more power)

Example:

Clinical Trial Planning

- Testing new drug, want power = 0.90 to detect meaningful effect
- Calculate required sample size during study design
- Power curve shows relationship between effect size and power for different sample sizes
- With n=100, power ≈ 0.75 for moderate effect
- With n=200, power ≈ 0.90 for same effect

3.7 Chi-Square Goodness-of-Fit Test

Definition:

A hypothesis test that determines if a sample data distribution matches a hypothesized population distribution.

Detailed Explanation:

The chi-square goodness-of-fit test compares observed frequencies with expected frequencies under the hypothesized distribution. It's used for categorical data and discrete probability distributions.

Example:

Random Number Generator Test Testing if a random number generator produces digits 0-9 equally often:

Digit	Observed	Expected (if uniform)
0	95	100
1	105	100
2	98	100
3	102	100
4	97	100
5	103	100
6	96	100

Digit	Observed	Expected (if uniform)
7	104	100
8	99	100
9	101	100

Test statistic:

- $\chi^2 = \sum(O-E)^2/E = (25+25+4+4+9+9+16+16+1+1)/100 = 110/100 = 1.10$
- Degrees of freedom: 10-1 = 9
- Critical value: $\chi^2(9, 0.05) = 16.92$

Decision: Since $1.10 < 16.92$, fail to reject H_0

Conclusion: No evidence that number generator is non-uniform

Chapter 4: Statistical Tests (Multiple Samples)

Definition: Hypothesis tests used to compare parameters across two or more different groups or populations.

4.1 Test Principles for Multiple Samples

Definition:

Extending hypothesis testing concepts to compare groups, including considerations for independent vs. paired samples.

Detailed Explanation:

Multiple sample tests account for between-group and within-group variability. Proper experimental design (randomization, blocking, controls) is crucial for valid comparisons.

Key Concepts:

- **Independent samples:** Different groups with no pairing
- **Paired samples:** Same subjects measured under different conditions
- **Blocking:** Controlling for extraneous variables by grouping similar subjects

Example:

Medication Effectiveness Study

- **Independent:** Different patients in treatment vs control groups
- **Paired:** Same patients measured before and after treatment
- **Blocking:** Grouping patients by age or severity before random assignment

4.2 Comparison of Two Variances

Definition:

A hypothesis test (F-test) to determine if the variances of two populations are equal.

Detailed Explanation:

The F-test compares two variances by taking their ratio. It's sensitive to non-normality, so it should be used cautiously. Some statisticians recommend always using variance versions of t-tests rather than testing variances separately.

Example:

Production Process Comparison

- Compare consistency of two manufacturing machines
- Machine A: $n_1=22$, $s_1^2=18.5$
- Machine B: $n_2=18$, $s_2^2=25.3$
- Test if variances differ ($\alpha=0.05$)

Hypotheses:

- $H_0: \sigma_1^2 = \sigma_2^2$
- $H_1: \sigma_1^2 \neq \sigma_2^2$

Test statistic:

- $F = s_2^2/s_1^2 = 25.3/18.5 = 1.368$ (larger variance in numerator)
- Critical values: $F(17,21,0.025) \approx 2.42$ and $F(17,21,0.975) \approx 0.41$

Decision: Since $0.41 < 1.368 < 2.42$, fail to reject H_0

Conclusion: No significant difference in variances between machines

4.3 Comparison of Two Means

Definition:

A hypothesis test (e.g., two-sample t-test or paired t-test) to determine if the means of two populations are equal.

Detailed Explanation:

The two-sample t-test compares means from independent groups. We must decide whether to assume equal variances (pooled t-test) or unequal variances (Welch's t-test). For paired data, we use the paired t-test on differences.

Example:

Teaching Method Comparison

- Method A: $n_1=35$, $\bar{X}_1=78.2$, $s_1=8.5$
- Method B: $n_2=32$, $\bar{X}_2=82.6$, $s_2=7.8$
- Test if Method B gives higher scores ($\alpha=0.05$)

Hypotheses:

- $H_0: \mu_1 = \mu_2$

- $H_1: \mu_1 < \mu_2$

Test statistic (assuming equal variances):

- Pooled variance: $s_p^2 = [(34\times72.25)+(31\times60.84)]/65 \approx 66.87$
- $t = (78.2-82.6)/\sqrt{[66.87(1/35+1/32)]} = -4.4/\sqrt{3.94} \approx -2.21$
- Critical value: $t(65,0.05) \approx -1.669$ (one-tailed)

Decision: Since $-2.21 < -1.669$, reject H_0

Conclusion: Method B gives significantly higher scores

4.4 Other Tests on Means

Definition:

Procedures like **Analysis of Variance (ANOVA)** for comparing means across three or more groups.

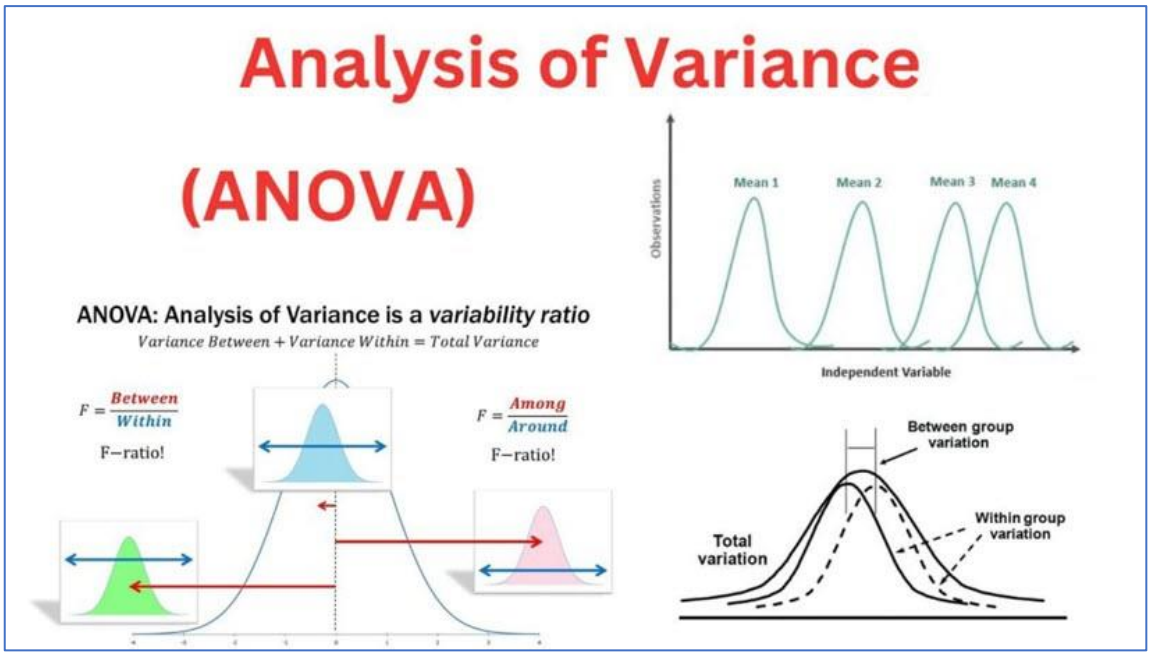

Figure 9: ANOVA Testing

Detailed Explanation:

Analysis of Variance (ANOVA) tests whether multiple group means are equal. If ANOVA is significant, post-hoc tests identify which specific means differ while controlling for multiple comparisons.

Example:

Employee Training Methods

- Compare four training methods (A, B, C, D)
- Performance scores for each method group
- **ANOVA tests:** H_0: $\mu\_A = \mu\_B = \mu\_C = \mu\_D$ vs H_1: at least one mean different

If ANOVA significant ($p < 0.05$):

- **Tukey's HSD:** Compares all pairs of means while controlling family-wise error rate
- **Bonferroni:** More conservative adjustment for multiple comparisons
- **Scheffé:** Most conservative, protects against all possible contrasts

4.5 Comparison of Two Proportions

Definition:

A hypothesis test (two-proportion z-test) to determine if the proportions of a characteristic are equal in two populations.

Detailed Explanation:

The two-proportion z-test compares proportions from independent groups. It uses a pooled proportion estimate under H_0 for the standard error calculation.

Example:

Marketing Campaign Effectiveness

- Campaign A: 220/800 conversions (27.5%)
- Campaign B: 280/800 conversions (35.0%)
- Test if Campaign B has higher conversion rate (α=0.05)
-

Hypotheses:

- H_0: $p_1 = p_2$
- H_1: $p_1 < p_2$

Test statistic:

- Pooled $\hat{p} = (220+280)/(800+800) = 500/1600 = 0.3125$
- SE = $\sqrt{[0.3125(1-0.3125)(1/800+1/800)]} \approx 0.0232$
- $z = (0.275-0.350)/0.0232 \approx -3.23$
- Critical value: $z(0.05) = -1.645$

Decision: Since -3.23 < -1.645, reject H_0

Conclusion: Campaign B has significantly higher conversion rate

4.6 Test of Independence - Chi-Square

Definition:

A hypothesis test that assesses whether two categorical variables are related (dependent) or unrelated (independent) in a population.

Detailed Explanation:

The chi-square test of independence assesses whether there's a relationship between two categorical variables. It compares observed frequencies with expected frequencies under the assumption of independence.

Example: Education Level vs Technology Adoption Survey of 400 people:

	Early Adopter	Mainstream	Resistant	Total
College	70	50	30	150
Some College	40	60	50	150
High School	20	40	40	100
Total	**130**	**150**	**120**	**400**

Expected frequencies: (row total × column total)/grand total

- $E_{11} = (150 \times 130)/400 = 48.75$, $E_{12} = (150 \times 150)/400 = 56.25$, etc.

Test statistic:

- $\chi^2 = \sum(O-E)^2/E \approx 28.6$
- Degrees of freedom: (3-1)(3-1) = 4
- Critical value: $\chi^2(4,0.05) = 9.49$

Decision: Since 28.6 > 9.49, reject H_0

Conclusion: Education level and technology adoption are related

4.7 Tests of Homogeneity - Chi-Square

Definition:

A hypothesis test that assesses whether different populations have the same distribution of a single categorical variable.

Detailed Explanation:

The chi-square test of homogeneity compares distributions across different populations. The calculations are identical to the independence test, but the sampling scheme and interpretation differ.

Example:

Regional Product Preference

- Compare product preference distribution across 4 regions
- Test if preference patterns are the same in all regions
- Similar calculations to independence test
- **Interpretation:** If significant, different regions have different preference patterns

Evaluation Method

- **Continuous Assessment:** 40%
- **Final Examination:** 60%

Learning Outcomes

Upon completion, students will be able to:

1. Apply combinatorial analysis to solve real-world counting problems
2. Define and work with various probability distributions in practical contexts
3. Conduct complete descriptive analysis of univariate and bivariate data
4. Perform point and interval estimation for population parameters
5. Conduct hypothesis tests for single and multiple samples
6. Interpret statistical results in practical, real-world contexts

References

[1] Y. Croissant, *Statistiques descriptives et probabilités*. Bruxelles, Belgique: De Boeck, 2010.

[2] J.-F. Delmas, *Introduction au calcul des probabilités et à la statistique*. Berlin, Allemagne: Springer, 2014.

[3] A. Hamon, *Statistique descriptive : exercices corrigés*. Rennes, France: Presses Universitaires de Rennes, 2008.

[4] D. J. Mercier, *Cahiers de mathématiques du supérieur*, vol. 1. Paris, France: Dunod, 2010.

[5] M. Mountassir, *Probabilités et statistique*, 2e éd. Paris, France: Ellipses, 2011.

[6] A. Oukacha, *Statistique descriptive et calcul de probabilités*. Alger, Algérie: Office des Publications Universitaires, 2010.

[7] A. Rebbouh, *Statistique descriptive et calculs de probabilités*. Alger, Algérie: Éditions Houma, 2009.

[8] M. Rouaud, *Probabilités, statistiques et analyses multicritères*. Paris, France: Dunod, 2012.

[9] G. Saporta, *Probabilités, analyse des données et statistique*, 3e éd. Paris, France: Technip, 2014.

[10] S. Schaum, *Théorie et applications de la statistique*. New York, NY, USA: McGraw-Hill, 1991.

[11] A.-J. Valleron, *Probabilités et statistique*, 2e éd. Paris, France: Flammarion, 2006.