Probability & Statistics for Engineers

Practice Exercises

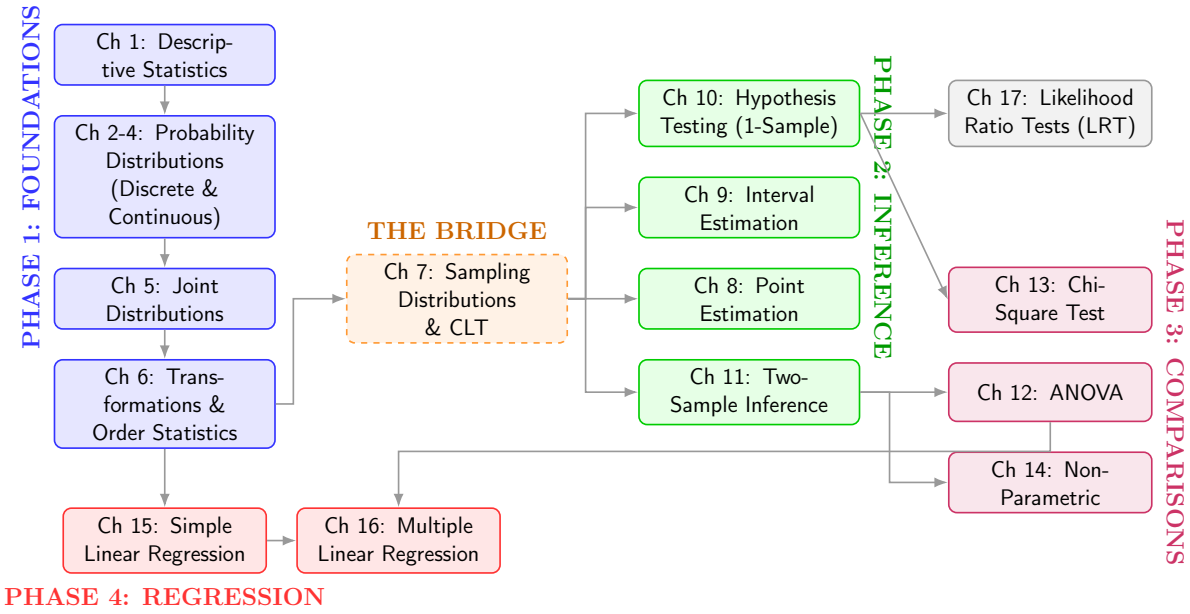
For Undergraduate Students

January 19, 2026

Contact:
If you have any questions, feel free to ask me via
thananop.klp@gmail.com
Note: Detailed solutions and source codes are provided via **GitHub**.

Learning Roadmap

How the 17 chapters connect to build your engineering statistics toolkit.



INSTRUCTIONS

This exercise book covers 18 chapters of Probability & Statistics for Engineers. Each chapter is divided into four sections:

- **Basic Concept:** Fundamental definitions.
- **Intermediate:** Calculation and interpretation.
- **Challenge:** Proofs and advanced problems.
- **Application:** Computational analysis.

Contents

Chapter 1

Descriptive Statistics

1.1 Basic Concept

1.1 Data Classification
Classify the following variables as qualitative, quantitative discrete, or quantitative continuous:

(a) Failure time of a light bulb (in hours).

(b) Number of defects on a silicon wafer.

(c) Alloy grade (Grade A, Grade B, Grade C).

(d) Temperature of a chemical reactor (°C).

(e) The zip code of a customer's address.

1.2 Measures of Center Properties
Compare Mean, Median, and Mode:

(a) Which measure is most robust against outliers? Explain why.

(b) In a positively skewed distribution (skewed right), what is the typical relationship between Mean, Median, and Mode?

(c) Can a dataset have more than one Mean? Can it have more than one Mode?

1.3 Parameter vs. Statistic
Identify whether the underlined value is a parameter or a statistic:

(a) The average salary of all 40 mayors in the country.

(b) The average weight of a sample of 100 bags of potatoes.

(c) A researcher interviews 50 voters and finds that 45% support the new policy.

1.2 Intermediate

1.4 Fizzy Drinks Analysis (Frequency Table)
The number of cans of fizzy drinks consumed by teenagers each day was recorded:

Cans per day (x)	0	1	2	3	4	5
No. of teenagers (f)	25	30	26	20	14	10

Calculate:

(a) The sample size n.

(b) The sample mean $\bar{x}$.

(c) The sample median.

(d) The sample variance S^2 and standard deviation S.

1.5 Stem-and-Leaf Plot & Outliers
Student grades on a chemistry exam were:

$$77, 78, 76, 81, 86, 51, 79, 82, 84, 99$$

(a) Calculate the Mean and Standard Deviation.

(b) Construct a stem-and-leaf plot.

(c) Identify any potential outliers using the $1.5 \times IQR$ rule.

(d) If the score 51 was a data entry error and should be 71, how would the mean and median change?

1.6 Box Plot Interpretation
Given a box plot where the five-number summary is: Min=10, $Q_1 = 20$, $Q_2 = 35$, $Q_3 = 50$, Max=90.

(a) Calculate the Interquartile Range (IQR).

(b) Determine the values of the Upper and Lower Fences (Whiskers limit).

(c) Is the value 90 considered an outlier?

(d) Describe the skewness of the data based on the box plot components.

1.7 Combined Mean and Variance
Two classes took the same engineering statistics exam:

- Class A: $n_A = 20$, $\bar{x}_A = 80$, $S_A^2 = 25$
- Class B: $n_B = 30$, $\bar{x}_B = 70$, $S_B^2 = 36$

Find the combined mean $\bar{x}_{combined}$ and the combined variance $S^2_{combined}$ of the 50 students.

1.8 Coefficient of Variation (CV)

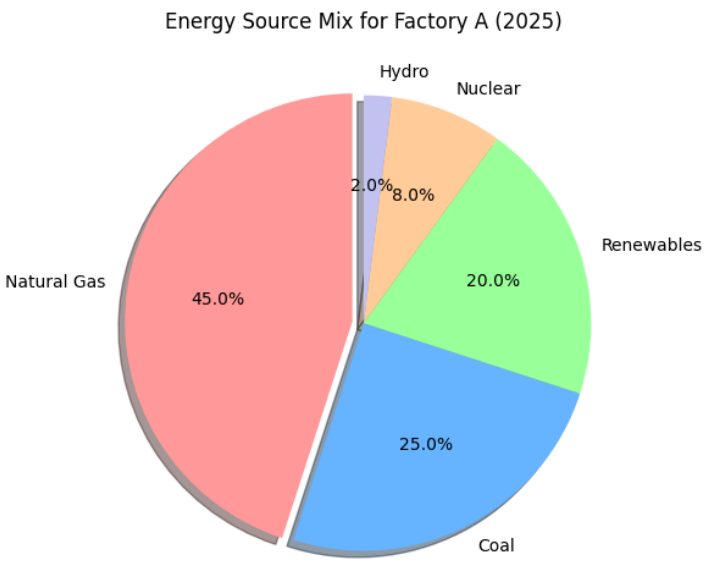
An engineer wants to compare the precision of two machines.

- Machine 1: Mean diameter = 10 mm, SD = 0.5 mm
- Machine 2: Mean weight = 100 kg, SD = 2 kg

Which machine is relatively more precise? Use the Coefficient of Variation to justify your answer.

1.9 Energy Source Analysis (Pie Chart)

The following pie chart represents the energy mix used by Factory A in 2025.



Questions:

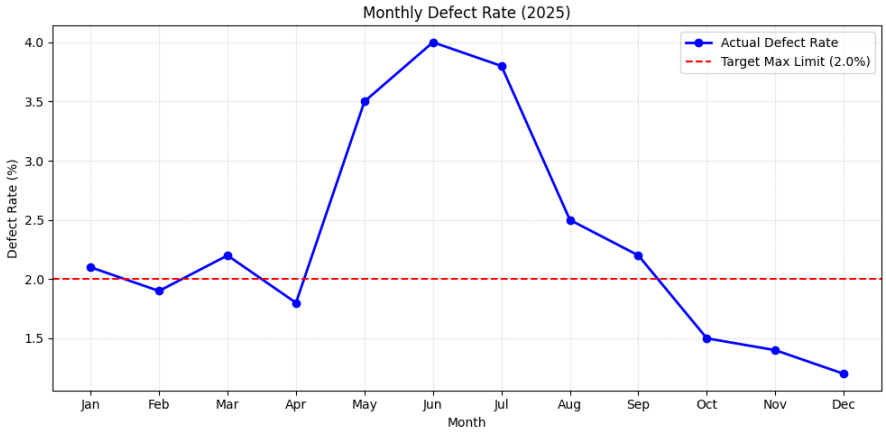
(a) Which energy source is the primary contributor to the factory's power supply?

(b) If the factory aims to have at least 30% of its energy from "Clean Sources" (Renewables + Nuclear + Hydro), have they met this target? Show your calculation.

(c) If the total energy consumption for the year was 10,000 MWh, calculate the amount of energy (in MWh) generated from **Coal**.

1.10 Process Stability Monitoring (Time Series)

An engineer tracks the monthly defect rate (%) of a production line. The target is to keep the defect rate below 2.0%.



Questions:

(a) Describe the trend of the defect rate from January to June. Is the process improving or deteriorating?

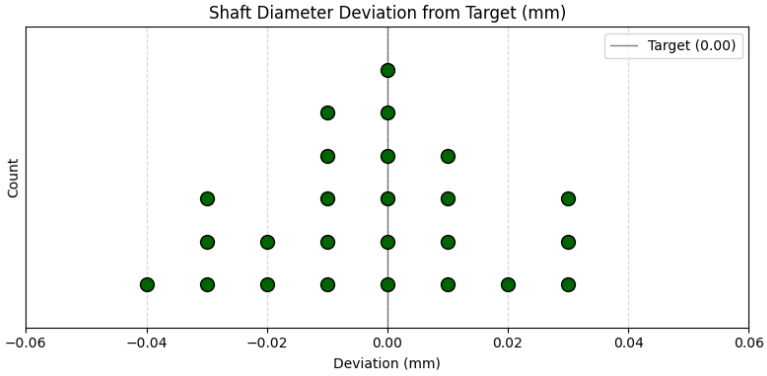
(b) Identify the month with the **highest** defect rate. What might have happened during this period? (Propose a logical engineering hypothesis, e.g., machine breakdown, new staff, raw material change).

(c) From September to December, the process seems to stabilize. Calculate the average defect rate for the last quarter (Oct, Nov, Dec).

(d) How many months did the process fail to meet the target specification ($>$ 2.0%)?

1.11 Precision Measurement (Dot Plot)

A quality inspector measures the deviation of shaft diameters from the target value (0.00 mm). A dot plot of 25 random samples is shown below.



Questions:

(a) Estimate the center of the distribution based on the plot. Is the process centered at the target (0.00 mm)?

(b) Identify the range of the deviations (Min to Max).

(c) Are there any potential outliers in the data? If so, at approximately what value?

(d) If the specification limits are ±0.05 mm, estimate the percentage of parts that are out of spec (defective).

1.3 Challenge

1.12 Minimizing Squared Deviations
Consider the function $g(a) = \sum_{i=1}^{n}(x_i - a)^2$. Prove using calculus that $g(a)$ is minimized when $a = \bar{x}$.

1.13 Sum of Deviations
Prove algebraically that the sum of deviations from the sample mean is always zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

1.14 Linear Transformation of Variance
Let $y_i = ax_i + b$ for constants a and b. Prove that:

$$S_y^2 = a^2 S_x^2$$

(Hint: Start with the definition $S_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$)

1.15 Computational Formula for Variance
Derive the "Computational Formula" for sample variance starting from the definition:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

1.4 Application

1.16 Excel Formulas & Functions

Suppose you have data in cells `A1:A50`. Write down the Excel formulas to calculate: (a) The arithmetic mean. (b) The sample standard deviation. (c) The median. (d) The count of values greater than 50. *(Hint: Use =COUNTIF)*

1.17 Data Analysis Toolpak Interpretation

You used the "Descriptive Statistics" tool in Excel and got the following output for a machine's production weight (grams):

- Mean: 50.05
- Standard Error: 0.02
- Median: 50.00
- Standard Deviation: 0.15
- Kurtosis: 1.5
- Skewness: 0.05
- Range: 0.8

Questions: (a) Is the distribution roughly symmetric or heavily skewed? Look at the Skewness value. (b) Does the data have "heavy tails" (more outliers than normal)? Look at the Kurtosis.

1.18 Exploratory Data Analysis (EDA)

The following Python code defines a dataset of sensor readings manually. Run the code to answer the questions.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# 1. Create the dataset manually
data = [10.2, 10.5, 9.8, 10.1, 10.3, 10.0, 9.9, 10.4,
        10.2, 10.3, 25.0, 10.1, 9.7, 10.2, 10.3]
df = pd.DataFrame(data, columns=['Reading'])

# 2. Calculate Statistics
print(df.describe())

# 3. Visualization
plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
df.boxplot()
plt.title("Boxplot")

plt.subplot(1,2,2)
df.hist(bins=5)
plt.title("Histogram")
plt.show()
```

Questions: (a) From the code output (describe), what is the Mean and Maximum value? (b) The value '25.0' is included in the list. How does this value affect the Mean versus the Median? (c) Based on the Boxplot generated, how would the value '25.0' be represented?

Chapter 2

Probability Distributions & Theorems

2.1 Basic Concept

2.1 Axioms of Probability
State Kolmogorov's three axioms of probability. Using these axioms, prove that $P(A^c) = 1 - P(A)$.

2.2 Independence vs. Mutually Exclusive

(a) Define "Independent Events" mathematically.

(b) Define "Mutually Exclusive Events" mathematically.

(c) Can two non-impossible events ($P(A) > 0, P(B) > 0$) be both independent and mutually exclusive? Explain.

2.3 Set Theory for Engineers
Let S be the sample space of component failures. A is the event "Capacitor fails", B is "Resistor fails". Express the following in set notation:

(a) Both components fail.

(b) At least one component fails.

(c) Only the Capacitor fails.

(d) Neither component fails.

2.2 Intermediate

2.4 Conditional Probability Calculation
Given $P(A) = 0.5, P(B) = 0.6$, and $P(A \cup B) = 0.8$. Find:

(a) $P(A \cap B)$

(b) $P(A|B)$

(c) $P(B|A)$

(d) Are A and B independent?

2.5 The Night Shift Crew (Combinatorics)
A company has 20 machinists. A night shift crew needs 3 machinists.

(a) How many different crews are possible?

(b) If machinists are ranked 1-20, how many crews would **not** have the rank #1 machinist?

(c) How many crews contain at least one of the top 5 machinists?

2.6 Concrete Strength (Conditional)
Data for 204 concrete samples:

	Compressive Strength	
Curing Time	Below Standard	Above Standard
3 days	12	40
14 days	44	16
24 days	56	36

Let A = Event curing time is ≤ 14 days. B = Event strength is Above Standard. Find:

(a) $P(A)$ and $P(B)$

(b) $P(A|B)$

(c) $P(B|A)$

(d) Event A and B are independent? Why?

2.7 Rare Disease (Bayes' Theorem)
A rare disease affects 1 in 500 people ($P(D) = 0.002$). A test is 95% accurate for sick people ($P(+|D) = 0.95$) and has a 1% false positive rate ($P(+|D^c) = 0.01$).

(a) If a person tests positive, what is the probability they actually have the disease?

(b) Why is this probability lower than most people intuitively expect?

2.8 System Reliability
Consider a system with 3 components.

(a) **Series:** If components are in series with reliabilities $0.9, 0.8, 0.95$, what is the system reliability?

(b) **Parallel:** If components are in parallel with the same reliabilities, what is the system reliability?

2.3 Challenge

2.9 Inclusion-Exclusion Principle
Prove the formula for the union of three events:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

2.10 Bonferroni's Inequality
Prove that for any two events A and B:

$$P(A \cap B) \geq P(A) + P(B) - 1$$

2.11 Conditional Independence Proof
If A and B are independent, prove that A^c and B^c are also independent.

2.4 Application

2.12 Birthday Paradox Simulation
Write a Python script to simulate the "Birthday Paradox".

(a) Function

```python
import random

def has_duplicate(n):
    """
    Generate n random birthdays (1  365  ).
    Return True if at least one duplicate birthday exists.
    """
    birthdays = [random.randint(1, 365) for _ in range(n)]
    return len(birthdays) != len(set(birthdays))

```

(b) Run the simulation 10,000 times for group sizes $n = 10$ to $n = 50$ What happened?.

Chapter 3

Discrete Random Variables & Distributions

3.1 Basic Concept

3.1 Validating PMF
Determine the value of the constant c so that the following functions represent valid Probability Mass Functions (PMF):

(a) $p(x) = c(x^2 + 1)$ for $x = 0, 1, 2, 3$.

(b) $p(x) = c \cdot (1/2)^x$ for $x = 1, 2, 3, \ldots$.

(c) $p(x) = c\binom{5}{x}$ for $x = 0, 1, 2, 3, 4, 5$.

3.2 Distribution Identification
Identify the most appropriate discrete probability distribution for each scenario (Bernoulli, Binomial, Poisson, Geometric, Hypergeometric, or Negative Binomial):

(a) The number of heads obtained in 10 tosses of a fair coin.

(b) The number of cars arriving at a toll booth between 12:00 PM and 1:00 PM.

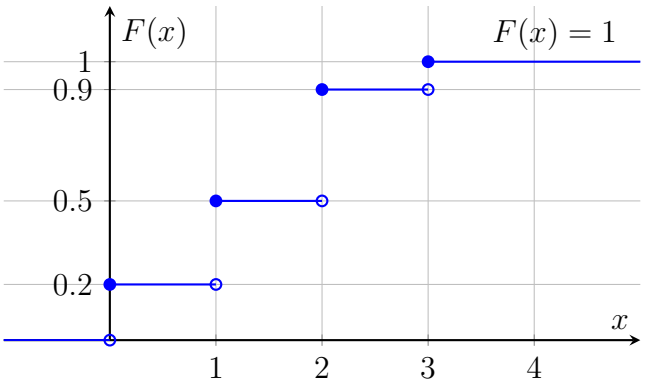
(c) The number of trials required to win a lottery for the first time.

(d) Selecting 5 defective parts from a box containing 20 parts (5 defective, 15 good) without replacement.

(e) The number of bits transmitted correctly until the 5th error occurs.

3.2 Intermediate

3.3 Visualizing the CDF (Step Function Analysis)
The Cumulative Distribution Function (CDF), $F(x) = P(X \leq x)$, of a discrete random variable X is plotted below.



(a) The "jump" size at each step corresponds to the probability mass at that point. Identify the support of X and calculate the Probability Mass Function (PMF) $P(X = x)$ for all values.

(b) Use the graph to determine the probability $P(0.5 < X \leq 2.5)$. *(Hint: Recall that for discrete variables, $P(a < X \leq b) = F(b) - F(a)$).*

(c) Calculate the expected value $E[X]$ using the probabilities found in (a).

3.4 Binomial: Quality Control
A manufacturing process produces 5% defective items. If a random sample of 20 items is selected:

(a) Calculate the probability of finding exactly 2 defective items.

(b) Calculate the probability of finding at least 1 defective item.

(c) Find the expected number and variance of defective items in the sample.

3.5 Hypergeometric vs. Binomial
A batch of 50 components contains 10 defectives. A sample of 5 components is drawn **without replacement**.

(a) Calculate the probability of getting exactly 1 defective using the Hypergeometric distribution.

(b) Approximate this probability using the Binomial distribution. Is the approximation good? Explain why.

3.6 Poisson: Web Server Traffic

Requests to a web server follow a Poisson process with an average rate of $\lambda = 5$ requests per second.

(a) Find the probability of receiving exactly 3 requests in a given second.

(b) Find the probability of receiving no requests in a 2-second interval.

(c) Determining the most likely number of requests (Mode) in a second.

3.7 Negative Binomial: Oil Drilling

The probability of hitting oil in a single drilling operation is 0.2. Drillings are independent.

(a) What is the probability that the first oil discovery occurs on the 4th drill?

(b) What is the probability that the third oil discovery occurs on the 10th drill?

(c) What is the expected number of drills needed to find oil 3 times?

3.8 Expectation and Variance Calculation

Let X be a random variable with the following PMF:

$$p(x) = \begin{cases} 0.1 & x = -2 \\ 0.2 & x = 0 \\ 0.4 & x = 1 \\ 0.3 & x = 3 \end{cases}$$

Calculate:

(a) $E[X]$ and $Var(X)$.

(b) $E[2X + 5]$ and $E[X^2]$.

(c) $Var(2X + 5)$.

3.9 Custom Discrete RV (Infinite Series)
Let X be a discrete random variable with the probability mass function (PMF) given by:

$$P(X = k) = c \cdot k \cdot \left(\frac{1}{3}\right)^k, \quad k = 1, 2, 3, \ldots$$

(where c is a normalization constant).

Hint: Use the following series summation formulas (for $|r| < 1$):

$$\sum_{k=1}^{\infty} kr^k = \frac{r}{(1-r)^2}, \quad \sum_{k=1}^{\infty} k^2 r^k = \frac{r(1+r)}{(1-r)^3}$$

(a) Determine the value of the constant c that makes this a valid PMF.

(b) Find the value of the Cumulative Distribution Function $F(2)$ (i.e., $P(X \leq 2)$).

(c) Calculate the Expected Value $E[X]$.

(d) Calculate the Variance $Var(X)$.

3.10 Binomial Approximation by Poisson
A high-speed fiber optic cable transmits data with a bit error rate of $p = 0.0002$ (probability of an error per bit). A packet consisting of $n = 10,000$ bits is transmitted. Let X be the number of error bits in a packet.

(a) Write down the exact expression for the probability of finding exactly 3 errors using the Binomial distribution formula (do not calculate the final value).

(b) Since n is large and p is small, approximate the probability of finding exactly 3 errors using the Poisson distribution.

(c) Calculate the absolute difference between the exact Mean (np) and Variance ($np(1-p)$) of the Binomial distribution. Why does this justify using Poisson ($\lambda = np$) as an approximation?

3.3 Challenge

3.11 Geometric Mean Derivation
Let $X \sim Geo(p)$ with PMF $P(X = k) = (1-p)^{k-1}p$ for $k = 1, 2, \ldots$. Prove algebraically that the expected value is:

$$E[X] = \frac{1}{p}$$

(Hint: Use the derivative of the geometric series sum formula $\sum x^k$.)

3.12 Poisson as a Limit of Binomial
Starting with the Binomial PMF $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Show that if $n \to \infty$ and $p \to 0$ such that $np = \lambda$ (constant), the limit is the Poisson PMF:

$$\lim_{n \to \infty} P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

3.13 Lack of Memory (Geometric)

Prove that the Geometric distribution is "memoryless". That is, for any integers $s, t > 0$:

$$P(X > s + t | X > s) = P(X > t)$$

Explain the physical meaning of this property in the context of tossing a coin.

3.14 Unknown Distribution Bounds

A manufacturing process produces parts with a mean diameter of $\mu = 5.0$ mm and a standard deviation of $\sigma = 0.1$ mm. The distribution of the diameter is **unknown** (it is NOT necessarily Normal).

(a) Using Chebyshev's Inequality, find the **minimum** probability that a randomly selected part has a diameter between 4.8 mm and 5.2 mm.

(b) Find the range in which at least 75% of the parts' diameters must fall.

3.15 Chebyshev vs. Normal

Let X be a random variable with $\mu = 100$ and $\sigma = 10$. We want to calculate $P(80 < X < 120)$.

(a) Calculate the lower bound of this probability using Chebyshev's Inequality.

(b) Assume $X \sim N(100, 100)$. Calculate the exact probability using the Standard Normal distribution.

(c) Compare the two results. Why is Chebyshev's bound much "looser" (more conservative)?

3.16 Derivation from Markov's Inequality (Challenge)

Markov's Inequality states that for a non-negative random variable Y, $P(Y \geq a) \leq \frac{E[Y]}{a}$.

Using this fact, prove Chebyshev's Inequality:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

(Hint: Let $Y = (X - \mu)^2$ and $a = (k\sigma)^2$. Since squared deviations are non-negative, Markov's inequality applies.)

3.4 Application

3.17 Binomial Convergence Simulation

Visualize how the Binomial distribution converges to the Normal distribution.

(a) Generate a Binomial distribution with $p = 0.5$ and varying $n \in \{10, 30, 100, 1000\}$.

(b) Plot the PMF (histogram) for each n.

(c) Overlay the corresponding Normal PDF with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom, norm
# Write your code here
```

3.18 Poisson vs. Real Data

Suppose you have data representing the number of defects per fabric roll: `data = [0, 1, 0, 2, 1, 0, 5, 1, 0, 2, 1, 0, 0, 3, 1]`

(a) Calculate the sample mean $\bar{x}$.

(b) Assume $\lambda = \bar{x}$. Calculate the theoretical Poisson probabilities for $k = 0, 1, 2, 3$.

(c) Compare the theoretical probabilities with the observed frequencies from the data.

Chapter 4

Continuous Random Variables & Distributions

4.1 Basic Concept

4.1 PDF Properties
The function $f(x)$ is a valid Probability Density Function (PDF). Answer True or False and explain:

(a) $f(x)$ can be greater than 1.

(b) The probability $P(X = c)$ is always 0 for any constant c.

(c) $\int_{-\infty}^{\infty} f(x)dx = 1$.

(d) If $f(x)$ is increasing, the Cumulative Distribution Function (CDF) must be convex.

4.2 Fundamental of Continuous Random Variables
Let the continuous random variable X have the PDF:

$$f(x) = \begin{cases} k(1-x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of k that makes this a valid PDF.

(b) Find the CDF $F(x)$.

(c) Calculate $P(-0.5 < X < 0.5)$.

(d) Calculate the expectation of X, and variance of $3X$.

4.2 Intermediate

4.3 Uniform Distribution: Rounding Error
A digital scale rounds weights to the nearest gram. The round-off error X is uniformly distributed between -0.5 and 0.5 grams.

(a) Write the PDF of X.

(b) Calculate the probability that the error is between -0.2 and 0.2.

(c) Find the mean and variance of the error.

4.4 Exponential: Component Life

The lifetime of a transistor is exponentially distributed with a mean of 10,000 hours ($\mu = 1/\lambda = 10000$).

(a) Find the parameter λ.

(b) What is the probability that a transistor lasts more than 12,000 hours?

(c) What is the probability that it fails within the first 5,000 hours?

(d) Find the median lifetime.

4.5 Normal Distribution: Grades

Exam scores are normally distributed with $N(\mu = 75, \sigma^2 = 100)$.

(a) Find the probability that a student scores between 60 and 80.

(b) If the top 10% of students get an A, what is the cutoff score?

(c) Six students are randomly selected. Find the probability that at least four of them score exactly 75.

4.6 Standard Normal Z

Using the standard normal table (or calculator):

(a) Find $P(Z > 1.645)$.

(b) Find $P(-1.96 < Z < 1.96)$.

(c) Find k such that $P(Z < k) = 0.95$.

4.3 Challenge

4.7 Mean of Exponential

Let $X \sim Exp(\lambda)$ with PDF $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Prove using integration by parts that:

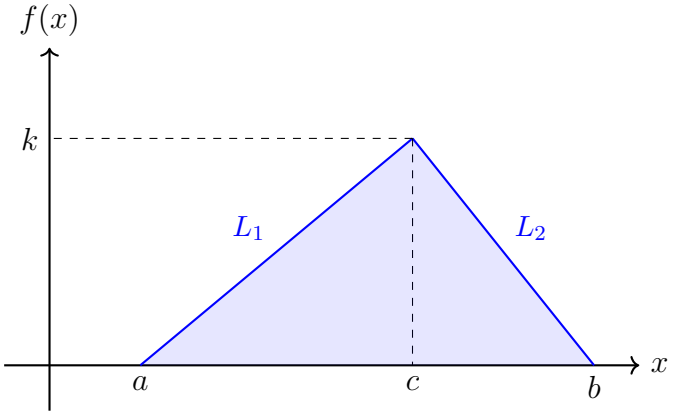
$$E[X] = \frac{1}{\lambda} \quad \text{and} \quad Var(X) = \frac{1}{\lambda^2}$$

4.8 Normal Inflection Points

Prove that the points of inflection of the Normal PDF $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ occur at $x = \mu \pm \sigma$. *(Hint: Find the second derivative $f''(x)$ and set it to zero.)*

4.9 Constructing a Triangular PDF

Consider a continuous random variable X whose probability density function $f(x)$ is shaped like a triangle. The density starts at $x = a$, rises linearly to a peak of height k at $x = c$, and then decreases linearly to $x = b$ (where $a < c < b$).



(a) Use the geometric property (Area under PDF = 1) to determine the value of k in terms of a and b.

(b) Find the equations for the two linear segments (L_1 and L_2) and define the piecewise function $f(x)$.

(c) Set up the integral and show that the expected value is:

$$E[X] = \frac{a+b+c}{3}$$

(d) Suppose specific values: $a = 0, c = 2, b = 4$. Calculate the probability $P(X > 3)$ using geometry (area of the small triangle) rather than integration.

4.4 Application

4.10 Area Under the Curve (Integration)

Calculate the probability $P(X < 2)$ for a Standard Normal distribution using numerical integration (Trapezoidal rule) in Python, and compare it with `scipy.stats.norm.cdf`.

```python
import numpy as np
from scipy.stats import norm

def standard_normal(x):
    return (1/np.sqrt(2*np.pi)) * np.exp(-0.5 * x**2)

# Implement Trapezoidal rule here
```

4.11 Fitting a Distribution

You are given a dataset of 1,000 machine failure times.

(a) Generate a histogram of the data.

(b) Use `scipy.stats.expon.fit` to estimate λ.

(c) Plot the fitted Exponential PDF over the histogram to visually assess the fit.

Chapter 5

Joint Probability Distributions

5.1 Basic Concept

5.1 Joint vs. Marginal Definitions
Let $f_{XY}(x, y)$ be the joint PDF of random variables X and Y.

(a) **Definition:** Write the formula to find the Marginal PDF $f_X(x)$ from the joint PDF.

(b) **Independence Check:** State the necessary and sufficient condition for X and Y to be statistically independent.

(c) **Concept:** If X and Y are independent, is it guaranteed that $Cov(X, Y) = 0$? Is the reverse true?

5.2 Discrete Joint Distribution
The joint PMF of X and Y is given by the table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.10	0.20	0.10
$X = 2$	0.10	0.30	0.20

(a) Find the marginal PMFs of X and Y.

(b) Find the conditional PMF of Y given $X = 1$.

(c) Are X and Y independent? Justify your calculation.

5.2 Intermediate

5.3 The Triangular Region (Exponential Density)
Let the continuous random variables X and Y have the joint PDF:

$$f(x,y) = ke^{-(2x+3y)} \quad \text{for } 0 < y < x < 1$$

and $f(x,y) = 0$ otherwise.

(a) Draw the region of non-zero probability in the xy-plane. Label the boundaries $y = 0$, $y = x$, and $x = 1$ clearly.

(b) Find the value of k that makes this a valid joint PDF.

(c) Find the marginal PDF of X, denoted $f_X(x)$.

(d) Calculate $P(Y < X/2)$. (Hint: Set up the double integral over the region where $0 < y < x/2$).

5.4 Conditional Probability & Independence
Using the joint PDF and the constant k found in the previous problem:

(a) Determine the conditional PDF $f_{Y|X}(y|x)$. State the valid range for y given a fixed x.

(b) Calculate $E[Y|X = x]$.

(c) Are X and Y independent? Justify your answer mathematically using the definition $f(x,y) \stackrel{?}{=} f_X(x)f_Y(y)$ or by examining the conditional PDF.

5.5 Linear Combination
Let X and Y be independent random variables with:

$$E[X] = 10, \quad Var(X) = 4, \quad E[Y] = 20, \quad Var(Y) = 9$$

Calculate the mean and variance of $W = 3X - 2Y + 5$.

5.6 Server Load Balancing (Discrete Joint Distribution)
Two servers, Server A (X) and Server B (Y), handle requests. The number of active tasks on each server is modeled by the following joint probability mass function table:

X (Server A) \ Y (Server B)	0	1	2
0	0.10	0.05	0.15
1	0.05	0.20	k
2	0.10	0.05	0.10

(a) Find the value of k.

(b) Find the marginal PMFs $P(X = x)$ and $P(Y = y)$. What is the probability that Server A has more tasks than Server B ($P(X > Y)$)?

(c) If Server B has exactly 1 active task, what is the expected number of tasks on Server A? (Find $E[X|Y = 1]$).

(d) Calculate $Cov(X, Y)$. Are the loads on the two servers independent?

5.7 Manufacturing Tolerance (Bivariate Normal)
The length (X) and width (Y) of a precision-machined part follow a Bivariate Normal distribution with parameters:

$$\mu_X = 100,\ \sigma_X = 2, \quad \mu_Y = 50,\ \sigma_Y = 1$$

The correlation coefficient between length and width is $\rho = 0.8$.

(a) Ignoring the width, what is the probability that a part is longer than 103 mm? (Use Standard Normal Z-table).

(b) If a part is measured to have a length of $x = 104$ mm (which is $+2\sigma_X$ from the mean), what is the expected width $E[Y|X = 104]$? *(Hint: For Bivariate Normal, $E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$).*

(c) Calculate the variance of the width given this length, $Var(Y|X = 104)$. Does the uncertainty reduce compared to the marginal variance $Var(Y)$? *(Hint: $Var(Y|X) = \sigma_Y^2(1 - \rho^2)$).*

5.3 Challenge

5.8 Correlation Bounds
Prove that the correlation coefficient ρ_{XY} lies between -1 and 1.

$$-1 \leq \rho_{XY} \leq 1$$

(Hint: Consider the variance of the random variable $Z = \frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}$ and use the fact that Variance ≥ 0.)

5.9 Variance of Sum
Prove the general formula for the variance of a linear combination:

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

5.4 Application

5.10 Manufacturing Tolerance Simulation

An engineer simulates the dimensions of a rectangular part. Let X be the length and Y be the width. Due to the manufacturing process, the width slightly depends on the length.

Run the following code and answer the questions:

```python
import numpy as np
import matplotlib.pyplot as plt

# Simulate Data
np.random.seed(42)
n = 1000
X = np.random.normal(10, 0.2, n)        # Length: Mean 10, SD 0.2
Y = 0.5 * X + np.random.normal(5, 0.1, n) # Width depends on X

# Calculate Statistics
correlation = np.corrcoef(X, Y)[0, 1]
print(f"Correlation: {correlation:.4f}")

# Plot
plt.scatter(X, Y, alpha=0.5)
plt.xlabel("Length (X)")
plt.ylabel("Width (Y)")
plt.title(f"Part Dimensions (Corr: {correlation:.2f})")
plt.show()

# Check Specifications
# Spec: Length must be > 9.8 AND Width must be > 9.8
spec_pass = np.sum((X > 9.8) & (Y > 9.8))
print(f"Parts passing spec: {spec_pass}/{n}")
```

Questions:

(a) Interpret the "Correlation" value. Is the relationship strong or weak? Positive or negative?

(b) From the code 'Y = 0.5 * X + ...', analytically, why are X and Y positively correlated?

(c) Based on the scatter plot logic, if we strictly control X to have less variance, would the variance of Y likely decrease? Explain.

Chapter 6

Method of Transformations & Order Statistics

6.1 Basic Concept

6.1 Moment Generating Function (Definition)
Let X be a random variable with PDF $f(x) = e^{-x}$ for $x > 0$ (Exponential with $\lambda = 1$).

(a) Find the Moment Generating Function (MGF), $M_X(t) = E[e^{tX}]$. For what values of t does it exist?

(b) Use $M_X(t)$ to calculate the first moment $E[X]$ and the second moment $E[X^2]$.

6.2 Identifying Distributions from MGF
Identify the distribution (name and parameters) of the random variable X associated with each of the following MGFs:

(a) $M_X(t) = (0.3 + 0.7e^t)^{10}$

(b) $M_X(t) = e^{5t+8t^2}$

6.3 Jacobian Method (1 RV)
Let X be a continuous random variable with PDF $f_X(x)$. Let $Y = g(X)$ be a one-to-one differentiable function of X.

(a) Write the formula for the PDF of Y, denoted $f_Y(y)$, involving the derivative of the inverse function $g^{-1}(y)$.

(b) Why do we need the absolute value of the Jacobian $|\frac{dx}{dy}|$?

6.4 Order Statistics Formulas
Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous population with PDF $f(x)$ and CDF $F(x)$. Write the formulas for:

(a) The PDF of the minimum, $X_{(1)} = \min(X_1, \ldots, X_n)$.

(b) The PDF of the maximum, $X_{(n)} = \max(X_1, \ldots, X_n)$.

6.2 Intermediate

6.5 Method of MGF (Sum of Normals)
Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent random variables. Use Moment Generating Functions to prove that $Y = X_1 + X_2$ follows a Normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. *(Hint: Recall that for independent variables, $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$.)*

6.6 Power Transformation of a Power Function
Let X be a random variable with the probability density function:

$$f_X(x) = 3x^2, \quad 0 < x < 1$$

Find the PDF of the random variable $Y = X^3$. *(Hint: First, determine the range of Y.)*

6.7 Inverse Transformation
Let the random variable X have the PDF:

$$f_X(x) = \frac{2}{x^3}, \quad x > 1$$

Find the distribution of $Y = \frac{1}{X}$. Identify the resulting distribution by name.

6.8 Rational Transformation
Let X be a random variable with PDF:

$$f_X(x) = 2x, \quad 0 < x < 1$$

Find the PDF of the random variable Y defined by:

$$Y = \frac{X^2}{1 - X^2}$$

Steps:

(a) Determine the support (valid range) of Y.

(b) Find the inverse function $x = g^{-1}(y)$.

(c) Calculate the Jacobian $|dx/dy|$.

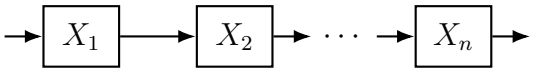
(d) Derive the final PDF $f_Y(y)$.

6.9 System Reliability (Order Statistics)

An electronic system consists of n independent components. The lifetime of each component X_i follows an Exponential Distribution with rate λ:

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

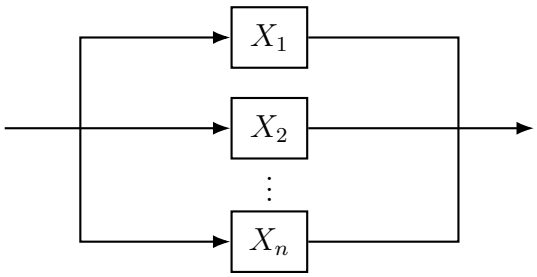
(a) **Series System:** The system fails if *any* component fails.



Series Configuration ($Y = \min$)

Let $Y = \min(X_1, \ldots, X_n)$. Find the PDF of Y. What is the distribution of the system lifetime?

(b) **Parallel System:** The system fails only if *all* components fail.



Parallel Configuration ($Z = \max$)

Let $Z = \max(X_1, \ldots, X_n)$. Write the formula for the CDF of Z, $F_Z(z)$.

(c) In general, which scenario (Series or Parallel) results in a higher expected system lifetime? Explain the physical reasoning behind your answer.

Try!: Prove your answer mathematically.
Hint: When evaluating the expectation for the Parallel case, you may use the algebraic identity:

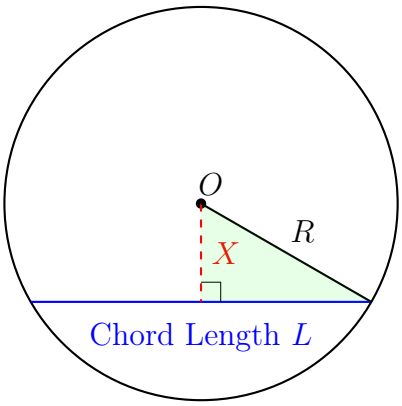
$$\frac{1-u^n}{1-u} = 1 + u + u^2 + \cdots + u^{n-1} = \sum_{k=0}^{n-1} u^k$$

6.10 Geometric Transformation (Random Chord)

Consider a circle with a fixed radius R. A chord is drawn such that its perpendicular distance from the center is a random variable X. Assume that X is uniformly distributed between 0 and R:

$$X \sim U(0, R) \implies f_X(x) = \frac{1}{R}, \quad 0 < x < R$$

Let L be the length of the chord.



(a) Using the Pythagorean theorem on the highlighted right triangle, write the equation for L in terms of X and the constant R.

(b) Determine the support (valid range) of the new random variable L.

(c) Find the Probability Density Function (PDF) of the chord length, $f_L(l)$.

(d) Which is more likely: finding a very short chord ($L \approx 0$) or a very long chord ($L \approx 2R$)? Explain based on your PDF.

6.3 Challenge

6.11 Sum of Independent Uniforms (Irwin-Hall)

Let X_1 and X_2 be independent random variables, both uniformly distributed on $[0, 1]$. Find the PDF of $Z = X_1 + X_2$. *(Hint: The result is a triangular distribution. You may need to split the domain of Z into $0 \leq z < 1$ and $1 \leq z \leq 2$.)*

6.12 Ratio of Normals (Cauchy Distribution)

Let X and Y be independent standard normal random variables $N(0, 1)$. Find the PDF of the ratio $V = X/Y$. *(Hint: The result should be the standard Cauchy distribution.)*

6.13 Box-Muller Transformation

Let U_1 and U_2 be independent random variables from $U(0, 1)$. Consider the transformation:

$$Z_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2)$$
$$Z_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2)$$

Prove that Z_1 and Z_2 are independent standard normal random variables.

6.14 Range of a Sample

Let $X_1, \ldots, X_n$ be a sample from a Uniform distribution $U(0,1)$. Let $R = X_{(n)} - X_{(1)}$ be the range. Derive the PDF of R. *(Hint: Find the joint PDF of $X_{(1)}$ and $X_{(n)}$ first.)*

6.4 Application

6.15 Verifying Transformation via Simulation

Simulate the "Rational Transformation" from Problem 6.8 ($Y = \frac{X^2}{1-X^2}$).

(a) Generate 10,000 samples of X from PDF $f(x) = 2x$ on $(0,1)$. (Hint: Use Inverse Transform Sampling: $X = \sqrt{U}$).

(b) Compute Y for each sample.

(c) Plot the histogram of Y (limit the x-axis to a reasonable range like 0 to 10).

(d) Overlay the theoretical PDF derived in Problem 6.9 to check for fit.

```
import numpy as np
import matplotlib.pyplot as plt

# 1. Generate X ~ f(x)=2x using Inverse Transform Method
u = np.random.uniform(0, 1, 10000)
x = np.sqrt(u)

# 2. Transform to Y
y = (x**2) / (1 - x**2)

# 3. Plotting
# Write your code here to plot histogram vs theoretical pdf
```

6.16 System Reliability Simulation (Series vs. Parallel)

Compare the reliability of Series and Parallel systems from Problem 6.9 using Monte Carlo simulation. Assume we have $n = 5$ components, each with an exponential lifetime ($\lambda = 0.5$).

(a) Generate a random matrix of size $(10000, 5)$ representing 10,000 systems with 5 components each.

(b) Calculate the system lifetime for **Series** configuration ($Y = \min$) and **Parallel** configuration ($Z = \max$).

(c) Plot the histograms of Y and Z on the same graph.

(d) Calculate and print the Mean Time To Failure (MTTF) for both systems. Does it match the theoretical expectation?

```
import numpy as np
import matplotlib.pyplot as plt

lam = 0.5
n_components = 5
n_sims = 10000

# Generate Data (Exponential)
```

```
# components = np.random.exponential(scale=1/lam, size=(n_sims,
    n_components))

# Calculate System Lifetimes
# series_life = np.min(components, axis=1)
# parallel_life = np.max(components, axis=1)

# Plot and Compare Means
```

6.17 Log-Transformation for Skewed Data

In engineering, data is often right-skewed (e.g., failure times, income, rainfall). We often transform it to be Normal.

(a) Generate 1,000 data points from a Lognormal distribution ($\mu = 0, \sigma = 1$).

(b) Plot the histogram. Observe the skewness.

(c) Apply the natural log transformation: $Y = \ln(X)$.

(d) Plot the histogram of Y. Does it look Normal?

(e) Use `scipy.stats.probplot` to generate a Q-Q plot for Y to verify normality visually.

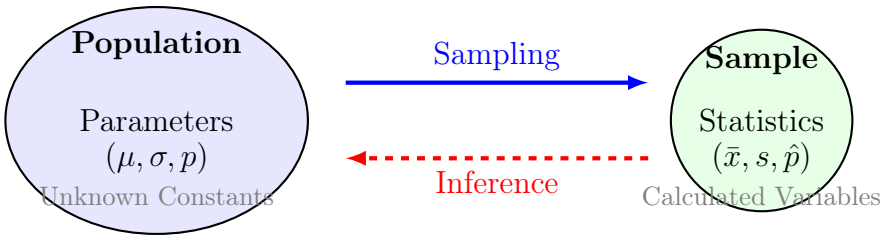
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Generate Lognormal Data
data = np.random.lognormal(mean=0, sigma=1, size=1000)

# Apply Transformation
transformed_data = np.log(data)

# Q-Q Plot
plt.figure()
stats.probplot(transformed_data, dist="norm", plot=plt)
plt.title("Q-Q Plot after Log Transformation")
plt.show()
```

Chapter 7

Sampling Distributions & The Central Limit Theorem

7.1 Basic Concept

7.1 Fundamental Definitions (The Big Picture)
The diagram below illustrates the core workflow of statistics. Use it to answer the questions.



[Image of statistical inference diagram population vs sample]

Explain the difference between the following pairs of terms. Give a concrete example for each.

(a) **Probability vs. Statistics:** Based on the diagram, which field moves from Left to Right (predicting the sample from a known population), and which moves from Right to Left?

(b) **Parameter vs. Statistic:** Identify which is a fixed constant (Truth) and which is a random variable that changes with every new sample.

(c) **Population vs. Sample:** Explain in terms of the set of all possible observations versus a subset used for analysis.

(d) **Random Sample vs. Sampling Distribution:** What is the difference between a single set of data $(x_1, \ldots, x_n)$ inside the green circle, and the probability distribution of the statistic $(\bar{X})$ if we repeated the blue arrow many times?

7.2 The Central Limit Theorem (CLT)

(a) State the Central Limit Theorem in your own words.

(b) What does the theorem say about the shape of the sampling distribution of $\bar{X}$ as $n \to \infty$, regardless of the population distribution?

(c) Does the CLT apply if the samples are not independent?

(d) It is highly recommended to **try** the Python application in Problem 7.10 to visually prove this theorem. Why is simulation often more convincing than mathematical proof for beginners?

7.2 Intermediate

7.3 Consistency of the Sample Mean

Explain the Weak Law of Large Numbers (WLLN) in your own words.

(a) What happens to the probability $P(|\bar{X}_n - \mu| > \epsilon)$ as the sample size $n \to \infty$?

(b) How does this justify using $\bar{x}$ from a large sample as an estimate for the population mean μ?

7.4 Sample Size Determination (Chebyshev's Approach)

You want to estimate the mean life of a battery. You know the variance is $\sigma^2 = 100$. You want your sample mean $\bar{X}$ to be within ± 2 hours of the true mean μ with a probability of at least 0.95.

Using Chebyshev's Inequality for the sample mean (where $Var(\bar{X}) = \sigma^2/n$), calculate the minimum sample size n required.

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

7.5 Sampling Distribution of the Mean (Normal Population)

The compressive strength of a specific concrete mix is normally distributed with $\mu = 4000$ psi and $\sigma = 200$ psi. A civil engineer takes a random sample of $n = 25$ specimens.

(a) **Parameters:** Calculate the expected value ($E[\bar{X}]$) and the standard error ($\sigma_{\bar{X}}$) of the sampling distribution.

(b) **Probability:** Find the probability that the sample mean is less than 3950 psi.

(c) **Inverse Calculation:** Find the value $\bar{x}$ such that 95% of the sample means exceed this value.

(d) **Sample Size Effect:** If the engineer wants to reduce the standard error by half, what sample size n_{new} is required?

7.6 Sampling Distribution of the Sum (Elevator Load)

The weight of an adult male is a random variable with mean $\mu = 170$ lbs and standard deviation $\sigma = 20$ lbs. An elevator has a maximum load limit of 2800 lbs. Suppose 16 males enter the elevator.

(a) Let $S = \sum_{i=1}^{16} X_i$ be the total weight. What are the mean and standard deviation of S?

(b) Based on the CLT, what is the approximate distribution of S?

(c) Calculate the probability that the total weight exceeds the maximum load limit.

(d) Is it necessary to assume that the weight of individual males is normally distributed to answer part (c)? Explain.

7.7 Sampling Distribution of Proportion (Quality Control)

A manufacturing process produces electronic chips with a defect rate of 8% ($p = 0.08$). A quality control inspector takes a random sample of $n = 400$ chips.

(a) Describe the sampling distribution of the sample proportion $\hat{P}$. Is the normal approximation valid? (Check np and $n(1-p)$).

(b) Calculate the mean and standard error of $\hat{P}$.

(c) Find the probability that the sample proportion of defects exceeds 10% ($\hat{P} > 0.10$).

(d) Find the probability that the sample proportion lies between 0.06 and 0.10.

7.8 Sampling Distribution of Variance (Chi-Square)

A precision instrument manufacturing process has a variance of $\sigma^2 = 0.01$ mm^2. A sample of $n = 20$ instruments is selected.

(a) Construct the statistic related to the sample variance S^2 that follows a Chi-square distribution. What are the degrees of freedom?

(b) Find the probability that the sample variance exceeds 0.015 mm^2.

(c) Find the critical values a and b such that $P(a < S^2 < b) = 0.95$ (Symmetric probability).

7.9 Ratio of Variances (F-Distribution)

Two independent machines are being compared. Machine A has a variance of $\sigma_A^2 = 10$ and Machine B has a variance of $\sigma_B^2 = 15$. Samples of sizes $n_A = 16$ and $n_B = 21$ are taken.

(a) Identify the distribution of the statistic $F = \frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2}$.

(b) If we simply calculate the ratio of sample variances $R = S_A^2/S_B^2$, what is the expected value of this ratio? (Qualitative answer).

(c) Using F-distribution tables (or software), find the probability $P(S_B^2 > 2S_A^2)$. *(Hint: Rearrange into the F-statistic form).*

7.10 Difference of Two Means (Comparison)

Two types of cooling fans are compared.

- Fan A: $\mu_A = 80$ CFM, $\sigma_A = 5$ CFM
- Fan B: $\mu_B = 75$ CFM, $\sigma_B = 3$ CFM

A sample of size $n_A = 50$ and $n_B = 50$ is taken.

(a) Find the probability that the sample mean of Fan A is greater than Fan B by at least 7 CFM ($P(\bar{X}_A - \bar{X}_B \geq 7)$).

(b) Find the probability that Fan B actually has a higher sample mean than Fan A ($P(\bar{X}_A < \bar{X}_B)$).

7.3 Application

7.11 Convergence of Sample Mean (Law of Large Numbers)

The Law of Large Numbers states that as $n \to \infty$, $\bar{X} \to \mu$. Chebyshev's Inequality gives us a theoretical bound on this convergence. Run the following code to visualize how the sample mean converges to the true mean.

```
import numpy as np
import matplotlib.pyplot as plt

# 1. Population: Exponential (Mean = 10)
true_mean = 10
pop_data = np.random.exponential(true_mean, 100000)

# 2. Simulation: Increasing Sample Size n
sample_sizes = range(1, 2000, 5)
sample_means = []

for n in sample_sizes:
    sample = np.random.choice(pop_data, n)
    sample_means.append(np.mean(sample))

# 3. Plot
plt.figure(figsize=(10, 6))
plt.plot(sample_sizes, sample_means, label='Sample Mean', alpha
    =0.6)
plt.axhline(y=true_mean, color='r', linestyle='--', label='True
    Mean (Parameter)')
plt.xlabel('Sample Size (n)')
plt.ylabel('Sample Mean')
plt.title('Convergence of Sample Mean to Population Mean')
plt.legend()
plt.show()
```

Interpret: What happens to the fluctuations of the sample mean as n increases? How does this relate to the concept of "Consistency"?

7.12 Visualizing the Central Limit Theorem

Simulate drawing samples from a **Uniform Distribution** (Flat shape) to see how the distribution of $\bar{X}$ becomes Bell-shaped.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# 1. Population (Uniform 0 to 10)
# Theoretical Mean = 5, Var = 100/12 = 8.33
pop_data = np.random.uniform(0, 10, 100000)

# 2. Draw Samples (n=30) and calculate Means
n = 30
num_simulations = 5000
means = [np.mean(np.random.choice(pop_data, n)) for _ in range(
    num_simulations)]

# 3. Plot Histogram vs Normal Curve
plt.hist(means, bins=50, density=True, alpha=0.6, color='g', label=
    'Sample Means')

# Overlay Normal PDF
mu = 5
sigma_xbar = np.sqrt(8.333 / n)
x = np.linspace(3, 7, 100)
plt.plot(x, norm.pdf(x, mu, sigma_xbar), 'r-', label='Theoretical
    Normal')

plt.title(f'Sampling Distribution of Mean (n={n})')
plt.legend()
plt.show()
```

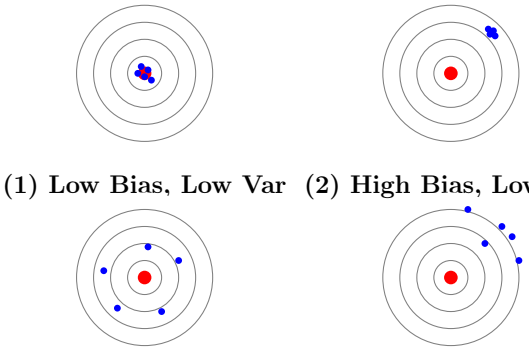
Chapter 8

Point Estimation

8.1 Basic Concept

8.1 The Goal of Estimation

Concept: In engineering, we estimate parameters (like the true center of a target) from data. The quality of an estimator is often described using the "Target Board" analogy.



(1) Low Bias, Low Var **(2) High Bias, Low Var**

(3) Low Bias, High Var **(4) High Bias, High Var**

(a) Define **Point Estimator** vs. **Estimate**. Which one is a random variable?

(b) Explain the concept of **Consistency**. Why is the sample mean $\bar{X}$ considered a consistent estimator for μ?

(c) Match the terms **Bias** (Accuracy) and **Variance** (Precision) to the target board diagrams above. Which scenario (1-4) corresponds to an estimator that is "Unbiased but Inefficient"?

8.2 Properties of Estimators

(a) **Unbiasedness:** Show that $E[\bar{X}] = \mu$ regardless of the population distribution.

(b) **Efficiency:** If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased, how do we choose the "better" one?

(c) **MSE:** Prove algebraically that $MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$.

8.2 Intermediate

8.3 MOM vs. MLE: Poisson Distribution
Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ (mean arrival rate).

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

(a) **MOM:** Find the Method of Moments estimator $\hat{\lambda}_{MOM}$.

(b) **MLE:** Write the log-likelihood function and derive the Maximum Likelihood Estimator $\hat{\lambda}_{MLE}$.

(c) **Invariance Property:** An engineer wants to estimate the probability of "zero defects" ($P(X = 0) = e^{-\lambda}$). What is the MLE for this probability?

8.4 MLE for Pareto Distribution (Heavy Tail)
The Pareto distribution is often used to model failure times or income. Its PDF is given by:

$$f(x; \alpha) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, \quad x \geq x_m$$

Assume the minimum value x_m is known (constant). We want to estimate the shape parameter α.

(a) Write the Likelihood function $L(\alpha)$ for a sample $x_1, \ldots, x_n$.

(b) Derive the Maximum Likelihood Estimator $\hat{\alpha}_{MLE}$.

(c) Interpret the result: How does $\hat{\alpha}_{MLE}$ relate to the geometric mean or the sum of logs of the data?

8.5 Parameter Estimation for a Power Function
The lifetime X of a certain mechanical part follows a distribution with PDF:

$$f(x; \theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0$$

(a) Find the population mean $E[X]$ and derive $\hat{\theta}_{MOM}$ in terms of $\bar{X}$.

(b) Derive the Maximum Likelihood Estimator $\hat{\theta}_{MLE}$.

(c) Suppose a sample of size $n = 3$ yields values: $0.5, 0.8, 0.9$. Calculate the estimates using both methods.

8.6 Rayleigh Distribution (Signal Processing)

The magnitude of a signal X follows a Rayleigh distribution with parameter σ^2:

$$f(x;\sigma^2) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x > 0$$

(a) Find the MLE for the parameter $\theta = \sigma^2$.

(b) Check if this estimator is Unbiased. (Hint: Recall that for Rayleigh, $E[X^2] = 2\sigma^2$).

8.7 Comparing Two Estimators (MSE Analysis)

Suppose we want to estimate the population mean μ. We have two proposed estimators based on a sample of size n:

- Estimator 1: $\hat{\mu}_1 = \bar{X}$ (Sample Mean)
- Estimator 2: $\hat{\mu}_2 = \frac{X_1 + X_n}{2}$ (Mid-range of first and last sample only)

Assume $X_i \sim N(\mu, \sigma^2)$.

(a) Show that both estimators are **Unbiased**.

(b) Find the Variance of both estimators. (Recall $Var(\bar{X}) = \sigma^2/n$ and independence).

(c) Which estimator is more efficient (has lower MSE) for $n > 2$? Explain why using all data is better than using just two points.

8.8 The Bias-Variance Trade-off

In engineering, an "Unbiased" estimator is not always the best choice. Sometimes, accepting a small Bias can significantly reduce the Variance, leading to a lower overall Mean Squared Error (MSE).

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

Consider a sample $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. We want to estimate the population variance σ^2. Let's consider a class of estimators defined by:

$$\hat{\sigma}_c^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where c is a constant. (Note: For sample variance S^2, $c = \frac{1}{n-1}$. For MLE, $c = \frac{1}{n}$).

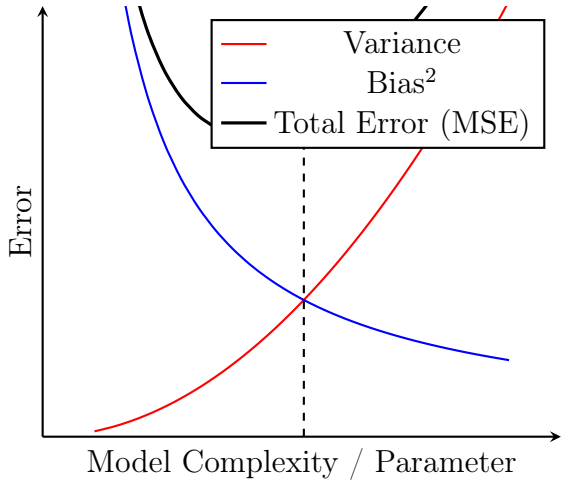
(a) Given that $E[\sum(X_i - \bar{X})^2] = (n-1)\sigma^2$ and $Var(\sum(X_i - \bar{X})^2) = 2(n-1)\sigma^4$. Show that the MSE of the estimator $\hat{\sigma}_c^2$ is:

$$MSE(c) = \sigma^4\left[2(n-1)c^2 + (c(n-1) - 1)^2\right]$$

(b) Find the value of c that **minimizes** the MSE. (Hint: Differentiate $MSE(c)$ with respect to c and set to 0).

(c) Does the value of c found in (b) correspond to the Unbiased Estimator (S^2) or the MLE? Or is it something else? What does this tell you about the "best" estimator in terms of MSE?

(d) Below is the classic Bias-Variance Trade-off graph representing model complexity (or parameter choice).



Question: Based on the graph, explain why minimizing **only** Bias (making it zero) might not lead to the lowest Total Error.

8.3 Challenge

8.9 MLE with Boundary Condition (Uniform)
Let $X_1, \ldots, X_n \sim U(0, \theta)$. The PDF is $f(x) = 1/\theta$ for $0 \leq x \leq \theta$.

(a) Write the Likelihood function $L(\theta)$. Specify the indicator function $\mathbb{I}(x_i \leq \theta)$.

(b) Explain why standard differentiation ($dL/d\theta = 0$) fails.

(c) Argue logically that $\hat{\theta}_{MLE} = X_{(n)}$ (the maximum of the sample).

(d) **Bias Check:** Given that $E[X_{(n)}] = \frac{n}{n+1}\theta$, find the bias. Construct an unbiased estimator from the MLE.

8.10 Shifted Exponential (Guarantee Time)
The time to failure follows a shifted exponential distribution (failures only occur after time δ):

$$f(x; \lambda, \delta) = \lambda e^{-\lambda(x-\delta)}, \quad x \geq \delta$$

(a) **Estimate δ:** Focusing on the constraint $x_i \geq \delta$, find the MLE $\hat{\delta}$.

(b) **Estimate λ:** Given $\hat{\delta}$, write the log-likelihood and maximize it to find $\hat{\lambda}$.

8.11 Optimal Scaling (MSE Trade-off)
Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. We want to estimate σ^2. Consider a class of estimators of the form $\hat{\sigma}^2_c = c \sum X_i^2$.

(a) Find the Bias and Variance of $\hat{\sigma}^2_c$ as a function of c.

(b) Find the value of c that minimizes the Mean Squared Error (MSE).

(c) Does the MSE-minimizing c correspond to the Unbiased estimator ($c = 1/n$) or the MLE?

8.4 Application

8.12 The Bias-Variance Trade-off (Variance Estimation)

In statistics, we often prefer *Unbiased* estimators (like S^2 with divisor $n-1$). However, minimizing the *Total Error* (MSE) is sometimes more important.

Run this simulation to compare two estimators for population variance $\sigma^2 = 100$ using a small sample size ($n = 5$).

- **Estimator A (Unbiased):** $S^2 = \frac{\sum(X-\bar{X})^2}{n-1}$
- **Estimator B (Biased MLE):** $\hat{\sigma}^2 = \frac{\sum(X-\bar{X})^2}{n}$

```python
import numpy as np
import matplotlib.pyplot as plt

# Configuration
true_var = 100
n = 5           # Small sample size highlights the trade-off
num_sims = 10000

# 1. Generate Data (Normal Distribution)
data = np.random.normal(0, np.sqrt(true_var), (num_sims, n))

# 2. Calculate Estimators
var_unbiased = np.var(data, axis=1, ddof=1) # Divisor n-1
var_mle = np.var(data, axis=1, ddof=0)      # Divisor n

# 3. Calculate Metrics
bias_unbiased = np.mean(var_unbiased) - true_var
bias_mle = np.mean(var_mle) - true_var

mse_unbiased = np.mean((var_unbiased - true_var)**2)
mse_mle = np.mean((var_mle - true_var)**2)

print(f"Unbiased (n-1): Bias = {bias_unbiased:.2f}, MSE = {
    mse_unbiased:.2f}")
print(f"MLE (n):        Bias = {bias_mle:.2f},       MSE = {mse_mle
    :.2f}")

# 4. Visualization
plt.figure(figsize=(10, 6))
plt.hist(var_unbiased, bins=50, alpha=0.5, label='Unbiased (n-1)',
    density=True)
plt.hist(var_mle, bins=50, alpha=0.5, label='MLE (n)', density=True
    )
plt.axvline(true_var, color='red', linestyle='dashed', linewidth=2,
     label='True Variance')
plt.title(f'Sampling Distributions of Variance Estimators (n={n})')
plt.legend()
plt.show()
```

Discussion:

(a) Look at the histograms: Which estimator is "centered" correctly at the red line?

(b) Look at the spread: Which estimator has a narrower width (smaller variance)?

(c) Check the MSE: Surprisingly, for small n, the biased estimator often has a *lower* MSE. Explain why this happens based on the formula $MSE = Variance + Bias^2$.

8.13 The German Tank Problem (Uniform Estimation)

During WWII, the Allies wanted to estimate the total number of German tanks (N) based on the serial numbers of captured tanks. This is equivalent to estimating θ in a Uniform distribution $U(0, \theta)$.

Compare three estimators using simulation:

- **MLE:** $\hat{\theta}_{MLE} = \max(X)$ (The maximum serial number seen).
- **MOM:** $\hat{\theta}_{MOM} = 2\bar{X}$ (Twice the sample mean).
- **Jackknife/Corrected:** $\hat{\theta}_{Corr} = \frac{n+1}{n}\max(X)$ (Unbiased version of MLE).

```python
import numpy as np
import matplotlib.pyplot as plt

true_theta = 1000
n = 10   # Captured tanks
num_sims = 5000

# 1. Generate Data (Uniform 0 to 1000)
data = np.random.uniform(0, true_theta, (num_sims, n))

# 2. Calculate Estimators
theta_mle = np.max(data, axis=1)
theta_mom = 2 * np.mean(data, axis=1)
theta_corr = ((n + 1) / n) * np.max(data, axis=1)

# 3. Visualization
plt.figure(figsize=(12, 6))
plt.boxplot([theta_mle, theta_mom, theta_corr], labels=['MLE (Max)', 'MOM (2*Mean)', 'Corrected Max'])
plt.axhline(true_theta, color='red', linestyle='--', label='True Total')
plt.title('Comparison of Estimators for German Tank Problem')
plt.legend()
plt.show()
```

Discussion:

(a) Why is the MLE (Max) always strictly less than or equal to the true value? (This is called negative bias).

(b) Compare the variance (spread) of MOM vs. Corrected Max. Which one is much more precise?

(c) Why is the "Corrected Max" generally considered the best estimator here?

Chapter 9

Interval Estimation

9.1 Basic Concept

9.1 True or False: The Meaning of Confidence
A 95% Confidence Interval (CI) for the mean μ is calculated to be $[10, 20]$. Determine whether the following statements are **True** or **False** and explain why.

(a) "There is a 95% probability that the true population mean μ lies between 10 and 20."

(b) "If we repeat the experiment many times and calculate the interval each time, approximately 95% of those intervals will contain the true mean μ."

(c) "95% of the data points in the sample lie between 10 and 20."

(d) "The interval $[10, 20]$ is a random variable before the data is collected, but a fixed interval after collection."

9.2 Factors Affecting Width
How would the width of a Confidence Interval change (Narrower, Wider, or Unchanged) in the following scenarios?

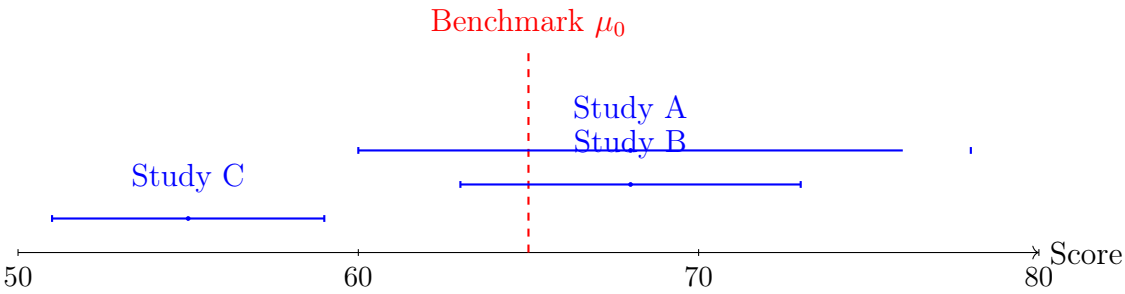
(a) Increasing the sample size n from 10 to 100.

(b) Increasing the confidence level $(1 - \alpha)$ from 90% to 99%.

(c) Using a sample with a larger standard deviation s.

9.3 Visual Forensics: Interpreting Confidence Intervals

The chart below visualizes the 95% Confidence Intervals for the mean score from three different independent studies (A, B, and C). The vertical dashed line represents the historical benchmark of $\mu_0 = 65$.



(a) Compare Study A and Study B. Both have the same sample mean ($\bar{x} = 68$) and assume they come from populations with similar variability (s). Which study likely used a **larger sample size (n)**? Explain your reasoning based on the interval width.

(b) Based on the interval for Study A, is the mean score **significantly different** from the benchmark ($\mu_0 = 70$) at the 5% significance level? Explain why using the concept of "containment".

(c) Look at the intervals for Study B and Study C. Do they overlap?

(d) Which study provides the most precise estimate of the population mean?

9.2 Intermediate

9.4 CI for Mean (Raw Data Processing)

An engineer tests the breaking strength (in MPa) of a new ceramic alloy. The observed data for $n = 10$ specimens is:

```
52, 48, 56, 45, 50, 53, 49, 51, 54, 42
```

(a) **Descriptive Stats:** Calculate the sample mean $\bar{x}$ and sample standard deviation s.

(b) **Interval Construction:** Construct a 95% Confidence Interval for the true mean strength μ.

(c) **Interpretation:** If the safety standard requires the mean strength to be at least 48 MPa, does this interval provide strong evidence that the alloy meets the standard?

9.5 Difference of Means (Independent Samples)

Two different chemical catalysts (Type A and Type B) are tested to see which one produces a higher yield.

- **Catalyst A:** 85, 88, 84, 86, 90, 83, 87, 85
- **Catalyst B:** 81, 78, 83, 82, 80, 79, 84, 81

Assume that the population variances are equal ($\sigma_A^2 = \sigma_B^2$).

(a) Calculate $\bar{x}_A, s_A^2$ and $\bar{x}_B, s_B^2$.

(b) Calculate the Pooled Variance S_p^2.

(c) Construct a 95% CI for the difference $\mu_A - \mu_B$.

(d) Does the interval suggest that Catalyst A produces a significantly higher yield than Catalyst B?

9.6 Variance Estimation (Precision Test)

A CNC machine is required to produce parts with very low variability. A random sample of 12 parts showed the following deviations from the target (in micrometers):

```
-2, 1, 0, -1, 2, -3, 0, 1, -2, 1, 3, -1
```

(a) Calculate the sample variance s^2.

(b) Construct a 90% CI for the true population variance σ^2.

(c) If the machine specification requires $\sigma^2 \leq 4$, does the data confirm that the machine is within spec with 90% confidence? (Check the Upper Confidence Limit).

9.7 Proportions (A/B Testing)

A website tests two button colors to see which gets more clicks.

- **Red Button:** 1000 views, 50 clicks.
- **Blue Button:** 1000 views, 70 clicks.

(a) Calculate the sample proportions $\hat{p}_{red}$ and $\hat{p}_{blue}$.

(b) Construct a 99% CI for the difference $p_{blue} - p_{red}$.

(c) Does the interval contain 0? What does this imply about the effectiveness of the Blue button?

9.3 Challenge

9.8 Introduction to Pivotal Quantity Method (Guided)

Concept: A "Pivotal Quantity" Q is a function of the data and the unknown parameter θ, whose distribution **does not depend on** θ.

Example: For $X \sim N(\mu, \sigma^2)$, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a pivot because $Z \sim N(0,1)$ (no μ in the distribution).

Task: Derive the CI for the mean μ (unknown σ) using the T-statistic as a pivot.

(a) Define the pivot $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$. State its distribution.

(b) Start with the probability statement:

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha$$

(c) Substitute T and rearrange the inequality to isolate μ in the center. Show that this leads to the standard formula $\bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$.

9.9 CI for Gamma Parameter (Pivotal Method)

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with known shape α but unknown scale β.

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

We want to find a Confidence Interval for β.

(a) **Identify the Pivot:** It is known that $Y = \frac{2}{\beta}\sum_{i=1}^{n} X_i$ follows a Chi-square distribution with $2n\alpha$ degrees of freedom ($\chi^2_{2n\alpha}$). Explain why Y is a valid pivotal quantity. (Does its distribution depend on β?).

(b) **Derive the Interval:** Start with $P(\chi^2_{1-\alpha/2} < Y < \chi^2_{\alpha/2}) = 1 - \alpha$. Substitute Y and solve for β to find the $100(1-\alpha)\%$ CI.

(c) **Application (Exponential):** If $\alpha = 1$ (Exponential distribution) and we observe $n = 10$ failure times with $\sum x_i = 500$ hours. Calculate a 95% CI for the mean time to failure ($\mu = \beta$).

9.10 Prediction Interval vs. Confidence Interval

We want to predict the value of the **next single observation** X_{n+1}, not just the mean. The error is $X_{n+1} - \bar{X}$.

(a) Find the variance of this error: $Var(X_{n+1} - \bar{X})$.

(b) Write the formula for a 95% Prediction Interval using the t-distribution.

9.4 Application

9.11 Visualizing "Confidence"

Simulate the construction of 100 Confidence Intervals to understand what "95% Confidence" actually means.

(a) Generate a population with $\mu = 50, \sigma = 10$.

(b) Draw 100 separate samples (each $n = 30$).

(c) Calculate the 95% CI for each sample.

(d) Plot all 100 intervals as horizontal lines. Draw a vertical line at the true $\mu = 50$.

(e) Count how many intervals "capture" the true mean. Is it exactly 95? Why or why not?

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t

mu = 50
sigma = 10
n = 30
num_intervals = 100
confidence = 0.95

# 1. Simulate Intervals
captured_count = 0
plt.figure(figsize=(8, 10))

for i in range(num_intervals):
    sample = np.random.normal(mu, sigma, n)
    x_bar = np.mean(sample)
    s = np.std(sample, ddof=1)

    # Calculate Margin of Error (T-dist)
    crit_val = t.ppf((1 + confidence) / 2, df=n-1)
    margin_error = crit_val * (s / np.sqrt(n))

    low = x_bar - margin_error
    high = x_bar + margin_error

    # Check capture
    is_captured = low <= mu <= high
    if is_captured:
        captured_count += 1
        color = 'green'
    else:
        color = 'red'

    # Plot
    plt.plot([low, high], [i, i], color=color)

plt.axvline(mu, color='black', linestyle='--', label='True Mean')
plt.title(f'100 CIs Simulation (Captured: {captured_count}/100)')
plt.xlabel('Value')
```

```
plt.ylabel('Interval Index')
plt.show()
```

Chapter 10

Hypothesis Testing for 1 Sample

10.1 Basic Concept

10.1 The Logic of Hypothesis Testing (Industrial Context)
Imagine you are a Quality Assurance (QA) Engineer for an automated assembly line. The line has an Automated Optical Inspection (AOI) machine that flags defective parts.

- H_0: The part is Good (No defect).
- H_1: The part is Defective.

The machine makes a decision based on statistical thresholds. Explain the following in this specific context:

(a) **Type I Error (α):** What actually happens on the production line if this error occurs? Is it "False Alarm" or "Missed Detection"? What is the cost associated with this?

(b) **Type II Error (β):** What actually happens if this error occurs? What is the impact on the customer?

(c) **Trade-off:** If you adjust the machine to be extremely strict (to catch every single defect), which error probability increases? Why?

10.2 P-value Interpretation and Decision Making
A structural engineer tests the hypothesis that the mean load capacity of a beam is greater than 50 tons. The calculated P-value is 0.042.

(a) **Misconception Check:** Does this mean there is a 95.8% probability that the mean load is greater than 50 tons? If false, provide the correct interpretation.

(b) **Decision:** If the safety regulation requires a significance level of $\alpha = 0.01$ (1% risk tolerance), should the engineer certify the beam?

(c) **Impact:** How does lowering α from 0.05 to 0.01 affect the probability of approving a beam that actually meets the requirements (Power)?

10.2 Intermediate

10.3 Z-Test: Cement Filling Process Control

Scenario: You are managing a cement packaging plant in Saraburi. The filling machine is set to fill bags with a target mean of $\mu_0 = 50$ kg. From historical process capability data, the standard deviation is known to be $\sigma = 1.2$ kg.

Recently, customers have complained that the bags feel lighter than usual. You decide to audit the process by randomly selecting 10 bags from the conveyor belt. The weights (in kg) are recorded as follows:

49.2	48.5	50.1	49.8	48.9
50.5	49.0	48.8	49.6	49.1

Analysis:

(a) alculate the sample mean $\bar{x}$. By looking at the value, does it seem "significantly" lower than 50, or could it be due to chance?

(b) Formulate the hypothesis to test if the mean weight is **significantly less** than 50 kg at $\alpha = 0.05$. Calculate the Z-statistic and make a decision.

(c) Find the critical weight x_{crit} (in kg) below which you would reject the lot. Does your $\bar{x}$ fall below this threshold?

(d) Suppose the machine calibration has indeed drifted, and the true mean is now $\mu_{true} = 49.0$ kg. Calculate the probability that your sampling plan ($n = 10$) fails to detect this drift (β). Is this risk acceptable?

10.4 T-Test: Aerospace Alloy Strength Qualification

Scenario: An aerospace company is qualifying a new Aluminum-Lithium alloy for wing structures. The manufacturer claims the mean Ultimate Tensile Strength (UTS) is greater than 600 MPa. Due to the high cost of destructive testing, only a small sample of $n = 8$ specimens is available.

The observed UTS values (MPa) are:

605	612	598	620
608	615	595	610

Analysis:

(a) What assumption must be made about the underlying distribution of the alloy strength to perform a T-test with such a small sample?

(b) Perform a one-tailed hypothesis test at $\alpha = 0.05$. Show the calculation of the sample mean, sample standard deviation, T-statistic, and degrees of freedom.

(c) Can you certify to the FAA (aviation authority) that the mean strength exceeds 600 MPa with 95% confidence?

(d) If the true mean strength were actually 602 MPa (only slightly better than the claim), would a sample of size $n = 8$ likely detect this? Discuss intuitively without calculation.

10.5 Z-Test for Proportion: Election Polling

Scenario: A political campaign manager claims that their candidate has secured "more than 55%" of the votes in a key district, which would guarantee a win without a runoff. To verify this, a media outlet conducts an independent exit poll of 500 randomly selected voters.

Data: Out of 500 voters, 290 indicated they voted for this candidate.

Analysis:

(a) Calculate the sample proportion $\hat{p}$.

(b) Test the campaign manager's claim ($p > 0.55$) at $\alpha = 0.05$. Compute the Z-statistic and P-value.

(c) Construct a 95% one-sided lower confidence bound for the true proportion. Does this bound exceed 0.55?

(d) If the true support level is actually a landslide 60% ($p = 0.60$), calculate the probability that this poll would **fail** to confirm the manager's claim (β).

10.6 Chi-Square Test: CNC Precision Machining

Scenario: A supplier of precision engine valves claims that their CNC process is highly capable, with a variance in valve diameter of no more than $\sigma^2 = 0.01$ mm^2. If the variance is higher, the valves might leak.

You inspect a random sample of 15 valves and measure the deviation from the target diameter (in mm):

```
-0.1, 0.2, 0.0, -0.2, 0.1, 0.3, -0.1, 0.0, 0.1, -0.3, 0.2, -0.1,
                        0.0, 0.1, -0.1
```

Analysis:

(a) Calculate the sample variance s^2 from the raw data.

(b) Test $H_0 : \sigma^2 \leq 0.01$ vs $H_1 : \sigma^2 > 0.01$ at a significance level of $\alpha = 0.05$.

(c) Do you have enough evidence to reject the supplier's shipment?

(d) In this scenario, define "Consumer's Risk" and "Producer's Risk". Which risk did you control by setting $\alpha = 0.05$?

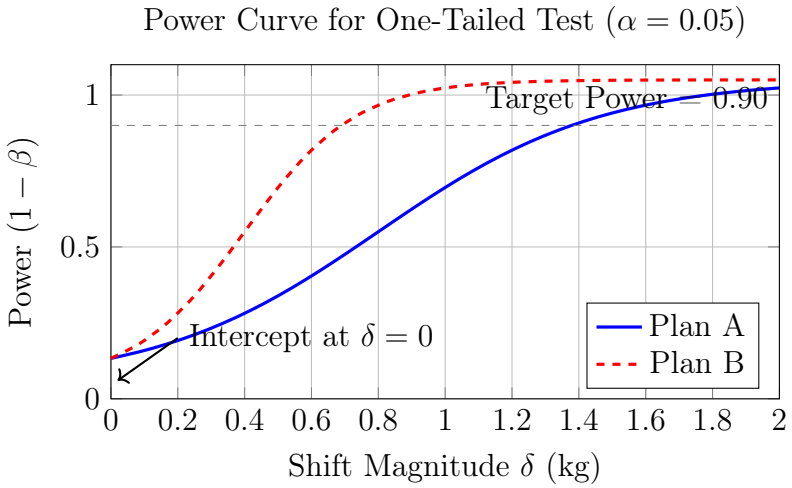
10.7 Power Function Analysis

Consider the cement filling process ($H_0 : \mu = 50$ vs. $H_1 : \mu < 50$).

Let δ denote the magnitude of the "true mean shift" from the target:

$$\delta = 50 - \mu_{true}$$

The graph below shows the **Power Curves** $(1-\beta)$ for two different sampling plans (Plan A and Plan B) as a function of the shift δ.



(a) Identify the y-intercept of the curves when $\delta = 0$ (i.e., when $\mu_{true} = 50$). What statistical probability does this value represent? (Hint: It is related to the significance level).

(b) Compare the steepness of Plan A vs. Plan B. Which plan corresponds to a **larger sample size** (n)? Explain your reasoning physically.

(c) Suppose management requires a 90% probability (Power = 0.9) of detecting a shift of $\delta = 1.0$ kg. Based on the graph, is Plan A sufficient, or must you use Plan B?

(d) If we relaxed the significance level from $\alpha = 0.05$ to $\alpha = 0.10$, how would the starting point (y-intercept) of these curves change?

10.3 Application

10.8 Visualizing the Power Curve

In engineering design, selecting the sample size (n) is a trade-off between cost and the ability to detect defects (Power). Run the following simulation to see how Power behaves.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# Configuration for Quality Control Z-Test (One-Tailed Lower)
# H0: mu >= 50 (Process is Good)
# H1: mu < 50  (Process Shifted/Bad)
mu0 = 50
sigma = 1.2
alpha = 0.05

# Compare two sample sizes
sample_sizes = [10, 50]
colors = ['blue', 'green']

plt.figure(figsize=(10, 6))

for n, col in zip(sample_sizes, colors):
    se = sigma / np.sqrt(n)
    # Critical Value: We reject if X_bar < crit_val
    crit_val = norm.ppf(alpha, loc=mu0, scale=se)

    # Simulate range of TRUE means (Process actually shifting down)
    true_means = np.linspace(48.5, 50.5, 200)
    powers = []

    for mu_true in true_means:
        # Power = P(Reject H0 | H1 is true)
        # Rejection region is (-inf, crit_val)
        power = norm.cdf(crit_val, loc=mu_true, scale=se)
        powers.append(power)

    plt.plot(true_means, powers, label=f'n={n}', color=col,
    linewidth=2)

plt.axvline(mu0, color='k', linestyle='--', label='Target Mean (50)
    ')
plt.axhline(0.05, color='r', linestyle=':', label='Alpha (0.05)')
plt.xlabel('True Process Mean (kg)')
plt.ylabel('Probability of Detecting Shift (Power)')
plt.title('Power Curve: Impact of Sample Size')
plt.legend()
plt.grid(True)
plt.show()
```

Discussion Questions:

(a) Observe the curve for $n = 10$ vs $n = 50$. If the true mean drops slightly to 49.5 kg, which sample size gives a higher probability of detecting this problem?

(b) Why does the power curve drop to 0.05 when the true mean is exactly 50?

(c) As a manager, if detecting a small shift of 0.5 kg is critical for customer safety, would you stick with $n = 10$ or pay more for $n = 50$?

Chapter 11

Statistical Inference for Two Samples

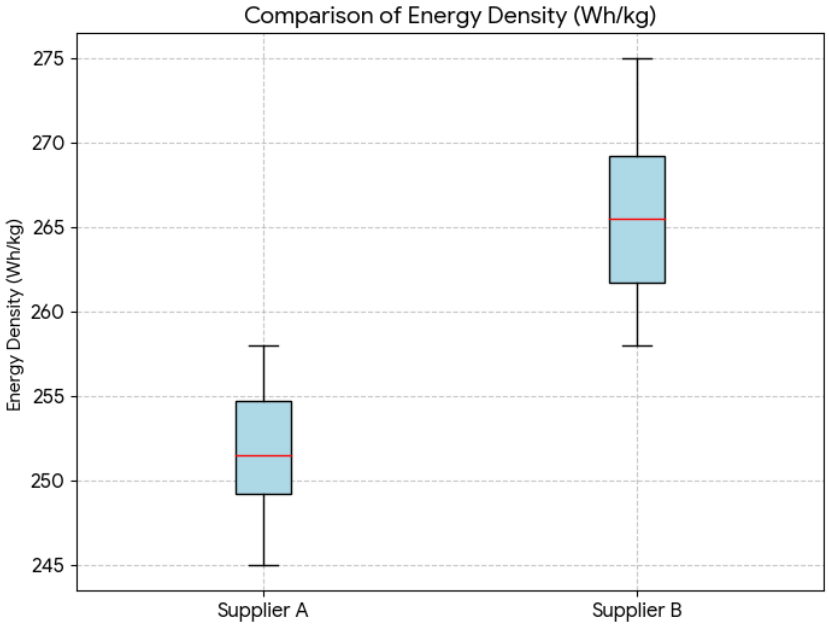
11.1 Intermidate

11.1 Battery Technology Comparison (Independent Samples)
Scenario: An Electric Vehicle (EV) manufacturer is choosing between two battery suppliers: *Supplier A* (Standard Li-ion) and *Supplier B* (New Solid-state composite). The critical performance metric is **Energy Density** (Wh/kg).

The engineering team collects random samples from both suppliers. The observed data (Wh/kg) is:

- **Supplier A** ($n_A = 10$)**:** 250, 255, 248, 252, 258, 245, 256, 251, 249, 254
- **Supplier B** ($n_B = 12$)**:** 265, 260, 272, 258, 268, 262, 275, 266, 270, 264, 261, 269

Visual Analysis: Refer to the boxplot generated from the lab data below:



Analysis:

(a) **Visual Interpretation:**

- Compare the **medians** (red lines). Which supplier appears to have higher density?
- Compare the **IQRs** (box heights). Which supplier has more consistent quality (less variability)?
- Is there any significant **overlap** between the distributions?

(b) Calculate the sample means ($\bar{x}_A, \bar{x}_B$) and sample variances (s_A^2, s_B^2).

(c) Test the claim that Supplier B provides a significantly **higher** mean energy density than Supplier A at $\alpha = 0.01$.

- Step 1: Check variance assumption ($F = s_B^2/s_A^2$). Assume equal variance if $P > 0.05$.
- Step 2: Calculate the T-statistic (using the Pooled Variance formula if appropriate).
- Step 3: Make a decision based on the critical value.

(d) Construct a 99% Confidence Interval for the difference $\mu_B - \mu_A$. Does the interval confirm your hypothesis test result?

11.2 VR Safety Training (Paired T-Test)

Scenario: A factory implements a new "Virtual Reality (VR) Safety Training" program. To evaluate its effectiveness, the safety knowledge scores (0-100) of 8 workers are recorded **Before** and **After** the training.

Worker ID	1	2	3	4	5	6	7	8
Before (X_1)	65	70	75	60	68	72	55	78
After (X_2)	75	78	74	70	80	79	65	82

Analysis:

(a) Why is an "Independent Two-Sample T-Test" inappropriate for this data? What characteristic of the data dictates the use of a "Paired T-Test"?

(b) Calculate the differences $d_i = \text{After}_i - \text{Before}_i$. Find the mean difference $\bar{d}$ and standard deviation of differences s_d.

(c) Test the hypothesis that the training significantly **increases** the safety score ($\mu_d > 0$) at $\alpha = 0.05$.

(d) Construct a 95% Confidence Interval for the mean improvement. Based on this, what is the *minimum* expected improvement for the average worker?

11.3 E-Commerce A/B Testing (Proportions)

Scenario: A marketing team performs an A/B test on a landing page to improve the "Click-Through Rate" (CTR).

- **Page A (Control):** Shown to 2,000 visitors, 160 clicked.
- **Page B (Variant):** Shown to 2,000 visitors, 200 clicked.

Analysis:

(a) Calculate the observed proportions $\hat{p}_A$ and $\hat{p}_B$.

(b) Can we state with 95% confidence that Page B performs better than Page A? Perform a Z-test for the difference of proportions.

(c) Construct a 95% Confidence Interval for the difference in proportions ($p_B - p_A$).

(d) If the website gets 100,000 visitors per month, and each click is worth \$5 on average, use the lower bound of the CI to estimate the *minimum additional monthly revenue* generated by switching to Page B.

11.2 Challenge

11.4 Derivation of Type II Error (β) for Two Samples

Consider the Z-test for the difference of two means with known variances σ_1^2, σ_2^2.

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 > \delta_0$$

We reject H_0 if the test statistic $Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > Z_\alpha$.

Task:

(a) **Definition:** State the definition of β in this context.

(b) **Derivation:** Derive the formula for β if the **true difference** is $\Delta = \mu_1 - \mu_2$ (where $\Delta > \delta_0$). Show that:

$$\beta = \Phi\left(Z_\alpha - \frac{\Delta - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

where Φ is the standard normal CDF.

(c) **Sample Size Formula:** Using the result above, show that for equal sample sizes ($n_1 = n_2 = n$), the required n for a specific power $(1 - \beta)$ is:

$$n = \frac{(Z_\alpha + Z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{(\Delta - \delta_0)^2}$$

11.3 Application

11.5 Interpreting Excel Output: Manufacturing Yield

An engineer uses Microsoft Excel (Data Analysis Toolpak) to compare the production yield of two different machines (Machine A and Machine B).

Scenario:

- $H_0 : \mu_A - \mu_B = 0$
- $H_1 : \mu_A \neq \mu_B$
- Significance Level $\alpha = 0.05$

Excel Output:

t-Test: Two-Sample Assuming Equal Variances		
	Machine A	*Machine B*
Mean	85.40	82.10
Variance	16.50	18.20
Observations	12	12
Pooled Variance	17.35	
Hypothesized Mean Difference	0	
df	22	
t Stat	2.129	
P(T<=t) one-tail	0.022	
t Critical one-tail	1.717	
P(T<=t) two-tail	0.044	
t Critical two-tail	2.074	

Questions:

(a) **Hypothesis Check:** Based on the hypothesis ($H_1 : \mu_A \neq \mu_B$), which P-value from the table should you use? (One-tail or Two-tail?)

(b) **Decision:** Compare the relevant P-value with $\alpha = 0.05$. Is there a statistically significant difference between the two machines?

(c) **Critical Value Approach:** Compare the absolute value of "t Stat" with "t Critical two-tail". Does this confirm your decision in (b)?

(d) **Pooled Variance:** Show the manual calculation of the "Pooled Variance" (17.35) using the provided Means and Variances to verify Excel's result.

11.6 Python: Determining Sample Size (Power Analysis)

Before running an expensive experiment to compare two drugs, a pharmaceutical company wants to know how many samples are needed.

Goal: Detect a "Small Effect Size" (Cohen's d = 0.5) with 80% Power ($1-\beta = 0.80$) at $\alpha = 0.05$. Run the following code to find the required sample size per group.

```python
from statsmodels.stats.power import TTestIndPower
import matplotlib.pyplot as plt
import numpy as np

# 1. Parameters
effect_size = 0.5  # Cohen's d (Mean Diff / Std Dev)
alpha = 0.05       # Significance Level
power = 0.80       # Desired Power
ratio = 1.0        # Ratio of sample sizes (n1/n2)

# 2. Calculate Sample Size
analysis = TTestIndPower()
sample_size = analysis.solve_power(effect_size=effect_size,
                                   power=power,
                                   alpha=alpha,
                                   ratio=ratio,
                                   alternative='two-sided')

print(f"Required Sample Size per Group: {sample_size:.2f}")

# 3. Plot Power Curve (varying n)
sample_sizes = np.arange(10, 100)
powers = analysis.power(effect_size=effect_size,
                        nobs1=sample_sizes,
                        alpha=alpha,
                        ratio=ratio)

plt.figure(figsize=(8, 5))
plt.plot(sample_sizes, powers, label='Power Curve (d=0.5)')
plt.axhline(0.8, color='r', linestyle='--', label='Target Power
    (0.8)')
plt.axvline(sample_size, color='g', linestyle='--', label=f'Req n={
    sample_size:.1f}')
plt.xlabel('Sample Size (n)')
plt.ylabel('Power')
plt.title('Sample Size Determination for Two-Sample T-Test')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()
```

Questions:

(a) Record the required sample size output by the code. Why must we round this number up?

(b) According to the graph, if we could only afford 20 samples per group, roughly what would be the Power of our test? (Is it acceptable?)

(c) If we wanted to detect a **smaller** difference (e.g., effect_size $= 0.2$), would the required sample size increase or decrease? Explain logically.

11.7 Python: A/B Testing for Proportions (Z-Test)

A marketing team runs an A/B test.

- Group A (Control): 1000 visitors, 45 conversions.
- Group B (Treatment): 1000 visitors, 65 conversions.

They want to know if Group B is significantly better ($p_B > p_A$).

```python
import numpy as np
from statsmodels.stats.proportion import proportions_ztest

# Data
count = np.array([65, 45])    # Successes (B, A)
nobs = np.array([1000, 1000]) # Trials (B, A)

# Z-Test for Proportions
# alternative='larger' means we test if Prop B > Prop A
stat, pval = proportions_ztest(count, nobs, alternative='larger')

print(f"Z-statistic: {stat:.4f}")
print(f"P-value: {pval:.4f}")

# Decision
alpha = 0.05
if pval < alpha:
    print("Reject H0: Design B is significantly better.")
else:
    print("Fail to Reject H0: No significant difference.")
```

Questions:

(a) Run the code and report the P-value.

(b) If the P-value is 0.026, what is the probability of observing such a difference (or more extreme) purely by chance if both designs were actually equal?

(c) Change the 'alternative' parameter to ''two-sided''. How does the P-value change? Why?

Chapter 12

Analysis of Variance (ANOVA)

12.1 Basic Concept

12.1 Why not multiple T-Tests?
Suppose we want to compare the means of $k = 5$ different manufacturing processes.

(a) If we perform pairwise T-tests for all combinations ($\binom{5}{2} = 10$ tests) at $\alpha = 0.05$, what happens to the overall Type I error probability (Family-wise error rate)?

(b) Explain how ANOVA solves this problem using the F-statistic.

12.2 Assumptions and Decomposition

(a) State the three key assumptions required for the ANOVA F-test to be valid.

(b) The fundamental identity of One-way ANOVA is $SST = SSTR + SSE$. Explain what each term represents in the context of "Signal" vs. "Noise".

12.2 Intermediate

12.3 Fill in the Blanks (One-Way ANOVA)
An engineer runs an experiment to compare the yield of 4 different chemical processes ($k = 4$). Five batches were run for each process ($n = 5$). Complete the missing values (A-F) in the ANOVA table below:

Source	**DF**	**SS**	**MS**	**F**	**P-value**
Treatments (Between)	(A)	150	(D)	(F)	< 0.05
Error (Within)	(B)	(C)	5		
Total	19	230			

Questions:

(a) Determine values for A, B, C, D, and F.

(b) Is there a significant difference between the process means?

12.4 Randomized Block Design (Tire Wear)

We want to compare the wear of 3 types of tires (A, B, C). However, the wear also depends on the *Car Model*. To control for this, we use "Car Model" as a **Block**.

Data (Wear in mm):

Block (Car)	Tire Type		
	Type A	Type B	Type C
Sedan	8	5	10
SUV	12	9	14
Truck	15	12	17
Sports	9	7	11

Task:

(a) Calculate SS_{Total}, $SS_{Treatments}$ (Tire), and SS_{Blocks} (Car).

(b) Construct the ANOVA table.

(c) Test if there is a significant difference between Tire Types at $\alpha = 0.05$.

(d) Why was Blocking important here? (Compare the MS_{Error} with Blocking vs. if you had ignored Car models).

12.5 Two-Way ANOVA (Interaction Effects)

A battery life experiment tests two factors:

- **Factor A (Material):** M1, M2
- **Factor B (Temperature):** Low, High

The ANOVA table is partially given:

Source	DF	SS	MS	F
Material	1	200	200	10.0
Temperature	1	180	180	9.0
Interaction (AxB)	1	80	80	4.0
Error	16	320	20	
Total	19	780		

Questions:

(a) Find the critical value $F_{0.05,1,16}$ (approx 4.49).

(b) Test for the significance of the **Interaction Effect** first. Is it significant?

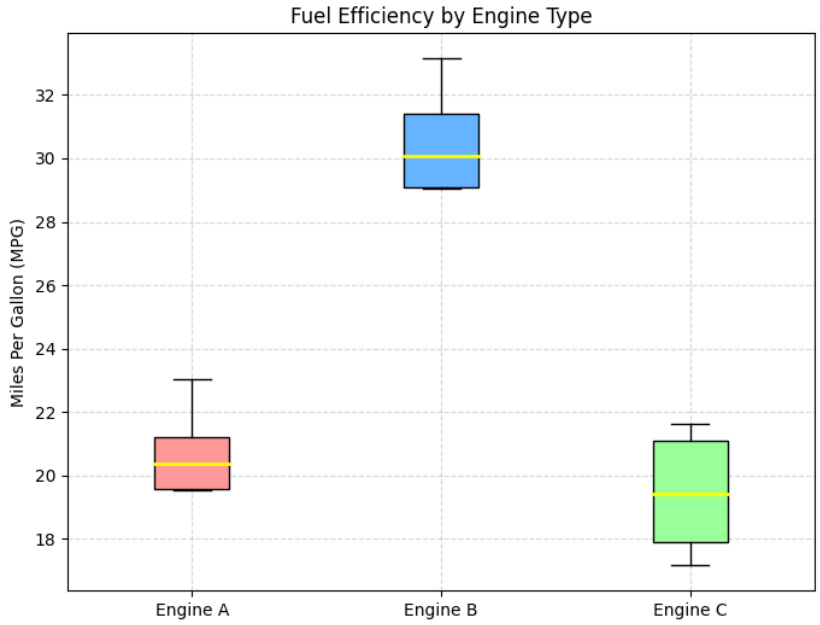
(c) Test for the Main Effects (Material and Temperature).

(d) Sketch a hypothetical "Interaction Plot" where lines cross each other. What does this crossing imply?

12.6 Visual Interpretation & ANOVA Table (Engine Efficiency)

An automotive engineer compares the fuel efficiency (mpg) of 3 different engine prototypes ($k = 3$). A sample of $n = 6$ cars is tested for each engine type.

Visual Inspection: Refer to the boxplot below:



Partial ANOVA Table:

Source	**DF**	**SS**	**MS**	**F**
Engines (Between)	(A)	350	(C)	(E)
Error (Within)	(B)	75	(D)	
Total	17	425		

Questions:

(a) Based on the boxplot, Engine B seems to have a much higher median than A and C. Do you expect the F-statistic to be large or small? Why?

(b) Fill in the missing values A, B, C, D, and E in the ANOVA table.

(c) If the critical value is $F_{crit} \approx 3.68$, is the difference statistically significant? Does this match your visual intuition?

12.7 Missing Values: Randomized Block (Agriculture)

An agronomist tests the effect of **4 types of fertilizer** on crop yield. The experiment is conducted across **5 different soil blocks** to control for soil quality variability.

Complete the missing values (A-F) in the Randomized Block ANOVA table:

Source	**DF**	**SS**	**MS**	**F**
Fertilizer (Treatments)	(A)	180	(D)	(F)
Soil Blocks	(B)	120	30	
Error	(C)	(E)	5	
Total	19	360		

Questions:

(a) Determine the values for A (Treatments), B (Blocks), and C (Error).

(b) Calculate the missing SSE (E) by using the identity $SS_{Total} = SS_{Trt} + SS_{Blk} + SS_{Err}$.

(c) Calculate the MS for Fertilizer (D) and the F-statistic (F).

(d) Is the effect of Fertilizer significant at $\alpha = 0.05$? (Assume $F_{crit} \approx 3.49$).

12.3 Applications

12.8 Visualizing ANOVA with Python

Before running ANOVA, it is crucial to visualize the data. Run this code to compare 3 teaching methods.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# 1. Simulate Data
data = {
    'Method A': [82, 85, 84, 86, 83, 85, 84],
    'Method B': [75, 78, 76, 79, 77, 76, 78],
    'Method C': [90, 92, 91, 93, 89, 91, 92]
}
df = pd.DataFrame(data)

# Melt for Boxplot
df_melt = df.melt(var_name='Method', value_name='Score')

# 2. Boxplot Interpretation
plt.figure(figsize=(8, 6))
sns.boxplot(x='Method', y='Score', data=df_melt)
plt.title('Student Scores by Teaching Method')
plt.grid(True, alpha=0.3)
plt.show()

# 3. One-Way ANOVA
f_stat, p_val = stats.f_oneway(df['Method A'], df['Method B'], df['
    Method C'])
print(f"F-Statistic: {f_stat:.2f}")
print(f"P-Value: {p_val:.4e}")
```

Questions:

(a) Look at the Boxplot. Do the boxes overlap? Which method appears to be the best?

(b) The P-value is extremely small. What does this confirm?

(c) **Post-hoc Logic:** Since ANOVA is significant, how do we know *specifically* which pairs differ? (Mention Tukey's Test).

12.9 Interpreting Excel Output: Marketing Strategy Analysis

A marketing manager tests 3 different advertising strategies (Strategy A, B, C) on 30 different regions (10 regions per strategy). The sales growth (%) is recorded.

Using Excel's "Data Analysis > Anova: Single Factor", the following output is generated:

SUMMARY

Groups	Count	Sum	Average	Variance
Strategy A	10	150	15.0	4.2
Strategy B	10	180	18.0	3.8
Strategy C	10	135	13.5	4.0

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	105.0	2	52.50	13.125	0.00011	3.354
Within Groups	108.0	27	4.00			
Total	213.0	29				

Questions:

(a) **Verification:** Show the manual calculation of the F-statistic using the MS values provided in the table. Does it match the "13.125" shown?

(b) **Hypothesis Test:**

- State H_0 and H_1.
- Compare the **P-value** with $\alpha = 0.05$. What is your conclusion?
- Compare the **F** with **F crit**. Does it lead to the same conclusion?

(c) **Understanding Variation:**

- Which value represents the variation caused by the different strategies? ($SS_{Between}$ or SS_{Within}?)
- Which value represents the random error or "noise"?

(d) **Conclusion:** Based on the "Average" column in the SUMMARY table, which strategy seems to be the most effective? Can we trust this observation statistically?

12.10 Two-Way ANOVA Output Interpretation (Interaction)

An engineer studies the effect of **Glass Type** (Type 1, Type 2) and **Temperature** (100C, 150C) on light output. The Excel output for "Anova: Two-Factor With Replication" is shown partially:

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample (Glass Type)	500	1	500	25.0	0.0001	4.49
Columns (Temp)	300	1	300	15.0	0.0012	4.49
Interaction	10	1	10	0.5	0.4895	4.49
Within	320	16	20			
Total	1130	19				

Questions:

(a) **Interaction Check:** Look at the "Interaction" row. Is the P-value less than 0.05? What does this imply about how Glass Type and Temperature affect the light output? (Do they depend on each other?)

(b) **Main Effects:** Are the main effects of Glass Type ("Sample") and Temperature ("Columns") significant?

(c) **Decision Strategy:** Since the Interaction is NOT significant, is it safe to interpret the Main Effects directly? (Yes/No).

Chapter 13

Chi-square Test

13.1 Basic Concept

13.1 Distinguishing the Tests

Match the scenario with the correct Chi-square test: (A) Goodness of Fit, (B) Test of Independence, (C) Test of Homogeneity.

(a) A Quality Engineer wants to check if the number of defects follows a Poisson distribution.

(b) A HR Manager wants to know if "Job Satisfaction" (High/Low) is related to "Work Shift" (Day/Night) using a single random sample of employees.

(c) A Production Manager samples 100 parts from Machine A, 100 from Machine B, and 100 from Machine C to see if the defect rate distribution is the same across all machines.

13.2 Intermediate

13.2 Goodness of Fit (Fair Die?)

A gambler suspects a die is loaded. He rolls it 60 times and observes the following:

Face	1	2	3	4	5	6
Observed (O_i)	5	8	12	15	12	8

Questions:

(a) **Hypothesis:** State H_0 regarding the distribution of the die faces.

(b) **Expectation:** Under H_0, what is the Expected Frequency (E_i) for each face? (Hint: Total $N = 60$).

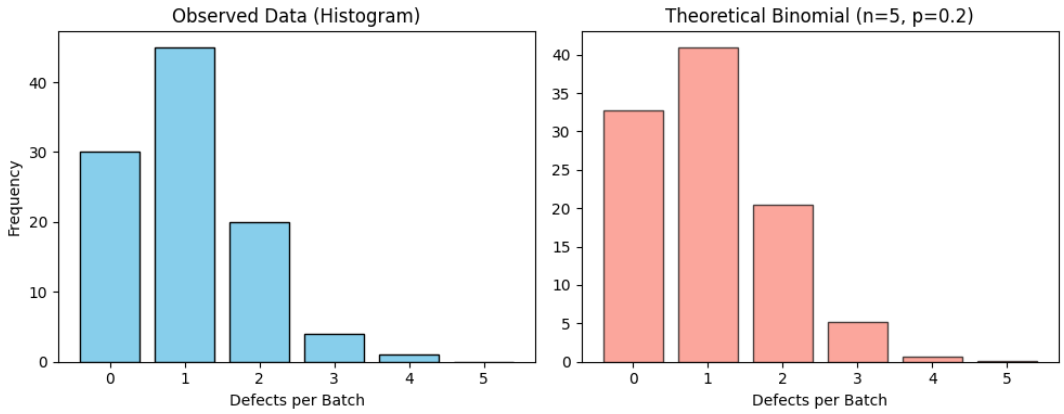
(c) **Calculation:** Calculate the Chi-square statistic $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

(d) **Decision:** At $\alpha = 0.05$ ($df = 5, \chi^2_{crit} = 11.07$), is the die biased?

13.3 Goodness of Fit: Binomial Distribution (Visual Comparison)

A Quality Control engineer inspects batches of products. Each batch contains $n = 5$ items. He collects data from $N = 100$ batches and counts the number of defective items in each.

Visual Analysis: Compare the Observed Data (Left) with the Theoretical Binomial Distribution (Right) generated assuming $p = 0.2$.



Data Table:

Defects (x)	0	1	2	3	4	5
Observed (O_i)	30	45	20	4	1	0
Expected (E_i)	32.8	41.0	20.5	5.1	0.6	0.0

Questions:

(a) **Visual Check:** Does the shape of the observed histogram roughly match the theoretical one? Are there any obvious discrepancies?

(b) **Parameter Estimation:** If we didn't know $p = 0.2$, how would you calculate $\hat{p}$ from the observed data? (Hint: Total defects divided by Total items).

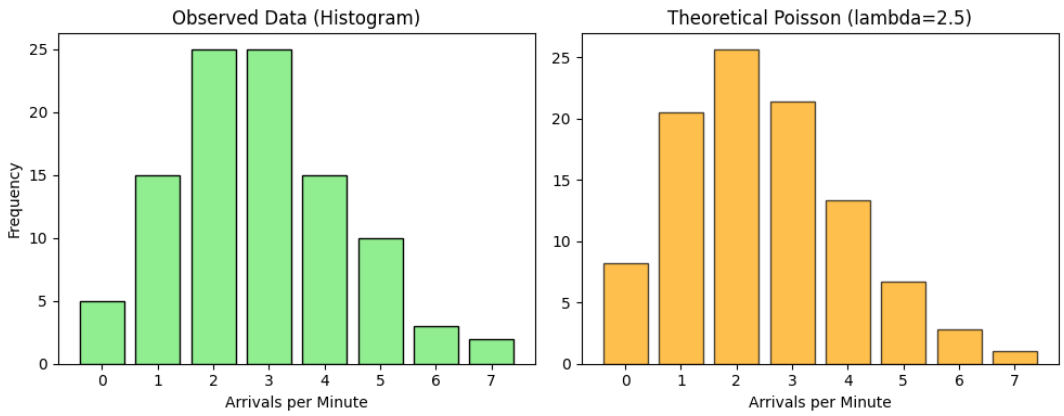
(c) **Rule of Thumb:** Notice that for $x = 4$ and $x = 5$, the Expected frequencies are very small (< 5). What should we do with these categories before calculating χ^2?

(d) **Degrees of Freedom:** If we group classes $x \geq 3$ together, and we estimated p from the data, what are the degrees of freedom? $(k - 1 - m)$.

13.4 Goodness of Fit: Poisson Distribution (Queueing)

The number of customers arriving at a service counter per minute is recorded for 100 minutes.

Visual Analysis: Compare the Observed Data (Left) with the Theoretical Poisson Distribution ($\lambda = 2.5$) (Right).



Questions:

(a) **Interpretation:** Poisson distribution is typically right-skewed. Does the observed data show this skewness?

(b) **Calculation:** Calculate the Expected Frequency for $x = 0$ using $P(X = 0) = e^{-2.5}\frac{2.5^0}{0!} \approx 0.082$. Does it match the graph (approx 8.2)?

(c) **Outliers:** If we observed a minute with 10 arrivals (which is very rare in the theoretical plot), how would that likely affect the Chi-square statistic?

13.5 Test of Independence (Partial Table)

A survey was conducted to see if there is an association between *Customer Age* (Young, Adult) and *Preferred Coffee Type* (Latte, Espresso). Total respondents $N = 200$.

Observed Frequencies (O_{ij}):

	Latte	**Espresso**	**Row Total**
Young	60	40	100
Adult	30	70	100
Col Total	90	110	200

Expected Frequencies (E_{ij}) - Partial:

	Latte	**Espresso**
Young	(A)	(B)
Adult	45	55

Questions:

(a) **Fill in the blanks:** Calculate the missing Expected values (A) and (B) using the formula $E_{ij} = \frac{R_i C_j}{N}$.

(b) **Degrees of Freedom:** What is the df for this 2×2 table?

(c) **Interpretation:** If the calculated $\chi^2 = 18.18$ and the critical value is 3.84, what is your conclusion? Is Age independent of Coffee Preference?

13.6 Test of Homogeneity (Defect Rates)

Three different shifts (Morning, Afternoon, Night) are monitored to see if the proportion of defective parts is the same.

Data:

- **Morning:** 200 parts, 10 defective.
- **Afternoon:** 200 parts, 15 defective.
- **Night:** 200 parts, 35 defective.

Task:

(a) Construct the 2×3 contingency table (Rows: Shift, Cols: Defect/Good).

(b) Calculate the expected number of defective parts for the "Night" shift under the Null Hypothesis (Homogeneity).

(c) Without calculating the full χ^2, looking at the data (10 vs 15 vs 35), does it look like the Night shift has a problem?

13.3 Applications

13.7 Chi-square Test with Python

Run the following code to perform a Test of Independence between "Gender" and "Product Preference".

```python
import pandas as pd
from scipy.stats import chi2_contingency

# Contingency Table
# Rows: Male, Female
# Cols: Product A, Product B, Product C
data = [[20, 30, 50],   # Male
        [40, 50, 10]]   # Female

# Perform Test
chi2, p, dof, expected = chi2_contingency(data)

print(f"Chi-square Statistic: {chi2:.2f}")
print(f"P-value: {p:.4e}")
print(f"Degrees of Freedom: {dof}")
print("Expected Frequencies:")
print(expected)
```

Questions:

(a) What is the P-value? Is it significant at $\alpha = 0.05$?

(b) Look at the 'expected' output. Are there any cells with $E_{ij} < 5$? Why is this check important?

(c) If the P-value is very small, what does it imply about Gender and Product Preference?

13.8 Interpreting Excel Output: Test of Independence

A factory manager wants to check if the **Type of Defect** (Crack, Scratch, Dent) depends on the **Machine** (Machine A, Machine B).

He performs a Chi-square Test in Excel (using a contingency table setup). The output is shown below:

Actual (Observed) Table:

	Crack	**Scratch**	**Dent**	**Total**
Machine A	15	20	5	40
Machine B	25	10	25	60
Total	40	30	30	100

Excel Calculation Block:

Metric	**Value**
Pearson Chi-square (χ^2)	16.875
Degrees of Freedom (df)	2
P-value	0.000216
Critical Value (0.05)	5.991

Questions:

(a) **Manual Check:** Calculate the Expected Frequency for (Machine A, Crack).

$$E = \frac{\text{Row Total} \times \text{Col Total}}{\text{Grand Total}}$$

Does the machine seem to produce fewer Cracks than expected (Observed 15 vs Expected)?

(b) **Hypothesis:**

- H_0: Defect Type is **independent** of Machine.
- H_1: Defect Type is **dependent** on Machine.

Based on the P-value (0.0002), what is your conclusion?

(c) **Business Insight:** Look at the Observed Table again. Machine B has 25 Dents compared to Machine A's 5. If the test is significant, what action should the engineer take regarding Machine B?

Chapter 14

Introduction to Non-parametric Tests

14.1 Basic Concept

14.1 Parametric vs. Non-parametric

(a) **Assumption Check:** Parametric tests (like T-test, ANOVA) rely on specific assumptions about the population distribution (e.g., Normality). What is the main advantage of Non-parametric tests regarding these assumptions?

(b) **Data Type:** If your data is "Ordinal" (e.g., Satisfaction Rating 1-5) or highly skewed with outliers, which type of test is preferred?

(c) **Efficiency:** If the data *is* actually Normal, why is using a T-test better than a Non-parametric test? (Discuss "Power of the Test").

14.2 The Art of Ranking (Handling Ties)

Non-parametric tests often use "Ranks" instead of raw values. Convert the following dataset into Ranks (1 = Smallest).

```
Data:  12, 5, 8, 5, 20, 8, 8, 15
```

Task:

(a) Sort the data from smallest to largest.

(b) Assign ranks. How do you handle "Ties" (e.g., the value 5 appears twice, value 8 appears three times)? Calculate the average rank for these ties.

(c) List the final ranks for the original data sequence.

14.2 Intermediate

14.3 The Sign Test (One Sample Median)

Scenario: A production manager claims the **Median** assembly time is less than 15 minutes. A sample of 10 workers shows the following times:

```
14, 12, 16, 13, 15, 11, 14, 18, 10, 13
```

Steps:

(a) Compare each value to the hypothesized median ($M_0 = 15$). Assign a (+) sign if > 15 and (-) if < 15. What happens to the value exactly equal to 15?

(b) Count the number of minus signs (x) and total valid observations (n).

(c) Test the hypothesis using the Binomial logic (or table). Is there enough evidence to say the median is less than 15?

14.4 Wilcoxon Signed-Rank Test (Paired Data)

Scenario: Testing a new pain-relief drug. Patients rate their pain (1-10) **Before** and **After** taking the drug.

Patient	1	2	3	4	5	6
Before	8	7	6	9	5	8
After	5	6	7	4	5	2

Steps:

(a) Calculate difference $d_i = \text{After} - \text{Before}$. Discard any $d_i = 0$.

(b) Rank the **absolute** differences $|d_i|$.

(c) Assign the sign of the original difference to the ranks.

(d) Calculate W_+ (Sum of positive ranks) and W_- (Sum of negative ranks).

(e) Which value (W_+ or W_-) represents the evidence *against* the drug being effective?

14.5 Mann-Whitney U Test (Two Independent Samples)

Scenario: Comparison of salaries (in \$1000s) between two departments. The data is skewed (not normal).

- **Dept A ($n_1 = 4$):** 40, 45, 50, 200 (Outlier)
- **Dept B ($n_2 = 5$):** 35, 38, 42, 44, 48

Steps:

(a) Combine all data and rank them from 1 to 9.

(b) Calculate the Sum of Ranks for Dept A (R_1).

(c) Calculate the U statistic: $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$.

(d) Explain why the outlier (200) affects the Rank-based test less than it would affect a T-test mean.

14.6 Kruskal-Wallis Test (Three Groups)

Scenario: Comparing the effectiveness of 3 different diets on weight loss. (Equivalent to One-Way ANOVA).

- **Diet A:** Ranks: 1, 3, 5
- **Diet B:** Ranks: 2, 4, 6
- **Diet C:** Ranks: 7, 8, 9

Question: Without calculating the full H-statistic, look at the ranks. Diet C has the highest ranks (7, 8, 9). Does this suggest a significant difference between groups? Explain the logic of "Sums of Ranks" in this test.

14.3 Applications

14.7 T-Test vs. Mann-Whitney on Skewed Data

Run the following Python code to see why we need Non-parametric tests when assumptions are violated.

```
import numpy as np
from scipy import stats

# 1. Create Skewed Data (e.g., Income)
# Group A: Normal-ish
group_A = [30, 32, 35, 33, 31, 34]
# Group B: Generally lower, but has one HUGE outlier
group_B = [25, 26, 24, 25, 27, 1000]

# 2. Parametric T-Test
t_stat, p_t = stats.ttest_ind(group_A, group_B)

# 3. Non-Parametric Mann-Whitney U Test
u_stat, p_u = stats.mannwhitneyu(group_A, group_B)

print(f"Mean A: {np.mean(group_A):.2f}, Mean B: {np.mean(group_B)
    :.2f}")
print(f"T-Test P-value: {p_t:.4f}")
print(f"Mann-Whitney P-value: {p_u:.4f}")
```

Questions:

(a) Look at the Means. Which group has a higher mean? Is this representative of the "typical" value?

(b) The T-test might fail to find a significant difference (High P-value) because the outlier inflates the Variance.

(c) The Mann-Whitney test likely yields a low P-value (Significant). Why is it able to detect that Group A is generally "larger" than Group B despite the outlier?

14.8 Spearman's Rank Correlation

We want to check the correlation between "Study Hours" and "Exam Rank".

- Study Hours: [10, 2, 8, 20, 5]
- Exam Rank: [2, 5, 3, 1, 4]

Since "Exam Rank" is already ordinal, we should use Spearman's ρ instead of Pearson's r.

Task: Convert "Study Hours" to Ranks. Compare the ranks with "Exam Rank". If they match perfectly, what is the value of ρ?

Chapter 15

Simple Linear Regression

15.1 Basic Concept

15.1 The Probabilistic Model

The Simple Linear Regression model is defined as $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

(a) **Deterministic vs. Stochastic:** Explain which part of the equation represents the "Signal" (Mean response) and which part represents the "Noise".

(b) **Assumptions on ϵ:** State the 4 key assumptions regarding the error term ϵ (often remembered as "LINE": Linearity, Independence, Normality, Equal Variance).

(c) **Residuals:** What is the difference between the true error term ϵ_i and the residual e_i?

15.2 Correlation vs. Regression

(a) If the correlation coefficient $r = 0.9$, does it imply that the slope β_1 is 0.9? What is the relationship between r and β_1?

(b) Why is it dangerous to use a regression line derived from data $x \in [10, 20]$ to predict Y at $x = 100$? (Concept of Extrapolation).

15.2 Intermediate

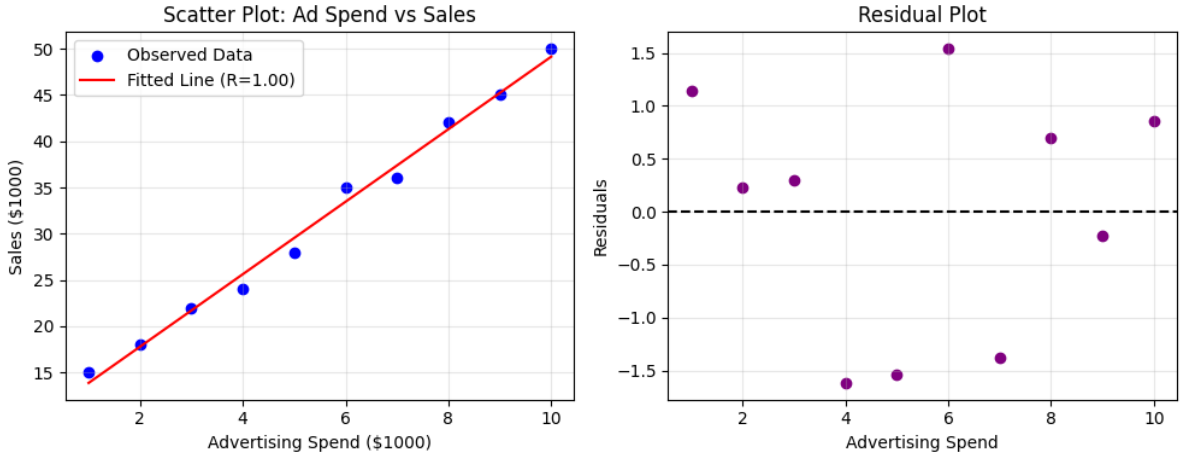
15.3 Sales Prediction Analysis ($n = 10$)

A marketing manager wants to predict **Sales (Y, in \$1000s)** based on **Advertising Spend (X, in \$1000s)**.

Data:

X	1	2	3	4	5	6	7	8	9	10
Y	15	18	22	24	28	35	36	42	45	50

Visual Inspection:



Tasks:

(a) **Sum of Squares Calculation:** Calculate the following terms manually:

$$\bar{x}, \bar{y}, S_{xx} = \sum (x - \bar{x})^2, S_{yy} = \sum (y - \bar{y})^2, S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

(b) **Model Estimation:** Calculate the least squares estimates b_1 (slope) and b_0 (intercept). Write the fitted regression equation $\hat{y}$.

(c) **Goodness of Fit:** Calculate the Coefficient of Determination (R^2) and Correlation Coefficient (r). Interpret R^2 in the context of sales variability.

(d) **ANOVA Table for Regression:** Complete the ANOVA table below to test the significance of the model.

Source	**DF**	**SS**	**MS**	**F**
Regression (SSR)	1	$b_1 S_{xy}$	?	?
Error (SSE)	$n - 2$	$S_{yy} - SSR$	?	
Total (SST)	$n - 1$	S_{yy}		

Is the regression significant at $\alpha = 0.05$? (Compare F_{calc} with $F_{1,8}$).

(e) **Hypothesis Testing (t-test):** Alternatively, test $H_0 : \beta_1 = 0$ using the t-statistic.

$$SE(b_1) = \sqrt{\frac{MSE}{S_{xx}}}, \quad t = \frac{b_1}{SE(b_1)}$$

Check if t^2 is equal to the F value from part (d).

(f) **Prediction & Intervals:** For a new branch spending $x_p = 5.5$:

- Predict the expected sales $\hat{y}$.
- Construct a 95% **Confidence Interval** for the mean sales.
- Construct a 95% **Prediction Interval** for an individual branch's sales.

(g) **Residual Check:** Look at the Residual Plot. Is there any pattern (e.g., U-shape, fanning out) that suggests a violation of assumptions?

15.3 Challenge

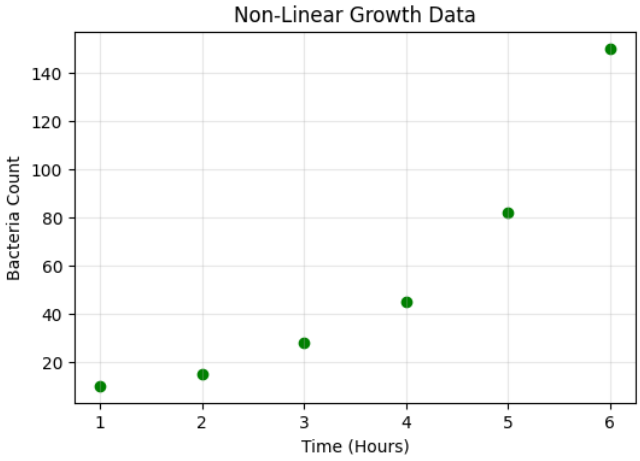
15.4 Exponential Growth & Probability Risk Assessment

A biologist studies bacteria growth. The data follows an exponential trend: $Y = \alpha e^{\beta x} \cdot \epsilon^*$.

Data:

Time (x)	1	2	3	4	5	6
Count (y)	10	15	28	45	82	150

Plot:



Linearized Model: By taking logs, we assume the transformed model is $\ln Y = \beta_0' + \beta_1' x + \epsilon'$, where $\epsilon' \sim N(0, \sigma^2)$.

Tasks:

(a) **Transformation:** Create a table of $(x, \ln y)$ and find the regression equation for $\ln Y$.

(b) **Estimation:** Calculate the Mean Squared Error (MSE or s^2) for the **transformed** data ($\ln y$). This estimates σ^2.

(c) **Probability (Normal):** At **Time** $x = 4$, suppose we want to know the probability that the bacteria count exceeds 50.

- Find the predicted mean of $\ln Y$ at $x = 4$.
- Convert the threshold $Y > 50$ to the log scale: $\ln Y > \ln 50$.
- Calculate $P(\ln Y > \ln 50)$ using the Z-score formula: $Z = \frac{\ln 50 - \widehat{\ln Y}}{\sqrt{MSE}}$.

(d) **Probability (Binomial):** Suppose we examine **10 independent petri dishes** at Time $x = 4$.

- Using the probability p calculated in part (c), what is the probability that **at least 2** dishes have a bacteria count greater than 50?
- Identify the distribution used here (e.g., $B(n, p)$).

15.5 Proof: BLUE Property

Prove that the Least Squares Estimator for the slope b_1 is the **Best Linear Unbiased Estimator (BLUE)**.

(a) Show $E[b_1] = \beta_1$ (Unbiased).

(b) Show $Var(b_1) = \frac{\sigma^2}{S_{xx}}$ and argue it is the minimum variance among linear estimators.

15.4 Applications

15.6 Interpreting Python Output (Statsmodels)

An engineer analyzes the relationship between **Temperature** (X) and **Yield** (Y). Instead of writing code, interpret the **standard output** provided below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.831
No. Observations:                  10   F-statistic:                     45.33
DF Residuals:                       8   Prob (F-statistic):           0.000148
==============================================================================

                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------

const         12.5000      2.100      5.952      0.000      7.657     17.343
Temperature    1.8000      0.267      6.733      0.000      1.184      2.416
==============================================================================

Omnibus:                        0.452   Durbin-Watson:
   2.100
Prob(Omnibus):                  0.798   Jarque-Bera (JB):
   0.521
Skew:                           0.125   Prob(JB):
   0.771
Kurtosis:                       1.915   Cond. No.
   15.4
==============================================================================
```

Questions:

(a) **Model Equation:** Write down the estimated linear equation $\hat{y} = b_0 + b_1 x$.

(b) **Significance:** Is the variable "Temperature" statistically significant at $\alpha = 0.01$? Which value tells you this?

(c) **Confidence Interval:** Interpret the range ‘[1.184, 2.416]‘ associated with Temperature. What does it represent?

(d) **Normality Assumption:** Look at the "Prob(JB)" (Jarque-Bera) value at the bottom. If H_0: Residuals are Normal, do we Reject or Fail to Reject H_0 (given P-value=0.771)? Is the normality assumption valid?

15.7 Interpreting Excel Output

A generic Excel Regression output for Y vs X is shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9200
R Square	0.8464
Standard Error	5.2000
Observations	20

ANOVA

	df	**SS**	**MS**	**F**	**Significance F**
Regression	1	2700.0	2700.0	99.85	0.0000
Residual	18	486.7	27.04		
Total	19	3186.7			

Questions:

(a) **Model Fit:** What percentage of the variance in Y is explained by the model?

(b) **Error Variance:** What is the estimated variance of the error term ($\hat{\sigma}^2$)? (Hint: Look at MS Residual).

(c) **Standard Error:** Verify the relationship: Standard Error = $\sqrt{\text{MS Residual}}$.

(d) **F-Test:** Does the "Significance F" (0.0000) imply that the model is useful or useless?

Chapter 16

Multiple Linear Regression

16.1 Basic Concept

16.1 The "Ceteris Paribus" Concept
The model is given by $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$.

(a) Interpret the meaning of b_1. (Hint: It involves the phrase "holding X_2 constant").

(b) Why is this different from Simple Linear Regression where we just ignore other variables?

16.2 R-squared vs. Adjusted R-squared

(a) In Simple Linear Regression, we look at R^2. In Multiple Regression, why do we prefer Adjusted R^2?

(b) What happens to the standard R^2 if you add a completely useless variable (e.g., "Shoe Size") to a model predicting "House Price"? Does it decrease?

(c) What happens to the Adjusted R^2 in the same scenario?

16.2 Intermediate

16.3 Real Estate Price Prediction (3D Visualization)

A real estate agent predicts House Price (Y) based on Size (X_1) and Age (X_2). The fitted model creates a "Plane" in 3D space.

Visual Inspection:

The Excel Output (Data Analysis > Regression) is provided below:

Regression Statistics

Multiple R	0.95
R Square	0.9025
Adjusted R Square	0.8850
Standard Error	15.2
Observations	30

ANOVA

	df	SS	MS	F	Significance F
Regression	3	60000	20000	86.58	0.0000
Residual	26	6006	231		
Total	29	66006			

Coefficients

	Coefficients	Standard Error	t Stat	P-value
Intercept	50.00	10.00	5.00	0.0000
Size (X_1)	0.15	0.02	7.50	0.0000
Age (X_2)	-2.00	0.50	-4.00	0.0004
Distance (X_3)	-1.50	1.00	-1.50	0.1450

Questions:

(a) **Model Equation:** Write the estimated regression equation.

(b) **Global Test (F-Test):** Test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (The whole model is useless) at $\alpha = 0.05$. Look at "Significance F".

(c) **Individual Tests (t-Test):** Which variables are statistically significant predictors at $\alpha = 0.05$? Identify the variable that is **not** significant.

(d) **Interpretation:** Interpret the coefficient for Age (X_2). What happens to the price for each additional year of age, assuming size and distance remain the same?

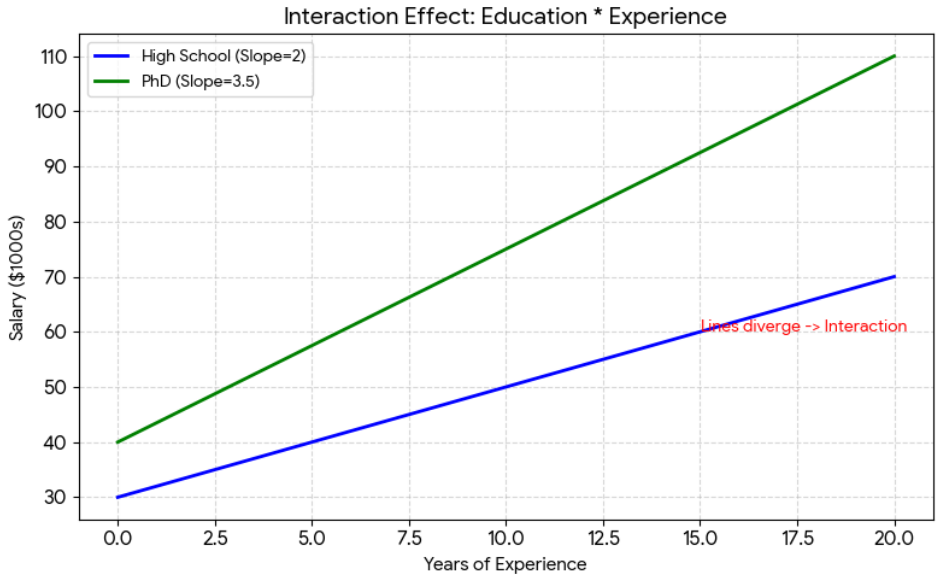
(e) **Prediction:** Predict the price of a house with: Size = 2000 sq.ft., Age = 10 years, Distance = 5 miles.

16.3 Challenge

16.4 Interaction Effects (Visualized)

An HR manager analyzes Salary based on Experience (X_1) and Education Level (X_2: High School vs. PhD). She suspects an Interaction Effect, meaning the value of experience depends on education.

Visual Inspection:



The model with interaction is: $\hat{Y} = 30 + 2X_1 + 10X_2 + 1.5(X_1 \cdot X_2)$. (Where $X_2 = 0$ for High School, $X_2 = 1$ for PhD).

Questions:

(a) **Visual Check:** Look at the plot. Are the two lines parallel? What does the "divergence" (spreading apart) imply about the return on experience for PhDs vs High School grads?

(b) **Slope Calculation:**

- For High School ($X_2 = 0$), the equation becomes $\hat{Y} = 30 + 2X_1$. What is the slope?
- For PhD ($X_2 = 1$), substitute $X_2 = 1$ into the full equation. What is the new slope for X_1?

(c) **Interpretation:** Does "1 year of experience" add the same dollar value to salary for both groups? Explain based on the interaction term coefficient ($+1.5$).

16.5 Multicollinearity (The Hidden Trap)

An engineer predicts "Fuel Consumption" using:

- X_1: Car Weight (kg)
- X_2: Car Weight (lbs)

(a) Intuitively, what is the correlation between X_1 and X_2?

(b) Why is this a problem for the regression calculation (Matrix Inversion)?

(c) How would this affect the P-values? (Can a variable be "important" but have an insignificant P-value?)

Chapter 17

Likelihood Ratio Tests (LRT)

17.1 Intermediate

17.1 Neyman-Pearson for Exponential Distribution
Let $X_1, X_2, \ldots, X_n$ be a random sample from an Exponential distribution with rate parameter θ:

$$f(x;\theta) = \theta e^{-\theta x}, \quad x > 0$$

We want to test the hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

where $\theta_1 > \theta_0$.

Tasks:

(a) **Likelihood Ratio:** Write down the likelihood function $L(\theta)$ and form the ratio $\Lambda = \frac{L(\theta_0)}{L(\theta_1)}$.

(b) **Derivation:** Show that the Neyman-Pearson critical region $\Lambda < k$ reduces to a condition on the sufficient statistic $T = \sum X_i$.

(c) **Result:** Prove that the rejection region is of the form:

$$\sum_{i=1}^{n} X_i < c$$

where c is a constant determined by the significance level α. (Hint: Take the logarithm of Λ and rearrange the inequality, noting that $\theta_0 - \theta_1 < 0$).

17.2 Neyman-Pearson for Custom PDF (Power Function)
Let X be a single observation ($n = 1$) from a population with probability density function:

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0$$

We test:

$$H_0 : \theta = 1 \quad \text{vs} \quad H_1 : \theta = 2$$

Tasks:

(a) **Calculation:** Calculate the Likelihood Ratio $\Lambda(x) = \frac{f(x;1)}{f(x;2)}$.

(b) **Rejection Region:** Show that the Most Powerful (MP) test of size $\alpha = 0.05$ rejects H_0 if:

$$x > \sqrt{0.95}$$

(c) **Power of Test:** Calculate the Power of the test $(1 - \beta)$ using the region derived in (b).

17.3 Neyman-Pearson for Normal Mean (Known Variance)

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$ (where $\mu_1 > \mu_0$).

Task: Prove that the rejection region for the Most Powerful test is equivalent to:

$$\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

(Show the steps of taking $\ln \Lambda$, canceling common terms $\sum x_i^2$, and isolating $\bar{X}$).

17.4 Neyman-Pearson for Uniform Distribution (Support Dependent)

Let $X_1, \ldots, X_n$ be a random sample from a Uniform distribution $U(0, \theta)$.

$$f(x; \theta) = \frac{1}{\theta}, \quad 0 < x < \theta, \quad (0 \text{ otherwise})$$

We want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_1 < \theta_0$.

Tasks:

(a) **Likelihood Function:** Write $L(\theta)$ using the indicator function $I(x_{(n)} < \theta)$, where $x_{(n)}$ is the maximum order statistic.

(b) **Ratio Evaluation:**

- Case 1: If $x_{(n)} > \theta_1$, what is the value of $L(\theta_1)$? What does the ratio Λ become?
- Case 2: If $x_{(n)} \leq \theta_1$, what is the value of the ratio Λ?

(c) **Rejection Region:** Show that the Most Powerful test rejects H_0 if $X_{(n)} \leq C$. Find the constant C such that $P(X_{(n)} \leq C | \theta_0) = \alpha$.

17.5 Neyman-Pearson for Geometric Distribution (Discrete)

Let $X_1, \ldots, X_n$ be a sample from a Geometric distribution ($P(X = k) = (1 - p)^{k-1}p$). Test $H_0 : p = p_0$ vs $H_1 : p = p_1$ (where $p_1 > p_0$).

Tasks:

(a) **Derivation:** Show that the Likelihood Ratio inequality $\frac{L(p_0)}{L(p_1)} < k$ simplifies to a condition on the sample mean $\bar{X}$.

(b) **Direction:** Does the rejection region correspond to "Reject if $\bar{X}$ is large" or "Reject if $\bar{X}$ is small"? (Hint: Consider the sign of $\ln(1-p_0) - \ln(1-p_1)$ given $p_1 > p_0$).

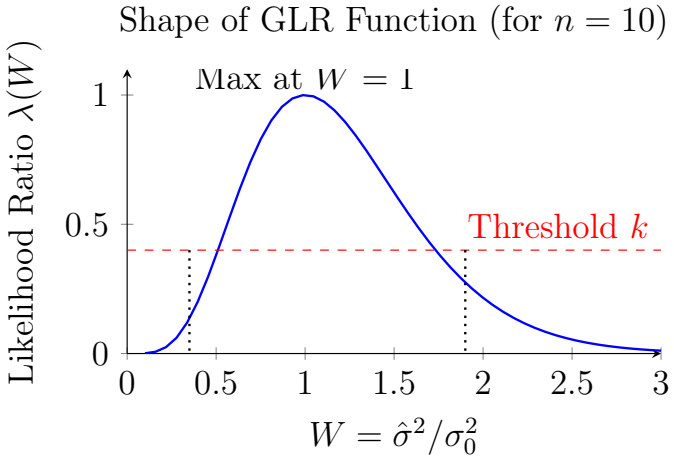
17.2 Challenge

17.6 GLRT for Normal Variance

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. We test $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$.

It can be shown that the Generalized Likelihood Ratio λ depends on the statistic $W = \frac{\hat{\sigma}^2}{\sigma_0^2}$:

$$\lambda(W) = W^{n/2} \exp\left(\frac{n}{2}(1 - W)\right)$$



Tasks:

(a) **MLEs:**

- Show that the unrestricted MLE for variance is $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.
- Explain why the restricted MLE (under H_0) is simply σ_0^2.

(b) Substitute these MLEs into the Normal Likelihood function to derive the expression for $\lambda(W)$ given above.

(c) Look at the graph. The test rejects H_0 when $\lambda(W) < k$ (below the red line). Explain why this leads to a rejection region of the form:

$$\text{Reject if } W < c_1 \text{ or } W > c_2$$

Does this correspond to observing a sample variance that is either "too small" or "too large"?

(d) **(Extremely Hard)** Using the **Neyman–Pearson** framework, consider testing

$$H_0 : \ \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1 : \ \sigma^2 \neq \sigma_0^2.$$

Show that the likelihood ratio for testing fixed alternatives $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$ is a monotone function of the statistic

$$\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Hence, argue that the most powerful test rejects H_0 for extreme values of this statistic. Explain how this leads to a two-sided rejection region and discuss whether this result is consistent with the rejection region obtained from the GLRT.

17.7 GLRT for Poisson Distribution (Count Data)

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Test $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$.

Tasks:

(a) **MLEs:** Identify the restricted MLE $\hat{\lambda}_0$ and the unrestricted MLE $\hat{\lambda}$.

(b) **Lambda Statistic:** Show that the generalized likelihood ratio is:

$$\lambda = \left(\frac{\lambda_0}{\bar{X}}\right)^{n\bar{X}} e^{-n(\lambda_0 - \bar{X})}$$

(c) **Visualization:** Define the function $g(t) = t\ln(1/t) + (t-1)$ where $t = \bar{X}/\lambda_0$. Show that this function implies rejection when $\bar{X}$ is far from λ_0 in either direction.

(d) **Wilks' Approximation:** Using $-2\ln\lambda \sim \chi^2_1$, write the approximate rejection condition.

17.8 GLRT for Rayleigh Distribution (Signal Processing)

The Rayleigh distribution is used to model signal intensity, with PDF:

$$f(x;\sigma) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x > 0$$

We test $H_0 : \sigma^2 = 1$ vs $H_1 : \sigma^2 \neq 1$.

Tasks:

(a) **Unrestricted MLE:** Find the MLE of σ^2 (let's call it $\hat{v}$) for the full parameter space. (Hint: Take log-likelihood, differentiate with respect to $v = \sigma^2$, set to 0).

(b) **Ratio Derivation:** Show that the likelihood ratio λ simplifies to a form involving $\sum x_i^2$.

$$\lambda = \left(\frac{\sum x_i^2}{2n}\right)^n e^{n - \frac{1}{2}\sum x_i^2}$$

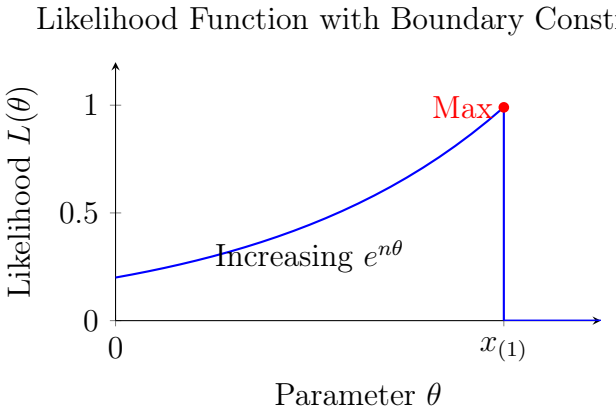
(Note: The exact constants might vary slightly depending on your MLE derivation steps, aim for the structure $A \cdot (\hat{v})^n e^{-n\hat{v}}$).

17.9 GLRT for Shifted Exponential (Boundary MLE)

Consider a sample from a Shifted Exponential distribution with PDF:

$$f(x;\theta) = e^{-(x-\theta)}, \quad x \geq \theta$$

We test $H_0 : \theta = 0$ vs $H_1 : \theta > 0$.



Tasks:

(a) Show that the likelihood function is given by $L(\theta)$:

$$L(\theta) = e^{-\sum x_i + n\theta}$$

(b) Based on the graph or the formula, explain why the unrestricted MLE is $\hat{\theta} = \min(X_1, \ldots, X_n)$ (denoted as $X_{(1)}$). Why can't $\hat{\theta}$ be larger than $X_{(1)}$?

(c) Under $H_0 : \theta = 0$, what is the likelihood value $L(0)$?

(d) Show that the generalized likelihood ratio simplifies to:

$$\lambda = \exp(-n\hat{\theta}) = \exp(-nX_{(1)})$$

(e) If we reject for $\lambda < k$, show that this corresponds to rejecting when $X_{(1)} > c$. Does this make intuitive sense? (i.e., if the smallest data point is far from 0, should we doubt $\theta = 0$?)

Chapter 18

Miscellaneous Problems

18.1 Application

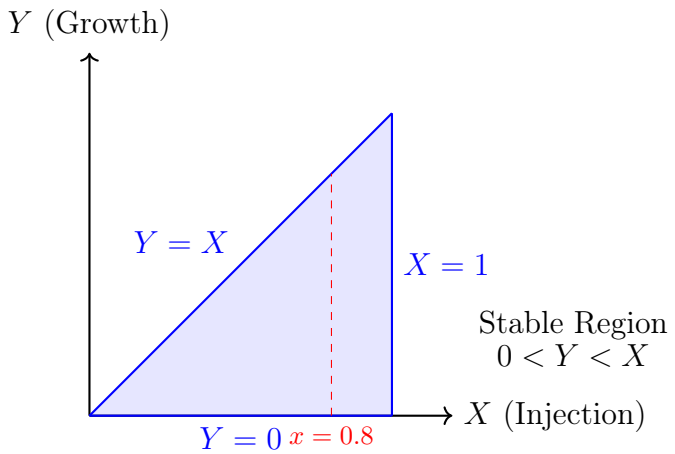
18.1 Thermodynamic Stability of Nano-Coating Process
Scenario Description:
In the advanced manufacturing of semiconductor wafers, the "Chemical Vapor Deposition" (CVD) process is critical. Unit-734, a prototype reactor, controls two highly sensitive variables:

- X: The **Precursor Injection Rate** (normalized unit, $0 < X < 1$).
- Y: The **Thin-Film Growth Rate** (normalized unit).

Due to the complex thermodynamics inside the chamber, the stability of the process is governed by strict physical constraints. Specifically, the Growth Rate (Y) depends on the Injection Rate (X) such that the system is only stable within the parabolic region $0 < Y < X < 1$.

Furthermore, entropy studies suggest that the probability density of the system state, denoted by $f(x, y)$, decays exponentially with injection rate but grows exponentially with growth rate, due to positive feedback loops in the reaction chamber.

The stability domain is visualized by the following:



Part I: Mathematical Modeling
Before any statistical analysis can begin, you must establish the probability model from first principles. The stability domain is visualized above:

(a) **Derivation from ODEs:** Let the probability density function $f(x, y)$ be separable, i.e., $f(x, y) = g(x)h(y)$. The physical laws governing the reactor states are described by the following First-Order Ordinary Differential Equations:

$$\frac{g'(x)}{g(x)} = -2 \quad \text{and} \quad \frac{h'(y)}{h(y)} = 3$$

Solve these **Separable Differential Equations** to show that the joint PDF takes the form:

$$f(x, y) = ke^{-2x+3y}$$

where k is a normalization constant.

(b) **Normalization (Calculus Challenge):** Given the stability domain D shown above, set up the double integral over the triangular region to find k.

$$\int_0^1 \int_0^x ke^{-2x+3y}\, dy\, dx = 1$$

(c) **Marginal Probability:** Find the Marginal PDF of the Injection Rate, $f_X(x)$, and specify its valid range.

(d) **Conditional Probability:** If the system is running at full injection capacity ($X \to 1$), does the probability of having high growth rate ($Y > 0.8$) increase? Calculate $P(Y > 0.8 \mid X = 1)$ using the Conditional PDF $f_{Y|X}(y|x)$.

(e) **Conditional Expectation (Process Optimization):** Engineers need to predict the *average* Growth Rate (Y) for any given Injection Rate setting ($X = x$).

- Derive the analytical expression for the regression function $g(x) = E[Y \mid X = x]$.
- *Hint:* You will need to perform **Integration by Parts** ($\int udv$) using the conditional PDF derived in the previous step.
- **Interpretation:** Based on your function $g(x)$, does the expected growth rate increase linearly with the injection rate?

Part II: Parameter Estimation
A critical defect called a "Micro-Void" (W) occurs during crystal growth. The size of these voids follows a **Triangular Distribution** dependent on an unknown parameter θ (the maximum possible void size):

$$f(w;\theta) = \begin{cases} \frac{2w}{\theta^2} & 0 \le w \le \theta \\ 0 & \text{otherwise} \end{cases}$$

You have a random sample of n void measurements: $W_1, W_2, \ldots, W_n$.

(f) **Method of Moments:**

- Derive the theoretical mean $E[W]$.
- Find the MOM estimator $\hat{\theta}_{MOM}$ in terms of the sample mean $\bar{W}$.

(g) **Maximum Likelihood Estimation:**

- Write down the Likelihood Function $L(\theta)$.
- **Crucial Step:** Explain why taking the derivative $\frac{dL}{d\theta} = 0$ fails to find the maximum in this case.
- By analyzing the domain constraint ($w_i \le \theta$), show that the MLE is the **Maximum Order Statistic**: $\hat{\theta}_{MLE} = W_{(n)} = \max(W_1, \ldots, W_n)$.

(h) **Distribution of the Estimator:**

- Using the CDF method or transformation formula, derive the PDF of the estimator $Y = W_{(n)}$.
- *Hint:* $F_Y(y) = [F_W(y)]^n$.

(i) **Properties of Estimators (Bias & MSE):**

- **Biasedness:** Calculate the expected value $E[\hat{\theta}_{MLE}]$. Is the estimator unbiased? If not, propose a correction factor to make it unbiased.
- **MSE Trade-off:** Write down the expression for the Mean Squared Error (MSE) of $\hat{\theta}_{MLE}$. Discuss conceptually why we might accept a slightly biased estimator if it yields a smaller variance compared to an unbiased one.

(j) **Invariance Principle:**

- The "Structural Integrity Index" is defined as $\eta = \ln(\theta^2 + 1)$.
- Use the Invariance Property of MLE to propose an estimator $\hat{\eta}_{MLE}$.

Part III: Statistical Inference
After establishing the theoretical model, you move to the production floor to validate the machine's performance using real data. The engineering specification states that the **Mean Growth Rate** (μ_Y) must be exactly 0.35 units to ensure chip reliability.

You collect a sample of $n = 25$ production runs.

Data Summary: $\bar{y} = 0.38$, $s_y = 0.08$.

(k) **Hypothesis Testing:** Test the hypothesis that the machine is off-target ($H_0 : \mu_Y = 0.35$ vs $H_1 : \mu_Y \neq 0.35$) at $\alpha = 0.05$.

- Calculate the Test Statistic.
- Find the P-value (approximate from T-table).
- State your engineering recommendation: Should we shut down Unit-734 for calibration?

(l) **Confidence Interval:** Construct a 99% Confidence Interval for the true mean growth rate. Does this interval include the target value of 0.35?

Part IV: Predictive Modeling
Finally, you investigate the relationship between the **Chamber Pressure** (P) and the **Film Thickness** (T). You suspect a linear relationship $T = \beta_0 + \beta_1 P + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

Regression Output ($n = 20$)**:**

$$\hat{T} = 12.5 + 0.45P$$

$$SE(\hat{\beta}_1) = 0.05, \quad MSE = 4.0, \quad R^2 = 0.85$$

(m) **Model Interpretation:** Interpret the coefficient $\beta_1 = 0.45$. What does this number physically represent in the CVD process?

(n) **Slope Significance:** Perform a T-test to determine if Chamber Pressure significantly affects Film Thickness ($H_0 : \beta_1 = 0$).

(o) **Probability of Defect:** A film thickness greater than 30 units is considered a "Defect".

- If the Chamber Pressure is set to $P = 35$, predict the mean thickness $\hat{T}$.
- Assuming the regression assumptions hold (Normal distribution of errors), calculate the **Probability** that a wafer produced at this pressure will be defective ($T > 30$).

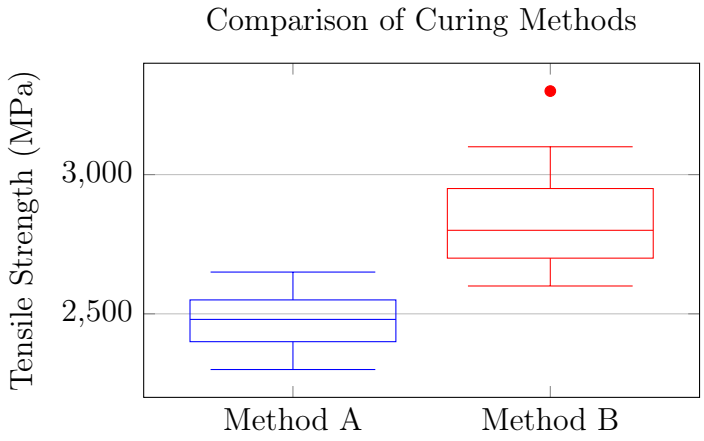
(p) **ANOVA Connection:** Given $R^2 = 0.85$ and $SST_{total} = 1000$, calculate the Regression Sum of Squares (SSR) and the Error Sum of Squares (SSE).

18.2 The Aerospace Composite Wing Project

Scenario: You are the Lead Material Scientist. Your goal is to select the best manufacturing process for a new Carbon Fiber Reinforced Polymer (CFRP). You must compare two Curing Methods, check for defect patterns, optimize parameters, and build a predictive model.

Part I: Comparative Forensics

You collected Tensile Strength data (MPa) from two different Curing Methods: **Method A** and **Method B**.



(a) Compare the two methods based on the Box Plots:

- Which method appears to have a higher median strength?
- Which method has a larger Interquartile Range (IQR)? What does this imply about process consistency?
- Identify the outlier in Method B. What is its approximate value?

Part II: Statistical Inference

Based on the data in Part I, the summary statistics are:

- **Method A:** $n_A = 15, \bar{x}_A = 2480, s_A = 100$
- **Method B:** $n_B = 15, \bar{x}_B = 2800, s_B = 250$

(b) **Assumption Check:** Based on the Box Plots and standard deviations (s_A vs s_B), is it safe to assume equal population variances ($\sigma_A^2 = \sigma_B^2$)? Explain your reasoning.

(c) **Hypothesis Testing:** Test if Method B yields a significantly higher mean strength than Method A ($\mu_B > \mu_A$).

- State the Null and Alternative Hypotheses.
- Calculate the Test Statistic (using unpooled degrees of freedom $\nu \approx 18$ or pooled if you justified it above).
- **Conclusion:** At $\alpha = 0.01$, should we switch to Method B?

Part III: Manufacturing Quality
Failures were tracked across three Mold Types. Is the defect rate independent of the mold used?

Contingency Table:

Result	Mold X	Mold Y	Mold Z	Total
Pass	90	85	65	240
Fail	10	15	35	60
Total	100	100	100	300

(d) **Risk Analysis:**

- Calculate the observed failure rate (%) for each Mold. Which Mold seems most problematic?
- Calculate the Expected Frequency of failures for Mold Z if they were independent.
- Perform a Chi-Square Test (H_0: Independent). Given $\chi^2_{calc} = 14.6$, compare with $\chi^2_{0.05,2} \approx 5.99$. What is your management recommendation?

Part IV: Process Optimization
You perform an ANOVA to test the effect of **Pressure** (Low, Med, High, Ultra) on strength. **Fill in the missing values (A, B, C, D):**

Source	DF	SS	MS	F	P-value
Treatment	3	**A** (=?)	800.0	**C** (=?)	0.004
Error	**B** (=?)	4800.0	**D** (=?)		
Total	27	7200.0			

(e) Reconstruct the table. What is the value of the F-statistic (**C**)?

(f) Does Pressure have a significant effect on strength at $\alpha = 0.01$?

Part V: Predictive Modeling

You study the relationship between **Fiber Density** (x) and **Strength** (y).

Raw Data Table ($n = 10$)**:**

Sample	Density (x)	Strength (y)	Sample	Density (x)	Strength (y)
1	1.5	2300	6	2.1	2900
2	1.6	2450	7	2.2	3100
3	1.8	2550	8	2.3	3150
4	1.9	2700	9	2.4	3300
5	2.0	2800	10	2.5	3400

Visual Diagnostics:

Figure A: Scatter Plot

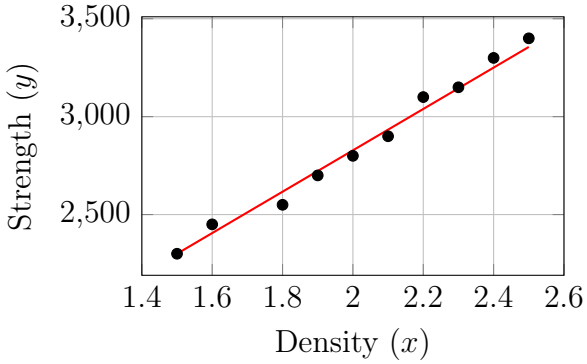
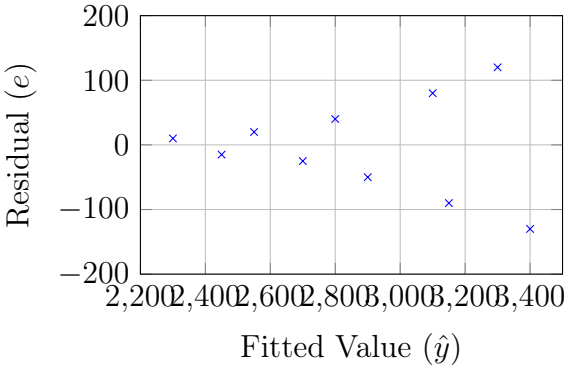


Figure B: Residual Plot



(g) **Model Fitting:** Calculate $\sum x, \sum y, S_{xx}, S_{xy}$ and find the regression equation $\hat{y} = \beta_0 + \beta_1 x$.

(h) **Hypothesis Testing on Slope:** Test whether Density has a statistically significant effect on Strength.

- State the null and alternative hypotheses for β_1.
- Specify the appropriate test statistic and its distribution.
- Based on the data, would you reject H_0 at $\alpha = 0.05$?

(i) **Goodness-of-Fit Assessment:**

- Define the coefficient of determination R^2.
- Explain what a high R^2 value implies in this physical context.
- Can a high R^2 coexist with model inadequacy? Explain briefly.

(j) **Residual Analysis (Critical):** Observe Figure B (Residual Plot).

- Does the variance of residuals remain constant as $\hat{y}$ increases?
- What is the technical term for this violation? (Homoscedasticity or Heteroscedasticity?)
- Suggest a data transformation (e.g., $\ln(y)$) that might fix this issue.

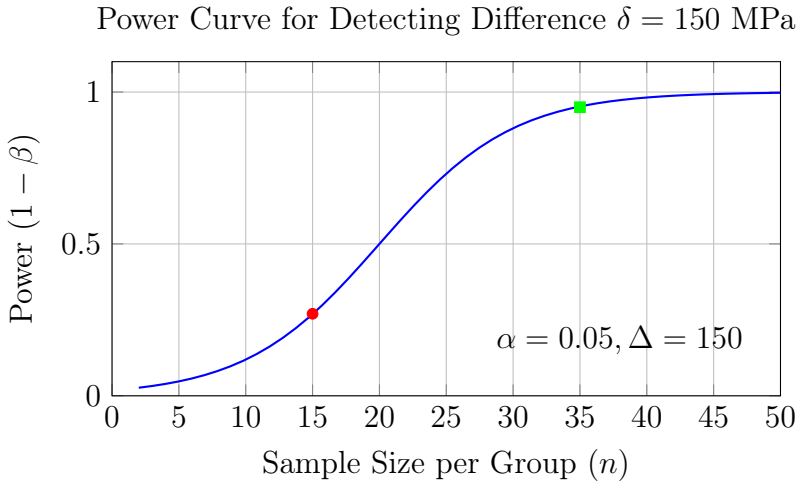
(k) **Probability of Success:** Suppose we target a Density of $x = 2.0$. The model predicts $\hat{y} \approx 2828$. Given $MSE \approx 2500$ ($\hat{\sigma} = 50$), calculate the probability that the actual strength will exceed 2750 MPa.

(l) **Engineering Decision Analysis:** Suppose the minimum acceptable tensile strength is 2800 MPa. Using your regression model and variance estimate,

- Estimate the probability that a randomly produced specimen at $x = 2.0$ meets the specification.
- Based on this probability, would you recommend using $x = 2.0$ as the operating point? Justify your answer statistically.

Part VI: Power Analysis & Sample Size

Before finalizing the report, you must verify if your sample size ($n = 15$ per group in Part I) was sufficient to detect a meaningful difference. You generated a Power Curve for the Two-Sample T-Test ($\alpha = 0.05$, $\sigma = 200$).



(m) **Visual Audit:** Look at the Power Curve above.

- Your current sample size was $n = 15$ (Red Dot). What was the approximate **Power** of your test?
- What is the probability that you committed a **Type II Error** (β) with this sample size? (Recall: Power $= 1 - \beta$).
- **Interpretation:** If the method truly improved strength by 150 MPa, was your experiment likely to detect it?

(n) **Strategic Planning:** Management requires a Power of at least **90% (0.90)** for the next certification phase.

- From the graph, approximately what sample size n is required?
- Use the formula for sample size (one-sided):

$$n \approx \frac{(z_\alpha + z_\beta)^2 (2\sigma^2)}{\delta^2}$$

Given $z_{0.05} = 1.645$, $z_{0.10} = 1.282$, $\sigma = 200$, and $\delta = 150$. Calculate the exact n required.

18.3 The Autonomous Drone Safety System

Context: You have not formally studied Reliability Engineering, but as an engineer, you must be able to understand and apply new technical specifications immediately.

Part I: Theoretical Background

1. Component Reliability: The reliability $R(t)$ is the probability that a component survives beyond time t. For electronic parts, this often follows an **Exponential Distribution**:

$$R(t) = e^{-\lambda t} = e^{-t/MTTF}$$

where λ is the failure rate and $MTTF$ is the Mean Time To Failure.

2. System Configurations:

- **Series System:** Fails if *any* component fails.

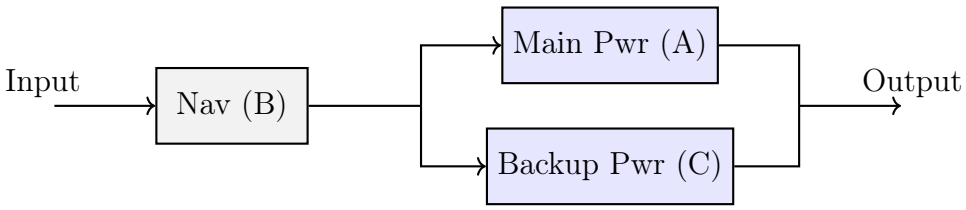
$$R_{series} = R_A \times R_B \times ...$$

- **Parallel System (Redundancy):** Fails only if *all* components fail.

$$R_{parallel} = 1 - (1 - R_A)(1 - R_B)...$$

Part II: The Scenario

You are designing a drone with three subsystems arranged in the Reliability Block Diagram (RBD) below. The drone flies if the **Navigation Module (B)** works **AND** the **Power System** (which consists of Main Power A and Backup Power C in parallel) works.



Specifications (Mission Time $t = 10$ hours):

- **Nav (B):** $MTTF = 100$ hours.
- **Main Pwr (A):** $MTTF = 50$ hours.
- **Backup Pwr (C):** $MTTF = 20$ hours (Low quality backup).

(a) **Component Analysis:** Using the formula $R(t) = e^{-t/MTTF}$, calculate the individual reliability for each component for a 10-hour mission (R_A, R_B, R_C).

(b) **Sub-System Calculation:** Calculate the reliability of the **Power System (Parallel Block)** consisting of A and C.

$$R_{Power} = 1 - (1 - R_A)(1 - R_C)$$

(c) **System Reliability:** Calculate the reliability of the entire drone. Note that the Nav Module is in **Series** with the Power System.

$$R_{System} = R_B \times R_{Power}$$

(d) **Design Improvement:** Your boss wants the system reliability to exceed **0.85**. You have budget to upgrade only ONE component:

- **Option X:** Replace Nav (B) with a better model ($MTTF = 200$).
- **Option Y:** Replace Backup Pwr (C) with a unit identical to Main Pwr A ($MTTF = 50$).

Recalculate R_{System} for both options. Which option achieves the goal?

18.2 Theoretical Concept

18.1 The Statistical Compendium (Conceptual Cloze Test)

Complete the following technical excerpts derived from the engineering handbook. Choose the most precise term from the Word Pools provided.

Theme A: The Mechanics of Modeling

Word Pool A:
Differential Equation, Probability Density Function, Cumulative Distribution Function, Stochastic Process, Deterministic Model, Boundary Condition, Joint Distribution

Unlike a _____ where inputs perfectly determine outputs, real-world engineering systems are inherently random, or a _____. To model the complex interaction between two variables, such as heat and pressure, we often derive the _____ from physical laws expressed as a _____. Solving this equation allows us to understand how probability mass is distributed across the state space.

Theme B: The Art of Estimation

Word Pool B:
Point Estimator, Unbiased, Consistent, Maximum Likelihood, Method of Moments, Mean Squared Error (MSE), Bias-Variance Tradeoff

When parameters are unknown, we calculate a _____ from data. Ideally, this value should be _____, meaning its expected value equals the true parameter. However, sometimes we prefer a biased estimator if it significantly reduces the variance, a concept known as the _____. The _____ method is particularly popular because it finds the parameter values that make the observed data most probable.

Theme C: Statistical Inference & Decision

Word Pool C:
Null Hypothesis, P-value, Type I Error, Type II Error, Power, Confidence Interval, Critical Region, Significance Level

In quality control, rejecting a good batch of products is a costly mistake known as a _____. To avoid this, we set a strict _____ (usually 0.05). If the calculated _____ is smaller than this threshold, we reject the _____. Conversely, the ability of our test to correctly identify a defective batch is called the _____.

Theme D: Regression Diagnostics

Word Pool D:
Residuals, Homoscedasticity, Multicollinearity, Extrapolation, Coefficient of Determination, Normal Probability Plot, Outlier

After fitting a regression line, we must validate assumptions. We examine the _____ to ensure they are randomly distributed. A funnel shape in the residual plot indicates a violation of _____. If we try to predict values far outside our data range, we risk errors from _____. Finally, to check if the error terms are normally distributed, we inspect the _____.

Theme E: Sampling Dynamics & Data Behavior

Word Pool E:
Central Limit Theorem, Standard Error, Skewness, Interquartile Range (IQR), Statistic, Parameter, Law of Large Numbers, Outlier, Robust

In exploratory analysis, the shape of the data tells a story. A distribution with a long tail extending to the right exhibits positive _____. When data contains extreme values, the Mean can be misleading, so engineers prefer the Median and the _____ as they are more _____ measures of center and spread. A data point falling far outside the expected range (typically $1.5 \times IQR$) is flagged as an _____.

Transitioning to inference, we distinguish between a numerical summary of a sample (known as a _____) and the true value of the population (known as a _____). The bridge between them is the _____, which states that the sampling distribution of the mean approaches Normality as sample size increases. The precision of this estimate is quantified by the _____, which decreases as $1/\sqrt{n}$.

18.2 Conceptual Checkpoint (True/False with Correction)

Determine whether the following statements are **True** or **False**. **If a statement is False, provide a brief correction or explanation.**

Part I: Descriptive Statistics

(1) The sample standard deviation S is calculated by dividing the sum of squared deviations by n, not $n-1$.

(2) In a perfectly symmetric distribution, the Mean, Median, and Mode are all equal.

(3) If a dataset is "Right-Skewed" (Positively Skewed), the Mean is typically greater than the Median.

(4) The Interquartile Range (IQR) contains exactly 50% of the data points, regardless of the distribution shape.

(5) Multiplying every data point by a constant c multiplies the Variance by c.

(6) A "Boxplot" can visualize the Mean, Standard Deviation, and Modality (number of peaks) of the data.

(7) The Coefficient of Variation (CV) allows us to compare the variability of two datasets with different units or vastly different means.

(8) If a z-score of an observation is 0, it means the observation is equal to the sample variance.

Part II: Random Variables & Probability Distributions

(9) If two events A and B are Mutually Exclusive ($P(A \cap B) = 0$), they must be Independent.

(10) The expected value of a constant c is $E[c] = c$, and the variance is $Var(c) = c$.

(11) For any two random variables X and Y, $E[X - Y] = E[X] - E[Y]$.

(12) For any two random variables X and Y, $Var(X - Y) = Var(X) - Var(Y)$.

(13) The Poisson Distribution has a unique property where its Mean is equal to its Standard Deviation.

(14) The Cumulative Distribution Function (CDF), $F(x)$, represents the probability $P(X = x)$.

(15) A linear combination of independent Normal random variables is always a Normal random variable.

Part III: Sampling Distributions & CLT

(16) The Central Limit Theorem (CLT) states that the population distribution becomes Normal as the sample size n increases.

(17) The "Standard Error of the Mean" is smaller than the population standard deviation (σ) for any sample size $n > 1$.

(18) If the population is already Normally distributed, the sampling distribution of $\bar{X}$ is Normal even for small sample sizes ($n < 30$).

(19) The Chi-Square distribution (χ^2) is used to model the sampling distribution of the sample variance (S^2) from a Normal population.

(20) As the degrees of freedom increase, the t-distribution approaches the Standard Normal (Z) distribution.

(21) The Law of Large Numbers implies that the sample mean $\bar{X}$ converges to the true mean μ as $n \to \infty$.

(22) The ratio of two independent Chi-Square variables divided by their degrees of freedom follows a t-distribution.

Part IV: Estimation (Point & Interval)

(23) An estimator $\hat{\theta}$ is "Unbiased" if its value in a single sample equals the true parameter θ.

(24) The Mean Squared Error (MSE) of an estimator is defined as $Variance + (Bias)^2$.

(25) A "Consistent" estimator gets closer to the true parameter value (in probability) as the sample size n increases.

(26) Maximum Likelihood Estimators (MLE) are always Unbiased.

(27) The Method of Moments (MOM) always yields the same estimator as the Method of Maximum Likelihood (MLE).

(28) A 95% Confidence Interval for μ means there is a 95% probability that the true parameter μ falls within the calculated interval.

(29) Increasing the sample size n (while keeping confidence level constant) will result in a wider confidence interval.

(30) Increasing the Confidence Level (e.g., from 90% to 99%) will result in a wider confidence interval.

(31) The "Margin of Error" of an interval estimate depends only on the sample standard deviation and the sample size.

(32) An "Efficient" estimator is the unbiased estimator that has the minimum possible variance.

Part V: Hypothesis Testing

(33) A Type I Error (α) occurs when we reject a Null Hypothesis that is actually True.

(34) A P-value of 0.04 means there is a 96% chance that the Alternative Hypothesis is true.

(35) The "Power" of a statistical test is the probability of failing to reject a true Null Hypothesis.

(36) Statistical Significance (low P-value) does not necessarily imply Practical Significance (large effect size).

(37) Failing to reject H_0 is mathematically equivalent to proving that H_0 is true.

(38) If we perform a hypothesis test at $\alpha = 0.05$ and the P-value is 0.06, we should reject H_0.

(39) A two-sided hypothesis test (e.g., $\mu \neq \mu_0$) rejects H_0 only if the test statistic falls in the upper tail.

Part VI: Linear Regression & ANOVA

(40) In Simple Linear Regression, the "Least Squares" method minimizes the sum of squared vertical distances (residuals) between observed points and the line.

(41) The residuals ϵ_i are assumed to be independent, normally distributed, and have constant variance (Homoscedasticity).

(42) If the Coefficient of Determination $R^2 = 1.0$, it implies that X causes Y.

(43) A negative slope ($\beta_1 < 0$) implies that the correlation coefficient r must be negative.

(44) In Simple Linear Regression ($Y = \beta_0 + \beta_1 X + \epsilon$), the F-test for the overall model yields the same p-value as the t-test for the slope β_1.

(45) Extrapolation (predicting Y outside the range of observed X) is safe as long as the R^2 is high.

(46) The sum of the residuals $\sum(y_i - \hat{y}_i)$ in a least-squares regression model is always zero.

(47) In One-Way ANOVA, the Null Hypothesis is that the sample means of all groups are equal.

(48) In ANOVA, the identity $SST = SSR + SSE$ (Total = Regression/Treatment + Error) always holds.

(49) If the F-statistic in ANOVA is close to 1, it suggests that there is a significant difference between the group means.

(50) "Multicollinearity" occurs when the dependent variable Y is highly correlated with the independent variable X.

Appendix A

Useful Formulas

1. Probability Foundations

Counting & Axioms

- **Permutations:** $P_{n,k} = \frac{n!}{(n-k)!}$
- **Combinations:** $\binom{n}{k} = C_{n,k} = \frac{n!}{k!(n-k)!}$
- **Addition Rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Conditional Probability:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Theorems

- **Total Probability Rule:** $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$
- **Bayes' Theorem:**

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

2. Random Variables & Expectations

Definitions

Concept	**Discrete** (Sum)	**Continuous** (Integral)
PMF / PDF	$P(X = x) = p(x)$	$P(a \leq X \leq b) = \int_a^b f(x)dx$
CDF $F(x)$	$\sum_{t \leq x} p(t)$	$\int_{-\infty}^{x} f(t)dt$
Expectation μ	$\sum xp(x)$	$\int_{-\infty}^{\infty} xf(x)dx$
Variance σ^2	$\sum (x-\mu)^2 p(x)$	$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

Key Properties

- $Var(X) = E[X^2] - (E[X])^2$

- $E[aX + b] = aE[X] + b$
- $Var(aX + b) = a^2 Var(X)$
- **Moment Generating Function (MGF):** $M_X(t) = E[e^{tX}]$
- **MGF Property:** $E[X^k] = M_X^{(k)}(0)$ (k-th derivative at $t = 0$)

Joint Distributions Independence

- **Covariance:** $Cov(X, Y) = E[XY] - E[X]E[Y]$
- **Correlation:** $\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \quad -1 \leq \rho \leq 1$
- **Variance of Linear Combination:**

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$$

- **Independence:** If $X \perp Y$, then $Cov(X, Y) = 0$ and $E[XY] = E[X]E[Y]$.

Transformation of Variables

- **One-to-One Function ($Y = g(X)$):**

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

- **CDF Technique:** $F_Y(y) = P(g(X) \leq y)$, then $f_Y(y) = F'_Y(y)$.

3. Common Probability Distributions

Discrete Distributions

Distribution	**PMF** $P(X = x)$	**Mean**	**Variance**
Bernoulli(p)	$p^x(1-p)^{1-x}, x \in \{0, 1\}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
Geometric(p)	$(1-p)^{x-1}p, x = 1, 2, ...$	$1/p$	$(1-p)/p^2$
Neg. Binomial(r, p)	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	r/p	$r(1-p)/p^2$
Hypergeometric	$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$	$n\frac{K}{N}$	$n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ

Continuous Distributions

Distribution	**PDF** $f(x)$	**Mean**	**Variance**

Uniform(a, b)	$\frac{1}{b-a}, a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(λ)	$\lambda e^{-\lambda x}, x > 0$	$1/\lambda$	$1/\lambda^2$
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Gamma(α, β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$
Chi-Square(ν)	(Gamma $\alpha = \nu/2, \beta = 2$)	ν	2ν
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

4. Statistical Inference Formulas

Point Estimation Properties

- **Likelihood Function:**
$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
- **Bias:** $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$
- **Mean Squared Error (MSE):** $MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$

One-Sample Hypothesis Tests

Parameter	**Assumption**	**Test Statistic (T_{calc})**	**Distribution**
Mean μ	σ Known	$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
Mean μ	σ Unknown	$T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t_{n-1}
Proportion p	Large n	$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$N(0,1)$
Variance σ^2	Normal Pop.	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	χ^2_{n-1}

Two-Sample Hypothesis Tests

Scenario	**Test Statistic**	**df / Notes**
Diff Means ($\sigma_{1,2}$ known)	$Z = \frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$N(0,1)$
Diff Means (Equal Var) (s_p^2 Pooled Var)	$T = \frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$df = n_1 + n_2 - 2$
Diff Means (Unequal Var)	$T = \frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	Welch-Satterthwaite eq.
Paired Difference (d_i)	$T = \frac{d-\delta_0}{s_d/\sqrt{n}}$	$df = n - 1$
Diff Proportions	$Z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ (Pooled)
Ratio of Variances	$F = \frac{s_1^2}{s_2^2}$	$df_1 = n_1 - 1, df_2 = n_2 - 1$

Confidence Intervals (General Form)

$$\text{Point Estimate} \pm (\text{Critical Value}) \times (\text{Standard Error})$$

- **Mean (σ unknown):** $\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$
- **Proportion:** $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Variance:** $\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$

5. Linear Regression & ANOVA

Simple Linear Regression ($y = \beta_0 + \beta_1 x + \epsilon$)

- $S_{xx} = \sum(x_i - \bar{x})^2, \quad S_{yy} = \sum(y_i - \bar{y})^2, \quad S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$
- **Slope:** $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ **Intercept:** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- **Sums of Squares:** $SST = S_{yy}, \quad SSR = \hat{\beta}_1 S_{xy}, \quad SSE = SST - SSR$
- **Coefficient of Determination:** $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

ANOVA Table (One-Way)

Source	DF	Sum of Squares (SS)	Mean Square (MS)	F
Treatment	$k-1$	$SSTr$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
Error	$N-k$	SSE	$MSE = \frac{SSE}{N-k}$	
Total	$N-1$	SST		

Chi-Square Goodness of Fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \quad df = k - 1 - (\#\text{est. params})$$

Likelihood Ratio Test (LRT)

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)}$$

Wilks' Theorem: As $n \to \infty$, $-2\ln\lambda \sim \chi^2_\nu$, where $\nu = \dim(\Theta) - \dim(\Theta_0)$.