STATISTICS FOR DATA SCIENCE WITH EXCEL

Exercises & Solutions

Tina Moré

Chapter 1 Practical Exercises and Challenges

Exercises and Challenges

These exercises are designed to strengthen your statistical thinking. Some require short written answers. Others ask you to reflect critically on real-world scenarios.

Take your time. Think carefully. Statistics begins with reasoning.

Exercise 1: Data Is Not Self-Explanatory

A company reports:

“After our website redesign, sales increased by 12%.”

List at least five statistical questions you would ask before accepting this conclusion.

Exercise 2: Descriptive or Inferential?

For each of the following scenarios, identify whether it is an example of descriptive statistics or inferential statistics. Briefly explain why.

1. Calculating the average monthly revenue for the past year.
2. Surveying 800 voters to predict a national election outcome.
3. Creating a histogram of customer ages.
4. Estimating the average income of all households in a city using a sample of 500 households.
5. Reporting the percentage of defective items in last week’s production batch.

Exercise 3: Thinking in Distributions

A report states:

“The average salary in Company X is $60,000.”

Explain why this statement may be incomplete or potentially misleading.
What additional statistical information would you request?

Exercise 4: Everyday Probability

The weather forecast predicts a 70% chance of rain tomorrow.

Explain:

1. What does “70% chance” actually mean?
2. What does it NOT mean?
3. How would a statistical thinker interpret this forecast when deciding whether to carry an umbrella?

Exercise 5: Model Evaluation Awareness

A machine learning model predicts customer churn with 95% accuracy.

What questions should you ask before concluding that this model is excellent?

Consider:

- Data imbalance
- Sample size
- Overfitting
- Real-world cost of errors

Challenge Exercise: The Meeting Scenario

You are in a meeting. A manager presents a chart showing that productivity increased sharply after a new policy was introduced.

Everyone agrees the policy caused the improvement.

As a statistical thinker:

1. What alternative explanations might you consider?
2. What additional data would you request?
3. What statistical methods might help clarify whether the policy truly caused the increase?

Write a structured response (at least one paragraph).

Solutions and Explanations

Use these explanations to check your reasoning. Do not worry if your answers differ slightly—focus on whether your thinking aligns with statistical principles.

Solution 1: Data Is Not Self-Explanatory

Strong statistical questions might include:

- How was "sales increase" measured? Revenue? Units sold? Profit?
- What time periods were compared?
- Was the comparison seasonally adjusted?
- Was the sample size sufficient?
- Were there other changes occurring simultaneously (marketing campaigns, pricing adjustments)?
- What was the variation in sales before and after?
- Is the increase statistically significant or within normal fluctuation?

This exercise reinforces: data does not interpret itself. Context and structure matter.

Solution 2: Descriptive or Inferential?

1. Average monthly revenue → Descriptive (summarizing existing data).
2. Surveying 800 voters → Inferential (using a sample to predict a larger population).
3. Histogram of ages → Descriptive (visual summary).
4. Estimating city-wide income from 500 households → Inferential (generalizing from sample).
5. Percentage of defective items last week → Descriptive (describing observed production).

Key distinction:
Descriptive = summarizes what is observed.
Inferential = extends conclusions beyond observed data.

Solution 3: Thinking in Distributions

The average alone may hide important details:

- Is income evenly distributed?
- Are a few executives earning very high salaries skewing the mean?
- What is the median salary?
- What is the salary range?
- What is the standard deviation?

Statistical thinkers do not rely on a single summary number. They ask about the distribution.

Solution 4: Everyday Probability

1. "70% chance of rain" typically means that, given similar atmospheric conditions in the past, rain occurred 70% of the time.
2. It does NOT mean it will rain for 70% of the day, nor that 70% of the area will receive rain.
3. A statistical thinker weighs probability against consequences. If getting wet is inconvenient, bringing an umbrella is rational.

This illustrates decision-making under uncertainty.

Solution 5: Model Evaluation Awareness

Important questions include:

- Is the dataset balanced? (If 95% of customers do not churn, predicting "no churn" always would achieve 95% accuracy.)
- How large was the training dataset?
- Was performance measured on unseen test data?

Accuracy alone rarely tells the full story.

Solution to Challenge Exercise

A statistical thinker would consider:

- Could productivity have increased due to seasonal effects?
- Were other operational changes introduced simultaneously?
- Is the increase consistent across departments?
- What was productivity variability before and after?

- Is the sample size sufficient?

Additional data might include:

- Control group comparisons
- Time-series data over longer periods
- Measures of variability
- External influencing factors

Statistical methods that could help:

- Hypothesis testing
- Difference-in-differences analysis
- Regression analysis controlling for confounders

This challenge reinforces the difference between association and causation.

Chapter 2 Practical Exercises and Challenges

Exercises and Challenges

Exercise 1: Identifying Variables and Observations

You are given the following dataset description:

A retail company records the following fields for each transaction: Transaction_ID, Customer_ID, Purchase_Date, Product_Category, Quantity, Unit_Price, Payment_Method.

Transaction_ID	Customer_ID	Purchase_Date	Product_Category	Quantity	Unit_Price	Payment_Method
TXN-1001	CUST-5822	2026-03-15	Electronics	1	450.00	Credit Card
TXN-1002	CUST-1290	2026-03-15	Home & Kitchen	3	25.50	PayPal
TXN-1003	CUST-3341	2026-03-16	Apparel	2	15.00	Cash
TXN-1004	CUST-5822	2026-03-16	Electronics	1	12.00	Credit Card
TXN-1005	CUST-9012	2026-03-17	Beauty	5	8.99	Mobile Wallet

Tasks:

1. Identify the observations.
2. Identify the variables.
3. Classify each variable as categorical or numerical.
4. Indicate which variables are identifiers and should not be used in calculations.

Exercise 2: Classifying Data Types

Classify the following variables:

1. Blood type (A, B, AB, O)
2. Temperature in °C
3. Number of children in a household
4. Customer satisfaction rating (1–5 scale)
5. Email address
6. Monthly income

For each variable, state:

- Categorical or numerical
- If numerical: discrete or continuous
- If categorical: nominal or ordinal

Exercise 3: Evaluating Data Sources

For each scenario below, identify:

- The data source
- One potential strength
- One possible limitation

a) Data collected from a national census

b) Social media sentiment scraped from posts

c) Sensor data from wearable fitness trackers

d) Sales data from a company's internal database

Exercise 4: Assessing Data Quality

A dataset contains customer records. During inspection, you notice:

- 15% of phone numbers are missing
- Some customers have negative ages

- The same customer appears multiple times with slightly different spellings
- Dates are recorded in multiple formats

Identify which data quality dimension is affected in each case:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Validity
- Uniqueness

Exercise 5: Ethical Evaluation

You are given a dataset collected for academic research on student performance. A company later requests access to use the data for targeted advertising.

1. What ethical concern arises?
2. Would consent need to be re-obtained? Why?
3. What lifecycle stage is involved in this decision?

Exercise 6: Mapping the Data Lifecycle

Match the following actions to the correct lifecycle stage:

a) Encrypting a database
b) Removing outdated records according to policy
c) Training a predictive model
d) Designing survey questions
e) Saving older datasets in long-term cloud storage
f) Cleaning missing values

Challenge 1: First Encounter Checklist

You receive a dataset with 200,000 rows and 15 columns from a colleague.

Write a structured plan describing the first five steps you would take before running any statistical analysis.

Challenge 2: Detecting Hidden Risk

A company reports impressive predictive model accuracy (98%). However:

- The training data came from a single region
- The dataset is five years old
- Documentation is incomplete
- No bias testing was conducted

Identify at least four risks and explain why they matter.

Solutions and Explanations

Solution 1: Variables and Observations

1. Observations: Each transaction (one row per purchase).
2. Variables**:**
    - Transaction_ID
    - Customer_ID
    - Purchase_Date
    - Product_Category
    - Quantity
    - Unit_Price
    - Payment_Method
3. Classification:

Variable	**Type**
Transaction_ID	Categorical (identifier)
Customer_ID	Categorical (identifier)
Purchase_Date	Date (treated as temporal)
Product_Category	Categorical (nominal)
Quantity	Numerical (discrete)
Unit_Price	Numerical (continuous)
Payment_Method	Categorical (nominal)

4. Identifiers:

Transaction_ID and Customer_ID should not be averaged or summed.

Identifiers label; they do not measure.

Solution 2: Data Type Classification

Variable	Category	Subtype
Blood type	Categorical	Nominal
Temperature	Numerical	Continuous
Number of children	Numerical	Discrete
Satisfaction rating	Categorical	Ordinal
Email address	Categorical	Nominal (identifier)
Monthly income	Numerical	Continuous

Key insight: The nature of a variable determines appropriate statistical treatment.

Solution 3: Data Sources

a) Census

- Source: Government survey
- Strength: Large coverage
- Limitation: Infrequent updates

b) Social media

- Source: Digital platform scraping
- Strength: Real-time sentiment
- Limitation: Sampling bias

c) Wearable sensors

- Source: IoT devices
- Strength: Continuous measurement
- Limitation: Device calibration errors

d) Internal sales database

- Source: Transactional system
- Strength: Accurate financial records
- Limitation: Limited to company customers

Understanding source clarifies reliability and bias.

Solution 4: Data Quality Dimensions

Issue	Dimension Affected

Missing phone numbers	Completeness
Negative ages	Validity
Duplicate customers	Uniqueness
Multiple date formats	Consistency

Each issue must be corrected before analysis.

Solution 5: Ethical Evaluation

1. Ethical concern: Purpose limitation violation.
2. Yes, consent must likely be re-obtained because data is being used for a new purpose.
3. Lifecycle stage: Sharing/Publication and Governance decision.

Ethical data use requires alignment with original intent.

Solution 6: Lifecycle Mapping

Action	**Stage**
Encrypting database	Storage
Removing outdated records	Deletion
Training predictive model	Analysis
Designing survey questions	Planning/Design
Saving older datasets	Archival
Cleaning missing values	Processing/Preparation

Challenge 1: First Encounter Plan (Model Answer)

1. Revisit analytical objective
2. Preview dataset structure
3. Review metadata
4. Verify data types and structure
5. Conduct basic integrity checks

This structured approach prevents premature modeling.

Challenge 2: Risk Identification

1. Sampling bias – Single region limits generalizability.
2. Timeliness issue – Data is outdated.

3. Lack of transparency – Incomplete documentation reduces reproducibility.
4. Fairness risk – No bias testing may lead to discriminatory outcomes.

High accuracy does not guarantee validity or fairness.

Chapter 3 Practical Exercises and Challenges

Exercises and Challenges

These exercises are designed to test both technical skill and analytical thinking. Do not rush to compute—first decide which method is appropriate and why.

Exercise 1: Coffee Shop Orders (Categorical Data)

A coffee shop records the preferred drink of 50 customers during a morning shift.

Drink preferences:

Latte (18), Espresso (12), Cappuccino (10), Tea (6), Mocha (4)

Tasks

1. Construct a frequency table.
2. Compute relative frequencies and percentages.
3. Identify the most and least popular drinks.
4. Write a short managerial interpretation (3–4 sentences).

Exercise 2: Household Internet Devices (Discrete Numerical Data)

A survey of 120 households records the number of internet-connected devices in each home:

Devices	Frequency
0	6
1	18
2	36
3	30
4	20
5 or more	10

Tasks

1. Compute relative frequencies.
2. Compute cumulative percentages.
3. What percentage of households have 3 or fewer devices?
4. Interpret the distribution pattern.

Exercise 3: Employee Monthly Salaries (Continuous Data)

A company groups 150 employee salaries into the following intervals:

Salary Range ($)	Frequency
30,000–39,999	12
40,000–49,999	28
50,000–59,999	42
60,000–69,999	36
70,000–79,999	20
80,000+	12

Tasks

1. Compute relative frequencies.
2. Compute cumulative percentages.
3. What percentage earn below $60,000?
4. Where is the concentration of salaries?

Exercise 4: Product Preference by Gender (Two-Way Table)

A company surveys 200 customers:

	Product A	Product B	Product C	Total
Male	30	40	20	90
Female	50	35	25	110
Total	80	75	45	200

Tasks

1. Compute column percentages.
2. Compute row percentages.
3. Which product is most popular within each gender?
4. Is there evidence of preference differences?

Exercise 5: Sales Analysis Using Pivot Logic

Use the sales dataset provided in this chapter's Excel workbook to practically complete the exercise.

You have 300 sales transactions with:

- Region (North, South, East, West)
- Sales Amount
- Salesperson

Tasks

1. Which pivot table structure would you use to find total sales by region?
2. Which structure would you use to find average sales by salesperson?
3. Why might averages be more informative than totals in some cases?
4. If one region has fewer transactions but higher average sales, what does that suggest?

Solutions and Explanations

Solution 1: Coffee Shop Orders

Frequency Table

Drink	Frequency	Relative Freq	Percentage
Latte	18	0.36	36%
Espresso	12	0.24	24%
Cappuccino	10	0.20	20%
Tea	6	0.12	12%
Mocha	4	0.08	8%
Total	50	1.00	100%

Interpretation:

Latte is clearly the dominant product (36%), while Mocha is the least popular (8%). Inventory and promotional strategy should prioritize high-demand items. However, niche products may still serve specific customer segments.

Solution 2: Household Devices

Relative Frequencies: Divide each frequency by 120.

Devices	Relative Freq	Percentage	Cumulative %
0	0.05	5%	5%
1	0.15	15%	20%

2	0.30	30%	50%
3	0.25	25%	75%
4	0.17	17%	92%
5+	0.08	8%	100%

Interpretation:

75% of households have 3 or fewer devices. The distribution centers around 2–3 devices, suggesting moderate digital penetration.

Solution 3: Salary Distribution

Relative Frequencies

Salary Range	Relative Freq	Percentage	Cumulative %
30–39k	0.08	8%	8%
40–49k	0.19	19%	27%
50–59k	0.28	28%	55%
60–69k	0.24	24%	79%
70–79k	0.13	13%	92%
80k+	0.08	8%	100%

Key Findings

- 55% earn below $60,000.
- The largest concentration (28%) is in the $50k–$59k range.
- Distribution appears centered around mid-level salaries.

This suggests a workforce concentrated in middle income bands.

Solution 4: Product Preference by Gender

Column Percentages (Within Product)

Example for Product A:

- Male: 30/80 = 37.5%
- Female: 50/80 = 62.5%

Compute similarly:

	Product A	Product B	Product C
Male	37.5%	53.3%	44.4%
Female	62.5%	46.7%	55.6%

Row Percentages (Within Gender)

- Male total = 90

- Female total = 110

Example:

- Male choosing Product A: 30/90 = 33.3%

	Product A	**Product B**	**Product C**
Male	33.3%	44.4%	22.2%
Female	45.5%	31.8%	22.7%

Interpretation:

- Males prefer Product B most (44.4%).
- Females prefer Product A most (45.5%).
- There are visible preference differences between genders.

This suggests targeted marketing opportunities.

Solution 5: Pivot Logic

1. Total Sales by Region
    - Rows → Region
    - Values → Sum of Sales Amount

2. Average Sales by Salesperson
    - Rows → Salesperson
    - Values → Average of Sales Amount

3. Why Averages Matter
   Totals reflect volume. Averages reflect performance efficiency. A salesperson with fewer transactions but higher average value may be strategically valuable.
4. Interpretation of Fewer but Higher Sales
   This suggests higher-value clients or premium sales strategies in that region.

Chapter 4 Practical Exercises and Challenges

Exercises and Challenges

The following exercises are designed to strengthen conceptual understanding and technical fluency. Use Excel where appropriate.

Exercise 1: Understanding the Center

A small business records the number of daily customer complaints over 12 days:

2, 3, 4, 3, 5, 2, 4, 3, 20, 3, 4, 2

1. Calculate the mean, median, and mode.
2. Which measure best represents the “typical” number of complaints?
3. Explain your reasoning.

Exercise 2: Investigating Spread

Using the same dataset from Exercise 1, use Excel functions to:

1. Calculate the range.
2. Calculate the interquartile range (IQR).
3. Calculate the sample standard deviation.
4. What does the variability tell you about complaint consistency?

Exercise 3: The Outlier Experiment

Remove the value 20 from the dataset.

1. Recalculate the mean and sample standard deviation.
2. Compare the new results with the original ones.
3. Which measures changed the most? Why?

Exercise 4: Population or Sample?

For each scenario below, decide whether you should use population or sample formulas. Justify your answer.

a) All 45 employees in a company complete a productivity test.
b) 180 voters are surveyed from a city of 210,000 residents.
c) You analyze the revenue of all 12 months in 2025.
d) You select 100 transactions from a database of 40,000.

Exercise 5: Comparing Relative Variability

Two departments report average monthly sales:

- Department A: Mean = 50,000; Standard Deviation = 5,000
- Department B: Mean = 10,000; Standard Deviation = 3,000

1. Calculate the coefficient of variation (CV) for both.
2. Which department has greater relative variability?
3. Interpret the result.

Challenge 1: Designing Your Own Dataset

Create two different datasets of 10 numbers each such that:

- Both datasets have the same mean.
- One dataset has a much larger standard deviation.

Explain how you constructed them and why their spreads differ.

Challenge 2: Symmetry vs Skewness

Construct a dataset where:

- The mean is significantly higher than the median.
- The IQR remains small.

Explain what this implies about the distribution.

Solutions and Explanations

Solution 1: Understanding the Center

Dataset:
2, 3, 4, 3, 5, 2, 4, 3, 20, 3, 4, 2

Excel Formulas:

Mean: =AVERAGE(range) → 4.58

Median: =MEDIAN(range) → 3

Mode: =MODE.SNGL(range) → 3

Interpretation:

The mean (4.58) is pulled upward by the extreme value 20.
The median and mode both equal 3, indicating that most days have about 3 complaints.

Best measure: The median, because the dataset contains an outlier.

Solution 2: Investigating Spread

Range: 20 − 2 = 18

Quartiles: =QUARTILE.INC(array, quart):

Q1 = 2.75

Q3 = 4

IQR = 1.25

Sample Standard Deviation:

=STDEV.S(range) → 4.94

Interpretation:

- The large range (18) reflects the extreme value.
- The IQR (1.25) shows that the middle 50% of days are tightly clustered.
- The high standard deviation confirms substantial overall variability due to the outlier.

Conclusion: Most days are stable, but one extreme day distorts the overall variability.

Solution 3: The Outlier Experiment

New dataset (without 20):

2, 3, 4, 3, 5, 2, 4, 3, 3, 4, 2

New Mean → 3.18
New Sample Standard Deviation → 0.98

Comparison

Measure	With Outlier	Without Outlier
Mean	4.58	3.18
Std Dev	4.99	0.98

Explanation:

The standard deviation changed dramatically, demonstrating its sensitivity to extreme values.

The median would barely change.

This confirms a key principle:
Outliers disproportionately affect mean and standard deviation.

Solution 4: Population or Sample?

a) All 45 employees → Population formula (complete group).

b) 180 voters from 210,000 → Sample formula (subset estimating larger population).

c) All 12 months in 2025 → Population formula (complete period).

d) 100 transactions from 40,000 → Sample formula (subset).

Rule applied:
Use N when describing everything.
Use n − 1 when estimating something larger.

Solution 5: Comparing Relative Variability

Department A:

CV = (5,000 / 50,000) × 100 = 10%

Department B:

CV = (3,000 / 10,000) × 100 = 30%

Interpretation:

Department B has greater relative variability.

Although its standard deviation is smaller in absolute terms, its fluctuations represent a much larger proportion of its average sales.

This demonstrates why the coefficient of variation is essential when comparing datasets with different scales.

Challenge 1: Example Construction

Dataset A:

10, 10, 10, 10, 10, 10, 10, 10, 10, 10

Mean = 10
Std Dev = 0

Dataset B:

1, 1, 1, 1, 1, 19, 19, 19, 19, 19

Mean = 10
Std Dev = High

Explanation:
Both datasets average to 10, but Dataset B spreads values far from the mean.

Challenge 2: Example Construction

Dataset:

5, 5, 5, 5, 5, 5, 5, 5, 5, 20

Mean ≈ 6.5
Median = 5

IQR = 0

Explanation:
Most values cluster tightly at 5 (small IQR), but one large outlier raises the mean substantially.

Chapter 5 Practical Exercises and Challenges

Exercises and Challenges

Use the Chapter 5 Excel dataset unless otherwise stated.

Part A: Skill-Building Exercises

Exercise 1: Comparing Central Tendencies

Using the household_income dataset:

1. Calculate the mean, median, and mode.
2. Arrange them from smallest to largest.
3. Based on their relationship, predict whether the distribution is left-skewed, right-skewed, or symmetric — before calculating skewness.

Exercise 2: Measuring Skewness

1. Use Excel's =SKEW() function to compute skewness.
2. Interpret the result:
    - Is the skew weak, moderate, or strong?
    - What does the direction tell you about high or low income values?

Exercise 3: Measuring Kurtosis

1. Use =KURT() in Excel.
2. State whether the distribution has:
    - Heavy tails
    - Light tails
    - Normal-like tails
3. Explain in plain language what this means for extreme income values.

Exercise 4: Histogram Analysis

Create a histogram using appropriate income bins.

Answer:

1. Where does the majority of households cluster?
2. Does the visual confirm your skewness calculation?
3. Is the tail longer on the left or the right?

Exercise 5: Mean vs Median Decision

Imagine you are preparing a public economic report.

1. Which measure would you report as "typical household income"?
2. Justify your answer using statistical reasoning.

Part B: Analytical Challenges

These require deeper reasoning.

Challenge 1: Outlier Removal

1. Remove the top 5% highest incomes from the dataset.
2. Recalculate:
    - Mean
    - Median
    - Skewness
3. Compare results with the original dataset.

Question: Which measure changes the most? Why?

Challenge 2: Income Cap Simulation

Create a new column where incomes above $250,000 are capped at $250,000.

1. Recalculate skewness and kurtosis.
2. Compare with the original values.

Question: What does this tell you about the sensitivity of skewness and kurtosis to extreme values?

Challenge 3: Designing a Fair Policy

Suppose a government wants to tax the top 10% of earners.

1. Use percentiles to determine the income cutoff.
2. Explain why percentiles are more appropriate than using the mean as a threshold.
3. Discuss how skewness supports your reasoning.

Solutions and Explanations

Solution 1: Comparing Central Tendencies

From the dataset:

Measure	Excel Formula	Result
Mean	=AVERAGE(B2:B1001)	$124,988.53
Median	=MEDIAN(B2:B1001)	$101,668.50
Mode	=MODE.MULT(B2:B1001)	$83,640 and $62,973

Ordered:

Mode < Median < Mean

This pattern suggests right skewness, because high-income outliers pull the mean upward.

Solution 2: Skewness

=SKEW(B2:B1001) → 1.6

A skewness of 1.6 indicates strong positive skew.

Interpretation:

- A small number of very high incomes stretch the distribution.
- Most households earn less than the mean.
- The distribution has a long right tail.

Solution 3: Kurtosis

=KURT(B2:B1001) → 2.8

Since Excel returns excess kurtosis, 2.8 indicates heavy tails relative to normal.

Meaning:

- Extreme incomes occur more often than expected under normal distribution.
- High-income outliers meaningfully influence the dataset.

Solution 4: Histogram

The histogram should show:

- Most incomes clustered between $40,000 and $80,000.
- A long right tail extending toward $500,000.

The visual confirms:

- Positive skew.
- Concentration in lower income ranges.
- Sparse but influential high earners.

Solution 5: Mean vs Median Decision

The median should be reported.

Reason:

- The distribution is strongly right-skewed.
- The mean is inflated by high earners.
- The median better represents the typical household.

This demonstrates why distribution shape must guide summary choice.

Solution to Challenge 1: Outlier Removal

After removing top 5%:

- Mean decreases significantly.
- Median changes slightly.
- Skewness reduces noticeably.

Conclusion:

The mean is more sensitive to extreme values than the median.

Skewness also decreases because the right tail shortens.

Solution to Challenge 2: Income Cap Simulation

After capping at $250,000:

- Skewness decreases.
- Kurtosis drops substantially.

Interpretation:

Skewness and kurtosis are highly sensitive to extreme values.
They are measures of tail behavior — when tails shrink, these measures respond.

Solution to Challenge 3: Policy Design

Using =PERCENTILE.INC(range, 0.90) gives the top 10% cutoff.

Percentiles are better because:

- They define position within distribution.
- They are not distorted by extreme values.

- They directly identify population segments.

Skewness supports this choice because a skewed distribution makes averages misleading for threshold design.

Chapter 5 Mastery Check

If you can:

- Predict skewness before calculating it,
- Explain kurtosis without using formulas,
- Identify when the mean is misleading,
- Connect distribution shape to policy decisions,

Then you are beginning to think like a data scientist.

Chapter 6 Practical Exercises and Challenges

Exercises and Challenges

These exercises are designed to help you apply the visualization principles discussed in this chapter using Excel. The tasks focus on selecting appropriate chart types, improving visual clarity, identifying misleading visualizations, and communicating insights effectively.

Exercise 1: Choosing the Appropriate Chart Type

You are given the following dataset representing quarterly sales for four products.

Product	Q1	Q2	Q3	Q4
A	120	135	150	170
B	95	110	115	130
C	80	90	100	105
D	60	75	85	95

Tasks

1. Enter the dataset into Excel.
2. Create a chart that compares the sales of the four products across the quarters.
3. Explain why your chosen chart type is appropriate.

Exercise 2: Visualizing a Trend Over Time

Suppose a company recorded monthly website visits for one year.

Month	Visits
Jan	1200
Feb	1350
Mar	1500
Apr	1700
May	1800
Jun	2100
Jul	2200
Aug	2300
Sep	2400
Oct	2600
Nov	2800
Dec	3000

Tasks

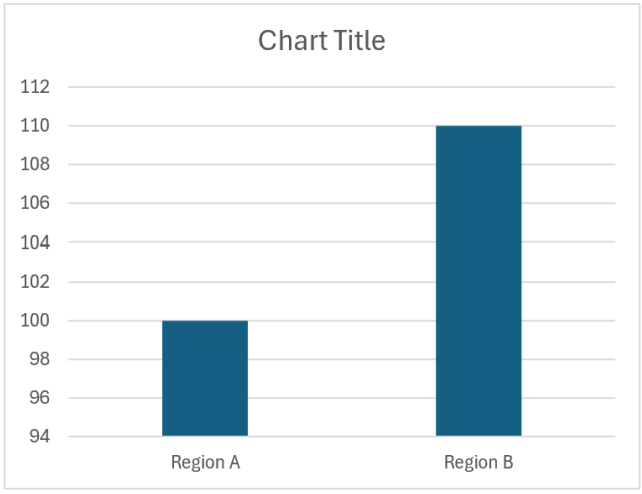
1. Create a visualization that clearly shows the trend in website visits.
2. Add a chart title and axis labels.
3. Describe the main trend visible in the chart.

Exercise 3: Identifying a Misleading Visualization

A column chart displays two values:

- Region A: 100
- Region B: 110

However, the vertical axis begins at 94 instead of zero.



Tasks

1. Explain why this chart is misleading.
2. Describe how to correct the chart in Excel.

Exercise 4: Reducing Visual Clutter

You are given a chart showing eight different product lines on the same line graph. The lines overlap heavily, making the chart difficult to interpret.

Tasks

1. Explain why this chart violates good visualization practices.
2. Suggest two alternative ways to present the information more clearly.

Exercise 5: Using Annotations to Highlight Insights

Create a small dataset showing quarterly profit growth for a company:

Quarter	Profit ($000)
Q1	50
Q2	65
Q3	70
Q4	120

Tasks

1. Create a line chart.
2. Add an annotation highlighting the most significant change in the data.
3. Write a one-sentence insight describing the key takeaway.

Challenge Exercise: Building a Data Story

A retail company records the following annual sales (in millions of dollars):

Year	Sales
2018	12
2019	13
2020	11
2021	14
2022	18

Tasks

1. Create a chart that shows the sales trend.

2. Identify the key insight revealed by the data.
3. Write a short data story using the Context → Insight → Action structure introduced in this chapter.

Solutions and Explanations

Solution 1: Choosing the Appropriate Chart Type

The most appropriate chart for comparing product sales across quarters is a clustered column chart.

Steps in Excel

1. Enter the dataset into a worksheet.
2. Highlight the entire table.
3. Go to Insert → Column Chart → Clustered Column.

Explanation:

Column charts are well suited for comparing categories across multiple groups. In this example:

- Products represent categories.
- Quarters represent grouped comparisons.

The chart allows viewers to easily compare both product performance and quarterly growth.

Solution 2: Visualizing a Trend Over Time

The best chart for displaying website visits across months is a line chart.

Steps in Excel

1. Select the dataset.
2. Go to Insert → Line Chart → 2-D Line Chart.
3. Add labels:
    - Chart title: Monthly Website Visits
    - X-axis: Month
    - Y-axis: Number of Visits

Explanation:

Line charts are ideal for displaying changes over time because they highlight trends and directional movement. In this example, the chart clearly shows a steady upward trend in website traffic throughout the year.

Solution 3: Identifying a Misleading Visualization

Why the Chart is Misleading

The chart exaggerates the difference between Region A and Region B by starting the vertical axis at 95 instead of zero. Although the actual difference is only **10 units**, the visual difference appears much larger.

How to Fix It:

1. Right-click the vertical axis.
2. Select Format Axis.
3. Set the minimum value to zero.

Explanation:

Bar and column charts represent values using length, and the baseline must begin at zero to preserve proportional comparisons.

Solution 4: Reducing Visual Clutter

Problem:

A chart with eight overlapping lines becomes difficult to interpret because viewers cannot easily distinguish individual trends.

Possible Solutions

Solution 1: Use Small Multiples

Create separate charts for each product line while keeping the same scale. This allows viewers to compare patterns across charts.

Solution 2: Filter Key Series

Display only the most important product lines in the main chart and place the remaining series in a secondary chart or appendix.

Explanation:

Effective visualization prioritizes clarity over completeness. Presenting too much information at once often reduces understanding.

Solution 5: Using Annotations to Highlight Insights

Chart Creation: Create a line chart showing quarterly profits.

Annotation: Add a text box near the Q4 data point stating:

> *"Sharp increase in profit during Q4."*

Insight: Profit increased dramatically in Q4, suggesting a strong seasonal sales effect.

Explanation: Annotations guide viewers directly to the most important element of the chart, reinforcing the narrative message.

Solution to the Challenge Exercise: Building a Data Story

Chart: A line chart displaying sales from 2018 to 2022 effectively shows the trend.

Key Insight: Sales declined in 2020 but increased significantly afterward, reaching the highest level in 2022.

Example Data Story

Context: Annual sales data from 2018 to 2022 shows how the company's revenue changed over time.

Insight: Sales dropped in 2020 but rebounded strongly in the following years, reaching $18 million in 2022.

Action: The company should analyze the strategies used during the recovery period and consider expanding the initiatives that contributed to the recent growth.

Chapter 7 Practical Exercises and Challenges

Exercises and Challenges

These exercises will help you apply the techniques learned in this chapter to analyze relationships between variables. The problems progress from basic interpretation to more analytical thinking.

Exercise 1: Interpreting Categorical Relationships

A retail company surveys 240 customers about their preferred shopping method.

Shopping Method	Online	In-Store	Mobile App	Total
Prefer Discounts	40	25	35	100
Prefer Convenience	30	15	45	90
Prefer Product Variety	20	25	5	50
Total	90	65	85	240

Tasks

1. Convert the table to column percentages.
2. Which shopping method has the highest percentage of customers who prefer convenience?
3. Do shopping method and customer preference appear independent or related?

Exercise 2: Comparing Numerical Values Across Categories

A company tracks monthly productivity scores for employees in three departments.

Department	Productivity Scores
Sales	72, 68, 75, 80, 74
Marketing	78, 82, 79, 81, 77
Engineering	85, 88, 84, 90, 87

Tasks

1. Calculate the mean productivity score for each department.
2. Which department has the highest average productivity?
3. Based on these averages, what preliminary conclusion can you draw about productivity differences?

Exercise 3: Calculating Correlation

A business records monthly marketing spending and sales revenue.

Marketing Spend ($1000s)	Sales ($1000s)
5	60
7	65
9	72
10	75
12	80
14	90

Tasks

1. Create a scatterplot of the data.
2. Calculate the correlation coefficient (r) using Excel.
3. Describe the strength and direction of the relationship.

Exercise 4: Interpreting r and r^2

Suppose a study finds the correlation between hours of exercise per week and resting heart rate is:

$r = -0.70$

Tasks

1. What does the negative sign indicate?
2. Calculate **r^2**.
3. Interpret the meaning of r^2 in this context.

Exercise 5: Identifying Misinterpretations

For each statement below, determine whether it is correct or incorrect, and explain why.

1. "If two variables are correlated, one must cause the other."
2. "A correlation of 0 means the variables have no relationship."
3. "Outliers can affect the correlation coefficient."
4. "Correlation measures the direction and strength of a linear relationship."

Exercise 6: Decision Framework Application

For each pair of variables below, identify the appropriate method and visualization.

Variable 1	Variable 2
Gender	Preferred Social Media Platform
Education Level	Annual Salary
Study Hours	Exam Score
Product Category	Customer Satisfaction Score

Tasks:

1. Identify the variable types.
2. Select the appropriate analysis method.
3. Suggest a visualization.

Solutions and Explanations

Solution 1: Categorical Relationships

Step 1: Convert to Column Percentages

Online (90 total)

Preference	Percentage
Discounts	44%
Convenience	33%
Variety	22%

In-Store (65 total)

Preference	Percentage
Discounts	38%
Convenience	23%
Variety	38%

Mobile App (85 total)

Preference	Percentage
Discounts	41%
Convenience	53%
Variety	6%

Step 2: Highest Convenience Preference

The Mobile App channel has the highest convenience preference at 53%.

Step 3: Independence or Relationship

The percentages differ noticeably across shopping methods. Because the distributions are not similar, shopping method and customer preference appear related.

Solution 2: Comparing Numerical Values Across Categories

Step 1: Calculate Means

Sales

Mean = (72 + 68 + 75 + 80 + 74) ÷ 5 = 73.8

Marketing

Mean = (78 + 82 + 79 + 81 + 77) ÷ 5 = 79.4

Engineering

Mean = (85 + 88 + 84 + 90 + 87) ÷ 5 = 86.8

Step 2: Highest Average Productivity

Engineering has the highest average productivity.

Step 3: Interpretation

Engineering employees appear to have higher productivity scores compared with Sales and Marketing. A boxplot would help visualize the differences in distributions.

Solution 3: Correlation Analysis

Step 1: Scatterplot

The scatterplot would show an upward trend, indicating that higher marketing spending tends to correspond with higher sales.

Step 2: Correlation Calculation

Using Excel:

=CORREL(A16:A21,B16:B21)

Result:

$r \approx +0.99$

Step 3: Interpretation

This indicates a very strong positive correlation. As marketing spending increases, sales revenue tends to increase.

Solution 4: Interpreting r and r^2

Negative Sign: The negative sign indicates an inverse relationship. As exercise hours increase, resting heart rate tends to decrease.

Calculate r^2

$r^2 = (-0.70)^2 = 0.49$

Interpretation: About 49% of the variation in resting heart rate is associated with variation in exercise levels. The remaining variation likely comes from other factors such as diet, genetics, and age.

Solution 5: Identifying Misinterpretations

1. **Incorrect**: Correlation does not imply causation. A third variable could influence both.
2. **Incorrect**: A correlation of zero means there is no linear relationship, but a nonlinear relationship could still exist.
3. **Correct**: Extreme values can significantly distort the correlation coefficient.
4. **Correct**: Correlation measures the strength and direction of a linear relationship.

Solution 6: Decision Framework Application

Variable Pair	Variable Types	Method	Visualization
Gender × Social Media Platform	Categorical × Categorical	Contingency table	Grouped bar chart
Education Level × Salary	Categorical × Numerical	Compare group means	Boxplots
Study Hours × Exam Score	Numerical × Numerical	Correlation analysis	Scatterplot
Product Category × Satisfaction Score	Categorical × Numerical	Group comparison	Boxplots

Explanation:

Each pair requires selecting a method based on variable types. This is the core idea behind the decision framework introduced in this chapter.

Chapter 8 Practical Exercises and Challenges

Exercises and Challenges

Exercise 1: Interpreting a Regression Equation

A regression analysis produced the following equation for predicting exam scores based on study hours:

$$\hat{Y} = 52 + 3.5X$$

Where:

- X = Number of hours studied
- $\hat{Y}$ = Predicted exam score

Questions

1. What does the intercept (52) represent in this context?
2. What does the slope (3.5) tell us about the relationship between study hours and exam score?
3. What exam score would you predict for a student who studies **6** hours?

Exercise 2: Calculating SS_{tot}, SS_{res}, and R^2

Suppose you have the following dataset:

Observation	Actual Y	Predicted Y
1	40	42
2	50	48
3	60	59
4	70	68

Tasks

1. Calculate the mean of Y.
2. Compute the Total Sum of Squares (SS_{tot}).
3. Compute the Residual Sum of Squares (SS_{res}).
4. Calculate the coefficient of determination (R^2).

Exercise 3: Interpreting R^2

A regression model predicting house prices from house size produces an R^2 value of 0.82.

Questions

1. What does this value tell us about the model?
2. How much variation in house prices remains unexplained by the model?
3. Is this generally considered a strong model? Explain your reasoning.

Exercise 4: Checking Regression Assumptions

You run a regression model and create a residual plot. The plot shows a clear fan-shaped pattern, where residuals spread out as X increases.

Questions

1. Which regression assumption is likely violated?
2. What does this pattern suggest about the variability of the residuals?
3. What might be one possible remedy?

Exercise 5: Regression vs Correlation

Consider two variables: hours spent exercising per week and resting heart rate.

The correlation between the two variables is −0.65.

Questions

1. What does the negative correlation indicate?
2. How does correlation differ from regression in analyzing this relationship?
3. If the correlation were 0, what would the slope of the regression line likely be?

Exercise 6: Identifying Model Limitations

A company builds a regression model predicting employee productivity based on coffee consumption. The model finds a strong positive relationship.

Questions

1. Why should analysts be cautious about concluding that coffee consumption causes higher productivity?
2. Suggest two other variables that might influence productivity.
3. What steps could analysts take to better understand the relationship?

Solutions and Explanations

Solution 1: Interpreting a Regression Equation

1. Intercept (52)
   The intercept represents the predicted exam score when study hours = 0. In this case, a student who does not study at all is predicted to score 52 points.
2. Slope (3.5)
   The slope indicates that for each additional hour studied, the predicted exam score increases by 3.5 points.
3. Prediction for 6 hours

   $\hat{Y} = 52 + 3.5(6)$
   $\hat{Y} = 52 + 21 = 73$
   Predicted score = 73

Solution 2: Calculating (SS_{tot}), (SS_{res}), and (R^2)

Step 1: Mean of Y

$\bar{Y} = (40 + 50 + 60 + 70) / 4$
$\bar{Y} = 55$

Step 2: Calculate (SStot)

$SS_{tot} = \sum(Y_i - \bar{Y})^2$

Y	(Y-$\bar{Y}$)	Square
40	−15	225
50	−5	25
60	5	25
70	15	225

$SS_{tot} = 225 + 25 + 25 + 225 = 500$

Step 3: Calculate (SSres)

$SS_{res} = \sum(Y_i - \hat{Y}_i)^2$

Actual Y	Predicted Y	Residual	Square
40	42	−2	4
50	48	2	4
60	59	1	1
70	68	2	4

$SS_{res} = 4 + 4 + 1 + 4 = 13$

Step 4: Calculate (R^2)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{13}{500}$$

$$R^2 = 1 - 0.026$$

$$R^2 = 0.974$$

Interpretation:

The model explains 97.4% of the variation in Y.

Solution 3: Interpreting R^2

1. An R^2 of 0.82 means that 82% of the variation in house prices is explained by house size.
2. The unexplained variation is:
   1 - 0.82 = 0.18
   So 18% of the variation is due to other factors.
3. Yes, this is generally considered a strong model, though interpretation depends on the context and complexity of the data.

Solution 4: Checking Regression Assumptions

1. The violated assumption is homoscedasticity.
2. The fan-shaped pattern indicates heteroscedasticity, meaning the variance of residuals increases as X increases.
3. Possible remedies include:
   - Transforming the dependent variable (for example using a log transformation)
   - Using a different regression model
   - Investigating whether another predictor variable should be included.

Solution 5: Regression vs Correlation

1. A correlation of −0.65 indicates a moderately strong negative relationship. As exercise hours increase, resting heart rate tends to decrease.

2. Correlation measures the strength and direction of the relationship. Regression models the relationship mathematically and allows predictions.
3. If correlation were 0, the regression slope would also be 0, producing a flat line at the mean of Y.

Solution 6: Identifying Model Limitations

1. Analysts should be cautious because correlation does not imply causation. Coffee consumption may be associated with productivity but may not cause it.
2. Other variables that might influence productivity include:
   - Employee experience
   - Work environment
   - Sleep quality
   - Motivation or job satisfaction
3. Analysts could:
   - Collect more variables and include them in a multiple regression model
   - Conduct controlled experiments
   - Use longitudinal data to observe changes over time.

The Analyst's Challenge

In these exercises, you will work with the used_cars_dataset available in Chapter 8's Excel workbook.

Dataset Variables (Features):

- Age (years)
- Mileage (thousands of miles)
- Price (dollars)

Exercise 1: Simple Regression

- Create a scatterplot of Age vs. Price
- Use Excel functions to calculate the slope and intercept
- Calculate R^2 using the correlation function
- Interpret your results: How much does a car's value decrease per year of age?

Exercise 2: Making Predictions

- Use your regression equation to predict the price of a 5-year-old car
- Use the FORECAST.LINEAR function to make the same prediction
- Calculate the residual if the actual price was $12,500

Exercise 3: Multiple Regression with ToolPak

- Use the Analysis ToolPak to perform multiple regression with both Age and Mileage as predictors
- Interpret the coefficients: What's the effect of each additional year of age? Each additional thousand miles?
- Compare the R^2 values from simple vs. multiple regression
- Which model would you trust more for making predictions?

Exercise 4: Model Evaluation

- Create residual plots to check regression assumptions
- Identify any outliers or patterns in the residuals
- Calculate MAE for your model manually using Excel functions
- What does this tell you about the typical accuracy of your predictions?

Exercise 5: Critical Thinking

- Would you use your model to predict the price of a 20-year-old car with 300,000 miles? Why or why not?
- If you found that more expensive cars tend to have higher mileage in your dataset, would you conclude that high mileage causes high prices? What alternative explanations might there be?
- How might you improve your model's predictive power?

Reflection Questions

After completing the exercises, consider these questions:

- How comfortable do you feel interpreting regression output?
- What surprised you most about the relationship between the variables?
- What questions do you still have about regression analysis?

- How might you apply regression analysis to a problem in your field of interest?

Remember, regression analysis is both an art and a science. The mathematical calculations are straightforward, but interpreting results, checking assumptions, and drawing appropriate conclusions require careful thought and domain expertise. Practicing with real data is the best way to develop these skills.

Chapter 9 Practical Exercises and Challenges

This section provides hands-on problems to help you apply the probability concepts introduced in this chapter. Try solving the problems on your own before reviewing the solutions.

Exercises and Challenges

Exercise 1: Identifying the Sample Space

A company tracks the outcome of a website visit. Each visitor can have one of the following outcomes:

- Purchase
- Add to Cart
- Browse Only
- Immediate Exit

Questions

1. Define the sample space for this experiment.
2. Identify two possible events from this sample space.
3. Is the event “Purchase” a simple event or a compound event?

Exercise 2: Types of Events

A six-sided die is rolled.

Questions

1. Define the event A = getting an even number.
2. Define the event B = getting a number greater than 4.
3. List the outcomes for events A and B.
4. Are events A and B mutually exclusive?

Exercise 3: Complement Rule

An email marketing campaign has a click rate of 12%.

Questions

1. What is the probability that a recipient does not click the link?
2. Express both probabilities in decimal form and percentage form.

Exercise 4: Addition Rule

A dataset of 1,000 customers shows the following information:

- 400 customers purchased Product A
- 300 customers purchased Product B
- 120 customers purchased both products

Questions

1. What is the probability that a randomly selected customer purchased Product A?
2. What is the probability that a customer purchased Product B?
3. What is the probability that a customer purchased A or B?

Exercise 5: Multiplication Rule (Independent Events)

A website experiment records two independent events:

- Probability a visitor clicks an advertisement: 0.30
- Probability the visitor signs up for a newsletter: 0.20

Question

What is the probability that a visitor both clicks the advertisement and signs up for the newsletter, assuming independence?

Exercise 6: Conditional Probability

A dataset of **1,000 website users** contains the following information:

- 300 users opened a promotional email
- 120 users clicked a link in the email
- 90 users both opened the email and clicked the link

Questions

1. What is the probability that a user opened the email?
2. What is the probability that a user clicked the link?
3. What is the probability that a user clicked given that they opened the email?

Exercise 7: Testing Independence

Using the information from Exercise 6:

Determine whether the events Open Email and Click Link are independent.

Exercise 8: Bayes' Theorem

A fraud detection system identifies suspicious transactions.

Given:

- 2% of transactions are fraudulent
- The system correctly detects fraud 90% of the time
- The system incorrectly flags 5% of legitimate transactions

Question

If a transaction is flagged as suspicious, what is the probability that it is actually fraudulent?

Exercise 9: Excel Application

You have an Excel dataset containing 1,000 email recipients with the following columns:

Column	Description
A	Customer ID
B	Opened Email (TRUE/FALSE)
C	Clicked Link (TRUE/FALSE)

Tasks

1. Calculate the probability that a recipient opened the email.
2. Calculate the probability that a recipient clicked the link.
3. Calculate the conditional probability that a user clicks the link given that they opened the email.

Solutions and Explanations

Solution 1: Identifying the Sample Space

Sample Space:

S = {Purchase, Add to Cart, Browse Only, Immediate Exit}

Example Events:

- Event A: Purchase
- Event B: Purchase or Add to Cart

Event Type:

"Purchase" is a simple event because it consists of one outcome.

Solution 2: Types of Events

Event definitions:

A = {2, 4, 6}
B = {5, 6}

Mutual Exclusivity: Events are not mutually exclusive because both include **6**.

Solution 3: Complement Rule

Click rate:

$$P(\text{Click}) = 0.12$$

Using the complement rule:

$$P(\text{No Click}) = 1 - P(\text{Click})$$

$$P(\text{No Click}) = 1 - 0.12 = 0.88$$

Results

- Click probability = 0.12 (12%)
- No-click probability = 0.88 (88%)

Solution 4: Addition Rule

Total customers = 1000

Step 1: P(A) = 400 / 1000 = 0.40

Step 2: P(B) = 300 / 1000 = 0.30

Step 3: Using the addition rule:

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

P(A or B) = 0.40 + 0.30 - 0.12

P(A ∪ B) = 0.58

Interpretation:

There is a 58% probability that a customer purchased Product A or Product B.

Solution 5: Multiplication Rule (Independent Events)

For independent events:

$$P(A \cap B) = P(A) \times P(B)$$

P(Click and Signup) = 0.30 * 0.20

= 0.06

Result

Probability = 0.06 (6%)

Solution 6: Conditional Probability

Total users = 1000

Email Open Probability:

P(Open) = 300 / 1000 = **0.30**

Click Probability:

P(Click) = 120 / 1000 = **0.12**

Conditional Probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

P(Click | Open) = P(Click and Open) / P(Open)

P(Click | Open) = (90/1000) / (300/1000) = 0.09 /0.3

P(Click | Open) = 0.30

Interpretation

Among users who opened the email, 30% clicked the link.

Solution 7: Testing Independence

To test independence:

Check whether:

P(Click | Open) = P(Click)

From earlier results:

P(Click | Open) = 0.30

P(Click) = 0.12

Since these values are not equal, the events are dependent.

Interpretation

Opening an email significantly increases the likelihood of clicking the link.

Solution 8: Bayes' Theorem

Step 1: Define the Events

Let:

- F: Transaction is fraudulent
- $\neg F$: Transaction is not fraudulent
- S: Transaction is flagged as suspicious

Given:

- $\mathrm{P(F)} = 0.02$
- $\mathrm{P(\neg F)} = 0.98$
- $\mathrm{P(S \mid F)} = 0.90$
- $\mathrm{P(S \mid \neg F)} = 0.05$

We want:

$$P(F \mid S)$$

Step 2: Apply Bayes' Theorem

$$\mathrm{P(F \mid S)} = \frac{\mathrm{P(S \mid F)P(F)}}{\mathrm{P(S)}}$$

Step 3: Compute $P(S)$

$$\mathrm{P(S)} = \mathrm{P(S \mid F)P(F)} + \mathrm{P(S \mid \neg F)P(\neg F)}$$
$$\mathrm{P(S)} = (0.90 \times 0.02) + (0.05 \times 0.98)$$
$$\mathrm{P(S)} = 0.018 + 0.049 = 0.067$$

Step 4: Compute Final Probability

$$P(F \mid S) = \frac{0.90 \times 0.02}{0.067}$$

$$P(F \mid S) = \frac{0.018}{0.067} \approx 0.2687$$

$$P(F \mid S) \approx 0.2687 \text{ (26.9\%)}$$

Interpretation (Critical Insight)

Even though the system is 90% accurate at detecting fraud, a flagged transaction is only about 27% likely to actually be fraudulent.

Why?

Because fraud is rare (only 2%), false positives from legitimate transactions accumulate.

Solution 9: Excel Application

Probability of Opening

=COUNTIF(B2:B1001,TRUE)/1000

Probability of Clicking

=COUNTIF(C2:C1001,TRUE)/1000

Conditional Probability

=COUNTIFS(B2:B1001,TRUE,C2:C1001,TRUE)/COUNTIF(B2:B1001,TRUE)

This formula calculates:

P(Click | Open)

which measures how likely a user is to click after opening the email.

Chapter 10: Practical Exercises and Challenges

Exercises and Challenges

A. Conceptual Understanding

Exercise 1: Define a probability distribution. What is the difference between a discrete and a continuous probability distribution?

Exercise 2: State two real-world examples where a Binomial Distribution would be appropriate.

Exercise 3: When is the Poisson Distribution preferred over the Binomial Distribution?

Exercise 4: List and explain three key characteristics of the Normal Distribution.

Exercise 5: Explain the Empirical Rule (68–95–99.7) in your own words.

B. Discrete Distributions

Exercise 6 (Binomial): A biased coin has a probability of heads = 0.6. It is tossed 8 times. What is the probability of getting exactly 5 heads?

Exercise 7 (Binomial – Excel): Using Excel, compute the probability of getting at most 3 successes in 10 trials when p = 0.4.

Exercise 8 (Poisson): A call center receives an average of 4 calls per hour. What is the probability of receiving exactly 6 calls in an hour?

C. Normal Distribution

Exercise 9: Exam scores are normally distributed with mean = 70 and standard deviation = 10. Find the probability that a student scores less than 85.

Exercise 10: Using the same distribution, find the probability that a student scores between 60 and 80.

Exercise 11: Find the score that corresponds to the top 5% of students.

D. Standardization and Z-Scores

Exercise 12: A student scored 78 in a test where the mean = 65 and standard deviation = 13. Calculate the z-score and interpret it.

Exercise 13: Convert a z-score of 1.5 into its corresponding probability.

Exercise 14: What proportion of values lie between z = −1 and z = 2?

E. Applied Data Science Challenges

Challenge 1: Quality Control

A factory produces light bulbs with an average lifetime of 1,000 hours and a standard deviation of 100 hours.

- What percentage of bulbs last more than 1,200 hours?
- What percentage last between 900 and 1,100 hours?

Challenge 2: Business Decision

A company's daily sales follow a normal distribution with mean = 500 units and standard deviation = 80 units.

- What is the probability sales exceed 650 units?
- What sales level corresponds to the top 10% of days?

Challenge 3: Risk Analysis

The number of system failures per week follows a Poisson distribution with $\lambda = 2$.

- What is the probability of zero failures in a week?
- What is the probability of more than 3 failures?

Solutions and Explanations

A. Conceptual Solutions

Solution 1: A probability distribution describes how probabilities are assigned to outcomes of a random variable.

- Discrete: countable outcomes (e.g., number of defects)
- Continuous: measurable outcomes (e.g., height, time)

Solution 2: Examples:

- Number of successful sales calls out of 20 attempts
- Number of defective items in a batch

Solution 3: Poisson is preferred when:

- Events occur over time/space
- Events are independent
- We count occurrences in an interval

Solution 4

- Symmetry
- Mean = Median = Mode
- Bell-shaped curve

Solution 5

- 68% within 1 SD
- 95% within 2 SD
- 99.7% within 3 SD

B. Discrete Distributions Solutions

Solution 6 (Binomial)

A biased coin has $p = 0.6, n = 8$. Find $P(X = 5)$.

We use the binomial formula:

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

$$P(X = 5) = \binom{8}{5} (0.6)^5 (0.4)^3$$

$$= 56 \times 0.07776 \times 0.064$$

$$= 56 \times 0.00497664 \approx 0.2787$$

Final Answer:

$$P(X = 5) \approx 0.279$$

Solution 7 (Excel)

Find $P(X \leq 3)$for $n = 10, p = 0.4$

Excel Formula:

=BINOM.DIST(3, 10, 0.4, TRUE)

Result:

$$P(X \leq 3) \approx 0.3823$$

Interpretation: ~38.23% chance of at most 3 successes.

Solution 8 (Poisson)

Average rate $\lambda = 4$, find $P(X = 6)$

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

$$P(X = 6) = \frac{e^{-4} \cdot 4^{6}}{6!}$$

$$= \frac{0.0183 \times 4096}{720} \approx \frac{74.96}{720} \approx 0.1041$$

Final Answer:

$$P(X = 6) \approx 0.104$$

C. Normal Distribution Solutions

Solution 9

=NORM.DIST(85, 70, 10, TRUE)

Result ≈ 0.9332

Interpretation: 93.32% score below 85.

Solution 10

=NORM.DIST(80, 70, 10, TRUE) - NORM.DIST(60, 70, 10, TRUE)

Result ≈ 0.6826

Interpretation: ~68.26% fall between 60 and 80.

Solution 11

=NORM.INV(0.95, 70, 10)

Result ≈ 86.45

Interpretation: Top 5% score above ~86.45.

D. Standardization Solutions

Solution 12

Given:

$$X = 78,\ \mu = 65,\ \sigma = 13$$

Compute Z-score

$$z = \frac{78 - 65}{13} = \frac{13}{13} = 1$$

Final Answer:

$$z = 1$$

Interpretation: The student's score is 1 standard deviation above the mean.

Meaning:

- The student performed better than average
- Roughly 84% of students scored below this mark

Solution 13

=NORM.S.DIST(1.5, TRUE)

Result ≈ 0.9332

Solution 14

=NORM.S.DIST(2, TRUE) - NORM.S.DIST(-1, TRUE)

Result ≈ 0.8186

Interpretation: ~81.86% of values fall in this range.

E. Applied Challenges Solutions

Challenge 1: Quality Control

Given:

- Mean $\mu = 1000$
- Standard deviation $\sigma = 100$

(a) Percentage of bulbs lasting more than 1,200 hours

Step 1: Standardize

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{1200 - 1000}{100} = 2$$

Step 2: Use Excel

=1 - NORM.S.DIST(2, TRUE)

Result ≈ 0.0228 (2.28%)

Final Answer:

$$\approx 2.28\%$$

(b) Percentage lasting between 900 and 1,100 hours

Step 1: Convert to Z-scores

$$z_1 = \frac{900 - 1000}{100} = -1$$

$$z_2 = \frac{1100 - 1000}{100} = 1$$

Step 2: Use symmetry

$$P(-1 < Z < 1) \approx 0.6826$$

Final Answer:

$$\approx 68.3\%$$

Insight: Most bulbs (about 68%) fall within ±1 standard deviation—indicating a stable production process.

Challenge 2: Business Decision

Given:

- $\mu = 500$
- $\sigma = 80$

(a) Probability sales exceed 650 units

Standardize

$$z = \frac{650 - 500}{80} = 1.875$$

Use Excel

=1 - NORM.S.DIST(1.875, TRUE)

Result ≈ 0.0304 (3.04%)

(b) Sales level for top 10%

Top 10% ⇒ upper tail ⇒

$$z \approx 1.28$$

Convert back

=NORM.INV(0.90, 500, 80)

Result ≈ 602.5 units

Final Answer:

$$\approx 602 \text{ units}$$

Insight**:** Only about 10% of days exceed ~602 units, which can guide:

- inventory planning
- staffing decisions
- performance targets

Challenge 3: Risk Analysis (Poisson)

Given:

$$\lambda = 2$$

(a) Probability of zero failures

=POISSON.DIST(0, 2, FALSE)

Result ≈ 0.1353

Answer:

$$\approx 13.5\%$$

(b) Probability of more than 3 failures

$$P(X > 3) = 1 - P(X \leq 3)$$

Compute cumulative:

=1 - POISSON.DIST(3, 2, TRUE)

Result ≈ 0.1429

Answer:

$$\approx 14.3\%$$

Insight:

Even with an average of only 2 failures; there is still a ~14% chance of 4+ failures, highlighting the importance of risk preparedness.

Chapter 11: Practical Exercises and Challenges

Exercises and Challenges

Exercise 1: Population vs Sample

A retail company has 5,000 customers. Due to time constraints, a data analyst selects 200 customers to study purchasing behavior.

Tasks

1. Identify the population and the sample.
2. Is the sample size reasonable? Briefly justify.
3. State one potential risk if the sample is not representative.

Exercise 2: Identifying Sampling Methods

For each scenario below, identify the sampling method used:

1. A researcher randomly selects 100 student IDs from a university database.
2. A company surveys the first 50 customers who enter a store.
3. A health study selects 30 patients from each age group.
4. A researcher surveys every 10th household on a street.

Exercise 3: Sampling Bias

A survey about internet usage is conducted using only responses from an online platform.

Tasks

1. Identify the type of bias present.
2. Explain how this bias affects the results.
3. Suggest one way to reduce the bias.

Exercise 4: Sample Size and Precision

A researcher wants to estimate a population mean with:

- Confidence level: 95%
- Standard deviation: 20
- Margin of error: 4

Task

Calculate the required sample size.

Exercise 5: Sampling Distribution

A population has:

- Mean = 100
- Standard deviation = 20
- Sample size = 25

Tasks

1. Compute the standard error of the mean.
2. Explain what this value represents.

Exercise 6: Central Limit Theorem (Conceptual)

A population is highly skewed.

Tasks

1. What does the Central Limit Theorem say about the sampling distribution of the mean?
2. What sample size is generally considered sufficient?

Exercise 7: Point Estimation

A sample of 80 households shows that 48 own a car.

Tasks

1. Compute the sample proportion.
2. Interpret this as a point estimate.

Exercise 8: Margin of Error (Proportion)

A survey finds that 60% of respondents support a policy.
Sample size = 400, confidence level = 95%.

Task

Compute the margin of error.

Exercise 9: Confidence Interval (Mean)

A sample of 50 students has:

- Mean score = 72
- Standard deviation = 10

Task

Construct a 95% confidence interval for the population mean.

Exercise 10: Confidence Interval (Proportion)

Out of 300 respondents, 180 prefer Product A.

Tasks

1. Compute the sample proportion.
2. Construct a 95% confidence interval.
3. Interpret the result.

Solutions and Explanations

Solution 1: Population vs Sample

1. Population = All 5,000 customers
   Sample = 200 selected customers
2. Yes, 200 can be reasonable depending on variability and desired precision.
3. Risk: Sampling bias → results may not reflect the entire customer base

Solution 2: Sampling Methods

1. Simple Random Sampling
2. Convenience Sampling
3. Stratified Sampling
4. Systematic Sampling

Solution 3: Sampling Bias

1. Selection Bias / Coverage Bias
2. Excludes individuals without internet access → results overrepresent online users
3. Use mixed data collection (offline + online)

Solution 4: Sample Size

Using:

$$n = \left(\frac{Z \cdot \sigma}{E}\right)^2$$

For 95% confidence, $Z = 1.96$:

$$n = \left(\frac{1.96 \cdot 20}{4}\right)^2 = (9.8)^2 = 96.04$$

Final Answer: $n = 97$ (round up)

Interpretation: At least 97 observations are needed to achieve the desired precision.

Solution 5: Sampling Distribution

Standard error:

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

Interpretation: Sample means typically vary by about ±4 units from the population mean.

Solution 6: Central Limit Theorem

1. The sampling distribution becomes approximately normal
2. Sample size: n ≥ 30

Solution 7: Point Estimation

$$\hat{p} = \frac{48}{80} = 0.6$$

Interpretation: Estimated proportion = 60% of households own a car

Solution 8: Margin of Error (Proportion)

Using:

$$MOE = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$MOE = 1.96 \cdot \sqrt{\frac{0.6(0.4)}{400}} = 1.96 \cdot \sqrt{0.0006}$$

$$= 1.96 \cdot 0.0245 \approx 0.048$$

Answer: ≈ 4.8%

Solution 9: Confidence Interval (Mean)

Using t-distribution (σ unknown):

$$CI = 72 \pm t \cdot \frac{10}{\sqrt{50}}$$

Approximate $t \approx 2.01$:

$$SE = \frac{10}{\sqrt{50}} \approx 1.414$$

$$MOE = 2.01 \cdot 1.414 \approx 2.84$$

$$CI = (69.16, 74.84)$$

Interpretation: We are 95% confident the true mean lies between 69.16 and 74.84.

Solution 10: Confidence Interval (Proportion)

1. Sample proportion:

$$\hat{p} = 180/300 = 0.6$$

2. Margin of error:

$$MOE = 1.96 \cdot \sqrt{\frac{0.6(0.4)}{300}} \approx 0.056$$

3. Confidence interval:

$$CI = (0.544, 0.656)$$

Interpretation: We are 95% confident that 54.4% to 65.6% of the population prefers Product A.

Chapter 13: Practical Exercises and Challenges

Exercises and Challenges

Exercise 1: One-Sample Z-Test (Mean)

A beverage company claims that its bottles contain 500 ml on average. The population standard deviation is known to be 20 ml. A random sample of 64 bottles has a mean volume of 495 ml.

At $\alpha = 0.05$, test whether the filling process is accurate.

Exercise 2: One-Sample Z-Test (Proportion)

A political party claims that 55% of voters support them. A survey of 400 voters shows that 200 support the party.

At $\alpha = 0.05$, test whether the support differs from the claim.

Exercise 3: Independent Two-Sample t-Test (Equal Variances)

Two machines produce metal rods. A sample from each machine gives:

- Machine A: $n_1 = 10$, $\bar{x}_1 = 50$, $s_1 = 4$
- Machine B: $n_2 = 12$, $\bar{x}_2 = 55$, $s_2 = 5$

Assume equal variances. At $\alpha = 0.05$, test whether the machines produce rods with different mean lengths.

Exercise 4: Paired Sample t-Test

A fitness program measures weight before and after training for 8 participants:

Before: [80, 85, 78, 90, 88, 76, 95, 89]
After: [78, 83, 77, 87, 85, 75, 92, 86]

At $\alpha = 0.05$, test whether the program leads to weight reduction.

Exercise 5: Chi-Square Goodness-of-Fit

A die is suspected to be biased. It is rolled 60 times, producing:

Face	**1**	**2**	**3**	**4**	**5**	**6**
Obs	8	9	10	11	12	10

Test at $\alpha = 0.05$ whether the die is fair.

Exercise 6: Chi-Square Test of Independence

A study examines whether education level is related to employment status:

	Employed	Unemployed
High School	40	20
Degree	60	10

At $\alpha = 0.05$, test whether the variables are independent.

Exercise 7: One-Way ANOVA

Three teaching methods yield the following test scores:

- Method A: [70, 72, 68, 71]
- Method B: [75, 78, 74, 77]
- Method C: [65, 67, 66, 64]

At $\alpha = 0.05$, test whether there is a difference in mean scores.

Exercise 8: Choosing the Right Test

For each scenario, identify the correct test:

a) Comparing average salary of one group to a known benchmark (σ unknown)
b) Testing whether gender is related to product preference
c) Comparing means of three independent groups
d) Comparing before-and-after measurements on the same individuals

Exercise 9: Interpretation Challenge

A hypothesis test produces a p-value of 0.08 at $\alpha = 0.05$.

- What is the decision?
- What does this mean in context?

Solutions and Explanations

Solution 1: One-Sample Z-Test (Mean)

Hypotheses: H_0: $\mu = 500$; H_1: $\mu \neq 500$ (Two-tailed)

Parameters: $\mu = 500$, $\sigma = 20$, $\bar{x} = 495$, $n = 64$

Compute z statistic:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

SE = 20 / √64 = 20 / 8 = 2.5

z = (495 − 500) / 2.5 = −2.0

Critical value (two-tailed, α = 0.05): ±1.96

Decision: |z| > 1.96 → Reject H_0

Conclusion: There is significant evidence that the mean fill volume differs from 500 ml — the filling process is inaccurate.

Solution 2: One-Sample Z-Test (Proportion)

Hypotheses: H_0: p = 0.55; H_1: p ≠ 0.55 (Two-tailed)

Parameters: p_0 = 0.55, p̂ = 200/400 = 0.50, σ = 20 , n = 400

p̂ = 0.50

SE = √[0.55 × 0.45 / 400] = 0.0248

z = (0.50 − 0.55) / 0.0248 ≈ −2.02

Decision: |z| > 1.96 Reject H_0

Conclusion: Support significantly differs from 55%.

Solution 3: Independent Two-Sample t-Test

Hypotheses: H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

Pooled variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{9(16) + 11(25)}{20} = \frac{144 + 275}{20} = 20.95$$

$$s_p = 4.58$$

SE = 4.58√(1/10 + 1/12) = 1.94

t = (50 − 55) / 1.94 = −2.58

df = 20

critical ≈ ±2.086

Decision: Reject H_0

Conclusion: Means differ significantly.

Solution 4: Paired t-Test

Differences: [2, 2, 1, 3, 3, 1, 3, 3]

Mean of Differences ($\bar{d}$): 2.25; Std Dev (s_d): 0.886

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

SE = 0.89 / $\sqrt{8}$ = 0.315

t = 2.25 / 0.315 ≈ 7.14

df = 7

$t_{crit} \approx 2.365$

Decision: Reject H_0

Conclusion: Strong evidence of weight reduction.

Solution 5: Chi-Square Goodness-of-Fit

Expected = 60 / 6 = 10 each

Compute χ^2:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} + \ldots = 0.4 + 0.1 + 0 + 0.1 + 0.4 + 0$$

$$\chi^2 = 1.0$$

$$df = 5$$

$$\chi^2_{crit} \approx 11.07$$

Decision: $1.0 < 11.07 \Rightarrow$ Fail to reject H_0

Conclusion: No evidence the die is biased.

Solution 6: Chi-Square Test of Independence

Compute Expected Values

Example:

$$E = \frac{Row \times Column}{Total}$$

HS employed: $\frac{60 \times 100}{130} = 46.15$

(Compute all cells similarly)

Compute Test Statistic

$$\chi^2 \approx 6.99$$

Compute Critical Value

df = 1

$$\chi^2_{crit} = 3.84$$

Decision: $6.99 > 3.84 \Rightarrow$ Reject H_0

Conclusion: Education level and employment status are not independent (they are related).

Solution 7: ANOVA

Means: A=70.25, B=76, C=65.5, Grand Mean: 70.58

$SS_{between}$: $= \sum n_i(\bar{x}_i - \bar{x}_{grand})^2$

$= 4[(70.25-70.58)^2 + (76-70.58)^2 + (65.5-70.58)^2] = 221.17$

$MS_{between}$: $= 221.17 / 2 = \mathbf{110.58}$

SS_{within}: $= \sum(x - \bar{x}_i)^2$ for each group

$= 8.75 + 10 + 5 = 23.75$

MS_{within}: $= 23.75/9 = \mathbf{2.6389}$

F: $= MS_{between} / MS_{within} = 110.58/2.6389 = 41.9$

df: = (2, 9)

F Critical: =F.INV.RT(0.05, 2, 9) = 4.26

Decision: 41.9 > 4.26. Reject H_0

Conclusion: At least one method differs significantly.

Solution 8: Choosing the Right Test

a) One-sample t-test
b) Chi-square test of independence
c) One-way ANOVA
d) Paired t-test

Solution 9: Interpretation Challenge

$p = 0.08 > 0.05 \rightarrow$ Fail to reject H_0

Interpretation:
There is not enough evidence to support the alternative hypothesis. This does not prove H_0 is true—it only indicates insufficient evidence against it.

Chapter 14 Practical Exercises and Challenges

Exercises and Challenges

Exercise 1: Testing Correlation Significance

A dataset of 25 students shows a correlation between hours studied and exam score of **r** = 0.52.

1. State the null and alternative hypotheses.
2. Compute the test statistic.
3. Determine the p-value ($\alpha = 0.05$).
4. Make a conclusion.

Exercise 2: Interpreting Correlation Output (Conceptual)

A researcher reports:

- $r = -0.72$
- p-value = 0.002

Answer the following:

1. Is the relationship strong or weak?
2. Is it positive or negative?

3. Is it statistically significant at $\alpha = 0.05$?
4. Interpret the result in plain language.

Exercise 3: Regression Slope Testing

A regression analysis yields:

- $b_1 = 4.5$
- $SE(b_1) = 1.2$
- $n = 18$

1. State hypotheses.
2. Compute the t-statistic.
3. Determine the critical value at $\alpha = 0.05$.
4. Make a decision.

Exercise 4: Correlation vs Regression Connection

Given:

- $r = 0.60$
- $n = 20$

1. Compute the t-statistic using the correlation formula.
2. Explain why this is equivalent to testing the regression slope.

Exercise 5: Excel-Based Regression Interpretation

You run a regression in Excel and obtain:

- Slope = 2.8
- p-value (slope) = 0.03
- $R^2 = 0.45$

1. Is the slope statistically significant at $\alpha = 0.05$?
2. Interpret the slope.
3. Interpret R^2.
4. Can this model be used for prediction?

Exercise 6: Assumption Checking

A residual plot shows a clear curved pattern.

1. Which assumption is violated?
2. What is the implication?
3. Suggest a solution.

Exercise 7: Real Data Mini-Analysis

A company tracks advertising (X) and sales (Y):

X	Y
2	40
3	50
4	65
5	70
6	80

1. Compute the correlation coefficient (use Excel).
2. Describe the relationship.
3. Would you expect the relationship to be significant? Why?

Exercise 8: Critical Thinking Challenge

A dataset shows:

- r = 0.25
- p-value = 0.001
- n = 500

1. Is the relationship statistically significant?
2. Is it practically strong?
3. Explain the difference between statistical significance and practical significance.

Solutions and Explanations

Solution 1: Testing Correlation

1. Hypotheses:

- H_0: ρ = 0
- H_1: ρ ≠ 0

2. Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.52\sqrt{25-2}}{\sqrt{1-0.52^2}} = \frac{0.52\sqrt{23}}{\sqrt{0.7296}} = \frac{0.52 \times 4.796}{0.854} = 2.92$$

   df = 23

3. p-value (Excel): =T.DIST.2T(2.92, 23)

   p ≈ 0.007

4. Conclusion: Since p < 0.05, reject H_0.

   There is a significant linear relationship.

Solution 2: Interpretation

1. Strength: Strong (|−0.72| is large)
2. Direction: Negative
3. Significance: Yes (0.002 < 0.05)
4. Interpretation:
   There is a strong, statistically significant negative relationship—when one variable increases, the other tends to decrease.

Solution 3: Regression Slope

1. Hypotheses:
   - H_0: $\beta_1 = 0$
   - H_1: $\beta_1 \neq 0$
2. Test statistic:

$$t = \frac{4.5}{1.2} = 3.75$$

   df = 16

3. Critical value: =T.INV.2T(0.05, 16) ≈ ±2.12
4. Decision: Since 3.75 > 2.12, **reject H_0**. The predictor is significant.

Solution 4: Correlation vs Regression

1. Test statistic:

$$t = \frac{0.60\sqrt{18}}{\sqrt{1-0.36}} = \frac{0.60 \times 4.243}{0.8} = 3.18$$

2. Explanation:
   Testing correlation ($\rho = 0$) is mathematically equivalent to testing slope ($\beta_1 = 0$). Both assess whether a linear relationship exists.

Solution 5: Regression Interpretation

1. Significance:
   Yes ($0.03 < 0.05$)
2. Slope interpretation:
   For every 1-unit increase in X, Y increases by 2.8 units.
3. R^2 interpretation:
   45% of the variation in Y is explained by X.
4. Prediction:
   Yes, but with moderate accuracy (since R^2 is not very high).

Solution 6: Assumption Violation

1. Violated assumption: Linearity
2. Implication: Linear regression model is inappropriate
3. Solution: Use polynomial or non-linear regression

Solution 7: Mini Analysis

1. Correlation (approximate): $r \approx 0.99$
2. Relationship: Very strong positive
3. Significance: Likely highly significant due to strong linear trend

Solution 8: Statistical vs Practical Significance

1. Statistical significance: Yes ($p = 0.001$)
2. Practical strength: Weak ($r = 0.25$)

3. Explanation:
   A large sample size can make even weak relationships statistically significant. However, the relationship may still have limited practical importance.