Statistics
in the Eye of a Nurse

A Clinical Approach to Data-Informed Practice

MAT 300 · Introduction to Statistics
Joyce University of Nursing and Health Sciences

Safaa Dabagh

Open Educational Resource · 2026

Statistics in the Eye of a Nurse: A Clinical Approach to Data-Informed Practice

Course: MAT 300 — Introduction to Statistics
Institution: Joyce University of Nursing and Health Sciences
Textbook alignment: Sullivan, *Statistics: Informed Decisions Using Data*, 6th Edition (Pearson)

All clinical examples, patient scenarios, and study data in this textbook are fictional and created solely for pedagogical purposes. Any resemblance to real patients, events, or institutions is coincidental.

Technology: Students are encouraged to use Desmos Scientific Calculator (`desmos.com/scientific`) and Microsoft Excel for all computations. Desmos is available free of charge and permitted on all assessments.

First edition, 2026.

For every nurse who ever wondered
why the researchers keep talking about p-values.

Now you know.

Preface

Statistics lives in every nursing action. The chart documenting a patient's vital signs every four hours is a time series. The hospital's infection rate is a proportion. The decision to change a treatment protocol based on a published study is applied hypothesis testing. The confidence interval in a clinical trial abstract is directly telling you how much uncertainty remains after the study was done.

This textbook was written because nursing students deserve a statistics course that meets them where they are: intelligent professionals with demanding clinical training who need statistics to make better decisions for patients, not to pass an abstract mathematics requirement.

Every example in this book involves patients, clinical outcomes, hospital data, or nursing practice. Every definition is followed by a clinical application. Every concept builds toward the final chapter, which asks: can you read a research article and judge whether its conclusions deserve to change what you do at the bedside?

How this book is organized.

Each section follows the same six-step rhythm:

1. **Read This First** — a clinical scenario that creates the problem
2. **Let's Talk About It** — intuition before formalism
3. **Now We Name It** — definitions, notation, and formulas
4. **Watch It Work** — a fully worked clinical example
5. **Your Turn** — practice problems with clinical context
6. **Think Like a Nurse** — interpretive close, connecting statistics to practice

Technology.

This course uses Desmos Scientific Calculator (`desmos.com/scientific`) for all probability and statistical computations. Desmos is free, runs in any browser, and is permitted on all assessments. Appendix A provides a complete reference guide to all Desmos commands used in this course. Excel alternatives are included throughout for students who prefer a spreadsheet environment.

A note on examples.

All patient scenarios, hospital names, study results, and clinical data in this book are entirely fictional. They were designed to be realistic and instructive, not to describe actual cases or institutions. When real statistical benchmarks are referenced (such as infection rate targets or NCLEX pass rates), these reflect general clinical knowledge at the time of writing.

Statistics will not make you a better nurse by itself. But knowing statistics will help you recognize a good study from a bad one, question a protocol that was never properly tested, and advocate for your patients with evidence. That is worth the work.

— Safaa Dabagh

Los Angeles, 2026

Contents

Preface 5

1 The Story Behind the Numbers 1
- 1.1 What Is Statistics and Why Should a Nurse Care? 1
- 1.2 What Kind of Data Are We Working With? 7
- 1.3 Levels of Measurement 11
- 1.4 Chapter Practice Problems 16
- Answer Key 19

2 How We Know What We Know: Study Design 22
- 2.1 Observational Studies vs. Experiments 22
- 2.2 The Structure of a Well-Designed Experiment 28
- 2.3 Types of Observational Studies 34
- 2.4 Chapter Practice Problems 39
- Answer Key 42

3 Collecting Good Data: Sampling 45
- 3.1 Why Sampling Matters 45
- 3.2 Five Sampling Methods 51
- 3.3 Sources of Bias in Sampling 57
- 3.4 Chapter Practice Problems 63
- Answer Key 65

4 Describing Data Visually 68
- 4.1 Organizing Data: Frequency Distributions 68
- 4.2 Graphs for Qualitative Data: Bar Charts and Pie Charts 74
- 4.3 Graphs for Quantitative Data: Histograms and More 79
- 4.4 Chapter Practice Problems 85
- Answer Key 87

5 Describing Data Numerically 90
- 5.1 Measures of Center: Mean, Median, and Mode 90
- 5.2 Measures of Spread: Range, Standard Deviation, and IQR 96

5.3 The Five-Number Summary, Boxplots, and Z-Scores . . . 102
5.4 Chapter Practice Problems . . . 107
Answer Key . . . 109

6 Correlation and Linear Regression 112
6.1 Scatter Plots and Correlation . . . 112
6.2 The Least-Squares Regression Line . . . 118
6.3 Chapter Practice Problems . . . 125
Answer Key . . . 128

7 Introduction to Probability 131
7.1 Probability Basics and the Addition Rule . . . 131
7.2 Conditional Probability and the Multiplication Rule . . . 137
7.3 Chapter Practice Problems . . . 143
Answer Key . . . 145

8 Probability Distributions 148
8.1 Discrete Probability Distributions . . . 148
8.2 The Binomial Distribution . . . 154
8.3 Chapter Practice Problems . . . 160
Answer Key . . . 162

9 The Normal Distribution 165
9.1 Properties of the Normal Distribution . . . 165
9.2 Finding Probabilities and Values Under the Normal Curve . . . 171
9.3 Chapter Practice Problems . . . 177
Answer Key . . . 179

10 The Central Limit Theorem 182
10.1 Sampling Distributions and the Standard Error . . . 182
10.2 Applying the Central Limit Theorem . . . 188
10.3 Chapter Practice Problems . . . 194
Answer Key . . . 196

11 Confidence Intervals 199
11.1 Confidence Intervals for a Mean . . . 199
11.2 Confidence Intervals for a Proportion and Sample Size . . . 205
11.3 Chapter Practice Problems . . . 210
Answer Key . . . 212

12 Hypothesis Testing: The Logic 215
12.1 The Logic of Hypothesis Testing . . . 215
12.2 Performing Hypothesis Tests . . . 222
12.3 Chapter Practice Problems . . . 228
Answer Key . . . 230

13 Hypothesis Testing: In Practice 233
13.1 Comparing Two Groups: t-Tests and Proportion Tests 233
13.2 The Chi-Square Test for Independence 240
13.3 Chapter Practice Problems 245
Answer Key 248

14 Comparing Groups: ANOVA 251
14.1 The Logic of ANOVA 251
14.2 Interpreting ANOVA in Clinical Research 258
14.3 Chapter Practice Problems 263
Answer Key 265

15 Reading the Research 268
15.1 Anatomy of a Research Article 268
15.2 Common Statistical Errors and Misleading Presentations 275
15.3 Chapter Practice Problems 280
Answer Key 282

Appendix A: Desmos Statistical Reference Guide 285
How to Use This Guide 285
A.1 Normal Distribution 285
A.2 t-Distribution 286
A.3 Chi-Square Distribution 287
A.4 F-Distribution 287
A.5 Quick Reference: Critical Values 288
A.6 Sample Size and Proportions 288
A.7 Exam-Day Checklist 288

Appendix B: Formula Sheet 290
B.1 Descriptive Statistics 290
B.2 Probability 291
B.3 Sampling Distributions and CLT 291
B.4 Confidence Intervals 292
B.5 Hypothesis Tests 292
B.6 ANOVA 293
B.7 Correlation and Regression 293
B.8 Clinical Research Measures 294

Appendix C: Statistical Tables 295
C.1 Standard Normal Table 295
C.2 t-Distribution Critical Values 297
C.3 Chi-Square Critical Values 299
C.4 F Critical Values 299

Glossary of Key Terms 301

CHAPTER 1

The Story Behind the Numbers

> *"Data is not numbers. Data is people."*
>
> — a principle worth remembering every time you open a chart

In this chapter, you will learn to:

- Explain what statistics is and why it matters in nursing practice
- Distinguish between a population and a sample
- Tell the difference between descriptive and inferential statistics
- Classify variables as qualitative or quantitative, discrete or continuous
- Identify the four levels of measurement: nominal, ordinal, interval, and ratio

1.1 What Is Statistics and Why Should a Nurse Care?

Step 1 Read This First

You are three months into your first job as a floor nurse at a large urban hospital. During a morning huddle, your charge nurse says:

"The hospital just published its quarterly report. Our unit's average patient satisfaction score dropped from 84% to 79%. Administration wants to meet with us next week."

Later that shift, a colleague hands you a printed study from a nursing journal:

"A randomized trial of 412 patients found that hourly rounding reduced call light usage by 38% ($p < 0.001$)."

By lunch, you have read three more numbers: a patient's sodium level of 148 mEq/L, a blood pressure of 142/88 mmHg, and a 30-day readmission rate of 12.4% for patients discharged from your unit last quarter.

Every one of those numbers is a statistical claim. Someone collected data, organized it, summarized it, and presented it as a fact that should influence a decision. Your job as a nurse is not just to receive those numbers — it is to understand them, question them, and act on them wisely.

That is what this course is about.

Step 2 Let's Talk About It

Before we define anything formally, think about the numbers from the scenario above.

When the charge nurse said satisfaction dropped from 84% to 79%, did that feel significant to you? What would you want to know before deciding whether to be worried?

When the journal article said hourly rounding reduced call light usage by 38%, did that convince you? What questions came to mind?

You already have statistical instincts. You already know that a single number rarely tells the whole story. You already suspect that *who was measured*, *how many were measured*, and *under what conditions* all matter.

Statistics gives precise language and reliable tools to those instincts. This course takes what you already do intuitively and makes it rigorous.

Step 3 Now We Name It

Definition: Statistics

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. When the information is numerical, we call it **data**.

Definition: Two Branches of Statistics

Descriptive statistics organizes and summarizes data that have already been collected. It describes what *was* observed.

Inferential statistics uses data from a **sample** to draw conclusions about a larger **population**. It goes beyond the data in hand to make predictions or generalizations.

Definition: Population and Sample

A **population** is the entire group of individuals or objects we want to study.

A **sample** is a subset of the population — the group we actually observe and measure.

In most real-world studies, we cannot observe the entire population. We study a sample and use what we learn to draw conclusions about the whole group.

Definition: Parameters and Statistics

A **parameter** is a numerical summary that describes a *population*. Parameters are usually unknown and are what we are trying to estimate.

A **statistic** is a numerical summary computed from a *sample*. Statistics are known from our data and are used to estimate parameters.

*Memory tip: **P**opulation → **P**arameter. **S**ample → **S**tatistic.*

Step 4 Watch It Work

Clinical Scenario: A hospital quality team wants to know the average wait time (in minutes) from triage to first nurse contact in the emergency department. They review the records of all 847 ED visits during February and find an average wait time of 23 minutes.

Meanwhile, a researcher studying the same question surveys a random sample of 60 patients from across the country and finds an average wait time of 31 minutes in that sample.

Classify each number and statement below.

Statement	Classification
The hospital reviewed all 847 February ED visits.	Population
The 847 visits produced an average wait of 23 minutes.	Parameter
The researcher's 60 patients are a subset of all patients.	Sample
The 31-minute average comes from those 60 patients.	Statistic
"The average ED wait time in our hospital in February was 23 minutes."	Descriptive statistics
"Based on our sample, we estimate the national average ED wait time is approximately 31 minutes."	Inferential statistics

Key observation: The hospital's statement is purely descriptive — it is reporting what actually happened in their own records. The researcher's statement is inferential — she is using 60 patients to say something about *all* patients, not just those 60.

Step 5 Your Turn

For each statement below, identify: (a) whether the group described is a population or a sample, and (b) whether the statement uses descriptive or inferential statistics.

1. A charge nurse surveys all 18 nurses on her floor and finds that 14 of them prefer 12-hour shifts. She reports: "78% of nurses on our floor prefer 12-hour shifts."
2. A researcher collects data from 200 randomly selected ICU patients across five hospitals and concludes: "ICU patients nationwide spend an average of 4.3 days on mechanical ventilation."
3. A hospital reviews the complete records of its 1,200 patients discharged last year and reports a 30-day readmission rate of 11.2%.

4. A public health researcher surveys 500 nurses in California and reports: “We estimate that approximately 68% of nurses in the United States have experienced burnout in the past year.”
5. A nurse manager pulls all shift reports from the past month and summarizes: “On average, our unit administered 14.7 medications per patient per shift.”

Step 6 Think Like a Nurse

A hospital administrator walks into your unit meeting and announces: "A new study found that nurses who use a structured handoff protocol make 22% fewer medication errors." She concludes the meeting by saying that your unit will adopt the new protocol starting next week.

Before the meeting ends, what statistical questions would a thoughtful nurse raise?

Consider: Where did the 22% come from? Was it a population or a sample? How large was the sample? Was the sample similar to your patient population? Is 22% a descriptive number (from the study's own nurses) or an inferential claim (applied to all nurses everywhere)?

These are not nitpicky questions. They are exactly the questions that determine whether a policy change is supported by solid evidence or by a number that sounds impressive but may not apply to your situation.

Statistics is not about being skeptical of everything. It is about knowing which questions to ask.

1.2 What Kind of Data Are We Working With?

Step 1 Read This First

You are preparing for a research summary assignment and pull up a dataset from a nursing outcomes study. The spreadsheet has the following columns:

Patient ID	Unit	Age	Pain Score	Temp (°F)
001	ICU	67	7	101.2
002	Med/Surg	43	3	98.6
003	ICU	71	9	102.4
004	Oncology	58	5	99.1
005	Med/Surg	29	2	98.2

Before you can analyze any of this data, you need to know what type each variable is — because the type of data determines which statistical tools are appropriate. Applying the wrong tool to the wrong type of data is one of the most common errors in nursing research.

Step 2 Let's Talk About It

Look at the five columns above. Which ones feel like they are in a different category from the others?

Patient ID and Unit seem different — they are labels or categories, not numbers you would average. But Age, Pain Score, and Temperature feel like numbers you could add up, find means for, and compare.

And yet, even among numbers, there are differences. The difference between a temperature of 98.6°F and 101.2°F is a real, meaningful difference in degrees. But is a Pain Score of 6 really *twice as much pain* as a Pain Score of 3? That is less clear.

These distinctions matter. Statistics has a careful vocabulary for them.

Step 3 Now We Name It

Definition: Qualitative vs. Quantitative Variables

A **qualitative variable** (also called a **categorical variable**) classifies individuals into groups or categories. These cannot be meaningfully added or averaged.

Examples: Blood type, nursing unit, insurance type, gender, shift assignment.

A **quantitative variable** is a numerical measurement where arithmetic operations (addition, averaging) make sense.

Examples: Age, temperature, heart rate, length of stay, number of medications.

Definition: Discrete vs. Continuous Variables

Among quantitative variables:

A **discrete variable** takes on countable values — usually whole numbers with no values possible in between.

Examples: Number of hospital admissions, number of medications administered, number of patients on a unit.

A **continuous variable** can take on any value within a range — including decimals. It is measured, not counted.

Examples: Blood pressure, temperature, weight, time (in hours or minutes).

Common Confusion: Is Pain Score Quantitative?

Pain scores (0–10) are technically **ordinal** — they rank pain levels, but the gap between 3 and 4 is not guaranteed to equal the gap between 7 and 8. In practice, healthcare researchers often treat them as quantitative for convenience. You will encounter both approaches. When in doubt, report medians and ranges for pain scores rather than means.

Step 4 Watch It Work

Returning to our dataset:

Variable	**Qualitative or Quantitative?**	**If Quantitative:**	**Reasoning**
Patient ID	Qualitative	—	A label; arithmetic is meani
Unit	Qualitative	—	Names a category
Age (years)	Quantitative	Discrete	Counted in whole years
Pain Score	Quantitative	Discrete	Counted values 0–10
Temperature (°F)	Quantitative	Continuous	Measured; 98.6, 98.61, etc. a

Why this matters clinically: If you want to know the "average unit," that question does not make sense — Unit is qualitative. But you can absolutely compute the average temperature or the average age of your patients.

Step 5 Your Turn

Classify each variable as **qualitative** or **quantitative**. If quantitative, state whether it is **discrete** or **continuous**.

1. Type of insurance coverage (Medicare, Medicaid, Private, Uninsured)
2. Number of falls reported on a unit during one month
3. Diastolic blood pressure reading (mmHg)
4. A nurse's years of experience
5. A patient's primary diagnosis (coded as ICD-10)
6. Number of IV lines currently in place
7. Body temperature recorded every four hours
8. Whether a patient was readmitted within 30 days (Yes / No)
9. Time (in minutes) between a patient pressing the call button and a nurse responding
10. Patient room number

Step 6 Think Like a Nurse

When a Number Is Not What It Seems

A hospital administrator distributes a report showing that the average patient acuity code on your unit last month was 2.8. The report uses this number to argue that your unit has a "moderate" patient load and that staffing levels do not need to increase.

Before accepting this conclusion, a statistically aware nurse asks: what type of variable is the acuity code, and is an average the right tool for summarizing it?

If acuity is coded 1 through 5 as a categorical label — with no guarantee that the distance from 2 to 3 equals the distance from 3 to 4 — then computing a mean of 2.8 is treating an ordinal (or even nominal) variable as if it were quantitative. The number looks precise, but the precision is an illusion.

Better questions to bring to that meeting: How many patients were coded at Level 4 or 5 (the highest acuity)? What percentage of shifts had a majority of high-acuity patients? A frequency breakdown tells a more honest story than a single average.

Classifying variables is not a formality. It is the first act of protection against misleading data summaries.

1.3 Levels of Measurement

Step 1 Read This First

You are reviewing patient records for a quality improvement project and come across the following measurements:

- **Room number:** 204, 207, 211
- **Satisfaction rating:** Poor, Fair, Good, Excellent
- **Temperature in Fahrenheit:** 97.8, 98.6, 101.2
- **Length of stay in days:** 0, 2, 4, 7

Someone on your team wants to compute averages and percentages for all four variables. Before you can agree or push back, you need to understand what each number actually means.

Step 2 Let's Talk About It

Room numbers look like numbers, but can you say Room 211 is "better" or "larger" in any meaningful sense? Can you say a patient in Room 211 waited 7 rooms longer than a patient in Room 204?

Satisfaction ratings have a natural order (Excellent is better than Poor), but is the distance between "Poor" and "Fair" the same as the distance between "Good" and "Excellent"?

Temperature in Fahrenheit has equal spacing between degrees, but does 0°F mean "no temperature"?

Length of stay in days has equal spacing *and* a true zero (0 days really means the patient was not admitted). That is the richest kind of data.

Statisticians classify measurement into four levels that capture these distinctions. The level of measurement determines which statistical operations are legitimate.

Step 3 Now We Name It

Definition: The Four Levels of Measurement

Nominal level: Data that are labels or names for categories. There is no natural order. Arithmetic operations are meaningless.

Examples: Blood type (A, B, AB, O), nursing unit (ICU, Med/Surg, Oncology), patient room number, insurance type.

Ordinal level: Data with a natural order, but the distances between categories are not uniform or meaningful.

Examples: Pain scale (mild, moderate, severe), satisfaction rating (Poor/Fair/-Good/Excellent), stage of cancer (Stage I, II, III, IV).

Interval level: Data with equal spacing between values, but no true zero point. A value of zero does not mean "none."

Examples: Temperature in Fahrenheit or Celsius (0°F does not mean no heat), calendar year (Year 0 does not mean no time).

Ratio level: Data with equal spacing between values *and* a true zero that means "none of the quantity." This is the most informative level.

Examples: Weight in pounds (0 lb = no weight), length of stay in days (0 days = not admitted), heart rate, medication dose in mg.

The Rule for Ratios

You can only make ratio statements ("twice as much," "three times as long") with **ratio-level** data. A patient with a heart rate of 120 bpm genuinely has twice the rate of a patient at 60 bpm. But you cannot say that 80°F is "twice as hot" as 40°F — temperature in Fahrenheit is interval, not ratio. (Temperature in Kelvin, with a true absolute zero, *is* ratio.)

Step 4 Watch It Work

Apply the four levels to common nursing data:

Variable	Level	Why?
Patient room number	Nominal	Labels; Room 210 is not "more" than Room 108
Primary diagnosis (ICD-10 code)	Nominal	Category labels; no meaningful order
Pain severity (mild/mod/severe)	Ordinal	Ordered, but gaps between levels are unequal
APGAR score (0–10)	Ordinal	Ranked, but equal spacing is not guaranteed
Temperature in Fahrenheit	Interval	Equal spacing, but 0°F $\neq$ no temperature
Temperature in Kelvin	Ratio	True zero exists (absolute zero = no heat)
Heart rate (bpm)	Ratio	0 bpm = no heartbeat (true zero)
Length of stay (days)	Ratio	0 days = not admitted (true zero)
Weight (kg)	Ratio	0 kg = no weight (true zero)

A diagnostic test example:
A clinical researcher codes patients' test results as: 1 = Negative, 2 = Inconclusive, 3 = Positive. A colleague suggests computing the mean of these codes and reporting it as the "average result."

Is this valid? No. The codes 1, 2, 3 are nominal labels even though they look like numbers. Computing a mean of 1.8 and calling it "average result" is statistically meaningless. The researcher should report counts and percentages instead.

Step 5 Your Turn

Identify the level of measurement for each variable. Be prepared to explain your reasoning.

1. A nurse records whether each patient's wound is: Healing Well, Stable, or Deteriorating.
2. The number of milligrams of morphine administered to a patient.
3. A patient's Social Security Number (used for identification purposes).
4. The Likert scale response to "How satisfied are you with your care?" (1 = Very Dissatisfied, 5 = Very Satisfied).
5. A patient's weight in kilograms.

6. The year a nurse graduated from nursing school.

7. A hospital assigns each nursing unit a numerical priority code (1 = highest priority, 5 = lowest) for equipment upgrades.

8. Time (in hours) from symptom onset to antibiotic administration in sepsis patients.

9. A researcher categorizes patients' insurance as: 1 = Private, 2 = Medicare, 3 = Medicaid, 4 = Uninsured.

Challenge problem:

10. A hospital rates nurses' clinical competency on a scale of 1 to 5, where 1 = Novice, 3 = Competent, and 5 = Expert. A manager computes the mean score for all nurses on her unit and gets 3.4. She concludes that the "average nurse is between Competent and Proficient."

What level of measurement is this scale? Is the manager's conclusion statistically appropriate? What would be a better way to summarize the data?

Step 6 Think Like a Nurse

The Coding Problem in Clinical Research

You are reading a published nursing study. In the methods section, the researchers write: “We coded patient acuity as 1 = Low, 2 = Moderate, 3 = High, and then computed the mean acuity score for each unit.”

Before accepting the study’s conclusions, ask: What level of measurement is acuity? Is “mean acuity” a meaningful statistic?

The honest answer is that acuity coded this way is *ordinal*. The mean of ordinal data can be misleading because it assumes equal spacing between categories — it assumes that the difference between Low and Moderate is identical to the difference between Moderate and High. That assumption is often untested and often wrong.

You will find this practice everywhere in healthcare research. It is not always wrong, but it is always worth noticing. A nurse who can spot this is a more critical reader of evidence.

1.4 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 1. Problems marked with ⋆ are more challenging.

Part A: Population, Sample, Parameter, Statistic

1. A hospital reviews the blood pressure readings of all 312 patients admitted during the month of March and calculates an average systolic pressure of 128 mmHg.
   (a) Is the group of 312 patients a population or a sample?
   (b) Is 128 mmHg a parameter or a statistic?
   (c) Is this descriptive or inferential statistics? Explain.

2. A researcher studying nurse burnout randomly selects 180 nurses from hospitals across three states. She finds that 62% of her respondents report feeling emotionally exhausted at least once a week. She concludes that burnout is prevalent among nurses nationwide.
   (a) What is the population of interest?
   (b) What is the sample?
   (c) Is the 62% figure a parameter or a statistic?
   (d) Is this descriptive or inferential statistics?

3. ⋆ A nursing school has 840 enrolled students. The dean surveys a randomly selected group of 120 students and finds that 74% are satisfied with the clinical placement process. She then reports to the board that "about three-quarters of our students are satisfied with clinical placements."
   (a) Identify the population and the sample.
   (b) The 74% from the survey — is this a parameter or a statistic?
   (c) What is the dean implicitly claiming about the population parameter?
   (d) Is there any risk in this inference? What would make the inference more reliable?

Part B: Variable Classification

4. Classify each variable as qualitative or quantitative. If quantitative, state discrete or continuous.
   (a) Shift type (Day, Evening, Night)
   (b) Number of patients a nurse cares for in a single shift
   (c) Oxygen saturation (SpO_2 as a percentage)
   (d) Whether a patient was discharged home, to a facility, or against medical advice
   (e) Serum creatinine level (mg/dL)
   (f) Number of pressure injuries reported per month on a unit

(g) A nurse's annual salary in dollars

(h) The method used to administer medication (oral, IV, subcutaneous, intramuscular)

5. ⋆ A hospital tracks "patient acuity" using a five-point scale where nurses rate each patient from 1 (requires minimal care) to 5 (requires constant monitoring).
(a) Is this variable qualitative or quantitative?
(b) Is it discrete or continuous?
(c) What level of measurement does this scale represent?
(d) A unit manager computes the average acuity score and finds it is 3.2. What are the limitations of this summary statistic for this type of variable?

Part C: Levels of Measurement

6. Identify the level of measurement. Justify your answer in one sentence.

(a) A patient's ABO blood type (A, B, AB, O)

(b) Number of hours a patient slept the previous night

(c) A nurse's shift satisfaction rated 1–5 (Strongly Dissatisfied to Strongly Satisfied)

(d) Systolic blood pressure in mmHg

(e) The temperature at which a medication must be stored (in degrees Celsius)

(f) A hospital's Star Rating (1 to 5 stars) from CMS

(g) The zip code of a patient's home address

(h) Time in hours from admission to first physician assessment

7. ⋆ A quality improvement coordinator assigns numerical codes to patient discharge destinations: 1 = Home, 2 = Skilled Nursing Facility, 3 = Rehabilitation Center, 4 = Hospice, 5 = Deceased.
(a) What level of measurement is this coding system?
(b) A data analyst computes the mean discharge destination code and gets 2.3. Interpret this number. Is it meaningful?
(c) What would be the appropriate way to summarize this data?
(d) Suppose the coordinator wants to know whether discharge destination is related to length of stay. What concerns should she raise about using this coded variable in a statistical analysis?

8. ⋆ **Research Application:** Read the following excerpt from a hypothetical study:

"We surveyed 95 nurses about their confidence level with electronic health record (EHR) documentation. Responses were coded as: 1 = Not Confident, 2 = Somewhat Confident, 3 = Confident, 4 = Very Confident. The mean confidence score was 2.87 (SD = 0.91)."
(a) What level of measurement is the confidence scale?

(b) What assumption must be true for the mean to be a meaningful summary of this data?

(c) If you were reviewing this study for a journal, what statistical concern would you raise?

(d) Suggest an alternative way to summarize the responses that does not require the assumption you identified in (b).

Answer Key — Selected Problems

Answer Key

Section 1.1 Practice — Your Turn

1. (a) Population (all 18 nurses on the floor). (b) Descriptive — reporting what was observed among those nurses.
2. (a) Sample (200 ICU patients from five hospitals, not all ICU patients). (b) Inferential — using the sample to draw conclusions about ICU patients nationwide.
3. (a) Population (all 1,200 discharged patients). (b) Descriptive — reporting what was observed in the complete records.
4. (a) Sample (500 California nurses). (b) Inferential — using California nurses to estimate a national figure.
5. (a) Population (all shift reports from the past month). (b) Descriptive — summarizing observed records.

Section 1.2 Practice — Your Turn

1. Qualitative (categories of coverage)
2. Quantitative, discrete (counted in whole numbers)
3. Quantitative, continuous (measured; decimals are possible)
4. Quantitative, discrete (whole years; can also argue continuous)
5. Qualitative (codes are category labels)
6. Quantitative, discrete (whole number count)
7. Quantitative, continuous (measured in degrees)
8. Qualitative (Yes/No categories)
9. Quantitative, continuous (measured time; fractions of minutes are possible)
10. Qualitative (room numbers are labels, not meaningful quantities)

Section 1.3 Practice — Your Turn

1. Ordinal (ordered categories: Healing Well > Stable > Deteriorating, but spacing is unequal)
2. Ratio (true zero exists: 0 mg = no medication administered)
3. Nominal (a label for identification; arithmetic is meaningless)
4. Ordinal (ordered 1–5, but equal spacing between satisfaction levels is not guaranteed)
5. Ratio (0 kg = no weight; true zero)
6. Interval (years are equally spaced, but Year 0 does not mean “no time”)
7. Ordinal (the codes 1–5 represent ordered priority, but the gaps between levels need

not be equal)

8. Ratio (0 hours = immediate administration; true zero)

9. Nominal (1, 2, 3, 4 are labels for insurance types; no meaningful order)

Problem 10 (Challenge): The scale is **ordinal**. The manager's conclusion is not statistically appropriate because ordinal data does not guarantee equal spacing between levels (the jump from Novice to Competent may not equal the jump from Competent to Expert). A better summary would be to report the **median** or the **frequency distribution** (e.g., "40% of nurses scored Competent or higher").

Chapter Practice — Selected

Problem 1: (a) Population — all 312 admitted patients were reviewed. (b) Parameter — computed from the entire population. (c) Descriptive — no generalization beyond those 312 patients.

Problem 2: (a) All nurses nationwide. (b) The 180 randomly selected nurses. (c) Statistic — from the sample only. (d) Inferential — using 180 nurses to make a claim about all nurses.

Problem 4e: Quantitative, continuous. Creatinine is measured and can take non-integer values.

Problem 7b (Challenge): The mean of 2.3 is not meaningful because the codes (1–5) represent nominal categories with no natural ordering. The number 2.3 cannot be interpreted as a destination between "Skilled Nursing Facility" and "Rehabilitation Center." The appropriate summary is a frequency table or bar chart showing how many patients fell into each category.

Section 1.1 — What Is Statistics?

- **Statistics** is the science of collecting, organizing, summarizing, and analyzing data to answer questions.
- **Descriptive statistics** summarizes observed data. **Inferential statistics** uses sample data to draw conclusions about a population.
- A **population** is the entire group of interest. A **sample** is the subset we actually observe.
- A **parameter** describes a population. A **statistic** describes a sample.
- Memory rule: **P**opulation → **P**arameter. **S**ample → **S**tatistic.

Section 1.2 — Types of Variables

- **Qualitative** variables classify into categories (blood type, unit, insurance type).
- **Quantitative** variables are numerical measurements where arithmetic makes sense.
- Quantitative variables are either **discrete** (counted: number of medications) or **continuous** (measured: temperature, weight).

Section 1.3 — Levels of Measurement

- **Nominal**: categories, no order (blood type, room number)
- **Ordinal**: ordered categories, unequal spacing (pain severity, satisfaction ratings)
- **Interval**: equal spacing, no true zero (temperature in °F, calendar year)
- **Ratio**: equal spacing + true zero (weight, length of stay, heart rate)
- Only ratio-level data supports statements like “twice as much” or “three times as long.”
- The level of measurement determines which statistical tools are appropriate.

The Nursing Connection

- Every chart, research article, and quality report you read contains statistical claims.
- Knowing whether a number describes a population or a sample — and whether the right statistical tools were applied — is a professional competency, not optional background knowledge.

CHAPTER 2

How We Know What We Know: Study Design

> *"A conclusion is only as trustworthy as the study behind it."*
>
> — a standard every nurse should apply before changing practice

In this chapter, you will learn to:

- Distinguish between observational studies and experiments
- Explain why only experiments can establish cause and effect
- Identify confounding variables and explain why they matter
- Recognize the three main experimental designs: completely randomized, randomized block, and matched pairs
- Explain the purpose of control groups, placebos, and blinding
- Critically evaluate study design in nursing research

2.1 Observational Studies vs. Experiments

Step 1 Read This First

Your unit has been piloting a new early mobility protocol for post-surgical patients. After three months, the charge nurse reports that patients on the protocol had an average length of stay of 3.8 days, compared to 5.1 days for patients on the standard protocol.

Before the hospital rolls out the protocol system-wide, a skeptical physician asks: "Were the patients in each group similar? Could the difference in stay be explained by something other than the protocol?"

Meanwhile, you are reading a nursing journal and find two studies on the same topic — nurse-to-patient ratios and medication error rates:

Study A: Researchers surveyed 40 hospitals and found that units with lower nurse-to-patient ratios had higher medication error rates.

Study B: Researchers randomly assigned 20 units to temporarily reduce their staffing ratios and 20 units to maintain current levels. Units with reduced ratios showed a 31% increase in errors.

Both studies found a relationship. But only one of them can tell you that the ratio *caused* the errors. Which one — and why?

Step 2 Let's Talk About It

Think about what is different between the two studies.

In Study A, the researchers simply *looked* at hospitals that already had different staffing levels. They did not change anything. The hospitals chose their own ratios for their own reasons — budget, union contracts, patient census, administrator preferences. The patients at low-ratio hospitals may have been sicker, older, or more complex. Any of those factors could explain the higher error rates.

In Study B, the researchers *randomly decided* which units got the reduced ratio. The random assignment means that, on average, the two groups of units started out similar in every way — patient complexity, nurse experience, unit culture. The only systematic difference between the groups was the ratio change. So if errors went up, the ratio is the most likely explanation.

This distinction — between watching and intervening — is the most important idea in research design. It determines whether a study can answer the question "Does X cause Y?" or only "Are X and Y related?"

Step 3 Now We Name It

Definition: Observational Study

In an **observational study**, the researcher observes and records data without intervening. No treatment is imposed. The researcher does not attempt to influence the subjects or change any conditions.

Examples: Surveying nurses about burnout levels; reviewing hospital records to compare infection rates across units; tracking patient outcomes without changing any care protocols.

Definition: Experiment

In an **experiment**, the researcher deliberately imposes a **treatment** on subjects and observes the effect. The researcher actively intervenes to change something.

Examples: Randomly assigning patients to receive a new wound care protocol or the standard protocol; testing whether a hydration reminder system reduces catheter-associated infections.

Definition: The Key Principle

Only experiments can establish cause and effect.

Observational studies can show that two variables are **associated** — that they tend to occur together — but association is not causation. The apparent relationship may be explained by a third variable that was not measured or controlled.

Definition: Confounding Variable

A **confounding variable** (also called a **lurking variable**) is a variable that is related to both the explanatory variable and the response variable, and that can distort or explain away the apparent relationship between them.

Confounding: The Most Common Error in Nursing Research

Confounding is everywhere in healthcare research. Patients who choose to exercise more are also more likely to follow other healthy behaviors. Hospitals with more resources tend to have both better nurse staffing *and* better patient outcomes. When you read a study that says "patients who did X had better outcomes," always ask: what else might be different about those patients?

Step 4 Watch It Work

Classifying Studies and Identifying Confounders

Scenario 1: A hospital notices that nurses who work night shifts have higher rates of self-reported medication errors. They conclude that night shifts cause more errors.

Classification: Observational study — no treatment was assigned.

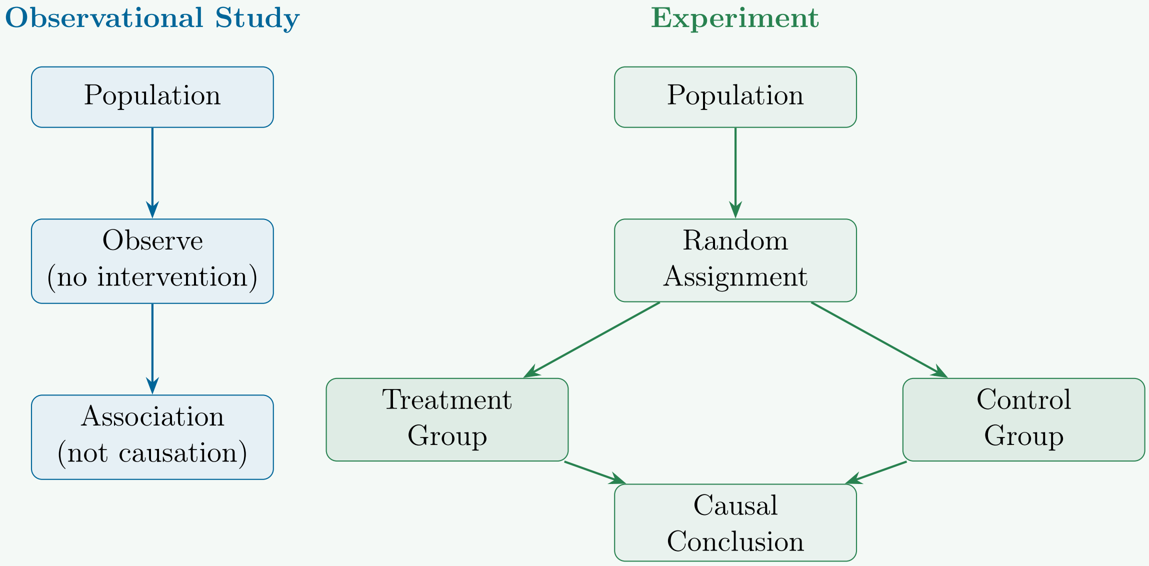
Possible confounder: Night-shift nurses may care for sicker or more unstable patients. Fatigue, lower staffing ratios at night, and less experienced staff may all contribute. The shift itself may not be the cause.

Scenario 2: A clinical researcher randomly assigns 60 post-operative patients to receive either standard pain management or a new multimodal approach. She measures pain scores and opioid use at 24 and 48 hours post-surgery.

Classification: Experiment — the treatment (pain protocol) was imposed by the researcher, and patients were randomly assigned.

Can it establish causation? Yes — because random assignment controls for confounders. Patients in both groups are, on average, similar in all other respects.

Diagram: Observational Study vs. Experiment



Step 5 Your Turn

For each scenario, state: (a) whether it is an observational study or an experiment, and (b) if observational, identify at least one possible confounding variable.

1. A research team reviews records from 500 hospitals and finds that hospitals with

Magnet nursing status have lower patient mortality rates than non-Magnet hospitals.

2. A nurse researcher randomly assigns patients with chronic back pain to either a structured stretching program or standard care, then compares pain scores after 8 weeks.

3. A hospital notices that patients who receive daily visits from a chaplain or counselor have shorter average lengths of stay than those who do not.

4. A pharmaceutical company recruits 200 hypertensive patients and randomly assigns half to receive a new ACE inhibitor and half to receive a placebo. Blood pressure is measured at 4, 8, and 12 weeks.

5. A public health researcher compares flu vaccination rates and flu hospitalization rates across counties and finds a negative association: counties with higher vaccination rates have fewer hospitalizations.

6. A hospital quality team compares fall rates before and after implementing a new hourly rounding protocol on one unit, without a control unit for comparison.

Step 6 Think Like a Nurse

Reading Research with a Critical Eye

A nursing journal publishes a study with the headline: *"Nurses who practice mindfulness meditation report 40% less burnout."* The hospital's wellness committee immediately proposes a mandatory mindfulness program for all staff.

Before you sign on, ask the design question: Was this an observational study or an experiment?

If nurses *chose* to practice mindfulness, the study is observational. Nurses who seek out mindfulness practices may already be more resilient, have more supportive home environments, work on less stressful units, or have greater access to mental health resources. The mindfulness is not the only difference between them and non-meditators. The 40% figure may reflect self-selection, not the effect of meditation itself.

To establish that mindfulness *causes* reduced burnout, you would need an experiment: randomly assign nurses to a mindfulness program or a control condition, measure burnout before and after, and compare.

The design of a study determines what its findings can and cannot tell you. A nurse who understands this is a far more effective advocate for evidence-based practice — and a far more useful voice when hospital administration wants to implement something based on weak evidence.

2.2 The Structure of a Well-Designed Experiment

Step 1 Read This First

Your hospital is testing a new early ambulation protocol for patients recovering from hip replacement surgery. The research team wants to know: does structured walking within 12 hours of surgery reduce time to discharge compared to standard care?

They have 120 patients enrolled and ready to begin. Now they face a series of decisions:

- How do we divide patients into groups — and how do we make the groups comparable?
- Should some patients receive a placebo? Can you even have a placebo for a walking protocol?
- Should the physical therapists assessing recovery know which protocol the patient received?
- What if older patients respond differently than younger ones? Should we account for that?

Each of these questions has a statistical answer. The answers together determine whether the study will produce results you can trust.

Step 2 Let's Talk About It

Think about the first question: how do you divide 120 patients into two groups?

If you let patients choose, some will pick the walking program (probably the more motivated ones) and some will choose standard care. Your groups will not be comparable.

If you assign the first 60 patients who enroll to the walking group and the next 60 to standard care, you may accidentally put all the younger or healthier patients in one group if they tend to schedule surgery at certain times.

The only way to guarantee that the groups are comparable — on average, in all the ways that matter — is to assign patients *randomly*. Random assignment is not a formality. It is the mechanism that allows you to attribute differences in outcomes to the treatment rather than to pre-existing differences between the groups.

Step 3 Now We Name It

Definition: Key Components of an Experiment

Experimental units (or **subjects** when human): the individuals being studied.

Treatment: the condition imposed on the experimental units. An experiment may have one treatment or several.

Control group: a group that receives either no treatment, the standard treatment, or a placebo. The control group provides a baseline for comparison.

Placebo: an inert or inactive treatment designed to look identical to the real treatment. Used to separate the psychological effect of receiving treatment from its actual physical effect.

Random assignment: using a chance mechanism to assign subjects to treatment or control groups. This is what makes the groups comparable at the start.

Definition: Blinding

Single-blind experiment: subjects do not know which treatment they received, but the researchers do.

Double-blind experiment: neither the subjects nor the researchers assessing outcomes know which treatment each subject received. This is the gold standard because it prevents both the placebo effect and researcher bias from distorting results.

Why Blinding Matters in Nursing Research

Nurses and patients are both susceptible to expectation effects. A nurse who knows a patient received the “new” protocol may unconsciously provide more encouragement or attention. A patient who knows they received the experimental treatment may report feeling better even before the treatment has had time to work. Double-blinding controls for both.

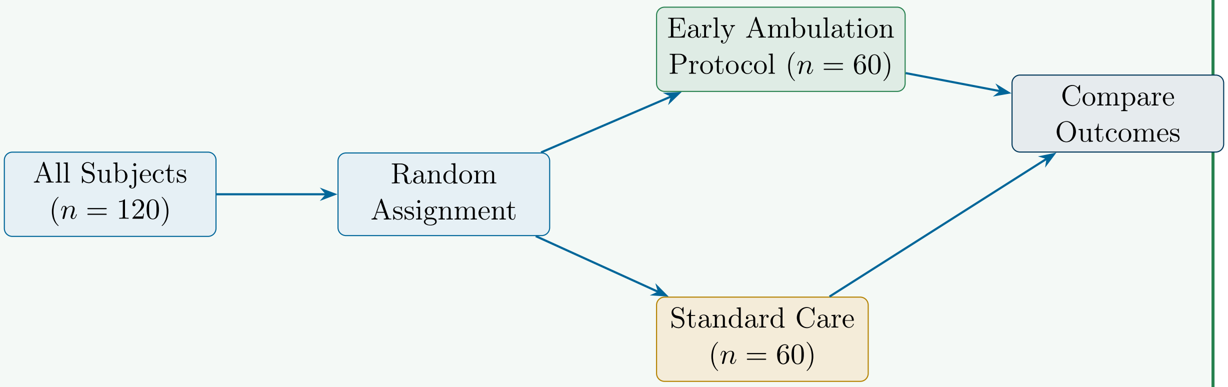
Step 4 Watch It Work

The Three Experimental Designs

The choice of experimental design depends on the research question, the characteristics of the subjects, and what sources of variability need to be controlled.

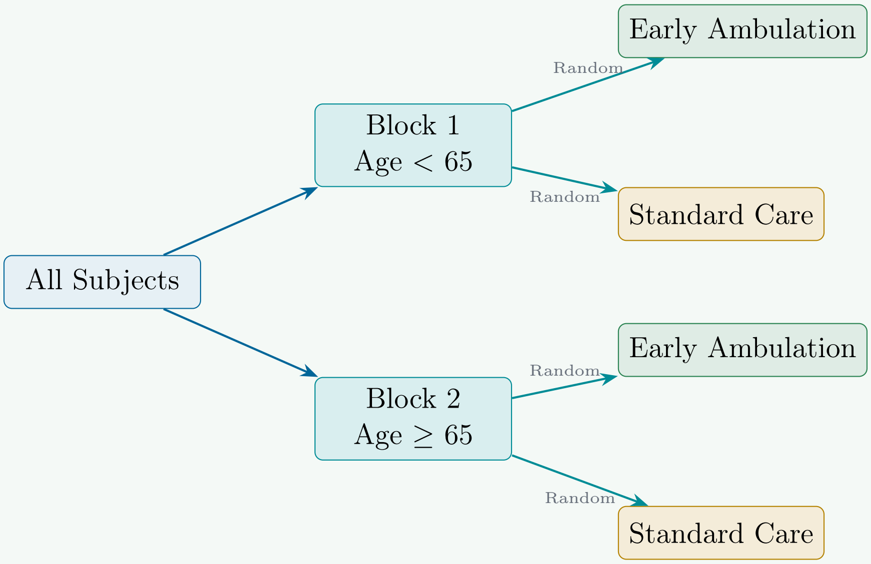
Design 1: Completely Randomized Design

All subjects are randomly assigned to one of the treatment groups. No blocking or pairing is used. This is the simplest design.



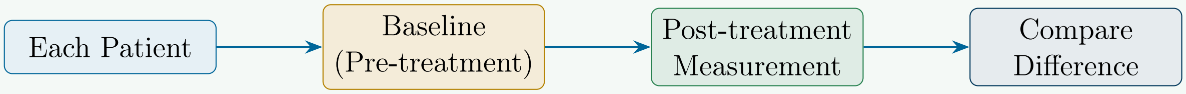
Design 2: Randomized Block Design

Subjects are first grouped into **blocks** based on a characteristic expected to affect the response (such as age or surgical risk). Then, within each block, subjects are randomly assigned to treatments. This controls for the blocking variable and increases precision.



Design 3: Matched Pairs Design

Each subject is compared to themselves (before vs. after) or matched with a very similar subject. This eliminates individual variability from the comparison.



Example: A patient's pain score is measured before and after implementing the new ambulation protocol. Each patient serves as their own control.

Step 5 Your Turn

1. Identify the experimental design (completely randomized, randomized block, or matched pairs) in each scenario.
   (a) A researcher wants to test whether a new hand hygiene reminder system reduces infection rates. She randomly selects 30 nursing units across the hospital system and assigns 15 to receive the reminders and 15 to continue standard protocols.

   (b) A study tests a new insulin delivery method. Patients are first grouped by their HbA1c level: well-controlled ($< 7\%$), moderately controlled (7–9%), and poorly controlled ($> 9\%$). Within each group, patients are randomly assigned to the new or standard delivery method.

   (c) A researcher measures nurses' perceived stress levels at the beginning and end of a 12-week mindfulness training program, comparing each nurse's own scores before and after.

   (d) Patients scheduled for cardiac surgery are matched in pairs based on age, sex, and ejection fraction. One patient from each pair is randomly assigned to an enhanced cardiac rehabilitation protocol; the other receives standard care.

2. A hospital plans to test a new sepsis early-warning alert system. The research team has 80 ICU patients enrolled.
   (a) Describe how you would conduct this study as a **completely randomized design**.
   (b) Explain why you might prefer a **randomized block design**, and identify an appropriate blocking variable.
   (c) Is a **matched pairs design** feasible here? Why or why not?
   (d) Should this study be blinded? If so, who should be blinded and why?

3. ⋆ A clinical trial is testing whether a new anticoagulant reduces stroke risk in patients with atrial fibrillation compared to the current standard. The trial is described as double-blind.

(a) Who is blinded in a double-blind trial?
(b) Why is blinding particularly important when the outcome measure is a patient-reported symptom (such as fatigue or dizziness)?
(c) The trial uses a placebo pill that looks identical to the new anticoagulant. What is the purpose of the placebo in this context?
(d) At the trial's midpoint, an independent safety board reviews the data. They find the new drug is dramatically more effective. Should the trial continue? What ethical considerations arise?

Step 6 Think Like a Nurse

Why "We Tried It and It Worked" Is Not Enough

A nurse manager introduces a new patient rounding checklist on her unit. Six months later, patient satisfaction scores have risen from 72% to 81%. She presents the results to hospital leadership as evidence that the checklist works, and recommends rolling it out hospital-wide.

The data looks compelling. But before the hospital invests in system-wide training and implementation, a statistically literate administrator asks: what kind of study was this?

This was not an experiment. There was no control group — no comparable unit that continued without the checklist. Satisfaction scores at many hospitals rose during that same period for other reasons: seasonal patterns, a reduction in patient census, improved staffing, a new cafeteria. Without a control group, there is no way to know how much of the improvement came from the checklist and how much would have happened anyway.

This is called a **pre-post design without a control**, and it is one of the most common — and most misleading — designs in quality improvement research.

A better design would have assigned some units to the checklist and kept others on standard practice, then compared the change in satisfaction scores across groups. That design would have isolated the effect of the checklist.

Before you advocate for a practice change, ask: compared to what? A finding without a comparison is a story, not evidence.

2.3 Types of Observational Studies

Step 1 Read This First

Not all research questions can or should be answered with an experiment. You cannot randomly assign people to smoke for 20 years to study lung cancer. You cannot randomly assign nurses to work 12-hour versus 8-hour shifts to study burnout — at least not easily or ethically. You cannot wait 10 years to see whether a dietary intervention prevents heart disease in a clinical trial.

For questions like these, researchers rely on carefully designed observational studies. Knowing the different types helps you understand what a study can and cannot tell you — and how much weight to give its conclusions.

Step 2 Let's Talk About It

Observational studies differ in one key dimension: *time*.

Some studies look backward — they start with an outcome (like a disease or an adverse event) and ask: what exposures preceded it? Others look forward — they start with an exposure (like a diet or a medication) and follow subjects over time to see what outcomes develop. Still others take a snapshot at a single point in time.

Each approach has strengths and weaknesses, and each is suited to certain types of research questions. You will see all three in nursing literature.

Step 3 Now We Name It

Definition: Cross-Sectional Study

A **cross-sectional study** collects data from a group of subjects at a **single point in time**. It is a snapshot — it can show associations but cannot establish time order or causation.

Nursing example: Surveying all nurses on a hospital's medical-surgical floors today to measure current burnout levels and current staffing ratios.

Definition: Case-Control Study

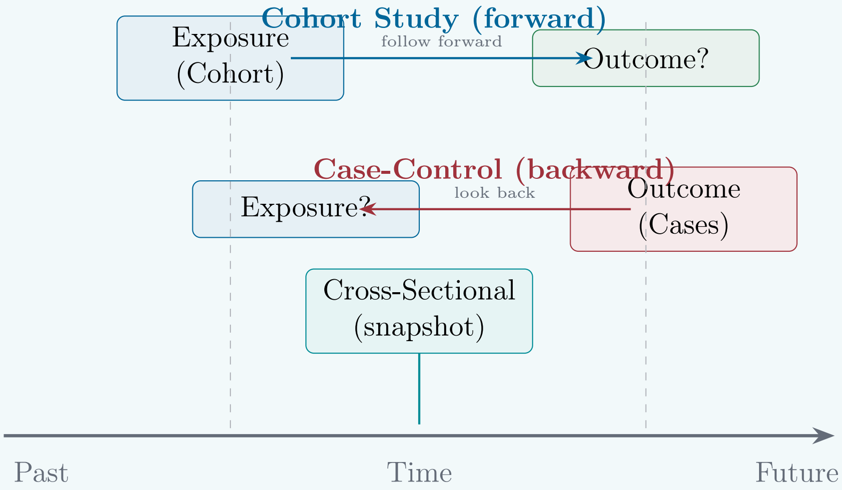
A **case-control study** starts with the **outcome** and looks **backward**. Researchers identify subjects who have the outcome of interest (cases) and subjects who do not (controls), then compare their past exposures.

Nursing example: Identifying 50 patients who developed a hospital-acquired infection (cases) and 50 patients who did not (controls), then examining prior antibiotic use, room assignments, and care procedures.

Definition: Cohort Study

A **cohort study** starts with the **exposure** and follows subjects **forward** in time to observe outcomes. It can establish time order (exposure preceded outcome) but not causation.

Nursing example: Following 300 newly hired nurses for two years, comparing burnout and turnover rates between those who work rotating shifts and those who work fixed shifts.



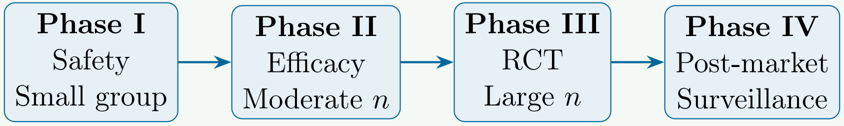
Step 4 Watch It Work

Identify the study type and evaluate its limitations.

Study Description	Type	Key Limitation
Researchers survey 800 hospital nurses today, measuring current shift length and current burnout scores.	Cross-sectional	Cannot determine which came first
A team identifies 40 nurses who left the profession within 2 years (cases) and 40 who stayed (controls), then reviews their original onboarding records.	Case-control	Relies on past records; recall bias
A hospital recruits 200 new graduate nurses and follows them for 3 years, comparing turnover between those mentored and those not mentored.	Cohort	Long time frame; loss to follow-up
Researchers select 100 patients who developed pressure injuries during hospitalization and 100 who did not, then compare their nursing care records.	Case-control	Cannot establish causation
A public health team surveys 2,000 nurses from across the country at one time point to assess current COVID-19 vaccination status and current health outcomes.	Cross-sectional	Temporal order unclear

Clinical Trial Phases: A special case of experiment

When a new drug or treatment is tested, the process follows regulated phases:



When you read "Phase III randomized controlled trial," you are reading about the gold standard of clinical evidence — a large, well-controlled experiment. Phase IV studies are observational, monitoring long-term safety after approval.

Step 5 Your Turn

1. Classify each study as cross-sectional, case-control, or cohort. Then identify one strength and one limitation of each.
   (a) Researchers follow 500 patients diagnosed with Type 2 diabetes for 5 years, comparing the rate of cardiovascular events between those who received

structured diabetes education and those who did not.

(b) A hospital surveys all staff nurses on a single day about their current work hours and their current level of compassion fatigue.

(c) A research team identifies 80 patients who developed delirium during hospitalization and 80 patients who did not, then reviews medical records for prior use of sleep aids, opioids, and anticholinergic medications.

2. ⋆ A study claims: "Nurses who eat breakfast daily have 35% lower rates of workplace errors." The study surveyed 1,200 nurses at a single point in time.

(a) What type of study is this?
(b) Can the researchers conclude that eating breakfast *causes* fewer errors? Why or why not?
(c) List two confounding variables that might explain the relationship.
(d) What type of study would be needed to establish causation?

3. ⋆ **Research Application:** The following description is from a hypothetical study abstract:

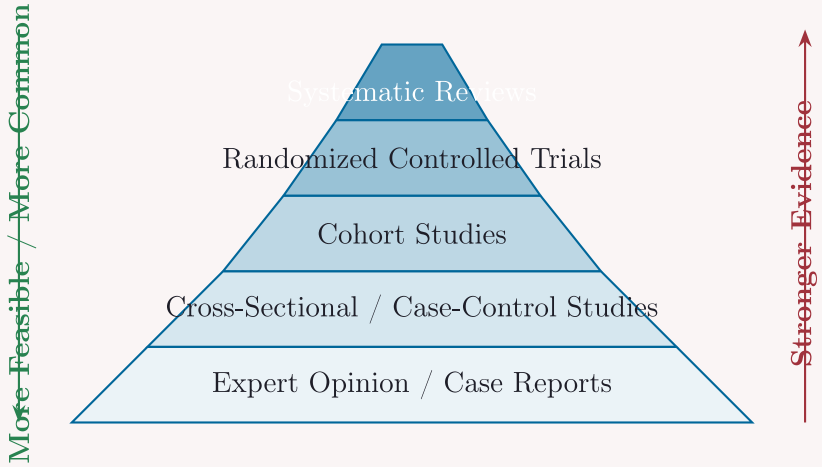
"We conducted a prospective cohort study of 348 registered nurses hired between January and June 2023. Participants completed validated burnout and intention-to-leave surveys at baseline, 6 months, and 12 months. We compared outcomes between nurses who participated in a formal mentorship program and those who did not."

(a) Is this an observational study or an experiment? Explain.
(b) What is the exposure variable? What is the outcome variable?
(c) The study is described as "prospective." What does this mean in the context of a cohort study?
(d) The researchers find that nurses in the mentorship program had significantly lower burnout at 12 months. Can they conclude that mentorship caused the reduction? What alternative explanations exist?

Step 6 Think Like a Nurse

The Hierarchy of Evidence

In evidence-based nursing practice, not all studies are created equal. The type of study design directly determines how much confidence you can place in its findings.



When a colleague hands you a study and says “the research shows,” your first question is: where does this study sit in the hierarchy? A cross-sectional survey and a randomized controlled trial are both “research” — but they are not equally persuasive. The higher up the pyramid, the more confidently you can apply the findings to your practice.

Evidence-based practice does not mean doing whatever the latest study says. It means knowing how much weight to give each study — and this chapter is where that skill begins.

2.4 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 2. Problems marked with ⋆ are more challenging.

Part A: Observational Studies vs. Experiments

1. For each scenario, state whether it is an observational study or an experiment. If it is an experiment, identify the treatment and the control condition. If it is observational, identify at least one possible confounding variable.

(a) A researcher asks nurses to self-report how many hours of sleep they get per night, then compares their scores on a cognitive performance test administered the next morning.

(b) A hospital randomly assigns patients admitted for pneumonia to receive either standard antibiotic therapy or a new combination therapy, then compares 30-day readmission rates.

(c) A study finds that patients in private rooms have lower rates of hospital-acquired infections than patients in shared rooms.

(d) A team randomly assigns ICU patients to receive either music therapy during wound care or silence, then measures pain scores on a validated scale.

(e) A nursing researcher examines whether nurses who completed a BSN degree have lower medication error rates than those with an ADN, using hospital records from the past three years.

2. A study reports: *"Patients treated at teaching hospitals have 12% lower 30-day mortality than patients treated at non-teaching hospitals."* The data come from Medicare claims records.

(a) Is this an observational study or an experiment?
(b) Can the researchers conclude that being treated at a teaching hospital *causes* lower mortality? Explain.
(c) Identify two confounding variables that could explain the association.
(d) What would need to be true for this finding to justify a policy of routing all complex patients to teaching hospitals?

Part B: Experimental Design

3. A hospital wants to test whether a new fall prevention checklist reduces fall rates in elderly inpatients. They have 90 patients aged 65 and older available for the study.

(a) Design this study as a **completely randomized experiment**. How would you assign patients to groups?
(b) A colleague suggests blocking by mobility level (independent, needs assis-

tance, or non-ambulatory). Redesign this as a **randomized block design**.
   (c) Could you use a **matched pairs design**? Describe how.
   (d) Should this study be blinded? Who could realistically be blinded, and who could not?

4. Identify the design of each experiment and explain your reasoning.
   (a) Blood glucose is measured in 40 diabetic patients before and after a 6-week dietary education program.
   (b) Thirty nursing units are randomly divided into two groups: 15 receive a new electronic documentation system, 15 continue with paper records.
   (c) Patients are grouped by diagnosis (cardiac, pulmonary, or neurological). Within each group, patients are randomly assigned to receive either standard discharge teaching or enhanced teach-back discharge teaching.
   (d) Twenty pairs of patients are matched by age, diagnosis, and length of stay. One patient in each pair receives acupuncture for post-operative nausea; the other receives standard antiemetics.

5. ⋆ A pharmaceutical company is testing a new antibiotic for hospital-acquired pneumonia. The trial is described as a *double-blind, placebo-controlled, randomized trial* with 300 patients across 15 hospitals.
   (a) Who is blinded and why?
   (b) What does "placebo-controlled" mean in the context of an antibiotic trial? Is a placebo ethical here?
   (c) Why is it important to recruit patients from 15 different hospitals rather than one?
   (d) The company analyzes results only for patients who completed the full course of antibiotics, excluding those who dropped out. What is the statistical concern with this approach?

Part C: Types of Observational Studies

6. Classify each study as cross-sectional, case-control, or cohort.
   (a) A research team identifies 60 nurses who left their jobs within 18 months of hire and 60 nurses who stayed, then reviews their pre-hire surveys for warning signs of burnout.
   (b) Researchers enroll 400 patients with newly diagnosed hypertension and follow them for 3 years, comparing cardiovascular events between those who adhere to their medication regimen and those who do not.
   (c) A hospital surveys all nurses on duty on a Tuesday morning about current sleep quality and current job satisfaction.
   (d) Researchers identify 100 patients who developed sepsis and 100 matched controls who did not, then compare the timing and type of antibiotic ad-

ministration in each group's prior care records.

7. ⋆ **Research Critique:** The following abstract is from a hypothetical study. Read it carefully and answer the questions.

   "We conducted a cross-sectional survey of 650 registered nurses working in acute care settings. Nurses who reported exercising at least 150 minutes per week had significantly lower burnout scores (mean 22.4) compared to those who exercised less (mean 31.7), $p < 0.001$. *We conclude that regular exercise is an effective intervention for preventing nurse burnout."*

   (a) What type of study is this?
   (b) Is the conclusion that exercise "prevents" burnout justified by this design? Explain.
   (c) Propose a study design that would allow a stronger causal conclusion.
   (d) The p-value is reported as $p < 0.001$. Without knowing what p-values mean yet, does a very small p-value change whether the causal conclusion is justified? Why or why not? (We will return to this question in Chapter 12.)

Answer Key — Selected Problems

Answer Key

Section 2.1 Practice — Your Turn

1. Observational. Possible confounders: Magnet hospitals tend to have better resources, higher nurse-to-patient ratios, and more experienced nursing staff. These factors may independently reduce mortality.
2. Experiment. Treatment: structured stretching program. Control: standard care. Random assignment allows causal inference.
3. Observational. Possible confounders: patients who receive chaplain visits may have stronger family support systems, longer planned stays (giving more time for visits), or less severe illness.
4. Experiment. Treatment: new ACE inhibitor. Control: placebo. Double-blind if neither patients nor outcome assessors know the assignment.
5. Observational (ecological). Possible confounders: counties with higher vaccination rates may have wealthier, healthier populations or better access to healthcare overall.
6. Observational (pre-post without control). No control group means observed changes cannot be attributed solely to the rounding protocol.

Section 2.2 Practice — Your Turn

1. (a) Completely randomized. (b) Randomized block — blocks by HbA1c level. (c) Matched pairs — each nurse serves as own control, before vs. after. (d) Matched pairs — patients paired on age, sex, and ejection fraction.
2. (a) Randomly assign 40 patients to the alert system and 40 to standard monitoring. (b) Block by APACHE score or admitting diagnosis, since severity likely affects outcomes. (c) Possible but difficult — ICU patients vary widely and finding true matches is challenging. (d) The outcome assessors should be blinded to group assignment; blinding patients is more difficult since the alert system is visible.
3. (a) Both patients and outcome assessors (physicians, nurses evaluating outcomes) are blinded. (b) A placebo pill ensures patients cannot tell which group they are in. The ethics are complex — a true placebo would mean some patients receive no antibiotic, which may be unethical for a serious infection; trials often compare new drug vs. current standard treatment instead. (c) Multi-site recruitment improves generalizability and reduces site-specific confounding. (d) Excluding dropouts creates *per-protocol bias* — patients who tolerate the full course may differ systematically from those who do not.

Section 2.3 Practice — Your Turn

1. (a) Cohort — begins with exposure, follows forward. Strength: establishes time

order. Limitation: long follow-up, costly, loss to follow-up. (b) Cross-sectional — single time point. Strength: quick, inexpensive. Limitation: cannot determine which came first. (c) Case-control — begins with outcome. Strength: efficient for rare outcomes. Limitation: relies on past records, subject to recall bias.

2. (a) Cross-sectional. (b) No — cross-sectional studies cannot establish causation. Nurses who eat breakfast may differ in many other health behaviors. (c) Overall health consciousness; socioeconomic status; working conditions (morning shift nurses are more likely to eat breakfast before work). (d) A randomized experiment assigning nurses to eat or skip breakfast.

Chapter Practice — Selected

Problem 1a: Observational. Possible confounders: nurses who sleep more may have better scheduling, fewer dependents at home, or work less demanding shifts.

Problem 1c: Observational. Confounders: private rooms may be associated with better insurance, higher socioeconomic status, less severe illness, or more attentive nursing ratios.

Problem 2: (a) Observational. (b) No — patients are not randomly assigned to hospital type. Sicker patients may self-select to teaching hospitals. (c) Illness severity; insurance type and socioeconomic status. (d) The groups would need to be comparable in all relevant ways except hospital type — virtually impossible without randomization.

Problem 5d (Challenge): Excluding dropouts creates selection bias. Patients who complete a full antibiotic course may be healthier, more adherent, or less severely ill than those who drop out. This makes the treatment appear more effective than it is for the full population. The correct approach is *intention-to-treat* analysis: analyze all patients in the group they were originally assigned to.

Chapter 2 Summary

Section 2.1 — Observational Studies vs. Experiments

- An **observational study** records data without imposing a treatment. It can show association but not causation.
- An **experiment** imposes a treatment and observes the effect. With random assignment, it can establish causation.
- A **confounding variable** is related to both the explanatory and response variables and can distort apparent relationships.
- **Association is not causation.** This is one of the most important principles in statistics.

Section 2.2 — Structure of a Well-Designed Experiment

- Key components: subjects, treatment, control group, placebo, random assignment.
- **Single-blind**: subjects do not know their group. **Double-blind**: neither subjects nor assessors know.
- **Completely randomized design**: all subjects randomly assigned with no blocking.
- **Randomized block design**: subjects grouped by a relevant characteristic first, then randomly assigned within blocks.
- **Matched pairs design**: subjects compared to themselves (before/after) or to a matched partner.

Section 2.3 — Types of Observational Studies

- **Cross-sectional**: data collected at one point in time; snapshot; cannot establish time order.
- **Case-control**: starts with outcome, looks backward at prior exposures; efficient for rare outcomes.
- **Cohort**: starts with exposure, follows forward to outcome; can establish time order but not causation.
- The **hierarchy of evidence** places randomized controlled trials above observational studies because random assignment controls for confounding.

The Nursing Connection

- Every practice change you implement or advocate for should be backed by evidence. Knowing what type of study produced that evidence tells you how much confidence to place in it.
- *Before changing practice, ask:* Was this an experiment or an observation? Were groups comparable? What confounders might explain the finding?

CHAPTER 3

Collecting Good Data: Sampling

> *"The sample is the window through which we see the population."*
>
> — how you choose that window determines everything you can see

In this chapter, you will learn to:

- Explain why sampling is necessary and what makes a sample trustworthy
- Describe and distinguish five sampling methods: simple random, stratified, systematic, cluster, and convenience
- Identify the sampling method used in a given study
- Recognize and explain four major sources of bias: sampling bias, nonresponse bias, response bias, and undercoverage
- Evaluate the quality of a sample used in a nursing research study

3.1 Why Sampling Matters

Step 1 Read This First

The state board of nursing wants to know how registered nurses across California feel about mandatory overtime policies. There are approximately 350,000 licensed RNs in California.

Option A: Survey all 350,000 nurses.

Option B: Randomly select 1,200 nurses, survey them carefully, and use their responses to draw conclusions about all 350,000.

Option A sounds more trustworthy — after all, more data is better, right? But consider the reality: contacting 350,000 nurses takes months and millions of dollars. Response rates fall. Data quality deteriorates when the collection process is rushed. By the time you finish, the policy landscape may have changed.

Option B, done well, can give you an answer that is just as accurate — and in far less time, at far less cost.

This is the power of sampling. But it only works if the sample is collected carefully. A poorly chosen sample of 1,200 nurses can produce results that are more misleading than no data at all.

Step 2 Let's Talk About It

Think about the last time someone cited a statistic about nurses or patients that surprised you. Did you wonder: who did they ask? How did they find those people? Could the people they asked be different from everyone else in some important way?

Those are exactly the right questions. A survey of nurses who attend a professional conference will not represent nurses who never attend conferences — probably the most burned-out, understaffed, and overworked ones. A patient satisfaction survey mailed to home addresses will miss patients who were discharged to skilled nursing facilities.

The fundamental challenge of sampling is this: we want to learn about a large group, but we can only observe a small piece of it. Whether that small piece accurately reflects the whole depends entirely on how we chose it.

Randomness is the key. When selection is left to chance, every member of the population has a fair shot at being included, and the resulting sample tends to reflect the diversity of the whole group.

Step 3 Now We Name It

Definition: Key Sampling Vocabulary

The **population** is the entire group we want to learn about.

The **sample** is the subset of the population we actually observe.

The **sampling frame** is the list or mechanism from which the sample is drawn. Ideally it includes every member of the population — but in practice it often misses some.

A **representative sample** mirrors the characteristics of the population. A good sample makes inferences from sample to population trustworthy.

Sampling error is the natural, unavoidable difference between a sample result and the true population value. It decreases as sample size increases.

Bias is a systematic error that consistently pushes results in one direction. Bias does not decrease with larger samples — it is a flaw in the method, not the size.

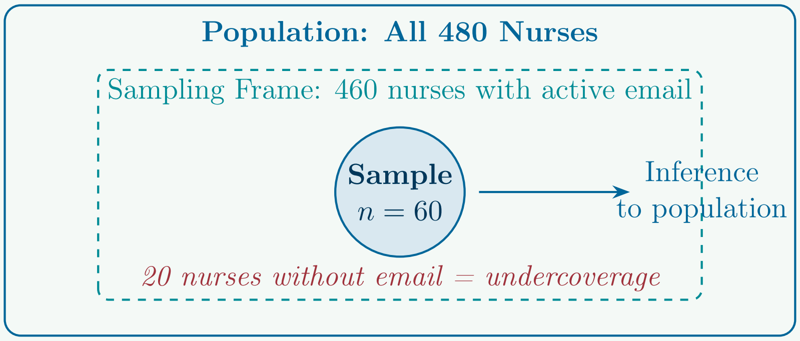
Bigger Is Not Always Better

A sample of 1,000 randomly selected nurses will give more reliable results than a sample of 10,000 self-selected volunteers. Size cannot fix a biased collection method. The infamous 1936 Literary Digest poll surveyed 2.4 million people and predicted the wrong winner of the presidential election — because the sample came from telephone directories and car registrations, systematically missing poorer voters. The lesson: method beats size every time.

Step 4 Watch It Work

Population vs. Sample Frame vs. Sample

A hospital wants to estimate the average job satisfaction score among all its nurses. There are 480 nurses employed across 12 units.



What to notice:

- The sampling frame (460 nurses with email) is smaller than the population (480 nurses). The 20 nurses without email addresses are excluded before sampling even begins — this is **undercoverage**.
- The sample of 60 is drawn from the frame. Inferences apply to the frame, and by extension to the population — but the 20 excluded nurses may systematically differ (perhaps older, part-time, or in lower-tech roles).
- If job satisfaction is related to tech access, the sample will overestimate satisfaction for the full workforce.

Step 5 Your Turn

1. A researcher wants to study nurse burnout among all nurses working in public hospitals in Los Angeles County. She obtains a list of nurses from the three largest public hospitals and draws her sample from that list.
   (a) What is the population of interest?
   (b) What is the sampling frame?
   (c) Are the sampling frame and population the same? If not, who is excluded?
   (d) Is there a risk of undercoverage? Explain.
2. A hospital surveys 200 patients about their care experience. The surveys are distributed by nurses at discharge. Response rate is 40%.
   (a) What is the population?
   (b) What is the potential source of bias in how surveys are distributed?
   (c) The 40% response rate means 120 patients did not respond. What type of bias does this introduce?
   (d) Would increasing the survey to 500 patients fix the bias issues you identified?

Explain.

3. True or False: A sample of 5,000 people is always more accurate than a sample of 500. Explain your answer.

Step 6 Think Like a Nurse

Who Is Missing from the Data?

A hospital publishes a report showing that 91% of patients rated their care as "Good" or "Excellent" last quarter. Administration uses this to argue that no staffing changes are needed.

Before accepting this conclusion, ask: who filled out the survey, and who did not?

Patients who died during admission did not complete a discharge survey. Patients who were discharged to ICU step-down or transferred to other facilities may have been missed. Patients who do not speak English may have been unable to complete the survey. Patients with the worst experiences may have left before staff had time to distribute surveys.

The 91% figure is not wrong — it accurately describes the patients who completed the survey. The problem is using it to describe *all* patients. The sample is almost certainly skewed toward patients who had smoother, less complicated stays.

When you see a satisfaction statistic, your first question is not "what does the number say?" It is "who was able to answer?" The missing voices often tell the most important story.

3.2 Five Sampling Methods

Step 1 Read This First

A nurse researcher at a large hospital system wants to study the relationship between shift length and compassion fatigue. The system employs 1,800 nurses across six hospitals. She has a budget that allows her to survey 300 nurses.

She has several options for selecting those 300 nurses. Some will give her a trustworthy, representative sample. Others will introduce systematic bias that no amount of analysis can fix. Her choice of method will determine how much she can actually conclude.

Here are five approaches she might consider. By the end of this section, you will know exactly which ones to trust — and which ones to question.

Step 2 Let's Talk About It

Before we name the methods, think about what you would do.

You could survey the nurses who happen to be in the break room when you visit — quick and easy, but are break-room nurses representative of everyone? Probably not.

You could assign every nurse a number and use a random number generator to pick 300 — fair to everyone, but what if night-shift nurses are never on the list because you only visited during the day?

What if you know that day-shift and night-shift nurses have very different experiences? Maybe you should make sure both groups are represented, rather than hoping they appear in the right proportion by chance.

Each of these instincts corresponds to a formal sampling method. Some are random. Some are not. The random ones can go wrong in different ways. Knowing the differences is what separates a critical reader of research from one who simply accepts the findings.

Step 3 Now We Name It

Definition: Method 1: Simple Random Sample (SRS)

Every individual in the population has an **equal and independent chance** of being selected. Selection is done using a random mechanism — a random number generator, a lottery, or a random number table.

How: Assign each of the 1,800 nurses a number from 1 to 1,800. Use a random number generator to select 300 numbers. Those nurses are your sample.

Strength: Eliminates selection bias. Every nurse has an equal chance.
Weakness: Requires a complete list of the population. By chance, some subgroups may be over- or under-represented.

Definition: Method 2: Stratified Random Sample

The population is divided into non-overlapping groups called **strata** based on a shared characteristic. Then a random sample is drawn from **each stratum**.

How: Divide the 1,800 nurses by shift (day, evening, night). Randomly select 100 from each shift group (proportional to shift size if needed).

Strength: Guarantees representation of every stratum. More precise than SRS when strata differ meaningfully.
Weakness: Requires knowledge of stratum membership for the entire population.

Definition: Method 3: Systematic Random Sample

Subjects are selected at **every kth interval** from an ordered list, where $k =$ population size ÷ sample size. A random starting point is chosen between 1 and k.

How: $k = 1800 \div 300 = 6$. Randomly select a starting point between 1 and 6 (say, 3). Then select nurses 3, 9, 15, 21, . . .

Strength: Simple and efficient when a list exists. Distributes selection evenly across the list.
Weakness: If the list has a hidden repeating pattern (e.g., every 6th nurse is always a charge nurse), the sample may be biased.

Definition: Method 4: Cluster Sample

The population is divided into **clusters** (often natural groupings like units, floors, or hospitals). A random sample of **clusters** is selected, and **all individuals** within selected clusters are surveyed.

How: Divide the 1,800 nurses into their 6 hospitals. Randomly select 3 hospitals. Survey all nurses at those 3 hospitals.

Strength: Cost-effective when the population is geographically spread out.
Weakness: Individuals within a cluster tend to be similar to each other, reducing diversity. Results are less precise than SRS of the same size.

Definition: Method 5: Convenience Sample

Subjects are selected because they are **easy to reach** — not through any random mechanism. Also called a **voluntary response sample** when subjects self-select.

How: Survey nurses who happen to be in the break room, or post a link on a nursing Facebook group and survey whoever clicks.

Strength: Fast and cheap.
Weakness: Almost always biased. Self-selected samples over-represent people with strong opinions. Break-room samples miss night-shift nurses, charge nurses, and those on leave.

Stratified vs. Cluster — The Most Tested Confusion

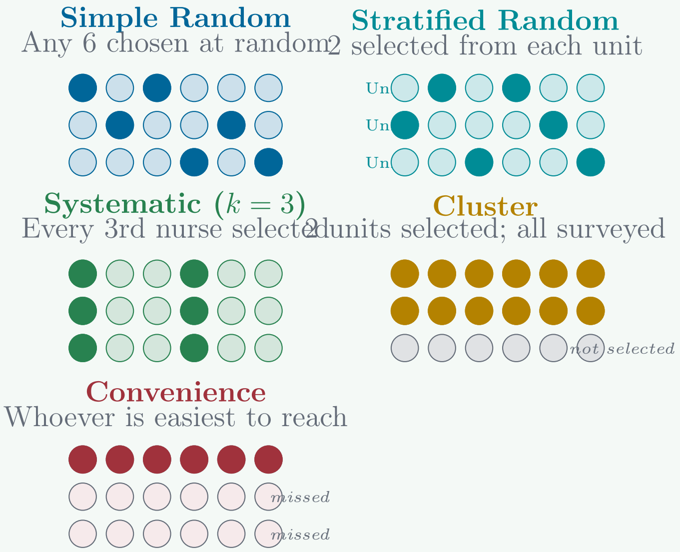
Stratified: divide into groups, randomly sample *from each* group. All groups are represented.
Cluster: divide into groups, randomly select *some* groups, survey *everyone* in those groups.

Memory trick: **S**tratified = **S**ome from each. **C**luster = **C**omplete groups selected.

Step 4 Watch It Work

Visual Comparison of the Five Methods

Imagine a population of 18 nurses represented as dots, organized into 3 units (rows) of 6 nurses each. We want to select 6 nurses.



Key observation: Only the simple random, stratified, and systematic methods guarantee that selection is governed by chance. Cluster introduces dependence within groups. Convenience is not random at all and should not be used for inference.

Step 5 Your Turn

1. Identify the sampling method used in each scenario.

(a) A hospital assigns each of its 600 nurses an ID number, then uses a random number table to select 60 nurses for a workplace safety survey.

(b) A researcher divides nurses into three groups by experience level: new graduates (0–2 years), mid-career (3–10 years), and experienced (>10 years). She randomly selects 40 nurses from each group.

(c) A state health department numbers all licensed nurses 1 through 45,000. They randomly select a starting number of 15, then survey every 150th nurse: 15, 165, 315, 465, . . .

(d) A large hospital system has 24 nursing units. Researchers randomly select 6 units and survey every nurse working on those units during the study week.

(e) A nursing school posts a link to a burnout survey on its alumni Facebook page and analyzes responses from the 312 alumni who click and complete it.

   (f) A quality improvement coordinator wants to assess medication safety. She stands at the nurses' station on the 7 AM to 3 PM shift every Monday and surveys the first 30 nurses she encounters.

2. For each of the following research questions, recommend the most appropriate sampling method and explain why.
   (a) A researcher wants to compare burnout levels across three types of nursing units: ICU, medical-surgical, and outpatient. She wants to make sure all three unit types are represented.

   (b) A hospital needs a quick estimate of how many nurses have completed their annual flu vaccination. Complete rosters are available.

   (c) A public health department wants to assess vaccination attitudes among nurses across 50 rural counties. Traveling to each county individually is too expensive.

3. ⋆ A nursing journal publishes a study on needle-stick injury rates. The sample consists of 800 nurses who responded to a survey posted on a major nursing online forum. The researchers conclude that 43% of nurses have experienced at least one needle-stick injury in the past year.
   (a) What sampling method was used?
   (b) Who is likely over-represented in this sample?
   (c) Who is likely under-represented?
   (d) Can the researchers validly conclude that 43% of *all* nurses have experienced needle-stick injuries? Why or why not?

Step 6 Think Like a Nurse

The Method You Choose Is the Conclusion You Can Draw
A hospital survey finds that 88% of nurses feel "supported by management." The survey was distributed by unit managers, who handed it to nurses before shift huddle. Completed surveys were collected by the same managers.

Think carefully about the sampling and data collection process here. Who selects the surveys? The managers. Who collects them? The managers. Who can see whether a nurse completed a survey? Quite possibly, the managers.

This is not a random sample of nurse opinions. Nurses who feel unsupported may be reluctant to say so when their manager is distributing and collecting the form. The 88% figure likely reflects a *social desirability bias* layered on top of a *convenience sample*. Neither the sampling method nor the data collection process protects against distortion.

A better design: distribute anonymous surveys via an independent department (HR, or an external vendor), collect them in sealed boxes that managers cannot access, and allow nurses to complete them privately.

Sampling method and data collection method are inseparable. A random sample undermined by a biased collection process is no longer a random sample in any meaningful sense.

3.3 Sources of Bias in Sampling

Step 1 Read This First

A hospital administrator presents the following findings at a board meeting:

- *"Our patient satisfaction survey shows 94% satisfaction." (Survey distributed only to patients discharged home; patients transferred to other facilities were excluded.)*
- *"Only 28% of nurses responded to our burnout questionnaire." (Of those, 71% reported high burnout.)*
- *"Nurses report an average of 1.2 medication errors per month." (Self-reported on anonymous forms — but nurses know the data may be used for performance reviews.)*
- *"The survey was sent to all 410 nurses on our roster." (The roster has not been updated in two years and includes 45 nurses who have since left.)*

Each of these findings has a bias problem buried in the methodology. By the end of this section you will be able to name each one — and explain why it matters.

Step 2 Let's Talk About It

Bias in data collection is not usually intentional. It creeps in through the small decisions that seem harmless at the time: where surveys are distributed, who gets to opt out, how questions are phrased, which patients are easy to reach.

What makes bias dangerous is that it is *systematic* — it does not average out the way random sampling error does. If the same type of person is consistently missed or consistently answers differently, no amount of data can fix it. The distortion compounds as the sample grows.

Understanding where bias comes from is the first step to detecting it in research you read — and avoiding it in research you conduct.

Step 3 Now We Name It

Definition: Sampling Bias (Selection Bias)

Sampling bias occurs when the method of selection systematically favors certain members of the population over others. The sample is not representative because the selection process was not random.

Example: Surveying only nurses who attend a voluntary wellness workshop to study overall nurse stress levels. Nurses who attend wellness events may be less stressed than those who do not attend.

Definition: Nonresponse Bias

Nonresponse bias occurs when individuals who do not respond to a survey differ systematically from those who do. Low response rates are a warning sign.

Example: A burnout survey sent to 500 nurses yields 140 responses (28%). If nurses with high burnout are too exhausted to complete surveys, the respondents will underrepresent the most burned-out group — and the true burnout rate will be underestimated.

Definition: Response Bias

Response bias occurs when respondents do not answer truthfully or accurately, often due to the phrasing of questions, the presence of an authority figure, or social desirability pressure.

Example: Nurses self-reporting the number of hand hygiene lapses per shift are likely to underreport — knowing the "right" answer and feeling social pressure to appear compliant.

Definition: Undercoverage

Undercoverage occurs when some members of the population are excluded from the sampling frame entirely, before any selection takes place.

Example: A satisfaction survey is emailed only to patients who provided an email address at registration. Elderly patients, non-English speakers, and lower-income patients are systematically less likely to have provided email addresses — their voices are absent from the data.

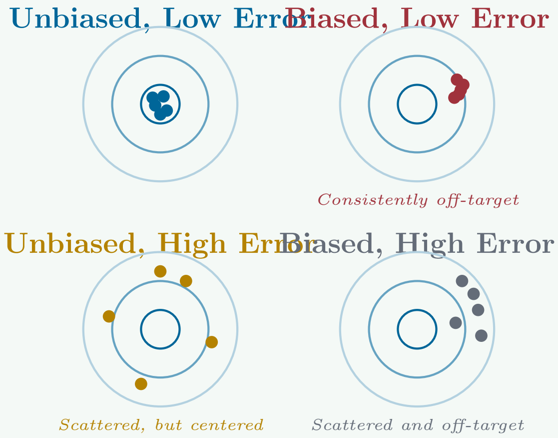
Step 4 Watch It Work

Diagnosing bias in the administrator's report

Return to the opening scenario. Each finding has a specific bias:

Finding and Method	Bias Type	Direction of Distortion
94% satisfaction — only home-discharged patients surveyed	Undercoverage	Overestimates satisfaction (sicker, more complex patients excluded)
28% response to burnout survey; 71% of respondents report high burnout	Nonresponse bias	Likely *underestimates* burnout (most burned-out nurses may not respond) or *overestimates* (only those motivated by high burnout respond)
Self-reported error rates on forms linked to performance reviews	Response bias	Underestimates true error rates (social desirability and fear of consequences)
Roster includes 45 departed nurses	Undercoverage / Outdated frame	Distorts denominators; surveys returned undeliverable

Visual: How Bias Differs from Sampling Error



Bias shifts all results in one direction. Larger samples reduce sampling error (scatter) *but do* ***not*** *fix bias (the off-center cluster stays off-center).*

Step 5 Your Turn

1. Identify the type of bias (sampling bias, nonresponse bias, response bias, or undercoverage) in each scenario. Explain how it could distort the results.
   (a) A hospital mails satisfaction surveys to patients after discharge. Only 22% of surveys are returned. The hospital reports an 89% satisfaction rate based on returned surveys.
   (b) A nurse researcher asks ICU nurses whether they have ever fallen asleep during a shift. The survey is administered face-to-face by the unit manager.
   (c) A study of medication errors uses self-reported data collected from incident report forms that nurses voluntarily submit.
   (d) A hospital surveys patients about their pain management using a questionnaire written only in English, in a community where 35% of patients speak primarily Spanish.
   (e) A nursing school studies student satisfaction by surveying students who attend an end-of-year banquet. Students who withdrew or failed out during the year are not surveyed.

2. ⋆ A pharmaceutical company surveys 2,000 patients who were prescribed their new blood pressure medication and asks: *"Have you experienced the following side effects: headache, dizziness, nausea, or fatigue?"* They report that 82% of patients experienced at least one side effect.
   (a) Is this a random sample? What type of bias might affect who is in the sample?
   (b) The question lists specific side effects. How might this influence the responses?
   (c) What type of bias does this create?
   (d) How should the question be reworded to reduce response bias?
   (e) The company claims this proves the medication causes side effects in 82% of users. Is this justified?

3. ⋆ **Research Application:** A study reports that 67% of nurses in the United States are satisfied with their current position. The data come from an online survey advertised on a nursing job board website. $n = 14,842$.
   (a) What type of sampling was used?
   (b) Who is likely to visit a nursing job board website? How does this affect the sample?
   (c) The sample size is nearly 15,000. Does this large sample size make the 67%

figure reliable? Explain.
(d) What would a better sampling method look like for this question?

Step 6 Think Like a Nurse

The Survey Behind the Statistic

The next time you read a statistic in a policy brief, a nursing journal, or a hospital report, run through this mental checklist before accepting the number:

1. Who is the population of interest?

2. What was the sampling frame? Who was excluded?

3. How were subjects selected? Was it random?

4. What was the response rate? Who did not respond?

5. How were questions phrased? Could they lead responses?

6. Is the conclusion proportionate to the study design?

If you cannot answer questions 1 through 4 from the methods section of a study, the study's findings should be held at arm's length. Not dismissed — but not adopted as policy either.

Statistics is not about being paralyzed by uncertainty. It is about calibrating how much confidence each piece of evidence deserves. A nurse who asks these six questions is a better advocate for patients than one who simply accepts any number with a decimal point in it.

3.4 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 3. Problems marked with ⋆ are more challenging.

Part A: Sampling Concepts

1. Define each term and give a nursing example of each.

   (a) Population

   (b) Sampling frame

   (c) Representative sample

   (d) Sampling error

   (e) Bias

2. A hospital system employs nurses across four campuses. A researcher wants to estimate average weekly overtime hours. She has access to the complete payroll database for all nurses.

   (a) What is the population?
   (b) What is the sampling frame?
   (c) Are the population and the sampling frame the same in this case?
   (d) Would it be appropriate to use a convenience sample here? Why or why not?

Part B: Sampling Methods

3. Identify the sampling method in each scenario. Then state one advantage and one disadvantage of the method used.

   (a) A researcher numbers all 920 nurses at a hospital 1–920 and uses a random number generator to select 80 nurses for a wellness survey.

   (b) A public health team divides the state into 15 health districts and randomly selects 4 districts. All nurses working in those districts are invited to participate.

   (c) A researcher groups nurses by specialty area (ICU, ED, Med/Surg, Oncology, Pediatrics) and randomly selects 25 nurses from each group.

   (d) A hospital manager surveys all nurses who attend the monthly staff meeting.

   (e) A researcher selects every 12th name from an alphabetical list of 2,400 nurses, starting with the 7th name.

4. ⋆ A state nursing association wants to survey nurses about their opinions on a proposed change to mandatory nurse-to-patient ratio legislation. They have a membership list of 18,500 nurses. They want a sample of 500.

(a) Design this study using a **simple random sample**. Describe the steps.
(b) Redesign using a **stratified sample**, stratifying by specialty area. What information would you need?
(c) Redesign using a **systematic sample**. What is k? What are the steps?
(d) The association decides to post a survey link in its monthly email newsletter instead. What sampling method does this become? What are the concerns?

Part C: Bias

5. For each scenario, identify the type of bias and explain how it likely distorts the results.

(a) A hospital reports that 95% of nursing staff are satisfied with their benefits package, based on surveys returned by 180 of the 600 nurses who received them.

(b) A researcher studying infection rates excludes patients who were transferred from other facilities because their prior care records are incomplete.

(c) A nurse satisfaction survey asks: *"Do you agree that the hospital provides excellent working conditions?"* with response options: Strongly Agree, Agree, Somewhat Agree.

(d) A study of patient outcomes uses a database of patients who were admitted to hospitals with electronic health records. Rural hospitals without EHR systems are not included.

6. ⋆ **Research Critique:** Read the following study description carefully.

"We assessed the prevalence of compassion fatigue among emergency department nurses. Surveys were distributed by charge nurses at the beginning of each shift over a two-week period. Of the 340 surveys distributed, 197 were completed and returned (58% response rate). Results show that 44% of respondents meet the clinical threshold for compassion fatigue."

(a) Is this a random sample? What type of sampling is this closest to?
(b) What is the response rate? What does a 42% non-response mean for the validity of the findings?
(c) Identify at least two specific sources of bias in this study.
(d) The researchers conclude that "44% of ED nurses experience compassion fatigue." Is this conclusion justified? What would a more careful statement of the finding look like?
(e) Propose one methodological change that would improve each bias you identified.

Answer Key — Selected Problems

Answer Key

Section 3.1 Practice — Your Turn

1. (a) All nurses working in public hospitals in LA County. (b) The nurse lists from the three largest public hospitals. (c) Not the same — nurses at smaller public hospitals are excluded. (d) Yes — nurses at smaller hospitals may have different working conditions and burnout levels.
2. (a) All patients at that hospital. (b) Nurse distribution creates potential social desirability effects; patients may feel they should respond positively to their nurse's face. (c) Nonresponse bias — the 120 non-responders may differ systematically. (d) No — increasing size does not fix a biased collection method.
3. False. A larger sample reduces sampling error but not bias. A biased sample of 5,000 will give a more consistently wrong answer than an unbiased sample of 500.

Section 3.2 Practice — Your Turn

1. (a) Simple random sample. (b) Stratified random sample. (c) Systematic random sample. (d) Cluster sample. (e) Convenience / voluntary response sample. (f) Convenience sample.
2. (a) Stratified — ensures all three unit types are represented, which is the research goal. (b) Systematic — quick and efficient when a complete roster exists. (c) Cluster — surveying entire counties avoids traveling to all 50; cost-effective for geographically spread populations.
3. (a) Voluntary response / convenience sample. (b) Nurses with strong opinions about needle-stick injuries — either those who have experienced them and want to report, or advocates for safety reform. (c) Nurses who do not use the forum, those who have not had injuries and feel less compelled to respond. (d) No — voluntary response samples cannot be used to make valid inferences about all nurses.

Section 3.3 Practice — Your Turn

1. (a) Nonresponse bias — 78% did not respond; if dissatisfied patients are less likely to return surveys, satisfaction is overestimated. (b) Response bias (social desirability) — nurses may underreport sleep events to avoid judgment from their manager. (c) Undercoverage combined with response bias — voluntary incident reports systematically undercount errors because not all errors are reported. (d) Undercoverage — non-English-speaking patients are excluded from the sampling frame. (e) Sampling bias — students who attend the banquet are likely more satisfied; those who left the program are never surveyed.
2. (a) Not random — patients prescribed the medication are not a random sample of all patients. Sampling bias. (b) Listing specific symptoms primes respondents

to think about them; they may report symptoms they would not have thought to mention otherwise. (c) Response bias (question wording / priming effect). (d) Reword as: “Please describe any changes in how you have felt since starting this medication.” (e) No — this is an observational sample without a comparison group. Side effects could be pre-existing.

Chapter Practice — Selected

Problem 3a: Simple random sample. Advantage: every nurse has equal chance; eliminates selection bias. Disadvantage: by chance, some specialty areas may be over- or under-represented.

Problem 3d: Convenience sample. Advantage: easy and fast. Disadvantage: nurses who attend staff meetings may be more engaged or satisfied than those who do not; results are not generalizable.

Problem 7 (Challenge): (a) Convenience/cluster (distributed at shifts, not randomly selected). (b) A 42% non-response is substantial; nurses not completing the survey may be those most affected by compassion fatigue or those least affected — both distort results in opposite directions. (c) Nonresponse bias (58% response); response bias (distributed by charge nurse, a supervisor figure). (d) Not fully justified — a more careful statement: “Among ED nurses who completed our survey, 44% met the threshold for compassion fatigue.” (e) Use anonymous drop boxes not accessible to supervisors; follow up with non-responders using a second distribution method.

Chapter 3 Summary

Section 3.1 — Why Sampling Matters

- A **representative sample** mirrors the population. Random selection is the mechanism that makes this happen.
- **Sampling error** is random and decreases with larger samples. **Bias** is systematic and does not decrease with larger samples.
- The **sampling frame** is the list from which the sample is drawn. If it excludes part of the population, **undercoverage** results.
- Bigger samples cannot fix a biased collection method.

Section 3.2 — Five Sampling Methods

- **Simple random sample (SRS):** every individual has an equal chance; requires a complete population list.
- **Stratified:** divide into strata, randomly sample from *each* — guarantees representation of all strata.
- **Systematic:** select every kth individual from an ordered list with a random start.
- **Cluster:** randomly select entire groups; survey everyone in those groups — efficient but less precise.
- **Convenience:** not random; produces biased results; should not be used for inference.
- Key distinction: Stratified samples *from* all groups. Cluster selects *some* groups entirely.

Section 3.3 — Sources of Bias

- **Sampling bias:** non-random selection systematically favors certain members.
- **Nonresponse bias:** non-respondents differ from respondents in ways that matter.
- **Response bias:** respondents answer inaccurately due to question wording, social pressure, or authority figures.
- **Undercoverage:** some members of the population are missing from the sampling frame entirely.

The Nursing Connection

- Before accepting any survey-based statistic, ask: Who was asked? Who was missed? Who might not have answered honestly?
- The method of data collection is as important as the method of sampling. A random sample undermined by a biased collection process produces biased results.

CHAPTER 4

Describing Data Visually

> *"A picture is worth a thousand data points — but only if you know how to read it."*
>
> — on why every nurse needs to understand graphs

In this chapter, you will learn to:

- Organize raw data into frequency distributions and relative frequency tables
- Construct and interpret bar charts and pie charts for qualitative data
- Construct and interpret histograms, stem-and-leaf plots, and dot plots for quantitative data
- Describe the shape of a distribution: symmetric, skewed left, skewed right
- Use Desmos and Excel to create graphs from clinical datasets
- Identify misleading graphs and explain what makes them deceptive

4.1 Organizing Data: Frequency Distributions

Step 1 Read This First

You are a charge nurse reviewing last month's incident reports for your 28-bed medical-surgical unit. The reports list the type of each incident: fall, medication error, pressure injury, patient elopement, IV infiltration, or other.

Here is the raw data — 30 incidents listed in the order they were reported:

Fall
Med Error
Pressure Injury
IV Infiltration
Fall
Med Error
Other
Fall
Fall
IV Infiltration
Fall
Fall
Med Error
Pressure Injury
Med Error
IV Infiltration
Med Error
Other
Fall
Fall
Pressure Injury
Fall
Fall
Med Error
Fall
Fall
IV Infiltration
Med Error
Other
Pressure Injury

Right now, this is just a list. You cannot quickly see which incident type is most common, whether falls are increasing, or how much of your team's time goes to each category. Before any of that analysis is possible, the data needs to be organized.

Step 2 Let's Talk About It

Look at the list above. Even with just 30 incidents, counting "falls" requires scanning every row. With 300 incidents or 3,000, it becomes impossible to see patterns in a raw list.

The solution is to *tally*: go through the list once and count how many times each category appears. Once you have counts, you can immediately see that falls are the dominant incident type, that medication errors are second, and that elopements did not occur at all last month.

But counts alone are not always enough. If you want to compare your unit's fall rate to another unit with 50 beds, you need to express each count as a *proportion* of the total. That way the comparison is fair regardless of how many total incidents each unit had.

This is the purpose of a frequency distribution — it transforms raw data into a structured table that makes patterns visible at a glance.

Step 3 Now We Name It

Definition: Frequency Distribution

A **frequency distribution** is a table that lists each possible value (or category) of a variable and the number of times it occurs in the dataset. The count for each value is called the **frequency**.

Definition: Relative Frequency and Cumulative Frequency

The **relative frequency** of a category is its frequency expressed as a proportion (or percentage) of the total:

$$\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Total } n}$$

The **cumulative frequency** is the running total of frequencies up to and including a given value. It is used for ordered (quantitative) data to show "how many observations fall at or below this value."

Definition: Class Width and Classes (for Quantitative Data)

When data are quantitative and continuous, we group values into **classes** (intervals). The **class width** is the length of each interval:

$$\text{Class Width} \approx \frac{\text{Maximum value} - \text{Minimum value}}{\text{Number of classes}}$$

A good rule of thumb: use between 5 and 20 classes. Classes must be non-overlapping and cover all data values.

Step 4 Watch It Work

Building the frequency distribution for the incident data

Counting each type from the raw list:

Incident Type	**Tally**	**Frequency**	**Relative Freq.**	**Cumul. Freq.**
Fall	*IIII IIII II*	12	$12/30 = 0.400$	12
Medication Error	*IIII II*	7	$7/30 = 0.233$	19
IV Infiltration	*IIII*	4	$4/30 = 0.133$	23
Pressure Injury	*IIII*	4	$4/30 = 0.133$	27
Other	*III*	3	$3/30 = 0.100$	30
Patient Elopement	—	0	$0/30 = 0.000$	30
Total		**30**	**1.000**	

What this table tells us immediately:

- Falls account for 40% of all incidents — nearly twice the rate of any other category.
- Medication errors are second at 23.3%.
- The top two categories together account for 63.3% of all incidents.
- Cumulative frequency shows that 23 of 30 incidents (76.7%) involved falls, medication errors, or IV infiltrations.

Now a quantitative example: Patient wait times (in minutes) from triage to first nurse contact for 20 ED patients:

8, 12, 5, 19, 23, 31, 14, 7, 27, 18, 9, 35, 22, 11, 16, 29, 6, 24, 33, 15

Using class width $= (35 - 5)/5 = 6$ minutes, rounded to width of 7 for clean intervals:

Wait Time (min)	**Frequency**	**Relative Freq.**	**Cumul. Freq.**
5 to < 12	6	0.300	6
12 to < 19	5	0.250	11
19 to < 26	4	0.200	15
26 to < 33	3	0.150	18
33 to < 40	2	0.100	20
Total	**20**	**1.000**	

Interpretation: 55% of patients waited less than 19 minutes. Two patients waited 33 minutes or more — a potential patient experience concern.

Step 5 Your Turn

1. A nurse records the primary reason for 24 call light activations during one shift:

Pain	Bathroom	Pain	Repositioning
Bathroom	Nausea	Pain	Bathroom
Repositioning	Pain	Other	Pain
Bathroom	Repositioning	Pain	Nausea
Pain	Bathroom	Repositioning	Pain
Other	Pain	Bathroom	Repositioning

(a) Construct a complete frequency distribution table including frequency, relative frequency, and cumulative frequency.

(b) What is the most common reason for a call light activation?

(c) What percentage of activations were for pain or bathroom needs combined?

(d) What cumulative frequency tells you that 75% of all activations have been accounted for?

2. The systolic blood pressures (mmHg) of 18 patients on a cardiac unit are:

118, 134, 142, 109, 127, 155, 138, 121, 148, 163, 116, 132, 145, 128, 137, 152, 119, 141

(a) Find the range (maximum − minimum).

(b) Using 5 classes of equal width, determine an appropriate class width.

(c) Construct the frequency distribution table.

(d) What proportion of patients have systolic BP of 140 mmHg or higher (hypertension stage 2 threshold)?

Step 6 Think Like a Nurse

What the Numbers Don't Show
Your frequency distribution shows that falls account for 40% of all incidents last month. Administration schedules a meeting to discuss fall prevention. Before that meeting, think about what the table tells you — and what it does not.

The table tells you *how many* falls occurred. It does not tell you:

- *When* they occurred (which shifts? which days of the week?)
- *Who* fell (patient demographics, acuity level, fall risk score)
- *Where* they fell (which rooms? near the call light? in the bathroom?)
- *Whether* the rate is improving or worsening over time

The frequency distribution is the starting point, not the conclusion. It tells you where to look next. A charge nurse who walks into the fall prevention meeting armed only with "we had 12 falls" is less prepared than one who has already cross-tabulated falls by shift, by patient acuity, and by room location.

Data organization is not the end of analysis. It is what makes deeper analysis possible.

4.2 Graphs for Qualitative Data: Bar Charts and Pie Charts

Step 1 Read This First

You need to present last month's incident data to hospital leadership. You have the frequency distribution table from Section 4.1. But leadership does not want to read a table — they want to see the story at a glance.

You have two standard tools for displaying qualitative (categorical) data visually:

- A **bar chart**, which shows the frequency or relative frequency of each category as the height of a bar
- A **pie chart**, which shows each category as a slice of a circle proportional to its relative frequency

Both are familiar. Both are commonly misused. By the end of this section you will know when each is appropriate, how to read them accurately, and how to spot when they are being used to mislead.

Step 2 Let's Talk About It

Think about which display would be more useful for leadership.

A bar chart makes it easy to compare the sizes of categories. The bars have a common baseline, so your eye can quickly judge that the fall bar is almost twice the height of the medication error bar.

A pie chart makes it easy to see how the whole is divided. It is natural for communicating parts-of-a-whole — "falls alone account for 40% of all incidents."

But pie charts struggle when there are many slices of similar size. If you have six categories and four of them are each around 10–15%, the slices look almost identical and are hard to compare. Bar charts handle this situation much better.

A good rule: use a bar chart when you want to *compare* categories. Use a pie chart when you want to show *composition* and there are few categories.

Step 3 Now We Name It

Definition: Bar Chart

A **bar chart** displays the frequency or relative frequency of each category of a qualitative variable as the height of a rectangular bar. Bars are separated by gaps (distinguishing them from histograms). The categories can be placed in any order, but ordering by frequency (highest to lowest) makes comparisons easiest — this is called a **Pareto chart**.

Definition: Pie Chart

A **pie chart** represents each category as a sector (slice) of a circle. The angle of each slice is proportional to the category's relative frequency:

$$\text{Angle} = \text{Relative Frequency} \times 360$$

Best used with 2–5 categories. Becomes difficult to read with more slices.

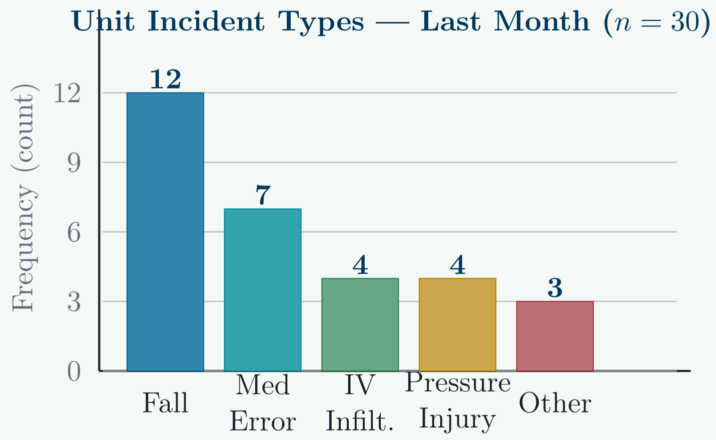
Side-by-Side Bar Charts for Comparisons

When comparing the same categories across two or more groups (e.g., incident types on Day shift vs. Night shift), a **side-by-side bar chart** places grouped bars next to each other for each category. This makes cross-group comparisons direct and clear. It is one of the most useful graphs in nursing quality improvement work.

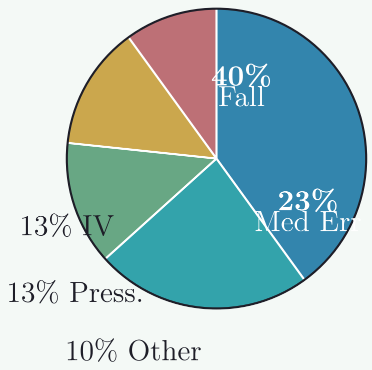
Step 4 Watch It Work

Bar chart and pie chart for the incident data

Bar Chart (ordered by frequency — Pareto style):


Unit Incident Types — Last Month $(n = 30)$

Pie Chart (showing proportional composition):


Incident Composition $(n = 30)$

Key takeaway: The bar chart makes it easier to compare category sizes. The pie chart communicates that falls alone make up nearly half the whole. Both are correct — the choice depends on the story you are telling.

Step 5 Your Turn

1. Use the call light data from Section 4.1 (Problem 1).
  (a) Sketch a bar chart ordered from most to least frequent (Pareto style). Label the axes and title your graph.
  (b) Calculate the angle for each category in a pie chart. Show your work.

   (c) Which graph would you show to a nurse manager who wants to know what is driving call light volume? Justify your choice.

2. A hospital tracks primary diagnosis at discharge for 120 patients over one month. The counts are: Cardiac 38, Respiratory 27, Neurological 19, Orthopedic 22, Other 14.
   (a) Construct a relative frequency table.
   (b) Calculate each angle for a pie chart.
   (c) Would a pie chart or bar chart be more appropriate here? Why?

3. ⋆ A side-by-side bar chart compares medication errors on Day shift vs. Night shift across four error types (Wrong Dose, Wrong Drug, Wrong Time, Wrong Route). Day shift counts are: 8, 3, 12, 2. Night shift counts are: 5, 6, 9, 7.
   (a) What does this chart tell you about the pattern of errors by shift?
   (b) Which error type shows the greatest day-to-night difference?
   (c) What might explain the higher Wrong Route errors at night?
   (d) If you were presenting this data at a safety committee meeting, what action would you recommend based on the chart?

Step 6 Think Like a Nurse

When Charts Lie

Hospital marketing produces a pie chart showing patient satisfaction scores for four categories: Excellent 48%, Very Good 31%, Good 15%, Fair/Poor 6%. The chart is displayed in the hospital lobby under the heading “Our Patients Love Us.”

The chart is not technically wrong. But look at what is being communicated versus what is being hidden.

The four slices are colored so that Excellent and Very Good are shown in deep blues and the Good and Fair/Poor slices are nearly the same shade of light gray, making them hard to distinguish. The 6% Fair/Poor slice — representing patients who had a negative experience — is visually minimized.

Now consider: what is 6% of your hospital’s annual patient volume? If you discharge 10,000 patients per year, that is 600 people who had a poor enough experience to say so on a formal survey. That is not a small number.

A more honest presentation would use a bar chart with a consistent color scheme that does not visually minimize negative outcomes, and would include the count next to the percentage.

A chart is an argument. Before you accept it, ask: what has been highlighted? What has been de-emphasized? What is missing entirely?

4.3 Graphs for Quantitative Data: Histograms and More

Step 1 Read This First

You are reviewing the ages of 30 patients admitted to a medical-surgical unit this week. You have the raw ages:

$$45, 72, 58, 61, 34, 83, 67, 49, 55, 78, 62, 41, 70, 53, 88, 66, 47, 59, 73, 38,$$
$$64, 51, 80, 57, 69, 44, 76, 52, 63, 71$$

You want to know: are these patients mostly middle-aged? Mostly elderly? Evenly spread? Are there a few very young patients or very old ones who might need special care planning?

A frequency table would organize the data, but a visual display would make the *shape* of the distribution immediately clear. This section introduces three graphs designed specifically for quantitative data.

Step 2 Let's Talk About It

The key difference between graphs for qualitative data (bar charts, pie charts) and graphs for quantitative data is that numerical data has an inherent order and the distance between values is meaningful.

When you look at a histogram, you are not just comparing bar heights — you are also reading the *shape* of the distribution. Is it bell-shaped? Skewed to one side? Does it have two separate peaks? The shape tells you something about the underlying phenomenon.

A roughly symmetric, bell-shaped distribution of patient ages would suggest a fairly typical general medical population. A distribution strongly skewed toward older ages would tell you that your unit tends to serve an elderly population — which has implications for fall risk, polypharmacy, length of stay, and discharge planning.

Step 3 Now We Name It

Definition: Histogram

A **histogram** is a graph for quantitative data where the horizontal axis represents the values of the variable (grouped into classes) and the vertical axis represents frequency or relative frequency. Unlike bar charts, the bars in a histogram **touch** — there are no gaps — because the data is continuous and the classes cover consecutive intervals without gaps.

The **shape** of a histogram is one of its most important features.

Definition: Distribution Shapes

Symmetric (bell-shaped): The left and right sides mirror each other. Most values cluster in the middle, with fewer values toward both extremes.

Clinical example: Resting heart rates in a healthy adult population.

Skewed right (positively skewed): The tail extends toward the right. Most values are low, with a few unusually high values pulling the tail out. The mean is pulled above the median.

Clinical example: Length of stay — most patients stay 2–5 days, but a few stay 30+ days.

Skewed left (negatively skewed): The tail extends toward the left. Most values are high, with a few unusually low values pulling the tail down.

Clinical example: APGAR scores at 5 minutes — most babies score 7–10, very few score below 4.

Definition: Stem-and-Leaf Plot

A **stem-and-leaf plot** displays quantitative data by splitting each value into a **stem** (the leading digit(s)) and a **leaf** (the last digit). It preserves all original data values while showing the distribution's shape. Best for small-to-medium datasets (roughly $n \leq 50$).

Definition: Dot Plot

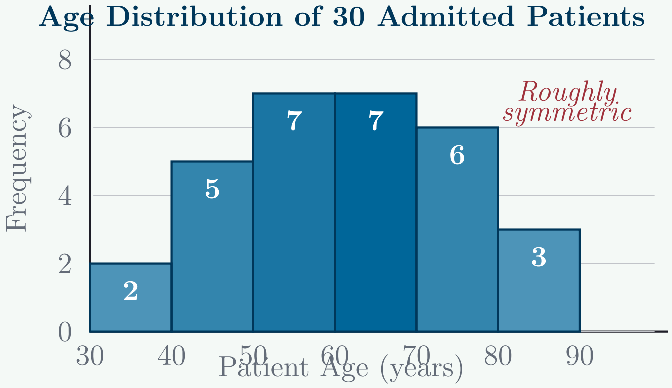
A **dot plot** places a dot above a number line for each data value. When multiple values are the same, dots stack vertically. Simple and useful for small datasets; reveals gaps, clusters, and outliers clearly.

Step 4 Watch It Work

Three displays for the patient age data

Data (sorted): 34, 38, 41, 44, 45, 47, 49, 51, 52, 53, 55, 57, 58, 59, 61, 62, 63, 64, 66, 67, 69, 70, 71, 72, 73, 76, 78, 80, 83, 88

Display 1: Histogram (class width = 10 years)



Display 2: Stem-and-Leaf Plot

Stem	Leaves
3	4 8
4	1 4 5 7 9
5	1 2 3 5 7 8 9
6	1 2 3 4 6 7 9
7	0 1 2 3 6 8
8	0 3 8

Key: Stem = tens digit, Leaf = ones digit. Example: 3 | 4 = age 34

What both displays show: The distribution is roughly symmetric and approximately bell-shaped. Most patients are between 50 and 79 years old, with the peak in the 50s and 60s. This is consistent with a general medical-surgical population. There are no extreme outliers, though the 34-year-old and 88-year-old are the most unusual ages.

Try This in Desmos

To create a histogram in Desmos Graphing Calculator from a list of ages:

1. Open `desmos.com/calculator`
2. Type: `A = [45, 72, 58, 61, 34, ...]` (enter all values)
3. In a new row type: `histogram(A)`
4. Desmos will display the histogram. Click the settings gear to adjust bin width.

Try This in Excel

To create a histogram in Excel from a data column:

1. Enter data in column A
2. Select the data range
3. Insert → Charts → Histogram
4. Right-click the horizontal axis → Format Axis → set Bin Width to 10
Note: Excel histograms use left-closed intervals [lower, upper) by default.

Step 5 Your Turn

1. The number of medications administered per patient during a 12-hour shift for 20 patients:

3, 7, 2, 9, 5, 11, 4, 8, 6, 3, 10, 5, 7, 2, 8, 6, 4, 9, 5, 7

- (a) Create a stem-and-leaf plot. Use stems 0 and 1.
- (b) What is the shape of this distribution?
- (c) Using Desmos or Excel, create a histogram using a bin width of 3. Describe what you see.
- (d) What is the most common number of medications per patient?

2. The length of stay (in days) for 25 patients discharged from an orthopedic unit:

2, 3, 2, 4, 5, 1, 3, 7, 2, 4, 3, 2, 6, 3, 2, 4, 8, 3, 2, 5, 3, 2, 14, 4, 3

- (a) Create a dot plot for this data.
- (b) Describe the shape of the distribution. What feature stands out most?
- (c) Is the distribution symmetric, skewed right, or skewed left? How do you know?
- (d) The patient with a 14-day stay is an outlier. How might this affect summary statistics? (We will revisit this in Chapter 5.)

3. ⋆ Match each clinical variable with the histogram shape you would most expect. Justify each answer.

- (a) Salary for all registered nurses in the United States
- (b) Gestation length (in weeks) for full-term pregnancies
- (c) Patient satisfaction scores at a highly-rated hospital (scale 1–10)
- (d) Time from symptom onset to emergency department arrival for stroke patients
- (e) Blood glucose levels in a general hospital population (many diabetic patients)

4. ⋆ **Desmos/Excel Application:** Open Desmos or Excel and use the following diastolic blood pressure readings for 24 hypertensive patients to create a histogram:

88, 94, 102, 86, 97, 108, 91, 85, 99, 112, 93, 87, 103, 96, 89, 101, 95, 92, 107, 98, 84, 100, 90, 106

(a) Create the histogram using a bin width of 5 mmHg.
(b) Describe the shape. Is it symmetric, skewed right, or skewed left?
(c) What percentage of patients have diastolic BP of 100 mmHg or higher (Stage 2 hypertension)?
(d) Would a stem-and-leaf plot work well for this dataset? Why or why not?

Step 6 Think Like a Nurse

The Shape of the Distribution Tells a Clinical Story

You are reviewing length-of-stay data for your unit. The histogram is strongly skewed right: most patients stay 2–4 days, but the tail extends out to 18 and even 25 days for a small number of patients.

Administration presents the *average* (mean) length of stay as 6.2 days and suggests that your unit is performing below the hospital benchmark of 5.5 days.

But look at the histogram. The 6.2-day mean is being pulled upward by those extreme outliers in the right tail. The *median* length of stay — the middle value, which is resistant to outliers — is 3 days. Half your patients go home in 3 days or less.

Which number better represents the typical patient experience? The median. The mean is being distorted by a handful of patients with complex, prolonged stays who are not representative of the usual case.

Before any summary statistic is applied, look at the shape. A right-skewed distribution with a long tail calls for the median, not the mean. A bell-shaped distribution with no extreme outliers can be fairly summarized by the mean.

The graph is not decoration. It is the prerequisite. Always visualize your data before computing summary statistics.

4.4 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 4. Problems marked with ⋆ are more challenging.

Part A: Frequency Distributions

1. A pediatric unit records the reason for 20 unplanned readmissions within 30 days:

Infection	Dehydration	Infection	Pain	Infection
Dehydration	Respiratory	Pain	Infection	Other
Respiratory	Infection	Dehydration	Pain	Infection
Other	Infection	Respiratory	Dehydration	Pain

(a) Construct a complete frequency distribution table (frequency, relative frequency, cumulative frequency).

(b) What percentage of readmissions involved infection?

(c) What is the cumulative relative frequency for the top two causes?

2. A hospital tracks overnight call-in rates (number of nurses who call in per night) over 30 nights:

0, 1, 2, 0, 3, 1, 2, 0, 1, 4, 2, 1, 0, 2, 1, 3, 1, 0, 2, 1, 1, 2, 0, 1, 3, 2, 1, 0, 1, 2

(a) Construct a frequency distribution.

(b) On what percentage of nights did at least 2 nurses call in?

(c) What is the relative frequency of zero call-ins?

Part B: Bar Charts and Pie Charts

3. Use the readmission data from Problem 1.

(a) Sketch a Pareto bar chart (ordered highest to lowest frequency). Label axes and provide a title.

(b) Calculate the pie chart angle for each category.

(c) Would you present a bar chart or pie chart to the quality improvement committee? Justify.

4. A hospital displays a pie chart in its annual report showing: *"91% of patients rated care as Excellent or Very Good; 9% rated it as Good, Fair, or Poor."* The 9% slice is drawn using a pale gray color, barely distinguishable from the white background.

(a) What presentation technique is being used to minimize the negative response?

(b) If the hospital serves 15,000 patients annually, how many patients gave a negative rating?

(c) Redesign this display in a more honest way. What would you change?

Part C: Histograms, Stem-and-Leaf, and Distribution Shape

5. The hemoglobin levels (g/dL) for 16 patients with anemia:

7.2, 8.8, 6.5, 9.1, 7.8, 8.3, 6.9, 9.5, 7.4, 8.1, 7.0, 9.8, 8.6, 6.7, 7.6, 8.9

- (a) Construct a stem-and-leaf plot (stems: 6, 7, 8, 9; leaves: tenths digit).
- (b) Describe the shape of the distribution.
- (c) What percentage of patients have hemoglobin below 8.0 g/dL?
- (d) Clinical reference: hemoglobin < 7.0 g/dL typically triggers transfusion consideration. How many patients are in this range?

6. Describe the shape you would expect for each histogram and explain your reasoning:

- (a) Ages at first diagnosis of Type 1 diabetes (mostly children and young adults)
- (b) Number of days since last continuing education course for a sample of nurses (many recently certified, some who haven't done it in years)
- (c) Systolic blood pressure in a random sample of healthy young adults
- (d) Patient age at admission to a hospice care unit

7. ⋆ The serum creatinine levels (mg/dL) for 22 patients in a renal unit:

1.2, 3.4, 1.8, 5.6, 2.1, 7.2, 1.5, 4.8, 2.3, 1.9, 6.1, 1.7, 3.2, 9.4, 2.6, 1.4, 4.3, 2.0, 1.6, 8.7, 3.8, 2.2

- (a) Sort the data and create a frequency distribution using class width of 2 (classes: 1–3, 3–5, 5–7, 7–9, 9–11).
- (b) Describe the shape of the distribution.
- (c) Normal creatinine range is approximately 0.6–1.2 mg/dL. What does the distribution tell you about this patient population?
- (d) Would the mean or median better represent the "typical" patient's creatinine level? Why?
- (e) Using Desmos or Excel, create a histogram and compare it to your frequency distribution.

Answer Key — Selected Problems

Answer Key

Section 4.1 Practice — Your Turn

1. (a) Frequency table: Pain=9, Bathroom=6, Repositioning=5, Nausea=2, Other=2; Relative frequencies: Pain=0.375, Bathroom=0.250, Repositioning=0.208, Nausea=0.083, Other=0.083. (b) Pain (9 activations, 37.5%). (c) Pain + Bathroom = 15/24 = 62.5%. (d) Cumulative frequency = 18 (Pain + Bathroom + Repositioning = 9+6+5).
2. (a) Range = 163 − 109 = 54. (b) Class width = 54/5 ≈ 11; round to 10 or 12 for clean intervals. Using width 10: classes 109–119, 119–129, 129–139, 139–149, 149–159, 159–169. (c) Counts: 3, 4, 4, 4, 2, 1. (d) Patients with BP ≥ 140: count values 141, 142, 145, 148, 152, 155, 163 = 7 patients; 7/18 = 38.9%.

Section 4.2 Practice — Your Turn

1. (b) Angles: Pain = 0.375 × 360 = 135; Bathroom = 90; Repositioning = 75; Nausea = 30; Other = 30. (c) Bar chart — comparing 5 categories of similar sizes is clearer with bars than with pie slices.
2. (a) Relative frequencies: Cardiac 0.317, Respiratory 0.225, Orthopedic 0.183, Neurological 0.158, Other 0.117. (b) Angles: Cardiac ≈ 114, Respiratory ≈ 81, Orthopedic ≈ 66, Neurological ≈ 57, Other ≈ 42. (c) Bar chart is preferable — 5 categories of similar sizes are difficult to distinguish as pie slices.
3. (a) Day shift has more Wrong Time errors (12 vs. 9); Night shift has more Wrong Drug and Wrong Route errors. (b) Wrong Route: Day = 2, Night = 7, difference = 5 — greatest day-to-night difference. (c) Night shift may have fewer available pharmacy consultations, reduced supervisor oversight, increased fatigue, or less familiar nursing staff. (d) Recommend a night-shift specific medication administration audit, particularly for route verification.

Section 4.3 Practice — Your Turn

1. (a) Stem 0: leaves 2, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9; Stem 1: leaves 0, 1. (b) Roughly symmetric or slightly right-skewed. (c) Most values cluster between 4–9. (d) Mode is 5 (appears 3 times).
2. (b) Right-skewed; the 14-day patient is an outlier extending the tail. (c) Skewed right — most values are clustered at 2–5 days with a long tail toward higher values. (d) The outlier will pull the mean upward, making it unrepresentative of the typical stay; the median would be more appropriate.
3. (a) Right-skewed (most nurses earn similar salaries but a small number earn very high salaries). (b) Symmetric/bell-shaped (gestation is tightly regulated biologically). (c) Left-skewed (high satisfaction; few very low scores). (d) Right-skewed (most stroke patients arrive late; few arrive immediately). (e) Bimodal or right-

skewed (one cluster for normal BG, another for diabetic patients with elevated levels).

Chapter Practice — Selected

Problem 1: Infection = 7 (35%), Dehydration = 4 (20%), Pain = 4 (20%), Respiratory = 3 (15%), Other = 2 (10%). Cumulative relative frequency for top two: Infection + Dehydration = 55%.

Problem 2b: At least 2 call-ins: count nights with 2, 3, or 4 = 8 + 3 + 1 = 12 nights; 12/30 = 40%.

Problem 5a: Stem-and-leaf: 6|5 7 9 7|0 2 4 6 8 8|1 3 6 8 9 9|1 5 8.

Problem 7 (Challenge): (b) Strongly right-skewed — most creatinine values are in the 1–3 range but a few very high values (7.2, 8.7, 9.4) extend the tail. (c) All patients are above the normal range (0.6–1.2 mg/dL), indicating a population with significant kidney dysfunction. (d) Median — the right skew and high outliers will pull the mean upward, making it unrepresentative of the typical patient.

Chapter 4 Summary

Section 4.1 — Frequency Distributions

- A **frequency distribution** organizes data into a table of categories and counts.
- **Relative frequency** = frequency ÷ total n; expresses each category as a proportion.
- **Cumulative frequency** shows how many observations fall at or below a given value.
- For quantitative data, choose 5–20 classes of equal width: width ≈ (max − min) ÷ number of classes.

Section 4.2 — Qualitative Data Graphs

- **Bar chart**: displays frequency/relative frequency of categories; bars do not touch; best for comparing categories.
- **Pareto chart**: bar chart ordered from highest to lowest frequency.
- **Pie chart**: sectors proportional to relative frequency; best for 2–5 categories showing composition.
- A chart is an argument — check what is highlighted, minimized, or missing.

Section 4.3 — Quantitative Data Graphs

- **Histogram**: bars touch (continuous data); reveals the shape of the distribution.
- **Stem-and-leaf plot**: preserves all data values; best for $n \leq 50$.
- **Dot plot**: one dot per value above a number line; reveals clusters, gaps, and outliers.
- **Symmetric**: mirror-image shape; mean ≈ median.
- **Skewed right**: long tail to the right; mean > median (e.g., length of stay).
- **Skewed left**: long tail to the left; mean < median (e.g., APGAR scores).

The Nursing Connection

- Always visualize data before computing summary statistics. The shape of the distribution determines which statistics are appropriate.
- Right-skewed clinical data (length of stay, wait times, creatinine) is better summarized by the median than the mean.
- Charts can mislead through color, scale manipulation, or selective display — read them critically.

CHAPTER 5

Describing Data Numerically

> *"A single number, chosen wisely, can summarize an entire patient population."*
>
> — chosen poorly, it can mislead everyone in the room

In this chapter, you will learn to:

- Compute and interpret the mean, median, and mode
- Explain when each measure of center is appropriate
- Compute and interpret range, variance, standard deviation, and IQR
- Construct and interpret a five-number summary and boxplot
- Identify and handle outliers using the IQR fence method
- Compute and interpret z-scores
- Apply the Empirical Rule to bell-shaped distributions
- Use Desmos and Excel for all calculations

5.1 Measures of Center: Mean, Median, and Mode

Step 1 Read This First

You are reviewing patient outcomes data for a quality improvement project. Your unit tracked the length of stay (in days) for 12 patients discharged last week:

$$3, 2, 5, 4, 3, 7, 2, 4, 3, 6, 3, 28$$

A colleague summarizes this as: *"The average length of stay on our unit is 5.8 days."* She presents this to administration as evidence that the unit needs more discharge planning resources.

You look at the numbers again. Eleven of the twelve patients went home in 7 days or less. One patient — who developed a serious complication — stayed 28 days. Does 5.8 days really represent the *typical* patient on your unit?

The answer depends entirely on which measure of center you choose — and whether that choice is appropriate for the shape of your data.

Step 2 Let's Talk About It

Look at the twelve values. Remove the 28 for a moment. The remaining eleven values — 2, 2, 3, 3, 3, 3, 4, 4, 5, 6, 7 — have an average of about 3.8 days. That feels like a much better description of the typical patient.

When you put 28 back in, the average jumps to 5.8. That one extreme value dragged the mean almost two full days higher. Yet only one of twelve patients stayed that long.

This is the fundamental tension in descriptive statistics: we want one number to represent the whole group, but that one number is always a simplification. The art is knowing which simplification is honest.

There are three common measures of center. Each makes a different promise about what it represents — and each breaks that promise under certain conditions.

Step 3 Now We Name It

Definition: Mean

The **mean** (arithmetic average) is the sum of all values divided by the number of values:

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{(sample mean)} \qquad \mu = \frac{\sum x_i}{N} \quad \text{(population mean)}$$

The mean uses every data value. This makes it sensitive to extreme values (outliers). It is the most appropriate measure of center for symmetric distributions without outliers.

Definition: Median

The **median** is the middle value when data are arranged in order. It divides the dataset so that half the values fall below and half above.

Finding the median:

- Sort the data from smallest to largest.
- If n is odd: the median is the middle value at position $\dfrac{n+1}{2}$.
- If n is even: the median is the average of the two middle values at positions $\dfrac{n}{2}$ and $\dfrac{n}{2}+1$.

The median is **resistant** to outliers. It is the preferred measure of center for skewed distributions.

Definition: Mode

The **mode** is the value (or values) that appear most frequently in the dataset. A dataset may have one mode (**unimodal**), two modes (**bimodal**), or no mode (all values appear equally often).

The mode is the only measure of center that can be used with qualitative data.

Example: The most common blood type in a sample of 50 patients is "O positive" — the mode for the qualitative variable Blood Type.

Which Measure of Center to Use?

Symmetric, no outliers: Mean and median will be close; use the mean.
Skewed or outliers present: The median is more representative; use the median.
Qualitative data: Only the mode applies.
Rule of thumb: If mean > median, the distribution is likely skewed right. If mean < median, likely skewed left.

Step 4 Watch It Work

Computing the three measures for the length-of-stay data

Data: 3, 2, 5, 4, 3, 7, 2, 4, 3, 6, 3, 28 $(n = 12)$

Step 1 — Sort: 2, 2, 3, 3, 3, 3, 4, 4, 5, 6, 7, 28

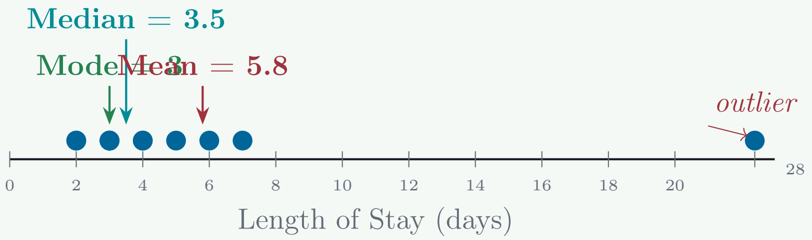
Step 2 — Mean:

$$\bar{x} = \frac{2+2+3+3+3+3+4+4+5+6+7+28}{12} = \frac{70}{12} \approx 5.8 \text{ days}$$

Step 3 — Median $(n = 12$, even$)$: average of 6th and 7th values:

$$\text{Median} = \frac{3+4}{2} = 3.5 \text{ days}$$

Step 4 — Mode: 3 appears four times — the mode is **3 days**.



Interpretation:

- The mean (5.8 days) is pulled well above the data cluster by the single 28-day outlier.
- The median (3.5 days) and mode (3 days) much better represent the *typical* patient experience.
- For this right-skewed dataset with an extreme outlier, the **median** is the honest summary.
- Reporting the mean to administration without mentioning the outlier misrepresents 11 of 12 patients.

Try This in Desmos

In Desmos Scientific Calculator, type:

`mean([2,2,3,3,3,3,4,4,5,6,7,28])` → returns 5.833

`median([2,2,3,3,3,3,4,4,5,6,7,28])` → returns 3.5

Try This in Excel

In Excel, with data in cells A1:A12:

`=AVERAGE(A1:A12)` → mean

`=MEDIAN(A1:A12)` → median

`=MODE(A1:A12)` → mode (returns first mode if multiple)

Step 5 Your Turn

1. The pain scores (0–10 scale) recorded for 11 patients at a post-operative check:

7, 3, 5, 8, 4, 6, 5, 9, 5, 2, 6

- (a) Sort the data and find the mean, median, and mode.
- (b) Are the mean and median close together or far apart? What does this suggest about the shape?
- (c) Which measure of center would you report to the charge nurse as the "typical" pain level? Why?

2. The diastolic blood pressures (mmHg) of 10 patients in a hypertension clinic:

92, 88, 105, 97, 84, 91, 118, 95, 89, 93

- (a) Compute the mean and median.
- (b) Remove the value 118 and recompute the mean and median.
- (c) How much did each measure change? Which is more resistant to the outlier?
- (d) Using Desmos or Excel, verify your calculations.

3. ⋆ A hospital reports the "average" nurse salary on its website as \$82,000. A newly hired nurse discovers that 8 of the 10 nurses on her unit earn between \$68,000 and \$75,000, but the two senior charge nurses each earn \$142,000.

- (a) Without computing exactly, explain why \$82,000 does not represent the typical nurse's salary.
- (b) Which measure of center would you recommend reporting? Why?
- (c) What might motivate a hospital to report the mean rather than the median for salary data?

Step 6 Think Like a Nurse

The Mean Can Be Technically Correct and Clinically Misleading

A hospital quality report states: *"Average time from ED triage to physician evaluation: 38 minutes."* This meets the national benchmark of 45 minutes, so the quality committee marks this indicator as "passing."

But when a nurse looks at the actual distribution, she finds: 80% of patients were seen in under 20 minutes. The remaining 20% waited between 90 and 180 minutes — patients who arrived during shift change or when the department was at capacity.

The mean of 38 minutes obscures a two-tiered system. Most patients are seen very quickly. A significant minority waits far too long. The mean hides this split because the low values and high values average out to a number that describes neither group accurately.

A more honest summary would report: *"Median wait time: 18 minutes. 20% of patients waited longer than 90 minutes."* That tells you something actionable.

When a single number is used to describe a system, ask: does this number describe what is actually happening to patients, or does it describe the arithmetic average of two very different experiences?

5.2 Measures of Spread: Range, Standard Deviation, and IQR

Step 1 Read This First

Two nursing units both have a median patient-to-nurse ratio of 5:1.

Unit A ratios over the last 14 shifts: 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6

Unit B ratios over the last 14 shifts: 1, 2, 3, 4, 5, 5, 5, 5, 5, 6, 7, 8, 9, 10

Both units have the same median. Are they equally safe?

Not even close. Unit A has almost no variation — the ratio is almost always exactly 5. Unit B swings wildly from 1 to 10. On some shifts, nurses on Unit B are managing 10 patients each. The median alone is dangerously misleading here.

This is why measures of *spread* are just as important as measures of center. The center tells you where the data is located. The spread tells you how much it varies around that center.

Step 2 Let's Talk About It

Think about what spread means in clinical terms.

A medication that reliably lowers blood pressure by about 10 mmHg in most patients is very different from one that lowers it by 10 mmHg on average but sometimes by 2 mmHg and sometimes by 35 mmHg. The average effect is the same. The *variability* in effect is completely different — and that variability is what determines whether the medication is safe to use broadly.

In statistics, spread is measured in several ways. The simplest is the range. The most commonly reported is the standard deviation. For skewed data, the interquartile range (IQR) is more robust. Each captures something different about how spread out the data is.

Step 3 Now We Name It

Definition: Range

The **range** is the difference between the maximum and minimum values:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Simple to compute but sensitive to outliers — one extreme value can make the range enormous.

Definition: Standard Deviation

The **standard deviation** measures how far, on average, each data value is from the mean. A larger standard deviation means the data is more spread out; a smaller value means the data clusters tightly around the mean.

Sample standard deviation:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Use s for sample data (which is almost always). The denominator $n - 1$ (not n) corrects for the tendency of samples to underestimate population variability.

Variance (s^2 or σ^2) is the square of the standard deviation. It is used in many statistical formulas but is harder to interpret directly because its units are squared.

Definition: Interquartile Range (IQR)

The **quartiles** divide ordered data into four equal parts:

- Q_1 (first quartile): 25th percentile — 25% of data falls below
- Q_2 (second quartile): 50th percentile — the median
- Q_3 (third quartile): 75th percentile — 75% of data falls below

The **interquartile range** is:

$$\text{IQR} = Q_3 - Q_1$$

The IQR captures the spread of the middle 50% of the data. It is **resistant to outliers**, making it the preferred spread measure for skewed distributions.

Step 4 Watch It Work

Computing spread measures for nurse-to-patient ratios on Unit B

Data (sorted): 1, 2, 3, 4, 5, 5, 5, 5, 5, 6, 7, 8, 9, 10 $(n = 14)$

Range: $10 - 1 = \mathbf{9}$

Quartiles: $n = 14$ values.
Lower half (first 7): 1, 2, 3, 4, 5, 5, 5 $\Rightarrow Q_1 = 4$ (4th value)
Upper half (last 7): 5, 5, 6, 7, 8, 9, 10 $\Rightarrow Q_3 = 7$ (4th value)

$$\textbf{IQR} = Q_3 - Q_1 = 7 - 4 = \mathbf{3}$$

Standard deviation:

$$\bar{x} = \frac{1+2+3+4+5+5+5+5+5+6+7+8+9+10}{14} = \frac{75}{14} \approx 5.36$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	
1	-4.36	18.98	
2	-3.36	11.27	
3	-2.36	5.56	
4	-1.36	1.84	
5	-0.36	0.13	×5
6	$+0.64$	0.41	
7	$+1.64$	2.70	
8	$+2.64$	6.98	
9	$+3.64$	13.27	
10	$+4.64$	21.56	
$\sum(x_i - \bar{x})^2 \approx 83.21$			

$$s = \sqrt{\frac{83.21}{14-1}} = \sqrt{\frac{83.21}{13}} = \sqrt{6.40} \approx \mathbf{2.53}$$

Compare Unit A vs. Unit B:

	Median	**IQR**	**Std Dev**
Unit A	5	0	0.38
Unit B	5	3	2.53

Same median. Very different spread. Unit B's standard deviation is nearly seven times larger. This is the story the center alone cannot tell.

Try This in Desmos

In Desmos, type `stdev([1,2,3,4,5,5,5,5,5,6,7,8,9,10])` for sample standard deviation, or
`stdevp([...])` for population standard deviation.

For quartiles: `quantile([...], 0.25)` gives Q_1; `quantile([...], 0.75)` gives Q_3.

Try This in Excel

`=STDEV(A1:A14)` → sample standard deviation (s)

`=STDEVP(A1:A14)` → population standard deviation (σ)

`=QUARTILE(A1:A14, 1)` → Q_1

`=QUARTILE(A1:A14, 3)` → Q_3

`=IQR` is not a built-in function; compute as `=QUARTILE(A1:A14,3)-QUARTILE(A1:A14,1)`

Step 5 Your Turn

1. The number of patient falls per month over 10 months on two nursing units:

Unit ICU: 0, 1, 0, 2, 1, 0, 1, 0, 1, 0
Unit Med/Surg: 3, 0, 5, 1, 7, 0, 4, 2, 6, 2

(a) Compute the mean and median for each unit.
(b) Compute the range and standard deviation for each unit (use Desmos or Excel for standard deviation).
(c) Both units have similar means. Which unit has a more serious fall problem? Justify using the spread measures.
(d) Which measure of spread is more appropriate here given the data shape?

2. Medication response times (in minutes) for 8 patients after an IV pain medication:

12, 8, 15, 10, 14, 9, 11, 13

(a) Compute the mean and standard deviation. Show your work for the standard deviation.
(b) What is the IQR?
(c) Interpret the standard deviation in context: what does it tell you about how consistently the medication works?

3. ⋆ A researcher reports that a new blood pressure medication has a mean reduction of 12 mmHg with a standard deviation of 1.2 mmHg. A second medication has a mean reduction of 12 mmHg with a standard deviation of 8.4 mmHg.

  (a) Which medication is more predictable in its effect? Explain.
  (b) For a patient with dangerously high BP where consistent reduction is critical, which medication would you prefer? Why?
  (c) For a patient whose BP is only mildly elevated, which concern about the second medication is most relevant?

Step 6 Think Like a Nurse

Standard Deviation Is a Safety Metric

The Joint Commission and hospital accreditation bodies do not just report average outcomes — they look at *variability* in outcomes. A hospital where 90% of patients receive their medication on time and 10% wait hours is more dangerous than one where every patient waits 15 minutes, even if both have the same mean wait time.

This is not abstract statistics. In clinical practice, high variability in:

- medication administration times ⇒ some patients receive drugs too early or too late
- nurse-to-patient ratios ⇒ some shifts are chronically unsafe
- handoff quality ⇒ some transitions miss critical information

When you see a quality report showing a mean that looks acceptable, always ask: what is the standard deviation? What does the distribution look like? A reassuring average can hide a dangerous tail.

In patient safety, the outliers are often the patients who are harmed. The mean protects the average patient. The standard deviation tells you about everyone else.

5.3 The Five-Number Summary, Boxplots, and Z-Scores

Step 1 Read This First

Your hospital's infection control team is comparing central line-associated bloodstream infection (CLABSI) rates (per 1,000 line-days) across eight ICUs in a regional hospital system:

$$0.8,\ 1.2,\ 0.5,\ 2.1,\ 0.9,\ 3.4,\ 1.1,\ 0.7$$

One of the eight ICUs has a rate of 3.4 — more than three times the rate of most units. Is this an outlier that should trigger an investigation, or is it within the normal range of variation? How do you decide?

Meanwhile, your own ICU has a rate of 2.1. A nurse asks: *"How do we compare to the other ICUs? Are we above average? By how much?"*

Two tools answer these questions precisely: the **boxplot** (which visualizes the distribution and flags potential outliers) and the **z-score** (which quantifies how far a value is from the mean in standard deviation units).

Step 2 Let's Talk About It

Think about the question: is 3.4 an outlier?

Intuitively, it seems high — six of the eight values are below 1.3. But "seems high" is not rigorous enough for a clinical investigation or a quality report.

We need a mathematical criterion for what counts as an outlier. The IQR gives us one: values that fall more than 1.5 times the IQR beyond Q_1 or Q_3 are flagged as potential outliers. This is called the **fence method**.

And for the question "how far above average is 2.1?" — the z-score answers this directly. It converts any data value to a universal scale measured in standard deviations, making comparisons meaningful even when the underlying data have different units.

Step 3 Now We Name It

Definition: Five-Number Summary

The **five-number summary** consists of five values that describe a dataset's distribution:

$$\text{Minimum}, \quad Q_1, \quad \text{Median}\,(Q_2), \quad Q_3, \quad \text{Maximum}$$

Together, these five numbers capture center, spread, and the extremes of the distribution.

Definition: Boxplot (Box-and-Whisker Plot)

A **boxplot** is a visual display of the five-number summary:

- A box spans from Q_1 to Q_3 (the IQR)
- A vertical line inside the box marks the median
- Whiskers extend from the box to the smallest and largest *non-outlier* values
- Outliers are plotted as individual dots beyond the whiskers

Definition: Outlier Fences (IQR Method)

A value is a **potential outlier** if it falls outside the fences:

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR} \qquad \text{Upper fence} = Q_3 + 1.5 \times \text{IQR}$$

Values below the lower fence or above the upper fence are flagged as outliers and plotted separately in a boxplot.

Definition: Z-Score (Standardized Score)

The **z-score** measures how many standard deviations a value is above or below the mean:

$$z = \frac{x - \bar{x}}{s}$$

- $z > 0$: the value is above the mean
- $z < 0$: the value is below the mean
- $|z| > 2$: the value is unusually far from the mean (approximately)
- $|z| > 3$: the value is extremely unusual

Z-scores have **no units**. They allow comparison across different variables or different datasets.

Step 4 Watch It Work

Five-number summary, boxplot, and z-scores for CLABSI rates

Data (sorted): 0.5, 0.7, 0.8, 0.9, 1.1, 1.2, 2.1, 3.4 $(n = 8)$

Five-number summary:

- **Min** = 0.5
- Q_1: median of lower half (0.5, 0.7, 0.8, 0.9) $= \frac{0.7+0.8}{2} = 0.75$
- **Median** $= \frac{0.9+1.1}{2} = 1.0$
- Q_3: median of upper half (1.1, 1.2, 2.1, 3.4) $= \frac{1.2+2.1}{2} = 1.65$
- **Max** = 3.4

IQR $= 1.65 - 0.75 = 0.90$

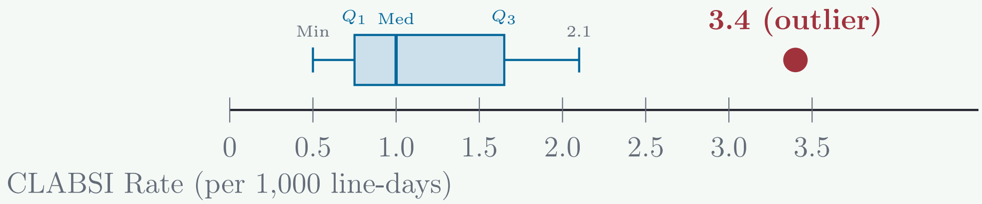
Fences:

$$\text{Lower fence} = 0.75 - 1.5(0.90) = 0.75 - 1.35 = -0.60$$

$$\text{Upper fence} = 1.65 + 1.5(0.90) = 1.65 + 1.35 = 3.00$$

Outlier check: $3.4 > 3.00 \Rightarrow$ **3.4 is a flagged outlier.**

Boxplot:



Z-score for your ICU's rate of 2.1:

Mean of 8 rates: $\bar{x} = \frac{0.5+0.7+0.8+0.9+1.1+1.2+2.1+3.4}{8} = \frac{10.7}{8} = 1.34$

Standard deviation (using Desmos): $s \approx 0.96$

$$z = \frac{2.1 - 1.34}{0.96} = \frac{0.76}{0.96} \approx +0.79$$

Interpretation: Your ICU's CLABSI rate of 2.1 is about 0.79 standard deviations *above* the group mean — above average but not extreme. The rate of 3.4 ($z \approx +2.15$) is more than 2 standard deviations above the mean and is flagged as an outlier by the fence method. That unit warrants investigation.

Step 5 Your Turn

1. The hemoglobin levels (g/dL) for 10 patients:

11.2, 9.8, 13.4, 8.7, 12.1, 10.5, 14.2, 9.1, 11.8, 7.2

(a) Find the five-number summary.
(b) Compute the IQR and determine the outlier fences.
(c) Are any values potential outliers?
(d) Sketch a boxplot. Label all five components.
(e) Compute the z-score for the value 7.2. Is it unusual?

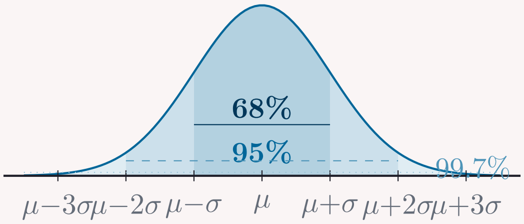
2. A hospital tracks 30-day readmission rates (%) for six hospital units: 8.2, 12.7, 9.1, 18.4, 10.3, 7.6.
(a) Find the five-number summary.
(b) Compute the IQR and outlier fences. Is 18.4 a flagged outlier?
(c) Compute the z-score for 18.4. Interpret it in context.
(d) If you were the nurse manager of the unit with 18.4% readmission, what would you do with this information?

3. $\star$ A nurse's resting heart rate is measured at 102 bpm. The mean resting heart rate for nurses on her unit is 72 bpm with a standard deviation of 9 bpm.
(a) Compute the z-score for 102 bpm.
(b) What does this z-score tell you about how unusual this reading is?
(c) If normal resting heart rate is 60–100 bpm, how does the z-score complement the clinical assessment?
(d) A second nurse has a heart rate of 58 bpm. Compute and interpret her z-score. Should this reading concern you clinically?

Step 6 Think Like a Nurse

The Empirical Rule: When Bell Shapes Tell Us Everything

When a distribution is approximately bell-shaped (symmetric, unimodal), an extraordinarily useful rule applies:



Interval	Contains approximately	Clinical meaning
$\mu \pm \sigma$	68% of data	Most patients fall in this range
$\mu \pm 2\sigma$	95% of data	Almost all patients fall here
$\mu \pm 3\sigma$	99.7% of data	Nearly every patient falls here

Clinical application: If resting heart rate in healthy adults is approximately normally distributed with $\mu = 72$ bpm and $\sigma = 9$ bpm, then:

- 68% of healthy adults have heart rates between 63 and 81 bpm
- 95% have rates between 54 and 90 bpm
- 99.7% have rates between 45 and 99 bpm
- A patient at 105 bpm has $z = (105 - 72)/9 = +3.67$ — this lies outside the 99.7% range and is clinically significant

The Empirical Rule transforms a standard deviation into a statement about individual patients. It is the bridge between the group's statistics and the individual's clinical picture.

5.4 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 5. Problems marked with ⋆ are more challenging.

Part A: Measures of Center

1. The oxygen saturation (SpO_2, %) readings for 9 patients on a respiratory unit:

94, 97, 91, 98, 88, 95, 96, 92, 98

(a) Find the mean, median, and mode.
(b) A tenth patient is added with SpO_2 of 72%. Recompute the mean and median.
(c) Which measure changed more? What does this illustrate?
(d) For the full 10-patient dataset, which measure of center is more appropriate? Why?

2. The number of patients each nurse cared for during 8 consecutive shifts:

5, 6, 5, 7, 5, 8, 5, 6

(a) Find the mean, median, and mode.
(b) The hospital reports "average nurse load of 5.9 patients." A nurse argues this overstates the typical load. Who is right?

Part B: Measures of Spread

3. Two nurses record medication administration times (minutes from order to administration) over 6 shifts:

Nurse A: 28, 32, 30, 29, 31, 30
Nurse B: 14, 45, 22, 38, 19, 52

(a) Compute the mean for each nurse.
(b) Compute the range and standard deviation for each (use Desmos or Excel for standard deviation).
(c) Both nurses have similar means. Which nurse is more consistent? Justify using the standard deviation.
(d) From a patient safety perspective, which variation pattern is more concerning?

4. Serum potassium levels (mEq/L) for 7 patients: 3.8, 4.2, 3.5, 4.7, 3.9, 4.1, 5.8.

(a) Compute the IQR.
(b) Compute the outlier fences. Is 5.8 a flagged outlier?
(c) Normal potassium range is 3.5–5.0 mEq/L. What does identifying 5.8 as an outlier add to the clinical picture?

Part C: Five-Number Summary, Boxplots, Z-Scores, and Empirical Rule

5. Patient wait times (minutes) from admission order to bed assignment for 10 patients:

45, 32, 67, 28, 51, 39, 88, 44, 35, 62

(a) Find the five-number summary.
(b) Compute the IQR and determine the outlier fences.
(c) Is 88 minutes a flagged outlier? Show your work.
(d) Sketch a labeled boxplot.
(e) Compute the z-score for the 88-minute wait. The mean is 49.1 min and standard deviation is 18.6 min.

6. ⋆ A hospital uses the Empirical Rule to set acceptable ranges for patient vital signs. Suppose systolic blood pressure in the general patient population at this hospital is approximately bell-shaped with $\mu = 128$ mmHg and $\sigma = 14$ mmHg.

(a) What range of systolic BP values contains approximately 68% of patients?
(b) What range contains approximately 95% of patients?
(c) A patient has systolic BP of 170 mmHg. Compute the z-score. What does it tell you?
(d) A patient has systolic BP of 100 mmHg. Compute the z-score. Is this clinically unusual?
(e) The Empirical Rule requires a bell-shaped distribution. What type of patients or conditions might violate this assumption in a hospital setting?

7. ⋆ **Research Application:** A study of nurse burnout scores (0–100 scale) in 120 ICU nurses reports: mean = 61.4, median = 58.2, standard deviation = 12.7, Q_1 = 52.1, Q_3 = 70.8.

(a) Is the distribution symmetric, skewed right, or skewed left? How do you know?
(b) Compute the IQR and outlier fences.
(c) A nurse with a burnout score of 95 is flagged. Compute her z-score. Is this consistent with the fence method?
(d) The study reports that the *mean* burnout score is 61.4. Given the skew you identified, is the mean or median a better summary? Why?
(e) If the hospital sets an intervention threshold at "burnout score above 2 standard deviations from the mean," what is the threshold score? How many nurses likely exceed it?

Answer Key — Selected Problems

Answer Key

Section 5.1 Practice — Your Turn

1. Sorted: 2, 3, 4, 5, 5, 5, 6, 6, 7, 8, 9. (a) Mean $= 60/11 \approx 5.5$; Median (6th value) $= 5$; Mode $= 5$. (b) Mean and median are close (5.5 vs. 5); the distribution is roughly symmetric. (c) Either mean or median is reasonable; mode (5) is also a valid answer since it represents the most common score.

2. (a) Mean $= (92+88+105+97+84+91+118+95+89+93)/10 = 952/10 = 95.2$; Sorted: 84, 88, 89, 91, 92, 93, 95, 97, 105, 118; Median $= (92+93)/2 = 92.5$. (b) Without 118: Mean $= 834/9 \approx 92.7$; Sorted: 84, 88, 89, 91, 92, 93, 95, 97, 105; Median $=$ 5th value $= 92$. (c) Mean changed by 2.5; median changed by 0.5. The median is more resistant. (d) The median.

Section 5.2 Practice — Your Turn

1. ICU: Mean $= 0.6$, Median $= 0.5$, Range $= 2$, SD ≈ 0.70. Med/Surg: Mean $= 3.0$, Median $= 2.5$, Range $= 7$, SD ≈ 2.49. (c) Med/Surg has a more serious problem despite a similar-looking structure — both the mean and spread are much higher. (d) The IQR, since the Med/Surg data is right-skewed.

2. Sorted: 8, 9, 10, 11, 12, 13, 14, 15. (a) Mean $= 92/8 = 11.5$. SD: $\sum(x_i - 11.5)^2 = 12.25+6.25+2.25+0.25+0.25+2.25+6.25+12.25 = 42.0$; $s = \sqrt{42/7} = \sqrt{6} \approx 2.45$ min. (b) $Q_1 = 9.5$, $Q_3 = 13.5$, IQR $= 4$. (c) Typical response time deviates from the mean by about 2.45 minutes — the medication is fairly consistent.

Section 5.3 Practice — Your Turn

1. Sorted: 7.2, 8.7, 9.1, 9.8, 10.5, 11.2, 11.8, 12.1, 13.4, 14.2. (a) Min=7.2, $Q_1 = 9.1$, Median$=\frac{10.5+11.2}{2} = 10.85$, $Q_3 = 12.1$, Max=14.2. (b) IQR=3.0; Lower fence$=9.1-4.5 = 4.6$; Upper fence$=12.1+4.5 = 16.6$. (c) No outliers ($7.2 > 4.6$; $14.2 < 16.6$). (e) Mean $= 10.8$, $s \approx 2.17$; $z = (7.2 - 10.8)/2.17 \approx -1.66$; below average but not extreme ($|z| < 2$).

2. Sorted: 7.6, 8.2, 9.1, 10.3, 12.7, 18.4. (a) Min=7.6, $Q_1 = 8.2$, Median=9.7, $Q_3 = 12.7$, Max=18.4. (b) IQR=4.5; Upper fence$=12.7 + 6.75 = 19.45$; $18.4 < 19.45$, so *not* a flagged outlier by the fence method. (c) Mean ≈ 11.05, SD ≈ 4.03; $z = (18.4 - 11.05)/4.03 \approx 1.82$ — above average but within 2 SDs.

Chapter Practice — Selected

Problem 1b: With 72% added: Mean $= (94 + 97 + 91 + 98 + 88 + 95 + 96 + 92 + 98 + 72)/10 = 921/10 = 92.1$; Median = average of 5th and 6th values in sorted set (91, 91, 92, 94, 95, 96, 97, 98, 98, 88 sorted = 72, 88, 91, 92, 94, 95, 96, 97, 98, 98) $= (94 + 95)/2 = 94.5$. Mean dropped by 2.6; median dropped by 0.5. Mean is far more affected.

Problem 5c: $z = (170 - 128)/14 = 42/14 = 3.0$. A z-score of 3.0 places this patient at the boundary of the 99.7% interval — only 0.15% of patients are expected to be this high or higher. This is clinically significant.

Problem 7 (Challenge): (a) Mean (61.4) > Median (58.2) ⇒ right-skewed. (b) IQR $= 70.8 - 52.1 = 18.7$; Lower fence $= 52.1 - 28.05 = 24.05$; Upper fence $= 70.8 + 28.05 = 98.85$. Score of 95 is below 98.85, so not flagged by fence. (c) $z = (95 - 61.4)/12.7 = 33.6/12.7 \approx 2.65$ — more than 2 SDs above mean, consistent with being extreme even if not flagged by IQR fence. (d) Median — the right skew means the mean is pulled up by high-burnout outliers. (e) Threshold $= 61.4 + 2(12.7) = 86.8$. Approximately 2.5% of nurses are expected above this threshold ≈ 3 nurses.

Chapter 5 Summary

Section 5.1 — Measures of Center

- **Mean** ($\bar{x}$): sum $\div$ n; sensitive to outliers; best for symmetric distributions.
- **Median**: middle value; resistant to outliers; best for skewed data.
- **Mode**: most frequent value; only measure for qualitative data.
- If mean $>$ median $\Rightarrow$ right-skewed. If mean $<$ median $\Rightarrow$ left-skewed.

Section 5.2 — Measures of Spread

- **Range** = Max − Min; simple but sensitive to outliers.
- **Standard deviation** (s): average distance from the mean; uses all data values.
- **IQR** $= Q_3 - Q_1$: spread of middle 50%; resistant to outliers.
- Use IQR with median (skewed data); use standard deviation with mean (symmetric data).

Section 5.3 — Five-Number Summary, Boxplots, Z-Scores

- **Five-number summary**: Min, Q_1, Median, Q_3, Max.
- **Outlier fences**: Lower $= Q_1 - 1.5 \times \text{IQR}$; Upper $= Q_3 + 1.5 \times \text{IQR}$.
- **Boxplot**: visual of five-number summary; outliers plotted as individual dots.
- **Z-score**: $z = (x - \bar{x})/s$; measures standard deviations from the mean; unitless.
- **Empirical Rule** (bell-shaped data): 68% within 1σ; 95% within 2σ; 99.7% within 3σ.

The Nursing Connection

- Center without spread is incomplete. Always report both.
- Skewed clinical data (length of stay, wait times, salary) requires the median + IQR.
- Z-scores translate any measurement into a universal scale — "how unusual is this patient?"
- The Empirical Rule connects standard deviation to clinical probability statements.

CHAPTER 6

Correlation and Linear Regression

> *"When two variables move together, the question is always: is one causing the other, or are they just traveling in the same direction?"*
>
> — the difference between correlation and causation, revisited

In this chapter, you will learn to:

- Construct and interpret scatter plots for two quantitative variables
- Compute and interpret the correlation coefficient r
- Distinguish between positive, negative, and no correlation
- Find the least-squares regression line and interpret its slope and intercept
- Use a regression equation to make predictions
- Compute and interpret the coefficient of determination r^2
- Explain why correlation does not imply causation
- Use Desmos and Excel for regression analysis

6.1 Scatter Plots and Correlation

Step 1 Read This First

A hospital quality team suspects that patients with higher body mass index (BMI) have longer post-surgical hospital stays. They collect data from 10 recent surgical patients:

Patient	1	2	3	4	5	6	7	8	9	10
BMI	22	27	31	25	35	29	38	24	33	28
LOS (days)	3	4	6	3	7	5	8	3	7	5

Before running any statistics, the team creates a scatter plot. In that plot, they can immediately see that as BMI increases, length of stay tends to increase as well. The relationship is not perfect — two patients with similar BMIs have different stays — but the general trend is clear.

This visual impression needs to be quantified. How strong is the relationship? Is it strong enough to act on? And does higher BMI *cause* longer stays, or could something else explain both?

Step 2 Let's Talk About It

Think about the scatter plot before the math.

If BMI and LOS were completely unrelated, the dots would form a random cloud with no discernible pattern. If the relationship were perfect, every dot would fall exactly on a straight line. Reality is somewhere in between.

The two things you are looking for in a scatter plot are *direction* (do the dots rise from left to right, or fall?) and *strength* (how closely do the dots cluster around a straight line, or are they spread out all over?).

A third thing worth checking is *form*: is the pattern linear? Or does it curve? The tools in this chapter assume linearity — they work best when the dots trend in a straight line rather than a curve.

Step 3 Now We Name It

Definition: Scatter Plot

A **scatter plot** is a graph that displays the relationship between two quantitative variables. Each dot represents one individual. The **explanatory variable** (the variable believed to drive the relationship) is placed on the x-axis; the **response variable** (the outcome) is placed on the y-axis.

Definition: Correlation Coefficient (r)

The **correlation coefficient** r (also called Pearson's r) measures the strength and direction of the *linear* relationship between two quantitative variables.

Properties of r:

- $-1 \leq r \leq 1$ always
- $r > 0$: positive association (as x increases, y tends to increase)
- $r < 0$: negative association (as x increases, y tends to decrease)
- $r = 0$: no linear relationship
- $|r|$ close to 1: strong linear relationship
- $|r|$ close to 0: weak or no linear relationship

Rough guide to strength:

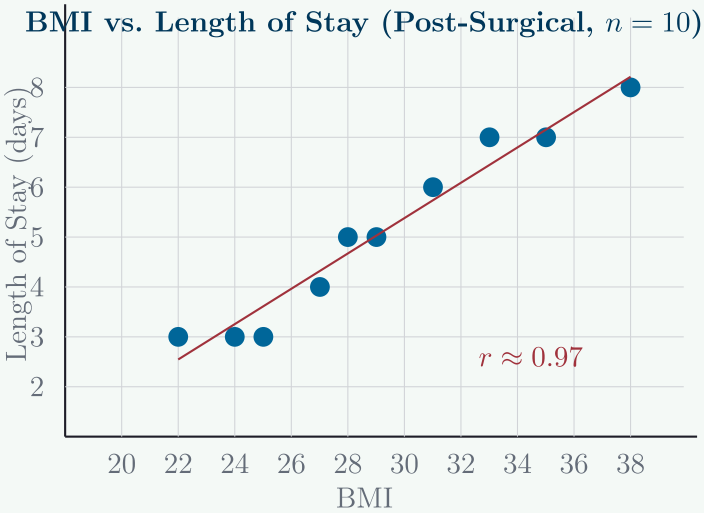
$\lvert r \rvert$	Interpretation	Clinical example
0.00–0.40	Weak	Shoe size and blood pressure
0.40–0.70	Moderate	Nurse age and job satisfaction
0.70–1.00	Strong	Height and weight

r Measures Only Linear Relationships

A correlation of $r = 0$ does not mean there is no relationship — it means there is no *linear* relationship. Two variables can have a strong curved relationship and still have $r \approx 0$. Always look at the scatter plot before interpreting r.

Step 4 Watch It Work

Scatter plot and correlation for the BMI–LOS data



BMI vs. Length of Stay (Post-Surgical, $n = 10$)

What we see:

- **Direction:** Positive — as BMI increases, LOS tends to increase.
- **Form:** Linear — the dots cluster reasonably well around a straight line.
- **Strength:** Strong — $r \approx 0.97$, very close to 1.0.

Try This in Desmos

Enter data as two lists in Desmos Graphing Calculator:

1. Type: `x1 = [22, 27, 31, 25, 35, 29, 38, 24, 33, 28]`

2. Type: `y1 = [3, 4, 6, 3, 7, 5, 8, 3, 7, 5]`

3. Type: `(x1, y1)` to create the scatter plot

4. Type: `corr(x1, y1)` to compute r

Try This in Excel

Enter BMI in column A and LOS in column B.

`=CORREL(A1:A10, B1:B10)` → correlation coefficient r

For a scatter plot: Select both columns → Insert → Charts → Scatter

Step 5 Your Turn

1. For each description, predict the direction of the correlation (positive, negative, or near zero) and estimate the approximate strength (weak, moderate, strong). Justify your answer.

   (a) A patient's age and their systolic blood pressure (in a general adult population)

   (b) The number of cigarettes smoked per day and forced expiratory volume (FEV1, a lung capacity measure)

   (c) A nurse's shoe size and their patient satisfaction scores

   (d) The number of hours of sleep a nurse reports and their self-reported cognitive performance score

   (e) Patient weight (kg) and daily insulin dose (units)

2. A nurse researcher records stress levels (scale 1–10) and hours of overtime worked per week for 8 nurses:

Overtime hrs	2	5	8	3	10	6	4	9
Stress level	3	5	7	4	9	6	4	8

   (a) Plot a sketch of the scatter plot. Label both axes.
   (b) Describe the direction, form, and strength of the relationship.
   (c) Using Desmos or Excel, compute the correlation coefficient r.
   (d) Interpret r in the context of nurse overtime and stress.

3. ⋆ Two scatter plots both show $r = 0.6$. In Plot A, the points form a clean oval cloud tilting upward. In Plot B, the points form a curved arc (like a quadratic). What caution does this illustrate about using r alone?

Step 6 Think Like a Nurse

Correlation Is Not Causation — Again, and in Detail

Researchers publish a study showing $r = 0.81$ between the number of nurses per unit and patient survival rates: units with more nurses have higher survival. The hospital board is impressed and approves a proposal to increase nurse staffing system-wide.

Is the causal claim justified? The correlation is real and strong. But consider what else might differ between well-staffed and understaffed units. Better-staffed units may also have newer equipment, better leadership, more experienced nurses, higher budgets, and lighter patient loads. Any or all of these factors could independently improve survival.

The correlation tells us that staffing and survival tend to move together. It cannot tell us whether changing staffing *causes* the change in survival, or whether both are driven by a third factor (hospital resources, management quality, patient population).

To establish causation, you would need a randomized experiment — which is ethically and practically difficult in this context. In the absence of randomization, strong correlation with a plausible mechanism and the elimination of major confounders is the best available evidence.

A correlation of 0.81 is compelling evidence. It is not proof. The distinction matters when policy decisions affect patient lives.

6.2 The Least-Squares Regression Line

Step 1 Read This First

The hospital team has confirmed a strong positive correlation between BMI and post-surgical LOS ($r = 0.97$). Now they want to go further.

The charge nurse asks: *"If a patient comes in with a BMI of 32, what length of stay should we plan for?"*

The scatter plot shows a strong linear trend, but the dots are not perfectly on a line. There are many lines you could draw through the data. Which one is the best? And once you have it, how do you use it to predict, and how much do you trust those predictions?

These are the questions that the least-squares regression line is designed to answer.

Step 2 Let's Talk About It

Think about what "best fit line" means. For any line you draw, each data point falls some distance above or below the line. Those vertical distances are called *residuals* — they measure how far each actual value is from the predicted value.

A good fit minimizes those residuals. The least-squares criterion minimizes the *sum of the squared residuals* — squaring the distances ensures that negative and positive deviations do not cancel, and penalizes large errors more heavily than small ones.

The result is the one unique line that fits the data better than any other straight line. And from that line, you can substitute any value of x (BMI) to predict the corresponding value of $\hat{y}$ (predicted LOS).

Step 3 Now We Name It

Definition: Least-Squares Regression Line

The **least-squares regression line** (also called the line of best fit) is the line that minimizes the sum of the squared vertical distances between each data point and the line.

Its equation is written as:

$$\hat{y} = b_1 x + b_0$$

where:

- $\hat{y}$ ("y-hat") is the **predicted value** of the response variable
- x is the explanatory variable
- b_1 is the **slope**: the predicted change in y for each one-unit increase in x
- b_0 is the **y-intercept**: the predicted value of y when $x = 0$

The slope and intercept are computed from the data as:

$$b_1 = r \cdot \frac{s_y}{s_x} \qquad\qquad b_0 = \bar{y} - b_1 \bar{x}$$

where s_x and s_y are the sample standard deviations of x and y, and $\bar{x}$, $\bar{y}$ are the sample means.

Definition: Residuals

A **residual** is the difference between an observed value and its predicted value:

$$\text{Residual} = y - \hat{y}$$

A positive residual means the actual value is *above* the regression line. A negative residual means it is *below*. The sum of all residuals equals zero for the least-squares line.

Definition: Coefficient of Determination (r^2)

The **coefficient of determination** r^2 measures the proportion of the variability in y that is explained by the linear relationship with x:

$$r^2 = \frac{\text{variation in } y \text{ explained by } x}{\text{total variation in } y}$$

r^2 ranges from 0 to 1. An r^2 of 0.80 means 80% of the variation in the response variable is explained by the explanatory variable. The remaining 20% is due to other factors or random variation.

Step 4 Watch It Work

Finding and using the regression line for BMI–LOS

From the data: $\bar{x} = 29.2$, $\bar{y} = 5.1$, $s_x \approx 5.07$, $s_y \approx 1.85$, $r = 0.97$

Step 1 — Slope:

$$b_1 = r \cdot \frac{s_y}{s_x} = 0.97 \times \frac{1.85}{5.07} \approx 0.97 \times 0.365 \approx 0.354$$

Step 2 — Intercept:

$$b_0 = \bar{y} - b_1\bar{x} = 5.1 - (0.354)(29.2) = 5.1 - 10.34 \approx -5.24$$

Regression equation:

$$\hat{y} = 0.354x - 5.24$$

Interpreting the equation:

- **Slope (0.354):** For each one-unit increase in BMI, the predicted length of stay increases by approximately **0.35 days**. A patient with a BMI 5 points higher is predicted to stay about 1.8 days longer.
- **Intercept (−5.24):** A patient with BMI = 0 is predicted to have a length of stay of −5.24 days. This is not clinically meaningful — BMI = 0 is impossible. The intercept is a mathematical necessity of the line, not a real prediction.

Making a prediction: For a patient with BMI = 32:

$$\hat{y} = 0.354(32) - 5.24 = 11.33 - 5.24 = \mathbf{6.09} \text{ days}$$

Coefficient of determination: $r^2 = (0.97)^2 = 0.941$
Interpretation: 94.1% of the variation in post-surgical length of stay is explained by the linear relationship with BMI. This is a very strong fit.

Checking a residual — Patient 3 (BMI = 31, actual LOS = 6):

$$\hat{y} = 0.354(31) - 5.24 = 10.97 - 5.24 = 5.73$$

$$\text{Residual} = 6 - 5.73 = +0.27 \text{ days}$$

Patient 3 stayed 0.27 days longer than predicted by the model — a small positive residual.

Try This in Desmos

In Desmos (after entering `x1` and `y1` lists):

Type: `y1 ~ m*x1 + b` → Desmos fits the line and displays m (slope), b (intercept), and r^2

Try This in Excel

Method 1 (formulas):

`=SLOPE(B1:B10, A1:A10)` → slope b_1

`=INTERCEPT(B1:B10, A1:A10)` → intercept b_0

`=RSQ(B1:B10, A1:A10)` → r^2

Method 2 (chart): Right-click on scatter plot data points → Add Trendline → check "Display Equation" and "Display R-squared"

Step 5 Your Turn

1. A study of 8 ICU nurses records average hours of sleep per night and number of self-reported clinical errors in the past month:

Sleep (hrs)	7.0	5.5	6.5	4.5	8.0	6.0	5.0	7.5
Errors	1	4	2	6	0	3	5	1

(a) Using Desmos or Excel, find the correlation coefficient r and the regression equation $\hat{y} = b_1 x + b_0$.
(b) Interpret the slope in context.
(c) Predict the number of errors for a nurse averaging 6.5 hours of sleep.
(d) Compute the residual for the nurse who sleeps 6.5 hours and reports 2 errors.
(e) Interpret r^2 in context.

2. A hospital tracks hand hygiene compliance rate (%) and hospital-acquired infection (HAI) rate (per 1,000 patient-days) across 9 units:

Compliance (%)	72	85	91	68	78	95	82	74	88
HAI rate	4.2	2.8	1.9	5.1	3.5	1.2	3.0	4.0	2.5

(a) Using Desmos or Excel, find r and the regression equation.
(b) What is the direction of the relationship? Is this clinically expected?
(c) Predict the HAI rate for a unit with 80% compliance.
(d) Interpret r^2 in context.
(e) The hospital sets a goal of reducing HAI rate to 1.5 per 1,000 patient-days. What compliance rate does the regression model predict would be needed? Identify any concerns with this extrapolation.

3. ⋆ A researcher plots patient age (x) vs. systolic blood pressure (y) for 200 patients and finds $r = 0.52$ and the regression equation $\hat{y} = 0.65x + 98$.
(a) Interpret the slope in clinical terms.
(b) Predict the systolic BP for a 70-year-old patient. For a newborn (age 0)?

(c) The newborn prediction is clearly wrong. What does this illustrate about the regression model?

(d) $r^2 = 0.27$. What percentage of BP variation is explained by age? What might explain the rest?

(e) A 65-year-old patient has an actual systolic BP of 195 mmHg. Compute the residual. What might this large residual indicate clinically?

Step 6 Think Like a Nurse

Regression Is a Tool for Estimation, Not Prophecy
The regression model predicts that a patient with BMI 32 will have a post-surgical LOS of about 6 days. Should the charge nurse book that bed for exactly 6 days?

No. The regression equation gives the *predicted average* outcome for patients with that BMI — the middle of the range of typical outcomes. Individual patients vary. The model explains 94% of the variation in LOS, which is excellent — but that remaining 6% is real. Some patients with BMI 32 will go home in 4 days. Others may stay 9 days.

The correct use of regression in clinical planning is: *"For a patient with BMI 32, our best estimate is approximately 6 days, but individual variation means we should plan flexibly and reassess daily."*

A second caution: the model was built on patients with BMIs ranging from 22 to 38. Using it to predict for a patient with BMI 50 is **extrapolation** — predicting beyond the range of the data. The linear relationship observed in the data may not continue outside that range. Extrapolation requires caution and clinical judgment.

Regression tells you the expected central tendency. Nursing care requires you to treat the individual, not the average.

6.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 6. Problems marked with ⋆ are more challenging.

Part A: Scatter Plots and Correlation

1. For each pair of variables, state the expected direction of correlation (positive, negative, or near zero) and the approximate strength. Justify each answer in one sentence.

- (a) Number of prenatal visits and birth weight of the baby
- (b) Patient age and recovery time after joint replacement surgery
- (c) Nurse height and patient satisfaction score
- (d) Daily exercise minutes and resting heart rate (in a healthy adult sample)
- (e) Number of medications a patient takes and number of reported drug interactions

2. A quality improvement nurse collects data on response time (minutes from call button press to nurse arrival) and patient fall rate (falls per 100 patient-days) for 7 nursing units:

Response time (min)	3	6	4	8	5	7	2
Fall rate	1.2	3.1	1.8	4.0	2.3	3.6	0.9

- (a) Sketch a scatter plot. Label axes and provide a title.
- (b) Describe the direction, form, and apparent strength.
- (c) Using Desmos or Excel, compute r. Interpret it in context.
- (d) Does this correlation prove that faster response times *cause* lower fall rates? What confounding variables might exist?

Part B: Regression Line and Prediction

3. Use the response time–fall rate data from Problem 2.

- (a) Using Desmos or Excel, find the regression equation $\hat{y} = b_1 x + b_0$.
- (b) Interpret the slope in context.
- (c) Predict the fall rate for a unit with a response time of 5.5 minutes.
- (d) Compute the residual for the unit with a 6-minute response time and a fall rate of 3.1.
- (e) The hospital aims for a fall rate of 1.0 per 100 patient-days. What response time does the model suggest is needed? Are you confident in this prediction?

4. A hospital tracks nursing staff overtime hours per week and medication error rate (errors per 100 administrations) across 6 units:

Overtime hrs/wk	0	4	8	12	16	20
Error rate	0.8	1.0	1.3	1.7	2.2	2.8

(a) Compute r using Desmos or Excel. Is this relationship strong?
(b) Find the regression equation.
(c) Interpret the slope. What does each additional hour of overtime predict?
(d) Predict the error rate at 10 hours of overtime per week.
(e) Interpret r^2. How much of the variation in error rate is explained by overtime?

5. ⋆ A public health study reports a strong positive correlation ($r = 0.88$) between the number of hospitals per county and the county's cancer mortality rate. A reporter writes: "More hospitals mean more cancer deaths. We should close hospitals."
(a) What is wrong with the reporter's conclusion?
(b) Propose at least two confounding variables that might explain the correlation.
(c) What type of study would be needed to support a causal claim about hospitals and mortality?
(d) This is an example of **ecological fallacy** — using group-level data to make claims about individuals. Explain why this reasoning is flawed.

Part C: Interpretation and Critique

6. Identify the error or limitation in each statement:
(a) "The correlation between nurse education level and patient outcomes is $r = 0.45$. This proves that higher education causes better outcomes."

(b) "We found $r = 0.02$ between patient weight and number of medications, so there is no relationship between these variables."

(c) "Our regression equation predicts a patient with BMI 65 will stay 17 days. We should plan accordingly."

(d) "Since $r^2 = 0.64$, the regression line is a perfect fit."

7. ⋆ **Research Application:** A study reports the following findings about a hospital's pain management program:

"We found a significant negative correlation between nurse-administered pain reassessment frequency (per shift) and patient-reported pain scores ($r = -0.71$, $r^2 = 0.50$). The regression equation was $\hat{y} = -1.4x + 8.2$, where x = number of reassessments per shift and y = average pain score."

(a) Interpret $r = -0.71$ in context. Is this a strong relationship?
(b) Interpret the slope (-1.4) in clinical terms.
(c) Interpret $r^2 = 0.50$. What does this tell you about the model?
(d) Predict the average pain score for a nurse who reassesses pain 4 times per

shift.

(e) The study recommends increasing reassessments to 6 per shift. The data range was 1–5 reassessments. What concern does this raise?

(f) Can you conclude from this study that more reassessments *cause* lower pain scores? What would be needed to establish causation?

Answer Key — Selected Problems

Answer Key

Section 6.1 Practice — Your Turn

1. (a) Positive, moderate-to-strong — BP tends to rise with age. (b) Negative, strong — smoking damages lung function. (c) Near zero — no plausible mechanism. (d) Positive, moderate — more sleep generally improves cognitive function. (e) Positive, moderate-to-strong — heavier patients typically need more insulin.

2. (c) Desmos will give $r \approx 0.99$. (d) Very strong positive relationship — nurses working more overtime hours report higher stress levels; 99% of variation in stress is associated with overtime hours.

3. r only measures linear association. Plot B has a strong relationship, but it is curved. Using $r = 0.6$ for Plot B would severely understate the true relationship. Always plot the data first.

Section 6.2 Practice — Your Turn

1. (a) Desmos gives approximately $r \approx -0.98$; $\hat{y} = -1.22x + 9.21$. (b) For each additional hour of sleep, the predicted number of errors decreases by about 1.22. (c) $\hat{y} = -1.22(6.5) + 9.21 = -7.93 + 9.21 = 1.28$ errors predicted. (d) Actual = 2, predicted = 1.28; residual = $2 - 1.28 = +0.72$. (e) $r^2 \approx 0.96$; 96% of the variation in clinical errors is explained by hours of sleep.

2. (a) $r \approx -0.99$; $\hat{y} \approx -0.107x + 12.43$ (values depend on method). (b) Negative — as compliance increases, infection rate decreases; clinically expected. (c) At 80%: $\hat{y} \approx -0.107(80) + 12.43 \approx 3.87$ per 1,000 patient-days. (d) $r^2 \approx 0.98$; 98% of variation in HAI rate is explained by compliance. (e) Set $1.5 = -0.107x + 12.43$, solve for $x \approx 101.7\%$ — impossible compliance; this illustrates the danger of extrapolation beyond the data range.

Chapter Practice — Selected

Problem 2c: Desmos gives approximately $r \approx 0.99$ — very strong positive linear relationship between response time and fall rate.

Problem 4e: $r^2 \approx 0.99$; approximately 99% of the variation in medication error rate is explained by overtime hours — an extremely strong linear fit.

Problem 5 (Challenge): (a) Correlation $\neq$ causation; counties with more hospitals likely have larger, older, or sicker populations. (b) Population size (more people = more hospitals and more deaths); age distribution; urban vs. rural; socioeconomic factors. (c) A randomized experiment where counties are randomly assigned to add or remove hospitals — practically impossible; this is a case where strong observational evidence must suffice. (d) Individual patients did not cause the county-level pattern; the relationship at the aggregate level does not imply the same relationship for individuals.

Problem 7b (Challenge): (a) $r = -0.71$ means a moderate-to-strong negative relationship; as reassessment frequency increases, pain scores tend to decrease. Yes, 0.71

is considered strong. (c) 50% of the variation in pain scores is explained by reassessment frequency — meaningful but not overwhelming; other factors matter too. (d) $\hat{y} = -1.4(4) + 8.2 = -5.6 + 8.2 = 2.6$. (e) Predicting at 6 is extrapolation beyond the data range (1–5); the linear relationship may not continue. (f) No — this is observational. Nurses who reassess more may also be more attentive in general. A randomized trial assigning nurses to different reassessment protocols would be needed.

Chapter 6 Summary

Section 6.1 — Scatter Plots and Correlation

- A **scatter plot** displays the relationship between two quantitative variables; the explanatory variable goes on the x-axis, the response variable on the y-axis.
- Describe scatter plots by **direction** (positive/negative), **form** (linear/curved), and **strength** (strong/moderate/weak).
- The **correlation coefficient** r measures the strength and direction of a linear relationship; $-1 \leq r \leq 1$.
- r only measures *linear* associations. Always plot the data first.
- **Correlation does not imply causation.**

Section 6.2 — Regression Line

- The **least-squares regression line** $\hat{y} = b_1 x + b_0$ minimizes the sum of squared residuals.
- **Slope** (b_1): predicted change in y for each one-unit increase in x.
- **Intercept** (b_0): predicted y when $x = 0$; often not clinically meaningful.
- **Residual** $= y - \hat{y}$: the difference between observed and predicted values.
- r^2: proportion of variation in y explained by the linear relationship with x.
- **Extrapolation** (predicting beyond the data range) is unreliable — use with caution.

The Nursing Connection

- Correlations between clinical variables (BMI and LOS, compliance and infection rates) guide planning but do not establish cause.
- Regression predictions describe the expected average for a group; individual patients vary around that prediction.
- A large residual signals that a patient deviates meaningfully from the expected pattern — worth clinical attention.
- r^2 tells you how useful the model is; a low r^2 means other variables matter more.

CHAPTER 7

Introduction to Probability

"Every clinical decision is a bet. The question is whether you know the odds."

— on why probability is the language of evidence-based practice

In this chapter, you will learn to:

- Define probability and explain the three approaches to computing it
- Apply the Addition Rule for probability (with and without overlap)
- Apply the Multiplication Rule for independent and dependent events
- Compute and interpret conditional probability
- Construct and use two-way tables for probability problems
- Compute sensitivity, specificity, positive predictive value, and negative predictive value
- Explain how disease prevalence affects the clinical value of a positive or negative test result

7.1 Probability Basics and the Addition Rule

Step 1 Read This First

You are a nurse in the emergency department. A patient arrives with chest pain. Based on your clinical experience and the population of patients who come to your ED with chest pain, you estimate there is a 30% chance this patient is having a myocardial infarction (MI).

What does that 30% actually mean? It does not mean this specific patient is 30% of the way through a heart attack. It means: if you saw 100 patients who looked exactly like this one — same age, same symptoms, same risk factors — about 30 of them would be having an MI.

Probability is always a statement about a *population of similar cases*, not a guarantee about one individual. And in nursing, understanding that distinction is the difference between using evidence well and misapplying it.

Step 2 Let's Talk About It

Think about how you already use probability without naming it.

When you read that a medication causes nausea in 15% of patients, you know that means most patients will be fine but some will not. When the pharmacy labels a drug "high alert," it means the probability of serious harm if given incorrectly is much higher than for other drugs.

When a physician orders a test, they are implicitly performing a probability calculation: how likely is this patient to have the disease, and how much will the test result change that estimate?

All of this is probability reasoning. This chapter gives it precise language and reliable rules — so that instead of relying on intuition, you can compute.

Step 3 Now We Name It

Definition: Probability

The **probability** of an event E is a number between 0 and 1 that represents the likelihood of the event occurring:

$$0 \le P(E) \le 1$$

$P(E) = 0$: the event is impossible. $P(E) = 1$: the event is certain.

Three approaches:

- **Classical:** Used when all outcomes are equally likely. $P(E) = \dfrac{\text{number of favorable outcomes}}{\text{total number of outcomes}}$
- **Empirical (relative frequency):** Based on observed data. $P(E) = \dfrac{\text{number of times } E \text{ occurred}}{\text{total number of trials}}$
- **Subjective:** Based on expert judgment or experience. Used when neither classical nor empirical is available.

Definition: Complement Rule

The **complement** of event E, written E^c (or $\bar{E}$), is the event that E does *not* occur.

$$P(E^c) = 1 - P(E)$$

Example: If the probability of a surgical complication is 0.08, then the probability of no complication is $1 - 0.08 = 0.92$.

Definition: Addition Rule

For any two events A and B:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are **mutually exclusive** (cannot both occur at the same time):

$$P(A \text{ or } B) = P(A) + P(B)$$

Example of mutually exclusive events: A patient's blood type is A or B (cannot be both simultaneously).

Step 4 Watch It Work

Probability in an Emergency Department

A busy ED tracks the primary complaints of 200 patients during one week:

Primary Complaint	Frequency
Chest pain	42
Abdominal pain	38
Shortness of breath	29
Injury/Trauma	51
Fever	24
Other	16
Total	**200**

Problem 1: What is the probability that a randomly selected patient has chest pain?

$$P(\text{chest pain}) = \frac{42}{200} = 0.210$$

Problem 2: What is the probability a patient presents with chest pain *or* shortness of breath?
Since these are mutually exclusive (each patient has one primary complaint):

$$P(\text{chest pain or SOB}) = \frac{42}{200} + \frac{29}{200} = \frac{71}{200} = 0.355$$

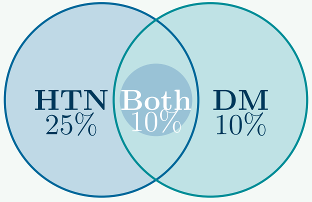
Problem 3: What is the probability a patient does *not* present with injury or trauma?

$$P(\text{not trauma}) = 1 - P(\text{trauma}) = 1 - \frac{51}{200} = 1 - 0.255 = 0.745$$

A general Addition Rule example:
At another hospital, 35% of patients have hypertension, 20% have diabetes, and 10% have both. What is the probability that a randomly selected patient has hypertension *or* diabetes?

$$P(H \text{ or } D) = P(H) + P(D) - P(H \text{ and } D) = 0.35 + 0.20 - 0.10 = 0.45$$

Without subtracting the overlap, you would count the 10% with both conditions twice.



$P(H \text{ or } D) = 0.25 + 0.10 + 0.10 = 0.45$

Step 5 Your Turn

1. A hospital reports that among 500 patients discharged last month:

- 120 had a surgical procedure
- 85 had a cardiac diagnosis
- 40 had both a surgical procedure and a cardiac diagnosis

(a) What is the probability a randomly selected patient had a surgical procedure?

(b) What is the probability a patient had a surgical procedure *or* a cardiac diagnosis?

(c) What is the probability a patient had *neither* a surgical procedure nor a cardiac diagnosis?

(d) Are surgical procedure and cardiac diagnosis mutually exclusive? How do you know?

2. A nurse knows that among patients with sepsis, 22% develop acute kidney injury (AKI). Among patients *without* sepsis, 5% develop AKI.

(a) If a patient has sepsis, what is the probability they do *not* develop AKI?

(b) If 60% of ICU patients have sepsis, what is the probability that a randomly selected ICU patient has sepsis *and* does not develop AKI? (Hint: use the Multiplication Rule from Section 7.2.)

3. $\star$ A study of 300 post-operative patients finds: 45 developed infection, 30 had a prolonged stay (> 7 days), and 15 had both.

(a) Find P(infection or prolonged stay).

(b) Find P(infection only, not prolonged stay).

(c) Would it be appropriate to use the simpler mutually exclusive Addition Rule here? Why or why not?

Step 6 Think Like a Nurse

Why "Rare" Does Not Mean "Impossible" at Scale

A serious medication adverse reaction has a probability of 0.002 — two in every thousand patients. That sounds negligible. A single nurse might work an entire career without seeing it.

But now think at a system level. A hospital administers 50,000 doses of this medication per year. The expected number of adverse reactions is:

$$0.002 \times 50{,}000 = 100 \text{ reactions per year}$$

That is not rare at a system level. That is roughly two per week. This is why hospital pharmacies track adverse reaction rates even for "rare" events, why drug safety monitoring systems exist, and why statistical thinking is essential for healthcare policy.

The complement rule makes this precise. If the probability of *no* adverse reaction for a single dose is 0.998, the probability of at least one adverse reaction occurring somewhere in 50,000 doses is:

$$1 - (0.998)^{50{,}000} \approx 1 - 1.8 \times 10^{-44} \approx 1.0$$

At that scale, it is essentially certain that some patients will be harmed.

Probability thinking scales. What is rare for one patient is predictable for a population. That is the core argument for population-level safety protocols.

7.2 Conditional Probability and the Multiplication Rule

Step 1 Read This First

You are reviewing records from 400 patients who were screened for depression using the PHQ-9 tool. You have two pieces of information about each patient: whether they tested positive on the PHQ-9 screen, and whether they were later confirmed to have clinical depression by a psychiatrist.

	Confirmed Depression	No Depression	Total
PHQ-9 Positive	80	20	100
PHQ-9 Negative	30	270	300
Total	110	290	400

A patient screens positive on the PHQ-9. What is the probability that this patient *actually* has depression? This is a different question from "what is the probability of a positive screen among all patients." The answer has to account for the fact that we already know the screen result.

This is a **conditional probability** question.

Step 2 Let's Talk About It

The key shift in conditional probability is this: once you know something has already happened, the sample space shrinks.

Before the screen result, there are 400 patients in our universe. After learning the screen is positive, there are only 100 patients in our universe (the PHQ-9 positive row). Of those 100, we want to know how many have confirmed depression: 80. So the probability is $80/100 = 0.80$.

Notice how this differs from $P(\text{depression}) = 110/400 = 0.275$ — the unconditional probability of depression in the whole sample. The screen result changed our estimate dramatically.

This updating of probability based on new information is the foundation of clinical reasoning. Every test result, every symptom, every risk factor you learn about a patient changes your probability estimate. Statistics gives this process a precise name and formula.

Step 3 Now We Name It

Definition: Conditional Probability

The **conditional probability** of event A given that event B has already occurred is:

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

Read $P(A \mid B)$ as "the probability of A given B."

Clinical meaning: If we know a patient tests positive (event B), what is the probability they truly have the disease (event A)?

Definition: Multiplication Rule

For any two events A and B:

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

If A and B are **independent** (knowing A occurred does not change the probability of B):

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Two events are independent if $P(B \mid A) = P(B)$.

Definition: Diagnostic Test Measures

Given a two-way table with disease status (rows) and test result (columns):

	Disease Present	**Disease Absent**
Test Positive	True Positive (TP)	False Positive (FP)
Test Negative	False Negative (FN)	True Negative (TN)

Sensitivity $= \dfrac{TP}{TP + FN}$ (probability test is positive given disease is present)

Specificity $= \dfrac{TN}{TN + FP}$ (probability test is negative given disease is absent)

Positive Predictive Value (PPV) $= \dfrac{TP}{TP + FP}$ (probability of disease given positive test)

Negative Predictive Value (NPV) $= \dfrac{TN}{TN + FN}$ (probability of no disease given negative test)

Sensitivity vs. PPV — The Critical Distinction

Sensitivity answers: Among patients who *have* the disease, how often does the test catch it?
PPV answers: If a patient *tests positive*, how likely are they to actually have the disease?

A test can have high sensitivity but low PPV if the disease is rare. This is the core of why mass screening programs require careful analysis.

Step 4 Watch It Work

Diagnostic measures for the PHQ-9 depression screening data

	Depression	**No Depression**	**Total**
PHQ-9 Positive	80 (TP)	20 (FP)	100
PHQ-9 Negative	30 (FN)	270 (TN)	300
Total	110	290	400

Sensitivity (probability of positive screen given depression):

$$P(\text{Pos} \mid \text{Dep}) = \frac{TP}{TP + FN} = \frac{80}{80 + 30} = \frac{80}{110} \approx 0.727$$

Specificity (probability of negative screen given no depression):

$$P(\text{Neg} \mid \text{No Dep}) = \frac{TN}{TN + FP} = \frac{270}{270 + 20} = \frac{270}{290} \approx 0.931$$

PPV (probability of depression given positive screen):

$$P(\text{Dep} \mid \text{Pos}) = \frac{TP}{TP + FP} = \frac{80}{80 + 20} = \frac{80}{100} = 0.800$$

NPV (probability of no depression given negative screen):

$$P(\text{No Dep} \mid \text{Neg}) = \frac{TN}{TN + FN} = \frac{270}{270 + 30} = \frac{270}{300} = 0.900$$

Interpreting these results clinically:

Measure	**Value**	**Clinical meaning**
Sensitivity	72.7%	Among depressed patients, the PHQ-9 correctly identifies 72.7%. About 27% of depressed patients will be missed (false negatives).
Specificity	93.1%	Among non-depressed patients, the PHQ-9 correctly clears 93.1%. About 7% will be flagged unnecessarily (false positives).
PPV	80.0%	If a patient screens positive, there is an 80% chance they truly have depression.
NPV	90.0%	If a patient screens negative, there is a 90% chance they truly do not have depression.

Conditional probability using the formula:

$$P(\text{Depression and Positive}) = \frac{80}{400} = 0.200$$

$$P(\text{Dep} \mid \text{Pos}) = \frac{P(\text{Dep and Pos})}{P(\text{Pos})} = \frac{0.200}{0.250} = 0.800 \quad \checkmark$$

Step 5 Your Turn

1. A hospital tracks fall risk assessment results and whether patients actually fell during hospitalization for 600 patients:

	Fell	**Did Not Fall**	**Total**
High Risk (assessed)	48	72	120
Low Risk (assessed)	18	462	480
Total	66	534	600

(a) Compute the sensitivity of the fall risk assessment.
(b) Compute the specificity.
(c) Compute the PPV and interpret it in clinical terms.
(d) Compute the NPV and interpret it in clinical terms.
(e) If a patient is assessed as high risk, what is the probability they will fall? Does this mean the assessment is not useful?

2. Two events: A = patient has diabetes, B = patient develops a foot wound.
(a) If $P(A) = 0.12$ and $P(B \mid A) = 0.25$, find $P(A \text{ and } B)$.
(b) If $P(B) = 0.04$ in the general population, does knowing a patient has diabetes change the probability of a foot wound? Are A and B independent?
(c) Find $P(B \mid A^c)$ if $P(A^c \text{ and } B) = 0.035$ and $P(A^c) = 0.88$.

3. ⋆ A COVID-19 rapid antigen test has sensitivity 85% and specificity 97%. In a community where the current prevalence (base rate) of active COVID is 8%:
(a) Build a two-way table for 10,000 hypothetical people.
(b) Compute PPV: among those who test positive, what fraction actually have COVID?
(c) Compute NPV: among those who test negative, what fraction are truly COVID-free?
(d) Repeat (b) and (c) with a prevalence of 40% (e.g., during a surge). How do PPV and NPV change?
(e) What does this illustrate about how the value of a test depends on the context in which it is used?

Step 6 Think Like a Nurse

The Base Rate Problem: Why a Positive Test Is Not the Same as Having the Disease

A hospital introduces mandatory screening for a rare but serious bloodstream infection. The test has excellent performance: 95% sensitivity and 99% specificity. A nurse's patient tests positive. The nurse tells the family: "The test is 95% accurate, so your father almost certainly has this infection."

But the infection is rare — it affects 1 in 1,000 patients in this hospital. Let us build the two-way table for 100,000 patients:

	Infected	**Not Infected**	**Total**
Test Positive	95 (TP)	999 (FP)	1,094
Test Negative	5 (FN)	98,901 (TN)	98,906
Total	100	99,900	100,000

PPV $= 95/1{,}094 \approx 8.7\%$

Despite the test's impressive sensitivity and specificity, a positive result means there is only an **8.7% chance** this patient actually has the infection. The nurse's reassurance was statistically incorrect — and may lead to unnecessary treatment, patient anxiety, and resource use.

The lesson: test performance statistics (sensitivity, specificity) describe the test in isolation. PPV and NPV describe the test in the context of the population being screened. When the disease is rare, even excellent tests produce mostly false positives.

Before interpreting any positive test result, ask: how common is this condition in patients like this one? The answer changes everything.

7.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 7. Problems marked with ⋆ are more challenging.

Part A: Basic Probability and Addition Rule

1. A nursing unit tracks adverse events over one quarter. Of 240 incidents: 72 were medication errors, 48 were patient falls, 36 were pressure injuries, 24 were IV complications, and 60 were classified as "other."

(a) What is the probability that a randomly selected incident was a medication error?
(b) What is the probability that an incident was a fall or a pressure injury?
(c) What is the probability an incident was *not* a medication error?
(d) Are fall and medication error mutually exclusive events? Explain.

2. In a hospital, 40% of patients have hypertension (H), 25% have obesity (O), and 15% have both.

(a) Find $P(H \text{ or } O)$.
(b) Find $P(\text{neither } H \text{ nor } O)$.
(c) Are hypertension and obesity independent? Explain using conditional probability.

Part B: Conditional Probability and Multiplication Rule

3. A study of 500 post-surgical patients tracked whether they received enhanced recovery protocols (ERP) and whether they were readmitted within 30 days:

	Readmitted	**Not Readmitted**	**Total**
ERP	18	232	250
No ERP	42	208	250
Total	60	440	500

(a) What is the probability of readmission for a patient who received ERP?
(b) What is the probability of readmission for a patient who did not receive ERP?
(c) Is readmission independent of receiving ERP? Show your reasoning.
(d) Find $P(\text{ERP and Readmitted})$ two ways: directly from the table, and using the Multiplication Rule.

4. A medication is effective in 70% of patients with a specific infection. If two patients with the infection are treated independently:

(a) What is the probability both patients respond to the medication?
(b) What is the probability neither patient responds?

(c) What is the probability at least one patient responds? (Use the complement rule.)

Part C: Diagnostic Testing

5. A new rapid strep test is evaluated in a clinic. Results for 200 patients:

	Strep Confirmed	**No Strep**	**Total**
Test Positive	54	12	66
Test Negative	6	128	134
Total	60	140	200

(a) Compute sensitivity, specificity, PPV, and NPV.

(b) Interpret each value in one clinical sentence.

(c) A parent asks: "The test was positive. Does my child have strep?" What would you tell them?

(d) Which is more costly in a pediatric setting: a false positive or a false negative? Explain.

6. ⋆ A hospital introduces a screening test for hospital-acquired pneumonia (HAP) with sensitivity 90% and specificity 85%. HAP affects 6% of ventilated patients.

(a) Build a two-way table for 1,000 ventilated patients.

(b) Compute PPV and NPV.

(c) A ventilated patient tests positive. Should the care team initiate antibiotic treatment immediately? What factors should they consider?

(d) If the specificity were increased to 95% (at the cost of reducing sensitivity to 80%), how would PPV and NPV change? Which change matters more clinically?

7. ⋆ **Research Application:** A meta-analysis reports that a biomarker test for sepsis has sensitivity 88% and specificity 76%. In a medical ICU, the baseline rate of sepsis is 30%.

(a) Build a two-way table for 10,000 ICU patients.

(b) Compute PPV. Interpret: if a patient tests positive, how confident should the team be in a sepsis diagnosis?

(c) Compute NPV. What does a negative test tell the team?

(d) The same test is used in a general hospital ward where sepsis prevalence is 5%. How does PPV change?

(e) What does the comparison between (b) and (d) tell you about the importance of pre-test probability in clinical decision making?

Answer Key — Selected Problems

Answer Key

Section 7.1 Practice — Your Turn

1. (a) $P(\text{surgical}) = 120/500 = 0.240$. (b) $P(\text{surgical or cardiac}) = 120/500 + 85/500 - 40/500 = 165/500 = 0.330$. (c) $P(\text{neither}) = 1 - 0.330 = 0.670$. (d) Not mutually exclusive — 40 patients had both, so both can occur simultaneously.

2. (a) $P(\text{no AKI} \mid \text{sepsis}) = 1-0.22 = 0.78$. (b) $P(\text{sepsis and no AKI}) = P(\text{sepsis}) \times P(\text{no AKI} \mid \text{sepsis}) = 0.60 \times 0.78 = 0.468$.

3. (a) $P(\text{infection or prolonged}) = 45/300 + 30/300 - 15/300 = 60/300 = 0.200$. (b) $P(\text{infection only}) = 45/300 - 15/300 = 30/300 = 0.100$. (c) No — 15 patients had both, so these events overlap and are not mutually exclusive.

Section 7.2 Practice — Your Turn

1. (a) Sensitivity $= 48/66 \approx 0.727$ (72.7%). (b) Specificity $= 462/534 \approx 0.865$ (86.5%). (c) PPV $= 48/120 = 0.400$ (40%) — only 40% of high-risk patients actually fell; however the assessment still provides useful risk stratification. (d) NPV $= 462/480 = 0.963$ (96.3%) — if assessed low risk, there is a 96.3% chance the patient will not fall. (e) $P(\text{fall} \mid \text{high risk}) = 0.400$; this does not mean the tool is useless — it identified a group with much higher fall probability than the overall rate of $66/600 = 11\%$.

2. (a) $P(A \text{ and } B) = 0.12 \times 0.25 = 0.030$. (b) $P(B \mid A) = 0.25 \neq P(B) = 0.04$, so A and B are *not* independent; diabetes substantially increases foot wound risk. (c) $P(B \mid A^c) = 0.035/0.88 \approx 0.040$.

3. (a) Table for 10,000: True positives $= 0.85 \times 800 = 680$; False positives $= 0.03 \times 9{,}200 = 276$; False negatives $= 0.15 \times 800 = 120$; True negatives $= 0.97 \times 9{,}200 = 8{,}924$. (b) PPV $= 680/(680{+}276) = 680/956 \approx 71.1\%$. (c) NPV $= 8{,}924/(8{,}924{+}120) = 8{,}924/9{,}044 \approx 98.7\%$. (d) At 40% prevalence: TP $= 3{,}400$, FP $= 180$, FN $= 600$, TN $= 5{,}820$; PPV $= 3{,}400/3{,}580 \approx 94.9\%$; NPV $= 5{,}820/6{,}420 \approx 90.7\%$. PPV rises dramatically; NPV falls. (e) A test's clinical usefulness depends heavily on disease prevalence; the same test performs very differently in a low-prevalence vs. high-prevalence setting.

Chapter Practice — Selected

Problem 1: (a) $72/240 = 0.300$. (b) $(48 + 36)/240 = 84/240 = 0.350$. (c) $1 - 0.300 = 0.700$. (d) Yes — a single incident is classified as one type; a fall cannot simultaneously be a medication error.

Problem 2: (a) $P(H \text{ or } O) = 0.40 + 0.25 - 0.15 = 0.50$. (b) $1 - 0.50 = 0.50$. (c) If independent, $P(H \text{ and } O) = 0.40 \times 0.25 = 0.10 \neq 0.15$; since the joint probability exceeds the independent prediction, H and O are positively associated (not independent).

Problem 5 (Challenge): (a) TP $= 0.90 \times 60 = 54$; FP $= 0.15 \times 940 = 141$; FN

$= 0.10 \times 60 = 6$; TN $= 0.85 \times 940 = 799$. (b) PPV $= 54/(54 + 141) = 54/195 \approx 27.7\%$; NPV $= 799/(799 + 6) = 799/805 \approx 99.3\%$. (c) PPV is only 27.7% — most positives are false positives; the team should consider clinical signs, cultures, and other evidence before initiating antibiotics. (d) Increasing specificity to 95% with sensitivity at 80%: TP $= 0.80 \times 60 = 48$; FP $= 0.05 \times 940 = 47$; FN $= 0.20 \times 60 = 12$; TN $= 0.95 \times 940 = 893$; PPV $= 48/(48 + 47) = 48/95 \approx 50.5\%$ (better than 27.7%); NPV $= 893/(893 + 12) = 893/905 \approx 98.7\%$ (essentially unchanged). In this context, improving PPV reduces unnecessary antibiotic use — the more important concern given antibiotic stewardship goals.

Chapter 7 Summary

Section 7.1 — Probability Basics and the Addition Rule

- $P(E)$ is a number from 0 to 1 representing the likelihood of event E.
- Three approaches: classical, empirical (relative frequency), subjective.
- **Complement Rule:** $P(E^c) = 1 - P(E)$.
- **Addition Rule:** $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.
- If A and B are mutually exclusive: $P(A \text{ or } B) = P(A) + P(B)$.

Section 7.2 — Conditional Probability and Multiplication Rule

- **Conditional probability:** $P(A \mid B) = P(A \text{ and } B)/P(B)$.
- **Multiplication Rule:** $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$.
- If independent: $P(A \text{ and } B) = P(A) \cdot P(B)$.
- **Sensitivity** $= P(\text{positive} \mid \text{disease})$; **Specificity** $= P(\text{negative} \mid \text{no disease})$.
- **PPV** $= P(\text{disease} \mid \text{positive})$; **NPV** $= P(\text{no disease} \mid \text{negative})$.
- PPV and NPV depend on disease prevalence — the same test performs differently in different populations.

The Nursing Connection

- Every test result updates your probability estimate of disease — this is Bayesian reasoning in clinical practice.
- A rare disease and a highly sensitive test can still produce mostly false positives (low PPV).
- Probability scales: what is rare per patient is predictable per population — the basis for system-level safety.
- Before acting on a test result, always ask: what is the pre-test probability of disease in this patient?

CHAPTER 8

Probability Distributions

"Not all outcomes are equally likely. A probability distribution tells you exactly how the possibilities are weighted."

— the mathematical model behind every clinical prediction

In this chapter, you will learn to:

- Define a discrete probability distribution and verify its properties
- Compute the mean (expected value) and standard deviation of a discrete distribution
- Identify situations that satisfy the binomial setting
- Compute binomial probabilities using the formula and Desmos
- Find the mean and standard deviation of a binomial distribution
- Apply binomial distributions to clinical quality and safety scenarios

8.1 Discrete Probability Distributions

Step 1 Read This First

A nurse manager reviews patient fall data for her 12-bed unit. Based on three years of records, she has compiled the following information about the number of falls that occur per week:

Falls per week (x)	0	1	2	3	4	5
Relative frequency	0.30	0.35	0.20	0.10	0.04	0.01

She wants to answer several practical questions:

- On a typical week, how many falls should she plan for?
- How much does the fall count vary from week to week?
- What is the probability of having *no* falls in a given week?
- What is the probability of having *three or more* falls?

The table above is a **probability distribution**. It is the tool that answers all of these questions in a single, organized framework.

Step 2 Let's Talk About It

Think about what makes this table a model rather than just data.

The relative frequencies came from historical records, but once we accept them as the probability of each outcome, we are making a model: we are saying "this is how the world works on this unit." The table is compact, it covers all possibilities, and the probabilities sum to 1.

From this table, we can do something powerful: instead of waiting to observe more falls and counting up, we can *calculate* what to expect. The mean — how many falls per week on average — can be computed directly from the distribution. So can the standard deviation.

This is the shift from describing data (Chapter 5) to modeling outcomes (this chapter). The distribution is the model; the mean and standard deviation are summaries of the model.

Step 3 Now We Name It

Definition: Discrete Probability Distribution

A **discrete probability distribution** lists all possible values of a discrete random variable X and the probability of each value. It must satisfy two conditions:

1. $0 \leq P(x) \leq 1$ for every value x
2. $\sum P(x) = 1$ (all probabilities sum to exactly 1)

Definition: Mean (Expected Value) of a Discrete Distribution

The **mean** of a discrete random variable, also called the **expected value**, is:

$$\mu = E(X) = \sum x \cdot P(x)$$

This is the long-run average value of X — what you would expect to observe per trial if you repeated the experiment many times. It does *not* have to be a whole number, even if X only takes whole number values.

Definition: Standard Deviation of a Discrete Distribution

The **standard deviation** of a discrete random variable measures the typical distance of X from its mean:

$$\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)}$$

A larger σ means the outcomes are more spread out (more variable) around the mean.

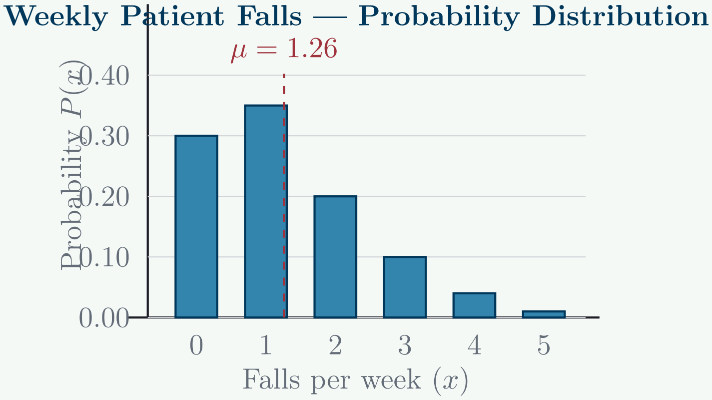
Step 4 Watch It Work

Computing the mean and standard deviation for the falls distribution

x	$P(x)$	$x \cdot P(x)$	$x - \mu$	$(x-\mu)^2$	$(x-\mu)^2 \cdot P(x)$
0	0.30	0.00	-1.26	1.588	0.476
1	0.35	0.35	-0.26	0.068	0.024
2	0.20	0.40	$+0.74$	0.548	0.110
3	0.10	0.30	$+1.74$	3.028	0.303
4	0.04	0.16	$+2.74$	7.508	0.300
5	0.01	0.05	$+3.74$	13.988	0.140
Total	**1.00**	$\boldsymbol{\mu = 1.26}$			$\boldsymbol{\sigma^2 = 1.353}$

$\mu = 1.26$ falls per week $\sigma = \sqrt{1.353} \approx 1.16$ falls per week

Visual: Probability distribution bar chart



Interpreting the results:

- On average, the unit expects **1.26 falls per week**.
- Falls vary by about **1.16 falls per week** around that mean.
- $P(x = 0) = 0.30$ — there is a 30% chance of a fall-free week.
- $P(x \geq 3) = 0.10 + 0.04 + 0.01 = 0.15$ — a 15% chance of three or more falls.

Step 5 Your Turn

1. A hospital pharmacy tracks the number of prescription errors caught by the verification system per shift. Based on the past year:

Errors per shift (x)	0	1	2	3	4
$P(x)$	0.25	0.40	0.20	0.10	0.05

(a) Verify that this is a valid probability distribution.
(b) Compute μ, the expected number of errors per shift.
(c) Compute σ, the standard deviation.
(d) What is the probability of catching *more than two* errors in a shift?
(e) Interpret μ and σ together in a sentence for the pharmacy director.

2. A home health agency knows from experience that when a nurse makes a home visit, the number of tasks requiring unplanned follow-up is:

Tasks (x)	0	1	2	3	4
$P(x)$	0.30	0.35	0.22	0.08	?

(a) What must $P(x = 4)$ equal? How do you know?
(b) Compute the expected number of follow-up tasks per visit.
(c) If the agency schedules 20 home visits on Tuesday, how many follow-up tasks should they expect in total that day?
(d) If the standard deviation is 1.02, what interval covers $\mu \pm 2\sigma$? What does this mean practically?

Step 6 Think Like a Nurse

Expected Value Is Not Always the Most Likely Outcome

The falls distribution has a mean of 1.26 falls per week. But look at the distribution: $P(x = 1) = 0.35$ is the single most likely outcome, and $P(x = 0) = 0.30$ is close behind. A week with exactly 1.26 falls is impossible — falls are whole numbers.

The expected value is a mathematical average over many weeks — the long-run mean. In any single week, the unit will have 0, 1, 2, or more falls, never 1.26. This matters when communicating with administrators.

Saying “we average 1.26 falls per week” is honest. Saying “we expect 1.26 falls next week” can be misunderstood. The expected value describes a pattern over time, not a prediction for a specific occasion.

This is also why the standard deviation matters. Reporting only the mean suggests the unit always has about 1.26 falls. Reporting mean 1.26, SD 1.16 communicates that in many weeks there will be zero falls, and in some weeks there will be three or four. The full picture is more actionable than the single number.

Expected value describes the center of a long-run pattern. Standard deviation describes how far individual weeks can stray from that center. Both are needed for honest planning.

8.2 The Binomial Distribution

Step 1 Read This First

A hospital tracks medication administration errors. Based on audit data, any given medication administration has a 2% chance of involving an error. On a busy shift, a nurse administers 15 medications.

Questions the charge nurse needs to answer:

- What is the probability of *zero* errors in those 15 administrations?
- What is the probability of *exactly one* error?
- What is the probability of *two or more* errors?
- How many errors should be expected per shift, on average?

Each administration is independent (one administration does not affect the next), there are only two outcomes for each (error or no error), and the probability of error is constant at 2%. This is a **binomial setting**, and the **binomial distribution** gives exact answers to all four questions.

Step 2 Let's Talk About It

Think about counting errors across 15 administrations.

The first question — zero errors — means every single administration must be correct. The probability of that is $(0.98)^{15}$, since each of the 15 independent administrations must avoid error.

The second question — exactly one error — is harder. The error could occur on the 1st administration, or the 2nd, or any of the 15. For each of those 15 cases, the error occurs once (probability 0.02) and the other 14 administrations are correct (probability 0.98^{14}). So we multiply by 15.

The third question — two or more errors — would require adding up all the cases with 2, 3, 4, ..., 15 errors. That is tedious. The complement is much easier: $P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$.

This counting logic is exactly what the binomial formula does — it packages all the cases systematically.

Step 3 Now We Name It

Definition: The Binomial Setting (BINS)

A random variable X follows a **binomial distribution** when all four conditions hold:

- **B**inary outcomes: each trial results in one of exactly two outcomes ("success" or "failure")
- **I**ndependent trials: the outcome of one trial does not affect any other trial
- **N**umber of trials is fixed: n trials are performed in advance
- **S**ame probability: the probability of success p is the same for every trial

We write $X \sim B(n, p)$ to mean "X follows a binomial distribution with n trials and success probability p."

Definition: Binomial Probability Formula

The probability of exactly k successes in n trials is:

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

where:

- $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$ is the number of ways to choose k successes from n trials
- p^k is the probability of exactly k successes
- $(1-p)^{n-k}$ is the probability of the remaining $n-k$ failures

Definition: Mean and Standard Deviation of a Binomial Distribution

If $X \sim B(n, p)$, then:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

These formulas are derived from the general discrete distribution formulas but specialized for the binomial structure. No table needed — just multiply.

Binomial vs. Not Binomial

Binomial: Flipping a coin 10 times and counting heads; screening 50 patients for hypertension and counting positives; administering 15 medications and counting errors.

Not binomial: Drawing cards without replacement (probabilities change); measuring blood pressure (not binary); counting the number of days until the first fall (not a fixed number of trials).

Step 4 Watch It Work

Solving the medication error problem: $X \sim B(15, 0.02)$

$n = 15$ administrations, $p = 0.02$ (probability of error per administration)

Problem 1 — $P(X = 0)$: probability of zero errors

$$P(X=0) = \binom{15}{0}(0.02)^0(0.98)^{15} = 1 \times 1 \times (0.98)^{15} \approx \mathbf{0.7386}$$

Problem 2 — $P(X = 1)$: probability of exactly one error

$$P(X=1) = \binom{15}{1}(0.02)^1(0.98)^{14} = 15 \times 0.02 \times (0.98)^{14} \approx 15 \times 0.02 \times 0.7536 \approx \mathbf{0.2261}$$

Problem 3 — $P(X \geq 2)$: probability of two or more errors

$$P(X \geq 2) = 1 - P(X=0) - P(X=1) = 1 - 0.7386 - 0.2261 = \mathbf{0.0353}$$

Problem 4 — Mean and standard deviation:

$$\mu = np = 15 \times 0.02 = \mathbf{0.30} \text{ errors per shift}$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{15 \times 0.02 \times 0.98} = \sqrt{0.294} \approx \mathbf{0.542}$$

Interpreting these results:

- There is a **73.9% chance** of a completely error-free shift.
- There is a **22.6% chance** of exactly one error — the most likely non-zero outcome.
- There is only a **3.5% chance** of two or more errors in a shift.
- On average, **0.30 errors per shift** are expected — meaning about 1 error every 3–4 shifts.

What if the error rate doubled to 4%?

$$\mu = 15 \times 0.04 = 0.60 \quad P(X=0) = (0.96)^{15} \approx 0.5421 \quad P(X \geq 2) \approx 0.1191$$

Doubling the error rate more than triples the probability of two or more errors (from 3.5% to 11.9%). Small changes in the per-administration error rate have outsized effects on multi-administration shifts.

Try This in Desmos

In Desmos Scientific Calculator:

$P(X = k)$: type `nCr(15, 0) * 0.02^0 * 0.98^15` (adjust k and n as needed)

Or for cumulative: `binomialdist(15, 0.02)` creates the distribution, then use `cdf` to find $P(X \leq k)$.

Try This in Excel

`=BINOM.DIST(k, n, p, FALSE)` → $P(X = k)$ (exact probability)

`=BINOM.DIST(k, n, p, TRUE)` → $P(X \leq k)$ (cumulative)

Example: `=BINOM.DIST(0, 15, 0.02, FALSE)` returns 0.7386

Step 5 Your Turn

1. A hospital reports that 8% of patients who receive a specific surgical procedure develop a post-operative infection. A surgeon performs 10 of these procedures in one week.
Let X = number of patients who develop a post-operative infection.
- (a) Does X satisfy the binomial conditions? Verify all four (BINS).
- (b) Find $P(X = 0)$: the probability no patients develop an infection.
- (c) Find $P(X = 1)$.
- (d) Find $P(X \geq 2)$ using the complement rule.
- (e) Find the expected number of infections and the standard deviation.
- (f) If the infection rate were reduced to 4% through a new sterilization protocol, how would the expected number of infections change? How would $P(X = 0)$ change?

2. A nursing licensure exam has a pass rate of 88% on the first attempt. A cohort of 20 nursing graduates takes the exam.
- (a) Find $P(X = 20)$: all 20 pass.
- (b) Find $P(X \geq 18)$: 18 or more pass. (Compute each and add.)
- (c) Find the expected number of first-time passers and the standard deviation.
- (d) Is it unusual (more than 2 standard deviations below the mean) for fewer than 14 students to pass? Show your work.

3. ⋆ A patient is prescribed a blood pressure medication with a 70% effectiveness rate. A physician treats 8 patients with this medication.
- (a) Find $P(X = 8)$: all 8 patients respond.
- (b) Find $P(X \leq 4)$ using Desmos or Excel.
- (c) Find μ and σ.
- (d) The physician claims: "I treated 3 patients and none responded. This medication doesn't work." Using the binomial distribution, evaluate this claim statistically. Is 0 successes in 3 trials surprising given $p = 0.70$?

Step 6 Think Like a Nurse

The Binomial Model in Quality Improvement

A hospital's CLABSI (central line-associated bloodstream infection) prevention team has been working to reduce infection rates. The current benchmark rate is 2% per line insertion. On a busy month, the ICU performs 80 central line insertions.

The binomial model gives us:

$$\mu = 80 \times 0.02 = 1.6 \text{ infections expected} \qquad \sigma = \sqrt{80 \times 0.02 \times 0.98} \approx 1.25$$

The quality team reports zero CLABSI events this month. Is this evidence of improvement, or just a lucky month?

Under the benchmark rate, $P(X = 0) = (0.98)^{80} \approx 0.198$. There is about a **20% chance** of zero infections even if nothing has improved. Zero infections is a good result, but it is not statistically surprising under the old rate.

Now suppose the team has genuinely reduced the rate to 0.5%. Then $P(X = 0) = (0.995)^{80} \approx 0.670$ — zero infections becomes much more likely. To distinguish improvement from luck, you need multiple months of data and formal hypothesis testing (Chapter 12).

The binomial distribution tells you what is expected by chance. Before celebrating a good outcome, ask: could this have happened anyway, even without improvement? How often?

8.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 8. Problems marked with ⋆ are more challenging.

Part A: Discrete Probability Distributions

1. Determine whether each table is a valid probability distribution. If not, explain why.

(a)

x	0	1	2	3	4
$P(x)$	0.20	0.35	0.25	0.15	0.05

(b)

x	1	2	3	4
$P(x)$	0.30	0.40	0.25	0.20

(c)

x	0	1	2	3
$P(x)$	0.50	0.30	0.15	−0.05

2. A neonatal unit records the number of admissions per day over 200 days:

Admissions (x)	0	1	2	3	4	5
Days observed	20	54	68	38	16	4

(a) Convert to a probability distribution using relative frequencies.
(b) Compute μ and σ.
(c) What is the probability of more than 3 admissions on a given day?
(d) The unit is fully staffed for up to 3 admissions per day. What is the probability that staffing will be strained (4 or more admissions)?

Part B: Binomial Distribution

3. Identify whether each scenario is binomial. If not, explain which condition is violated.

(a) A nurse draws blood from 12 patients. Each draw either succeeds on the first attempt or does not. Based on her skill level, each attempt succeeds with probability 0.92.

(b) A hospital randomly selects patients from its database without replacement until it has selected 50 patients with diabetes.

(c) A clinical trial enrolls 200 patients. Each patient independently either responds to the new drug (probability 0.65) or does not.

(d) A nurse makes 10 home visits. The number of miles driven per visit varies. We count the number of visits where she drives more than 15 miles.

4. A rapid diagnostic test for influenza has a false positive rate of 5% (specificity = 95%). A clinic tests 20 patients who do not have influenza.
Let X = number of false positives.
  (a) Identify n, p, and the meaning of "success" in this context.
  (b) Find $P(X = 0)$: no false positives.
  (c) Find $P(X \geq 2)$.
  (d) Find μ and σ. Interpret μ in context.

5. A hospital-acquired infection (HAI) prevention program aims for a VAP (ventilator-associated pneumonia) rate of 1% per ventilator day. An ICU has 25 ventilated patient-days in one week.
  (a) Find $P(X = 0)$: no VAP events.
  (b) Find $P(X = 1)$.
  (c) Find $P(X \geq 2)$.
  (d) Find μ and σ.
  (e) The ICU reports 3 VAP events this week. Using the model, is this result statistically surprising? Find $P(X \geq 3)$.

6. $\star$ A medication adherence program shows that patients who enroll are adherent 85% of the time (take medications as prescribed). A nurse case manager follows 15 enrolled patients.
  (a) Find the expected number of adherent patients and the standard deviation.
  (b) Find $P(X = 15)$: all patients are adherent.
  (c) Find $P(X \leq 10)$ using Desmos or Excel.
  (d) Is it unusual (more than 2 standard deviations below the mean) to have 10 or fewer adherent patients? Compute $\mu - 2\sigma$ and compare.
  (e) If the program's effectiveness is questioned and the true adherence rate is actually 60% (not 85%), find $P(X \geq 13)$ under both rates. What does the difference tell you?

Answer Key — Selected Problems

Answer Key

Section 8.1 Practice — Your Turn

1. (a) All $P(x) \geq 0$ and $0.25 + 0.40 + 0.20 + 0.10 + 0.05 = 1.00$. Valid. (b) $\mu = 0(0.25)+1(0.40)+2(0.20)+3(0.10)+4(0.05) = 0+0.40+0.40+0.30+0.20 = 1.30$ errors per shift. (c) $\sigma^2 = 1.69(0.25) + 0.09(0.40) + 0.49(0.20) + 2.89(0.10) + 7.29(0.05) = 0.4225 + 0.036 + 0.098 + 0.289 + 0.3645 = 1.21$; $\sigma = \sqrt{1.21} \approx 1.10$. (d) $P(x > 2) = P(3) + P(4) = 0.10 + 0.05 = 0.15$. (e) On average, the pharmacy catches 1.30 errors per shift, with typical variation of about 1.10 errors above or below that mean.

2. (a) $P(4) = 1 - (0.30 + 0.35 + 0.22 + 0.08) = 1 - 0.95 = 0.05$. (b) $\mu = 0(0.30) + 1(0.35)+2(0.22)+3(0.08)+4(0.05) = 0+0.35+0.44+0.24+0.20 = 1.23$ tasks per visit. (c) $20 \times 1.23 = 24.6$ follow-up tasks expected. (d) $\mu \pm 2\sigma = 1.23 \pm 2(1.02) = 1.23 \pm 2.04$; range: $(-0.81, 3.27)$, meaning practically 0 to 3 tasks covers the typical range.

Section 8.2 Practice — Your Turn

1. (a) Binary (infection / no infection); Independent (different patients); $n = 10$ fixed; $p = 0.08$ constant. All four conditions met. (b) $P(X = 0) = (0.92)^{10} \approx 0.4344$. (c) $P(X = 1) = \binom{10}{1}(0.08)^1(0.92)^9 = 10(0.08)(0.4722) \approx 0.3777$. (d) $P(X \geq 2) = 1 - 0.4344 - 0.3777 = 0.1879$. (e) $\mu = 10(0.08) = 0.80$; $\sigma = \sqrt{10(0.08)(0.92)} \approx 0.858$. (f) New $\mu = 10(0.04) = 0.40$; $P(X = 0) = (0.96)^{10} \approx 0.6648$, rising from 43.4% to 66.5%.

2. (a) $P(X = 20) = (0.88)^{20} \approx 0.0776$. (b) $P(X = 18) = \binom{20}{18}(0.88)^{18}(0.12)^2 = 190 \times 0.10016 \times 0.0144 \approx 0.2740$; $P(X = 19) = \binom{20}{19}(0.88)^{19}(0.12)^1 = 20 \times 0.08814 \times 0.12 \approx 0.2115$; $P(X = 20) \approx 0.0776$; $P(X \geq 18) \approx 0.5631$. (c) $\mu = 20(0.88) = 17.6$; $\sigma = \sqrt{20(0.88)(0.12)} \approx 1.453$. (d) $\mu - 2\sigma = 17.6 - 2.906 = 14.69$; fewer than 14 is below $\mu - 2\sigma$, so yes, it would be unusual.

Chapter Practice — Selected

Problem 1a: Valid — all probabilities between 0 and 1, and they sum to 1.00.
Problem 1b: Not valid — probabilities sum to $0.30 + 0.40 + 0.25 + 0.20 = 1.15 > 1$.
Problem 1c: Not valid — $P(3) = -0.05 < 0$; probabilities cannot be negative.
Problem 2: (a) Relative frequencies: 0.10, 0.27, 0.34, 0.19, 0.08, 0.02. (b) $\mu = 0(0.10) + 1(0.27)+2(0.34)+3(0.19)+4(0.08)+5(0.02) = 0+0.27+0.68+0.57+0.32+0.10 = 1.94$; $\sigma \approx 1.11$. (c) $P(x > 3) = 0.08 + 0.02 = 0.10$. (d) $P(x \geq 4) = 0.10$.
Problem 4d: $\mu = 20(0.05) = 1.0$ false positive per 20 non-infected patients.
Problem 5e: $P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$. $P(0) = (0.99)^{25} \approx 0.7778$; $P(1) \approx 25(0.01)(0.99)^{24} \approx 0.1964$; $P(2) \approx \binom{25}{2}(0.01)^2(0.99)^{23} \approx 0.0238$; $P(X \geq 3) \approx 1 - 0.7778 - 0.1964 - 0.0238 = 0.0020$. Yes, 3 events is highly unusual under the 1% rate model ($P \approx 0.2\%$); this warrants investigation.

Problem 6 (Challenge): (a) $\mu = 15(0.85) = 12.75$; $\sigma = \sqrt{15(0.85)(0.15)} \approx 1.382$. (b) $P(X = 15) = (0.85)^{15} \approx 0.0874$. (d) $\mu - 2\sigma = 12.75 - 2.764 = 9.986$; since $10 \approx \mu - 2\sigma$, having 10 or fewer is borderline unusual. (e) At $p = 0.85$: $P(X \geq 13) \approx P(13) + P(14) + P(15) \approx 0.2184 + 0.1539 + 0.0874 = 0.4597$; at $p = 0.60$: $P(X \geq 13) \approx 0.0124 + 0.0022 + 0.0003 = 0.0149$. Under the lower rate, 13 or more adherent patients is very rare (1.5% vs 46%) — strong evidence that if 13+ are adherent, the 85% rate is plausible.

Chapter 8 Summary

Section 8.1 — Discrete Probability Distributions

- A **discrete probability distribution** assigns probabilities to all possible values of a discrete random variable; probabilities must be non-negative and sum to 1.
- **Expected value (mean):** $\mu = \sum x \cdot P(x)$ — the long-run average outcome.
- **Standard deviation:** $\sigma = \sqrt{\sum (x-\mu)^2 \cdot P(x)}$ — typical spread around the mean.
- The expected value is a mathematical average over many trials, not a prediction for a single trial.

Section 8.2 — The Binomial Distribution

- Binomial conditions (BINS): Binary outcomes, Independent trials, fixed Number of trials, Same probability.
- $X \sim B(n,p)$: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$
- **Mean:** $\mu = np$ **Standard deviation:** $\sigma = \sqrt{np(1-p)}$
- Use the complement rule for "at least" probabilities: $P(X \geq k) = 1 - P(X \leq k-1)$.
- Use `BINOM.DIST` in Excel or `binomialdist` in Desmos for cumulative probabilities.

The Nursing Connection

- Medication error rates, infection rates, and test positivity rates all follow binomial logic when applied to a fixed number of independent events.
- Small changes in the per-event probability (p) have large effects on the probability of multiple events across many trials.
- The binomial model establishes what is *expected by chance* — the baseline against which quality improvement efforts should be measured.
- Always ask: is an observed count within the expected range, or is it statistically unusual enough to warrant investigation?

CHAPTER 9

The Normal Distribution

> *"Nature does not distribute its gifts randomly — it clusters them around a center, thinning out toward the extremes."*
>
> — why the bell curve appears everywhere in clinical data

In this chapter, you will learn to:

- Describe the properties of the normal distribution
- Compute and interpret z-scores in clinical contexts
- Find probabilities (areas) under the normal curve using Desmos and Excel
- Find values corresponding to given probabilities (inverse normal)
- Apply the normal distribution to lab values, vital signs, and clinical thresholds
- Assess whether a dataset is approximately normally distributed

9.1 Properties of the Normal Distribution

Step 1 Read This First

A hospital's laboratory reports that hemoglobin levels in healthy adult women are normally distributed with a mean of 13.8 g/dL and a standard deviation of 1.2 g/dL.

You are reviewing three patients:

- Patient A: hemoglobin = 15.0 g/dL
- Patient B: hemoglobin = 12.4 g/dL
- Patient C: hemoglobin = 10.8 g/dL

You want to know: how unusual is each of these values? Is Patient C's level rare enough to warrant immediate clinical attention? What percentage of healthy women have hemoglobin this low or lower?

These are questions about areas under a normal curve — and the normal distribution gives exact, quantitative answers. After this chapter, you will be able to answer them for any normally distributed clinical variable.

Step 2 Let's Talk About It

Think about what you already know about bell-shaped distributions from Chapter 4.

When a variable is normally distributed, most values cluster near the mean, and values become progressively rarer as you move further from the center in either direction. The distribution is perfectly symmetric. The mean, median, and mode are all identical.

In Chapter 5, you learned the Empirical Rule: approximately 68% of values fall within one standard deviation of the mean, 95% within two, and 99.7% within three. That rule was a approximation based on the normal shape.

In this chapter, we go further: using the exact normal distribution, we can find the probability of any range of values, not just the round numbers covered by the Empirical Rule. We can ask: what fraction of patients have hemoglobin below 11.0? What fraction have it between 12.5 and 14.5? What value separates the lowest 5% from the rest?

Step 3 Now We Name It

Definition: The Normal Distribution

A continuous random variable X has a **normal distribution** with mean μ and standard deviation σ, written $X \sim N(\mu, \sigma)$, if its distribution has the characteristic bell shape:

- Symmetric about μ
- Mean = Median = Mode = μ
- The total area under the curve equals 1
- Approximately 68% of values lie within $\mu \pm \sigma$
- Approximately 95% of values lie within $\mu \pm 2\sigma$
- Approximately 99.7% of values lie within $\mu \pm 3\sigma$

Definition: The Standard Normal Distribution

The **standard normal distribution** is a special case with $\mu = 0$ and $\sigma = 1$, written $Z \sim N(0, 1)$.

Any normally distributed variable $X \sim N(\mu, \sigma)$ can be converted to the standard normal using the **z-score**:

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

The z-score measures how many standard deviations x is above $(z > 0)$ or below $(z < 0)$ the mean.

Probabilities as areas: $P(a < X < b)$ equals the area under the normal curve between a and b. Since the total area is 1, all probabilities are between 0 and 1.

Notation Conventions

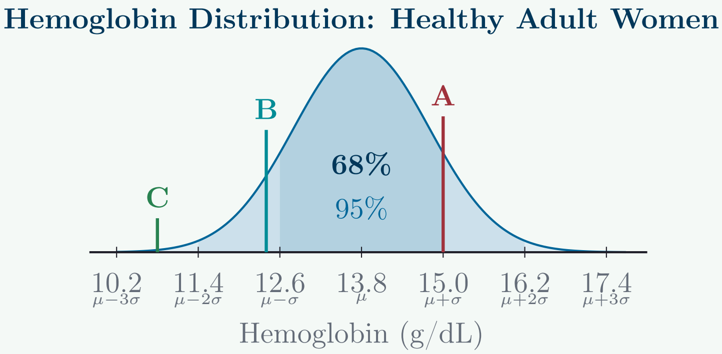
$P(X < x)$: probability that X is *less than* x — the area to the *left* of x.
$P(X > x)$: probability that X is *greater than* x — the area to the *right* $= 1 - P(X < x)$.
$P(a < X < b)$: area *between* a and $b = P(X < b) - P(X < a)$.

For a continuous distribution, $P(X = x) = 0$ for any exact value — so $P(X \leq x) = P(X < x)$.

Step 4 Watch It Work

Visualizing the normal distribution for hemoglobin levels

$X \sim N(13.8,\ 1.2)$: hemoglobin in healthy adult women (g/dL)



Z-scores for the three patients:

$$z_A = \frac{15.0 - 13.8}{1.2} = \frac{1.2}{1.2} = +1.00 \qquad z_B = \frac{12.4 - 13.8}{1.2} = \frac{-1.4}{1.2} \approx -1.17 \qquad z_C = \frac{10.8 - 13.8}{1.2} = \frac{-3.0}{1.2} =$$

Clinical interpretation:

- Patient A ($z = +1.00$): slightly above average — unremarkable.
- Patient B ($z = -1.17$): somewhat below average but within one standard deviation below the mean — within common variation.
- Patient C ($z = -2.50$): 2.5 standard deviations below the mean. By the Empirical Rule, fewer than 2.5% of healthy women have hemoglobin this low or lower. This warrants investigation.

Step 5 Your Turn

1. Systolic blood pressure in healthy adults is approximately normally distributed with $\mu = 120$ mmHg and $\sigma = 12$ mmHg.
   (a) Compute the z-score for a patient with systolic BP of 144 mmHg.
   (b) Compute the z-score for a patient with systolic BP of 102 mmHg.
   (c) Which patient is further from the population mean? How do you know?
   (d) Using the Empirical Rule, approximately what percentage of healthy adults have systolic BP between 96 and 144 mmHg?
   (e) A patient has $z = -2.8$. What is their actual systolic BP value?

2. The resting heart rate of healthy nurses is approximately normally distributed with $\mu = 70$ bpm and $\sigma = 8$ bpm.

   (a) A nurse has a resting heart rate of 88 bpm. Compute and interpret her z-score.
   (b) A second nurse has $z = -1.5$. What is her resting heart rate?
   (c) Based on the Empirical Rule, what range of heart rates is considered "typical" (within 2 SDs of the mean)?
   (d) A heart rate of 55 bpm: is this unusually low? Compute z and interpret.

3. ⋆ A patient's serum sodium is 148 mEq/L. The normal distribution for serum sodium in healthy adults has $\mu = 140$ mEq/L and $\sigma = 3$ mEq/L. The clinical threshold for hypernatremia is sodium > 145 mEq/L.
   (a) Compute the z-score for this patient.
   (b) Is this value above or below the clinical threshold? What does the z-score tell you that the clinical label alone does not?
   (c) Compute the z-score for the threshold value of 145 mEq/L.
   (d) By the Empirical Rule, approximately what percentage of healthy adults would exceed the hypernatremia threshold?

Step 6 Think Like a Nurse

Lab Reference Ranges Are Built on the Normal Distribution

When a laboratory report shows a reference range like "Hemoglobin: 12.0–16.0 g/dL (female)," that range was not chosen arbitrarily. It was constructed to capture the middle 95% (or sometimes 99%) of values from a large sample of healthy individuals.

For a normally distributed variable, the middle 95% falls between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$. For hemoglobin in women ($\mu = 13.8$, $\sigma = 1.0$ in some reference populations):

$$\text{Lower bound} = 13.8 - 1.96(1.0) = 11.84 \approx 12.0 \text{ g/dL}$$

$$\text{Upper bound} = 13.8 + 1.96(1.0) = 15.76 \approx 16.0 \text{ g/dL}$$

This means that approximately **5% of perfectly healthy women** will fall outside the reference range by chance alone. A "flagged" lab value does not mean the patient is sick — it means the value is outside the range observed in most healthy individuals.

A nurse who understands this is a better interpreter of lab reports. A borderline low hemoglobin in an otherwise healthy patient is very different from the same value in a patient who is symptomatic, trending downward, or in a high-risk group.

Reference ranges are statistical constructs, not clinical verdicts. The normal distribution tells you how rare a value is. Clinical context tells you what to do about it.

9.2 Finding Probabilities and Values Under the Normal Curve

Step 1 Read This First

The hemoglobin distribution for healthy adult women: $X \sim N(13.8,\ 1.2)$

You need to answer three types of questions that the Empirical Rule alone cannot answer exactly:

Type 1 (Left-tail): What fraction of women have hemoglobin *below* 12.0 g/dL?
Type 2 (Right-tail): What fraction have hemoglobin *above* 15.5 g/dL?
Type 3 (Between): What fraction have hemoglobin *between* 12.5 and 14.8 g/dL?

And a fourth type — working backwards:
Type 4 (Inverse): What hemoglobin value separates the lowest 10% from the rest?

These are the four fundamental normal distribution calculations. Every clinical probability question you encounter falls into one of these four types.

Step 2 Let's Talk About It

The key insight is that every normal probability is an area under the bell curve — and every area can be converted to a standard normal area by computing a z-score.

Once you have the z-score, the probability can be found using Desmos, Excel, or a standard normal table. The technology does the integration for you. Your job is to set up the problem correctly: find the right z-score, decide whether you want the left tail, right tail, or middle area, and interpret the result.

The most common mistake is forgetting to use the complement for right-tail questions. Technology gives left-tail probabilities by default ($P(Z < z)$). For right-tail questions, you subtract from 1.

Step 3 Now We Name It

Definition: Four Types of Normal Probability Calculations

Type 1 — Left tail: $P(X < a)$

$$z = \frac{a-\mu}{\sigma} \qquad P(X < a) = P(Z < z) = \Phi(z)$$

Type 2 — Right tail: $P(X > a)$

$$P(X > a) = 1 - P(X < a) = 1 - \Phi(z)$$

Type 3 — Between: $P(a < X < b)$

$$P(a < X < b) = P(X < b) - P(X < a) = \Phi(z_b) - \Phi(z_a)$$

Type 4 — Inverse (find x given probability):

$$\text{Given } P(X < x) = p, \quad \text{find } x = \mu + z^* \cdot \sigma$$

where z^* is the z-score corresponding to cumulative probability p.

Try This in Desmos

In Desmos Scientific Calculator, use the normal cumulative distribution function:

Left tail: `normaldist(mu, sigma).cdf(x)` gives $P(X < x)$

Right tail: `1 - normaldist(mu, sigma).cdf(x)`

Between: `normaldist(mu, sigma).cdf(b) - normaldist(mu, sigma).cdf(a)`

Inverse: `normaldist(mu, sigma).inversecdf(p)` gives x such that $P(X < x) = p$

Example: `normaldist(13.8, 1.2).cdf(12)` gives $P(X < 12)$

Try This in Excel

`=NORM.DIST(x, mean, std_dev, TRUE)` $\rightarrow P(X < x)$ (left tail)

`=1 - NORM.DIST(x, mean, std_dev, TRUE)` $\rightarrow P(X > x)$ (right tail)

`=NORM.DIST(b,...,TRUE) - NORM.DIST(a,...,TRUE)` $\rightarrow$ between

`=NORM.INV(p, mean, std_dev)` $\rightarrow x$ such that $P(X < x) = p$ (inverse)

For standard normal: `=NORM.S.DIST(z, TRUE)` and `=NORM.S.INV(p)`

Step 4 Watch It Work

Solving all four types for hemoglobin: $X \sim N(13.8,\ 1.2)$

Type 1 — Left tail: $P(X < 12.0)$

$$z = \frac{12.0 - 13.8}{1.2} = \frac{-1.8}{1.2} = -1.50$$

Using Desmos: `normaldist(13.8,1.2).cdf(12)` $\approx$ **0.0668**
About 6.7% of healthy women have hemoglobin below 12.0 g/dL.

Type 2 — Right tail: $P(X > 15.5)$

$$z = \frac{15.5 - 13.8}{1.2} = \frac{1.7}{1.2} \approx 1.417$$

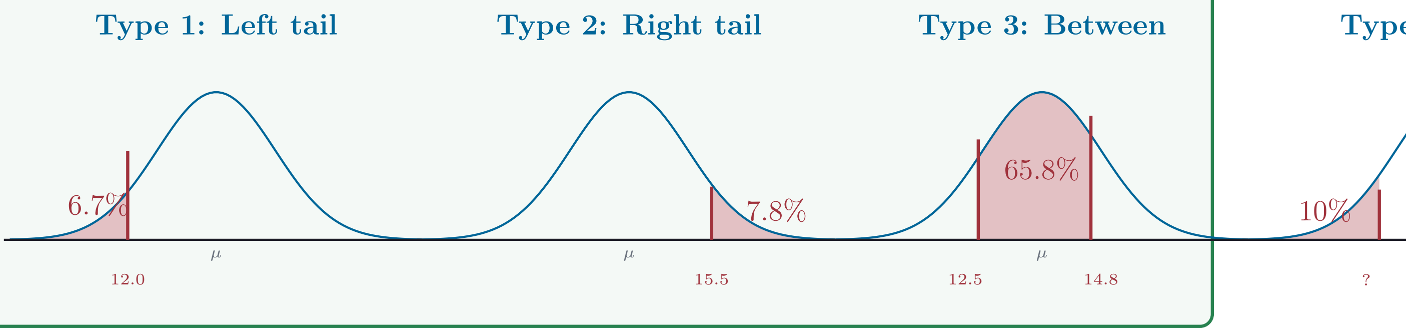
$P(X > 15.5) = 1 - P(X < 15.5) \approx 1 - 0.9218 = \mathbf{0.0782}$
About 7.8% of healthy women have hemoglobin above 15.5 g/dL.

Type 3 — Between: $P(12.5 < X < 14.8)$

$$z_1 = \frac{12.5 - 13.8}{1.2} \approx -1.083 \qquad z_2 = \frac{14.8 - 13.8}{1.2} \approx 0.833$$

$P(12.5 < X < 14.8) = P(X < 14.8) - P(X < 12.5) \approx 0.7977 - 0.1394 = \mathbf{0.6583}$
About 65.8% of healthy women have hemoglobin between 12.5 and 14.8 g/dL.

Type 4 — Inverse: What value separates the lowest 10%?
We need x such that $P(X < x) = 0.10$.
Using Desmos: `normaldist(13.8,1.2).inversecdf(0.10)` $\approx$ **12.26** g/dL
Verification: $z = (12.26 - 13.8)/1.2 = -1.28$. And indeed, $P(Z < -1.28) \approx 0.10$. ✓
The lowest 10% of healthy women have hemoglobin below about 12.26 g/dL.



Type 1: Left tail **Type 2: Right tail** **Type 3: Between** **Type**

Step 5 Your Turn

Use $X \sim N(\mu, \sigma)$ and Desmos or Excel for all calculations. Show the z-score(s) and identify the probability type.

1. Diastolic blood pressure in adults with controlled hypertension is approximately $N(82, 9)$ mmHg.
   (a) Find $P(X < 70)$: the probability of a reading below 70 mmHg (potential hypotension concern).
   (b) Find $P(X > 95)$: the probability of a reading above 95 mmHg (elevated).
   (c) Find $P(70 < X < 90)$: the probability of a reading in the well-controlled range.
   (d) Find the 90th percentile: what diastolic BP separates the lowest 90% from the highest 10%?
   (e) A patient on this medication has diastolic BP of 65 mmHg. Compute z and interpret the clinical concern.

2. Birth weights for full-term infants in the US are approximately $N(3400, 490)$ grams.
   (a) Find $P(X < 2500)$: the probability of low birth weight ($<$ 2,500 g).
   (b) Find $P(X > 4000)$: the probability of macrosomia ($>$ 4,000 g).
   (c) Find $P(2500 < X < 4000)$: the probability of normal birth weight range.
   (d) What birth weight marks the 5th percentile? (This is sometimes used as a clinical threshold for very low birth weight.)
   (e) A newborn weighs 2,200 g. Compute and interpret the z-score.

3. $\star$ A hospital sets drug dosing based on a patient's creatinine clearance (CrCl), which is approximately normally distributed in adult patients with $\mu = 85$ mL/min and $\sigma = 22$ mL/min.
   (a) Find $P(X < 60)$: the proportion of patients who would require dose adjustment (CrCl < 60 is the typical threshold for moderate renal impairment).
   (b) Find $P(X < 30)$: the proportion requiring severe dose adjustment.
   (c) What CrCl value marks the 25th percentile of this population?
   (d) A pharmacist uses $\mu \pm 2\sigma$ as an approximate reference range. What are those bounds? How do they compare to the clinical thresholds of 30 and 60 mL/min?
   (e) Is a CrCl of 38 mL/min statistically unusual in this population? Find the probability of a value this low or lower.

Step 6 Think Like a Nurse

Assessing Normality: When Can You Trust the Normal Model?
The normal distribution is enormously useful, but only when the data are actually approximately normally distributed. Applying it to skewed or bimodal data gives wrong answers.

How to assess normality in practice:

1. **Look at the histogram.** Is it roughly bell-shaped and symmetric? Does it have a single peak? A strongly skewed distribution is not approximately normal.
2. **Check the mean vs. median.** If mean ≈ median, the distribution is approximately symmetric. A large gap suggests skewness.
3. **Use the Empirical Rule as a rough check.** In a normal distribution, approximately 68% of values should fall within one SD of the mean. If you find 90% or 50%, the normal model is suspect.
4. **Normal probability plot (Q-Q plot).** A plot of the data against expected normal quantiles should be roughly linear if the data are normal. Curves or outlier points signal departures.

Clinical variables that are often approximately normal: resting heart rate in healthy adults, birth weight in full-term infants, adult height, hemoglobin in a healthy reference population, body temperature.

Clinical variables that are often *not* normal: length of stay (right-skewed), medical costs (right-skewed), number of falls per unit (discrete, often skewed), bacterial colony counts (may require log transformation).

The normal distribution is a model. Like all models, it is useful when it fits and misleading when it does not. Your first step before computing any normal probability is to look at the data.

9.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 9. Problems marked with ⋆ are more challenging. Use Desmos or Excel for all probability calculations.

Part A: Z-Scores and the Normal Distribution

1. Serum potassium in healthy adults is approximately $N(4.0,\ 0.4)$ mEq/L.
   (a) Compute the z-score for a potassium level of 3.2 mEq/L.
   (b) Compute the z-score for a potassium level of 5.1 mEq/L.
   (c) Normal range is 3.5–5.0 mEq/L. Compute the z-scores for both thresholds.
   (d) By the Empirical Rule, approximately what percentage of healthy adults have potassium in the normal range?
   (e) A patient has $z = +3.2$. What is their potassium level? Is this a clinical emergency?

2. A clinical lab measures fasting glucose in healthy adults as $N(92,\ 12)$ mg/dL.
   (a) Pre-diabetes is diagnosed when fasting glucose is 100–125 mg/dL. Compute z-scores for both thresholds.
   (b) Diabetes is diagnosed at ≥ 126 mg/dL. Compute the z-score for this threshold.
   (c) A patient has $z = -2.3$. What is their fasting glucose? Should you be clinically concerned?
   (d) Find the 95th percentile of fasting glucose. What does this value represent clinically?

Part B: Computing Normal Probabilities

3. Oxygen saturation (SpO_2) in healthy resting adults is approximately $N(97.5,\ 1.5)$%.
   (a) Find $P(\text{SpO}_2 < 95)$: the proportion with low-normal saturation.
   (b) Find $P(\text{SpO}_2 > 99)$.
   (c) Find $P(95 < \text{SpO}_2 < 99)$.
   (d) What SpO_2 value separates the lowest 2.5% from the rest?
   (e) A patient has $\text{SpO}_2 = 93\%$. Compute z. Is this unusual in a healthy population?

4. The time (in minutes) for nurses to complete electronic documentation after a shift is approximately $N(18,\ 4)$ minutes at one hospital.
   (a) Find the probability that a nurse finishes documentation in under 12 minutes.
   (b) Find the probability it takes more than 25 minutes.
   (c) Find the probability it takes between 14 and 22 minutes.
   (d) Hospital policy flags documentation taking more than 30 minutes for review.

What percentage of nurses are flagged?

(e) What documentation time marks the 75th percentile?

Part C: Inverse Normal and Applications

5. ⋆ Nurse exam scores at a hospital are approximately $N(78,\ 8)$.

(a) What score separates the bottom 15% from the rest?

(b) What score marks the 90th percentile?

(c) A passing score is 70. What proportion of nurses pass?

(d) The hospital wants to identify the top 5% for a leadership program. What minimum score qualifies?

(e) A nurse scores 62. Compute and interpret her z-score. Should the hospital require remediation?

6. ⋆ **Research Application:** A clinical study measures C-reactive protein (CRP) levels in patients admitted with suspected infection. CRP in this population is approximately $N(45, 18)$ mg/L.

(a) Find $P(\text{CRP} < 10)$: proportion of patients with near-normal CRP (possible misdiagnosis concern).

(b) Find $P(\text{CRP} > 80)$: proportion with severely elevated CRP.

(c) The clinical team uses a threshold of $\text{CRP} > 20$ to confirm active infection. What proportion of admitted patients meet this threshold?

(d) What CRP level marks the 95th percentile? What might this indicate clinically?

(e) A patient has CRP of 90 mg/L. The study team wants to classify this as "extreme." Is it more than 2 standard deviations above the mean? Show your work.

Answer Key — Selected Problems

Answer Key

Section 9.1 Practice — Your Turn

1. (a) $z = (144 - 120)/12 = 2.00$. (b) $z = (102 - 120)/12 = -1.50$. (c) Patient A is further from the mean ($|z| = 2.00 > 1.50$). (d) 96 to 144 spans $\mu \pm 2\sigma$; Empirical Rule gives approximately 95%. (e) $x = 120 + (-2.8)(12) = 120 - 33.6 = 86.4$ mmHg.

2. (a) $z = (88 - 70)/8 = 2.25$; 2.25 SDs above the mean — at approximately the 98.8th percentile of healthy resting HR, warrants evaluation. (b) $x = 70 + (-1.5)(8) = 70 - 12 = 58$ bpm. (c) $70 \pm 2(8) = 54$ to 86 bpm. (d) $z = (55 - 70)/8 = -1.875$; moderately below average but within 2 SDs; not statistically extreme, though clinically worth noting if symptomatic.

3. (a) $z = (148 - 140)/3 = 2.67$. (b) Above the threshold; the z-score reveals it is 2.67 SDs above normal mean — in the top 0.4% of healthy adults. (c) $z = (145 - 140)/3 = 1.67$. (d) The threshold at $z = 1.67$ corresponds to about the top 5% of healthy adults — roughly 5% would be flagged by this threshold.

Section 9.2 Practice — Your Turn

1. (a) $z = (70 - 82)/9 = -1.33$; $P(X < 70) \approx 0.0912$ (9.1%). (b) $z = (95 - 82)/9 = 1.44$; $P(X > 95) \approx 1 - 0.9251 = 0.0749$ (7.5%). (c) $z = (90 - 82)/9 = 0.889$; $P(X < 90) \approx 0.8133$; $P(70 < X < 90) \approx 0.8133 - 0.0912 = 0.7221$ (72.2%). (d) `normaldist(82,9).inversecdf(0.90)` ≈ 93.5 mmHg. (e) $z = (65 - 82)/9 = -1.89$; probability of this low or lower is about 2.9% — diastolic BP of 65 is approaching hypotension territory; monitor for symptoms, particularly lightheadedness or syncope.

2. (a) $z = (2500 - 3400)/490 = -1.837$; $P(X < 2500) \approx 0.0332$ (3.3%). (b) $z = (4000 - 3400)/490 = 1.224$; $P(X > 4000) \approx 1 - 0.8895 = 0.1105$ (11.1%). (c) $P(2500 < X < 4000) \approx 0.8895 - 0.0332 = 0.8563$ (85.6%). (d) `normaldist(3400,490).inversecdf(0.05)` ≈ 2594 g. (e) $z = (2200 - 3400)/490 = -2.45$; fewer than 1% of full-term infants weigh this little — very unusual, warrants immediate neonatal attention.

Chapter Practice — Selected

Problem 1e: $x = 4.0 + 3.2(0.4) = 4.0 + 1.28 = 5.28$ mEq/L. Yes, this is severe hyperkalemia and a potential cardiac emergency.

Problem 3d: `normaldist(97.5,1.5).inversecdf(0.025)` $\approx 94.6\%$; values below 94.6% SpO_2 are in the lowest 2.5% of healthy adults — clinically significant hypoxemia.

Problem 3e: $z = (93 - 97.5)/1.5 = -3.0$; $P(X < 93) \approx 0.0013$. Only 0.13% of healthy adults have SpO_2 this low — extremely unusual and clinically urgent.

Problem 4b: $z = (25 - 18)/4 = 1.75$; $P(X > 25) = 1 - P(X < 25) \approx 1 - 0.9599 = 0.0401$ (4.0%).
Problem 5d: `normaldist(78,8).inversecdf(0.95)` ≈ 91.2. Nurses scoring above approximately 91 are in the top 5%.
Problem 6e: $z = (90 - 45)/18 = 2.50$; yes, CRP of 90 is exactly 2.5 SDs above the mean — more than 2 SDs, so it qualifies as "extreme" by the conventional criterion. $P(X > 90) \approx 0.0062$ (0.6% of admitted patients have CRP this high or higher).

Chapter 9 Summary

Section 9.1 — Properties of the Normal Distribution

- $X \sim N(\mu, \sigma)$: symmetric, bell-shaped, total area = 1.
- Mean = Median = Mode = μ.
- Empirical Rule: 68% within 1σ, 95% within 2σ, 99.7% within 3σ.
- **Z-score:** $z = (x - \mu)/\sigma$ — measures standard deviations from the mean.
- The standard normal $Z \sim N(0, 1)$ is the reference distribution.
- Lab reference ranges are built from the normal distribution: they capture the middle 95% of healthy values, so 5% of healthy individuals fall outside them by chance.

Section 9.2 — Finding Probabilities and Values

- **Left tail:** $P(X < a)$ — use `normaldist(μ,σ).cdf(a)` in Desmos.
- **Right tail:** $P(X > a) = 1 - P(X < a)$.
- **Between:** $P(a < X < b) = P(X < b) - P(X < a)$.
- **Inverse:** Given probability p, find x using `inversecdf(p)` or `NORM.INV` .
- Always identify the probability type before calculating.
- Assess normality before applying the model: check the histogram, compare mean and median, verify the Empirical Rule holds approximately.

The Nursing Connection

- Z-scores translate any lab value into a universal scale: how unusual is this patient relative to a reference population?
- Lab reference ranges are statistical — 5% of healthy people fall outside them. A flagged value needs clinical interpretation, not automatic alarm.
- Not all clinical variables are normal: length of stay, costs, and count data are often skewed.
- The normal distribution answers the question: "How rare is this value?" Clinical judgment answers: "What should we do about it?"

CHAPTER 10

The Central Limit Theorem

> *"The most important theorem in statistics is one that most people have never heard of — and it explains why averages are so much more reliable than individual measurements."*
>
> — on why sample means behave so predictably

In this chapter, you will learn to:

- Define a sampling distribution and explain why it matters
- State the Central Limit Theorem and identify when it applies
- Compute and interpret the standard error of the mean
- Find probabilities for sample means using the normal distribution
- Explain why larger samples produce more reliable estimates
- Connect the CLT to the logic of confidence intervals and hypothesis tests

10.1 Sampling Distributions and the Standard Error

Step 1 Read This First

A hospital quality team wants to estimate the average length of stay (LOS) for patients on their medical-surgical unit. They know the unit's true population mean is $\mu = 4.8$ days with a standard deviation of $\sigma = 2.1$ days, but in practice they only have data from samples.

Nurse A takes a random sample of 36 patients and finds $\bar{x} = 4.6$ days.
Nurse B takes a different random sample of 36 patients and finds $\bar{x} = 5.1$ days.
Nurse C takes yet another sample and finds $\bar{x} = 4.9$ days.

All three nurses sampled from the same population. Why are their sample means different? Which one should the team trust? And if they took 1,000 such samples and plotted all the sample means, what would that distribution look like?

These questions lead directly to one of the most powerful ideas in all of statistics: the **sampling distribution of the sample mean**.

Step 2 Let's Talk About It

Think about what happens when you take many samples from the same population.

Each sample produces a different mean — because each sample contains different patients. Some means will be a bit above the true population mean, some a bit below. Rarely, a sample will happen to include mostly long-stay patients and produce a high mean, or mostly short-stay patients and produce a low mean.

But here is the key insight: the sample means are not scattered randomly all over the place. They cluster around the true population mean μ. And as sample size increases, they cluster more tightly. A sample of 100 patients produces a mean much closer to μ than a sample of 5 patients does.

The distribution of all possible sample means — if you could take every possible sample of a given size from the population — is called the **sampling distribution**. Understanding its shape, center, and spread is the foundation of all statistical inference.

Step 3 Now We Name It

Definition: Sampling Distribution of $\bar{x}$

The **sampling distribution of the sample mean** $\bar{x}$ is the probability distribution of all possible values of $\bar{x}$ that could be obtained by taking all possible samples of size n from a population.

Key properties:

- **Center:** The mean of the sampling distribution equals the population mean: $\mu_{\bar{x}} = \mu$
- **Spread:** The standard deviation of the sampling distribution, called the **standard error**, is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- **Shape:** Determined by the Central Limit Theorem (see below)

Definition: Standard Error of the Mean

The **standard error** (SE) measures how much a sample mean is expected to vary from the population mean:

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

A larger sample size n produces a smaller standard error — sample means cluster more tightly around μ. This is the mathematical expression of why larger samples are more reliable.

Key relationship: Doubling the sample size does *not* halve the standard error. It is reduced by a factor of $\sqrt{2} \approx 1.41$. To cut the standard error in half, you must *quadruple* the sample size.

Definition: Central Limit Theorem (CLT)

If a random sample of size n is drawn from *any* population with mean μ and finite standard deviation σ, then the sampling distribution of $\bar{x}$ is approximately normal:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{for sufficiently large } n$$

Practical guideline: The normal approximation is generally reliable when $n \geq 30$. For populations that are already approximately normal, even small samples (e.g., $n \geq 10$) work well.

The CLT applies *regardless of the shape of the population distribution* — even skewed or non-normal populations produce approximately normal sampling distributions when n is large enough. This is why the normal distribution plays such a central role in statistical inference.

Step 4 Watch It Work

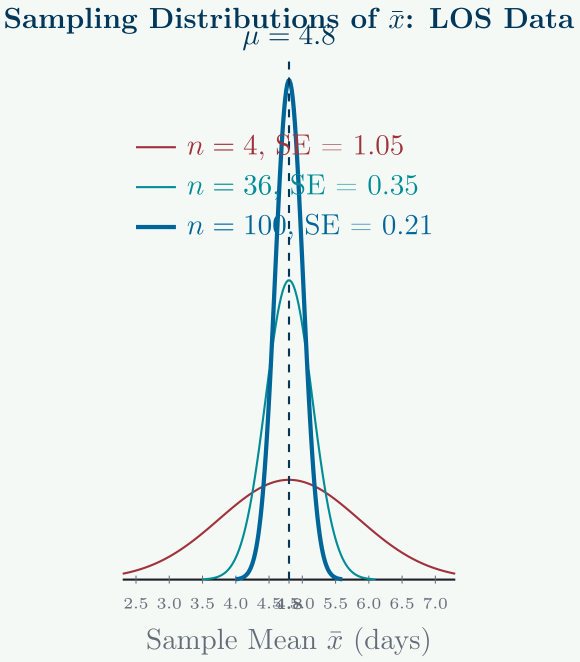
Applying the CLT to hospital length-of-stay data

Population: LOS on a medical-surgical unit, $\mu = 4.8$ days, $\sigma = 2.1$ days.
The population distribution is right-skewed (most stays are short, a few are very long).

Standard errors for different sample sizes:

Sample size n	$SE = \sigma/\sqrt{n}$	Calculation	Interpretation
4	$2.1/\sqrt{4} = 2.1/2.0$	1.050 days	Very wide; sample means vary a lot
9	$2.1/\sqrt{9} = 2.1/3.0$	0.700 days	Moderate spread
36	$2.1/\sqrt{36} = 2.1/6.0$	0.350 days	Much tighter
100	$2.1/\sqrt{100} = 2.1/10.0$	0.210 days	Very precise

Visual: The CLT at work — three sampling distributions



What the graph shows:

- All three sampling distributions are centered at the same population mean $\mu = 4.8$ days.
- The original LOS data is right-skewed, but all three sampling distributions are approximately normal (CLT).
- Larger n produces a narrower, taller curve — sample means cluster more tightly around μ.
- With $n = 100$, nearly all sample means fall between 4.4 and 5.2 days — a tight

window around the true mean.

Step 5 Your Turn

1. Patient wait times at a hospital emergency department have a mean of $\mu = 42$ minutes and standard deviation $\sigma = 18$ minutes. The distribution is right-skewed.

   (a) A quality analyst takes a random sample of $n = 36$ patients. What is the mean and standard error of the sampling distribution of $\bar{x}$?

   (b) For $n = 81$, what is the standard error?

   (c) By what factor does the standard error decrease when going from $n = 36$ to $n = 144$?

   (d) For which sample sizes can you use the normal distribution to model $\bar{x}$? Why?

   (e) If a nurse claims her sample of $n = 36$ had a mean wait time of 52 minutes, is this plausible? What would have to be true about her sample?

2. A hospital tracks hemoglobin levels in its post-surgical patient population: $\mu = 11.2$ g/dL, $\sigma = 1.8$ g/dL.

   (a) Compute the standard error for samples of size $n = 16$, $n = 25$, and $n = 49$.

   (b) Fill in the table:

n	SE	Range covering $\approx 95\%$ of $\bar{x}$ values ($\mu \pm 2 \cdot SE$)
16		
25		
49		

   (c) What sample size would reduce the standard error to 0.2 g/dL?

Step 6 Think Like a Nurse

Why Individual Measurements and Sample Averages Behave Differently
A patient's systolic blood pressure is known to be approximately normally distributed with $\mu = 120$ mmHg and $\sigma = 15$ mmHg. There are two very different questions you could ask:

Question 1 (individual): What is the probability that a single patient has a systolic BP above 140 mmHg?
Question 2 (sample mean): What is the probability that the *average* BP of a sample of 25 patients is above 140 mmHg?

These questions use entirely different distributions:

	Individual X	**Sample mean $\bar{x}$ $(n = 25)$**
Distribution	$N(120, 15)$	$N(120,\ 15/\sqrt{25}) = N(120, 3)$
SD used	$\sigma = 15$	$SE = 3$
z for value 140	$(140 - 120)/15 = 1.33$	$(140 - 120)/3 = 6.67$
Probability above 140	$\approx 9.2\%$	$\approx 0.000\%$

A single patient having BP above 140 is unusual but not rare — about 1 in 11 patients. But getting a *sample average* of 25 patients above 140 is essentially impossible unless the underlying population mean has shifted dramatically. The average of 25 measurements is far more stable than any single measurement.

This is the practical power of the CLT. When you see a hospital reporting an average outcome, you can judge how reliable that average is based on the sample size. A report based on 5 patients is nearly meaningless. A report based on 500 patients is highly precise.

Clinical intuition applies to individual patients. Statistical inference applies to sample means. Knowing which distribution to use is half the battle.

10.2 Applying the Central Limit Theorem

Step 1 Read This First

A hospital's nutrition team wants to assess whether post-surgical patients are meeting caloric targets. They know from previous research that daily caloric intake among post-surgical patients follows a right-skewed distribution with $\mu = 1{,}850$ kcal and $\sigma = 320$ kcal.

A dietitian reviews a random sample of $n = 64$ post-surgical patients from this month and finds a mean intake of $\bar{x} = 1{,}780$ kcal.

Two questions arise:

1. If nothing has changed, how likely is it to observe a sample mean as low as 1,780 kcal or lower just by random chance?
2. Does a mean of 1,780 kcal suggest that this month's patients are systematically under-eating, or is it within the normal range of sampling variation?

Answering these questions requires computing a probability for a sample mean — using the CLT and the normal distribution.

Step 2 Let's Talk About It

Once the CLT tells us that $\bar{x}$ is approximately normally distributed, finding probabilities for $\bar{x}$ works exactly like finding probabilities in Chapter 9 — with one critical change.

In Chapter 9, the standard deviation in the z-score denominator was σ (the population SD for individual observations). Now it is the *standard error* $\sigma/\sqrt{n}$ (the SD for sample means).

Everything else is the same: compute a z-score, identify the probability type (left-tail, right-tail, between, or inverse), use Desmos or Excel. The z-score for a sample mean is:

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

The denominator is the standard error, not the population standard deviation.

Step 3 Now We Name It

Definition: Z-Score for a Sample Mean

When the sampling distribution of $\bar{x}$ is approximately normal (by the CLT), probabilities are computed using:

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

This z-score measures how many *standard errors* the sample mean $\bar{x}$ is above or below the population mean μ.

Probabilities: Use the same four types from Chapter 9 (left-tail, right-tail, between, inverse), but with the normal distribution $N(\mu,\ \sigma/\sqrt{n})$ rather than $N(\mu, \sigma)$.

The Most Common CLT Error

Always use the standard error $\sigma/\sqrt{n}$ when working with sample means.

Using σ instead of $\sigma/\sqrt{n}$ when computing a z-score for $\bar{x}$ is the single most common error in this chapter. Before computing any z-score, ask: am I working with an individual value or a sample mean? The answer determines which standard deviation to use.

Individual X: $z = \dfrac{x - \mu}{\sigma}$ Sample mean $\bar{x}$: $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Definition: Conditions for Applying the CLT

Before using the normal distribution for $\bar{x}$, verify:

1. **Random sample:** The data are a random sample from the population of interest.
2. **Independence:** Observations are independent. If sampling without replacement from a finite population, the sample should be at most 10% of the population size.
3. **Sample size:** Either $n \geq 30$ (for any population), or the population is approximately normal (for smaller n).

Step 4 Watch It Work

Finding probabilities for sample means: the caloric intake problem

Population: $\mu = 1{,}850$ kcal, $\sigma = 320$ kcal (right-skewed)
Sample: $n = 64$, observed $\bar{x} = 1{,}780$ kcal

Step 1 — Check CLT conditions:

- Random sample: assumed.
- Independence: $n = 64 \ll$ total hospital population. ✓
- Sample size: $n = 64 \geq 30$. ✓

Step 2 — Standard error:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{320}{\sqrt{64}} = \frac{320}{8} = 40 \text{ kcal}$$

Step 3 — Z-score for $\bar{x} = 1{,}780$:

$$z = \frac{\bar{x} - \mu}{SE} = \frac{1{,}780 - 1{,}850}{40} = \frac{-70}{40} = -1.75$$

Step 4 — Probability: $P(\bar{x} \leq 1{,}780) = P(Z \leq -1.75) \approx 0.0401$

Interpretation: If the population mean truly is 1,850 kcal, there is only about a **4%** chance of observing a sample mean as low as 1,780 kcal (or lower) in a sample of 64. This is somewhat unusual. The dietitian has reason to investigate whether intake has genuinely declined.

Additional question: What range contains 95% of all sample means?

$$\mu \pm 1.96 \cdot SE = 1{,}850 \pm 1.96(40) = 1{,}850 \pm 78.4$$
$$\Rightarrow \quad (1{,}771.6 \text{ kcal},\ 1{,}928.4 \text{ kcal})$$

If nothing has changed, 95% of all sample means from samples of $n = 64$ will fall between approximately 1,772 and 1,928 kcal. The observed $\bar{x} = 1{,}780$ is just inside this interval — borderline.

Try This in Desmos

Using Desmos for the sample mean probability:

Method 1: Use the sampling distribution directly:

`normaldist(1850, 40).cdf(1780)` ≈ 0.0401

Method 2: Convert to z first:
Compute $z = (1780 - 1850)/40 = -1.75$, then `normaldist(0,1).cdf(-1.75)` ≈ 0.0401

Try This in Excel

`=NORM.DIST(1780, 1850, 40, TRUE)` → $P(\bar{x} \leq 1780) \approx 0.0401$

Note: enter the *standard error* (40) as the standard deviation argument, not $\sigma = 320$.

Step 5 Your Turn

1. A large hospital's average patient satisfaction score is $\mu = 72$ (out of 100) with $\sigma = 14$. A quality analyst surveys $n = 49$ patients after a new patient experience initiative.

  (a) What is the standard error of $\bar{x}$ for $n = 49$?
  (b) The analyst observes $\bar{x} = 75.8$. Find $P(\bar{x} \geq 75.8)$ if the population mean is still 72.
  (c) Is the observed sample mean unusually high? What might this suggest about the initiative?
  (d) Find the range of sample means that would contain 95% of all possible $\bar{x}$ values if $\mu = 72$.
  (e) What $\bar{x}$ value would represent the top 5% of all possible sample means (i.e., the 95th percentile of the sampling distribution)?

2. The blood glucose levels of patients admitted to a diabetic care unit are approximately $N(178, 35)$ mg/dL.

  (a) What is the probability that a single randomly selected patient has blood glucose above 200 mg/dL?
  (b) A nurse reviews a random sample of $n = 25$ patients. What is the probability the sample mean exceeds 200 mg/dL?
  (c) Why is the answer to (b) so much smaller than the answer to (a)?
  (d) Find the probability the sample mean falls between 170 and 190 mg/dL for $n = 25$.
  (e) If n were increased to 100, would the probability in (d) increase or decrease? Recalculate and confirm.

3. ⋆ A hospital's pharmacy records show that the time from medication order to patient administration has $\mu = 28$ minutes and $\sigma = 9$ minutes. A new electronic ordering system is implemented. After the change, a random sample of $n = 81$ orders has a mean time of $\bar{x} = 25.5$ minutes.

  (a) Compute the standard error.
  (b) Compute the z-score for $\bar{x} = 25.5$.
  (c) Find $P(\bar{x} \leq 25.5)$ if the true population mean is still 28 minutes.
  (d) Based on this probability, does the new system appear to have reduced

medication administration time? Justify your answer.

(e) What assumption(s) are you making when you apply the CLT here?

Step 6 Think Like a Nurse

The CLT Is the Engine Behind Confidence Intervals and Hypothesis Tests
You have now learned the Central Limit Theorem, but you may be wondering: why does this matter? The answer is that everything in the next four chapters — confidence intervals and hypothesis tests — depends entirely on the CLT.

Here is the connection. In Section 10.2, you calculated that if nothing has changed, there is only a 4% chance of observing a sample mean as low as 1,780 kcal. That 4% is the seed of a *p-value* (Chapter 12): the probability of observing what you saw, or something more extreme, if the null situation (nothing has changed) is true.

And the interval $1{,}850 \pm 1.96 \times 40 = (1{,}772, 1{,}928)$ is the seed of a *confidence interval* (Chapter 11): a range of plausible values for the true population mean, constructed from the sampling distribution.

Both of these tools rest on the same foundation:

1. The CLT guarantees that $\bar{x}$ is approximately normal.
2. The standard error tells us how wide the spread of sample means is.
3. Probabilities under the normal distribution tell us how likely or unlikely an observed result is.

The Central Limit Theorem is not the end of the story. It is the beginning of statistical inference. Every time a researcher reports a confidence interval or a p-value, the CLT is silently working in the background.

10.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 10. Problems marked with ⋆ are more challenging.

Part A: Sampling Distributions and Standard Error

1. The weights of patients admitted to a bariatric surgery program are approximately normally distributed with $\mu = 285$ lbs and $\sigma = 42$ lbs.
   (a) Compute the standard error for $n = 9$, $n = 36$, and $n = 100$.
   (b) For each sample size, describe the sampling distribution of $\bar{x}$.
   (c) By what factor does the standard error change when n increases from 9 to 36?
   (d) What sample size is needed to achieve a standard error of 5 lbs?

2. A hospital tracks the number of overtime hours worked by nurses each month. The distribution is right-skewed with $\mu = 12.4$ hours and $\sigma = 5.6$ hours.
   (a) Is the population distribution approximately normal? Does this matter for the sampling distribution of $\bar{x}$ when $n = 40$? Explain.
   (b) Compute the standard error for $n = 40$.
   (c) What is the approximate shape of the sampling distribution for $n = 40$?
   (d) What is the approximate shape for $n = 5$? Could you use the normal distribution here? Why or why not?

Part B: Probabilities for Sample Means

3. The duration (in minutes) of patient nursing assessments has $\mu = 22$ minutes and $\sigma = 6$ minutes. A supervisor reviews a random sample of $n = 36$ assessments.
   (a) What is the standard error?
   (b) Find $P(\bar{x} < 20)$: what is the probability the sample mean is below 20 minutes?
   (c) Find $P(\bar{x} > 24)$.
   (d) Find $P(21 < \bar{x} < 23)$.
   (e) The supervisor considers a sample mean above 25 minutes to indicate a staffing concern. How likely is this to occur by chance alone?

4. Post-operative pain scores (0–10 scale) in a surgical unit follow a distribution with $\mu = 5.2$ and $\sigma = 2.0$. A researcher samples $n = 64$ patients.
   (a) Find the standard error.
   (b) Find $P(\bar{x} \leq 4.8)$.
   (c) Find $P(\bar{x} \geq 5.7)$.
   (d) What range of sample means would be considered “unusual” (outside 2 standard errors of μ)?
   (e) The researcher observes $\bar{x} = 4.5$. Is this statistically unusual? Find the

probability.

Part C: CLT in Context

5. ⋆ A large health system's average patient discharge time (from discharge order to patient leaving) has $\mu = 3.2$ hours and $\sigma = 0.9$ hours. A hospital in the system implements a new discharge checklist. Over the next month, they record $n = 100$ discharges with $\bar{x} = 3.0$ hours.

(a) Compute the standard error.
(b) Compute the z-score for $\bar{x} = 3.0$.
(c) Find $P(\bar{x} \leq 3.0)$ under the assumption that nothing has changed.
(d) Is this result statistically unusual? At what probability threshold do you consider it surprising?
(e) The hospital administrator claims the checklist reduced discharge time. Does the statistical analysis support this claim? What would be needed to make a stronger claim?

6. ⋆ **Individual vs. Sample Mean:** Serum creatinine in hospitalized patients is approximately normally distributed with $\mu = 1.3$ mg/dL and $\sigma = 0.5$ mg/dL.

(a) Find the probability that a single patient has creatinine above 2.0 mg/dL.
(b) Find the probability that the mean creatinine for a sample of $n = 16$ patients exceeds 2.0 mg/dL.
(c) Find the probability that the mean creatinine for a sample of $n = 64$ patients exceeds 2.0 mg/dL.
(d) Explain in plain language why the probabilities in (a), (b), and (c) are so different.
(e) If a unit's average creatinine for the month was reported as 2.0 mg/dL based on 64 patients, what would you conclude about this unit's patient population?

7. ⋆ **Research Application:** A nurse researcher wants to detect a clinically meaningful decrease in average pain scores from $\mu = 5.2$ to $\mu = 4.7$ (a reduction of 0.5 points). She plans to use a sample mean from $n = 64$ patients. Assume $\sigma = 2.0$.

(a) Under the original mean ($\mu = 5.2$), find $P(\bar{x} \leq 4.7)$.
(b) Under the new mean ($\mu = 4.7$), find $P(\bar{x} \leq 4.7)$.
(c) What does the comparison between (a) and (b) tell you about the ability to detect this change?
(d) If instead $n = 16$, recompute (a) and (b). What happens to the ability to detect the change?
(e) This question is previewing the concept of statistical *power*. Describe in your own words what "power" means in the context of detecting a clinical improvement.

Answer Key — Selected Problems

Answer Key

Section 10.1 Practice — Your Turn

1. (a) $\mu_{\bar{x}} = 42$ min; $SE = 18/\sqrt{36} = 3.0$ min. (b) $SE = 18/\sqrt{81} = 2.0$ min. (c) From $n = 36$ to $n = 144$: $SE_{36} = 18/6 = 3.0$; $SE_{144} = 18/12 = 1.5$; the SE is halved (factor of 2). (d) For all three sample sizes mentioned ($n = 36$, 81, 144); all are ≥ 30, so the CLT guarantees an approximately normal sampling distribution regardless of the skewed population shape. (e) $z = (52-42)/3 = 3.33$; $P(\bar{x} \geq 52) \approx 0.04\%$ — extremely unlikely. Her sample would have to be a very unusual collection of mostly long-wait patients.

2. (a) $SE_{16} = 1.8/4 = 0.450$; $SE_{25} = 1.8/5 = 0.360$; $SE_{49} = 1.8/7 \approx 0.257$ g/dL. (b) Table: $n = 16$: SE=0.450, range $(11.2 \pm 0.900) = (10.30, 12.10)$; $n = 25$: SE=0.360, range $(10.48, 11.92)$; $n = 49$: SE=0.257, range $(10.69, 11.71)$. (c) $SE = 1.8/\sqrt{n} = 0.2 \Rightarrow \sqrt{n} = 9 \Rightarrow n = 81$.

Section 10.2 Practice — Your Turn

1. (a) $SE = 14/\sqrt{49} = 2.0$. (b) $z = (75.8-72)/2.0 = 1.90$; $P(\bar{x} \geq 75.8) = 1-P(Z \leq 1.90) \approx 1 - 0.9713 = 0.0287$ (2.9%). (c) Only about a 3% chance of observing this high a mean by chance; the initiative may be having a positive effect, though formal testing is needed. (d) $72 \pm 1.96(2) = 72 \pm 3.92 \Rightarrow (68.08, 75.92)$. (e) `normaldist(72,2).inversecdf(0.95)` ≈ 75.3.

2. (a) Individual: $z = (200 - 178)/35 = 0.629$; $P(X > 200) \approx 1 - 0.7353 = 0.2647$ (26.5%). (b) $SE = 35/\sqrt{25} = 7$; $z = (200 - 178)/7 = 3.14$; $P(\bar{x} > 200) \approx 1-0.9992 = 0.0008$ (0.08%). (c) Sample means are far less variable than individual measurements; the sampling distribution is much narrower ($SE = 7$ vs $\sigma = 35$). (d) $SE = 7$; $z_1 = (170 - 178)/7 = -1.14$; $z_2 = (190 - 178)/7 = 1.71$; $P(170 < \bar{x} < 190) \approx 0.9564 - 0.1271 = 0.8293$ (82.9%). (e) $n = 100$: $SE = 35/10 = 3.5$; $z_1 = (170-178)/3.5 = -2.29$, $z_2 = (190-178)/3.5 = 3.43$; $P \approx 0.9997-0.0110 = 0.9887$ (98.9%) — increases substantially.

Chapter Practice — Selected

Problem 3b: $SE = 6/\sqrt{36} = 1.0$; $z = (20 - 22)/1.0 = -2.0$; $P(\bar{x} < 20) \approx 0.0228$ (2.3%).

Problem 3e: $z = (25 - 22)/1.0 = 3.0$; $P(\bar{x} > 25) \approx 0.0013$ (0.13%) — very unlikely by chance alone; a mean above 25 minutes would be a meaningful signal.

Problem 5b: $SE = 0.9/\sqrt{100} = 0.09$; $z = (3.0 - 3.2)/0.09 = -2.22$; $P(\bar{x} \leq 3.0) \approx 0.0132$ (1.3%) — unusual if nothing has changed.

Problem 6a: $z = (2.0 - 1.3)/0.5 = 1.4$; $P(X > 2.0) \approx 0.0808$ (8.1%).

Problem 6b: $SE = 0.5/\sqrt{16} = 0.125$; $z = (2.0 - 1.3)/0.125 = 5.6$; $P(\bar{x} > 2.0) \approx 0.000\%$ — essentially impossible.

Problem 7a: $SE = 2.0/\sqrt{64} = 0.25$; $z = (4.7 - 5.2)/0.25 = -2.0$; $P(\bar{x} \leq 4.7 \mid \mu = 5.2) \approx 0.0228$ (2.3%).
Problem 7b: Under new $\mu = 4.7$: $z = (4.7 - 4.7)/0.25 = 0$; $P(\bar{x} \leq 4.7 \mid \mu = 4.7) = 0.50$ (50%).

Chapter 10 Summary

Section 10.1 — Sampling Distributions and the Standard Error

- The **sampling distribution of** $\bar{x}$ describes how sample means vary across all possible samples of size n.
- **Center:** $\mu_{\bar{x}} = \mu$ (sample means are centered at the population mean).
- **Spread:** $SE = \sigma/\sqrt{n}$ (standard error decreases as n increases).
- **CLT:** For $n \geq 30$, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ regardless of population shape.
- To halve the standard error, quadruple the sample size.

Section 10.2 — Applying the CLT

- Z-score for a sample mean: $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$.
- **Always use $SE = \sigma/\sqrt{n}$, not σ, when working with $\bar{x}$.**
- Conditions: random sample, independence, $n \geq 30$ (or approximately normal population).
- Use `normaldist(μ, σ/√n).cdf(x̄)` in Desmos, or `NORM.DIST` with SE in Excel.

The Nursing Connection

- Individual measurements are noisy; sample averages are much more stable and predictable.
- The CLT is why small hospitals can still make valid statistical comparisons — the sampling distribution is normal even when the underlying data are not.
- An unusual sample mean (one that is far from μ in terms of standard errors) is a signal worth investigating — but chance alone can produce unusual results, and context matters.
- The CLT is the mathematical foundation for all the inference tools in the chapters ahead: confidence intervals, hypothesis tests, and ANOVA.

CHAPTER 11

Confidence Intervals

> *"We never know the truth exactly. But we can know how wrong we are likely to be — and that is almost as good."*
>
> — on what a confidence interval actually tells you

In this chapter, you will learn to:

- Explain what a confidence interval is and what it means
- Construct and interpret confidence intervals for a population mean (σ known)
- Construct and interpret confidence intervals for a population mean (σ unknown, using t)
- Construct and interpret confidence intervals for a population proportion
- Determine the sample size needed to achieve a desired margin of error
- Correctly interpret and avoid common misinterpretations of confidence intervals

11.1 Confidence Intervals for a Mean

Step 1 Read This First

A hospital quality team wants to estimate the true mean recovery time for patients who undergo a new minimally invasive knee procedure. They collect a random sample of $n = 40$ patients and record the days to full ambulation. The sample mean is $\bar{x} = 8.4$ days and the sample standard deviation is $s = 3.1$ days.

The team knows that 8.4 days is the best single estimate of the true population mean μ — but they also know it is almost certainly not exactly right. Every sample gives a slightly different mean. The real question is: *how wrong could 8.4 be?*

A **confidence interval** answers this question directly. Rather than reporting a single point estimate, it reports a range of plausible values for μ, along with a specified level of confidence that the interval contains the true value.

By the end of this section, the team will be able to say: *"We are 95% confident that the true mean recovery time is between 7.4 and 9.4 days."* That statement is precise, honest, and clinically actionable.

Step 2 Let's Talk About It

Think about what it means to be "95% confident."

Suppose you repeated this study 100 times, each time drawing a fresh random sample of 40 patients and computing a 95% confidence interval. About 95 of those 100 intervals would contain the true population mean μ. About 5 would miss it entirely.

You do not know whether *your* particular interval is one of the 95 that captures μ or one of the 5 that does not. What you know is that the *procedure* you used is correct 95% of the time.

This is subtle. The confidence level describes the reliability of the method, not the probability about this specific interval. Once the interval is computed, the true mean either is or is not inside it — there is no probability about that. The 95% refers to how often the method works in the long run.

Step 3 Now We Name It

Definition: Confidence Interval for μ (Large Sample or σ Known)

When σ is known *or* $n \geq 30$, a $(1 - \alpha) \times 100\%$ confidence interval for μ is:

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

where z^* is the **critical value** from the standard normal distribution corresponding to the desired confidence level:

Confidence Level	α	z^*
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Definition: Confidence Interval for μ (Small Sample, σ Unknown — the t-interval)

In practice, σ is almost never known. When $n < 30$ or when using s instead of σ, use the **t-distribution**:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

where t^* is the critical value from the t-distribution with $df = n - 1$ degrees of freedom.

The t-distribution is similar to the standard normal but has heavier tails, reflecting additional uncertainty from estimating σ with s. As n increases, the t-distribution approaches the standard normal.

Conditions for the t-interval:

- Random sample from the population of interest
- Either $n \geq 30$, or the population is approximately normal
- Observations are independent

Definition: Margin of Error

The **margin of error** (ME) is half the width of the confidence interval:

$$ME = z^* \cdot \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad ME = t^* \cdot \frac{s}{\sqrt{n}}$$

The confidence interval can be written as: $\bar{x} \pm ME$
A smaller margin of error means a more precise estimate. The ME decreases when the sample size increases or when the confidence level decreases.

Step 4 Watch It Work

Building the 95% confidence interval for recovery time

$n = 40$, $\bar{x} = 8.4$ days, $s = 3.1$ days
Since $n = 40 \geq 30$ and σ is unknown, use the t-interval with $df = 39$.

Step 1 — Critical value: For 95% confidence with $df = 39$:
Using Desmos: `tdist(39).inversecdf(0.975)` $= t^* \approx 2.023$

Step 2 — Standard error:

$$SE = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{40}} \approx \frac{3.1}{6.325} \approx 0.490 \text{ days}$$

Step 3 — Margin of error:

$$ME = t^* \cdot SE = 2.023 \times 0.490 \approx 0.991 \approx 0.99 \text{ days}$$

Step 4 — Confidence interval:

$$\bar{x} \pm ME = 8.4 \pm 0.99 \quad \Rightarrow \quad (\mathbf{7.41},\ \mathbf{9.39}) \text{ days}$$

Interpretation: We are 95% confident that the true mean days to full ambulation for patients undergoing this procedure is between **7.41 and 9.39 days**.

What happens when we change the confidence level?

Confidence	t^* ($df = 39$)	ME	Interval	Trade-off
90%	1.685	0.83	(7.57, 9.23)	Narrower, less confident
95%	2.023	0.99	(7.41, 9.39)	Standard choice
99%	2.708	1.33	(7.07, 9.73)	Wider, more confident

Higher confidence requires a wider interval. You cannot have a narrow interval with high confidence unless the sample size is large. This is the fundamental trade-off.

Try This in Desmos

Critical value: `tdist(39).inversecdf(0.975)` → t^* for 95% CI with $df = 39$

For 90%: `tdist(39).inversecdf(0.95)` For 99%: `tdist(39).inversecdf(0.995)`

Note: use 0.975 (not 0.95) for a 95% CI because 2.5% goes in each tail.

Try This in Excel

Critical value: `=T.INV.2T(0.05, 39)` → t^* for 95% CI (two-tailed, $\alpha = 0.05$)

`=CONFIDENCE.T(0.05, s, n)` → margin of error directly

Then: Lower bound = $\bar{x}-$ ME Upper bound = $\bar{x}+$ ME

Step 5 Your Turn

1. A clinical researcher measures daily step counts for a random sample of $n = 50$ hospitalized patients. The sample mean is $\bar{x} = 1{,}240$ steps and the sample standard deviation is $s = 480$ steps.
   (a) Find the standard error.
   (b) Find t^* for a 95% confidence interval with $df = 49$ using Desmos.
   (c) Compute the margin of error and the 95% confidence interval.
   (d) Interpret the interval in context.
   (e) What sample size would be needed to reduce the margin of error to 100 steps at 95% confidence?

2. Post-operative blood pressure measurements (systolic, mmHg) for 12 patients: 142, 138, 155, 129, 148, 161, 135, 144, 152, 139, 158, 146.
   (a) Compute $\bar{x}$ and s for this sample.
   (b) Find t^* for a 95% confidence interval ($df = 11$).
   (c) Construct the 95% confidence interval.
   (d) Construct the 99% confidence interval and compare it to the 95% CI.
   (e) A clinical benchmark for this procedure is a mean systolic BP below 150 mmHg post-operatively. Is the benchmark plausible given your confidence interval?

Step 6 Think Like a Nurse

The Most Common Misinterpretation of a Confidence Interval

A clinical trial reports: *"We are 95% confident that the new drug reduces mean blood pressure by between 4.2 and 8.7 mmHg."*

A nurse reads this and concludes: *"There is a 95% chance that the true effect is between 4.2 and 8.7 mmHg."*

This is wrong — and the distinction matters.

The interval (4.2, 8.7) is fixed. Once the data are collected and the interval is computed, the true effect either *is* or *is not* inside that interval. There is no probability involved for this specific interval. The probability was used in the method before the data were collected.

The correct interpretation: *If this procedure were repeated many times, 95% of the resulting intervals would contain the true population mean reduction.*

Three statements that are always wrong:

- "There is a 95% probability that the true mean is in (4.2, 8.7)." *Wrong.*
- "95% of patients experience effects between 4.2 and 8.7 mmHg." *Wrong — this is about the mean, not individual patients.*
- "The true mean is definitely somewhere in (4.2, 8.7)." *Wrong — 5% of intervals miss entirely.*

A confidence interval is a statement about the reliability of a method, not a probability statement about a particular number. The 95% refers to the process, not the outcome.

11.2 Confidence Intervals for a Proportion and Sample Size

Step 1 Read This First

A hospital infection control team wants to estimate the proportion of nurses on their unit who are fully compliant with the hand hygiene protocol. They conduct an unannounced audit and observe 80 randomly selected nurses, finding that 68 are fully compliant.

The sample proportion is $\hat{p} = 68/80 = 0.85$. But this is just an estimate. The team wants to know: *how much does this estimate vary from sample to sample?* And *how large a sample do they need* if they want their estimate to be within 5 percentage points of the truth with 95% confidence?

The proportion version of the confidence interval is structurally identical to the mean version — but uses different formulas for the standard error and requires checking different conditions.

Step 2 Let's Talk About It

Think of each nurse observation as a coin flip: compliant (success, probability p) or not compliant (failure, probability $1 - p$). With 80 independent observations, the sample proportion $\hat{p}$ behaves like a sample mean of 0s and 1s.

By the CLT, $\hat{p}$ is approximately normally distributed when the sample is large enough. The standard error of $\hat{p}$ comes from the binomial formula: $\sqrt{p(1-p)/n}$. Since we do not know p, we substitute $\hat{p}$.

The logic is the same as before: construct an interval around the point estimate $\hat{p}$ by adding and subtracting a margin of error. The margin of error gets smaller when n increases or when the confidence level decreases — the same trade-off as for the mean.

Step 3 Now We Name It

Definition: Confidence Interval for a Population Proportion p

A $(1-\alpha)\times 100\%$ confidence interval for a population proportion p is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = x/n$ is the sample proportion, x is the number of successes, and z^* is the standard normal critical value.

Conditions:

- Random sample
- Independence (or $n \leq 10\%$ of population)
- At least 10 successes and 10 failures: $n\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$

Definition: Sample Size for Estimating a Proportion

To achieve a margin of error of at most E with $(1-\alpha)\times 100\%$ confidence, the required sample size is:

$$n = \left(\frac{z^*}{E}\right)^2 \hat{p}(1-\hat{p})$$

If no prior estimate of p is available, use $\hat{p} = 0.5$ (this gives the largest possible — most conservative — required sample size).

Always round n up to the nearest whole number.

Definition: Sample Size for Estimating a Mean

To achieve a margin of error of at most E for a population mean with $(1-\alpha)\times 100\%$ confidence:

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

A prior estimate of σ (from a pilot study or previous research) is required. Always round up.

Step 4 Watch It Work

Confidence interval and sample size for the hand hygiene compliance data

$n = 80$, $x = 68$ compliant, $\hat{p} = 68/80 = 0.850$

Step 1 — Check conditions:

- Random sample: unannounced audit, assumed random. ✓
- Independence: 80 nurses $\ll$ total nursing staff. ✓
- Successes: $n\hat{p} = 80(0.850) = 68 \geq 10$. ✓
- Failures: $n(1-\hat{p}) = 80(0.150) = 12 \geq 10$. ✓

Step 2 — Standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.850 \times 0.150}{80}} = \sqrt{\frac{0.1275}{80}} = \sqrt{0.001594} \approx 0.0399$$

Step 3 — Margin of error (95% confidence, $z^* = 1.960$):

$$ME = 1.960 \times 0.0399 \approx 0.0782$$

Step 4 — Confidence interval:

$$0.850 \pm 0.0782 \quad \Rightarrow \quad (\mathbf{0.772},\ \mathbf{0.928})$$

Interpretation: We are 95% confident that the true proportion of nurses who are fully compliant with hand hygiene protocol is between **77.2% and 92.8%**.

Sample size calculation: How large a sample is needed for ME ≤ 0.05 (within 5 percentage points)?
Using the current estimate $\hat{p} = 0.85$:

$$n = \left(\frac{1.960}{0.05}\right)^2 (0.85)(0.15) = (39.2)^2 \times 0.1275 = 1{,}536.64 \times 0.1275 \approx 195.9$$

Round up to $n = 196$ nurses.

If no prior estimate of compliance were available, use $\hat{p} = 0.5$:

$$n = \left(\frac{1.960}{0.05}\right)^2 (0.5)(0.5) = 1{,}536.64 \times 0.25 = 384.2 \Rightarrow n = 385$$

The conservative estimate requires nearly twice as many observations.

Try This in Desmos

Standard error for proportion: `sqrt(0.85*0.15/80)` ≈ 0.0399
Sample size: $\lceil (1.96/0.05)^2 \times 0.85 \times 0.15 \rceil = 196$
Conservative: $\lceil (1.96/0.05)^2 \times 0.5 \times 0.5 \rceil = 385$

Try This in Excel

`=CONFIDENCE.NORM(0.05, SQRT(0.85*0.15/80), 1)` is not direct; instead use:

`=1.96*SQRT(0.85*0.15/80)` → margin of error ≈ 0.0782

Then: Lower $= \hat{p}-$ ME, Upper $= \hat{p}+$ ME

Step 5 Your Turn

1. A hospital surveys $n = 120$ patients about whether they received adequate pain management during their stay. Of these, 96 report that their pain was adequately managed.
  (a) Compute $\hat{p}$.
  (b) Check all conditions for the proportion confidence interval.
  (c) Construct a 95% confidence interval for the true proportion of patients with adequate pain management.
  (d) Construct a 90% confidence interval. Is it wider or narrower? Why?
  (e) The hospital's target is 85% or higher. Does the confidence interval support concluding that the target is met?

2. A nurse researcher wants to estimate the proportion of ICU patients who develop hospital-acquired infections. From previous studies, the estimated proportion is 12%.
  (a) How large a sample is needed to estimate the proportion within 3 percentage points (ME ≤ 0.03) at 95% confidence? Use $\hat{p} = 0.12$.
  (b) How large if no prior estimate is available (use $\hat{p} = 0.5$)?
  (c) Why does the conservative estimate ($\hat{p} = 0.5$) require a larger sample?
  (d) If budget constraints limit the study to $n = 200$ patients, what margin of error can be achieved at 95% confidence using $\hat{p} = 0.12$?

3. ⋆ A quality improvement team wants to estimate the mean time (in hours) nurses spend on documentation per shift. They plan to conduct a pilot study to determine the sample size needed for their full study. A prior pilot of 10 nurses gives $s = 1.4$ hours.
  (a) How large a sample is needed to estimate the population mean within 0.3 hours at 95% confidence?
  (b) At 99% confidence?
  (c) How does the required sample size change when the desired precision doubles (ME reduced to 0.15 hours)?
  (d) If the true mean is approximately 2.8 hours and the team wants the interval width (not ME) to be at most 0.6 hours at 95% confidence, what n is required?

Step 6 Think Like a Nurse

Confidence Intervals in Clinical Research: What the Width Tells You

A newly published randomized trial reports: *"The new wound care protocol reduced average healing time by 2.1 days (95% CI: 0.1 to 4.1 days)."*

A second trial of the same protocol reports: *"Average healing time reduced by 2.1 days (95% CI: 1.5 to 2.7 days)."*

Both trials found the same point estimate: 2.1 days. But the clinical implications are very different.

Trial 1 has a very wide interval. The true effect could be as small as 0.1 days (barely noticeable) or as large as 4.1 days (clinically meaningful). The study cannot tell us much. This width suggests a small sample size — the result is statistically positive but practically uncertain.

Trial 2 has a narrow interval. Even the lower bound (1.5 days) represents a clinically meaningful reduction. The large sample provides enough precision to be confident the effect is real and practically important.

The lesson for evidence-based nursing practice: *always check the confidence interval width, not just the point estimate.* A study can show an effect of 2.1 days and still tell you almost nothing useful if the sample is too small. The width of the interval is the direct measure of how much information the study actually provides.

A narrow confidence interval is not just a statistical achievement. It is evidence that the study was designed carefully enough to support the conclusions it draws.

11.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 11. Problems marked with ⋆ are more challenging.

Part A: Confidence Intervals for a Mean

1. A nurse researcher collects resting heart rate data from a random sample of $n = 36$ ICU nurses. The sample mean is $\bar{x} = 74.2$ bpm and the sample standard deviation is $s = 9.6$ bpm.
   (a) Compute the standard error.
   (b) Find t^* for 95% confidence ($df = 35$) using Desmos.
   (c) Construct the 95% confidence interval.
   (d) Interpret the interval in context.
   (e) A colleague argues: "We have a 95% probability that the true mean is in your interval." Is this correct? Explain.

2. The albumin levels (g/dL) in a random sample of 8 malnourished patients: 3.1, 2.8, 2.4, 3.3, 2.9, 2.6, 3.0, 2.7.
   (a) Find $\bar{x}$ and s.
   (b) Identify the appropriate critical value (z^* or t^*) and justify your choice.
   (c) Construct a 95% confidence interval.
   (d) The clinical threshold for normal albumin is ≥ 3.5 g/dL. What does the interval tell you about this patient population?

3. A hospital reports: "Based on a sample of 100 patients, we estimate the mean LOS is 5.3 days with a 95% CI of (4.8, 5.8)."
   (a) What is the margin of error?
   (b) What is the standard error?
   (c) Estimate the sample standard deviation s.
   (d) If the hospital wanted to narrow the interval to ±0.3 days, approximately what sample size would be needed?

Part B: Confidence Intervals for a Proportion

4. A random sample of 200 patients is surveyed about readmission risk communication. Of these, 134 report that their discharge instructions were clear and complete.
   (a) Compute $\hat{p}$ and verify the conditions for the proportion CI.
   (b) Construct a 95% confidence interval.
   (c) Construct a 99% confidence interval.
   (d) The hospital's benchmark is 75% clarity. Does your interval provide evidence the benchmark is met?
   (e) How does the interval change if only 100 patients were sampled with the

same $\hat{p}$?

5. In an audit of 150 medication administrations, 12 contained an error.

(a) Compute $\hat{p}$.

(b) Check the conditions for the proportion CI.

(c) Construct a 95% confidence interval for the true medication error rate.

(d) The hospital's target is an error rate below 5%. What does the CI suggest?

Part C: Sample Size and Interpretation

6. $\star$ A nursing school wants to estimate the proportion of its graduates who pass the NCLEX-RN on the first attempt. From national data, the pass rate is approximately 85%.

(a) How many graduates must be sampled to estimate the proportion within 4 percentage points at 95% confidence? Use $\hat{p} = 0.85$.

(b) How many if the margin of error must be 2 percentage points?

(c) How many using the conservative $\hat{p} = 0.5$ for a 4-point margin of error?

(d) A prior estimate of $\hat{p} = 0.85$ vs. $\hat{p} = 0.5$ leads to very different sample sizes. Why is using a prior estimate preferred when available?

7. $\star$ **Research Application:** A clinical pharmacist wants to estimate the mean daily number of medication reconciliation discrepancies per patient in a large ICU. A pilot study of 20 patients gave $\bar{x} = 3.4$ and $s = 1.8$ discrepancies.

(a) Compute the 95% CI from the pilot study.

(b) The pharmacist wants the full study's margin of error to be at most 0.4 discrepancies. How large a sample is needed at 95% confidence? Use $\sigma \approx s = 1.8$ for the calculation.

(c) How does the required n change if the desired ME is halved to 0.2?

(d) If the 95% CI from the full study (with the n from part b) were reported as $(3.1, 3.9)$, what is the correct interpretation?

(e) A colleague says: "The true mean is probably around 3.4 because that was the pilot result." Is this a valid statistical interpretation? Explain.

Answer Key — Selected Problems

Answer Key

Section 11.1 Practice — Your Turn

1. (a) $SE = 480/\sqrt{50} \approx 67.9$ steps. (b) `tdist(49).inversecdf(0.975)` ≈ 2.010. (c) $ME = 2.010 \times 67.9 \approx 136$ steps; CI: $(1{,}240 \pm 136) = (1{,}104,\ 1{,}376)$ steps. (d) We are 95% confident the true mean daily step count for hospitalized patients is between 1,104 and 1,376 steps. (e) $n = (t^* \cdot s/E)^2 \approx (1.96 \times 480/100)^2 = (9.408)^2 \approx 88.5 \Rightarrow n = 89$ (use $z^* = 1.96$ for large-sample planning).

2. (a) $\sum x = 1{,}747$; $\bar{x} = 1{,}747/12 \approx 145.6$ mmHg; $s \approx 9.68$ mmHg. (b) t^*, $df = 11$: `tdist(11).inversecdf(0.975)` ≈ 2.201. (c) $SE = 9.68/\sqrt{12} \approx 2.794$; $ME = 2.201 \times 2.794 \approx 6.15$; 95% CI: $(139.4,\ 151.7)$ mmHg. (d) 99% CI: $t^* \approx 3.106$; $ME \approx 8.68$; CI: $(136.9,\ 154.3)$ mmHg — wider, as expected. (e) The 95% CI $(139.4, 151.7)$ crosses 150, so the benchmark is plausible but not certain; the data do not conclusively show the mean is below 150 mmHg.

Section 11.2 Practice — Your Turn

1. (a) $\hat{p} = 96/120 = 0.800$. (b) Conditions: random sample ✓; $n\hat{p} = 96 \geq 10$ ✓; $n(1-\hat{p}) = 24 \geq 10$ ✓. (c) $SE = \sqrt{0.8 \times 0.2/120} \approx 0.0365$; $ME = 1.96 \times 0.0365 \approx 0.0715$; 95% CI: $(0.729,\ 0.872)$ or (72.9%, 87.2%). (d) 90% CI: $z^* = 1.645$; $ME \approx 0.060$; CI: $(0.740,\ 0.860)$ — narrower because lower confidence. (e) The lower bound of 72.9% falls below 85%; the interval does not conclusively support the target being met.

2. (a) $n = (1.96/0.03)^2(0.12)(0.88) = (65.33)^2(0.1056) = 4{,}268.4 \times 0.1056 \approx 450.7 \Rightarrow n = 451$. (b) $n = (1.96/0.03)^2(0.5)(0.5) = 4{,}268.4 \times 0.25 \approx 1{,}067.1 \Rightarrow n = 1{,}068$. (c) $\hat{p} = 0.5$ gives maximum $p(1-p) = 0.25$, larger than $0.12 \times 0.88 = 0.1056$; more uncertainty means more observations needed. (d) $ME = 1.96\sqrt{0.12 \times 0.88/200} = 1.96 \times 0.0230 \approx 0.045$ (4.5 percentage points).

Chapter Practice — Selected

Problem 1c: $SE = 9.6/\sqrt{36} = 1.6$; $t^* \approx 2.030$ ($df = 35$); $ME = 2.030 \times 1.6 \approx 3.25$; 95% CI: $(71.0,\ 77.5)$ bpm.

Problem 2a: $\bar{x} = (3.1+2.8+2.4+3.3+2.9+2.6+3.0+2.7)/8 = 22.8/8 = 2.85$ g/dL; $s \approx 0.289$ g/dL.

Problem 2c: $t^* \approx 2.365$ ($df = 7$, 95%); $SE = 0.289/\sqrt{8} \approx 0.102$; $ME \approx 0.242$; 95% CI: $(2.61,\ 3.09)$ g/dL. The entire interval is below 3.5 — this population is clearly below normal albumin.

Problem 3d: $ME = 0.5$ (half the CI width: $(5.8 - 4.8)/2 = 0.5$); $SE = ME/z^* \approx 0.5/1.96 \approx 0.255$; $s \approx SE \times \sqrt{n} = 0.255 \times 10 = 2.55$. For ME $= 0.3$: $n \approx (1.96 \times 2.55/0.3)^2 = (16.66)^2 \approx 278$.

Problem 6a: $n = (1.96 \times 0.85 \times 0.15)^{1/2}$... use formula: $n = (1.96/0.04)^2(0.85)(0.15) = (49)^2(0.1275) = 2{,}401 \times 0.1275 \approx 306.1 \Rightarrow n = 307$.

Problem 7b: $n = (1.96 \times 1.8/0.4)^2 = (8.82)^2 \approx 77.8 \Rightarrow n = 78$.

Chapter 11 Summary

Section 11.1 — Confidence Intervals for a Mean

- A **confidence interval** provides a range of plausible values for μ at a stated level of confidence.
- z**-interval** (large n or σ known): $\bar{x} \pm z^* \cdot (\sigma/\sqrt{n})$.
- t**-interval** (σ unknown, practical default): $\bar{x} \pm t^* \cdot (s/\sqrt{n})$, with $df = n - 1$.
- Critical values: $z^* = 1.645$ (90%), 1.960 (95%), 2.576 (99%).
- Higher confidence $\Rightarrow$ wider interval. Larger $n \Rightarrow$ narrower interval.
- **Interpretation:** "We are 95% confident that μ is in this interval" — not a probability about a fixed interval.

Section 11.2 — Confidence Intervals for a Proportion and Sample Size

- **Proportion CI:** $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$; requires $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
- **Sample size for proportion:** $n = (z^*/E)^2 \hat{p}(1-\hat{p})$; use $\hat{p} = 0.5$ if no prior estimate.
- **Sample size for mean:** $n = (z^* \sigma/E)^2$; requires a prior estimate of σ.
- Always round n **up** to the nearest whole number.

The Nursing Connection

- A confidence interval is more informative than a point estimate alone — always report both.
- The *width* of the CI reflects how much information the study provided; a wide CI means low precision.
- For clinical decisions, check whether the entire CI falls on one side of a threshold — or whether it straddles it.
- Sample size planning is an ethical obligation in research: too small a sample wastes resources and produces unreliable intervals; too large wastes patient participation unnecessarily.

CHAPTER 12

Hypothesis Testing: The Logic

> *"Before you can prove something, you must first take seriously the possibility that you are wrong."*
>
> — the philosophical core of the hypothesis test

In this chapter, you will learn to:

- State the null and alternative hypotheses for a research question
- Define and interpret the p-value
- Describe Type I and Type II errors and their clinical consequences
- Perform and interpret a one-sample z-test for a population mean
- Perform and interpret a one-sample t-test for a population mean
- Perform and interpret a one-sample z-test for a population proportion
- Connect hypothesis testing to confidence intervals

12.1 The Logic of Hypothesis Testing

Step 1 Read This First

A hospital has used a standard wound care protocol for years. The known mean healing time under the old protocol is $\mu_0 = 14.2$ days. A nurse researcher believes a new protocol will reduce healing time. She trains 30 nurses in the new method and tracks outcomes for 30 patients.

The sample mean under the new protocol is $\bar{x} = 12.8$ days — 1.4 days faster. But the researcher knows that random variation alone could produce a difference of 1.4 days even if the new protocol does nothing at all. The sample just happened to include faster-healing patients.

The central question of hypothesis testing: *Is the observed difference (12.8 vs. 14.2 days) large enough to be convincing evidence that the new protocol actually works — or is it plausibly just sampling variation?*

This is not a question answered by intuition. It requires a precise probability calculation.

Step 2 Let's Talk About It

The logic of hypothesis testing is the logic of a courtroom.

In a criminal trial, the defendant is presumed innocent until proven guilty. The prosecution must provide evidence strong enough to overcome that presumption. "Innocent until proven guilty" is the *default* position; the burden of proof lies with those who want to claim otherwise.

In hypothesis testing, the default position is that *nothing has changed*: the new protocol works the same as the old one, $\mu = 14.2$ days. This is the **null hypothesis**. The researcher's claim — that the new protocol is faster — is the **alternative hypothesis**.

The test asks: *If the null hypothesis were true (nothing has changed), how likely is it to observe a sample mean as low as 12.8 days just by chance?* If the answer is "very unlikely," the data constitute strong evidence against the null hypothesis. If the answer is "quite plausible," the data do not provide sufficient evidence for the researcher's claim.

Notice what is *not* happening: we are not proving the null hypothesis is false. We are assessing whether the data are consistent with it.

Step 3 Now We Name It

Definition: Null and Alternative Hypotheses

The **null hypothesis** H_0 is the default claim — the assumption of no effect, no difference, or no change. It always contains an equality ($=$, $\leq$, or $\geq$).

The **alternative hypothesis** H_a (also written H_1) is the researcher's claim — what they are trying to find evidence for. It contains a strict inequality ($\neq$, $<$, or $>$).

Types of tests based on H_a:

- **Two-tailed:** $H_a : \mu \neq \mu_0$ (effect could be in either direction)
- **Left-tailed:** $H_a : \mu < \mu_0$ (testing for a decrease)
- **Right-tailed:** $H_a : \mu > \mu_0$ (testing for an increase)

The direction of H_a determines which tail of the distribution is used to compute the p-value.

Definition: The P-Value

The **p-value** is the probability of obtaining a test statistic as extreme as (or more extreme than) the observed value, *assuming the null hypothesis is true.*

- A **small p-value** (typically $p < 0.05$) means the observed result is unlikely under H_0. This is evidence against H_0 — we **reject** H_0.
- A **large p-value** ($p \geq 0.05$) means the observed result is plausible under H_0. We **fail to reject** H_0.

Critical reminder: Failing to reject H_0 does *not* mean H_0 is true. It means the data do not provide sufficient evidence to conclude otherwise.

Definition: Significance Level α and Decision Rule

The **significance level** α is the threshold for deciding whether a p-value is "small enough" to reject H_0. The most common choice is $\alpha = 0.05$.

Decision rule:

- If p-value $< \alpha$: reject H_0. The result is **statistically significant**.
- If p-value $\geq \alpha$: fail to reject H_0. The result is **not statistically significant**.

Definition: Type I and Type II Errors

	H_0 is actually true	H_0 is actually false
Reject H_0	**Type I error** (α)	Correct decision (power)
Fail to reject H_0	Correct decision	**Type II error** (β)

Type I error: Concluding there is an effect when there is none. Probability $= \alpha$.
Type II error: Missing a real effect. Probability $= \beta$.
Power $= 1 - \beta$: the probability of correctly detecting a real effect.

In clinical settings: a Type I error might lead to adopting an ineffective treatment. A Type II error might lead to abandoning an effective one. The relative costs of each error should influence the choice of α.

Step 4 Watch It Work

Setting up the wound care hypothesis test

Research question: Does the new wound care protocol reduce mean healing time below 14.2 days?

Step 1 — Hypotheses:
$H_0 : \mu = 14.2$ days (the new protocol is no different from the old)
$H_a : \mu < 14.2$ days (the new protocol reduces healing time)
This is a **left-tailed** test.

Step 2 — Significance level: $\alpha = 0.05$

Step 3 — Identify the error types in this context:

	New protocol is actually the same	New protocol actually works
We conclude it works	**Type I error:** adopt a useless protocol	Correct
We conclude it doesn't work	Correct	**Type II error:** miss an effective protocol

Which error is more harmful here? Adopting a protocol that wastes resources (Type I) vs. missing a protocol that genuinely helps patients heal faster (Type II). In most quality improvement contexts, the Type II error is more clinically costly — a false negative means patients continue receiving a less effective treatment.

Step 4 — What would the p-value tell us?
If $p < 0.05$: The observed sample mean ($\bar{x} = 12.8$ days) is unlikely enough under H_0 that we reject H_0 and conclude the new protocol significantly reduces healing time.
If $p \geq 0.05$: The observed difference could plausibly be due to chance. We do not have sufficient evidence to conclude the new protocol works — but we cannot conclude it does not work either.

Step 5 Your Turn

1. For each research question, state H_0 and H_a, identify the test direction, and describe the Type I and Type II errors in clinical terms.
   (a) A hospital wants to know whether a new medication reduces mean systolic blood pressure below the current mean of 148 mmHg in hypertensive patients.

   (b) A nurse researcher wants to test whether the proportion of patients who fall is different from the hospital's historical rate of 8%.

(c) A quality improvement team tests whether average medication administration time has increased above the target of 20 minutes since a new electronic system was implemented.

(d) A clinical pharmacist tests whether the mean creatinine clearance for patients on a new drug is greater than 60 mL/min.

2. A p-value of 0.032 is obtained for a hypothesis test. The significance level is $\alpha = 0.05$.

(a) What decision do you make about H_0?
(b) Does this p-value prove that H_a is true? Explain.
(c) What is the probability you are making a Type I error if you reject H_0?
(d) If instead $\alpha = 0.01$, what decision would you make? Does the data change?

3. ⋆ A clinical trial tests a new antibiotic. The null hypothesis is that the drug has no effect on infection clearance rate. The significance level is set at $\alpha = 0.01$.

(a) Why might researchers use $\alpha = 0.01$ rather than $\alpha = 0.05$ here?
(b) Setting $\alpha = 0.01$ reduces the probability of which type of error?
(c) What happens to the probability of the other type of error as a result?
(d) Describe the clinical consequences of each type of error in this specific context.
(e) When would you want a higher α (e.g., 0.10)? Give a clinical example.

Step 6 Think Like a Nurse

Statistical Significance Is Not Clinical Significance

A large hospital system conducts a study of $n = 2{,}500$ patients comparing a new fall prevention protocol to the standard protocol. The results show a statistically significant reduction in fall rates: $p = 0.003$.

The actual difference? The new protocol reduced falls from 8.2% to 7.9% — a reduction of 0.3 percentage points. In absolute terms, across 2,500 patients, that is about 7 fewer falls.

The result is statistically significant. Is it clinically meaningful?

That depends on costs. If the new protocol requires 2 additional hours of staff training per week, costs $50,000 to implement system-wide, and disrupts existing workflows for months — the marginal benefit of preventing 7 falls per 2,500 patients may not justify the resources.

The p-value only tells you whether the observed difference is unlikely to be due to chance. It says nothing about whether the difference is large enough to matter in practice. A very large sample can detect extremely small differences as statistically significant. A very small sample may fail to detect a large, clinically meaningful difference as statistically significant.

Always pair the p-value with the effect size and a confidence interval:

- P-value: is the result real or due to chance?
- Effect size: how large is the difference?
- Confidence interval: how precisely have we estimated the effect?

Statistical significance is a threshold, not a measure of importance. Clinical significance requires judgment that no formula can supply.

12.2 Performing Hypothesis Tests

Step 1 Read This First

Returning to the wound care study: the researcher has 30 patients treated with the new protocol. The results are $\bar{x} = 12.8$ days and $s = 4.2$ days. The historical mean under the old protocol is $\mu_0 = 14.2$ days.

Hypotheses: $H_0 : \mu = 14.2$ vs. $H_a : \mu < 14.2$ at $\alpha = 0.05$.

Now it is time to compute the p-value and make a decision. The procedure is the same framework used throughout Chapter 10: compute a z- or t-score, then find the corresponding probability.

Step 2 Let's Talk About It

The test statistic measures how far the sample mean is from the null hypothesis value — in standard error units. The further the sample mean is from μ_0, the more evidence against H_0.

For a left-tailed test, a very negative test statistic (sample mean much lower than μ_0) provides evidence against H_0. The p-value is the probability of getting a test statistic *this negative or more negative* if H_0 is true.

This is identical to the probability calculations in Chapters 9 and 10 — with one key difference: instead of asking about an observed data value or a sample mean from a process we know, we are asking how unusual the observed sample mean would be under a hypothetical world where H_0 is true.

Step 3 Now We Name It

Definition: One-Sample z-Test for μ (σ Known)

When σ is known and either the population is normal or $n \geq 30$:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

P-value depends on the direction of H_a:

- Left-tailed ($H_a : \mu < \mu_0$): $p = P(Z \leq z)$
- Right-tailed ($H_a : \mu > \mu_0$): $p = P(Z \geq z) = 1 - P(Z \leq z)$
- Two-tailed ($H_a : \mu \neq \mu_0$): $p = 2 \times P(Z \leq -|z|)$

Definition: One-Sample t-Test for μ (σ Unknown)

When σ is unknown (the practical default), use the sample standard deviation s:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

with $df = n - 1$. The p-value is computed from the t-distribution with $df = n - 1$.

Conditions: Random sample; either $n \geq 30$ or population approximately normal; independent observations.

Definition: One-Sample z-Test for a Proportion

To test $H_0 : p = p_0$ when $np_0 \geq 10$ and $n(1 - p_0) \geq 10$:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Note: Use p_0 (the null value), not $\hat{p}$, in the standard error formula — because we are computing probabilities *under the assumption that* H_0 is true.

Definition: Five-Step Hypothesis Testing Framework

1. **State hypotheses:** Write H_0 and H_a clearly.
2. **Check conditions:** Random sample, independence, sample size.
3. **Compute test statistic:** z or t formula.
4. **Find p-value:** Use Desmos, Excel, or a table.
5. **Make a decision and interpret:** Compare p-value to α; state conclusion in context.

Step 4 Watch It Work

Three complete hypothesis tests

Test 1 — One-sample t-test: wound care healing time
$H_0 : \mu = 14.2 \quad H_a : \mu < 14.2 \quad \alpha = 0.05$
$n = 30$, $\bar{x} = 12.8$, $s = 4.2$, $df = 29$

Conditions: Random sample ✓; $n = 30 \geq 30$ ✓; independent ✓.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{12.8 - 14.2}{4.2/\sqrt{30}} = \frac{-1.4}{0.7669} \approx -1.826$$

P-value (left-tailed): `tdist(29).cdf(-1.826)` ≈ 0.0392
Decision: $p = 0.039 < 0.05 \quad \Rightarrow$ **Reject H_0.**
Conclusion: There is sufficient evidence at $\alpha = 0.05$ to conclude that the new protocol reduces mean healing time below 14.2 days.

Test 2 — One-sample z-test for proportion: fall rate
A hospital's historical fall rate is 8%. After a new fall prevention program, 12 of 200 patients fell.
$H_0 : p = 0.08 \quad H_a : p < 0.08 \quad \alpha = 0.05$
$n = 200$, $\hat{p} = 12/200 = 0.060$

Conditions: $np_0 = 200(0.08) = 16 \geq 10$ ✓; $n(1 - p_0) = 184 \geq 10$ ✓.

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.060 - 0.080}{\sqrt{0.08 \times 0.92/200}} = \frac{-0.020}{0.01918} \approx -1.043$$

P-value (left-tailed): $P(Z \leq -1.043) \approx 0.1484$
Decision: $p = 0.148 \geq 0.05 \quad \Rightarrow$ **Fail to reject H_0.**
Conclusion: Insufficient evidence at $\alpha = 0.05$ to conclude the program reduced the fall rate below 8%.

Test 3 — Two-tailed t-test: medication administration time
Target mean is 20 minutes. A sample of $n = 25$ gives $\bar{x} = 22.3$ min, $s = 5.8$ min.
$H_0 : \mu = 20 \quad H_a : \mu \neq 20 \quad \alpha = 0.05, df = 24$

$$t = \frac{22.3 - 20}{5.8/\sqrt{25}} = \frac{2.3}{1.16} \approx 1.983$$

P-value (two-tailed): $2 \times P(T_{24} \geq 1.983) \approx 2 \times 0.0294 = 0.0588$
Decision: $p = 0.059 \geq 0.05 \quad \Rightarrow$ **Fail to reject H_0.**
Conclusion: Insufficient evidence at $\alpha = 0.05$ that mean administration time differs from 20 minutes. The result is borderline ($p = 0.059$); a larger sample would provide more conclusive evidence.

Try This in Desmos

Left-tailed t: `tdist(29).cdf(-1.826)` ≈ 0.039

Two-tailed t: `2*(1 - tdist(24).cdf(1.983))` ≈ 0.059

Left-tailed z: `normaldist(0,1).cdf(-1.043)` ≈ 0.148

Try This in Excel

`=T.DIST(-1.826, 29, TRUE)` ≈ 0.039 (left-tailed t)

`=T.DIST.2T(ABS(1.983), 24)` ≈ 0.059 (two-tailed t)

`=NORM.S.DIST(-1.043, TRUE)` ≈ 0.148 (left-tailed z)

Step 5 Your Turn

1. A nursing home reports that their mean resident satisfaction score is 78 out of 100. A quality improvement team suspects scores have declined after a staffing change. They survey $n = 40$ residents and find $\bar{x} = 74.2$ and $s = 11.6$.
   (a) State H_0 and H_a.
   (b) Check all conditions.
   (c) Compute the t-test statistic.
   (d) Find the p-value using Desmos or Excel.
   (e) State your conclusion at $\alpha = 0.05$ in context.

2. A hospital's historical rate of 30-day readmission is 12%. After implementing a new discharge education program, 22 of 250 patients were readmitted.
   (a) State H_0 and H_a.
   (b) Check the conditions for the proportion test.
   (c) Compute the z-test statistic.
   (d) Find the p-value and state your conclusion at $\alpha = 0.05$.
   (e) Does the program appear effective? Is the p-value alone sufficient to answer this question?

3. $\star$ ICU nurses at a hospital are tested on sepsis protocol knowledge. The national benchmark is a mean score of 72%. A random sample of 20 ICU nurses scores a mean of 76.4% with $s = 8.9\%$.
   (a) State H_0 and H_a for a two-tailed test.
   (b) Check all conditions. Is the t-test appropriate here?
   (c) Compute the t-statistic and find the p-value ($df = 19$).
   (d) State your conclusion at $\alpha = 0.05$.

(e) Construct a 95% CI and verify it agrees with the test result.

Step 6 Think Like a Nurse

The Hypothesis Test and the Confidence Interval Are Two Sides of the Same Coin

Every hypothesis test has a corresponding confidence interval, and they will always agree.

For two-tailed tests, the agreement is precise: if a value μ_0 falls *outside* the 95% confidence interval ($\bar{x} \pm t^* \cdot s/\sqrt{n}$), then the two-tailed test at $\alpha = 0.05$ will reject $H_0 : \mu = \mu_0$. If μ_0 falls *inside* the interval, the test will fail to reject.

In Test 3 above, the two-tailed t-test failed to reject $H_0 : \mu = 20$ at $\alpha = 0.05$. The corresponding 95% CI is:

$$22.3 \pm 2.064 \times \frac{5.8}{\sqrt{25}} = 22.3 \pm 2.394 = (19.9,\ 24.7)$$

The null value 20 is *inside* the interval $(19.9,\ 24.7)$ — consistent with failing to reject.

This equivalence means a confidence interval gives you more information than a hypothesis test: instead of just "reject" or "fail to reject," it tells you the entire range of values consistent with the data.

Practical guideline:

- Use the hypothesis test for a binary decision: is there evidence of an effect?
- Use the confidence interval to quantify the effect: how large is it, and how precisely?
- Use both together for complete communication of evidence.

A p-value of 0.039 says "we found something." A confidence interval of (11.2, 14.4) days says "we found something about this big, with this much uncertainty." The interval is almost always the more useful communication.

12.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 12. Problems marked with ⋆ are more challenging.

Part A: Hypotheses, P-Values, and Errors

1. State the null and alternative hypotheses for each scenario. Identify the test direction.
   - (a) A hospital tests whether a new triage system changes mean ED wait time from the current 48 minutes (it could increase or decrease).
   - (b) A vaccine trial tests whether the proportion who develop immunity exceeds 85%.
   - (c) A study tests whether average nurse overtime has fallen below 10 hours per week since a new scheduling policy.
   - (d) A researcher tests whether mean pain scores after a new analgesic differ from the standard drug's mean of 4.8.

2. A study reports $p = 0.072$ for a test at $\alpha = 0.05$. Identify each statement as correct or incorrect.
   - (a) "There is a 7.2% probability that the null hypothesis is true."
   - (b) "The result is not statistically significant at $\alpha = 0.05$."
   - (c) "We have proven that there is no effect."
   - (d) "If $\alpha = 0.10$, the result would be statistically significant."
   - (e) "The probability of a Type I error, if we reject H_0 at $\alpha = 0.05$, is 0.05."

Part B: One-Sample Tests for a Mean

3. The standard central line insertion procedure takes a mean of 38 minutes. After a simulation training program, $n = 36$ nurses perform the procedure with $\bar{x} = 35.1$ min and $s = 8.4$ min. Test whether training reduced procedure time.
   - (a) State H_0 and H_a.
   - (b) Compute the t-statistic ($df = 35$).
   - (c) Find the p-value and state your conclusion at $\alpha = 0.05$.
   - (d) Construct a 95% CI and verify it is consistent with the test result.

4. A random sample of 16 NICU nurses has a mean stress score of 68.4 (out of 100) with $s = 12.1$. The national mean for ICU nurses is 62.0.
   - (a) State H_0 and H_a for a two-tailed test.
   - (b) Check all conditions.
   - (c) Compute the t-statistic ($df = 15$) and find the p-value.
   - (d) State your conclusion at $\alpha = 0.05$.

  (e) At what α would this result first become statistically significant?

Part C: One-Sample Tests for a Proportion

5. A hospital reports that 62% of discharged patients are fully medication-compliant at 30 days. After a pharmacist counseling intervention, 85 of 120 patients are compliant.
  (a) State H_0 and H_a (test whether compliance has increased).
  (b) Check the conditions.
  (c) Compute the z-statistic.
  (d) Find the p-value and state your conclusion at $\alpha = 0.05$.
  (e) Construct a 95% CI and compare to the test result.

6. $\star$ A nurse researcher claims fewer than 20% of chronic pain patients use opioids as their only pain management strategy. In a sample of 300 patients, 48 use opioids only.
  (a) State H_0 and H_a.
  (b) Check conditions and compute the z-statistic.
  (c) Find the p-value. State your conclusion at $\alpha = 0.01$.
  (d) Compute the 99% CI and comment on alignment with the test.

7. $\star$ **Research Application:** Before a hand hygiene reminder system, compliance was $p_0 = 0.72$. After three months, 183 of 240 observations showed compliance.
  (a) Test whether compliance has changed (two-tailed) at $\alpha = 0.05$.
  (b) Compute the absolute and relative change in proportion.
  (c) Construct a 95% CI for the new compliance proportion.
  (d) Using the CI, discuss whether the change, if real, is clinically meaningful.
  (e) The study was funded by the company that makes the reminder device. How does this affect your interpretation?

Answer Key — Selected Problems

Answer Key

Section 12.1 Practice — Your Turn

1. (a) $H_0 : \mu = 148$; $H_a : \mu < 148$; left-tailed. Type I: conclude drug lowers BP when it doesn't — patient takes unnecessary medication. Type II: miss a drug that genuinely lowers BP — patient denied effective treatment. (b) $H_0 : p = 0.08$; $H_a : p \neq 0.08$; two-tailed. Type I: declare fall rate changed when it hasn't. Type II: miss a genuine change. (c) $H_0 : \mu = 20$; $H_a : \mu > 20$; right-tailed. Type I: conclude time increased when it hasn't — unwarranted intervention. Type II: miss a real increase — safety concern unaddressed. (d) $H_0 : \mu = 60$; $H_a : \mu > 60$; right-tailed.

2. (a) Reject H_0 since $p = 0.032 < 0.05$. (b) No — a small p-value only says the result is unlikely under H_0; it does not prove H_a or quantify the effect size. (c) The probability of a Type I error equals $\alpha = 0.05$, not the p-value of 0.032. (d) At $\alpha = 0.01$: $p = 0.032 \geq 0.01$, so fail to reject H_0. The data do not change; only the decision threshold changes.

Section 12.2 Practice — Your Turn

1. (a) $H_0 : \mu = 78$; $H_a : \mu < 78$. (b) $n = 40 \geq 30$ ✓; random sample assumed ✓; independent ✓. (c) $SE = 11.6/\sqrt{40} \approx 1.835$; $t = (74.2 - 78)/1.835 \approx -2.071$; $df = 39$. (d) `tdist(39).cdf(-2.071)` ≈ 0.0224. (e) $p = 0.022 < 0.05$; reject H_0. There is sufficient evidence that satisfaction scores declined below 78 after the staffing change.

2. (a) $H_0 : p = 0.12$; $H_a : p < 0.12$. (b) $np_0 = 250(0.12) = 30 \geq 10$ ✓; $n(1 - p_0) = 220 \geq 10$ ✓. (c) $\hat{p} = 22/250 = 0.088$; $z = (0.088 - 0.12)/\sqrt{0.12 \times 0.88/250} = -0.032/0.02055 \approx -1.557$. (d) $P(Z \leq -1.557) \approx 0.060$; fail to reject H_0. The observed reduction is not statistically significant; p-value alone is insufficient — we should also consider the effect size (reduction from 12% to 8.8%) and a CI.

Chapter Practice — Selected

Problem 3b: $SE = 8.4/\sqrt{36} = 1.4$; $t = (35.1 - 38)/1.4 = -2.9/1.4 \approx -2.071$; $df = 35$.

Problem 3c: `tdist(35).cdf(-2.071)` ≈ 0.023; $p < 0.05$; reject H_0. Training significantly reduced procedure time.

Problem 3d: 95% CI: $t^* \approx 2.030$ ($df = 35$); $ME = 2.030 \times 1.4 \approx 2.84$; CI: $(32.3,\ 37.9)$ min. The value 38 lies above the upper bound, consistent with rejecting $H_0 : \mu = 38$.

Problem 4c: $SE = 12.1/\sqrt{16} = 3.025$; $t = (68.4 - 62)/3.025 \approx 2.116$; two-tailed p-value: `2*(1 - tdist(15).cdf(2.116))` ≈ 0.051.

Problem 4d: $p \approx 0.051 \geq 0.05$; fail to reject H_0 at $\alpha = 0.05$ (borderline).

Problem 5c: $\hat{p} = 85/120 \approx 0.7083$; $z = (0.7083 - 0.62)/\sqrt{0.62 \times 0.38/120} = 0.0883/0.04431 \approx 1.994$.

Problem 5d: Right-tailed: $P(Z \geq 1.994) \approx 0.023$; $p < 0.05$; reject H_0. Significant evidence compliance increased above 62%.
Problem 7a: $\hat{p} = 183/240 = 0.7625$; $z = (0.7625 - 0.72)/\sqrt{0.72 \times 0.28/240} = 0.0425/0.02898 \approx 1.467$; two-tailed: $p \approx 2(0.0713) = 0.143$; fail to reject H_0 at $\alpha = 0.05$.

Chapter 12 Summary

Section 12.1 — The Logic of Hypothesis Testing

- H_0 is the default (no effect); H_a is the research claim.
- **P-value:** probability of data as extreme as observed, if H_0 is true.
- Reject H_0 when $p < \alpha$; fail to reject when $p \geq \alpha$.
- **Type I error** (α): reject H_0 when it is true (false positive).
- **Type II error** (β): fail to reject H_0 when it is false (false negative).
- **Power** $= 1 - \beta$: probability of correctly detecting a real effect.
- Statistical significance $\neq$ clinical significance. Always check effect size and CI.

Section 12.2 — Performing Hypothesis Tests

- Five steps: state hypotheses, check conditions, compute statistic, find p-value, decide and interpret.
- **One-sample t-test:** $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, $df = n - 1$; use when σ unknown.
- **Proportion z-test:** $z = (\hat{p} - p_0)/\sqrt{p_0(1 - p_0)/n}$; use p_0 (not $\hat{p}$) in SE.
- Two-tailed CI and two-tailed test always agree: μ_0 outside CI $\Leftrightarrow$ test rejects H_0.

The Nursing Connection

- Every quality improvement study implicitly tests a hypothesis: did this intervention change outcomes?
- A p-value below 0.05 says “the effect is real”; it does not say “the effect is large enough to matter.”
- Type I and Type II errors carry clinical costs that should inform the choice of α.
- Report the CI alongside the p-value — it communicates how large the effect is, not just whether it exists.

CHAPTER 13

Hypothesis Testing: In Practice

> *"Most clinical questions involve comparing two groups, not one group to a fixed number. The tools change; the logic does not."*
>
> — extending hypothesis testing to two-sample and categorical settings

In this chapter, you will learn to:

- Perform and interpret a two-sample t-test for comparing two independent group means
- Perform and interpret a paired t-test for matched or before-after data
- Distinguish between independent and paired designs and choose the appropriate test
- Perform and interpret a two-sample z-test for comparing two proportions
- Perform and interpret a chi-square test for independence in a two-way table
- Apply all tests using the five-step framework from Chapter 12

13.1 Comparing Two Groups: t-Tests and Proportion Tests

Step 1 Read This First

A hospital is evaluating two wound care dressings. Patients are randomly assigned to either the standard dressing (Group A, $n_1 = 35$) or a new hydrocolloid dressing (Group B, $n_2 = 33$). The number of days to wound closure is recorded for each patient.

Group A (standard): $\bar{x}_1 = 14.8$ days, $s_1 = 4.1$ days
Group B (hydrocolloid): $\bar{x}_2 = 12.3$ days, $s_2 = 3.6$ days

Group B healed 2.5 days faster on average. But is this difference real, or could it be explained by random variation between the two groups? This is a **two-sample** problem: we are comparing two independent groups rather than comparing one group to a known population value.

Similarly, suppose a hospital wants to know whether the proportion of patients who develop pressure injuries differs between day and night shifts. Both require comparing two groups — a setting that Chapter 12 could not handle.

Step 2 Let's Talk About It

The logic is exactly the same as in Chapter 12 — we just need to extend it.

Instead of asking "is the sample mean different from a fixed number?" we now ask "are the two sample means different from each other?" The null hypothesis becomes $H_0 : \mu_1 = \mu_2$, or equivalently, $\mu_1 - \mu_2 = 0$.

The test statistic measures how far the observed difference $(\bar{x}_1 - \bar{x}_2)$ is from the null value of 0, measured in standard error units. The standard error for a difference in means combines the variability from both groups.

For two proportions, the same idea applies: $H_0 : p_1 = p_2$. A pooled estimate of the common proportion is used in the standard error, because under the null hypothesis the two groups share the same proportion.

Step 3 Now We Name It

Definition: Two-Sample t-Test for $\mu_1 - \mu_2$ (Independent Groups)

To test whether two independent population means differ:
$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 \neq \mu_2$ (or $<$ or $>$)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Degrees of freedom: use the conservative $df = \min(n_1 - 1,\ n_2 - 1)$ for hand calculation, or Welch's formula via Desmos/Excel for a more precise value.

Conditions: Both samples are random and independent of each other; each sample is either $n \geq 30$ or from an approximately normal population.

Definition: Paired t-Test

When each observation in Group 1 is naturally matched with one in Group 2 (e.g., before-after measurements on the same patient, or matched pairs), compute the **difference** $d_i = x_{1i} - x_{2i}$ for each pair and apply a one-sample t-test to the differences:

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

where $\bar{d}$ is the mean of the differences, s_d is the standard deviation of the differences, and n is the number of pairs. Degrees of freedom: $df = n - 1$.

Use paired when: the same subjects are measured twice; subjects are matched on important characteristics; natural pairing reduces extraneous variation.
Use independent when: subjects are randomly assigned to different groups with no natural link between individuals across groups.

Definition: Two-Sample z-Test for $p_1 - p_2$

To test whether two population proportions differ:
$H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$ (or $<$ or $>$)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1 - \hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_c = \dfrac{x_1 + x_2}{n_1 + n_2}$ is the **pooled proportion** — the combined proportion across both groups.

Conditions: $n_1\hat{p}_c \geq 10$, $n_1(1 - \hat{p}_c) \geq 10$, $n_2\hat{p}_c \geq 10$, $n_2(1 - \hat{p}_c) \geq 10$.

Step 4 Watch It Work

Two complete tests: two-sample t-test and paired t-test

Test 1 — Two-sample t-test: wound dressing comparison

$H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 > \mu_2$ (standard takes longer) $\quad \alpha = 0.05$

$n_1 = 35,\ \bar{x}_1 = 14.8,\ s_1 = 4.1; \quad n_2 = 33,\ \bar{x}_2 = 12.3,\ s_2 = 3.6$

Standard error:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{16.81}{35} + \frac{12.96}{33}} = \sqrt{0.4803 + 0.3927} = \sqrt{0.8730} \approx 0.9343$$

Test statistic: $df = \min(34, 32) = 32$

$$t = \frac{14.8 - 12.3}{0.9343} = \frac{2.5}{0.9343} \approx 2.676$$

P-value (right-tailed): `1 - tdist(32).cdf(2.676)` ≈ 0.0058

Decision: $p = 0.006 < 0.05 \quad \Rightarrow \quad$ **Reject H_0.**

Conclusion: There is significant evidence that the standard dressing takes longer to achieve wound closure than the hydrocolloid dressing ($p = 0.006$). The new dressing reduces mean healing time by approximately 2.5 days.

Test 2 — Paired t-test: blood pressure before and after medication

Six patients have their systolic BP measured before and after starting a new antihypertensive:

Patient	1	2	3	4	5	6
Before	158	162	147	175	153	168
After	143	150	138	159	145	152
d_i (Before − After)	15	12	9	16	8	16

$H_0 : \mu_d = 0 \quad H_a : \mu_d > 0$ (medication reduces BP) $\quad \alpha = 0.05,\ df = 5$

$$\bar{d} = \frac{15 + 12 + 9 + 16 + 8 + 16}{6} = \frac{76}{6} \approx 12.67 \text{ mmHg}$$

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n - 1}} \approx 3.56 \text{ mmHg}$$

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{12.67}{3.56/\sqrt{6}} = \frac{12.67}{1.453} \approx 8.72$$

P-value (right-tailed): `1 - tdist(5).cdf(8.72)` ≈ 0.0002

Decision: $p < 0.05 \quad \Rightarrow \quad$ **Reject H_0.**

Conclusion: There is highly significant evidence ($p \approx 0.0002$) that the medication reduces mean systolic BP. The average reduction is approximately 12.67 mmHg.

Try This in Desmos

Two-sample t: `1 - tdist(32).cdf(2.676)` ≈ 0.006

Paired t: `1 - tdist(5).cdf(8.72)` ≈ 0.0002

For two-proportion test: `normaldist(0,1).cdf(z)` for left-tailed; adjust for direction.

Try This in Excel

Two-sample t (right-tailed): `=T.DIST.RT(2.676, 32)` ≈ 0.006

Paired t (right-tailed): `=T.DIST.RT(8.72, 5)` ≈ 0.0002

Two-proportion z: `=1 - NORM.S.DIST(z, TRUE)` for right-tailed

Step 5 Your Turn

1. A hospital compares medication error rates on two units. Unit A ($n_1 = 200$ administrations): 14 errors. Unit B ($n_2 = 180$ administrations): 7 errors. Test whether the error rates differ between units (two-tailed, $\alpha = 0.05$).
   (a) State H_0 and H_a.
   (b) Compute $\hat{p}_1$, $\hat{p}_2$, and the pooled proportion $\hat{p}_c$.
   (c) Check all conditions.
   (d) Compute the z-statistic.
   (e) Find the p-value and state your conclusion.

2. A nurse researcher measures patient anxiety scores (0–100) before and after a nurse-led relaxation intervention for 8 patients:

Patient	1	2	3	4	5	6	7	8
Before	72	65	81	58	74	69	83	61
After	60	55	71	52	62	58	70	54

   (a) Should you use a paired or independent samples t-test? Justify.
   (b) Compute the differences $d_i = \text{Before} - \text{After}$.
   (c) Find $\bar{d}$ and s_d.
   (d) Perform the paired t-test at $\alpha = 0.05$ (one-tailed: test whether anxiety decreased).
   (e) Interpret the result clinically.

3. $\star$ Two hospitals are compared on 30-day readmission rates after hip replacement surgery. Hospital 1: 18 readmitted out of 150 patients. Hospital 2: 9 readmitted out of 120 patients.

(a) State H_0 and H_a for a two-tailed test.
(b) Compute $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_c$.
(c) Perform the two-proportion z-test at $\alpha = 0.05$.
(d) Construct 95% CIs for each proportion separately. What do the overlapping intervals suggest?
(e) A hospital administrator concludes "our readmission rate is higher, therefore our care is worse." What statistical and clinical cautions would you raise?

Step 6 Think Like a Nurse

Paired vs. Independent: Why the Design Choice Matters
A hospital wants to test whether a new shift handoff protocol reduces adverse events. They consider two designs:

Design A (Independent): Compare 30 patients on units using the old protocol vs. 30 patients on units using the new protocol. Different patients in each group.
Design B (Paired/Before-After): On the same 30 units, count adverse events for one month under the old protocol, then count them for one month under the new protocol.

Both designs address the same question. But Design B (paired) will almost always produce a more powerful test — and here is why.

In Design A, the two groups likely differ in patient acuity, staffing levels, unit culture, and many other factors. These differences add noise to the comparison. The standard error $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ includes all this between-group variation.

In Design B, each unit serves as its own control. The differences d_i (new minus old) capture only the within-unit change. All the fixed differences between units cancel out. The resulting standard error $s_d/\sqrt{n}$ is often much smaller.

A smaller standard error produces a larger test statistic and a smaller p-value for the same true effect. The paired design has more *statistical power* to detect a real improvement.

When possible, design studies so that patients or units serve as their own controls. Matching reduces noise. Reduced noise reveals signal.

13.2 The Chi-Square Test for Independence

Step 1 Read This First

An infection control nurse is investigating whether the rate of hospital-acquired infections (HAI) differs across three patient care units. She audits 300 patients and records both the unit they were on and whether they developed an HAI.

	HAI	**No HAI**	**Total**
Medical Unit	18	82	100
Surgical Unit	24	76	100
ICU	30	70	100
Total	72	228	300

The ICU has the highest HAI rate (30%) and the medical unit has the lowest (18%). But are these differences statistically significant, or could they arise from chance variation alone?

This is neither a one-sample test nor a two-sample test. There are three groups and two outcomes. The tool for this situation is the **chi-square test for independence**.

Step 2 Let's Talk About It

The key idea of the chi-square test is to compare what we *observed* with what we would *expect* if there were no relationship between the two categorical variables.

If HAI rates were identical across all three units, each unit should have roughly the same proportion of HAI cases. Since 72 of 300 patients overall (24%) developed an HAI, we would expect about 24 HAI cases per unit if unit and HAI status were unrelated.

The observed counts are 18, 24, and 30 — Medical is below 24, Surgical is at 24, and ICU is above 24. The chi-square statistic measures how far the observed counts are from these expected counts, weighted by their magnitude.

A large chi-square statistic means the observed pattern deviates substantially from what independence would predict. The p-value answers: how likely is this large a deviation by chance if the two variables truly are independent?

Step 3 Now We Name It

Definition: Chi-Square Test for Independence

The **chi-square test for independence** tests whether two categorical variables are associated in a two-way (contingency) table.

H_0: The two variables are independent (no association).
H_a: The two variables are associated (not independent).

Expected count for each cell:

$$E_{ij} = \frac{(\text{row total}_i) \times (\text{column total}_j)}{\text{grand total}}$$

Chi-square test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed count and E_{ij} is the expected count for each cell.

Degrees of freedom: $df = (r-1)(c-1)$, where r = number of rows and c = number of columns.

Conditions:

- Random sample (or random assignment)
- All expected counts ≥ 5
- Observations are independent

Chi-Square Tests Are Always Right-Tailed

The chi-square statistic χ^2 is always non-negative (it sums squared quantities). Large values indicate poor fit between observed and expected. The p-value is always the right-tail area: $P(\chi^2 \geq \text{observed})$. There is no left-tailed or two-tailed chi-square test for independence.

Step 4 Watch It Work

Chi-square test of independence: HAI rates across three units

H_0: Unit type and HAI status are independent.
H_a: There is an association between unit type and HAI status.
$\alpha = 0.05$, $df = (3-1)(2-1) = 2$

Step 1 — Expected counts:
Overall HAI rate: $72/300 = 0.24$; no-HAI rate: 0.76. Each unit has 100 patients, so:

$$E(\text{HAI}) = 100 \times 0.24 = 24 \qquad E(\text{No HAI}) = 100 \times 0.76 = 76$$

All expected counts $= 24$ or $76 \geq 5$. ✓

Step 2 — Chi-square statistic:

Unit	Category	Observed (O)	Expected (E)	$(O-E)^2/E$
Medical	HAI	18	24	$36/24 = 1.500$
Medical	No HAI	82	76	$36/76 = 0.474$
Surgical	HAI	24	24	$0/24 = 0.000$
Surgical	No HAI	76	76	$0/76 = 0.000$
ICU	HAI	30	24	$36/24 = 1.500$
ICU	No HAI	70	76	$36/76 = 0.474$
Chi-square statistic				$\chi^2 = 3.947$

Step 3 — P-value:
Using Desmos: `1 - chisquaredist(2).cdf(3.947)` ≈ 0.139
Decision: $p = 0.139 \geq 0.05 \quad \Rightarrow \quad$ **Fail to reject H_0.**
Conclusion: There is not sufficient evidence at $\alpha = 0.05$ to conclude that HAI rates differ significantly across the three units. The observed differences (18%, 24%, 30%) could plausibly arise from random variation if the true rates were equal.

Note: This does not mean the ICU's higher rate is unimportant clinically. With a larger sample or over more months, the same pattern might reach significance. The chi-square test tells us about statistical evidence, not clinical priority.

Try This in Desmos

`1 - chisquaredist(2).cdf(3.947)` ≈ 0.139

Or compute directly: `chisquaredist(2).cdf(3.947)` gives $P(\chi^2 \leq 3.947)$; subtract from 1.

Try This in Excel

`=CHISQ.TEST(observed_range, expected_range)` → p-value directly

`=CHISQ.DIST.RT(3.947, 2)` → right-tail p-value ≈ 0.139

Step 5 Your Turn

1. A quality improvement nurse surveys 240 patients about their satisfaction with pain management and records their age group:

	Satisfied	**Not Satisfied**	**Total**
Under 50	48	32	80
50–70	60	40	100
Over 70	42	18	60
Total	150	90	240

(a) State H_0 and H_a.
(b) Compute all six expected counts.
(c) Check the conditions for the chi-square test.
(d) Compute χ^2 and the degrees of freedom.
(e) Find the p-value using Desmos or Excel and state your conclusion at $\alpha = 0.05$.

2. A hospital records the day-of-week and whether patients were readmitted within 30 days. Of 350 patients: 70 admitted Monday–Wednesday with 14 readmissions; 140 admitted Thursday–Friday with 21 readmissions; 140 admitted on weekends with 35 readmissions.

(a) Construct the two-way table.
(b) State H_0 and H_a.
(c) Compute expected counts and χ^2 ($df = 2$).
(d) Find the p-value and state your conclusion at $\alpha = 0.05$.
(e) If there is a significant association, what clinical or operational factors might explain it?

Step 6 Think Like a Nurse

What the Chi-Square Test Cannot Tell You

The chi-square test is a test of *independence* — it tells you whether two categorical variables are associated. It does not tell you:

- **Which groups differ.** If the test is significant with three or more groups, you know at least one pair differs, but not which one. Post-hoc comparison methods (beyond this course) are needed to identify specific group differences.
- **The direction of the association.** A chi-square test of the 2×3 table above rejects H_0, but the χ^2 statistic does not tell you whether higher-acuity patients are more likely to be dissatisfied or less likely.
- **The strength of the association.** A chi-square statistic of 12 vs. 3 tells you little about effect size without context. Standardized measures such as Cramér's V quantify association strength independently of sample size.
- **Causation.** Even a highly significant chi-square result does not establish that one variable causes the other. A large sample with confounding variables can produce significant associations that disappear after adjustment.

In the HAI example, the test did not reach significance with $n = 300$. But clinical judgment tells you that ICU patients are sicker and more invasively managed — the 30% HAI rate may reflect case mix rather than care quality. Statistical non-significance does not make the 12-percentage-point spread between medical and ICU rates clinically unimportant.

The chi-square test opens a conversation. The clinician's job is to ask the follow-up questions that statistics alone cannot answer.

13.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 13. Problems marked with ⋆ are more challenging.

Part A: Two-Sample and Paired Tests

1. Two ICUs are compared on mean ventilator days per patient. ICU A ($n_1 = 40$): $\bar{x}_1 = 6.2$ days, $s_1 = 2.8$ days. ICU B ($n_2 = 38$): $\bar{x}_2 = 5.1$ days, $s_2 = 2.3$ days. Test whether mean ventilator days differ (two-tailed, $\alpha = 0.05$).

(a) State H_0 and H_a.
(b) Compute the test statistic using $df = \min(39, 37) = 37$.
(c) Find the p-value and state your conclusion.
(d) Construct a 95% CI for $\mu_1 - \mu_2$ and verify consistency with the test.

2. A nurse researcher measures hemoglobin levels in 10 patients before and after a 4-week iron supplementation program:

Patient	1	2	3	4	5	6	7	8	9	10
Before	9.8	10.2	8.9	11.1	9.4	10.7	8.6	11.3	9.0	10.5
After	11.2	11.8	10.4	12.3	11.0	12.1	10.1	12.9	10.8	12.0

(a) Compute d_i = After − Before for each patient.
(b) Find $\bar{d}$ and s_d.
(c) Test whether iron supplementation increased mean hemoglobin (right-tailed, $\alpha = 0.05$, $df = 9$).
(d) State your conclusion clinically.
(e) Why would a paired test be preferable to an independent samples test here?

3. Two nursing units are compared on the proportion of patients who experience a fall. Unit 1: 12 falls in 200 patients. Unit 2: 8 falls in 180 patients.

(a) State H_0 and H_a (two-tailed).
(b) Compute $\hat{p}_c$ and the z-statistic.
(c) Find the p-value and state your conclusion at $\alpha = 0.05$.
(d) Is this result practically meaningful? What additional information would help?

Part B: Chi-Square Tests

4. A hospital surveys 400 patients about whether they received discharge instructions in a language they fully understood. Results by patient education level:

	Understood	**Did Not Understand**	**Total**
Less than high school	42	38	80
High school diploma	88	32	120
Some college or more	158	42	200
Total	288	112	400

(a) State H_0 and H_a.
(b) Compute all expected counts.
(c) Compute χ^2 and find the p-value ($df = 2$).
(d) State your conclusion at $\alpha = 0.05$ and interpret the clinical implications.
(e) If significant, which cell contributes most to the chi-square statistic?

5. ⋆ A clinical study examines the relationship between nurse shift length (8-hour vs. 12-hour) and patient safety events (medication error, fall, or pressure injury):

	Medication Error	**Fall**	**Pressure Injury**
8-hour shift	24	18	12
12-hour shift	36	30	18

(a) Add row and column totals. State H_0 and H_a.
(b) Compute expected counts and verify all are ≥ 5.
(c) Compute χ^2 ($df = 2$) and find the p-value.
(d) State your conclusion. Are any specific safety events more associated with one shift length?
(e) The total number of events on 12-hour shifts is higher. Does the chi-square test tell you this? What does it actually test?

6. ⋆ **Research Application:** A multicenter study examines whether the type of vascular access (peripheral IV, PICC line, central line) is associated with infection type (local vs. systemic) in 360 patients.

	Local Infection	**Systemic Infection**	**Total**
Peripheral IV	36	24	60
PICC line	42	78	120
Central line	42	138	180
Total	120	240	360

(a) State H_0 and H_a. Compute all expected counts.
(b) Compute χ^2 ($df = 2$).
(c) Find the p-value and state your conclusion at $\alpha = 0.01$.
(d) Identify which cell(s) deviate most from expected values. What does this tell you clinically?
(e) Can you conclude from this test that central lines *cause* more systemic

infections? What type of study would be needed?

Answer Key — Selected Problems

Answer Key

Section 13.1 Practice — Your Turn

1. (b) $\hat{p}_1 = 14/200 = 0.070$; $\hat{p}_2 = 7/180 = 0.039$; $\hat{p}_c = (14+7)/(200+180) = 21/380 = 0.05526$. (c) $n_1\hat{p}_c = 200(0.05526) = 11.1 \geq 10$ ✓; $n_2\hat{p}_c = 180(0.05526) = 9.95 \approx 10$ — borderline but acceptable. (d) $SE = \sqrt{0.05526(0.94474)(1/200+1/180)} = \sqrt{0.05222 \times 0.01056} \approx \sqrt{0.000552} \approx 0.02349$; $z = (0.070-0.039)/0.02349 \approx 1.320$. (e) Two-tailed: $p = 2(1 - P(Z \leq 1.320)) \approx 2(0.0934) = 0.187$; fail to reject H_0. Insufficient evidence that medication error rates differ between the units.

2. Differences (Before − After): 12, 10, 10, 6, 12, 11, 13, 7. (c) $\bar{d} = 81/8 = 10.125$; $s_d \approx 2.475$. (d) $t = 10.125/(2.475/\sqrt{8}) = 10.125/0.875 \approx 11.57$; $df = 7$; right-tailed: $p \approx 0.000003$; reject H_0. Highly significant evidence that the intervention reduced anxiety scores.

Section 13.2 Practice — Your Turn

1. Expected counts: Under 50: $E(\text{sat}) = 50$, $E(\text{not}) = 30$; 50–70: $E(\text{sat}) = 62.5$, $E(\text{not}) = 37.5$; Over 70: $E(\text{sat}) = 37.5$, $E(\text{not}) = 22.5$. All ≥ 5 ✓. $\chi^2 = (48-50)^2/50 + (32-30)^2/30 + (60-62.5)^2/62.5 + (40-37.5)^2/37.5 + (42-37.5)^2/37.5 + (18-22.5)^2/22.5 = 0.080+0.133+0.100+0.167+0.540+0.900 = 1.920$; $df = 2$; $p \approx 0.383$; fail to reject H_0. No significant association between age group and pain management satisfaction.

2. Table: Mon–Wed: 14 readmitted, 56 not (total 70); Thu–Fri: 21 readmitted, 119 not (total 140); Weekend: 35 readmitted, 105 not (total 140); Total: 70 readmitted, 280 not, $n = 350$. Overall readmission rate: $70/350 = 0.20$. Expected readmissions: $70 \times 70/350 = 14$, $140 \times 70/350 = 28$, $140 \times 70/350 = 28$; expected non-readmissions: 56, 112, 112. $\chi^2 = 0/14 + 0/56 + (21-28)^2/28 + (119-112)^2/112 + (35-28)^2/28 + (105-112)^2/112 = 0+0+1.750+0.438+1.750+0.438 = 4.375$; $df = 2$; $p \approx 0.112$; fail to reject H_0. No significant association between day-of-week admission and readmission.

Chapter Practice — Selected

Problem 1b: $SE = \sqrt{2.8^2/40 + 2.3^2/38} = \sqrt{0.196+0.139} = \sqrt{0.335} \approx 0.579$; $t = (6.2-5.1)/0.579 \approx 1.899$; $df = 37$.

Problem 1c: Two-tailed: `2*(1-tdist(37).cdf(1.899))` ≈ 0.065; $p \geq 0.05$; fail to reject H_0 at $\alpha = 0.05$.

Problem 2b: Differences (After − Before): 1.4, 1.6, 1.5, 1.2, 1.6, 1.4, 1.5, 1.6, 1.8, 1.5. $\bar{d} = 15.1/10 = 1.51$; $s_d \approx 0.160$.

Problem 2c: $t = 1.51/(0.160/\sqrt{10}) = 1.51/0.0506 \approx 29.84$; $df = 9$; right-tailed $p \approx 0.000$; reject H_0. Highly significant evidence that iron supplementation increased hemoglobin.

Problem 4c: Expected counts: Less than HS: $E(\text{und}) = 57.6$, $E(\text{not}) = 22.4$; HS: $E(\text{und}) = 86.4$, $E(\text{not}) = 33.6$; College+: $E(\text{und}) = 144$, $E(\text{not}) = 56$. $\chi^2 = (42 - 57.6)^2/57.6 + (38-22.4)^2/22.4 + (88-86.4)^2/86.4 + (32-33.6)^2/33.6 + (158-144)^2/144 + (42-56)^2/56 \approx 4.224 + 10.857 + 0.030 + 0.076 + 1.361 + 3.500 = 20.05$; $df = 2$; $p < 0.001$; reject H_0. Significant association between education level and comprehension of discharge instructions.
Problem 6b (Challenge): Totals: 8-hr row = 54, 12-hr row = 84; Med Error col = 60, Fall col = 48, Pressure Injury col = 30; Grand total = 138. Expected: 8-hr/Med Error = $54 \times 60/138 = 23.48$; 8-hr/Fall = $54 \times 48/138 = 18.78$; 8-hr/PI = $54 \times 30/138 = 11.74$; 12-hr cells = $36.52, 29.22, 18.26$. $\chi^2 \approx 0.012+0.033+0.006+0.007+0.021+0.004 = 0.082$; $df = 2$; $p \approx 0.960$; fail to reject H_0. No significant association between shift length and safety event type — the distribution of event types is similar across shift lengths.

Chapter 13 Summary

Section 13.1 — Two-Sample Tests

- **Two-sample t-test:** $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$; $df = \min(n_1 - 1, n_2 - 1)$; for independent groups.
- **Paired t-test:** Compute differences d_i; $t = \bar{d}/(s_d/\sqrt{n})$; $df = n - 1$; for matched or before-after data.
- Use paired design when possible — it eliminates between-subject variation and increases power.
- **Two-proportion z-test:** $z = (\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{p}_c(1 - \hat{p}_c)(1/n_1 + 1/n_2)}$; uses pooled $\hat{p}_c$.

Section 13.2 — Chi-Square Test for Independence

- Tests whether two categorical variables are associated in a contingency table.
- H_0: independence; H_a: association.
- Expected counts: $E_{ij} = (\text{row total} \times \text{column total})/n$.
- $\chi^2 = \sum (O - E)^2/E$; $df = (r - 1)(c - 1)$; always right-tailed.
- Condition: all expected counts ≥ 5.
- A significant chi-square test identifies an association but does not specify which cells drive it, the direction, or the cause.

The Nursing Connection

- Most clinical comparisons involve two groups: new vs. old treatment, day vs. night shift, high-risk vs. low-risk patients.
- Before-after designs on the same patients are inherently paired — use the paired t-test to gain power.
- Chi-square tests are ideal for audits and surveys with categorical outcomes: satisfaction yes/no, complication type, unit category.
- Statistical significance does not resolve causal questions. Case mix, confounding, and study design all require clinical interpretation alongside any p-value.

CHAPTER 14

Comparing Groups: ANOVA

> *"When you have more than two groups and one hypothesis test, ANOVA does the job of many without inflating your chances of being wrong."*
>
> — why we need a new method when comparing three or more means

In this chapter, you will learn to:

- Explain why multiple t-tests inflate the Type I error rate
- State the hypotheses for a one-way ANOVA
- Identify the between-group and within-group sources of variation
- Compute and interpret the ANOVA F-statistic
- Use Desmos and Excel to perform one-way ANOVA
- Interpret ANOVA results in clinical and research contexts
- Understand the role of post-hoc tests when ANOVA is significant

14.1 The Logic of ANOVA

Step 1 Read This First

A hospital quality team is comparing patient satisfaction scores across four nursing units: Medical, Surgical, ICU, and Oncology. They want to know: do patients on different units rate their care differently?

A random sample from each unit yields:

	Medical	Surgical	ICU	Oncology
n	15	15	15	15
$\bar{x}$	78.4	82.1	71.6	85.3
s	9.2	8.7	10.1	7.8

There are four groups. To compare all possible pairs using two-sample t-tests, you would need $\binom{4}{2} = 6$ separate tests. But running six tests at $\alpha = 0.05$ each means the probability of making *at least one* Type I error across all tests is no longer 5% — it compounds to nearly 26%.

Analysis of Variance (ANOVA) solves this by testing all groups simultaneously with a single test, keeping the overall Type I error rate at exactly α.

Step 2 Let's Talk About It

Think about where the variation in the data comes from.

Some of the variation in satisfaction scores is *between groups*: the ICU mean (71.6) is noticeably lower than the Oncology mean (85.3). If all units truly had the same population mean, we would not expect to see differences this large.

But some variation is *within groups*: even within a single unit, individual patients rate their care differently. Some Medical unit patients score 65, others score 90. This within-group variation is just natural individual differences — noise.

ANOVA asks: is the between-group variation large *relative to* the within-group variation? If the groups truly have the same mean, then the between-group variability should be roughly the same magnitude as the within-group variability. If the between-group variation is much larger than the within-group variation, that is evidence the group means are genuinely different.

The ratio of these two sources of variation is the F-statistic.

Step 3 Now We Name It

Definition: One-Way ANOVA Hypotheses

One-way ANOVA tests whether the means of $k \geq 3$ groups are all equal.

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ (all group means are equal)
H_a: At least one μ_i differs from the others

Note: H_a does *not* say all means differ — only that at least one is different.

Definition: Sources of Variation in ANOVA

The total variability in the data is partitioned into two components:

Between-group variation (treatment) — $SS_{Between}$: measures how much the group means differ from the overall (grand) mean.

$$SS_{Between} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{grand})^2$$

Within-group variation (error) — SS_{Within}: measures how much individual observations vary around their own group mean.

$$SS_{Within} = \sum_{i=1}^{k} (n_i - 1)s_i^2$$

Mean Squares are the sums of squares divided by their degrees of freedom:

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}} \qquad df_{Between} = k - 1$$

$$MS_{Within} = \frac{SS_{Within}}{df_{Within}} \qquad df_{Within} = N - k$$

where $N = \sum n_i$ is the total sample size across all groups.

Definition: The F-Statistic

The ANOVA test statistic is the F-ratio:

$$F = \frac{MS_{Between}}{MS_{Within}}$$

- If H_0 is true: $F \approx 1$ (between-group variation $\approx$ within-group variation)
- If H_0 is false: $F > 1$ (between-group variation exceeds within-group variation)
- The p-value is always the **right-tail** area of the F-distribution with $df_1 = k - 1$ and $df_2 = N - k$

Conditions for ANOVA:

- Independent random samples from each group
- Approximately normal distributions within each group (or $n_i \geq 30$ for each group)
- Equal (or roughly equal) variances across groups — a good rule of thumb is that the largest s should be no more than twice the smallest s

Step 4 Watch It Work

One-way ANOVA: patient satisfaction across four nursing units

Data: $k = 4$ groups, $n_i = 15$ each, $N = 60$

	Medical	Surgical	ICU	Oncology
$\bar{x}_i$	78.4	82.1	71.6	85.3
s_i	9.2	8.7	10.1	7.8

Step 1 — Grand mean:

$$\bar{x}_{grand} = \frac{15(78.4) + 15(82.1) + 15(71.6) + 15(85.3)}{60} = \frac{4{,}761}{60} = 79.35$$

Step 2 — Between-group sum of squares ($df_{Between} = 3$):
Deviations from the grand mean: $(78.4 - 79.35)^2 = (-0.95)^2 = 0.9025$
$(82.1 - 79.35)^2 = (2.75)^2 = 7.5625$
$(71.6 - 79.35)^2 = (-7.75)^2 = 60.0625$
$(85.3 - 79.35)^2 = (5.95)^2 = 35.4025$

$$SS_{Between} = 15(0.9025 + 7.5625 + 60.0625 + 35.4025) = 15(103.93) = 1{,}558.95$$

$$MS_{Between} = \frac{1{,}558.95}{3} = 519.65$$

Step 3 — Within-group sum of squares ($df_{Within} = 56$):

$$SS_{Within} = (15 - 1)(9.2^2 + 8.7^2 + 10.1^2 + 7.8^2) = 14(84.64 + 75.69 + 102.01 + 60.84)$$

$$= 14(323.18) = 4{,}524.52$$

$$MS_{Within} = \frac{4{,}524.52}{56} \approx 80.795$$

Step 4 — F-statistic:

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{519.65}{80.795} \approx 6.43$$

Step 5 — P-value:
Using Desmos: `1 - fdist(3, 56).cdf(6.43)` ≈ 0.0007
Decision: $p = 0.0007 < 0.05 \quad \Rightarrow$ **Reject H_0.**
Conclusion: There is significant evidence ($p = 0.0007$) that mean patient satisfaction scores differ across the four nursing units. At least one unit has a different true mean than the others.

Source	SS	df	MS	F	P-value
Between groups	1,558.95	3	519.65	6.43	0.0007
Within groups	4,524.52	56	80.80		
Total	6,083.47	59			

Try This in Desmos

P-value from F-distribution: `1 - fdist(3, 56).cdf(6.43)` ≈ 0.0007

Note: `fdist(`df_1`, `df_2`)` requires numerator df first, denominator df second.

Try This in Excel

Method 1 (built-in): Data → Data Analysis → ANOVA: Single Factor

Method 2 (formula): `=F.DIST.RT(6.43, 3, 56)` $\approx$ 0.0007 (right-tail p-value)

Step 5 Your Turn

1. A hospital compares mean LOS (days) for patients with three primary diagnoses: cardiac ($n_1 = 20$, $\bar{x}_1 = 5.8$, $s_1 = 2.1$), respiratory ($n_2 = 20$, $\bar{x}_2 = 4.2$, $s_2 = 1.8$), and orthopedic ($n_3 = 20$, $\bar{x}_3 = 6.9$, $s_3 = 2.4$).
   (a) State H_0 and H_a.
   (b) Compute the grand mean.
   (c) Compute $SS_{Between}$ and $MS_{Between}$.
   (d) Compute SS_{Within} and MS_{Within}.
   (e) Compute the F-statistic. Find the p-value ($df_1 = 2$, $df_2 = 57$) using Desmos.
   (f) State your conclusion at $\alpha = 0.05$.

2. A nurse manager notices that nursing documentation time (minutes) appears to differ by shift. A random sample from each shift:

Shift	n	$\bar{x}$ (min)	s (min)
Day	12	22.4	4.1
Evening	12	18.9	5.2
Night	12	25.1	4.8

   (a) Check the equal variance condition. Is it satisfied?
   (b) State H_0 and H_a.
   (c) Using Desmos or Excel, perform the one-way ANOVA and report the F-statistic and p-value.

(d) State your conclusion at $\alpha = 0.05$.

(e) If the result is significant, what follow-up analysis would identify *which* shifts differ?

Step 6 Think Like a Nurse

ANOVA Tells You Something Differs — But Not What or By How Much
The ANOVA for patient satisfaction across four units returned $F = 6.43$ and $p = 0.0007$. This is compelling evidence that at least one unit mean differs from the others. But which unit? By how much? Is the difference clinically meaningful?

ANOVA answers none of these questions. It is an omnibus test — it detects that something is different somewhere but does not identify where.

To find out *which* pairs of means differ, you need **post-hoc tests**. These are pairwise comparisons (like t-tests) that are corrected for the fact that multiple comparisons are being made. Common post-hoc procedures include:

- **Tukey's HSD** (Honest Significant Difference): compares all pairs while controlling the family-wise error rate. The most commonly used in nursing research.
- **Bonferroni correction**: divides α by the number of comparisons. Simple but conservative.

In our example, a post-hoc analysis would likely show that the ICU (mean 71.6) and Oncology (mean 85.3) differ significantly from each other, and possibly from the Medical and Surgical units. The 13.7-point difference between ICU and Oncology is clinically substantial on a 100-point scale.

The ANOVA is the gatekeeper: if F is not significant, there is no statistical justification for any pairwise comparisons. If F is significant, post-hoc tests identify which specific pairs differ.

A significant ANOVA is the beginning of the analysis, not the end. The clinical question is always: which groups differ, and does that difference matter for patient care?

14.2 Interpreting ANOVA in Clinical Research

Step 1 Read This First

A nurse researcher reads the following in a published clinical study:

"We compared mean serum albumin levels across four patient groups: well-nourished, mildly malnourished, moderately malnourished, and severely malnourished. A one-way ANOVA revealed significant differences among the groups ($F(3, 116) = 42.7$, $p < 0.001$). Post-hoc comparisons (Tukey's HSD) showed that all pairwise differences were statistically significant except between the well-nourished and mildly malnourished groups."

The nurse needs to read this result critically. She needs to understand what $F(3, 116) = 42.7$ means, why $p < 0.001$ is reported the way it is, and what the post-hoc result tells her about clinical practice.

Reading ANOVA output from published research is a core skill for evidence-based practice. This section teaches you how.

Step 2 Let's Talk About It

The notation $F(3, 116) = 42.7$ is shorthand. The numbers in parentheses are the two degrees of freedom: $df_{Between} = 3$ (four groups minus one) and $df_{Within} = 116$ (total observations minus number of groups). The F-statistic is 42.7.

A very large F-statistic (42.7 is enormous compared to the critical value around 2.7) means the between-group variation is about 43 times larger than the within-group variation. This is not plausible if all four groups have the same true mean — the result is highly significant.

The post-hoc result is crucial: not all groups are different from each other. Well-nourished and mildly malnourished patients have similar albumin levels. The other pairs — well-nourished vs. moderately and severely malnourished, mild vs. moderate and severe, moderate vs. severe — are all significantly different. This suggests clinically meaningful albumin differences begin at moderate malnutrition.

Step 3 Now We Name It

Definition: Reading ANOVA Output from Published Research

Standard ANOVA notation in published studies:

$$F(df_{Between},\ df_{Within}) = F\text{-statistic}, \quad p = \text{p-value}$$

Key quantities to identify:

- $df_{Between} = k - 1$: number of groups minus 1
- $df_{Within} = N - k$: total N minus number of groups
- F-statistic: the ratio $MS_{Between}/MS_{Within}$
- P-value: right-tail area of the F-distribution
- Post-hoc results: which specific pairs of means differ

Definition: ANOVA Table Structure

Source	**SS**	**df**	**MS**	F	**P-value**
Between groups	SS_B	$k-1$	$SS_B/(k-1)$	MS_B/MS_W	Right-tail
Within groups	SS_W	$N-k$	$SS_W/(N-k)$		
Total	SS_T	$N-1$			

Note: $SS_{Total} = SS_{Between} + SS_{Within}$ and $df_{Total} = df_{Between} + df_{Within}$.

Effect Size in ANOVA: η^2

As with all hypothesis tests, a significant ANOVA does not automatically imply a practically meaningful difference. The effect size statistic η^2 (eta-squared) measures the proportion of total variability explained by group membership:

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

Interpretation: $\eta^2 = 0.06$ means 6% of variance in the outcome is explained by group. Rough guidelines: $0.01 =$ small, $0.06 =$ medium, $0.14 =$ large.
A large F with a small η^2 in a very large sample should prompt caution about clinical meaningfulness.

Step 4 Watch It Work

Reading and extending published ANOVA output

Suppose the albumin study reports this partial ANOVA table:

Source	**SS**	**df**	**MS**	F	**P-value**
Between groups	72.44	3	24.147	42.7	< 0.001
Within groups	65.60	116	0.566		
Total	138.04	119			

What we can read directly:

- $k = df_{Between} + 1 = 3 + 1 = 4$ groups ✓
- $N = df_{Total} + 1 = 119 + 1 = 120$ total patients
- $n = 120/4 = 30$ patients per group (if balanced)
- $F = MS_{Between}/MS_{Within} = 24.147/0.566 = 42.7$ ✓
- $SS_B + SS_W = 72.44 + 65.60 = 138.04 = SS_T$ ✓

Effect size:

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}} = \frac{72.44}{138.04} \approx 0.525$$

52.5% of the variance in albumin levels is explained by malnutrition status — a very large effect. This is clinically meaningful, not just statistically significant.

Interpreting the post-hoc result:
The Tukey HSD found all pairwise comparisons significant *except* well-nourished vs. mildly malnourished. This means:

Well-nourished	Mildly malnourished	Moderately malnourished
Not significantly different ($p > 0.05$)		
All other pairs are significantly different ($p < 0.05$)		

Clinical takeaway: Nutritional interventions should be targeted at patients who are moderately or severely malnourished, where albumin levels are measurably and significantly lower. Mildly malnourished patients resemble well-nourished patients statistically.

Try This in Desmos

Verify the p-value: `1 - fdist(3, 116).cdf(42.7)` ≈ 0.000 (essentially zero)

Check F-statistic: `24.147 / 0.566` $= 42.66 \approx 42.7$ ✓

Try This in Excel

`=F.DIST.RT(42.7, 3, 116)` ≈ 0 (p-value essentially zero)

η^2: `=72.44/138.04` ≈ 0.525

Step 5 Your Turn

1. A published study reports ANOVA results for pain scores after three types of analgesic treatment (Drug A, Drug B, and Placebo): $F(2, 87) = 8.43$, $p = 0.0004$.
  - (a) How many groups are there? How many total patients?
  - (b) What are $df_{Between}$ and df_{Within}?
  - (c) Is the result statistically significant at $\alpha = 0.05$? At $\alpha = 0.01$?
  - (d) Can you conclude from the F-statistic alone that Drug A and Drug B have different effects? Explain.
  - (e) If $SS_{Between} = 112.4$ and $SS_{Total} = 641.8$, compute η^2 and interpret the effect size.

2. A hospital tracks response time (minutes) from nurse call to bedside across five units. An ANOVA yields the following partial table:

Source	**SS**	**df**	**MS**	F	**P-value**
Between groups	184.2	?	?	?	?
Within groups	892.5	95	?		
Total	?	?			

  - (a) Fill in all missing values in the table.
  - (b) How many units are in this study? How many total observations?
  - (c) Compute the F-statistic and find the p-value using Desmos or Excel.
  - (d) State your conclusion at $\alpha = 0.05$.
  - (e) Compute η^2 and interpret.

Step 6 Think Like a Nurse

ANOVA Assumptions Matter More Than You Might Think

ANOVA rests on three assumptions: independent samples, approximate normality within each group, and equal variances across groups. Violations of these assumptions can produce misleading results — a non-significant ANOVA that misses a real effect, or a significant result that is an artifact of unequal variances.

Checking equal variances in practice:

The rule of thumb "largest s should be no more than twice the smallest s" is quick and practical. In the patient satisfaction example, the standard deviations ranged from 7.8 to 10.1 — a ratio of $10.1/7.8 = 1.30$, well within the acceptable range.

When variances differ substantially, Welch's ANOVA (which does not assume equal variances) is preferred. Most statistical software will compute it alongside the standard ANOVA.

Non-normality: ANOVA is fairly robust to moderate non-normality when sample sizes are large. The central limit theorem provides some protection. For small samples from clearly non-normal populations, the Kruskal–Wallis test (a non-parametric alternative) may be more appropriate.

Independence: The most critical assumption. If measurements within a group are correlated — for example, patients from the same physician tend to have similar outcomes — the ANOVA F-statistic is invalid. In nursing research, clustering by unit or provider is common and must be addressed in the study design.

Statistical methods come with terms and conditions. Reading those conditions is not optional — it is part of interpreting any result responsibly.

14.3 Chapter Practice Problems

Step 5 Your Turn

The following problems cover all topics in Chapter 14. Problems marked with ⋆ are more challenging.

Part A: Computing One-Way ANOVA

1. A researcher compares systolic blood pressure (mmHg) in three diet groups after 8 weeks: Low-sodium ($n = 25$, $\bar{x} = 128.4$, $s = 12.1$), Mediterranean ($n = 25$, $\bar{x} = 122.7$, $s = 10.8$), and Control ($n = 25$, $\bar{x} = 135.2$, $s = 13.4$).

(a) State H_0 and H_a.
(b) Check the equal variance condition.
(c) Compute the grand mean.
(d) Compute $SS_{Between}$, $MS_{Between}$, SS_{Within}, MS_{Within}.
(e) Compute F and find the p-value ($df_1 = 2$, $df_2 = 72$).
(f) State your conclusion at $\alpha = 0.05$ and compute η^2.

2. A hospital quality team measures patient fall rates (per 1,000 patient-days) on six units over one quarter. Using Desmos or Excel with the following data, perform a one-way ANOVA:

Unit	Med-Surg A	Med-Surg B	ICU	Oncology	Ortho	Neuro
n	10	10	10	10	10	10
$\bar{x}$	2.8	3.1	1.4	2.2	3.8	2.6
s	0.9	1.1	0.6	0.8	1.2	0.9

(a) State H_0 and H_a.
(b) Check the equal variance condition.
(c) Compute F and find the p-value ($df_1 = 5$, $df_2 = 54$).
(d) State your conclusion at $\alpha = 0.05$.
(e) Compute η^2 and interpret the effect size.

Part B: Reading ANOVA Output

3. A clinical study compares nurse burnout scores across four hospital types: Community ($n = 30$), Academic ($n = 30$), Safety-net ($n = 30$), and Veterans ($n = 30$). The ANOVA table is:

Source	**SS**	**df**	**MS**	F	**P-value**
Between groups	1,248.6	3	416.2	4.21	0.007
Within groups	11,480.4	116	98.97		
Total	12,729.0	119			

(a) Verify: does $SS_{Between} + SS_{Within} = SS_{Total}$?
(b) Verify the F-statistic using MS values.

(c) State H_0 and H_a and give the conclusion at $\alpha = 0.05$.
(d) Compute η^2 and interpret the effect size.
(e) The researchers report Tukey post-hoc results showing that Safety-net hospitals have significantly higher burnout than all other types. What does this mean clinically?

4. ⋆ A published study reports: "Patients were randomized to one of three nursing intervention frequencies (once, twice, or three times daily). Mean pain scores at 72 hours were compared using one-way ANOVA: $F(2, 147) = 11.24$, $p < 0.001$, $\eta^2 = 0.13$."

(a) How many total patients were in the study? How many per group (if balanced)?
(b) Interpret $\eta^2 = 0.13$. Is this a large, medium, or small effect?
(c) Reconstruct the partial ANOVA table (SS, df, MS, F). Use $SS_{Total} = 1{,}420$.
(d) Post-hoc tests show that twice-daily and three-times-daily interventions both differ significantly from once-daily, but not from each other. Draw a diagram showing which groups differ.
(e) What would you recommend to hospital administration based on these results? Consider both statistical evidence and operational feasibility.

5. ⋆ **Research Application:** A multicenter nursing study compares time to ambulation (hours post-surgery) across five surgical protocols. The ANOVA yields $F(4, 245) = 6.84$, $p < 0.001$. Post-hoc Tukey tests reveal:

- Protocols A and B are not significantly different from each other.
- Protocols C and D are not significantly different from each other.
- Protocol E differs significantly from all others.
- All other pairs are significantly different.

(a) How many total patients? What is N?
(b) Sketch a diagram grouping protocols that are NOT significantly different.
(c) If Protocol E has the lowest mean ambulation time, what are the clinical implications?
(d) The study reports $\eta^2 = 0.10$. Is this effect clinically meaningful? How would you decide?
(e) What additional analyses (beyond ANOVA and post-hoc tests) would strengthen the clinical recommendation?

Answer Key — Selected Problems

Answer Key

Section 14.1 Practice — Your Turn

1. (b) $\bar{x}_{grand} = (20 \times 5.8 + 20 \times 4.2 + 20 \times 6.9)/60 = (116 + 84 + 138)/60 = 338/60 \approx 5.633$. (c) $SS_{Between} = 20(5.8 - 5.633)^2 + 20(4.2 - 5.633)^2 + 20(6.9 - 5.633)^2 = 20(0.028) + 20(2.054) + 20(1.604) = 0.556 + 41.089 + 32.089 = 73.733$; $MS_{Between} = 73.733/2 = 36.867$. (d) $SS_{Within} = 19(2.1^2 + 1.8^2 + 2.4^2) = 19(4.41 + 3.24 + 5.76) = 19(13.41) = 254.79$; $MS_{Within} = 254.79/57 = 4.470$. (e) $F = 36.867/4.470 \approx 8.25$; `1 - fdist(2,57).cdf(8.25)` ≈ 0.0007. (f) $p = 0.0007 < 0.05$; reject H_0. Significant evidence that mean LOS differs by diagnosis group.

2. (a) Variance check: s values are 4.1, 5.2, 4.8; largest/smallest $= 5.2/4.1 = 1.27 < 2$ ✓. (b) $H_0 : \mu_{day} = \mu_{eve} = \mu_{night}$; H_a: at least one shift mean differs. (e) Post-hoc tests (e.g., Tukey HSD) comparing all pairs: Day vs. Evening, Day vs. Night, Evening vs. Night.

Section 14.2 Practice — Your Turn

1. (a) 3 groups; $N = df_{Within} + k = 87 + 3 = 90$ total patients. (b) $df_{Between} = 2$; $df_{Within} = 87$. (c) Yes, significant at both $\alpha = 0.05$ and $\alpha = 0.01$ (p = 0.0004). (d) No — ANOVA only tells you that *at least one* mean differs; post-hoc tests are needed to determine which specific pairs differ. (e) $\eta^2 = 112.4/641.8 \approx 0.175$ — large effect.

2. (a) $df_{Between} = 4$ (5 units); $MS_{Between} = 184.2/4 = 46.05$; $MS_{Within} = 892.5/95 = 9.395$; $F = 46.05/9.395 \approx 4.90$; $SS_{Total} = 1{,}076.7$; $df_{Total} = 99$. (b) $k = 5$ units; $N = df_{Total} + 1 = 100$ observations. (c) `1 - fdist(4,95).cdf(4.90)` ≈ 0.0013. (d) $p = 0.001 < 0.05$; reject H_0. (e) $\eta^2 = 184.2/1{,}076.7 \approx 0.171$ — large effect; unit membership accounts for about 17% of variation in response time.

Chapter Practice — Selected

Problem 1c: Grand mean $= (25 \times 128.4 + 25 \times 122.7 + 25 \times 135.2)/75 = (3{,}210 + 3{,}067.5 + 3{,}380)/75 = 9{,}657.5/75 = 128.767$.

Problem 1d: $SS_{Between} = 25(128.4 - 128.767)^2 + 25(122.7 - 128.767)^2 + 25(135.2 - 128.767)^2$

$= 25(0.135) + 25(36.806) + 25(41.387) = 3.375 + 920.156 + 1{,}034.675 = 1{,}958.2$.

$MS_{Between} = 1{,}958.2/2 = 979.1$.

$SS_{Within} = 24(12.1^2 + 10.8^2 + 13.4^2) = 24(146.41 + 116.64 + 179.56) = 24(442.61) = 10{,}622.6$.

$MS_{Within} = 10{,}622.6/72 = 147.5$.

Problem 1e: $F = 979.1/147.5 \approx 6.64$; `1 - fdist(2,72).cdf(6.64)` ≈ 0.002; reject H_0.

Problem 1f: $\eta^2 = 1{,}958.2/(1{,}958.2 + 10{,}622.6) = 1{,}958.2/12{,}580.8 \approx 0.156$ — large

effect.
Problem 3d: $\eta^2 = 1{,}248.6/12{,}729.0 \approx 0.098$ — medium effect (nearly large).
Problem 4c: $SS_{Between} = \eta^2 \times SS_{Total} = 0.13 \times 1{,}420 = 184.6$; $MS_{Between} = 184.6/2 = 92.3$; $SS_{Within} = 1{,}420 - 184.6 = 1{,}235.4$; $MS_{Within} = 1{,}235.4/147 = 8.404$; $F = 92.3/8.404 = 10.98 \approx 11.24$ (slight rounding). $N = df_{Within} + k = 147 + 3 = 150$ patients; 50 per group if balanced.

Chapter 14 Summary

Section 14.1 — The Logic of ANOVA

- Use ANOVA (not multiple t-tests) when comparing three or more group means — multiple t-tests inflate the Type I error rate.
- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$; H_a: at least one mean differs.
- $F = MS_{Between}/MS_{Within}$; always right-tailed; $df_1 = k - 1$, $df_2 = N - k$.
- Conditions: independent samples, approximate normality, equal variances (ratio of largest to smallest s should be < 2).
- A significant F means *at least one* group mean differs; post-hoc tests identify which pairs.

Section 14.2 — Interpreting ANOVA in Research

- ANOVA notation: $F(df_{Between}, df_{Within}) = F$-value, $p =$ p-value.
- From the ANOVA table: recover k, N, MS values, and F.
- $\eta^2 = SS_{Between}/SS_{Total}$: proportion of variance explained; 0.01 = small, 0.06 = medium, 0.14 = large.
- Post-hoc tests (Tukey HSD, Bonferroni) identify specific pairs that differ while controlling the family-wise error rate.
- A significant ANOVA is the starting point; effect size and post-hoc comparisons complete the picture.

The Nursing Connection

- Quality improvement comparing more than two units, shifts, or protocols requires ANOVA, not multiple t-tests.
- A large F-statistic in a big sample may reflect a trivially small effect clinically — always report η^2.
- Post-hoc tests guide targeted action: if only Safety-net hospitals show high burnout, the intervention can be targeted.
- Check the equal variance assumption before interpreting ANOVA results from published research.

CHAPTER 15

Reading the Research

"Statistics does not live in textbooks. It lives in the studies that shape what nurses do at the bedside every day."

— connecting everything you have learned to the literature you will read

In this chapter, you will learn to:

- Identify study design from an abstract or methods section
- Locate and interpret key statistical results in published research
- Distinguish between statistical significance and clinical significance
- Evaluate the quality of evidence using a systematic framework
- Recognize common statistical errors and misleading presentations
- Apply the statistical tools from this course to critically read a full research article

15.1 Anatomy of a Research Article

Step 1 Read This First

A hospital is considering switching from a standard patient positioning protocol to a new pressure-redistribution protocol aimed at reducing pressure injuries in long-term ICU patients. The nursing director asks her team to review the evidence.

A staff nurse finds a relevant article: *"Effect of a Pressure-Redistribution Positioning Protocol on Pressure Injury Incidence in Mechanically Ventilated Adults: A Randomized Controlled Trial"* (*American Journal of Critical Care*, 2023).

She opens the article and sees: an abstract with numbers she half-recognizes, a methods section with sampling language, a results section full of p-values and confidence intervals, and a discussion that seems to claim the protocol works. She needs to know: does it actually work? How confident should she be? Should her hospital adopt this protocol?

Answering these questions requires reading the article statistically — not just reading the words.

Step 2 Let's Talk About It

Every research article has the same skeleton, regardless of the journal or topic. Once you recognize the skeleton, you know exactly where to look for each piece of statistical information.

The abstract gives you the headline: design, sample size, primary outcome, and main result. Read it first, but not last. The abstract is a marketing document — it presents the authors' preferred interpretation. Your job is to check whether the full article supports those claims.

The methods section tells you what was actually done: how patients were selected, how groups were formed, what was measured, and which statistical tests were chosen. This is where you evaluate whether the study design justifies the conclusions.

The results section contains the numbers. Every p-value, confidence interval, sample mean, and test statistic you have learned to interpret in this course will appear here. Read it carefully and literally — a table is more honest than prose.

The discussion interprets the results. Authors are allowed to be optimistic here, but they are also supposed to acknowledge limitations. A discussion that does not mention limitations is a warning sign.

Step 3 Now We Name It

Definition: The Five Questions to Ask Any Study

Before reading a single number, ask:

1. **What is the study design?** (RCT, cohort, cross-sectional, case-control, systematic review) — the design determines what causal claims are valid.
2. **Who is in the sample?** (inclusion/exclusion criteria, how patients were selected) — determines to whom the results generalize.
3. **What is the primary outcome?** (the one variable the study was powered to detect) — every other result is secondary.
4. **Was the sample size adequate?** (was there a power calculation?) — a negative result from an underpowered study does not mean the treatment has no effect.
5. **Were appropriate statistical methods used?** (does the analysis match the design and data type?) — mismatch is a common source of invalid conclusions.

Definition: Study Designs and the Hierarchy of Evidence

Not all studies are created equal. The hierarchy from strongest to weakest causal evidence:

Design	What it can establish
Systematic review / meta-analysis	Pooled evidence across multiple studies
Randomized controlled trial (RCT)	Causation (with proper blinding and allocation)
Cohort study	Association; temporal ordering of exposure and outcome
Case-control study	Association; efficient for rare outcomes
Cross-sectional survey	Prevalence; correlation (not causation)
Case report / expert opinion	Hypothesis generation only

A statistically significant result from a cross-sectional study does not justify the same clinical confidence as the same result from a well-conducted RCT.

Definition: Key Statistical Terms in Published Research

- **Incidence / prevalence:** incidence = new cases over time; prevalence = existing cases at a point in time.
- **Relative risk (RR):** $P(\text{outcome}|\text{exposed})/P(\text{outcome}|\text{unexposed})$. $\text{RR} = 1$ means no difference; $\text{RR} > 1$ means higher risk in exposed group.
- **Odds ratio (OR):** commonly used in case-control studies. $\text{OR} \approx \text{RR}$ when the outcome is rare ($< 10\%$).
- **Number needed to treat (NNT):** $1/(\text{absolute risk reduction})$. $\text{NNT} = 20$ means you must treat 20 patients to prevent 1 additional adverse outcome.
- **Confidence interval for RR or OR:** if the CI includes 1, the result is not statistically significant.

Step 4 Watch It Work

Reading a published abstract: the pressure-redistribution protocol study

Abstract (abridged):

Background: Pressure injuries affect up to 25% of mechanically ventilated ICU patients and increase mortality, cost, and LOS.
Objective: To evaluate whether a nurse-led pressure-redistribution positioning protocol reduces pressure injury incidence compared to standard care.
Methods: Randomized controlled trial. Adults mechanically ventilated ≥ 48 *hours were randomized 1:1 to the intervention* ($n = 112$) *or control* ($n = 108$)*. Primary outcome: pressure injury (Stage II or higher) within 14 days. Secondary outcomes: LOS, 28-day mortality. Analysis: chi-square for proportions, independent-samples t-test for continuous outcomes.*
Results: Pressure injury incidence was 11.6% in the intervention group vs. 24.1% in the control group ($\chi^2 = 6.84$, $p = 0.009$, *95% CI for difference:* -21.2% *to* -3.8%)*. Mean ICU LOS was 9.2 days (SD = 4.1) vs. 10.8 days (SD = 5.3;* $p = 0.018$)*. Mortality did not differ (18.8% vs. 20.4%,* $p = 0.76$)*.*
Conclusion: The pressure-redistribution protocol significantly reduced pressure injury incidence and ICU LOS without affecting mortality.

Step-by-step reading:

1. Study design: RCT. Strongest causal design. Random allocation means confounding is controlled.

2. Sample: $n = 220$ total $(112 + 108)$. Adults on mechanical ventilation ≥ 48 hours. Note: results apply to this specific population, not all ICU patients.

3. Primary outcome: Pressure injury within 14 days. The study was powered for this.

4. Verify the chi-square result for the primary outcome:
$\hat{p}_1 = 13/112 \approx 0.116$, $\hat{p}_2 = 26/108 \approx 0.241$
$\hat{p}_c = (13 + 26)/(112 + 108) = 39/220 = 0.1773$

$$z = \frac{0.116 - 0.241}{\sqrt{0.1773(0.8227)(1/112 + 1/108)}} = \frac{-0.125}{0.0515} \approx -2.43$$

Two-tailed: $p = 2 \times P(Z \leq -2.43) \approx 0.015$. The paper reports $p = 0.009$ via chi-square (a related but not identical procedure) — both indicate significance at $\alpha = 0.05$.

5. Interpret the CI for the difference in proportions:
95% CI: $(-21.2\%, -3.8\%)$. The entire interval is negative — the intervention group always had fewer injuries. The minimum plausible benefit is 3.8 percentage points, maximum is 21.2. Both are clinically meaningful in an ICU context.

6. Absolute risk reduction and NNT:

$ARR = 24.1\% - 11.6\% = 12.5$ percentage points
$NNT = 1/0.125 = 8$
Treat 8 ICU patients with the new protocol to prevent 1 pressure injury.

7. LOS result: $p = 0.018$ (significant). But does the CI contain only clinically meaningful differences? LOS reduction of 1.6 days is meaningful in an ICU.

8. Mortality result: $p = 0.76$ (not significant). The protocol does not affect mortality — neither helps nor harms.

Step 5 Your Turn

1. A study reports: "The intervention group had a 30-day readmission rate of 8.4% compared to 14.7% in the control group (RR = 0.57, 95% CI: 0.38 to 0.87, $p = 0.009$)."
   (a) What does RR = 0.57 mean in plain language?
   (b) Does the 95% CI support statistical significance? Explain.
   (c) Compute the absolute risk reduction (ARR) and NNT.
   (d) The study enrolled $n = 300$ per group. Verify the approximate number of events in each group.
   (e) A colleague says "The intervention cuts the readmission rate nearly in half — we should adopt it." What statistical and clinical questions would you ask before agreeing?

2. A cross-sectional survey of 500 nurses reports: "Nurses working 12-hour shifts had significantly higher burnout scores than those working 8-hour shifts ($t(498) = 3.12$, $p = 0.002$, Cohen's $d = 0.28$)."
   (a) What study design is this? What causal claims does it support?
   (b) Interpret Cohen's $d = 0.28$. (For t-tests: small = 0.2, medium = 0.5, large = 0.8.)
   (c) The result is statistically significant. Is it clinically meaningful?
   (d) What would you need to see to recommend a shift-length policy change based on this study?

Step 6 Think Like a Nurse

Three Questions Every Evidence-Based Nurse Should Ask

When you read a published study, three questions should run in parallel to the numbers:

1. Is this study valid?

Validity means the study measured what it claimed to measure and reached conclusions that follow from the data. Threats to validity include: non-random sampling, uncontrolled confounders, outcome measurement bias, and selective reporting. An RCT that was not blinded is less valid than one that was. A cohort study that failed to adjust for severity of illness is weaker than one that did.

2. Are these results real?

Statistical significance tells you the result is unlikely to be due to chance. But "unlikely due to chance" is not the same as "definitely real." A p-value of 0.049 is one data point. Reproducibility matters. A result replicated across multiple studies and settings is more credible than a single significant finding. When you see $p < 0.05$ once, be interested. When you see it in five independent studies, be convinced.

3. Do these results apply to my patients?

Generalizability means the study's findings extend to patients like yours. Check the inclusion and exclusion criteria. A trial conducted in a tertiary academic medical center with strict exclusion criteria may not generalize to a community hospital with a different case mix. Sample demographics, comorbidity burdens, and treatment settings all affect applicability.

Evidence-based nursing is not the same as "the study was significant, so we should change practice." It is the disciplined integration of research evidence, clinical expertise, and patient values — and it requires knowing how to read the evidence critically enough to know when it earns trust.

15.2 Common Statistical Errors and Misleading Presentations

Step 1 Read This First

A hospital quality dashboard shows: “Patient satisfaction improved by 15% after implementing bedside rounding.” A quality improvement report claims: “Medication errors are down — only 0.3% of administrations involved an error.” A published abstract reports: “The new drug produced a statistically significant improvement in pain scores ($p = 0.042$).”

Each of these statements contains a number. Each could be misleading. The 15% improvement — is that relative or absolute? From what baseline? The 0.3% error rate — is that good or bad? What does “only” mean? The p-value of 0.042 — what was the actual effect size? How many outcomes were tested? Was this the primary or a secondary endpoint?

Statistical literacy is not just knowing how to compute a confidence interval. It is knowing how to read other people’s numbers skeptically.

Step 2 Let’s Talk About It

Most statistical errors in published research are not fraud — they are honest mistakes, optimistic interpretations, or presentation choices that favor the authors’ preferred narrative. Knowing the common patterns helps you spot them without needing to read every article with suspicion.

The most important lesson is this: *never let a p-value substitute for thinking.* A p-value answers exactly one question: how likely is this result (or something more extreme) if the null hypothesis is true? Everything else — effect size, clinical importance, generalizability, replicability — requires judgment.

Numbers invite the illusion of precision. An odds ratio of 1.47 sounds exact. But if the confidence interval runs from 0.98 to 2.21, the truth is consistent with no effect at all.

Step 3 Now We Name It

Definition: Common Statistical Errors in Clinical Research

- **Confusing relative and absolute risk.** "The drug reduced mortality by 50%" sounds dramatic. If baseline mortality was 2% and the new rate is 1%, the absolute reduction is 1 percentage point (NNT = 100). Always ask: reduction from what?
- **P-hacking (multiple comparisons without correction).** A study that tests 20 outcomes without correcting for multiple comparisons expects to find about 1 false positive at $\alpha = 0.05$. Look for primary vs. secondary endpoints; be skeptical of significant results on outcomes not declared as primary.
- **Underpowered studies.** A non-significant result from a study of 30 patients per group tells you almost nothing. Small studies have wide confidence intervals. Absence of evidence is not evidence of absence.
- **Confusing statistical and clinical significance.** With a large enough sample, trivially small effects become statistically significant. A mean pain score that drops from 6.0 to 5.9 on a 10-point scale is clinically meaningless, but $p < 0.001$ if $n = 10{,}000$.
- **Misleading graphs.** A y-axis that does not start at zero can make a small difference appear dramatic. Bar charts without error bars conceal variability. Always look at the axis scale.
- **Survivor bias and missing data.** Patients who dropped out of the study were not random — they were often sicker or more dissatisfied. An intention-to-treat analysis (analyzing all patients as randomized, regardless of completion) is more valid than a per-protocol analysis.
- **Correlation presented as causation.** A significant chi-square or Pearson correlation does not establish cause and effect. In observational studies, unmeasured confounders are always a threat.

Step 4 Watch It Work

Spotting errors and evaluating two contrasting reports

Report 1 — Misleading relative risk framing:

"Our new hand hygiene campaign doubled compliance. Before the campaign, only 43 of 200 observed encounters showed proper technique. After, 86 of 200 did. That is a 100% improvement!"

What is actually true:
Before: $\hat{p}_1 = 43/200 = 21.5\%$ After: $\hat{p}_2 = 86/200 = 43.0\%$
Relative improvement: $(43.0 - 21.5)/21.5 \times 100\% = 100\%$ — technically correct.
Absolute improvement: $43.0\% - 21.5\% = 21.5$ percentage points.
The problem: Compliance went from 21.5% to 43% — still less than halfway to the target of 80%. The relative framing sounds transformational; the absolute framing reveals how much work remains.

Report 2 — Evaluating a secondary outcome significant result:

"A trial of a new sepsis protocol reports: primary outcome (28-day mortality) was not significantly reduced ($p = 0.23$). Among the 15 secondary outcomes tested, ICU length of stay was significantly shorter in the intervention group ($p = 0.038$)."

What this means statistically:
The study failed on its primary endpoint. Testing 15 secondary outcomes without correction means about 0.75 false positives are expected at $\alpha = 0.05$. The significant LOS result is almost certainly one of them — a false positive that would disappear under proper Bonferroni correction ($\alpha = 0.05/15 = 0.0033$).
Clinical bottom line: Do not change practice based on a secondary outcome from a trial that failed its primary endpoint. Treat it as a hypothesis for a future study.

Try This in Desmos

Verify the hand hygiene two-proportion z-statistic:
$\hat{p}_c = (43 + 86)/400 = 0.3225$; $SE = \sqrt{0.3225(0.6775)(1/200 + 1/200)} \approx 0.0467$
$z = (0.430 - 0.215)/0.0467 \approx 4.60$; `1 - normaldist(0,1).cdf(4.60)` ≈ 0.000002
The before-after change is highly significant even though the absolute compliance level remains low.

Step 5 Your Turn

1. A hospital report states: "Catheter-associated urinary tract infections (CAUTIs) dropped by 33% after our new insertion bundle." Further investigation reveals: there were 9 CAUTIs in the pre-bundle period and 6 in the post-bundle period,

each measured over 3 months in a 20-bed unit.

(a) Verify the 33% claim. Is this relative or absolute?
(b) The baseline rate was 9 infections per quarter. Is 9 → 6 a convincing change? What concerns would you raise about this analysis?
(c) What statistical test might you apply here, and what would you need to make a proper inference?
(d) If the post-bundle rate is 6 CAUTIs per quarter and the national benchmark is 2 per quarter, what does the statistical result tell us about the adequacy of the improvement?

2. A published study reports: "Nurses who received therapeutic communication training showed significantly higher patient satisfaction scores than untrained nurses (82.3 vs. 79.1, $p = 0.04$, 95% CI: 0.1 to 6.3)."

(a) Is the result statistically significant? Explain.
(b) What does the confidence interval tell you about the precision of the estimate?
(c) The difference is 3.2 points on a 100-point scale. Is this clinically meaningful? What additional information would you want?
(d) The study was not randomized — nurses self-selected into training. What threat to validity does this create? What would an appropriate study design look like?

3. ⋆ A meta-analysis of 12 RCTs examining nurse staffing ratios and patient mortality reports: "Pooled relative risk of in-hospital mortality for patients on understaffed units vs. adequately staffed units: RR = 1.31 (95% CI: 1.18 to 1.45, $p < 0.001$, $I^2 = 38\%$)."

(a) Interpret RR = 1.31 in plain language.
(b) The 95% CI does not include 1. What does this tell you?
(c) $I^2 = 38\%$ is a measure of heterogeneity across the 12 studies. A rough guideline: $I^2 < 25\%$ = low, 25–50% = moderate, $> 75\%$ = high heterogeneity. What does $I^2 = 38\%$ indicate about the consistency of results across studies?
(d) This is a meta-analysis of RCTs — the strongest form of evidence. What does this result mean for nursing policy advocacy?
(e) Even with RR = 1.31 and $p < 0.001$, what additional factors would a hospital administrator need to weigh before changing staffing policy?

Step 6 Think Like a Nurse

You Are Now a Statistical Reader
When you began this course, a table of means and p-values was a wall of numbers. Now each number speaks.

A mean of 79.1 tells you where a group sits. A standard deviation of 12.3 tells you how spread out they are. A confidence interval of (0.1, 6.3) tells you the result is real but the effect might be trivially small. A p-value of 0.04 tells you the result would be unusual by chance — but it does not tell you whether it matters.

An RCT tells you something about cause. A cohort study tells you about association. A cross-sectional survey tells you about prevalence at one moment. A meta-analysis pools many voices. You now know what each voice can and cannot say.

Statistics you have mastered in this course:

- Describing data: means, medians, standard deviations, quartiles, graphs
- Probability and distributions: normal, binomial, t, chi-square, F
- The Central Limit Theorem: why sample means are predictable
- Confidence intervals for means and proportions
- Hypothesis testing: one-sample, two-sample, paired, proportion, chi-square, ANOVA
- Effect sizes: Cohen's d, η^2, relative risk, odds ratio, NNT
- Critical reading: study design, validity, clinical vs. statistical significance

What comes next:
This course is the foundation. Regression analysis, survival analysis, multilevel modeling, Bayesian methods, and systematic review methodology are the floors above it. Every one of them is built on what you have learned here.

Every policy decision at your hospital should have a number attached — and someone who can read it. That person is you now. Use it.

15.3 Chapter Practice Problems

Step 5 Your Turn

The following problems ask you to read and critically evaluate short research excerpts. Problems marked with ⋆ are more challenging.

Part A: Identifying Design and Interpreting Results

1. Read the following abstract excerpt and answer the questions:

"We conducted a retrospective cohort study of 1,240 adult patients admitted to three hospitals. Patients who received nurse-led discharge education ($n = 620$) were compared to those who did not ($n = 620$). The primary outcome was 30-day readmission. The intervention group had a readmission rate of 11.3% vs. 18.5% in the control group (OR = 0.56, 95% CI: 0.39 to 0.80, $p = 0.001$)."

(a) What study design is this? What causal claims does it support?
(b) Interpret OR = 0.56.
(c) The CI does not include 1. Is the result statistically significant?
(d) Compute the absolute risk reduction and NNT.
(e) The study is observational. Name two confounders that might explain the result without a causal effect of the intervention.

2. Read the following results excerpt and answer the questions:

"Mean length of stay was 5.2 days (SD = 2.4) in the intervention group and 5.9 days (SD = 2.7) in the control group ($t(238) = 2.18$, $p = 0.030$, 95% CI for difference: -1.33 to -0.07 days)."

(a) Is the result statistically significant at $\alpha = 0.05$?
(b) What is the point estimate of the difference in means?
(c) The CI is $(-1.33, -0.07)$ days. What does the wide interval tell you?
(d) How many total patients are in the study?
(e) A hospital administrator says: "We should implement this immediately — it cut LOS by 0.7 days." How would you respond using the confidence interval?

Part B: Evaluating Evidence Quality

3. For each of the following study descriptions, identify the design, state the level of evidence (from the hierarchy in Section 15.1), and describe one major limitation:

(a) A nurse surveys 200 patients about their pain management experience on a single hospital unit in one month.

(b) Researchers follow 500 patients with hypertension for 5 years and compare outcomes between those who have a regular nurse practitioner and those who do not.

(c) A pilot study assigns 15 patients to a new wound care gel and 15 to standard care using coin-flip randomization.

(d) A systematic review synthesizes findings from 22 RCTs examining the effect of early mobilization on ICU outcomes.

(e) A nurse anecdotally describes two patients who recovered faster after receiving music therapy.

4. ⋆ A study reports the following ANOVA table comparing pain scores across four analgesic protocols:

Source	**SS**	**df**	**MS**	F	**P-value**
Between groups	48.6	3	16.2	4.39	0.005
Within groups	575.8	156	3.691		
Total	624.4	159			

(a) How many total patients? How many groups?
(b) Verify the F-statistic from the MS values.
(c) Compute η^2 and classify the effect size.
(d) The authors conclude: "Analgesic protocol significantly affects pain scores." Is this conclusion supported?
(e) Post-hoc Tukey tests find that only Protocol A differs significantly from Protocol D. The mean difference is 1.1 points on a 10-point pain scale. Is this clinically meaningful? How would you decide?

Part C: Integrated Analysis

5. ⋆ **Full Article Analysis.** Examine the following result set from a hypothetical RCT of a structured nurse handoff protocol ($n = 180$ per group):

Outcome	**Intervention**	**Control**	**P-value**	**95% CI**
Adverse events (prop.)	8.3%	14.4%	0.041	−11.9% to −0.3%
Near-miss reports (mean)	1.2	1.5	0.18	−0.7 to 0.1
Nurse satisfaction (mean/100)	74.1	71.3	0.063	−0.2 to 5.8
LOS (days, mean)	5.8	6.3	0.009	−0.9 to −0.1
30-day readmission (prop.)	9.4%	11.1%	0.54	−5.7% to 2.3%

(a) Which outcomes are statistically significant at $\alpha = 0.05$? Which are not?
(b) The primary outcome was adverse events. Compute the ARR and NNT.
(c) The nurse satisfaction p-value is 0.063. The CI is $(-0.2, 5.8)$. Interpret this carefully: what does the CI tell you that the p-value alone does not?
(d) Five outcomes were tested. Apply Bonferroni correction. Which results survive at the corrected α?
(e) Based on the full pattern of results, write a 3-sentence summary that a nursing director could use to decide whether to implement this protocol.

Answer Key — Selected Problems

Answer Key

Section 15.1 Practice — Your Turn

1. (a) RR = 0.57 means patients in the intervention group had 57% of the readmission risk of controls — a 43% relative reduction. (b) The CI (0.38, 0.87) excludes 1, confirming statistical significance at $\alpha = 0.05$. (c) ARR = 14.7% − 8.4% = 6.3 percentage points; NNT = 1/0.063 ≈ 16. Treat 16 patients to prevent 1 readmission. (d) Approximate counts: $0.084 \times 300 \approx 25$ events in intervention; $0.147 \times 300 \approx 44$ in control. (e) What is the control group's usual care? Who was excluded? Is the reduction sustained beyond 30 days? Does NNT = 16 justify the cost and burden of the intervention?

2. (a) Cross-sectional survey. This design supports correlation/association only; it cannot establish that shift length *causes* burnout. (b) Cohen's $d = 0.28$ is a small effect: the means of the two groups are about 0.28 standard deviations apart. (c) Statistically significant but likely not clinically meaningful with $d = 0.28$ — a small effect. (d) An RCT or natural experiment comparing nurses randomly assigned (or quasi-randomly assigned by scheduling) to different shift lengths, with controlled follow-up. Observational designs are insufficient to justify a policy change.

Section 15.2 Practice — Your Turn

1. (a) $9 \to 6$ is a relative reduction of $(9-6)/9 = 33\%$. This is a relative measure; the absolute reduction is 3 infections per quarter. (b) A reduction of 3 infections in a 20-bed unit is a small absolute number. With counts this small, random variation is large — one extra case next quarter would reverse the trend. (c) A proportion test or Poisson rate comparison would be appropriate; the main issue is that $n = 9$ and $n = 6$ are raw counts and confidence intervals would be very wide. (d) Reaching 6 CAUTIs per quarter is still 3 times the benchmark rate of 2 per quarter — the improvement is real but the target has not been met.

2. (a) Yes, statistically significant at $\alpha = 0.05$ since $p = 0.04 < 0.05$. (b) The CI $(0.1, 6.3)$ is very wide, ranging from a trivial 0.1-point difference to a potentially meaningful 6.3-point difference. The result is consistent with anywhere in that range — the estimate is imprecise. (c) A 3.2-point difference on a 100-point scale is clinically debatable; the minimum clinically important difference (MCID) for patient satisfaction tools is typically 5–10 points. (d) Threat: nurses who chose training may differ systematically from those who did not — more motivated, less burned out, with more supportive managers. A proper design would randomize units (not individuals) to training or control, with stratification on baseline satisfaction.

Chapter Practice — Selected

Problem 1b: OR = 0.56 means the odds of readmission in the educated group were

56% of the odds in the control group — a substantial reduction.
Problem 1d: ARR $= 18.5\% - 11.3\% = 7.2$ percentage points; NNT $= 1/0.072 \approx 14$.
Problem 2b: Point estimate: $5.2 - 5.9 = -0.7$ days.
Problem 2d: From the t-test notation $t(238)$: $df = n_1 + n_2 - 2 = 238$, so $N = 240$.
Problem 4b: $F = 16.2/3.691 \approx 4.39 \approx 4.38$ ✓.
Problem 4c: $\eta^2 = 48.6/624.4 \approx 0.078$ — medium effect.
Problem 5a: Adverse events ($p = 0.041$) and LOS ($p = 0.009$) are statistically significant. Near-miss, satisfaction, and readmission are not.
Problem 5b: ARR $= 14.4\% - 8.3\% = 6.1$ percentage points; NNT $= 1/0.061 \approx 16$.
Problem 5d: Bonferroni-corrected $\alpha = 0.05/5 = 0.01$. Only LOS ($p = 0.009$) survives. Adverse events ($p = 0.041 > 0.01$) no longer significant under correction.

Chapter 15 Summary

Section 15.1 — Anatomy of a Research Article

- Ask five questions before reading numbers: design, sample, primary outcome, sample size adequacy, and statistical appropriateness.
- Study design determines causal claims: RCTs > cohort > case-control > cross-sectional > case report.
- Absolute risk reduction (ARR) and NNT are more actionable than relative risk alone.
- If the confidence interval for RR or OR includes 1, the result is not statistically significant.

Section 15.2 — Common Statistical Errors and Misleading Presentations

- Relative vs. absolute risk: always ask "reduction from what baseline?"
- P-hacking: significant secondary outcomes from trials that failed their primary endpoint require extreme caution.
- Underpowered studies: non-significant $\neq$ no effect when n is small.
- Statistical significance $\neq$ clinical significance: effect size and clinical context are always required.
- Survivor bias and missing data: intention-to-treat analysis is the gold standard for RCTs.
- Correlation is not causation: observational studies require adjustment for confounders.

The Nursing Connection

- Every clinical protocol, medication, and procedure at your hospital was validated — or should have been — through research using the tools in this course.
- Your ability to read that research critically is a patient safety skill, not just an academic one.
- Evidence-based practice requires three things: valid evidence, clinical judgment, and patient values. Statistics provides the first; nursing provides the other two.
- The question is never just "is $p < 0.05$?" — it is "does this number, in this study, with these patients, tell me something I should act on?"

Appendix A: Desmos Statistical Reference Guide

Desmos (`desmos.com/scientific`) is your graphing calculator for this course.
All functions below work in the Desmos Scientific Calculator.

How to Use This Guide

Type each command exactly as shown into the Desmos expression bar. Replace values in **bold** with your actual numbers. Desmos is case-sensitive: use lowercase for all distribution names.

A.1 Normal Distribution

Task	Desmos Command	Chapter
Area to the left of x under $N(\mu, \sigma)$	`normaldist(mu, sigma).cdf(x)`	9
Area to the right of x	`1 - normaldist(mu, sigma).cdf(x)`	9
Area between a and b	`normaldist(mu, sigma).cdf(b) - normaldist(mu, sigma).cdf`	9
Inverse: find x given left-tail area p	`normaldist(mu, sigma).inversecdf(p)`	9
Standard normal $P(Z \leq z)$	`normaldist(0,1).cdf(z)`	9, 12
Standard normal inverse	`normaldist(0,1).inversecdf(p)`	11

Common Mistake: σ vs. SE

When working with **individual values** use σ as the second argument. When working with **sample means**, use $SE = \sigma/\sqrt{n}$ instead. Same command, different input.

Examples:

- $P(X \leq 135)$ for $N(120, 15)$: `normaldist(120, 15).cdf(135)` $= 0.8413$
- $P(\bar{x} \leq 122)$ for $n = 25$, $\mu = 120$, $\sigma = 15$, $SE = 3$: `normaldist(120, 3).cdf(122)` $= 0.7475$
- 95th percentile of $N(100, 12)$: `normaldist(100, 12).inversecdf(0.95)` $= 119.74$

A.2 *t*-Distribution

Task	**Desmos Command**	**Chapter**
Left-tail area $P(T \leq t)$	`tdist(df).cdf(t)`	11, 12
Right-tail p-value	`1 - tdist(df).cdf(t)`	12
Two-tailed p-value	`2*(1 - tdist(df).cdf(abs(t)))`	12, 13
Critical value t^* for 95% CI	`tdist(df).inversecdf(0.975)`	11
Critical value t^* for 90% CI	`tdist(df).inversecdf(0.95)`	11
Critical value t^* for 99% CI	`tdist(df).inversecdf(0.995)`	11

Two-Tailed Critical Values

For a 95% CI, use `inversecdf(0.975)`, not `inversecdf(0.95)`. The 0.975 leaves 2.5% in each tail, for a total of 5% outside the interval.

Examples:

- Left-tail p-value, $t = -2.07$, $df = 39$: `tdist(39).cdf(-2.07)` ≈ 0.023
- Right-tail p-value, $t = 1.98$, $df = 24$: `1 - tdist(24).cdf(1.98)` ≈ 0.030
- Two-tailed p-value, $t = 2.45$, $df = 15$: `2*(1 - tdist(15).cdf(2.45))` ≈ 0.027
- t^* for 95% CI, $df = 29$: `tdist(29).inversecdf(0.975)` $= 2.045$

A.3 Chi-Square Distribution

Task	Desmos Command	Chapter
Right-tail p-value $P(\chi^2 \geq x)$	`1 - chisquaredist(df).cdf(x)`	13
Left-tail area $P(\chi^2 \leq x)$	`chisquaredist(df).cdf(x)`	13

Note: Chi-square tests are *always right-tailed*. Use `1 - chisquaredist(df).cdf(value)`.

Examples:

- $P(\chi^2 \geq 3.95)$, $df = 2$: `1 - chisquaredist(2).cdf(3.95)` ≈ 0.139
- $P(\chi^2 \geq 20.05)$, $df = 2$: `1 - chisquaredist(2).cdf(20.05)` ≈ 0.000

A.4 F-Distribution (ANOVA)

Task	Desmos Command	Chapter
Right-tail p-value $P(F \geq x)$	`1 - fdist(df1, df2).cdf(x)`	14
Left-tail area	`fdist(df1, df2).cdf(x)`	14

Note: Enter numerator df $(k-1)$ first, then denominator df $(N-k)$. ANOVA p-values are always right-tailed.

Examples:

- $P(F \geq 6.43)$, $df_1 = 3$, $df_2 = 56$: `1 - fdist(3, 56).cdf(6.43)` ≈ 0.001
- $P(F \geq 42.7)$, $df_1 = 3$, $df_2 = 116$: `1 - fdist(3, 116).cdf(42.7)` ≈ 0.000

A.5 Quick Reference: Critical Values

Distribution	**90% CI**	**95% CI**	**99% CI**
Standard normal z^*	1.645	1.960	2.576
t^*, $df = 10$	1.812	2.228	3.169
t^*, $df = 20$	1.725	2.086	2.845
t^*, $df = 30$	1.697	2.042	2.750
t^*, $df = 40$	1.684	2.021	2.704
t^*, $df = 60$	1.671	2.000	2.660
t^*, $df = 120$	1.658	1.980	2.617
t^*, $df = \infty$	1.645	1.960	2.576

A.6 Desmos for Sample Size and Proportions

Task	**Desmos Expression**
SE for sample proportion	`sqrt(phat*(1-phat)/n)`
SE for sample mean	`sigma/sqrt(n)`
n for proportion, ME $= E$, z^*, prior $\hat{p}$	`ceil((1.96/E)^2 * phat * (1-phat))`
n for mean, ME $= E$, z^*, σ	`ceil((1.96*sigma/E)^2)`
Margin of error (95%, t-interval)	`tdist(df).inversecdf(0.975) * s/sqrt(n)`

Note: The `ceil()` function rounds up to the nearest integer. Always use it for sample size calculations.

A.7 Exam-Day Checklist

Before pressing Enter in Desmos during an exam, ask yourself:

1. **Individual or sample mean?** Use σ for individuals; $\sigma/\sqrt{n}$ (or $s/\sqrt{n}$) for sample means.
2. **Which tail?** Left-tail: use `.cdf()` directly. Right-tail: `1 - .cdf()`. Two-tailed: `2*(1 - .cdf(abs(t)))`.
3. **Which critical value quantile?** For 95% CI: `inversecdf(0.975)`. For 90%: `inversecdf(0.95)`. For 99%: `inversecdf(0.995)`.

4. **Correct distribution?** Normal for z-tests and proportions; t for means with unknown σ; chi-square for independence; F for ANOVA.

5. **Correct degrees of freedom?** t: $n-1$ (one sample) or $\min(n_1-1, n_2-1)$ (two sample). χ^2: $(r-1)(c-1)$. F: $(k-1, N-k)$.

Appendix B: Formula Sheet

This sheet may be provided on exams. All symbols are defined as used in the textbook.

B.1 Descriptive Statistics

Quantity	**Formula**
Sample mean	$\bar{x} = \dfrac{\sum x_i}{n}$
Sample variance	$s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$
Sample std. deviation	$s = \sqrt{s^2}$
Population mean	$\mu = \dfrac{\sum x_i}{N}$
Population std. deviation	$\sigma = \sqrt{\dfrac{\sum (x_i - \mu)^2}{N}}$
z-score (individual)	$z = \dfrac{x - \mu}{\sigma}$
IQR	$IQR = Q_3 - Q_1$

B.2 Probability

Rule	Formula
Addition (general)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Addition (mutually exclusive)	$P(A \cup B) = P(A) + P(B)$
Multiplication (general)	$P(A \cap B) = P(A) \cdot P(B \mid A)$
Multiplication (independent)	$P(A \cap B) = P(A) \cdot P(B)$
Complement	$P(A^c) = 1 - P(A)$
Conditional	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
Binomial probability	$P(X = k) = \dbinom{n}{k} p^k (1-p)^{n-k}$
Binomial mean	$\mu = np$
Binomial std. deviation	$\sigma = \sqrt{np(1-p)}$

B.3 Sampling Distributions and the CLT

Quantity	Formula
Standard error of $\bar{x}$	$SE = \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
z-score for sample mean	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
CLT (approximate normality)	$\bar{x} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$ for $n \geq 30$

B.4 Confidence Intervals

Interval	**Formula**
Mean (σ known, large n)	$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$
Mean (σ unknown, t-interval)	$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}, \quad df = n-1$
Proportion	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Margin of error	$ME = t^* \cdot \frac{s}{\sqrt{n}} \quad \text{or} \quad ME = z^* \cdot \frac{\sigma}{\sqrt{n}}$
Critical values	$z^* = 1.645\ (90\%),\ 1.960\ (95\%),\ 2.576\ (99\%)$
Sample size: mean	$n = \left(\frac{z^*\sigma}{E}\right)^2 \quad \text{(round up)}$
Sample size: proportion	$n = \left(\frac{z^*}{E}\right)^2 \hat{p}(1-\hat{p}) \quad \text{(round up)}$

B.5 Hypothesis Tests

Test	**Test Statistic**
One-sample z (mean, σ known)	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
One-sample t (mean, σ unknown)	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n-1$
Paired t-test	$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad df = n-1$
Two-sample t (independent)	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad df = \min(n_1 - 1,\ n_2 - 1)$
One-sample z (proportion)	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$
Two-sample z (proportions)	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n_1 + 1/n_2)}}, \quad \hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$
Chi-square (independence)	$\chi^2 = \sum \frac{(O-E)^2}{E}, \quad df = (r-1)(c-1), \quad E = \frac{\text{row total} \times \text{col total}}{n}$

B.6 ANOVA

Quantity	**Formula**
Grand mean	$\bar{x}_{grand} = \dfrac{\sum n_i \bar{x}_i}{N}$
Between-group SS	$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{grand})^2, \quad df_B = k - 1$
Within-group SS	$SS_W = \sum_{i=1}^{k} (n_i - 1)s_i^2, \quad df_W = N - k$
F-statistic	$F = \dfrac{MS_B}{MS_W} = \dfrac{SS_B/df_B}{SS_W/df_W}$
Effect size	$\eta^2 = \dfrac{SS_B}{SS_B + SS_W}$

B.7 Correlation and Regression

Quantity	**Formula**
Pearson correlation	$r = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$
Regression slope	$b_1 = r \cdot \dfrac{s_y}{s_x}$
Regression intercept	$b_0 = \bar{y} - b_1\bar{x}$
Predicted value	$\hat{y} = b_0 + b_1 x$
Coefficient of determination	$R^2 = r^2$

B.8 Clinical Research Measures

Measure	**Formula**
Relative risk	$RR = \dfrac{P(\text{outcome}\mid\text{exposed})}{P(\text{outcome}\mid\text{unexposed})}$
Absolute risk reduction	$ARR = p_{\text{control}} - p_{\text{treatment}}$
Number needed to treat	$NNT = \dfrac{1}{ARR}$
Odds ratio	$OR = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$

Appendix C: Statistical Tables

All values were computed to four decimal places. Use Desmos for values not listed here.

C.1 Standard Normal Distribution (Left-Tail Areas)

The table gives $P(Z \leq z)$ — the area to the **left** of z under the standard normal curve.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

C.2 t-Distribution Critical Values

Values shown are t^* such that $P(T \leq t^*) = p$ for the given df.

	Upper-tail probability α						
df	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
	(Two-tailed confidence level: 80%	*90%*	*95%*	*98%*	*99%*	*99.8%*	*99.9%)*
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

C.3 Chi-Square Distribution Critical Values

Values shown are χ^2_α such that $P(\chi^2 \geq \chi^2_\alpha) = \alpha$ (right-tail area).

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209
11	17.275	19.675	21.920	24.725
12	18.549	21.026	23.337	26.217
13	19.812	22.362	24.736	27.688
14	21.064	23.685	26.119	29.141
15	22.307	24.996	27.488	30.578
16	23.542	26.296	28.845	32.000
17	24.769	27.587	30.191	33.409
18	25.989	28.869	31.526	34.805
19	27.204	30.144	32.852	36.191
20	28.412	31.410	34.170	37.566
25	34.382	37.652	40.646	44.314
30	40.256	43.773	46.979	50.892

C.4 F-Distribution Critical Values ($\alpha = 0.05$)

Values shown are $F_{0.05}$ such that $P(F \geq F_{0.05}) = 0.05$.
Rows = df_{Within} (df_2), Columns = $df_{Between}$ ($df_1 = k - 1$).

$df_2 \backslash df_1$	1	2	3	4	5	6	8	10
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.83

Glossary of Key Terms

Terms are listed alphabetically. Chapter references indicate where each term is first introduced.

Alternative hypothesis (H_a)

The research claim being tested; contains a strict inequality ($<$, $>$, or $\neq$). Evidence against H_0 constitutes evidence for H_a. Ch. 12

ANOVA (Analysis of Variance)

A method for comparing means across three or more groups simultaneously using a single F-test, avoiding inflation of the Type I error rate from multiple comparisons. Ch. 14

Bar graph

A graphical display for categorical data showing counts or proportions as rectangular bars of equal width. Ch. 4

Bias

A systematic error in data collection or analysis that causes estimates to consistently over- or underestimate the true value. Ch. 2

Binomial distribution

The probability distribution for the number of successes in n independent trials, each with the same probability p of success. Ch. 8

Boxplot

A graphical display of five-number summary (min, Q_1, median, Q_3, max), useful for comparing distributions and identifying outliers. Ch. 4

Central Limit Theorem (CLT)

For any population with mean μ and finite standard deviation σ, the sampling distribution of $\bar{x}$ is approximately normal for sufficiently large n ($n \geq 30$). Ch. 10

Chi-square test

A hypothesis test for independence between two categorical variables in a contingency table, using the statistic $\chi^2 = \sum(O - E)^2/E$. Ch. 13

Cluster sampling

A sampling method in which the population is divided into groups (clusters); a random sample of clusters is selected and all members of chosen clusters are surveyed. Ch. 3

Coefficient of determination (R^2)

The proportion of variability in the response variable explained by the linear regression model; equals r^2. Ch. 6

Confidence interval

A range of plausible values for a population parameter, constructed so that the procedure captures the true value at the stated confidence level (e.g., 95%) in repeated sampling. Ch. 11

Confounding variable

A variable associated with both the explanatory variable and the response variable that can distort the apparent relationship between them. Ch. 2

Continuous variable

A quantitative variable that can take any value in an interval; measured rather than counted. Ch. 1

Control group

In an experiment, the group that receives no treatment or the standard treatment, used as a baseline for comparison. Ch. 2

Convenience sample

A non-probability sample composed of individuals who are easy to reach; prone to bias and not representative of the population. Ch. 3

Correlation coefficient (r)

A measure of the strength and direction of the linear relationship between two quantitative variables; ranges from -1 to 1. Ch. 6

Critical value

The threshold value of a test statistic (z^* or t^*) that separates the rejection region from the non-rejection region; also used to construct confidence intervals. Ch. 11, 12

Degrees of freedom (df)

A parameter that determines the shape of t-, chi-square, and F-distributions; generally related to sample size minus the number of estimated parameters. Ch. 11–14

Discrete variable

A quantitative variable that takes only countable values (e.g., number of patients). Ch. 1

Effect size

A standardized measure of the practical magnitude of an effect, independent of sample size; examples include Cohen's d, η^2, and relative risk. Ch. 12, 14, 15

Expected value

The long-run average outcome of a random variable; for a discrete variable, $E(X) = \sum x \cdot P(X = x)$. Ch. 8

Experimental study

A study in which the researcher randomly assigns subjects to treatments and observes outcomes; the only design that can establish causation. Ch. 2

F-statistic

The ratio $MS_{Between}/MS_{Within}$ in ANOVA; large values indicate the between-group variation exceeds within-group variation, providing evidence against H_0. Ch. 14

Five-number summary

The minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum of a data set. Ch. 5

Frequency distribution

A table displaying the count (frequency) or proportion (relative frequency) of observations in each category or class interval. Ch. 4

Histogram

A graphical display of the distribution of a quantitative variable, using adjacent bars whose heights represent frequencies or relative frequencies. Ch. 4

Hypothesis test

A formal statistical procedure for evaluating evidence against a null hypothesis, yielding a p-value and a reject/fail-to-reject decision. Ch. 12

Independent events

Two events A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$; the occurrence of one does not affect the probability of the other. Ch. 7

Inferential statistics
Methods for drawing conclusions about a population based on data from a sample. Ch. 1

Interquartile range (IQR)
The difference $Q_3 - Q_1$; measures the spread of the middle 50% of the data and is resistant to outliers. Ch. 5

Least-squares regression line
The line $\hat{y} = b_0 + b_1 x$ that minimizes the sum of squared residuals; the "best fit" line for predicting y from x. Ch. 6

Level of measurement
The type of information encoded by a variable: nominal (categories), ordinal (ordered categories), interval (equal spacing, no true zero), or ratio (equal spacing with true zero). Ch. 1

Lurking variable
A variable not included in the analysis that may affect the relationship between the variables under study. Ch. 6

Margin of error (ME)
Half the width of a confidence interval; measures the precision of the estimate. Ch. 11

Mean ($\bar{x}$ or μ)
The arithmetic average; the sum of all values divided by the number of values. Ch. 5

Median The middle value in an ordered data set; resistant to outliers. Ch. 5

Mode The most frequently occurring value in a data set. Ch. 5

Mutually exclusive events
Events that cannot occur simultaneously; $P(A \cap B) = 0$. Ch. 7

Normal distribution
A symmetric, bell-shaped probability distribution completely described by its mean μ and standard deviation σ; the most important distribution in statistics. Ch. 9

Number needed to treat (NNT)
The number of patients who must receive a treatment to prevent one additional adverse outcome; $NNT = 1/ARR$. Smaller NNT values indicate a more effective intervention. Ch. 15

Null hypothesis (H_0)
The default claim of no effect, no difference, or no change; always contains an equality. Ch. 12

Observational study
A study in which the researcher observes subjects without intervention; cannot establish causation due to potential confounding. Ch. 2

Outlier An observation that lies far from the other values; may indicate a measurement error or a genuinely unusual case. Ch. 5

Odds ratio (OR)
The ratio of the odds of an outcome in one group to the odds in another; used in case-control studies and logistic regression. OR $\approx$ RR when the outcome is rare ($< 10\%$). Ch. 15

Paired t-test
A one-sample t-test applied to the differences $d_i = x_{1i} - x_{2i}$ for matched pairs or before-after measurements. Ch. 13

Parameter
A numerical summary of a population (e.g., μ, σ, p); typically unknown and estimated from sample data. Ch. 1

Percentile
A value below which a given percentage of observations fall; the kth percentile has $k\%$ of the data below it. Ch. 5

P-value The probability of obtaining a test statistic as extreme as or more extreme than the observed value, assuming H_0 is true; smaller values constitute stronger evidence against H_0. Ch. 12

Point estimate
A single value used to estimate a population parameter (e.g., $\bar{x}$ estimates μ). Ch. 11

Population
The entire group of individuals or objects of interest in a study. Ch. 1

Power The probability that a hypothesis test correctly rejects a false null hypothesis; equals $1 - \beta$. Ch. 12

Probability

A number between 0 and 1 expressing the long-run relative frequency of an event. Ch. 7

Qualitative variable

A variable that classifies subjects into categories; also called a categorical variable. Ch. 1

Quantitative variable

A variable that takes numerical values for which arithmetic operations make sense. Ch. 1

Random sample

A sample in which every member of the population has a known, non-zero probability of being selected. Ch. 3

Randomized controlled trial (RCT)

An experimental study in which subjects are randomly assigned to treatment and control groups; the gold standard for establishing causation. Ch. 2, 15

Range The difference between the maximum and minimum values; a simple but non-resistant measure of spread. Ch. 5

Regression line

See *Least-squares regression line.* Ch. 6

Relative risk (RR)

The ratio of the probability of an outcome in the exposed group to the probability in the unexposed group; $RR = 1$ means no difference. Ch. 15

Residual The difference between an observed value and its predicted value in regression: $e = y - \hat{y}$. Ch. 6

Sample A subset of the population selected for study. Ch. 1

Sampling distribution

The probability distribution of a statistic (e.g., $\bar{x}$) over all possible samples of size n from a population. Ch. 10

Sampling variability
The natural variation in a statistic from sample to sample due to random selection. Ch. 10

Scatterplot
A graphical display showing the relationship between two quantitative variables; each point represents one observation. Ch. 6

Significance level (α)
The threshold probability below which a p-value is considered small enough to reject H_0; typically $\alpha = 0.05$. Ch. 12

Simple random sample (SRS)
A sample in which every possible sample of size n has an equal probability of being selected. Ch. 3

Skewness
A measure of asymmetry in a distribution; right-skewed distributions have a long tail to the right, left-skewed to the left. Ch. 4

Standard deviation
The square root of the variance; measures spread in the original units of the data. Ch. 5

Standard error (SE)
The standard deviation of a sampling distribution; for $\bar{x}$: $SE = \sigma/\sqrt{n}$. Measures how much a sample statistic varies from sample to sample. Ch. 10

Standard normal distribution
A normal distribution with $\mu = 0$ and $\sigma = 1$; denoted $N(0, 1)$ or Z. Ch. 9

Statistic A numerical summary of a sample (e.g., $\bar{x}$, s, $\hat{p}$); used to estimate population parameters. Ch. 1

Statistical significance
A result is statistically significant at level α if the p-value $< \alpha$; does not imply clinical or practical importance. Ch. 12

Stratified sampling
A sampling method in which the population is divided into strata (subgroups) and a random sample is drawn from each stratum. Ch. 3

Systematic sampling
A sampling method in which subjects are selected at regular intervals (every kth

individual) from an ordered list. Ch. 3

t-distribution

A family of symmetric, bell-shaped distributions with heavier tails than the normal; used when σ is unknown and estimated by s; parameterized by degrees of freedom. Ch. 11

Treatment

In an experiment, the condition or intervention assigned to subjects by the researcher. Ch. 2

Type I error

Rejecting H_0 when it is actually true (false positive); probability equals α. Ch. 12

Type II error

Failing to reject H_0 when it is actually false (false negative); probability equals β. Ch. 12

Undercoverage

A form of bias in which some members of the population have no chance of being selected into the sample. Ch. 3

Variance

The average of the squared deviations from the mean; $s^2 = \sum(x_i - \bar{x})^2/(n-1)$ for a sample. Ch. 5

Voluntary response sample

A sample composed of individuals who choose to respond; typically biased toward strong opinions. Ch. 3

z-score The number of standard deviations an observation lies above (positive) or below (negative) the mean: $z = (x - \mu)/\sigma$. Ch. 9