STAT340 Intuitive Study Guide

Understanding Probability, Statistics, and Monte Carlo Methods

How to Use This Guide: This guide explains concepts in plain English first, then gives you the math. Think of it as your friend explaining stats over coffee, not a textbook.

Table of Contents

---

1. Probability Basics

What is Probability Anyway?

The Big Idea: Probability is just a fancy way of measuring uncertainty. It's a number between 0 and 1 that tells you how likely something is to happen. - 0 = impossible (like rolling a 7 on a normal 6-sided die) - 1 = certain (like the sun rising tomorrow) - 0.5 = 50-50 chance (like flipping heads on a fair coin)

Sample Space (the "stuff that could happen")

Plain English: The sample space is just ALL the possible things that could happen in your experiment.

Symbol: Ω (omega - looks fancy but just means "all possibilities")

Examples to Remember: - Flip a coin: $\Omega = \{\text{Heads}, \text{Tails}\}$ - Roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$ - Measure someone's height: $\Omega = \text{all positive numbers}$

Memory Aid: Think of omega as "everything that could possibly happen."

Events (the "stuff we care about")

Plain English: An event is just a collection of outcomes we're interested in. It's a subset of the sample space.

Examples: - Event A = "rolling an even number" = $\{2, 4, 6\}$ - Event B = "rolling more than 4" = $\{5, 6\}$

Why This Matters: We calculate probabilities for events, not individual outcomes (well, sometimes individual outcomes too, but you get it).

Basic Probability Rules

Rule 1: Addition Rule (the "OR" rule)

Plain English: If you want to know the probability that EITHER event A OR event B happens, you add their probabilities but subtract the overlap (because you'd be counting it twice otherwise).

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

Read this as: "Prob of A or B = Prob of A + Prob of B - Prob of both"

Why subtract the overlap? Imagine you're counting people who like pizza OR burgers. If someone likes both, you'd count them twice if you just added. Subtracting the overlap fixes this.

Example: - P(roll even) = 3/6 = 0.5 - P(roll > 4) = 2/6 = 0.333 - P(roll even AND > 4) = P(roll 6) = 1/6 = 0.167 - P(roll even OR > 4) = 0.5 + 0.333 - 0.167 = 0.666

Rule 2: Complement Rule (the "NOT" rule)

Plain English: The probability something DOESN'T happen is just 1 minus the probability it DOES happen.

$$P(A^c) = 1 - P(A)$$

Read this as: "Prob of NOT A = 1 - Prob of A"

Memory Aid: Total probability = 1 (something has to happen). So if A takes up 0.3 of that, NOT-A gets the remaining 0.7.

Example: If P(rain tomorrow) = 0.3, then P(no rain) = 1 - 0.3 = 0.7

Rule 3: Conditional Probability (the "GIVEN" rule)

Plain English: This answers "What's the probability of A happening, GIVEN that we already know B happened?" You're narrowing down your sample space to just the cases where B is true.

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

Read this as: "Prob of A given B = Prob of both A and B / Prob of B"

Why This Makes Sense: You're zooming in on just the world where B happened, then asking how much of THAT world also has A.

Example That Clicks: - You roll a die and someone tells you "it's even" (that's B) - Now you're only considering {2, 4, 6} - What's the chance it's also greater than 4? (that's A) - Only 6 satisfies both, so P(A|B) = 1/3

The "Zoom In" Analogy: Conditional probability is like zooming in on just the part of the sample space where B is true, then calculating probability within that smaller world.

Independence - When Knowing One Thing Tells You Nothing

Plain English: Two events are independent if knowing one happened doesn't change the probability of the other. They don't affect each other AT ALL.

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

OR equivalently: $P(A \mid B) = P(A)$ (knowing B doesn't change A's probability)

Examples: - **Independent**: Coin flip 1 and coin flip 2 (first flip doesn't affect second) - **NOT Independent**: Drawing cards without replacement (first card affects what's left) - **Independent**: Your height and tomorrow's weather (these don't affect each other)

Memory Aid: Independent = "Info about one tells you NOTHING about the other"

---

2. Random Variables

What Even IS a Random Variable?

Plain English: A random variable is just a way to turn outcomes into numbers. That's it. It's a function that assigns a number to each possible outcome.

Symbol: We use capital letters like X, Y, Z

Why We Do This: Math is easier with numbers than with words. Instead of saying "the outcome was heads," we can say "$X = 1$".

Example: Flip a coin. Define X = 1 if heads, 0 if tails. Now we can do math!

Two Types of Random Variables

Discrete Random Variables (counting stuff)

Plain English: Takes on separate, distinct values. You can COUNT the possibilities (even if there are infinitely many).

Examples: - Number of heads in 10 coin flips: {0, 1, 2, ..., 10} - Number of customers in an hour: {0, 1, 2, 3, ...} - Roll of a die: {1, 2, 3, 4, 5, 6}

Memory Aid: Discrete = you can list them out (even if the list is infinite)

Continuous Random Variables (measuring stuff)

Plain English: Can take ANY value in a range. You MEASURE these things. Between any two values, there are infinitely many other values.

Examples: - Height: could be 5.7 or 5.73 or 5.732… feet - Time: could be 2.5 or 2.51 or 2.512… seconds - Temperature: any real number

Memory Aid: Continuous = measuring tape (can be any value in a range)

PMF - Probability Mass Function (for DISCRETE variables)

What It Is: PMF = “Probability Mass Function” (just a fancy name for “probability of each specific value”)

Plain English: For discrete random variables, the PMF tells you the probability of each possible value. It’s literally just $P(X = x)$ for each x.

Notation: $p_X(x) = P(X = x)$

Properties You Should Know: 1. All probabilities are between 0 and 1: $p_X(x) \geq 0$ 2. All probabilities add to 1: $\sum_{\text{all } x} p_X(x) = 1$

Example - Fair Die: - $p_X(1) = 1/6$ - $p_X(2) = 1/6$ - … and so on

Visualize It: Think of a bar graph where each bar’s height is the probability.

PDF - Probability Density Function (for CONTINUOUS variables)

What It Is: PDF = “Probability Density Function” (NOT probability of a specific point!)

Plain English: For continuous variables, the PDF tells you the “density” of probability at each point. The HEIGHT of the PDF at a point tells you how likely values near that point are.

IMPORTANT: For continuous random variables, $P(X = \text{exactly } 5) = 0$. Why? Because there are infinitely many possible values, any single exact value has probability zero.

What We Actually Calculate: Probabilities of INTERVALS (ranges):

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

Read this as: “Probability X is between a and b equals the AREA under the PDF curve from a to b”

The Area Analogy: Think of the PDF as a landscape. Probability = area under the curve. The total area under the entire PDF = 1.

Properties: 1. $f_X(x) \geq 0$ (density can’t be negative) 2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (total area = 1)

Memory Aid: - PMF = exact probabilities (discrete) - PDF = probability density (continuous, need to integrate to get probability)

CDF - Cumulative Distribution Function (works for BOTH types)

What It Is: CDF = "Cumulative Distribution Function"

Plain English: The CDF tells you "What's the probability that X is LESS THAN OR EQUAL TO x?" It's the "accumulated" probability up to that point.

$$F_X(x) = P(X \leq x)$$

Read this as: "The CDF at x equals the probability that X is at most x"

Why This Is Useful: - The CDF always exists (for both discrete and continuous) - Going from 0 to 1 as x increases - Can calculate any probability: $P(a < X \leq b) = F_X(b) - F_X(a)$

The "Accumulation" Analogy: Imagine walking left to right along the number line, collecting probability as you go. The CDF tells you how much you've collected so far.

Properties to Remember: 1. Always between 0 and 1 2. Never decreases (only goes up or stays flat) 3. Approaches 0 as $x \to -\infty$ 4. Approaches 1 as $x \to \infty$

Connection to PMF/PDF: - **Discrete**: $F_X(x) = \sum_{k \leq x} p_X(k)$ (add up all PMF values up to x) - **Continuous**: $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ (integrate PDF up to x)

---

3. Expected Value and Variance

Expected Value - The "Long Run Average"

What It Is: $E[X]$ = "Expected value" (also called the mean or average)

Plain English: If you repeated this random process MANY times, what would the average value be? That's the expected value.

Formula: - **Discrete**: $E[X] = \sum_{\text{all } x} x \cdot p_X(x)$ (multiply each value by its probability, add them up) - **Continuous**: $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$ (same idea, but integrate)

Simple Example: Fair die

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Interpretation: "On average, you'd roll 3.5." (Even though you can never actually roll 3.5!)

Memory Aid: Expected value = center of mass. If you made the probability distribution out of cardboard, $E[X]$ is where you'd balance it.

Linearity of Expectation (SUPER USEFUL!)

The Rule: $E[aX + bY] = aE[X] + bE[Y]$

Plain English: Expected values behave nicely with addition and multiplication by constants.

THE MAGIC: This works EVEN IF X and Y are dependent! You don't need independence!

Example: If $E[X] = 10$ and $E[Y] = 20$, then: - $E[X + Y] = 10 + 20 = 30$ - $E[3X] = 3 \times 10 = 30$ - $E[2X + 5Y] = 2(10) + 5(20) = 120$

Why This Matters: Makes complex calculations way easier.

Variance - How "Spread Out" Things Are

What It Is: $\text{Var}(X)$ = "Variance" (measures spread/variability)

Plain English: Variance tells you how far values typically are from the mean. High variance = lots of spread, low variance = clustered around mean.

Formula:

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

where $\mu = E[X]$

Read this as: "Variance = expected value of the squared distance from the mean"

Computational Formula (easier to use):

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

Read this as: "Variance = average of squares minus square of average"

Memory Aid: - $E[X^2]$ = "average of the squares" - $(E[X])^2$ = "square of the average" - Variance = average of squares - square of average

Properties: 1. Always ≥ 0 (can't have negative spread) 2. $\text{Var}(X) = 0$ only if X is constant (no randomness) 3. $\text{Var}(aX + b) = a^2\text{Var}(X)$ (adding constant doesn't change spread, multiplying scales variance by a^2)

Why a^2? Because variance is in squared units. If you double X, the spread doubles, so the squared spread quadruples.

Standard Deviation - Variance's Friendlier Cousin

What It Is: $\text{SD}(X) = \sigma = \sqrt{\text{Var}(X)}$

Plain English: Standard deviation is just the square root of variance. Why? Because it's in the SAME UNITS as X, making it easier to interpret.

Example: If X = height in inches, variance is in “squared inches” (weird!), but standard deviation is in inches (makes sense!).

Memory Aid: SD = typical distance from the mean. About 68% of values are within 1 SD of the mean (for normal distributions).

---

4. Common Distributions - The Usual Suspects

Think of distributions as “templates” for random situations. Each has a story and specific formulas.

DISCRETE DISTRIBUTIONS (Counting Things)

Bernoulli - The “Yes/No” Distribution

The Story: One trial with two outcomes. Success or failure. 1 or 0. That’s it.

Notation: $X \sim \text{Bernoulli}(p)$ means “X follows a Bernoulli distribution with probability p”

Parameter: p = probability of success

PMF:

$$P(X=1)=p, \quad P(X=0)=1-p$$

When to Use: Single coin flip, single yes/no question, one trial of anything with two outcomes.

Key Facts: - $E[X]=p$ (makes sense: if $p=0.3$, average value is 0.3) - $\text{Var}(X)=p(1-p)$ (variance is biggest when $p=0.5$)

Example: Flip a fair coin. $X=1$ if heads, $X=0$ if tails. Then $X \sim \text{Bernoulli}(0.5)$.

Memory Aid: Bernoulli = one trial, two outcomes.

Binomial - Counting Successes in Multiple Trials

The Story: You do n independent Bernoulli trials (like flipping a coin n times). Count how many successes you get.

Notation: $X \sim \text{Binomial}(n,p)$

Parameters: - n = number of trials - p = probability of success on each trial

PMF:

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

Read this as: "Probability of exactly k successes = (# ways to arrange k successes) × (prob of that specific arrangement)"

The Binomial Coefficient: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ = "n choose k" = number of ways to pick k things from n things

When to Use: - Number of heads in 10 coin flips - Number of people who show up out of 20 invited - Number of defective items in a sample

Key Facts: - $E[X] = np$ (makes sense: if you do 10 trials with $p = 0.3$, expect 3 successes) - $\text{Var}(X) = np(1-p)$

Memory Aid: Binomial = bunch of Bernoulli trials, count the successes.

Geometric - "How Long Until Success?"

The Story: You keep doing Bernoulli trials UNTIL you get your first success. Count the number of FAILURES before that first success.

Notation: $X \sim \text{Geometric}(p)$

Parameter: p = probability of success on each trial

PMF:

$$P(X = k) = (1-p)^k p$$

Read this as: "k failures (each with prob $1-p$) then one success (prob p)"

When to Use: - Number of times you have to roll a die before getting a 6 - Number of job applications before getting an offer - Number of failed login attempts before success

Key Facts: - $E[X] = \frac{1-p}{p}$ (if $p = 0.1$, expect 9 failures before first success) - $\text{Var}(X) = \frac{1-p}{p^2}$

Memoryless Property (weird but cool): If you've already failed 10 times, the expected number of ADDITIONAL failures is still $\frac{1-p}{p}$. It "forgets" the past.

Memory Aid: Geometric = keep trying until first success.

NOTE: Some books define this as "number of trials until first success" (so $E[X] = \frac{1}{p}$). Check your book!

Poisson - Counting Rare Events

The Story: Count events that happen over some interval of time or space. Events happen independently at a constant average rate.

Notation: $X \sim \text{Poisson}(\lambda)$

Parameter: λ = average rate (e.g., 5 emails per hour, 3 customers per day)

PMF:

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

When to Use: - Number of emails in an hour - Number of earthquakes in a year - Number of typos on a page - Approximation to Binomial when n is large, p is small, and np is moderate

Key Facts: - $E[X] = \lambda$ - $\text{Var}(X) = \lambda$ (mean EQUALS variance - this is special!)

Memory Aid: Poisson = counting events over time/space. Mean = variance = λ.

Connection to Exponential: If events follow Poisson, the TIME BETWEEN events follows Exponential (see below).

CONTINUOUS DISTRIBUTIONS (Measuring Things)

Uniform - "All Values Equally Likely"

The Story: Every value in the interval $[a, b]$ is equally likely. Flat probability density.

Notation: $X \sim \text{Uniform}(a, b)$

Parameters: a = minimum, b = maximum

PDF:

$$f_X(x) = \frac{1}{b - a} \text{ for } a \leq x \leq b$$

(and 0 outside that interval)

Read this as: "Constant density = $\frac{1}{\text{width of interval}}$"

When to Use: - Random number between 0 and 1 - Random location along a route - "I have no idea, all values in this range seem equally plausible"

Key Facts: - $E[X] = \frac{a+b}{2}$ (midpoint - makes sense!) - $\text{Var}(X) = \frac{(b-a)^2}{12}$

Memory Aid: Uniform = flat, all values equally likely in a range.

Exponential - "Waiting Time for an Event"

The Story: How long until the next event happens? Time between events in a Poisson process.

Notation: $X \sim \text{Exponential}(\lambda)$

Parameter: λ = rate (same as Poisson rate)

PDF:

$$f_X(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

CDF (useful!):

$$F_X(x) = 1 - e^{-\lambda x} \text{ for } x \geq 0$$

When to Use: - Time until next customer arrives - Lifetime of a light bulb - Time until next earthquake

Key Facts: - $E[X] = \frac{1}{\lambda}$ (if $\lambda = 2$ per hour, expect to wait 0.5 hours) - $\text{Var}(X) = \frac{1}{\lambda^2}$

Memoryless Property: If you've already waited 5 minutes, the expected ADDITIONAL wait is still $\frac{1}{\lambda}$ minutes. Past waiting doesn't matter!

Memory Aid: Exponential = waiting time. Mean = $\frac{1}{\text{rate}}$.

Normal (Gaussian) - The "Bell Curve"

The Story: THE most important distribution. Symmetric bell-shaped curve. Shows up EVERYWHERE due to the Central Limit Theorem.

Notation: $X \sim N(\mu, \sigma^2)$

Parameters: - μ = mean (center of the bell) - σ^2 = variance (how spread out the bell is)

PDF (you don't need to memorize this):

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

When to Use: - Heights, weights, test scores - Measurement errors - Anything that's the SUM of many small random effects (CLT!)

Key Facts: - $E[X] = \mu$ - $\text{Var}(X) = \sigma^2$ - Symmetric around μ

The 68-95-99.7 Rule (MEMORIZE THIS): - About 68% of values within $\mu \pm \sigma$ - About 95% of values within $\mu \pm 2\sigma$ - About 99.7% of values within $\mu \pm 3\sigma$

Standard Normal: $Z \sim N(0, 1)$ (mean=0, variance=1) - Use for all calculations with normal distributions - Convert any normal to standard normal: $Z = \frac{X-\mu}{\sigma}$

Memory Aid: Normal = bell curve. 95% within 2 standard deviations.

Important Properties: 1. **Sum of Normals is Normal**: If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then:

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

2. **Linear Transformation**: If $X \sim N(\mu, \sigma^2)$, then:

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

---

5. Independence and Conditional Probability

Conditional Probability - "Given That We Know…"

The Big Idea: Conditional probability is about updating our beliefs when we get new information.

Formula:

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

The "Zoom In" Interpretation: 1. Start with the full sample space 2. Someone tells you "B happened" - now you zoom in to just the B part 3. Within that zoomed-in B world, what fraction also has A? 4. That fraction is $P(A \mid B)$

Example That Makes It Click: You have 100 students: - 60 are CS majors (let's call this event C) - 40 like statistics (event S) - 30 are CS majors AND like statistics

Question: What's $P(\text{likes stats} \mid \text{CS major})$?

Answer: Among the 60 CS majors, 30 like stats. So $P(S \mid C) = \frac{30}{60} = 0.5$

Using the formula: $P(S \mid C) = \frac{P(S \cap C)}{P(C)} = \frac{30/100}{60/100} = \frac{30}{60} = 0.5$ ✓

Bayes' Rule - THE MOST IMPORTANT FORMULA

THIS IS HUGE: Bayes' Rule is one of the most important concepts in all of statistics. It shows up EVERYWHERE: medical testing, machine learning, decision making, updating beliefs with new evidence.

What It Does

Plain English: Bayes' Rule lets you "flip" a conditional probability. It answers: "I know $P(B \mid A)$, but I actually need $P(A \mid B)$. How do I get it?"

The Classic Mix-Up: - You know: "If you have the disease, probability of testing positive" - You want: "If you test positive, probability of having the disease" - These are NOT the same! Bayes' Rule connects them.

The Formula

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

In Words: “posterior = (likelihood × prior) / evidence”

The Extended Version (when you need to calculate $P(B)$):

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)}$$

Why the Extended Version? Often we don’t know $P(B)$ directly, but we can calculate it using the Law of Total Probability (see below).

The Terms (MEMORIZE THESE NAMES)

- $P(A)$ = **PRIOR**: What you believed about A before seeing any evidence B
- $P(B \mid A)$ = **LIKELIHOOD**: How likely the evidence B is, assuming A is true
- $P(A \mid B)$ = **POSTERIOR**: What you believe about A after seeing evidence B
- $P(B)$ = **MARGINAL/EVIDENCE**: Overall probability of seeing evidence B

The Story: Start with prior belief → See evidence → Update to posterior belief

Law of Total Probability (The Missing Piece)

Before we can use Bayes, we often need to calculate $P(B)$:

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)$$

Plain English: The total probability of B happening is the sum of: - (Prob of B when A happens) × (Prob A happens) - (Prob of B when A doesn’t happen) × (Prob A doesn’t happen)

The Partition Idea: A and A^c split the world into two pieces. B can happen in either piece, so we add them up.

Why Bayes Works - The Intuition

Think of it like this: 1. Out of 1000 people, how many have A? → That’s $P(A) \times 1000$ 2. Of those with A, how many show evidence B? → That’s $P(B \mid A) \times [\text{people with A}]$ 3. But we also get B from people WITHOUT A → That’s $P(B \mid A^c) \times [\text{people without A}]$ 4. Total people with B = people from step 2 + people from step 3 5. Of all people with B, what fraction actually have A? → Step 2 / Step 4

That’s Bayes’ Rule!

Example 1: Medical Testing (The Classic)

Setup: - 1% of people have disease (very rare!) - If you have disease: test is 95% accurate (true positive rate) - If you’re healthy: test has 5% false positive rate

Question: You test positive. What’s the probability you actually have the disease?

Most People Think: 95% (WRONG!)

Let's Use Bayes:

Let D = "has disease", + = "test positive"

Want: $P(D \mid +)$

Know: - $P(D) = 0.01$ (prior: 1% have disease) - $P(+ \mid D) = 0.95$ (likelihood: test is 95% sensitive) - $P(+ \mid D^c) = 0.05$ (false positive rate)

Step 1: Calculate $P(+)$ using Law of Total Probability:

$$P(+) = P(+ \mid D) \cdot P(D) + P(+ \mid D^c) \cdot P(D^c)$$

$$= 0.95 \times 0.01 + 0.05 \times 0.99 = 0.0095 + 0.0495 = 0.0590$$

Step 2: Apply Bayes:

$$P(D \mid +) = \frac{P(+ \mid D) \cdot P(D)}{P(+)} = \frac{0.95 \times 0.01}{0.0590} = \frac{0.0095}{0.0590} \approx 0.161$$

Answer: Only 16%!

Why So Low? Out of 1000 people: - 10 have disease → 9.5 test positive (true positives) - 990 are healthy → 49.5 test positive (false positives) - Total positives: 59 - True positives: 9.5 - Probability: 9.5/59 ≈ 16%

The Lesson: When the disease is RARE (low prior), most positive tests are false alarms! The base rate matters A LOT.

Example 2: Backyard Animals (From Your Homework)

Setup: - I see a squirrel 80% of days (4/5 of days) - When squirrel present: I see hawk 65% of time - When squirrel absent: I see hawk 15% of time

Question: I see a hawk today. What's the probability the squirrel also visited?

Let's Use Bayes:

Let S = "squirrel visits", H = "hawk visits"

Want: $P(S \mid H)$

Know: - $P(S) = 0.8$ (prior: squirrel usually visits) - $P(H \mid S) = 0.65$ (likelihood: hawk often comes with squirrel) - $P(H \mid S^c) = 0.15$ (hawk sometimes comes without squirrel)

Step 1: Calculate $P(H)$:

$$P(H) = P(H \mid S) \cdot P(S) + P(H \mid S^c) \cdot P(S^c)$$

$$= 0.65 \times 0.8 + 0.15 \times 0.2 = 0.52 + 0.03 = 0.55$$

Step 2: Apply Bayes:

$$P(S \mid H) = \frac{P(H \mid S) \cdot P(S)}{P(H)} = \frac{0.65 \times 0.8}{0.55} = \frac{0.52}{0.55} \approx 0.945$$

Answer: 94.5% chance the squirrel was there!

Interpretation: Seeing a hawk is strong evidence the squirrel was there, because hawks are much more likely when the squirrel is present.

Example 3: Multiple Categories

Setup: Email spam filter. Emails can be "Spam" (60%), "Important" (10%), or "Regular" (30%). Word "FREE" appears in: - 80% of spam emails - 5% of important emails - 20% of regular emails

Question: Email contains "FREE". What's the probability it's spam?

Let's Use Bayes:

Let S = Spam, I = Important, R = Regular, F = contains "FREE"

Want: $P(S \mid F)$

Extended Bayes for Multiple Categories:

$$P(S \mid F) = \frac{P(F \mid S) \cdot P(S)}{P(F \mid S) \cdot P(S) + P(F \mid I) \cdot P(I) + P(F \mid R) \cdot P(R)}$$

Calculate: - Numerator: $0.80 \times 0.60 = 0.48$ - Denominator: $0.80 \times 0.60 + 0.05 \times 0.10 + 0.20 \times 0.30 = 0.48 + 0.005 + 0.06 = 0.545$

$$P(S \mid F) = \frac{0.48}{0.545} \approx 0.88$$

Answer: 88% chance it's spam!

The Bayes Strategy - Step by Step

Every Bayes Problem: 1. **Identify what you want**: $P(A \mid B)$ - what's the question asking? 2. **Identify what you know**: - $P(A)$ - the prior - $P(B \mid A)$ - likelihood when A is true - $P(B \mid A^c)$ - likelihood when A is false 3. **Calculate** $P(B)$ using Law of Total Probability (if not given) 4. **Plug into Bayes formula** 5. **Check your answer**: Does it make intuitive sense?

Common Bayes Patterns

Pattern 1: Medical Test - Prior: disease prevalence - Likelihood: test accuracy rates - Question: Given positive test, probability of disease?

Pattern 2: Source of Signal - Prior: probability of each source - Likelihood: probability of signal from each source - Question: Given signal, which source?

Pattern 3: Updating Beliefs - Prior: initial belief - Likelihood: how much evidence supports/contradicts - Question: Updated belief after seeing evidence?

Key Insights About Bayes

1. **Base Rates Matter**: Even with accurate tests, rare conditions stay rare after positive tests
2. **Rare Events Stay Rare**: If prior is very small (like 0.01), posterior usually won't be huge unless likelihood is extremely strong
3. **Strong Evidence Updates Beliefs**: If $P(B \mid A)$ is much bigger than $P(B \mid A^c)$, seeing B strongly suggests A
4. **Symmetric Information**: Bayes is "reversible" - you can flip conditionals back and forth

Practice Problems

Problem 1: Factory machines - Machine A makes 70% of products, Machine B makes 30% - Machine A: 5% defect rate - Machine B: 10% defect rate - You find a defective product. Probability it came from Machine A?

▶ Solution

Problem 2: Weather prediction - Rain tomorrow: 30% chance (prior) - If rain coming: clouds today 90% of time - If no rain coming: clouds today 40% of time - There are clouds today. Probability of rain tomorrow?

▶ Solution

Memory Aids for Bayes

The Formula in Plain English: > "Probability of cause given effect = (Probability of effect given cause × Prior probability of cause) / Total probability of effect"

Mnemonic: "Please Be Accurate Please" = Prior × Likelihood / Total

Key Things to Remember: - Always use Law of Total Probability to get the denominator - Check the base rate (prior) - it matters a lot! - If prior is low, posterior will be lower than you think - Bayes "flips" $P(B \mid A)$ to $P(A \mid B)$

When You See These Words, Think Bayes: - "Given that [evidence], what's the probability of [cause]?" - "If we observe [result], what's the chance it came from [source]?" - "Update your belief based on new information" - Medical tests, diagnostic problems, source identification

DON'T CONFUSE: $P(A \mid B) \neq P(B \mid A)$ - they're usually very different!

Independence - "Knowing One Tells You Nothing About the Other"

Definition: Events A and B are independent if:

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

Equivalently: $P(A \mid B) = P(A)$ (knowing B doesn't change A's probability)

Plain English: Learning that B happened gives you ZERO information about whether A happened.

Examples: - **Independent**: Flipping a coin twice (first flip doesn't affect second) - **NOT Independent**: Drawing cards without replacement (first card affects what's left) - **Independent**: Height and social security number (no relationship) - **NOT Independent**: Height and weight (taller people tend to weigh more)

For Random Variables: X and Y are independent if knowing the value of one tells you nothing about the probability distribution of the other.

Key Property: If X and Y are independent: - $E[XY] = E[X] \cdot E[Y]$ - $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ (variances ADD)

Memory Aid: Independent = no connection, no relationship, totally separate.

Covariance and Correlation - Measuring Relationship

Covariance - Do They Move Together?

Plain English: Covariance measures whether two variables tend to move in the same direction. - Positive covariance: When X goes up, Y tends to go up - Negative covariance: When X goes up, Y tends to go down - Zero covariance: No linear relationship

Formula:

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

The Second Formula is Easier: $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$

Key Facts: - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ (covariance with yourself is your variance) - If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ (but converse isn't always true!) - Used in the formula: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

Problem with Covariance: Units are weird (if X is in meters and Y is in kilograms, covariance is in meter-kilograms). Hard to interpret scale.

Correlation - Standardized Covariance

Plain English: Correlation is just covariance scaled to always be between -1 and +1. Much easier to interpret!

Formula:

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Read as: "rho = correlation = covariance divided by the product of standard deviations"

Properties: - Always between -1 and +1 - $\rho = 1$: Perfect positive linear relationship (Y = aX + b with a > 0) - $\rho = -1$: Perfect negative linear relationship (Y = aX + b with a < 0) - $\rho = 0$: No LINEAR relationship (but could still have non-linear relationship!)

Interpretation Guide: - $|\rho| > 0.8$: Strong linear relationship - $0.5 < |\rho| < 0.8$: Moderate linear relationship - $|\rho| < 0.5$: Weak linear relationship

IMPORTANT: Correlation only measures LINEAR relationships. Y could equal X^2 (strongly related!) but have correlation near 0.

Memory Aid: Correlation is standardized covariance. Always between -1 and 1. Measures linear relationship strength.

---

6. Monte Carlo Simulations

What Is Monte Carlo?

Plain English: Monte Carlo = using random sampling to solve problems. Instead of solving something with calculus or complex math, you just simulate it a bunch of times and see what happens.

The Name: Named after the Monte Carlo casino in Monaco (because it involves randomness, like gambling).

The Core Idea: If you want to know a probability or expected value, just: 1. Simulate the random process many times (maybe 10,000 times) 2. Count how often the event happens 3. That proportion is your estimate

When to Use: - Calculating probabilities that are hard to find with formulas - Estimating expected values - Computing integrals - Testing if your estimator/test/confidence interval works well

Law of Large Numbers - Why Monte Carlo Works

The LLN (Law of Large Numbers):

Plain English: If you repeat something random many times, the average of your results gets closer and closer to the true expected value.

Formula:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} E[X]$$

Read as: "Sample average approaches true mean as sample size grows"

Example: - True probability of heads = 0.5 - Flip 10 times: might get 0.6 heads (not great) - Flip 100 times: might get 0.52 heads (better) - Flip 10,000 times: might get 0.5003 heads (very close!)

Why This Matters: This is THE reason Monte Carlo works. With enough simulations, your estimates get close to the truth.

Standard Error - How Accurate Is Your Estimate?

Formula:

$$\text{SE}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Plain English: Standard error tells you how accurate your estimate is. It's the standard deviation of your estimator.

Key Insights: - SE decreases as $\sqrt{n}$ (so to cut error in half, need 4× more simulations) - Your estimate is typically within $2 \times \text{SE}$ of the truth (95% confidence) - If SE = 0.01, your estimate is probably within ±0.02 of the truth

Rule of Thumb: - 1,000 simulations: SE ≈ $0.032/\sqrt{1000} \approx 0.01$ (error ≈ 1%) - 10,000 simulations: SE ≈ $0.032/\sqrt{10000} \approx 0.003$ (error ≈ 0.3%)

Memory Aid: To cut error in half, need 4× more simulations. SE = $\frac{\text{SD}}{\sqrt{n}}$

How to Do Monte Carlo

Estimating a Probability

Problem: Find $P(\text{event})$

Algorithm: 1. Simulate the random process n times (like $n = 10{,}000$) 2. Count how many times the event occurred: k 3. Estimate: $\hat{p} = \frac{k}{n}$

Example - Birthday Problem: What's the probability that in a group of 23 people, at least 2 share a birthday?

```
n_sims <- 10000
shared_birthday <- 0
```

```
for(i in 1:n_sims) {
  birthdays <- sample(1:365, size=23, replace=TRUE)
  if(length(unique(birthdays)) < 23) {
    shared_birthday <- shared_birthday + 1
  }
}

prob_estimate <- shared_birthday / n_sims
# Answer: about 0.507 (50.7%)!
```

Estimating an Expected Value

Problem: Find $E[g(X)]$ where X has some distribution

Algorithm: 1. Generate $X_1, X_2, \ldots, X_n$ from the distribution of X 2. Calculate $g(X_1), g(X_2), \ldots, g(X_n)$ 3. Estimate: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$

Example: Estimate $E[X^2]$ where $X \sim \text{Exponential}(2)$

```
n_sims <- 10000
x <- rexp(n_sims, rate=2)
estimate <- mean(x^2)
# True value: Var(X) + (E[X])^2 = 1/4 + 1/4 = 0.5
```

Monte Carlo Integration

Problem: Calculate $\int_a^b g(x)\,dx$

Key Insight: This integral equals $(b-a) \times E[g(U)]$ where $U \sim \text{Uniform}(a, b)$

Algorithm: 1. Generate $U_1, \ldots, U_n \sim \text{Uniform}(a, b)$ 2. Calculate $g(U_1), \ldots, g(U_n)$ 3. Estimate: $\hat{I} = (b-a) \times \frac{1}{n}\sum_{i=1}^{n} g(U_i)$

Example: Estimate π using the unit circle

Area of quarter circle = $\frac{\pi}{4}$

Area of square = 1

So $\pi = 4 \times P(\text{random point in square is in circle})$

```
n_sims <- 100000
x <- runif(n_sims, -1, 1)
y <- runif(n_sims, -1, 1)
in_circle <- (x^2 + y^2 <= 1)
pi_estimate <- 4 * mean(in_circle)
# Estimate ≈ 3.14
```

Memory Aid: Monte Carlo = simulate many times, calculate average. Works because of Law of Large Numbers.

When Monte Carlo Fails

Problem: If expected value doesn't exist, Monte Carlo doesn't work.

Example: Cauchy distribution has no defined mean. If you sample from Cauchy and compute the average, it will never converge - it just bounces around forever.

Lesson: Make sure the thing you're trying to estimate actually exists!

---

7. Hypothesis Testing

The Big Picture - What Are We Doing?

Scenario: You have some data. You want to know if an effect is real or just due to chance.

Examples: - Does this drug actually work, or did patients just get lucky? - Did sales increase because of our marketing, or is it random variation? - Is this coin actually fair, or is it biased?

The Hypothesis Testing Framework: 1. **Null Hypothesis** (H_0): The "boring" hypothesis. No effect, nothing special happening. 2. **Alternative Hypothesis** (H_a): The "interesting" hypothesis. There IS an effect.

Goal: Determine if the data provides enough evidence to reject H_0.

KEY POINT: We never "accept" H_0. We either: - **Reject** H_0: Strong evidence against it - **Fail to reject** H_0: Not enough evidence (doesn't mean H_0 is true!)

The Process (Step by Step)

Step 1: State Your Hypotheses

Example: Testing if a coin is fair

- $H_0 : p = 0.5$ (coin is fair)
- $H_a : p \neq 0.5$ (coin is biased)

Types of Alternative Hypotheses: - **Two-sided**: $H_a : \theta \neq \theta_0$ (could be bigger OR smaller) - **One-sided (upper)**: $H_a : \theta > \theta_0$ (only interested in increases) - **One-sided (lower)**: $H_a : \theta < \theta_0$ (only interested in decreases)

Step 2: Choose a Test Statistic

Test Statistic: A number calculated from your data that measures how far your data is from what H_0 predicts.

Common Test Statistics: - Difference in means: $\bar{X}_1 - \bar{X}_2$ - Sample mean: $\bar{X}$ - Sample proportion: $\hat{p}$ - Standardized version: $\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Example: If testing whether mean = 100, use $T = \bar{X}$ (how far is sample mean from 100?)

Step 3: Calculate the P-value

P-value: THE most important concept in hypothesis testing.

Definition: P-value = probability of seeing data at least as extreme as what we observed, ASSUMING H_0 is true.

$$\text{p-value} = P(\text{observe test statistic this extreme} \mid H_0 \text{ true})$$

Interpretation: - **Small p-value** (< 0.05): Data is very unlikely under H_0, so reject H_0 - **Large p-value** (> 0.05): Data is consistent with H_0, so fail to reject

THE KEY IDEA: If H_0 is true, you should get a "typical" result. If your result is super unlikely under H_0 (small p-value), then maybe H_0 isn't true!

How to Calculate (for different tails): - **Two-sided**: $p = P(|T| \geq |T_{\text{obs}}| \mid H_0)$ - **Upper tail**: $p = P(T \geq T_{\text{obs}} \mid H_0)$ - **Lower tail**: $p = P(T \leq T_{\text{obs}} \mid H_0)$

Memory Aid: Small p-value = surprising result under H_0 = reject H_0

Step 4: Make a Decision

Significance Level: α (alpha) = threshold for rejecting H_0

Common Choices: - $\alpha = 0.05$ (standard, reject if p < 0.05) - $\alpha = 0.01$ (conservative, reject if p < 0.01) - $\alpha = 0.10$ (liberal, reject if p < 0.10)

Decision Rule: - If $\text{p-value} < \alpha$: Reject H_0 (statistically significant) - If $\text{p-value} \geq \alpha$: Fail to reject H_0 (not significant)

What α Means: α is the probability of Type I error (rejecting true H_0). Choosing $\alpha = 0.05$ means "I'm willing to be wrong 5% of the time when H_0 is true."

Common Misunderstandings About P-values

❌ **WRONG**: "P-value is the probability that H_0 is true" ✓ **CORRECT**: "P-value is the probability of seeing data this extreme IF H_0 is true"

❌ **WRONG**: "P-value is the probability I'm making a mistake" ✓ **CORRECT**: "P-value is computed ASSUMING H_0 is true, then seeing how weird the data is"

❌ **WRONG**: "P-value = 0.001 means H_a is 99.9% likely to be true" ✓ **CORRECT**: "P-value = 0.001 means if H_0 were true, there's only a 0.1% chance of seeing data this extreme"

Permutation Tests - The Monte Carlo Way

When to Use: Comparing two groups, don't want to assume a specific distribution.

The Idea: Under H_0, group labels don't matter. So we can randomly shuffle labels and see what test statistics we'd get by chance.

Algorithm: 1. Calculate observed test statistic (e.g., difference in means) 2. Pool all data together 3. Randomly reassign group labels (shuffle!) 4. Calculate test statistic for this shuffled data 5. Repeat steps 3-4 many times (e.g., 10,000) 6. P-value = proportion of shuffled statistics at least as extreme as observed

Example: Testing if treatment group has higher mean than control group

```r
# Observed data
treatment <- c(23, 25, 27, 24, 28)
control <- c(18, 20, 19, 21, 17)

# Observed test statistic
T_obs <- mean(treatment) - mean(control)

# Permutation test
all_data <- c(treatment, control)
n_treatment <- length(treatment)
n_perms <- 10000

T_perm <- replicate(n_perms, {
  shuffled <- sample(all_data)
  new_treatment <- shuffled[1:n_treatment]
  new_control <- shuffled[(n_treatment+1):length(all_data)]
  mean(new_treatment) - mean(new_control)
})

# P-value (two-sided)
p_value <- mean(abs(T_perm) >= abs(T_obs))
```

Why This Works: If treatment has no effect, shuffling labels shouldn't matter - we should see similar differences. If we rarely see differences as big as the observed one, that's evidence the treatment worked!

Type I and Type II Errors

The Truth Table:

	H_0 **True**	H_0 **False**
Reject H_0	Type I Error ❌	Correct ✓
Fail to Reject	Correct ✓	Type II Error ❌

Type I Error (False Positive): - You reject H_0 when it's actually true - You think you found an effect, but it's just random chance - Probability = α (significance level)

Type II Error (False Negative): - You fail to reject H_0 when it's actually false - There's a real effect, but you missed it - Probability = β (depends on true effect size, sample size, etc.)

The Tradeoff: Decreasing α (being more conservative) increases β (more likely to miss real effects).

Power:

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

Plain English: Power = probability of correctly detecting a real effect.

What Affects Power: 1. **Sample size**: Bigger n → more power 2. **Effect size**: Bigger real effect → easier to detect → more power 3. **Significance level**: Bigger α → more power (but more Type I errors) 4. **Variability**: Less noise → easier to see signal → more power

Memory Aid: - Type I = false alarm - Type II = missed detection - Power = ability to detect real effects

---

8. Estimation

Point Estimation - Making Your Best Guess

What Is It: Using data to estimate an unknown parameter (like population mean, proportion, etc.)

Terminology: - **Parameter** (θ): Unknown true value we want to know - **Estimator** ($\hat{\theta}$): Formula/rule for estimating θ from data (random variable) - **Estimate**: Specific number you get from your data

Example: - Parameter: μ (true population mean height) - Estimator: $\bar{X}$ (sample mean) - Estimate: If your data is {65, 68, 70}, then $\bar{x} = 67.67$

Properties of Estimators

Bias - Is It Right on Average?

Definition:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Plain English: If you could repeat your sampling infinitely many times, would your estimator give the right answer on average?

Unbiased: $E[\hat{\theta}] = \theta$ (right on average)

Examples: - $\bar{X}$ is unbiased for μ ✓ - $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ is unbiased for σ^2 ✓ - Sample max is biased for population max (tends to underestimate) ❌

Why Sample Variance Uses $n-1$: Using n gives a biased (too small) estimate because we're measuring deviations from $\bar{X}$ (which is calculated from the same data) rather than from the true μ. The $n-1$ corrects for this.

Memory Aid: Unbiased = right on average (but could be off for any single sample).

Variance - How Much Does It Bounce Around?

Definition: $\mathrm{Var}(\hat{\theta})$ = variance of the estimator

Plain English: How much does the estimator vary from sample to sample?

For Sample Mean:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Key Insight: Variance decreases with sample size! Bigger n → more stable estimate.

Memory Aid: Low variance = consistent estimates across different samples.

Mean Squared Error - Combining Bias and Variance

Definition:

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Plain English: Average squared error of your estimator.

Decomposition:

$$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Variance}$$

The Tradeoff: Sometimes accepting a little bias can greatly reduce variance, giving lower MSE overall.

Example: Estimating population max: - Sample max is biased but low variance - Could use an unbiased estimator but with higher variance - Which is better? Depends on MSE!

Consistency - Does It Get Better with More Data?

Definition: $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \to \theta$ as $n \to \infty$

Plain English: With enough data, the estimator gets arbitrarily close to the true value.

Examples: - $\bar{X}_n$ is consistent for μ (by Law of Large Numbers) - Even biased estimators can be consistent if bias → 0 as n grows

Memory Aid: Consistent = gets it right eventually (with enough data).

Interval Estimation - Giving a Range

The Idea: Instead of one number (point estimate), give a range that likely contains the true parameter.

Confidence Interval: A range $[L, U]$ such that

$$P(L \leq \theta \leq U) = 1 - \alpha$$

Common Confidence Levels: - 90% CI: $\alpha = 0.10$ - 95% CI: $\alpha = 0.05$ (most common) - 99% CI: $\alpha = 0.01$

What Does "95% Confidence" Mean?

CORRECT Interpretation: If you repeated this procedure many times (new samples each time), about 95% of the resulting intervals would contain the true parameter.

WRONG Interpretation: "There's a 95% probability the true parameter is in this specific interval" (The parameter is fixed! It's either in there or not.)

The Procedure Has 95% Confidence: Think of it like a factory making intervals. The factory's success rate is 95%. Any specific interval either worked or didn't (but you don't know which).

Analogy: It's like a basketball player with a 95% free throw percentage. Any specific shot either goes in or doesn't, but in the long run, 95% go in.

Monte Carlo Confidence Intervals

When to Use: You have an estimator but don't know its exact distribution.

Algorithm: 1. From your data, calculate point estimate $\hat{\theta}$ 2. Generate many datasets by simulating from the model with parameter $\hat{\theta}$ 3. For each simulated dataset, calculate the estimator 4. Take the 2.5th and 97.5th percentiles of these simulated values 5. That's your 95% CI

Example - Exponential Rate:

```
# Real data
x <- rexp(50, rate=2)
lambda_hat <- 1 / mean(x)

# Monte Carlo CI
n_sims <- 10000
n <- length(x)

lambda_sim <- replicate(n_sims, {
  x_sim <- rexp(n, rate=lambda_hat)
```

```r
  1 / mean(x_sim)
})

ci <- quantile(lambda_sim, c(0.025, 0.975))
```

Central Limit Theorem - The Magic Behind CIs

The CLT (Central Limit Theorem):

Plain English: No matter what distribution your data comes from, if you average enough observations, that average will be approximately normal.

Formula:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1) \text{ for large } n$$

Why This Is HUGE: Even if your data is skewed, weird, whatever - the sample mean $\bar{X}$ is approximately normal! This lets us use normal distribution methods.

How Large is "Large"?: - Symmetric distributions: $n \geq 20$ usually fine - Moderately skewed: $n \geq 30$ - Very skewed: $n \geq 50$ or $n \geq 100$

Applications: This is why we can make confidence intervals and do tests even when we don't know the underlying distribution!

Standard Confidence Intervals

Z-Interval (Known σ, or Large n)

Formula:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

When to Use: - You somehow know the population standard deviation σ (rare!) - OR sample size is large ($n > 30$) and you use s instead of σ

Critical Values: - 90% CI: $z_{0.05} = 1.645$ - 95% CI: $z_{0.025} = 1.96$ (memorize this one!) - 99% CI: $z_{0.005} = 2.576$

Example: Sample of $n = 36$, $\bar{x} = 100$, $\sigma = 15$, want 95% CI:

$$100 \pm 1.96 \times \frac{15}{\sqrt{36}} = 100 \pm 4.9 = [95.1, 104.9]$$

T-Interval (Unknown σ, Small n)

Formula:

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

When to Use: - Don't know σ (use sample SD s instead) - Sample size is small - Data is approximately normal

The T-Distribution: - Similar to normal but with heavier tails (more uncertainty) - As n increases, t-distribution approaches normal - Use $n-1$ degrees of freedom

Example: Sample of $n = 16$, $\bar{x} = 100$, $s = 15$, want 95% CI: - Look up: $t_{0.025,15} = 2.131$ - CI: $100 \pm 2.131 \times \frac{15}{\sqrt{16}} = 100 \pm 7.99 = [92.0, 108.0]$

Memory Aid: Unknown σ + small n = use t. T has heavier tails = wider interval (more uncertainty).

Proportion Interval

Formula:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

When to Use: Estimating a proportion (like percentage who vote yes, success rate, etc.)

Requirements: $n\hat{p} \geq 5$ AND $n(1-\hat{p}) \geq 5$ (need enough successes and failures)

Example: Out of $n = 100$ people, 60 say yes. Find 95% CI for true proportion: - $\hat{p} = 60/100 = 0.6$ - SE = $\sqrt{\frac{0.6\times0.4}{100}} = 0.049$ - CI: $0.6 \pm 1.96 \times 0.049 = 0.6 \pm 0.096 = [0.504, 0.696]$

Margin of Error

Definition: Margin of error (ME) is half the width of the confidence interval.

For means: $\mathrm{ME} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

For proportions: $\mathrm{ME} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

What Affects Width: 1. **Confidence level**: Higher confidence → wider interval (tradeoff: confidence vs precision) 2. **Sample size**: Larger n → narrower interval (more data → more precise) 3. **Variability**: Larger σ → wider interval (more noise → less precise)

Sample Size Calculation: If you want ME = m, solve for n:

$$n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$$

Example: Want to estimate mean within ±2 units (95% confidence), $\sigma = 10$:

$$n = \left(\frac{1.96 \times 10}{2}\right)^2 = 96.04 \approx 97$$

Memory Aid: Want half the margin of error? Need 4× the sample size!

---

9. R Programming Quick Reference

Probability Distributions in R

The System: `prefix + distribution_name`

Prefixes: - `d` = density/mass (PDF/PMF): gives you $P(X = x)$ or $f(x)$ - `p` = probability (CDF): gives you $P(X \leq x)$ - `q` = quantile (inverse CDF): gives you the value x such that $P(X \leq x) = p$ - `r` = random: generates random values

Common Distributions: - Binomial: `binom(x, size=n, prob=p)` - Geometric: `geom(x, prob=p)` - Poisson: `pois(x, lambda)` - Uniform: `unif(x, min=a, max=b)` - Exponential: `exp(x, rate=lambda)` - Normal: `norm(x, mean=mu, sd=sigma)`

Examples:

```
# Normal distribution
dnorm(0)           # PDF at 0 for standard normal
pnorm(1.96)        # P(Z ≤ 1.96) ≈ 0.975
qnorm(0.975)       # Value where CDF = 0.975 (answer: 1.96)
rnorm(100, 5, 2)   # 100 random values from N(5, 4)

# Binomial distribution
dbinom(3, 10, 0.5)     # P(X = 3) when X ~ Binom(10, 0.5)
pbinom(3, 10, 0.5)     # P(X ≤ 3)
rbinom(1000, 10, 0.5) # 1000 random values

# Exponential distribution
rexp(100, rate=2)      # 100 random values from Exp(2)
pexp(1, rate=2)        # P(X ≤ 1) when X ~ Exp(2)
```

Basic Monte Carlo Template

```
set.seed(123)  # For reproducibility
n_sims <- 10000

# Estimate E[g(X)]
results <- replicate(n_sims, {
```

```r
  # Generate data
  x <- rnorm(50, mean=10, sd=2)

  # Calculate thing you care about
  g_x <- mean(x^2)

  return(g_x)
})

# Analysis
estimate <- mean(results)
se <- sd(results) / sqrt(n_sims)
ci <- quantile(results, c(0.025, 0.975))

cat("Estimate:", estimate, "\n")
cat("95% CI: [", ci[1], ",", ci[2], "]\n")
```

Hypothesis Testing Template

Permutation Test:

```r
set.seed(123)

# Data
group1 <- c(12, 15, 18, 14, 16)
group2 <- c(20, 22, 19, 25, 21)

# Observed test statistic
T_obs <- mean(group2) - mean(group1)

# Permutation test
all_data <- c(group1, group2)
n1 <- length(group1)
n_perms <- 10000

T_perm <- replicate(n_perms, {
  shuffled <- sample(all_data)
  mean(shuffled[(n1+1):length(all_data)]) - mean(shuffled[1:n1])
})

# P-value (two-sided)
p_value <- mean(abs(T_perm) >= abs(T_obs))

# Visualize
hist(T_perm, breaks=50)
abline(v=T_obs, col="red", lwd=2)
```

Parametric Test (testing Poisson mean):

```r
# Observed data
x <- rpois(20, lambda=28)  # Your actual data
x_bar_obs <- mean(x)

# Null hypothesis: λ = 25
lambda0 <- 25
n <- length(x)

# Simulate under H0
x_bar_sim <- replicate(10000, mean(rpois(n, lambda=lambda0)))

# P-value (two-sided)
p_value <- 2 * min(mean(x_bar_sim >= x_bar_obs),
                   mean(x_bar_sim <= x_bar_obs))

cat("P-value:", p_value, "\n")
```

Confidence Interval Templates

Monte Carlo CI:

```r
# Observed data
data <- rexp(50, rate=2)
theta_hat <- 1 / mean(data)  # Point estimate

# Monte Carlo CI
n_sims <- 10000
n <- length(data)

theta_sim <- replicate(n_sims, {
  sim_data <- rexp(n, rate=theta_hat)
  1 / mean(sim_data)
})

# 95% CI
ci <- quantile(theta_sim, c(0.025, 0.975))
cat("95% CI: [", ci[1], ",", ci[2], "]\n")
```

Normal-based CI:

```r
# Sample statistics
x_bar <- mean(data)
s <- sd(data)
n <- length(data)
alpha <- 0.05

# T-interval (unknown σ)
t_crit <- qt(1 - alpha/2, df=n-1)
```

```r
ci <- x_bar + c(-1, 1) * t_crit * s / sqrt(n)

# Proportion interval
p_hat <- mean(binary_data)  # For 0/1 data
z_crit <- qnorm(1 - alpha/2)
ci_prop <- p_hat + c(-1, 1) * z_crit * sqrt(p_hat*(1-p_hat)/n)
```

Useful Functions

Summary statistics:

```r
mean(x)      # Average
median(x)    # Middle value
var(x)       # Variance (n-1 denominator)
sd(x)        # Standard deviation
IQR(x)       # Interquartile range
quantile(x, c(0.25, 0.75))  # Quartiles
summary(x)   # Five-number summary + mean
```

Data manipulation:

```r
sum(x)         # Add everything up
length(x)      # How many elements
unique(x)      # Distinct values
table(x)       # Count each value
sort(x)        # Put in order
sample(x, size=10, replace=TRUE)  # Random sample
rep(5, times=10)   # Repeat value
seq(1, 10, by=0.5) # Sequence
c(x, y)       # Combine vectors
```

Logical operations:

```r
x > 5                 # Which are > 5?
x >= 5 & x <= 10      # AND
x < 3 | x > 7         # OR
!condition            # NOT
sum(x > 5)            # Count how many
mean(x > 5)           # Proportion
which(x > 5)          # Which indices
```

Visualization:

```r
hist(x, breaks=50)           # Histogram
plot(x, y)                   # Scatterplot
barplot(table(x))           # Bar chart
```

```
abline(h=10, col="red")      # Horizontal line
abline(v=5, col="blue")      # Vertical line
```

Quick Formula Reference

Probability: - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ - $P(A^c) = 1 - P(A)$ - $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ - Bayes: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$

Expected Value & Variance: - $E[aX + bY] = aE[X] + bE[Y]$ - $\text{Var}(X) = E[X^2] - (E[X])^2$ - $\text{Var}(aX + b) = a^2\text{Var}(X)$ - $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

Common Distributions ($E[X]$, $\text{Var}(X)$): - Bernoulli(p): p, $p(1-p)$ - Binomial(n, p): np, $np(1-p)$ - Geometric(p): $\frac{1-p}{p}$, $\frac{1-p}{p^2}$ - Poisson(λ): λ, λ - Uniform(a, b): $\frac{a+b}{2}$, $\frac{(b-a)^2}{12}$ - Exponential(λ): $\frac{1}{\lambda}$, $\frac{1}{\lambda^2}$ - Normal (μ, σ^2): μ, σ^2

Confidence Intervals: - Mean (σ known): $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ - Mean (σ unknown): $\bar{X} \pm t_{0.025,n-1}\frac{s}{\sqrt{n}}$ - Proportion: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Standard Error: - $\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Memory Aids and Tricks

68-95-99.7 Rule (Normal Distribution): - 68% within 1 SD - 95% within 2 SD - 99.7% within 3 SD

Critical Values (memorize these): - 90% CI: $z = 1.645$ - 95% CI: $z = 1.96$ ← most common - 99% CI: $z = 2.576$

PMF vs PDF vs CDF: - PMF: Discrete, exact probabilities - PDF: Continuous, need to integrate - CDF: Both, cumulative "up to x"

Bias vs Variance: - Bias: Wrong on average - Variance: Inconsistent across samples - MSE: Combines both

Type I vs Type II: - Type I: False alarm (reject true H_0) - Type II: Missed detection (fail to reject false H_0) - α : P(Type I) - Power: 1 - P(Type II)

P-value: - Small p-value (< 0.05) = reject H_0 - P-value = prob of seeing data this extreme IF H_0 true - NOT the prob that H_0 is true!

Good luck on your exam!