Day 6 Lecture Notes: Probability, Simulation, and Randomness

Mathematical Modeling & Computational Projects Camp

Morning Structure (10:00–12:30)

- 10:00–10:30: Lecture Block 1
- 10:35–11:05: Problem Solving Session 1 (30 minutes)
- 11:10–11:40: Lecture Block 2
- 11:40–12:10: Problem Solving Session 2 (shorter) + 12:10–12:30 synthesis/review

Lecture Block 1 (10:00–10:30): Probability Basics

Probability measures uncertainty from 0 to 1.

$$P(\text{event}) = \frac{\text{favorable outcomes}}{\text{total equally likely outcomes}}.$$

Expected value is the long-run average:

$$E(X) = \sum xP(X = x).$$

Instructor prompt: "If an event has probability 0.25, does that mean it happens exactly once every 4 trials?"

Core reminders:

- Probability is a long-run proportion, not a short-run guarantee.
- Complement rule: $P(\text{not } A) = 1 - P(A)$.
- Expected value is an average over many repetitions.

Mini example (expected value): A game pays \$20 with probability 0.1 and \$0 otherwise.

$$E = 20(0.1) + 0(0.9) = 2.$$

Interpretation: average payout is \$2 per play over many plays.

New Topic: Basic Descriptive Statistics

Descriptive statistics summarize data using a few key numbers.

Mean (arithmetic average):

$$\text{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Variance and standard deviation measure spread (how far values are from the mean):

$$\text{population variance } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2, \qquad \text{population SD } \sigma = \sqrt{\sigma^2}.$$

$$\text{sample variance } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad \text{sample SD } s = \sqrt{s^2}.$$

High-school-friendly distinction:

- Use population formulas when you truly have every value in the group of interest.
- Use sample formulas when your data are only part of a larger group.
- Sample formulas divide by $n-1$ to slightly correct for using an estimated mean.

Worked example 1 (mean): for data $\{2,4,6,8\}$,

$$\bar{x} = \frac{2+4+6+8}{4} = 5.$$

Worked example 2 (spread): for data $\{3,5,7\}$,

$$\bar{x} = 5, \quad \text{squared deviations } (3-5)^2, (5-5)^2, (7-5)^2 = 4, 0, 4.$$

Population variance:

$$\sigma^2 = \frac{4+0+4}{3} = \frac{8}{3} \approx 2.67, \qquad \sigma \approx 1.63.$$

Sample variance:

$$s^2 = \frac{4+0+4}{2} = 4, \qquad s = 2.$$

Interpretation:

- Small variance/SD means values are clustered near the mean.
- Large variance/SD means more variability (more spread/noise).
- In simulations and measurements, noise is natural random fluctuation.

Problem Solving Session 1 (10:35–11:05): Practice Set A (30 minutes)

Goal: Compute exact probabilities, interpret expected value, and summarize data with basic statistics.

1. Roll one die: find $P(\text{even})$.
2. Flip 3 fair coins: find $P(\text{exactly 2 heads})$.
3. For a fair die, find $P(\text{not even})$ using complement rule.
4. Game pays 10 with probability 0.2 and 0 otherwise. Find expected payout.

5. A raffle has 1 winning ticket out of 200 tickets and pays $500. What is expected value of one ticket from payout only?
6. Explain in one or two sentences why expected value may not be an actual outcome.
7. For data $\{4, 6, 8, 12\}$, compute the mean.
8. For data $\{2, 2, 6, 6\}$, compute the population variance and population standard deviation.
9. For sample data $\{5, 7, 9\}$, compute the sample variance and sample standard deviation.

Instructor checkpoint: ask students to separate "chance of winning" from "average value."

Lecture Block 2 (11:10–11:40): Simulation and Monte Carlo

When exact formulas are hard, we simulate many trials and estimate probabilities by frequencies.

Worked example: Simulate many two-dice rolls and estimate

$$P(\text{sum} \geq 10).$$

Larger trial counts usually give more stable estimates.

Exact vs simulated comparison:

Method	$P(\text{sum} \geq 10)$
Exact calculation	$6/36 \approx 0.167$
Simulation (200 trials)	0.180 (example run)
Simulation (5000 trials)	0.169 (example run)

Second simple example (estimating π with random points): draw random points in a 1×1 square and count how many fall in a quarter-circle of radius 1.

$$\pi \approx 4 \cdot \frac{\text{points in quarter-circle}}{\text{total random points}}.$$

As total points increase, this estimate usually stabilizes.

Another real-world simulation example: cafeteria queue modeling can simulate student arrivals and service times to estimate average wait and decide how many serving lines are needed.

Randomness, sampling, and convergence:

- Randomness means each trial can differ, so short runs can bounce around.
- Sampling means we observe only a finite number of trials from a random process.
- Convergence means estimates often move closer to a stable value as trial count grows.

High-level simulation error notes:

- **Random error:** finite samples create natural wiggle in estimates.
- **Model error:** wrong assumptions can shift results in a biased direction.
- More trials reduce random error, but do not automatically fix model error.

Simulation workflow for students:

1. Define random process and event.
2. Run many trials (e.g., 1000, then 10000).
3. Count event frequency.
4. Estimate probability as count/trials.
5. Compare estimates across trial sizes.

Important interpretation:

- Simulation output is an estimate, not exact truth.
- More trials usually reduce random noise.
- Poor assumptions still give poor estimates, even with many trials.

Bias warning: if a model assumes unrealistic arrival rates, wrong probabilities, or missing groups, running more trials only repeats the same bias more precisely; it does not fix the model design.

Transition: From Mean to Least Squares

In Session 1, mean was an arithmetic average. It also has a fitting meaning: the mean is the number m that minimizes the sum of squared errors

$$S(m) = \sum_{i=1}^{n} (x_i - m)^2.$$

So the same idea that summarizes noisy data also leads to model fitting. For one number, best fit is the mean; for a trend line, we use least squares.

New Topic: Linear Least Squares

Suppose data are pairs (x_i, y_i) and we want a line

$$y \approx ax + b.$$

Residual for point i is

$$r_i = y_i - (ax_i + b).$$

Least squares chooses a, b to minimize

$$S(a, b) = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2.$$

At a high-school algebra level, the minimizing values satisfy the normal equations:

$$\sum y_i = a \sum x_i + nb,$$

$$\sum x_i y_i = a \sum x_i^2 + b \sum x_i.$$

These are two linear equations in unknowns a and b.

Worked example: use points $(1,2),(2,3),(3,5)$. Then

$$n=3, \quad \sum x_i = 6, \quad \sum y_i = 10, \quad \sum x_i^2 = 14, \quad \sum x_i y_i = 23.$$

Normal equations become

$$10 = 6a + 3b, \qquad 23 = 14a + 6b.$$

Solve:

$$a = 1.5, \qquad b = \frac{1}{3}.$$

Best-fit line:

$$\hat{y} = 1.5x + \frac{1}{3}.$$

Geometric interpretation: the best-fit line is the line that makes the vertical distances (residuals) from points to line as small as possible overall in the squared-error sense.

Problem Solving Session 2 (11:40–12:10): Practice Set B (shorter)

Goal: Design simulation plans, evaluate simulation quality, and connect to linear least squares.

1. Design a simulation to estimate P(at least one head in 4 flips) and list the trial steps clearly.
2. You run 200 trials and observe event count 78. What is your estimate?
3. You run 5000 trials and observe event count 1970. What is your estimate? Compare with Problem 2.
4. Briefly explain why two simulation runs (for the same event) can give 0.49 and 0.52.
5. Give one process where simulation is better than hand counting, and explain why.
6. For points $(1,2),(2,3),(3,5)$, verify that $a = 1.5$ and $b = 1/3$ satisfy both normal equations.
7. For line $\hat{y} = 1.5x + 1/3$, interpret slope and intercept in context of these data.
8. Compare total squared residuals for lines $\hat{y} = x+1$ and $\hat{y} = 1.5x+1/3$ on points $(1,2),(2,3),(3,5)$. Which fits better?

Instructor move for fast finishers: have teams design two different simulation methods for the same event and compare strengths.

Synthesis and Review (12:10–12:30)

- Compare exact probability and simulated estimate.
- Revisit terms: random variable, expected value, mean, variance, standard deviation, Monte Carlo, least squares.
- Exit check: "How does number of trials affect reliability?"
- Exit check 2: "Why can expected value be non-integer while outcomes are integers?"
- Exit check 3: "What is one assumption that could invalidate your simulation?"
- Exit check 4: "Why does minimizing squared residuals produce a best-fit line?"

Python Preview (Afternoon)

Use NumPy random generators for simulation and Matplotlib for histogram summaries.