Applied Probability and Statistics I

STAT400 / DATA400 — Lecture Notes

Shixin Zheng
University of Maryland

Fall 2026

These notes follow the course as taught in Fall 2025 and Spring 2026. The reference text is Devore, *Probability and Statistics for Engineering and the Sciences*, 9th edition; section numbers in the lecture headers point there.

Contents

Chapter 1

Introduction: Five Questions

Lecture 1 | Devore 9e: Ch. 1 (overview)

The whole course is previewed by five questions. None of them can be answered today. Each one names the tool that will answer it, and the lecture where that tool arrives.

1.1 Question 1: Poker

Five cards are dealt from a standard 52-card deck. The hands, from best to worst, begin

$$\text{royal flush} \;>\; \text{straight flush} \;>\; \text{four of a kind} \;>\; \text{full house} \;>\; \text{flush} \;>\; \text{straight} \;>\; \cdots$$

Why is that the ranking? Why is a flush better than a straight?

The ranking is not a convention — it is a *fact*, forced by how rare each hand is. To see that, we must be able to count how many five-card hands of each type exist. Counting well is harder than it looks.

→ **Lecture 3** (counting and combinatorics), where we compute all of these.

1.2 Question 2: How Buffon computed π

Rule a floor with parallel lines a distance a apart, and drop a needle of length $\ell < a$ on it N times. Let n be the number of times the needle crosses a line. Then — as we will prove —

$$\frac{n}{N} \approx \frac{2\ell}{\pi a} \qquad \Longrightarrow \qquad \pi \approx \frac{2\ell N}{a\, n}.$$

You can compute π by dropping a needle on the floor. Why does that work?

Note that the sample space here is not a finite list of outcomes — the needle can land in a continuum of positions and angles. Counting is useless; we need to measure *area* instead.

→ **Lecture 4** (geometric probability), where this is derived.

1.3 Question 3: The flu test

A test for the flu is

- correct 99% of the time on people who *have* the flu;
- correct 95% of the time on people who *do not*.

You take the test and it comes back positive.

What is the probability that you have the flu?

Most people answer "about 99%". That answer is not just wrong, it is not even well posed: **the question cannot be answered from the information given.** You also need to know how common the flu is right now. If the flu is rare, a positive result may well mean less than a coin flip.

→ **Lectures 5–7** (conditional probability and Bayes' theorem).

1.4 Question 4: Two ways to summarize an experiment

(a) Toss a coin three times. The outcomes are $HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$ — eight of them. But usually we do not care *which* sequence occurred, only how many heads it contained. Let X be that number. Then X takes the values $0, 1, 2, 3$ with probabilities

$$\frac{1}{8}, \quad \frac{3}{8}, \quad \frac{3}{8}, \quad \frac{1}{8}.$$

Eight outcomes have been compressed into four numbers.

(b) Wait for a bus. Let X be the number of minutes you wait. Now X can be 3, or 3.5, or $3.14159\ldots$ — any real number in an interval. The probability of waiting *exactly* 3 minutes, to infinite precision, is zero. So the list-the-values-and-their-probabilities approach breaks down completely.

What is the right way to describe a numerical quantity that is random?

→ **Lectures 7–8** for case (a), **Lecture 13** for case (b).

1.5 Question 5: What score do you need?

Suppose the STAT400 final has mean 88 and variance 15.

You want to finish in the top 10% *of the class. What score should you aim for?*

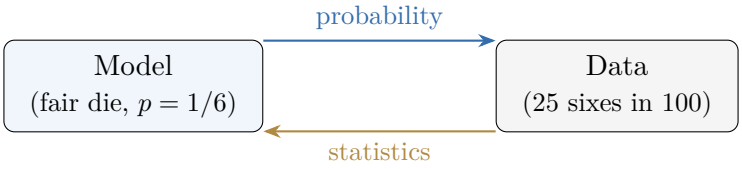
The mean and the variance alone are not enough — you also need to know the *shape* of the distribution of scores. Once we know that shape is the normal curve, the answer is a short computation:

$$\sigma = \sqrt{15} \approx 3.87, \qquad x = \mu + z_{0.90}\,\sigma \approx 88 + (1.28)(3.87) \approx 93.$$

→ **Lecture 14** (the normal distribution).

1.6 The shape of the course

Two of those five questions (1 and 2) are about *probability*: the mechanism is known, and we ask what the data will look like. The last part of the course reverses the arrow — the data is observed and the mechanism is unknown.



Lectures	Block	Content
2–7	Probability	sets, counting, axioms, conditional probability, independence, Bayes
8–12	Random variables	pmf, cdf, expectation, variance; the named discrete families
13–17	Continuous rv's	pdf, transformations; normal, exponential, gamma, Weibull, beta
18–21	Joint behavior	joint distributions, covariance, correlation, CLT
22–26	Statistics	estimators, bias, method of moments, maximum likelihood

Remark 1.1 (Notation)**.** S is the sample space, ω a single outcome, A, B, E events, $\mathbb{P}$ the probability. Capital letters near the end of the alphabet (X, Y, Z) are random variables; the matching lower case (x, y, z) are the values they take. In lecture I often write ξ (xi) for a random variable; it means the same as X.

Chapter 2

Basic Set Theory, Experiments and Events

Lecture 2 | Devore 9e: 2.1

Today there is still *no probability*. We only build the language: a random experiment produces outcomes, outcomes form a set, and events are subsets of that set. Every probability statement later in the course is a statement about sets, so it pays to be fluent here first.

2.1 Sets

Definition 2.1 (Set)**.** A **set** is a collection of *distinct* objects. The objects are its **elements**; we write $x \in A$ if x is an element of A, and $x \notin A$ otherwise.

Example 2.2. $\{\text{“apple”}, \text{“orange”}, \text{“car”}\}$, $\{0, 1, \pi, \frac{1}{2}\}$, $\mathbb{N} = \{0, 1, 2, \dots\}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$. A set can also be described by a property rather than listed:

$$\big\{(x, y) : 0 \le x \le 1,\ 0 \le y \le 1,\ x, y \in \mathbb{R}\big\}$$

is the unit square.

Definition 2.3 (Empty set)**.** The **empty set** $\emptyset$ is the set with nothing in it.

Remark 2.4. $\{\emptyset\} \neq \emptyset$. The left side is a set with one element (that element happening to be the empty set); the right side has no elements. A bag containing an empty bag is not an empty bag.

2.1.1 Relations and operations

Throughout, A and B are sets, and all of them sit inside a fixed universal set $\mathcal{U}$.

Definition 2.5 (Subset)**.** A is a **subset** of B, written $A \subset B$, if $\forall x \in A$ we have $x \in B$.

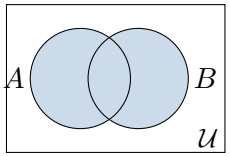
Remark 2.6. Some books write $A \subseteq B$ for this and reserve $A \subset B$ for a *proper* subset. In this course $A \subset B$ always allows $A = B$.

Definition 2.7 (Equality)**.** If $A \subset B$ and $B \subset A$, then $A = B$.

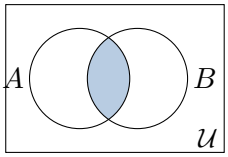
This is the standard way to prove two sets are equal: show each contains the other.

Definition 2.8 (Complement, union, intersection, difference)**.**

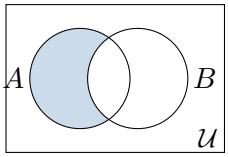
$$\begin{aligned} A^c = \overline{A} &:= \{x : x \notin A\} && \text{(complement)} \\ A \cup B &:= \{x : x \in A \text{ or } x \in B\} && \text{(union)} \\ A \cap B &:= \{x : x \in A \text{ and } x \in B\} && \text{(intersection)} \\ A - B = A \setminus B &:= \{x : x \in A \text{ and } x \notin B\} = A \cap B^c && \text{(difference)} \end{aligned}$$



$A \cup B$



$A \cap B$



$A - B$

Theorem 2.9 (De Morgan's rules)**.**

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c.$$

In words: "not (A or B)" means "not A and not B". These are the identities that let us convert a hard union into an easy intersection, and we will use them constantly ("at least one" problems are almost always done by complementing to "none").

2.1.2 Cardinality

Definition 2.10 (Cardinality)**.** $|A|$ is the number of elements of A.

The three-way split below is exactly what forces the three definitions of probability in Lectures 3–4.

$\lvert A\rvert < \infty$	**finite**	$\{1, 2, 3, 4, 5, 6\}$
$\lvert A\rvert = \infty$	**countably** infinite	$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$
	uncountably infinite	$(0, 1)$, $\mathbb{R}$, the unit square

Remark 2.11. Countable means "can be listed as a sequence $a_1, a_2, a_3, \ldots$". Surprisingly $|\mathbb{Q}| = |\mathbb{Z}|$ — the rationals can be listed — while $(0, 1)$ cannot. A number like $\pi/10 \in (0, 1)$ is not on anybody's list.

2.2 Collections of sets

Definition 2.12 (Collection of sets)**.** Let Γ be an **index set**. A **collection** of sets indexed by Γ is written $\{A_\alpha\}_{\alpha\in\Gamma}$.

Example 2.13.

- $\Gamma = \mathbb{N}$: $\{A_i\}_{i\in\mathbb{N}} = \{A_1, A_2, A_3, \dots\}$.
- $\Gamma = (0,1)$ and $A_\alpha = (-\alpha, \alpha)$: an uncountable collection of open intervals.

Definition 2.14 (Countable union and intersection)**.** For $A_1, A_2, A_3, \dots$,

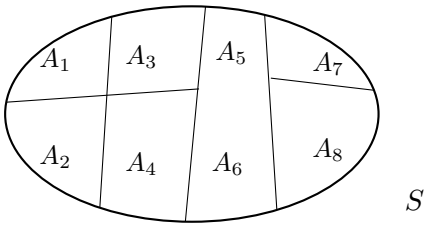
$$\bigcup_{i\in\mathbb{N}} A_i = \{x : x \in A_i \text{ for } \textit{some } i\}, \qquad \bigcap_{i\in\mathbb{N}} A_i = \{x : x \in A_i \text{ for } \textit{every } i\}.$$

Definition 2.15 (Pairwise disjoint)**.** $\{A_i\}_{i\in\mathbb{N}}$ is **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 2.16 (Partition)**.** $\{A_i\}_{i\in\mathbb{N}}$ is a **partition** of a set S if

1. the A_i are pairwise disjoint, and
2. $\bigcup_{i\in\mathbb{N}} A_i = S$.

A partition cuts S into non-overlapping pieces that use up all of S. This is the single most useful picture in the course — the law of total probability (Lecture 6) is nothing but "compute on each piece of a partition, then add up".



2.3 Experiments, outcomes, events

Definition 2.17 (Experiment)**.** An **experiment** is a repeatable task with well-defined outcomes. Individual outcomes are denoted $\omega_1, \omega_2, \dots$, and the set of all of them is the **sample space** S.

Definition 2.18 (Event)**.** Any subset $E \subset S$ is an **event**. We say E **happens** if the outcome of the experiment lies in E. An event with exactly one outcome in it is a **simple event**.

Example 2.19 (Finite sample space)**.** Roll a six-sided die once: $S = \{1, 2, 3, 4, 5, 6\}$.

- "You got an even number": $E = \{2, 4, 6\}$.
- "You got the number 7": $E = \emptyset$.

Roll a die, then toss a coin:

$$S = \{(1, H), (1, T), (2, H), (2, T), \ldots, (6, H), (6, T)\}, \qquad |S| = 12.$$

Example 2.20 (Countably infinite sample space)**.** Record the number of whole minutes until the first call arrives at a call center:

$$S = \{0, 1, 2, 3, \ldots\}.$$

Example 2.21 (Uncountable sample space)**.** Pick a point at random from the unit square:

$$S = [0, 1] \times [0, 1] = \{(x, y) : 0 \le x \le 1,\ 0 \le y \le 1\}.$$

The event "the point is within distance 1 of the origin" is $E = \{(x, y) \in S : x^2 + y^2 \le 1\}$, a quarter disc.

Example 2.22 (Uber Eats)**.** You order fries and a pizza. Let x be the delivery time of the fries in minutes and y that of the pizza, so an outcome is a pair and

$$S = \{(x, y) : x \ge 0,\ y \ge 0\}.$$

Now translate. Let

$A = \{(x, y) \in S : x \le 10\}$ (fries arrive within 10 min), $B = \{(x, y) \in S : y \le 15\}$ (pizza within 15 min).

Then

$A \cap B$	the whole meal is ready within 15 minutes
$A \cup B$	at least one of the two arrives on time
$A - B$	the fries arrive but you wait past 15 minutes for the pizza
$(A \cup B)^c$	neither arrives on time

Read the third row carefully: $A - B$ is the event that you have fries and nothing else for at least five minutes. Getting from an English sentence to the right set is the skill this lecture is for.

2.4 The dictionary

This table is the whole point of the lecture. Read it left to right when you are formalizing a word problem, and right to left when you are interpreting a formula.

Set notation	**Colloquial description**
S	sure event (something happens)
$\emptyset$	impossible event (nothing happens)
A	event A happens
A^c	A does not happen
$A \cup B$	A *or* B (or both)
$A \cap B$	A *and* B
$A \setminus B$	A happens but B does not
$A \cap B = \emptyset$	A and B are mutually exclusive
$\bigcup_{i \in \mathbb{N}} A_i$	at least one of the events happens
$\bigcap_{i \in \mathbb{N}} A_i$	all of the events happen

Remark 2.23. Note what is *not* on this page: any number between 0 and 1. We have a language for events but no way yet to say how likely one is. That starts next lecture.

Chapter 3

Classical Probability, Counting and Combinatorics

Lecture 3 | Devore 9e: 2.2–2.3

Last time: set theory, experiments, events — and no probability anywhere. Today we assign numbers to events for the first time, in the oldest and simplest way.

3.1 Classical (Laplace) probability

Roll a fair die. $S = \{1, 2, 3, 4, 5, 6\}$. What is "the probability" here? We assign each outcome a number, and symmetry says they should all get the same one:

$$\mathbb{P}(1) = \mathbb{P}(2) = \cdots = \mathbb{P}(6) = \tfrac{1}{6}.$$

Definition 3.1 (Classical probability)**.** Let $S = \{\omega_1, \ldots, \omega_n\}$ be *finite* with all outcomes equally likely, $\mathbb{P}(\omega_1) = \cdots = \mathbb{P}(\omega_n) = \frac{1}{n}$. For an event $A \subset S$ containing m outcomes,

$$\boxed{\mathbb{P}(A) = \frac{m}{n} = \frac{|A|}{|S|}}$$

The triple $(S, \text{"events"}, \mathbb{P})$ — where "events" means all subsets of S — is our first probability model. Two things to notice:

- **Computing $\mathbb{P}$ is exactly counting.** There is no probability theory left in the formula; there is only m and n.
- **Counting can be very tricky.** That is the real content of this lecture, and of most exam problems in the first month.

3.2 Two fundamental principles of counting

Theorem 3.2 (Addition principle)**.** *If a task can be done in m ways or in n ways, and the two groups of ways do not overlap, the total is $m + n$.*

Theorem 3.3 (Multiplication principle)**.** *If a task is done in two steps, the first in m ways and the second in n ways, the total is $m \times n$. With k steps, multiply all k counts.*

Example 3.4 (Dining halls — addition)**.** Three dining halls on campus offer 10, 8 and 6 entrées. If you will eat at exactly one of them, you have

$$10 + 8 + 6 = 24$$

choices. "Or" means add — and note this is only correct because no entrée is counted twice.

Example 3.5 (Passcode — multiplication)**.** A 4-digit phone passcode is built by choosing each digit in turn, 10 ways each:

$$10 \times 10 \times 10 \times 10 = 10^4 = 10{,}000.$$

"And then" means multiply.

Remark 3.6. "Or" $\Rightarrow$ add; "and then" $\Rightarrow$ multiply. Most counting errors come from using the wrong one, or from steps that are not independent of each other (the count at step 2 must not depend on *which* choice was made at step 1 — only on how many there were).

3.3 Permutations and combinations

The organizing question: *you have n distinct objects and select k of them. How many ways?* The answer depends on two yes/no questions:

Does **order** matter?
Is the selection **with replacement**?

That is four cases. We do all four.

3.3.1 Case 1: order matters, without replacement

Build the selection in k steps:

$$\begin{aligned} &\text{step 1:} && n \text{ choices} \\ &\text{step 2:} && n-1 \text{ choices} \\ &\vdots \\ &\text{step } k\text{:} && n-k+1 = n-(k-1) \text{ choices} \end{aligned}$$

Definition 3.7 (Permutation)**.** The number of ways to arrange k objects out of n distinct

objects, order mattering, without replacement, is

$$A_n^k := n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!},$$

read "*n permute k*".

Taking $k = n$ gives $A_n^n = n!$: the number of ways to order all n objects.

3.3.2 Case 2: order does not matter, without replacement

Here is the trick that makes this case free. Every unordered selection of k objects can be ordered in $A_k^k = k!$ ways, so

$$\underbrace{A_n^k}_{\text{order matters}} = \underbrace{(\#\text{ order does not matter})}_{\text{what we want}} \times k!.$$

Definition 3.8 (Combination)**.**

$$C_n^k = \binom{n}{k} := \frac{A_n^k}{k!} = \frac{n!}{k!\,(n-k)!},$$

read "*n choose k*". Equivalently $A_n^k = \binom{n}{k} \times k!$.

Remark 3.9 (Two notations)**.** I write C_n^k on the board and $\binom{n}{k}$ in typed material; they are the same number. Likewise A_n^k is often written $P(n,k)$ or nPk elsewhere.

Example 3.10. Take $n = 4$ objects $\{1,2,3,4\}$ and $k = 2$. Ordered: $A_4^2 = 12$ pairs $(1,2),(2,1),(1,3),(3,1),\ldots$ Unordered: $\binom{4}{2} = 6$ sets $\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$. And indeed $12 = 6 \times 2!$.

Example 3.11 (Candy — the relation made concrete)**.** There are 10 different candies.

- *Pick* 3 *to eat for dessert.* You will eat all three, so the order does not matter: {red, green, blue} is the same dessert as {green, red, blue}. That is $\binom{10}{3} = 120$ ways.
- *Pick* 3 *and hand them to Alice, Bob and Carol.* Now it matters who gets which. Each of the 120 selections can be handed out in $3! = 6$ ways, so

$$\binom{10}{3} \times 3! = 120 \times 6 = 720 = A_{10}^3.$$

The second bullet *is* the identity $A_n^k = \binom{n}{k}\, k!$: choosing an ordered list is the same as choosing the set and then ordering it.

3.3.3 Case 3: order matters, with replacement

Each of the k steps now has all n objects available again:

$$\underbrace{n \times n \times \cdots \times n}_{k} = n^k.$$

3.3.4 Case 4: order does not matter, with replacement

This one is genuinely tricky, and the standard trick is worth knowing.

Since order does not matter, all that is recorded is *how many times each object was selected*. Say object i was chosen x_i times. Then we need the number of integer solutions of

$$\sum_{i=1}^{n} x_i = k, \qquad x_i \geq 0.$$

The balls-and-bars model. Draw the k selections as k balls in a row. To split them into n groups we need $n-1$ bars:

$$\underbrace{\circ\circ}_{x_1} \mid \underbrace{\circ\circ}_{x_2} \mid \underbrace{\circ}_{x_3} \mid \underbrace{\circ\circ\circ}_{x_4}$$

There are $n+k-1$ positions in total (balls plus bars), and choosing which $n-1$ of them are bars determines everything:

$$\boxed{\binom{n+k-1}{n-1}}$$

3.3.5 The summary table

	without replacement	**with replacement**
order matters	$A_n^k = \dfrac{n!}{(n-k)!}$	n^k
order does not matter	$\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$	$\binom{n+k-1}{n-1}$

Remark 3.12. Memorize the table, but more importantly memorize the two *questions*. On an exam the hard part is never the formula — it is deciding which box you are in.

3.4 Poker: answering Lecture 1's question

A five-card hand is dealt from a standard 52-card deck (13 ranks $\times$ 4 suits). Order does not matter and there is no replacement, so we are in the lower-left box of the table:

$$|S| = \binom{52}{5} = 2{,}598{,}960.$$

Every hand is equally likely, so classical probability applies and each computation below is just "count the favorable hands".

Royal flush — $10, J, Q, K, A$ all in one suit. There is exactly one such hand per suit, so 4 hands:

$$\mathbb{P} = \frac{4}{2{,}598{,}960} = 0.000154\%.$$

Straight flush — five consecutive ranks, all one suit. A straight is determined by its lowest card, which can be $A, 2, 3, \ldots, 10$ (with A playing low in $A2345$ and high in $10JQKA$), so there are 10 possible runs per suit and $4 \times 10 = 40$ straight flushes. Excluding the 4 royal flushes, which are already ranked above:

$$\mathbb{P} = \frac{40 - 4}{2{,}598{,}960} = \frac{36}{2{,}598{,}960} = 0.001385\%.$$

Four of a kind — build it in two steps. Choose the rank that appears four times (13 ways); all four of its cards are then forced. Choose the remaining fifth card from the other 48 cards (48 ways):

$$\mathbb{P} = \frac{13 \times 48}{2{,}598{,}960} = \frac{624}{2{,}598{,}960} = 0.024010\%.$$

Full house — three of one rank plus two of another. Four steps, multiplied:

$$\underbrace{13}_{\text{rank of the triple}} \times \underbrace{\binom{4}{3}}_{\text{which 3 suits}} \times \underbrace{12}_{\text{rank of the pair}} \times \underbrace{\binom{4}{2}}_{\text{which 2 suits}} = 13 \times 4 \times 12 \times 6 = 3{,}744,$$

$$\mathbb{P} = \frac{3{,}744}{2{,}598{,}960} = 0.144058\%.$$

Note the 12, not 13: the pair must be a *different* rank from the triple.

Flush — all five cards one suit. Choose the suit (4) and then 5 of its 13 cards, $\binom{13}{5} = 1287$. That gives $4 \times 1287 = 5{,}148$, but this count includes the 40 straight flushes, which rank higher:

$$\mathbb{P} = \frac{5{,}148 - 40}{2{,}598{,}960} = \frac{5{,}108}{2{,}598{,}960} = 0.196540\%.$$

Straight — five consecutive ranks, suits arbitrary. Choose the run (10 ways), then a suit for each of the five cards independently ($4^5 = 1024$), giving 10,240; again remove the 40 straight flushes:

$$\mathbb{P} = \frac{10{,}240 - 40}{2{,}598{,}960} = \frac{10{,}200}{2{,}598{,}960} = 0.392465\%.$$

Hand	**Count**	**Probability**
Royal flush	4	0.000154%
Straight flush	36	0.001385%
Four of a kind	624	0.024010%
Full house	3,744	0.144058%
Flush	5,108	0.196540%
Straight	10,200	0.392465%

Remark 3.13. There is the answer to Question 1 of Lecture 1. The ranking is not a convention: each hand beats the one below it because it is strictly rarer. A flush outranks a straight because there are 5,108 flushes and 10,200 straights — a straight is almost exactly twice as likely.

Chapter 4

Geometric Probability and the Modern Definition

Lecture 4 | Devore 9e: 2.1–2.2

Last time: classical probability, where S is finite and all outcomes are equally likely, so $\mathbb{P}(A) = |A|/|S|$ and everything reduces to counting.

Today: what happens when S is uncountably infinite, so that $\mathbb{P}(\omega_i) = \mathbb{P}(\omega_j) =$ "$1/\infty$"? Counting is useless. We patch it, watch the patch fail, and then rebuild from scratch.

4.1 Geometric probability

Suppose S has a nice geometry — an interval, a rectangle, a disc. Then measure size by length-/area/volume instead of by counting:

Definition 4.1 (Geometric probability)**.** For $A \subset S$,

$$\mathbb{P}(A) := \frac{\text{Area}(A)}{\text{Area}(S)}$$

(with length or volume in place of area as the dimension requires).

4.1.1 A rendezvous

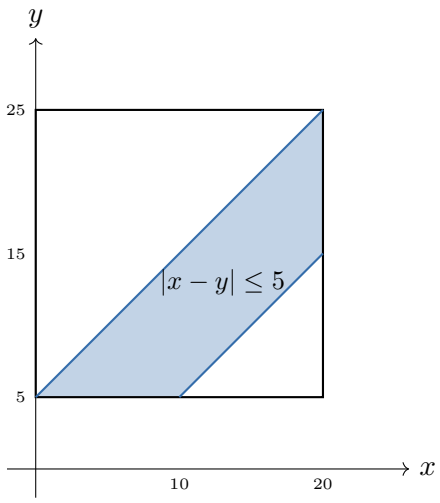
Example 4.2 (Will we meet?)**.** You arrive at the meeting point at a time uniformly distributed between 11:50 and 12:10; the person you are meeting arrives uniformly between 11:55 and 12:15, independently. Whoever arrives first waits 5 minutes and then leaves. What is the probability that you meet?

Measure time in minutes after 11:50 and let x be your arrival, y theirs. Then

$$S = \{(x, y) : 0 \le x \le 20,\ 5 \le y \le 25\},$$

a 20×20 square, so $\text{Area}(S) = 400$. You meet exactly when neither has left before the other arrives, i.e. when

$$A = \{(x, y) \in S : |x - y| \le 5\} = \{(x, y) \in S : y - 5 \le x \le y + 5\}.$$



The complement is easier to measure: it is two right triangles. Above the band, $y > x + 5$ cuts off the triangle with corners $(0, 5), (0, 25), (20, 25)$ — the line $y = x + 5$ runs from one corner of the square to the opposite one, so both legs are 20. Below the band, $y < x - 5$ cuts off the triangle with corners $(10, 5), (20, 5), (20, 15)$, whose legs are both 10. Hence

$$\text{Area}(A) = 400 - \tfrac{1}{2}(20)^2 - \tfrac{1}{2}(10)^2 = 400 - 200 - 50 = 150,$$

$$\mathbb{P}(A) = \frac{150}{400} = \boxed{0.375}.$$

The asymmetry is real: their window is shifted five minutes later than yours, so the two triangles are *not* the same size, and the meeting region is a trapezoid rather than the symmetric hexagon you would get from identical windows.

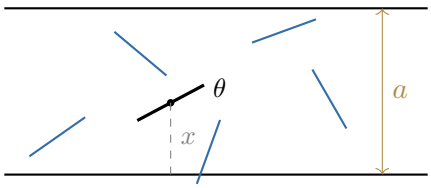
Remark 4.3. Two things to notice. First, the answer came from *areas*, not counting — there are infinitely many arrival times. Second, we computed the complement because two triangles are easier than the four-sided region itself; that move is worth remembering.

4.1.2 Buffon's needle

Example 4.4 (Buffon, 1777)**.** The plane is ruled with parallel lines a distance a apart. A needle of length $\ell < a$ is dropped at random. What is the probability it crosses a line?

Parametrize a landing by two numbers: x, the distance from the needle's center to the nearest line, and θ, the angle the needle makes with the lines. Then

$$S = \left\{(x, \theta) \ : \ 0 \le \theta \le \pi, \ \ 0 \le x \le \tfrac{a}{2}\right\}, \qquad \text{Area}(S) = \frac{a}{2} \cdot \pi.$$



The needle crosses a line exactly when its half-projection onto the perpendicular reaches the line, i.e.

$$A = \Big\{(x, \theta) \ : \ x \le \tfrac{\ell}{2} \sin\theta\Big\}.$$

Therefore

$$\mathbb{P}(A) = \frac{\int_0^{\pi} \frac{\ell}{2} \sin\theta \, \mathrm{d}\theta}{\frac{a}{2} \cdot \pi} = \frac{\frac{\ell}{2} \cdot 2}{\frac{a}{2}\pi} = \boxed{\frac{2\ell}{\pi a}}.$$

Remark 4.5. Solving for π turns this into an experiment: drop a needle many times, and $\pi \approx 2\ell N/(a\, N_{\text{cross}})$. This is arguably the first Monte Carlo method.

4.1.3 Bertrand's paradox — where "at random" breaks

Example 4.6 (Bertrand, 1889)**.** Choose a chord of the unit circle *at random*. What is the probability that its length exceeds the side of the inscribed equilateral triangle (which is $\sqrt{3}$)?

Three perfectly reasonable readings of "at random":

How the chord is chosen	**Answer**
Fix one endpoint, pick the other endpoint uniformly on the circle. The chord is long when the second endpoint is in the far arc, which is 1/3 of the circumference.	$\frac{1}{3}$
Pick the chord's *midpoint* distance from the center uniformly on $[0, 1]$. The chord is long when that distance is $< \frac{1}{2}$.	$\frac{1}{2}$
Pick the chord's *midpoint* uniformly in the disc. The chord is long when the midpoint lands in the disc of radius $\frac{1}{2}$, of area $\pi/4$ out of π.	$\frac{1}{4}$

All three are correct. The phrase "at random" does not specify a probability model, and geometric probability silently assumed one.

4.1.4 So where are we

	requires	
Classical $\mathbb{P}$	S finite,	$\mathbb{P}(\omega_i) = \mathbb{P}(\omega_j)$
Geometric $\mathbb{P}$	S a nice shape,	$\mathbb{P}(\omega_i) = \mathbb{P}(\omega_j)$

Limitations. Neither handles

- uneven probabilities across S (a loaded die),
- countably infinite S,
- uncountable S with a complicated shape.

We need a definition that never asks where the numbers came from.

4.2 The modern definition

The change of viewpoint: stop assigning probability to *outcomes* and start assigning it to *events*. So $\mathbb{P}$ becomes a function whose input is a set.

$$\text{experiment} \Rightarrow \text{outcomes } \omega_i \Rightarrow S \Rightarrow \text{events (subsets of } S) \Rightarrow \mathbb{P} \text{ a function on events.}$$

But which subsets are allowed to be events? The natural first answer is "all of them", i.e. the power set $\mathcal{P}(S)$ (also written 2^S). For a finite or countable S that works. For an uncountable S it does not: one can construct subsets of $[0,1]$ so pathological that *no* sensible assignment of length to them exists.

So we retreat, deliberately: **do not try to define $\mathbb{P}$ on every subset — choose a modest family of subsets and define $\mathbb{P}$ only there.** The family must be closed under the operations we actually perform on events (complement, and countable unions), or the theory would be useless. That requirement, and nothing more, is the next definition.

Definition 4.7 (σ-algebra)**.** Let $\mathcal{F}$ be a collection of subsets of S such that

(i) $S \in \mathcal{F}$;

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;

(iii) $A_i \in \mathcal{F}$ for $i = 1, 2, \cdots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Then $\mathcal{F}$ is a **σ-algebra** (or *event field*), and its elements are called **events**.

Remark 4.8. Note (iii) uses *countable* unions — this is exactly why Lecture 2 spent time on countable versus uncountable. Also note that $\mathcal{F}$ is a set *of sets*; $\mathcal{F} = \emptyset \cup S$ is a type error, the correct smallest example is $\mathcal{F} = \{\emptyset, S\}$.

Example 4.9.

- $\mathcal{F} = \{\emptyset, S\}$ — the smallest σ-algebra.
- For a fixed $A \subset S$: $\mathcal{F} = \{\emptyset, A, A^c, S\}$.
- $\mathcal{F} = \mathcal{P}(S)$, all subsets — the largest σ-algebra. This is the right choice whenever S is finite or countable.

The pair $(S, \mathcal{F})$ is now fixed, and we put a function on it.

Definition 4.10 (Probability)**.** A **probability** is a non-negative function $\mathbb{P}$ defined on $\mathcal{F}$ such that

(i) $\mathbb{P}(S) = 1$;

(ii) (*countable additivity*) if $A_1, A_2, \cdots \in \mathcal{F}$ are pairwise disjoint, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

The triple $(S, \mathcal{F}, \mathbb{P})$ is a **probability space**.

That is the entire foundation. Everything else in the course is a consequence.

4.3 Properties of probability

Theorem 4.11 (Basic properties)**.** *For events $A, B, A_1, \dots, A_n$:*

(i) $\mathbb{P}(\emptyset) = 0$*;*

(ii) (finite additivity) if $A_1, \dots, A_n$ are pairwise disjoint then $\mathbb{P}\big(\bigcup_{i=1}^n A_i\big) = \sum_{i=1}^n \mathbb{P}(A_i)$*;*

(iii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(iv) $\mathbb{P}(A - B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$*;*

(v) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$*.*

Proof.

(i) Write $S = S \cup \emptyset \cup \emptyset \cup \cdots$, a disjoint union. Countable additivity gives $\mathbb{P}(S) = \mathbb{P}(S) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots$, so $\sum_{i \geq 1} \mathbb{P}(\emptyset) = 0$, forcing $\mathbb{P}(\emptyset) = 0$.

(ii) Pad the finite union with empty sets: $\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_n \cup \emptyset \cup \cdots) = \sum_{i=1}^n \mathbb{P}(A_i) + \sum \mathbb{P}(\emptyset) = \sum_{i=1}^n \mathbb{P}(A_i)$.

(iii) $S = A \cup A^c$ disjointly, so $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$.

(iv) $A = (A - B) \cup (A \cap B)$ disjointly.

(v) $A \cup B = (A - B) \cup B$ disjointly, so by (iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B)$. □

Theorem 4.12 (Inclusion–exclusion)**.**

$$\mathbb{P}\Big(\bigcup_{i=1}^n A_i\Big) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathbb{P}(A_1 \cap \cdots \cap A_n).$$

Proof. Induction on n, with (v) as the base case. (See Homework 1.) □

4.4 Finite and countable sample spaces

The axioms cover both cases classical probability could not touch.

4.4.1 A loaded die

Example 4.13. A die is weighted so that

$$\mathbb{P}(1) = \tfrac{1}{2}, \quad \mathbb{P}(2) = \tfrac{1}{6}, \quad \mathbb{P}(3) = \mathbb{P}(4) = \mathbb{P}(5) = \mathbb{P}(6) = \tfrac{1}{12}.$$

This is legitimate: the six numbers are non-negative and $\frac{1}{2} + \frac{1}{6} + 4 \cdot \frac{1}{12} = \frac{1}{2} + \frac{1}{6} + \frac{1}{3} = 1$. Take $\mathcal{F} = \mathcal{P}(S)$; then for any event, additivity gives the answer, e.g.

$$\mathbb{P}(\{1, 3, 4\}) = \tfrac{1}{2} + \tfrac{1}{12} + \tfrac{1}{12} = \tfrac{8}{12} = \tfrac{2}{3}.$$

Classical probability cannot express this at all — it assumes all outcomes are equally likely, and here they are not. The axioms never ask *where* the numbers came from, only that they are non-negative and add to 1.

4.4.2 Countably infinite sample spaces

Let $S = \{\omega_1, \omega_2, \dots\}$ be countably infinite and take $\mathcal{F} = \mathcal{P}(S)$. Assign each outcome a number $\mathbb{P}(\omega_i) \geq 0$ with $\sum_{i=1}^{\infty} \mathbb{P}(\omega_i) = 1$; then for any event A,

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i).$$

No equal-likelihood assumption is needed — and indeed *no* uniform assignment on a countably infinite set is possible.

Example 4.14 (Toss until the first head)**.** The outcome is the number of tosses, so $S = \{1, 2, 3, \dots\}$ and $\mathbb{P}(n) = \dfrac{1}{2^n}$. This is legitimate because $\sum_{n=1}^{\infty} 2^{-n} = 1$. For instance $\mathbb{P}(\text{even number of tosses}) = \sum_{k \geq 1} 2^{-2k} = \frac{1}{3}$.

Chapter 5

Conditional Probability and Independence

Lecture 5 | Devore 9e: 2.4

Last week we finished the axioms: a probability space $(S, \mathcal{F}, \mathbb{P})$ and a function $\mathbb{P} : \mathcal{F} \to [0, 1]$. This week: what happens to $\mathbb{P}$ when partial information arrives.

5.1 Conditional probability

Example 5.1 (The motivating computation)**.** Toss a coin twice, $S = \{(H, H), (H, T), (T, H), (T, T)\}$. Then $\mathbb{P}(2 \text{ heads}) = \frac{1}{4}$.

Now suppose someone tells you *there was at least one head.* Three outcomes remain possible, all still equally likely, so

$$\mathbb{P}(2 \text{ heads} \mid \text{at least 1 head}) = \tfrac{1}{3}.$$

What happened: the information shrank the sample space from S to B, and we renormalized so that B has total probability 1.

Definition 5.2 (Conditional probability)**.** Let $(S, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. For any $A \in \mathcal{F}$, the **conditional probability of A given B** is

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

Example 5.3 (Die)**.** Roll a die, $S = \{1, \dots, 6\}$. Let $A = \{\text{outcome is even}\} = \{2, 4, 6\}$ and $B = \{\text{outcome} > 3\} = \{4, 5, 6\}$. Then $A \cap B = \{4, 6\}$, so

$$\mathbb{P}(A \mid B) = \frac{2/6}{3/6} = \frac{2}{3}, \qquad \mathbb{P}(B \mid A) = \frac{2/6}{3/6} = \frac{2}{3}.$$

(The two happen to agree here because $\mathbb{P}(A) = \mathbb{P}(B)$; in general they differ, and confusing them is the single most common error in this course.)

Example 5.4 (Disease testing — set up now, solved in Lecture 7)**.** A disease affects 1% of the population. Let $A = \{\text{person has the disease}\}$, so $\mathbb{P}(A) = 0.01$ and $\mathbb{P}(A^c) = 0.99$. A test satisfies

- if the person has the disease, the result is positive 99% of the time: $\mathbb{P}(B \mid A) = 0.99$;
- if the person does not, the result is negative 95% of the time: $\mathbb{P}(B^c \mid A^c) = 0.95$;

where $B = \{\text{test result is positive}\}$. The patient's question is not $\mathbb{P}(B \mid A)$ — it is $\mathbb{P}(A \mid B)$. These are wildly different numbers, and Bayes' theorem is what converts one into the other.

5.2 Properties

Theorem 5.5 (Multiplication rule)**.** *If $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A) = \mathbb{P}(B)\,\mathbb{P}(A \mid B).$$

More generally, for events $A_1, \dots, A_n$ with $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$,

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1A_2) \cdots \mathbb{P}(A_n \mid A_1 \cdots A_{n-1}).$$

Here juxtaposition means intersection: $A_1A_2 := A_1 \cap A_2$.

This is the tool for anything described *sequentially*: draws from an urn, cards dealt one at a time, components failing in order.

Theorem 5.6 (Conditioning gives a probability)**.** *Fix B with $\mathbb{P}(B) > 0$. Then $A \mapsto \mathbb{P}(A \mid B)$ is itself a probability on $(S, \mathcal{F})$. In particular*

$$\mathbb{P}(A \mid B) \geq 0, \qquad \mathbb{P}(S \mid B) = 1, \qquad \mathbb{P}(\emptyset \mid B) = 0, \qquad \mathbb{P}(A^c \mid B) = 1 - \mathbb{P}(A \mid B),$$

and for pairwise disjoint $A_1, A_2, \dots$, $\mathbb{P}\big(\bigcup_i A_i \mid B\big) = \sum_i \mathbb{P}(A_i \mid B)$.

Everything you know about $\mathbb{P}$ therefore holds for $\mathbb{P}(\cdot \mid B)$ as well. Note carefully that this is a statement about the *first* slot only: $\mathbb{P}(A \mid B^c)$ has nothing to do with $\mathbb{P}(A \mid B)$, and $\mathbb{P}(A \mid B) + \mathbb{P}(A \mid B^c)$ is not 1 in general.

5.2.1 Pólya's urn model

Example 5.7 (Pólya's urn)**.** An urn holds r red and b black balls. Repeatedly: draw one ball at random, put it back, and add c more balls of the color just drawn. After n draws, what is the probability that the first n_1 draws are black and the remaining $n_2 = n - n_1$ are red?

Let A_i be the event that draw i comes out the prescribed color. The multiplication rule handles it, because after each draw we know exactly what the urn contains:

$$\mathbb{P}(A_1) = \frac{b}{b+r}, \qquad \mathbb{P}(A_2 \mid A_1) = \frac{b+c}{b+r+c}, \quad \mathbb{P}(A_3 \mid A_1A_2) = \frac{b+2c}{b+r+2c}, \quad \dots$$

$$\mathbb{P}(A_{n_1} \mid A_1 \cdots A_{n_1-1}) = \frac{b+(n_1-1)c}{b+r+(n_1-1)c},$$

and then the reds begin, with the urn already holding $b + n_1c$ black balls:

$$\mathbb{P}(A_{n_1+1} \mid \cdots) = \frac{r}{b+r+n_1c}, \qquad \mathbb{P}(A_{n_1+2} \mid \cdots) = \frac{r+c}{b+r+(n_1+1)c}, \ \dots, \ \mathbb{P}(A_n \mid \cdots) = \frac{r+(n_2-1)c}{b+r+(n-1)c}.$$

Multiplying,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \frac{\prod_{i=0}^{n_1-1}(b+ic)\ \prod_{j=0}^{n_2-1}(r+jc)}{\prod_{k=0}^{n-1}(b+r+kc)}.$$

Remark 5.8. The answer depends only on *how many* blacks and reds, not on the order — so any sequence with n_1 blacks and n_2 reds has the same probability. Pólya's urn is the standard model for "success breeds success" (contagion, popularity, reinforcement). Setting $c = 0$ recovers sampling with replacement.

5.3 Does conditioning raise or lower the probability?

Neither, in general. Draw one card from a deck, and let $A = \{\text{card is red}\}$, so $\mathbb{P}(A) = \frac{1}{2}$.

B	$\mathbb{P}(A \mid B)$	
card is a heart	1	information helps
card is a spade	0	information hurts
card is an ace	$\frac{1}{2}$	information is irrelevant

The third row is the interesting one, and it deserves a name.

5.4 Independence

Definition 5.9 (Independence of two events)**.** Events A and B are **independent**, written $A \perp\!\!\!\perp B$, if

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \iff \mathbb{P}(B \mid A) = \mathbb{P}(B) \iff \boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)}.$$

The boxed form is the official definition: it is symmetric in A and B, and it makes sense even when $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Proposition 5.10. *If $A \perp\!\!\!\perp B$ then $A^c \perp\!\!\!\perp B$, $A \perp\!\!\!\perp B^c$, and $A^c \perp\!\!\!\perp B^c$.*

Proof. $\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(A^c)\mathbb{P}(B)$. Apply this twice for the last claim. □

Remark 5.11. Independent $\neq$ mutually exclusive. If $A \cap B = \emptyset$ and both have positive probability, then $\mathbb{P}(A \cap B) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$, so disjoint events are as *dependent* as they get: knowing A happened tells you B did not.

Example 5.12 (Drawing with replacement)**.** A bag has a black and b white balls. Draw twice *with replacement*. Let $A = \{\text{first draw black}\}$, $B = \{\text{second draw black}\}$. Then

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\frac{a}{a+b} \cdot \frac{a}{a+b}}{\frac{a}{a+b}} = \frac{a}{a+b},$$

while, splitting on the first draw,

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) = \frac{a^2}{(a+b)^2} + \frac{ba}{(a+b)^2} = \frac{a(a+b)}{(a+b)^2} = \frac{a}{a+b}.$$

So $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ and the draws are independent — as they should be, since we put the ball back. Without replacement they are not.

5.5 More than two events

For three or more events there are two competing notions, and they are not the same.

Definition 5.13 (Pairwise vs. mutual independence)**.** Events $A_1, \ldots, A_n$ are

- **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$;
- **mutually independent** if for *every* sub-collection $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Mutual independence $\Rightarrow$ pairwise independence. The converse is **false**, and here is the standard counterexample — next lecture.

Chapter 6

Independent Trials, Total Probability, Bayes' Theorem

Lecture 6 | Devore 9e: 2.4–2.5

Last time: conditional probability, the multiplication rule, and independence of two events. We ended by defining pairwise and mutual independence and claiming they differ. Let us settle that first.

6.1 Pairwise does not imply mutual

Example 6.1 (Bernstein's tetrahedron)**.** A regular tetrahedron has four faces, painted

$$\text{red}, \quad \text{white}, \quad \text{black}, \quad \text{red} + \text{white} + \text{black (all three)}.$$

Roll it and look at the face it lands on. Let R, W, K be the events that the resting face contains red, white, black respectively. Each color appears on exactly two of the four faces, so

$$\mathbb{P}(R) = \mathbb{P}(W) = \mathbb{P}(K) = \tfrac{2}{4} = \tfrac{1}{2}.$$

Two colors appear together on exactly one face (the tricolor one), so

$$\mathbb{P}(R \cap W) = \mathbb{P}(R \cap K) = \mathbb{P}(W \cap K) = \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2}.$$

So the three events are **pairwise independent**. But

$$\mathbb{P}(R \cap W \cap K) = \tfrac{1}{4} \neq \tfrac{1}{8} = \mathbb{P}(R)\mathbb{P}(W)\mathbb{P}(K),$$

so they are **not mutually independent**.

The moral: pairwise independence lets you multiply two at a time and nothing more. Whenever a problem says "independent" about three or more events, it means mutually independent.

6.2 Independent trials

So far independence has been a property of events inside one experiment. Usually we want it to be a property of *experiments*: repeated coin tosses, repeated measurements, repeated draws. That requires building a new probability space out of two old ones.

Let Experiment 1 have space $(S_1, \mathcal{F}_1, \mathbb{P}_1)$ and Experiment 2 have $(S_2, \mathcal{F}_2, \mathbb{P}_2)$. Take the **product**

$$S := S_1 \times S_2, \qquad \mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2,$$

and define $\mathbb{P}$ on rectangles $A = A_1 \times A_2$ by

$$\mathbb{P}(A_1 \times A_2) := \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2).$$

Definition 6.2 (Independent trials)**.** With the above $\mathbb{P}$, the triple $(S, \mathcal{F}, \mathbb{P})$ is a probability space, and Experiments 1 and 2 are called **independent trials**.

Example 6.3. $S_1 = \{1, 2, 3, 4, 5, 6\}$ (a die), $S_2 = \{\text{red}, \text{black}\}$ (a card color). Then $S_1 \times S_2 = \{(1, \text{red}), (1, \text{black}), (2, \text{red}), \dots\}$, with twelve outcomes, and $\mathbb{P}\big((3, \text{red})\big) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$.

Example 6.4 (Calm gambler)**.** One game: $\mathbb{P}(W) = 0.6$, $\mathbb{P}(L) = 0.4$, and games are independent trials. Play three games, so $S = \{W, L\}^3$. Then

$$\mathbb{P}(WWW) = \mathbb{P}(W)\mathbb{P}(W \mid W)\mathbb{P}(W \mid WW) = \mathbb{P}(W)^3 = 0.6^3 = 0.216,$$

$$\mathbb{P}(LLW) = 0.4 \times 0.4 \times 0.6 = 0.096.$$

Independence is exactly what lets us drop the conditioning bars.

Example 6.5 (Emotional gambler)**.** The same chances in the *first* game, $\mathbb{P}(W) = 0.6$, $\mathbb{P}(L) = 0.4$, but the player is rattled by wins and motivated by losses:

$$\mathbb{P}(W \mid W) = 0.2, \quad \mathbb{P}(L \mid W) = 0.8, \qquad \mathbb{P}(W \mid L) = 0.8, \quad \mathbb{P}(L \mid L) = 0.2,$$

and the player reacts only to the game just played, so $\mathbb{P}(W \mid WW) = \mathbb{P}(W \mid W)$ and so on. These are *not* independent trials, so we must keep the multiplication rule in full:

$$\mathbb{P}(WWW) = 0.6 \times 0.2 \times 0.2 = 0.024, \qquad \mathbb{P}(LLW) = 0.4 \times 0.2 \times 0.8 = 0.064.$$

Compare with the calm gambler: same first-game odds, very different three-game behavior.

6.3 The law of total probability

Theorem 6.6 (Law of total probability)**.** *Let $\{A_i\}_{i \geq 1}$ be a partition of S with $\mathbb{P}(A_i) > 0$ for every i. Then for any event B,*

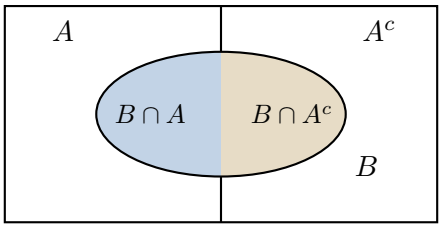
$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\, \mathbb{P}(B \mid A_i).$$

In the simplest case ($A_1 = A$, $A_2 = A^c$): $\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c)$.

Proof. Since $\bigcup_i A_i = S$,

$$\mathbb{P}(B) = \mathbb{P}\Big(B \cap \bigcup_i A_i\Big) = \mathbb{P}\Big(\bigcup_i (B \cap A_i)\Big) = \sum_i \mathbb{P}(B \cap A_i) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B \mid A_i),$$

where the third equality is countable additivity: the sets $B \cap A_i$ are pairwise disjoint because the A_i are. □



Remark 6.7. This is the workhorse. Whenever a problem has a hidden "which case are we in?" — which machine made the part, which route the person took, which box the coin came from — condition on the cases and add up.

Example 6.8 (Lost umbrella)**.** You lost your umbrella in one of three places, with prior probabilities

$$\mathbb{P}(A_1) = 0.50 \text{ (library)}, \quad \mathbb{P}(A_2) = 0.30 \text{ (classroom)}, \quad \mathbb{P}(A_3) = 0.20 \text{ (bus)},$$

and if it is there, the chance you actually recover it is

$$\mathbb{P}(B \mid A_1) = 0.80, \qquad \mathbb{P}(B \mid A_2) = 0.60, \qquad \mathbb{P}(B \mid A_3) = 0.05,$$

where $B = \{\text{umbrella is found}\}$. Then

$$\mathbb{P}(B) = 0.50(0.80) + 0.30(0.60) + 0.20(0.05) = 0.40 + 0.18 + 0.01 = 0.59.$$

6.4 Bayes' theorem

The law of total probability computes $\mathbb{P}(B)$ from the pieces. Bayes' theorem runs the arrow backwards: given that B happened, which piece were we in?

Theorem 6.9 (Bayes' theorem)**.** *Let $\{A_i\}$ be a partition of S with $\mathbb{P}(A_i) > 0$, and let $\mathbb{P}(B) > 0$. Then*

$$\boxed{\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\sum_j \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)}}$$

Proof. By the definition of conditional probability and the multiplication rule,

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_i)\mathbb{P}(B \mid A_i)}{\mathbb{P}(B)},$$

and expand the denominator with the law of total probability. □

6.4.1 Vocabulary

$\mathbb{P}(A_i)$	**prior**	what you believed before the evidence
$\mathbb{P}(B \mid A_i)$	**likelihood**	how well cause i explains the evidence
$\mathbb{P}(A_i \mid B)$	**posterior**	what you believe after the evidence

Example 6.10 (Lost umbrella, continued)**.** With $\mathbb{P}(B) = 0.59$ from above,

$$\mathbb{P}(A_1 \mid B) = \frac{0.40}{0.59} \approx 0.678, \quad \mathbb{P}(A_2 \mid B) = \frac{0.18}{0.59} \approx 0.305, \quad \mathbb{P}(A_3 \mid B) = \frac{0.01}{0.59} \approx 0.017.$$

The bus started at 20% and dropped to under 2%: recovering it at all is strong evidence it was never on the bus.

Example 6.11 (n heads in a row)**.** A coin is two-headed with probability $\mathbb{P}(A) = 0.05$, and fair with probability $\mathbb{P}(A^c) = 0.95$. You toss it n times and get all heads; call that B. Then $\mathbb{P}(B \mid A) = 1$ and $\mathbb{P}(B \mid A^c) = (1/2)^n$, so

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c)} = \frac{1 \cdot 0.05}{1 \cdot 0.05 + (1/2)^n \cdot 0.95} \xrightarrow[n\to\infty]{} 1.$$

n	1	2	3	5	10	20
$\mathbb{P}(A \mid B)$	0.095	0.174	0.296	0.627	0.982	0.99998

Remark 6.12. Even a heavily skeptical prior is overwhelmed by enough evidence — but notice how slowly it moves at first. After three heads the two-headed hypothesis is still under 30%. Cheap evidence does not flip strong priors; that is the whole content of the disease-testing example we set up last lecture and finish next time.

Chapter 7

Bayes in Practice; Random Variables

Lecture 7 | Devore 9e: 2.4–3.1

Recall from last time:

$$\underbrace{\mathbb{P}(A_i \mid B)}_{\text{posterior}} = \frac{\overbrace{\mathbb{P}(B \mid A_i)}^{\text{likelihood}}\overbrace{\mathbb{P}(A_i)}^{\text{prior}}}{\sum_j \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}.$$

7.1 Disease testing, finished

Example 7.1 (Rare disease)**.** A disease affects 1% of the population:

$$A = \{\text{person has the disease}\}, \qquad \mathbb{P}(A) = 0.01.$$

The test satisfies $\mathbb{P}(B \mid A) = 0.99$ (sensitivity) and $\mathbb{P}(B^c \mid A^c) = 0.95$ (specificity), where $B = \{\text{test is positive}\}$; hence $\mathbb{P}(B \mid A^c) = 0.05$. A patient tests positive. What is $\mathbb{P}(A \mid B)$?

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c)} = \frac{(0.01)(0.99)}{(0.01)(0.99) + (0.99)(0.05)} = \frac{0.0099}{0.0594} = \boxed{\frac{1}{6} \approx 16.7\%}$$

Remark 7.2 (The base-rate trap)**.** The test is right 99% of the time on sick people, yet a positive result means only a 1-in-6 chance of being sick. Here is the back-of-envelope version, and it is worth teaching this way:

$$\text{positive result} \implies \begin{array}{ll} \text{either} & \text{the person is sick} \quad (1\% \text{ of people}) \\ \text{or} & \text{the test is wrong } (5\% \text{ of the other } 99\%) \end{array}$$

and $\frac{1}{1+5} = \frac{1}{6}$. Because the disease is rare, the (small) false-positive *rate* acts on a (huge) healthy population, and the false positives outnumber the true positives five to one.

Now check the negative result. With $\mathbb{P}(B^c \mid A) = 0.01$,

$$\mathbb{P}(A^c \mid B^c) = \frac{(0.99)(0.95)}{(0.99)(0.95) + (0.01)(0.01)} = \frac{0.9405}{0.9406} \approx 99.99\%.$$

So for a *rare* disease:

positive ⇒ has the disease × (only 16.7%)
negative ⇒ does not ✓ (99.99%)

Example 7.3 (Common disease — the pattern reverses)**.** Same test, but now $\mathbb{P}(A) = 0.99$. Then

$$\mathbb{P}(A \mid B) = \frac{(0.99)(0.99)}{(0.99)(0.99) + (0.01)(0.05)} = \frac{0.9801}{0.9806} \approx 99.9\%,$$

$$\mathbb{P}(A^c \mid B^c) = \frac{(0.01)(0.95)}{(0.01)(0.95) + (0.99)(0.01)} = \frac{0.0095}{0.0194} \approx 49\%.$$

Now a positive result is conclusive and a *negative* result is a coin flip.

Remark 7.4. The test never changed. What changed is $\mathbb{P}(A)$. This is the sentence to remember: $\mathbb{P}(A \mid B)$ *is not a property of the test; it is a property of the test* ***and*** *the population.*

Remark 7.5 (Careful: the shortcut does not survive the switch)**.** The "$\frac{1}{1+5}$" reasoning above works for the *rare* disease because $\mathbb{P}(A) = 1\%$ is small enough that the healthy group is essentially the whole population, so its false-positive mass is just 5%. Running the same shortcut on the common disease gives

$$\frac{99}{99+5} = \frac{99}{104} \approx 95\%,$$

which is **wrong** — it charges the full 5% false-positive rate to a group that is only 1% of the population. The correct weighting is $\mathbb{P}(A^c)\mathbb{P}(B \mid A^c) = (0.01)(0.05) = 0.0005$, giving 99.9% as computed above. When the base rate is not small, do the arithmetic in full.

7.1.1 Testing twice

Example 7.6. Rare disease again, $\mathbb{P}(A) = 0.01$. Test the patient twice; let B_1, B_2 be the two positive results, and assume the two tests are *conditionally independent* given the true status:

$$\mathbb{P}(B_1 \cap B_2 \mid A) = \mathbb{P}(B_1 \mid A)\mathbb{P}(B_2 \mid A), \qquad \mathbb{P}(B_1 \cap B_2 \mid A^c) = \mathbb{P}(B_1 \mid A^c)\mathbb{P}(B_2 \mid A^c).$$

Then

$$\mathbb{P}(A \mid B_1 \cap B_2) = \frac{\mathbb{P}(B_1 \mid A)\mathbb{P}(B_2 \mid A)\mathbb{P}(A)}{\mathbb{P}(B_1 \mid A)\mathbb{P}(B_2 \mid A)\mathbb{P}(A) + \mathbb{P}(B_1 \mid A^c)\mathbb{P}(B_2 \mid A^c)\mathbb{P}(A^c)} = \frac{(0.99)^2(0.01)}{(0.99)^2(0.01) + (0.05)^2(0.99)}.$$

Numerically $\dfrac{0.009801}{0.009801 + 0.002475} \approx \boxed{80\%}$.

One positive: 16.7%. Two positives: 80%. This is why screening programs re-test rather than trusting a single result.

7.2 Random variables

We rarely care about the outcome ω itself; we care about some *number* attached to it. That number is a random variable.

Example 7.7. $S = \{\text{heart}, \text{diamond}, \text{spade}, \text{club}\}$ and

$$X(\text{heart}) = X(\text{diamond}) = 0, \qquad X(\text{spade}) = X(\text{club}) = 1.$$

The new "sample space" — the set of values — is $\widehat{S} = \{0, 1\}$.

Example 7.8. A gambler plays n games: $S = \{W, L\}^n$, so $|S| = 2^n$. Let X be the number of wins. Then $X(WLL\cdots L) = 1$, $X(LWLL\cdots L) = 1$, and $\widehat{S} = \{0, 1, 2, \ldots, n\}$, of size $n + 1$.

Note the compression: 2^n outcomes collapse to $n + 1$ values. That is the point — a random variable throws away the information we do not need.

Example 7.9 (Continuous)**.** Throw a dart at the board $S = \{(x, y) : |x| \le 1,\ |y| \le 1\}$ and record the distance from the center:

$$X(x, y) = \sqrt{x^2 + y^2}, \qquad \widehat{S} = [0, \sqrt{2}\,].$$

Definition 7.10 (Random variable)**.** Let $(S, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : S \to \mathbb{R}$ is a **random variable** if for every $x \in \mathbb{R}$,

$$\{\omega \in S : X(\omega) \le x\} \in \mathcal{F}.$$

Remark 7.11. Why the condition? Because $\mathbb{P}$ is only defined on $\mathcal{F}$. If the set $\{X \le x\}$ were not an event, the symbol $\mathbb{P}(X \le x)$ would be meaningless — and that symbol is the entire subject of the next lecture. When $\mathcal{F} = \mathcal{P}(S)$ (any finite or countable S) the condition is automatic, so in this course it is never an obstacle; it is there so the definition also works on $\mathbb{R}$.

Remark 7.12 (Notation)**.** In lecture I often write ξ (xi) for a random variable, and so do several homework problems; the textbook and the exams write X. They mean the same thing. Uppercase is the random variable (a *function*); lowercase x is a particular value it might take (a *number*). Writing $\mathbb{P}(X \le x)$ is meaningful; writing $\mathbb{P}(x \le x)$ is not.

Chapter 8

Distributions of Random Variables: pmf and cdf

Lecture 8 | Devore 9e: 3.1–3.2

Last time we defined a random variable $X : S \to \mathbb{R}$. Today: how to describe one completely, without ever mentioning S again.

Definition 8.1 (Discrete and continuous)**.** If X takes values in a finite or countable set $x_1, x_2, x_3, \ldots$, it is a **discrete** random variable. If it takes values filling an interval such as $[a, b]$ or (a, ∞), it is a **continuous** random variable.

Discrete this week; continuous starting in Lecture 13.

8.1 The probability mass function

Definition 8.2 (pmf)**.** The **probability mass function** of a discrete random variable X is

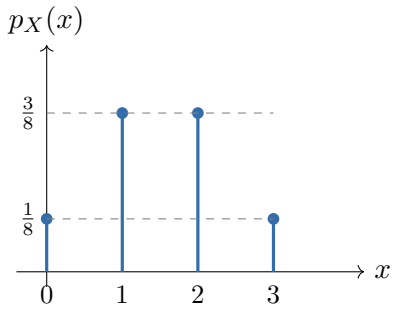
$$p_X(x) := \mathbb{P}\{X = x\} = \mathbb{P}\big(\{\omega \in S : X(\omega) = x\}\big).$$

Example 8.3 (Toss a coin three times)**.** $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ and $X = \#\text{heads}$, so X takes values $0, 1, 2, 3$. Pull back each value to an event:

$$\{X = 0\} = \{TTT\}, \qquad \{X = 1\} = \{HTT, THT, TTH\},$$
$$\{X = 2\} = \{THH, HTH, HHT\}, \quad \{X = 3\} = \{HHH\}.$$

All eight outcomes are equally likely, so

x	0	1	2	3
$p_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



Theorem 8.4 (Characterization of a pmf)**.** *A function p is the pmf of some discrete random variable if and only if*

$$p(x_i) \geq 0 \quad \text{for all } i, \qquad \sum_i p(x_i) = 1.$$

This is what makes "find the constant c so that this is a pmf" a well-posed question: the constant is whatever makes the sum equal 1.

Example 8.5 (Shooting at a target — the geometric pmf)**.** Each shot hits with probability $\beta \in (0,1)$, independently, and you repeat until you hit. Writing H for a hit and M for a miss, the experiment stops at the first H, so the sample space is

$$S = \{H,\ MH,\ MMH,\ MMMH,\ \dots\},$$

which is countably infinite. Let X be the number of shots fired, so $X \in \{1, 2, 3, \dots\}$. Each value pulls back to a single outcome,

$$\{X = 1\} = \{H\}, \qquad \{X = 2\} = \{MH\}, \qquad \{X = 3\} = \{MMH\},\ \dots$$

and by independence

$$\mathbb{P}\{X = 1\} = \beta, \quad \mathbb{P}\{X = 2\} = (1-\beta)\beta, \quad \mathbb{P}\{X = 3\} = (1-\beta)^2\beta,\ \dots$$

so

$$\boxed{p_X(n) = (1-\beta)^{n-1}\beta, \qquad n = 1, 2, 3, \dots}$$

Check: $\sum_{n\geq 1}(1-\beta)^{n-1}\beta = \beta \cdot \frac{1}{1-(1-\beta)} = 1$. ✓

This is the **geometric distribution**; we return to it in Lecture 10.

Example 8.6 (Indicator random variable)**.** For an event $A \in \mathcal{F}$ define

$$\mathbf{1}_{\{A\}}(\omega) = \mathbf{1}_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

Then $\mathbb{P}\{\mathbf{1}_A = 1\} = \mathbb{P}(A)$ and $\mathbb{P}\{\mathbf{1}_A = 0\} = 1 - \mathbb{P}(A) = \mathbb{P}(A^c)$:

x	0	1
$p(x)$	$\mathbb{P}(A^c)$	$\mathbb{P}(A)$

The indicator is the bridge between events and random variables: every event becomes a $\{0,1\}$-valued random variable, and we will use this repeatedly.

Remark 8.7. Once you have the pmf you can compute the probability of *any* event about X by adding up. For the shooting example with only 10 bullets available,

$$\mathbb{P}\{X \le 10\} = \mathbb{P}\Big(\bigcup_{k=1}^{10}\{X = k\}\Big) = \sum_{k=1}^{10}(1-\beta)^{k-1}\beta = 1-(1-\beta)^{10},$$

the events $\{X = k\}$ being disjoint.

8.1.1 The recipe

Every pmf computation in this course follows the same five steps. When a problem stalls, it is almost always because one of them was skipped.

1. Identify the **experiment**: its outcomes and its sample space S.
2. State the **rule** $X : S \to \mathbb{R}$ — what number is being recorded.
3. List the **values** X can take.
4. For each value x, find the **pre-image** $\{X = x\} = \{\omega \in S : X(\omega) = x\}$. This is an event, i.e. a subset of S.
5. Compute $\mathbb{P}$ of each of those events. The resulting table is the **pmf**.

Steps 1–3 are bookkeeping; step 4 is where the thinking happens; step 5 is the probability from the first half of the course. In the three-coin example, step 4 was the list $\{X = 0\} = \{TTT\}$, $\{X = 1\} = \{HTT, THT, TTH\}$, and so on — and once that list was written down, step 5 was immediate.

8.2 The cumulative distribution function

The pmf works only for discrete variables. The cdf works for all of them, which is why it is the more fundamental object.

Definition 8.8 (cdf)**.** The **cumulative distribution function** of a random variable X is

$$F_X(x) := \mathbb{P}\{X \le x\}, \qquad x \in \mathbb{R}.$$

For a discrete X this is computed from the pmf by

$$F_X(x) = \sum_{y\,:\,y \le x} p_X(y).$$

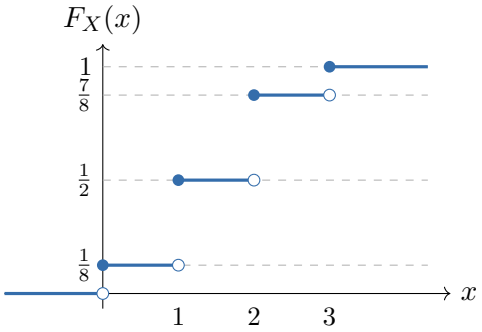
Example 8.9 (cdf of an indicator)**.** With $p(0) = \mathbb{P}(A^c)$ and $p(1) = \mathbb{P}(A)$:

- $x < 0$: no values of X are $\le x$, so $F(x) = 0$;
- $0 \le x < 1$: only the value 0 qualifies, so $F(x) = \mathbb{P}(A^c)$;
- $x \ge 1$: both qualify, so $F(x) = \mathbb{P}(A^c) + \mathbb{P}(A) = 1$.

$$F_{\mathbf{1}_A}(x) = \begin{cases} 0, & x < 0, \\ \mathbb{P}(A^c), & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Example 8.10 (cdf of #heads in three tosses)**.** From the pmf $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{8}, & 0 \le x < 1, \\ \frac{1}{2}, & 1 \le x < 2, \\ \frac{7}{8}, & 2 \le x < 3, \\ 1, & x \ge 3. \end{cases}$$



The picture shows the two features that characterize every discrete cdf: it is a **step function**, and it is **right continuous** — at a jump the value is the upper dot, because $\{X \le x\}$ uses $\le$.

Theorem 8.11 (Classification of cdfs)**.** *Let X be a random variable on $(S, \mathcal{F}, \mathbb{P})$ and $F_X(x) = \mathbb{P}\{X \le x\}$. Then*

(i) F_X is non-decreasing: $a < b \Rightarrow F_X(a) \le F_X(b)$;

(ii) $\lim_{x \to +\infty} F_X(x) = 1$ and $\lim_{x \to -\infty} F_X(x) = 0$;

(iii) F_X is right continuous: $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$.

Conversely, if a function F satisfies (i)–(iii), then there exist a probability space and a random variable X on it whose cdf is exactly F.

Remark 8.12. The converse is what makes distributions first-class objects. From now on we can say "let X be geometric with parameter β" and never mention what S was. The underlying probability space disappears from the course at this point, and only distributions remain.

8.2.1 Reading probabilities off a cdf

For any $a < b$,

$$\mathbb{P}\{a < X \le b\} = F_X(b) - F_X(a), \qquad \mathbb{P}\{X = a\} = F_X(a) - F_X(a^-),$$

where $F_X(a^-) = \lim_{x \to a^-} F_X(x)$. So the jump of the cdf at a *is* the mass at a; where the cdf is continuous, $\mathbb{P}\{X = a\} = 0$. This is the identity that will let continuous random variables make sense in Lecture 13.

Chapter 9

Expected Value of a Discrete Random Variable

Lecture 9 | Devore 9e: 3.3

9.1 Leftovers on the cdf

Everything about X can be read off $F_X(x) = \mathbb{P}\{X \le x\}$, provided you are careful about which endpoints are included. Write

$$F_X(x^-) := \lim_{y \to x^-} F_X(y)$$

for the left limit (the value "just before" any jump). Then

$$\begin{aligned} \mathbb{P}\{X > x\} &= 1 - F_X(x), \\ \mathbb{P}\{X < x\} &= F_X(x^-), \\ \mathbb{P}\{X = x\} &= F_X(x) - F_X(x^-), \\ \mathbb{P}\{a < X \le b\} &= F_X(b) - F_X(a), \\ \mathbb{P}\{a \le X \le b\} &= F_X(b) - F_X(a^-). \end{aligned}$$

Derivation of the last one.

$$\{a \le X \le b\} = \{X \ge a\} \cap \{X \le b\} = \{X \le b\} \cap \{X < a\}^c = \{X \le b\} \setminus \{X < a\},$$

and $\{X < a\} \subset \{X \le b\}$, so the probabilities subtract: $\mathbb{P}\{a \le X \le b\} = \mathbb{P}\{X \le b\} - \mathbb{P}\{X < a\} = F_X(b) - F_X(a^-)$. □

Remark 9.1. For a discrete X the left limit is easy to evaluate: if X takes only the values $1, 2, 3, \ldots$ then $\mathbb{P}\{X < 3\} = \mathbb{P}\{X \le 2\} = F_X(2)$. If X can take the value 2.8 but nothing in $(2.8, 2.9]$, then $\mathbb{P}\{X < 2.9\} = \mathbb{P}\{X \le 2.8\} = F_X(2.8)$. In words: step back to the largest value X can actually attain.

9.2 Expected value

Example 9.2 (Where the definition comes from)**.** Among 100 students, the number registered for $0, 1, 2, 3, 4$ courses is $5, 10, 30, 30, 25$ respectively. On average, how many courses does a student

take?

$$\frac{\text{total courses}}{\text{total students}} = \frac{5\cdot 0 + 10\cdot 1 + 30\cdot 2 + 30\cdot 3 + 25\cdot 4}{100} = 0.05{\cdot}0{+}0.10{\cdot}1{+}0.30{\cdot}2{+}0.30{\cdot}3{+}0.25{\cdot}4 = 2.6.$$

Look at the middle expression: dividing through by 100 turned the counts into *proportions*, i.e. into a pmf. So the average is a **probability-weighted sum of the values**.

Definition 9.3 (Expected value)**.** Let X be a discrete random variable taking values in $D = \{x_1, x_2, \dots\}$ with $p_i = \mathbb{P}\{X = x_i\}$. The **expected value** (mean) of X is

$$\mathbb{E}[X] = \mu_X := \sum_{x_i \in D} x_i\, p_i = \sum_{x_i \in D} x_i\, \mathbb{P}\{X = x_i\},$$

provided the sum converges absolutely.

Remark 9.4. X is a *function*; $\mathbb{E}[X]$ is a *number*. Also, $\mathbb{E}[X]$ need not be a value X can take — no student registers for 2.6 courses.

Example 9.5 (Expectation generalizes the plain average)**.** Given numbers $a_1 \le a_2 \le \cdots \le a_n$, define X by $\mathbb{P}\{X = a_i\} = \frac{1}{n}$. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \cdot \frac{1}{n} = \frac{a_1 + \cdots + a_n}{n},$$

the ordinary mean. The general definition just allows unequal weights.

9.2.1 A worked series: the geometric mean

Example 9.6 (Shooting at a target)**.** $\mathbb{P}\{X = n\} = (1-\beta)^{n-1}\beta$ for $n \ge 1$. Then

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n(1-\beta)^{n-1}\beta = \beta \sum_{n=1}^{\infty} n(1-\beta)^{n-1}.$$

The trick is to recognize $n(1-\beta)^{n-1} = -\frac{d}{d\beta}(1-\beta)^n$:

$$\mathbb{E}[X] = -\beta \frac{d}{d\beta} \sum_{n=1}^{\infty} (1-\beta)^n = -\beta \frac{d}{d\beta}\left(\frac{1-\beta}{1-(1-\beta)}\right) = -\beta \frac{d}{d\beta}\left(\frac{1-\beta}{\beta}\right) = -\beta \frac{d}{d\beta}\left(\frac{1}{\beta} - 1\right) = -\beta\left(-\frac{1}{\beta^2}\right) = \boxed{\frac{1}{\beta}}.$$

(We used $\sum_{n\ge 1} a^n = \frac{a}{1-a}$ for $|a| < 1$, and differentiated term by term, which is legitimate inside the radius of convergence.)

Sanity check: if you hit 10% of the time you expect to need 10 shots. Good.

9.2.2 When the mean does not exist

Example 9.7 (Heavy tail)**.** Let X be the number of interviews needed to get a job, with

$$\mathbb{P}\{X = x\} = a \cdot \frac{1}{x^2}, \qquad x = 1, 2, 3, \dots$$

First find a. We need $\sum_{x\ge 1} a/x^2 = 1$, and $\sum_{x\ge 1} x^{-2} = \pi^2/6$, so $a = 6/\pi^2$. That is a perfectly legitimate pmf. But

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{6}{\pi^2}\frac{1}{x^2} = \frac{6}{\pi^2}\sum_{x=1}^{\infty} \frac{1}{x} = \infty.$$

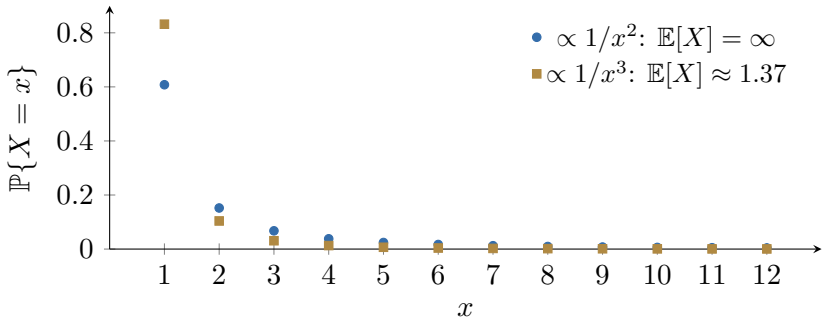
Remark 9.8. The mean is $+\infty$: the distribution has a **heavy tail**. Large values are individually unlikely but not unlikely *enough* to keep the weighted sum finite. This is not a pathology invented by mathematicians — insurance losses, file sizes, city populations, and income all behave this way, and for such data the sample mean is a badly behaved summary.

Example 9.9 (How much lighter must the tail be?)**.** Keep the same shape but decay one power faster: let

$$\mathbb{P}\{X = x\} = \frac{b}{x^3}, \qquad x = 1, 2, 3, \dots$$

Normalizing needs $\sum_{x\geq 1} x^{-3} = \zeta(3) \approx 1.2021$, so $b = 1/\zeta(3) \approx 0.832$. Now

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{b}{x^3} = b \sum_{x=1}^{\infty} \frac{1}{x^2} = b \cdot \frac{\pi^2}{6} \approx 0.832 \times 1.645 \approx 1.37 \ < \ \infty.$$



Remark 9.10. Both pmfs look like they decay to zero, and on a plot they are nearly indistinguishable past $x = 5$. Yet one has an infinite mean and the other does not. "Heavy tail" is therefore not something you can reliably eyeball — it is a statement about how fast $\sum x\, p(x)$ accumulates out in the tail, and the boundary sits exactly at the $1/x^2$ rate: for $p(x) \propto x^{-s}$ the sum $\sum x\, p(x)$ behaves like $\sum x^{1-s}$, which converges precisely when $s > 2$.

9.3 Functions of a random variable

If $X : S \to \mathbb{R}$ is a random variable and $h : \mathbb{R} \to \mathbb{R}$, then the composition $h(X) = h \circ X : S \to \mathbb{R}$ is again a random variable. Its expectation can be computed without ever finding the distribution of $h(X)$:

Theorem 9.11 (Law of the unconscious statistician)**.** *If X takes values in $D = \{x_1, x_2, \dots\}$ and $h : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[h(X)] = \sum_{x_i \in D} h(x_i)\, \mathbb{P}\{X = x_i\}.$$

Example 9.12. Shooting at a target, X = number of shots. Each bullet costs 5, and maintenance costs $\log X$. Total expense is $h(X) = 5X + \log X$:

X	1	2	3	$\cdots$
$h(X)$	$5 + \log 1$	$10 + \log 2$	$15 + \log 3$	$\cdots$
$\mathbb{P}$	β	$(1-\beta)\beta$	$(1-\beta)^2\beta$	$\cdots$

so $\mathbb{E}[h(X)] = \sum_{n\geq 1}(5n + \log n)(1-\beta)^{n-1}\beta$. Note we did *not* have to work out the pmf of $h(X)$ — that is exactly what the theorem buys us.

Corollary 9.13 (Linearity in the simplest case)**.** *For constants $a, b \in \mathbb{R}$,*

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b.$$

Proof.

$$\mathbb{E}[aX+b] = \sum_{x_i \in D}(ax_i + b)\mathbb{P}\{X = x_i\} = a\underbrace{\sum_{x_i\in D} x_i\mathbb{P}\{X = x_i\}}_{=\,\mathbb{E}[X]} + b\underbrace{\sum_{x_i\in D}\mathbb{P}\{X = x_i\}}_{=\,1} = a\mathbb{E}[X] + b. \quad \square$$

Remark 9.14. Linearity is what makes expectation usable. Note carefully that $\mathbb{E}[h(X)] \neq h(\mathbb{E}[X])$ in general — for instance $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$, and the gap between them is exactly the variance, which is next lecture.

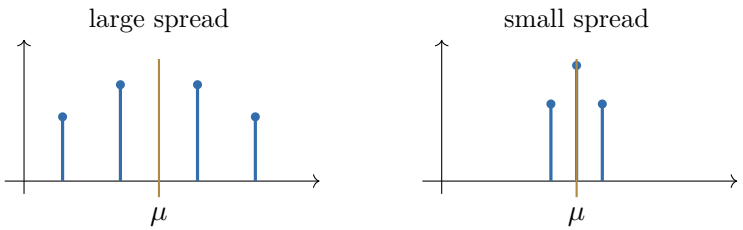
Chapter 10

Variance; Binomial, Geometric and Hypergeometric

Lecture 10 | Devore 9e: 3.3–3.5

10.1 Variance

The mean locates the distribution. It says nothing about how spread out it is: the two pmfs below have the same μ and could hardly be more different.



The natural measure of spread is the average squared distance from the mean.

Definition 10.1 (Variance and standard deviation)**.** Let X have mean $\mu = \mathbb{E}[X]$. The **variance** is

$$V(X) = \mathrm{Var}(X) = \sigma_X^2 := \mathbb{E}\big[(X-\mu)^2\big],$$

and the **standard deviation** is $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Note that $\mathrm{Var}(X)$ is just $\mathbb{E}[h(X)]$ for $h(x) = (x-\mu)^2$, so the previous lecture already tells us how to compute it: $\mathrm{Var}(X) = \sum_{x_i \in D} (x_i - \mu)^2 \mathbb{P}\{X = x_i\}$.

Remark 10.2. Why square rather than take absolute values? Squaring is smooth (differentiable), it makes the algebra below work, and it is what turns independence into additivity later (Lecture 22). The price is that σ^2 has the *squared* units of X — which is why we also report σ, which is in the original units.

Theorem 10.3 (Computational formula)**.**

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

Proof. Expand and use linearity, remembering μ is a constant:

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \quad \square$$

So computing a variance is always the same three steps:

$$(1)\ \mathbb{E}[X], \qquad (2)\ \mathbb{E}[X^2], \qquad (3)\ \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Example 10.4 (Both routes on the same numbers)**.** A loaded die has pmf

x	1	2	3	4	5	6
$p(x)$	0.30	0.25	0.15	0.05	0.10	0.15

(The probabilities sum to 1. ✓) First the mean:

$$\mathbb{E}[X] = 1(0.30)+2(0.25)+3(0.15)+4(0.05)+5(0.10)+6(0.15) = 0.30+0.50+0.45+0.20+0.50+0.90 = 2.85.$$

Route 1 — straight from the definition.

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{x=1}^{6}(x-2.85)^2 p(x)\\ &= (-1.85)^2(0.30) + (-0.85)^2(0.25) + (0.15)^2(0.15)\\ &\quad + (1.15)^2(0.05) + (2.15)^2(0.10) + (3.15)^2(0.15)\\ &= 1.026750 + 0.180625 + 0.003375 + 0.066125 + 0.462250 + 1.488375\\ &= 3.2275.\end{aligned}$$

Route 2 — the computational formula.

$$\mathbb{E}[X^2] = 1(0.30)+4(0.25)+9(0.15)+16(0.05)+25(0.10)+36(0.15) = 0.30+1.00+1.35+0.80+2.50+5.40 = 11.35,$$

$$\mathrm{Var}(X) = 11.35 - (2.85)^2 = 11.35 - 8.1225 = 3.2275. \quad \checkmark$$

Either way $\sigma = \sqrt{3.2275} \approx 1.80$.

Remark 10.5. Route 2 is almost always less work — six squarings of small integers instead of six squarings of awkward decimals — and it is far less error-prone. Use the definition to understand what variance *means*; use $\mathbb{E}[X^2] - \mu^2$ to compute it.

Proposition 10.6. $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$ *and* $\sigma_{aX+b} = |a|\,\sigma_X$.

Proof. $\mathbb{E}[aX+b] = a\mu+b$, so $(aX+b)-(a\mu+b) = a(X-\mu)$, and squaring gives $a^2(X-\mu)^2$. Taking expectations, $\mathrm{Var}(aX+b) = a^2\mathbb{E}[(X-\mu)^2] = a^2\,\mathrm{Var}(X)$; the standard deviation is the square root, so it picks up $|a|$ rather than a. $\square$

Shifting does not change spread; rescaling scales it. That is exactly what a measure of spread should do.

Example 10.7 (Variance of the shooting variable)**.** With $\mathbb{P}\{X=n\}=(1-\beta)^{n-1}\beta$ and $\mathbb{E}[X]=1/\beta$, we need $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2]=\beta\sum_{n=1}^{\infty}n^2(1-\beta)^{n-1}.$$

Differentiate the geometric series twice. Since $\frac{d^2}{d\beta^2}(1-\beta)^{n+1}=(n+1)n(1-\beta)^{n-1}=n^2(1-\beta)^{n-1}+n(1-\beta)^{n-1}$, summing over $n\geq 1$ and differentiating term by term reduces everything to one series we can add up in closed form:

$$\sum_{n\geq 1}(1-\beta)^{n+1}=\frac{(1-\beta)^2}{\beta}=\frac{1}{\beta}-2+\beta,\qquad \frac{d^2}{d\beta^2}\left(\frac{1}{\beta}-2+\beta\right)=\frac{2}{\beta^3},$$

so $\sum_{n\geq 1}\big(n^2+n\big)(1-\beta)^{n-1}=\frac{2}{\beta^3}$. The first-derivative result of Lecture 9 already gave $\sum_{n\geq 1}n(1-\beta)^{n-1}=\frac{1}{\beta^2}$, and subtracting it leaves $\sum_{n\geq 1}n^2(1-\beta)^{n-1}=\frac{2}{\beta^3}-\frac{1}{\beta^2}$. Putting the pieces together,

$$\mathbb{E}[X^2]=\beta\left(\frac{2}{\beta^3}-\frac{1}{\beta^2}\right)=\frac{2}{\beta^2}-\frac{1}{\beta}=\frac{2-\beta}{\beta^2},\qquad \text{hence}\qquad \mathrm{Var}(X)=\frac{2-\beta}{\beta^2}-\frac{1}{\beta^2}=\boxed{\frac{1}{\beta^2}-\frac{1}{\beta}=\frac{1-\beta}{\beta^2}}.$$

10.2 The binomial distribution

From here on we stop building distributions by hand and start naming the ones that occur over and over.

Definition 10.8 (Binomial experiment)**.** An experiment consisting of n *independent* trials, where

- each trial has exactly two outcomes, success (S) and failure (F);
- $\mathbb{P}(S)=p$ and $\mathbb{P}(F)=1-p\overset{\text{def}}{=}q$, the same on every trial.

Definition 10.9 (Binomial distribution)**.** In a binomial experiment let $X:=$ the number of successes among the n trials. Then $X\sim\text{Binomial}(n,p)$, and its pmf is written $b(x;n,p)$.

Theorem 10.10.

$$b(x;n,p)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}, & x=0,1,2,\dots,n,\\ 0, & \textit{otherwise.}\end{cases}$$

Proof. The sample space consists of all strings of length n over $\{S,F\}$, e.g. $SSF\cdots FS$. By independence, any *particular* string with x successes has probability

$$\mathbb{P}(S)\mathbb{P}(S)\mathbb{P}(F)\cdots\mathbb{P}(F)\mathbb{P}(S)=p^x(1-p)^{n-x},$$

regardless of the order the letters appear in. The event $\{X=x\}$ is the disjoint union of all such strings, and there are $\binom{n}{x}$ of them (choose which x of the n positions hold the successes). Adding up, $\mathbb{P}\{X=x\}=\binom{n}{x}p^x(1-p)^{n-x}$. □

Theorem 10.11. *If* $X \sim \text{Binomial}(n, p)$ *then* $\mathbb{E}[X] = np$ *and* $\text{Var}(X) = np(1-p)$.

(The slick proof writes $X = \mathbf{1}_1 + \cdots + \mathbf{1}_n$ as a sum of indicators; we give it in Lecture 22 once we have additivity of variance.)

10.3 The geometric distribution

Definition 10.12 (Geometric distribution)**.** In a binomial experiment, let $X :=$ the number of trials up to and including the first success. Then $X \sim \text{Geometric}(p)$ with pmf

$$g(x; p) = (1-p)^{x-1}p, \qquad x = 1, 2, 3, \ldots$$

This is the shooting-at-a-target variable from Lectures 8–9, so we already know

$$\mathbb{E}[X] = \frac{1}{p}, \qquad \text{Var}(X) = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

Remark 10.13. Binomial and geometric come from the *same* experiment; they differ in what is fixed and what is random. Binomial: n fixed, count the successes. Geometric: one success required, count the trials. It is named after the geometric series, which is where its normalization and its mean came from.

Remark 10.14 (Warning)**.** Some books (and software) define the geometric as the number of *failures* before the first success, i.e. shifted down by one, with mean $(1-p)/p$. We always use the "number of trials" convention, starting at $x = 1$.

10.4 The hypergeometric distribution

Binomial sampling is *with* replacement (or from an inexhaustible population), so p never changes. What if we sample without replacement?

Definition 10.15 (Hypergeometric distribution)**.** A population has N objects, of which M are "good" and $N - M$ are "bad". Select n objects without replacement (all at once). Let $X :=$ the number of good objects selected. Then $X \sim \text{Hypergeometric}(n, M, N)$ with

$$\mathbb{P}\{X = x\} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

valid for integers x with $0 \le x \le M$, $0 \le n - x \le N - M$ (and $n \le N$).

The formula is pure counting: choose which x of the M good ones you got, choose which $n - x$ of the $N - M$ bad ones you got, and divide by the total number of samples.

Theorem 10.16. *With $p := M/N$ (the population fraction of good objects),*

$$\mathbb{E}[X] = n \cdot \frac{M}{N} = np, \qquad \mathrm{Var}(X) = \underbrace{\frac{N-n}{N-1}}_{\text{finite population correction}} \cdot np(1-p).$$

Remark 10.17 (Hypergeometric versus binomial)**.** The mean is *identical* to the binomial mean. The variance carries an extra factor $\frac{N-n}{N-1} \leq 1$: sampling without replacement is more informative, so the count is less variable.

If $n \ll N$ then $\frac{N-n}{N-1} \approx 1$, and removing a few objects barely changes the composition of the population. In that regime

$$\mathbb{P}\{X_{\text{hyp}} = x\} \approx \mathbb{P}\{X_{\text{bin}} = x\},$$

which is why polling 1000 people out of 300 million is treated as a binomial problem. (Homework 5 asks you to make this precise.)

Chapter 11

Review for Midterm 1

Lecture 11 | Devore 9e: Ch. 2–3.5

Midterm 1 covers Lectures 1–10: probability spaces, counting, conditional probability, independence, Bayes, and discrete random variables through variance.

11.1 What you must be able to do

1. **Write down a sample space** and express a verbal event as a subset (Lecture 2). Translate "at least one", "none", "exactly two" into unions, complements, intersections.
2. **Count** (Lecture 3): decide *order?* and *replacement?*, then use the right entry of the table.
3. **Compute geometric probabilities** as ratios of length/area (Lecture 4).
4. **Use the probability axioms** (Lecture 4) to prove small identities: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, inclusion–exclusion.
5. **Condition** (Lectures 5–7): multiplication rule for sequential experiments, law of total probability when there is a hidden case split, Bayes when the question reverses the given.
6. **Check independence** from the definition $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ — and know that pairwise does not imply mutual (Lectures 5–6).
7. **Work with a discrete random variable** (Lectures 8–10): find the pmf, build the cdf, compute $\mathbb{E}[X]$, $\mathbb{E}[h(X)]$, and $\mathrm{Var}(X)$.
8. **Recognize the named families** so far: Bernoulli (the indicator variable of Lecture 8), binomial, geometric, hypergeometric.

11.2 Formula sheet

Counting

	without replacement	with replacement
order matters	$A_n^k = \frac{n!}{(n-k)!}$	n^k
order doesn't matter	$\binom{n}{k}$	$\binom{n+k-1}{n-1}$

Probability

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \qquad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$$
$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B \mid A_i) \quad \mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B \mid A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B \mid A_j)}$$

$A \perp\!\!\!\perp B \iff \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Discrete random variables

$$F_X(x) = \mathbb{P}\{X \le x\}, \qquad \mathbb{E}[X] = \sum_i x_i p_i, \qquad \mathbb{E}[h(X)] = \sum_i h(x_i) p_i,$$
$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2, \qquad \sigma_X = \sqrt{\mathrm{Var}(X)},$$
$$\mathbb{E}[aX+b] = a\mathbb{E}[X] + b, \qquad \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X), \qquad \sigma_{aX+b} = |a|\,\sigma_X.$$

Families so far

Name	pmf	support	$\mathbb{E}$	Var
Bernoulli(p)	$p^k(1-p)^{1-k}$	$k \in \{0,1\}$	p	$p(1-p)$
Binomial(n,p)	$\binom{n}{k}p^k(1-p)^{n-k}$	$k = 0,\dots,n$	np	$np(1-p)$
Geometric(p)	$(1-p)^{k-1}p$	$k = 1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric(n,M,N)	$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$	k feasible	$n\frac{M}{N}$	$\frac{N-n}{N-1}n\frac{M}{N}\big(1-\frac{M}{N}\big)$

11.3 The three mistakes that cost the most points

1. **Confusing $\mathbb{P}(A \mid B)$ with $\mathbb{P}(B \mid A)$.** Read the sentence again: what is *given*, and what is being *asked*? The disease example (Lecture 7) exists entirely to make this vivid.
2. **Treating disjoint as independent.** Disjoint events with positive probability are maximally *dependent*. "Mutually exclusive" is a statement about $A \cap B$; "independent" is a statement about $\mathbb{P}(A \cap B)$.
3. **Forgetting that $\mathbb{E}[h(X)] \neq h(\mathbb{E}[X])$.** In particular $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$; their difference is the variance.

Remark 11.1. Practice problems for this midterm are in `exams/2026-spring/tex/` — see `practice1.tex` and `make-up-exam.tex`, both with the same topic coverage. (`mid2.tex` and `mid2-makeup.tex` are Midterm 2 material.)

Chapter 12

Negative Binomial, Poisson, and the Poisson Process

Lecture 12 | Devore 9e: 3.5–3.6

So far the discrete families are geometric, binomial, hypergeometric. Today we add the last two, and then collect everything in one table.

12.1 The negative binomial distribution

Definition 12.1 (Bernoulli random variable)**.** A single trial with two outcomes S, F and $\mathbb{P}(S) = p$, $\mathbb{P}(F) = 1 - p$. Coded as a random variable, $X \sim \text{Bernoulli}(p)$ takes the value 1 on S and 0 on F.

Definition 12.2 (Negative binomial distribution)**.** Take a sequence of i.i.d. (*independent and identically distributed*) Bernoulli trials and repeat until a total of r successes has occurred. Let

$$X := \text{the number of failures before the } r\text{-th success.}$$

Then $X \sim \text{NegBinomial}(r, p)$.

Theorem 12.3.

$$\mathbb{P}\{X = x\} = \binom{x + r - 1}{x}(1 - p)^x p^r, \qquad x = 0, 1, 2, \ldots$$

Proof. The experiment stops at the r-th success, so the *last* trial is forced to be S. Before it there are $x + r - 1$ trials containing exactly x failures and $r - 1$ successes, in any order:

$$\underbrace{\cdots}_{x\ F\text{'s and } (r-1)\ S\text{'s, any order}} \boxed{S}$$

There are $\binom{x+r-1}{x}$ such arrangements, each of probability $(1 - p)^x p^{r-1}$, and then the final S contributes another factor p. □

Remark 12.4 (Consistency check)**.** Set $r = 1$: $\mathbb{P}\{X = x\} = (1-p)^x p$, the geometric distribution counted by *failures*. So the geometric is NegBinomial$(1, p)$, up to the usual shift between "number of trials" and "number of failures".

Theorem 12.5.

$$\mathbb{E}[X] = r\Big(\frac{1}{p} - 1\Big) = \frac{r(1-p)}{p}, \qquad \mathrm{Var}(X) = \frac{r\big(\frac{1}{p} - 1\big)}{p} = \frac{r(1-p)}{p^2}.$$

The mean is worth seeing intuitively. Reaching one success takes $1/p$ trials on average, of which one is the success itself, so $\frac{1}{p} - 1$ are failures. Reaching r successes just repeats that r times.

12.2 The Poisson distribution

Recall $b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$. Now suppose n is very large while the mean $np = \mu$ stays moderate — so p is tiny. Many trials, each almost certainly a failure. What happens to the pmf?

Theorem 12.6 (Poisson limit)**.** *If $n \to \infty$ and $p \to 0$ with $np = \mu$ fixed, then for each $x = 0, 1, 2, \ldots$*

$$\binom{n}{x} p^x (1-p)^{n-x} \;\longrightarrow\; \frac{\mu^x}{x!} e^{-\mu}.$$

Sketch (this is Homework 5). Write $p = \mu/n$ and expand:

$$\binom{n}{x}\Big(\frac{\mu}{n}\Big)^x \Big(1 - \frac{\mu}{n}\Big)^{n-x} = \underbrace{\frac{n(n-1)\cdots(n-x+1)}{n^x}}_{\to\, 1} \cdot \frac{\mu^x}{x!} \cdot \underbrace{\Big(1 - \frac{\mu}{n}\Big)^n}_{\to\, e^{-\mu}} \cdot \underbrace{\Big(1 - \frac{\mu}{n}\Big)^{-x}}_{\to\, 1},$$

using $\lim_{n\to\infty}(1 + \frac{a}{n})^n = e^a$. □

Definition 12.7 (Poisson distribution)**.** $X \sim \mathrm{Poisson}(\mu)$ if

$$\mathbb{P}\{X = x\} = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, 2, \ldots$$

Remark 12.8 (Sanity check)**.** $\displaystyle\sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} = e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} = e^{-\mu} e^{\mu} = 1.$ ✓

Theorem 12.9. *If $X \sim \mathrm{Poisson}(\mu)$ then $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \mu$.*

Proof. $\mathbb{E}[X] = \sum_{x\ge1} x \frac{\mu^x}{x!} e^{-\mu} = \mu e^{-\mu} \sum_{x\ge1} \frac{\mu^{x-1}}{(x-1)!} = \mu e^{-\mu} e^{\mu} = \mu$ (the $x = 0$ term contributes nothing, and $x/x! = 1/(x-1)!$). The same trick twice gives

$$\mathbb{E}[X(X-1)] = \sum_{x\ge2} x(x-1) \frac{\mu^x}{x!} e^{-\mu} = \mu^2 e^{-\mu} \sum_{x\ge2} \frac{\mu^{x-2}}{(x-2)!} = \mu^2 e^{-\mu} e^{\mu} = \mu^2,$$

so $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \mu^2 + \mu$ and $\mathrm{Var}(X) = \mu^2 + \mu - \mu^2 = \mu$. □

Remark 12.10. Mean = variance is the fingerprint of the Poisson. Count data whose variance noticeably exceeds its mean ("overdispersed") is telling you the Poisson model is wrong.

12.3 The Poisson process

The Poisson distribution arose above as a limit of counting. It also arises directly, from a short list of assumptions about events in *time*.

Think of calls arriving at a call center. Suppose:

- In a short interval of length Δt, the probability that *exactly one* call arrives is
$$\alpha\,\Delta t + o(\Delta t),$$
where $o(\Delta t)$ denotes any quantity with $\lim_{\Delta t\to 0} \frac{o(\Delta t)}{\Delta t} = 0$;
- the probability that *two or more* calls arrive in Δt is $o(\Delta t)$;
- the number of calls arriving in Δt is independent of the number that arrived before this interval.

Definition 12.11 (Poisson process)**.** A stream of events satisfying the three conditions above is a **Poisson process** with rate α.

Theorem 12.12. *Let $P_k(t)$ be the probability that exactly k events occur in an interval of length t. Then*
$$P_k(t) = \frac{(\alpha t)^k}{k!} e^{-\alpha t},$$
i.e. the count over an interval of length t is Poisson(μ) *with $\mu = \alpha t$.*

Remark 12.13. So α is the expected number of events *per unit time*, and $\mu = \alpha t$ is the expected number in a window of length t. The two roles of the parameter are the usual source of confusion: "4 calls per hour" is $\alpha = 4$; over 15 minutes the count is Poisson(1), not Poisson(4).

Example 12.14. A call center receives calls as a Poisson process with $\alpha = 4$ per hour.
$$\mathbb{P}\{\text{no calls in 30 minutes}\} = P_0(0.5) = e^{-2} \approx 0.135,$$
$$\mathbb{P}\{\text{at least 2 calls in an hour}\} = 1 - P_0(1) - P_1(1) = 1 - e^{-4} - 4e^{-4} \approx 0.908.$$

Remark 12.15. The waiting *time* between consecutive events of a Poisson process is exponentially distributed — we prove this in Lecture 15, and it is the cleanest link between the discrete and continuous halves of the course.

12.4 Summary: the discrete families

Distribution	**Scenario**	**Support**	**pmf** $\mathbb{P}\{X=k\}$	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	single trial, success rate p	$k \in \{0,1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial(n,p)	number of S in n i.i.d. Bernoulli trials	$k=0,\dots,n$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Hypergeometric(n,M,N)	sampling *without* replacement: N total, M good, n drawn	k feasible	$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$	$n\frac{M}{N}$	$\frac{N-n}{N-1}\,n\frac{M}{N}\Big(1-\frac{M}{N}\Big)$
Geometric(p)	number of trials until the 1st S	$k=1,2,\dots$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
NegBinomial(r,p)	number of F before the r-th S	$k=0,1,2,\dots$	$\binom{k+r-1}{k}p^r(1-p)^k$	$r\Big(\frac{1}{p}-1\Big)$	$\frac{r(1-p)}{p^2}$
Poisson(μ)	number of occurrences in a window of a Poisson process	$k=0,1,2,\dots$	$\frac{\mu^k}{k!}e^{-\mu}$	μ	μ

Remark 12.16 (How they connect)**.**

$$\text{Bernoulli}(p) \xrightarrow[\text{count } S]{n \text{ trials}} \text{Binomial}(n,p) \xrightarrow[np=\mu]{n\to\infty,\ p\to 0} \text{Poisson}(\mu), \qquad \text{Binomial}(n,p) \xleftarrow[n\ll N]{} \text{Hypergeometric}(n,M,N$$

Geometric(p) = NegBinomial($1,p$) (up to the shift), NegBinomial(r,p) = sum of r independent geometrics (c

Everything on this page comes from one experiment — a coin with bias p — and differs only in what is held fixed and what is counted.

Chapter 13

Continuous Random Variables: pdf, cdf, Percentiles

Lecture 13 | Devore 9e: 4.1–4.2

13.1 What goes wrong with a pmf

A random variable $X : S \to \mathbb{R}$ is **discrete** if its values form a finite or countable set, and **continuous** if they fill intervals — $[a, b]$, $[a, \infty)$, or a union such as $(a, b) \cup (c, \infty)$.

Example 13.1 (Height in the Himalayas)**.** Pick a point at random in a mountain region and let X be its elevation, so $X \in [0 \text{ ft}, 29\,029 \text{ ft}]$. What is $\mathbb{P}\{X = 10\,000 \text{ ft}\}$?

Exactly $10\,000.000\ldots$ feet, to infinite precision, is a contour line of zero area. The answer is 0. And the same argument applies to *every* value:

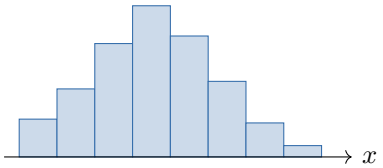
$$\mathbb{P}\{X = c\} = 0 \quad \text{for every } c.$$

So a pmf is useless here — it would be identically zero. But the variable is certainly not trivial: some elevation bands are far more likely than others. The information lives in *intervals*, not in points.

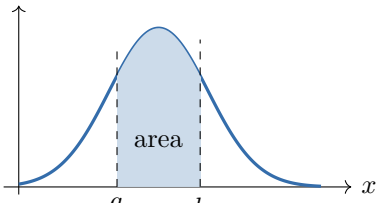
> **Definition 13.2** (Continuous random variable)**.** X is **continuous** if its possible values form intervals and $\mathbb{P}\{X = c\} = 0$ for every $c \in \mathbb{R}$.

From histogram to density

Chop the range into bands $[0, 100]$, $(100, 200]$, $(200, 300]$, … and record the probability of each. That gives a histogram. Now refine the bands: each bar gets shorter, so instead plot *probability per unit length*, i.e. bar height divided by bar width. As the width goes to 0 the tops of the bars trace out a curve. That curve is the density.



coarse bands



the limiting density f

13.2 The probability density function

Definition 13.3 (pdf)**.** The **probability density function** of a continuous random variable X is a function f such that for all $a < b$,

$$\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)\,\mathrm{d}x.$$

Theorem 13.4 (Properties)**.**

$$f(x) \ge 0 \quad \textit{for all } x, \qquad \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Conversely, any f with these two properties is the pdf of some continuous random variable.

Two consequences worth stating separately.

Proposition 13.5. $\mathbb{P}\{X = c\} = 0$ *for every c, and therefore*

$$\mathbb{P}\{a \le X \le b\} = \mathbb{P}\{a < X \le b\} = \mathbb{P}\{a \le X < b\} = \mathbb{P}\{a < X < b\}.$$

Proof. $\mathbb{P}\{X = c\} = \lim_{h\to 0} \int_c^{c+h} f(x)\,\mathrm{d}x = 0.$ □

Remark 13.6 (The crucial warning)**.**

$$\boxed{f(x) \ne \mathbb{P}\{X = x\}}$$

A pdf is *not* a probability. It is a probability *per unit length*, so it can exceed 1 — $f(x) = 5$ on $[0, 0.2]$ is a perfectly good density. What f measures is relative likelihood: for small δ,

$$\mathbb{P}\{c - \delta \le X \le c + \delta\} = \int_{c-\delta}^{c+\delta} f(x)\,\mathrm{d}x \approx 2\delta f(c),$$

so a large $f(c)$ means X is likely to land *near* c, not *at* c.

Discrete	**Continuous**
values countable	values fill intervals
pmf $p(x) = \mathbb{P}\{X = x\}$	pdf $f(x) \ne \mathbb{P}\{X = x\}$
$\sum_i \mathbb{P}\{X = x_i\} = 1$	$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
$\mathbb{P}(A) = \sum_{x_i \in A} p(x_i)$	$\mathbb{P}(A) = \int_A f(x)\,\mathrm{d}x$

13.3 The cdf, again

The cdf is defined exactly as before, and it is what unifies the two cases.

Definition 13.7 (cdf of a continuous random variable)**.**

$$F_X(x) = \mathbb{P}\{X \le x\} = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Theorem 13.8. *Where f is continuous,*

$$f(x) = F_X'(x),$$

by the fundamental theorem of calculus. Moreover

$$\mathbb{P}\{a < X < b\} = F_X(b) - F_X(a), \qquad \mathbb{P}\{X > a\} = 1 - F_X(a).$$

Remark 13.9. Compare with the discrete case: there, F was a step function and the pmf was the size of the jumps. Here F is continuous and the pdf is its slope. Same object, two regimes — which is why Lecture 8 insisted that the cdf, not the pmf, is the fundamental description.

13.4 Percentiles

Definition 13.10 (Percentile)**.** For $0 < p < 1$, the $(100p)$-th percentile of a continuous random variable X is the value x_p with

$$\mathbb{P}\{X \le x_p\} = p, \qquad \text{i.e.} \qquad F_X(x_p) = p.$$

So $x_p = F_X^{-1}(p)$: percentiles are what you get by reading the cdf backwards.

median	50th percentile
first quartile	25th percentile
third quartile	75th percentile

Example 13.11 (Percentiles of a uniform)**.** For $X \sim \text{Uniform}[A, B]$ we will see below that $F(x) = \frac{x-A}{B-A}$ on $[A, B]$. To find the 30th percentile, set $F(x_{0.3}) = 0.3$ and solve:

$$\frac{x_{0.3} - A}{B - A} = 0.3 \qquad \Longrightarrow \qquad x_{0.3} = A + 0.3\,(B - A).$$

So it sits 30% of the way along the interval — exactly what "uniform" should mean. The median is $x_{0.5} = A + \frac{1}{2}(B - A) = \frac{A+B}{2}$.

13.5 Two examples

Example 13.12 (Uniform distribution on $[A, B]$)**.** "Equally likely anywhere in $[A, B]$" means constant density:

$$f(x) = \begin{cases} \dfrac{1}{B-A}, & A \le x \le B, \\ 0, & \text{elsewhere,} \end{cases} \qquad F(x) = \begin{cases} 0, & x < A, \\ \dfrac{x-A}{B-A}, & A \le x \le B, \\ 1, & x > B. \end{cases}$$

The constant $\frac{1}{B-A}$ is forced by $\int f = 1$. Note it exceeds 1 whenever $B - A < 1$.

Example 13.13 (Throwing a dart — density is not uniform even when the dart is)**.** A dart lands uniformly on a disc of radius 1. Let X be the distance from the center. *The dart is uniform on the disc, but X is not uniform on* $[0, 1]$.

Go through the cdf, which is where geometric probability from Lecture 4 pays off:

$$F(x) = \mathbb{P}\{X \le x\} = \frac{\text{area of disc of radius } x}{\text{area of disc of radius } 1} = \frac{\pi x^2}{\pi \cdot 1^2} = x^2, \qquad 0 \le x \le 1.$$

Differentiate:

$$f(x) = F'(x) = 2x, \qquad 0 \le x \le 1.$$

Check: $\int_0^1 2x\,\mathrm{d}x = 1$. ✓ The density increases linearly, because there is more area far from the center than near it.

Remark 13.14. This example is the template for the change-of-variable method in Lecture 15: to find the distribution of a *function* of a random variable, compute its cdf first and differentiate at the end. Do not try to transform densities directly.

13.6 Expectation in the continuous case

Every sum becomes an integral and every pmf becomes a pdf:

$$\underbrace{\sum_i x_i\, p(x_i)}_{\text{discrete}} \qquad \rightsquigarrow \qquad \boxed{\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x}$$

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x, \qquad \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \mu^2.$$

Remark 13.15 (Physical reading)**.** Since $\int f = 1$,

$$\mathbb{E}[X] = \frac{\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x}{\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x},$$

which is precisely the **center of mass** of a rod with density f: the balance point of the pdf. This is a genuinely useful picture — for a symmetric density the mean is the axis of symmetry, no integration required.

Example 13.16. For the dart, $f(x) = 2x$ on $[0, 1]$:

$$\mathbb{E}[X] = \int_0^1 x \cdot 2x\,\mathrm{d}x = \frac{2}{3}, \qquad \mathbb{E}[X^2] = \int_0^1 x^2 \cdot 2x\,\mathrm{d}x = \frac{1}{2}, \qquad \mathrm{Var}(X) = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}.$$

The average distance from the center is 2/3, not 1/2.

13.7 A worked example start to finish

Example 13.17. A continuous random variable has cdf $F(x) = c\,x^2$ for $0 \le x \le 2$ (and the obvious 0 and 1 outside). Find c, the pdf, $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Step 1: find c. A cdf must reach 1 at the top of the range, so $F(2) = 1$:

$$c \cdot 2^2 = 4c = 1 \qquad \Longrightarrow \qquad c = \tfrac{1}{4}, \qquad F(x) = \frac{x^2}{4}, \quad 0 \le x \le 2.$$

Step 2: differentiate to get the pdf.

$$f(x) = F'(x) = \frac{2x}{4} = \frac{x}{2}, \qquad 0 \le x \le 2,$$

and $f(x) = 0$ elsewhere. Check: $\int_0^2 \frac{x}{2}\,\mathrm{d}x = \frac{x^2}{4}\big|_0^2 = 1$. ✓

Step 3: the mean.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \underbrace{\int_{-\infty}^{0} x \cdot 0\,\mathrm{d}x}_{=0} + \int_0^2 x \cdot \frac{x}{2}\,\mathrm{d}x + \underbrace{\int_2^{\infty} x \cdot 0\,\mathrm{d}x}_{=0} = \frac{1}{2} \cdot \frac{x^3}{3}\bigg|_0^2 = \frac{8}{6} = \frac{4}{3}.$$

Step 4: the second moment.

$$\mathbb{E}[X^2] = \int_0^2 x^2 \cdot \frac{x}{2}\,\mathrm{d}x = \frac{1}{2} \cdot \frac{x^4}{4}\bigg|_0^2 = \frac{16}{8} = 2.$$

Step 5: the variance.

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = 2 - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = \frac{2}{9}.$$

Remark 13.18. Step 3 is written out in full to make one point: the integral $\int_{-\infty}^{\infty}$ always splits at the edges of the support, and outside the support $f = 0$ so those pieces vanish. After you have done it once, just integrate over the support — but be sure you know *what* the support is. Note also the sanity check in step 2; a pdf that does not integrate to 1 means an error upstream, and catching it there costs nothing.

Chapter 14

The Normal Distribution

Lecture 14 | Devore 9e: 4.3

Recall the continuous toolkit from last lecture:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \quad \mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x, \quad \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x = \mathbb{E}[X^2] - \mu^2.$$

14.1 Why this distribution and not another

Take $X_1, X_2, \ldots, X_N$ independent and identically distributed — *whatever* their common distribution — and form the average

$$Z = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

As N grows, the distribution of Z looks like the same bell-shaped curve, no matter what the X_i were. Complicated inputs, one universal output. That is the central limit theorem (Lecture 21), and it is the reason the normal distribution is everywhere: any quantity that is the accumulation of many small independent effects — measurement error, height, exam totals — comes out approximately normal.

14.2 Definition

Definition 14.1 (Normal distribution)**.** $X \sim \mathrm{Normal}(\mu, \sigma^2)$ if its pdf is

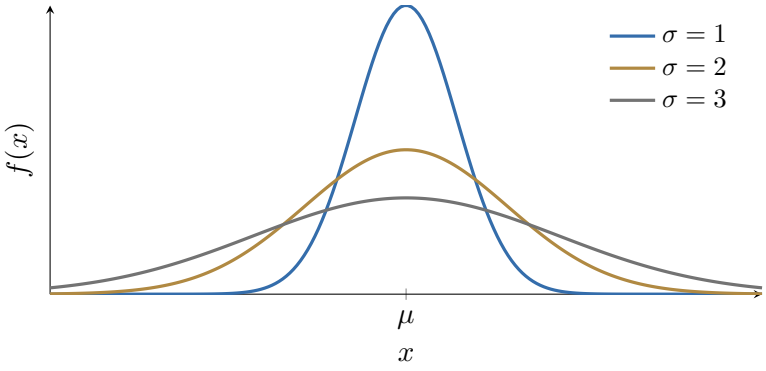
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$.

Three facts to check (all are Homework):

$$\int_{-\infty}^{\infty} f(x;\mu,\sigma)\,\mathrm{d}x = 1, \qquad \mathbb{E}[X] = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

The factor $\frac{1}{\sqrt{2\pi}\sigma}$ is exactly the *normalization* needed to make the total area 1; the shape is carried entirely by $\exp\!\left(-\frac{1}{2}z^2\right)$ with $z = \frac{x-\mu}{\sigma}$.



The curve is symmetric about μ (so the mean, median and mode all coincide there), and larger σ means more spread. Note that the peak height is $\frac{1}{\sqrt{2\pi}\sigma}$, so a very small σ gives a very tall narrow spike — again a reminder that a density is not a probability and may exceed 1.

14.3 The standard normal, and standardization

Definition 14.2 (Standard normal)**.** $Z \sim \text{Normal}(0,1)$, with pdf and cdf

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad \Phi(z) = \mathbb{P}\{Z \le z\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

Remark 14.3. Φ has no closed form in elementary functions — you cannot antidifferentiate $e^{-t^2/2}$ by hand. That is why the standard normal table exists, and why every normal computation is routed through it.

Theorem 14.4 (Standardization)**.** *If* $X \sim \text{Normal}(\mu, \sigma^2)$ *and*

$$Z := \frac{X - \mu}{\sigma},$$

then $Z \sim \text{Normal}(0,1)$.

Proof. $F_Z(z) = \mathbb{P}\{\frac{X-\mu}{\sigma} \le z\} = \mathbb{P}\{X \le \mu + \sigma z\} = F_X(\mu + \sigma z)$. Differentiate in z: $f_Z(z) = \sigma f_X(\mu + \sigma z) = \sigma \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2} = \varphi(z)$. □

This single fact reduces every normal question to a table lookup:

$$\boxed{\mathbb{P}\{X \le x\} = \mathbb{P}\left\{\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right)}$$

and consequently

$$\mathbb{P}\{a \le X \le b\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Remark 14.5 (Reading the table backwards)**.** Percentiles work the same way. If z_p solves $\Phi(z_p) = p$, then the p-th percentile of X is $x_p = \mu + \sigma z_p$. Values worth memorizing:

$$z_{0.90} = 1.28, \qquad z_{0.95} = 1.645, \qquad z_{0.975} = 1.96, \qquad z_{0.99} = 2.33.$$

By symmetry $\Phi(-z) = 1 - \Phi(z)$, which is how negative arguments are handled.

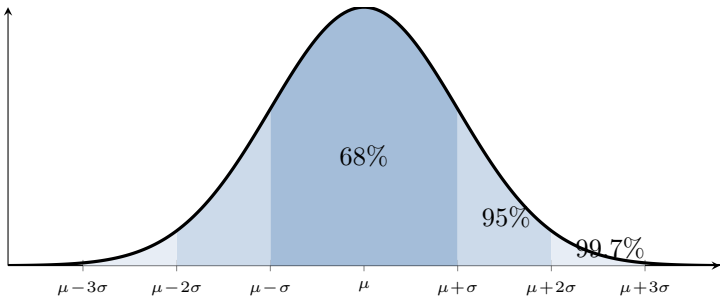
Example 14.6. Exam scores are Normal$(72, 8^2)$. What fraction of students score between 60 and 90?

$$\mathbb{P}\{60 \le X \le 90\} = \Phi\left(\frac{90-72}{8}\right) - \Phi\left(\frac{60-72}{8}\right) = \Phi(2.25) - \Phi(-1.5) = 0.9878 - 0.0668 = 0.9210.$$

14.4 The empirical rule

Theorem 14.7. *If* $X \sim \text{Normal}(\mu, \sigma^2)$ *then*

$$\begin{aligned}\mathbb{P}\{\mu - \sigma \le X \le \mu + \sigma\} &\approx 68\%,\\ \mathbb{P}\{\mu - 2\sigma \le X \le \mu + 2\sigma\} &\approx 95\%,\\ \mathbb{P}\{\mu - 3\sigma \le X \le \mu + 3\sigma\} &\approx 99.7\%.\end{aligned}$$



Remark 14.8. This is worth internalizing because it turns σ into an interpretable number. If a process has $\mu = 60$ and $\sigma = 10$, then roughly 95% of output lies in $[40, 80]$, and a value of 100 is a four-sigma event — about 1 in 30,000. The rule is also the fastest sanity check on an exam: if your answer says $\mathbb{P}\{X > \mu + \sigma\} = 0.40$, you have made an arithmetic error, because it must be about 0.16.

14.5 Looking ahead: the normal approximation to the binomial

If $X \sim \text{Binomial}(n, p)$ then $\mu = np$ and $\sigma^2 = np(1-p)$. For large n the binomial pmf traces out a normal curve, and

$$\mathbb{P}\{X \le x\} \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right).$$

The $+0.5$ is the **continuity correction**: we are approximating a lattice-valued variable by a continuous one, so the discrete point x is represented by the interval $[x - 0.5,\, x + 0.5]$. The usual rule of thumb is that the approximation is good when $np \ge 10$ and $n(1-p) \ge 10$.

We will see in Lecture 21 that this is just the central limit theorem applied to $X = \mathbf{1}_1 + \cdots + \mathbf{1}_n$, where $\mathbf{1}_i$ is the indicator of a success on the i-th trial: a binomial count *is* a sum of n i.i.d. terms.

Chapter 15

Transformations; the Exponential and Gamma Distributions

Lecture 15 | Devore 9e: 4.4

15.1 Change of variable

Given X with density f_X and a function g, set $Y = g(X)$. What is f_Y?

Remark 15.1 (The method)**. Never transform densities directly.** Always go through the cdf:

$$\text{find } F_Y(y) = \mathbb{P}\{Y \le y\} \;\longrightarrow\; \text{rewrite as an event about } X \;\longrightarrow\; \text{differentiate.}$$

Suppose g is monotone, so that g^{-1} exists. Then

$$F_Y(y) = \mathbb{P}\{Y \le y\} = \mathbb{P}\{g(X) \le y\} = \begin{cases} \mathbb{P}\{X \le g^{-1}(y)\} = F_X\big(g^{-1}(y)\big), & g \text{ increasing,} \\ \mathbb{P}\{X \ge g^{-1}(y)\} = 1 - F_X\big(g^{-1}(y)\big), & g \text{ decreasing.} \end{cases}$$

Differentiating with the chain rule,

$$f_Y(y) = \begin{cases} f_X\big(g^{-1}(y)\big)\dfrac{d}{dy}g^{-1}(y), & g \text{ increasing,} \\ -f_X\big(g^{-1}(y)\big)\dfrac{d}{dy}g^{-1}(y), & g \text{ decreasing.} \end{cases}$$

In the increasing case $\frac{d}{dy}g^{-1}(y) > 0$ and in the decreasing case it is < 0, so in both cases the right-hand side is $f_X(g^{-1}(y))$ times the *absolute value* of that derivative. The two cases therefore collapse into one formula.

Theorem 15.2 (Change of variable)**.** *If g is monotone and differentiable with differentiable inverse, then*

$$\boxed{f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|}$$

on the range of g, and $f_Y(y) = 0$ elsewhere.

Remark 15.3. The absolute value of the derivative is a stretching factor: where g compresses the x-axis, probability piles up and the density rises. Do not forget to state the new range — half the lost points on this topic come from a correct formula attached to the wrong interval.

Example 15.4. $X \sim \text{Uniform}[0,1]$ and $Y = -X$, so Y ranges over $[-1,0]$. Directly:

$$F_Y(y) = \mathbb{P}\{-X \le y\} = \mathbb{P}\{X \ge -y\} = 1 - F_X(-y),$$

so

$$f_Y(y) = \frac{d}{dy}\big(1 - F_X(-y)\big) = -F_X'(-y)\cdot(-1) = f_X(-y) = 1$$

for $-1 \le y \le 0$, and 0 elsewhere. So $Y \sim \text{Uniform}[-1,0]$, as it must be.

Example 15.5 (A non-monotone case)**.** If g is not monotone the boxed formula does not apply, but the method still does. For $X \sim \text{Normal}(0,1)$ and $Y = X^2$, with $y > 0$,

$$F_Y(y) = \mathbb{P}\{X^2 \le y\} = \mathbb{P}\{-\sqrt{y} \le X \le \sqrt{y}\} = 2\Phi(\sqrt{y}) - 1,$$

so, differentiating and using the chain rule factor $\frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}$,

$$f_Y(y) = 2\varphi(\sqrt{y})\cdot\frac{1}{2\sqrt{y}} = \frac{\varphi(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \qquad y > 0.$$

This is the χ^2 distribution with one degree of freedom.

15.2 The exponential distribution

Recall the Poisson process with rate λ: the number of arrivals in a window of length t is $\text{Poisson}(\lambda t)$. The Poisson distribution counts arrivals. Now ask the complementary question: *how long until the first arrival?*

Let X be that waiting time. For $t \ge 0$,

$$\mathbb{P}\{X \le t\} = \mathbb{P}\{\text{at least one arrival in } (0,t]\} = 1 - \mathbb{P}\{\text{no arrival in } (0,t]\} = 1 - \frac{(\lambda t)^0}{0!}e^{-\lambda t} = 1 - e^{-\lambda t}.$$

Definition 15.6 (Exponential distribution)**.** $X \sim \text{Exponential}(\lambda)$ if

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0, \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Theorem 15.7. $\mathbb{E}[X] = \dfrac{1}{\lambda}$ *and* $\text{Var}(X) = \dfrac{1}{\lambda^2}$.

Remark 15.8. Both the discrete and continuous pictures of the same process agree: at rate $\lambda = 5$ calls per hour, you expect 5 calls in an hour, and you expect to wait $1/5$ hour $= 12$ minutes for the first one.

15.2.1 The memoryless property

Theorem 15.9 (Memorylessness)**.** *For $X \sim$ Exponential(λ) and any $t_0, t \geq 0$,*

$$\mathbb{P}\{X \geq t_0 + t \mid X \geq t_0\} = \mathbb{P}\{X \geq t\} = e^{-\lambda t}.$$

Proof. Since $\{X \geq t_0 + t\} \subset \{X \geq t_0\}$,

$$\mathbb{P}\{X \geq t_0 + t \mid X \geq t_0\} = \frac{\mathbb{P}\{X \geq t_0 + t\}}{\mathbb{P}\{X \geq t_0\}} = \frac{1 - F(t_0 + t)}{1 - F(t_0)} = \frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}} = e^{-\lambda t}. \qquad \square$$

Remark 15.10. In words: having already waited 10 minutes tells you nothing about how much longer you must wait. The distribution has no memory of the past — which makes it the right model for the lifetime of a component that does not age (electronics, radioactive decay) and the *wrong* model for one that wears out (bearings, tires). The Weibull distribution in Lecture 16 is what you use for the latter.

The exponential is the only continuous distribution with this property; the geometric is its discrete counterpart.

15.3 The gamma distribution

The exponential is the waiting time until the *first* arrival. What about the waiting time until the α-th?

Definition 15.11 (Gamma function)**.** For $\alpha > 0$,

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} \, \mathrm{d}x.$$

Proposition 15.12.

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ *for* $\alpha > 1$*;*
2. $\Gamma(n) = (n-1)!$ *for positive integers* n*;*
3. $\Gamma(1/2) = \sqrt{\pi}$.

Property (1) is integration by parts; (2) follows by induction from (1) with $\Gamma(1) = 1$. So Γ interpolates the factorial to non-integers, which is exactly what is needed to allow a non-integer number of "arrivals".

Definition 15.13 (Gamma distribution)**.** $X \sim$ Gamma(α, β) with shape $\alpha > 0$ and scale $\beta > 0$ if

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} \, x^{\alpha-1} e^{-x/\beta}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The case $\beta = 1$ is the **standard gamma distribution**, $f(x; \alpha, 1) = \dfrac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)}$.

Theorem 15.14. $\mathbb{E}[X] = \alpha\beta$ *and* $\mathrm{Var}(X) = \alpha\beta^2$.

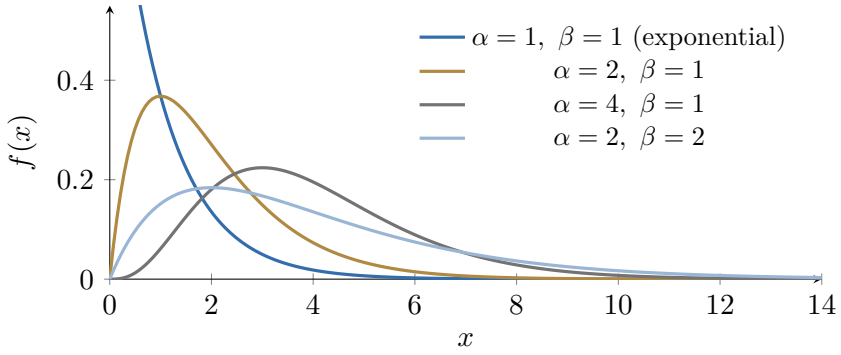
Remark 15.15 (How to read the parameters)**.** Take $\alpha = 1$ and $\beta = 1/\lambda$:

$$f(x; 1, 1/\lambda) = \frac{1}{(1/\lambda)\Gamma(1)} x^0 e^{-\lambda x} = \lambda e^{-\lambda x},$$

the exponential. So Exponential(λ) = Gamma($1, 1/\lambda$), and in general

α	the number of arrivals being waited for
β	the mean waiting time for each single arrival
X	the total waiting time

which makes $\mathbb{E}[X] = \alpha\beta$ and $\mathrm{Var}(X) = \alpha \cdot \beta^2$ immediate: a sum of α independent exponential waits, each with mean β and variance β^2.



Increasing α pushes the mass to the right and makes the shape more symmetric (it is a sum of more waits); increasing β stretches the whole picture horizontally. For $\alpha \leq 1$ the density is decreasing with a spike at the origin; for $\alpha > 1$ it has an interior mode.

Remark 15.16. Two special cases are worth naming now: Gamma($1, 1/\lambda$) = Exponential(λ), and Gamma($\nu/2, 2$) = χ^2_ν, the chi-square distribution with ν degrees of freedom, which reappears in the statistics half of the course.

Chapter 16

Chi-Square, Weibull, Lognormal, Beta

Lecture 16 | Devore 9e: 4.4–4.5

Last time: change of variable, exponential, gamma. Today we finish the gamma and collect the remaining named continuous families. Every one of them is built out of pieces we already have.

16.1 Gamma, continued

Devore writes the gamma density in the equivalent form

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\beta\,\Gamma(\alpha)}\left(\dfrac{x}{\beta}\right)^{\alpha-1}e^{-x/\beta}, & x\ge 0,\\[2mm] 0, & x<0,\end{cases}\qquad \mathbb{E}[X]=\alpha\beta,\quad \operatorname{Var}(X)=\alpha\beta^2.$$

Proposition 16.1 (Scaling)**.** *If* $X\sim\text{Gamma}(\alpha,\beta)$ *then* $\dfrac{X}{\beta}\sim\text{Gamma}(\alpha,1)$.

This is the change of variable of Lecture 15 with $g(x)=x/\beta$: the inverse is $g^{-1}(u)=\beta u$ with $\left|\frac{d}{du}g^{-1}(u)\right|=\beta$, so

$$f_{X/\beta}(u)=\beta f_X(\beta u)=\beta\cdot\frac{(\beta u)^{\alpha-1}e^{-u}}{\beta^{\alpha}\Gamma(\alpha)}=\frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}.$$

So, just as with the normal, one table suffices.

Definition 16.2 (Incomplete gamma function)**.** For the *standard* gamma ($\beta=1$),

$$\text{IGF}(x;\alpha):=\mathbb{P}\{X\le x\}=\frac{1}{\Gamma(\alpha)}\int_0^x y^{\alpha-1}e^{-y}\,\mathrm{d}y.$$

Then for general β, by the scaling property,

$$\mathbb{P}\{X\le x\}=\mathbb{P}\left\{\frac{X}{\beta}\le\frac{x}{\beta}\right\}=\text{IGF}\left(\frac{x}{\beta};\alpha\right),$$

which is looked up in a table exactly as Φ was.

Theorem 16.3 (Additivity)**.** *If* $X_1\sim\text{Gamma}(\alpha_1,\beta)$ *and* $X_2\sim\text{Gamma}(\alpha_2,\beta)$ *are indepen-*

dent (same β!), then

$$X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta).$$

This is just the waiting-time reading: waiting for α_1 arrivals and then for α_2 more is waiting for $\alpha_1 + \alpha_2$ arrivals. It also explains what a *fractional* α means: any Gamma(α, β) splits as a sum of a Gamma$(\lfloor\alpha\rfloor, \beta)$ piece and a Gamma$(\alpha - \lfloor\alpha\rfloor, \beta)$ piece.

Remark 16.4 (Shape)**.** $0 < \alpha < 1$ gives a density with a spike at 0 — the "early failure" regime, appropriate for items that fail immediately or not at all. $\alpha > 1$ gives an interior mode.

16.2 The chi-square distribution

Definition 16.5 (Chi-square)**.** Let $Z_1, \ldots, Z_k$ be i.i.d. Normal$(0, 1)$ and set

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_k^2.$$

Then $Y \sim \chi^2_k$, chi-square with k **degrees of freedom**, with pdf

$$f(y; k) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)}\, y^{k/2-1} e^{-y/2}, & y > 0, \\ 0, & y \le 0. \end{cases}$$

Comparing with the gamma density:

$$\boxed{\chi^2_k = \text{Gamma}\left(\frac{k}{2},\, 2\right)} \qquad \Longrightarrow \qquad \mathbb{E}[Y] = \frac{k}{2} \cdot 2 = k, \qquad \text{Var}(Y) = \frac{k}{2} \cdot 2^2 = 2k.$$

(For $k = 1$ this matches the transformation $Y = Z^2$ we did in Lecture 15.)

Remark 16.6 (Intuition: energy)**.** A gas molecule has velocity $\vec{V} = (V_x, V_y, V_z)$ with each component approximately Normal$(0, 1)$ after scaling. Its kinetic energy is proportional to

$$V_x^2 + V_y^2 + V_z^2 \sim \chi^2_3.$$

Chi-square is what you get when you add up squared fluctuations — which is exactly what a sample variance does, and why this distribution runs the statistics half of the course.

16.3 The Weibull distribution

Definition 16.7 (Weibull)**.** $X \sim \text{Weibull}(\alpha, \beta)$ with shape $\alpha > 0$ and scale $\beta > 0$ if

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^\alpha}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Its cdf is unusually simple:

$$F(x) = 1 - e^{-(x/\beta)^\alpha}, \qquad x \ge 0.$$

Remark 16.8 (Weibull versus gamma — look at the exponent)**.**

$$\text{Weibull:}\quad \frac{\alpha}{\beta}\Big(\frac{x}{\beta}\Big)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\boxed{\alpha}}}$$

$$\text{Gamma:}\quad \frac{1}{\beta\Gamma(\alpha)}\Big(\frac{x}{\beta}\Big)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\boxed{1}}}$$

The polynomial factor is the same and the constants out front differ only in what it takes to normalize; the families differ in whether α appears in the exponent of the exponential. That one change is what makes the Weibull cdf elementary while the gamma cdf requires a table.

Theorem 16.9 (Where it comes from)**.** *If* $T \sim$ Exponential(1) *and* $X := \beta T^{1/\alpha}$*, then* $X \sim$ Weibull(α, β).

Proof. $\mathbb{P}\{X \le x\} = \mathbb{P}\{\beta T^{1/\alpha} \le x\} = \mathbb{P}\{T \le (x/\beta)^\alpha\} = 1 - e^{-(x/\beta)^\alpha}$. □

So the Weibull is a *stretched* exponential waiting time. Setting $\alpha = 1$ recovers Exponential$(1/\beta)$ exactly.

Theorem 16.10.

$$\mathbb{E}[X] = \beta\,\Gamma\Big(1 + \frac{1}{\alpha}\Big), \qquad \mathrm{Var}(X) = \beta^2\left[\Gamma\Big(1 + \frac{2}{\alpha}\Big) - \Gamma\Big(1 + \frac{1}{\alpha}\Big)^2\right].$$

Remark 16.11 (Why engineers use it)**.** The exponential is memoryless, so it cannot describe wear-out. The Weibull can, through α:

$\alpha < 1$	failure rate *decreasing* — infant mortality
$\alpha = 1$	constant failure rate — exponential, no aging
$\alpha > 1$	failure rate *increasing* — wear-out

This is the standard lifetime model in reliability engineering.

16.4 The lognormal distribution

Definition 16.12 (Lognormal)**.** $X \sim$ Lognormal(μ, σ^2) if $\ln X \sim$ Normal(μ, σ^2).

Derive the density by the usual cdf route: for $x > 0$,

$$\mathbb{P}\{X \le x\} = \mathbb{P}\{\ln X \le \ln x\} = \Phi\Big(\frac{\ln x - \mu}{\sigma}\Big),$$

and differentiating (chain rule, $\frac{d}{dx}\ln x = \frac{1}{x}$),

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \cdot \frac{1}{x}, \qquad x > 0.$$

Theorem 16.13.

$$\mathbb{E}[X] = e^{\mu+\sigma^2/2}, \qquad \mathrm{Var}(X) = \left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}.$$

Remark 16.14 (Warning)**.** μ and σ^2 are the mean and variance of $\ln X$, *not* of X. Note $\mathbb{E}[X] = e^{\mu+\sigma^2/2} > e^{\mu}$, which is the median. The lognormal is right-skewed.

Example 16.15 (Multiplicative growth)**.** A stock grows by a random percentage each day:

$$X = X_0(1+R_1)(1+R_2)\cdots(1+R_n).$$

Take logs:

$$\ln X = \ln X_0 + \sum_{i=1}^{n} \ln(1+R_i).$$

The right-hand side is a *sum* of many small independent terms, so by the central limit theorem it is approximately normal — and therefore X itself is approximately lognormal.

Remark 16.16. So: normal is what you get from adding many small effects; lognormal is what you get from *multiplying* them. Prices, incomes, particle sizes, and file sizes are lognormal for this reason.

16.5 The beta distribution

Everything so far has had unbounded support. For a random *proportion* we need a distribution living on $[0,1]$.

Definition 16.17 (Standard beta)**.** $X \sim \mathrm{Beta}(\alpha,\beta)$ if

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

To place it on a general interval $[A,B]$: say Y is beta on $[A,B]$ if $\frac{Y-A}{B-A} \sim \mathrm{Beta}(\alpha,\beta)$.

Theorem 16.18.

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Theorem 16.19 (Where it comes from)**.** *Let* $Y_1 \sim \mathrm{Gamma}(\alpha,1)$ *and* $Y_2 \sim \mathrm{Gamma}(\beta,1)$ *be independent. Then*

$$X := \frac{Y_1}{Y_1+Y_2} \sim \mathrm{Beta}(\alpha,\beta).$$

Remark 16.20. That is the right way to read the parameters: α and β are the "sizes" of two competing gamma quantities, and X is the *share* taken by the first. For example, if spring rainfall is $\mathrm{Gamma}(\alpha,1)$ and the rest of the year is $\mathrm{Gamma}(\beta,1)$, then the fraction of the annual total falling in spring is $\mathrm{Beta}(\alpha,\beta)$. Note $\mathrm{Beta}(1,1) = \mathrm{Uniform}[0,1]$.

16.6 Summary: the continuous families

Distribution	Relation	What it models
Uniform$[A, B]$	—	equally likely anywhere in an interval
Normal(μ, σ^2)	—	sums of many small independent effects
Exponential(λ)	$=$ Gamma$(1, 1/\lambda)$	waiting time for the 1st arrival; memoryless lifetime
Gamma(α, β)	—	waiting time for α arrivals
χ^2_k	$=$ Gamma$(k/2, 2)$	sum of k squared standard normals; "energy"
Weibull(α, β)	$\beta T^{1/\alpha}$, $T \sim$ Exponential(1)	stretched waiting time; lifetimes *with* aging
Lognormal(μ, σ^2)	$\ln X \sim$ Normal(μ, σ^2)	multiplicative growth
Beta(α, β)	$\frac{Y_1}{Y_1+Y_2}$, Y_i gamma	a random proportion in $[0, 1]$

Remark 16.21. Notice how few genuinely independent objects there are. Starting from a Poisson process you get exponential, gamma, chi-square, Weibull and beta; starting from the normal you get chi-square and lognormal. Learning the *relations* is far more economical than memorizing eight densities.

Chapter 17

Review for Midterm 2

Lecture 17 | Devore 9e: Ch. 3.5–4.5

Midterm 2 covers Lectures 12–16 — the remaining discrete families, continuous random variables, and the named continuous distributions — together with the hypergeometric distribution from Lecture 10, since Devore 3.5 treats it alongside the negative binomial. Note that joint distributions are *not* on this exam — they start next lecture.

17.1 What you must be able to do

1. **Change of variable.** Given X with known cdf/pdf and $Y = g(X)$, find the cdf and then the pdf of Y. Always go through the cdf; always state the new range.

$$F_Y(y) = \mathbb{P}\{Y \le y\} = \mathbb{P}\{g(X) \le y\} \;\longrightarrow\; f_Y(y) = F_Y'(y).$$

2. **Use a pdf.** Compute $\mathbb{P}\{a \le X \le b\} = \int_a^b f$, normalize an unknown constant via $\int f = 1$, and find $\mathbb{E}[X]$, $\mathbb{E}[h(X)]$, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
3. **Normal distribution.** Standardize, use the z-table for both probabilities and percentiles, exploit symmetry ($\Phi(-z) = 1 - \Phi(z)$).
4. **Derive** the negative binomial (Lecture 12) and hypergeometric (Lecture 10) pmfs — not just quote them. Both are counting arguments.
5. **Poisson process.** Know which of Poisson / exponential / gamma answers which question:

how many events in time t?	Poisson(λt)
how long until the 1st event?	Exponential(λ) = Gamma$(1, 1/\lambda)$
how long until the α-th event?	Gamma$(\alpha, 1/\lambda)$

17.2 Formula sheet

Continuous random variables

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \qquad f(x) = F'(x), \qquad \mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x,$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2, \qquad f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\tfrac{d}{dy} g^{-1}(y)\right|.$$

Named continuous families

	pdf	$\mathbb{E}$, Var
Uniform$[A, B]$	$\frac{1}{B-A}$ on $[A, B]$	$\frac{A+B}{2}$, $\frac{(B-A)^2}{12}$
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$	μ, σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$, $x \geq 0$	$\frac{1}{\lambda}$, $\frac{1}{\lambda^2}$
Gamma(α, β)	$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$, $x \geq 0$	$\alpha\beta$, $\alpha\beta^2$
$\chi^2_k = \text{Gamma}(\frac{k}{2}, 2)$	—	k, $2k$
Weibull(α, β)	$\frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1} e^{-(x/\beta)^\alpha}$	$\beta\Gamma(1+\frac{1}{\alpha})$, $\beta^2[\Gamma(1+\frac{2}{\alpha}) - \Gamma(1+\frac{1}{\alpha})^2]$
Lognormal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x - \mu)^2/2\sigma^2}$	$e^{\mu+\sigma^2/2}$, $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$, $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Facts worth quoting

$$F_{\text{Exponential}}(x) = 1 - e^{-\lambda x}, \qquad F_{\text{Weibull}}(x) = 1 - e^{-(x/\beta)^\alpha},$$

memorylessness: $\mathbb{P}\{X > s + t \mid X > s\} = \mathbb{P}\{X > t\}$ for the exponential only.

17.3 The three mistakes that cost the most points

1. **Treating $f(x)$ as a probability.** It is a density; it may exceed 1, and $\mathbb{P}\{X = x\} = 0$ always.
2. **Losing the range in a change of variable.** A correct algebraic formula on the wrong interval is worth very little. Track where g maps the support of X.
3. **Mixing up λ and β.** The exponential is often written with rate λ; the gamma with scale $\beta = 1/\lambda$. Write down which convention the problem uses before computing anything.

Remark 17.1. Practice problems: `exams/2026-spring/practice2.tex`, and the Fall and Spring Midterm 2 papers in `exams/`.

Chapter 18

Joint Distributions, Marginals, Conditionals

Lecture 18 | Devore 9e: 5.1–5.2

18.1 Warm-up: linear transformations

Example 18.1 (Curving an exam)**.** Scores X_{old} are to be rescaled by $X_{\text{new}} = aX_{\text{old}} + b$ so that the top score of 100 stays 100 and the class average of 75 becomes 83 (a B). Solve

$$\begin{cases} 100a + b = 100 \\ 75a + b = 83 \end{cases} \quad \Longrightarrow \; 25a = 17, \quad a = 0.68, \quad b = 32.$$

Recall from Lectures 9–10 that this shifts the mean to $a\mu + b$ and *shrinks* the standard deviation to $a\sigma = 0.68\sigma$ — a curve of this kind compresses the class together as well as moving it up.

18.2 From one random variable to two

Everything so far concerned a single X. But the interesting questions are usually about relationships: height *and* weight, auto premium *and* home premium, almonds *and* cashews in the same can. So we now study pairs (X, Y) defined on the same sample space.

18.2.1 The discrete case

Definition 18.2 (Joint pmf)**.** For discrete X, Y on the same sample space,

$$p(x, y) := \mathbb{P}\{X = x \text{ and } Y = y\},$$

which satisfies $p \geq 0$ and $\sum_x \sum_y p(x, y) = 1$. For any set A of pairs,

$$\mathbb{P}\{(X, Y) \in A\} = \sum_{(x,y) \in A} p(x, y).$$

Example 18.3 (Insurance)**.** X is the auto deductible and Y the homeowner's deductible for a randomly chosen customer:

$p(x,y)$	$y=500$	$y=1000$	$y=5000$	$p_X(x)$
$x=100$	0.30	0.05	0	0.35
$x=500$	0.15	0.20	0.05	0.40
$x=1000$	0.10	0.10	0.05	0.25
$p_Y(y)$	0.55	0.35	0.10	1

Definition 18.4 (Marginal pmf)**.**

$$p_X(x)=\sum_y p(x,y), \qquad p_Y(y)=\sum_x p(x,y).$$

Marginals are the row and column sums — literally written in the margins of the table, which is where the name comes from. Note that the marginals do *not* determine the joint pmf: many tables have the same margins.

18.2.2 The continuous case

Definition 18.5 (Joint pdf)**.** $f(x,y)$ is a **joint pdf** if $f\ge 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y=1$; then

$$\mathbb{P}\{(X,Y)\in A\}=\iint_{(x,y)\in A} f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Where a single-variable pdf was a curve and probability was area under it, a joint pdf is a *surface* and probability is the volume under it over a region.

Definition 18.6 (Marginal pdf)**.**

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y, \qquad f_Y(y)=\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x.$$

Example 18.7. $f(x,y)=\frac{6}{5}(x+y^2)$ for $0\le x\le 1$, $0\le y\le 1$. First check that this is a density: $\int_0^1\int_0^1 \frac{6}{5}(x+y^2)\,\mathrm{d}y\,\mathrm{d}x=\frac{6}{5}\left(\frac{1}{2}+\frac{1}{3}\right)=1$. ✓ Then

$$\mathbb{P}\Big\{0\le X\le \tfrac14,\ 0\le Y\le \tfrac14\Big\}=\int_0^{1/4}\int_0^{1/4}\tfrac{6}{5}(x+y^2)\,\mathrm{d}y\,\mathrm{d}x=\tfrac{6}{5}\left(\tfrac{1}{128}+\tfrac{1}{768}\right)=\frac{7}{640}\approx 0.0109.$$

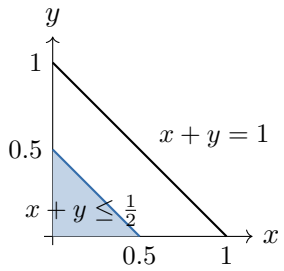
Example 18.8 (Can of nuts — the running example)**.** A 1-lb can of "deluxe mixed nuts" contains almonds, cashews and peanuts. Let X be the weight of almonds and Y the weight of cashews, so $1-X-Y$ is the weight of peanuts. Take

$$f(x,y)=24xy \quad \text{for } x,y\ge 0,\ x+y\le 1,$$

and 0 otherwise. Note the support is a *triangle*, not a rectangle — this is where most of the difficulty lives. It really is a density:

$$\int_0^1 \int_0^{1-x} 24xy \, dy \, dx = \int_0^1 12x(1-x)^2 \, dx = 1. \quad \checkmark$$

What is $\mathbb{P}\{X + Y \le 0.5\}$? The region is the triangle below the line $y = -x + \frac{1}{2}$:



$$\mathbb{P}\{X + Y \le 0.5\} = \int_0^{1/2} \left(\int_0^{-x+1/2} 24xy \, dy \right) dx = \int_0^{1/2} 12x\left(\tfrac{1}{2} - x\right)^2 dx = \frac{1}{16} = 0.0625.$$

Remark 18.9. The hard part of every joint-pdf problem is the *limits of integration*, not the integrand. Draw the region first. Always.

18.3 Independence

Definition 18.10 (Independent random variables)**.** X and Y are **independent** if

$$\text{(discrete)} \quad p(x,y) = p_X(x)\, p_Y(y) \qquad \text{or} \qquad \text{(continuous)} \quad f(x,y) = f_X(x)\, f_Y(y)$$

for all x, y. More generally $X_1, \dots, X_n$ are independent if every subcollection has joint pmf/pdf equal to the product of its marginals.

Example 18.11. $X \sim \text{Exponential}(\lambda_1)$ and $Y \sim \text{Exponential}(\lambda_2)$ independent gives

$$f(x,y) = \lambda_1 e^{-\lambda_1 x} \cdot \lambda_2 e^{-\lambda_2 y} = \lambda_1 \lambda_2 e^{-(\lambda_1 x + \lambda_2 y)}, \qquad x, y \ge 0,$$

so, for instance, $\mathbb{P}\{X \ge 1500,\, Y \ge 1500\} = e^{-1500\lambda_1} e^{-1500\lambda_2}$.

Remark 18.12 (A fast test for dependence)**.** In the can-of-nuts example $f(x,y) = 24xy$ *looks* like a product — but the support $\{x + y \le 1\}$ is not a rectangle, so knowing $X = 0.9$ forces $Y \le 0.1$. The variables are dependent. Rule of thumb: if the support is not a product region, the variables cannot be independent no matter how the formula factors.

Example 18.13 (Maxima and minima)**.** If $X_1, \dots, X_n$ are independent, then for $T_{\max} = \max(X_1, \dots, X_n)$,

$$\mathbb{P}\{T_{\max} \le t\} = \mathbb{P}\{X_1 \le t, \dots, X_n \le t\} = \prod_{i=1}^{n} F_{X_i}(t),$$

and similarly $\mathbb{P}\{T_{\min} > t\} = \prod_i \big(1 - F_{X_i}(t)\big)$. This is the standard way to handle "the first component to fail" and "the last one standing".

18.4 Conditional distributions

Definition 18.14 (Conditional pdf/pmf)**.** For $f_X(x) > 0$,

$$f_{Y|X}(y \mid x) := \frac{f(x,y)}{f_X(x)}, \qquad \text{and symmetrically} \qquad f_{X|Y}(x \mid y) := \frac{f(x,y)}{f_Y(y)}.$$

The discrete version replaces f by p throughout.

Compare with $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ from Lecture 5: same shape, with the joint density playing the role of $\mathbb{P}(A \cap B)$ and the marginal that of $\mathbb{P}(B)$.

For each fixed x, $f_{Y|X}(\cdot \mid x)$ is a genuine pdf in y, so it can be integrated like any other:

$$\mathbb{P}\{Y \le 0.5 \mid X = 0.2\} = \int_{-\infty}^{0.5} f_{Y|X}(y \mid 0.2)\,\mathrm{d}y.$$

Remark 18.15. Conditioning on $\{X = 0.2\}$ is conditioning on an event of probability zero, which Lecture 5 did not allow. The definition above is how that is made legitimate: it is a limit of conditioning on $\{0.2 - \delta \le X \le 0.2 + \delta\}$.

Proposition 18.16. *X and Y are independent if and only if $f_{Y|X} = f_Y$ and $f_{X|Y} = f_X$; that is, conditioning changes nothing.*

Example 18.17 (Can of nuts, conditionals)**.** With $f(x,y) = 24xy$ on the triangle,

$$f_X(x) = \int_0^{1-x} 24xy\,\mathrm{d}y = 12x(1-x)^2, \qquad 0 \le x \le 1,$$

so for $0 < x < 1$,

$$f_{Y|X}(y \mid x) = \frac{24xy}{12x(1-x)^2} = \frac{2y}{(1-x)^2}, \qquad 0 \le y \le 1 - x.$$

Given that the can holds x pounds of almonds, the cashew weight is Beta$(2,1)$-shaped on $[0, 1-x]$ — and the range depends on x, which is dependence made visible.

Chapter 19

Expected Values, Covariance and Correlation

Lecture 19 | Devore 9e: 5.2

19.1 Expectation of a function of two variables

Theorem 19.1. *For $h : \mathbb{R}^2 \to \mathbb{R}$,*

$$\mathbb{E}[h(X,Y)] = \begin{cases} \displaystyle\sum_x \sum_y h(x,y)\, p(x,y), & \textit{discrete,} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y, & \textit{continuous.} \end{cases}$$

Example 19.2 (Seats)**.** Five seats in a row are numbered $1, \dots, 5$. You and one other person sit in two distinct seats chosen at random; X is your seat, Y theirs. So

$$p(x,y) = \begin{cases} \frac{1}{20}, & x, y \in \{1,\dots,5\},\ x \neq y, \\ 0, & \text{else,} \end{cases}$$

since there are $5 \times 4 = 20$ equally likely ordered pairs. Let $h(x,y) = |x-y| - 1$ be the number of seats *between* you. Then

$$\mathbb{E}[h(X,Y)] = \sum_x \sum_y h(x,y)p(x,y) = \frac{1}{20}\sum_{x\neq y}\big(|x-y|-1\big) = \frac{20}{20} = 1.$$

The sum is easiest to organize by the gap $d = |x-y|$: there are $2(5-d)$ ordered pairs with gap d, each contributing $d-1$, so

$$\sum_{x\neq y}\big(|x-y|-1\big) = \sum_{d=1}^{4} 2(5-d)(d-1) = 8\cdot 0 + 6\cdot 1 + 4\cdot 2 + 2\cdot 3 = 20.$$

Example 19.3 (Cost of a can of nuts)**.** Almonds cost \$1.50/lb, cashews \$2.25/lb, peanuts \$0.75/lb. With X, Y the almond and cashew weights and $1 - X - Y$ the peanut weight, the cost of a can is

$$h(X,Y) = 1.5X + 2.25Y + 0.75(1 - X - Y) = 0.75 + 0.75X + 1.5Y.$$

With $f(x,y) = 24xy$ on the triangle,

$$\mathbb{E}[h(X,Y)] = \int_0^1 \int_0^{1-x} \big(0.75 + 0.75x + 1.5y\big)\, 24xy \,\mathrm{d}y \,\mathrm{d}x = \$1.65.$$

Remark 19.4 (Shortcut)**.** Because h is *linear*, the double integral was avoidable. Using the marginal means $\mathbb{E}[X] = \mathbb{E}[Y] = \frac{2}{5}$ computed later this lecture,

$$\mathbb{E}[h(X,Y)] = 0.75 + 0.75\,\mathbb{E}[X] + 1.5\,\mathbb{E}[Y] = 0.75 + 0.30 + 0.60 = \$1.65,$$

and note that this needed *no* assumption of independence — expectation is linear regardless. Doing the integral once in full is still worth it, so that you trust the shortcut when you use it later.

19.2 Conditional expectation

Definition 19.5 (Conditional expectation)**.**

$$\mathbb{E}[h(X) \mid Y = y] = \int_{-\infty}^{\infty} h(x)\, f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Remark 19.6. This is a *function of* y, not a number: for each value of Y you get a different average. Writing $\mathbb{E}[h(X) \mid Y]$ (with a capital Y) therefore denotes a random variable.

Theorem 19.7 (Law of total expectation / tower property)**.**

$$\mathbb{E}\big[\ \mathbb{E}[h(X) \mid Y]\ \big] = \mathbb{E}[h(X)].$$

Proof. The $f_Y(y)$ cancels, and then the order of integration is swapped:

$$\mathbb{E}\big[\mathbb{E}[h(X) \mid Y]\big] = \int \left(\int h(x) \frac{f(x,y)}{f_Y(y)} \,\mathrm{d}x \right) f_Y(y)\,\mathrm{d}y = \iint h(x) f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int h(x) \left(\int f(x,y)\,\mathrm{d}y \right) \mathrm{d}x = \int h(x) f$$

□

This is the expectation version of the law of total probability: average within each case, then average the case averages.

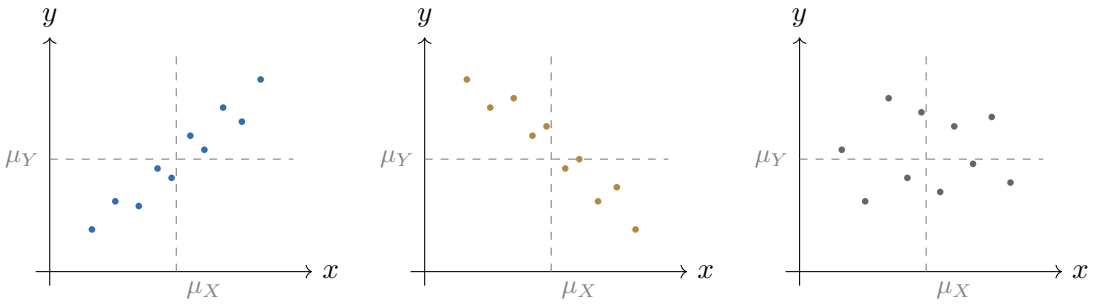
19.3 Covariance

Variance measured how far X strays from its own mean. Covariance measures whether X and Y stray *together*.

Definition 19.8 (Covariance)**.**

$$\mathrm{Cov}(X,Y) := \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big] = \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y) p(x,y), & \text{discrete,} \\ \displaystyle\iint (x - \mu_X)(y - \mu_Y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y, & \text{continuous.} \end{cases}$$

Remark 19.9 (Intuition)**.** The product $(X-\mu_X)(Y-\mu_Y)$ is positive when both variables are on the same side of their means and negative when they are on opposite sides. Averaging picks out which happens more often.



$\mathrm{Cov} > 0$ $\mathrm{Cov} < 0$ $\mathrm{Cov} = 0$

Theorem 19.10 (Properties)**.** *1.* $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$*;*

2. $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y$ *(the computational formula);*

3. $\mathrm{Cov}(aX+b,\, cY+d) = ac\,\mathrm{Cov}(X,Y)$*;*

4. $X \perp\!\!\!\perp Y \implies \mathrm{Cov}(X,Y) = 0$. ***The converse is false.***

Property (2) gives the recipe: from the joint pdf get $\mathbb{E}[XY]$ by a double integral; from the marginals get μ_X and μ_Y; subtract.

$$f(x,y) \longrightarrow \mathbb{E}[XY], \qquad f(x,y) \longrightarrow f_X, f_Y \longrightarrow \mu_X, \mu_Y.$$

Example 19.11 (Can of nuts)**.** $f(x,y) = 24xy$ on $\{x, y \geq 0,\ x+y \leq 1\}$.

$$\mathbb{E}[XY] = \int_0^1 \int_0^{1-x} xy \cdot 24xy \,\mathrm{d}y\,\mathrm{d}x = 8\int_0^1 x^2(1-x)^3\,\mathrm{d}x = \frac{2}{15},$$

$$f_X(x) = 12x(1-x)^2, \quad f_Y(y) = 12y(1-y)^2 \implies \mu_X = \mu_Y = \int_0^1 12x^2(1-x)^2\,\mathrm{d}x = 12\cdot\frac{1}{30} = \frac{2}{5}.$$

Therefore

$$\mathrm{Cov}(X,Y) = \frac{2}{15} - \frac{2}{5}\cdot\frac{2}{5} = \frac{10}{75} - \frac{12}{75} = \boxed{-\frac{2}{75}}.$$

Negative, as it must be: the can weighs one pound, so more almonds means fewer cashews.

19.4 Correlation

Covariance has units — change pounds to grams and it changes by a factor of 454^2. The strength of a relationship should not depend on the choice of units.

Definition 19.12 (Correlation coefficient)**.**

$$\mathrm{Corr}(X,Y) = \rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X,X)}\sqrt{\mathrm{Cov}(Y,Y)}}.$$

Theorem 19.13 (Properties). *1.* $\mathrm{Corr}(aX+b,\, cY+d) = \mathrm{Corr}(X,Y)$ *when* $a, c > 0$ — *units do not matter;*

2. $-1 \le \rho_{X,Y} \le 1$*;*

3. $\rho_{X,Y} = \pm 1$ *if and only if* $Y = aX + b$ *exactly, for some* $a \neq 0$*;*

4. $X \perp\!\!\!\perp Y \Rightarrow \rho = 0$ *("uncorrelated"), but not conversely.*

Example 19.14 (Property 3 in action — the slope is invisible). Let $Y = 100X + 2$. Then $\mathrm{Cov}(X,Y) = 100\,\mathrm{Var}(X)$ by property 3 of covariance, and $\sigma_Y = 100\,\sigma_X$, so

$$\rho_{X,Y} = \frac{100\,\mathrm{Var}(X)}{\sigma_X \cdot 100\,\sigma_X} = \frac{100\sigma_X^2}{100\sigma_X^2} = 1.$$

Now let $Y = -0.00001X + 10{,}000$. The slope is minuscule and the intercept enormous, yet by the same cancellation ($\sigma_Y = 0.00001\,\sigma_X$)

$$\rho_{X,Y} = \frac{-0.00001\,\mathrm{Var}(X)}{\sigma_X \cdot 0.00001\,\sigma_X} = -1.$$

Correlation records only *whether* the relationship is exactly linear and in which direction — never how steep it is. A nearly flat line and a nearly vertical one both give $|\rho| = 1$.

Remark 19.15. The last expression in the definition should look familiar: it is

$$\cos\theta = \frac{\langle u, v\rangle}{\|u\|\,\|v\|}$$

with covariance as the inner product. Correlation is the cosine of the angle between the centered variables — which is exactly why it lands in $[-1,1]$ and why ± 1 means perfect alignment.

Example 19.16 (Uncorrelated but dependent). Let (X,Y) take the four values $(-4,1), (4,-1), (2,2), (-2,-2)$, each with probability $\frac{1}{4}$. Then $\mu_X = \mu_Y = 0$ and

$$\mathbb{E}[XY] = \tfrac{1}{4}\big(-4 + (-4) + 4 + 4\big) = 0,$$

so $\mathrm{Cov}(X,Y) = 0$ and $\rho = 0$. But X and Y are wildly dependent: knowing $X = -4$ tells you $Y = 1$ exactly.

Remark 19.17 (The one-sentence summary). ρ measures the strength of a **linear** relationship only. A perfect non-linear relationship can have $\rho = 0$. So "uncorrelated" is strictly weaker than "independent", and $\rho = 0$ is never evidence of independence.

Remark 19.18 (Looking ahead). Writing $Y = aX + b + \varepsilon$ with ε a noise term and asking for the a, b that fit best is **linear regression**. The answer turns out to be $a = \rho\,\sigma_Y/\sigma_X$, which is the sense in which ρ "is" the slope.

Chapter 20

Samples, Statistics, and Sampling Distributions

Lecture 20 | Devore 9e: 5.3

This lecture is the hinge of the course. Until now the distribution was given and we computed probabilities. From here on the distribution is *unknown* and we have only data — but the data is itself random, so probability is exactly the tool we need to describe it.

20.1 Samples and statistics

Definition 20.1 (Sample and sample data)**.** A **sample** is a collection of random variables $X_1, X_2, \dots, X_n$. The **sample data** $x_1, x_2, \dots, x_n \in \mathbb{R}$ is a collection of observed values of that sample.

The distinction is the whole point. Before you run the experiment, X_i is a random variable; afterwards, x_i is a number. Capital letters are random, lower case are data.

Definition 20.2 (Statistic)**.** A **statistic** is any quantity whose value can be computed from the sample data — equivalently, any random variable expressible as a function of $X_1, \dots, X_n$.

Example 20.3. The two we care about most:

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n} \qquad \text{(sample mean)},$$

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{(sample variance)}, \qquad S = \sqrt{S^2}.$$

But there are many others — $\frac{X_1}{X_1+X_2}$, $e^{X_1+X_2^2}$, $\left(\sum_i X_i^k\right)^{1/k}$ — and each is a legitimate statistic.

Remark 20.4 (Why $n-1$?)**.** Dividing by $n-1$ rather than n makes $\mathbb{E}[S^2] = \sigma^2$ exactly, for every n. We prove this in Lecture 23; the short reason is that $\bar{X}$ has already been estimated from the same data, using up one "degree of freedom".

Remark 20.5 (The central observation)**.**

A statistic is a random variable, so it has a distribution.

That distribution is called the **sampling distribution** of the statistic, and essentially all of statistics consists of working out what it is.

20.2 What the sampling distribution depends on

(1) The underlying distribution of the sample

If every X_i is identically 0, then $\bar{X} = 0$ always — a degenerate sampling distribution. If $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$, the answer is quite different.

(2) The method of sampling

Suppose you measure the heights $H_1, \ldots, H_{100}$ of 100 people and form $\bar{H}$. If you sampled 100 strangers, $\bar{H}$ behaves one way. If you sampled 100 members of the same family, the H_i are correlated and $\bar{H}$ is far more variable than it looks.

Push it to the extreme to see how badly this can go. Suppose $X_1 \sim \text{Normal}(0, 1)$ but you accidentally record the *same* person n times, so

$$X_2 = X_1, \quad X_3 = X_1, \quad \ldots, \quad X_n = X_1.$$

Then

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{nX_1}{n} = X_1,$$

and therefore

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_1 - X_1)^2 = 0 \qquad \Longrightarrow \qquad S^2 = 0.$$

The sample reports zero variability, with total confidence, about a population whose variance is 1. Nothing in the formula for S^2 went wrong; the *sampling* did. No amount of data fixes this — taking $n = 10^6$ copies of the same person still gives $S^2 = 0$.

Definition 20.6 (Random sample of size n)**.** $X_1, \ldots, X_n$ is a **random sample** (equivalently: the X_i are **i.i.d.**) if

1. $X_1, \ldots, X_n$ are independent, and
2. each X_i has the same distribution.

From now on "sample" means random sample unless stated otherwise. Both conditions are assumptions about how the data was *collected*, and no amount of clever computation can repair a violation of them.

20.3 Two ways to find a sampling distribution

1. **Closed-form derivation.** Works when the underlying distribution and the statistic are both simple.
2. **Numerical simulation.** Draw many samples, compute the statistic each time, histogram the results. This is the only option once things get complicated — and it always works.

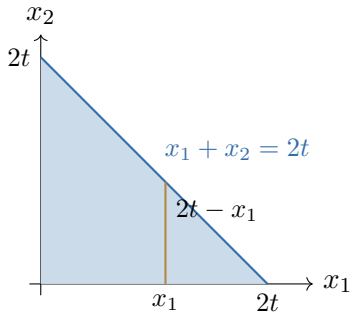
20.3.1 A closed-form example

Example 20.7. Let X_1, X_2 be i.i.d. Exponential(λ) and $\bar{X} = \frac{X_1+X_2}{2}$. Find the cdf and pdf of $\bar{X}$.

As always, go through the cdf. For $t > 0$,

$$F_{\bar{X}}(t) = \mathbb{P}\{\bar{X} \le t\} = \mathbb{P}\{X_1 + X_2 \le 2t\} = \iint_{\{x_1+x_2\le 2t\}} f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2.$$

By independence $f(x_1, x_2) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2}$, and the region is the triangle $\{x_1, x_2 \ge 0,\ x_1 + x_2 \le 2t\}$:



Slice vertically:

$$\begin{aligned} F_{\bar{X}}(t) &= \int_0^{2t} \left(\int_0^{2t-x_1} \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2}\, \mathrm{d}x_2 \right) \mathrm{d}x_1 \\ &= \int_0^{2t} \lambda e^{-\lambda x_1} \left(1 - e^{-\lambda(2t-x_1)}\right) \mathrm{d}x_1 \\ &= \int_0^{2t} \left(\lambda e^{-\lambda x_1} - \lambda e^{-2\lambda t}\right) \mathrm{d}x_1 \\ &= 1 - e^{-2\lambda t} - 2\lambda t\, e^{-2\lambda t}, \qquad t \ge 0. \end{aligned}$$

Differentiating,

$$f_{\bar{X}}(t) = F'_{\bar{X}}(t) = 4\lambda^2 t\, e^{-2\lambda t}, \qquad t \ge 0.$$

Remark 20.8. Notice what happened: the average of two exponentials is *not* exponential — its density vanishes at $t = 0$ and has an interior peak. In fact $X_1 + X_2 \sim \text{Gamma}(2, 1/\lambda)$, exactly as the additivity property of Lecture 16 predicts, and $\bar{X}$ is that divided by 2.

Also notice how much work this was for $n = 2$ with the simplest possible underlying distribution. Doing it for $n = 100$, or for a distribution without a nice closed form, is hopeless. That is why the next lecture's theorem is so valuable: it gives the approximate sampling distribution of $\bar{X}$ for *every* underlying distribution, for free.

Remark 20.9 (Sanity checks that cost nothing)**.** Whatever the underlying distribution, if $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2 then

$$\mathbb{E}[\bar{X}] = \mu, \qquad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

We prove these next lecture. Check them against the example above: Exponential(λ) has $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$, so we should find $\mathbb{E}[\bar{X}] = 1/\lambda$ and $\text{Var}(\bar{X}) = 1/(2\lambda^2)$. Integrating $f_{\bar{X}}(t) = 4\lambda^2 t e^{-2\lambda t}$ against t and t^2, using $\int_0^\infty t^k e^{-2\lambda t}\, \mathrm{d}t = k!/(2\lambda)^{k+1}$,

$$\mathbb{E}[\bar{X}] = 4\lambda^2 \cdot \frac{2}{(2\lambda)^3} = \frac{1}{\lambda}, \qquad \mathbb{E}[\bar{X}^2] = 4\lambda^2 \cdot \frac{6}{(2\lambda)^4} = \frac{3}{2\lambda^2},$$

so $\text{Var}(\bar{X}) = \frac{3}{2\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{2\lambda^2} = \sigma^2/2$. ✓

Chapter 21

Simulation and the Central Limit Theorem

Lecture 21 | Devore 9e: 5.4–5.5

Last time: samples, statistics, and the closed-form derivation of the sampling distribution of $\bar{X}$ for two exponentials. That derivation was already unpleasant. Today: what to do when it is impossible.

21.1 Numerical simulation

Closed form works when both the underlying distribution and the statistic are simple. It fails as soon as

- n is large — imagine the region $\{x_1 + \cdots + x_{100} \le 100t\}$;
- the underlying distribution is complicated;
- the statistic is awkward, e.g. $|X_1| + |X_2| + \cdots + |X_n|$ or the rank of X_1 among $X_1, \ldots, X_n$.

Simulation handles all of them. The setup:

- a statistic $T = T(X_1, \ldots, X_n)$ — e.g. $\bar{X}$ or S;
- a sample distribution, which we can *draw from* even if we cannot integrate against it (a black box);
- the goal: $F_T(t) = \mathbb{P}\{T \le t\}$, or the pdf of T.

The algorithm. Repeat k times:

$$\begin{aligned} t_1 &= T(x_1, \ldots, x_n) \quad \text{from a freshly generated sample} \\ t_2 &= T(x_1, \ldots, x_n) \quad \text{from another fresh sample} \\ &\vdots \\ t_k &= T(x_1, \ldots, x_n) \end{aligned}$$

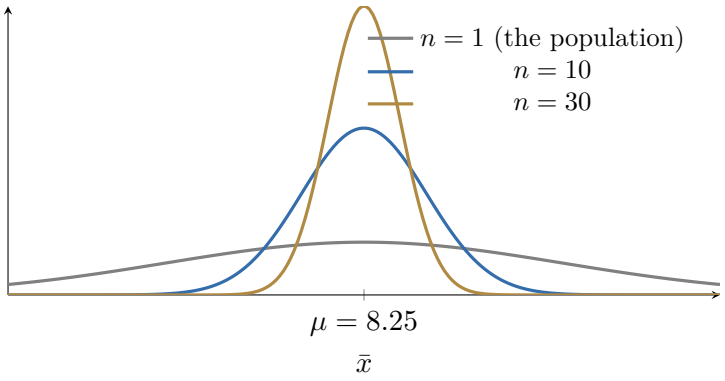
Then histogram $t_1, \ldots, t_k$. That histogram *is* an estimate of the pdf of T, and the proportion of t_i's below t estimates $F_T(t)$.

Example 21.1. Take $X_i \sim \text{Normal}(\mu, \sigma^2)$ with $\mu = 8.25$, $\sigma = 0.75$, sample size $n = 10$, and $k = 500$ repetitions. Each repetition gives one number $\bar{x}_j = (x_1 + \cdots + x_{10})/10$; we get 500 of them — 8.0, 8.1, 7.9, …, 8.6, 7.5 — and histogram them.

Remark 21.2 (Two knobs, two different effects)**.**

larger k	more *accuracy*: the histogram converges to the true pdf of T, whatever it is.
larger n	the pdf of T *itself changes*: the peak stays at μ but the spread shrinks.

Confusing these two is the classic error. Increasing k tells you more about the same object; increasing n changes the object.



21.2 Mean and variance of the sample mean

Before the theorem, two computations that need no assumptions beyond i.i.d.

Theorem 21.3. *If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then*

$$\mathbb{E}[\bar{X}] = \mu, \qquad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Proof. By linearity of expectation (which needs no independence),

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu.$$

For the variance we *do* need independence, which makes variances add:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\,\text{Var}(X_1 + \cdots + X_n) = \frac{\text{Var}(X_1) + \cdots + \text{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$ □

Remark 21.4 (The $\sqrt{n}$ law)**.** The precision of an average improves like $\sqrt{n}$, not like n. To halve your error bar you must *quadruple* the sample size. This single fact governs the cost of every survey and every experiment.

21.3 The central limit theorem

Simulation shows something striking: whatever you start with — Weibull, uniform, something lumpy and skewed — the histogram of $\bar{X}$ comes out bell-shaped once n is large enough.

Theorem 21.5 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then for large n,*

$$\bar{X} \overset{approx}{\sim} \text{Normal}\left(\mu,\ \frac{\sigma^2}{n}\right), \qquad \textit{equivalently} \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} \text{Normal}(0,1).$$

The mean and variance are exactly what we computed above; the content of the theorem is that the *shape* is normal, regardless of the underlying distribution.

Remark 21.6 (Rule of thumb)**.** $n > 30$ is the usual working threshold. It is only a rule of thumb: a nearly symmetric underlying distribution needs far fewer, and a heavily skewed one (or one with a heavy tail) needs far more. Note the hypothesis $\sigma^2 < \infty$ — for the heavy-tailed pmf of Lecture 9, where even the mean is infinite, the theorem simply does not apply.

Example 21.7 (Why this matters)**.** $X_1, \ldots, X_{1000}$ are i.i.d. from some complicated Weibull. What is $\mathbb{P}\{\bar{X} \le t\}$? Deriving it exactly is hopeless. By the CLT,

$$\mathbb{P}\{\bar{X} \le t\} \approx \Phi\left(\frac{t-\mu}{\sigma/\sqrt{1000}}\right),$$

which is a table lookup.

Example 21.8 (Normal approximation to the binomial, explained)**.** $X \sim \text{Binomial}(n,p)$ is the sum of n i.i.d. Bernoulli variables, each with mean p and variance $p(1-p)$. Applying the CLT to that sum recovers exactly the approximation quoted in Lecture 14:

$$X \overset{\text{approx}}{\sim} \text{Normal}\big(np,\ np(1-p)\big).$$

21.4 The exactly-normal case

When the sample is drawn from a normal population, the approximations become identities and the sample variance gets a distribution too.

Theorem 21.9. *Let $X_1, \ldots, X_n$ be i.i.d.* Normal(μ, σ^2)*. Then*

1. $\bar{X} \sim \text{Normal}\left(\mu, \dfrac{\sigma^2}{n}\right)$ *exactly, for every n;*
2. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$*;*
3. $\bar{X}$ *and* S^2 *are independent.*

Remark 21.10. Part (2) is where the $n-1$ finally earns its keep, and where the name "degrees of freedom" comes from: the n deviations $X_i - \bar{X}$ satisfy one linear constraint, $\sum_i (X_i - \bar{X}) = 0$, so only $n-1$ of them are free. Part (3) — that the sample mean and the sample variance are independent — is special to the normal distribution, and it is what makes the t-statistic work in a later course.

Chapter 22

Linear Combinations; Point Estimation and Bias

Lecture 22 | Devore 9e: 5.5, 6.1

22.1 Linear combinations

Let $X_1, \dots, X_n$ have means $\mathbb{E}[X_i] = \mu_i$ and known covariances $\mathrm{Cov}(X_i, X_j)$, and consider the statistic

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n.$$

Theorem 22.1 (Mean)**.**

$$\mathbb{E}[Y] = \sum_{i=1}^{n} a_i \, \mathbb{E}[X_i] = \sum_{i=1}^{n} a_i \mu_i.$$

This needs **no** independence assumption — expectation is always linear. Variance is different.

Theorem 22.2 (Variance)**.**

$$\mathrm{Var}(Y) = \mathrm{Cov}(Y,Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \, \mathrm{Cov}(X_i, X_j) = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \cdots & \mathrm{Cov}(X_n,X_n) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.$$

If the X_i are independent, all off-diagonal terms vanish and

$$\boxed{\mathrm{Var}\Big(\sum_i a_i X_i\Big) = \sum_{i=1}^{n} a_i^2 \, \sigma_{X_i}^2}$$

Remark 22.3. Note the a_i^2: coefficients enter the mean linearly and the variance *quadratically*. In particular $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ for independent X, Y — subtracting does not subtract variance, it adds it.

Theorem 22.4 (Normal is closed under linear combinations)**.** *If $X_1, \dots, X_n$ are independent and each $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$, then*

$$\sum_{i=1}^{n} a_i X_i \;\sim\; \text{Normal}\Big(\sum_i a_i \mu_i,\; \sum_i a_i^2 \sigma_i^2\Big)$$

exactly — no approximation, no large-n requirement.

This is the fact behind "$\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$ exactly for a normal sample" from last lecture.

22.2 Point estimation

Definition 22.5 (Point estimator and point estimate)**.** Let θ be an unknown parameter of a distribution. A **point estimator** $\hat{\theta}$ of θ is a statistic — a random variable computed from $X_1, \dots, X_n$. The number it produces from actual data $x_1, \dots, x_n$ is the **point estimate**.

Remark 22.6. Same distinction as always: estimat*or* = random variable (capital letters in); estimat*e* = number (data in). An estimator has a distribution; an estimate does not.

Example 22.7 (The mean)**.** $X_1, \dots, X_n$ i.i.d. with mean μ and variance σ^2. Then $\hat{\mu} = \bar{X}$ is a point estimator of μ, and $\mathbb{E}[\bar{X}] = \mu$.

Example 22.8 (A proportion)**.** Estimate the pass rate p of STAT400. Sample n students and let

$$X_i = \begin{cases} 1, & \text{student } i \text{ passes,} \\ 0, & \text{fails.} \end{cases}$$

Then $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$. Since $\mathbb{E}[X_i] = p$, we get $\mathbb{E}[\hat{p}] = p$.

Example 22.9 (A difference of means)**.** iPhone battery life has mean μ_X; Samsung's has mean μ_Y. To compare them, estimate $\mu_X - \mu_Y$ by

$$\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

By the linear-combination rules, $\mathbb{E}[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$ and (for independent samples) $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$.

Example 22.10 (Two estimators of the variance)**.** For an i.i.d. sample, which of

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{or} \qquad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

should estimate σ^2? The next definition decides it.

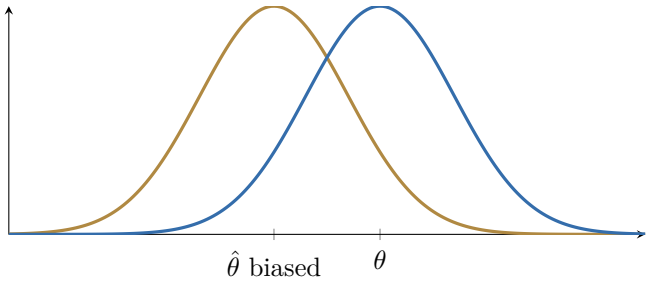
22.3 Bias

Definition 22.11 (Unbiased estimator)**.** $\hat{\theta}$ is an **unbiased** estimator of θ if

$$\mathbb{E}[\hat{\theta}] = \theta.$$

In general $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

Remark 22.12. Unbiased means: *on average over repeated samples*, the estimator hits the target. It says nothing about any single sample — an unbiased estimator can be wildly wrong every time, as long as the errors cancel. That is why the next lecture adds variance to the picture.



Theorem 22.13. *For an i.i.d. sample with variance σ^2,*

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sigma^2,$$

so S^2 is unbiased, while $\frac{1}{n}\sum_i (X_i - \bar{X})^2$ has expectation $\frac{n-1}{n}\sigma^2$ and is biased downward.

Proof sketch (this is Homework). Use the identity $\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2$, take expectations, and use $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ and $\text{Var}(\bar{X}) = \sigma^2/n$:

$$\mathbb{E}\Big[\sum_i (X_i - \bar{X})^2\Big] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2. \qquad \square$$

22.4 A worked example: estimating the endpoint of a uniform

Example 22.14 (Apple Watch water resistance)**.** The maximum time a watch survives under water is an unknown constant θ. You have n identical watches and test each one, obtaining $X_1, \ldots, X_n$ i.i.d. Uniform$[0, \theta]$. Estimate θ.

The obvious estimator is

$$\hat{\theta} = \max(X_1, \ldots, X_n),$$

and note immediately that $0 \le \hat{\theta} \le \theta$ always — it can never overshoot, which already smells biased. Let us compute.

For $0 \le t \le \theta$, using independence,

$$F_{\hat{\theta}}(t) = \mathbb{P}\{\max(X_1, \ldots, X_n) \le t\} = \mathbb{P}\{X_1 \le t\} \cdots \mathbb{P}\{X_n \le t\} = \left(\frac{t}{\theta}\right)^n,$$

so

$$f_{\hat{\theta}}(t) = \frac{n\, t^{n-1}}{\theta^n}, \qquad 0 \le t \le \theta.$$

Therefore

$$\mathbb{E}[\hat{\theta}] = \int_0^\theta t \cdot \frac{n t^{n-1}}{\theta^n}\, \mathrm{d}t = \frac{n}{\theta^n} \int_0^\theta t^n\, \mathrm{d}t = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\,\theta \;\neq\; \theta.$$

So max is **biased**, with bias $-\theta/(n+1)$ — it systematically underestimates, exactly as the picture suggested. But the bias vanishes as $n \to \infty$, and the fix is immediate: rescale.

Theorem 22.15.

$$\boxed{\hat{\theta}_{unbiased} = \frac{n+1}{n} \max(X_1, \ldots, X_n)}$$

is an unbiased estimator of θ.

Remark 22.16. This is the classic "German tank problem": estimating the size of a population of serial numbers from the largest one observed. Notice that the naive estimator was not wrong so much as *systematically* wrong, and a systematic error is exactly the kind you can correct once you have computed it.

Chapter 23

Minimum Variance Estimators; Method of Moments

Lecture 23 | Devore 9e: 6.1–6.2

23.1 The sample variance is unbiased

We owe a proof from last lecture. First an algebraic identity.

Lemma 23.1.

$$\sum_{i=1}^{n}(X_i-\bar X)^2=\sum_{i=1}^{n}X_i^2-n\bar X^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\Big(\sum_{i=1}^{n}X_i\Big)^2.$$

Proof. Expand the square and use $\sum_i X_i = n\bar X$:

$$\sum_i (X_i-\bar X)^2=\sum_i X_i^2-2\bar X\sum_i X_i+n\bar X^2=\sum_i X_i^2-2n\bar X^2+n\bar X^2=\sum_i X_i^2-n\bar X^2. \qquad \square$$

Theorem 23.2. *If $X_1,\dots,X_n$ are i.i.d. with mean μ and variance σ^2, then $\mathbb{E}[S^2]=\sigma^2$.*

Proof. The two ingredients are $\mathbb{E}[Z^2]=\operatorname{Var}(Z)+\mathbb{E}[Z]^2$ applied twice:

$$\mathbb{E}[X_i^2]=\sigma^2+\mu^2, \qquad \mathbb{E}\Big[\Big(\sum_i X_i\Big)^2\Big]=\operatorname{Var}\Big(\sum_i X_i\Big)+\Big(\mathbb{E}\Big[\sum_i X_i\Big]\Big)^2=n\sigma^2+(n\mu)^2,$$

the middle step using independence. Then by the lemma,

$$\begin{aligned}\mathbb{E}[S^2]&=\frac{1}{n-1}\,\mathbb{E}\Big[\sum_i X_i^2-\tfrac{1}{n}\big(\textstyle\sum_i X_i\big)^2\Big]\\&=\frac{1}{n-1}\Big\{n(\sigma^2+\mu^2)-\tfrac{1}{n}\big(n\sigma^2+n^2\mu^2\big)\Big\}\\&=\frac{1}{n-1}\Big\{n\sigma^2+n\mu^2-\sigma^2-n\mu^2\Big\}=\frac{(n-1)\sigma^2}{n-1}=\sigma^2. \qquad \square\end{aligned}$$

23.2 Which unbiased estimator is best?

Unbiasedness is a weak requirement: many estimators have it, and they can be very different. Recall the Uniform$[0, \theta]$ problem.

Example 23.3 (Two unbiased estimators of θ)**.** $X_1, \dots, X_n$ i.i.d. Uniform$[0, \theta]$.

- From last lecture, $\hat{\theta}_1 = \dfrac{n+1}{n} \max(X_1, \dots, X_n)$ is unbiased.
- Also $\hat{\theta}_2 = 2\bar{X}$ is unbiased, since $\mathbb{E}[X_i] = \theta/2$ gives $\mathbb{E}[\hat{\theta}_2] = 2 \cdot \frac{\theta}{2} = \theta$.

Both hit the target on average. Now compare their variances. Write $M = \max(X_1, \dots, X_n)$, whose density $f_M(t) = nt^{n-1}/\theta^n$ we found last lecture. Then

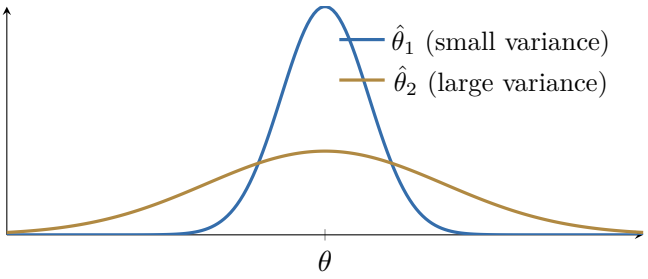
$$\mathbb{E}[M^2] = \int_0^\theta t^2 \cdot \frac{nt^{n-1}}{\theta^n} \, \mathrm{d}t = \frac{n\theta^2}{n+2}, \qquad \mathrm{Var}(M) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2},$$

the last step because $(n+1)^2 - n(n+2) = 1$. Since $\hat{\theta}_1 = \frac{n+1}{n} M$,

$$\mathrm{Var}(\hat{\theta}_1) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}(M) = \frac{\theta^2}{n(n+2)}, \qquad \mathrm{Var}(\hat{\theta}_2) = \frac{4}{n} \mathrm{Var}(X_i) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Since $n(n+2) > 3n$ for $n > 1$,

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2) \qquad \Longrightarrow \qquad \hat{\theta}_1 \text{ is the better estimator.}$$



Remark 23.4. The difference is dramatic: $\mathrm{Var}(\hat{\theta}_1)$ shrinks like $1/n^2$ while $\mathrm{Var}(\hat{\theta}_2)$ shrinks only like $1/n$. Using the maximum extracts far more information about θ than the average does — which makes sense, since only the largest observation carries information about where the interval ends.

Definition 23.5 (MVUE)**.** Among all unbiased estimators of θ, the one with the smallest variance is the **minimum variance unbiased estimator** (MVUE).

Remark 23.6 (Is a biased estimator always worse?)**.** No. An estimator with a small bias and a much smaller variance can beat an unbiased one in every practical sense. The standard way to trade the two off is the **mean squared error**

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \big(\mathrm{bias}(\hat{\theta})\big)^2,$$

which penalizes both. Unbiasedness is a convenient constraint, not a sacred one.

23.3 Constructing estimators

So far we have *checked* estimators that were pulled out of thin air. Two systematic recipes produce them:

- the **method of moments** (today);
- the **method of maximum likelihood** (Lecture 24).

Definition 23.7 (Population and sample moments)**.** The **k-th (population) moment** of a distribution is $\mathbb{E}[X^k]$; the first moment is the mean. The **k-th sample moment** of $X_1, \dots, X_n$ is

$$\overline{X^k} := \frac{X_1^k + X_2^k + \cdots + X_n^k}{n}, \qquad \text{so in particular} \qquad \overline{X^1} = \bar{X}.$$

Definition 23.8 (Method of moments)**.** Let $X_1, \dots, X_n$ be i.i.d. from a distribution with m unknown parameters $\theta_1, \dots, \theta_m$. Set the first m population moments equal to the corresponding sample moments,

$$\mathbb{E}[X^k] = \overline{X^k}, \qquad k = 1, 2, \dots, m,$$

and solve the resulting m equations for $\theta_1, \dots, \theta_m$.

The idea in one line:

$$\underbrace{\mathbb{E}[X^k]}_{\text{an expression in } \theta_1,\dots,\theta_m} = \underbrace{\overline{X^k}}_{\text{a number computed from the data}}.$$

The left side is a formula in the unknowns; the right side is something you can evaluate. Write down as many such equations as you have unknowns, then solve. The justification is the law of large numbers: $\overline{X^k} \to \mathbb{E}[X^k]$ as n grows, so matching them is matching quantities that genuinely become equal.

23.3.1 Gamma

Example 23.9. $X_1, \dots, X_n$ i.i.d. Gamma(α, β), so $m = 2$ unknowns. We know

$$\mathbb{E}[X] = \alpha\beta, \qquad \mathrm{Var}(X) = \alpha\beta^2,$$

hence $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2 = \alpha\beta^2 + \alpha^2\beta^2$. The two equations are

$$\begin{cases} \bar{X} = \alpha\beta, \\ \overline{X^2} = \alpha\beta^2 + \alpha^2\beta^2. \end{cases}$$

Subtract the square of the first equation from the second. Since $\bar{X}^2 = (\alpha\beta)^2 = \alpha^2\beta^2$, the $\alpha^2\beta^2$ terms cancel:

$$\overline{X^2} - \bar{X}^2 = \alpha\beta^2.$$

Now divide this by the first equation to eliminate α:

$$\frac{\overline{X^2} - \bar{X}^2}{\bar{X}} = \frac{\alpha\beta^2}{\alpha\beta} = \beta,$$

and substitute back into $\bar{X} = \alpha\beta$ to recover α:

$$\boxed{\hat{\beta} = \frac{\overline{X^2} - \bar{X}^2}{\bar{X}}, \qquad \hat{\alpha} = \frac{\bar{X}}{\hat{\beta}} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2}}$$

Remark 23.10. Sanity check on the structure: $\overline{X^2} - \bar{X}^2$ is the (biased) sample variance, so the estimators read $\hat{\beta} = \widehat{\mathrm{Var}}/\widehat{\mathbb{E}}$ and $\hat{\alpha} = \widehat{\mathbb{E}}^2/\widehat{\mathrm{Var}}$ — exactly the inversion of $\mathbb{E} = \alpha\beta$, $\mathrm{Var} = \alpha\beta^2$.

With the ten observations $152, 115, 109, 94, 88, 125, 40, 128, 123, 136$ you get $\bar{x} = 1110/10 = 111$ and $\overline{x^2} = 132024/10 = 13202.4$, so $\overline{x^2} - \bar{x}^2 = 13202.4 - 12321 = 881.4$ and

$$\hat{\beta} = \frac{881.4}{111} \approx 7.94, \qquad \hat{\alpha} = \frac{12321}{881.4} \approx 13.98.$$

23.3.2 Negative binomial

Example 23.11. $X_1, \ldots, X_n$ i.i.d. NegBinomial(r, p), where *both* r and p are unknown. Here

$$\mathbb{E}[X] = \frac{r(1-p)}{p}, \qquad \mathbb{E}[X^2] = \frac{r(1-p)(r - rp + 1)}{p^2}.$$

Set these equal to $\bar{X}$ and $\overline{X^2}$. As with the gamma, the system is easier after subtracting the square of the first equation from the second, which replaces the second moment by the variance:

$$\bar{X} = \frac{r(1-p)}{p}, \qquad \overline{X^2} - \bar{X}^2 = \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

Dividing the first by the second cancels r and $1 - p$ outright:

$$\frac{\bar{X}}{\overline{X^2} - \bar{X}^2} = \frac{r(1-p)/p}{r(1-p)/p^2} = p \qquad \Longrightarrow \qquad \hat{p} = \frac{\bar{X}}{\overline{X^2} - \bar{X}^2}.$$

Solving the first equation for r gives $r = \bar{X}\, p/(1-p)$; substituting $1 - \hat{p} = \dfrac{\overline{X^2} - \bar{X}^2 - \bar{X}}{\overline{X^2} - \bar{X}^2}$ yields

$$\hat{r} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2 - \bar{X}}.$$

Remark 23.12. Note that $\hat{p}$ and $\hat{r}$ are random variables (functions of $X_1, \ldots, X_n$); plugging in the observed data turns them into numbers.

23.4 Where the method of moments runs out

1. **Computing $\mathbb{E}[X^k]$ can be hard.** The method needs closed-form population moments. For the negative binomial above, the second moment was already unpleasant; for many distributions there is no elementary formula at all.
2. **Solving the system can be hard.** With $m \geq 2$ parameters the moment equations are non-linear and need not have a closed-form solution — or any solution.
3. **The answer can fall outside the parameter space.** In the negative binomial, if the sample happens to have variance smaller than its mean, then $\overline{X^2} - \bar{X}^2 - \bar{X} < 0$ and $\hat{r}$ comes out *negative* — impossible, since r counts successes. Nothing in the method prevents this.

Remark 23.13. Method-of-moments estimators are cheap to compute and are a good way to get a starting value, but they carry no guarantee: not of landing in the parameter space, and not of being efficient. Maximum likelihood, next lecture, is built to fix both.

Chapter 24

Maximum Likelihood Estimation

Lecture 24 | Devore 9e: 6.2

The method of moments matched summary statistics. Maximum likelihood asks a different and more direct question:

Which parameter value makes the data I actually saw as likely as possible?

24.1 A discrete warm-up

Example 24.1 (Candy jar)**.** A jar holds 10 candies, k of them red, with k unknown. You draw 4 candies without replacement and observe 3 red. Estimate k.

A first instinct is to match proportions: $\frac{k}{10} = \frac{3}{4}$, so $k = 7.5$ — which is not an integer, and not obviously right. Instead, write down the probability of what happened, as a function of k:

$$\mathbb{P}(3 \text{ red out of } 4) = \frac{\binom{k}{3}\binom{10-k}{1}}{\binom{10}{4}}, \qquad 3 \le k \le 9.$$

(The range is forced: we saw 3 reds so $k \ge 3$, and we saw a non-red so $k \le 9$.) The denominator does not involve k, and $\binom{k}{3}\binom{10-k}{1} = \frac{k(k-1)(k-2)}{6}(10-k)$, so we maximize $k(k-1)(k-2)(10-k)$ over the integers:

k	3	4	5	6	7	$\boxed{8}$	9
$k(k-1)(k-2)(10-k)$	42	144	300	480	630	$\boxed{672}$	504

So $\hat{k}_{\text{MLE}} = 8$.

Remark 24.2. The estimate is the value of the parameter under which the observed data is least surprising. Note that this required no derivative — for a discrete parameter you simply tabulate.

24.2 The continuous case, and the likelihood function

Example 24.3 (Light bulbs)**.** Lifetimes follow Exponential(λ). You observe 5 bulbs:

$$x_1 = 2, \quad x_2 = 1.5, \quad x_3 = 3, \quad x_4 = 0.5, \quad x_5 = 4.$$

Estimate λ.

We cannot copy the discrete recipe, because $\mathbb{P}\{X_1 = 2,\ X_2 = 1.5,\ \dots\} = 0$ for a continuous variable. Use the joint *density* instead: it plays exactly the role that probability played above.

Definition 24.4 (Likelihood function)**.** Let $X_1, \dots, X_n$ be a sample from a distribution with parameters $\theta_1, \dots, \theta_m$, and let $x_1, \dots, x_n$ be the observed data. The **likelihood function** is the joint pdf (or pmf) viewed as a function of the parameters:

$$L(\theta_1, \dots, \theta_m) := f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m).$$

The **maximum likelihood estimates** $\hat{\theta}_1, \dots, \hat{\theta}_m$ are the values maximizing L.

Remark 24.5. Read the roles carefully. The joint pdf is normally a function of the data with the parameters fixed. The likelihood is the *same formula* regarded as a function of the parameters with the data fixed. Same expression, opposite point of view.

24.2.1 The log trick

For an i.i.d. sample the joint density is a product, $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$, and products are unpleasant to differentiate. Since log is strictly increasing, maximizing L is the same as maximizing

$$\ell(\theta) := \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta),$$

which turns the product into a sum. Always take logs first.

24.2.2 Exponential, worked

$$L(\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i}, \qquad \log L(\lambda) = 5 \log \lambda - \lambda \sum_{i=1}^{5} x_i.$$

Differentiate and set to zero:

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{5}{\lambda} - \sum_{i=1}^{5} x_i \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\lambda}_{\text{MLE}} = \frac{5}{\sum_i x_i}.$$

With $\sum x_i = 11$, $\hat{\lambda} = \frac{5}{11} = \frac{1}{2.2}$. In general

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}}$$

which is reassuring, since $\mathbb{E}[X] = 1/\lambda$ for the exponential.

24.3 Normal

Example 24.6 (Unknown mean)**.** X_i i.i.d. Normal$(\mu, 1)$ with data $8, 9, 10, 11, 12$.

$$L(\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu)^2/2}, \qquad \log L(\mu) = -\frac{5}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{5} (x_i - \mu)^2.$$

Then

$$\frac{d}{d\mu}\log L(\mu) = \sum_{i=1}^{5}(x_i - \mu) \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\mu}_{\text{MLE}} = \frac{\sum_i x_i}{5} = \bar{x} = 10.$$

Remark 24.7. Notice that maximizing the likelihood here is the same as *minimizing* $\sum_i (x_i - \mu)^2$. Least squares is maximum likelihood under normal errors — that identity is the foundation of regression.

Example 24.8 (Unknown variance)**.** X_i i.i.d. Normal$(10, \sigma^2)$ with the same data. Writing $v = \sigma^2$,

$$\log L(v) = -\frac{5}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{5}(x_i - 10)^2,$$

$$\frac{d}{dv}\log L(v) = -\frac{5}{2}\cdot\frac{1}{v} + \frac{\sum_i (x_i - 10)^2}{2v^2} \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{5}(x_i - 10)^2}{5} = \frac{10}{5} = 2.$$

Remark 24.9 (MLE can be biased)**.** In general, when μ is also estimated from the data, the MLE of the variance is

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2,$$

the *biased* sample variance from Lecture 22. So maximum likelihood does not automatically produce unbiased estimators. It has other virtues — consistency, asymptotic efficiency, and invariance — which is why it is the default method anyway.

24.4 When calculus does not apply

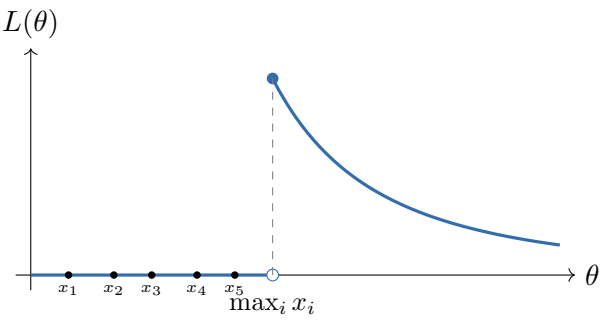
Example 24.10 (Uniform $[0, \theta]$, one more time)**.** $X_1, \ldots, X_n$ i.i.d. Uniform$[0, \theta]$. The likelihood is

$$L(\theta) = f(x_1)\cdots f(x_n) = \begin{cases} \dfrac{1}{\theta^n}, & 0 \le x_i \le \theta \text{ for all } i, \\ 0, & \text{otherwise.} \end{cases}$$

The condition "$x_i \le \theta$ for all i" is just $\theta \ge \max_i x_i$. So

$$L(\theta) = \begin{cases} 0, & \theta < \max_i x_i, \\ \theta^{-n}, & \theta \ge \max_i x_i, \end{cases}$$

which is 0, jumps up at $\theta = \max_i x_i$, and then decreases.



There is no interior critical point — the derivative is never zero. The maximum sits at the left edge of the allowed region:

$$\boxed{\hat{\theta}_{\text{MLE}} = \max(X_1, \ldots, X_n)}$$

Remark 24.11. So maximum likelihood recovers the biased estimator from Lecture 22, not the bias-corrected $\frac{n+1}{n}$ max. Two lessons:

1. Always check whether the support of the distribution depends on the parameter. If it does, do not reach for the derivative — sketch L instead.
2. MLE and unbiasedness are different criteria and can disagree.

24.5 Recipe

1. Write the likelihood $L(\theta) = \prod_i f(x_i; \theta)$, being careful about the range of validity.
2. Take logs: $\ell(\theta) = \sum_i \log f(x_i; \theta)$.
3. Differentiate, set to zero, solve.
4. Check it is a maximum (second derivative, or the shape of ℓ).
5. If the support depends on θ, skip steps 2–4 and argue directly.

Chapter 25

Properties of MLEs; Confidence Intervals

Lecture 25 | Devore 9e: 6.2, 7.1

25.1 More maximum likelihood

25.1.1 Several parameters at once

When there are m unknown parameters, set all m partial derivatives of the log-likelihood to zero and solve the system.

Example 25.1 (Normal, both parameters unknown)**.** $X_1, \dots, X_n$ i.i.d. Normal(μ, σ^2). With $v = \sigma^2$,

$$\ell(\mu, v) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{v}\sum_i (x_i - \mu) = 0 \implies \hat{\mu} = \bar{x},$$

$$\frac{\partial \ell}{\partial v} = -\frac{n}{2v} + \frac{\sum_i (x_i - \mu)^2}{2v^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2.$$

So the MLE of σ^2 is the *biased* sample variance, $\frac{n-1}{n}S^2$.

25.1.2 Three properties worth knowing

Theorem 25.2 (Invariance)**.** *If $\hat{\theta}$ is the MLE of θ and g is any function, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

This is a genuine convenience: having found $\hat{\sigma}^2$, the MLE of σ is $\sqrt{\hat{\sigma}^2}$ — no new optimization. (Note that unbiasedness has no such property: $\mathbb{E}[\hat{\theta}] = \theta$ does *not* give $\mathbb{E}[g(\hat{\theta})] = g(\theta)$.)

Theorem 25.3 (Large-sample behavior)**.** *Under mild regularity conditions, as $n \to \infty$ the MLE is*

- ***consistent**: $\hat{\theta} \to \theta$;*
- ***asymptotically unbiased**: $\mathbb{E}[\hat{\theta}] \to \theta$;*
- ***asymptotically normal and efficient**: no other estimator has smaller variance in the limit.*

Remark 25.4 (MLE versus method of moments)**.**

	Method of moments	**Maximum likelihood**
effort	solve algebraic equations	maximize a function
uses	the first m moments only	the whole density
guarantees	few	consistent, asymptotically efficient
risk	can leave the parameter space	usually well behaved

Method of moments is a good way to get a starting value; maximum likelihood is what you report.

25.2 From point estimates to intervals

A point estimate is a single number and is therefore almost surely wrong. Reporting $\hat{\mu} = 8.3$ says nothing about whether the truth is plausibly 8.2 or plausibly 40. What we want is an interval, together with a statement of how often that interval works.

Definition 25.5 (Confidence interval)**.** A $100(1-\alpha)\%$ **confidence interval** for θ is a pair of statistics $L = L(X_1, \dots, X_n)$ and $U = U(X_1, \dots, X_n)$ with

$$\mathbb{P}\{L \le \theta \le U\} = 1 - \alpha.$$

25.2.1 The normal mean, σ known

Let $X_1, \dots, X_n$ be i.i.d. Normal(μ, σ^2) with σ known. From Lecture 21, $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$ exactly, so

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

Let $z_{\alpha/2}$ be the value with $\mathbb{P}\{Z > z_{\alpha/2}\} = \alpha/2$. By symmetry,

$$\mathbb{P}\left\{-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1 - \alpha.$$

Now solve the inequality for μ — this is the only real step:

$$\boxed{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;\le\; \mu \;\le\; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

confidence	α	$z_{\alpha/2}$	interval
90%	0.10	1.645	$\bar{x} \pm 1.645\,\sigma/\sqrt{n}$
95%	0.05	1.96	$\bar{x} \pm 1.96\,\sigma/\sqrt{n}$
99%	0.01	2.576	$\bar{x} \pm 2.576\,\sigma/\sqrt{n}$

Example 25.6. A machine fills bottles with $\sigma = 0.5$ ml. A sample of $n = 25$ bottles has $\bar{x} = 500.2$ ml. A 95% confidence interval for μ is

$$500.2 \pm 1.96 \cdot \frac{0.5}{\sqrt{25}} = 500.2 \pm 0.196 = [500.00,\ 500.40].$$

Remark 25.7 (What "95% confident" means — and does not)**.** μ is a fixed unknown number; it is not random. The *interval* is random, because $\bar{X}$ is. The correct reading is:

if we repeated the whole experiment many times, about 95% *of the intervals so constructed would contain* μ*.*

It is **not** correct to say "μ has a 95% chance of lying in $[500.00, 500.40]$" — once the data is in, either it does or it does not.

Remark 25.8 (Width)**.** The width is $2z_{\alpha/2}\sigma/\sqrt{n}$. So:

- more confidence $\Rightarrow$ wider interval (you cannot get both);
- to halve the width, take 4 times as many observations — the $\sqrt{n}$ law again.

25.2.2 Large samples, σ unknown

In practice σ is unknown too. For large n, replace it by S and appeal to the CLT: for *any* underlying distribution with finite variance,

$$\mu \ \approx\ \bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}} \qquad (n \text{ large}).$$

For small n from a normal population, $z_{\alpha/2}$ must be replaced by a t-distribution quantile — which is the starting point of STAT401.

Remark 25.9. Everything in this course has been building to this sentence: *data* $\to$ *estimator* $\to$ *sampling distribution* $\to$ *interval with a guarantee*. The probability half of the course exists precisely to make the third arrow possible.

Chapter 26

Review for the Final Exam

Lecture 26 | Devore 9e: Ch. 2–6

The final is cumulative, with the weight on Lectures 18–25 (joint distributions, CLT, estimation). This chapter is the review class: a checklist, a formula sheet, and the worked problems.

26.1 Checklist

Topic	**You should be able to**
Bayes	invert a conditional; recognize the base-rate trap
Transformations	get f_Y from f_X via the cdf
Named distributions	recognize the scenario and quote pdf, mean, variance
Joint distributions	normalize, marginalize, condition; *draw the region*
Covariance/correlation	compute from a joint pmf/pdf; interpret the sign
CLT	approximate $\bar{X}$; know when it applies
Estimation	check unbiasedness; method of moments; MLE
Confidence intervals	build $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$; say what 95% means

26.2 Master formula sheet

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad \mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B \mid A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B \mid A_j)}$$

$$\mathbb{E}[h(X)] = \int h(x) f(x)\,\mathrm{d}x, \qquad \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$f_Y(y) = f_X(g^{-1}(y))\left|\tfrac{d}{dy} g^{-1}(y)\right|, \qquad f_X(x) = \int f(x,y)\,\mathrm{d}y$$

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}, \qquad \mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X \mu_Y$$

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \qquad \mathrm{Var}\Big(\sum a_i X_i\Big) = \sum a_i^2 \sigma_i^2 \text{ (independent)}$$

$$\mathbb{E}[\bar{X}] = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \bar{X} \overset{\text{approx}}{\sim} \mathrm{Normal}\Big(\mu, \frac{\sigma^2}{n}\Big)$$

$$\mathbb{E}[S^2] = \sigma^2, \qquad \ell(\theta) = \sum_i \log f(x_i;\theta)$$

$$\mathbb{E}[X^k] = \overline{X^k} \text{ (method of moments)}, \qquad \mu \in \bar{x} \pm z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} \ \ (\sigma \text{ known})$$

26.3 Worked problems

Problem 1 (Bayes)

A disease affects 3% of the population. A test satisfies $\mathbb{P}(+ \mid D) = 0.92$ and $\mathbb{P}(- \mid D^c) = 0.88$. *(a)* Find $\mathbb{P}(+)$. *(b)* Find $\mathbb{P}(D \mid +)$.

Solution. $\mathbb{P}(D) = 0.03$, $\mathbb{P}(D^c) = 0.97$, and $\mathbb{P}(+ \mid D^c) = 1 - 0.88 = 0.12$.

$$\mathbb{P}(+) = \mathbb{P}(+ \mid D)\mathbb{P}(D) + \mathbb{P}(+ \mid D^c)\mathbb{P}(D^c) = (0.92)(0.03) + (0.12)(0.97) = 0.0276 + 0.1164 = 0.144.$$

$$\mathbb{P}(D \mid +) = \frac{\mathbb{P}(+ \mid D)\mathbb{P}(D)}{\mathbb{P}(+)} = \frac{0.0276}{0.144} = 0.1917.$$

Again: a positive result raises the probability from 3% only to 19%. □

Problem 2 (transformation)

$X \sim \mathrm{Normal}(0,1)$ and $Y = e^X$. Find $\mathbb{P}\{Y > e^{1.5}\}$.

Solution. exp is increasing, so the event is unchanged by taking logs:

$$\mathbb{P}\{Y > e^{1.5}\} = \mathbb{P}\{e^X > e^{1.5}\} = \mathbb{P}\{X > 1.5\} = 1 - \Phi(1.5) = 0.0668.$$

No density calculation is needed — that is the point of working with the cdf. □

Problem 3 (Weibull)

$X \sim \text{Weibull}(\alpha = 3, \beta = 10)$, so $F(x) = 1 - e^{-(x/10)^3}$ for $x > 0$. *(a)* Find the median $t_{0.5}$. *(b)* Find $\mathbb{P}\{5 \le X \le 15\}$.

Solution. (a) Solve $F(t_{0.5}) = 0.5$:

$$1 - e^{-(t_{0.5}/10)^3} = 0.5 \;\Rightarrow\; e^{-(t_{0.5}/10)^3} = \tfrac{1}{2} \;\Rightarrow\; \left(\tfrac{t_{0.5}}{10}\right)^3 = \ln 2 \;\Rightarrow\; t_{0.5} = 10(\ln 2)^{1/3} \approx 8.85.$$

(b) $\mathbb{P}\{5 \le X \le 15\} = F(15) - F(5) = e^{-(5/10)^3} - e^{-(15/10)^3} = e^{-1/8} - e^{-27/8} \approx 0.8825 - 0.0342 = 0.8483$. □

Problem 4 (beta)

$X \sim \text{Beta}(3, 3)$, so $f(x) = 30x^2(1-x)^2$ on $(0, 1)$. *(a)* Find $\mathbb{E}[X]$ and $\text{Var}(X)$. *(b)* Find $\mathbb{P}\{0.4 \le X \le 0.6\}$.

Solution. (a) By the formulas for $\text{Beta}(\alpha, \beta)$,

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} = \frac{3}{6} = \frac{1}{2}, \qquad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{9}{36 \cdot 7} = \frac{1}{28}.$$

(b) $\displaystyle\int_{0.4}^{0.6} 30x^2(1-x)^2\,\mathrm{d}x = 30\Big[\tfrac{x^3}{3} - \tfrac{x^4}{2} + \tfrac{x^5}{5}\Big]_{0.4}^{0.6} = 0.6826 - 0.3174 = 0.3651.$ □

Problem 5 (joint pdf)

$f_{X,Y}(x, y) = c(x + y)$ for $0 < x < 1$, $0 < y < 1$. *(a)* Find c. *(b)* Find f_X.

Solution. (a)

$$1 = \int_0^1\int_0^1 c(x+y)\,\mathrm{d}x\,\mathrm{d}y = c\int_0^1 \Big(\tfrac{1}{2} + y\Big)\,\mathrm{d}y = c\Big(\tfrac{1}{2} + \tfrac{1}{2}\Big) = c \qquad \Longrightarrow \qquad c = 1.$$

(b) $f_X(x) = \displaystyle\int_0^1 (x+y)\,\mathrm{d}y = x + \tfrac{1}{2}$ for $0 < x < 1$.

The density is symmetric in x and y, so the same computation gives $f_Y(y) = y + \frac{1}{2}$ on $(0, 1)$. Note $f_X(x)f_Y(y) = (x + \frac{1}{2})(y + \frac{1}{2}) \neq x + y$, so X and Y are dependent. □

Problem 6 (covariance and correlation from a table)

(X, Y) is uniform on the six pairs $(0,1), (1,1), (0,2), (2,2), (1,3), (2,3)$, each with probability $\frac{1}{6}$. Find $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\text{Cov}(X, Y)$, $\rho_{X,Y}$.

Solution. Marginals: each of $X = 0, 1, 2$ has probability $\frac{2}{6} = \frac{1}{3}$, and likewise each of $Y = 1, 2, 3$. So

$$\mathbb{E}[X] = \tfrac{1}{3}(0+1+2) = 1, \qquad \mathbb{E}[Y] = \tfrac{1}{3}(1+2+3) = 2.$$

$$\mathbb{E}[XY] = \tfrac{1}{6}\big(0\cdot 1 + 1\cdot 1 + 0\cdot 2 + 2\cdot 2 + 1\cdot 3 + 2\cdot 3\big) = \tfrac{14}{6} = \tfrac{7}{3},$$

$$\text{Cov}(X, Y) = \tfrac{7}{3} - (1)(2) = \tfrac{1}{3}.$$

$\mathbb{E}[X^2] = \frac{1}{3}(0+1+4) = \frac{5}{3} \Rightarrow \text{Var}(X) = \frac{5}{3} - 1 = \frac{2}{3}$, $\quad \mathbb{E}[Y^2] = \frac{1}{3}(1+4+9) = \frac{14}{3} \Rightarrow \text{Var}(Y) = \frac{14}{3} - 4 = \frac{2}{3}$.

$$\rho_{X,Y} = \frac{1/3}{\sqrt{(2/3)(2/3)}} = \frac{1/3}{2/3} = \frac{1}{2}.$$

□

Problem 7 (MLE)

$X_1, \dots, X_n$ i.i.d. Exponential(λ), $f(x;\lambda) = \lambda e^{-\lambda x}$. *(a)* Write the likelihood. *(b)* Find $\hat{\lambda}_{\text{MLE}}$. *(c)* Is it unbiased?

Solution. (a) $L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big)$.

(b) $\ell(\lambda) = n \log \lambda - \lambda \sum_i x_i$, so $\ell'(\lambda) = \frac{n}{\lambda} - \sum_i x_i = 0$ gives

$$\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_i X_i} = \frac{1}{\bar{X}}.$$

(c) No. We know $\mathbb{E}[\bar{X}] = 1/\lambda$, but $\mathbb{E}\big[1/\bar{X}\big] \neq 1/\mathbb{E}[\bar{X}]$ — the reciprocal is a non-linear function. Concretely, $\sum_i X_i \sim \text{Gamma}(n, 1/\lambda)$ by gamma additivity, so for $n \geq 2$

$$\mathbb{E}\Big[\frac{1}{\sum_i X_i}\Big] = \int_0^\infty \frac{1}{s} \cdot \frac{\lambda^n s^{n-1} e^{-\lambda s}}{\Gamma(n)}\, \mathrm{d}s = \frac{\lambda\, \Gamma(n-1)}{\Gamma(n)} = \frac{\lambda}{n-1},$$

and therefore $\mathbb{E}[\hat{\lambda}] = n \cdot \frac{\lambda}{n-1} = \frac{n}{n-1}\lambda > \lambda$: the MLE overshoots on average. It is, however, asymptotically unbiased, and $\frac{n-1}{n}\hat{\lambda}$ is exactly unbiased. □

Remark 26.1 (Final advice)**.** Three habits that recover the most points:

1. **Draw the region** before setting up any double integral.
2. **Go through the cdf** for anything involving a transformation.
3. **Sanity-check the answer**: probabilities lie in $[0, 1]$, variances are non-negative, $|\rho| \leq 1$, and a mean should land inside the support.