Stats

Combined Lecture Notes

Lectures 2 through 30

UVA Statistics • Spring 2026

April 26, 2026

A single bound volume of every lecture in the course, with continuous part, section, page, and definition numbering across the whole text. The volume opens with sample spaces and the axioms of probability, builds up conditional probability, counting, and random variables, catalogues every named discrete and continuous distribution family, develops moment generating functions and variable transformations, covers diagnostic probability plots and the universal inequalities of Markov, Chebychev, and Jensen, and closes with the linear-algebra arc that runs from vectors through matrix diagonalization.

Contents

Lecture 2 — Sample Spaces and Events

Lecture 3 — Probability: Definition, Axioms, and Rules

Lecture 4 — More Probability Rules and Applications

Lecture 5 — Conditional Probability, Bayes' Rule, and Independence

Lecture 9 — Continuous Random Variables: pdf, cdf, and Percentiles

Lecture 12 — Moment Generating Functions

Lecture 13 — Variable Transformations

Lecture 18 — The Uniform and Normal Distributions

Lecture 19 — The Exponential, Gamma, and Beta Distributions

Lecture 20 — Probability Plots: Poisson-ness and QQ Plots

Lecture 24 — Vector Spaces and Subspaces

Sample Spaces and Events

Lecture 2 • UVA Stats

January 2026

Probability begins with what we are uncertain about: the sample space S of all possible outcomes, and the events — subsets of S — that probability assigns numbers to. We develop the language of unions, intersections, and complements via Venn diagrams, distinguish mutually exclusive from exhaustive events, and close with practice problems that translate everyday questions into precise event notation.

PART 1

Motivation

1. Random Phenomena

Term — Random Phenomenon

A *phenomenon* is some result we would like to observe. A phenomenon is **random** if individual outcomes are uncertain, but there is still a regular distribution of outcomes in a large number of repetitions.

In other words, chance behavior is unpredictable in the *short run*, but it has a regular (predictable) pattern in the *long run*.

1.1 Examples of random phenomena

- Flipping a coin.
- Tossing a die.
- While statisticians love the cliché examples of a coin flip and a die toss, the possible outcomes of a random phenomenon need not be equally likely.

In statistics, we often think of the outcomes in a sample as random phenomena. For instance, suppose I randomly select a University of Virginia student. The student's year at UVA, the student's major, and the student's age are all examples of random phenomena.

Key Idea 1

The whole subject of probability is the language we use to describe random phenomena precisely. Today we lay down the two most fundamental objects: the *sample space*, which lists every outcome the phenomenon could produce, and an *event*, which is any subcollection of those outcomes we care to ask a question about.

PART 2

Methods and Examples

2. Sample Space

Term — Sample Space

The **sample space** S of a random phenomenon is the set of all possible outcomes, where the outcomes are mutually exclusive (the phenomenon produces exactly one of them on each repetition).

Worked examples

- Flipping a coin: $S = \{H, T\}$.
- Tossing a die: $S = \{1, 2, 3, 4, 5, 6\}$.

2.1 Countable vs. uncountable sample spaces

A sample space can be either *countable* or *uncountable*.

- A sample space is **countable** when the total number of possible outcomes can be enumerated and written out (possibly as an infinite list).
- A sample space is **uncountable** when the total number of possible outcomes is infinite and cannot be written as such a list.

The coin-flip and die-toss sample spaces above are both countable. An example of an uncountable sample space is to choose a real number at random between 0 and 1:

$$S = \{\omega : \omega \in [0, 1]\}.$$

3. Events

Term — Event

An **event** is any collection of outcomes of a random phenomenon — equivalently, any subset of the sample space.

3.1 Simple vs. compound events

An event can be a single outcome (a **simple** event) or several outcomes (a **compound** event).

Example. The result of a die toss.

- Event A: the result is no greater than 3 $\rightarrow$ $A = \{1, 2, 3\}$ (compound)
- Event B: the result is even $\rightarrow$ $B = \{2, 4, 6\}$ (compound)
- Event C: the result is a 4 $\rightarrow$ $C = \{4\}$ (simple)

3.2 Complement of an event

The **complement** of an event A, denoted A^C or A', is the set of all outcomes in the sample space S that are *not* in event A.

Example. The result of a die toss.

- $A = \{1, 2, 3\} \implies A^C = \{4, 5, 6\}$.
- $B = \{2, 4, 6\} \implies B^C = \{1, 3, 5\}$.

3.3 Containment

We say event A is a *subset* of event B if all outcomes in A are also outcomes in B; equivalently, B **contains** A, written $A \subset B$.

Example. For the die toss, $B = \{2, 4, 6\}$ and $C = \{4\}$. Since every outcome of C is in B, we have $C \subset B$, so B contains C.

3.4 Equality of events

Two events A and B are **equal**, denoted $A = B$, if and only if they consist of the exact same outcomes. To prove $A = B$, one shows both $A \subset B$ and $B \subset A$.

4. Union of Events

The **union** of two events A and B, denoted $A \cup B$, is the event consisting of all outcomes that are in event A or in event B (or both).

Example. Die toss with $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$:

$$A \cup B = \{1, 2, 3, 4, 6\}.$$

Two useful properties:

- For any event A, $A \cup A^C = S$. (For the die toss, $A \cup A^C = \{1, 2, 3\} \cup \{4, 5, 6\} = S$.)
- If $C \subset B$, then $B \cup C = B$. (For the die toss, $B = \{2, 4, 6\}$ and $C = \{4\}$ give $B \cup C = \{2, 4, 6\} = B$.)

5. Intersection of Events

The **intersection** of two events A and B, denoted $A \cap B$, is the event consisting of all outcomes that are in *both* event A and event B.

Example. Die toss with $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$:

$$A \cap B = \{2\}.$$

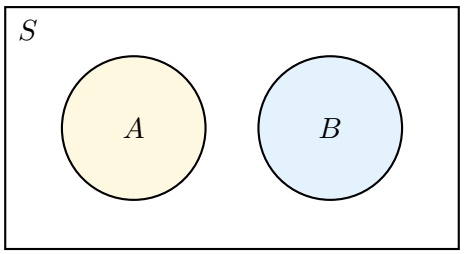
A useful property: if $C \subset B$, then $B \cap C = C$. For the die toss with $B = \{2, 4, 6\}$ and $C = \{4\}$, we have $B \cap C = \{4\} = C$.

6. Disjoint (Mutually Exclusive) Events

If events A and B have no outcomes in common, then their intersection is empty:

$$A \cap B = \varnothing.$$

In this case we say events A and B are **disjoint** or **mutually exclusive**.



Example. Die toss with $A = \{1, 2, 3\}$ and $C = \{4\}$:

$$A \cap C = \varnothing,$$

so A and C are disjoint.

A useful property: for any event A, $A \cap A^C = \varnothing$ (an event and its complement are always disjoint).

7. Partitions

If the sample space is divided into several events that are all pairwise disjoint and whose union is S, the events are said to form a **partition** of the sample space.

General template. If $A \cup B \cup C = S$ and $A \cap B = A \cap C = B \cap C = \varnothing$, then $\{A, B, C\}$ partitions S.

Example. Die toss with $A = \{1, 2, 3\}$, $C = \{4\}$, $D = \{5, 6\}$: since $A \cup C \cup D = S$ and the three events are pairwise disjoint, $\{A, C, D\}$ partitions S.

A useful property: for any event A, $\{A, A^C\}$ always partitions the sample space.

8. Difference of Events

The **difference** of events A and B, denoted $A - B$, is the event consisting of all outcomes that are in A but not in B:

$$A - B = A \cap B^C.$$

Example. Die toss with $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$:

$$A - B = \{1, 3\}, \qquad B - A = \{4, 6\}.$$

Note that in general $A - B \neq B - A$.

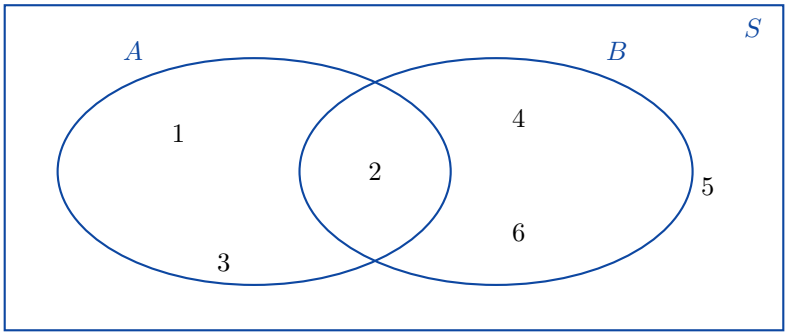
A useful property: if A and C are disjoint, then $A - C = A$ (subtracting an event you do not overlap with removes nothing).

9. Venn Diagrams

It is often helpful to picture the relationship between events visually using **Venn diagrams**: a rectangle represents the sample space, and overlapping circles represent the events.

9.1 Two events on the die-toss sample space

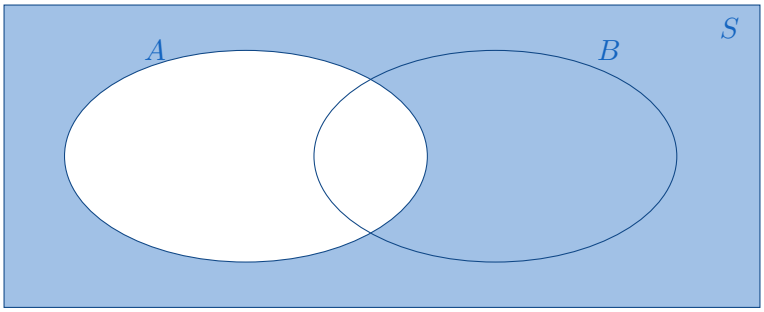
Example. Die toss with $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$:



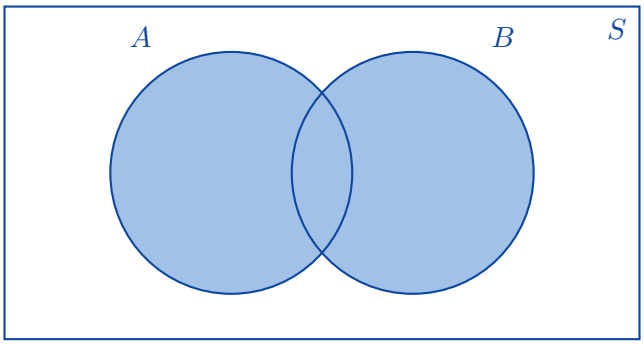
9.2 Visualizing set operations

The set operations introduced above all have a clean Venn-diagram picture (shaded region is the operation's result).

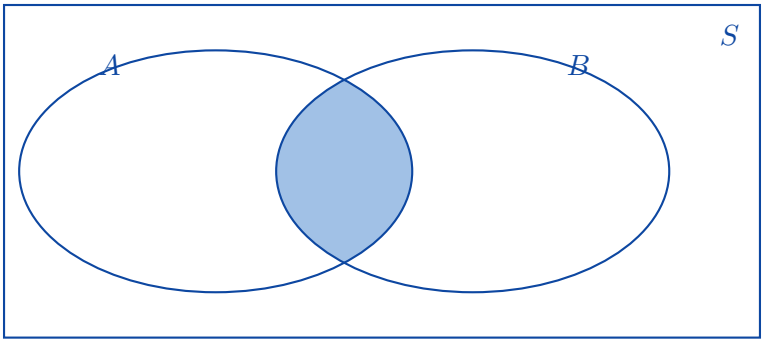
Complement A^C.



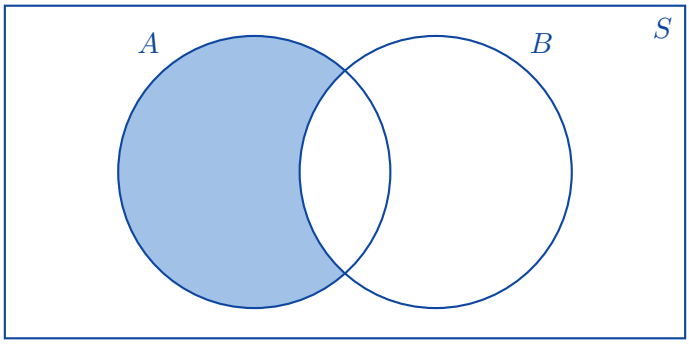
Union $A \cup B$.



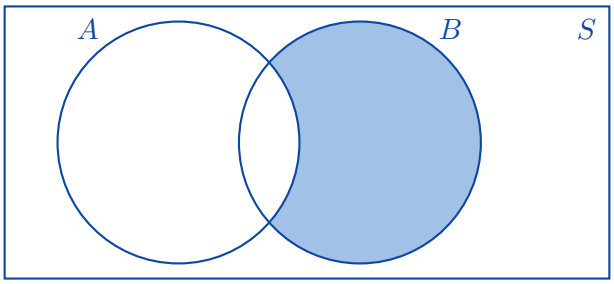
Intersection $A \cap B$.



Difference $A - B$.



Difference $B - A$.



PART 3

Practice Problems

10. Example #1 — Two Four-Sided Dice

Example #1

Suppose that you are playing a board game that involves tossing two four-sided dice consecutively.

(a) What is the sample space for the *sequence* of dice toss results?
(b) What is the sample space for the *sum* of the two dice toss results?
(c) Are these sample spaces countable or uncountable?

For the remaining parts, work with the sum sample space and define the events

$$A = \{\text{ sum is between 4 and 6 }\} = \{4,5,6\}, \qquad B = \{\text{ sum is at least 5 }\} = \{5,6,7,8\}.$$

(d) What is B^C?
(e) What is $A \cup B$?
(f) What is $A^C \cup B^C$?
(g) What is $A \cap B$?
(h) What is $A \cap B^C$?
(i) What is $A - B$?
(j) What is $B - A$?

Solutions

(a) Sample space for the sequence of results. Each die independently shows one of $\{1,2,3,4\}$, so we list the $4 \times 4 = 16$ ordered pairs:

$$\begin{aligned} S = \{&(1,1),(1,2),(1,3),(1,4),\\ &(2,1),(2,2),(2,3),(2,4),\\ &(3,1),(3,2),(3,3),(3,4),\\ &(4,1),(4,2),(4,3),(4,4)\}. \end{aligned}$$

(b) Sample space for the sum of the results. The smallest possible sum is $1+1=2$ and the largest is $4+4=8$, and every integer in between is achievable:

$$S = \{2,3,4,5,6,7,8\}.$$

(c) Countable or uncountable? Both are countable. The first sample space has 16 outcomes and the second has 7 outcomes; both are finite, so in particular countable.

(d) B^C**.** $B = \{5,6,7,8\}$, so within the sum sample space $S = \{2,3,4,5,6,7,8\}$,

$$\boldsymbol{B^C = \{2,3,4\}}.$$

(e) $A \cup B$**.** $A = \{4,5,6\}$ and $B = \{5,6,7,8\}$, so

$$\boldsymbol{A \cup B = \{4,5,6,7,8\}}.$$

(f) $A^C \cup B^C$**.** $A^C = \{2, 3, 7, 8\}$ and $B^C = \{2, 3, 4\}$, so

$$\boldsymbol{A^C \cup B^C = \{2, 3, 4, 7, 8\}}.$$

(g) $A \cap B$**.** The outcomes that lie in $A = \{4, 5, 6\}$ and also in $B = \{5, 6, 7, 8\}$ are

$$\boldsymbol{A \cap B = \{5, 6\}}.$$

(h) $A \cap B^C$**.** $A = \{4, 5, 6\}$ intersected with $B^C = \{2, 3, 4\}$ gives

$$\boldsymbol{A \cap B^C = \{4\}}.$$

(i) $A - B$**.** By definition, $A - B = A \cap B^C$, which we just computed:

$$\boldsymbol{A - B = \{4\}}.$$

(j) $B - A$**.** The outcomes in $B = \{5, 6, 7, 8\}$ but not in $A = \{4, 5, 6\}$ are

$$\boldsymbol{B - A = \{7, 8\}}.$$

Probability: Definition, Axioms, and Rules

Lecture 3 • UVA Stats

January 2026

We turn the intuition of "chance" into a formal mathematical object. Three axioms — non-negativity, normalisation, and countable additivity — generate every rule we will ever use, including the complement rule, the inclusion-exclusion rule for unions, and the rule for differences. Each axiom is proved from the definition, and each rule is illustrated by a worked example.

PART 1

Motivation

1. Where Last Lecture Left Us

The previous lecture established the language we use for any random situation:

- A *random phenomenon* is a process whose individual outcomes are uncertain, but whose outcomes exhibit a regular distribution over many repetitions. Examples include the result of a coin flip, the result of a die toss, the major of a randomly selected UVA student, and the percent change in the value of a stock today.
- The *sample space* S is the set of all possible outcomes of the random phenomenon, where the outcomes are mutually exclusive.
- An *event* is any collection of outcomes of the random phenomenon. We define events for which we wish to calculate probabilities.

We have the language. We do not yet have a way to *number* an event — to say how likely it is. That is the gap this lecture fills.

2. What Should "Probability" Mean?

Suppose the random phenomenon can be repeated identically and independently. Observe it n times and let n_A be the number of times the outcome lands in event A.

Term — Relative Frequency / Empirical Probability

If a random phenomenon is observed n times and event A occurs n_A times, the **relative frequency** (or **empirical probability**) of A is

$$\frac{n_A}{n}.$$

The empirical probability of an event is not the same for every sequence of n observations. For example, if I flip a coin $n = 10$ times and count heads, it is feasible that my first $n = 10$ flips give $n_A = 6$ heads (an empirical probability of 0.6), and my next $n = 10$ flips give $n_A = 3$ heads (an empirical probability of 0.3). The relative frequency wobbles in the short run.

In the long run, however, it settles down. As n grows, the relative frequency n_A/n approaches a limiting value.

PART 2

Methods and Examples

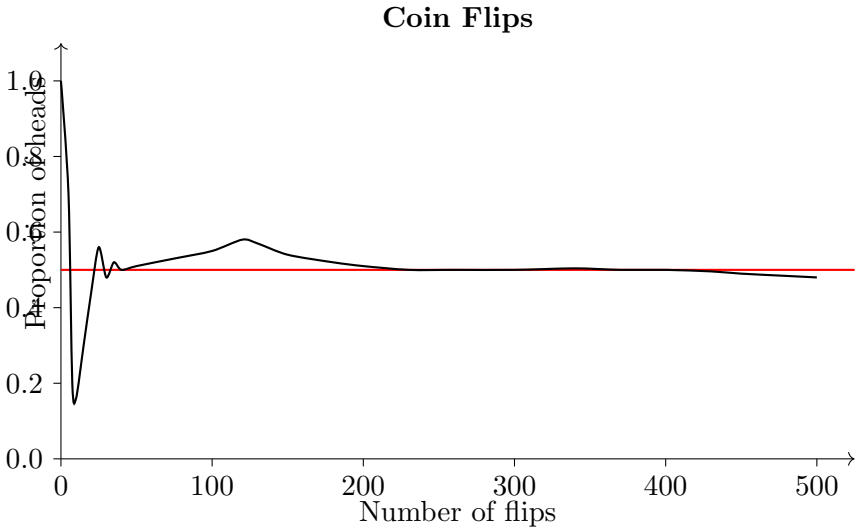
3. Limiting Relative Frequency

Term — Probability of an Event

The **probability** of event A, denoted $P(A)$, is the limit of the relative frequency of A as the number of repetitions grows without bound:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

The picture for a fair coin: as the number of flips grows, the running proportion of heads stops fluctuating wildly and settles toward 0.5.



4. Estimating Probabilities in Practice

In reality, computing this limiting relative frequency exactly is rarely possible. Two practical workarounds:

- **Empirical estimate.** Use a sample's relative frequency as our estimate of the true probability. *Example.* In a random sample of 100 UVA students, 72 are from Virginia, so we estimate the probability that a UVA student is from in-state to be 0.72.

- **Subjective estimate.** Use intuition or expert judgement. *Example.* "There is a 70% chance that UVA wins its basketball game this weekend against UNC."

Either way, the *interpretation* of the resulting number stays the same:

Key Idea 2

When we say "the probability of event A is $P(A)$," we mean: *event A will occur $P(A) \cdot 100\%$ of the time over a large number of repeated trials of the random phenomenon.*

5. Probability as a Function

Mathematically, probability is a function: the input is an event, and the output is the probability of that event. To make this function behave the way limiting relative frequencies behave, we need to constrain it. What rules should it satisfy?

6. The Three Axioms of Probability

The Russian mathematician Andrey Kolmogorov (1933) wrote down the three axioms below. These are the only "rules" of probability we need to *assume*; every other rule follows from them.

Term — Axioms of Probability

For any events defined on a sample space S:

(1) $P(A) \geq 0$ for every event A.

(2) $P(S) = 1$.

(3) If $A_1, A_2, A_3, \ldots$ is a countably infinite collection of pairwise disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i).$$

This third axiom is called **countable additivity**.

Note: You will often see the third axiom written in the simpler two-event form $P(A \cup B) = P(A) + P(B)$ for disjoint A and B. Kolmogorov's axiom is the more general statement: it covers any countably infinite collection at once.

The next five sections derive the everyday "rules" of probability directly from these axioms.

7. Rule #1: Probability of the Empty Set

Key Idea 3

Rule #1. $P(\varnothing) = 0$.

Proof.

- Take the countably infinite collection $A_1 = \varnothing$, $A_2 = \varnothing$, $A_3 = \varnothing, \ldots$
- Because $\varnothing \cap \varnothing = \varnothing$, these events are pairwise disjoint.
- Furthermore, $\varnothing \cup \varnothing \cup \varnothing \cup \cdots = \varnothing$.
- By countable additivity,

$$P(\varnothing) \;=\; P(\varnothing \cup \varnothing \cup \cdots) \;=\; \sum_{i=1}^{\infty} P(\varnothing).$$

The only value of $P(\varnothing)$ that satisfies this is 0. (Axiom 1 forbids negative values, so the only finite x satisfying $x = nx$ for all $n \geq 2$ is $x = 0$.) $\square$

8. Rule #2: Finite Additivity

Key Idea 4

Rule #2. If $A_1, A_2, \ldots, A_k$ is a *finite* collection of pairwise disjoint events, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) \;=\; \sum_{i=1}^{k} P(A_i).$$

This rule is called **finite additivity**.

Proof.

- Take a finite disjoint collection $A_1, \ldots, A_k$. Append empty events $A_{k+1} = \varnothing$, $A_{k+2} = \varnothing, \ldots$ to make a countably infinite collection.
- The infinite union equals the finite one, since appending empties changes nothing: $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{k} A_i$.
- By countable additivity and Rule #1,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{k} P(A_i) + \sum_{i=k+1}^{\infty} P(\varnothing) = \sum_{i=1}^{k} P(A_i). \qquad \square$$

9. Rule #3: Complement Rule

Key Idea 5

Rule #3. For any event A,

$$P(A^C) \;=\; 1 - P(A).$$

Proof.

- $A \cap A^C = \varnothing$, so A and A^C are disjoint.
- $A \cup A^C = S$.
- By finite additivity, $P(S) = P(A \cup A^C) = P(A) + P(A^C)$.
- Axiom (2) gives $P(S) = 1$, so $P(A) + P(A^C) = 1$, and therefore $P(A^C) = 1 - P(A)$. □

10. Rule #4: Probability is at Most 1

Key Idea 6

Rule #4. For any event A, $P(A) \leq 1$.

Proof.

- By Rule #3, $P(A) = 1 - P(A^C)$.
- Axiom (1) says $P(A^C) \geq 0$.
- Therefore $P(A) = 1 - P(A^C) \leq 1$. □

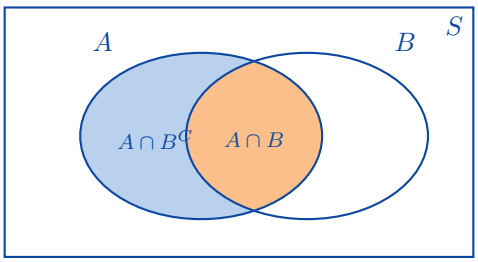
11. Rule #5: Law of Total Probability

Key Idea 7

Rule #5 — Law of Total Probability. For any events A and B,

$$P(A) \;=\; P(A \cap B) + P(A \cap B^C).$$

The intuition is that any outcome of A either lies in B or it does not, and these two pieces are disjoint, so their probabilities add up to $P(A)$.



Proof.

- $(A \cap B) \cap (A \cap B^C) = \varnothing$, so the two events are disjoint.
- $(A \cap B) \cup (A \cap B^C) = A$. (Every element of A is either in B or in B^C.)
- By finite additivity,

$$P(A) = P\big((A \cap B) \cup (A \cap B^C)\big) = P(A \cap B) + P(A \cap B^C). \qquad \square$$

More Probability Rules and Applications

Lecture 4 • UVA Stats

January 2026

Building on the axioms from the previous lecture, we prove a wider toolkit of probability rules — the law of total probability, partition arguments, and an extended inclusion-exclusion — then apply them to genuine problems where computing $P(A)$ directly is hopeless but a clever decomposition is straightforward.

PART 1

Motivation

1. Where Last Lecture Left Us

Last lecture defined a probability $P(A)$ as the limiting relative frequency of event A over many independent trials,

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

and required every probability assignment to satisfy three axioms:

(1) $P(A) \geq 0$ for every event A,

(2) $P(S) = 1$,

(3) *Countable additivity:* for any countable collection of pairwise disjoint events $A_1, A_2, A_3, \ldots$, $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$.

From these we derived five everyday rules: $P(\varnothing) = 0$, finite additivity, the complement rule $P(A^C) = 1 - P(A)$, the upper bound $P(A) \leq 1$, and the Law of Total Probability $P(A) = P(A \cap B) + P(A \cap B^C)$.

Today we add four more rules to that toolkit and put everything to work on two extended practice problems.

PART 2

Methods and Examples

2. Rule #6: The Subset Rule

Key Idea 8

Rule #6 (Subset Rule). If $A \subset B$, then $P(A) \leq P(B)$.

Proof.

- Because $A \subset B$, we have $A \cap B = A$, so $P(A \cap B) = P(A)$.
- Apply the Law of Total Probability to B:

$$P(B) \;=\; P(B \cap A) + P(B \cap A^C) \;=\; P(A) + P(B \cap A^C).$$

- Axiom (1) gives $P(B \cap A^C) \geq 0$, so $P(B) \geq P(A)$. □

3. Rule #7: The General Addition Rule

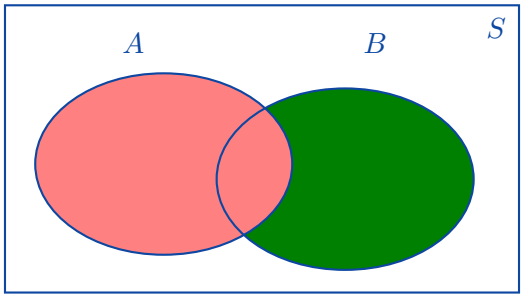
Key Idea 9

Rule #7 (General Addition Rule). For any events A and B,

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

The intuition: $P(A) + P(B)$ double-counts the overlap $A \cap B$, so we subtract it once to get the correct probability of $A \cup B$. Splitting $A \cup B$ into two disjoint pieces makes this precise:

$$A \cup B \;=\; A \;\cup\; (B \cap A^C).$$



Proof.

$$\begin{aligned} P(A \cup B) &= P\big(A \cup (B \cap A^C)\big) \\ &= P(A) + P(B \cap A^C) && \text{(finite additivity, the two pieces are disjoint)} \\ &= P(A) + \big[\,P(B) - P(A \cap B)\,\big] && \text{(Law of Total Probability applied to } B) \\ &= P(A) + P(B) - P(A \cap B). && \square \end{aligned}$$

3.1 Three-event extension

The General Addition Rule extends to three events by inclusion–exclusion:

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ + P(A \cap B \cap C). \end{aligned}$$

4. Rule #8: Countable Subadditivity

Key Idea 10

Rule #8 (Countable Subadditivity). For any countable collection of events $A_1, A_2, A_3, \ldots$ (disjoint or not),

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) \leq \sum_{i=1}^{\infty} P(A_i).$$

The key trick is to convert the original collection into a disjoint one with the same union, then apply countable additivity.

Proof.

- Define a new collection that "carves out" each A_i into the part not yet covered: $B_1 = A_1$, $B_2 = A_2 - A_1$, $B_3 = A_3 - (A_1 \cup A_2)$, $B_4 = A_4 - (A_1 \cup A_2 \cup A_3)$, and so on.
- By construction the B_i are pairwise disjoint and have the same union as the A_i:

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i, \quad \text{so} \quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right).$$

- Countable additivity on the disjoint B_i gives $P(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i)$.
- Each $B_i \subset A_i$, so by the Subset Rule $P(B_i) \leq P(A_i)$ for every i, hence $\sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i)$.
- Stringing these together:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i). \qquad \square$$

5. De Morgan's Laws

Term — De Morgan's Laws

For any finite collection of events $A_1, \ldots, A_k$,

(1) $\left(\bigcup_{i=1}^{k} A_i\right)^C = \bigcap_{i=1}^{k} A_i^C$ "not (any) is the same as (all not)"

(2) $\left(\bigcap_{i=1}^{k} A_i\right)^C = \bigcup_{i=1}^{k} A_i^C$ "not (all) is the same as (some not)"

Taking probabilities of both sides gives the corresponding identities for P.

These are set-theoretic identities, not probability identities, so the proof works at the level of outcomes ω, not probabilities.

Proof of (1). We prove the two inclusions $\left(\bigcup A_i\right)^C \subset \bigcap A_i^C$ and $\bigcap A_i^C \subset \left(\bigcup A_i\right)^C$ separately.

Forward direction: $\left(\bigcup A_i\right)^C \subset \bigcap A_i^C$

- Let $\omega \in \left(\bigcup_{i=1}^k A_i\right)^C$. Then $\omega \notin \bigcup_{i=1}^k A_i$.
- If $\omega \notin \bigcup A_i$, then $\omega \notin A_i$ for every i.
- If $\omega \notin A_i$ for every i, then $\omega \in A_i^C$ for every i.
- If $\omega \in A_i^C$ for every i, then $\omega \in \bigcap_{i=1}^k A_i^C$.

Reverse direction: $\bigcap A_i^C \subset \left(\bigcup A_i\right)^C$

- Let $\omega \in \bigcap_{i=1}^k A_i^C$. Then $\omega \in A_i^C$ for every i.
- If $\omega \in A_i^C$ for every i, then $\omega \notin A_i$ for every i.
- If $\omega \notin A_i$ for every i, then $\omega \notin \bigcup A_i$.
- Therefore $\omega \in \left(\bigcup A_i\right)^C$.

The two inclusions together give equality, so $\left(\bigcup A_i\right)^C = \bigcap A_i^C$, and taking probabilities:

$$P\left(\left(\bigcup_{i=1}^k A_i\right)^C\right) = P\left(\bigcap_{i=1}^k A_i^C\right). \qquad \square$$

Proof of (2). Identical structure, just swapping the roles of $\cup$ and $\cap$. (Or: apply (1) to the collection $\{A_i^C\}$ and complement once more.)

PART 3

Practice Problems

6. Applying the Rules

Applying probability rules requires two skills:

(1) Translating a word problem into probability notation (turning "50% have a Visa" into "$P(V) = 0.5$", etc.).

(2) Identifying which rule to deploy given the information you have.

The next two examples drill both skills.

7. Example #1 — Computing Probabilities from Three Givens

Example #1

Suppose that the following probabilities are true:

$$P(A \cup B) = 0.7, \qquad P(A) = 0.4, \qquad P(B) = 0.5.$$

(a) What is $P(A \cap B)$?
(b) What is $P(A \cap B^C)$?
(c) What is $P(A^C \cap B^C)$?
(d) What is $P(A \cup B^C)$?

Solutions

(a) $P(A \cap B)$ via the General Addition Rule.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ 0.7 &= 0.4 + 0.5 - P(A \cap B) \\ 0.7 &= 0.9 - P(A \cap B) \\ P(A \cap B) &= \mathbf{0.2}. \end{aligned}$$

(b) $P(A \cap B^C)$ via the Law of Total Probability. Apply the Law of Total Probability to A:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^C) \\ 0.4 &= 0.2 + P(A \cap B^C) \\ P(A \cap B^C) &= \mathbf{0.2}. \end{aligned}$$

(c) $P(A^C \cap B^C)$ via De Morgan + complement. By De Morgan's Law (1),

$$P(A^C \cap B^C) \;=\; P\big((A \cup B)^C\big).$$

By the complement rule,

$$P\big((A \cup B)^C\big) \;=\; 1 - P(A \cup B) \;=\; 1 - 0.7 \;=\; \mathbf{0.3}.$$

(d) $P(A \cup B^C)$ via the General Addition Rule. First, $P(B^C) = 1 - P(B) = 1 - 0.5 = 0.5$. Part (b) gave $P(A \cap B^C) = 0.2$. The General Addition Rule applied to A and B^C:

$$\begin{aligned} P(A \cup B^C) &= P(A) + P(B^C) - P(A \cap B^C) \\ &= 0.4 + 0.5 - 0.2 \\ &= \mathbf{0.7}. \end{aligned}$$

8. Example #2 — Visa and MasterCard

Example #2

Suppose that
- 50% of adults have a Visa credit card: $P(V) = 0.50$.
- 40% of adults have a MasterCard: $P(M) = 0.40$.
- 25% of adults have both: $P(V \cap M) = 0.25$.

(a) What is the probability that an adult has at least one of these cards?
(b) What is the probability that an adult does not have a MasterCard?
(c) What is the probability that an adult has neither of these cards?
(d) What is the probability that an adult has a MasterCard but not a Visa?

Solutions

(a) "At least one" = union. "At least one card" is $V \cup M$. General Addition Rule:

$$\begin{aligned} P(V \cup M) &= P(V) + P(M) - P(V \cap M) \\ &= 0.50 + 0.40 - 0.25 \\ &= \mathbf{0.65}. \end{aligned}$$

(b) "Does not have MasterCard" = complement. Complement rule:

$$P(M^C) \;=\; 1 - P(M) \;=\; 1 - 0.40 \;=\; \mathbf{0.60}.$$

(c) "Neither" = intersection of complements. By De Morgan's Law (1), $P(V^C \cap M^C) = P\big((V \cup M)^C\big)$. Complement rule on the answer to part (a):

$$P\big((V \cup M)^C\big) \;=\; 1 - P(V \cup M) \;=\; 1 - 0.65 \;=\; \mathbf{0.35}.$$

(d) "MasterCard but not Visa." This is $M \cap V^C$. Apply the Law of Total Probability to M:

$$\begin{aligned} P(M) &= P(M \cap V) + P(M \cap V^C) \\ 0.40 &= 0.25 + P(M \cap V^C) \\ P(M \cap V^C) &= \mathbf{0.15}. \end{aligned}$$

Conditional Probability, Bayes' Rule, and Independence

Lecture 5 • UVA Stats

January 2026

What is the probability of A given that we already know B? Conditional probability is the engine of statistical reasoning. We derive the multiplication rule, expand it into Bayes' rule (the formula behind every "given the test was positive…" problem), and define independence as the case where conditioning leaves probabilities untouched. Three worked medical-testing examples close the lecture.

PART 1

Motivation

1. Drawing an Ace

Define the event $A = \{\text{the next card I draw is an ace}\}$. A standard deck has 52 cards and 4 aces, so

$$P(A) \;=\; \frac{4}{52} \;\approx\; 0.077.$$

Now suppose I have *already* drawn one card from the deck and it was an ace. Only 51 cards remain and just 3 of them are aces, so my probability of drawing an ace on the next pick is

$$P(A) \;=\; \frac{3}{51} \;\approx\; 0.059.$$

The number $P(A)$ changed because new information arrived.

2. What This Lecture Adds

So far we have only worked with *unconditional* probabilities. We now want to update an event's probability when we learn something. That is the role of the **conditional probability**

$$P(A \mid B) \;=\; \text{“probability of } A \text{ given that } B \text{ has occurred,”}$$

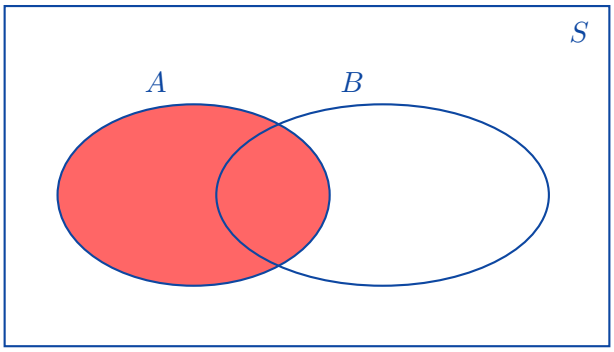
and of two consequences of it: **Bayes' Rule** (which swaps the roles of the two events) and **independence** (when learning B does *not* move $P(A)$).

PART 2

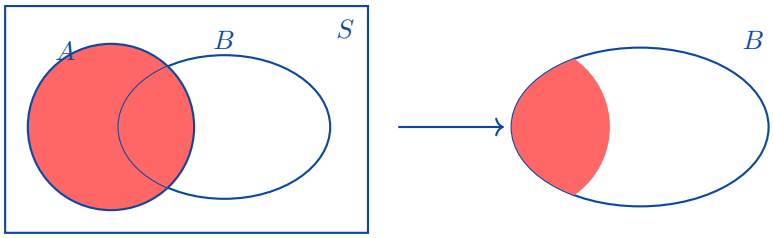
Methods and Examples

3. Conditional Probability

For the unconditional probability $P(A)$, we evaluate the limiting relative frequency of A over *all* realizations of the random phenomenon.



For the *conditional* probability $P(A \mid B)$, we restrict attention to realizations in which B has occurred: B becomes the new sample space, and we ask how much of A lies inside it.



Term — Conditional Probability

For any two events A and B with $P(B) > 0$, the conditional probability of A given B is

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

Worked example: country and rap

Suppose 45% of people like country (A), 25% like rap (B), and 10% like both ($A \cap B$).

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.10}{0.25} = 0.40, \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.10}{0.45} \approx 0.22.$$

Caution: The two conditional probabilities are different objects. In general $P(A \mid B) \neq P(B \mid A)$, $P(A \mid B) \neq P(A)$, and $P(B \mid A) \neq P(B)$.

4. The General Multiplication Rule

Rearranging the definition of conditional probability gives a formula for the probability of an intersection:

Key Idea 11

General Multiplication Rule.

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A).$$

5. Bayes' Rule

Suppose we know $P(A \mid B)$ but want $P(B \mid A)$. Bayes' Rule lets us reverse the conditional:

Key Idea 12

Bayes' Rule. For events A and B with $P(A) > 0$,

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

Proof.

$$\begin{aligned} P(B \mid A) &= \frac{P(A \cap B)}{P(A)} && \text{definition of conditional probability} \\ &= \frac{P(B) \cdot P(A \mid B)}{P(A)}. && \text{General Multiplication Rule on the numerator} \end{aligned}$$

□

Worked example: medical test

Suppose 1 in 20 adults are infected with disease A, so $P(A) = 0.05$. A test has true-positive rate $P(T \mid A) = 0.98$ and false-positive rate $P(T \mid A^C) = 0.04$. What is the probability that an individual has the disease given that they test positive, $P(A \mid T)$?

Bayes' Rule says $P(A \mid T) = \dfrac{P(A) \cdot P(T \mid A)}{P(T)}$. We need $P(T)$. Use the Law of Total Probability:

$$\begin{aligned} P(T) &= P(T \cap A) + P(T \cap A^C) \\ &= P(T \mid A) \cdot P(A) + P(T \mid A^C) \cdot P(A^C) \\ &= (0.98)(0.05) + (0.04)(0.95) \\ &= 0.049 + 0.038 = 0.087. \end{aligned}$$

Then

$$P(A \mid T) \;=\; \frac{(0.05)(0.98)}{0.087} \;\approx\; 0.563.$$

Key Idea 13

Even with a 98%-accurate test, a positive result only gives a 56.3% chance the person has the disease — because the disease is rare to begin with. The base rate matters.

6. Independence

Term — Independence

Two events A and B are **independent**, written $A \perp B$, if the occurrence or non-occurrence of one has no bearing on the probability of the other:

$$P(A \mid B) \;=\; P(A).$$

If $P(A \mid B) \neq P(A)$, the events are **dependent**.

6.1 Symmetry: one direction implies the other

Why does it suffice to verify $P(A \mid B) = P(A)$ and not also $P(B \mid A) = P(B)$? Independence is symmetric in A and B:

$$\begin{aligned} P(B \mid A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(A \mid B) \cdot P(B)}{P(A)} && \text{General Multiplication Rule} \\ &= \frac{P(A) \cdot P(B)}{P(A)} && \text{using } P(A \mid B) = P(A) \\ &= P(B). \end{aligned}$$

6.2 Multiplicative form: a second test for independence

The General Multiplication Rule under independence simplifies dramatically: if $P(A \mid B) = P(A)$, then

$$P(A \cap B) \;=\; P(A \mid B) \cdot P(B) \;=\; P(A) \cdot P(B).$$

This gives us a second equivalent test:

Key Idea 14

Two ways to verify $A \perp B$.
(1) $P(A \mid B) = P(A)$, or equivalently $P(B \mid A) = P(B)$.
(2) $P(A \cap B) = P(A) \cdot P(B)$.

6.3 Independence with complements

If $A \perp B$, then automatically $A \perp B^C$, $A^C \perp B$, and $A^C \perp B^C$.

Proof that $A \perp B \implies A \perp B^C$.

$$\begin{aligned} P(A \mid B^C) &= \frac{P(A \cap B^C)}{P(B^C)} && \text{definition of conditional probability} \\ &= \frac{P(A) - P(A \cap B)}{P(B^C)} && \text{Law of Total Probability on the numerator} \\ &= \frac{P(A) - P(A) \cdot P(B)}{P(B^C)} && \text{using } P(A \cap B) = P(A)P(B) \text{ from independence} \\ &= \frac{P(A)\big(1 - P(B)\big)}{P(B^C)} \\ &= P(A). && \text{since } 1 - P(B) = P(B^C) \end{aligned}$$

□

6.4 Disjoint events are not independent

Suppose A and B are disjoint with $P(A) > 0$ and $P(B) > 0$. Then

$$P(A \cap B) \;=\; P(\varnothing) \;=\; 0, \qquad \text{but} \qquad P(A) \cdot P(B) \;>\; 0.$$

The two values disagree, so disjoint events with positive probability *cannot* be independent. (Knowing B occurred tells you A did *not* — that's a lot of information.)

7. Conditional Independence

It is possible for A and B to be dependent in general but independent given another event C. This is called **conditional independence**.

Term — Conditional Independence

Events A and B are *conditionally independent given* C if

$$P\big((A \cap B) \mid C\big) \;=\; P(A \mid C) \cdot P(B \mid C).$$

Worked example: a die

Toss a fair six-sided die. Let $A = \{1, 2, 3\}$, $B = \{2, 4, 6\}$, $C = \{2, 3, 4, 5\}$.

Are A and B independent?

$$P(A) = \tfrac{1}{2}, \quad P(B) = \tfrac{1}{2}, \quad P(A \cap B) = P(\{2\}) = \tfrac{1}{6}, \quad P(A)P(B) = \tfrac{1}{4}.$$

$\frac{1}{6} \neq \frac{1}{4}$, so A and B are *not* independent.

Are A and B conditionally independent given C? Inside $C = \{2,3,4,5\}$:

$$P(A \mid C) = \frac{|\{2,3\}|}{|\{2,3,4,5\}|} = \tfrac{1}{2}, \qquad P(B \mid C) = \frac{|\{2,4\}|}{|\{2,3,4,5\}|} = \tfrac{1}{2},$$

$$P\big((A \cap B) \mid C\big) = \frac{|\{2\}|}{|\{2,3,4,5\}|} = \tfrac{1}{4} = P(A \mid C) \cdot P(B \mid C).$$

Yes — given C, the two events *are* independent.

PART 3

Practice Problems

8. Example #1 — Visa and MasterCard, Revisited

Example #1

Suppose that $P(V) = 0.50$, $P(M) = 0.40$, and $P(V \cap M) = 0.25$ (V = "has a Visa"; M = "has a MasterCard").

(a) Given an adult has a Visa, what is the probability they have a MasterCard?

(b) Given an adult has a MasterCard, what is the probability they have a Visa?

(c) Are "having a Visa" and "having a MasterCard" independent?

Solutions

(a) $P(M \mid V)$**.**

$$P(M \mid V) = \frac{P(M \cap V)}{P(V)} = \frac{0.25}{0.50} = \mathbf{0.50}.$$

(b) $P(V \mid M)$**.**

$$P(V \mid M) = \frac{P(V \cap M)}{P(M)} = \frac{0.25}{0.40} = \mathbf{0.625}.$$

(c) Independence? By either of the two tests:

$$P(M \mid V) = 0.50 \neq 0.40 = P(M),$$

$$P(V \cap M) = 0.25 \neq 0.50 \cdot 0.40 = P(V) \cdot P(M).$$

So V and M are not independent.

9. Example #2 — Two Dice Rolls

Example #2

Roll a fair six-sided die twice in a row. The two rolls are independent.
(a) What is the probability of rolling a 6 on both tosses?
(b) What is the probability of rolling no 6 on either toss?
(c) What is the probability of rolling a 1 and a 2 (in either order)?

Solutions

(a) Two sixes. By independence,

$$\begin{aligned} P(\text{both sixes}) &= P(\text{toss 1 is 6}) \cdot P(\text{toss 2 is 6}) \\ &= \tfrac{1}{6} \cdot \tfrac{1}{6} = \tfrac{1}{36} \approx \mathbf{0.028}. \end{aligned}$$

(b) Neither toss is a six.

$$\begin{aligned} P(\text{no six}) &= P(\text{toss 1 not 6}) \cdot P(\text{toss 2 not 6}) \\ &= \tfrac{5}{6} \cdot \tfrac{5}{6} = \tfrac{25}{36} \approx \mathbf{0.694}. \end{aligned}$$

(c) One 1 and one 2, in either order. The event splits into two disjoint cases (a 1 on toss 1 followed by a 2 on toss 2, or vice versa):

$$\begin{aligned} P(\text{a 1 and a 2}) &= P\big((\text{toss 1 is 1} \cap \text{toss 2 is 2}) \cup (\text{toss 1 is 2} \cap \text{toss 2 is 1})\big) \\ &= P(\text{toss 1 is 1} \cap \text{toss 2 is 2}) + P(\text{toss 1 is 2} \cap \text{toss 2 is 1}) && \text{(disjoint cases)} \\ &= \tfrac{1}{6} \cdot \tfrac{1}{6} + \tfrac{1}{6} \cdot \tfrac{1}{6} && \text{(independence within each case)} \\ &= \tfrac{2}{36} = \tfrac{1}{18} \approx \mathbf{0.056}. \end{aligned}$$

Counting and Combinatorics in Probability

Lecture 6 • UVA Stats

January 2026

Equally-likely outcomes reduce probability to counting, and counting in turn rests on the multiplication principle, permutations, and combinations. We derive each formula from first principles and apply them to deck-of-cards, lottery, and birthday problems where the right combinatorial structure makes the answer obvious.

PART 1

Motivation

1. Where Last Lecture Left Us

We defined the probability of an event A as the limiting relative frequency in a long sequence of trials,

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

In practice, we cannot run a phenomenon infinitely many times, so probabilities are usually estimated empirically or proposed subjectively.

2. The Easy Special Case

There is one situation in which $P(A)$ can be computed *exactly* with no infinite limit at all: when every outcome of the random phenomenon is equally likely.

Key Idea 15

When all outcomes are equally likely, computing $P(A)$ reduces to counting — count how many outcomes are in A, divide by how many are in S. The whole subject of this lecture is therefore *counting*: tools to count the size of a sample space when listing it by hand is hopeless.

The catch: even simple problems can have huge sample spaces.

- How many possible *sequences* are there for five dice tosses?
- How many ways can we sample ten students from a class of thirty-two?

We need systematic counting tools: the *product rule*, *permutations*, and *combinations*.

PART 2

Methods and Examples

3. Equally Likely Outcomes

Term — Equally Likely Outcomes

Suppose the sample space S of a random phenomenon consists of N equally likely outcomes. Then

$$P(\text{any single outcome}) \;=\; \frac{1}{N},$$

and for any event A made up of n_A of these outcomes,

$$P(A) \;=\; \frac{n_A}{N}.$$

Worked example: one die toss

Toss a fair six-sided die. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$ and each outcome has probability $1/6$. Let A be the event "the result is even," so $A = \{2, 4, 6\}$ with $n_A = 3$. Then

$$P(A) \;=\; \frac{n_A}{N} \;=\; \frac{3}{6} \;=\; 0.5.$$

We have used this rule informally for coin flips and dice tosses already. The rest of the lecture builds tools for the hard part: *counting* N *and* n_A when the sample space is too big to list.

4. The Product Rule

Term — Product Rule (two objects, ordered)

Suppose we form an ordered pair (O_1, O_2), where (O_1, O_2) is treated as different from (O_2, O_1). If O_1 has n_1 possible results and O_2 has n_2 possible results, the total number of pairs is

$$n_{1,2} \;=\; n_1 \cdot n_2.$$

Worked example: two dice

Toss two six-sided dice. The first die has $n_1 = 6$ results; the second has $n_2 = 6$. The total number of ordered outcomes is

$$n_{1,2} \;=\; 6 \cdot 6 \;=\; 36.$$

The Product Rule treats order as significant, so $(1, 2)$ is a different outcome from $(2, 1)$.

4.1 Extension to k-tuples

The Product Rule extends to any number of objects.

Term — Product Rule (k-tuple)

An ordered collection of k objects is a *k-tuple*. If O_i has n_i possible results for $i = 1, 2, \ldots, k$, the total number of k-tuples is

$$n_{1,2,\ldots,k} \;=\; n_1 \cdot n_2 \cdot \cdots \cdot n_k.$$

Worked example: five dice

Toss five six-sided dice. Each die has $n_i = 6$, so the total number of ordered outcomes is

$$n_{1,2,3,4,5} \;=\; 6^5 \;=\; 7\,776.$$

4.2 When does the Product Rule apply?

The Product Rule requires one of two conditions to hold:

(1) Each object in the sequence is drawn from a *distinct* set, or

(2) All objects are drawn from the *same* set *with replacement*.

Term — Sampling with Replacement

Sampling with replacement occurs when each element drawn is returned to the set before the next draw, so the same element can be sampled more than once.

5. Permutations

What if we sample *without* replacement?

Term — Sampling without Replacement

Sampling without replacement occurs when elements are not returned to the set before subsequent elements are drawn, so the same element cannot be sampled twice.

Term — Permutation

A **permutation of size** k is an ordered sequence of k distinct objects taken from n distinct objects without replacement.

5.1 Counting permutations

How many permutations of size k exist from a set of n elements? Walk through the choices one slot at a time:

(1) There are n options for the first slot.

(2) Because the first element is now removed from the pool, there remain $n - 1$ options for the second slot.

(3) ...

(4) There remain $n-k+2$ options for the $(k-1)$-th slot.

(5) There remain $n-k+1$ options for the k-th slot.

By the Product Rule, the total is

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1).$$

5.2 Factorial notation

This product collapses neatly using *factorials*. Recall that

$$m! \;=\; m \cdot (m-1) \cdot (m-2) \cdots 2 \cdot 1, \qquad 0! \;=\; 1.$$

So the number of permutations of size k from n elements is

$$n \cdot (n-1) \cdots (n-k+1) \;=\; \frac{n!}{(n-k)!}.$$

Worked example: presentation orderings

Suppose I divide a class into ten groups for a final project. Five groups present on the second-to-last day and five present on the last day. How many ordered sequences of five presentations are possible on the second-to-last day?

$$\frac{n!}{(n-k)!} \;=\; \frac{10!}{(10-5)!} \;=\; \frac{10!}{5!} \;=\; 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \;=\; 30\,240.$$

5.3 Special case: $k = n$

How many permutations of size k from a set of k elements? Substituting $n = k$:

$$\frac{n!}{(n-k)!} \;=\; \frac{k!}{0!} \;=\; k!.$$

Key Idea 16

A set of k distinct elements admits $k!$ different orderings. This fact is the bridge from permutations to combinations: every unordered collection of k elements corresponds to exactly $k!$ ordered permutations.

6. Combinations

Term — Combination

A **combination of size** k is an *unordered* set of k distinct objects drawn without replacement from n distinct objects.

6.1 Counting combinations

Each combination corresponds to $k!$ permutations (its possible orderings). So we divide:

$$\binom{n}{k} = \frac{\text{permutations of size } k}{k!} = \frac{n!/(n-k)!}{k!} = \frac{n!}{k!\,(n-k)!}.$$

$\binom{n}{k}$ is read as "n choose k."

Worked example: small set comparison

Take $\{A, B, C\}$ and $k = 2$.

- Permutations of size 2:
$$\frac{3!}{(3-2)!} = 6,$$
namely $\{(A,B),(A,C),(B,A),(B,C),(C,A),(C,B)\}$.
- Combinations of size 2:
$$\binom{3}{2} = \frac{3!}{2!\,1!} = 3,$$
namely $\{\{A,B\},\{A,C\},\{B,C\}\}$.

Worked example: presentation groupings (revisited)

How many *combinations* of five second-to-last-day presenters are there from ten groups (order does not matter)?

$$\binom{10}{5} = \frac{10!}{5!\,5!} = 252.$$

PART 3

Practice Problems

7. Example #1 — Pick Three Lottery

Example #1

In the Virginia Lottery's "Pick Three" game, a random sequence of three numbers is drawn. Each number is an integer between 0 and 9 (so digits can repeat). To win the "Exact Order" prize, your three guesses must match the drawn sequence *in order*.
What is the probability of winning the Exact Order prize?

Solution

Each digit is independent and can be 0 through 9, i.e. this is sampling *with replacement*. Apply the Product Rule:

$$N = n_1 \cdot n_2 \cdot n_3 = 10 \cdot 10 \cdot 10 = 1\,000.$$

Exactly one of these 1 000 sequences wins, so

$$P(\text{win}) = \frac{1}{1\,000} = \mathbf{0.001}.$$

(For reference: the payout is \$500 on a \$1 ticket, which is below the fair value of \$1·1000/500 = \$2 per dollar bet — so this game has a negative expected return for the player.)

8. Example #2 — Factory Shift Sampling

Example #2

A small factory employs 15 people on the day shift and 10 on the night shift. Human resources will randomly select 4 of these 25 workers for in-depth interviews.

(a) What is the probability that all four workers come from the day shift?
(b) What is the probability that all four workers come from the night shift?
(c) What is the probability that two workers from each shift are sampled?
(d) What is the probability that both shifts are represented in the sample of four workers?

Solutions

The total number of ways to choose 4 workers from 25 (order doesn't matter) is

$$N = \binom{25}{4} = \frac{25!}{4! \cdot 21!} = 12\,650.$$

This denominator appears in every part below.

(a) All four from the day shift. Choose 4 of the 15 day-shift workers:

$$n_{\text{all day}} = \binom{15}{4} = \frac{15!}{4! \cdot 11!} = 1\,365.$$

Therefore

$$P(\text{all day}) = \frac{1\,365}{12\,650} \approx \mathbf{0.108}.$$

(b) All four from the night shift. Choose 4 of the 10 night-shift workers:

$$n_{\text{all night}} = \binom{10}{4} = \frac{10!}{4! \cdot 6!} = 210.$$

Therefore

$$P(\text{all night}) = \frac{210}{12\,650} \approx \mathbf{0.017}.$$

(c) Two from each shift. Choose 2 of the 15 day-shift workers *and* 2 of the 10 night-shift workers. By the Product Rule (the two choices are independent):

$$n_{2/2} \;=\; \binom{15}{2} \cdot \binom{10}{2} \;=\; 105 \cdot 45 \;=\; 4\,725.$$

Therefore

$$P(\text{two of each}) \;=\; \frac{4\,725}{12\,650} \;\approx\; \mathbf{0.374}.$$

(d) Both shifts represented. The complement of "both shifts represented" is "all four from one shift," which is the disjoint union of (a) and (b). So

$$\begin{aligned} P(\text{both shifts}) &= 1 - P(\text{all same shift}) \\ &= 1 - \big[P(\text{all day}) + P(\text{all night})\big] \\ &= 1 - (0.108 + 0.017) \\ &\approx \mathbf{0.875}. \end{aligned}$$

Random Variables and Probability Distributions

Lecture 7 • UVA Stats

January 2026

A random variable is a function from the sample space to the real line — a way of summarising "what happened" with a single number. We distinguish discrete from continuous random variables, define the probability mass function (pmf) for the discrete case, and establish the two properties every pmf must satisfy. Examples from coin-flips and dice anchor the abstraction.

PART 1

Motivation

1. Where Last Lecture Left Us

We defined the probability of an event A as the limiting relative frequency

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

and used the Kolmogorov axioms, the rules derived from them, conditional probability, independence, Bayes' Rule, and the counting tools of Lecture 6 to compute $P(A)$ in practice.

2. From Sample Spaces to Numbers

The sample space S lists every possible outcome and carries a total probability of 1. The natural next question is: *how* is that probability of 1 spread across the outcomes?

To answer this we replace "the random phenomenon" — which might produce non-numeric outcomes like "heads" or "in state" — with a function that always produces a number. That function is a **random variable**, and the rule that records how its probability is distributed is its **probability distribution**. Today we set up both.

PART 2

Methods and Examples

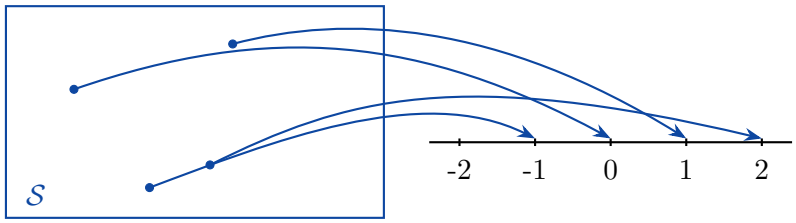
3. Random Variables

Term — Random Variable

A **random variable** is a function that associates a real number with every outcome in the sample space:

$$X : S \to \mathbb{R}.$$

For an outcome $s \in S$ we write $X(s) = x$. We typically denote random variables with capital letters X, Y, Z and their realised values with lower-case x, y, z.



3.1 Choosing the function

- When the outcomes are already numeric, just take $x = s$.
  Example. Let X be the GPA of a randomly selected UVA student.
- When the outcomes are non-numeric, we choose a numeric encoding.
  Example. Let Y be the result of a coin flip:

$$Y = \begin{cases} 0 & \text{if heads } (s = H) \\ 1 & \text{if tails } (s = T). \end{cases}$$

4. Discrete vs. Continuous Random Variables

Term — Discrete Random Variable

A **discrete** random variable is one whose set of possible values is either finite or can be listed as an infinite sequence $x_1, x_2, x_3, \ldots$ (a countable set). *Example:* the result of a coin flip.

Term — Continuous Random Variable

A **continuous** random variable is one whose set of possible values consists of an interval (or union of intervals) on the number line, e.g. $(-\infty, \infty)$, $[0, 1]$, or $[0, 10] \cup [20, 30]$. *Example:* the GPA of a randomly selected UVA student, with possible values $[0, 4]$.

The distinction matters because it changes *how* we assign probabilities.

5. Probability Distributions

Term — Probability Distribution

The **probability distribution** of a random variable X describes how the total probability of 1 is distributed across the values x.

- For *discrete* X we use a **probability mass function** (pmf).
- For *continuous* X we use a **probability density function** (pdf).

Term — Support

The **support** of X is the set of all values x for which the pmf or pdf is defined — equivalently, the range of X.

Example. The coin-flip variable Y above has support $\{0, 1\}$.

6. The Probability Mass Function (pmf)

Term — Probability Mass Function (pmf)

For a discrete random variable X, the **probability mass function** assigns a probability to every value x in the support of X:

$$f(x) \;=\; P(X = x).$$

A function f is a valid pmf if and only if

(1) $f(x) \geq 0$ for all x, and

(2) $\displaystyle\sum_{x} f(x) \;=\; 1.$

Worked example: a fair coin flip

For Y defined above (0 = heads, 1 = tails),

$$f(y) = \begin{cases} 0.5 & y = 0 \\ 0.5 & y = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The optional "0 otherwise" line records explicitly that any value outside the support has zero probability.

Worked example: UVA student residency

Let Z be "a UVA student is in-state." Encode 0 = out-of-state, 1 = in-state. If 67% of UVA students are in-state,

$$f(z) = \begin{cases} 0.33 & z = 0 \\ 0.67 & z = 1. \end{cases}$$

7. Parameters and Families of Distributions

The two pmf's above share a shape. They both look like

$$f(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases}$$

for some $p \in [0, 1]$. The number p is a **parameter**: a single value that picks out which member of a whole family of pmf's we are using ($p = 0.5$ for the coin, $p = 0.67$ for residency).

Term — Parameter and Family

A **parameter** of a probability distribution is a value that pins down which specific pmf or pdf in a family we are working with. The collection of distributions across all admissible parameter values is called a **family** of distributions. *Example:* the coin-flip Y and the residency Z belong to the same family (parameterised by p).

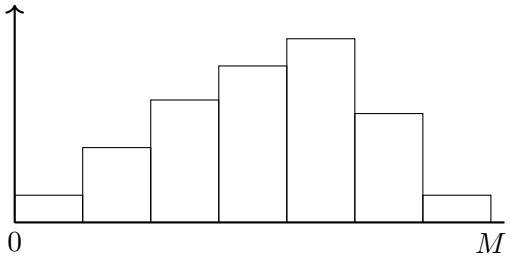
8. The Probability Density Function (pdf)

A discrete random variable has a countable support, so we can list every outcome and assign each its own probability. A continuous random variable has an uncountable support, so no such listing is possible. Instead, probabilities are described by a *density curve*.

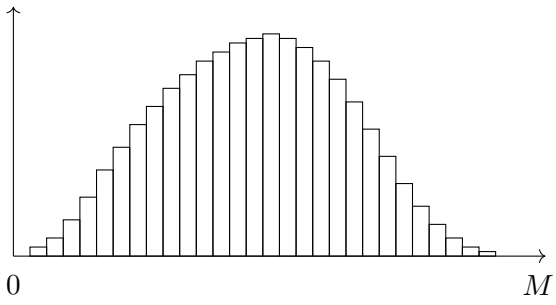
Illustration: pond depth

Let X be the depth of a pond at a randomly chosen surface point, with M the maximum depth, so the support is $[0, M]$. Imagine constructing a histogram of the depth and refining the binning:

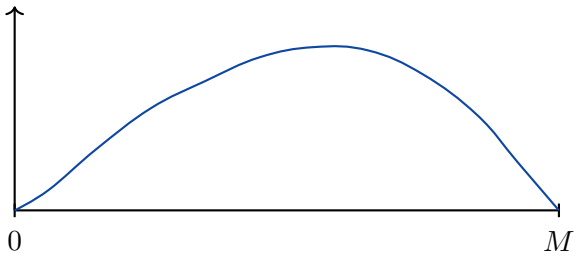
Round to the nearest whole foot:



Round to the nearest tenth of a foot:



Measure exactly (no rounding):



In the limit, the histogram becomes a smooth density curve. Probabilities are no longer bar heights; they are areas under that curve.

Term — Probability Density Function (pdf)

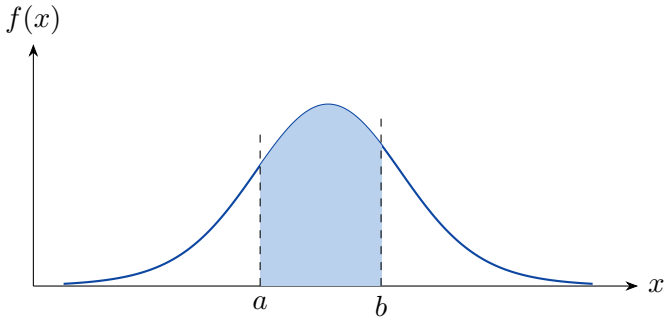
For a continuous random variable X, the **probability density function** is a function $f(x)$ such that, for any $a \le b$,

$$P(a \le X \le b) \;=\; \int_a^b f(x)\,dx.$$

A function f is a valid pdf if and only if

(1) $f(x) \ge 0$ for all x, and

(2) $\displaystyle\int_{\text{support}} f(x)\,dx = 1.$



8.1 A subtle consequence: $P(X = c) = 0$

Even though every outcome x has $f(x) \ge 0$, each *individual* value of a continuous random variable has probability *zero*:

$$P(X = c) \;=\; \int_c^c f(x)\,dx \;=\; 0.$$

Among uncountably many possible outcomes, no single one appears with positive probability. Probability is only positive over intervals (regions of width).

PART 3

Practice Problems

9. Example #1 — A Discrete pmf

Example #1

X is a discrete random variable with pmf

$$f(x) = \begin{cases} 0.5 & x = 1 \\ 0.3 & x = 2 \\ 0.2 & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is the support of X?
(b) Is this a valid pmf?

Solutions

(a) Support of X. The support is the set of *inputs* on which the pmf is nonzero — the possible outcomes of X, not their probabilities. So

$$\text{support}(X) = \mathbf{\{1, 2, 3\}}.$$

(It is *not* $\{0.5, 0.3, 0.2\}$ — those are output values.)

(b) Valid pmf? Check the two conditions:

(1) $f(x) \geq 0$ for all x — yes, every value of f is in $\{0, 0.2, 0.3, 0.5\}$.

(2) $\sum_x f(x) = 0.5 + 0.3 + 0.2 = 1$ — yes.

Both conditions hold, so this is a valid pmf.

10. Example #2 — The Exponential pdf

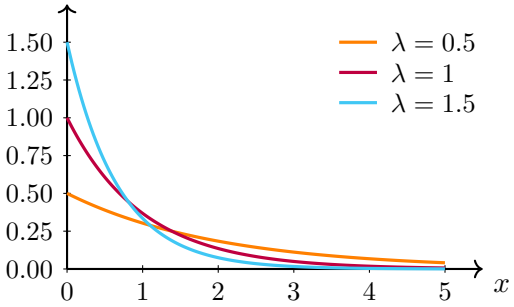
Example #2

X is a continuous random variable with support $[0, \infty)$ and pdf

$$f(x) \;=\; \lambda e^{-\lambda x}, \qquad \lambda > 0.$$

Is f a valid probability density function?

The shape of the pdf depends on λ: larger λ means a steeper drop and shorter expected waiting time.



Solution

Condition (1): non-negativity. $\lambda > 0$ by assumption, and $e^u > 0$ for any real u, so $f(x) = \lambda \cdot e^{-\lambda x} > 0$ for all x.

Condition (2): integrates to 1.

$$\begin{aligned}
\int_{\text{support}} f(x)\,dx &= \int_0^\infty \lambda e^{-\lambda x}\,dx \\
&= \lambda \int_0^\infty e^{-\lambda x}\,dx \\
&= \lambda \left[-\tfrac{1}{\lambda} e^{-\lambda x} \right]_{x=0}^{x=\infty} \\
&= \lambda \left[0 - \left(-\tfrac{1}{\lambda} \cdot 1 \right) \right] = \lambda \cdot \tfrac{1}{\lambda} = 1.
\end{aligned}$$

Both conditions are satisfied. So $\boldsymbol{f(x) = \lambda e^{-\lambda x}}$ is a valid pdf on $[0, \infty)$. This family is the *exponential distribution* with rate parameter λ.

11. Example #3 — Normalizing Constants

Example #3

Y is a continuous random variable with support $[0, 1]$ and proposed pdf

$$f(y) \;=\; y^2 + y.$$

(a) Is $f(y)$ a valid pdf?
(b) If not, find a constant c such that $c \cdot (y^2 + y)$ *is* a valid pdf.

Solutions

(a) Is $f(y) = y^2 + y$ valid? Check non-negativity first: for $y \in [0, 1]$, $y \geq 0$ and $y^2 \geq 0$, so $f(y) = y^2 + y \geq 0$. ✓

Now check the integral:

$$\begin{aligned}\int_{\text{support}} f(y)\,dy &= \int_0^1 (y^2 + y)\,dy \\ &= \left[\frac{y^3}{3} + \frac{y^2}{2}\right]_0^1 \\ &= \tfrac{1}{3} + \tfrac{1}{2} = \tfrac{5}{6}.\end{aligned}$$

The total area is $\frac{5}{6} \neq 1$, so f is not a valid pdf.

(b) Find the normalizing constant c. We want $c \cdot \int_0^1 (y^2 + y)\,dy = 1$. Setting $c = \dfrac{1}{5/6} = \dfrac{6}{5}$,

$$\int_0^1 \tfrac{6}{5}(y^2 + y)\,dy \;=\; \tfrac{6}{5} \cdot \tfrac{5}{6} \;=\; 1.$$

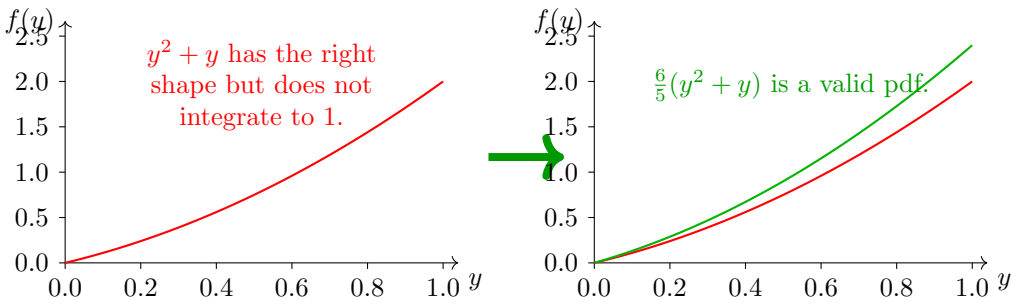
Therefore

$$f(y) \;=\; \tfrac{6}{5}\big(y^2 + y\big), \qquad 0 \leq y \leq 1$$

is a valid pdf.

Key Idea 17

The number $c = \frac{6}{5}$ is called a **regularizing constant** (or **normalizing constant**). Its role is to rescale a function with the right shape but the wrong total area into a valid pdf. Whenever you are handed a non-negative function on a bounded support, computing $c = 1/\int f$ is the standard way to turn it into a density.


$f(y)$
2.5
2.0
1.5
1.0
0.5
0.0
$y^2 + y$ has the right shape but does not integrate to 1.
0.0
0.2
0.4
0.6
0.8
1.0
y
$f(y)$
2.5
2.0
1.5
1.0
0.5
0.0
$\frac{6}{5}(y^2 + y)$ is a valid pdf.
0.0
0.2
0.4
0.6
0.8
1.0
y

Probability Calculations and the Cumulative Distribution Function

Lecture 8 • UVA Stats

January 2026

We extend the pmf machinery from the previous lecture to compute probabilities of arbitrary events, then define the cumulative distribution function (cdf) $F(x) = P(X \leq x)$. The cdf is monotone, right-continuous, and takes values in $[0, 1]$ — properties we verify on several worked discrete examples before computing tail and interval probabilities.

PART 1

Motivation

1. Where Last Lecture Left Us

Last lecture defined a **random variable** X as a function from a sample space S to the real line, split into *discrete* (countable support) and *continuous* (uncountable support). The **probability distribution** of X describes how the total probability of 1 is allocated across its values:

- For a discrete X, the pmf is $f(x) = P(X = x)$.
- For a continuous X, the pdf $f(x)$ gives probabilities by area: $P(a \le X \le b) = \int_a^b f(x)\,dx$.

2. Today's Question

Given the pmf or pdf, how do we actually *compute* $P(\text{event})$ for events written as inequalities — $P(X \le a)$, $P(a < X < b)$, and so on? And is there a single object that encodes *every* such question about X?

The answer to the second question is the **cumulative distribution function** (cdf), and most of this lecture is about working back and forth between the pmf and cdf.

PART 2

Methods and Examples

3. Probability Calculations from a Discrete pmf

For a discrete random variable, an event probability is just a sum of point probabilities:

$$P(a \le X \le b) \;=\; P(X = a) + P(X = a + 1) + \cdots + P(X = b).$$

This is finite or countable additivity applied to the disjoint single-outcome events.

Worked example

Let X be discrete with pmf

x	1	2	3	4	5
$f(x)$	0.10	0.15	0.20	0.25	0.30

Then

$$P(2 \le X \le 4) = f(2) + f(3) + f(4) = 0.15 + 0.20 + 0.25 = 0.60.$$

4. Strict vs. Non-Strict Inequalities

For a discrete X, each value carries positive mass $P(X = a)$, so strict and non-strict inequalities give *different* probabilities:

$$P(X \le a) \;=\; P(X < a) + P(X = a).$$

Define a^- as the next value in the support *below* a, and a^+ as the next value *above* a. Then we can rewrite strict inequalities as non-strict ones:

$$\begin{aligned} P(X < a) &= P(X \le a^-) \\ P(X > a) &= P(X \ge a^+) \\ P(a < X < b) &= P(a^+ \le X \le b^-). \end{aligned}$$

Worked example

Using the same pmf,

$$P(X \le 3) = f(1) + f(2) + f(3) = 0.10 + 0.15 + 0.20 = 0.45,$$

$$P(X < 3) = P(X \le 2) = f(1) + f(2) = 0.10 + 0.15 = 0.25.$$

The two answers differ by $f(3) = 0.20$.

5. The Cumulative Distribution Function

Term — Cumulative Distribution Function (cdf)

The **cumulative distribution function** of a random variable X, denoted $F(x)$, is

$$F(x) \;=\; P(X \le x).$$

The cdf uses a *capital* letter (F); the pmf or pdf uses a lower-case letter (f).

A function F is a valid cdf if and only if it satisfies all four:

(1) F is non-decreasing in x.

(2) F is right-continuous: $\lim_{x \to x_0^+} F(x) = F(x_0)$.

(3) $\lim_{x \to -\infty} F(x) = 0$.

(4) $\lim_{x \to \infty} F(x) = 1$.

6. Discrete CDF as a Sum

For a discrete random variable, the cdf is the running total of the pmf:

$$F(x) \;=\; P(X \le x) \;=\; \sum_{u \,:\, u \le x} f(u).$$

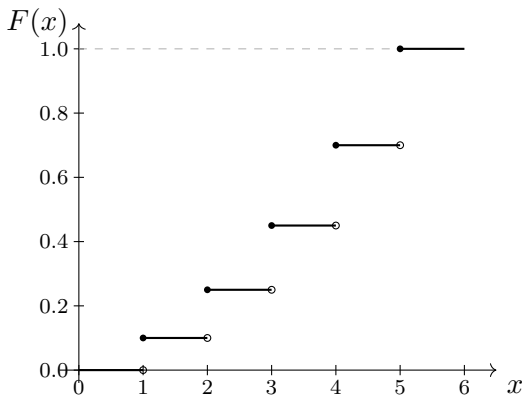
Worked example: pmf to cdf

Continuing with the pmf above:

x	$x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x < 4$	$4 \le x < 5$	$x \ge 5$
$F(x)$	0	0.10	0.25	0.45	0.70	1

7. The CDF as a Step Function

The discrete cdf is a **step function**: flat between support values, jumping by $f(x_i)$ exactly at each support point x_i. The right-continuity in the definition matches the closed circle on the right of each step.



8. Probability Calculations from the CDF

Once we have F, every common probability statement becomes a difference (or one-minus) of cdf values.

Key Idea 18

The CDF dictionary.

$$\begin{aligned} P(X \le a) &= F(a) \\ P(X < a) &= F(a^-) \\ P(X \ge a) &= 1 - F(a^-) \\ P(X > a) &= 1 - F(a) \\ P(a \le X \le b) &= F(b) - F(a^-) \\ P(a \le X < b) &= F(b^-) - F(a^-) \\ P(a < X \le b) &= F(b) - F(a) \\ P(a < X < b) &= F(b^-) - F(a). \end{aligned}$$

Worked example: cdf-driven probabilities

Using the cdf table from the previous worked example:

$$P(2 \le X \le 4) = F(4) - F(1) = 0.70 - 0.10 = 0.60,$$

$$P(X \le 3) = F(3) = 0.45, \qquad P(X < 3) = F(2) = 0.25.$$

9. Going the Other Way: pmf from the cdf

To recover the pmf from the cdf, look at the *jump heights* of the step function. The pmf at a support point x_i is the size of the jump there:

$$f(x_i) \;=\; F(x_i) - F(x_i^-).$$

Worked example

From the cdf table above,

$$f(1) = F(1) - F(<1) = 0.10 - 0 = 0.10, \quad f(2) = 0.25 - 0.10 = 0.15, \quad \ldots$$

which recovers the original pmf.

PART 3

Practice Problems

10. Example #1 — pmf to cdf

Example #1

X is a discrete random variable with pmf

x	1	2	3	4
$f(x)$	0.1	0.2	0.3	0.4

(a) What is the cdf $F(x)$?
(b) What is $P(X \leq 3)$?
(c) What is $P(X < 3)$?
(d) What is $P(2 \leq X < 4)$?

Solutions

(a) Compute $F(x)$ by running totals.

x	$x < 1$	$1 \leq x < 2$	$2 \leq x < 3$	$3 \leq x < 4$	$x \geq 4$
$F(x)$	0	0.1	0.3	0.6	1

(b) $P(X \leq 3) = F(3)$**.**

$$P(X \leq 3) \;=\; F(3) \;=\; \mathbf{0.6}.$$

(c) $P(X < 3) = F(2)$**.**

$$P(X < 3) \;=\; P(X \leq 2) \;=\; F(2) \;=\; \mathbf{0.3}.$$

(d) $P(2 \leq X < 4)$**.**

$$\begin{aligned} P(2 \leq X < 4) &= P(X < 4) - P(X < 2) \\ &= P(X \leq 3) - P(X \leq 1) \\ &= F(3) - F(1) \\ &= 0.6 - 0.1 = \mathbf{0.5}. \end{aligned}$$

11. Example #2 — cdf to pmf

Example #2

Y is a discrete random variable with cdf

y	$y < 0$	$0 \leq y < 5$	$5 \leq y < 10$	$y \geq 10$
$F(y)$	0	0.3	0.8	1

(a) What is the pmf $f(y)$?
(b) What is $P(Y > 5)$?
(c) What is $P(Y \geq 5)$?
(d) What is $P(2 \leq Y \leq 7)$?

Solutions

(a) Read jumps to recover the pmf. The cdf jumps at $y = 0, 5, 10$:

$$\begin{aligned} f(0) &= F(0) - F(<0) = 0.3 - 0 = 0.3, \\ f(5) &= F(5) - F(<5) = 0.8 - 0.3 = 0.5, \\ f(10) &= F(10) - F(<10) = 1.0 - 0.8 = 0.2. \end{aligned}$$

So

$$f(y) = \begin{cases} 0.3 & y = 0 \\ 0.5 & y = 5 \\ 0.2 & y = 10 \\ 0 & \text{otherwise.} \end{cases}$$

(b) $P(Y > 5) = 1 - F(5)$**.**

$$P(Y > 5) \;=\; 1 - F(5) \;=\; 1 - 0.8 \;=\; \mathbf{0.2}.$$

(c) $P(Y \geq 5) = 1 - F(5^-)$**.** The next support point below 5 is 0, so $F(5^-) = F(0) = 0.3$:

$$P(Y \geq 5) \;=\; 1 - F(5^-) \;=\; 1 - 0.3 \;=\; \mathbf{0.7}.$$

(d) $P(2 \leq Y \leq 7)$**.**

$$\begin{aligned} P(2 \leq Y \leq 7) &= F(7) - F(2^-) \\ &= F(7) - F(0) \\ &= 0.8 - 0.3 = \mathbf{0.5}. \end{aligned}$$

Sanity check: the only support point in $[2, 7]$ is $y = 5$, with $f(5) = 0.5$ — agrees.

Continuous Random Variables: pdf, cdf, and Percentiles

Lecture 9 • UVA Stats

January 2026

For continuous random variables, point probabilities vanish and we must work with the probability density function (pdf). We derive the relationship between pdf and cdf via integration, define and compute percentiles (including the median), and close with worked examples on uniform and triangular densities.

PART 1

Motivation

1. Where Last Lecture Left Us

Last lecture introduced the **cumulative distribution function** $F(x) = P(X \le x)$ and used it to compute event probabilities for both discrete and continuous random variables. The probability distribution of a continuous random variable X is described by a **probability density function** (pdf) $f(x)$, with probabilities recovered by area:

$$P(a \le X \le b) \;=\; \int_a^b f(x)\,dx, \qquad P(X = c) \;=\; 0.$$

The cdf packages every such question into one object, $F(x) = P(X \le x)$.

2. Today's Question

This lecture zooms in on the *continuous* case and develops the full toolkit:

- Compute event probabilities directly from the pdf by integration.
- Move from pdf to cdf by integrating, and from cdf back to pdf by differentiating (the Fundamental Theorem of Calculus).
- Use the cdf to compute one-sided, two-sided, and complement probabilities.
- Solve for percentiles by inverting the cdf.

We close with two extended practice problems that drill every operation: starting from a pdf and going forward to probabilities (Example #1), and starting from a cdf and going backward to a pdf and a percentile (Example #2).

PART 2

Methods and Examples

3. Probability Calculations from a pdf

Term — Probability via Density

For a continuous random variable X with pdf f, the probability of an event written as an interval is the area under the density curve over that interval:

$$P(a \le X \le b) \;=\; \int_a^b f(x)\,dx.$$

This stems from the *definition* of a pdf, not from any of the probability axioms or rules.

Worked example

Let X be continuous with pdf $f(x) = 2x$ for $0 \le x \le 1$.

Then

$$P(0.25 \le X \le 0.75) \;=\; \int_{0.25}^{0.75} 2x\,dx \;=\; \big[x^2\big]_{0.25}^{0.75} \;=\; 0.75^2 - 0.25^2 \;=\; \tfrac{9}{16} - \tfrac{1}{16} \;=\; \tfrac{8}{16} \;=\; 0.5,$$

and

$$P(X \le 0.5) \;=\; \int_0^{0.5} 2x\,dx \;=\; \big[x^2\big]_0^{0.5} \;=\; 0.5^2 - 0 \;=\; 0.25.$$

4. Strict vs. Non-Strict Inequalities (Continuous)

When X is continuous, every single point has zero probability mass:

$$P(X = a) \;=\; 0 \quad \text{for every } a.$$

This follows from the integral interpretation: $P(X = a) = \int_a^a f(x)\,dx = 0$, since the integral of any function over a single point is zero. Consequently

$$P(X \le a) \;=\; P(X < a), \qquad P(a < X < b) \;=\; P(a \le X \le b),$$

and so on — strict and non-strict inequalities give the *same* probability. Contrast this with the discrete case from last lecture, where the two differ by $P(X = a)$.

5. The Cumulative Distribution Function

Term — Cumulative Distribution Function (cdf)

The **cumulative distribution function** of a random variable X, denoted $F(x)$, is

$$F(x) \;=\; P(X \le x).$$

The cdf uses a capital letter F; the pmf/pdf uses a lowercase f.

For F to be a valid cdf, it must satisfy four requirements:

1. F is a non-decreasing function of x.
2. F is right-continuous: $\lim_{x \to x_0^+} F(x) = F(x_0)$ for every x_0.
3. $\lim_{x \to -\infty} F(x) = 0$.
4. $\lim_{x \to \infty} F(x) = 1$.

6. Computing a cdf from a pdf

For a continuous random variable, the cdf is obtained by integrating the pdf from the lower bound of the support up to x:

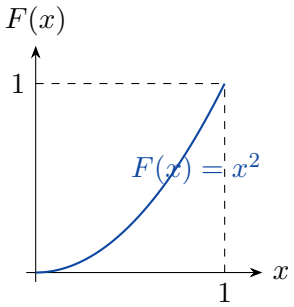
Key Idea 19

$$F(x) \;=\; P(X \le x) \;=\; \int_{\min}^{x} f(u)\,du.$$

Worked example

Let X be continuous with pdf $f(x) = 2x$ on $[0, 1]$. Then

$$F(x) \;=\; \int_0^x 2u\,du \;=\; \big[u^2\big]_0^x \;=\; x^2.$$



7. Probability Calculations from a cdf

Once we have F, every interval probability becomes a difference (or complement) of F-values:

Key Idea 20

$$\begin{aligned} P(X \le a) &= F(a), \\ P(X < a) &= F(a) \quad \text{(continuous case)}, \\ P(X \ge a) = P(X > a) &= 1 - F(a), \\ P(a \le X \le b) &= F(b) - F(a). \end{aligned}$$

Worked example

For $F(x) = x^2$ on $[0, 1]$,

$$P(0.25 \le X \le 0.75) \;=\; F(0.75) - F(0.25) \;=\; 0.75^2 - 0.25^2 \;=\; \tfrac{9}{16} - \tfrac{1}{16} \;=\; 0.5,$$

$$P(X \le 0.5) \;=\; F(0.5) \;=\; 0.5^2 \;=\; 0.25.$$

8. Recovering a pdf from a cdf

Going the other direction, the pdf is the derivative of the cdf. This is the Fundamental Theorem of Calculus:

Key Idea 21

If $F(x) = \int_a^x f(u)\,du$ on $[a, b]$, then

$$\frac{d}{dx}F(x) \;=\; f(x) \quad \text{for all } x \in (a, b).$$

Worked example

Let X be continuous with cdf $F(x) = x^2$ on $[0, 1]$. Differentiate to recover the pdf:

$$f(x) \;=\; \frac{d}{dx}F(x) \;=\; \frac{d}{dx}x^2 \;=\; 2x.$$

9. Percentiles via the cdf

The cdf also lets us answer the *inverse* question: which x-value has a given cumulative probability?

Term — Percentile Function

Define $Q(p)$ to be the p^{th} percentile of the distribution of X. It is determined by

$$P\big(X \le Q(p)\big) \;=\; p, \qquad P\big(X \ge Q(p)\big) \;=\; 1-p.$$

Because $P(X \le Q(p)) = F(Q(p)) = p$, we solve for $Q(p)$ by setting the cdf equal to p and inverting:

$$F\big(Q(p)\big) \;=\; p \quad\Longrightarrow\quad Q(p) \;=\; F^{-1}(p).$$

Worked examples

Let $F(x) = x^2$ on $[0,1]$.

25$^{\text{th}}$ percentile.

$$\begin{aligned} F(x) &= 0.25 \\ x^2 &= 0.25 \\ x &= \sqrt{0.25} \;=\; 0.5, \end{aligned}$$

so $Q(0.25) = 0.5$.

Median (50$^{\text{th}}$ percentile).

$$\begin{aligned} F(x) &= 0.5 \\ x^2 &= 0.5 \\ x &= \sqrt{0.5} \;\approx\; 0.71, \end{aligned}$$

so $Q(0.5) \approx 0.71$.

PART 3

Practice Problems

Example #1 — Going forward from the pdf

Key Idea 22

X is a continuous random variable with support $[0,1]$ and pdf

$$f(x) \;=\; \frac{3x^2}{2} \;+\; x.$$

(a) What is the cdf $F(x)$?

(b) Find $P(X < 0.5)$.

(c) Find $P(X > 0.75)$.

(d) Find $P(0.25 < X < 0.5)$.

(e) Find $P(X > 0.25 \mid X < 0.5)$.

Solutions

(a) The cdf $F(x)$. Integrate the pdf from 0 to x:

$$\begin{aligned} F(x) &= \int_0^x \left(\frac{3u^2}{2} + u\right) du \\ &= \left[\frac{u^3}{2} + \frac{u^2}{2}\right]_{u=0}^{u=x} \\ &= \frac{x^3}{2} + \frac{x^2}{2}. \end{aligned}$$

(b) $P(X < 0.5)$. For a continuous X, $P(X < 0.5) = F(0.5)$:

$$\begin{aligned} P(X < 0.5) = F(0.5) &= \frac{0.5^3}{2} + \frac{0.5^2}{2} \\ &= \frac{1}{16} + \frac{1}{8} = \frac{1}{16} + \frac{2}{16} = \frac{3}{16} = 0.1875. \end{aligned}$$

(c) $P(X > 0.75)$. Use the complement:

$$\begin{aligned} P(X > 0.75) &= 1 - F(0.75) \\ &= 1 - \left[\frac{0.75^3}{2} + \frac{0.75^2}{2}\right] \\ &= 1 - \left[\frac{27}{128} + \frac{9}{32}\right] \\ &= 1 - \frac{27 + 36}{128} = 1 - \frac{63}{128} = \frac{65}{128} \approx 0.51. \end{aligned}$$

(d) $P(0.25 < X < 0.5)$. Difference of cdf values:

$$\begin{aligned} P(0.25 < X < 0.5) &= F(0.5) - F(0.25) \\ &= \left[\frac{1}{16} + \frac{1}{8}\right] - \left[\frac{0.25^3}{2} + \frac{0.25^2}{2}\right] \\ &= \frac{3}{16} - \left[\frac{1}{128} + \frac{1}{32}\right] \\ &= \frac{3}{16} - \frac{5}{128} = \frac{24}{128} - \frac{5}{128} = \frac{19}{128} \approx 0.15. \end{aligned}$$

(e) $P(X > 0.25 \mid X < 0.5)$**.** Apply the definition of conditional probability and reuse parts (b) and (d):

$$\begin{aligned} P(X > 0.25 \mid X < 0.5) &= \frac{P(X > 0.25 \cap X < 0.5)}{P(X < 0.5)} \\ &= \frac{P(0.25 < X < 0.5)}{P(X < 0.5)} \\ &= \frac{19/128}{3/16} = \frac{19}{128} \cdot \frac{16}{3} = \frac{19}{24} \approx 0.79. \end{aligned}$$

The intersection $\{X > 0.25\} \cap \{X < 0.5\}$ is just the interval $0.25 < X < 0.5$. Dividing by $P(X < 0.5) = 3/16$ rescales that probability to live inside the conditioning event.

Example #2 — Going backward from the cdf

Key Idea 23

Y is a continuous random variable with support $[-1, 1]$ and cdf

$$F(y) = \frac{y+1}{2}.$$

(a) What is the pdf $f(y)$?

(b) Find $P(Y > 0.25)$.

(c) Find $P(-0.5 \le Y \le 0.5)$.

(d) Find the 90^{th} percentile of Y.

Solutions

(a) The pdf $f(y)$**.** Differentiate the cdf:

$$\begin{aligned} f(y) = \frac{d}{dy}F(y) &= \frac{d}{dy}\left(\frac{y+1}{2}\right) \\ &= \frac{1}{2}\frac{d}{dy}(y+1) = \frac{1}{2}(1) = \frac{1}{2} \quad \text{for } -1 \le y \le 1. \end{aligned}$$

So Y is uniform on $[-1, 1]$ (constant density $\frac{1}{2}$).

(b) $P(Y > 0.25)$**.**

$$\begin{aligned} P(Y > 0.25) &= 1 - F(0.25) \\ &= 1 - \frac{0.25+1}{2} = 1 - 0.625 = 0.375. \end{aligned}$$

(c) $P(-0.5 \le Y \le 0.5)$**.**

$$\begin{aligned} P(-0.5 \le Y \le 0.5) &= F(0.5) - F(-0.5) \\ &= \frac{0.5+1}{2} - \frac{-0.5+1}{2} \\ &= 0.75 - 0.25 = 0.5. \end{aligned}$$

(d) $90^{\textbf{th}}$ **percentile.** Solve $F(y) = 0.9$:

$$\begin{aligned} \frac{y+1}{2} &= 0.9 \\ y+1 &= 1.8 \\ y &= 0.8. \end{aligned}$$

So $Q(0.9) = 0.8$: 90% of the probability mass of Y lies at or below $y = 0.8$.

Expected Value, Variance, and Moments

Lecture 11 • UVA Stats

February 2026

Once we have a distribution, what summary statistics describe its shape? The expected value $E[X]$ gives the centre, the variance $\mathrm{Var}(X)$ gives the spread, and the higher moments capture skewness and kurtosis. We derive each from the pmf/pdf, prove linearity of expectation, and work through several examples including a discrete random variable in a casino game.

PART 1

Motivation

1. Where Last Lecture Left Us

A **random variable** X is a function from a sample space S to the real line, classified as *discrete* (countable support) or *continuous* (uncountable support). Its **probability distribution** describes how the total probability of 1 is allocated:

- For a discrete X, the pmf is $f(x) = P(X = x)$.
- For a continuous X, the pdf $f(x)$ gives probabilities by area: $P(a \le X \le b) = \int_a^b f(x)\,dx$.

2. Describing a Distribution

Beyond computing probabilities, we summarise distributions along three axes: **shape**, **center**, and **spread**.

- **Shape** captures *modality* (the number of peaks), *skewness* (symmetric vs. tailing right or left), and *tail weight* (light or heavy tails).
- **Center** locates the "middle": median or mean.
- **Spread** measures how far probability spreads from the center: range, percentiles, or variance.

3. Today's Question

Last lecture, we used the cdf to compute medians and other percentiles. But two of the most informative summaries — the *mean* and the *variance* — are weighted averages, not order statistics, so the cdf approach does not apply directly.

Key Idea 24

How can we calculate measures that are an *average*, like the mean and variance?

The answer is the **expected value** operator $E[\cdot]$, which takes a weighted average of any quantity computed from X. From it we will derive the mean, the variance, the higher moments, and a clean algebra for linear transformations.

PART 2

Methods and Examples

4. Expected Value

Term — Expected Value

The **expected value** is a function $E[\,\cdot\,]$ that takes a weighted average of whatever is input to it. When the input is a random variable X, the weights are the probabilities (mass or density) of the outcomes:

$$\text{discrete:} \quad E[X] = \sum_{x} x \cdot f(x),$$

$$\text{continuous:} \quad E[X] = \int_{\text{support}} x \cdot f(x)\,dx.$$

Worked example (discrete)

Let X have pmf

x	1	2	3	4
$f(x)$	0.05	0.25	0.30	0.40

Then

$$E[X] = (1)(0.05) + (2)(0.25) + (3)(0.30) + (4)(0.40) = 0.05 + 0.50 + 0.90 + 1.60 = 3.05.$$

Worked example (continuous)

Let X have pdf $f(x) = \frac{3}{2}(1 - x^2)$ for $x \in [0, 1]$. Then

$$E[X] = \int_0^1 x \cdot \frac{3}{2}(1 - x^2)\,dx = \frac{3}{2}\int_0^1 (x - x^3)\,dx$$
$$= \frac{3}{2}\left[\frac{x^2}{2} - \frac{x^4}{4}\right]_0^1 = \frac{3}{2}\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{3}{2} \cdot \frac{1}{4} = \frac{3}{8}.$$

5. Law of the Unconscious Statistician (LOTUS)

We can take expectations not just of X itself, but of any function $h(X)$:

Key Idea 25

LOTUS. For any reasonable function h,

$$\text{discrete:} \quad E[h(X)] = \sum_x h(x)\, f(x),$$

$$\text{continuous:} \quad E[h(X)] = \int_{\text{support}} h(x)\, f(x)\, dx.$$

The name reflects the fact that we use $f(x)$ directly without first deriving the distribution of $h(X)$.

6. Linear Transformations of $E[X]$

A particularly common case is $h(X) = a + bX$: a *linear transformation* (e.g. converting feet to inches via $h(X) = 12X$, or Celsius to Fahrenheit via $h(Y) = 32 + \frac{9}{5}Y$).

Key Idea 26

$$E[a + bX] = a + b\, E[X].$$

Proof (discrete case)

$$\begin{aligned} E[a + bX] &= \sum_x (a + bx)\, f(x) \\ &= a \sum_x f(x) + b \sum_x x\, f(x) \\ &= a\,(1) + b\, E[X] = a + b\, E[X]. \end{aligned}$$

The continuous case is identical with sums replaced by integrals.

Special case: expectation of a constant

Setting $b = 0$,

$$E[a] = E[a + 0 \cdot X] = a + 0 \cdot E[X] = a.$$

The weighted average of a single non-random value equals itself.

7. Variance

Term — Variance

The **variance** of X is the expected squared deviation from $E[X]$:

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

This is just LOTUS applied to $h(X) = (X - E[X])^2$:

$$\text{discrete:} \quad \operatorname{Var}(X) = \sum_x \big(x - E[X]\big)^2 f(x),$$

$$\text{continuous:} \quad \operatorname{Var}(X) = \int_{\text{support}} \big(x - E[X]\big)^2 f(x)\,dx.$$

Alternative formula

Because $E[X]$ is a constant once X's distribution is fixed, we can pull it through expectations:

$$\begin{aligned}\operatorname{Var}(X) &= E\big[(X - E[X])^2\big] \\ &= E\big[X^2 - 2X\,E[X] + E[X]^2\big] \\ &= E[X^2] - 2\,E[X]\,E[X] + E[X]^2 \\ &= E[X^2] - 2\,E[X]^2 + E[X]^2\end{aligned}$$

Key Idea 27

$$\operatorname{Var}(X) \;=\; E[X^2] - E[X]^2.$$

This formula is usually faster: compute the second moment, square the first moment, subtract.

Worked example

Continuing with $f(x) = \frac{3}{2}(1 - x^2)$ on $[0, 1]$ (we already have $E[X] = \frac{3}{8}$),

$$\begin{aligned}E[X^2] &= \int_0^1 x^2 \cdot \frac{3}{2}(1 - x^2)\,dx = \frac{3}{2}\int_0^1 (x^2 - x^4)\,dx \\ &= \frac{3}{2}\left[\frac{x^3}{3} - \frac{x^5}{5}\right]_0^1 = \frac{3}{2}\left(\frac{1}{3} - \frac{1}{5}\right) = \frac{3}{2} \cdot \frac{2}{15} = \frac{1}{5}.\end{aligned}$$

Then

$$\operatorname{Var}(X) \;=\; E[X^2] - E[X]^2 \;=\; \tfrac{1}{5} - \left(\tfrac{3}{8}\right)^2 \;=\; \tfrac{1}{5} - \tfrac{9}{64} \;=\; \tfrac{19}{320} \;\approx\; 0.06.$$

8. Variance under a Linear Transformation

Key Idea 28

$$\operatorname{Var}(a + bX) \;=\; b^2 \operatorname{Var}(X).$$

The shift a drops out (shifting all values does not change spread); the scale b appears squared because variance is in squared units. This identity follows from linearity of expectation applied twice: shifting by a does not change spread, and scaling by b multiplies deviations by b, hence variances by b^2.

Proof

Using the alternative formula,

$$\begin{aligned}\mathrm{Var}(a+bX) &= E\big[(a+bX)^2\big] - E[a+bX]^2 \\ &= E\big[a^2 + 2abX + b^2X^2\big] - (a + b\,E[X])^2 \\ &= \big(a^2 + 2ab\,E[X] + b^2\,E[X^2]\big) - \big(a^2 + 2ab\,E[X] + b^2\,E[X]^2\big) \\ &= b^2\big(E[X^2] - E[X]^2\big) = b^2\,\mathrm{Var}(X).\end{aligned}$$

9. Moments and Central Moments

Term — Moments

The k^{th} **moment** of X is $E[X^k]$. The k^{th} **central moment** is $E\big[(X - E[X])^k\big]$.

k	**moment** $E[X^k]$	**central moment**
1	$E[X]$	$E[X - E[X]] = 0$
2	$E[X^2]$	$E[(X - E[X])^2]$
3	$E[X^3]$	$E[(X - E[X])^3]$
k	$E[X^k]$	$E[(X - E[X])^k]$

Moments and central moments package up familiar descriptors:

- **Mean.** $E[X]$ (first moment).
- **Variance.** $\mathrm{Var}(X) = E[(X - E[X])^2]$ (second central moment).
- **Skewness.** $\dfrac{E[(X - E[X])^3]}{\mathrm{Var}(X)^{3/2}}$ — a function of the third central moment; measures asymmetry.
- **Kurtosis.** $\dfrac{E[(X - E[X])^4]}{\mathrm{Var}(X)^2}$ — a function of the fourth central moment; informative about tail weight.

The moments of a distribution are *not* parameters of the distribution: they are descriptive measures. Moments are often functions of the parameters — as we will see in the practice problems — but the moments themselves are derived quantities, not parameters.

PART 3

Practice Problems

Example #1 — A Bernoulli random variable

Key Idea 29

X is discrete with pmf

$$f(x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases}$$

(a) Solve for the expected value of X.

(b) Solve for the variance of X.

Solutions

(a) $E[X]$**.**

$$\begin{aligned} E[X] &= \sum_x x \cdot f(x) \\ &= (0)(1-p) + (1)(p) \\ &= p. \end{aligned}$$

(b) $\mathrm{Var}(X)$**.** Apply the definition directly:

$$\begin{aligned} \mathrm{Var}(X) &= \sum_x \big(x - E[X]\big)^2 f(x) \\ &= (0-p)^2(1-p) + (1-p)^2\, p \\ &= p^2(1-p) + (1-p)^2\, p \\ &= p(1-p)\big[\, p + (1-p)\,\big] \\ &= p(1-p). \end{aligned}$$

So a Bernoulli(p) random variable has mean p and variance $p(1-p)$ — both expressed as functions of the single parameter p, as foreshadowed at the end of Part 2.

Example #2 — A continuous triangular density

Key Idea 30

Y is continuous with support $[0, 4]$ and pdf

$$f(y) = \tfrac{1}{8}\, y.$$

(a) Solve for the expected value of Y.

(b) Solve for the variance of Y.

Solutions

(a) $E[Y]$**.**

$$\begin{aligned} E[Y] &= \int_0^4 y \cdot \tfrac{1}{8}\, y \, dy = \tfrac{1}{8} \int_0^4 y^2 \, dy \\ &= \tfrac{1}{8} \left[\tfrac{y^3}{3} \right]_0^4 = \tfrac{1}{8} \cdot \tfrac{64}{3} = \tfrac{8}{3}. \end{aligned}$$

(b) $\mathrm{Var}(Y)$**.** First the second moment:

$$\begin{aligned} E[Y^2] &= \int_0^4 y^2 \cdot \tfrac{1}{8}\, y \, dy = \tfrac{1}{8} \int_0^4 y^3 \, dy \\ &= \tfrac{1}{8} \left[\tfrac{y^4}{4} \right]_0^4 = \tfrac{1}{8} \cdot 64 = 8. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{Var}(Y) &= E[Y^2] - E[Y]^2 \\ &= 8 - \left(\tfrac{8}{3}\right)^2 = 8 - \tfrac{64}{9} \\ &= \tfrac{72}{9} - \tfrac{64}{9} = \tfrac{8}{9}. \end{aligned}$$

Moment Generating Functions

Lecture 12 • UVA Stats

February 2026

The moment generating function $M_X(t) = E[e^{tX}]$ is a single function that, when differentiated and evaluated at zero, hands us every moment of X for free. We derive the MGF for several named families, prove the moment-extraction formula, and use the MGF to identify distributions and compute means and variances cleanly.

PART 1

Motivation

1. Where Last Lecture Left Us

Last lecture introduced the **expected value** operator $E[\,\cdot\,]$, which takes a weighted average of any function of X. From it we built up:

- the mean $E[X]$,
- the variance $\mathrm{Var}(X) = E[X^2] - E[X]^2$, and
- the higher moments $E[X^k]$ and central moments $E[(X - E[X])^k]$.

Computing each moment from scratch via LOTUS requires its own integral or sum:

$$E[X^k] \;=\; \int_{\text{support}} x^k f(x)\,dx \quad \text{or} \quad \sum_x x^k f(x).$$

2. Today's Question

What if a single function of X could package up *every* moment at once, so that we could read off $E[X], E[X^2], E[X^3], \ldots$ by repeated differentiation rather than fresh integration?

Key Idea 31

The **moment generating function** $M_X(t)$ encodes the entire moment sequence of X, and turns moment computation into a derivative-and-evaluate routine.

We will define the MGF, prove that its k^{th} derivative at 0 returns the k^{th} moment, and use it on two extended practice problems — including recovering the mean and variance of an exponential distribution and of a random variable specified *only* by its MGF.

PART 2

Methods and Examples

3. The Moment Generating Function

Term — Moment Generating Function (MGF)

The **moment generating function** of a random variable X is

$$M_X(t) \;=\; E\big[e^{tX}\big].$$

This is just LOTUS applied to $h(X) = e^{tX}$:

$$\text{discrete:}\quad M_X(t) \;=\; \sum_x e^{tx} f(x),$$

$$\text{continuous:}\quad M_X(t) \;=\; \int_{\text{support}} e^{tx} f(x)\,dx.$$

Worked example: a Bernoulli MGF

Let X be discrete with pmf $f(0) = 1 - p,\ f(1) = p$. Then

$$\begin{aligned} M_X(t) &= \sum_x e^{tx} f(x) \\ &= e^{t\cdot 0}(1-p) + e^{t\cdot 1}\, p \\ &= (1-p) + pe^t. \end{aligned}$$

We will reuse this MGF below to recover $E[X] = p$.

4. Recovering Moments by Differentiation

Key Idea 32

$$E[X^k] \;=\; M_X^{(k)}(0) \;=\; \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}.$$

The k^{th} derivative of the MGF, evaluated at $t = 0$, equals the k^{th} moment.

Proof that $M_X^{(1)}(0) = E[X]$ (discrete case)

Differentiate inside the sum:

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\sum_x e^{tx}f(x)\\ &= \sum_x \frac{d}{dt}\left[e^{tx}f(x)\right]\\ &= \sum_x f(x)\cdot\frac{d}{dt}e^{tx} \quad (f(x) \text{ does not depend on } t)\\ &= \sum_x f(x)\cdot xe^{tx} \;=\; \sum_x xe^{tx}f(x).\end{aligned}$$

Substituting $t = 0$:

$$M_X^{(1)}(0) \;=\; \sum_x xe^0 f(x) \;=\; \sum_x xf(x) \;=\; E[X].$$

The continuous case is identical with sums replaced by integrals, and the higher-order derivatives produce $\sum x^k e^{tx}f(x)$ under the same logic, yielding $E[X^k]$ at $t = 0$.

Worked example: the Bernoulli mean from its MGF

We had $M_X(t) = (1-p) + pe^t$. Differentiate:

$$\frac{d}{dt}M_X(t) \;=\; \frac{d}{dt}\left[(1-p)+pe^t\right] \;=\; 0 + pe^t \;=\; pe^t.$$

Hence

$$E[X] \;=\; M_X^{(1)}(0) \;=\; pe^0 \;=\; p.$$

This matches the direct calculation from Lecture 11.

5. Properties of the MGF

The MGF has two structural properties that we will use repeatedly:

Key Idea 33

P1. The MGF *uniquely determines* a random variable's distribution. Two random variables with the same MGF on a neighbourhood of $t = 0$ have the same distribution — just like two variables with the same pmf/pdf or cdf.

P2. If $Y = a + bX$, then

$$M_Y(t) \;=\; e^{at}\,M_X(bt).$$

The proof of P1 requires complex analysis and is beyond our scope; we will use the result throughout to identify distributions from their MGFs.

Proof of P2

$$\begin{aligned} M_Y(t) &= E\big[e^{tY}\big] = E\Big[e^{t(a+bX)}\Big] \\ &= E\Big[e^{ta}\cdot e^{tbX}\Big] = e^{ta}\,E\Big[e^{(tb)X}\Big] \\ &= e^{ta}\,M_X(tb). \end{aligned}$$

PART 3

Practice Problems

Example #1 — An exponential MGF

Key Idea 34

X is continuous with pdf

$$f(x) \;=\; \lambda e^{-\lambda x} \quad \text{for } x \in [0,\infty),\ \lambda > 0.$$

(a) Derive the MGF for X.

(b) Use the MGF to solve for the first moment.

(c) Use the MGF to solve for the second moment.

Solutions

(a) The MGF.

$$\begin{aligned} M_X(t) = E\big[e^{tX}\big] &= \int_0^\infty e^{tx}\cdot\lambda e^{-\lambda x}\,dx \\ &= \lambda\int_0^\infty e^{(t-\lambda)x}\,dx = \lambda\int_0^\infty e^{-(\lambda-t)x}\,dx \\ &= \lambda\left[-\frac{1}{\lambda-t}\,e^{-(\lambda-t)x}\right]_{x=0}^{x=\infty} \\ &= \lambda\left[0-\left(-\frac{1}{\lambda-t}\right)\right] = \frac{\lambda}{\lambda-t}. \end{aligned}$$

This is only valid for $\lambda > t$: if $t \geq \lambda$, the integrand fails to decay and the integral diverges. The MGF is undefined for $t \geq \lambda$. For the moment-recovery trick we only need a neighbourhood of $t = 0$, so for any positive λ the MGF is well defined near $t = 0$.

(b) The first moment $E[X]$. Write $M_X(t) = \lambda(\lambda - t)^{-1}$ and apply the chain rule:

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \lambda \cdot (-1)(\lambda - t)^{-2} \cdot (-1) \\ &= \frac{\lambda}{(\lambda - t)^2}.\end{aligned}$$

Evaluate at $t = 0$:

$$E[X] \;=\; M_X^{(1)}(0) \;=\; \frac{\lambda}{(\lambda - 0)^2} \;=\; \frac{\lambda}{\lambda^2} \;=\; \frac{1}{\lambda}.$$

(c) The second moment $E[X^2]$. Differentiate again:

$$\begin{aligned}\frac{d^2}{dt^2}M_X(t) &= \frac{d}{dt}\left[\lambda(\lambda - t)^{-2}\right] \\ &= \lambda \cdot (-2)(\lambda - t)^{-3} \cdot (-1) = \frac{2\lambda}{(\lambda - t)^3}.\end{aligned}$$

Then

$$E[X^2] \;=\; M_X^{(2)}(0) \;=\; \frac{2\lambda}{\lambda^3} \;=\; \frac{2}{\lambda^2}.$$

> As a bonus, this gives $\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$, the textbook variance of the Exponential(λ) distribution.

Example #2 — Recovering moments from a given MGF

Key Idea 35

Y is continuous with support $[0, \infty)$ and MGF

$$M_Y(t) \;=\; (1 - 5t)^{-2}.$$

(a) Solve for $E[Y]$.

(b) Solve for $\mathrm{Var}(Y)$.

Solutions

(a) $E[Y]$. Differentiate:

$$\begin{aligned}\frac{d}{dt}M_Y(t) &= \frac{d}{dt}\left[(1 - 5t)^{-2}\right] \\ &= (-2)(1 - 5t)^{-3} \cdot (-5) = 10(1 - 5t)^{-3}.\end{aligned}$$

At $t = 0$:

$$E[Y] \;=\; M_Y^{(1)}(0) \;=\; 10\,(1 - 0)^{-3} \;=\; 10.$$

(b) $\mathrm{Var}(Y)$**.** Differentiate again:

$$\begin{aligned}\frac{d^2}{dt^2}M_Y(t) &= \frac{d}{dt}\left[10(1-5t)^{-3}\right] \\ &= 10\cdot(-3)(1-5t)^{-4}\cdot(-5) = 150(1-5t)^{-4}.\end{aligned}$$

At $t=0$:

$$E[Y^2] \;=\; M_Y^{(2)}(0) \;=\; 150\,(1-0)^{-4} \;=\; 150.$$

Then

$$\mathrm{Var}(Y) \;=\; E[Y^2]-E[Y]^2 \;=\; 150-10^2 \;=\; 150-100 \;=\; 50.$$

We solved for both summary statistics without ever knowing the pdf of Y — the MGF carried all the information we needed. (For the curious: $M_Y(t)=(1-5t)^{-2}$ is the MGF of a Gamma($\alpha=2,\beta=5$) distribution.)

Variable Transformations

Lecture 13 • UVA Stats

February 2026

If $Y = g(X)$, what is the distribution of Y? We develop the cdf method for general transformations and the change-of-variables (Jacobian) formula for monotone transformations of continuous random variables. Worked examples include linear, square, and log transformations.

PART 1

Motivation

1. Where We Are Heading

So far we have studied a single random variable X through its pmf/pdf, cdf, and MGF. But many statistical questions involve a *transformation* of X:

- Convert a height X (in feet) to inches: $Y = 12X$.
- Square a normal variable to study its energy: $Y = X^2$.
- Take the negative logarithm of a uniform draw to simulate an exponential variable: $Y = -\ln X$.

Term — Univariate Variable Transformation

A **(univariate) variable transformation** is a new random variable $Y = g(X)$ defined as a function of a single random variable X.

2. Why a New Tool Is Needed

Last lecture we used the *Law of the Unconscious Statistician* to compute $E[g(X)]$ directly from f_X — but LOTUS only gives us the *expected value* of Y, not its distribution.

To specify the distribution of Y we need its pmf, pdf, cdf, or mgf. We already have one rule for the special case of a linear transformation: if $Y = a + bX$, then $M_Y(t) = e^{at} M_X(bt)$ and (by uniqueness of the MGF) the distribution of Y is determined.

But what if g is *not* linear — e.g. $Y = \sqrt{X}$, $Y = 1/X$, $Y = X^2$, or $Y = \ln X$? Today's machinery handles all of those.

Key Idea 36

The **Transformation Theorem** converts the pdf of X into the pdf of $Y = g(X)$ whenever g is monotonic and differentiable, and a small extension covers the non-monotonic case (e.g. $Y = X^2$) by splitting the domain.

PART 2

Methods and Examples

3. The Transformation Theorem

Term — Transformation Theorem

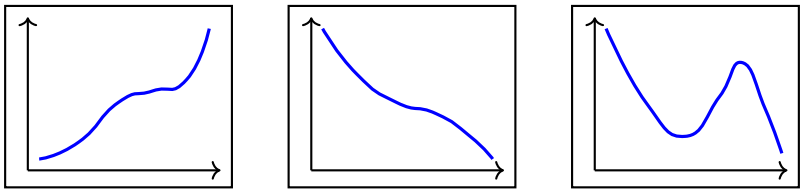
Let X be a continuous random variable, and let $Y = g(X)$ where g is *monotonic* and *differentiable*. Then

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

To find $f_Y(y)$: take f_X, substitute $g^{-1}(y)$ wherever x appears, and multiply by $\left|\frac{d}{dy}g^{-1}(y)\right|$.

What "monotonic" and "differentiable" mean

- **Monotonic.** g is strictly decreasing on its entire domain, or strictly increasing on its entire domain — never both. Strict monotonicity guarantees that g has a well-defined inverse g^{-1}, so the substitution $x = g^{-1}(y)$ makes sense in the change-of-variables formula.
- **Differentiable.** g' exists at every point of the domain.



Yes Yes No

4. Proof of the Transformation Theorem

We prove the formula by working through the cdf of Y and differentiating. The two cases (increasing, decreasing) merge into the single absolute-value formula.

Case 1: g increasing

Then g^{-1} is also increasing, so $\{Y \le y\}$ corresponds to $\{X \le g^{-1}(y)\}$:

$$F_Y(y) = P(Y \le y) = P\big(g(X) \le y\big) = P\big(X \le g^{-1}(y)\big)$$
$$= \int_{x_{\min}}^{g^{-1}(y)} f_X(x)\,dx = F_X\big(g^{-1}(y)\big)\,.$$

Differentiate via the chain rule:

$$f_Y(y) \;=\; \frac{d}{dy} F_X\big(g^{-1}(y)\big) \;=\; f_X\big(g^{-1}(y)\big) \cdot \frac{d}{dy} g^{-1}(y).$$

Because g^{-1} is increasing, the derivative is non-negative, so the formula equals $f_X(g^{-1}(y))\,|\frac{d}{dy} g^{-1}(y)|$.

Case 2: g decreasing

Then g^{-1} is decreasing, so $\{Y \le y\}$ corresponds to $\{X \ge g^{-1}(y)\}$:

$$F_Y(y) = P\big(X \ge g^{-1}(y)\big) = 1 - F_X\big(g^{-1}(y)\big)\,.$$

Differentiate:

$$f_Y(y) \;=\; -f_X\big(g^{-1}(y)\big) \cdot \frac{d}{dy} g^{-1}(y).$$

Because g^{-1} is decreasing, its derivative is non-positive, so the leading minus sign cancels and we again obtain $f_X(g^{-1}(y))\,|\frac{d}{dy} g^{-1}(y)|$.

The absolute-value bars unify both cases.

5. How to Apply the Theorem

Key Idea 37

Three steps:
1. Solve for $X = g^{-1}(Y)$ from the relation $Y = g(X)$.
2. Plug $g^{-1}(y)$ and $\left|\frac{d}{dy} g^{-1}(y)\right|$ into the Transformation Theorem.
3. Determine the support of Y by mapping the support of X through g.

Worked example: $Y = \sqrt{X}$

Let X be continuous with $f_X(x) = \dfrac{2}{x^3}$ for $x > 1$, and let $Y = \sqrt{X}$.

Step 1. $Y = g(X) = \sqrt{X}$, so $X = g^{-1}(Y) = Y^2$.

Step 2.

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \left|\frac{d}{dy} g^{-1}(y)\right|$$
$$= \frac{2}{(y^2)^3} \left|\frac{d}{dy} y^2\right| = \frac{2}{y^6} \cdot |2y| = \frac{4}{y^5}.$$

Step 3. $x > 1 \Rightarrow \sqrt{x} > 1 \Rightarrow y > 1$.

Answer. Y is continuous with $f_Y(y) = \dfrac{4}{y^5}$ for $y > 1$.

6. Non-Monotonic Transformations

What if g is *not* monotonic on the support of X? Then we cannot apply the theorem directly, but the fix is simple:

Key Idea 38

Split the support of X into pieces on which g is monotonic and differentiable, apply the Transformation Theorem on each piece, and add the resulting density contributions over each y-value where multiple x-values map.

Worked example: chi-squared from a standard normal

Let Z be continuous on $(-\infty, \infty)$ with

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

and let $Y = Z^2$. The function $g(z) = z^2$ is *not* monotonic: it decreases on $(-\infty, 0)$ and increases on $(0, \infty)$.

Piece 1: $z < 0$. $Z = g^{-1}(Y) = -\sqrt{Y}$, so

$$\begin{aligned} f_Y^{(1)}(y) &= f_Z(-\sqrt{y}) \left| \frac{d}{dy}(-\sqrt{y}) \right| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left| -\frac{1}{2\sqrt{y}} \right| = \frac{1}{2\sqrt{2\pi y}} e^{-y/2}. \end{aligned}$$

Piece 2: $z > 0$. $Z = g^{-1}(Y) = \sqrt{Y}$, so similarly

$$f_Y^{(2)}(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{2\sqrt{2\pi y}} e^{-y/2}.$$

Combining. Both pieces map to $y > 0$, and each y is hit from *both* $z = -\sqrt{y}$ and $z = +\sqrt{y}$. The two density contributions add:

$$f_Y(y) = \frac{1}{2\sqrt{2\pi y}} e^{-y/2} + \frac{1}{2\sqrt{2\pi y}} e^{-y/2} = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad \text{for } y > 0.$$

This is the χ^2_1 (chi-squared with 1 degree of freedom) density. Squaring a standard normal yields a chi-squared — a fact we will rely on later in distribution theory.

PART 3

Practice Problems

Example #1 — Reciprocal of X

Key Idea 39

X is continuous with $f_X(x) = 2x$ for $0 < x < 1$. If $Y = 1/X$, what is $f_Y(y)$?

Solution

Step 1. $Y = 1/X \Rightarrow X = 1/Y$, so $g^{-1}(y) = 1/y$.

Step 2.

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|$$
$$= 2\left(\frac{1}{y}\right)\left|\frac{d}{dy}\left(\frac{1}{y}\right)\right| = \frac{2}{y}\cdot\left|-\frac{1}{y^2}\right| = \frac{2}{y^3}.$$

Step 3. $x \in (0,1) \Rightarrow 1/x > 1 \Rightarrow y > 1$.

Answer.

$$f_Y(y) = \frac{2}{y^3} \quad \text{for } y > 1.$$

Example #2 — Logarithm of a Uniform

Key Idea 40

X is continuous with $f_X(x) = 1$ for $0 < x < 1$ (a Uniform$(0,1)$ variable). What distribution does $Y = \ln X$ follow?

Solution

Step 1. $Y = \ln X \Rightarrow X = e^Y$, so $g^{-1}(y) = e^y$.

Step 2.

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

Step 3. $0 < x < 1 \Rightarrow 0 < e^y < 1 \Rightarrow y < 0.$

Answer.

$$f_Y(y) \;=\; e^y \quad \text{for } y < 0.$$

This is the density of $-$ Exponential(1): if $W \sim$ Exponential(1) then $-W$ has the density e^y on $y < 0$. Equivalently, $Y = \ln X$ has the same distribution as $-$Exponential(1), which means $-Y = -\ln X$ is itself Exponential(1). This is the inverse-transform sampling trick used to generate exponential variables from uniforms.

Discrete Distribution Families

Lecture 14 • UVA Stats

February 2026

Many real-world experiments share a common mechanism: repeated independent trials with two outcomes per trial. We catalogue four families that live in this setting — Bernoulli, Binomial, Geometric, and Negative Binomial — and for each derive the pmf, MGF, mean, and variance. Three extended practice problems on a shared allergy-medication scenario illustrate how to identify the right family from a problem statement.

PART 1

Motivation

1. Distributions, Parameters, and Families

A **discrete random variable** takes values from a finite or countably infinite set. The pmf $f(x)$ tells us how the total probability of 1 is distributed across that support.

Term — Parameter and Distribution Family

A **parameter** of a probability distribution
- can be assigned any one of a number of possible values,
- specifies the pmf $f(x)$,
- determines a different probability distribution for each value.

A collection of all probability distributions for different values of the parameter(s) is called a **family** of distributions.

2. Today's Question

So far we have computed properties of X one-off: write down a pmf, integrate or sum to find $E[X]$ and $\text{Var}(X)$, differentiate the MGF, etc.

Many real-world situations share the *same* underlying mechanism — repeated independent trials with two possible outcomes — and recur often enough to deserve named distribution families.

Key Idea 41

Today we catalogue four discrete distribution families that all live inside the "n independent success/failure trials" setting: **Bernoulli**, **Binomial**, **Geometric**, and **Negative Binomial**. Each has a recognisable data-generating story, a closed-form pmf, MGF, mean, and variance, and a typical scenario where it is the natural model.

PART 2

Methods and Examples

3. Bernoulli Distribution

Term — Bernoulli(p)

A **Bernoulli random variable** X models a random phenomenon with two possible outcomes — "success" ($X = 1$) and "failure" ($X = 0$). The single parameter is p, the probability of success. Write $X \sim \text{Bern}(p)$.

$$f(x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases}$$

- *Example.* $X = \mathbb{1}\{\text{die toss is a 6}\} \sim \text{Bern}(p = 1/6)$.

Mean, variance, MGF

$$\begin{aligned} E[X] &= p, \\ \text{Var}(X) &= p(1-p), \\ M_X(t) &= (1-p) + pe^t. \end{aligned}$$

(Last lecture's worked example derived all three; repeat the calculations on the right-hand side as needed.)

4. Binomial Distribution

A **binomial random variable** counts the number of successes in a fixed number of independent Bernoulli trials — the sum of n iid $\text{Bern}(p)$ variables.

Term — Binomial Setting

Four criteria define the binomial setting:
1. A fixed number n of trials.
2. Each trial has two outcomes (success/failure).
3. Trials are independent.
4. The success probability p is constant across trials.

Term — Binomial(n, p)

$X \sim \text{Binom}(n, p)$ has pmf

$$P(X = x) \;=\; \binom{n}{x} p^x \,(1-p)^{\,n-x}, \qquad x \in \{0, 1, \ldots, n\},$$

where $\binom{n}{x} = \dfrac{n!}{x!\,(n-x)!}$.

- *Example.* X = number of 6's in five die tosses $\sim \text{Binom}(n = 5,\ p = 1/6)$.

Mean, variance, MGF

$$\begin{aligned} M_X(t) &= \big[(1-p) + pe^t\big]^n, \\ E[X] &= np, \\ \text{Var}(X) &= np(1-p). \end{aligned}$$

The MGF follows from $X = \sum_{i=1}^n X_i$ with $X_i \sim \text{Bern}(p)$ independent: the MGF of a sum of independent variables is the product of MGFs, so $M_X(t) = \prod_{i=1}^n M_{X_i}(t) = [(1-p) + pe^t]^n$.

Worked example: deriving $E[X] = np$ from the MGF

$$\frac{d}{dt} M_X(t) = n\big[(1-p) + pe^t\big]^{n-1} \cdot pe^t.$$

Evaluate at $t = 0$:

$$E[X] \;=\; n\big[(1-p) + p\big]^{n-1} \cdot p \;=\; n \cdot 1 \cdot p \;=\; np.$$

5. Geometric Distribution

A **geometric random variable** counts the number of independent trials *until* the first success. Same independent-trials story as the binomial, except n is no longer fixed — the sequence ends on the first success.

Term — Geometric Setting

1. Sequence of trials with no fixed length; the sequence ends on the first success.
2. Each trial has two outcomes.
3. Trials are independent.
4. Constant success probability p.

Term — Geometric(p)

$X \sim \text{Geom}(p)$ has pmf

$$P(X = x) \;=\; p\,(1-p)^{x-1}, \qquad x \in \{1, 2, 3, \ldots\}.$$

- *Example.* $X =$ number of die tosses until the first 6 $\sim$ Geom$(p = 1/6)$.

Mean, variance, MGF

$$M_X(t) = \frac{pe^t}{1 - e^t(1-p)},$$
$$E[X] = \frac{1}{p},$$
$$\mathrm{Var}(X) = \frac{1-p}{p^2}.$$

6. Negative Binomial Distribution

A **negative binomial random variable** counts the number of *failures* in a sequence of independent trials before the r^{th} success.

Term — Negative Binomial Setting

1. Sequence of trials with no fixed length; the sequence ends on the r^{th} success.
2. Two outcomes per trial.
3. Trials are independent.
4. Constant success probability p.

Term — Negative Binomial(r, p)

$X \sim NB(r, p)$ has pmf

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

- *Example.* $X =$ number of non-6's tossed until the second 6 $\sim NB(r = 2,\ p = 1/6)$.

Mean, variance, MGF

$$M_X(t) = \left(\frac{p}{1 - (1-p)e^t}\right)^r,$$
$$E[X] = \frac{r(1-p)}{p},$$
$$\mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

Geometric vs. negative binomial: a comparison

The geometric and negative binomial distributions are close cousins. The difference is two-dimensional:

	Sequence ends on 1st success	**Sequence ends on r^{th} success**
Counts *failures*	— (special case of NB with $r = 1$)	Negative Binomial
Counts *trials*	Geometric	—

7. Calculation Help (R)

These probabilities are usually computed with software rather than by hand. The standard R helpers are:

Family	$P(X = x)$	$P(X \le x)$
Binomial	`dbinom(x, n, p)`	`pbinom(x, n, p)`
Geometric	`dgeom(x - 1, p)`	`pgeom(x - 1, p)`
Negative Binomial	`dnbinom(x, r, p)`	`pnbinom(x, r, p)`

Caveat for `dgeom/pgeom`: R's geometric helpers want the number of *failures* before the first success, not the number of trials. If your X counts trials, pass `x - 1` as the first argument.

PART 3

Practice Problems

Setup (shared across Examples 1–3). Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. For each scenario:

(a) Specify the distribution and parameter(s).

(b) Solve for $E[X]$.

(c) Solve for $P(X = 3)$.

Example #1 — Failures before the 10th success

Key Idea 42

X = number of patients for which the medication was ineffective *before* it worked effectively ten times.

(a) Distribution. We are counting failures before the $r = 10$ success. This is the negative binomial setting:

$$X \sim NB(r = 10,\ p = 0.8).$$

(b) $E[X]$**.**

$$E[X] = \frac{r(1-p)}{p} = \frac{10\,(1-0.8)}{0.8} = \frac{2}{0.8} = 2.5.$$

(c) $P(X = 3)$**.**

$$P(X=3) = \binom{10+3-1}{3}(0.8)^{10}(0.2)^3$$
$$= \binom{12}{3}(0.8)^{10}(0.2)^3 = 220\,(0.8)^{10}(0.2)^3 \approx 0.19.$$

Example #2 — Successes in five trials

Key Idea 43

In a sample of five new patients, X = number for which the medication was effective.

(a) Distribution. A fixed number of trials ($n = 5$) and we count successes. Binomial setting:

$$X \sim \text{Binom}(n = 5,\ p = 0.8).$$

(b) $E[X]$**.**

$$E[X] = np = 5 \cdot 0.8 = 4.$$

(c) $P(X = 3)$**.**

$$P(X=3) = \binom{5}{3}(0.8)^3(0.2)^{5-3}$$
$$= 10\,(0.8)^3(0.2)^2 \approx 0.20.$$

Example #3 — Ineffective patients before the first success

Key Idea 44

X = number of new patients for which the medication was ineffective *before* the first effective case.

(a) Distribution. We are counting failures before the first success — this is the negative binomial with $r = 1$:

$$X \sim NB(r = 1,\ p = 0.8).$$

Equivalently, $X+1$ (the trial number of the first success) is Geom$(p = 0.8)$, a useful re-parameterisation if you prefer geometric formulas.

(b) $E[X]$. Either using the NB mean directly, or via the geometric-shift relation:

$$E[X+1] = \frac{1}{p} = \frac{1}{0.8} = 1.25$$
$$E[X] = E[X+1] - 1 = 1.25 - 1 = 0.25.$$

The NB formula gives the same answer: $E[X] = r(1-p)/p = 1 \cdot 0.2/0.8 = 0.25$.

(c) $P(X = 3)$. Convert to the geometric (trial-count) random variable $X + 1$:

$$\begin{aligned} P(X = 3) &= P\big((X+1) = 4\big) \\ &= (1 - 0.8)^{4-1} \cdot 0.8 \\ &= (0.2)^3 \cdot 0.8 = 0.0064. \end{aligned}$$

Hypergeometric, Poisson, and Discrete Uniform Distributions

Lecture 17 • UVA Stats

March 2026

We continue the tour of named discrete distributions with three families that live outside the Bernoulli-trial setting: Hypergeometric (sampling without replacement), Poisson (counts of rare events in a fixed interval), and Discrete Uniform (equally-likely outcomes on a finite set). For each we derive the pmf, MGF, mean, and variance, and close with two extended practice problems — a tagged-deer ecology study and a hospital-births scenario.

PART 1

Motivation

1. Beyond Bernoulli trials

A **discrete random variable** is a random variable whose possible values either (i) constitute a finite set or (ii) can be listed in an infinite sequence in which there is a first element, second element, and so on. A **parameter** for a probability distribution can be assigned any one of a number of possible values; it specifies the pmf $f(x)$ and determines a different probability distribution for each value. A collection of all probability distributions for different values of the parameter is called a **family** of distributions.

So far in the course we have built up a family of distributions centred on the **Bernoulli distribution**, which models a random phenomenon with two outcomes ("success" and "failure") with success probability p. Several other common families count outcomes in a series of independent Bernoulli trials:

- **Binomial distribution**: number of successes in n trials.
- **Geometric distribution**: number of trials until the first success.
- **Negative binomial distribution**: number of failures before the r^{th} success.

Lecture 17 introduces three additional families that live *outside* the Bernoulli-trial setting: the **Hypergeometric** distribution (sampling without replacement), the **Poisson** distribution (counting events in a fixed interval), and the **Discrete Uniform** distribution (equally likely outcomes).

Key Idea 45

Different sampling structures call for different distribution families — replacement vs. no replacement, fixed-trial vs. fixed-interval counts, equally-likely outcomes.

PART 2

Methods and Examples

2. The Hypergeometric Distribution

Term — Hypergeometric Distribution

Suppose there exists a finite population of N elements, each of which can be categorised as either a success or a failure, so that there are M successes and $N - M$ failures. A random sample of n elements is selected from this population without replacement. This is the **hypergeometric setting**.

A **hypergeometric random variable** X models the number of successes in such a sample. We can think of X as the sum of n Bernoulli random variables, but these Bernoulli random variables are not independent: the probability of success for the first selection is M/N, while the probability of success for the second selection is $(M-1)/(N-1)$ if the first selection was a success and $M/(N-1)$ if the first selection was a failure.

The distribution is defined by three parameters:

- n — the number of elements sampled,
- M — the total number of successes in the population,
- N — the total population size.

We write $X \sim \mathit{Hyper}(n, M, N)$. For example,

$$X = \textit{number of clubs in a hand of five cards} \sim \mathit{Hyper}(n = 5,\ M = 13,\ N = 52).$$

Common real-world settings for the hypergeometric distribution include:

- Dealing a hand of five cards from a standard deck of 52 cards and asking how many clubs appear in the hand.
- Selecting a sample of size n from a finite population (such as UVA students) and asking how many statistics majors appear in the sample.

Key Idea 46

For $X \sim \mathit{Hyper}(n, M, N)$, the probability mass function is

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

where $\binom{a}{b} = \dfrac{a!}{b!\,(a-b)!}$.

The expected value of a hypergeometric random variable is

$$E[X] = \frac{nM}{N}.$$

The variance of a hypergeometric random variable is

$$\mathrm{Var}(X) = \left(\frac{N-n}{N-1}\right)\left(\frac{nM}{N}\right)\left(1 - \frac{M}{N}\right).$$

Worked example: clubs in a poker hand

What does data look like for a hypergeometric random variable? Consider $X = \textit{clubs in a hand} \sim \textit{Hyper}(n = 5,\ M = 13,\ N = 52)$.

Random Phenomenon	Random Variable
	$X = 1$
	$X = 0$
	$X = 5$

With $X = \textit{number of clubs in a hand of five cards} \sim \text{Hyper}(n = 5,\ M = 13,\ N = 52)$, what is $P(X = 1)$?

- There are $\binom{13}{1}$ ways to choose 1 club from the 13 clubs in the deck.
- There are $\binom{39}{4}$ ways to choose 4 non-clubs from the 39 non-clubs in the deck.
- There are $\binom{13}{1} \cdot \binom{39}{4}$ ways to choose 1 club and 4 non-clubs.
- There are a total of $\binom{52}{5}$ ways to choose 5 cards from 52.

Therefore,

$$P(X = 1) = \frac{\binom{13}{1} \cdot \binom{39}{4}}{\binom{52}{5}}.$$

3. The Poisson Distribution

Term — Poisson Distribution

All of the discrete distribution families discussed so far count the number of successes or failures in a certain number of trials. A **Poisson random variable** counts the number of successes in a *fixed interval* (time, space, etc.). Examples include the number of emails received during a workday and the number of scratches on a car-rental return.

Three assumptions must be satisfied for the **Poisson setting**:

1. The number of successes that occur in any unit of measure is independent of the number of successes that occur in any non-overlapping unit of measure.
2. The probability that a success will occur in any unit of measure is the same for all units of equal size.
3. The probability that two or more successes occur in a unit of measure approaches 0 as the size of the unit becomes smaller.

Worked example: are emails Poisson?

Consider $X =$ *the number of emails I receive during the workday.* We check each of the three Poisson assumptions in turn.

Assumption 1 (Independence). The number of successes that occur in any unit of measure is independent of the number of successes that occur in any non-overlapping unit of measure. Is the number of emails received today between 1:00 and 2:00 independent of the number of emails received today between 3:30 and 4:00?

Assumption 2 (Equal probability per unit). The probability that a success will occur in any unit of measure is the same for all units of equal size. Is the probability of receiving an email today between 1:00 and 2:00 the same as the probability of receiving an email today between 3:00 and 4:00?

Assumption 3 (Rare events in small units). The probability that two or more successes occur in a unit of measure approaches 0 as the size of the unit becomes smaller. Is the probability of two or more emails arriving in the same second equal to 0? What about the probability of two or more emails arriving in the same millisecond?

The Poisson distribution is defined by <u>one</u> parameter: λ, often called a **rate parameter**. λ specifies how many successes are observed per interval on average. We write $X \sim Pois(\lambda)$. For example,

$$X = number\ of\ emails\ per\ hour \sim Pois\,(\lambda = 5).$$

Key Idea 47

For $X \sim Pois(\lambda)$, the probability mass function is

$$f(x) = P(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}, \qquad x \in \{0, 1, 2, 3, \ldots\}.$$

It is not intuitive to derive this pmf as it has been for some of the other discrete distribution families.

Worked example: emails per hour

Let $X = number\ of\ emails\ per\ hour \sim \text{Pois}(\lambda = 5)$. What is $P(X = 1)$?

$$P(X = 1) = \frac{e^{-5}\,5^1}{1!} = e^{-5} \cdot 5 \approx 0.034.$$

The moment generating function of a Poisson random variable is

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

(This mgf was derived on Day 8 Homework.)

The expected value of a Poisson random variable is

$$E[X] = \lambda.$$

(This expected value was derived using the Poisson mgf on Day 8 Homework.)

The variance of a Poisson random variable is

$$\text{Var}(X) = \lambda.$$

4. The Discrete Uniform Distribution

Term — Discrete Uniform Distribution

A **discrete uniform random variable** is used to model a random phenomenon with N equally likely outcomes. The random variable assigns each of these outcomes to the integer values between 1 and N. Because all N outcomes are equally likely, each has a probability of $1/N$.
This distribution is defined by one parameter: N, the number of outcomes. We write $X \sim \mathit{Unif}\{N\}$. For example, $X = \textit{the result of a dice toss} \sim \mathit{Unif}\{6\}$.

Key Idea 48

For $X \sim \mathit{Unif}\{N\}$, the probability mass function is

$$f(x) = P(X = x) = \begin{cases} \dfrac{1}{N} & \text{if } x \in \{1, \ldots, N\}, \\ 0 & \text{otherwise.} \end{cases}$$

Worked example: a fair die

Let $X = \textit{the result of a dice toss} \sim \text{Unif}\{6\}$. Then

$$P(X = 1) = \frac{1}{6}$$

and

$$P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

To derive the moment generating function of the discrete uniform distribution:

$$\begin{aligned} M_X(t) &= E\left[e^{tX}\right] \\ &= \sum_{x=1}^{N} e^{tx} \cdot \frac{1}{N} \\ &= \frac{\displaystyle\sum_{x=1}^{N} e^{tx}}{N}. \end{aligned}$$

To derive the expected value of the discrete uniform distribution:

$$\begin{aligned} E[X] &= \sum_{x=1}^{N} x \cdot \frac{1}{N} \\ &= \frac{1}{N} \sum_{x=1}^{N} x \\ &= \frac{1}{N} \left(\frac{N(N+1)}{2} \right) \\ &= \frac{N+1}{2}. \end{aligned}$$

To derive the variance of the discrete uniform distribution:

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{N} x^2 \cdot \frac{1}{N} \\ &= \frac{1}{N} \sum_{x=1}^{N} x^2 \\ &= \frac{1}{N} \left(\frac{N(N+1)(2N+1)}{6} \right) \\ &= \frac{(N+1)(2N+1)}{6} \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E[X^2] - E[X]^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2} \right)^2 \\ &= \frac{2N^2+3N+1}{6} - \frac{N^2+2N+1}{4} \\ &= \frac{4N^2+6N+2}{12} - \frac{3N^2+6N+3}{12} \\ &= \frac{N^2-1}{12} \end{aligned}$$

5. Computing discrete probabilities in R

We often do not calculate these discrete probabilities by hand. Among other calculator options, R software can help:

- Hypergeometric:
  - $f(x) = P(X = x) =$ `dhyper`$(x, M, N-M, n)$
  - $F(x) = P(X \leq x) =$ `phyper`$(x, M, N-M, n)$
- Poisson:
  - $f(x) = P(X = x) =$ `dpois`(x, λ)
  - $F(x) = P(X \leq x) =$ `ppois`(x, λ)

PART 3

Practice Problems

6. Example 1 — Tagged animals (Hypergeometric)

Suppose that 4 animals near extinction have been caught, tagged, and released to mix back into the population. After these animals have had the opportunity to mix, a random sample of 10 animals is selected. If there are actually 25 animals in the population:

(a) What is the distribution of X = the number of tagged animals in the sample?

(b) What is the expected number of tagged animals in the sample?

(c) What is the probability that there are 2 tagged animals in the sample?

Solutions

(a) Distribution of X. $X \sim \textit{Hyper}(n = 10,\ M = 4,\ N = 25)$, where $n = 10$ animals are selected in the sample, $M = 4$ tagged animals serve as "successes," and $N = 25$ is the total number of animals.

(b) Expected value $E[X]$.

$$\begin{aligned} E[X] &= \frac{nM}{N} \\ &= \frac{10(4)}{25} \\ &= \frac{40}{25} \\ &= 1.6. \end{aligned}$$

(c) $P(X = 2)$. $P(X = 2)$

$$\begin{aligned} &= \frac{\binom{4}{2}\binom{21}{8}}{\binom{25}{10}} \\ &= \frac{6 \cdot 203{,}490}{3{,}268{,}760} \\ &= 0.37. \end{aligned}$$

7. Example 2 — Hospital births (Poisson)

Suppose that births occur randomly in a hospital at a rate of 1.8 per hour.

(a) If X = the number of births within an hour, then $X \sim Pois(1.8)$. What is $P(X = 1)$?

(b) If Y = the number of births within a day, then $Y \sim Pois(43.2)$. What is $P(Y = 24)$?

Solutions

(a) $P(X = 1)$**.**

$$P(X = 1) = \frac{e^{-1.8}\, 1.8^1}{1!} = e^{-1.8}\,(1.8) = 0.30.$$

(b) $P(Y = 24)$**.**

$$P(Y = 24) = \frac{e^{-43.2}\, 43.2^{24}}{24!} = 0.0005.$$

The Uniform and Normal Distributions

Lecture 18 • UVA Stats

March 2026

We turn from named discrete families to the first two named continuous families: the Uniform on $[a, b]$ and the Normal $N(\mu, \sigma^2)$. For each we derive the pdf, cdf, MGF, mean, and variance, introduce the standardised version ($U(0,1)$ and the standard normal Z), and link them via the transformation theorem. Closing examples cover a chemical-reaction temperature and standardising a normal score.

PART 1

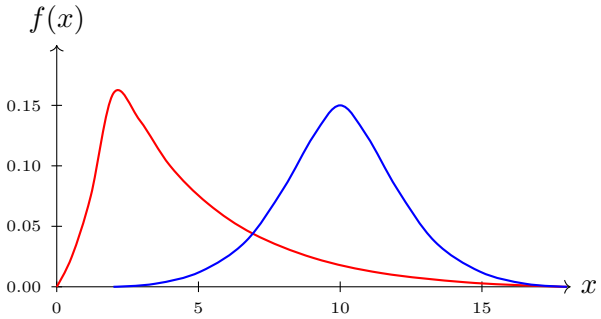
Motivation

1. From discrete families to continuous families

A **continuous random variable** is a random variable whose possible values consist of an interval or union of intervals on the number line (uncountably infinite number of outcomes). For continuous random variables, the probability distribution is described using a **probability density function (pdf)** $f(x)$, where

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

The pdf for a continuous distribution can look many different ways, varying in shape (modality, skewness, tail weight), center, and spread. There are many common distribution families that satisfy these various scenarios. The figure below illustrates two contrasting shapes: a right-skewed density (red) and a roughly symmetric, bell-shaped density (blue).



This lecture introduces two of the most important named continuous families: the **Uniform** distribution, which models equal-probability-density phenomena, and the **Normal** distribution, which models the ubiquitous bell-shaped phenomena encountered across science, engineering, and everyday life.

Key Idea 49

Continuous distributions are described by a probability density function $f(x) \geq 0$ whose total area equals 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

The shape of $f(x)$ — its modality, symmetry, tail weight, center, and spread — determines the distribution family.

PART 2

Methods and Examples

2. The Uniform Distribution

Term — Uniform Distribution

A **uniform random variable** is used to model a random phenomenon where event probabilities are proportional to the size of the event, i.e. the probability density is *constant*.

Examples:

- Suppose that the subway arrives at a station every fifteen minutes. If I arrive to the station at random, how long will I wait for my train?
- Suppose you are playing a board game with a spinner. When I flick the spinner, at what angle will it land?

The uniform distribution is defined by two parameters:

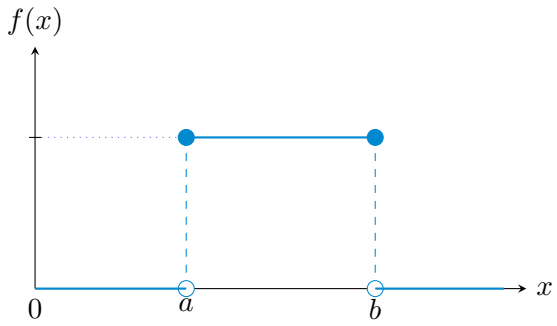
- a, the minimum possible value of the random variable
- b, the maximum possible value of the random variable

All outcomes between a and b have equal probability density; all outcomes below a or above b are impossible. We write $X \sim \text{Unif}(a, b)$.

Example: $X = \textit{wait time for train} \sim \text{Unif}(0, 15)$.

PDF Derivation

Because all outcomes between a and b have equal probability density, $f(x)$ must be constant for all $x \in [a, b]$.



To find the constant value c that $f(x)$ takes, we use the fact that any pdf must integrate to 1:

$$\int_a^b f(x)\,dx = 1$$

$$\int_a^b c\,dx = 1$$

$$\left[cx\right]_{x=a}^{x=b} = 1$$

$$bc - ac = 1$$

$$c(b-a) = 1$$

$$c = \frac{1}{b-a}$$

Key Idea 50

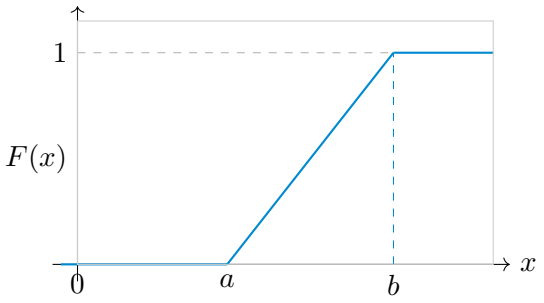
The pdf of $X \sim \text{Unif}(a, b)$ is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

CDF Derivation

The cumulative distribution function is obtained by integrating the pdf from a up to x:

$$\begin{aligned} F(x) &= \int_{\min}^{x} f(u)\,du \\ &= \int_a^x \frac{1}{b-a}\,du \\ &= \left[\frac{u}{b-a}\right]_{u=a}^{u=x} \\ &= \frac{x-a}{b-a} \end{aligned}$$



Moment Generating Function

The moment generating function of a uniform random variable is:

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

Expected Value

$$\begin{aligned}
E[X] &= \int_{\text{support}} x \cdot f(x)\, dx \\
&= \int_a^b x \cdot \frac{1}{b-a}\, dx \\
&= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_{x=a}^{x=b} \\
&= \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) \\
&= \frac{1}{b-a}\left(\frac{(b+a)(b-a)}{2}\right) \\
&= \frac{b+a}{2}
\end{aligned}$$

Variance

The variance of a uniform random variable is:

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Standard Uniform

We often refer to the $\text{Unif}(0,1)$ distribution as a **standard uniform random variable**. If $U \sim \text{Unif}(0,1)$, then

$$f(u) = \begin{cases} 1 & \text{if } 0 \le u \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Standardising a Uniform Random Variable

Term — Uniform Distribution

If $X = a + (b-a)U$, then $X \sim \text{Unif}(a,b)$.

Proof: Transformation Theorem!

1. If $X = a + (b-a)U$, then $U = g^{-1}(X) = \dfrac{x-a}{b-a}$.

2. $f_X(x) = f_U\big(g^{-1}(x)\big)\left|\dfrac{d}{dx}\,g^{-1}(x)\right| = 1 \cdot \left|\dfrac{d}{dx}\left(\dfrac{x-a}{b-a}\right)\right| = 1 \cdot \left|\dfrac{1}{b-a}\right| = \dfrac{1}{b-a}.$

3. $0 \le u \le 1 \;\Rightarrow\; 0 \le \dfrac{x-a}{b-a} \le 1 \;\Rightarrow\; 0 \le x-a \le b-a \;\Rightarrow\; a \le x \le b$

Key Idea 51

If $X = a + (b-a)U$, then $X \sim \text{Unif}(a, b)$.

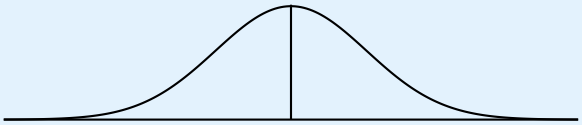
This means that we can standardise and un-standardise any uniform random variable:

- $X = a + (b-a)U$
- $U = \dfrac{X-a}{b-a}$.

3. The Normal Distribution

Term — Normal Distribution

The **Normal distribution** is a unimodal, symmetric, and light-tailed probability density function used to model random phenomena with a bell-shaped distribution.



Real-world examples:

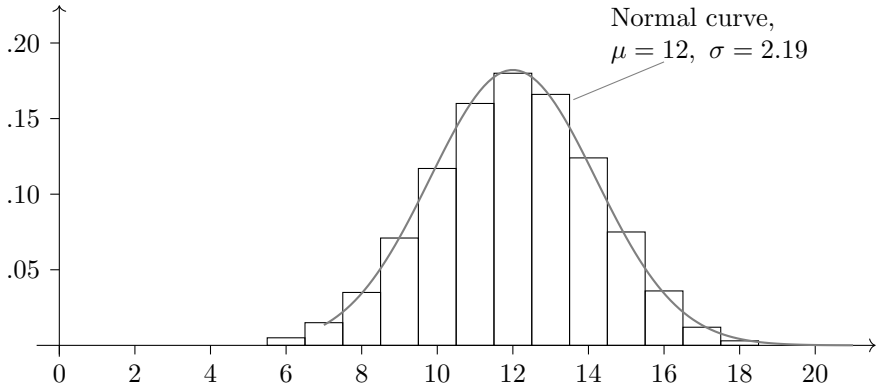
- Standardised test scores are often formulated to have a Normal distribution, and IQ scores are a particularly good example.
- Adult heights are known to have a relatively Normal distribution.

Why do statisticians love the Normal distribution?

- "All distributions look roughly Normal at the middle."
- Under certain conditions, many distributions will start to look even more Normal!
  - Example: For large values of the parameter n, the binomial distribution will be approximately Normal.
- Many statistics that we calculate from sample data, like the sample mean or sample variance, have a Normal distribution when the sample size is large.

Worked Example: Binomial vs. Normal

The figure below overlays the Binom($n = 20$, $p = 0.6$) probability mass function (bars) with the $N(\mu = 12,\, \sigma^2 = 4.8)$ density curve, illustrating how closely the two distributions agree for large n.



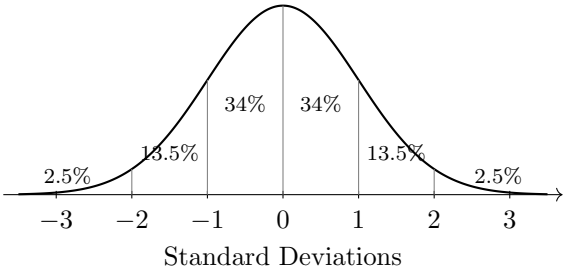
Parameters and Notation

Any Normal distribution will always have the same shape: unimodal, symmetric, and light-tailed. The two parameters of the Normal distribution specify the center and spread of the distribution:

- μ is referred to as the mean of the Normal distribution
- σ^2 is referred to as the variance of the Normal distribution
- We write $X \sim N(\mu, \sigma^2)$.

The parameters μ and σ^2 specify the center and spread more precisely:

- μ is the mean, median, and mode of the distribution.
- σ^2 specifies how far the distribution spreads from μ:
  - 68% of the distribution is between $\mu - \sigma$ and $\mu + \sigma$
  - 95% of the distribution is between $\mu - 2\sigma$ and $\mu + 2\sigma$
  - 99.7% of the distribution is between $\mu - 3\sigma$ and $\mu + 3\sigma$



PDF Formula

Key Idea 52

The probability density function of a $N(\mu, \sigma^2)$ random variable is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in (-\infty, \infty).$$

The Normal distribution always has infinite support.

Moment Generating Function

Key Idea 53

The moment generating function of a $N(\mu, \sigma^2)$ random variable is

$$M_X(t) = e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}$$

MGF-Based Derivation of $E[X] = \mu$

We can use the moment generating function to show that $E[X] = \mu$:

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)} \\
&= e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)} \cdot \left(\mu + \frac{\sigma^2}{2}(2t)\right) \\
&= e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)} \cdot (\mu + \sigma^2 t)
\end{aligned}$$

$$\begin{aligned}
E[X] &= \frac{d}{dt} M_X(t)\Big|_{t=0} \\
&= e^{\left(\mu(0) + \frac{\sigma^2 \cdot 0^2}{2}\right)} \cdot (\mu + \sigma^2(0)) \\
&= e^0 \cdot (\mu + 0) \\
&= \mu
\end{aligned}$$

Standard Normal

Term — Standard Normal Random Variable

We often refer to the $N(0, 1)$ distribution as a **standard Normal random variable**.

- The mean is 0, and the variance is 1.

If $Z \sim N(0, 1)$, then

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Standardising a Normal Random Variable

Term — Normal Distribution

If $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$.

Proof: Transformation Theorem!

1. If $X = \mu + \sigma Z$, then $Z = g^{-1}(X) = \dfrac{X - \mu}{\sigma}$.
2. $$f_X(x) = f_Z\big(g^{-1}(x)\big)\left|\frac{d}{dx}\,g^{-1}(x)\right| = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\cdot\left|\frac{d}{dx}\left(\frac{x-\mu}{\sigma}\right)\right|$$
$$= \frac{1}{\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot\left|\frac{1}{\sigma}\right| = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Key Idea 54

If $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$.

This means that we can standardise and un-standardise any Normal random variable:

- $X = \mu + \sigma Z$
- $Z = \dfrac{X - \mu}{\sigma}$ $\quad\rightarrow\quad$ often referred to as a **z-score**

4. Computing Continuous Probabilities in R

We often don't calculate these probabilities by hand. Among other calculator options, R software can help:

- **Uniform:**
  - $f(x) = \texttt{dunif}(x, a, b)$
  - $F(x) = P(X \le x) = \texttt{punif}(x, a, b)$
- **Normal:**
  - $f(x) = \texttt{dnorm}(x, \mu, \sigma)$ $\quad\rightarrow\quad$ make sure to input σ and not σ^2!
  - $F(x) = P(X \le x) = \texttt{pnorm}(x, \mu, \sigma)$ $\quad\rightarrow\quad$ make sure to input σ and not σ^2!

PART 3

Practice Problems

5. Example 1 — Chemical Reaction Temperature (Uniform)

Suppose that the reaction temperature for a chemical process is uniformly distributed between $-5°C$ and $5°C$. Let X denote the reaction temperature.

(a) What is the pdf for X?

(b) What is the expected value for X?

(c) What is $P(X < 0)$?

(d) What is $P(-2 < X < 2)$?

Solutions

(a) The uniform pdf is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

In our example, $a = -5$ and $b = 5$. So

$$\frac{1}{b-a} = \frac{1}{5-(-5)} = \frac{1}{10} = 0.1.$$

Thus,

$$f(x) = \begin{cases} 0.1 & \text{if } -5 \le x \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

(b)

$$E[X] = \frac{b+a}{2} = \frac{5+(-5)}{2} = \frac{0}{2} = 0$$

(c) The uniform cdf is

$$F(x) = \frac{x-a}{b-a}.$$

For our example, this is

$$F(x) = \frac{x-a}{b-a} = \frac{x-(-5)}{5-(-5)} = \frac{x+5}{10}.$$

Therefore,

$$P(X < 0) = F(0) = \frac{0+5}{10} = \frac{5}{10} = 0.5.$$

(d)

$$P(-2 < X < 2) = P(X < 2) - P(X < -2) = F(2) - F(-2) = \frac{2+5}{10} - \frac{-2+5}{10} = \frac{7}{10} - \frac{3}{10} = \frac{4}{10} = 0.4.$$

6. Example 2 — Standardising a Normal Variable

The heights of adult men in the United States are approximately Normal with a mean of 70 inches and a variance of 9.

(a) UVA guard Sam Lewis is 79 inches tall. What is the standardised value of his height?

(b) The 90^{th} percentile of the standard Normal distribution is 1.28. What is the 90^{th} percentile of adult male heights?

Solutions

(a)

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{79-70}{\sqrt{9}} \\ &= \frac{9}{3} \\ &= 3 \end{aligned}$$

Sam's height is 3 standard deviations above average (i.e. *much* taller than average).

(b) Solving for the raw value x:

$$\begin{aligned} x &= \mu + \sigma z \\ &= 70 + \sqrt{9}\,(1.28) \\ &= 70 + 3\,(1.28) \\ &= 70 + 3.84 \\ &= 73.84 \end{aligned}$$

90% of adult men are 73.84 inches or shorter.

The Exponential, Gamma, and Beta Distributions

Lecture 19 • UVA Stats

March 2026

Three more named continuous families: the Exponential (waiting time between Poisson events), the Gamma (a generalisation with shape parameter α), and the Beta (a flexible distribution on $[0, 1]$ for proportions). For each we derive the pdf, MGF, mean, and variance, and we link the Exponential to the Poisson process through the memoryless property. We close with a Poisson/Exponential fish-arrival practice example.

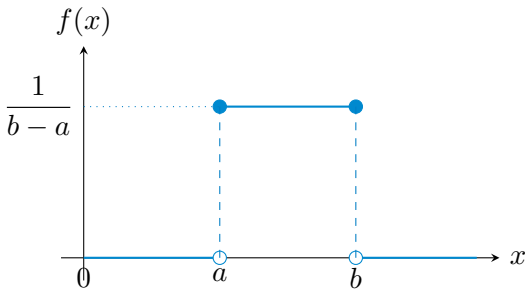
PART 1

Motivation

1. Three more continuous families

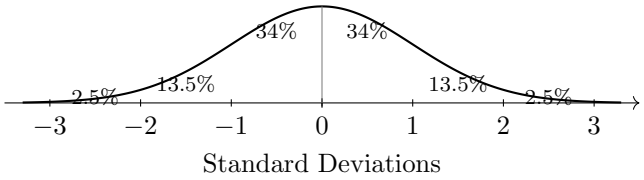
A **continuous random variable** is a random variable whose possible values consist of an interval or union of intervals on the number line (uncountably infinite number of outcomes). The pdf for a continuous distribution can take many shapes—differing in modality, skewness, tail weight, center, and spread—and there are many common distribution families that describe these various scenarios.

In the previous lecture we met two such families. A **uniform random variable** is used to model a random phenomenon where event probabilities are proportional to the size of the event:



Term — Normal Distribution

The **Normal distribution** is a unimodal, symmetric, and light-tailed probability density function used to model random phenomena with bell-shaped distributions.



This lecture introduces three new continuous families: the **Exponential**, **Gamma**, and **Beta** distributions.

Key Idea 55

Each new family extends or generalises a family the reader already knows: the Exponential distribution is the continuous counterpart of the Poisson (it models waiting times in a Poisson process); the Gamma distribution generalises the Exponential by adding a shape parameter α;

and the Beta distribution generalises the Uniform on $[0, 1]$ by adding two shape parameters that allow the distribution to take a wide variety of forms.

PART 2

Methods and Examples

2. The Exponential Distribution

Term — Exponential Distribution

The **exponential distribution** is used to model the time until an event occurs in the Poisson setting.

Reminder: a Poisson random variable counts the number of events that occur within a fixed interval of time. These events occur independently and at a constant average rate λ. Suppose that these events occur at times $T_1, T_2, T_3, \ldots$; then the inter-arrival times T_1, $T_2 - T_1$, $T_3 - T_2, \ldots$ each have an exponential distribution.

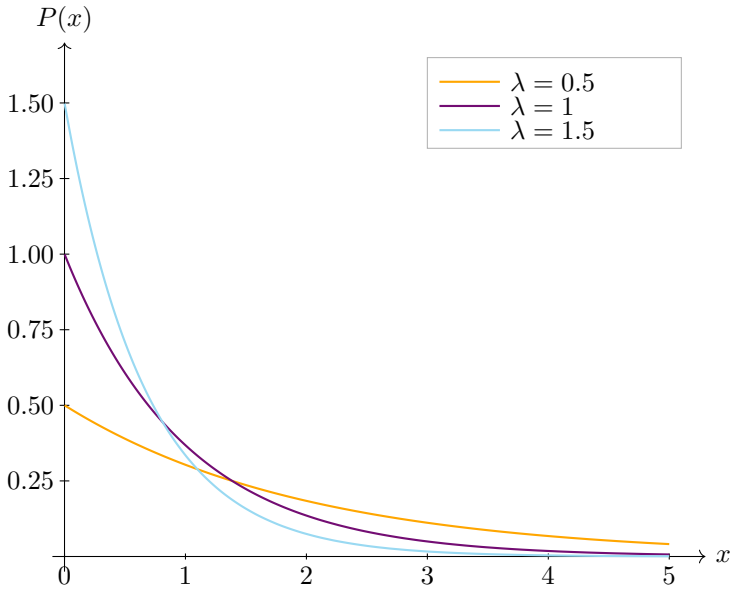
Because of this relationship between exponential and Poisson random variables, the exponential distribution is defined by the same parameter λ, which specifies how many events are observed per interval on average. We write $X \sim \text{Expo}(\lambda)$.

Example: if $X =$ number of emails per hour $\sim \text{Pois}(\lambda = 5)$, then $Y =$ time between emails $\sim \text{Expo}(\lambda = 5)$.

Key Idea 56

The probability density function for an exponential random variable is

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



We can derive the exponential cdf from this pdf:

$$\begin{aligned} F(x) &= \int_{\min}^{x} f(u)\,du \\ &= \int_{0}^{x} \lambda e^{-\lambda u}\,du \\ &= \left[-e^{-\lambda u}\right]_{u=0}^{u=x} \\ &= -e^{-\lambda x} - \left(-e^{-\lambda(0)}\right) \\ &= -e^{-\lambda x} - (-1) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Using the exponential cdf, we can show the relationship between exponential and Poisson random variables. Suppose that $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Expo}(\lambda)$. If no events occur within an interval of time, then $X = 0$ and $Y > 1$.

- $P(X = 0) = f_X(0) = \dfrac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$
- $P(Y > 1) = 1 - P(Y \leq 1) = 1 - F_Y(1) = 1 - \left(1 - e^{-\lambda(1)}\right) = e^{-\lambda}$

The two probabilities are equal, confirming the intimate connection between the Poisson count and the exponential waiting time.

Memoryless property

Exponential random variables have a special property called the **memoryless property**. Suppose that the time between events is exponentially distributed with rate parameter λ, and the event has still not occurred by time t. What is the probability that the event will still not occur for an additional length of time s?

$$
\begin{aligned}
&P(X > s+t \mid X > t) \\
&= \frac{P(X > s+t \cap X > t)}{P(X > t)} \\
&= \frac{P(X > s+t)}{P(X > t)} \\
&= \frac{1 - F(s+t)}{1 - F(t)} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\
&= e^{-\lambda s}
\end{aligned}
$$

$= P(X > s)$ ⟶ The conditional probability is the same as the original, so it is "memoryless."

The conditional probability that the event has not yet occurred, given that it has not occurred up to time t, equals the unconditional probability that it does not occur within a further interval s. The past waiting time carries no information about the future.

In the Day 12 lecture we already derived the exponential moment generating function:

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

and used it to obtain the mean and variance:

$$E[X] = \frac{1}{\lambda} \qquad \text{Var}(X) = \frac{1}{\lambda^2}.$$

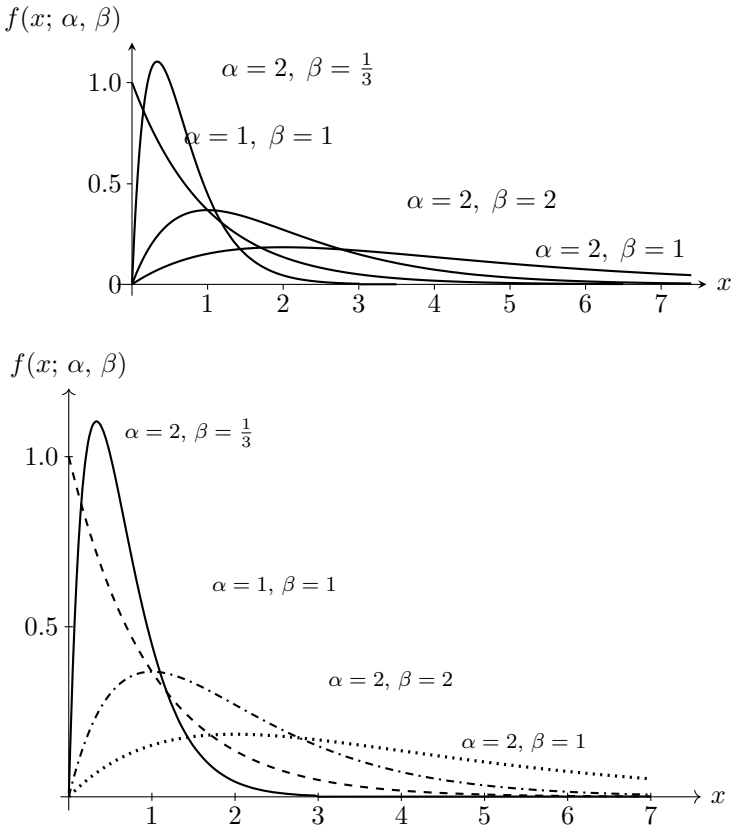
3. The Gamma Distribution

Term — Gamma Distribution

The **Gamma distribution** is a flexible, unimodal, and skewed-right distribution with a support of $x \geq 0$. It is often used to model time periods like exponential random variables, though with a more flexible shape; it also commonly models the size of events such as rainfall totals or insurance claims.

The Gamma distribution is specified by two parameters:

- A shape parameter α (controls the skewness of the distribution).
- A scale parameter β (controls the spread of the distribution).



Key Idea 57

The probability density function for a Gamma random variable is

$$f(x) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\,x^{\alpha-1}e^{-\frac{x}{\beta}} \quad \text{for } x \geq 0,$$

where $\Gamma(\alpha)$ is the **Gamma function**:

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

It is included in the pdf so that the area under the curve equals 1.

The moment generating function for a Gamma random variable is

$$M_X(t) = (1-\beta t)^{-\alpha}.$$

We can use the Gamma mgf to derive its expected value:

$$\begin{aligned}&\frac{d}{dt}M_X(t)\\&=\frac{d}{dt}\left[(1-\beta t)^{-\alpha}\right]\\&=-\alpha(1-\beta t)^{-\alpha-1}*(-\beta)\\&=\alpha\beta(1-\beta t)^{-\alpha-1}\end{aligned}$$

$$\begin{aligned}&E[X]\\&=\frac{d}{dt}M_X(0)\\&=\alpha\beta\big(1-\beta(0)\big)^{-\alpha-1}\\&=\alpha\beta(1)^{-\alpha-1}\\&=\alpha\beta\end{aligned}$$

The variance of a Gamma random variable is

$$\text{Var}(X)=\alpha\beta^2.$$

4. The Beta Distribution

Term — Beta Distribution

The **beta distribution** is a flexible distribution with a support of $[0,1]$. Because of this support, it is often used to model proportions and percentages.

The beta distribution is specified by two shape parameters:

- A shape parameter α.
- A shape parameter β.

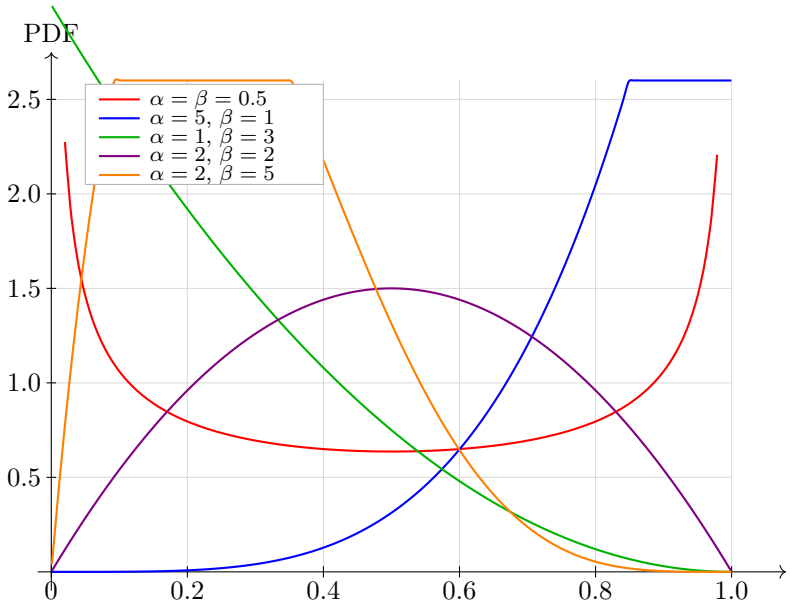
Note: the symbols α and β here are *not* the same as the Gamma parameters. The parameters jointly determine the shape:

- $\alpha=\beta$? Symmetric
- $\alpha>\beta$? Skewed left
- $\alpha<\beta$? Skewed right
- As α and β get larger, the spread of the distribution decreases.

Worked example: Beta PDF shapes

Key Idea 58

The **Beta distribution** is a continuous distribution on $[0,1]$ with shape parameters $\alpha>0$ and $\beta>0$. Its PDF changes dramatically depending on the parameter values.



Key Idea 59

The Beta pdf, cdf, and mgf do not have simple closed forms, so we work directly with the mean and variance formulas. The expected value of a Beta random variable is

$$E[X] = \frac{\alpha}{\alpha + \beta},$$

and the variance is

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

5. Computing continuous probabilities in R

We often do not calculate continuous probabilities by hand. Among other options, R software provides built-in functions for each family:

- **Exponential:**
  - $f(x) =$ `dexp`$(x,\ \lambda)$
  - $F(x) = P(X \leq x) =$ `pexp`$(x,\ \lambda)$
- **Gamma:**
  - $f(x) =$ `dgamma`$(x,\ \alpha,\ 1/\beta)$ $\rightarrow$ make sure to input $1/\beta$ and not β!
  - $F(x) = P(X \leq x) =$ `pgamma`$(x,\ \alpha,\ 1/\beta)$ $\rightarrow$ make sure to input $1/\beta$ and not β!
- **Beta:**
  - $f(x) =$ `dbeta`$(x,\ \alpha,\ \beta)$
  - $F(x) = P(X \leq x) =$ `pbeta`$(x,\ \alpha,\ \beta)$

PART 3

Practice Problems

6. Example 1 — Fish arrivals at a river observation point

Suppose that X measures the amount of time (in minutes) between fish swimming by a particular location in a river and Y measures the total number of fish that swim by in a minute. Suppose that fish swim independently, and the mean rate of the process is that 1 fish passes every 5 minutes.

(a) What is the distribution of Y?

(b) What is the distribution of X?

(c) What is the probability that a fish will swim by in the next minute, i.e. $P(X < 1)$?

(d) If we do not see a fish in the next five minutes, what is the probability that a fish will not swim by in the following minute, i.e. $P(X > 6 \mid X > 5)$?

Solutions

(a) Y counts the number of events (fish) in a fixed interval (one minute), so Y will be a Poisson random variable.

- The rate is 1 fish per 5 minutes, which is 0.2 fish per minute.
- So $Y \sim \text{Pois}(\lambda = 0.2)$.

(b) X measures the time between events (fish) in a Poisson process, so X will be an exponential random variable.

- The rate is 1 fish per 5 minutes, which is 0.2 fish per minute.
- So $X \sim \text{Expo}(\lambda = 0.2)$.

(c) Because $X \sim \text{Expo}(\lambda = 0.2)$, we know that $F(x) = 1 - e^{-0.2x}$.

$$\begin{aligned} F(1) & \\ &= 1 - e^{-0.2(1)} \\ &= 1 - e^{-0.2} \\ &= 1 - 0.82 \\ &= 0.18. \end{aligned}$$

(d)

$$P(X > 6 \mid X > 5) = P(X > (1+5) \mid X > 5) = P(X > 1)$$
$$s+t \qquad\qquad t \qquad\qquad s$$

This is true from the memoryless property of exponential random variables.

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - 0.18 = 0.82.$$

Probability Plots: Poisson-ness and QQ Plots

Lecture 20 • UVA Stats

April 2026

Once we have catalogued the named distributions, the next question is whether a given dataset actually fits one. This lecture introduces two diagnostic tools: the Poisson-ness plot for discrete count data and the quantile-quantile (QQ) plot for continuous data. We derive the linear form each plot should take when the model fits, interpret systematic deviations, and apply the diagnostics to a London-blitz bomb-strike dataset and an exponential / normal / uniform comparison.

PART 1

Motivation

1. From distributions to goodness-of-fit

Over the past several lectures we have built up a catalogue of named distribution families.

Discrete:	Continuous:
Bernoulli	Uniform
Binomial	Normal
Geometric	Exponential
Negative binomial	Gamma
Hypergeometric	Beta
Poisson	
Discrete uniform	

We have also discussed many random phenomena that are often modelled by these families. In practice, however, when we are handed a dataset we face a harder question: is a chosen distribution family actually a *good fit*? A variable's distribution might look unimodal and symmetric, but is it unimodal and symmetric *enough* to be modelled by the Normal family? Simply eyeballing a histogram is rarely decisive.

The answer is **probability plots**: graphical tools that make the goodness-of-fit question precise and visual. Different probability-plot designs exist for different families; we study two in this lecture—the *Poisson-ness plot* for discrete count data, and the *quantile-quantile (QQ) plot* for continuous data.

Key Idea 60

A probability plot transforms the candidate-fit question into a "is this scatter linear?" question. Linear patterns are something the human eye is extremely good at detecting, so the plot provides an immediate and reliable visual diagnostic: a roughly straight point cloud supports the assumed family, while systematic curvature or bowing rejects it.

PART 2

Methods and Examples

2. The Poisson-ness Plot

Term — Poisson-ness Plot

For most discrete random variables the setting itself tells us which family to use (e.g. Binomial: number of successes in n independent trials; Geometric: number of trials until the first success). The Poisson setting is harder to verify: when we count events within an interval, are they truly occurring independently at a constant average rate?
In a sample of N observations, define n_k as the number of observations equal to k. If $X \sim \text{Pois}(\lambda)$ then we expect

$$n_k \approx N \cdot \frac{e^{-\lambda}\lambda^k}{k!}.$$

The **Poisson-ness plot** is a graphical check: if all $n_k \approx N \cdot \frac{e^{-\lambda}\lambda^k}{k!}$, the Poisson model is valid. The plot turns this comparison into a linearity check (see derivation below).

Derivation of the Poisson-ness plot

Starting from the Poisson probability mass function,

$$P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!},$$

in a sample of size N we expect $N \cdot P[X = k]$ observations with value k. Setting $n_k \approx N \cdot \frac{e^{-\lambda}\lambda^k}{k!}$ and taking the natural logarithm of both sides:

$$\ln n_k \approx \ln\left(N \cdot \frac{e^{-\lambda}\lambda^k}{k!}\right)$$

$$\ln n_k \approx \ln N + \ln e^{-\lambda} + \ln \lambda^k - \ln k!$$

$$\ln n_k \approx \ln N - \lambda + k \ln \lambda - \ln k!$$

Rearranging so that the data-derived quantity sits on the left:

$$\ln n_k - \ln N + \ln k! \approx -\lambda + k \ln \lambda.$$

Key Idea 61

Define the *log result* for outcome k as

$$\ln n_k - \ln N + \ln k!.$$

If the data follow a Poisson distribution then, when we plot this quantity against k, the points should fall on a straight line with

- **slope** $= \ln \lambda$, and
- **intercept** $= -\lambda$.

Non-linearity in the plot is evidence against the Poisson model.

Interpreting the linearity check

The relationship

$$\ln n_k - \ln N + \ln k! \;\approx\; -\lambda + k \ln \lambda$$

tells us precisely how to read off λ from the fitted line:

- Plot $(\ln n_k - \ln N + \ln k!)$ on the vertical axis versus k on the horizontal axis.
  - If there is a *linear* relationship, the Poisson distribution is a good model for the data.
    - The intercept of the line equals $-\lambda$.
    - The slope of the line equals $\ln \lambda$.
  - If the plot is *non-linear*, the Poisson distribution is not a good fit.

Note: $\ln n_k$ is undefined when $n_k = 0$, so we plot only the values of k that are actually observed in the data.

Worked example: London-blitz bomb strikes

The city of London is 6 square miles in area, which can be divided into 576 $\frac{1}{4}$-by-$\frac{1}{4}$ mile square blocks. During World War II, the number of bombs to land in each square block was recorded.

#bomb hits (k)	0	1	2	3	4	5	6	7
#blocks (n_k)	229	211	93	35	7	0	0	1

We want to plot $(\ln n_k - \ln N + \ln k!)$ versus k, where $N = 576$.

Since $n_5 = n_6 = 0$, the log result is undefined for $k = 5$ and $k = 6$, so those points are excluded. The single observation at $k = 7$ appears to be an extreme outlier and is also excluded. The plotted points are computed as follows.

$k = 0$**:**

$$\ln n_0 - \ln N + \ln 0! \;=\; \ln 229 - \ln 576 + \ln 1 \;=\; 5.43 - 6.36 + 0 \;=\; -0.93.$$

Add the point $(0, -0.93)$ to the plot.

$k = 1$**:**

$$\ln 211 - \ln 576 + \ln 1! \;=\; 5.35 - 6.36 + 0 \;=\; -1.01 \;\approx\; -1.005.$$

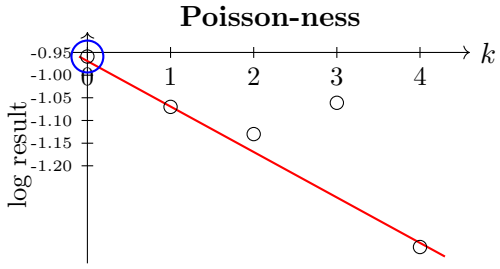
Add the point $(1, -1.005)$.

$k = 2$**:**

$$\ln 93 - \ln 576 + \ln 2! \;=\; 4.53 - 6.36 + 0.69 \;=\; -1.13.$$

Add the point $(2, -1.13)$.

Continuing for $k = 3$ and $k = 4$ in the same way gives the full set of five plotted points. The resulting Poisson-ness plot is shown below.

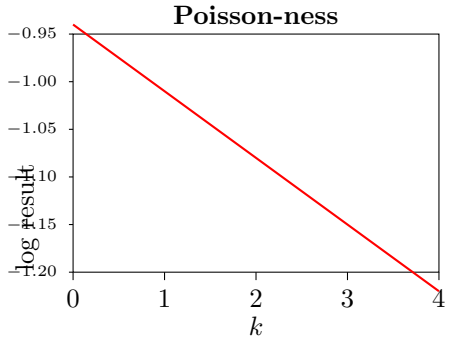


Reading off the fitted line:

Slope: -0.07 **Intercept:** -0.94

- Slope estimate for λ:
  - $\ln \lambda = -0.07 \rightarrow \lambda = 0.93$
- Intercept estimate for λ:
  - $-\lambda = -0.94 \rightarrow \lambda = 0.94$

Both methods give $\hat{\lambda} \approx 0.93$–$0.94$. The points track the line reasonably well (noting the $k = 0$ point marked in blue as a slight outlier), so the Poisson model is a plausible fit.



3. The Quantile-Quantile (QQ) Plot

Term — QQ Plot

For a discrete random variable like the Poisson, a small number of outcomes are each expected to appear many times, so we compare observed versus expected *frequencies*. For a *continuous* random variable we do not expect to see duplicates—there are uncountably many possible outcomes—so frequency comparison is not feasible.

Instead, we compare *percentiles*. If my data follow a Normal distribution then the 10^{th} percentile of my sample should roughly equal the 10^{th} percentile of a Normal distribution, the 40^{th} percentile of my sample should roughly equal the 40^{th} percentile of a Normal distribution, and so on.

A **quantile-quantile plot** (or **QQ plot**) graphs the sample data (sample percentiles) versus the corresponding percentile values of the plausible distribution family (theoretical percentiles).

If the data come from the assumed distribution, the points will fall approximately on a straight line.

Identifying sample percentiles

To construct a QQ plot we first need to assign a percentile rank to each data point.

1. Sort the n sample data points from smallest to largest; call the i^{th} smallest value $X_{(i)}$.
2. The i^{th} smallest observation is the $\left(\dfrac{i-0.5}{n}\right)^{\text{th}}$ sample percentile.

Worked example: percentile ranks

Example: Suppose I have a dataset with $n = 10$ observations.

- For the smallest observation ($i = 1$): $\dfrac{i-0.5}{n} = \dfrac{1-0.5}{10} = \dfrac{0.5}{10} = 0.05$.
  This is the sample estimate of the 5^{th} percentile.
- For the second smallest observation ($i = 2$): $\dfrac{i-0.5}{n} = \dfrac{2-0.5}{10} = \dfrac{1.5}{10} = 0.15$.
  This is the sample estimate of the 15^{th} percentile.

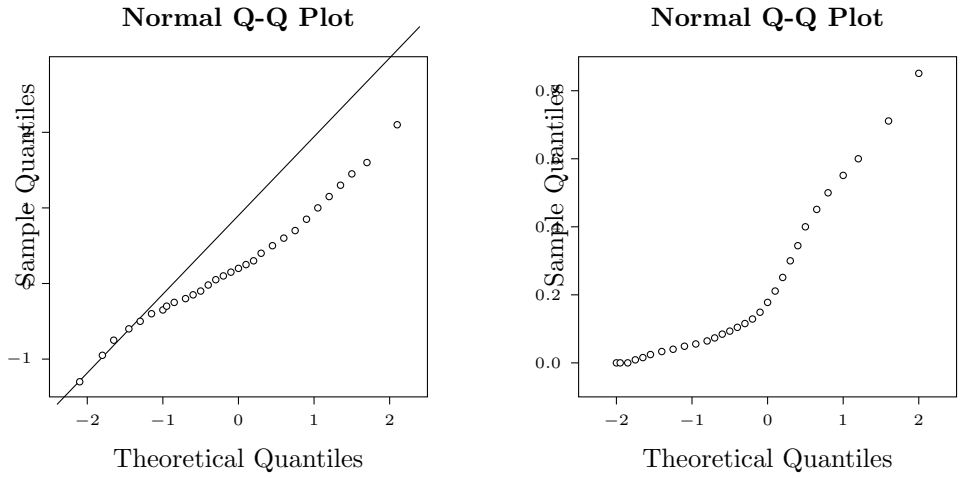
Once each $X_{(i)}$ has been assigned percentile rank $\frac{i-0.5}{n}$, we find the corresponding theoretical percentile from the candidate distribution and plot the pair. For example, if the candidate is Normal, the first plot point has x-coordinate equal to the 5^{th} percentile of a Normal distribution and y-coordinate $X_{(1)}$.

Key Idea 62

If the points in a QQ plot fall in a straight line, then the presumed distribution is valid. If the points deviate from a linear trend, then the presumed distribution is not a good fit.

Worked example: Normal vs. non-Normal

If the data come from the assumed distribution the points will fall approximately on a straight line.



- **Left plot**: Points track closely along the reference line ⇒ data are approximately Normal.
- **Right plot**: Points curve away from the line (concave upward) ⇒ data are right-skewed (heavy right tail); normality is not supported.

Scale invariance and the Standard Normal

When constructing a Normal QQ plot, one might ask: which Normal distribution should supply the theoretical percentiles? The 10^{th} percentile of an $N(1,4)$ distribution is -1.56, while the 10^{th} percentile of the $N(6,1)$ distribution is 4.72—so the choice seems to matter. The answer, however, is that it does *not* matter.

Any $N(\mu, \sigma^2)$ random variable is a linear transformation of the standard Normal:

$$X = \mu + \sigma Z.$$

This means that the p^{th} percentile of the $N(\mu, \sigma^2)$ distribution is

$$X_p = \mu + \sigma Z_p,$$

where Z_p is the p^{th} percentile of $N(0,1)$. The Normal distribution shape is always the same; the parameters only change the center and spread.

Key Idea 63

Changing μ or σ^2 will only change the scale of the x-axis in the QQ plot, but the *pattern* in the plot (linear or not) will not change. Therefore, the theoretical percentiles in the Normal QQ plot can always be calculated from the standard Normal distribution for simplicity.

This scale-invariance property holds for any family with a location/scale parameterisation:

- If $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0,1)$, then $X = \mu + \sigma Z$.
- If $X \sim \text{Unif}(a,b)$ and $U \sim \text{Unif}(0,1)$, then $X = a + (b-a)U$.
- If $X \sim \text{Expo}(\lambda)$ and $Y \sim \text{Expo}(1)$, then $X = \dfrac{1}{\lambda} Y$.

QQ plots for flexible distributions

For distributions with more flexible shapes—such as the Gamma and Beta families—it is not as straightforward to generate a QQ plot. The parameters must be estimated first in order to compute the theoretical percentiles. We will not focus on these cases in this course.

PART 3

Practice Problems

4. Example 1 — Exponential model for graduate-school waiting times

The data table below lists the number of statistics majors graduating from UVA each year 2007–2021.

2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
1	8	7	5	7	27	23	41
2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	
60	93	86	98	149	203	248	

Can the number of graduates per year be modelled using an exponential distribution? There are $n = 15$ observations. Answer the following questions.

a) 248 majors graduated in 2021, which is the highest observed value in the 15-year dataset. 248 is which sample percentile?

b) What is the corresponding percentile of a standard exponential distribution Expo($\lambda = 1$)?

c) Looking at the QQ plot, is the exponential distribution a good fit?

Solutions

(a) For the largest observation ($i = 15$, $n = 15$):

$$\frac{i - 0.5}{n} = \frac{15 - 0.5}{15} = \frac{14.5}{15} = 0.967.$$

Therefore $x = 248$ is the sample 96.7^{th} percentile.

(b) When $\lambda = 1$ the exponential CDF is $F(x) = 1 - e^{-x}$. We solve for the 96.7^{th} percentile of Expo(1):

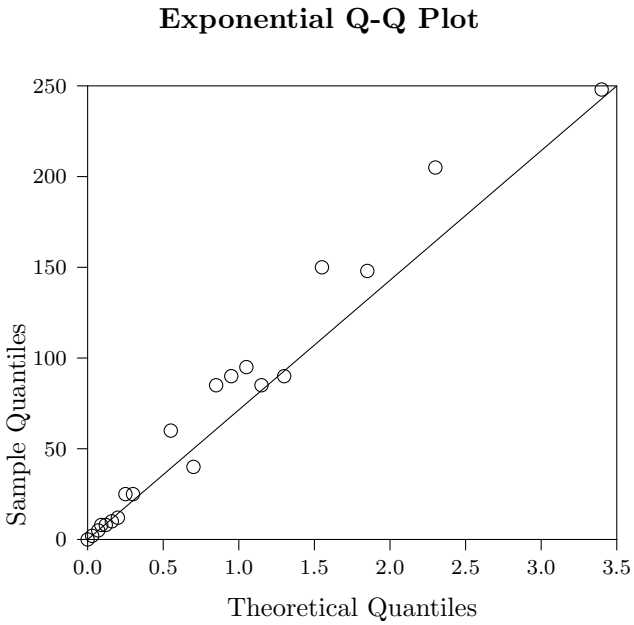
$$\begin{aligned} 1 - e^{-x} &= 0.967 \\ -e^{-x} &= -0.033 \\ e^{-x} &= 0.033 \\ -x &= \ln 0.033 \\ -x &= -3.40 \\ x &= 3.40. \end{aligned}$$

This is the 96.7^{th} percentile of the standard exponential distribution.

The final point in the QQ plot therefore has:

- x-coordinate 3.40 (theoretical 96.7^{th} percentile), and
- y-coordinate 248 (observed 96.7^{th} percentile).

(c) The full QQ plot, with all 15 points computed analogously, is shown below.



The points lie reasonably close to the reference line, providing moderate support for an exponential model for the annual graduate counts.

5. Example 2 — Identifying the right family from QQ plots

For each of the following plots, assess whether the suggested distribution is a good fit.

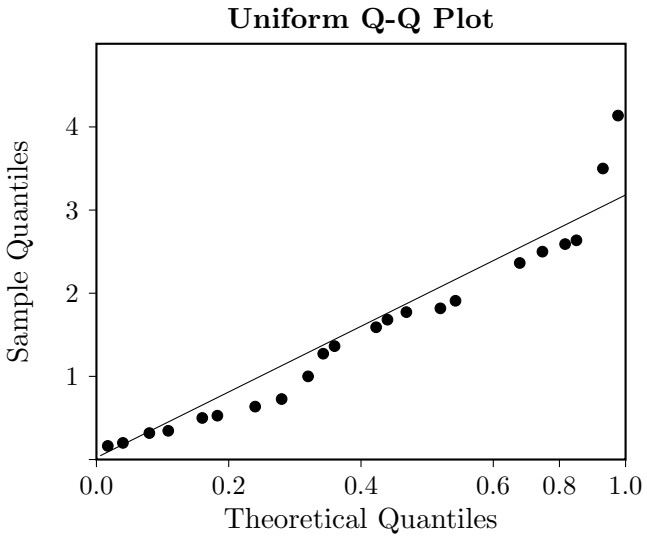
a) Uniform QQ plot

b) Normal QQ plot

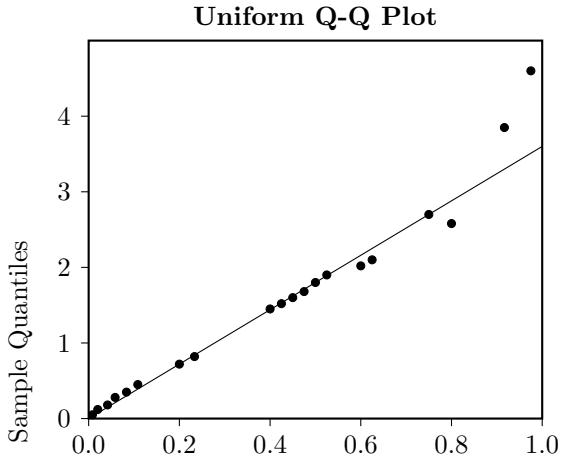
c) Poisson-ness plot

Solutions

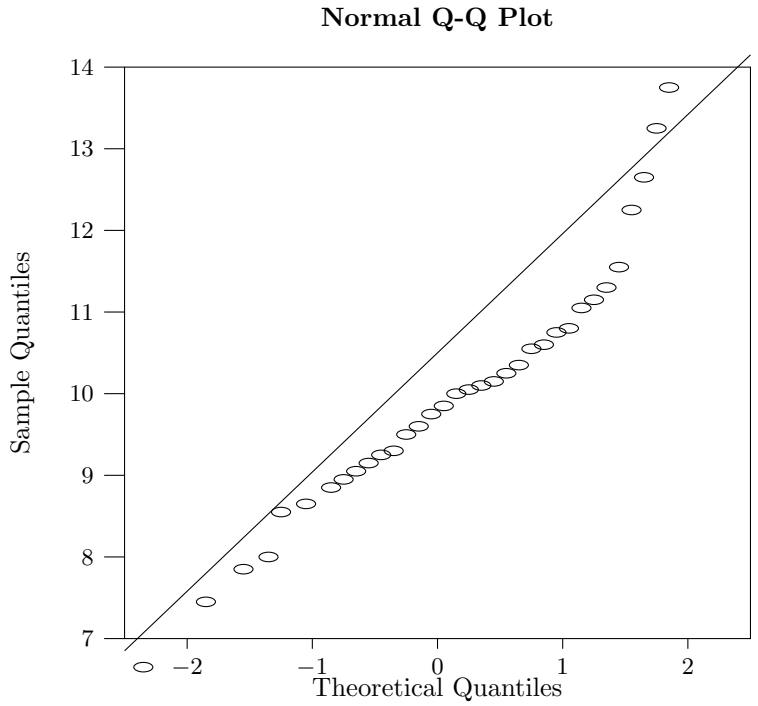
(a) Does the plot below suggest that the data could come from a uniform distribution?



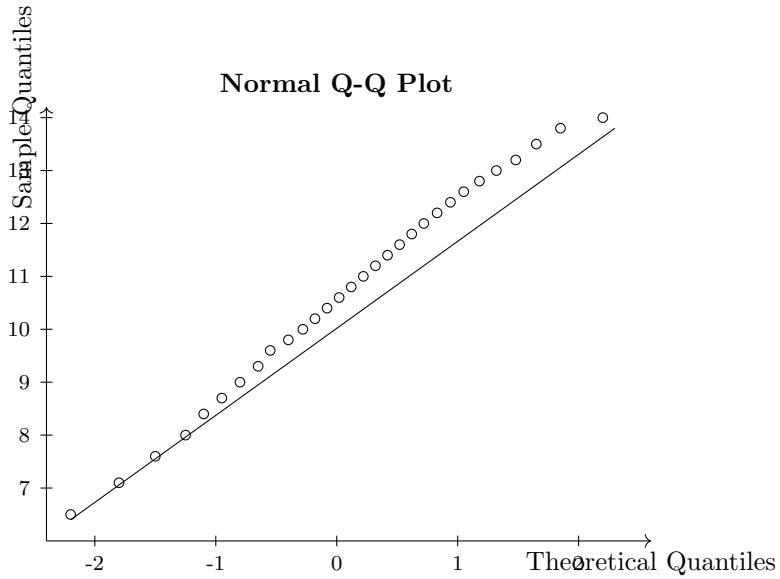
As it turns out, this is 20 simulated observations from an Expo($\lambda = 0.75$) distribution. The upper-right tail of the points curves sharply away from the reference line, indicating that the uniform model is *not* a good fit: the data have a much heavier right tail than the Uniform distribution would predict.



(b) Does the plot suggest that the data could come from a Normal distribution?

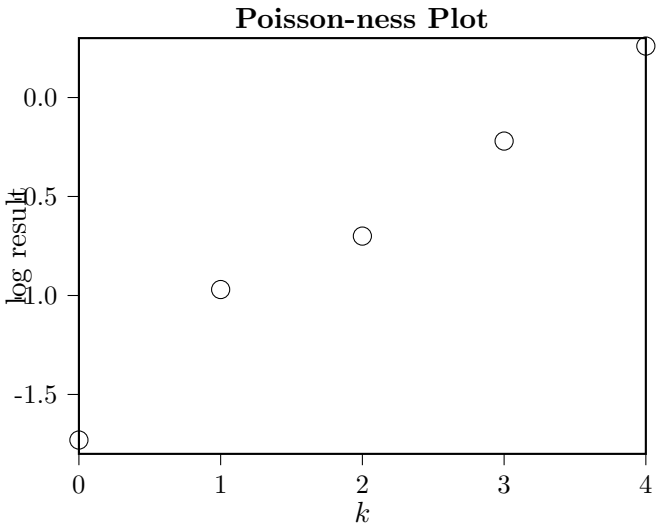


As it turns out, this is 30 simulated observations from an $N(\mu = 10, \sigma^2 = 4)$ distribution. The points follow the reference line very closely, providing strong support for normality.

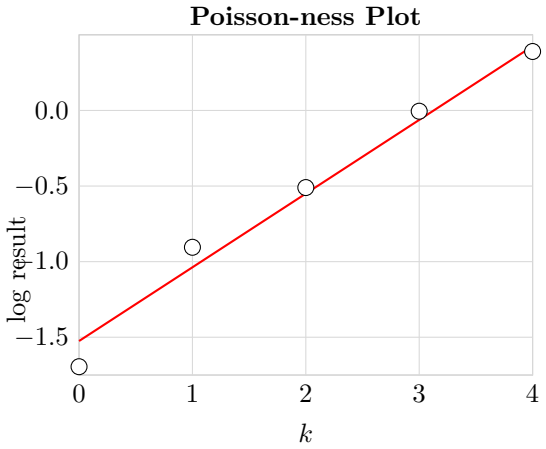


(c) Is the following data Poisson distributed?

k	0	1	2	3	4
n_k	9	19	12	7	3



As it turns out, this is 50 simulated observations from a Pois($\lambda = 1.5$) distribution. The Poisson-ness plot with the fitted regression line is shown below.



The five points fall very close to the fitted line (shown in red), confirming that the Poisson model is a good fit for these data.

Important Inequalities: Markov, Chebychev, and Jensen

Lecture 21 • UVA Stats

April 2026

Often we cannot compute a probability or expectation in closed form, but we can still bound it. This lecture proves three universal inequalities: Markov (tail bounded by mean over threshold), Chebychev (concentration around the mean in units of standard deviation), and Jensen (the expectation of a convex function bounds the function of the expectation). Each is proved from first principles and applied to the standard exponential distribution.

PART 1

Motivation

1. Three universal inequalities

Probability and statistics frequently demand that we say something useful about a random variable when we do not know its full distribution. Three classical inequalities make this possible: **Markov's Inequality**, which bounds the probability that a non-negative random variable exceeds a threshold; **Chebychev's Inequality**, which bounds the probability that any random variable deviates from its mean by more than a given multiple of its standard deviation; and **Jensen's Inequality**, which relates the expected value of a function of a random variable to the function of the expected value.

Each inequality converts a moment we can compute (a mean, a variance) into a bound on something harder to pin down (a tail probability, a transformed expectation). They appear constantly in proofs precisely because they impose no structural assumptions on the distribution beyond the existence of the relevant moment.

Key Idea 64

Markov's, Chebychev's, and Jensen's inequalities hold for *all* distribution families. The only requirements are:

- Markov: $Y \geq 0$ and $E[Y] < \infty$.
- Chebychev: $\text{Var}(Y) < \infty$.
- Jensen: $E[X] < \infty$ and the function g is convex (or concave).

This universality is exactly what makes them useful when closed-form computation is hopeless or the distribution is unknown.

PART 2

Methods and Examples

2. Markov's Inequality

Term — Markov's Inequality

Let X be a non-negative random variable, i.e. the support of X is some subset of $\{x \geq 0\}$, with a finite mean $E[X] < \infty$.

Then, for any value $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof. Define a new random variable

$$Z = \begin{cases} a & \text{if } X \geq a, \\ 0 & \text{if } X < a. \end{cases}$$

Because Z takes the value a whenever X reaches a and 0 otherwise, we have $Z \leq X$ everywhere on the sample space, and therefore $E[Z] \leq E[X]$.

Computing $E[Z]$ directly:

$$\begin{aligned} E[Z] &= \big(a \cdot P(X \geq a)\big) + \big(0 \cdot P(X < a)\big) \\ &= a \cdot P(X \geq a). \end{aligned}$$

Substituting into $E[Z] \leq E[X]$ gives

$$\begin{aligned} a \cdot P(X \geq a) &\leq E[X] \\ P(X \geq a) &\leq \frac{E[X]}{a}. \end{aligned}$$

Key Idea 65

For any non-negative random variable X with finite mean and any $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

The bound depends only on the mean of X, not on any other feature of its distribution.

Worked example: bounding income above a threshold

Define X as the income of an American worker. Suppose we want to know the probability that a randomly chosen worker earns at least five times the national average, i.e.

$$P\big(X \geq 5 \cdot E[X]\big).$$

We do not know the full income distribution, but Markov's Inequality gives an immediate upper bound: setting $a = 5 \cdot E[X]$,

$$P\big(X \geq 5 \cdot E[X]\big) \leq \frac{E[X]}{5 \cdot E[X]} = \frac{1}{5}.$$

Therefore, no more than 20% of workers can earn more than five times the national average, regardless of how income is distributed.

3. Chebychev's Inequality

Term — Chebychev's Inequality

Let Y be a random variable with finite variance $\mathrm{Var}(Y) < \infty$.

Then, for any $a > 0$,

$$P\big(|Y - E[Y]| \geq a\big) \;\leq\; \frac{\mathrm{Var}(Y)}{a^2}.$$

Proof. Define $X = (Y - E[Y])^2$. Since $X \geq 0$ we can apply Markov's Inequality:

$$\begin{aligned}
&P(|Y - E[Y]| \geq a) && \\
&= P\big((Y - E[Y])^2 \geq a^2\big) && \text{square both sides of the inequality} \\
&= P(X \geq a^2) && \text{substitute } X = (Y - E[Y])^2 \\
&\leq E[X]/a^2 && \text{apply Markov's Inequality} \\
&= E[(Y - E[Y])^2]/a^2 && \text{re-substitute } X = (Y - E[Y])^2 \\
&= \mathrm{Var}(Y)/a^2 && \mathrm{Var}(Y) = E[(Y - E[Y])^2]
\end{aligned}$$

Key Idea 66

For any random variable Y with finite variance and any $a > 0$,

$$P\big(|Y - E[Y]| \geq a\big) \;\leq\; \frac{\mathrm{Var}(Y)}{a^2}.$$

Chebychev's Inequality is Markov's Inequality applied to the squared deviation $(Y - E[Y])^2$. It converts information about the variance into a bound on how far the variable can stray from its mean.

Worked example: two standard deviations

For any random variable X, what is the probability that X is within two standard deviations of the mean? We seek a lower bound on

$$P\Big(|X - E[X]| < 2\sqrt{\mathrm{Var}(X)}\Big).$$

Chebychev's Inequality bounds the complementary event. Setting $a = 2\sqrt{\mathrm{Var}(X)}$:

$$P\Big(\,|X - E[X]| \geq 2\sqrt{\mathrm{Var}(X)}\,\Big) \;\leq\; \frac{\mathrm{Var}(X)}{\big(2\sqrt{\mathrm{Var}(X)}\big)^2} \;=\; \frac{\mathrm{Var}(X)}{4\,\mathrm{Var}(X)} \;=\; \frac{1}{4}.$$

By the complement rule, if $P\Big(\,|X - E[X]| \geq 2\sqrt{\mathrm{Var}(X)}\,\Big) \leq \frac{1}{4}$, then

$$P\Big(\,|X - E[X]| < 2\sqrt{\mathrm{Var}(X)}\,\Big) \;\geq\; \frac{3}{4}.$$

At least 75% of the probability mass of *any* distribution lies within two standard deviations of the mean.

4. Jensen's Inequality

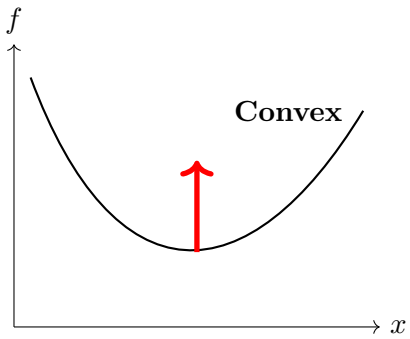
Term — Jensen's Inequality

Let X be a random variable with finite mean $E[X] < \infty$.

Then, for any concave up (convex) function $g(\cdot)$,

$$E[g(X)] \geq g(E[X]).$$

A function is *convex* (concave up) if it curves upward; equivalently, $g''(x) \geq 0$ everywhere on its domain.



Key Idea 67

Jensen's Inequality has two directions:

- If g is **convex** ($g''(x) \geq 0$), then $E[g(X)] \geq g(E[X])$.
- If g is **concave** ($g''(x) \leq 0$), then $E[g(X)] \leq g(E[X])$.

Equality holds in both cases if and only if g is linear on the support of X (or equivalently, X is essentially constant).

Why this is true. A convex function lies above every one of its tangent lines; in particular, above the tangent at $\mu = E[X]$, which has equation $\ell(x) = g(\mu) + g'(\mu)(x - \mu)$. Taking expectations, $E[g(X)] \geq E[\ell(X)] = g(\mu) + g'(\mu) \cdot E[X - \mu] = g(\mu)$, since the second term vanishes.

Worked example: $g(x) = x^2$ and the variance identity

Define $g(x) = x^2$. Because $g'(x) = 2x$ and $g''(x) = 2 \geq 0$, g is convex everywhere. Jensen's Inequality therefore gives

$$E[X^2] \geq \big(E[X]\big)^2.$$

This single inequality has an important consequence for the definition of variance. Recall that

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

Key Idea 68

Because $E[X^2] \geq \big(E[X]\big)^2$ (Jensen), we have $\mathrm{Var}(X) = E[X^2] - E[X]^2 \geq 0$. Jensen's Inequality thus provides a clean proof that variance is always non-negative.

Worked example: $g(x) = \log x$

Consider $g(x) = \ln x$. Here $g'(x) = x^{-1}$ and $g''(x) = -x^{-2} < 0$, so $\ln x$ is concave (concave down).

Since $g^*(x) = -\ln x$ is the negation of a concave function, g^* is convex. Applying Jensen's Inequality to g^*:

$$E[-\ln X] \geq -\ln E[X].$$

Multiplying both sides by -1 (and flipping the inequality):

$$E[\ln X] \leq \ln E[X].$$

This is the concave version of Jensen's Inequality applied directly to $g(x) = \ln x$: the expected log is no greater than the log of the expected value.

PART 3

Practice Problems

5. Example 1 — Markov on the Standard Exponential

We evaluate Markov's Inequality using the standard exponential distribution. Let

$$Y \sim \text{Expo}(\lambda = 1).$$

For the standard exponential distribution:

- The probability density function is $f(y) = \lambda e^{-\lambda y} = e^{-y}$.
- The cumulative distribution function is $F(y) = 1 - e^{-\lambda y} = 1 - e^{-y}$.
- The expected value is $E[Y] = 1/\lambda = 1$.
- The variance is $\text{Var}(Y) = 1/\lambda^2 = 1$.

(a) According to Markov's Inequality, what is the probability that *any* nonnegative random variable Y with a finite mean of $E[Y] = 1$ produces an outcome greater than 5, i.e.

$$P(Y \geq 5)\,?$$

(b) For the standard exponential distribution $Y \sim \text{Expo}(\lambda = 1)$, what is $P(Y \geq 5)$? Does the result satisfy Markov's Inequality, i.e. is $P(Y \geq 5) \leq \dfrac{1}{5}$?

Solutions

(a) Markov's Inequality states that $P(Y \geq a) \leq \dfrac{E[Y]}{a}$, so with $a = 5$ and $E[Y] = 1$:

$$P(Y \geq 5) \;\leq\; \frac{E[Y]}{5} \;=\; \frac{1}{5}.$$

(b) For the standard exponential we can compute the tail probability exactly:

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y < 5) \\ &= 1 - F(5) \\ &= 1 - (1 - e^{-5}) \\ &= e^{-5} \\ &\approx 0.0067. \end{aligned}$$

Since $0.0067 \leq 0.2 = \frac{1}{5}$, the exact probability does satisfy Markov's Inequality. Notice how much tighter the exact value is compared to the bound: the Markov bound is about 30 times larger than the true probability. This illustrates that Markov's bound is universal but can be quite loose for specific distributions.

6. Example 2 — Chebychev on the Standard Exponential

We continue with $Y \sim \text{Expo}(\lambda = 1)$, so $E[Y] = 1$ and $\text{Var}(Y) = 1$.

(a) According to Chebychev's Inequality, what is the probability that *any* random variable Y with $E[Y] = 1$ and $\text{Var}(Y) = 1$ produces an outcome more than two units from the mean, i.e.

$$P\big(|Y - E[Y]| \geq 2\big) = P\big(|Y - 1| \geq 2\big)\,?$$

(b) For the standard exponential distribution $Y \sim \text{Expo}(\lambda = 1)$, what is $P(|Y - 1| \geq 2)$? Does the result satisfy Chebychev's Inequality, i.e. is $P(|Y - 1| \geq 2) \leq \frac{1}{4}$?

Solutions

(a) Chebychev's Inequality states that $P(|Y - E[Y]| \geq a) \leq \dfrac{\text{Var}(Y)}{a^2}$. With $a = 2$, $E[Y] = 1$, and $\text{Var}(Y) = 1$:

$$P(|Y - 1| \geq 2) \;\leq\; \frac{\text{Var}(Y)}{2^2} \;=\; \frac{1}{4}.$$

(b) For the standard exponential we split the event using the absolute value:

$$\begin{aligned} P(|Y - 1| \geq 2) &= P\big((Y - 1) \leq -2\big) &+ \quad & P\big((Y - 1) \geq 2\big) \\ &= P(Y \leq -1) &+ \quad & P(Y \geq 3) \\ &= 0 &+ \quad & P(Y \geq 3) \\ &= \;\; P(Y \geq 3). \end{aligned}$$

The event $\{Y \leq -1\}$ has probability zero because the support of the exponential is $[0, \infty)$. We now compute $P(Y \geq 3)$ exactly:

Key Idea 69

$$\begin{aligned} P(Y \geq 3) &= 1 - P(Y < 3) \\ &= 1 - F(3) \\ &= 1 - (1 - e^{-3}) \\ &= e^{-3} \\ &= 0.050. \end{aligned}$$

Since $0.050 \leq 0.25 = \frac{1}{4}$, the exact probability satisfies Chebychev's Inequality, as guaranteed. As with the Markov comparison, the exact probability (0.050) is substantially smaller than the universal bound (0.25), confirming that these inequalities are conservative but reliable across all distributions.

Introduction to Linear Algebra

Lecture 23 • UVA Stats

April 22, 2026

An introduction to the objects and operations that underpin statistical computing: vectors, their geometry, and the four multiplications that connect them. We motivate the subject from a statistician's viewpoint, develop the methods with worked examples, and close with a set of practice problems covering every operation introduced.

PART 1

Motivation

1. What is Linear Algebra?

Term — Linear Algebra

Linear algebra is the branch of mathematics concerned with *vectors* and *matrices*, their linear combinations, and operations acting upon them.

1.1 Why it matters for statisticians

Data are almost never a single number. A survey gives a list of responses; an image gives a grid of pixels; a panel study gives a table indexed by person and time. Each of these is naturally a vector or a matrix, and every statistical method you will learn — regression, PCA, classification, clustering — is ultimately a sequence of linear-algebra operations acting on that structure.

Key Idea 70

Data live in matrices. As statisticians we care about linear algebra because our data are typically stored in matrices, and analysis of this data is carried out using linear algebra operations. Mastering the vocabulary of vectors and their operations is therefore a prerequisite for everything that follows.

PART 2

Methods and Examples

2. Vectors

Term — Vector

A **vector** in dimension p is an ordered list of p real numbers. Geometrically it is represented as a directed line segment (an arrow) from the origin, characterised by a *direction* and a *magnitude* (also called the norm, or length).

A vector can be written with parentheses, square brackets, or angle brackets — they all refer to the same object:

$$(2,3) \;=\; [2,3] \;=\; \langle 2,3\rangle.$$

2.1 Dimension, Direction, and Magnitude

Three attributes describe every vector:

- **Dimension** — the number of entries in the vector, which fixes the coordinate system. The vector (a,b,c) has three entries, so it lives in 3-D.
- **Direction** — the number of units the vector travels in each coordinate direction. The vector (a,b,c) travels a units in direction 1, b units in direction 2, and c units in direction 3.
- **Magnitude (norm)** — the overall length of the vector:

$$\|(a,b,c)\| = \sqrt{a^2+b^2+c^2}.$$

Worked example

For $\mathbf{v} = (2,3)$:

- dimension is 2, so the coordinate system is 2-D;
- travels 2 units in x and 3 units in y;
- magnitude $\|\mathbf{v}\| = \sqrt{2^2+3^2} = \sqrt{13} \approx 3.60$.

2.2 Standard Position

Vectors do not have a designated starting point. The vector $(1,-2)$ can be drawn anywhere in the plane, and in every case it represents the same vector. For interpretability we usually plot vectors in *standard position*, with the tail at the origin.

2.3 Row and Column Vectors

Vectors are typically labelled with a lowercase letter, and can be written as a **row vector** or a **column vector**:

$$\mathbf{v} = (2,3) \qquad \text{or} \qquad \mathbf{v} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

The interpretation is the same in either case.

2.4 The Transpose

Term — Transpose

The **transpose** of a vector flips its rows and columns.

$$(a,b,c)^T = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \qquad \begin{pmatrix} a \\ b \\ c \end{pmatrix}^T = (a,b,c).$$

3. Scalar Multiplication

Term — Scalar

A **scalar** is a single number (a constant). Scalars are typically denoted by Greek letters, e.g. $\alpha = 5$ or $\lambda = -1$.

Scalar multiplication of a vector $\mathbf{v} = (a,b,c)^T$ by λ is defined element-wise:

$$\lambda\mathbf{v} = \begin{pmatrix} \lambda a \\ \lambda b \\ \lambda c \end{pmatrix}.$$

3.1 Geometric effect

Scalar multiplication changes only the magnitude of the vector, *not* its dimension or its axis of direction:

- If $\lambda > 0$, $\lambda\mathbf{v}$ keeps the direction of $\mathbf{v}$.
- If $\lambda < 0$, $\lambda\mathbf{v}$ reverses the direction.
- If $|\lambda| > 1$, $\lambda\mathbf{v}$ is longer than $\mathbf{v}$.
- If $|\lambda| < 1$, $\lambda\mathbf{v}$ is shorter than $\mathbf{v}$.

3.2 Unit Vectors

Term — Unit Vector

Any vector with magnitude 1 is a **unit vector**, written $\mathbf{u}$.

We can build a unit vector in the direction of any non-zero $\mathbf{v}$ using scalar multiplication:

$$\mathbf{u}_v = \frac{1}{\|\mathbf{v}\|}\mathbf{v}.$$

Worked example

Let $\mathbf{v} = (4, 7, 4)$.

$$\|\mathbf{v}\| = \sqrt{4^2 + 7^2 + 4^2} = \sqrt{81} = 9,$$

so

$$\mathbf{u}_v = \tfrac{1}{9}(4, 7, 4) = \left(\tfrac{4}{9},\ \tfrac{7}{9},\ \tfrac{4}{9}\right).$$

4. Vector Addition and Subtraction

4.1 Addition

Vector addition is carried out entry-by-entry (so the two vectors must share a dimension):

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1,\ a_2 + b_2,\ \ldots,\ a_p + b_p).$$

Geometrically, place the two vectors head-to-tail; the sum runs from the first tail to the last head.

Worked example

Let $\mathbf{v}_1 = (1, 2)$ and $\mathbf{v}_2 = (2, 0)$.

$$\mathbf{v}_1 + \mathbf{v}_2 = (1 + 2,\ 2 + 0) = (3, 2).$$

4.2 Subtraction

Vector subtraction is analogous; equivalently, add the negative of the second vector to the first:

$$\mathbf{a} - \mathbf{b} = (a_1 - b_1,\ a_2 - b_2,\ \ldots,\ a_p - b_p).$$

Worked example

With $\mathbf{v}_1 = (1, 2)$ and $\mathbf{v}_2 = (2, 0)$,

$$\mathbf{v}_1 - \mathbf{v}_2 = (-1, 2), \qquad \mathbf{v}_2 - \mathbf{v}_1 = (1, -2).$$

Note that $\mathbf{v}_2 - \mathbf{v}_1$ points in the exact opposite direction from $\mathbf{v}_1 - \mathbf{v}_2$.

5. Vector Multiplication

Key Idea 71

There are *four* distinct notions of multiplication for vectors, and they do different things:
(1) Dot product (inner product) — returns a scalar.
(2) Outer product — returns a matrix.
(3) Hadamard product (element-wise) — returns a vector.
(4) Cross product — returns a vector (in 3-D only).

5.1 Dot Product

Term — Dot Product

The **dot product** of two vectors of the same dimension is the sum of the products of corresponding entries. It returns a scalar.

For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$,

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = \sum_{i=1}^{p} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_p b_p.$$

Worked example

Let $\mathbf{a} = (-3, 1, 4)^T$ and $\mathbf{b} = (2, -2, 1)^T$. Then

$$\mathbf{a} \cdot \mathbf{b} = (-3)(2) + (1)(-2) + (4)(1) = -6 - 2 + 4 = -4.$$

Linear combinations.

If $\mathbf{c}$ holds prices and $\mathbf{q}$ holds quantities, $\mathbf{c} \cdot \mathbf{q}$ is the total bill. A group orders four drinks (\$3), three burgers (\$10), and two nachos (\$12):

$$\mathbf{c} = (3, 10, 12), \quad \mathbf{q} = (4, 3, 2), \quad \mathbf{c} \cdot \mathbf{q} = 12 + 30 + 24 = 66.$$

Weighted averages.

For data $\mathbf{v} = (v_1, \ldots, v_n)$ and weights $\mathbf{w} = (w_1, \ldots, w_n)$, the weighted average is simply $\mathbf{v} \cdot \mathbf{w}$.

Geometry.

Projecting $\mathbf{a}$ onto $\mathbf{b}$ gives the relation

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta_{ab},$$

where θ_{ab} is the angle between the two vectors. This identity follows from the law of cosines applied to the triangle with sides $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a} - \mathbf{b}$. *If $\mathbf{a} \cdot \mathbf{b} = 0$ and neither vector is zero, the two vectors are perpendicular.*

Illustration.

Let $\mathbf{a} = (2, 3)$ and $\mathbf{b} = (4, 0)$.

$$\|\mathbf{a}\| = \sqrt{13}, \qquad \|\mathbf{b}\| = 4, \qquad \mathbf{a} \cdot \mathbf{b} = (2)(4) + (3)(0) = 8.$$

The projection of $\mathbf{a}$ onto $\mathbf{b}$ is $\mathbf{a}_b = (2, 0)$ with $\|\mathbf{a}_b\| = 2$, and indeed $\|\mathbf{a}_b\| \, \|\mathbf{b}\| = 2 \cdot 4 = 8$. To recover the angle,

$$\cos\theta_{ab} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|} = \frac{8}{4\sqrt{13}} = \frac{2}{\sqrt{13}}, \qquad \theta_{ab} = \cos^{-1}\!\left(\tfrac{2}{\sqrt{13}}\right) \approx 56.3^\circ.$$

5.2 Outer Product

Term — Outer Product

The **outer product** of two vectors creates a matrix of the products of all possible combinations of entries. It does *not* require the two vectors to share a dimension.

For $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$, the outer product is the $m \times n$ matrix

$$\mathbf{a}\mathbf{b}^T = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_mb_1 & a_mb_2 & \cdots & a_mb_n \end{bmatrix}.$$

Worked example

With $\mathbf{a} = (-3, 1, 4)^T$ and $\mathbf{b} = (2, -2, 1)^T$,

$$\mathbf{a}\mathbf{b}^T = \begin{bmatrix} -6 & 6 & -3 \\ 2 & -2 & 1 \\ 8 & -8 & 4 \end{bmatrix}.$$

5.3 Hadamard (Element-wise) Product

Term — Hadamard Product

The **Hadamard product** $\mathbf{a} \odot \mathbf{b}$ is element-wise multiplication of two vectors of the same dimension.

$$\mathbf{a} \odot \mathbf{b} = (a_1b_1,\ a_2b_2,\ \ldots,\ a_pb_p).$$

Worked example

With $\mathbf{a} = (-3, 1, 4)^T$ and $\mathbf{b} = (2, -2, 1)^T$,

$$\mathbf{a} \odot \mathbf{b} = \begin{pmatrix} -6 \\ -2 \\ 4 \end{pmatrix}.$$

Key Idea 72

The Hadamard product is the natural operation for *assigning weights* or *injecting random error* into data stored as a vector — each entry is scaled independently.

5.4 Cross Product

Term — Cross Product

The **cross product** is defined only for two 3-D vectors. It returns a vector perpendicular to both inputs.

For $\mathbf{a} = (a_1, a_2, a_3)^T$ and $\mathbf{b} = (b_1, b_2, b_3)^T$,

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_3b_1 - a_1b_3 \\ a_1b_2 - a_2b_1 \end{pmatrix}.$$

Worked example

With $\mathbf{a} = (-3, 1, 4)^T$ and $\mathbf{b} = (2, -2, 1)^T$,

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} (1)(1) - (4)(-2) \\ (4)(2) - (-3)(1) \\ (-3)(-2) - (1)(2) \end{pmatrix} = \begin{pmatrix} 9 \\ 11 \\ 4 \end{pmatrix}.$$

Sanity check: the standard basis.

For the unit basis vectors in $\mathbb{R}^3$,

$$\mathbf{x} = (1, 0, 0), \quad \mathbf{y} = (0, 1, 0), \quad \mathbf{x} \times \mathbf{y} = (0, 0, 1) = \mathbf{z},$$

which is perpendicular to both $\mathbf{x}$ and $\mathbf{y}$, as advertised.

PART 3

Practice Problems

6. Example #1

Define the vector

$$\mathbf{a} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}.$$

(a) What is the dimension and magnitude of $\mathbf{a}$?

(b) Find a unit vector in the direction of $\mathbf{a}$.

Now introduce a second vector, $\mathbf{b} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$, and compute:

(c) $\mathbf{a} + \mathbf{b}$;

(d) $\mathbf{a} - \mathbf{b}$;

(e) the dot product $\mathbf{a} \cdot \mathbf{b}$;

(f) the angle θ_{ab} between $\mathbf{a}$ and $\mathbf{b}$;

(g) the outer product $\mathbf{a}\mathbf{b}^T$;

(h) the Hadamard product $\mathbf{a} \odot \mathbf{b}$.

Solutions

(a) Dimension and magnitude of a. $\mathbf{a}$ has two entries, so its dimension is 2.

$$\|\mathbf{a}\| = \sqrt{4^2 + 3^2} = \sqrt{25} = 5.$$

(b) Unit vector in the direction of a.

$$\mathbf{u}_a = \frac{1}{\|\mathbf{a}\|}\,\mathbf{a} = \frac{1}{5}\begin{pmatrix}4\\3\end{pmatrix} = \begin{pmatrix}0.8\\0.6\end{pmatrix}.$$

(c) $\mathbf{a} + \mathbf{b}$.

$$\mathbf{a} + \mathbf{b} = \begin{pmatrix}4 + (-1)\\3 + 2\end{pmatrix} = \begin{pmatrix}3\\5\end{pmatrix}.$$

(d) $\mathbf{a} - \mathbf{b}$.

$$\mathbf{a} - \mathbf{b} = \begin{pmatrix}4 - (-1)\\3 - 2\end{pmatrix} = \begin{pmatrix}5\\1\end{pmatrix}.$$

(e) Dot product.

$$\mathbf{a} \cdot \mathbf{b} = (4)(-1) + (3)(2) = -4 + 6 = 2.$$

(f) Angle between a and b. First $\|\mathbf{b}\| = \sqrt{(-1)^2 + 2^2} = \sqrt{5}$. Then

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= \|\mathbf{a}\|\,\|\mathbf{b}\|\,\cos\theta_{ab}\\ 2 &= (5)(\sqrt{5})\cos\theta_{ab}\\ \cos\theta_{ab} = \frac{2}{5\sqrt{5}} &\implies \theta_{ab} = \cos^{-1}\!\left(\tfrac{2}{5\sqrt{5}}\right) \approx 79.7^\circ.\end{aligned}$$

(g) Outer product.

$$\mathbf{a}\mathbf{b}^T = \begin{bmatrix}(4)(-1) & (4)(2)\\(3)(-1) & (3)(2)\end{bmatrix} = \begin{bmatrix}-4 & 8\\-3 & 6\end{bmatrix}.$$

(h) Hadamard product.

$$\mathbf{a} \odot \mathbf{b} = \begin{pmatrix}(4)(-1)\\(3)(2)\end{pmatrix} = \begin{pmatrix}-4\\6\end{pmatrix}.$$

Vector Spaces and Subspaces

Lecture 24 • UVA Stats

April 22, 2026

Where do vectors live? This lecture develops the language of fields, vector spaces, *and* subspaces *that situates last lecture's operations, then introduces the notion of an* ambient space *and the algebraic test —* linear independence *— that tells us the dimensionality of the subspace a collection of vectors spans. We close with two worked practice problems, each illustrated by the geometric picture.*

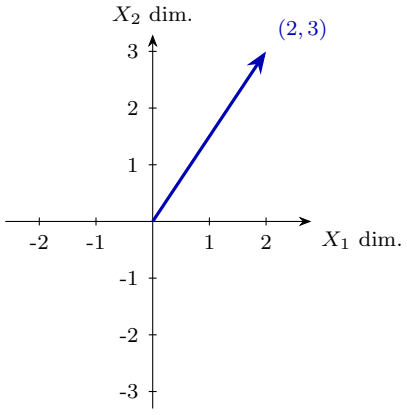
PART 1

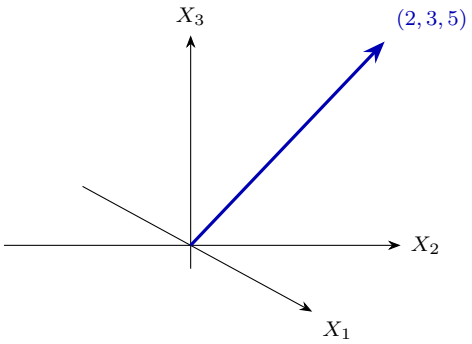
Motivation

1. Where Last Lecture Left Us

Term — Vector (review)

A **vector** is an ordered list of values, like $\boldsymbol{v} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ or $\boldsymbol{a} = (2, 3, 5)$. Geometrically, a vector in dimension p is a directed arrow from the origin, characterised by a *direction* and a *magnitude* (also called the norm, or length).





In the previous lecture we discussed seven operations on vectors: *scalar multiplication*, *addition*, *subtraction*, the *dot product*, the *outer product*, *Hadamard* (element-wise) multiplication, and the *cross product*.

Key Idea 73

Today's goal. Understand the *mathematical spaces that vectors occupy* — and the spaces that vectors can *generate* when we combine them via the operations above. Those spaces are what the rest of statistics (regression, PCA, classification, ...) quietly lives inside.

PART 2

Methods and Examples

2. Fields

Term — Field

A **field** is a set of numbers for which the four basic arithmetic operations (addition, subtraction, multiplication, division) are all defined.

Two fields you already know:

- $\mathbb{R}$ — the real numbers.
- $\mathbb{Q}$ — the rational numbers, i.e. numbers expressible as a fraction $\frac{a}{b}$ of two integers a and b.

Fields are how we specify the *dimensionality* and the *entry type* of a vector.

Worked example

$\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ is the set of all real-number-valued 2-dimensional vectors (a, b) with $a, b \in \mathbb{R}$. More generally, $\mathbb{R}^n$ denotes the real-valued n-dimensional vectors.

3. Vector Spaces

Term — Vector Space

A **vector space** is any set of objects (*vectors*) for which addition and scalar multiplication are defined — and the set is *closed* under both operations.

Worked example

Consider

$$\{(v_1, v_2, v_3) : v_1, v_2, v_3 \in \mathbb{R}\} \;=\; \mathbb{R}^3.$$

Addition is defined:

$$(v_1, v_2, v_3) + (x_1, x_2, x_3) = (v_1 + x_1,\ v_2 + x_2,\ v_3 + x_3).$$

The sum of two real numbers is still a real number, so $(v_1 + x_1), (v_2 + x_2), (v_3 + x_3) \in \mathbb{R}$ and the result stays in $\mathbb{R}^3$.

Scalar multiplication is defined:

$$\lambda(v_1, v_2, v_3) = (\lambda v_1,\ \lambda v_2,\ \lambda v_3).$$

The product of two real numbers is still real, so $\lambda v_1, \lambda v_2, \lambda v_3 \in \mathbb{R}$ and the result again stays in $\mathbb{R}^3$.

Vector spaces can be defined in much more convoluted ways than this, but that is not particularly relevant for our study of vectors. We have pinned the concept down so that we can distinguish it from a vector *sub*space.

4. Subspaces

Term — Subspace

A **subspace** is the set of all points obtainable by combining a particular collection of vectors from a larger vector space through addition and scalar multiplication. A subspace is itself a vector space, and is generally written with an italicised capital letter (V, W, ...).

4.1 Span of a single vector

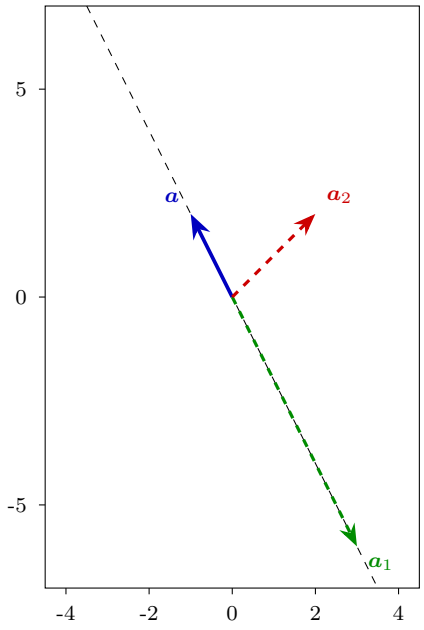
For the vector $\boldsymbol{a} = (-1, 2)$, the subspace is

$$V = \{\, \lambda \boldsymbol{a} = \lambda(-1, 2) = (-\lambda, 2\lambda) : \lambda \in \mathbb{R} \,\}.$$

Because there is only a single vector in the collection we only have to worry about scalar multiplication. Geometrically V is the line through the origin with direction $\boldsymbol{a}$, i.e. the line $y = -2x$.

Two quick test points:

- $\boldsymbol{a}_1 = (3, -6)$ *is* in the subspace of $\boldsymbol{a}$, with $\lambda = -3$ (since $-3 \cdot (-1, 2) = (3, -6)$).
- $\boldsymbol{a}_2 = (2, 2)$ is <u>not</u> in the subspace of $\boldsymbol{a}$: no scalar λ sends $(-1, 2)$ to $(2, 2)$.



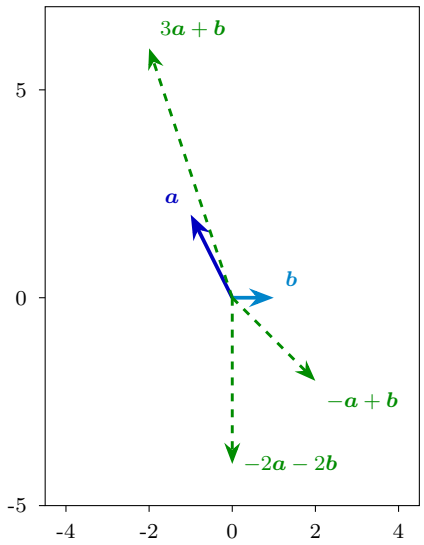
4.2 Span of two vectors

For $\boldsymbol{a} = (-1, 2)$ and $\boldsymbol{b} = (1, 0)$, the subspace is

$$W = \{\, \lambda_1 \boldsymbol{a} + \lambda_2 \boldsymbol{b} = (-\lambda_1 + \lambda_2,\ 2\lambda_1) : \lambda_1, \lambda_2 \in \mathbb{R} \,\}.$$

Three representative linear combinations (drawn dashed in green) illustrate how the span fills the plane:

$$3\boldsymbol{a} + \boldsymbol{b} = (-2, 6), \qquad -\boldsymbol{a} + \boldsymbol{b} = (2, -2), \qquad -2\boldsymbol{a} - 2\boldsymbol{b} = (0, -4).$$



Key Idea 74

Every subspace contains the origin. Taking $\lambda_1 = \cdots = \lambda_m = 0$ in any linear combination yields the zero vector, so $\mathbf{0}$ is unavoidably in every subspace. For W above, $\lambda_1 = \lambda_2 = 0 \Longrightarrow \lambda_1 \boldsymbol{a} + \lambda_2 \boldsymbol{b} = (0, 0)$. This gives a fast way to *disqualify* a candidate set: if it misses the origin, it isn't a subspace.

5. Ambient Space

Term — Ambient Space

The **ambient space** is the larger space in which the subspace is embedded. Its dimension equals the number of *components* of each vector — e.g. $\mathbb{R}^n$ for vectors with n entries — regardless of how many vectors are used to generate the subspace.

Examples:

- For $\boldsymbol{a} = (-1, 2)$, the subspace sits in a 2-D ambient space, perhaps $\mathbb{R}^2$.
- For $\boldsymbol{v}_1 = (4, 1, 5)$, $\boldsymbol{v}_2 = (-2.5, 2, 2)$, $\boldsymbol{v}_3 = (-1, -1, -7)$, the subspace sits in a 3-D ambient space, perhaps $\mathbb{R}^3$.

A subspace inside an n-dimensional ambient space can itself have dimensionality anywhere between 0 and n. To pin that dimension down we need one more tool.

6. Linear Independence

Term — Linear Independence

A collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ is **linearly independent** if no vector in the collection is a linear combination of the others:

$$\boldsymbol{v}_i \neq \sum_{j \neq i} \lambda_j \boldsymbol{v}_j \qquad \text{for every } i.$$

Otherwise, the collection is **linearly dependent**.

Equivalent definition via the zero vector

A collection is linearly dependent iff some non-trivial linear combination produces the zero vector:

$$\mathbf{0} = \sum_{j=1}^{m} \lambda_j \boldsymbol{v}_j, \qquad \text{with at least one } \lambda_j \neq 0.$$

Worked example

Let $\boldsymbol{a} = (1, -2)$, $\boldsymbol{b} = (1, 0)$, $\boldsymbol{c} = (-5, -1)$.

$\{\boldsymbol{a}, \boldsymbol{b}\}$ is linearly independent. No scalar λ sends $(1, 0)$ to $(1, -2)$, so $\boldsymbol{a} \neq \lambda \boldsymbol{b}$ for all $\lambda \in \mathbb{R}$.

$\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ is linearly dependent. We can express $\boldsymbol{a}$ as a combination of $\boldsymbol{b}$ and $\boldsymbol{c}$:

$$11\boldsymbol{b} + 2\boldsymbol{c} = 11(1, 0) + 2(-5, -1) = (11, 0) + (-10, -2) = (1, -2) = \boldsymbol{a}.$$

Equivalently, in the zero-vector formulation:

$$\boldsymbol{a} - 11\boldsymbol{b} - 2\boldsymbol{c} = (1, -2) + (-11, 0) + (10, 2) = (0, 0).$$

Counting rule

For m vectors of dimension n:

- If $m > n$, the collection is automatically *linearly dependent* — there are more vectors than the ambient space can accommodate independently.
- If $m \leq n$, the collection *may or may not* be linearly independent; we must check directly whether any vector is a linear combination of the others.

Key Idea 75

Dimensionality of a subspace. A collection of m vectors of dimension n that contains exactly d linearly independent members spans a *d-dimensional* subspace. Redundant (dependent)

vectors add no new directions; they just retrace ground already covered.

PART 3

Practice Problems

7. Example #1 — Is This a Subspace?

We know that $\{(v_1, v_2) : v_1, v_2 \in \mathbb{R}\} = \mathbb{R}^2$ is a vector space. Define the candidate subset

$$V = \{(v_1, v_2) : v_1, v_2 \geq 0\}$$

(the closed first quadrant).

Question. Is V a subspace of $\mathbb{R}^2$?

Solution

Addition is defined: $(v_1, v_2) + (x_1, x_2) = (v_1 + x_1,\ v_2 + x_2)$.

We know that the sum of two non-negative numbers v_i and x_i is still non-negative, so $(v_1 + x_1), (v_2 + x_2) \geq 0$. (✓)

Scalar multiplication is **not** defined: $\lambda(v_1, v_2) = (\lambda v_1,\ \lambda v_2)$.

For $\lambda < 0$ and $v_i > 0$, we have $\lambda v_i < 0$. A vector with a negative-valued element is not a part of V, so V is *not* closed under scalar multiplication. (×)

Conclusion. V is *not* a subspace of $\mathbb{R}^2$.

8. Example #2 — Dimensionality of a Span

Define the collection

$$\boldsymbol{S} = \left\{ \begin{pmatrix} -2 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 \\ \alpha \end{pmatrix} \right\}.$$

(a) For what value of α will this collection of vectors be linearly dependent?

(b) When $\alpha = -15$, what is the dimensionality of the subspace spanned by $\boldsymbol{S}$?

(c) When $\alpha = 0$, what is the dimensionality of the subspace spanned by $\boldsymbol{S}$?

(d) At any value of α, what is the ambient space in which this subspace is embedded?

Solutions

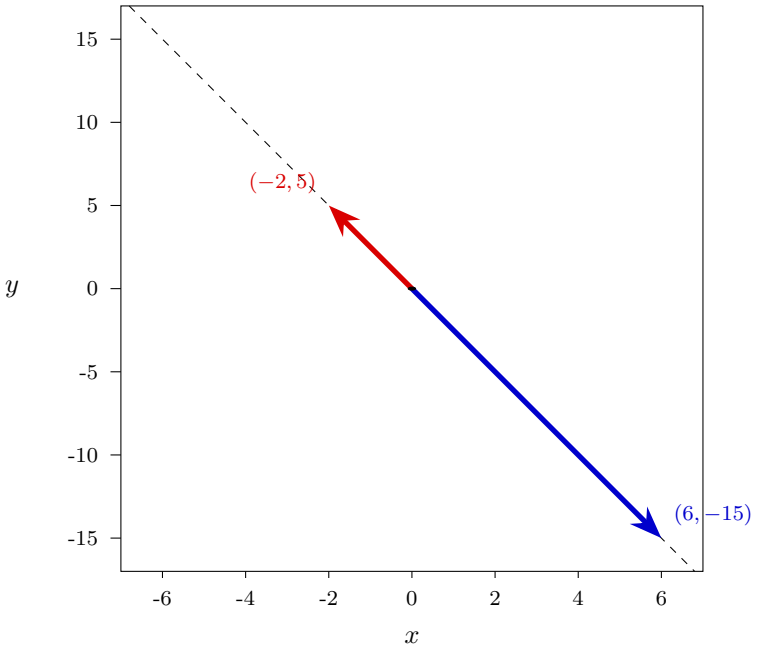
(a) Value of α making $\boldsymbol{S}$ linearly dependent. Using the zero-vector formulation, seek a scalar λ (and a value of α) satisfying

$$\begin{pmatrix}-2\\5\end{pmatrix}+\lambda\begin{pmatrix}6\\\alpha\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix} \qquad -2+6\lambda=0 \qquad 5+\alpha\lambda=0$$
$$\begin{pmatrix}-2\\5\end{pmatrix}+\begin{pmatrix}6\lambda\\\alpha\lambda\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix} \qquad 6\lambda=2 \qquad 5+\tfrac{1}{3}\alpha=0$$
$$\begin{pmatrix}-2+6\lambda\\5+\alpha\lambda\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix} \qquad \lambda=\tfrac{1}{3} \qquad \tfrac{1}{3}\alpha=-5$$
$$\alpha=-15$$

So $\boldsymbol{S}$ is linearly dependent precisely when $\alpha=-15$.

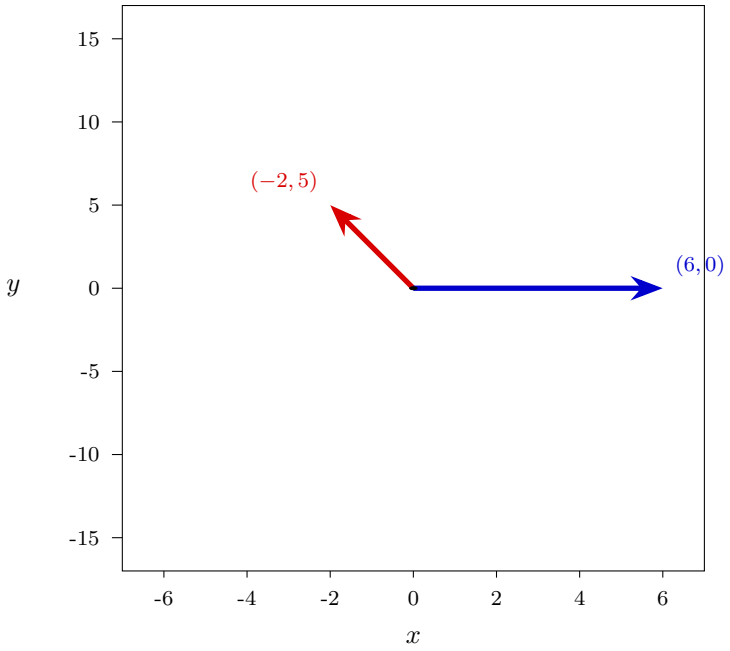
(b) Dimensionality when $\alpha=-15$. $\begin{pmatrix}-2\\5\end{pmatrix}$ and $\begin{pmatrix}6\\-15\end{pmatrix}$ are linearly *dependent* vectors. The subspace will only be **one-dimensional**.

Geometrically, both vectors lie on the line $y=-\frac{5}{2}x$ through the origin (a quick check: $-\frac{5}{2}\cdot(-2)=5$ and $-\frac{5}{2}\cdot 6=-15$), so the span is that single line:



(c) Dimensionality when $\alpha=0$. $\begin{pmatrix}-2\\5\end{pmatrix}$ and $\begin{pmatrix}6\\0\end{pmatrix}$ are linearly *independent* vectors. Two linearly independent vectors will form a **two-dimensional** subspace.

Indeed, neither vector is a scalar multiple of the other (the second has zero second-coordinate, the first does not), so the span is all of $\mathbb{R}^2$:



(d) Ambient space. Because both vectors in the collection $\boldsymbol{S}$ contain two elements — i.e. the vectors are two-dimensional — the subspace spanned by $\boldsymbol{S}$ is embedded in a two-dimensional ambient space, which we call $\mathbb{R}^2$ for real-number-valued vectors.

Span and Basis

Lecture 25 • UVA Stats

April 22, 2026

Having built up vector spaces, subspaces, and linear independence last lecture, we now pin down two more ideas that sit at the heart of multivariate statistics: the span *of a collection of vectors and a* basis *for a subspace. A basis is a minimal, linearly independent "ruler" for its subspace; every point inside admits unique coordinates in that ruler. We close with a three-part practice problem that exercises span, membership, and the basis criterion.*

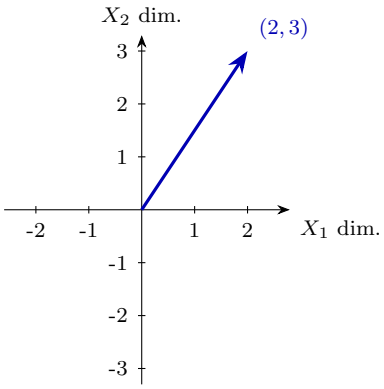
PART 1

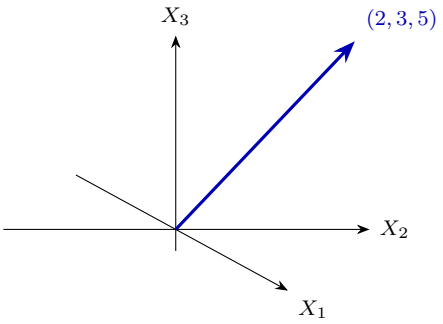
Motivation

1. Where Last Lecture Left Us

Term — Vector, Subspace, Ambient Space (review)

A **vector** is an ordered list of values like $\boldsymbol{v} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ or $\boldsymbol{a} = (2, 3, 5)$, determined by a direction and a magnitude. A **subspace** is the set of all points obtainable by combining a particular collection of vectors from a larger vector space through addition and scalar multiplication. The **ambient space** is the larger space in which the subspace is embedded — think of it as the graphical or plotting space. A collection of m vectors in $\mathbb{R}^n$, of which d are linearly independent, forms a d-dimensional subspace.





Key Idea 76

Today's goal. Two words enter the picture: *span* (the verb for how a collection of vectors generates a subspace) and *basis* (a minimal, linearly independent ruler for that subspace). Most of multivariate statistics is ultimately about choosing *good* bases for the subspaces your data inhabit.

PART 2

Methods and Examples

2. Span

Term — Span

A collection of vectors **spans** the subspace that is the set of all points obtainable by combining them through addition and scalar multiplication. "Subspace" is the noun; "span" is the verb.

Worked example: span of the standard basis in $\mathbb{R}^2$

For $\boldsymbol{a} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\boldsymbol{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$,

$$V = \left\{ \lambda_1 \boldsymbol{a} + \lambda_2 \boldsymbol{b} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} : \lambda_1, \lambda_2 \in \mathbb{R} \right\} = \mathbb{R}^2.$$

We say the *subspace* $V = \mathbb{R}^2$ is formed by the *span* of $\{\boldsymbol{a}, \boldsymbol{b}\}$.

Membership: is a new vector in the span?

A common question: given a collection S, is some new vector $\boldsymbol{w}$ in $\text{span}(S)$?

Example. Is $\boldsymbol{w} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$ in the span of $\left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix} \right\}$?

Yes. Observe that

$$-2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \boldsymbol{w}.$$

Key Idea 77

Membership in a span is not always easy to verify by hand. The matrix-algebra algorithms from the next few lectures — row reduction and the rank of an augmented matrix — turn this guess-and-check into a mechanical procedure.

3. Basis

Term — Basis

A collection of n-dimensional vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ forms a **basis** for some subspace of $\mathbb{R}^n$ if
(1) $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ spans that entire subspace, and
(2) $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ is linearly independent.

Think of a basis as a *ruler* or a *coordinate system* for the subspace.

The Cartesian basis for $\mathbb{R}^2$

The set $\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$ is a basis for all of $\mathbb{R}^2$:

(1) To reach any point (x, y) in $\mathbb{R}^2$ we take $x\begin{pmatrix}1\\0\end{pmatrix} + y\begin{pmatrix}0\\1\end{pmatrix}$.

(2) $\begin{pmatrix}1\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$ are linearly independent: $\lambda_1\begin{pmatrix}1\\0\end{pmatrix} + \lambda_2\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}$ forces $\lambda_1 = \lambda_2 = 0$.

This basis is valued for its simplicity: both vectors are unit vectors, and they are perpendicular (their dot product is zero).

A basis need not span the whole ambient space

Consider $V = \left\{\begin{pmatrix}x\\0\end{pmatrix} : x \in \mathbb{R}\right\}$, a 1-dimensional subspace of $\mathbb{R}^2$ (the x-axis). The set $\left\{\begin{pmatrix}1\\0\end{pmatrix}\right\}$ alone is a basis for V.

Bases are not unique

The same subspace admits many different bases. For instance, $\left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$ is *also* a basis for $\mathbb{R}^2$:

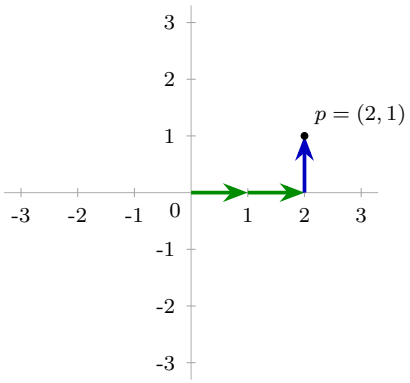
(1) To reach any (x, y): $x\begin{pmatrix}1\\1\end{pmatrix} + (y - x)\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}x\\x + (y - x)\end{pmatrix} = \begin{pmatrix}x\\y\end{pmatrix}$. ✓

(2) Linearly independent: neither vector is a scalar multiple of the other.

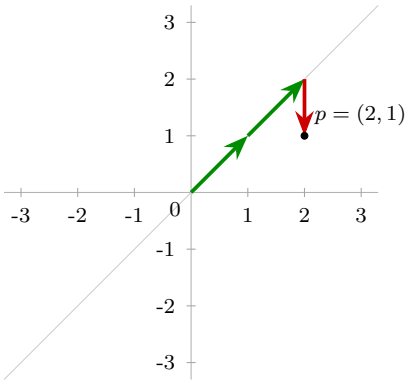
3.1 Coordinates inside a basis

To identify a point p in a subspace we determine the set of coefficients that multiplies each basis vector to get from the origin to p.

Cartesian basis. Let $p = (2, 1)$. Coefficients: $(2, 1)$. Two unit steps along $\begin{pmatrix}1\\0\end{pmatrix}$, then one step along $\begin{pmatrix}0\\1\end{pmatrix}$:



Alternate basis $\left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$**.** Same point $p = (2, 1)$, different coefficients: $(2, -1)$. Two steps along $\begin{pmatrix}1\\1\end{pmatrix}$ to $(2, 2)$, then minus one step along $\begin{pmatrix}0\\1\end{pmatrix}$ back down to $(2, 1)$:



Key Idea 78

Coordinates in a basis are unique. Although there are infinitely many bases for the same subspace, once a basis is fixed every point/vector has *exactly one* coordinate tuple in it. Uniqueness follows from the linear independence of the basis vectors.

Why care about multiple bases?

If coordinates are unique once we pick a basis, why allow different bases at all? Why not just stick with the Cartesian system $\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$ for $\mathbb{R}^2$?

- Some problems become dramatically easier in a well-chosen basis.
- Finding the "optimal" basis for a given dataset is a *central* focus of multivariate analysis — PCA, for instance, is exactly the problem of choosing a basis in which the coordinates of the data are as informative as possible.

PART 3

Practice Problems

4. Example #1 — Span, Membership, and Basis

Define the collection

$$\boldsymbol{S} = \left\{ \begin{pmatrix} -2 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 \\ \alpha \end{pmatrix} \right\}.$$

Let $\boldsymbol{v} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$.

(a) When $\alpha = -15$, is $\boldsymbol{v}$ in span($\boldsymbol{S}$)?

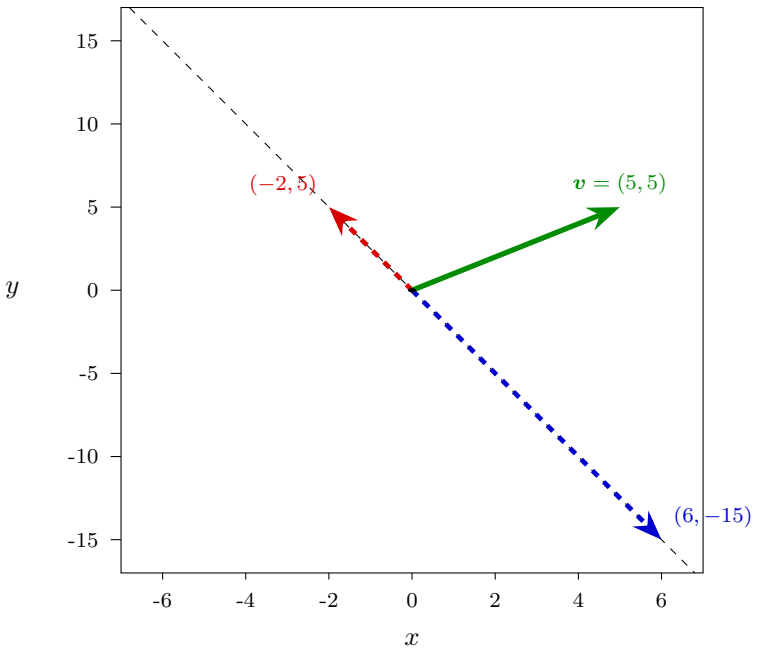
(b) When $\alpha = 0$, is $\boldsymbol{v}$ in span($\boldsymbol{S}$)?

(c) When $\alpha = 0$, is $\boldsymbol{S}$ a basis for all of $\mathbb{R}^2$?

Solutions

(a) Membership when $\alpha = -15$. Recall from Lecture 24 that $\alpha = -15$ makes the two vectors linearly dependent; they both sit on the line $y = -\frac{5}{2}x$ through the origin, so span($\boldsymbol{S}$) is that single line.

Because $\begin{pmatrix} -2 \\ 5 \end{pmatrix}$ and $\begin{pmatrix} 6 \\ -15 \end{pmatrix}$ are linearly *dependent*, we can proceed with just one vector in the collection. Take $\boldsymbol{S} = \left\{ \begin{pmatrix} -2 \\ 5 \end{pmatrix} \right\}$. Then we need $\begin{pmatrix} 5 \\ 5 \end{pmatrix} = \lambda \begin{pmatrix} -2 \\ 5 \end{pmatrix}$ for some $\lambda \in \mathbb{R}$. The second entry forces $\lambda = 1$; but with $\lambda = 1$ the first entry gives $-2 \neq 5$, so the two equations cannot be satisfied simultaneously. No such λ exists, so $\begin{pmatrix} 5 \\ 5 \end{pmatrix} \neq \lambda \begin{pmatrix} -2 \\ 5 \end{pmatrix}$ for any λ. Therefore $\boldsymbol{v} \notin$ span($\boldsymbol{S}$).



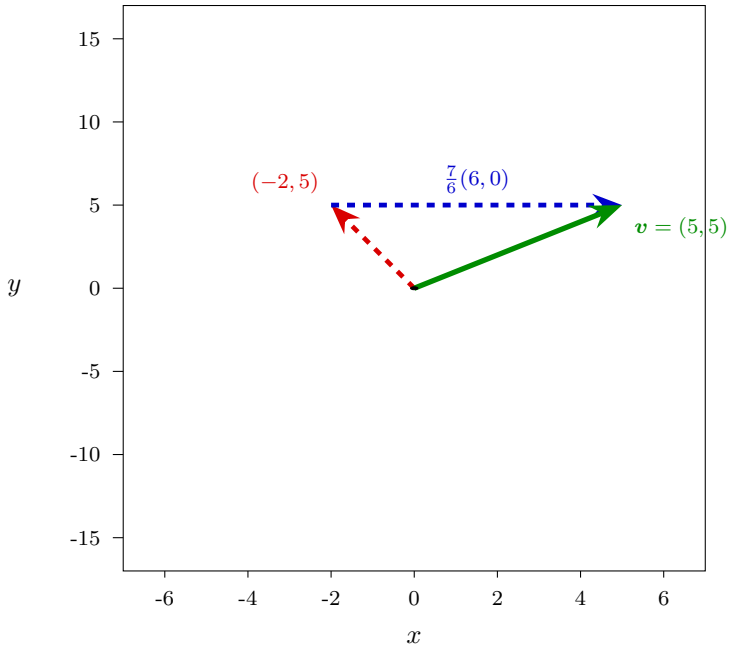
(b) Membership when $\alpha = 0$. Now $\boldsymbol{S} = \left\{ \begin{pmatrix} -2 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 \\ 0 \end{pmatrix} \right\}$ is linearly independent, so $\text{span}(\boldsymbol{S}) = \mathbb{R}^2$ and any vector is in the span. Explicitly, seek λ_1, λ_2 with

$$\begin{pmatrix} 5 \\ 5 \end{pmatrix} = \lambda_1 \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \lambda_2 \begin{pmatrix} 6 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} 5 \\ 5 \end{pmatrix} = \begin{pmatrix} -2\lambda_1 \\ 5\lambda_1 \end{pmatrix} + \begin{pmatrix} 6\lambda_2 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} 5 \\ 5 \end{pmatrix} = \begin{pmatrix} -2\lambda_1 + 6\lambda_2 \\ 5\lambda_1 \end{pmatrix}$$

$$5\lambda_1 = 5$$
$$\lambda_1 = 1$$

$$-2\lambda_1 + 6\lambda_2 = 5$$
$$-2(1) + 6\lambda_2 = 5$$
$$6\lambda_2 = 7$$
$$\lambda_2 = \tfrac{7}{6}$$

So $\boldsymbol{v} = \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \frac{7}{6} \begin{pmatrix} 6 \\ 0 \end{pmatrix} \in \text{span}(\boldsymbol{S})$. Geometrically, walk once along $(-2, 5)$ to get to $(-2, 5)$, then $\frac{7}{6}$ times along $(6, 0)$ to land on $(5, 5)$:



(c) Is $\boldsymbol{S}$ a basis for $\mathbb{R}^2$ when $\alpha = 0$? Need to show **(1)** $\boldsymbol{S}$ spans all of $\mathbb{R}^2$, and **(2)** $\boldsymbol{S}$ is linearly independent.

(1) Spanning. We can reach any $(x, y) \in \mathbb{R}^2$ by solving

$$\lambda_1 \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \lambda_2 \begin{pmatrix} 6 \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$5\lambda_1 = y$$

$$\lambda_1 = \tfrac{y}{5}$$

$$-2\lambda_1 + 6\lambda_2 = x$$

$$-2\left(\tfrac{y}{5}\right) + 6\lambda_2 = x$$

$$6\lambda_2 = x + \tfrac{2}{5}y$$

$$\lambda_2 = \tfrac{1}{6}x + \tfrac{1}{15}y$$

$$x = -2\lambda_1 + 6\lambda_2$$

$$y = 5\lambda_1$$

So for any $(x, y) \in \mathbb{R}^2$,

$$\tfrac{y}{5} \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \left(\tfrac{1}{6}x + \tfrac{1}{15}y\right) \begin{pmatrix} 6 \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix},$$

which demonstrates that $\boldsymbol{S}$ spans all of $\mathbb{R}^2$. (✓)

(2) Linear independence. We need $\lambda_1 \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \lambda_2 \begin{pmatrix} 6 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ to force $\lambda_1 = \lambda_2 = 0$.

The component equations are

$$0 = -2\lambda_1 + 6\lambda_2 \implies \lambda_1 = 3\lambda_2, \qquad 0 = 5\lambda_1 \implies \lambda_1 = 0.$$

Together these give $\lambda_1 = 0$ and hence $\lambda_2 = 0$. So $\lambda_1 \begin{pmatrix} -2 \\ 5 \end{pmatrix} + \lambda_2 \begin{pmatrix} 6 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ only if $\lambda_1 = \lambda_2 = 0$, i.e. the collection is linearly independent. (✓)

Conclusion. Both criteria are satisfied, so *yes* — when $\alpha = 0$, $\boldsymbol{S}$ is a basis for $\mathbb{R}^2$.

Matrices: Notation, Types, and Operations

Lecture 26 • *UVA Stats*

April 24, 2026

Last lecture wrapped up vectors with span and basis. Today we move one step further and introduce the matrix *— a rectangular array of numbers that stacks row- or column-vectors together. We cover the notation, the canonical families (square, diagonal, identity, symmetric, triangular, dense versus sparse), the scalar summaries (diagonal, trace, Frobenius norm), and every arithmetic operation that will matter in later courses: transposition, addition and subtraction, scalar multiplication, the proper matrix product and its identity/diagonal/orthogonal variants, the element-wise Hadamard product, and the Frobenius dot product. A consolidation practice problem at the end exercises the full toolbox on a single pair of matrices.*

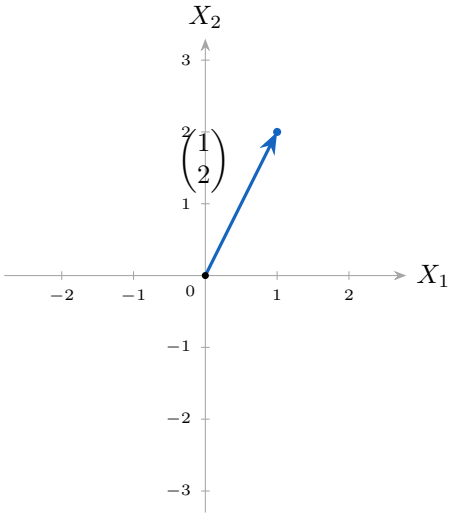
PART 1

Motivation

A matrix is the natural next step after a vector: once we allow ourselves to stack several vectors side by side, we get a rectangular object that can carry *two* kinds of structure at once — one per row and one per column. This is the data structure behind tabular datasets, systems of linear equations, and every design matrix you will ever see in a statistical model.

Recall: what a vector looked like

Last lecture a vector was an ordered list of values, drawn geometrically as an arrow from the origin with a given direction and magnitude. For example, the vector $\begin{pmatrix}1\\2\end{pmatrix}$ — which we could also write as $\langle 1, 2\rangle$ — lives in the 2-D plane as the arrow from $(0, 0)$ to $(1, 2)$:



From vectors to matrices

Term — Matrix

A *matrix* is a rectangular array of numbers, symbols, or expressions arranged in rows and columns. Think of a matrix as a system of stacked row vectors or adjacent column vectors. Matrices are typically represented using boldface capital letters.

As archetypes, we will use the following two matrices throughout the lecture:

$$\boldsymbol{A} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 5 & -1 & 1 \end{bmatrix}.$$

Some common uses of matrices include:

- storing data in rows (observations) and columns (variables);
- solving systems of linear equations;
- representing the regressors in a statistical model.

Key Idea 79

A matrix is just a *bundle of vectors*. Every operation we meet in Part 2 — transpose, trace, norm, addition, scalar and matrix multiplication, Hadamard and Frobenius products — is either an element-wise lift of a scalar operation or a pattern of dot products between the constituent rows and columns.

PART 2

Methods and Examples

1. Notation for Matrices

Term — Size of a matrix

The size of a matrix is given by the pair (M, N), where M is the number of rows and N is the number of columns. If every entry belongs to the same field (typically $\mathbb{R}$), we write $\boldsymbol{A} \in \mathbb{R}^{M \times N}$.

For our two archetypes,

$$\boldsymbol{A} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \quad \text{is a } 3 \times 3 \text{ matrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 5 & -1 & 1 \end{bmatrix} \quad \text{is a } 2 \times 3 \text{ matrix}.$$

Individual entries are named with double-subscripted lowercase letters: $a_{i,j}$ is the entry in row i, column j of $\boldsymbol{A}$. For instance,

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & \mathbf{2} \\ \mathbf{5} & -1 & 1 \end{bmatrix}, \qquad b_{2,1} = 5, \qquad b_{1,3} = 2.$$

2. Transpose Operation

Term — Transpose

The *transpose* of a matrix flips its rows and columns: row i of $\boldsymbol{A}$ becomes column i of $\boldsymbol{A}^T$. Equivalently, $(\boldsymbol{A}^T)_{i,j} = a_{j,i}$.

Worked example

Taking $\boldsymbol{B}$ from above,

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & \mathbf{2} \\ \mathbf{5} & -1 & 1 \end{bmatrix} \qquad \Longrightarrow \qquad \boldsymbol{B}^T = \begin{bmatrix} 1 & \mathbf{5} \\ 0 & -1 \\ \mathbf{2} & 1 \end{bmatrix}.$$

The entries that were in positions $b_{2,1}$ and $b_{1,3}$ now live at $b^T_{1,2} = 5$ and $b^T_{3,1} = 2$ — exactly as the definition promises.

3. Common Types of Matrices

Several families of matrices appear so frequently they have their own names. In each case the defining property is a statement about the positions of the zero entries, the symmetry across the main diagonal, or the squareness of the shape.

Term — Square matrix

A matrix with the same number of rows and columns: size $N \times N$.

Example: $\boldsymbol{A} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}$ is a 3×3 square matrix.

Term — Zero matrix

A matrix whose entries are all zero; usually written $\mathbf{0}$ or $\mathbf{0}_N$ for the square $N \times N$ version.

Example: $\mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{0}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

Term — Diagonal matrix

A square matrix in which every off-diagonal entry is zero — that is, $a_{ij} = 0$ whenever $i \neq j$.

Examples:

$$\boldsymbol{C} = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}, \qquad \boldsymbol{D} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Term — Identity matrix

A square diagonal matrix whose main-diagonal entries are all ones; denoted $\boldsymbol{I}$, or $\boldsymbol{I}_N$ to emphasise the size.

Examples:

$$\boldsymbol{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Term — Symmetric matrix

A square matrix that is mirrored across the main diagonal, i.e. $\boldsymbol{A} = \boldsymbol{A}^T$.

Examples:

$$\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}, \qquad \boldsymbol{G} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 3 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

Term — Triangular matrix

A matrix in which all the entries either above or below the main diagonal are zero. An *upper-triangular* matrix has zeros below the diagonal; a *lower-triangular* matrix has zeros above.

Examples of upper-triangular matrices:

$$\boldsymbol{H} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ 0 & h_{2,2} \end{bmatrix}, \qquad \boldsymbol{J} = \begin{bmatrix} 6 & 0 & 1 \\ 0 & -2 & 5 \end{bmatrix}.$$

Examples of lower-triangular matrices:

$$\boldsymbol{K} = \begin{bmatrix} k_{1,1} & 0 \\ k_{2,1} & k_{2,2} \end{bmatrix}, \qquad \boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & -3 & 0 \end{bmatrix}.$$

Term — Dense vs. sparse matrix

A *dense* matrix is one whose entries are mostly non-zero; a *sparse* matrix is one whose entries are mostly zero.

Examples:

$$\boldsymbol{M} = \begin{bmatrix} 5 & 8 & 9 \\ 2 & 1 & 0 \\ 3 & 6 & 4 \end{bmatrix} \text{ (dense)}, \qquad \boldsymbol{N} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 \end{bmatrix} \text{ (sparse)}.$$

4. Diagonal and Trace

Term — Diagonal of a matrix

The *diagonal* of a matrix is the vector formed by reading off the entries on its main diagonal.

Term — Trace of a matrix

The *trace* of a matrix is the sum of the entries on its main diagonal. It is only defined for square matrices — the diagonal is defined for any shape, but trace is not.

Worked example

For $\boldsymbol{M} = \begin{bmatrix} 5 & 8 & 9 \\ 2 & 1 & 0 \\ 3 & 6 & 4 \end{bmatrix}$,

$$\operatorname{diag}(\boldsymbol{M}) = \begin{pmatrix} 5 \\ 1 \\ 4 \end{pmatrix}, \qquad \operatorname{trace}(\boldsymbol{M}) = 5 + 1 + 4 = 10.$$

For the non-square $\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 5 & -1 & 1 \end{bmatrix}$, $\operatorname{diag}(\boldsymbol{B}) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, but $\operatorname{trace}(\boldsymbol{B})$ is not defined.

5. Matrix Norms

Term — Frobenius matrix norm

A scalar that quantifies the magnitude of a matrix:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(a_{i,j}\right)^2}.$$

Equivalent to the Euclidean norm of the vectorised matrix.

Once we have a norm, we immediately get a distance between any two matrices of the same size:

$$\|\boldsymbol{A}-\boldsymbol{B}\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(a_{i,j}-b_{i,j}\right)^2}.$$

Worked example

For $\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}$,

$$\|\boldsymbol{F}\|_F = \sqrt{0^2+(-2)^2+(-2)^2+3^2} = \sqrt{0+4+4+9} = \sqrt{17}.$$

And for $\boldsymbol{F}$ and $\boldsymbol{C} = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$,

$$\begin{aligned}\|\boldsymbol{F}-\boldsymbol{C}\|_F &= \sqrt{(0-3)^2+(-2-0)^2+(-2-0)^2+(3-(-2))^2} \\ &= \sqrt{(-3)^2+(-2)^2+(-2)^2+5^2} = \sqrt{9+4+4+25} = \sqrt{42}.\end{aligned}$$

6. Matrix Addition

Term — Matrix addition

Add two matrices by adding their corresponding entries. Both matrices must have the same size.

In full generality,

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} + \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} = \begin{bmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} & a_{1,3}+b_{1,3} \\ a_{2,1}+b_{2,1} & a_{2,2}+b_{2,2} & a_{2,3}+b_{2,3} \\ a_{3,1}+b_{3,1} & a_{3,2}+b_{3,2} & a_{3,3}+b_{3,3} \end{bmatrix}.$$

Worked example

For $\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$,

$$\boldsymbol{F} + \boldsymbol{C} = \begin{bmatrix} 0+3 & -2+0 \\ -2+0 & 3+(-2) \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ -2 & 1 \end{bmatrix}.$$

7. Matrix Subtraction

Term — Matrix subtraction

Subtract two matrices by subtracting the second matrix's entry from the corresponding entry in the first. Both matrices must have the same size.

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} - \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} = \begin{bmatrix} a_{1,1}-b_{1,1} & a_{1,2}-b_{1,2} & a_{1,3}-b_{1,3} \\ a_{2,1}-b_{2,1} & a_{2,2}-b_{2,2} & a_{2,3}-b_{2,3} \\ a_{3,1}-b_{3,1} & a_{3,2}-b_{3,2} & a_{3,3}-b_{3,3} \end{bmatrix}.$$

Worked example

For the same $\boldsymbol{F}$ and $\boldsymbol{C}$ as above,

$$\boldsymbol{F} - \boldsymbol{C} = \begin{bmatrix} 0-3 & -2-0 \\ -2-0 & 3-(-2) \end{bmatrix} = \begin{bmatrix} -3 & -2 \\ -2 & 5 \end{bmatrix}.$$

8. Matrix Scalar Multiplication

Term — Scalar multiplication

Multiply a matrix by a scalar $\lambda \in \mathbb{R}$ by multiplying every entry by λ. The operation commutes: $\lambda \boldsymbol{A} = \boldsymbol{A}\lambda$.

$$\lambda \boldsymbol{A} = \lambda \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} = \begin{bmatrix} \lambda a_{1,1} & \lambda a_{1,2} & \lambda a_{1,3} \\ \lambda a_{2,1} & \lambda a_{2,2} & \lambda a_{2,3} \\ \lambda a_{3,1} & \lambda a_{3,2} & \lambda a_{3,3} \end{bmatrix}.$$

Worked example

For $\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}$,

$$5\boldsymbol{F} = \begin{bmatrix} 5 \cdot 0 & 5 \cdot (-2) \\ 5 \cdot (-2) & 5 \cdot 3 \end{bmatrix} = \begin{bmatrix} 0 & -10 \\ -10 & 15 \end{bmatrix}.$$

9. Matrix Multiplication

Matrix multiplication is the centrepiece of matrix algebra — and the point where size *really* matters.

Term — Matrix product (size rule)

For the product $\boldsymbol{AB}$ to be defined, the number of *columns* of $\boldsymbol{A}$ must equal the number of *rows* of $\boldsymbol{B}$. If $\boldsymbol{A}$ is $M \times N$ and $\boldsymbol{B}$ is $N \times K$, then $\boldsymbol{AB}$ is an $M \times K$ matrix.

Worked example (size check)

For $\boldsymbol{D} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (3×3) and $\boldsymbol{E} = \begin{bmatrix} 6.7 & 0 & 0 \\ 0 & 4.2 & 0 \end{bmatrix}$ (2×3):

- $\boldsymbol{E}_{2\times 3}\,\boldsymbol{D}_{3\times 3}$ works; the result is 2×3.
- $\boldsymbol{D}_{3\times 3}\,\boldsymbol{E}_{2\times 3}$ does not work (inner dimensions $3 \neq 2$).
- $\boldsymbol{D}_{3\times 3}\,\boldsymbol{E}^{T}_{3\times 2}$ works; the result is 3×2.

Term — Matrix product (entry rule)

If the sizes align, then $(ab)_{i,j}$, the entry in row i, column j of $\boldsymbol{AB}$, is the dot product of row i of $\boldsymbol{A}$ with column j of $\boldsymbol{B}$:

$$(ab)_{i,j} \;=\; \boldsymbol{a}_{i,\cdot} \cdot \boldsymbol{b}_{\cdot,j}.$$

Worked example (the product $\boldsymbol{FC}$)

Let $\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$.

$$(\boldsymbol{FC})_{1,1} = [0,\,-2] \cdot \begin{bmatrix} 3 \\ 0 \end{bmatrix} = (0)(3) + (-2)(0) = 0,$$

$$(\boldsymbol{FC})_{1,2} = [0,\,-2] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = (0)(0) + (-2)(-2) = 4,$$

$$(\boldsymbol{FC})_{2,1} = [-2,\,3] \cdot \begin{bmatrix} 3 \\ 0 \end{bmatrix} = (-2)(3) + (3)(0) = -6,$$

$$(\boldsymbol{FC})_{2,2} = [-2,\,3] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = (-2)(0) + (3)(-2) = -6.$$

So $\boldsymbol{FC} = \begin{bmatrix} 0 & 4 \\ -6 & -6 \end{bmatrix}$.

Worked example (the product CF)

Now swap the order:

$$(\boldsymbol{CF})_{1,1} = [3,\,0] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = 0,$$
$$(\boldsymbol{CF})_{1,2} = [3,\,0] \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = -6,$$
$$(\boldsymbol{CF})_{2,1} = [0,\,-2] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = 4,$$
$$(\boldsymbol{CF})_{2,2} = [0,\,-2] \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = -6.$$

So $\boldsymbol{CF} = \begin{bmatrix} 0 & -6 \\ 4 & -6 \end{bmatrix}$.

Notice that $\boldsymbol{CF} \neq \boldsymbol{FC}$: matrix multiplication is *not* commutative.

Worked example (matrix times vector)

Vectors are just $N \times 1$ matrices, so the same size-rule applies. For $\boldsymbol{O} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$ and $\boldsymbol{p} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$,

$$\boldsymbol{Op} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \end{bmatrix} = \begin{bmatrix} (4)(5) + (2)(2) \\ (1)(5) + (3)(2) \end{bmatrix} = \begin{bmatrix} 20 + 4 \\ 5 + 6 \end{bmatrix} = \begin{bmatrix} 24 \\ 11 \end{bmatrix}.$$

Algebraic properties

When the sizes line up, matrix multiplication satisfies:

- **Associativity:** $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$.
- **Distributivity over addition:** $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$.
- **Not commutative:** in general, $\boldsymbol{AB} \neq \boldsymbol{BA}$.

Multiplication by the identity

For any $M \times N$ matrix $\boldsymbol{A}$,

$$\boldsymbol{A}_{M\times N}\,\boldsymbol{I}_N \;=\; \boldsymbol{I}_M\,\boldsymbol{A}_{M\times N} \;=\; \boldsymbol{A}_{M\times N}.$$

So the identity matrix plays the same role as the scalar 1. As a sanity check with $\boldsymbol{F}$ and $\boldsymbol{I}_2$,

$$(\boldsymbol{FI})_{1,1} = [0,\,-2] \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 0, \qquad (\boldsymbol{FI})_{1,2} = [0,\,-2] \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = -2,$$
$$(\boldsymbol{FI})_{2,1} = [-2,\,3] \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -2, \qquad (\boldsymbol{FI})_{2,2} = [-2,\,3] \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3,$$

so $\boldsymbol{FI}_2 = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix} = \boldsymbol{F}$.

Multiplication by a diagonal matrix

A diagonal matrix acts as a simple *rescaler*. For any matrix $\boldsymbol{A}$ and any diagonal matrix $\boldsymbol{D}$:

- $\boldsymbol{DA}$ scales the *rows* of $\boldsymbol{A}$ by the diagonal entries of $\boldsymbol{D}$.
- $\boldsymbol{AD}$ scales the *columns* of $\boldsymbol{A}$ by the diagonal entries of $\boldsymbol{D}$.

We already saw this implicitly: with $\boldsymbol{C} = \text{diag}(3, -2)$,

$$\boldsymbol{CF} = \begin{bmatrix} 0 & -6 \\ 4 & -6 \end{bmatrix} \quad \text{(rows of } \boldsymbol{F} \text{ scaled by 3 and } -2\text{)},$$

$$\boldsymbol{FC} = \begin{bmatrix} 0 & 4 \\ -6 & -6 \end{bmatrix} \quad \text{(columns of } \boldsymbol{F} \text{ scaled by 3 and } -2\text{)}.$$

Key Idea 80

The matrix product $\boldsymbol{AB}$ is just a grid of dot products: row i of $\boldsymbol{A}$ meets column j of $\boldsymbol{B}$ to deliver entry (i, j). Everything else — associativity, the role of the identity, the scaling behaviour of diagonal matrices — follows from that single recipe.

10. Orthogonal Matrices

Term — Orthogonal matrix

An $N \times N$ square matrix $\boldsymbol{Q}$ is *orthogonal* if $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_N$. Equivalently:
- the columns of $\boldsymbol{Q}$ are pairwise perpendicular — $\boldsymbol{q}_{\cdot,i} \cdot \boldsymbol{q}_{\cdot,j} = 0$ for $i \neq j$; and
- each column of $\boldsymbol{Q}$ has unit length — $\|\boldsymbol{q}_{\cdot,i}\| = 1$.

The identity matrix is orthogonal.

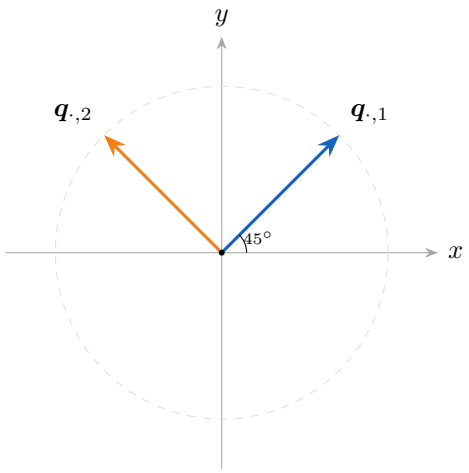
Worked example

Let $\boldsymbol{Q} = \begin{bmatrix} \sqrt{0.5} & -\sqrt{0.5} \\ \sqrt{0.5} & \sqrt{0.5} \end{bmatrix}$. Check the two conditions:

$$\begin{aligned} \boldsymbol{q}_{\cdot,1} \cdot \boldsymbol{q}_{\cdot,2} &= \left(\sqrt{0.5}\right)\left(-\sqrt{0.5}\right) + \left(\sqrt{0.5}\right)\left(\sqrt{0.5}\right) = -0.5 + 0.5 = 0, \\ \|\boldsymbol{q}_{\cdot,1}\| &= \sqrt{\left(\sqrt{0.5}\right)^2 + \left(\sqrt{0.5}\right)^2} = \sqrt{0.5 + 0.5} = 1, \\ \|\boldsymbol{q}_{\cdot,2}\| &= \sqrt{\left(-\sqrt{0.5}\right)^2 + \left(\sqrt{0.5}\right)^2} = \sqrt{0.5 + 0.5} = 1. \end{aligned}$$

Both conditions hold, so $\boldsymbol{Q}$ is orthogonal.

Geometrically, the columns of $\boldsymbol{Q}$ are the unit vectors at 45° and 135° in the plane:



11. Hadamard Multiplication

Term — Hadamard product

The *Hadamard* (element-wise) product $\boldsymbol{A} \odot \boldsymbol{B}$ multiplies each entry of $\boldsymbol{A}$ by the corresponding entry of $\boldsymbol{B}$. Both matrices must have the same size.

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \odot \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} = \begin{bmatrix} a_{1,1}b_{1,1} & a_{1,2}b_{1,2} & a_{1,3}b_{1,3} \\ a_{2,1}b_{2,1} & a_{2,2}b_{2,2} & a_{2,3}b_{2,3} \\ a_{3,1}b_{3,1} & a_{3,2}b_{3,2} & a_{3,3}b_{3,3} \end{bmatrix}.$$

Worked example

For $\boldsymbol{F}$ and $\boldsymbol{C}$ as before,

$$\boldsymbol{F} \odot \boldsymbol{C} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix} \odot \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} (0)(3) & (-2)(0) \\ (-2)(0) & (3)(-2) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -6 \end{bmatrix}.$$

Note how radically different this is from $\boldsymbol{FC}$: Hadamard multiplication forgets about the row/column structure entirely.

12. Frobenius Dot Product

Term — Vectorising a matrix

The *vectorisation* of an $M \times N$ matrix stacks its columns on top of one another to form a single column vector of length MN.

$$\operatorname{vec}\left(\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}\right) = \begin{bmatrix} a \\ d \\ b \\ e \\ c \\ f \end{bmatrix}.$$

Term — Frobenius dot product

The *Frobenius dot product* of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size is the ordinary dot product of their vectorisations:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{vec}(\boldsymbol{A}) \cdot \operatorname{vec}(\boldsymbol{B}).$$

Equivalently,

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{trace}\big(\boldsymbol{A}^T \boldsymbol{B}\big) = \sum_{i,j} a_{i,j}\, b_{i,j}$$

— the sum of all entries of $\boldsymbol{A} \odot \boldsymbol{B}$.

The two expressions are equal because the trace of $\boldsymbol{A}^T\boldsymbol{B}$ sums the diagonal entries $\sum_i (\boldsymbol{A}^T\boldsymbol{B})_{ii} = \sum_i \sum_j a_{j,i}\, b_{j,i} = \sum_{i,j} a_{i,j}\, b_{i,j}$, exactly the entrywise inner product.

Worked example

For $\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}$ and $\boldsymbol{O} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$,

$$\operatorname{vec}(\boldsymbol{F}) = \begin{bmatrix} 0 \\ -2 \\ -2 \\ 3 \end{bmatrix}, \qquad \operatorname{vec}(\boldsymbol{O}) = \begin{bmatrix} 4 \\ 1 \\ 2 \\ 3 \end{bmatrix}.$$

Therefore

$$\begin{aligned} \langle \boldsymbol{F}, \boldsymbol{O} \rangle_F &= (0)(4) + (-2)(1) + (-2)(2) + (3)(3) \\ &= 0 - 2 - 4 + 9 \\ &= 3. \end{aligned}$$

PART 3

Practice Problems

Having met every operation individually, let us exercise them all on a single pair of matrices.

Example #1. Let

$$\boldsymbol{F} = \begin{bmatrix} 0 & -2 \\ -2 & 3 \end{bmatrix}, \qquad \boldsymbol{O} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}.$$

(a) State the size of each matrix and identify which of square, diagonal, symmetric, upper-triangular, and lower-triangular families they belong to.

(b) Compute the transpose $\boldsymbol{O}^T$ and the sum $\boldsymbol{F} + \boldsymbol{O}$.

(c) Compute the scalar multiple $3\boldsymbol{O}$ and the Frobenius norm $\|\boldsymbol{F}\|_F$.

(d) Compute $\text{diag}(\boldsymbol{O})$ and $\text{trace}(\boldsymbol{O})$.

(e) Compute the matrix product $\boldsymbol{FO}$ and check whether $\boldsymbol{FO} = \boldsymbol{OF}$.

(f) Compute the Hadamard product $\boldsymbol{F} \odot \boldsymbol{O}$ and the Frobenius dot product $\langle \boldsymbol{F}, \boldsymbol{O} \rangle_F$. Verify that $\langle \boldsymbol{F}, \boldsymbol{O} \rangle_F$ equals the sum of all entries of $\boldsymbol{F} \odot \boldsymbol{O}$.

Solutions

(a) Sizes and types. Both $\boldsymbol{F}$ and $\boldsymbol{O}$ are 2×2 matrices, hence both are *square*. Neither is diagonal (each has non-zero off-diagonal entries). $\boldsymbol{F}$ is *symmetric* because its off-diagonal entries agree: $f_{1,2} = f_{2,1} = -2$, so $\boldsymbol{F} = \boldsymbol{F}^T$. $\boldsymbol{O}$ is *not* symmetric since $o_{1,2} = 2 \neq 1 = o_{2,1}$. Neither matrix is triangular (each has a non-zero entry both above and below the main diagonal).

(b) Transpose and sum. Flip rows and columns of $\boldsymbol{O}$:

$$\boldsymbol{O}^T = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}.$$

Add entry-wise:

$$\boldsymbol{F} + \boldsymbol{O} = \begin{bmatrix} 0+4 & -2+2 \\ -2+1 & 3+3 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ -1 & 6 \end{bmatrix}.$$

(c) Scalar multiple and Frobenius norm. Multiply every entry of $\boldsymbol{O}$ by 3:

$$3\boldsymbol{O} = \begin{bmatrix} 3 \cdot 4 & 3 \cdot 2 \\ 3 \cdot 1 & 3 \cdot 3 \end{bmatrix} = \begin{bmatrix} 12 & 6 \\ 3 & 9 \end{bmatrix}.$$

For the Frobenius norm of $\boldsymbol{F}$, square all entries and take the square root of the sum:

$$\|\boldsymbol{F}\|_F = \sqrt{0^2 + (-2)^2 + (-2)^2 + 3^2} = \sqrt{0 + 4 + 4 + 9} = \sqrt{17}.$$

(d) Diagonal and trace of $\boldsymbol{O}$. Read off the main-diagonal entries:

$$\operatorname{diag}(\boldsymbol{O}) = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \qquad \operatorname{trace}(\boldsymbol{O}) = 4 + 3 = 7.$$

(e) Matrix product $\boldsymbol{FO}$ and commutativity check. Using the dot-product rule on every row–column pair:

$$(\boldsymbol{FO})_{1,1} = [0,\, -2] \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = (0)(4) + (-2)(1) = -2,$$
$$(\boldsymbol{FO})_{1,2} = [0,\, -2] \cdot \begin{bmatrix} 2 \\ 3 \end{bmatrix} = (0)(2) + (-2)(3) = -6,$$
$$(\boldsymbol{FO})_{2,1} = [-2,\, 3] \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = (-2)(4) + (3)(1) = -5,$$
$$(\boldsymbol{FO})_{2,2} = [-2,\, 3] \cdot \begin{bmatrix} 2 \\ 3 \end{bmatrix} = (-2)(2) + (3)(3) = 5,$$

so

$$\boldsymbol{FO} = \begin{bmatrix} -2 & -6 \\ -5 & 5 \end{bmatrix}.$$

Computing the other order,

$$(\boldsymbol{OF})_{1,1} = [4,\, 2] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = -4,$$
$$(\boldsymbol{OF})_{1,2} = [4,\, 2] \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = -8 + 6 = -2,$$
$$(\boldsymbol{OF})_{2,1} = [1,\, 3] \cdot \begin{bmatrix} 0 \\ -2 \end{bmatrix} = -6,$$
$$(\boldsymbol{OF})_{2,2} = [1,\, 3] \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = -2 + 9 = 7,$$

giving

$$\boldsymbol{OF} = \begin{bmatrix} -4 & -2 \\ -6 & 7 \end{bmatrix}.$$

Since $\boldsymbol{FO} \neq \boldsymbol{OF}$, matrix multiplication *fails* to commute here ($\times$), as it generally does.

(f) Hadamard and Frobenius dot products. Element-wise multiplication gives

$$\boldsymbol{F} \odot \boldsymbol{O} = \begin{bmatrix} (0)(4) & (-2)(2) \\ (-2)(1) & (3)(3) \end{bmatrix} = \begin{bmatrix} 0 & -4 \\ -2 & 9 \end{bmatrix}.$$

The Frobenius dot product is the sum of the entries of the Hadamard product (equivalently, the dot product of the vectorisations):

$$\langle \boldsymbol{F}, \boldsymbol{O} \rangle_F = 0 + (-4) + (-2) + 9 = 3.$$

As a consistency check,

$$\sum_{i,j} \big(\boldsymbol{F} \odot \boldsymbol{O}\big)_{i,j} = 0 + (-4) + (-2) + 9 = 3 = \langle \boldsymbol{F}, \boldsymbol{O} \rangle_F \quad (\checkmark)$$

— the Hadamard sum identity confirms our answer.

Rank, Determinant, and Inverse

Lecture 27 • UVA Stats

April 24, 2026

Having met the matrix and its arithmetic last lecture, we now ask three sharper questions about it. How many genuinely independent directions does it carry — its rank*? What scalar summarises whether those directions collapse — its* determinant*? And when the matrix is square and non-degenerate, what multiplicative undo — its* inverse *— solves the linear system $\boldsymbol{Ax} = \boldsymbol{b}$? A consolidation practice problem at the end asks for rank, determinant, and inverse of a single 2×2 matrix $\boldsymbol{G}$.*

PART 1

Motivation

Where we are

Last lecture ended with a full toolbox for manipulating matrices: transposition, addition, scalar multiplication, matrix multiplication, the Hadamard product, and the Frobenius dot product. As a reminder, a matrix is just a rectangular array of numbers, interpreted either as a column of row vectors or a row of column vectors:

$$\boldsymbol{A} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 5 & -1 & 1 \end{bmatrix}.$$

Three new questions

Today we move beyond arithmetic and ask three *structural* questions about a matrix:

- **Rank.** How many linearly independent directions are really captured by the rows (or columns) of $\boldsymbol{A}$? Are any of them redundant?
- **Determinant.** Is there a single scalar summary telling us whether those directions collapse onto one another — i.e. whether the matrix is singular?
- **Inverse.** When $\boldsymbol{A}$ is square and non-degenerate, can we find a matrix $\boldsymbol{A}^{-1}$ that "undoes" it, so that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$?

These three notions are tightly linked, as we will see: a square matrix is invertible if and only if it is full rank, which is if and only if its determinant is non-zero.

Term — Rank of a matrix

The *rank* of a matrix is the largest number of its columns that can form a linearly independent set. Equivalently, it is the largest number of rows that form a linearly independent set, or — geometrically — the dimension of the subspace spanned by the columns (or rows) of the matrix.

Key Idea 81

Rank counts the independent directions; the determinant measures whether those directions collapse; the inverse exists exactly when they do not. Everything in this lecture — from the 2×2 formula through Sarrus' rule through the *shifting* trick — is built around this single chain of equivalences.

PART 2

Methods and Examples

1. Matrix Rank

For an $M \times N$ matrix $\boldsymbol{A}$, the rank always sits inside the interval

$$0 \;\le\; \text{rank}(\boldsymbol{A}) \;\le\; \min\{M, N\}.$$

- When $\text{rank}(\boldsymbol{A}) = \min\{M, N\}$, we say $\boldsymbol{A}$ is *full rank*.
- When $\text{rank}(\boldsymbol{A}) < \min\{M, N\}$, we say $\boldsymbol{A}$ is *reduced rank* or *low rank*. (For a *square* matrix in this situation we also say $\boldsymbol{A}$ is *singular*.)

Why we care

A matrix can only be *inverted* if it is full rank, and many downstream operations in regression and PCA require full-rank inputs. When a matrix is reduced rank but we still need to invert it, we can sometimes rescue it by *shifting* (Section 2).

Worked example

Let $\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 5 & -1 & 1 \end{bmatrix}$. Its three columns are

$$\boldsymbol{b}_{\cdot,1} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \quad \boldsymbol{b}_{\cdot,2} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad \boldsymbol{b}_{\cdot,3} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

The three columns together are *linearly dependent*, because

$$\mathbf{0} = 2\begin{bmatrix} 1 \\ 5 \end{bmatrix} + 9\begin{bmatrix} 0 \\ -1 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 - 2 \\ 10 - 9 - 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Drop any one of them — say the third — and the remaining pair $\left\{ \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$ is linearly independent. Therefore $\text{rank}(\boldsymbol{B}) = 2$. Since $\min\{2, 3\} = 2$, $\boldsymbol{B}$ is in fact full rank.

2. Shifting a Matrix to Full Rank

Term — Shifting (regularisation / smoothing)

Shifting a square matrix $\boldsymbol{A}$ means adding a small multiple of the identity to it:

$$\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \lambda \boldsymbol{I}.$$

The off-diagonal entries are unchanged; only the diagonal gets a small bump of size λ. In statistics and machine learning this same operation is called *regularisation* or *matrix smoothing*.

Worked example

Consider

$$\boldsymbol{C} = \begin{bmatrix} 1 & 3 & -19 \\ 5 & -7 & 59 \\ -5 & 2 & -24 \end{bmatrix}.$$

Its columns satisfy

$$\mathbf{0} = 2\begin{bmatrix} 1 \\ 5 \\ -5 \end{bmatrix} - 7\begin{bmatrix} 3 \\ -7 \\ 2 \end{bmatrix} - \begin{bmatrix} -19 \\ 59 \\ -24 \end{bmatrix} = \begin{bmatrix} 2 - 21 + 19 \\ 10 + 49 - 59 \\ -10 - 14 + 24 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

so the three columns are linearly dependent, hence $\operatorname{rank}(\boldsymbol{C}) = 2 < 3 = \min\{3, 3\}$. We say $\boldsymbol{C}$ is reduced rank.

Shift $\boldsymbol{C}$ by $\lambda = 0.01$:

$$\widetilde{\boldsymbol{C}} = \boldsymbol{C} + 0.01\,\boldsymbol{I} = \begin{bmatrix} 1 & 3 & -19 \\ 5 & -7 & 59 \\ -5 & 2 & -24 \end{bmatrix} + \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.01 \end{bmatrix} = \begin{bmatrix} 1.01 & 3 & -19 \\ 5 & -6.99 & 59 \\ -5 & 2 & -23.99 \end{bmatrix}.$$

$\widetilde{\boldsymbol{C}}$ is barely changed — only the diagonal moved by 0.01 — but it is now full rank ($\operatorname{rank}(\widetilde{\boldsymbol{C}}) = 3$), which is exactly what we need if we want to invert it downstream.

3. Determinant

Term — Determinant

The *determinant* is a scalar summary of a square matrix, written $\det(\boldsymbol{A})$ or $|\boldsymbol{A}|$. Unlike rank it does not admit a simple verbal interpretation, but it carries a crucial structural fact: $\det(\boldsymbol{A}) = 0$ exactly when $\boldsymbol{A}$ is reduced rank (singular).

Why we care

A square matrix is invertible if and only if its determinant is non-zero. Determinants also appear throughout the *eigen-decomposition* of a matrix, which underlies principal component analysis and many other statistical techniques.

4. Determinant of a 2×2 Matrix

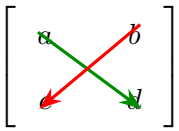
If $\boldsymbol{A}$ is a 2×2 matrix,

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

then

$$\det(\boldsymbol{A}) \;=\; |\boldsymbol{A}| \;=\; ad - bc.$$

The recipe is: multiply down the main diagonal (ad), subtract the product of the anti-diagonal (bc):



$$\det(\boldsymbol{A}) = ad - bc.$$

Worked example

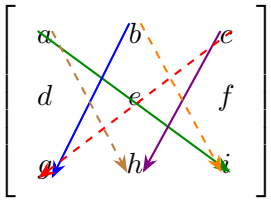
Let $\boldsymbol{D} = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$. Then

$$\det(\boldsymbol{D}) = (1)(8) - (2)(4) = 8 - 8 = 0.$$

The second column is just 2 times the first ($[2, 8]^T = 2[1, 4]^T$), so the columns are linearly dependent and $\boldsymbol{D}$ is reduced rank — which is precisely why the determinant vanished.

5. Determinant of a 3×3 Matrix

For $\boldsymbol{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$, *Sarrus' rule* adds the three "down-right" diagonal products and subtracts the three "down-left" diagonal products (wrapping around the matrix as needed):



$$\det(\boldsymbol{A}) \;=\; aei + bfg + cdh \;-\; ceg - bdi - afh.$$

Worked example

Let $\boldsymbol{E} = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 3 & 4 \\ 1 & 5 & -2 \end{bmatrix}$. Applying Sarrus' rule with $(a,b,c,d,e,f,g,h,i) = (2,-1,0,0,3,4,1,5,-2)$:

$$\begin{aligned} \det(\boldsymbol{E}) &= \big[(2)(3)(-2)\big] + \big[(-1)(4)(1)\big] + \big[(0)(0)(5)\big] \\ &\quad - \big[(0)(3)(1)\big] - \big[(-1)(0)(-2)\big] - \big[(2)(4)(5)\big] \\ &= -12 + (-4) + 0 - 0 - 0 - 40 \\ &= -56. \end{aligned}$$

Since $\det(\boldsymbol{E}) \neq 0$, $\boldsymbol{E}$ is full rank and invertible.

6. Matrix Inverse

Term — Matrix inverse

The *inverse* of a square matrix $\boldsymbol{A}$, written $\boldsymbol{A}^{-1}$, satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

The inverse exists only when $\boldsymbol{A}$ is square, full rank, and has $\det(\boldsymbol{A}) \neq 0$.

7. Inverse of a 2×2 Matrix

For $\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ the inverse has a clean closed form:

$$\boldsymbol{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{\det(\boldsymbol{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Swap the two diagonal entries, flip the sign of the two off-diagonal entries, and divide every entry by the determinant.

Worked example

Let $\boldsymbol{F} = \begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix}$. Then $\det(\boldsymbol{F}) = (1)(5) - (3)(2) = -1$, and

$$\boldsymbol{F}^{-1} = \frac{1}{(1)(5) - (3)(2)} \begin{bmatrix} 5 & -3 \\ -2 & 1 \end{bmatrix} = \frac{1}{-1} \begin{bmatrix} 5 & -3 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -5 & 3 \\ 2 & -1 \end{bmatrix}.$$

Worked example (verification)

Does $\boldsymbol{F}\boldsymbol{F}^{-1} = \boldsymbol{I}_2$?

$$(\boldsymbol{F}\boldsymbol{F}^{-1})_{1,1} = [1,\,3] \cdot \begin{bmatrix} -5 \\ 2 \end{bmatrix} = (1)(-5) + (3)(2) = -5 + 6 = 1,$$

$$(\boldsymbol{F}\boldsymbol{F}^{-1})_{1,2} = [1,\,3] \cdot \begin{bmatrix} 3 \\ -1 \end{bmatrix} = (1)(3) + (3)(-1) = 3 - 3 = 0,$$

$$(\boldsymbol{F}\boldsymbol{F}^{-1})_{2,1} = [2,\,5] \cdot \begin{bmatrix} -5 \\ 2 \end{bmatrix} = (2)(-5) + (5)(2) = -10 + 10 = 0,$$

$$(\boldsymbol{F}\boldsymbol{F}^{-1})_{2,2} = [2,\,5] \cdot \begin{bmatrix} 3 \\ -1 \end{bmatrix} = (2)(3) + (5)(-1) = 6 - 5 = 1.$$

So $\boldsymbol{F}\boldsymbol{F}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{I}_2$, as required.

8. Inverse of a Diagonal Matrix

Diagonal matrices are pleasantly easy to invert: flip each diagonal entry. If $\boldsymbol{A} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$, then

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{a} & 0 & 0 \\ 0 & \frac{1}{b} & 0 \\ 0 & 0 & \frac{1}{c} \end{bmatrix},$$

provided no diagonal entry is zero. (If any diagonal entry is zero the matrix is reduced rank and has no inverse.)

For square matrices larger than 2×2 with no special structure, there is no pleasant hand-formula; in practice we let a computer invert the matrix for us.

9. Systems of Linear Equations

One headline application of the matrix inverse is *solving a linear system*. Suppose we have

$$\begin{aligned} 2x + 3y - 5z &= 8, \\ -2y + 2z &= -3, \\ 5x \qquad\quad - 4z &= 3. \end{aligned}$$

We can bundle the coefficients, the unknowns, and the right-hand side into matrices and vectors:

$$\begin{bmatrix} 2 & 3 & -5 \\ 0 & -2 & 2 \\ 5 & 0 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 8 \\ -3 \\ 3 \end{bmatrix}, \quad \text{i.e.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

Key Idea 82

Any linear system can be written as $\boldsymbol{Ax} = \boldsymbol{b}$. If $\boldsymbol{A}$ is square and invertible, then left-multiplying by $\boldsymbol{A}^{-1}$ gives the unique solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.

The algebra behind that key idea:

$$\begin{aligned} \boldsymbol{Ax} &= \boldsymbol{b} \\ \boldsymbol{A}^{-1}\boldsymbol{Ax} &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ \boldsymbol{Ix} &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ \boldsymbol{x} &= \boldsymbol{A}^{-1}\boldsymbol{b}. \end{aligned}$$

Worked example

Take the 2×2 system

$$\begin{aligned} 2x + 5y &= -4, \\ x - y &= 5. \end{aligned}$$

In matrix form,

$$\underbrace{\begin{bmatrix} 2 & 5 \\ 1 & -1 \end{bmatrix}}_{\boldsymbol{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} -4 \\ 5 \end{bmatrix}}_{\boldsymbol{b}}.$$

The determinant is $\det(\boldsymbol{A}) = (2)(-1) - (5)(1) = -7$, so the inverse exists:

$$\boldsymbol{A}^{-1} = \frac{1}{-7}\begin{bmatrix} -1 & -5 \\ -1 & 2 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 1 & 5 \\ 1 & -2 \end{bmatrix}.$$

Applying $\boldsymbol{A}^{-1}$ to $\boldsymbol{b}$ gives

$$\begin{aligned} \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} &= \frac{1}{7}\begin{bmatrix} 1 & 5 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} -4 \\ 5 \end{bmatrix} \\ &= \frac{1}{7}\begin{bmatrix} (1)(-4) + (5)(5) \\ (1)(-4) + (-2)(5) \end{bmatrix} \\ &= \frac{1}{7}\begin{bmatrix} -4 + 25 \\ -4 - 10 \end{bmatrix} \\ &= \frac{1}{7}\begin{bmatrix} 21 \\ -14 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}. \end{aligned}$$

So $x = 3$ and $y = -2$. A quick plug-in check: $2(3) + 5(-2) = 6 - 10 = -4$ (✓), and $3 - (-2) = 5$ (✓).

PART 3

Practice Problems

Having met rank, determinant, and inverse separately, we now string all three together on a single 2×2 matrix.

Example #1. Define

$$\boldsymbol{G} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

(a) Solve for $\text{rank}(\boldsymbol{G})$.

(b) Solve for $\det(\boldsymbol{G})$.

(c) Solve for $\boldsymbol{G}^{-1}$, and verify that $\boldsymbol{G}\boldsymbol{G}^{-1} = \boldsymbol{I}_2$.

Solutions

(a) Rank of $\boldsymbol{G}$. Write the columns of $\boldsymbol{G}$ as $\boldsymbol{g}_{\cdot,1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\boldsymbol{g}_{\cdot,2} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, and ask whether

$$\boldsymbol{0} = \lambda_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + 2\lambda_2 \\ \lambda_1 + 3\lambda_2 \end{bmatrix}$$

is possible for some non-trivial (λ_1, λ_2). Subtracting the two equations gives $\lambda_2 = 0$, which then forces $\lambda_1 = 0$. So the only solution is the trivial one, the columns are linearly *independent*, and $\text{rank}(\boldsymbol{G}) = 2$. Since $\min\{2, 2\} = 2$, $\boldsymbol{G}$ is full rank.

(b) Determinant of $\boldsymbol{G}$. Apply the 2×2 determinant formula:

$$\det(\boldsymbol{G}) = g_{1,1}\, g_{2,2} - g_{1,2}\, g_{2,1} = (1)(3) - (2)(1) = 3 - 2 = 1.$$

Good: $\det(\boldsymbol{G}) = 1 \neq 0$, consistent with $\boldsymbol{G}$ being full rank (as we found in part (a)) and therefore invertible.

(c) Inverse of $\boldsymbol{G}$. Using the closed-form 2×2 inverse,

$$\begin{aligned} \boldsymbol{G}^{-1} &= \frac{1}{\det(\boldsymbol{G})} \begin{bmatrix} g_{2,2} & -g_{1,2} \\ -g_{2,1} & g_{1,1} \end{bmatrix} \\ &= \frac{1}{1} \begin{bmatrix} 3 & -2 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ -1 & 1 \end{bmatrix}. \end{aligned}$$

Verify $\boldsymbol{G}\boldsymbol{G}^{-1} = \boldsymbol{I}_2$:

$$(\boldsymbol{G}\boldsymbol{G}^{-1})_{1,1} = [1,\,2] \cdot \begin{bmatrix} 3 \\ -1 \end{bmatrix} = (1)(3) + (2)(-1) = 3 - 2 = 1,$$

$$(\boldsymbol{G}\boldsymbol{G}^{-1})_{1,2} = [1,\,2] \cdot \begin{bmatrix} -2 \\ 1 \end{bmatrix} = (1)(-2) + (2)(1) = -2 + 2 = 0,$$

$$(\boldsymbol{G}\boldsymbol{G}^{-1})_{2,1} = [1,\,3] \cdot \begin{bmatrix} 3 \\ -1 \end{bmatrix} = (1)(3) + (3)(-1) = 3 - 3 = 0,$$

$$(\boldsymbol{G}\boldsymbol{G}^{-1})_{2,2} = [1,\,3] \cdot \begin{bmatrix} -2 \\ 1 \end{bmatrix} = (1)(-2) + (3)(1) = -2 + 3 = 1.$$

Hence $\boldsymbol{G}\boldsymbol{G}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{I}_2$ (✓).

Gaussian Elimination

Lecture 28 • UVA Stats

April 24, 2026

Last lecture showed that a square, invertible coefficient matrix $\boldsymbol{A}$ *gives us the clean formula* $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ *for the linear system* $\boldsymbol{Ax} = \boldsymbol{b}$*. But many real systems — including the span-membership questions from two lectures ago — have a* non-square *or* singular *coefficient matrix. Today we learn the general-purpose tool that works for* any *shape of* $\boldsymbol{A}$*:* Gaussian elimination*. We augment* $\boldsymbol{A}$ *with* $\boldsymbol{b}$*, perform row operations to reach* echelon form*, and then* back substitute *to read off* $\boldsymbol{x}$*. We classify the three possible outcomes (one solution, no solution, infinitely many solutions) and close with a full* 3×3 *practice problem.*

PART 1

Motivation

Where we left off

Last lecture, a system of linear equations of the form

$$\begin{aligned} 2x + 3y - 5z &= 8, \\ -2y + 2z &= -3, \\ 5x \qquad\quad - 4z &= 3 \end{aligned}$$

could be bundled into the matrix equation

$$\underbrace{\begin{bmatrix} 2 & 3 & -5 \\ 0 & -2 & 2 \\ 5 & 0 & -4 \end{bmatrix}}_{\boldsymbol{A}} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \underbrace{\begin{bmatrix} 8 \\ -3 \\ 3 \end{bmatrix}}_{\boldsymbol{b}}, \qquad \text{i.e.} \quad \boldsymbol{Ax} = \boldsymbol{b}.$$

If $\boldsymbol{A}$ is square *and* invertible, then left-multiplying both sides by $\boldsymbol{A}^{-1}$ gives the clean one-line answer

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}.$$

But $\boldsymbol{A}^{-1}$ may not exist

We need to acknowledge that the matrix of coefficients $\boldsymbol{A}$ may not be invertible! Two obstructions appear in practice:

- $\boldsymbol{A}$ is not *square* — so $\boldsymbol{A}^{-1}$ is not even defined.
- $\boldsymbol{A}$ is square but *reduced rank* (equivalently $\det(\boldsymbol{A}) = 0$), so the inverse simply does not exist.

The canonical non-square example is a *span-membership question*: given a vector $\boldsymbol{v}$ and a collection of candidate vectors, is $\boldsymbol{v}$ in the span of that collection, and if so, with what coefficients?

A concrete obstruction

Is $\boldsymbol{v} = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix}$ in the span of $\left\{ \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ -3 \end{pmatrix} \right\}$?

We need λ_1, λ_2 such that

$$\begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix} = \lambda_1 \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ -2 \\ -3 \end{pmatrix}, \qquad \text{i.e.} \qquad \underbrace{\begin{bmatrix} 2 & 1 \\ 1 & -2 \\ -1 & -3 \end{bmatrix}}_{\boldsymbol{A} \text{ is } 3\times 2} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 4 \end{bmatrix}.$$

The coefficient matrix here is *not square*, so there is no $\boldsymbol{A}^{-1}$ to multiply by.

Key Idea 83

We need a general-purpose procedure that solves $\boldsymbol{Ax} = \boldsymbol{b}$ for *any* shape and rank of $\boldsymbol{A}$, not just square invertible ones. That procedure is *Gaussian elimination*: augment, row-reduce, back-substitute.

PART 2

Methods and Examples

1. Gaussian Elimination and the Augmented Matrix

Term — Gaussian elimination

Gaussian elimination is an application of matrix *row reduction* to solve a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$. The procedure has three stages:
1. form the *augmented matrix* by gluing $\boldsymbol{b}$ onto $\boldsymbol{A}$;
2. *row-reduce* the left block into echelon form;
3. *back-substitute* to read off $\boldsymbol{x}$.

Term — Augmented matrix

The *augmented matrix* for $\boldsymbol{Ax} = \boldsymbol{b}$ is the matrix formed by placing $\boldsymbol{b}$ as an extra column at the right of $\boldsymbol{A}$, typically separated by a vertical bar.

Worked example (setting up)

For the span-membership system above, $\begin{bmatrix} 2 & 1 \\ 1 & -2 \\ -1 & -3 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 4 \end{bmatrix}$, the augmented matrix is

$$\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 1 & -2 & 6 \\ -1 & -3 & 4 \end{array}\right).$$

2. Row Reduction and Echelon Form

Term — Row reduction

Row reduction modifies rows of a matrix using three elementary operations:
- scalar multiplication of a row, $R_i \leftarrow c\,R_i$ (with $c \neq 0$);
- vector addition of rows, $R_i \leftarrow R_i + c\,R_j$;
- row swap, $R_i \leftrightarrow R_j$ (used to bring a non-zero entry into a pivot position).

Each operation corresponds to left-multiplying by an invertible matrix, so all three preserve the solution set of the system, and the answer to $\boldsymbol{Ax} = \boldsymbol{b}$ is the same before and after.

Term — Echelon form

A matrix is in *echelon form* when its left block is upper triangular in the sense that:

- the first non-zero entry in each row is strictly to the right of the first non-zero entry in the row above; and
- any all-zero row, if present, sits below all rows that have a non-zero entry.

Schematically, an echelon-form left block looks like

$$\begin{bmatrix} \blacksquare & * & * & * \\ 0 & \blacksquare & * & * \\ 0 & 0 & \blacksquare & * \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(the } \blacksquare\text{'s are non-zero } \textit{pivots}\text{)}.$$

The further reduction in which every pivot equals 1 and every entry above each pivot is also zero is called *reduced row echelon form (RREF)*; it is unique for a given matrix, whereas plain echelon form is not.

Worked example (row-reducing our system)

Take the augmented matrix from above and clear the first column below the top entry. First eliminate the leading entry of row 2 using row 1:

$$\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 1 & -2 & 6 \\ -1 & -3 & 4 \end{array}\right) \xrightarrow{R_2 = R_1 - 2R_2} \left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ -1 & -3 & 4 \end{array}\right).$$

Next clear the leading entry of row 3:

$$\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ -1 & -3 & 4 \end{array}\right) \xrightarrow{R_3 = R_1 + 2R_3} \left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ 0 & -5 & 10 \end{array}\right).$$

Finally add row 2 to row 3 to zero it out:

$$\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ 0 & -5 & 10 \end{array}\right) \xrightarrow{R_3 = R_2 + R_3} \left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ 0 & 0 & 0 \end{array}\right).$$

The left block is now upper triangular, so this matrix is in echelon form.

3. Back Substitution

Term — Back substitution

Once the left block of the augmented matrix is in echelon form, convert it back into a system of equations and solve from the bottom row *upward*: each equation is in one fewer unknown than the one above it.

Worked example (finishing our system)

The echelon form $\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ 0 & 0 & 0 \end{array}\right)$ corresponds to

$$\begin{aligned} 2\lambda_1 + \lambda_2 &= 2, \\ 5\lambda_2 &= -10, \\ 0 &= 0 \quad \text{(trivially satisfied).} \end{aligned}$$

From the second equation, $\lambda_2 = -2$. Substituting back into the first:

$$2\lambda_1 + (-2) = 2 \quad \Longrightarrow \quad 2\lambda_1 = 4 \quad \Longrightarrow \quad \lambda_1 = 2.$$

So the unique solution is $(\lambda_1, \lambda_2) = (2, -2)$. A quick membership check: $2\begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} - 2\begin{pmatrix} 1 \\ -2 \\ -3 \end{pmatrix} = \begin{pmatrix} 4-2 \\ 2+4 \\ -2+6 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix}$ (✓). So $\boldsymbol{v}$ *is* in the span.

4. Pivots and Non-Uniqueness of Echelon Form

Term — Pivot

For a matrix in echelon form, the *pivot* of a row is its left-most non-zero entry.

For our example, the echelon form $\left(\begin{array}{cc|c} \boxed{2} & 1 & 2 \\ 0 & \boxed{5} & -10 \\ 0 & 0 & 0 \end{array}\right)$ has pivots 2 and 5.

Echelon form is *not* unique: we can always rescale a row to change its pivot. For simplicity and consistency it is common to scale so that every pivot equals 1:

$$\left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 5 & -10 \\ 0 & 0 & 0 \end{array}\right) \xrightarrow{R_2 = \frac{1}{5}R_2} \left(\begin{array}{cc|c} 2 & 1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{array}\right) \xrightarrow{R_1 = \frac{1}{2}R_1} \left(\begin{array}{cc|c} 1 & 0.5 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{array}\right).$$

Reading off the new system:

$$\begin{aligned} \lambda_1 + 0.5\lambda_2 &= 1, \\ \lambda_2 &= -2. \end{aligned}$$

Back-substitution: $\lambda_2 = -2$, then $\lambda_1 + 0.5(-2) = 1 \Rightarrow \lambda_1 = 2$. The solution matches the previous echelon form — as promised, the *solution* is invariant even though the echelon form is not.

5. Number of Solutions

A system $\boldsymbol{Ax} = \boldsymbol{b}$ can have exactly one solution, no solution, or infinitely many solutions. Gaussian elimination diagnoses all three cases automatically.

One solution (the generic case)

Our running example had exactly one solution: the number of rows with at least one non-zero entry in the row-reduced augmented matrix *equalled* the number of unknowns (two non-zero rows, two unknowns).

No solution

Consider the inconsistent system

$$\begin{aligned} x + y &= 2, \\ x + y &= 4. \end{aligned}$$

The augmented matrix is $\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & 1 & 4 \end{array}\right)$. Row-reducing:

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & 1 & 4 \end{array}\right) \xrightarrow{R_2 = R_1 - R_2} \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 0 & -2 \end{array}\right).$$

Back-substituting, the bottom row reads $0 = -2$ ($\times$), a false statement. A system with no solution always produces a contradictory row during back substitution.

Infinitely many solutions

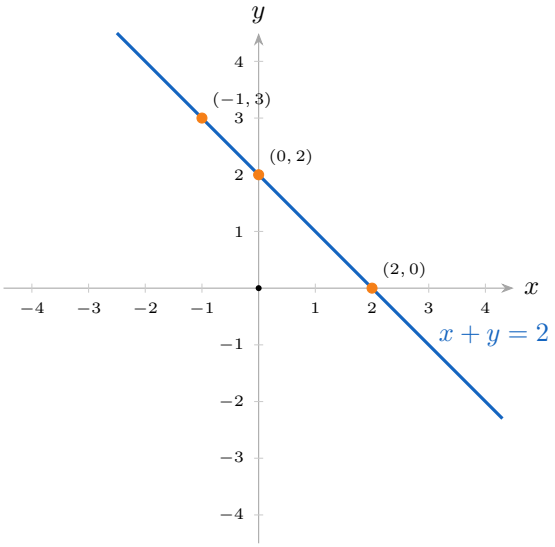
Consider the degenerate system

$$\begin{aligned} x + \ \ y &= 2, \\ 2x + 2y &= 4. \end{aligned}$$

The second equation is 2 times the first; they describe the same line. Row-reducing:

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 2 & 2 & 4 \end{array}\right) \xrightarrow{R_2 = 2R_1 - R_2} \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 0 & 0 \end{array}\right).$$

Back-substitution leaves a single equation $x + y = 2$ in two unknowns, so any (x, y) on that line is a solution — $(0, 2), (2, 0), (10, -8), \ldots$ all work:



Every point of that line is a valid (x, y), hence infinitely many solutions.

Reading off the count from the row-reduced matrix

Key Idea 84

Let r be the number of rows with a non-zero entry in the row-reduced augmented matrix, and let u be the number of unknowns in $\boldsymbol{x}$.

- If back substitution yields a false statement like $0 = -2$: the system has **no solution**.
- Otherwise, if $r = u$: the system has **exactly one solution**.
- Otherwise (so $r < u$): the system has **infinitely many solutions**, and the "missing" $u - r$ equations become free parameters.

PART 3

Practice Problems

Now we apply the full procedure end-to-end on a square 3×3 system.

Example #1. Apply Gaussian elimination to solve

$$\begin{aligned} x + \ y + \ z &= 10, \\ x + 2y + 3z &= 24, \\ x - \ y - \ z &= \ 0. \end{aligned}$$

Solutions

Step 1 — matrix form. Bundle the coefficients, unknowns, and right-hand side:

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & -1 & -1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 10 \\ 24 \\ 0 \end{bmatrix}}_{b}.$$

Step 2 — augmented matrix.

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 1 & 2 & 3 & 24 \\ 1 & -1 & -1 & 0 \end{array}\right).$$

Step 3 — row-reduce to echelon form. Remember that the choice of row operations is not unique — any correct sequence will do. One efficient sequence:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 1 & 2 & 3 & 24 \\ 1 & -1 & -1 & 0 \end{array}\right) \xrightarrow{R_2 = R_2 - R_1} \left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 0 & 1 & 2 & 14 \\ 1 & -1 & -1 & 0 \end{array}\right)$$

$$\xrightarrow{R_3 = R_1 - R_3} \left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 0 & 1 & 2 & 14 \\ 0 & 2 & 2 & 10 \end{array}\right) \xrightarrow{R_3 = 2R_2 - R_3} \left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 0 & 1 & 2 & 14 \\ 0 & 0 & 2 & 18 \end{array}\right)$$

$$\xrightarrow{R_3 = \frac{1}{2}R_3} \left(\begin{array}{ccc|c} 1 & 1 & 1 & 10 \\ 0 & 1 & 2 & 14 \\ 0 & 0 & 1 & 9 \end{array}\right).$$

The left block is now upper triangular with pivots $1, 1, 1$.

Step 4 — back substitute. Reading off the equivalent system,

$$\begin{aligned} x + y + z &= 10, \\ y + 2z &= 14, \\ z &= 9. \end{aligned}$$

From the bottom, $z = 9$. Substituting into the middle equation,

$$y + 2(9) = 14 \quad \Longrightarrow \quad y + 18 = 14 \quad \Longrightarrow \quad y = -4.$$

Substituting both into the top equation,

$$x + (-4) + 9 = 10 \quad \Longrightarrow \quad x + 5 = 10 \quad \Longrightarrow \quad x = 5.$$

So $(x, y, z) = (5, -4, 9)$.

Step 5 — plug-in check.

$$\begin{aligned} 5 + (-4) + 9 &= 10 \ (\checkmark), \\ 5 + 2(-4) + 3(9) &= 5 - 8 + 27 = 24 \ (\checkmark), \\ 5 - (-4) - 9 &= 5 + 4 - 9 = 0 \ (\checkmark). \end{aligned}$$

All three equations are satisfied, and the row-reduced matrix had $r = 3$ non-zero rows equal to $u = 3$ unknowns, confirming the solution is unique.

Eigenvalues and Eigenvectors

Lecture 29 • UVA Stats

April 24, 2026

Last lecture closed the chapter on solving linear systems $\boldsymbol{Ax} = \boldsymbol{b}$. *Today we turn the matrix into the unknown of a different question:* which directions does $\boldsymbol{A}$ leave invariant, and by what factor does it stretch them? *The answers are the* eigenvalues *and* eigenvectors *of* $\boldsymbol{A}$ *— the building blocks of* matrix eigendecomposition *and the foundation of principal component analysis, spectral clustering, and countless other tools we will meet later in the course. We motivate the special case in which matrix-vector multiplication acts like scalar multiplication, derive the characteristic equation* $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, *develop the Gaussian-elimination procedure for the corresponding eigenvectors, and close with a full* 2×2 *practice problem.*

PART 1

Motivation

1. What is matrix eigendecomposition?

Term — Matrix Eigendecomposition

Matrix eigendecomposition is the procedure of extracting two paired features from a square matrix $\boldsymbol{A}$: its *eigenvalues* (scalars) and its *eigenvectors* (vectors). An $N \times N$ matrix has exactly N eigenvalues and N eigenvectors, and each eigenvalue is matched with a particular eigenvector.

Before formally defining what those features are, it helps to sketch the framework.

- Eigendecomposition only works for square matrices.
- The purpose is to extract two important features from a matrix: **eigenvalues** (scalars) and **eigenvectors** (vectors).
- An $N \times N$ matrix yields N eigenvalues and N eigenvectors. They come in pairs: each eigenvalue is matched with a particular eigenvector.

For any linguists in the room, "eigen-" is a German prefix meaning "self," "own," or "characteristic."

Key Idea 85

So, what exactly *are* eigenvalues and eigenvectors, and how do we identify them inside a matrix? To answer that, we revisit the idea of *matrix-vector multiplication*.

PART 2

Methods and Examples

2. Matrix-vector multiplication revisited

We already know that multiplying a matrix by a vector produces another vector. *Usually* the new vector points in a different direction — the matrix has *rotated* the input. *Occasionally*, though, the direction is unchanged and the vector is merely rescaled. That special case is the doorway into eigendecomposition.

Worked example — the typical case

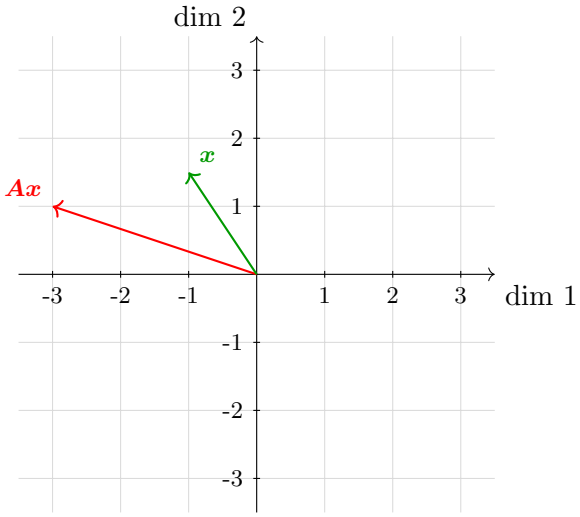
Let

$$\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}, \qquad \boldsymbol{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Then

$$\boldsymbol{Ax} = \begin{bmatrix} (4)(-1) + (1)(1) \\ (2)(-1) + (3)(1) \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}.$$

The angle of $\boldsymbol{x}$ and the angle of $\boldsymbol{Ax}$ are different: $\boldsymbol{A}$ has *rotated* $\boldsymbol{x}$.



Worked example — the special case

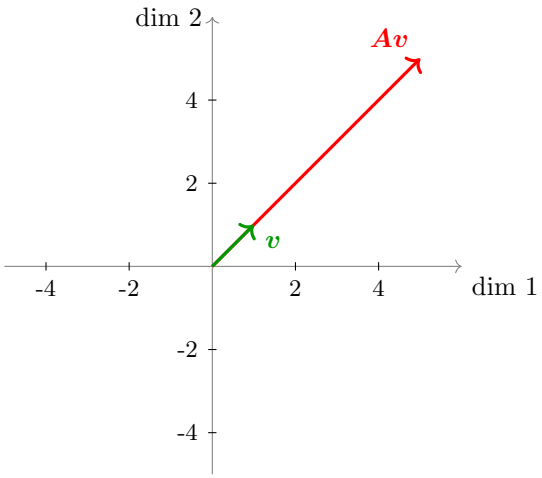
Now let

$$\boldsymbol{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Then

$$\boldsymbol{Av} = \begin{bmatrix} (4)(1) + (1)(1) \\ (2)(1) + (3)(1) \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \end{bmatrix} = 5\,\boldsymbol{v}.$$

The output points in the *same* direction as $\boldsymbol{v}$, only longer. Matrix-vector multiplication has acted like scalar multiplication: there exists a scalar $\lambda = 5$ for which $\boldsymbol{Av} = \lambda\boldsymbol{v}$.



3. The fundamental eigenvalue equation

Term — Eigenvalue / Eigenvector

For a square matrix $\boldsymbol{A}$, a non-zero vector $\boldsymbol{v}$ is an **eigenvector** of $\boldsymbol{A}$ with corresponding **eigenvalue** λ if

$$\boldsymbol{A}\boldsymbol{v} = \lambda\,\boldsymbol{v}.$$

That is, multiplying $\boldsymbol{v}$ by $\boldsymbol{A}$ has the same effect as multiplying $\boldsymbol{v}$ by the scalar λ — the direction of $\boldsymbol{v}$ is preserved and only its length changes.

In the example above, $\boldsymbol{v} = (1, 1)^\top$ is an eigenvector of $\boldsymbol{A}$ with eigenvalue $\lambda = 5$. The natural next question is how to find such pairs without having to guess.

4. Solving for eigenvalues

Starting from the fundamental eigenvalue equation, we manipulate it into a form that no longer mentions $\boldsymbol{v}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\,\boldsymbol{v} \implies \boldsymbol{A}\boldsymbol{v} - \lambda\,\boldsymbol{v} = \boldsymbol{0} \implies \boldsymbol{A}\boldsymbol{v} - \lambda\,(\boldsymbol{I}\boldsymbol{v}) = \boldsymbol{0} \implies (\boldsymbol{A} - \lambda\boldsymbol{I})\,\boldsymbol{v} = \boldsymbol{0}.$$

Sparing further derivation: for this to have a non-zero solution $\boldsymbol{v}$, the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$ must be *singular*. Equivalently,

$$\boxed{\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0.}$$

We can apply the determinant calculation to solve for eigenvalues!

Worked example

For $\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$,

$$\boldsymbol{A} - \lambda \boldsymbol{I} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}.$$

The determinant is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (4-\lambda)(3-\lambda) - (1)(2) = \lambda^2 - 7\lambda + 12 - 2 = \lambda^2 - 7\lambda + 10 = (\lambda - 5)(\lambda - 2).$$

Setting this equal to zero gives the two eigenvalues

$$\lambda_1 = 5 \quad \text{and} \quad \lambda_2 = 2.$$

Key Idea 86

The calculation grows quickly with the size of $\boldsymbol{A}$ — the determinant of an $N \times N$ matrix involves $N!$ terms. We will only compute eigenvalues by hand for 2×2 matrices; for larger ones, we lean on software.

5. Solving for eigenvectors

Once an eigenvalue λ is known, the corresponding eigenvector $\boldsymbol{v}$ is the unknown in the linear system

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\, \boldsymbol{v} = \boldsymbol{0},$$

where the matrix on the left has known entries and the right-hand side is the zero vector. Because $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, the matrix is *not* invertible, so we solve the system using **Gaussian elimination**.

Worked example — $\lambda_1 = 5$

For $\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$ and $\lambda = 5$,

$$\boldsymbol{A} - 5\boldsymbol{I} = \begin{bmatrix} 4-5 & 1 \\ 2 & 3-5 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}.$$

The system $(\boldsymbol{A} - 5\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}$ becomes the augmented matrix

$$\left(\begin{array}{cc|c} -1 & 1 & 0 \\ 2 & -2 & 0 \end{array}\right) \xrightarrow{R_2 \to 2R_1 + R_2} \left(\begin{array}{cc|c} -1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

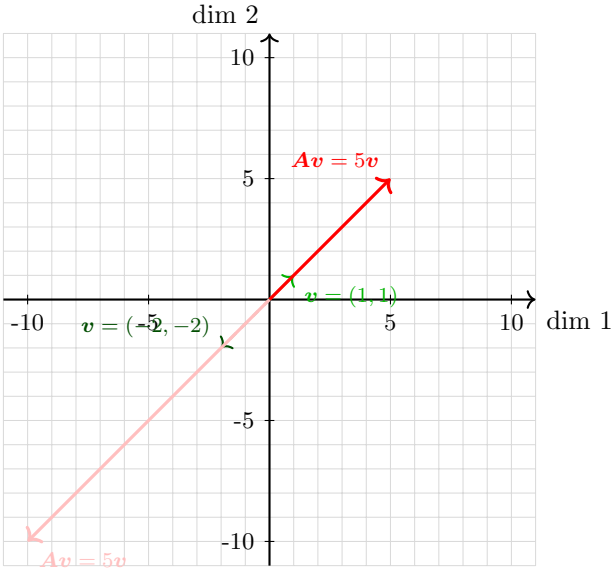
Back substitution gives

$$-v_1 + v_2 = 0 \quad \Longleftrightarrow \quad v_1 = v_2.$$

This is one equation in two unknowns, so there are *infinitely many* solutions: $(1, 1)$, $(-2, -2)$, $(10.25, 10.25)$, … — any vector satisfying $v_1 = v_2$.

Key Idea 87

Eigenvectors are **not unique**. For a fixed matrix $\boldsymbol{A}$ and a fixed eigenvalue λ, there are infinitely many eigenvectors paired with λ, but they all share the same direction (they lie on a single line through the origin). To pick a canonical representative, we conventionally report the eigenvector as a *unit vector* ($\|\boldsymbol{v}\| = 1$).



Starting from $\boldsymbol{v} = (1, 1)^\top$ with $\|\boldsymbol{v}\| = \sqrt{2}$, the unit eigenvector is

$$\boldsymbol{u}_{\boldsymbol{v}_1} = \frac{1}{\|\boldsymbol{v}\|}\boldsymbol{v} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

Worked example — $\lambda_2 = 2$

For the same $\boldsymbol{A}$ and $\lambda = 2$,

$$\boldsymbol{A} - 2\boldsymbol{I} = \begin{bmatrix} 4-2 & 1 \\ 2 & 3-2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}.$$

The augmented system row-reduces as

$$\left(\begin{array}{cc|c} 2 & 1 & 0 \\ 2 & 1 & 0 \end{array}\right) \xrightarrow{R_2 \to R_2 - R_1} \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right),$$

giving $2v_1 + v_2 = 0$, i.e. $v_2 = -2v_1$. Picking $\boldsymbol{v} = (1, -2)^\top$,

$$\|\boldsymbol{v}\| = \sqrt{1^2 + (-2)^2} = \sqrt{5}, \qquad \boldsymbol{u}_{\boldsymbol{v}_2} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}.$$

So the eigendecomposition of $\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$ is

$$\lambda_1 = 5,\ \boldsymbol{u}_{\boldsymbol{v}_1} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}; \qquad \lambda_2 = 2,\ \boldsymbol{u}_{\boldsymbol{v}_2} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}.$$

PART 3

Practice Problems

6. Example 1

Consider the matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix}.$$

(a) Solve for the two eigenvalues of $\boldsymbol{B}$.

(b) For each eigenvalue, solve for the corresponding unit eigenvector.

(c) Report the complete eigendecomposition of $\boldsymbol{B}$.

Solutions

(a) Eigenvalues of $\boldsymbol{B}$.

$$\boldsymbol{B} - \lambda \boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1-\lambda & 0 \\ 5 & 3-\lambda \end{bmatrix}.$$

The determinant of an upper- or lower-triangular matrix is the product of its diagonal entries, so

$$\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = (1-\lambda)(3-\lambda) - (0)(5) = (1-\lambda)(3-\lambda).$$

Setting this equal to zero gives

$$\lambda_1 = 3 \quad \text{and} \quad \lambda_2 = 1.$$

(b) Eigenvector for $\lambda_1 = 3$.

$$\boldsymbol{B} - 3\boldsymbol{I} = \begin{bmatrix} 1-3 & 0 \\ 5 & 3-3 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 5 & 0 \end{bmatrix}.$$

The system $(\boldsymbol{B} - 3\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}$ becomes

$$\left(\begin{array}{cc|c} -2 & 0 & 0 \\ 5 & 0 & 0 \end{array}\right) \xrightarrow{R_2 \to \frac{5}{2}R_1 + R_2} \left(\begin{array}{cc|c} -2 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

Back substitution gives $-2v_1 = 0$, so $v_1 = 0$ while v_2 is free. Taking $v_2 = 1$ already produces a unit vector:

$$\boldsymbol{u}_{v_1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

(b, continued) Eigenvector for $\lambda_2 = 1$.

$$\boldsymbol{B} - 1\boldsymbol{I} = \begin{bmatrix} 1-1 & 0 \\ 5 & 3-1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 5 & 2 \end{bmatrix}.$$

Swapping rows so the leading non-zero entry is on top,

$$\left(\begin{array}{cc|c} 0 & 0 & 0 \\ 5 & 2 & 0 \end{array}\right) \xrightarrow{R_1 \leftrightarrow R_2} \left(\begin{array}{cc|c} 5 & 2 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

Back substitution gives $5v_1 + 2v_2 = 0$, so $v_2 = -\frac{5}{2}v_1$. Choosing $\boldsymbol{v} = (2, -5)^\top$ to clear the fraction,

$$\|\boldsymbol{v}\| = \sqrt{2^2 + (-5)^2} = \sqrt{4 + 25} = \sqrt{29}, \qquad \boldsymbol{u}_{\boldsymbol{v}_2} = \frac{1}{\sqrt{29}} \begin{bmatrix} 2 \\ -5 \end{bmatrix} = \begin{bmatrix} 2/\sqrt{29} \\ -5/\sqrt{29} \end{bmatrix}.$$

(c) Complete eigendecomposition of $\boldsymbol{B}$.

$$\lambda_1 = 3,\ \boldsymbol{u}_{\boldsymbol{v}_1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 1,\ \boldsymbol{u}_{\boldsymbol{v}_2} = \begin{bmatrix} 2/\sqrt{29} \\ -5/\sqrt{29} \end{bmatrix}.$$

Matrix Diagonalization

Lecture 30 • UVA Stats

April 24, 2026

Last lecture extracted, one at a time, the eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$. Today we package those N separate eigenvalue equations into a single matrix identity: $\boldsymbol{AV} = \boldsymbol{V\Lambda}$, and — whenever the eigenvectors are linearly independent — factor $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{V\Lambda V}^{-1}$. This is matrix diagonalization*: the workhorse identity behind principal component analysis, spectral clustering, powers $\boldsymbol{A}^k$, and a closed-form expression for $\boldsymbol{A}^{-1}$ that costs only the inverse of a diagonal matrix. We verify the identity end-to-end on a 2×2 example, motivate why it matters, and close with a complete diagonalization practice problem.*

PART 1

Motivation

1. Recap: eigendecomposition

Term — Fundamental Eigenvalue Equation

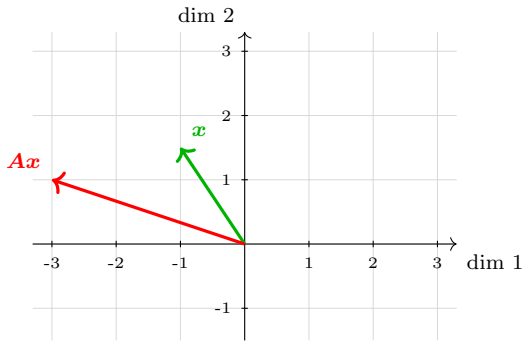
For a square matrix $\boldsymbol{A}$, a non-zero vector $\boldsymbol{v}$ is an **eigenvector** of $\boldsymbol{A}$ with corresponding **eigenvalue** λ if

$$\boldsymbol{A}\boldsymbol{v} = \lambda\,\boldsymbol{v}.$$

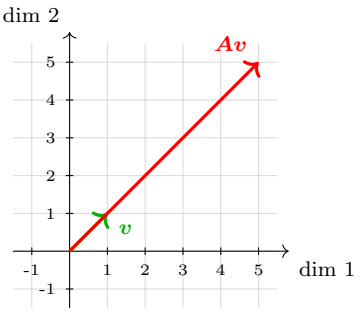
The matrix $\boldsymbol{A}$ scales $\boldsymbol{v}$ by λ without changing its direction.

The contrast with the generic case is the whole point: most vectors are *rotated* by $\boldsymbol{A}$, but eigenvectors are merely *rescaled*.

$\boldsymbol{A}$ rotates $\boldsymbol{x}$ ($\boldsymbol{x}$ is not an eigenvector)



$\boldsymbol{A}$ scales $\boldsymbol{v}$ ($\boldsymbol{v}$ is an eigenvector)



To find such pairs, we manipulate the equation into a form that no longer involves $\boldsymbol{v}$ on its right-hand side:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \implies (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}.$$

Solve $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ for the eigenvalues; for each λ, plug back in and run Gaussian elimination to get the corresponding eigenvector. Eigenvectors are not unique — only their direction is — so any non-zero scalar multiple is also an eigenvector.

Key Idea 88

An $N \times N$ square matrix has N eigenvalue/eigenvector pairs. Writing one fundamental eigenvalue equation per pair gives N separate matrix-vector identities. Today's question is whether we can collapse all N of them into a single matrix-matrix identity.

PART 2

Methods and Examples

2. From N equations to one: $\boldsymbol{AV = V\Lambda}$

We have N separate equations $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, one for each pair $(\lambda_i, \boldsymbol{v}_i)$. Stack the eigenvectors as the columns of a single matrix $\boldsymbol{V}$, and place the eigenvalues on the diagonal of a single matrix $\boldsymbol{\Lambda}$:

$$\boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_N \\ | & | & & | \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}.$$

Then all N eigenvalue equations are summarised by the single matrix identity

$$\boxed{\boldsymbol{AV} = \boldsymbol{V\Lambda}.}$$

To see why $\boldsymbol{V\Lambda}$ (rather than $\boldsymbol{\Lambda V}$) sits on the right, recall how multiplication by a diagonal matrix behaves:

- $\boldsymbol{D\,M}$ (diagonal × matrix) scales the i-th *row* of $\boldsymbol{M}$ by d_{ii}.
- $\boldsymbol{M\,D}$ (matrix × diagonal) scales the j-th *column* of $\boldsymbol{M}$ by d_{jj}.

In $\boldsymbol{V\Lambda}$, the diagonal sits on the right, so the j-th column of $\boldsymbol{V}$ — i.e. $\boldsymbol{v}_j$ — is rescaled by λ_j. That is exactly what each fundamental equation $\boldsymbol{A}\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j$ demands.

Worked example

From the previous lecture, the eigendecomposition of

$$\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

is

$$\lambda_1 = 5,\ \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 2,\ \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

(Any non-zero scalar multiple of an eigenvector is also an eigenvector, so we use these convenient integer-entry choices rather than the unit forms.) Stacking,

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

Verifying $\boldsymbol{AV} = \boldsymbol{V\Lambda}$:

$$\boldsymbol{AV} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} (4)(1)+(1)(1) & (4)(1)+(1)(-2) \\ (2)(1)+(3)(1) & (2)(1)+(3)(-2) \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 5 & -4 \end{bmatrix},$$

$$\boldsymbol{V\Lambda} = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} (1)(5)+(1)(0) & (1)(0)+(1)(2) \\ (1)(5)+(-2)(0) & (1)(0)+(-2)(2) \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 5 & -4 \end{bmatrix}.$$

The two products agree, so $\boldsymbol{AV} = \boldsymbol{V\Lambda}$ indeed holds.

3. Diagonalization: $\boldsymbol{A = V\Lambda V^{-1}}$

Term — Matrix Diagonalization

If the eigenvectors of $\boldsymbol{A}$ are linearly independent ($\det(\boldsymbol{V}) \neq 0$), then $\boldsymbol{V}$ is invertible and we can multiply both sides of $\boldsymbol{AV} = \boldsymbol{V\Lambda}$ on the right by $\boldsymbol{V}^{-1}$ to obtain

$$\boldsymbol{A} = \boldsymbol{V}\,\boldsymbol{\Lambda}\,\boldsymbol{V}^{-1}.$$

This factorisation is called the **diagonalization** of $\boldsymbol{A}$.

The derivation is one line of bookkeeping:

$$\boldsymbol{AV} = \boldsymbol{V\Lambda} \implies \boldsymbol{AVV}^{-1} = \boldsymbol{V\Lambda V}^{-1} \implies \boldsymbol{AI} = \boldsymbol{V\Lambda V}^{-1} \implies \boldsymbol{A} = \boldsymbol{V\Lambda V}^{-1}.$$

Worked example (continued)

We continue with $\boldsymbol{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$, $\boldsymbol{V} = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}$, and $\boldsymbol{\Lambda} = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}$.

Check $\boldsymbol{V}$ is invertible.

$$\det(\boldsymbol{V}) = (1)(-2) - (1)(1) = -3 \neq 0,$$

so $\boldsymbol{V}^{-1}$ exists.

Compute $\boldsymbol{V}^{-1}$. For a 2×2 matrix,

$$\boldsymbol{V}^{-1} = \frac{1}{\det(\boldsymbol{V})} \begin{bmatrix} v_{22} & -v_{12} \\ -v_{21} & v_{11} \end{bmatrix} = \frac{1}{-3} \begin{bmatrix} -2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{bmatrix}.$$

Recover $\boldsymbol{A}$ from $\boldsymbol{V\Lambda V}^{-1}$. We already computed $\boldsymbol{V\Lambda} = \begin{bmatrix} 5 & 2 \\ 5 & -4 \end{bmatrix}$, so

$$(\boldsymbol{V\Lambda})\boldsymbol{V}^{-1} = \begin{bmatrix} 5 & 2 \\ 5 & -4 \end{bmatrix} \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{bmatrix}.$$

Working entry by entry:

- $(1,1)$: $(5)(2/3) + (2)(1/3) = 10/3 + 2/3 = 12/3 = 4$.

- $(1,2)$: $(5)(1/3) + (2)(-1/3) = 5/3 - 2/3 = 3/3 = 1$.
- $(2,1)$: $(5)(2/3) + (-4)(1/3) = 10/3 - 4/3 = 6/3 = 2$.
- $(2,2)$: $(5)(1/3) + (-4)(-1/3) = 5/3 + 4/3 = 9/3 = 3$.

$$\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} = \boldsymbol{A}.$$

We have recovered $\boldsymbol{A}$ from its eigenvalues and eigenvectors alone.

4. Why bother?

Key Idea 89

Diagonalization is the engine behind a long list of statistical and numerical tools. For a data matrix $\boldsymbol{X}$, the diagonalization of $\boldsymbol{X}^\top\boldsymbol{X}$ is the mathematical core of *principal component analysis* and spectral methods more broadly — the eigenvectors point along the directions of greatest variance, and the eigenvalues quantify that variance. Whenever we want to project, compress, or rotate a dataset into the most informative coordinates, we are leaning on the identity $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$.

There is also a clean numerical pay-off: *matrix inversion through diagonalization.* Because $\boldsymbol{\Lambda}$ is diagonal, its inverse is just the reciprocals of its diagonal entries:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \implies \boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix}.$$

Combined with $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$, this gives

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{V}^{-1},$$

a closed-form inverse that costs only one diagonal inversion plus two matrix multiplications. The same trick generalises to integer powers: $\boldsymbol{A}^k = \boldsymbol{V}\,\boldsymbol{\Lambda}^k\,\boldsymbol{V}^{-1}$, where $\boldsymbol{\Lambda}^k$ is again obtained entry-wise. By induction, $\boldsymbol{A}^k = (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1})^k = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}\cdot\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}\cdots\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} = \boldsymbol{V}\,\boldsymbol{\Lambda}^k\,\boldsymbol{V}^{-1}$, since the inner $\boldsymbol{V}^{-1}\boldsymbol{V}$ pairs collapse to $\boldsymbol{I}$.

PART 3

Practice Problems

5. Example 1

Consider the matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix},$$

whose eigendecomposition (from the previous lecture) is

$$\lambda_1 = 3,\ \boldsymbol{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 1,\ \boldsymbol{v}_2 = \begin{bmatrix} 2 \\ -5 \end{bmatrix}.$$

(a) Form the matrices $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ and compute $\boldsymbol{V\Lambda}$.

(b) Compute $\boldsymbol{V}^{-1}$ (after first checking that $\boldsymbol{V}$ is invertible).

(c) Verify the diagonalization $\boldsymbol{B} = \boldsymbol{V\Lambda V}^{-1}$.

Solutions

(a) $\boldsymbol{V}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{V\Lambda}$. Using the eigendecomposition,

$$\boldsymbol{V} = \begin{bmatrix} 0 & 2 \\ 1 & -5 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

Multiplying,

$$\boldsymbol{V\Lambda} = \begin{bmatrix} 0 & 2 \\ 1 & -5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

Entry by entry:

- $(1,1)$: $(0)(3) + (2)(0) = 0$.
- $(1,2)$: $(0)(0) + (2)(1) = 2$.
- $(2,1)$: $(1)(3) + (-5)(0) = 3$.
- $(2,2)$: $(1)(0) + (-5)(1) = -5$.

$$\boldsymbol{V\Lambda} = \begin{bmatrix} 0 & 2 \\ 3 & -5 \end{bmatrix}.$$

(b) $\boldsymbol{V}^{-1}$. First, check invertibility:

$$\det(\boldsymbol{V}) = (0)(-5) - (2)(1) = -2 \neq 0,$$

so $\boldsymbol{V}^{-1}$ exists. Applying the 2×2 inverse formula,

$$\boldsymbol{V}^{-1} = \frac{1}{\det(\boldsymbol{V})} \begin{bmatrix} v_{22} & -v_{12} \\ -v_{21} & v_{11} \end{bmatrix} = \frac{1}{-2} \begin{bmatrix} -5 & -2 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 2.5 & 1 \\ 0.5 & 0 \end{bmatrix}.$$

(c) $\boldsymbol{V\Lambda V}^{-1}$.

$$(\boldsymbol{V\Lambda})\boldsymbol{V}^{-1} = \begin{bmatrix} 0 & 2 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 2.5 & 1 \\ 0.5 & 0 \end{bmatrix}.$$

Entry by entry:

- $(1,1)$: $(0)(2.5) + (2)(0.5) = 0 + 1 = 1$.
- $(1,2)$: $(0)(1) + (2)(0) = 0$.
- $(2,1)$: $(3)(2.5) + (-5)(0.5) = 7.5 - 2.5 = 5$.

- $(2,2)$: $(3)(1) + (-5)(0) = 3$.

$$\boldsymbol{V\Lambda V}^{-1} = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix} = \boldsymbol{B}.$$

The diagonalization checks out: $\boldsymbol{B}$ has been recovered from its eigenvalues and eigenvectors alone.