

Conditional Probability & Independence

Bayes' rule, total probability, and the backbone of Naïve Bayes

Probability Theory • Lecture 2

Applied Data Analysis — Module I

September 2026

Roadmap & learning objectives

Today we learn how probabilities *change* when we receive information.

- Define **conditional probability** and prove it is itself a probability measure.
- Derive the **multiplication rule** and the **chain rule**.
- Prove the **law of total probability** and **Bayes' theorem**; learn the odds form.
- Define **independence** (pairwise, mutual, conditional) and prove its basic properties.
- Recognize the classic traps: base-rate neglect, the prosecutor's fallacy, Simpson's paradox.
- ML tie-in: the **Naïve Bayes** classifier, and why precision and recall are conditional probabilities.



Where we stand after Lecture 1

$$(\Omega, \mathcal{F}, \mathbb{P}): \quad (\text{K1}) \mathbb{P}(A) \geq 0; \quad (\text{K2}) \mathbb{P}(\Omega) = 1; \quad (\text{K3}) \text{disjoint } A_n \Rightarrow \mathbb{P}\left(\bigsqcup_n A_n\right) = \sum_n \mathbb{P}(A_n).$$

Consequences we will use constantly today:

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A), \quad A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B).$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$
- $A = (A \cap B) \sqcup (A \cap B^c), \text{ hence } \mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$

The new question

All of this describes probabilities *before* anything is observed. What happens to $\mathbb{P}(A)$ once we learn that some event B has occurred?



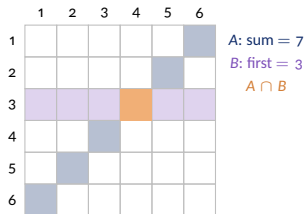


1.

Conditional probability

Motivating example — two dice

Roll two fair dice; Ω has 36 equally likely outcomes. Let $A = \{\text{sum} = 7\}$ and $B = \{\text{first die shows } 3\}$.



- Before rolling: $\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}$.
- We are told the first die shows 3. Only the 6 outcomes of B remain possible.
- Among them, exactly one (namely $(3, 4)$) has sum 7.

$$\mathbb{P}(A \text{ given } B) = \frac{1}{6} = \frac{|A \cap B|}{|B|} = \frac{1/36}{6/36} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Conditioning = **shrinking the sample space to B and renormalizing.**



Definition

Definition (Conditional probability)

Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The conditional probability of A given B is

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

- Read: “the probability of A , given *that* B has occurred”.
- The denominator renormalizes so that $\mathbb{P}(B | B) = 1$: the new “certain event” is B .
- If $\mathbb{P}(B) = 0$ the ratio is undefined — conditioning on impossible events needs more theory (measure-theoretic conditioning, beyond this course).

Two dice again

$$\mathbb{P}(\text{sum} = 7 | \text{first} = 3) = \frac{1/36}{1/6} = \frac{1}{6}, \quad \mathbb{P}(\text{first} = 3 | \text{sum} = 7) = \frac{1/36}{1/6} = \frac{1}{6}. \text{ Equal here —}$$

but not in general!

Conditional probability is a probability measure

Theorem

Fix B with $\mathbb{P}(B) > 0$. Then $Q(A) := \mathbb{P}(A \mid B)$ satisfies Kolmogorov's axioms (K1)–(K3) on $(\Omega, \mathcal{F})$.

Proof.

(K1) $Q(A) = \mathbb{P}(A \cap B)/\mathbb{P}(B) \geq 0$ as a ratio of nonnegative numbers.

(K2) $Q(\Omega) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.

(K3) Let $A_1, A_2, \dots$ be disjoint. Then the sets $A_n \cap B$ are also disjoint, so by (K3) for $\mathbb{P}$:

$$Q\left(\bigsqcup_n A_n\right) = \frac{\mathbb{P}\left(\bigsqcup_n (A_n \cap B)\right)}{\mathbb{P}(B)} = \frac{\sum_n \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_n Q(A_n).$$

□

Why this matters

Every theorem of Lecture 1 holds with $\mathbb{P}(\cdot \mid B)$ in place of $\mathbb{P}(\cdot)$ — for free.

Consequences — the usual rules, conditioned

From the theorem, for any B with $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A^c | B) = 1 - \mathbb{P}(A | B)$.
- $\mathbb{P}(A_1 \cup A_2 | B) = \mathbb{P}(A_1 | B) + \mathbb{P}(A_2 | B) - \mathbb{P}(A_1 \cap A_2 | B)$.
- $A_1 \subseteq A_2 \Rightarrow \mathbb{P}(A_1 | B) \leq \mathbb{P}(A_2 | B)$.
- $A \supseteq B \Rightarrow \mathbb{P}(A | B) = 1$; $A \cap B = \emptyset \Rightarrow \mathbb{P}(A | B) = 0$.

Check the last two

$A \supseteq B$: $A \cap B = B$, so $\mathbb{P}(A | B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$. $A \cap B = \emptyset$: $\mathbb{P}(A \cap B) = 0$, so $\mathbb{P}(A | B) = 0$.

What does *not* hold

Nothing links $\mathbb{P}(A | B)$ and $\mathbb{P}(A | B^c)$ in general: they may be equal, or wildly different. And $\mathbb{P}(A | B) + \mathbb{P}(A | B^c) \neq 1$ in general.



The multiplication rule

Theorem (Multiplication rule)

If $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \cap B) = \mathbb{P}(B) \mathbb{P}(A | B)$. Symmetrically, if $\mathbb{P}(A) > 0$, $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B | A)$.

Proof.

Multiply both sides of the definition $\mathbb{P}(A | B) = \mathbb{P}(A \cap B) / \mathbb{P}(B)$ by $\mathbb{P}(B)$. □

Trivial to prove — but it turns the definition into a **computational tool**: joint probabilities are built from a first-stage probability and a second-stage conditional probability.

Two cards without replacement

Draw two cards from a standard deck.

$$\mathbb{P}(\text{both aces}) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1) = \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221} \approx 0.0045.$$



The chain rule

Theorem (Chain rule)

If $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$, then

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

Induction on n .

$n = 2$ is the multiplication rule. Assume the formula for $n - 1$ events and set $C = A_1 \cap \cdots \cap A_{n-1}$. By the multiplication rule $\mathbb{P}(C \cap A_n) = \mathbb{P}(C) \mathbb{P}(A_n | C)$; expand $\mathbb{P}(C)$ by the induction hypothesis. All the conditioning events have positive probability because they contain C . □

- Any ordering of the A_i works — choose the one whose conditionals are easy.
- Note the positivity condition: it guarantees every conditional probability is defined.



Worked example — the chain rule in action

Three aces in a row

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{5525}.$$

An urn

An urn contains 5 red and 3 blue balls. Draw three without replacement.

$$\mathbb{P}(R, R, B \text{ in this order}) = \frac{5}{8} \cdot \frac{4}{7} \cdot \frac{3}{6} = \frac{60}{336} = \frac{5}{28}.$$

Any other order of two reds and one blue gives the same value (check!), so

$$\mathbb{P}(\text{exactly two red}) = 3 \cdot \frac{5}{28} = \frac{15}{28}.$$

Birthday problem, revisited

“All n birthdays distinct” = $D_2 \cap \dots \cap D_n$, where D_k = “person k differs from persons 1, $k-1$ ”. Chain rule: $\mathbb{P} = \prod_{k=2}^n \mathbb{P}(D_k | D_2 \cap \dots \cap D_{k-1}) = \prod_{k=2}^n \frac{365-(k-1)}{365}$ — exactly the



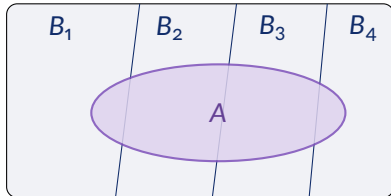
2.

Total probability and Bayes' theorem

Partitions

Definition (Partition)

Events $B_1, \dots, B_n$ (or countably many) form a **partition** of Ω if they are pairwise disjoint and $\bigsqcup_i B_i = \Omega$. We also require $\mathbb{P}(B_i) > 0$ for conditioning.



- Simplest partition: $\{B, B^c\}$.
- Any event A is cut into disjoint slices $A \cap B_i$:

$$A = \bigsqcup_i (A \cap B_i).$$

- Partitions model “cases” or “scenarios”:
which factory made the item, which
disease the patient has, which class an
email belongs to.

The law of total probability

Theorem (Law of total probability)

If $\{B_i\}$ is a partition of Ω with $\mathbb{P}(B_i) > 0$, then for every event A

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A | B_i) \mathbb{P}(B_i).$$

Proof.

The slices $A \cap B_i$ are disjoint and their union is A , so by (K3) $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$. Apply the multiplication rule to each term: $\mathbb{P}(A \cap B_i) = \mathbb{P}(A | B_i) \mathbb{P}(B_i)$. $\square$

- A **weighted average** of the conditional probabilities, weights = $\mathbb{P}(B_i)$. In particular $\min_i \mathbb{P}(A | B_i) \leq \mathbb{P}(A) \leq \max_i \mathbb{P}(A | B_i)$.
- Two-case version: $\mathbb{P}(A) = \mathbb{P}(A | B) \mathbb{P}(B) + \mathbb{P}(A | B^c) \mathbb{P}(B^c)$.



Worked example — two production lines

A plant has two lines. Line 1 makes 60% of the output with a 2% defect rate; line 2 makes 40% with a 5% defect rate. An item is picked at random.

Total probability

B_1 = “from line 1”, B_2 = “from line 2” — a partition. D = “defective”.

$$\mathbb{P}(D) = \mathbb{P}(D | B_1)\mathbb{P}(B_1) + \mathbb{P}(D | B_2)\mathbb{P}(B_2) = 0.02 \cdot 0.6 + 0.05 \cdot 0.4 = 0.012 + 0.020 = 0.032.$$

- 3.2% lies between 2% and 5%, closer to line 1 because line 1 has more weight.
- Natural follow-up question: the item *is* defective — which line did it probably come from? That is the *inverse* question, answered by Bayes.



Bayes' theorem

Theorem (Bayes)

Let $\{B_i\}$ be a partition with $\mathbb{P}(B_i) > 0$, and $\mathbb{P}(A) > 0$. Then for each j

$$\mathbb{P}(B_j | A) = \frac{\mathbb{P}(A | B_j) \mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A | B_j) \mathbb{P}(B_j)}{\sum_i \mathbb{P}(A | B_i) \mathbb{P}(B_i)}.$$

Proof.

By the multiplication rule written both ways, $\mathbb{P}(B_j \cap A) = \mathbb{P}(A | B_j) \mathbb{P}(B_j) = \mathbb{P}(B_j | A) \mathbb{P}(A)$. Divide by $\mathbb{P}(A)$; then expand $\mathbb{P}(A)$ by the law of total probability. $\square$

Vocabulary (the language of Bayesian statistics)

$\mathbb{P}(B_j)$ — **prior**; $\mathbb{P}(A | B_j)$ — **likelihood**; $\mathbb{P}(B_j | A)$ — **posterior**; $\mathbb{P}(A)$ — **evidence** (normalizing constant).

posterior $\propto$ likelihood $\times$ prior.

Worked example — which line produced the defect?

Bayes on the production lines

$$\mathbb{P}(B_2 | D) = \frac{\mathbb{P}(D | B_2)\mathbb{P}(B_2)}{\mathbb{P}(D)} = \frac{0.05 \cdot 0.4}{0.032} = \frac{0.020}{0.032} = 0.625.$$

Prior $\mathbb{P}(B_2) = 0.4 \rightarrow$ posterior 0.625: the defect is evidence for line 2.

- Check:

$\mathbb{P}(B_1 | D) = 0.012/0.032 = 0.375$; the posteriors sum to 1.

- The direction of update is decided by the *likelihood ratio*

$$\mathbb{P}(D | B_2)/\mathbb{P}(D | B_1) = 2.5 > 1.$$

	line 1	line 2
prior	0.60	0.40
likelihood	0.02	0.05
product	0.012	0.020
posterior	0.375	0.625



Worked example — a medical test and the base rate

Prevalence $\mathbb{P}(D) = 0.001$. Test sensitivity $\mathbb{P}(+ | D) = 0.99$, false-positive rate $\mathbb{P}(+ | D^c) = 0.05$. A random person tests positive. What is $\mathbb{P}(D | +)$?

$$\mathbb{P}(+) = 0.99 \cdot 0.001 + 0.05 \cdot 0.999 = 0.00099 + 0.04995 = 0.05094, \quad \mathbb{P}(D | +) = \frac{0.00099}{0.05094}$$

Base-rate neglect

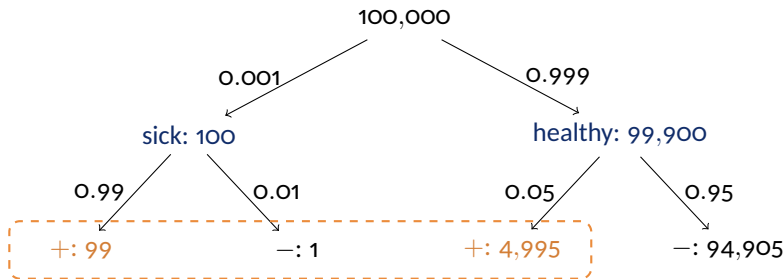
A “99% accurate” test, yet a positive result means under 2% chance of disease. The rare disease is swamped by false positives from the huge healthy population.

Natural frequencies — the same computation with 100,000 people

100 are sick, 99 of them test positive. 99,900 are healthy, 4,995 of them test positive. Among the 5,094 positives, only 99 are sick: $99/5094 \approx 1.9\%$.



Bayes as a tree



$$\mathbb{P}(D | +) = \frac{\text{sick and } +}{\text{all } +} = \frac{99}{99 + 4995} \approx 0.019.$$

A probability tree makes both total probability (sum along the dashed box) and Bayes (one branch over the box) visible.

How the base rate drives the answer

Same test (0.99 sensitivity, 0.05 false-positive rate), different populations:

prevalence $\mathbb{P}(D)$	0.001	0.01	0.10	0.50
$\mathbb{P}(D \mid +)$	0.019	0.167	0.688	0.952

- The posterior is *not* a property of the test alone; it depends on *whom* you test.
- Screening a random population vs. testing a symptomatic patient are different problems with the same test.
- Data-science version: a classifier's precision changes when the class balance of the deployment data changes.



The odds form of Bayes' theorem

Theorem (Bayes in odds)

For a hypothesis H and evidence E with $\mathbb{P}(E) > 0$,

$$\underbrace{\frac{\mathbb{P}(H|E)}{\mathbb{P}(H^c|E)}}_{\text{posterior odds}} = \underbrace{\frac{\mathbb{P}(E|H)}{\mathbb{P}(E|H^c)}}_{\text{likelihood ratio (LR)}} \cdot \underbrace{\frac{\mathbb{P}(H)}{\mathbb{P}(H^c)}}_{\text{prior odds}}.$$

Proof.

Write Bayes for H and for H^c ; both have the same denominator $\mathbb{P}(E)$. Divide one by the other. □

Medical test in odds

Prior odds $0.001/0.999 \approx 1:999$. $LR = 0.99/0.05 = 19.8$. Posterior odds $\approx 19.8:999 \approx 1:50.5$, i.e. $\mathbb{P}(D|+) \approx 1/51.5 \approx 0.019$. The evidence multiplies the odds by ≈ 20 — but $20 \times$ tiny is still small

Sequential updating

Observe two pieces of evidence E_1, E_2 that are *conditionally independent given the hypothesis* (defined formally later): $\mathbb{P}(E_1 \cap E_2 | H) = \mathbb{P}(E_1 | H)\mathbb{P}(E_2 | H)$, same for H^c .

Theorem

Posterior odds after $E_1, E_2 = LR_1 \cdot LR_2 \cdot \text{prior odds}$.

Proof.

Apply the odds form with $E = E_1 \cap E_2$; the likelihood ratio factorizes by conditional independence. □

- **Order does not matter:** updating on E_1 then E_2 equals updating on E_2 then E_1 , equals updating on both at once. Batch Bayes = online Bayes.
- In logs: $\log(\text{posterior odds}) = \log(\text{prior odds}) + \sum_k \log LR_k$ — evidence *adds up*. This is exactly the structure of Naïve Bayes (Section 5).

Second, independent test

After one positive: odds 1:50.5. A second positive (same LR 19.8): odds 19.8:50.5 ≈ 0.39 , 22 / 49

The prosecutor's fallacy

The mistake

Confusing $\mathbb{P}(\text{evidence} \mid \text{innocent})$ with $\mathbb{P}(\text{innocent} \mid \text{evidence})$.

A DNA match

A DNA profile found at the scene matches 1 in 1,000,000 people, and it matches the suspect. “The chance he is innocent is one in a million” — **wrong**. That number is $\mathbb{P}(\text{match} \mid \text{innocent})$.

In a city of 5,000,000 people, about 5 innocent people also match. If the suspect was found by searching a database with no other evidence, the prior is $1/5,000,000$ and

$$\mathbb{P}(\text{guilty} \mid \text{match}) = \frac{1 \cdot 1}{1 \cdot 1 + 5,000,000 \cdot 10^{-6}} = \frac{1}{1 + 5} \approx 0.17.$$

Same lesson as the medical test: a small $\mathbb{P}(E \mid H^c)$ does not mean a small $\mathbb{P}(H^c \mid E)$ unless the prior of H is substantial. **Always ask: what is the base rate?**



Worked example — which coin?

A box holds three coins: fair ($p = \frac{1}{2}$), biased ($p = \frac{3}{4}$), and double-headed ($p = 1$). Pick one at random and flip it twice: **both heads**. Which coin is it?

	fair	biased	two heads
prior $\mathbb{P}(C_i)$	$1/3$	$1/3$	$1/3$
likelihood $\mathbb{P}(HH C_i)$	$1/4$	$9/16$	1
product	$1/12$	$3/16$	$1/3$
posterior	0.138	0.310	0.552

Evidence $\mathbb{P}(HH) = \frac{1}{12} + \frac{3}{16} + \frac{1}{3} = \frac{29}{48}$; posteriors are products divided by it.

- Three hypotheses, one rule. The likelihood of the data under each hypothesis does all the work.
- With uniform priors, the posterior is just the *normalized likelihood*.





3.

Independence

Independence — definition

When does information about B tell us *nothing* about A ? When $\mathbb{P}(A | B) = \mathbb{P}(A)$.

Definition (Independence)

Events A and B are **independent** (written $A \perp\!\!\!\perp B$) if

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

Theorem

If $\mathbb{P}(B) > 0$, then $A \perp\!\!\!\perp B \iff \mathbb{P}(A | B) = \mathbb{P}(A)$. If also $\mathbb{P}(A) > 0$, this is $\iff \mathbb{P}(B | A) = \mathbb{P}(B)$.

Proof.

$\mathbb{P}(A | B) = \mathbb{P}(A \cap B) / \mathbb{P}(B)$ equals $\mathbb{P}(A)$ iff $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$. The second claim is symmetric. □



Properties of independence

Theorem

- 1 Ω and $\emptyset$ are independent of every event.
- 2 If $A \perp\!\!\!\perp B$, then $A \perp\!\!\!\perp B^c$, $A^c \perp\!\!\!\perp B$, $A^c \perp\!\!\!\perp B^c$.
- 3 If $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ with $B \cap C = \emptyset$, then $A \perp\!\!\!\perp (B \cup C)$.

Proof.

$$(1) \mathbb{P}(A \cap \Omega) = \mathbb{P}(A) = \mathbb{P}(A)\mathbb{P}(\Omega); \mathbb{P}(A \cap \emptyset) = 0 = \mathbb{P}(A)\mathbb{P}(\emptyset).$$

$$(2) \mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c).$$

The other two follow by symmetry and by applying this twice.

$$(3) \mathbb{P}(A \cap (B \cup C)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(C) = \mathbb{P}(A)\mathbb{P}(B \cup C). \quad \square$$

But

$A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ do not imply $A \perp\!\!\!\perp (B \cap C)$ or $A \perp\!\!\!\perp (B \cup C)$ in general (next slides).



Disjoint is not independent

A very common confusion

Disjoint ($A \cap B = \emptyset$) and independent ($\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$) are almost *opposites*.

Theorem

If $A \cap B = \emptyset$ and $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$, then A and B are **dependent**.

Proof.

$\mathbb{P}(A \cap B) = 0$ while $\mathbb{P}(A)\mathbb{P}(B) > 0$. □

- Intuition: if A and B cannot both happen, learning B tells you A is now impossible — maximal information, not zero information. $\mathbb{P}(A | B) = 0 \neq \mathbb{P}(A)$.
- Disjointness is a statement about *sets*; independence about *numbers*. You cannot read independence off a Venn diagram (recall the rectangle picture from Lecture 1).



Worked example — checking independence

Two fair dice. A = “first die even”, B = “sum = 7”, C = “sum = 8”.

A and B

$\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{6}{36} = \frac{1}{6}$. $A \cap B = \{(2, 5), (4, 3), (6, 1)\}$, so $\mathbb{P}(A \cap B) = \frac{3}{36} = \frac{1}{12} = \frac{1}{2} \cdot \frac{1}{6}$.

Independent. Knowing the first die is even does not change the chance of a 7.

A and C

$\mathbb{P}(C) = \frac{5}{36}$ (pairs $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$). $A \cap C = \{(2, 6), (4, 4), (6, 2)\}$, so $\mathbb{P}(A \cap C) = \frac{3}{36}$, while $\mathbb{P}(A)\mathbb{P}(C) = \frac{1}{2} \cdot \frac{5}{36} = \frac{5}{72} \neq \frac{3}{36}$. **Dependent:** $\mathbb{P}(A | C) = \frac{3}{5} > \frac{1}{2}$.

Independence is a *numerical coincidence* to be checked, not a visual property — and it can hold for events that look related (here, A and B both involve the first die).



Mutual independence of several events

Definition (Mutual independence)

Events $A_1, \dots, A_n$ are **mutually independent** if for every subset of indices $\{i_1 < \dots < i_k\}$, $k \geq 2$,

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

- For $n = 3$ this is **four** equations: three pairwise ones and $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.
- In general $2^n - n - 1$ equations; pairwise independence is only $\binom{n}{2}$ of them.
- Consequence (exercise): if $A_1, \dots, A_n$ are mutually independent, so are any events obtained by replacing some A_i by A_i^c .

Independent trials

“Repeat an experiment n times independently” means: events determined by different repetitions are mutually independent. This is the model behind coin flips, dice, i.i.d. samples.



Pairwise independence does not imply mutual independence

Two fair coins

A = “first coin H ”, B = “second coin H ”, C = “the two coins agree”.

- $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$.
 - $A \cap C = \{HH\}$, so $\mathbb{P}(A \cap C) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C)$; likewise $B \perp\!\!\!\perp C$; and $A \perp\!\!\!\perp B$. **All three pairs are independent.**
 - But $A \cap B \cap C = \{HH\}$, so $\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.
-
- Here C is a deterministic function of (A, B) : knowing *both* A and B fixes C , although knowing either one alone says nothing about C .
 - This is also the promised counterexample: $C \perp\!\!\!\perp A$, $C \perp\!\!\!\perp B$, yet $C \not\perp\!\!\!\perp (A \cap B)$.



Independent trials — a first look at the Binomial

Theorem

Perform n independent trials, each succeeding with probability p . Then

$$\mathbb{P}(\text{exactly } k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n.$$

Proof.

Fix a pattern with successes at k specified trials and failures elsewhere. By mutual independence its probability is $p^k (1-p)^{n-k}$, whatever the positions. There are $\binom{n}{k}$ such patterns (Lecture 1), they are disjoint events, so add. $\square$

Check

$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1$ by the binomial theorem — a valid distribution.
It gets a name, $\text{Bin}(n, p)$, in Lecture 4.

Worked example — reliability of a system

Each of n components works independently with probability p .

Series (all must work)

$$\mathbb{P}(\text{system works}) = p^n.$$

$n = 5, p = 0.9$: $0.9^5 \approx 0.59$. Reliability decays fast.

Parallel (at least one must work)

$$\mathbb{P}(\text{system works}) = 1 - (1 - p)^n.$$

$n = 5, p = 0.9$: $1 - 0.1^5 = 0.99999$. Redundancy is powerful.

Both use independence: the parallel case via the complement (“all fail”) and property (2) of independence — complements of independent events are independent.

When independence fails

A shared power supply, a common bug, correlated market shocks: if failures are *dependent*, $1 - (1 - p)^n$ is wildly optimistic. Independence is an **assumption** to be justified, not a fact.



Conditional independence

Definition

A and B are **conditionally independent given C** (with $\mathbb{P}(C) > 0$), written $A \perp\!\!\!\perp B \mid C$, if

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

Independence under the conditional measure $\mathbb{P}(\cdot \mid C)$ — all properties above transfer.

Two tests for the same disease

Given the patient's true status, the two test results are (plausibly) independent: each test reacts only to the disease, not to the other test. This is what made sequential updating multiply likelihood ratios.

Neither implies the other

- $A \perp\!\!\!\perp B \mid C$ but $A \not\perp\!\!\!\perp B$: the two test results are *dependent* unconditionally — a positive first test raises the chance of a positive second (both point to disease).
- $A \perp\!\!\!\perp B$ but $A \not\perp\!\!\!\perp B \mid C$: two coins A, B with $C =$ “they agree” — independent, but



4.

Paradoxes and pitfalls

Simpson's paradox — the numbers

Two treatments, two hospitals. Recovery rates:

	Hospital 1	Hospital 2	Combined
Treatment A	$\frac{90}{100} = 0.90$	$\frac{30}{100} = 0.30$	$\frac{120}{200} = 0.60$
Treatment B	$\frac{870}{1000} = 0.87$	$\frac{2}{10} = 0.20$	$\frac{872}{1010} = 0.86$

- In *each* hospital, A beats B ($0.90 > 0.87$ and $0.30 > 0.20$).
- Pooled, B beats A ($0.86 > 0.60$).

Which treatment would you choose?

A — the within-hospital comparison holds the confounder (hospital, i.e. patient severity) fixed. The pooled numbers compare a mostly-easy-case group to a mostly-hard-case group.



Simpson's paradox — why it happens

Let $R = \text{recovery}$, $T \in \{A, B\}$, $H \in \{1, 2\}$. By total probability, conditioning on treatment,

$$\mathbb{P}(R | T) = \mathbb{P}(R | T, H=1) \mathbb{P}(H=1 | T) + \mathbb{P}(R | T, H=2) \mathbb{P}(H=2 | T).$$

- The pooled rate is a weighted average of the within-hospital rates — but with **treatment-specific weights** $\mathbb{P}(H | T)$.
- Treatment A was given mostly in hospital 2 (hard cases): weights (0.5, 0.5).
Treatment B mostly in hospital 1 (easy cases): weights (0.99, 0.01).
- Different weights can reverse the order of two averages even when each pair of terms is ordered the same way.

Averages of conditional probabilities depend on the weights. Compare like with like: condition on the confounder.



Data-science relevance: aggregated dashboards, A/B tests with unequal traffic allocation, and any observational comparison are all exposed to this reversal.

A catalogue of conditional-probability traps

- 1 **Inverse confusion.** $\mathbb{P}(A | B) \neq \mathbb{P}(B | A)$. Prosecutor's fallacy; "most accidents happen near home".
- 2 **Base-rate neglect.** Ignoring the prior. A strong likelihood ratio on a rare hypothesis still yields a small posterior.
- 3 **Disjoint $\neq$ independent.** Mutually exclusive events with positive probability are dependent.
- 4 **Pairwise $\neq$ mutual.** Check all $2^n - n - 1$ equations, or trust a trial-based model.
- 5 **Assuming independence.** Multiplying probabilities is only legal if independence holds (or is conditional on the right thing).
- 6 **Simpson's reversal.** Pooled comparisons can flip; condition on the confounder.
- 7 **Conditioning on the wrong event.** "Given at least one head" $\neq$ "given the first coin is a head": $\mathbb{P}(HH | \geq 1H) = \frac{1}{3}$ but $\mathbb{P}(HH | \text{first } H) = \frac{1}{2}$.





5.

Machine learning: Naïve Bayes

Classification is a conditional probability

Data: features $x = (x_1, \dots, x_d)$, label $y \in \{1, \dots, K\}$ (e.g. spam / ham). A classifier wants $\mathbb{P}(y | x)$ — the posterior of the class given what we observe.

Bayes' theorem, applied

$$\mathbb{P}(y | x) = \frac{\mathbb{P}(x | y) \mathbb{P}(y)}{\mathbb{P}(x)} \propto \mathbb{P}(x | y) \mathbb{P}(y).$$

$\mathbb{P}(y)$: class prior (how common is spam). $\mathbb{P}(x | y)$: how likely this email is, if it is spam.

$\mathbb{P}(x)$: the same for all classes — irrelevant for choosing the best y .

The problem

x lives in a huge space (2^d binary feature patterns). We can never estimate $\mathbb{P}(x | y)$ for every pattern from data. We need a simplifying assumption.



The naïve assumption and the classifier

Naïve conditional independence

Given the class, the features are conditionally independent:

$$\mathbb{P}(x | y) = \mathbb{P}(x_1, \dots, x_d | y) = \prod_{j=1}^d \mathbb{P}(x_j | y).$$

This is the chain rule $\mathbb{P}(x | y) = \prod_j \mathbb{P}(x_j | y, x_1, \dots, x_{j-1})$ with every conditional simplified by $x_j \perp\!\!\!\perp (x_1, \dots, x_{j-1}) | y$. Now only $d \cdot K$ one-dimensional tables are needed.

Theorem (Naïve Bayes decision rule)

$$\hat{y}(x) = \arg \max_y \left[\log \mathbb{P}(y) + \sum_{j=1}^d \log \mathbb{P}(x_j | y) \right].$$



Worked example — a tiny spam filter

Vocabulary: {free, meeting}; $x_j = 1$ if the word occurs. From training data:

	spam	ham
$\mathbb{P}(y)$	0.3	0.7
$\mathbb{P}(\text{free} = 1 y)$	0.6	0.05
$\mathbb{P}(\text{meeting} = 1 y)$	0.1	0.5

New email contains free but not meeting: $x = (1, 0)$.

$$\text{spam score} = 0.3 \cdot 0.6 \cdot 0.9 = 0.162, \quad \text{ham score} = 0.7 \cdot 0.05 \cdot 0.5 = 0.0175.$$

$$\mathbb{P}(\text{spam} | x) = \frac{0.162}{0.162 + 0.0175} \approx 0.90 \Rightarrow \hat{y} = \text{spam}.$$

In log-odds: $\log \frac{0.3}{0.7} + \log \frac{0.6}{0.05} + \log \frac{0.9}{0.5} = -0.85 + 2.48 + 0.59 = 2.22$; each word adds its own log-likelihood ratio — the sequential-updating picture from Section 2.



Estimation and the zero-count problem

The tables $\mathbb{P}(x_j | y)$ are estimated from counts: $\hat{\mathbb{P}}(x_j=1 | y) = \frac{\#\{\text{class } y \text{ with } x_j = 1\}}{\#\{\text{class } y\}}$.

Zero counts are fatal

If a word never appeared in training spam, $\hat{\mathbb{P}}(x_j=1 | \text{spam}) = 0$, and a single occurrence forces the spam score to *exactly* 0 — no amount of other evidence can recover. One factor kills the product.

Laplace (additive) smoothing

$$\hat{\mathbb{P}}(x_j=1 | y) = \frac{\#\{y, x_j = 1\} + \alpha}{\#\{y\} + 2\alpha}, \quad \alpha > 0 \text{ (often 1)}.$$

Pretend every outcome was seen α extra times. This is a Bayesian posterior mean under a $\text{Beta}(\alpha, \alpha)$ prior — Lecture 7 makes this precise.

 Takeaway: the model is a product of conditional probabilities, so estimation errors in any factor propagate multiplicatively.

Precision and recall are conditional probabilities

Binary case: true label $y \in \{0, 1\}$, prediction $\hat{y} \in \{0, 1\}$. Over a random test point:

recall (sensitivity, TPR) $\mathbb{P}(\hat{y} = 1 \mid y = 1)$

precision $\mathbb{P}(y = 1 \mid \hat{y} = 1)$

false-positive rate $\mathbb{P}(\hat{y} = 1 \mid y = 0)$

specificity $\mathbb{P}(\hat{y} = 0 \mid y = 0)$

Theorem (Precision via Bayes)

$$\text{precision} = \frac{\text{recall} \cdot \mathbb{P}(y=1)}{\text{recall} \cdot \mathbb{P}(y=1) + \text{FPR} \cdot \mathbb{P}(y=0)}.$$

Proof.

Bayes' theorem with $A = \{\hat{y} = 1\}$ and partition $\{y = 1\}, \{y = 0\}$. □

Why accuracy misleads under class imbalance

The lazy classifier

If 99% of mail is ham, predicting “ham” always has accuracy $\mathbb{P}(\hat{y} = y) = 0.99$ — and recall 0 for spam.

By total probability,

$$\text{accuracy} = \mathbb{P}(\hat{y} = y) = \mathbb{P}(\hat{y}=1 \mid y=1)\mathbb{P}(y=1) + \mathbb{P}(\hat{y}=0 \mid y=0)\mathbb{P}(y=0),$$

a weighted average dominated by the majority class. The minority-class term barely moves it.

- Report the conditional quantities separately (recall, precision, FPR), or a summary that weights classes equally (balanced accuracy, F_1).
- Precision computed on a balanced validation set will *not* transfer to deployment with a different prior — recompute it with the formula on the previous slide.



The Bayes classifier is optimal

Theorem (Bayes decision rule)

Among all classifiers $h : \mathcal{X} \rightarrow \{1, \dots, K\}$, the misclassification probability $\mathbb{P}(h(X) \neq Y)$ is minimized by

$$h^*(x) = \arg \max_y \mathbb{P}(Y = y | X = x).$$

Sketch.

Condition on $X = x$: $\mathbb{P}(h(X) \neq Y | X = x) = 1 - \mathbb{P}(Y = h(x) | X = x)$. For each x this is smallest when $h(x)$ is the most probable class. Averaging over x (total probability) preserves the inequality. □

- $\mathbb{P}(h^*(X) \neq Y)$ is the **Bayes error** — the irreducible error of the problem.
- Naïve Bayes *approximates* h^* by approximating $\mathbb{P}(y | x)$ under the naïve assumption. It is optimal when the assumption holds, and often surprisingly good when it does not — because the argmax only needs the *ranking* of posteriors to be right, not their values.





6.

Wrap-up

Summary

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}; \quad \mathbb{P}(A) = \sum_i \mathbb{P}(A | B_i) \mathbb{P}(B_i); \quad \mathbb{P}(B_j | A) = \frac{\mathbb{P}(A | B_j) \mathbb{P}(B_j)}{\mathbb{P}(A)}; \quad A \perp\!\!\!\perp B \iff \mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B).$$

- $\mathbb{P}(\cdot | B)$ is a probability measure: all of Lecture 1 applies after conditioning.
- Multiplication and chain rules build joint probabilities stage by stage.
- Total probability averages over cases; Bayes inverts the conditioning. In odds: posterior = LR $\times$ prior.
- Independence is a numerical property; disjoint $\neq$ independent; pairwise $\neq$ mutual; conditional independence is a different notion again.
- Traps: base rates, inverse confusion, Simpson's reversal.
- ML: Naïve Bayes = Bayes + conditional independence; precision depends on the prior; the Bayes rule is optimal.



Seminar problems & next lecture

- 1 Prove: if $\mathbb{P}(A \mid B) > \mathbb{P}(A)$ then $\mathbb{P}(B \mid A) > \mathbb{P}(B)$ and $\mathbb{P}(A \mid B^c) < \mathbb{P}(A)$.
- 2 Two dice. Find $\mathbb{P}(\text{sum} \geq 10 \mid \text{first} \geq 4)$ and $\mathbb{P}(\text{first} \geq 4 \mid \text{sum} \geq 10)$.
- 3 Three boxes contain (2 gold), (1 gold, 1 silver), (2 silver). A box is chosen at random and a gold coin is drawn. What is the probability the other coin in that box is gold? (Bertrand's box.)
- 4 Show that $A_1, \dots, A_n$ mutually independent implies $A_1^c, A_2, \dots, A_n$ mutually independent.
- 5 A test has sensitivity 0.95 and specificity 0.90. For which prevalence is $\mathbb{P}(D \mid +) = \frac{1}{2}$?
- 6 Build a Simpson's-paradox table with two universities and admission rates by gender.
- 7 Naïve Bayes with three binary words: compute the posterior for a given email and show the effect of $\alpha = 1$ smoothing when one count is zero.

Next: random variables, expectation, and variance.

