



The Probability Space

Axioms, Interpretations & Combinatorics

Probability Theory • Lecture 1

Applied Data Analysis — Module I

September 2026

Welcome to Probability Theory!

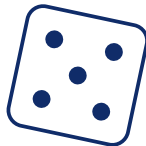
- Two modules, twenty lectures: from coin flips to the Central Limit Theorem, and on to estimators, likelihood, and EM.
- Every lecture ends with a machine-learning tie-in — probability *for data analysts*, not for its own sake.
- Definitions, theorems, *and* proofs — and you will be asked to the board. A brave wrong attempt earns more credit than silence.

Ground rule

Probability is the only course where “I don’t know” is a rigorous answer — provided you can say *exactly how much* you don’t know.



Six? Again?!
*I'm not biased — you're just
bad at small samples.*



Warning: this course has probability 1 of
containing at least one dice joke.
That was it.

Course overview: two modules, 20 lectures

Module I — Probability Theory

September–December 2026

- 1 The probability space, axioms & combinatorics
- 2 Conditional probability, Bayes & independence
- 3 Discrete random variables: $\mathbb{E}$, Var , indicators
- 4 Essential discrete distributions
- 5 Generating functions & Chernoff bounds
- 6 Continuous random variables & inverse sampling
- 7 Gamma, Beta, conjugacy & heavy tails
- 8 The Normal distribution & least squares
- 9 Simulation & Monte Carlo
- 10 Joint distributions, covariance & PCA
- 11 Conditional expectation & bias–variance
- 12 Famous problems & Benford's law
- 13 Concentration inequalities & the LLN
- 14 The Central Limit Theorem

Module II — Mathematical Statistics

Winter 2027

- 1 Sampling distributions: χ^2 , t , F
- 2 Sufficient statistics & exponential families
- 3 What makes a “good” estimator
- 4 Building estimators: MOM & MLE
- 5 The EM algorithm
- 6 Fisher information, Cramér–Rao & optimality

Running thread

Every lecture closes with an ML tie-in: from Naïve Bayes and PCA to generalization bounds, GLMs, and Gaussian mixtures.



Course assessment

The final grade is a weighted sum of four components:

$$O_{\text{final}} = 0.15 \cdot O_{\text{HW}} + 0.15 \cdot O_{\text{seminar}} + 0.25 \cdot O_{\text{midterm}} + 0.45 \cdot O_{\text{exam}}$$

- O_{HW} — homework assignments (15%).
- O_{seminar} — seminar participation: substantive work at the board, not attendance (15%).
- O_{midterm} — written midterm test (25%).
- O_{exam} — final written exam (45%).

All grades are on the 10-point scale; the weighted sum is rounded arithmetically.

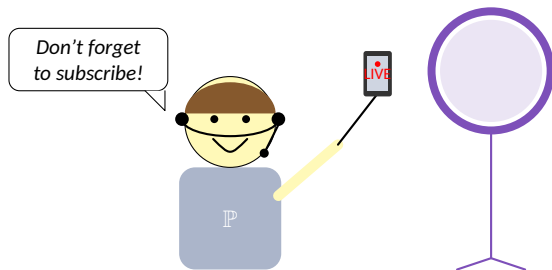


Course Telegram group



Scan to join

Slides, problem sets, announcements,
and deadlines are posted there first.



Main textbooks

- Newbold, P., Carlson, W.L., Thorne, B.M. *Statistics for Business and Economics*.
- Gmurman, V.E. *Fundamentals of Probability Theory and Mathematical Statistics*.
- Freedman, D., Pisani, R., Purves, R. *Statistics*.
- Grimmett, G., Stirzaker, D. *Probability and Random Processes*, 3rd ed. OUP, 2001.
- Casella, G., Berger, R.L. *Statistical Inference*, 2nd ed. Duxbury, 2002.

Supplementary — University of London / LSE subject guides

- Abdey, J.S. *Statistics 1* (ST104a), 2014.
- Abdey, J.S. *Statistics 2* (ST104b), 2014.



Roadmap & learning objectives

We build the object every later result rests on: a *probability measure* on a σ -*algebra* of events.

- State **Kolmogorov's axioms** and the triple $(\Omega, \mathcal{F}, \mathbb{P})$.
- Derive the basic **properties** of $\mathbb{P}$ (complement, monotonicity, inclusion-exclusion, continuity).
- Compare the three **interpretations**: classical, frequentist, subjective.
- Review the **counting** tools used all term.
- First ML tie-in: **counting hypotheses** and why $\log |\mathcal{H}|$ appears in learning bounds.





1.

Foundations

Why a data analyst needs measure-theoretic care

A dataset is a **sample** from an unknown distribution:

$$(x_1, y_1), \dots, (x_n, y_n) \stackrel{\text{i.i.d.}}{\sim} \mathbb{P} \quad \text{on } \mathcal{X} \times \mathcal{Y}.$$

- “Learning” = using the sample to infer properties of $\mathbb{P}$.
- Every generalization guarantee — *test error* $\approx$ *train error* — is a statement about $\mathbb{P}$.
- To make such statements provable we first need a rigorous notion of $\mathbb{P}$.

Takeaway

Probability is the language in which *all* learning guarantees are written.



What is probability? Three everyday examples

A fair die

Six faces, all alike.

$\mathbb{P}(\text{even}) = \frac{3}{6} = \frac{1}{2}$: favorable outcomes over all outcomes.

A basketball player

412 of 500 free throws this season: her chance on the next one is about 0.82 — a *long-run frequency*.

Tomorrow's weather

“70% chance of rain.”

Tomorrow happens only once — a *degree of belief* based on today's data.

Three very different situations, one common language:

- a number between 0 (impossible) and 1 (certain);
- “one of two things that can't both happen” gets the *sum* of the two numbers;
- everything that can happen, taken together, gets exactly 1.

Where we are heading

A dataset is like the free-throw record: a sample from something unknown. To say anything reliable about it, we first need precise rules for these numbers.



Sample space and events

Definition (Sample space)

The **sample space** Ω is the set of all outcomes ω ; an **event** is a subset $A \subseteq \Omega$.

- Ω = certain event; $\emptyset$ = impossible event.
- Elementary event $\{\omega\}$ vs. compound event.

Example — two coin flips

$\Omega = \{HH, HT, TH, TT\}$; “at least one head” is $A = \{HH, HT, TH\}$.



From examples to rules: what every probability must satisfy

Two coin flips, $\Omega = \{HH, HT, TH, TT\}$, each outcome equally likely, so $\mathbb{P}(\{\omega\}) = \frac{1}{4}$.

Observation 1 — never negative

Every event gets a number ≥ 0 :

$$\mathbb{P}(\text{two heads}) = \frac{1}{4}, \mathbb{P}(\emptyset) = 0.$$

Observation 2 — the whole space is certain

$$\mathbb{P}(\Omega) = \mathbb{P}(\{HH, HT, TH, TT\}) = \frac{4}{4} = 1.$$

Observation 3 — disjoint events add

$A = \{HH\}$, $B = \{TT\}$ cannot happen together:

$$\mathbb{P}(A \cup B) = \frac{2}{4} = \mathbb{P}(A) + \mathbb{P}(B).$$

But $A = \{HH, HT\}$ and $C = \{HH, TH\}$ overlap in

$$HH: \mathbb{P}(A \cup C) = \frac{3}{4} \neq \frac{1}{2} + \frac{1}{2}.$$

- The same three facts hold for the die, the free throws, the weather — for *any* sensible assignment.
- Kolmogorov (1933): take these three facts as the **definition** of probability; derive everything else.
- Notation: $\mathcal{F}$ = the collection of events we assign probabilities to. For finite Ω it is simply *all* subsets; subtleties appear only for infinite Ω .



Kolmogorov's axioms

Definition (Probability measure)

A function $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ satisfying:

$$(K1) \mathbb{P}(A) \geq 0; \quad (K2) \mathbb{P}(\Omega) = 1; \quad (K3) A_i \cap A_j = \emptyset \Rightarrow \mathbb{P}\left(\bigsqcup_n A_n\right) = \sum_n \mathbb{P}(A_n).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a **probability space**. Everything else is a theorem.



Properties derived from the axioms

Theorem

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(A^c) = 1 - \mathbb{P}(A), \quad A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof.

$\Omega = A \sqcup A^c$, so by (K3),(K2): $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$, giving $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and (with $A = \Omega$) $\mathbb{P}(\emptyset) = 0$. Monotonicity: $B = A \sqcup (B \setminus A)$ and $\mathbb{P} \geq 0$. □



Inclusion-exclusion

Theorem

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i < j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1} \mathbb{P}\left(\bigcap_i A_i\right).$$

Case $n = 2$.

$A \cup B = A \sqcup (B \setminus A)$ gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$; and $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
Hence $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$; induction for general n . $\square$



Continuity of probability

Theorem (Continuity from below)

$$A_1 \subseteq A_2 \subseteq \cdots, A = \bigcup_n A_n \Rightarrow \mathbb{P}(A_n) \rightarrow \mathbb{P}(A).$$

Proof.

Disjointify $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$. Then $A = \bigsqcup_k B_k$, $A_n = \bigsqcup_{k \leq n} B_k$, so
 $\mathbb{P}(A) = \sum_k \mathbb{P}(B_k) = \lim_n \sum_{k \leq n} \mathbb{P}(B_k) = \lim_n \mathbb{P}(A_n).$ □

A symmetric argument gives continuity from above.



Three interpretations, one axiom system

Classical (Laplace)

Equally likely outcomes: $\mathbb{P}(A) = |A|/|\Omega|$. Pure counting.

Frequentist

$\mathbb{P}(A) = \lim_{n \rightarrow \infty} \#\{A \text{ in } n\}/n$ (justified later by the LLN).

Subjective (Bayesian)

Degree of belief, updated by data via Bayes (Lecture 2) — the engine of Bayesian ML.

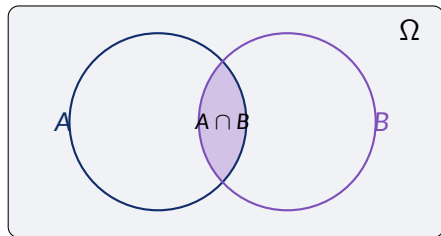




2.

Counting

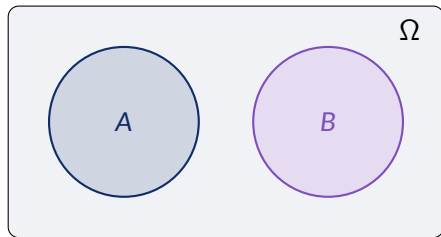
Geometric intuition — probability as area



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

Inclusion-exclusion: adding the two areas double-counts the overlap.

Disjoint events — no overlap at all



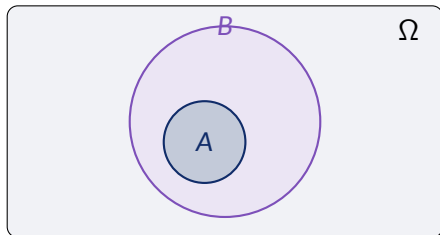
$A \cap B = \emptyset \iff A \text{ and } B \text{ are mutually exclusive.}$

The empty set

$\emptyset = \{ \}$ is the unique set with no elements — the *impossible* (null) event, and $\mathbb{P}(\emptyset) = 0$.



Subevents — one event inside another



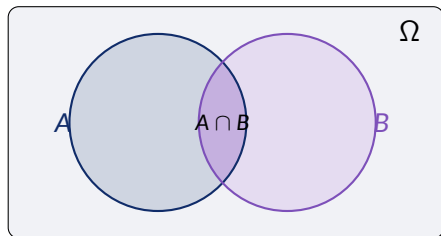
$A \subseteq B$: every outcome of A is also in B , so if A happens, B happens too.

Example — expulsion vs. failing

Mark asks which is more likely: to fail Statistics (F : grade < 3) or to be expelled (E). You can only be expelled *after* failing, so $E \subseteq F$ and hence $\mathbb{P}(E) \leq \mathbb{P}(F)$ — no numbers needed.



Union probability — the additive law



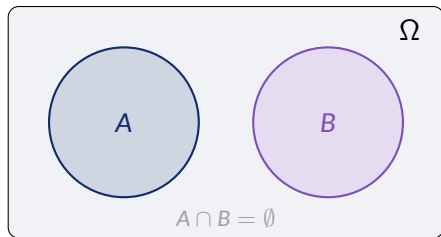
For any events A, B : $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

Split $A \cup B$ into disjoint pieces: $A \cup B = A \sqcup (B \setminus A)$ and $B = (A \cap B) \sqcup (B \setminus A)$. By (K3), $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$ and $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Substitute. $\square$



Disjoint events — the additive law simplifies

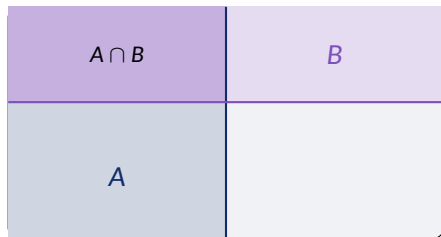


If $A \cap B = \emptyset$: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

No overlap means nothing is double-counted: the subtracted term $\mathbb{P}(A \cap B)$ is simply 0.
This is axiom (K3) for two events — and the reason we can add probabilities of a die showing 1 or 2.



Independent events — the multiplicative law



If A and B are **independent**: $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Independence is not visible as “no overlap” — it is a statement about *numbers*: knowing that B happened does not change the chance of A . In the picture, A takes half the width and B takes $\frac{0.7}{1.7}$ of the height, so $A \cap B$ takes exactly the product of the two fractions.



Sampling — four ways to draw objects

Definition

Sampling is choosing elements (objects) from a set of outcomes; the result is a **sample**, e.g. (A, B, C) .

with replacement each selected object is put back before the next draw	ordered (A, B, C) and (A, C, B) are not the same
without replacement objects are not returned	unordered (A, B, C) and (A, C, B) are the same

Two independent choices give four counting situations — the next slides count each of them.



Counting I — the multiplication principle

Multiplication principle

k independent stages with $n_1, \dots, n_k$ choices give $n_1 \cdots n_k$ outcomes.

- Permutations of n items: $n!$; k -permutations: $P(n, k) = \frac{n!}{(n-k)!}$.

Check

Ordered top-3 among 8 runners: $P(8, 3) = 8 \cdot 7 \cdot 6 = 336$.



Counting II — combinations & multinomials

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad \binom{n}{k_1, \dots, k_m} = \frac{n!}{k_1! \cdots k_m!}.$$

- Combinations: unordered selections; multinomial: split n into labeled groups.
- **Stars & bars:** nonnegative solutions of $x_1 + \cdots + x_m = n$ number $\binom{n+m-1}{m-1}$.



Two identities, proved by counting

Theorem (Binomial theorem)

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

Theorem (Vandermonde)

$$\sum_{k=0}^r \binom{m}{k} \binom{n}{r-k} = \binom{m+n}{r}.$$

Combinatorial proof of Vandermonde.

Choose r from m men and n women: the RHS counts directly; the LHS splits by the number k of men chosen. The counts agree. □



Worked example 1 — the birthday problem

n people, 365 equally likely birthdays:

$$\mathbb{P}(\text{all distinct}) = \prod_{i=0}^{n-1} \frac{365-i}{365}, \quad \mathbb{P}(\text{a match}) = 1 - \prod_{i=0}^{n-1} \frac{365-i}{365}.$$

- Complementary counting turns a hard event into an easy product.
- At $n = 23$: $\mathbb{P}(\text{match}) \approx 0.507$; the threshold sits near $n \approx 1.18\sqrt{365}$ — a $\sqrt{\cdot}$ scale we meet again (hash collisions, Lecture 12).



Worked example 2 — sampling a batch (quality control)

Batch of N with K defective; draw n **without replacement**:

$$\mathbb{P}(\text{none defective}) = \frac{\binom{N-K}{n}}{\binom{N}{n}}, \quad \mathbb{P}(\geq 1) = 1 - \frac{\binom{N-K}{n}}{\binom{N}{n}}.$$

Numbers

$N = 100, K = 5, n = 10$: $\mathbb{P}(\text{none}) = \binom{95}{10} / \binom{100}{10} \approx 0.584$, so $\mathbb{P}(\geq 1) \approx 0.416$.

This is the *hypergeometric* model (Lecture 4).



Worked example 3 — count it two ways

From 10 people, form a committee of 4 and name a chair:

$$\underbrace{\binom{10}{4} \cdot 4}_{4, \text{ then a chair}} = \underbrace{10 \cdot \binom{9}{3}}_{\text{chair, then 3}} = 840.$$

Two counting stories, one number — a built-in correctness check.





3.

From counting to machine learning

ML application — the size of a hypothesis space

Binary features $x \in \{0, 1\}^d$: 2^d distinct inputs.

- **All labelings**: 2^{2^d} Boolean functions.
- **Conjunctions** over d literals: 3^d (each feature positive / negated / ignored).
- **Decision stumps** on m features: $O(mn)$ candidates.

The point

Model *capacity* is a counting quantity.



Summary

$$(\Omega, \mathcal{F}, \mathbb{P}); \quad \mathbb{P}(A^c) = 1 - \mathbb{P}(A); \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

- Axioms (K1)–(K3) $\Rightarrow$ complement, monotonicity, inclusion–exclusion, continuity.
- Three interpretations, one axiom system; counting is the classical engine.
- ML link: capacity is a count; $\log |\mathcal{H}|$ drives generalization.



Seminar problems & next lecture

- 1 Derive $\mathbb{P}(A \cup B \cup C)$ from the $n = 2$ case.
- 2 Fair die twice: $\mathbb{P}(\text{sum} = 7)$, $\mathbb{P}(\text{max} = 4)$.
- 3 Show $\sum_k \binom{n}{k} = 2^n$ by counting.
- 4 Smallest n with birthday-match probability ≥ 0.9 ?

Next: conditional probability, Bayes, and independence.

