ECON 671 — Metrics

Expanded Notes

Week 1 - Class 1

Sample space and events

Definition (Sample space)**.** The set of all possible outcomes of an experiment is the *sample space* S.

Definition (Event)**.** An *event* is any subset $A \subseteq S$. Event A occurs if the realized outcome $s \in S$ lies in A.

Example. Fair die: $S = \{1, 2, 3, 4, 5, 6\}$. Two coin tosses: $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$.

Countable and uncountable sets

Definition (At most countable and countably infinite)**.** A set A is *at most countable* if it is finite or there exists a bijection $f : A \to B$ with some subset $B \subseteq \mathbb{N}$ (equivalently, a subset of $\mathbb{Z}$). If A is infinite and there is a bijection $A \to \mathbb{N}$, then A is *countably infinite*.

Proposition ($\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality)**.** *There exists a bijection $g : \mathbb{N} \to \mathbb{Z}$, for instance*

$$g(0) = 0, \qquad g(2k-1) = k, \qquad g(2k) = -k \quad (k \in \mathbb{N},\ k \geq 1).$$

Proof. Surjectivity: every $z \in \mathbb{Z}$ is hit by g (positives via $2z - 1$, negatives via $2|z|$, and 0 via 0). Injectivity: distinct n map to distinct elements because the images fall in disjoint blocks $\{0\}$, $\{1, 2, 3, \ldots\}$, and $\{-1, -2, -3, \ldots\}$. □

Theorem 1 (Cantor: $(0, 1)$ is uncountable)**.** *There is no bijection between $(0, 1)$ and $\mathbb{N}$. In particular, $\mathbb{R}$ is uncountable.*

Diagonal argument. Assume $(0, 1) = \{x_1, x_2, \ldots\}$ is a list. Write $x_n = 0.d_{n1}d_{n2}d_{n3}\ldots$ in decimal form, choosing representations that do not end with a tail of 9's. Define a new number $y = 0.c_1c_2c_3\ldots$ by taking $c_n \in \{1, 2\}$ with $c_n \neq d_{nn}$. Then $y \in (0, 1)$ and y differs from each x_n in the n-th digit, so $y \neq x_n$ for all n, a contradiction. Hence $(0, 1)$ is uncountable. □

Remark. Any nondegenerate interval $[a, b]$ is uncountable (there is a bijection with $(0, 1)$ via an affine map).

Set operations

For $A, B, C \subseteq S$:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}, \quad A \cap B = \{x : x \in A \text{ and } x \in B\}, \quad A^c = \{x \in S : x \notin A\}.$$

Theorem 2 (Algebra of sets)**.** *For all $A, B, C \subseteq S$:*

a) ***Commutativity:*** *$A \cup B = B \cup A$ and $A \cap B = B \cap A$.*

b) ***Associativity:*** *$A \cup (B \cup C) = (A \cup B) \cup C$ and similarly for $\cap$.*

c) ***Distributive laws:*** *$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.*

d) ***De Morgan (finite):*** *$(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.*

Proof of (a): commutativity. Show $A \cup B \subseteq B \cup A$. If $x \in A \cup B$, then $x \in A$ or $x \in B$, hence $x \in B \cup A$. The reverse inclusion is identical. For intersections: if $x \in A \cap B$ then $x \in A$ and $x \in B$, so $x \in B \cap A$, and conversely. □

Proof of (d): De Morgan (finite). We prove $(A \cup B)^c = A^c \cap B^c$ by double inclusion. If $x \in (A \cup B)^c$, then $x \notin A$ and $x \notin B$, so $x \in A^c \cap B^c$. Conversely, if $x \in A^c \cap B^c$ then $x \notin A$ and $x \notin B$, hence $x \notin A \cup B$, i.e., $x \in (A \cup B)^c$. The other identity is analogous. □

Countable unions and intersections

For a family $\{A_i\}_{i \geq 1}$ of subsets of S:

$$\bigcup_{i=1}^{\infty} A_i = \{x : \exists i,\ x \in A_i\}, \qquad \bigcap_{i=1}^{\infty} A_i = \{x : \forall i,\ x \in A_i\}.$$

Example. In $S = (0, 1]$, let $A_i = [1/i, 1]$. Then

$$\bigcup_{i=1}^{\infty} A_i = (0, 1] \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i = \{1\}.$$

Indeed, any $x \in (0, 1]$ belongs to A_i for large enough i, while any $x < 1$ eventually falls outside A_i when $i > 1/x$.

Theorem 3 (De Morgan (general))**.** *For any index set Γ and family $\{A_i\}_{i \in \Gamma}$:*

$$\Big(\bigcup_{i \in \Gamma} A_i\Big)^c = \bigcap_{i \in \Gamma} A_i^c, \qquad \Big(\bigcap_{i \in \Gamma} A_i\Big)^c = \bigcup_{i \in \Gamma} A_i^c.$$

Proof. Using quantifiers:

$$x \in \Big(\bigcup_i A_i\Big)^c \iff \neg(\exists i:\ x \in A_i) \iff (\forall i:\ x \notin A_i) \iff x \in \bigcap_i A_i^c.$$

The other identity follows by negating the universal quantifier. □

Disjointness and partitions

Definition (Disjoint sets)**.** A and B are *disjoint* if $A \cap B = \varnothing$. A family $\{B_i\}$ is *pairwise disjoint* if $B_i \cap B_j = \varnothing$ for $i \neq j$.

Definition (Partition)**.** A family $\{B_i\}_{i\in I}$ is a *partition* of S if (i) it is pairwise disjoint and (ii) $\bigcup_{i\in I} B_i = S$.

Theorem 4 (Partitioning theorem)**.** *If $\{B_i\}_{i\in I}$ is a partition of S, then for every $A \subseteq S$:*

1) $A = \bigcup_{i\in I}(A \cap B_i)$.
2) *The sets $A_i := A \cap B_i$ are pairwise disjoint.*

Proof of 1). ($\subseteq$) Take $x \in A$. Since $\{B_i\}$ partitions S, there is a unique i with $x \in B_i$. Then $x \in A \cap B_i \subseteq \bigcup_i (A \cap B_i)$. ($\supseteq$) If $x \in \bigcup_i (A \cap B_i)$, some i satisfies $x \in A \cap B_i$, hence $x \in A$. □

Proof of 2). If $i \neq j$ and $x \in (A \cap B_i) \cap (A \cap B_j)$, then $x \in B_i \cap B_j = \varnothing$, a contradiction. Thus the intersections are empty. □

Images and preimages

Definition (Image and preimage)**.** Let $f : A \to B$. For $Y \subseteq A$ and $X \subseteq B$,

$$f(Y) = \{f(y) : y \in Y\}, \qquad f^{-1}(X) = \{a \in A :\ f(a) \in X\}.$$

The preimage is *always* defined, even if f is not invertible.

Example. If $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$, then $f^{-1}(\{-1\}) = \varnothing$, $f^{-1}(\{0\}) = \{0\}$, and $f^{-1}([1,4]) = [-2,-1] \cup [1,2]$.

Proposition (Image/preimage laws)**.** *For $Y, Z \subseteq A$ and $X, W \subseteq B$:*

$$f(Y \cup Z) = f(Y) \cup f(Z), \qquad f(Y \cap Z) \subseteq f(Y) \cap f(Z),$$
$$f^{-1}(X \cup W) = f^{-1}(X) \cup f^{-1}(W), \qquad f^{-1}(X \cap W) = f^{-1}(X) \cap f^{-1}(W),$$
$$f^{-1}(X^c) = \big(f^{-1}(X)\big)^c.$$

Moreover, $f(Y \cap Z) \subseteq f(Y) \cap f(Z)$ can be strict when f is not injective.

Proof. All identities (and the inclusion) follow by double inclusion from the definitions. For instance, if $a \in f^{-1}(X \cup W)$, then $f(a) \in X \cup W$, i.e., $f(a) \in X$ or $f(a) \in W$, hence $a \in f^{-1}(X)$ or $a \in f^{-1}(W)$, so $a \in f^{-1}(X) \cup f^{-1}(W)$. $\square$

Week 1 — Class 2

Sets, maps, image and preimage

Let $f : A \to B$ be any map between sets (read carefully the domain and codomain).

- For $Y \subseteq A$, the **image** is $f(Y) = \{f(y) : y \in Y\} \subseteq B$.
- For $X \subseteq B$, the **preimage** is $f^{-1}(X) = \{a \in A : f(a) \in X\} \subseteq A$.

Preimages exist for any f (no invertibility needed) and are the key notion in measurability.

Example. Let $f : \{1,2,3\} \to \{a,b\}$ with $f(1) = a$, $f(2) = a$, $f(3) = b$. Then $f(\{1,3\}) = \{a,b\}$ and $f^{-1}(\{a\}) = \{1,2\}$.

The σ-algebras and power sets

Let S be a base set. Its power set $\mathcal{P}(S)$ is the collection of *all* subsets of S. If S has a countable number of elements, say N, the power set has 2^N elements.

A collection $\mathcal{B} \subseteq \mathcal{P}(S)$ is a σ**-algebra** if:

1. $\varnothing \in \mathcal{B}$ and $S \in \mathcal{B}$,
2. if $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$,
3. if $A_1, A_2, \dots \in \mathcal{B}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{B}$.

Key 1: The first to properties means $S \in \mathcal{B}$ because $\varnothing^c = S$.

Key 2: By De Morgan, $\mathcal{B}$ is also closed under countable intersections.

Proof. By De Morgan's law,

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c \right)^c .$$

Since $A_n \in \mathcal{B}$, we have $A_n^c \in \mathcal{B}$ (closure under complements), and since $\mathcal{B}$ is closed under countable unions, $\bigcup_{n=1}^{\infty} A_n^c \in \mathcal{B}$. Taking the complement once more keeps us in $\mathcal{B}$, proving the claim. □

Smallest and largest. The smallest σ-algebra on S is $\{\varnothing, S\}$; the largest is $\mathcal{P}(S)$.

Key question: "Do we need $\mathcal{P}(S)$ to be countable?" No. "Countable" in the definition refers to the *operations* (countable unions/intersections), not to the *size* of the collection. $\mathcal{P}(S)$ can be uncountable and still be a perfectly valid σ-algebra. In practice, when S is uncountable (e.g., $S = \mathbb{R}$), we typically do *not* use $\mathcal{P}(S)$ because it contains non-measurable sets; we work with a manageable σ-algebra such as the Borel σ-algebra (or its completion under Lebesgue measure).

Example (Finite S)**.** If $S = \{1, 2, 3\}$, then $\mathcal{B} = \{\varnothing, \{1\}, \{2, 3\}, S\}$ is a σ-algebra: check complements and (finite/ countable) unions.

Proposition (Intersection of σ-algebras)**.** *Let $\{\mathcal{A}_i\}_{i \in I}$ be σ-algebras on the same base set S, and define*

$$\mathcal{B} := \bigcap_{i \in I} \mathcal{A}_i = \{A \subseteq S : A \in \mathcal{A}_i \text{ for all } i \in I\}.$$

Then $\mathcal{B}$ is a σ-algebra on S.

Proof. We verify the three axioms algebraically for $\mathcal{B}$.

(1) $\varnothing, S \in \mathcal{B}$. Since each $\mathcal{A}_i$ is a σ-algebra on S, $\varnothing \in \mathcal{A}_i$ and $S \in \mathcal{A}_i$ for every $i \in I$. Because they are in every set, $\varnothing, S \in \bigcap_{i \in I} \mathcal{A}_i = \mathcal{B}$.

(2) Closure under complements. Let $A \in \mathcal{B}$. Because A is in every set, $A \in \mathcal{A}_i$ for every $i \in I$. Because each $\mathcal{A}_i$ is a σ-algebra, $A^c \in \mathcal{A}_i$ for every $i \in I$. It follows that, $A^c \in \bigcap_{i \in I} \mathcal{A}_i = \mathcal{B}$.

(3) Closure under countable unions. Let $(A_n : n \in \mathbb{N}) \subseteq \mathcal{B}$. For each n and each $i \in I$ we have $A_n \in \mathcal{A}_i$. Since every $\mathcal{A}_i$ is a σ-algebra, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_i$ for every $i \in I$. By last, $\bigcup_{n=1}^{\infty} A_n \in \bigcap_{i \in I} \mathcal{A}_i = \mathcal{B}$.

$\mathcal{B}$ is a σ-algebra on S. □

Generated σ-algebras

Two levels. Fix a base set S.

- Elements of S are points $x \in S$.
- Elements of $\mathcal{P}(S)$ are *sets of points* $A \subseteq S$.
- A σ-algebra $\mathcal{A}$ is a *set of sets of points*, i.e. $\mathcal{A} \subseteq \mathcal{P}(S)$.
- A generator $\mathcal{M} \subseteq \mathcal{P}(S)$ is also a *set of sets of points*.

So the relation $\mathcal{M} \subseteq \mathcal{A}$ is a subset relation between two *collections of subsets of S*.

Definition (generated σ-algebra). Given $\mathcal{M} \subseteq \mathcal{P}(S)$, define the family of all σ-algebras that contain $\mathcal{M}$:

$$\mathsf{S}(\mathcal{M}) := \Big\{ \mathcal{A} \subseteq \mathcal{P}(S) \ : \ \mathcal{A} \text{ is a } \sigma\text{-algebra on } S \text{ and } \mathcal{M} \subseteq \mathcal{A} \Big\}.$$

The σ**-algebra generated by** $\mathcal{M}$ is

$$\sigma(\mathcal{M}) := \bigcap_{\mathcal{A} \in \mathsf{S}(\mathcal{M})} \mathcal{A}.$$

Why this intersection is "the smallest". (i) $\mathcal{M} \subseteq \sigma(\mathcal{M})$ because every $\mathcal{A} \in \mathsf{S}(\mathcal{M})$ contains $\mathcal{M}$.

(ii) If $\mathcal{B}$ is any σ-algebra with $\mathcal{M} \subseteq \mathcal{B}$, then $\mathcal{B} \in \mathsf{S}(\mathcal{M})$, hence $\sigma(\mathcal{M}) \subseteq \mathcal{B}$.
Together, (i)–(ii) show $\sigma(\mathcal{M})$ is the unique *smallest* σ-algebra containing $\mathcal{M}$.

Example (Singleton generator on a finite set)**.** Let $S = \{1,2,3\}$ and $\mathcal{M} = \{\{1\}\}$. Start with $\{\varnothing, S\} \cup \mathcal{M} = \{\varnothing, S, \{1\}\}$. Close under complements: add $\{1\}^c = \{2,3\}$. Close under unions/intersections: with $\{\varnothing, S, \{1\}, \{2,3\}\}$, any union/intersection stays in the same four sets. No new sets appear, hence

$$\sigma(\mathcal{M}) = \{\varnothing, \{1\}, \{2,3\}, S\}.$$

What is an element of what?

$$\underbrace{1,2,3}_{\in S} \quad \in \quad \underbrace{\{1\},\{2,3\}}_{\in \mathcal{P}(S)} \quad \in \quad \underbrace{\{\varnothing,\{1\},\{2,3\},S\}}_{=\sigma(\mathcal{M}) \subseteq \mathcal{P}(S)}.$$

Here, $\{1\}$ is an *element* of $\sigma(\mathcal{M})$; $\sigma(\mathcal{M})$ is a *subset* of $\mathcal{P}(S)$.

Why more generators can explode to $\mathcal{P}(S)$? If $\mathcal{M} = \{\{1\}, \{2\}\}$ on the same S, then complements add $\{2,3\}$ and $\{1,3\}$; unions/intersections generate $\{3\}$ and every other subset; hence $\sigma(\mathcal{M}) = \mathcal{P}(S)$.

Borel as a generated $\sigma - algebra$ (notation mirror). Let $\mathcal{G} = \{(a,b) : a < b,\ a,b \in \mathbb{R}\}$ (all open intervals). Then

$$\mathsf{S}(\mathcal{G}) = \{\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}) \ : \ \mathcal{A} \text{ is a } \sigma\text{-algebra and } \mathcal{G} \subseteq \mathcal{A}\}, \qquad \mathcal{B}(\mathbb{R}) = \bigcap_{\mathcal{A} \in \mathsf{S}(\mathcal{G})} \mathcal{A}.$$

This explicitly encodes "the smallest σ-algebra containing all open intervals".[1]

Borel on $\mathbb{R}$. The **Borel σ-algebra** $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by all open intervals (a,b). It contains open and closed sets, half-open intervals, countable unions/intersections of those, and all sets obtainable from them by taking complements. (In metric spaces, one may equivalently generate with open balls.)

Why not use $\mathcal{P}(\mathbb{R})$? Because it is "too large": it contains pathological non-measurable sets for which a reasonable measure (like Lebesgue) cannot be defined consistently. Borel sets strike a balance between expressiveness and tractability.

[1] Note to future me: I am not fully sure of understanding this properly. It might be helpful to revise and make an intuition.

Measures and measure spaces

Definition. (Measurable space, measurable sets)
Fix a sample space S. If $\mathcal{B}$ is a σ-algebra, then we call the pair $(S, \mathcal{B})$ a *measurable space*, and the elements of $\mathcal{B}$ are called *measurable sets*.

A **measure space** is a triple $(S, \mathcal{B}, \mu)$ where $\mathcal{B}$ is a σ-algebra on S and $\mu : \mathcal{B} \to [0, \infty) \cup \{\infty\}$ satisfies:

1. $\mu(\varnothing) = 0$
2. $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ for disjoint $A_n \in \mathcal{B}$

Example. Some common measure spaces:

- **Counting measure** on a countable S: $\mu(A) = |A|$ is the cardinality (possibly ∞).
- **Dirac measure** at $x \in S$: $\varepsilon_x(A) = \mathbf{1}\{x \in A\}$ (this is a probability measure). Don't get it.
- **Lebesgue measure** λ on $\mathbb{R}^n$: generalizes length/area/volume; e.g. $\lambda((a, b]) = b - a$ on $\mathbb{R}$.

Measurable maps

Definition. Let $(S_1, \mathcal{B}_1)$ and $(S_2, \mathcal{B}_2)$ be measurable spaces.[2] A map $f : S_1 \to S_2$ is **measurable (w.r.t. $\mathcal{B}_1, \mathcal{B}_2$)** if

$$f^{-1}(A_2) \in \mathcal{B}_1 \qquad \text{for all } A_2 \in \mathcal{B}_2.$$

That is, the preimages of measurable sets are measurable.

Remark (Terminology)**.** When the codomain is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (or $\overline{\mathbb{R}}$), one usually says *measurable function* rather than measurable map.

Minimal intuition. "Events" live in the codomain: $A_2 \in \mathcal{B}_2$. Measurability says: pulling events back through f gives events in the domain: $f^{-1}(A_2) \in \mathcal{B}_1$. This is why a random variable $X : \Omega \to \mathbb{R}$ is defined as a measurable map $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Preimage calculus. For any map f and sets A, B:

$$f^{-1}\Big(\bigcup_i A_i\Big) = \bigcup_i f^{-1}(A_i), \qquad f^{-1}\Big(\bigcap_i A_i\Big) = \bigcap_i f^{-1}(A_i), \qquad f^{-1}(A^c) = \big(f^{-1}(A)\big)^c.$$

Hence if f^{-1} sends a *generator* of $\mathcal{B}_2$ into $\mathcal{B}_1$, then it sends all of $\mathcal{B}_2$ into $\mathcal{B}_1$ (closure under countable unions/complements).

[2] Typo check for future me: it's *metric* spaces and *Borel* σ-algebra.

Example (Indicator functions). Let $(\Omega, \mathcal{B}_1)$ be measurable and $(\mathbb{R}, \mathcal{B}_2)$ with Borel sets. For a measurable set $B \in \mathcal{B}_1$, define $I_B : \Omega \to \mathbb{R}$ by

$$I_B(\omega) = \begin{cases} 1, & \omega \in B, \\ 0, & \omega \notin B. \end{cases}$$

Since I_B only takes values in $\{0, 1\}$, for any $A_2 \in \mathcal{B}_2$,

$$I_B^{-1}(A_2) = \begin{cases} \varnothing, & A_2 \cap \{0,1\} = \varnothing, \\ B, & A_2 \cap \{0,1\} = \{1\}, \\ B^c, & A_2 \cap \{0,1\} = \{0\}, \\ \Omega, & A_2 \cap \{0,1\} = \{0,1\}. \end{cases}$$

Each right-hand set is in $\mathcal{B}_1$ (since $B \in \mathcal{B}_1$ and $\mathcal{B}_1$ is closed under complements; $\varnothing, \Omega \in \mathcal{B}_1$), so I_B is measurable. *Equivalently:* $\{I_B = 1\} = B \in \mathcal{B}_1$ and $\{I_B = 0\} = B^c \in \mathcal{B}_1$.

Simple (step) functions

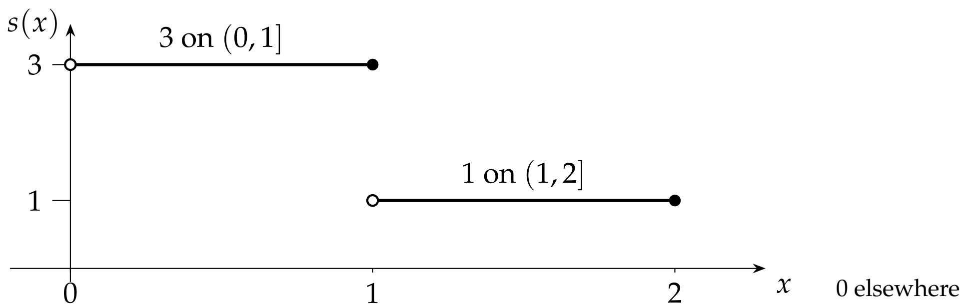
Definition (Simple function). Fix a measurable space $(S, \mathcal{B})$. A function $s : S \to \mathbb{R}$ is a *simple function* (or *step function*) if it takes only finitely many values. Equivalently, there exist pairwise disjoint measurable sets $A_1, \ldots, A_n \in \mathcal{B}$ and scalars $c_1, \ldots, c_n \in \mathbb{R}$ such that

$$s(x) = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}(x) \qquad \text{for all } x \in S.$$

We call s *nonnegative* if $s(x) \geq 0$ for all x; in that case we can choose the representation with all $c_i \geq 0$.

Intuition. A simple function is *piecewise constant* on a measurable partition $\{A_1, \ldots, A_n\}$ of S.

Example. On $S = \mathbb{R}$, the map $s(x) = 3\,\mathbf{1}_{(0,1]}(x) + 1\,\mathbf{1}_{(1,2]}(x)$ is simple and nonnegative.



Remark (What we use it for). Simple nonnegative functions are the building blocks of the

Lebesgue integral: for $s = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ with $c_i \geq 0$,

$$\int_S s \, d\mu \; = \; \sum_{i=1}^{n} c_i \, \mu(A_i).$$

General nonnegative measurable functions are defined/integrated by approximating them from below with simple ones.

Lebesgue integral

Let $(S, \mathcal{B}, \mu)$ be a measure space and let $A \in \mathcal{B}$ be a *measurable set*. Define the indicator $I_A : S \to \mathbb{R}$ by $I_A(x) = 1$ if $x \in A$ and 0 otherwise which is measurable.

Definition. (Integral of an indicator) The Lebesgue integral of I_A with respect to μ is

$$\int_S I_A \, d\mu \; := \; \mu(A).$$

Interpretation.

- Counting measure: $\int I_A \, d\# = \#(A)$ (número de elementos de A).
- Lebesgue measure: $\int I_A \, d\lambda = \lambda(A)$ (longitud/área/volumen de A).
- Probability: $\int_S I_A \, d\mathbb{P} = \mathbb{P}(A)$, i.e., the probability of A under $\mathbb{P}$, where $\mathcal{B}$ determines which sets (events) are measurable.

From indicator functions, we can naturally extend to simple functions. If $s = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ with pairwise disjoint $A_i \in \mathcal{B}$ and $c_i \geq 0$, define

$$\int_S s \, d\mu \; := \; \sum_{i=1}^{n} c_i \, \mu(A_i).$$

A *simple function* is a finite–valued measurable function. Equivalently,[3] it can be written as a finite linear combination of indicators:

$$s(x) = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}(x), \qquad A_i \in \mathcal{B}.$$

Such s is measurable because sums and scalar multiples of measurable maps are measurable (addition and scaling are continuous; compose with $(f, g) \mapsto f + g$).

We first restrict attention to the set of *nonnegative* simple functions,

$$S^+ := \{ \, s : S \to \mathbb{R} \mid s \text{ simple and } s \geq 0 \, \}.$$

[3]If s takes finitely many values $\{c_1, \ldots, c_n\}$, set $A_i := s^{-1}(\{c_i\})$; then $A_i \in \mathcal{B}$, are pairwise disjoint, and $s = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$. Conversely, any finite sum $\sum_i c_i \mathbf{1}_{A_i}$ with $A_i \in \mathcal{B}$ is measurable and takes only values in $\{c_1, \ldots, c_n\}$.

This avoids undefined expressions like $\infty - \infty$ and matches the way general nonnegative functions will be built as limits from below. Moreover, if $s \in S^+$ admits a decomposition with pairwise disjoint A_i, then necessarily $c_i \geq 0$ (because $s(x) = c_i$ on A_i).

Definition (Integral of a nonnegative simple function)**.** If $s = \sum_{i=1}^{n} c_i \, \mathbf{1}_{A_i}$ with $A_i \in \mathcal{B}$ pairwise disjoint and $c_i \geq 0$, define

$$\int_S s \, d\mu := \sum_{i=1}^{n} c_i \, \mu(A_i) \in [0, \infty) \cup \{\infty\}.$$

This value is well defined (independent of the particular representation): if $s = \sum_i c_i \mathbf{1}_{A_i} = \sum_j d_j \mathbf{1}_{B_j}$, refine to the disjoint partition $\{A_i \cap B_j\}_{i,j}$ and both sums coincide.

Proposition (Basic properties on S^+)**.** *Let $s, t \in S^+$ (nonnegative simple functions) and $a, b \geq 0$. Then:*

1. ***Nonnegativity and nullity:*** *$\int s \, d\mu \geq 0$, and $\int s \, d\mu = 0$ iff $s = 0$ a.e.*
2. ***Homogeneity:*** *$\displaystyle\int (as) \, d\mu = a \int s \, d\mu$.*
3. ***Additivity:*** *$\displaystyle\int (s+t) \, d\mu = \int s \, d\mu + \int t \, d\mu$.*
4. ***Monotonicity:*** *If $s \leq t$ a.e., then $\displaystyle\int s \, d\mu \leq \int t \, d\mu$.*
5. ***Restriction to a set:*** *For $E \in \mathcal{B}$,*

$$\int_E s \, d\mu \;:=\; \int_S s \, \mathbf{1}_E \, d\mu \;=\; \sum_{i=1}^{n} c_i \, \mu(A_i \cap E).$$

6. ***Well-definedness (independence of representation):*** *If $s = \sum_i c_i \mathbf{1}_{A_i} = \sum_j d_j \mathbf{1}_{B_j}$, then both formulas give the same value.*

Proof sketch (a bit beyond scope). Write $s = \sum_i c_i \mathbf{1}_{A_i}$ with A_i disjoint and $c_i \geq 0$.

1. Since $\mu(A_i) \geq 0$, the sum is ≥ 0; if it equals 0, then $\mu(A_i) = 0$ whenever $c_i > 0$, hence $s = 0$ a.e.
2. $\int (as) = \int \sum_i (ac_i) \mathbf{1}_{A_i} = \sum_i (ac_i) \mu(A_i) = a \sum_i c_i \mu(A_i)$.
3. Refine to a disjoint partition $\{A_i \cap B_j\}_{i,j}$ for representations of s and t; use finite additivity (a consequence of countable additivity of μ).
4. $t - s \geq 0$ implies $\int (t - s) \geq 0$, hence $\int s \leq \int t$.
5. Replace A_i by $A_i \cap E$ in the definition.
6. Use the common refinement $\{A_i \cap B_j\}_{i,j}$; both sums reduce to $\sum_{i,j}$(value on $A_i \cap B_j$) $\mu(A_i \cap B_j)$.

□

Lebesgue integral (beyond simple functions)

So far we can integrate *nonnegative simple functions*. For a *general* nonnegative measurable function $f : S \to [0, \infty]$, we "integrate from below": we build simple functions that sit under f and climb up to it.

"From below" (what it means). We construct a sequence $(s_n)_{n \geq 1}$ of nonnegative simple functions such that

$$0 \leq s_1 \leq s_2 \leq \cdots \leq f \quad \text{and} \quad s_n(x) \uparrow f(x) \text{ for each } x \in S.$$

Think of s_n as a staircase with finer and finer steps that never overshoots f.

A concrete construction you can always use. For each $n \in \mathbb{N}$, set

$$f^{(n)}(x) := \min\{f(x), n\}, \qquad s_n(x) := 2^{-n} \lfloor 2^n f^{(n)}(x) \rfloor.$$

Then each s_n is simple (it only takes values in $\{0, 1/2^n, \ldots, n\}$), $0 \leq s_n \leq f$, and $s_n(x) \uparrow f(x)$ pointwise. This is exactly "approximate f from below by simple functions."

Definition (Integral of a nonnegative measurable function)**.** Let S^+ be the set of nonnegative simple functions on S. For a measurable $f : S \to [0, \infty]$,

$$\int_S f \, d\mu \ := \ \sup \left\{ \int_S h \, d\mu \ : \ h \in S^+, \, 0 \leq h \leq f \right\}.$$

Reading it: take all simple functions that fit *under* f, integrate each, and keep the largest value (the supremum).

Remark (Integrability (as on the slide))**.** We say that f is μ-integrable if $\int_S f \, d\mu < \infty$.

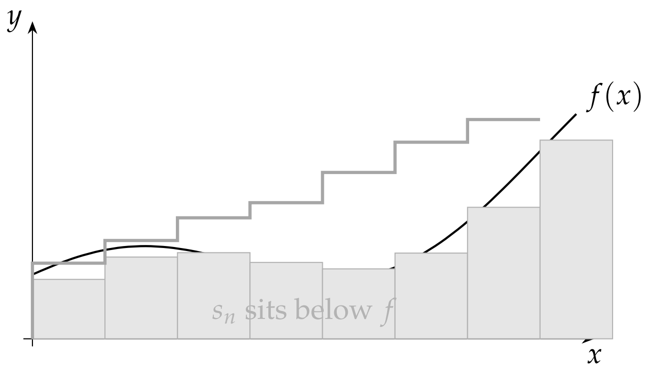


Figure 1: Approximating f from below by simple "staircases" s_n. The integral of f is the supremum of the integrals of all such staircases.

Lebesgue integral for general (signed) functions

Assume a measure space $(S, \mathcal{B}, \mu)$. Define the space of integrable (absolutely integrable) functions

$$\mathcal{L}(\mu) \ := \ \{f \text{ measurable} : \int_S |f| \, d\mu < \infty\}.$$

Positive/negative parts. For a measurable $f : S \to \mathbb{R}$ set

$$f^+(x) := \max\{f(x), 0\}, \qquad f^-(x) := \max\{-f(x), 0\}.$$

Then $f^+, f^- \geq 0$ are measurable and

$$f = f^+ - f^-, \qquad |f| = f^+ + f^-, \qquad f^+ f^- = 0 \text{ a.e.}$$

(*Reason:* $t \mapsto \max\{t, 0\}$ and $t \mapsto \max\{-t, 0\}$ are continuous, hence preserve measurability; the identities are pointwise algebra.)

Definition (Integral of $f \in \mathcal{L}(\mu)$)**.** If $f \in \mathcal{L}(\mu)$, define

$$\int_S f \, d\mu \ := \ \int_S f^+ \, d\mu \ - \ \int_S f^- \, d\mu.$$

Example. If $A, B \in \mathcal{B}$ are disjoint and $f = 2\,\mathbf{1}_A - 3\,\mathbf{1}_B$, then

$$f^+ = 2\mathbf{1}_A, \quad f^- = 3\mathbf{1}_B, \qquad \int_S f \, d\mu = 2\,\mu(A) - 3\,\mu(B).$$

Well-definedness: since $|f| = f^+ + f^-$,

$$\int_S f^+ \, d\mu \leq \int_S |f| \, d\mu < \infty, \qquad \int_S f^- \, d\mu \leq \int_S |f| \, d\mu < \infty,$$

so no $\infty - \infty$ ambiguity arises.[4]

[4]Here it might be helpful to remember the triangle inequality in L^1: for $f \in L^1(\mu)$,

$$\left| \int f \, d\mu \right| \leq \int |f| \, d\mu.$$

Proof. Write $f = f^+ - f^-$ with $f^+ = \max\{f, 0\}$ and $f^- = \max\{-f, 0\}$. Then $f^\pm \geq 0$, $|f| = f^+ + f^-$, and (since $f \in L^1$) both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite. Hence

$$\left| \int f \, d\mu \right| = \left| \int f^+ \, d\mu - \int f^- \, d\mu \right| \leq \int f^+ \, d\mu + \int f^- \, d\mu = \int |f| \, d\mu.$$

Equality μ-almost everywhere

Definition. Fix two measurable functions $f, g : S \to \mathbb{R}_+$. We say

$$f = g \ \mu\text{-a.e.} \quad \Longleftrightarrow \quad \mu(\{x \in S : f(x) \neq g(x)\}) = 0.$$

Thus: they may differ only on a μ-null set. The phrase "almost" is always with respect to the underlying measure μ. In L^p spaces we identify functions that are equal μ-a.e.

Example (Graphical intuition)**.** Let $S = [0, 5]$ with Lebesgue measure. Take any measurable curve f; define $g(x) = f(x)$ for all $x \neq x_0$ and set $g(x_0) = f(x_0) + 1$. Since $\mu(\{x_0\}) = 0$, we have $f = g$ μ-a.e.

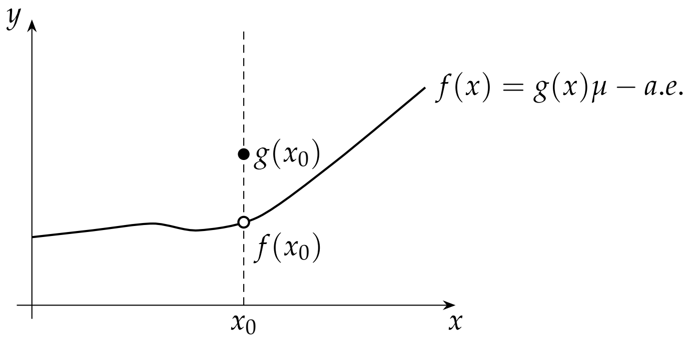


Figure 2: Equal almost everywhere: g equals f except at a single point x_0 (a μ-null set).

Properties of the Lebesgue integral. Let $f, g \geq 0$ be measurable (more generally, $f, g \in L^1(\mu)$).

1. If $f = g$ μ-a.e., then $\int_S f \, d\mu = \int_S g \, d\mu$.
   Proof: *out of scope.* ***Intuition:*** *If two functions differ only on a set of measure zero, that set has no mass, so their integrals coincide.*

2. If $f \leq g$ μ-a.e., then $\int_S f \, d\mu \leq \int_S g \, d\mu$.
   Proof: *out of scope.* ***Intuition:*** *The integral is a measure–weighted sum. If $f \leq g$ almost everywhere, then pointwise f contributes no more than g except on a null set (which contributes nothing), so the total cannot exceed $\int_S g \, d\mu$.*

3. For $f \geq 0$,
   $$f = 0 \ \mu\text{-a.e.} \quad \Longleftrightarrow \quad \int_S f \, d\mu = 0.$$
   Proof: *out of scope.* ***Intuition:*** *A nonnegative function has nonnegative "area." If $\int_S f \, d\mu = 0$, there cannot be any region of positive measure where f stays above some $\varepsilon > 0$; otherwise the area would be at least $\varepsilon \, \mu(\text{that region}) > 0$. Hence $f = 0$ except on a μ-null set. Conversely, if $f = 0$ a.e., its integral is clearly 0.*

Convergence theorems (when can we swap limit and integral?)

We collect the three results used throughout the course to pass limits through the Lebesgue integral.

Theorem 5 (Monotone Convergence Theorem (MCT) - Beppo Levi)**.** *Let $(S, \mathcal{B}, \mu)$ be a measure space. Suppose $f_n : S \to [0, \infty)$ are measurable functions with $f_1 \le f_2 \le \cdots$ μ-a.e., and let $f := \lim_{n\to\infty} f_n$ (pointwise a.e.).*

Then

$$\lim_{n\to\infty} \int_S f_n \, d\mu \;=\; \int_S f \, d\mu.$$

Proof. **Step 1 (upper bound).** Since $0 \le f_n \le f$ for each n, by monotonicity of the integral, $\int f_n \, d\mu \le \int f \, d\mu$. Hence

$$\limsup_{n\to\infty} \int f_n \, d\mu \;\le\; \int f \, d\mu. \tag{$*$}$$

Step 2 (lower bound via simple under-approximations). Pick any nonnegative simple function $h \le f$. Write it as a finite staircase $h = \sum_{k=1}^{m} c_k \mathbf{1}_{A_k}$ with $c_k \ge 0$ and disjoint A_k.

For each "level" c_k, look at the part where f_n already reaches that level:

$$B_{k,n} \;:=\; A_k \cap \{f_n \ge c_k\} \quad \text{(the portion of } A_k \text{ where } f_n \text{ has caught up with } h\text{)}.$$

Because $f_n \uparrow f$ and $h \le f$, these sets expand: $B_{k,n} \uparrow A_k$. Therefore their measures grow: $\mu(B_{k,n}) \uparrow \mu(A_k)$ (continuity from below of measures). On $B_{k,n}$ we have $f_n \ge c_k$, so by monotonicity of the integral

$$\int_S f_n \, d\mu \;\ge\; \sum_{k=1}^{m} \int_{B_{k,n}} c_k \, d\mu \;=\; \sum_{k=1}^{m} c_k \, \mu(B_{k,n}).$$

Letting $n \to \infty$ and using that the sum is finite,

$$\liminf_{n\to\infty} \int_S f_n \, d\mu \;\ge\; \sum_{k=1}^{m} c_k \, \mu(A_k) \;=\; \int_S h \, d\mu.$$

Since this holds for *every* simple $h \le f$, taking the supremum over such h yields

$$\int_S f \, d\mu \;\le\; \liminf_{n\to\infty} \int_S f_n \, d\mu. \tag{$**$}$$

Remark (liminf/limsup intuition)**.** For a sequence (a_n), look at each *tail* $\{a_k : k \ge n\}$ and take its infimum:

$$\ell_n := \inf_{k \ge n} a_k \quad (\ell_n \text{ is increasing in } n).$$

Then the liminf is the supremum of these tail infima (the "best eventual lower bound"):

$$\liminf_{n\to\infty} a_n \;=\; \lim_{n\to\infty} \ell_n \;=\; \sup_n \inf_{k\geq n} a_k.$$

Dually, with $u_n := \sup_{k\geq n} a_k$ (decreasing),

$$\limsup_{n\to\infty} a_n \;=\; \lim_{n\to\infty} u_n \;=\; \inf_n \sup_{k\geq n} a_k.$$

Step 3 (combine). From $(*)$ and $(**)$,

$$\int f\,d\mu \;\leq\; \liminf_n \int f_n\,d\mu \;\leq\; \limsup_n \int f_n\,d\mu \;\leq\; \int f\,d\mu,$$

so all three quantities are equal and the limit exists with value $\int f\,d\mu$. □

Read it. If f_n increases pointwise to f and all are nonnegative, we may *swap limit and integral*. The proof mirrors the "integrate from below" idea: any simple $h \leq f$ eventually gets captured under the f_n's in measure, forcing the integrals of f_n up to $\int f$.[5]

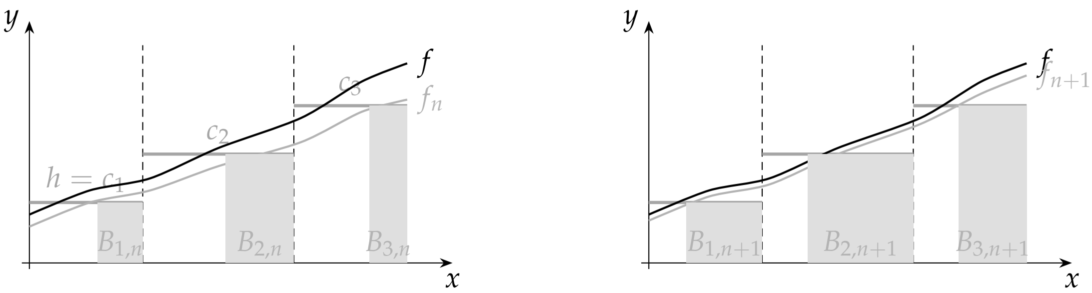


Figure 3: Step 2 (visual): for each step of h, the region where f_n already exceeds that level grows, forcing $\int f_n$ to eventually exceed $\int h$. Since this holds for any $h \leq f$, taking the supremum yields $\int f \leq \liminf_n \int f_n$.

Proof sketch (intuition). Approximate f by simple functions from below and note that, for nonnegative, increasing f_n, the integrals of these approximations also increase to $\int f\,d\mu$. The key is that $\int(\cdot)$ is continuous along monotone increases of nonnegative functions (no cancellations from negative parts).

[5] **Intuition for future me:** If you keep adding area from below without ever overshooting, the accumulated area increases and eventually equals the total area. That's the MCT: for a sequence of nonnegative functions with $f_n \uparrow f$ pointwise, the integrals also increase and reach $\int f\,d\mu$. Picture f as a mountain and each f_n as a staircase filling it from below. Each step adds nonnegative volume on top of f_{n-1}, so the integral cannot go down; and since the staircase never exceeds f, it cannot surpass $\int f\,d\mu$. As n refines the staircase, you "touch" the whole mountain: $\int f_n\,d\mu \uparrow \int f\,d\mu$.

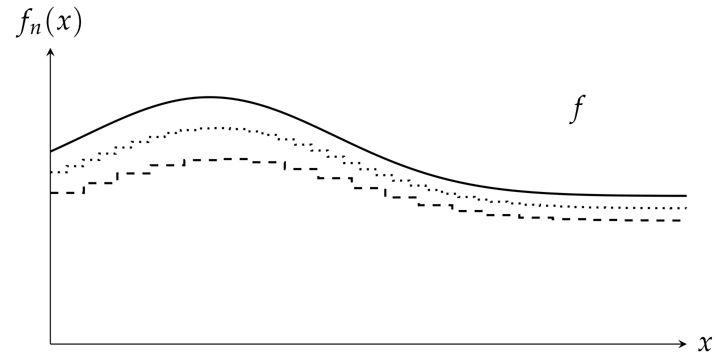


Figure 4: Monotone nonnegative approximations $f_n \uparrow f$: area increases to $\int f\, d\mu$.

Week 1 – Discussion

Problem 1. For any three events, A, B,and C, defined on a sample space S:

a. **Commutativity.** $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

Proof (sketch). For any x,

$$x \in A \cup B \iff (x \in A \text{ or } x \in B) \iff x \in B \cup A.$$

Similarly, $x \in A \cap B \iff (x \in A \text{ and } x \in B) \iff x \in B \cap A$. □

b. **Associativity.** $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$.

Proof (sketch). For any x,

$$x \in A \cup (B \cup C) \iff (x \in A) \text{ or } (x \in B) \text{ or } (x \in C) \iff x \in (A \cup B) \cup C.$$

The intersection case is identical with "or" replaced by "and." □

c. **Distributive laws.** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Proof (sketch). For any x,

$$\begin{aligned} x \in A \cap (B \cup C) &\iff (x \in A) \text{ and } \big((x \in B) \text{ or } (x \in C)\big) \\ &\iff \big((x \in A \cap B) \text{ or } (x \in A \cap C)\big) \iff x \in (A \cap B) \cup (A \cap C). \end{aligned}$$

For the second identity,

$$\begin{aligned} x \in A \cup (B \cap C) &\iff (x \in A) \text{ or } \big((x \in B) \text{ and } (x \in C)\big) \\ &\iff \big((x \in A \text{ or } x \in B) \text{ and } (x \in A \text{ or } x \in C)\big) \\ &\iff x \in (A \cup B) \cap (A \cup C). \end{aligned}$$

□

d. **De Morgan's laws.** $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.

Proof (sketch). For any x,

$$x \in (A \cup B)^c \iff \neg(x \in A \cup B) \iff (\neg x \in A) \text{ and } (\neg x \in B) \iff x \in A^c \cap B^c.$$

Likewise,

$$x \in (A \cap B)^c \iff \neg(x \in A \cap B) \iff (\neg x \in A) \text{ or } (\neg x \in B) \iff x \in A^c \cup B^c.$$

□

Problem 2. Verify the following identities:

a. $A \setminus B = A \setminus (A \cap B) = A \cap B^c$.

Proof (sketch). By definition, $A \setminus B = A \cap B^c$. Also,

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

Hence all three sets coincide. □

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

Proof (sketch). If $x \in B$, then either $x \in A$ or $x \in A^c$. Thus $x \in (B \cap A) \cup (B \cap A^c)$, so $B \subseteq (B \cap A) \cup (B \cap A^c)$. Conversely, every element of $(B \cap A)$ or $(B \cap A^c)$ lies in B, so $(B \cap A) \cup (B \cap A^c) \subseteq B$. Therefore equality holds. □

c. $B \setminus A = B \cap A^c$.

Proof (sketch). This is the definition of set difference: $B \setminus A := \{x : x \in B \text{ and } x \notin A\} = B \cap A^c$. □

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

Proof (sketch). Using distributivity and complements,

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

Equivalently, by (b), $B = (B \cap A) \cup (B \cap A^c)$, hence $A \cup B = A \cup \big((B \cap A) \cup (B \cap A^c)\big) = (A \cup (B \cap A)) \cup (B \cap A^c) = A \cup (B \cap A^c)$ since $B \cap A \subseteq A$. □

Problem 3. Provide an example of two σ-algebras such that their union is not a σ-algebra.
Solution. Let $S = \{1, 2, 3\}$ and define

$$\mathcal{A} = \{\varnothing, \{1\}, \{2, 3\}, S\}, \qquad \mathcal{C} = \{\varnothing, \{2\}, \{1, 3\}, S\}.$$

Each of $\mathcal{A}$ and $\mathcal{C}$ is a σ-algebra on S (they contain $\varnothing$, are closed under complements in S, and—being finite—are closed under countable unions).

Consider their union:

$$\mathcal{A} \cup \mathcal{C} = \{\varnothing, \{1\}, \{2\}, \{1,3\}, \{2,3\}, S\}.$$

This family is *not* a σ-algebra because it is not even closed under finite unions:

$$\{1\} \in \mathcal{A} \cup \mathcal{C}, \quad \{2\} \in \mathcal{A} \cup \mathcal{C}, \quad \text{but} \quad \{1\} \cup \{2\} = \{1,2\} \notin \mathcal{A} \cup \mathcal{C}.$$

Therefore $\mathcal{A} \cup \mathcal{C}$ fails to be a σ-algebra. □

Problem 4.) Prove that if $\mathcal{B}$ is a σ-algebra on S and $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcap\limits_{n=1}^{\infty} A_n \in \mathcal{B}$.

Proof. Because $\mathcal{B}$ is closed under complements, for each $n \in \mathbb{N}$ we have $C_n := A_n^c \in \mathcal{B}$. Since $\mathcal{B}$ is closed under countable unions,

$$U := \bigcup_{n=1}^{\infty} C_n \in \mathcal{B}.$$

Again using closure under complements and De Morgan's law,

$$U^c = \left(\bigcup_{n=1}^{\infty} C_n\right)^c = \bigcap_{n=1}^{\infty} C_n^c = \bigcap_{n=1}^{\infty} A_n \in \mathcal{B}.$$

Hence $\bigcap_{n=1}^{\infty} A_n \in \mathcal{B}$, as claimed. □

Week 2 — Class 3

Theorem 6 (Faotu's Lemma)**.** *Let $(S, \mathcal{B}, \mu)$ be a measure space. Suppose $f_n : S \to [0, \infty) \cup \{\infty\}$ are measurable functions $\forall n \in \mathbb{N}$.*

Then

$$\int_S \liminf_{n\to\infty} f_n \, d\mu \;\leq\; \liminf_{n\to\infty} \int_S f_n \, d\mu.$$

Let's build intuition. When dealing with nonnegative functions (which can be "ugly" or even infinite on some parts of the domain), the integral is *lower semicontinuous* with respect to pointwise limits. In other words, if we look at the *eventual floor* of the sequence at each point (the pointwise lim inf) and integrate it, the result will never exceed the best possible lower limit of the integrals of the original sequence. This is the essence of Fatou's Lemma.

Why it is useful and what it does *not* require.

- It does **not require convergence** of f_n to a function f; nonnegativity is enough.
- It does **not require domination** (that assumption appears in the Dominated Convergence Theorem, which is stronger but demands an integrable bound).
- It provides a robust **lower bound** when passing to limits: very useful when one can only control "tails" or "eventual minima."

Remark. By contrast, the "reverse Fatou" (with lim sup) **does** require extra conditions (e.g., domination) in order for the inequality to hold in the opposite direction.

Proposition (How MCT yields Fatou's Lemma (mechanism))**.** *For nonnegative measurable (f_n), define $g_n(x) := \inf_{k\geq n} f_k(x)$. Then $g_n \uparrow \liminf_n f_n$ and, for each n, $g_n \leq f_k$ for all $k \geq n$. By MCT,*

$$\int_S \liminf_{n} f_n \, d\mu \;=\; \lim_{n\to\infty} \int_S g_n \, d\mu \;\leq\; \liminf_{n\to\infty} \int_S f_n \, d\mu,$$

which is *Fatou's Lemma.*

Intuition. Replace the sequence by its "eventual floor" g_n—now monotone. Integrate first along this monotone path (by MCT), then compare to the original integrals using $g_n \leq f_k$ for large k.

Theorem 7 (Reverse Fatou under domination)**.** *Let (f_n) be measurable and suppose there exists $g \in L^1(\mu)$ with $f_n \leq g$ a.e. for all n. Then*

$$\limsup_{n\to\infty} \int_S f_n \, d\mu \;\leq\; \int_S \limsup_{n\to\infty} f_n \, d\mu.$$

More generally, for signed f_n, if $f_n^- \leq h \in L^1$ uniformly (uniformly integrable negative parts), the same inequality holds after splitting into positive/negative parts.

Intuition. The domination $f_n \le g$ prevents mass from "escaping upward" on small sets. Apply Fatou to $g - f_n \ge 0$:

$$\int \liminf (g - f_n) \le \liminf \int (g - f_n) \quad \Rightarrow \quad \int g - \int \limsup f_n \le \int g - \limsup \int f_n,$$

and rearrange.

Theorem 8 (Dominated Convergence Theorem (DCT))**.** *Let $(S, \mathcal{B}, \mu)$ be a measure space. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions $f_n : S \to \mathbb{R}$ and let $f : S \to \mathbb{R}$ be measurable such that*

$$f_n(x) \to f(x) \quad \text{for } \mu\text{-almost every } x \in S.$$

Assume there exists a dominating function $g : S \to [0, \infty)$ with $g \in L^1(\mu)$ such that

$$|f_n(x)| \le g(x) \quad \text{for all } n \in \mathbb{N} \text{ and } \mu\text{-almost every } x \in S.$$

Then:

1. *$f_n \in L^1(\mu)$ for every n, and $f \in L^1(\mu)$;*
2. $\lim_{n\to\infty} \int_S f_n \, d\mu \;=\; \int_S f \, d\mu.$

   Let's build intuition. If you have a sequence of measurable functions f_n that converge pointwise to f, and all of them are uniformly bounded in magnitude by some integrable "guardian" function $g \in L^1$, then you can safely *swap limit and integral*:

$$\lim_{n\to\infty} \int f_n \, d\mu \;=\; \int f \, d\mu.$$

   The role of g is to prevent the functions from "exploding" in sets of small measure, ensuring that no mass is lost or gained when passing the limit inside the integral.

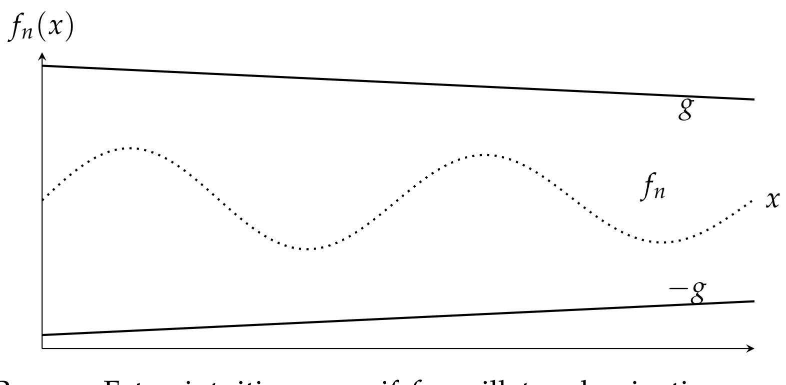


Figure 5: DCT/Reverse-Fatou intuition: even if f_n oscillates, domination prevents "mass leakage."

Remark. How does this differ from the Monotone Convergence Theorem (MCT)?

MCT applies only to *monotone increasing* nonnegative sequences $f_n \uparrow f$, and in that case no dominating function is needed: the monotonicity alone guarantees safety in swapping limit and integral. The Dominated Convergence Theorem (DCT) is strictly more general: it drops the monotonicity requirement but demands the existence of an integrable bound g.

Remark (MCT vs. Fatou vs. DCT—when to use which?)**.** Compact comparison:

- **MCT** (Beppo Levi): Nonnegative and *monotone increasing*. Then $\int \lim = \lim \int$.
- **Fatou:** Nonnegative, no convergence needed. Gives a *lower bound:* $\int \liminf \leq \liminf \int$.
- **Reverse Fatou (dominated):** If $f_n \leq g \in L^1$, then $\limsup \int \leq \int \limsup$.
- **DCT:** Pointwise a.e. convergence *and* $|f_n| \leq g \in L^1$. Then full swap: $\lim \int f_n = \int f$.

Remark (Quick checklist for swapping limit and integral)**.** When you want $\lim \int f_n = \int \lim f_n$, check:

1. *Is it monotone and nonnegative?* $\Rightarrow$ MCT applies.
2. *Is there an L^1 dominator g for $|f_n|$?* $\Rightarrow$ DCT applies.
3. *None of the above?* Use Fatou to get a one-sided inequality (often enough for bounds).

Probability Theory

Definition (Probability function / measure)**.** Given a sample space S and an associated σ-algebra $\mathcal{B}$, a *probability function* is a map $\mathbb{P} : \mathcal{B} \to [0, 1]$ that satisfies:

1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$;
2. $\mathbb{P}(S) = 1$;
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

These are the **Axioms of Probability** (Kolmogorov).[6]

Theorem 9 (Basic consequences of the axioms)**.** *If $\mathbb{P}$ is a probability function and $A \in \mathcal{B}$, then:*

(a) $\mathbb{P}(\varnothing) = 0$*;*

(b) $\mathbb{P}(A) \leq 1$ *(and by (i) also* $\mathbb{P}(A) \geq 0$*);*

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

Proof (using only the axioms). One by one:

(a) Since $S = \varnothing \cup S$ and the pieces are disjoint, additivity gives $1 = \mathbb{P}(S) = \mathbb{P}(\varnothing) + \mathbb{P}(S)$, hence $\mathbb{P}(\varnothing) = 0$.

(b) Monotonicity follows from additivity: if $A \subseteq B$, then $B = A \cup (B \setminus A)$, so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$. Since $A \subseteq S$, we get $\mathbb{P}(A) \leq \mathbb{P}(S) = 1$.

(c) $S = A \cup A^c$ with disjoint parts; thus $1 = \mathbb{P}(S) = \mathbb{P}(A) + \mathbb{P}(A^c)$, i.e., $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

□

Theorem 10 (Two-set identities and monotonicity)**.** *If $\mathbb{P}$ is a probability function and $A, B \in \mathcal{B}$, then*

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

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

(c) If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$*.*

Short proof from the axioms. One by one:

(a) Partition B as a disjoint union: $B = (B \cap A) \,\dot{\cup}\, (B \cap A^c)$. By additivity, $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$, hence the identity.

[6] *Intuition for future me.* This is just a normalized measure: (i) forbids negative mass, (ii) fixes total mass to 1, (iii) guarantees additivity over countable disjoint unions.

(b) Decompose $A \cup B$ into three disjoint pieces: $(A \setminus B)$, $(B \setminus A)$, and $(A \cap B)$. Then $\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$. Also, $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$ and $\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$. Combine and rearrange to obtain the formula.

(c) If $A \subseteq B$, then $B = A \,\dot\cup\, (B \setminus A)$, so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.

□

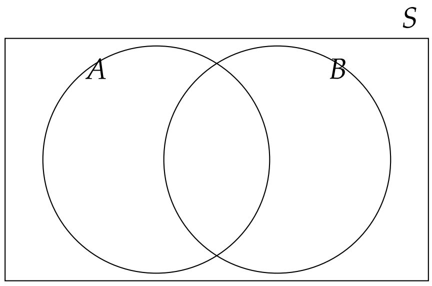


Figure 6: Identities in Thm. 10 follow by partitioning into disjoint pieces and using additivity.

Theorem 11 (Partition identity and union bound). *If $\mathbb{P}$ is a probability function, then:*

(a) For any partition $(C_i)_{i\geq 1}$ of S,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i).$$

(b) For any sets $A_1, A_2, \ldots$,

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

Proof. Using only the axioms:

(a) Because (C_i) is a partition, the sets $(A \cap C_i)$ are pairwise disjoint and $A = \bigcup_{i=1}^{\infty}(A \cap C_i)$. Countable additivity gives $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap C_i)$.

(b) The issue is that the A_i need not be disjoint. "Disjointify" them by

$$A_1^* = A_1, \qquad A_k^* = A_k \setminus \bigcup_{j<k} A_j \quad (k \geq 2).$$

Then (A_i^*) are pairwise disjoint and $\bigcup_i A_i = \biguplus_i A_i^*$, so $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i^*)$. Since $A_i^* \subseteq A_i$, by monotonicity $\mathbb{P}(A_i^*) \leq \mathbb{P}(A_i)$, hence the inequality.

□

Proposition (Continuity from below). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(A_n)_{n\geq 1} \subset \mathcal{F}$ be an increasing sequence, i.e., $A_1 \subseteq A_2 \subseteq \cdots$. Then*

$$\lim_{n\to\infty} P(A_n) \;=\; P\left(\bigcup_{n=1}^{\infty} A_n\right).$$

Proposition (Continuity from above). *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and let* $(B_n)_{n\geq 1} \subset \mathcal{F}$ *be a decreasing sequence, i.e.,* $B_1 \supseteq B_2 \supseteq \cdots$. *Then*

$$\lim_{n\to\infty} P(B_n) \;=\; P\left(\bigcap_{n=1}^{\infty} B_n\right).$$

Counting

Counting is about computing the *total number of ways* an outcome can occur in a finite sample space. Always check:

- **with vs. without replacement**;
- **with order vs. without order**.

Notation. $n! = n \times (n-1) \times \cdots \times 2 \times 1$, $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ (for $n \geq r$).

Selections of size r from n objects.

- *Ordered, without replacement:*

$$\frac{n!}{(n-r)!} = n\,(n-1)\cdots(n-r+1).$$

- *Ordered, with replacement:* n^r.
- *Unordered, without replacement:* $\binom{n}{r}$ (divide out the $r!$ orderings).
- *Unordered, with replacement:* $\binom{n+r-1}{r}$. Unordered with replacement is like assignin integer values to $x_1, ..., x_n$ s.t.

$$x_1 + ... + x_n = r$$

  Instead of integers, let map this into vertical lines. Complete this part and fully understand it.

Why counting matters (equally likely outcomes). If $S = \{s_1, \ldots, s_N\}$ and each outcome has probability $1/N$, then for any $A \subseteq S$,

$$\mathbb{P}(A) = \sum_{s_i \in A} \mathbb{P}(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\text{\#elements in A}}{\text{\#elements in S}}.$$

Example. Poker hands How many distinct 5-card hands can be dealt from a standard 52-card deck?

- Order does not matter (a hand is a set).
- Cards are drawn without replacement.
- Formula: $\binom{52}{5} = \frac{52!}{5!\,47!} = 2{,}598{,}960.$

Interpretation. Every possible 5-card poker hand is one of these $\approx$ 2.6 million outcomes.

Example. PIN codes How many different 4-digit PIN codes can be formed using digits 0–9?

- Order matters ($1234 \neq 4321$).
- Digits can repeat (with replacement).
- Formula: $10^4 = 10{,}000$.

Interpretation. A random guess at a PIN has probability $1/10{,}000$.

Conditional Probability

Example. Roll a fair six-sided die; $S = \{1, 2, 3, 4, 5, 6\}$. The *events* then are:

$$A = \{\text{even}\} = \{2, 4, 6\}, \qquad B = \{\text{greater than } 3\} = \{4, 5, 6\}.$$

Probabilities:

$$\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{3}{6} = \frac{1}{2}, \quad \mathbb{P}(A \cap B) = \mathbb{P}(\{4, 6\}) = \frac{2}{6} = \frac{1}{3}.$$

$$\text{Conditional:} \quad \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Knowing B occurred restricts the sample space to $\{4, 5, 6\}$, where two of three outcomes are even, i.e. are in A.

Bayes' Rule

Using the definition

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (\mathbb{P}(B) > 0),$$

we obtain the *product rule*

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

Similarly,

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

Equating both expressions yields Bayes' rule:

$$\boxed{\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\, \frac{\mathbb{P}(A)}{\mathbb{P}(B)}} \qquad (\mathbb{P}(B) > 0).$$

Bayes' Rule (partition form)

Theorem 12 (Bayes' Rule). *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any event. Then, for each $i = 1, 2, \ldots$,*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\, \mathbb{P}(A_j)}.$$

Appears often in economics: Monty Hall; Bayesian updating in micro; Bayesian econometrics.

Bayes' Rule: Monty Hall

Setup: Three doors; one has a prize. You pick door A. Monty opens one of the other two doors (call it C), showing it is empty. You may switch to the remaining closed door B.

Goal. Compute $\mathbb{P}(A \text{ has prize} \mid C \text{ open})$.

Unconditional probabilities.

$$\mathbb{P}(A \text{ has prize}) = \mathbb{P}(B \text{ has prize}) = \mathbb{P}(C \text{ has prize}) = \tfrac{1}{3}.$$

Key conditionals.

$$\mathbb{P}(C \text{ open} \mid A \text{ has prize}) = \tfrac{1}{2}, \quad \mathbb{P}(C \text{ open} \mid B \text{ has prize}) = 1, \quad \mathbb{P}(C \text{ open} \mid C \text{ has prize}) = 0.$$

Bayes.

$$\mathbb{P}(A \text{ has prize} \mid C \text{ open}) = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 0\cdot\frac{1}{3}} = \frac{1}{3},$$

$$\mathbb{P}(B \text{ has prize} \mid C \text{ open}) = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 0\cdot\frac{1}{3}} = \frac{2}{3}.$$

Conclusion. Switch.

Law of Total Probability

Theorem 13 (Law of Total Probability). *If $\{B_1, B_2, \ldots\}$ is a partition of S and $\mathbb{P}(B_i) > 0$ for all i, then for any event A,*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i).$$

Proof. Since (B_i) is a partition, the sets $(A \cap B_i)$ are pairwise disjoint and $A = \bigcup_{i=1}^{\infty}(A \cap B_i)$. By countable additivity,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i),$$

where the last equality uses $\mathbb{P}(A \mid B_i) = \mathbb{P}(A \cap B_i)/\mathbb{P}(B_i)$. □

Independence

Sometimes an event A may not be affected by event B, i.e. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. By the definition of conditional probability, this implies

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

Definition (Independence)**.** Two events A and B are *statistically independent* if

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

Equivalences for Independence

Proposition (Equivalent characterizations of independence)**.** *Let* $A, B \in \mathcal{F}$ *be events with* $\mathbb{P}(A), \mathbb{P}(B) > 0$*. The following are equivalent:*

1. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ *(definition).*
2. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.
3. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

Remark. If $\mathbb{P}(B) = 0$, the conditional probability $\mathbb{P}(A \mid B)$ is undefined, so items 2–3 do not apply. The definition in item 1 remains valid.

Remark (Disjointness vs. independence)**.** If A and B are disjoint with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then they are *not* independent because $\mathbb{P}(A \cap B) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$.

Example. Independence in a Deck of Cards

Setup. A standard deck has 52 cards, 4 suits (spades, hearts, diamonds, clubs), each with 13 cards.

Events.

$$A = \{\text{"the card is an ace } (\mathbf{1})\text{"}\}, \qquad B = \{\text{"the card is a spade } \spadesuit\text{"}\}.$$

Computations.

$$\mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}, \qquad \mathbb{P}(B) = \frac{13}{52} = \frac{1}{4}, \qquad \mathbb{P}(A \cap B) = \mathbb{P}(\text{"ace of spades"}) = \frac{1}{52}.$$

Hence

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13} = \mathbb{P}(A),$$

so A and B are independent because $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Interpretation. Knowing that the card is a spade reduces the sample space from 52 to 13 equally likely outcomes; exactly one of those is an ace, so the chance remains $1/13$. Learning B provides no information about A, which is precisely independence.

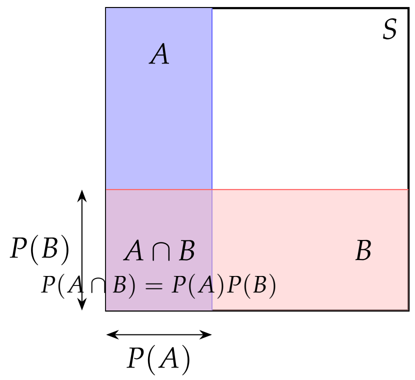


Figure 7: Independence: the intersection area factors as $P(A)P(B)$.

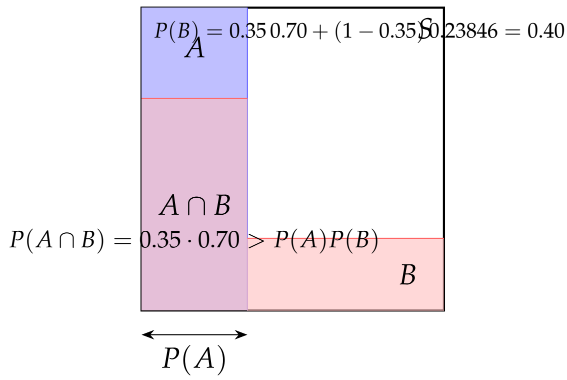


Figure 8: Dependence: B is concentrated inside A, so $P(A \cap B) > P(A)P(B)$.

Independence: complements and collections

Theorem 14 (Closure under complements). *If A and B are independent events, then the following pairs are also independent:*

1. *A and B^c,*
2. *A^c and B,*
3. *A^c and B^c.*

Proof of (a). By additivity, $P(A \cap B^c) = P(A) - P(A \cap B)$. Independence of A and B gives $P(A \cap B) = P(A)P(B)$, hence

$$P(A \cap B^c) = P(A)\big(1 - P(B)\big) = P(A)P(B^c),$$

which is the required factorization. Parts (b)–(c) are analogous. □

Reminder: collections and subcollections. A *collection (family)* of events is any set $\mathcal{A} \subseteq \mathcal{P}(\Omega)$. A *subcollection* is any subset $\mathcal{A}' \subseteq \mathcal{A}$ (i.e., a selection of some of the events in $\mathcal{A}$).

Definition (Mutual independence of a collection)**.** Events $A_1, \ldots, A_n$ are *mutually independent* if for every nonempty index set $I \subseteq \{1, \ldots, n\}$ (equivalently, for every subcollection $\{A_i : i \in I\}$)

we have

$$P\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} P(A_i).$$

It is common to check this for all I with $|I| \geq 2$ (the case $|I| = 1$ is tautological).

Remark (Pairwise vs. mutual independence)**.** Pairwise independence does *not* imply mutual independence. For instance, let $\Omega = \{1,2,3,4\}$ with the uniform probability and define

$$A = \{1,2\}, \quad B = \{1,3\}, \quad C = \{1,4\}.$$

Then $P(A) = P(B) = P(C) = \frac{1}{2}$ and each pair intersects with probability $1/4 = \frac{1}{2} \cdot \frac{1}{2}$, so pairs are independent; but

$$P(A \cap B \cap C) = P(\{1\}) = \tfrac{1}{4} \neq \tfrac{1}{8} = \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2},$$

so A, B, C are not mutually independent.

Conversely, having only $P(A \cap B \cap C) = P(A)P(B)P(C)$ is *not* sufficient for mutual independence, because pairwise factorizations may fail. For example, with $\Omega = \{1,\ldots,8\}$ uniform, take

$$A = \{1,2,3,4\}, \quad B = \{1,2,3,5\}, \quad C = \{1,5,6,7\}.$$

Then $P(A) = P(B) = P(C) = \frac{1}{2}$, $P(A \cap B \cap C) = \frac{1}{8}$ (so the triple product holds), but $P(A \cap B) = \frac{3}{8} \neq \frac{1}{4}$ and $P(A \cap C) = \frac{1}{8} \neq \frac{1}{4}$, hence not mutually independent.

Week 2 — Class 4

Random Variables

Double-Check and re-do this subsection.

Definition and basic construction

Definition (Random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A (real-valued) *random variable* is a measurable function

$$X : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B}),$$

i.e., for every Borel set $A \in \mathcal{B}$ we have $X^{-1}(A) \in \mathcal{F}$.

Remark (Why measurability?)**.** Measurability guarantees that *preimages* $X^{-1}(A)$ are events, so we can assign probabilities to statements about X. Equivalently, X lets us *push* the probability from $(\Omega, \mathcal{F}, \mathbb{P})$ to the real line. The condition of measurability requires that these preimages belong to $\mathcal{F}$ so that $\mathbb{P}$ can assign probabilities to them. **Without that condition, you could have a set of outcomes Ω to which you don't know how to assign probabilities.**

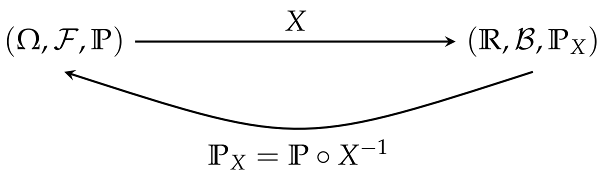
Definition (Induced (pushforward) distribution of X)**.** The *distribution* of X is the probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B})$ defined by

$$\mathbb{P}_X(A) \; := \; \mathbb{P}\big(X \in A\big) = \mathbb{P}\big(X^{-1}(A)\big), \qquad A \in \mathcal{B}.$$

Equivalently, $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$.

Proposition. *$\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ (nonnegativity, normalization $\mathbb{P}_X(\mathbb{R}) = 1$, and countable additivity).*

Remark (Sample space for X)**.** Sometimes one restricts to the range $\mathcal{X} := X(\Omega) \subseteq \mathbb{R}$ and equips it with the σ-algebra $\mathcal{B} \cap \mathcal{X}$; then $(\mathcal{X}, \mathcal{B} \cap \mathcal{X}, \mathbb{P}_X)$ is the "new" probability space on which X lives.



[6]**Intuition for future me.** An outcome $\omega \in \Omega$ is the physical result of the experiment. A random variable is a *question* about ω whose answer is a real number. Different questions (e.g., parity of a die vs. the square of the face) are different random variables on the *same* underlying outcome. Probabilities about X are computed by looking at preimages in Ω.

Finite sample space: induced law on the range

Setup. Let $\Omega = \{s_1, \dots, s_n\}$ with probability function $\mathbb{P}$ on 2^Ω, and let $X : \Omega \to \mathbb{R}$ be a random variable with (finite) range

$$\mathcal{X} := X(\Omega) = \{x_1, \dots, x_m\} \subset \mathbb{R}.$$

Each x_i is a distinct value of the random variable X. Therefore, the events

$$\{\omega \in \Omega : X(\omega) = x_i\}$$

are **pairwise disjoint** subsets of Ω.
The intuition is clear:

- A single ω cannot make X take two different values simultaneously.
- Hence, the sets defining each x_i are incompatible.
- Their union is Ω (since X always takes one of these values).

Induced probability on the range. Define the probability of each value in the range by

$$p_X(x_i) \;:=\; \mathbb{P}_X(\{x_i\}) = \mathbb{P}\big(\{\omega \in \Omega : \; X(\omega) = x_i\}\big), \qquad i = 1, \dots, m.$$

Then $p_X : \mathcal{X} \to [0, 1]$ is a probability mass function (pmf): (i) $p_X(x_i) \geq 0$ and (ii) $\sum_{i=1}^{m} p_X(x_i) = 1$. For any $A \subseteq \mathcal{X}$,

$$\mathbb{P}_X(A) = \sum_{x_i \in A} p_X(x_i).$$

Key takeaway. A random variable does not bring "new" probabilities; it *inherits* them via preimages: $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ on $\mathcal{X}$.

Example. Parity of a die Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ with $\mathbb{P}(\{\omega\}) = 1/6$ and define

$$X(\omega) = \begin{cases} 0, & \text{if } \omega \text{ is even,} \\ 1, & \text{if } \omega \text{ is odd.} \end{cases}$$

Then $\mathcal{X} = \{0, 1\}$ and

$$p_X(0) = \mathbb{P}(\{2, 4, 6\}) = \tfrac{3}{6} = \tfrac{1}{2}, \qquad p_X(1) = \mathbb{P}(\{1, 3, 5\}) = \tfrac{3}{6} = \tfrac{1}{2}.$$

Hence $X \sim \text{Bernoulli}(1/2)$ (induced by the die roll).

Remark. Different questions on the same experiment

From the same outcome ω we can define other r.v.'s, e.g.

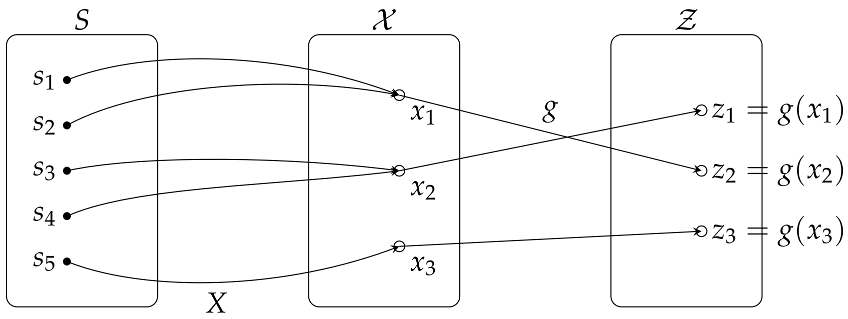
$$Y(\omega) = \mathbf{1}\{\omega \text{ is prime}\}, \qquad Z(\omega) = \omega^2.$$

For Z, the range is $\{1, 4, 9, 16, 25, 36\}$ and, by preimages,

$$\mathbb{P}(Z = 36) = \mathbb{P}(\{\omega = 6\}) = \tfrac{1}{6}, \quad \mathbb{P}(Z \in \{1,4,9\}) = \mathbb{P}(\{1,2,3\}) = \tfrac{3}{6} = \tfrac{1}{2}, \text{ etc.}$$

Even if Z is not one-to-one, probabilities are always computed via sets of the form $\{\omega : Z(\omega) \in A\}$ in Ω.

$$X : S \to \mathcal{X} \qquad g : \mathcal{X} \to \mathcal{Z}, \quad Z = g \circ X$$



Intuition: On the left sits the sample space $S = \{s_1, \ldots, s_n\}$ (the physical outcomes). On the right sits the range $\mathcal{X} = X(S) = \{x_1, \ldots, x_\ell\}$ (the numerical values). Arrows represent the function $X : S \to \mathbb{R}$: each outcome s_j is sent to exactly one value x_i. Several outcomes may land on the same x_i (a many-to-one map). The set

$$F_i := X^{-1}(\{x_i\}) = \{\, s \in S : X(s) = x_i \,\}$$

is the preimage of x_i. Intuition for $\mathcal{Z}$ TDB.

Uncountable range: pushforward via sets

When the range $\mathcal{X} = X(\Omega)$ is uncountable, we define the induced law on *Borel sets*. For any $A \in \mathcal{B} \cap \mathcal{X}$,

$$\mathbb{P}_X(A) \;=\; \mathbb{P}(X \in A) \;=\; \mathbb{P}(\{\omega \in \Omega :\; X(\omega) \in A\}) \;=\; \mathbb{P}\Big(X^{-1}(A)\Big).$$

Remark (Why sets instead of points?)**.** For continuous distributions one typically has $\mathbb{P}(X = x) = 0$ for all x, so probabilities are assigned to *sets* (intervals, unions of intervals, etc.), not to singletons.

Example (Uniform). Let $\Omega = [0,1]$ with Lebesgue measure $\mathbb{P}$ and define $X(\omega) = \omega$. Then for any interval $A = [a,b] \subset [0,1]$,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in [a,b]) = \mathbb{P}(\{\omega : \omega \in [a,b]\}) = b - a, \qquad \text{while} \quad \mathbb{P}(X = x) = 0 \ \forall x.$$

Distribution function (cdf)

Definition (Cumulative distribution function)**.** For a random variable X, the cdf is the function $F_X : \mathbb{R} \to [0,1]$ given by

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

Proposition (Basic properties)**.** *Every cdf F_X satisfies:*

- *F_X is nondecreasing;*
- *F_X is right–continuous;*
- $\lim_{x\to-\infty} F_X(x) = 0$ *and* $\lim_{x\to\infty} F_X(x) = 1$.

This properties are not only necessary but sufficient, i.e. every function $F_X : \mathbb{R} \to [0,1]$ that holds those three properties ***is also the cdf of some random variable****. If X is discrete, F_X has jumps and $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$.*

Proof. TBD. □

Example. Discrete cdf (fair die). Let X be the outcome of a fair die, $\mathbb{P}(X = k) = 1/6$ for $k = 1,\ldots,6$. Then

$$F_X(x) = \begin{cases} 0, & x < 1, \\ \frac{k}{6}, & k \le x < k+1, \quad k = 1,\ldots,5, \\ 1, & x \ge 6. \end{cases}$$

Intuition 1. Each step *adds* the mass $P(X = k) = 1/6$; the graph is right–continuous (closed dot at the right end of each step). Remember that we are choosing an x and asking *how many values are less or equal to x*.

Intuition 2. The cdf is a running total of probability mass: as x moves to the right, $F_X(x)$ increases only when x passes a value that X can actually take (an atom). For the fair die, on each interval $[k, k+1)$ the cdf is constant $F_X(x) = k/6$, and at the integer k it jumps by exactly $P(X = k) = 1/6$. Right–continuity means $F_X(k) = k/6$ while the left limit is $F_X(k^-) = (k-1)/6$.

Example. Continous cdf. Let $X \sim Exp(\lambda)$ with $\lambda = 1$. Then

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 0, \\ 1 - \exp^{-x}, & x \geq 0. \end{cases}$$

Add plots?

Equality in distribution

Definition (Identically distributed)**.** Let X and Y be real-valued random variables (possibly on different probability spaces). We say that X and Y are *identically distributed*, written $X \stackrel{d}{=} Y$, if

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A) \qquad \text{for every } A \in \mathcal{B}.$$

Equivalently, their pushforward laws coincide: $\mathbb{P}_X = \mathbb{P}_Y$ on $(\mathbb{R}, \mathcal{B})$.

Remark. Equality in distribution is *not* equality as random variables: X and Y need not be equal almost surely, nor defined on the same sample space. They may also be dependent or independent; independence is unrelated to equality in distribution.

Theorem 15 (Characterizations)**.** *For real random variables X and Y, the following are equivalent:*

1. $X \stackrel{d}{=} Y$.
2. $F_X(x) = F_Y(x)$ *for every* $x \in \mathbb{R}$.

Proof sketch. (1)⇒(2): If $\mathbb{P}_X = \mathbb{P}_Y$, then for each x, $F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}_Y((-\infty, x]) = F_Y(x)$. (2)⇒(1): The family $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system generating $\mathcal{B}$. If two probability measures agree on this generator (the cdfs are equal), they agree on $\mathcal{B}$. TBD. □

Remark (Useful corollaries)**.** If X and Y are discrete, then $X \stackrel{d}{=} Y$ iff $p_X(x) = p_Y(x)$ for all x. If they admit densities, then $X \stackrel{d}{=} Y$ iff $f_X = f_Y$ almost everywhere.

Example. Let X be the number of heads in n fair tosses and $Y := n - X$ the number of tails. Then $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n, \frac{1}{2})$, hence $X \stackrel{d}{=} Y$, but generally $X \neq Y$.

Week 2 – Discussion

Problem 1. If P is a probability on $\mathcal{B}$ and $A, B \in \mathcal{B}$, then:

a) $P(B \cap A^c) = P(B) - P(A \cap B)$.

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

c) If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof.

a) Partition B as a disjoint union: $B = (B \cap A) \dot\cup (B \cap A^c)$. Hence

$$P(B) = P(B \cap A) + P(B \cap A^c) \quad \Rightarrow \quad P(B \cap A^c) = P(B) - P(A \cap B).$$

b) Note $A \cup B = A \dot\cup (B \cap A^c)$, so

$$P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B),$$

using part (a).

c) From $A \subseteq B$ we have $B = A \dot\cup (B \cap A^c)$, hence

$$P(B) = P(A) + P(B \cap A^c) \ \geq \ P(A).$$

Thus $P(A) \leq P(B)$. □

Problem 2. Prove that if A and B are independent, then A^c and B^c are independent.

Proof. Independence gives $P(A \cap B) = P(A)P(B)$. Then

$$\begin{aligned} P(A^c \cap B^c) &= 1 - P(A \cup B) = 1 - \big(P(A) + P(B) - P(A \cap B)\big) \\ &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) \\ &= P(A^c)\, P(B^c), \end{aligned}$$

so A^c and B^c are independent. □

Problem 3. Provide an alternative proof of Fatou's Lemma using the Dominated Convergence Theorem (DCT). *Make any assumptions necessary to apply DCT.*

Proof (via DCT). Let $(S, \mathcal{S}, \mu)$ be a measure space and let $(f_n)_{n \geq 1}$ be measurable with $0 \leq f_n \leq h$ a.e. for some $h \in L^1(\mu)$. Define for each n the *running lower envelope*

$$g_n(x) := \inf_{k \geq n} f_k(x), \qquad x \in S.$$

Then (g_n) is nondecreasing and $g_n(x) \uparrow g(x) := \liminf_{n\to\infty} f_n(x)$ for a.e. x. Moreover $0 \le g_n \le h$ a.e., so by DCT,

$$\int g\,d\mu \;=\; \lim_{n\to\infty} \int g_n\,d\mu.$$

For each fixed n we have $g_n \le f_k$ for all $k \ge n$, hence

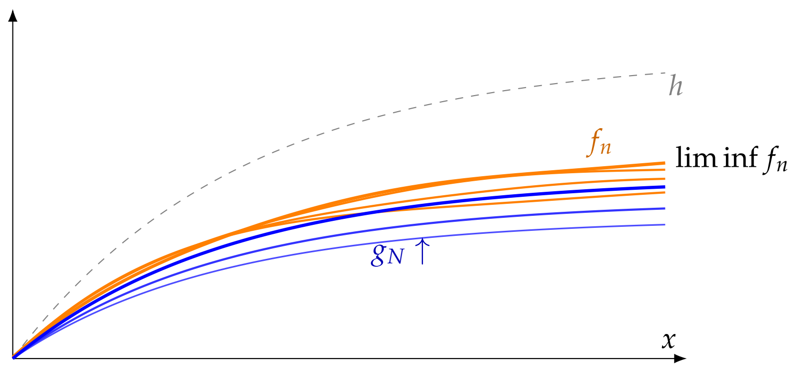
$$\int g_n\,d\mu \;\le\; \inf_{k\ge n} \int f_k\,d\mu.$$

Taking $n \to \infty$ yields

$$\int \liminf_{n\to\infty} f_n\,d\mu \;=\; \lim_n \int g_n\,d\mu \;\le\; \liminf_n \inf_{k\ge n} \int f_k\,d\mu \;=\; \liminf_{n\to\infty} \int f_n\,d\mu.$$

This is Fatou's inequality under the stated domination. □

Remark. If $f_n \uparrow f$ a.e. (monotone increase), then $\int f = \lim_n \int f_n$ by the Monotone Convergence Theorem, so Fatou's inequality holds with equality. If $f_n \to f$ a.e. and $0 \le f_n \le h \in L^1$, then $\int f = \lim_n \int f_n$ by DCT, again giving equality.



Orange curves are f_n; thin blue curves are the increasing lower envelopes $g_N = \inf_{k\ge N} f_k$; the thick blue curve is $\liminf f_n = \lim_{N\uparrow\infty} g_N$. A dominating $h \in L^1$ (dashed) ensures DCT applies. Early terms f_n may lie below $\liminf f_n$; the key property is that, for every fixed x and every $\varepsilon > 0$, there exists N such that $f_n(x) \ge \liminf_{k\to\infty} f_k(x) - \varepsilon$ for all $n \ge N$.

Problem 4. Let X have the standard Cauchy density $f_X(x) = \dfrac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.

(a) Show that $Y = \dfrac{1}{X}$ also has the standard Cauchy distribution.

(b) What is the expected value of X?

Solution. TBW.

(a) $Y = 1/X$ **is Cauchy.** Since $P(X=0) = 0$, the map $g(x) = 1/x$ is invertible a.s. with inverse $g^{-1}(y) = 1/y$ and $\left|\dfrac{d}{dy} g^{-1}(y)\right| = \dfrac{1}{y^2}$. By the change-of-variables formula,

$$f_Y(y) = f_X(g^{-1}(y)) \left|\tfrac{d}{dy} g^{-1}(y)\right| = \frac{1}{\pi(1+(1/y)^2)} \cdot \frac{1}{y^2} = \frac{1}{\pi(1+y^2)}, \qquad y \in \mathbb{R}.$$

Hence $Y \sim \text{Cauchy}(0,1)$.

(CDF check, piecewise). With $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$: for $y > 0$,

$$F_Y(y) = P(1/X \le y) = P(X \ge 1/y,\, X > 0) + P(X < 0) = \tfrac{1}{2} + \Big[1 - F_X(1/y)\Big] = \tfrac{1}{2} + \tfrac{1}{\pi}\arctan y.$$

For $y < 0$,

$$F_Y(y) = P(1/y \le X < 0) = F_X(0) - F_X(1/y) = -\tfrac{1}{\pi}\arctan(1/y) = \tfrac{1}{2} + \tfrac{1}{\pi}\arctan y.$$

Thus $F_Y(y) = \frac{1}{2} + \frac{1}{\pi}\arctan y$ for all y, the Cauchy CDF.

(b) $E[X]$ **does not exist.**

$$\int_{\mathbb{R}} |x|\, f_X(x)\, dx = \frac{2}{\pi}\int_1^\infty \frac{x}{1+x^2}\, dx = \frac{1}{\pi}\Big[\ln(1+x^2)\Big]_1^\infty = \infty.$$

Since $X \notin L^1$, the (Lebesgue) expectation $E[X]$ is *undefined* (the improper symmetric integral yields 0 as a principal value, but this is not a finite expectation). □

Week 3 — Class 5

Quantiles, PMFs, and a Geometric Example

Definition (Quantile function)**.** Let $F(x) = \mathbb{P}(X \le x)$ be a cdf on $\mathbb{R}$. The (generalized) α–quantile is

$$q(\alpha) \;:=\; \inf\{x \in \mathbb{R} : \; F(x) \ge \alpha\}, \qquad \alpha \in [0,1].$$

Equivalently, $F(q(\alpha)) \ge \alpha$ and $F(x) < \alpha$ for all $x < q(\alpha)$. The mapping $q : [0,1] \to \mathbb{R}$ sends probability levels to values of X.

Remark (Why the infimum?)**.** If F is strictly increasing and continuous, then $q(\alpha) = F^{-1}(\alpha)$ in the usual sense. When F has flat regions or jumps (discrete or mixed distributions), a strict inverse need not exist; the infimum definition always works and returns the leftmost value hitting level α.

Example (Named quantiles)**.** Median $= q(0.5)$. The pth percentile is $100p$ and equals $q(p)$. Quintiles are $q(0.2), q(0.4), q(0.6), q(0.8)$; deciles are $q(0.1), \ldots, q(0.9)$. Quantiles compactly summarize the distribution's location and spread.

Definition (Probability mass function (pmf))**.** A random variable X is *discrete* if it takes values in a countable set $\mathcal{X} \subset \mathbb{R}$. Its probability mass function is

$$f_X(x) \;:=\; \mathbb{P}(X = x), \qquad x \in \mathcal{X},$$

with $f_X(x) \ge 0$ and $\sum_{x\in\mathcal{X}} f_X(x) = 1$. For any $A \subseteq \mathcal{X}$, $\mathbb{P}(X \in A) = \sum_{x\in A} f_X(x)$ and $F(x) = \sum_{y\in\mathcal{X}:\, y\le x} f_X(y)$.

Example (Geometric distribution: "number of tosses until first head")**.** Let independent tosses have $\mathbb{P}(\text{head}) = p \in (0,1)$. Define $X = \min\{n \ge 1 : \text{the } n\text{th toss is head}\}$. Then X takes values in $\{1, 2, \ldots\}$ and

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1}p, \qquad x = 1, 2, \ldots$$

(the first $x-1$ are tails, then a head). Its cdf is

$$F_X(x) = \mathbb{P}(X \le x) = \sum_{i=1}^{\lfloor x\rfloor}(1-p)^{i-1}p = p\,\frac{1-(1-p)^{\lfloor x\rfloor}}{1-(1-p)} = 1-(1-p)^{\lfloor x\rfloor},$$

so for integer $x \ge 1$, $F_X(x) = 1-(1-p)^x$.

Remark (Parameterizations and support)**.** We can also define X on $\{0, 1, 2, \ldots\}$ ("number of failures before the first success"), with pmf $\tilde{f}(k) = (1-p)^k p$ and cdf $\tilde{F}(k) = 1-(1-p)^{k+1}$. We just need to be clear on the interpretation.

Continuous r.v.s: pmf vs. pdf and the cdf link

Proposition (Point probabilities for continuous X). *If X has a continuous cdf F_X, then for every $x \in \mathbb{R}$, $\mathbb{P}(X = x) = 0$.*

Proof. For any $\varepsilon > 0$,

$$\{X = x\} \subset (x - \varepsilon < X \le x) \quad \Rightarrow \quad \mathbb{P}(X = x) \le \mathbb{P}(x - \varepsilon < X \le x) = F_X(x) - F_X(x - \varepsilon).$$

By continuity of F_X, $F_X(x - \varepsilon) \to F_X(x)$ as $\varepsilon \downarrow 0$, so $0 \le \mathbb{P}(X = x) \le 0$, hence $\mathbb{P}(X = x) = 0$. $\square$

Definition (Probability density function (pdf)). A function $f_X : \mathbb{R} \to [0, \infty)$ is a pdf of X if

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{for all } x \in \mathbb{R},$$

equivalently, for any Borel A, $\mathbb{P}(X \in A) = \int_A f_X(t)\, dt$.

Remark (Fundamental link). If f_X is (Lebesgue) integrable and continuous at x, then

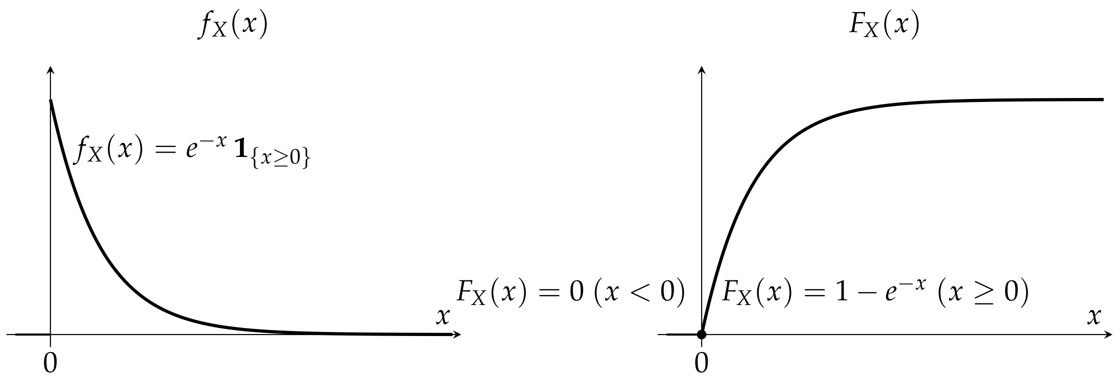
$$\frac{d}{dx} F_X(x) = f_X(x).$$

In discrete cases F_X is a step function and the derivative is not a useful notion; there we work with the pmf $p_X(x) = \mathbb{P}(X = x)$ and $F_X(x) = \sum_{y \le x} p_X(y)$.

Example (Exponential(λ)). For $\lambda > 0$, the pdf and cdf are

$$f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x \ge 0\}}, \qquad F_X(x) = \mathbb{P}(X \le x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

Indeed, $F_X(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}$ for $x \ge 0$ and $\frac{d}{dx} F_X(x) = \lambda e^{-\lambda x} = f_X(x)$ for $x > 0$.



Theorem 16 (Characterization of pmf/pdf). *Let X be a random variable.*

(i) (Discrete) A function $p : \mathcal{X} \to [0, \infty)$ is a pmf of X iff $\sum_{x \in \mathcal{X}} p(x) = 1$. Then $F_X(x) =$

$\sum_{y \le x} p(y)$.

(ii) (Continuous) A function $f : \mathbb{R} \to [0, \infty)$ with $\int_{\mathbb{R}} f(t)\, dt = 1$ is a pdf of X in the sense that

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t)\, dt, \qquad x \in \mathbb{R}.$$

Conversely, if F_X is a cdf that admits such a representation, then $f \ge 0$ a.e. and $\int_{\mathbb{R}} f = 1$.

Sketch for the continous case (ii) TBW. Each side of the if and only if: ($\Rightarrow$) If $F_X(x) = \int_{-\infty}^{x} f(t)\, dt$, then for $a < b$, $0 \le F_X(b) - F_X(a) = \int_a^b f(t)\, dt$, hence $f \ge 0$ a.e. (Lebesgue lemma). Also, by monotone convergence, $\int_{\mathbb{R}} f = \lim_{x\to\infty} \int_{-\infty}^{x} f = \lim_{x\to\infty} F_X(x) = 1$.

($\Leftarrow$) If $f \ge 0$ and $\int_{\mathbb{R}} f = 1$, define $F(x) = \int_{-\infty}^{x} f(t)\, dt$. Then F is nondecreasing since $F(b) - F(a) = \int_a^b f \ge 0$. Moreover, $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Finally, F is right–continuous: if $x_n \downarrow x$, then $F(x_n) - F(x) = \int_{(x,x_n]} f \to 0$ by absolute continuity of the Lebesgue integral. Thus F is a cdf and f is a pdf. $\square$

Uniform. If $X \sim U[a, b]$ with $a < b$, then

$$f_X(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x), \qquad F_X(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

Basic facts: $\mathbb{E}[X] = \frac{a+b}{2}$, $\mathbf{Var}(X) = \frac{(b-a)^2}{12}$.

Normal. If $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$, then

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \qquad F_X(x) = \Phi\Big(\frac{x-\mu}{\sigma}\Big),$$

where Φ is the standard normal cdf and $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ its pdf. When $\mu = 0$, $\sigma = 1$ we write $X \sim \mathcal{N}(0, 1)$, with pdf ϕ and cdf Φ.

Transformations of random variables

Pushforward definition (general measurable mapping). Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a r.v., and let $g : \mathcal{X} \to \mathcal{Y}$ be measurable. Define $Y = g(X)$. For any $A \in \mathcal{B}_{\mathcal{Y}}$,

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}\big(X \in g^{-1}(A)\big),$$

so the law of Y is the pushforward measure $\mathbb{P}_Y = \mathbb{P}_X \circ g^{-1}$.

Discrete case (countable support). If X is discrete with pmf p_X, then $Y = g(X)$ is discrete with

$$p_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{-1}(\{y\})} p_X(x), \qquad \text{and } \ p_Y(y) = 0 \text{ if } y \notin g(\mathcal{X}).$$

Recipe: enumerate the preimage $g^{-1}(y)$ and sum the appropriate masses.

Continuous case: cdf method. For any $y \in \mathbb{R}$,

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(X \in g^{-1}((-\infty, y])\big).$$

When g is strictly increasing and continuous, g^{-1} exists and $F_Y(y) = F_X\big(g^{-1}(y)\big)$ for all y; if g is strictly decreasing, $F_Y(y) = 1 - F_X\big(g^{-1}(y)\big)$ (right-limits understood when needed).

Example (Discrete transformation via preimages)**.** Let X take values $\{-2, -1, 0, 1, 2\}$ with pmf

$$f_X(-2) = 0.10, \quad f_X(-1) = 0.20, \quad f_X(0) = 0.40, \quad f_X(1) = 0.20, \quad f_X(2) = 0.10.$$

Define $Y = g(X) = |X|$. Then $\mathcal{Y} = \{0, 1, 2\}$ and

$$g^{-1}(0) = \{0\}, \qquad g^{-1}(1) = \{-1, 1\}, \qquad g^{-1}(2) = \{-2, 2\}.$$

By $f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x)$,

$$\begin{aligned} f_Y(0) &= f_X(0) = 0.40, \\ f_Y(1) &= f_X(-1) + f_X(1) = 0.20 + 0.20 = 0.40, \\ f_Y(2) &= f_X(-2) + f_X(2) = 0.10 + 0.10 = 0.20, \end{aligned}$$

and $\sum_{y \in \mathcal{Y}} f_Y(y) = 1$. *Recipe:* find $\mathcal{Y} = g(\mathcal{X})$, compute each preimage $g^{-1}(y)$, and sum f_X over it.

Continuous r.v.s: cdf method. If X has pdf f_X and $Y = g(X)$, then for any $y \in \mathbb{R}$,

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) = \mathbb{P}\big(\{x \in \mathcal{X} : g(x) \le y\}\big) = \int_{\{x \in \mathcal{X} : g(x) \le y\}} f_X(x)\, dx.$$

Note. Identifying the region $\{x : g(x) \le y\}$ may be hard when g is not monotone.

Monotone g: explicit cdf. If g is monotone so that g^{-1} is single-valued:

- If g is increasing,

$$\{x : g(x) \le y\} = \{x : x \le g^{-1}(y)\} \quad \Rightarrow \quad F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\, dx = F_X\big(g^{-1}(y)\big).$$

- If g is decreasing,

$$\{x : g(x) \leq y\} = \{x : x \geq g^{-1}(y)\} \quad \Rightarrow \quad F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\, dx = 1 - F_X\big(g^{-1}(y)\big).$$

Theorem 17 (Change of variables for a monotone g)**.** *Let X have pdf $f_X(x)$ and let $Y = g(X)$ where g is monotone. Let $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = g(\mathcal{X})$. Assume f_X is continuous and g^{-1} has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is*

$$f_Y(y) = \begin{cases} f_X\big(g^{-1}(y)\big) \left| \dfrac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise.} \end{cases}$$

Week 3 – Class 6

Transformations (Continuation)

Theorem 18. *(Piecewise Monotone) Theorem 2.1.8 in Casella & Berger deals with the case when g is monotone over certain intervals. In particular, suppose there are partitions $\{A_i\}_{i=1}^k$ of $\mathcal{X}$ and functions g_i defined on those partitions for which $g(x) = g_i(x)$ for $x \in A_i$, where $g_i(x)$ is monotone on A_i, and g_i^{-1} has a continuous derivative. Then we can still derive the pdf:*

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{Y} = \bigcup_{i=1}^k g(A_i)$. This can be useful, for example, if we have a squared transformation.

Example. (Casella & Berger, Ex. 2.1.9) Let $X \sim N(0,1)$, i.e., "standard normal distribution,"

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

Consider the transformation $Y = X^2$. $g(x) = x^2$ is monotone on $(-\infty, 0)$ and $(0, \infty)$. We use sets:

$$A_0 = \{0\}, \qquad A_1 = (-\infty, 0),\ g_1(x) = x^2,\ g_1^{-1}(y) = -\sqrt{y}, \qquad A_2 = (0, \infty),\ g_2(x) = x^2,\ g_2^{-1}(y) = \sqrt{y}.$$

This gives the pdf:

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

$$\Rightarrow \quad f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}, \qquad 0 < y < \infty \quad \text{"Chi-squared r.v. with 1 degree of freedom".}$$

Theorem 19. *(A second transformation: inverse transform sampling)*

- *If $X \sim F_X(x)$ and $Y = F_X(X)$, then $Y \sim U(0,1)$, i.e., $P(Y \le y) = y$, $0 < y < 1$.*
- *This tells us that if we want to generate (simulate) an observation X from a population with cdf $F_X(x)$, we can simulate a uniform random number $V \sim U(0,1)$ with realization u and solve for x in the equation $F_X(x) = u$.*

Proof. (Quick proof omitting some of the details about end-points and such.) We define $F_X^{-1}(y) = \inf\{x : F_X(x) \ge y\}$ to deal with F_X potentially being constant on some intervals and not being strictly increasing.

$$P(Y \le y) = P(F_X(X) \le y) = P(F_X^{-1}(F_X(X)) \le F_X^{-1}(y)) = P(X \le F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

□

Expected value

Definition. The expected value or mean of a random variable $g(X)$, denoted by $\mathbb{E}(g(X))$, is

$$\mathbb{E}(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx, & \text{if } X \text{ is continuous,} \\ \sum_{x \in \mathcal{X}} g(x)\, f_X(x) = \sum_{x \in \mathcal{X}} g(x)\, P(X = x), & \text{if } X \text{ is discrete,} \end{cases}$$

provided that the integral or sum exists. If $\mathbb{E}(g(X)) = \infty$, we say $\mathbb{E}(g(X))$ does not exist.

Theorem 20. *Let X be a random variable, and let a, b, c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

a. $\mathbb{E}\big(ag_1(X) + bg_2(X) + c\big) = a\,\mathbb{E}(g_1(X)) + b\,\mathbb{E}(g_2(X)) + c.$

b. If $g_1(x) \geq 0$ for all x, then $\mathbb{E}(g_1(X)) \geq 0$.

c. If $g_1(x) \geq g_2(x)$ for all x then $\mathbb{E}(g_1(X)) \geq \mathbb{E}(g_2(X))$.

d. If $a \leq g_1(x) \leq b$ for all x, then $a \leq \mathbb{E}(g_1(X)) \leq b$.

These are useful when computing expectations.

Proof. (discrete case) Let $p(x) = P(X = x)$ and $\mathcal{X} = \{x : \ p(x) > 0\}$.

(a)

$$\begin{aligned} \mathbb{E}\big(ag_1(X) + bg_2(X) + c\big) &= \sum_{x \in \mathcal{X}} \big(ag_1(x) + bg_2(x) + c\big)\, p(x) \\ &= a \sum_x g_1(x)p(x) + b \sum_x g_2(x)p(x) + c \sum_x p(x) \\ &= a\,\mathbb{E}(g_1(X)) + b\,\mathbb{E}(g_2(X)) + c, \end{aligned}$$

since $\sum_x p(x) = 1$.

(b) If $g_1(x) \geq 0$ and $p(x) \geq 0$ for all x, then each term $g_1(x)p(x) \geq 0$, hence $\mathbb{E}(g_1(X)) = \sum_x g_1(x)p(x) \geq 0$.

(Continuous case: replace $\sum_x g(x)p(x)$ by $\int g(x) f_X(x)\, dx$; the same algebra holds.)

(c) If $g_1(x) \geq g_2(x)$ for all x, then $g_1(x) - g_2(x) \geq 0$ for all x, so by (b)

$$\mathbb{E}\big(g_1(X) - g_2(X)\big) \geq 0 \ \Rightarrow\ \mathbb{E}(g_1(X)) - \mathbb{E}(g_2(X)) \geq 0,$$

using (a).

(d) If $a \leq g_1(x) \leq b$ for all x, multiply by $p(x) \geq 0$ and sum over x:

$$a \sum_x p(x) \ \leq\ \sum_x g_1(x)p(x) \ \leq\ b \sum_x p(x).$$

Since $\sum_x p(x) = 1$, this gives $a \le \mathbb{E}(g_1(X)) \le b$.

(Continuous case: $a \le g_1(x) \le b$ a.e. $\Rightarrow a \int f_X = a \le \int g_1 f_X = \mathbb{E}[g_1(X)] \le b \int f_X = b$.) $\square$

Nonlinear transformations and expectations. When working with nonlinear functions $g(x)$, one can either try to compute

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

directly, or do a transformation and find $f_Y(y)$ of $Y = g(X)$ and have

$$\mathbb{E}(g(X)) = E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy.$$

Example (Uniform-to-exponential via transformation)**.** Let $X \sim \mathrm{Unif}(0,1)$ and $g(x) = -\ln(1-x)$. We want $\mathbb{E}[g(X)]$.

Transformation. Define $Y = g(X) = -\ln(1-X)$. Then $g : (0,1) \to (0,\infty)$ is strictly increasing with

$$g^{-1}(y) = 1 - e^{-y}, \qquad \frac{d}{dy} g^{-1}(y) = e^{-y}.$$

Since $f_X(x) = 1$ on $(0,1)$,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = e^{-y}\, \mathbf{1}_{\{y \ge 0\}},$$

so $Y \sim \mathrm{Exp}(1)$ and

$$\mathbb{E}[g(X)] = \mathbb{E}[Y] = \int_0^{\infty} y e^{-y}\, dy = 1.$$

Direct computation.

$$\mathbb{E}[g(X)] = \int_0^1 -\ln(1-x)\, dx = \int_0^1 -\ln u\, du \quad (u = 1-x) = \big[-u \ln u + u\big]_0^1 = 1.$$

Definition (Convexity/concavity)**.** A function $g(x)$ is *convex* if for any $\lambda \in [0,1]$ and all x, y,

$$g\big(\lambda x + (1-\lambda) y\big) \;\le\; \lambda g(x) + (1-\lambda) g(y).$$

The function $g(x)$ is *concave* if

$$g\big(\lambda x + (1-\lambda) y\big) \;\ge\; \lambda g(x) + (1-\lambda) g(y).$$

Theorem 21 (Jensen's Inequality)**.** *For any random variable X, if g is convex then*

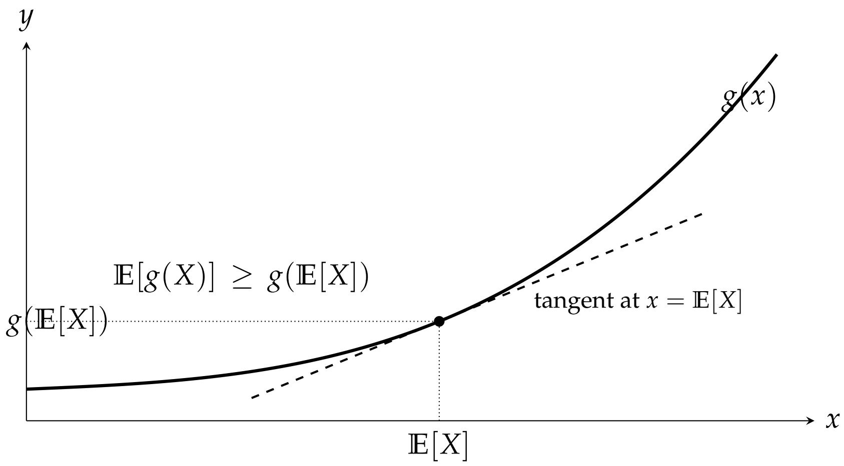
$$g\big(\mathbb{E}(X)\big) \;\le\; \mathbb{E}\big(g(X)\big).$$

If g is concave, then

$$\mathbb{E}\big(g(X)\big) \;\le\; g\big(\mathbb{E}(X)\big).$$

Example (Consequences via Jensen). • $\exp(\mathbb{E}X) \;\le\; \mathbb{E}\big(\exp X\big)$.

- $\big(\mathbb{E}X\big)^2 \;\le\; \mathbb{E}\big(X^2\big)$.
- If $X > 0$, then $\mathbb{E}(\log X) \;\le\; \log \mathbb{E}(X)$.
- If $X \ge 0$, then $\mathbb{E}\big(X^{1/2}\big) \;\le\; \big(\mathbb{E}X\big)^{1/2}$.
- $\big|\,\mathbb{E}X\,\big| \;\le\; \mathbb{E}|X|$.



Proof. Sketch: Compute the tangent (supporting) line at $m = \mathbb{E}(X)$. We know that $g(x) \ge a + b * x$ (by convexity). Taking expectations on both sides, $\mathbb{E}(g(X)) \;\ge\; a + b * \mathbb{E}(X)$. By construction of the tangent line at $\mathbb{E}(X)$, we have $a + b\mathbb{E}(X) \;=\; g(\mathbb{E}(X))$. It follows that $\mathbb{E}(g(X)) \ge g(\mathbb{E}(X))$. □

Theorem 22 (Markov; basis for Chebyshev). *Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,*

$$\Pr\big(g(X) \ge r\big) \;\le\; \frac{\mathbb{E}[g(X)]}{r}.$$

Proof. Using $g \ge 0$ and splitting the integral over the sets $\{x : g(x) < r\}$ and $\{x : g(x) \ge r\}$,

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int g(x)\, f_X(x)\, dx \\
&= \int_{\{g(x)<r\}} g(x)\, f_X(x)\, dx + \int_{\{g(x)\ge r\}} g(x)\, f_X(x)\, dx \\
&\ge\; r \int_{\{g(x)\ge r\}} f_X(x)\, dx \\
&= r\,\Pr(g(X) \ge r)
\end{aligned}$$

□

Remark. Equivalently, $\mathbf{1}_{\{g(X)\geq r\}} \leq g(X)/r$ (since $g \geq 0$); taking expectations yields Markov. This inequality is often used to obtain *conservative* probability bounds.

Corollary (Chebyshev)**.** For any $t > 0$,

$$\Pr(\,|X - \mathbb{E}X| \geq t\,) \;\leq\; \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. Apply Theorem 22 with $g(x) = (x - \mathbb{E}X)^2$ (nonnegative) and $r = t^2$. □

Moment Generating Functions

Definition (Moments)**.** The expected value (the *mean*) is the first moment of X:

$$\mu = \mathbb{E}(X).$$

For each integer $n \geq 1$, the nth (raw) moment of X is

$$\mu'_n \;=\; \mathbb{E}(X^n)\,.$$

The nth *central* moment of X is

$$\mu_n \;=\; \mathbb{E}[(X-\mu)^n]\,, \qquad \text{where } \mu = \mu'_1 = \mathbb{E}(X).$$

Remark. For $n > 1$ these are often called *higher-order moments*.

Definition (Variance and standard deviation)**.** The *variance* of a random variable X is its second central moment,

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

The positive square root of $\mathrm{Var}(X)$ is the *standard deviation* of X.

Remark (Interpretation)**.** Small variance (and hence small standard deviation) means X is very likely to be close to its mean $\mathbb{E}(X)$. The standard deviation has the same units as X, which aids interpretation.

Theorem 23 (Variance of an affine transformation)**.** *If X is a random variable with finite variance, then for any constants a, b,*

$$\mathbf{Var}(aX + b) = a^2\,\mathbf{Var}(X).$$

Proof.

$$\begin{aligned}\mathbf{Var}(aX+b) &= \mathbb{E}\Big[\big((aX+b)-\mathbb{E}(aX+b)\big)^2\Big]\\ &= \mathbb{E}\Big[\big(aX+b-a\,\mathbb{E}(X)-b\big)^2\Big]\\ &= \mathbb{E}\Big[\big(a(X-\mathbb{E}X)\big)^2\Big]\\ &= a^2\,\mathbb{E}\Big[\big(X-\mathbb{E}X\big)^2\Big] = a^2\,\mathbf{Var}(X).\end{aligned}$$

□

Proposition (Useful relationship)**.**

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

Definition (Symmetry about 0)**.** A distribution is *symmetric* about 0 if its cdf satisfies

$$F(x) = 1 - F(-x) \quad \text{for all } x.$$

If X has a density f, then this is equivalent to $f(x) = f(-x)$ (an even density).

Proposition (Odd moments vanish under symmetry)**.** *If a random variable X is symmetric about 0 and $\mathbb{E}|X|^m < \infty$ for an odd integer m, then*

$$\mathbb{E}\big[X^m\big] = 0.$$

Proof (continuous case). Write

$$\mathbb{E}\big[X^m\big] = \int_{-\infty}^{\infty} x^m f(x)\,dx = \int_0^{\infty} x^m f(x)\,dx + \int_{-\infty}^{0} x^m f(x)\,dx.$$

In the second integral substitute $x = -t$ (so $t \geq 0$ and $dx = -dt$):

$$\int_{-\infty}^{0} x^m f(x)\,dx = \int_{\infty}^{0} (-t)^m f(-t)\,(-dt) = -\int_0^{\infty} (-t)^m f(-t)\,dt.$$

Since m is odd, $(-t)^m = -t^m$, and by symmetry $f(-t) = f(t)$. Hence

$$\int_{-\infty}^{0} x^m f(x)\,dx = -\int_0^{\infty} \big(-t^m\big) f(t)\,dt = \int_0^{\infty} t^m f(t)\,dt.$$

Therefore the two halves cancel:

$$\mathbb{E}\big[X^m\big] = \int_0^{\infty} x^m f(x)\,dx - \int_0^{\infty} x^m f(x)\,dx = 0.$$

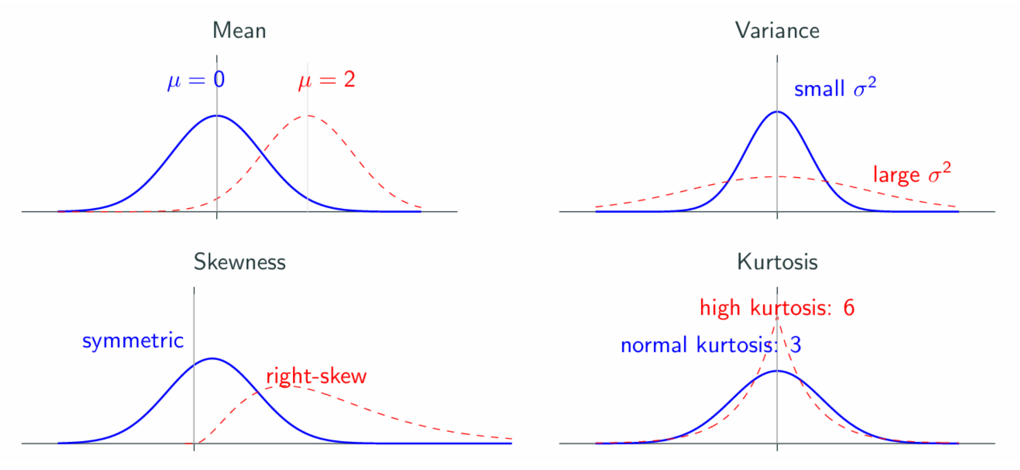
One-line viewpoint (your "easier road"): when f is even and m is odd, the integrand $x^m f(x)$ is an odd function, so its integral over the symmetric domain $(-\infty, \infty)$ is 0 (provided $\mathbb{E}|X|^m < \infty$). □

Remark (Higher-order moments)**.** Traditionally, most focus is on the first and second moments. One reason is that the normal distribution can be fully characterized by its first two moments. Higher-order moments are increasingly used in economics/finance, particularly:

- **Skewness**: third central moment — measures the asymmetry of a distribution. The normal distribution has skewness 0; it is symmetric.
- **Kurtosis**: fourth central moment — measures the thickness of the tails of a distribution. For the normal distribution, the kurtosis is 3.

We often talk about *excess kurtosis*, which is kurtosis -3, i.e., the excess relative to the normal distribution.

Figure 9: MHigher Order Moments



Definition (Moment generating function (mgf))**.** Let X be a random variable with cdf F_X. The *moment generating function* (mgf) of X, denoted $M_X(t)$, is

$$M_X(t) \;=\; \mathbb{E}(e^{tX}),$$

provided the expectation exists for some real t in a neighborhood of 0. That is, there exists $h > 0$ such that for all $t \in (-h, h)$, $\mathbb{E}(e^{tX})$ exists. (Here t is a real parameter/argument of the mgf.)

Given the expected value definition, we can also write

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx \quad \text{if } X \text{ is continuous,}$$

and

$$M_X(t) = \sum_x e^{tx}\,\mathbb{P}(X = x) \quad \text{if } X \text{ is discrete.}$$

If the expectation does not exist in a neighborhood of 0, we say the mgf does not exist.

Example (MGF of a Bernoulli(p)). Let $X \sim \text{Bernoulli}(p)$, so $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. By the discrete mgf definition,

$$M_X(t) \;=\; \sum_x e^{tx}\,\mathbb{P}(X = x) \;=\; (1-p)e^{t\cdot 0} + p\,e^{t\cdot 1} \;=\; 1 - p + p\,e^t.$$

Theorem 24 (MGF and moments). *If X has moment generating function $M_X(t)$, then for any integer $n \geq 1$,*

$$\mathbb{E}[X^n] \;=\; M_X^{(n)}(0),$$

where

$$M_X^{(n)}(0) \;=\; \frac{d^n}{dt^n}\,M_X(t)\bigg|_{t=0}.$$

That is, the nth moment of X is the nth derivative of its mgf, evaluated at $t = 0$.

Example (Bernoulli(p) via Theorem 24). Recall that for $X \sim \text{Bernoulli}(p)$ we obtained

$$M_X(t) = 1 - p + pe^t.$$

Applying Theorem 24:

- First derivative at $t = 0$:
$$M_X'(0) = p = \mathbb{E}[X].$$
- Second derivative at $t = 0$:
$$M_X''(0) = p, \quad \Rightarrow \quad \mathbf{Var}(X) = M_X''(0) - \big(M_X'(0)\big)^2 = p - p^2 = p(1-p).$$

Thus, the mgf reproduces the mean and variance of the Bernoulli distribution.

Proof of Theorem 24. Consider the continuous case (the discrete case is analogous).

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx}\,f_X(x)\,dx.$$

Differentiate under the integral (justified if M_X exists in a neighborhood of 0):

$$\frac{d}{dt}M_X(t) = \int_{-\infty}^{\infty} xe^{tx}\,f_X(x)\,dx = \mathbb{E}\left[Xe^{tX}\right].$$

Evaluating at $t = 0$ gives $\mathbb{E}[X]$.

By induction, differentiating n times yields

$$\frac{d^n}{dt^n}M_X(t) = \mathbb{E}\left[X^n e^{tX}\right].$$

Evaluating at $t = 0$ gives

$$M_X^{(n)}(0) = \mathbb{E}[X^n],$$

which establishes the result. $\square$

Moment–Generating Function: interchange and uniqueness

Remark (Flipping differentiation and integration)**.** From the Leibniz rule,

$$\frac{d}{d\theta}\int_a^b f(x,\theta)\,dx \;=\; \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

When the range of integration may be infinite, it is safer to rewrite the derivative as a limit,

$$\frac{\partial}{\partial\theta} f(x,\theta) \;=\; \lim_{\delta\to 0}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta},$$

so that

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)\,dx \;=\; \lim_{\delta\to 0}\int_{-\infty}^{\infty}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}\,dx.$$

Thus, the question reduces to *interchanging* an integral and a limit; this is where one tries to apply Lebesgue's Dominated Convergence Theorem (from the previous lecture).

Lemma (A DCT-ready bound (Lipschitz condition near θ_0))**.** *Suppose there exist a function $g(x,\theta_0)$ and a constant $\delta_0 > 0$ such that*

$$\left|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\right| \;\le\; g(x,\theta_0), \qquad \text{for all } x \text{ and } |\delta| \le \delta_0.$$

Then the difference quotients are dominated and one may apply the Dominated Convergence Theorem to justify interchanging the limit and the integral near θ_0.

Note (for Theorem 2.3.7): *in our mgf setting we require the integrand $e^{tx} f_X(x)$ to satisfy such a domination near $t = 0$. The book treats the discrete and continuous cases separately (via Theorem 2.4.8 for sums and differentiation), but a discrete pmf can also be viewed as a simple function so that the same DCT logic applies.*

Remark (Do moments determine the distribution?)**.** If the mgf exists, it can generate (infinitely many) moments. Do moments uniquely determine the cdf?

- **Not in general:** two distinct random variables can share all moments.
- **Yes, with bounded support:** moments determine the distribution.
- **Yes, with an mgf near** 0**:** existence of $M_X(t)$ in a neighborhood of 0 pins down the distribution.

Theorem 25 (2.3.11)**.** *Let $F_X(x)$ and $F_Y(y)$ be two cdfs for which all moments exist.*

a) If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if $\mathbb{E}X^r = \mathbb{E}Y^r$ for all $r = 0, 1, 2, \ldots$

b) If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Theorem 26 (Convergence of mgfs). *Suppose $\{X_i\}_{i\geq 1}$ is a sequence of random variables with mgfs $M_{X_i}(t)$. Assume that for some $h > 0$,*

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t) \qquad \text{for all } t \in (-h, h),$$

and that the pointwise limit $M_X(t)$ is itself an mgf. Then there exists a unique cdf F_X whose moments are determined by M_X, and for every continuity point x of F_X,

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x).$$

Equivalently, $X_i \overset{d}{\Rightarrow} X$.

Idea of proof. By assumption, M_X exists on a neighborhood of 0, so it uniquely determines a distribution F_X. For any bounded, continuous f, approximate f by polynomials and then by exponentials e^{tx} for small t—objects controlled by mgfs. The pointwise convergence $M_{X_i}(t) \to M_X(t)$ on $(-h, h)$ transfers to convergence of integrals against these approximants, which yields convergence of cdfs at continuity points of F_X. □

Remark. Convergence of mgfs on a neighborhood of 0 is *sufficient* (but not necessary) for convergence in distribution of a sequence of random variables.

Example (Binomial→Poisson via mgfs). Let $X_n \sim \text{Binomial}\big(n, \frac{\lambda}{n}\big)$ with fixed $\lambda > 0$. Then $\mathbb{E}[X_n] = \lambda$ for all n, and the mgf of X_n is

$$M_{X_n}(t) = \mathbb{E}\left[e^{tX_n}\right] = \sum_{k=0}^{n} e^{tk} \binom{n}{k} \left(\tfrac{\lambda}{n}\right)^k \left(1 - \tfrac{\lambda}{n}\right)^{n-k}.$$

Recognize a binomial expansion with

$$a = 1 - \frac{\lambda}{n}, \qquad b = \frac{\lambda}{n} e^t.$$

Therefore,

$$M_{X_n}(t) = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^t\right)^n = \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n.$$

Using the classical limit

$$\lim_{n\to\infty} \left(1 + \frac{y}{n}\right)^n = e^y \qquad (\text{fixed } y \in \mathbb{R}),$$

with $y = \lambda(e^t - 1)$, we obtain

$$\lim_{n\to\infty} M_{X_n}(t) = \exp\big(\lambda(e^t - 1)\big) =: M(t).$$

But $M(t) = \exp\big(\lambda(e^t - 1)\big)$ is the mgf of Poisson(λ). Since M exists in a neighborhood of 0, by Theorem 26 we conclude

$$X_n \overset{d}{\Rightarrow} \text{Poisson}(\lambda).$$

Lemma. *For any fixed* $y \in \mathbb{R}$, $\lim_{n\to\infty} \left(1 + \frac{y}{n}\right)^n = e^y$.

Proof. Take logs: $n \log\big(1 + \frac{y}{n}\big) \to y$ by $\log(1+u) = u + o(u)$ as $u \to 0$; exponentiate. □

Remark (What to watch for in this example)**.** 1. The binomial theorem step is purely algebraic; it packages the sum defining the mgf into $(a+b)^n$.

2. The limit uses Lemma with $y = \lambda(e^t - 1)$, which is valid for all t.

3. The limit function is an *actual* mgf (Poisson), so the hypothesis of Theorem 26 is satisfied.

Characteristic Functions and the Normal Distribution

Definition (Characteristic Function)**.** For a random variable X, the *characteristic function* is

$$\phi_X(t) \;=\; \mathbb{E}\Big(e^{itX}\Big), \qquad i = \sqrt{-1}.$$

Characteristic functions always exist (the integrand has modulus 1), they completely determine the distribution, and each cdf has a unique characteristic function.

Remark. Unlike mgfs, $\phi_X(t)$ exists even when moments (or the mgf) do not. This is why characteristic functions are especially useful for convergence results.

Theorem 27 (Lévy continuity theorem: convergence via c.f.'s)**.** *Let* X_k, $k = 1, 2, \ldots$ *be random variables with characteristic functions* $\phi_{X_k}(t)$. *Suppose that for all* t *in a neighborhood of* 0,

$$\lim_{k\to\infty} \phi_{X_k}(t) \;=\; \phi_X(t),$$

and that $\phi_X(t)$ *is a characteristic function. Then, for every* x *at which* F_X *is continuous,*

$$\lim_{k\to\infty} F_{X_k}(x) \;=\; F_X(x).$$

Remark. In words: convergence of characteristic functions (near 0) implies convergence of cdfs (i.e., convergence in distribution).

Definition (Normal (Gaussian) distribution)**.** A random variable X is said to have a normal distribution with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its pdf is

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

Remark. The normal plays a central role in statistics and economics: it is tractable, has the familiar bell shape, and can well-approximate many distributions in large samples.

Proposition (Standardization and tractability)**.** *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the standardized variable*

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

has the standard normal distribution, $Z \sim \mathcal{N}(0,1)$.

Remark. This is convenient: probabilities and expectations can be computed for Z and then transformed back to $X \sim \mathcal{N}(\mu, \sigma^2)$. Variances transform analogously. The two parameters (μ, σ) fully describe the location and scale (shape and location) of the distribution, making the normal part of the location–scale family.

Proposition (Empirical 68–95–99.7 rule)**.** *For $X \sim \mathcal{N}(\mu, \sigma^2)$,*

$$\mathbb{P}(|X-\mu| \le \sigma) \approx 0.68, \qquad \mathbb{P}(|X-\mu| \le 2\sigma) \approx 0.95, \qquad \mathbb{P}(|X-\mu| \le 3\sigma) \approx 0.997.$$

Equivalently, for $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(|Z| \le 1) \approx 0.68$, $\mathbb{P}(|Z| \le 2) \approx 0.95$, and $\mathbb{P}(|Z| \le 3) \approx 0.997$.

Location and Scale Families

Theorem 28 (Location and scale transformation)**.** *Let $f(x)$ be any pdf and let $\mu \in \mathbb{R}$ and $\sigma > 0$ be constants. Define*

$$g(x \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$

Then $g(x \mid \mu, \sigma)$ is a pdf.

Proof. Since f is a pdf, $f(x) \ge 0$ for all x. Hence $g(x \mid \mu, \sigma) \ge 0$. Next, check normalization:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx = \int_{-\infty}^{\infty} f(y)\, dy = 1,$$

where the change of variable $y = (x-\mu)/\sigma$ was used. Thus g is a valid pdf. □

Remark (Families of distributions)**.**

- **Location family:** If $f(x)$ is a pdf, then $\{f(x-\mu) : \mu \in \mathbb{R}\}$ is the location family with standard pdf $f(x)$. The parameter μ shifts the distribution left/right.
- **Scale family:** If $f(x)$ is a pdf, then $\left\{\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) : \sigma > 0\right\}$ is the scale family with standard pdf $f(x)$. The parameter σ stretches ($\sigma > 1$) or contracts ($\sigma < 1$) the distribution.

- **Location–scale family:** If $f(x)$ is a pdf, then

$$\left\{\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma > 0\right\}$$

is the location–scale family. Here μ is the location parameter and σ is the scale parameter.

Remark. As with the normal distribution, calculations can often be carried out using the *standard* pdf $f(x)$ and then transferred to the whole family via the location and scale transformation.

Week 3 – Discussion

Transformations of random variables: Theorems 2.1.5 and 2.1.8

Let X be a real-valued r.v. with continuous pdf f_X supported on $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$, and let $Y = g(X)$. Write $\mathcal{Y} = \{y \in \mathbb{R} : y = g(x) \text{ for some } x \in \mathcal{X}\}$.

Theorem 29 (Monotone transformation (2.1.5))**.** *Suppose $g : \mathcal{X} \to \mathcal{Y}$ is strictly monotone and differentiable with g^{-1} differentiable on $\mathcal{Y}$. Then Y has pdf*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\,\left|\frac{d}{dy}g^{-1}(y)\right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently, writing $x = g^{-1}(y)$, $f_Y(y) = \dfrac{f_X(x)}{|g'(x)|}$ for $y \in \mathcal{Y}$.

Sketch. If g is strictly increasing, then $F_Y(y) = P(Y \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$. Differentiate and use the chain rule to obtain $f_Y(y) = f_X(g^{-1}(y))\,(g^{-1})'(y)$. Since $(g^{-1})'(y) = 1/g'(x) > 0$, this equals $f_X(x)/g'(x)$. If g is strictly decreasing, the inequality reverses, $F_Y(y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y))$, and differentiation introduces a minus sign; taking absolute values gives the stated formula in both cases. □

Theorem 30 (Non–monotone transformation (2.1.8))**.** *Suppose g is differentiable on $\mathcal{X}$ and there exist disjoint intervals $\mathcal{X}_1, \dots, \mathcal{X}_m$ that cover $\mathcal{X}$ such that each restriction $g|_{\mathcal{X}_j}$ is strictly monotone with a differentiable inverse $x_j(\cdot)$ onto its image. Then $Y = g(X)$ has pdf, for $y \in \mathcal{Y}$,*

$$f_Y(y) = \sum_{j=1}^{m} f_X(x_j(y))\,\left|\frac{d}{dy}x_j(y)\right| = \sum_{x \in g^{-1}(\{y\})} \frac{f_X(x)}{|g'(x)|},$$

and $f_Y(y) = 0$ for $y \notin \mathcal{Y}$.

Remark (Why monotonicity matters)**.** If g is not one-to-one globally, $g^{-1}(y)$ is multi-valued. Theorem 30 says: split the domain into monotone branches, invert on each branch, and sum the Jacobian-adjusted contributions. The absolute value accounts for the sign of g' (increasing vs. decreasing branch).

Cookbook procedure for Theorem 29.

1. Verify that $g(\cdot)$ is strictly monotone on $\mathcal{X}$ (hence invertible onto $\mathcal{Y}$).
2. Compute the inverse $x = g^{-1}(y)$ and its derivative $(g^{-1})'(y) = 1/g'(x)$.
3. Evaluate f_X at the inverse: $f_X(g^{-1}(y))$.
4. Multiply by the Jacobian factor: $f_Y(y) = f_X(g^{-1}(y))\,|(g^{-1})'(y)|$.

5. Set $f_Y(y) = 0$ for $y \notin \mathcal{Y}$ and check that $\int_{\mathcal{Y}} f_Y(y)\,dy = 1$.

Cookbook for non–monotone g (Theorem 30).

1. Partition $\mathcal{X}$ into disjoint intervals where g is strictly monotone.
2. For a given y, solve $g(x) = y$ on each branch to get the preimages $x_j(y)$.
3. Sum the branchwise contributions: $f_Y(y) = \sum_j f_X\big(x_j(y)\big)\,\big|1/g'(x_j(y))\big|$.
4. Declare $f_Y(y) = 0$ when no preimage exists (i.e. $y \notin \mathcal{Y}$).

Theorem 31 (Monotone transformation; cf. Thm. 2.1.5)**.** *Let X have pdf f_X with support $\mathcal{X} = \{x : f_X(x) > 0\}$ and let $Y = g(X)$, where $g : \mathcal{X} \to \mathbb{R}$ is monotone (either increasing or decreasing) and invertible on $\mathcal{X}$. Assume f_X is continuous on $\mathcal{X}$ and g^{-1} is continuously differentiable on $\mathcal{Y} := g(\mathcal{X})$. Then the pdf of Y is*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\ \left|\frac{d}{dy}g^{-1}(y)\right|, & y \in \mathcal{Y}, \\ 0, & \textit{otherwise.} \end{cases}$$

Proof sketch. If g is increasing, $F_Y(y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$; if decreasing, $F_Y(y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y))$. Differentiating and using the chain rule yields the stated density with the absolute derivative. □

Cookbook for Thm. 2.1.5

- **Inputs:** pdf f_X and a monotone g.
- **Goal:** derive f_Y of $Y = g(X)$.
- **Steps:**
  1. Verify g is monotone and invertible on $\mathcal{X}$; set $\mathcal{Y} = g(\mathcal{X})$.
  2. Compute $g^{-1}(y)$ for $y \in \mathcal{Y}$.
  3. Compute $\frac{d}{dy}g^{-1}(y)$.
  4. Plug into $f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$ and restrict to $\mathcal{Y}$.

Remark (Why monotonicity matters)**.** If g is not monotone on $\mathcal{X}$, it is not globally invertible on $\mathcal{X}$. You must partition $\mathcal{X}$ into regions where g *is* monotone and sum the branch contributions (Theorem 32 below).

Theorem 32 (Piecewise monotone transformation; cf. Thm. 2.1.8)**.** *Let X have pdf f_X with support $\mathcal{X}$. Suppose there is a finite partition $\{A_0, A_1, \dots, A_k\}$ of $\mathcal{X}$ with $P(X \in A_0) = 0$ and, for $i = 1, \dots, k$, functions $g_i : A_i \to \mathbb{R}$ such that:*

(i) $g(x) = g_i(x)$ *for* $x \in A_i$ *(so* g *agrees with* g_i *on* A_i*);*

(ii) g_i *is monotone on* A_i*;*

(iii) The image set $\mathcal{Y} := \{y : \exists x \in A_i \text{ s.t. } y = g_i(x)\}$ *is the same for all* i*;*

(iv) g_i^{-1} *exists on* $\mathcal{Y}$ *and is continuously differentiable there.*

Then the pdf of $Y = g(X)$ *is*

$$f_Y(y) = \begin{cases} \displaystyle\sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise.} \end{cases}$$

Cookbook for Thm. 2.1.8

- **Inputs:** pdf f_X and $g(\cdot)$ satisfying the piecewise conditions.
- **Typical classroom setting:** $k = 2$, $A_0 = \{0\}$, $A_1 \subset \mathbb{R}_{--}$, $A_2 \subset \mathbb{R}_{++}$, often with $A_1 = -A_2$ (e.g., even g).
- **Goal:** derive f_Y for $Y = g(X)$.
- **Steps:**
  1. Determine the partition $A_0, \ldots, A_k$ and $\mathcal{Y}$.
  2. Check (i)–(iv) hold.
  3. For each branch i, compute $g_i^{-1}(y)$ on $\mathcal{Y}$.
  4. Evaluate $f_X(g_i^{-1}(y))$ for each i.
  5. Compute $\frac{d}{dy} g_i^{-1}(y)$ for each i.
  6. Sum the branch contributions to obtain $f_Y(y)$ on $\mathcal{Y}$.

Example (Even transform; $k = 2$). If $Y = X^2$ and $P(X = 0) = 0$, then with $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$, $g_1^{-1}(y) = -\sqrt{y}$, $g_2^{-1}(y) = \sqrt{y}$, and

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} \mathbf{1}_{\{y>0\}}.$$

Problem 1. In each of the following, find the pdf of Y and show that it integrates to 1.

(a) $f_X(x) = \frac{1}{2} e^{-|x|}$ for $x \in \mathbb{R}$ (Laplace), and $Y = |X|^3$.

Solution. TBW Partition $A_0 = \{0\}$, $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$, with $g_1(x) = -x^3$ on A_1 and $g_2(x) = x^3$ on A_2. Then $g_1^{-1}(y) = -y^{1/3}$ and $g_2^{-1}(y) = y^{1/3}$, with $\left| \frac{d}{dy} g_i^{-1}(y) \right| =$

$\frac{1}{3}y^{-2/3}$. Hence, for $y > 0$,

$$f_Y(y) = \sum_{i=1}^{2} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| = \left(\tfrac{1}{2}e^{-y^{1/3}} + \tfrac{1}{2}e^{-y^{1/3}}\right) \cdot \frac{1}{3}y^{-2/3} = \frac{1}{3}e^{-y^{1/3}}y^{-2/3}.$$

Otherwise $f_Y(y) = 0$.

Check $\int f_Y = 1$. With $t = y^{1/3}$, $dy = 3t^2\,dt$:

$$\int_0^\infty \frac{1}{3}e^{-y^{1/3}}y^{-2/3}\,dy = \int_0^\infty e^{-t}\,dt = 1.$$

(b) $f_X(x) = \frac{3}{8}(x+1)^2$ for $-1 < x < 1$, and $Y = 1 - X^2$.

Solution. TBW Partition $A_1 = (-1,0)$ and $A_2 = (0,1)$ with $g_1(x) = 1 - x^2$ (increasing on A_1), $g_2(x) = 1 - x^2$ (decreasing on A_2). For $y \in (0,1)$,

$$g_1^{-1}(y) = -\sqrt{1-y}, \quad g_2^{-1}(y) = +\sqrt{1-y}, \quad \left| \frac{d}{dy} g_i^{-1}(y) \right| = \frac{1}{2\sqrt{1-y}}.$$

Thus

$$\begin{aligned} f_Y(y) &= \frac{3}{8}\frac{(1-\sqrt{1-y})^2}{2\sqrt{1-y}} + \frac{3}{8}\frac{(1+\sqrt{1-y})^2}{2\sqrt{1-y}} \\ &= \frac{3}{8}\Big((1-y)^{-1/2} + (1-y)^{1/2}\Big), \qquad 0 < y < 1, \end{aligned}$$

and $f_Y(y) = 0$ otherwise.

Check $\int f_Y = 1$.

$$\int_0^1 \frac{3}{8}(1-y)^{-1/2}\,dy = \frac{3}{8} \cdot 2 = \frac{3}{4}, \qquad \int_0^1 \frac{3}{8}(1-y)^{1/2}\,dy = \frac{3}{8} \cdot \frac{2}{3} = \frac{1}{4}.$$

Sum $= 1$.

(c) $f_X(x) = \frac{3}{8}(x+1)^2$ for $-1 < x < 1$, and

$$Y = \begin{cases} 1 - X^2, & X \le 0, \\ 1 - X, & X > 0. \end{cases}$$

Solution. TBW. Take $A_1 = (-1,0]$ with $g_1(x) = 1 - x^2$ (increasing), and $A_2 = (0,1)$ with $g_2(x) = 1 - x$ (decreasing). For $y \in (0,1)$,

$$g_1^{-1}(y) = -\sqrt{1-y}, \quad \left| \frac{d}{dy} g_1^{-1}(y) \right| = \frac{1}{2\sqrt{1-y}}, \qquad g_2^{-1}(y) = 1 - y, \quad \left| \frac{d}{dy} g_2^{-1}(y) \right| = 1.$$

Therefore

$$\boxed{f_Y(y) = \frac{3}{16}\left(\frac{1}{\sqrt{1-y}} - 2 + \sqrt{1-y}\right) + \frac{3}{8}(2-y)^2, \quad 0 < y < 1,}$$

(and $f_Y(y) = 0$ otherwise). Equivalently,

$$f_Y(y) = \frac{3}{16}\frac{(1-\sqrt{1-y})^2}{\sqrt{1-y}} + \frac{3}{8}(2-y)^2.$$

Check $\int f_Y = 1$. Let $s = \sqrt{1-y}$, $dy = -2s\,ds$:

$$\int_0^1 \frac{3}{16}\left(\frac{1}{\sqrt{1-y}} - 2 + \sqrt{1-y}\right) dy = \frac{3}{8}\int_0^1 (1-s)^2\,ds = \frac{1}{8}.$$

Also,

$$\int_0^1 \frac{3}{8}(2-y)^2\,dy = \frac{3}{8}\left[4y - 2y^2 + \frac{y^3}{3}\right]_0^1 = \frac{7}{8}.$$

Sum $= 1$.

Problem 2. Show the following (a) Let X be a continuous, nonnegative random variable with cdf F_X. Show that

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\,dx.$$

(b) Let X be a nonnegative, integer–valued random variable with cdf $F_X(k) = \mathbb{P}(X \le k)$. Show that

$$\mathbb{E}[X] = \sum_{k=0}^{\infty}(1 - F_X(k)) = \sum_{k=0}^{\infty}\mathbb{P}(X > k).$$

Solution. TBW.

(a) Continuous case. Since $X \ge 0$ and f_X is its pdf,

$$\begin{aligned}
\int_0^\infty (1 - F_X(x))\,dx &= \int_0^\infty \mathbb{P}(X > x)\,dx \\
&= \int_0^\infty \int_x^\infty f_X(y)\,dy\,dx \\
&= \int_0^\infty \int_0^y dx\, f_X(y)\,dy \\
&= \int_0^\infty y\, f_X(y)\,dy \\
&= \mathbb{E}[X],
\end{aligned}$$

where the change in the order of integration is justified by Tonelli/Fubini (since the integrand is nonnegative).

(b) Discrete case. For $X \in \{0, 1, 2, \dots\}$ and $X \geq 0$,

$$X = \sum_{k=0}^{\infty} \mathbf{1}\{X > k\} \quad \text{a.s.}$$

Taking expectations and using Monotone Convergence,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbf{1}\{X > k\}\right] = \sum_{k=0}^{\infty} \mathbb{P}(X > k) = \sum_{k=0}^{\infty} (1 - F_X(k)).$$

This matches the continuous formula with integrals replaced by sums. □

Problem 3. Let X have pdf $f_X(x) = \frac{1}{2}(1 + x)$ for $-1 < x < 1$ and 0 otherwise.

(a) Find the pdf of $Y = X^2$.

(b) Compute $\mathbb{E}[Y]$ and $\text{Var}(Y)$.

Solution. *TBW.*

(a) Pdf of $Y = X^2$. Partition $\mathcal{X} = (-1, 1)$ as $A_0 = \{0\}$, $A_1 = (-1, 0)$, $A_2 = (0, 1)$. On A_1 and A_2 the map $g(x) = x^2$ is monotone with

$$g_1^{-1}(y) = -\sqrt{y}, \qquad g_2^{-1}(y) = +\sqrt{y}, \qquad \left|\frac{d}{dy} g_i^{-1}(y)\right| = \frac{1}{2\sqrt{y}}.$$

By Theorem 2.1.8, for $0 < y < 1$,

$$\begin{aligned} f_Y(y) &= \sum_{i=1}^{2} f_X(g_i^{-1}(y)) \left|\frac{d}{dy} g_i^{-1}(y)\right| \\ &= \left[\frac{1}{2}(1 - \sqrt{y}) + \frac{1}{2}(1 + \sqrt{y})\right] \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2} y^{-1/2}, \end{aligned}$$

and $f_Y(y) = 0$ otherwise. (Note that $\int_0^1 \frac{1}{2} y^{-1/2} dy = 1$.)

(b) Mean and variance.

$$\mathbb{E}[Y] = \int_0^1 y\, f_Y(y)\, dy = \frac{1}{2} \int_0^1 y^{1/2} dy = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

$$\mathbb{E}[Y^2] = \int_0^1 y^2\, f_Y(y)\, dy = \frac{1}{2} \int_0^1 y^{3/2} dy = \frac{1}{2} \cdot \frac{2}{5} = \frac{1}{5}.$$

Hence

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \big(\mathbb{E}[Y]\big)^2 = \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}.$$

□

Problem 4. Suppose X has geometric pmf $P(X = x) = \frac{1}{3}\left(\frac{2}{3}\right)^x$ for $x = 0, 1, 2, \dots$. Define $Y = \dfrac{X}{X+1}$. Determine the pmf of Y.

Solution. TBW. The mapping $x \mapsto y = \dfrac{x}{x+1}$ is strictly increasing on $\{0, 1, 2, \ldots\}$ and takes values

$$\mathcal{S}_Y = \left\{0,\ \frac{1}{2},\ \frac{2}{3},\ \frac{3}{4},\ \ldots,\ \frac{x}{x+1},\ \ldots\right\} \subset [0, 1).$$

It is one-to-one, with inverse on $\mathcal{S}_Y$ given by $x = \dfrac{y}{1-y}$. Therefore, for $y \in \mathcal{S}_Y$,

$$P(Y = y) = P\left(X = \frac{y}{1-y}\right) = \frac{1}{3}\left(\frac{2}{3}\right)^{\frac{y}{1-y}}, \qquad \text{and} \qquad P(Y = y) = 0 \text{ for } y \notin \mathcal{S}_Y.$$

Equivalently, writing $y_x = \dfrac{x}{x+1}$,

$$P(Y = y_x) = P(X = x) = \frac{1}{3}\left(\frac{2}{3}\right)^x, \quad x = 0, 1, 2, \ldots$$

and the probabilities sum to 1 since $\sum_{x \geq 0} P(Y = y_x) = \sum_{x \geq 0} P(X = x) = 1$. □

Problem 5. (a) Let $X \sim \mathcal{N}(m, \sigma^2)$ with $\sigma = \sqrt{\sigma^2} > 0$. Show that $Z = \dfrac{X - m}{\sigma} \sim \mathcal{N}(0, 1)$.

(b) Let $Z \sim \mathcal{N}(0, 1)$ and $m \in \mathbb{R}$, $\sigma \neq 0$. Show that $X = m + \sigma Z \sim \mathcal{N}(m, \sigma^2)$.

(c) Let $X \sim \mathcal{N}(m, \sigma^2)$ and $a \in \mathbb{R}$, $b \neq 0$. Prove that $Y = a + bX$ is normal and find its parameters.

Solution. TBW.

(a) Standardization. For $z \in \mathbb{R}$,

$$F_Z(z) = \mathbb{P}\left(\frac{X - m}{\sigma} \leq z\right) = \mathbb{P}(X \leq m + \sigma z) = F_X(m + \sigma z).$$

Differentiating (chain rule) gives the pdf of Z:

$$f_Z(z) = \sigma\, f_X(m + \sigma z) = \sigma \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(m + \sigma z - m)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

the $\mathcal{N}(0, 1)$ density. (Equivalently, $\mathbb{E}[Z] = (\mathbb{E}[X] - m)/\sigma = 0$ and $\mathrm{Var}(Z) = \mathrm{Var}(X)/\sigma^2 = 1$.)

(b) Affine build-up from a standard normal. Let $X = m + \sigma Z$. For any $x \in \mathbb{R}$,

$$F_X(x) = \mathbb{P}(m + \sigma Z \leq x) = \mathbb{P}\left(Z \leq \frac{x - m}{\sigma}\right) = \Phi\left(\frac{x - m}{\sigma}\right),$$

so X has cdf of $\mathcal{N}(m, \sigma^2)$ and hence $X \sim \mathcal{N}(m, \sigma^2)$. Differentiating yields the familiar pdf $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2})$.

(c) Closure under affine maps. Define $Y = a + bX$. Using the change-of-variables formula (with monotone linear $y \mapsto x = (y - a)/b$), for $y \in \mathbb{R}$,

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y - a}{b}\right) = \frac{1}{|b|} \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{y-a}{b} - m\right)^2}{2\sigma^2}\right) = \frac{1}{(|b|\sigma)\sqrt{2\pi}} \exp\left(-\frac{(y - (a + bm))^2}{2(b\sigma)^2}\right).$$

Hence $Y \sim \mathcal{N}(a+bm,\ (b\sigma)^2)$ (note the variance $b^2\sigma^2$; the standard deviation is $|b|\sigma$). In particular, $\mathbb{E}[Y] = a + bm$ and $\mathrm{Var}(Y) = b^2\sigma^2$. □

Problem 6. Let $B_n \sim \mathrm{Bin}(n, p_n)$ with $p_n = \lambda/n$ for some fixed $\lambda > 0$. Show that for each fixed $k = 0, 1, 2, \ldots,$

$$\lim_{n\to\infty} \mathbb{P}(B_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. TBW. For $k \in \{0, 1, \ldots, n\}$,

$$\mathbb{P}(B_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k}.$$

With $p_n = \lambda/n$,

$$\begin{aligned}\mathbb{P}(B_n = k) &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\longrightarrow 1}\ \frac{\lambda^k}{k!}\ \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\longrightarrow e^{-\lambda}}\ \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\longrightarrow 1}.\end{aligned}$$

Taking limits as $n \to \infty$ and using the standard limits $(1-\lambda/n)^n \to e^{-\lambda}$ and $(1-\lambda/n)^{-k} \to 1$, we obtain

$$\lim_{n\to\infty} \mathbb{P}(B_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

□

Remark (More general hypothesis)**.** The same conclusion holds under the weaker assumption $n\,p_n \to \lambda$ and $p_n \to 0$:

$$\mathbb{P}(B_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \left[\prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right)\right]\frac{(np_n)^k}{k!}\,(1-p_n)^n\,(1-p_n)^{-k} \longrightarrow e^{-\lambda}\frac{\lambda^k}{k!}.$$

Problem 7. In a production line, a device is defective with probability 0.2 (it fails immediately), while a non–defective device has failure–free time T that is exponential with rate $\lambda = 0.05$ hour^{-1}.

(a) Find the distribution function of the device's failure–free operation time.

(b) Find the mean and variance of the device's uptime.

Solution. TBW. Let τ denote the failure–free time of a randomly chosen device. Then $\mathbb{P}(\tau = 0) = 0.2$ and, conditional on $\{\tau > 0\}$, $\tau \sim \mathrm{Exp}(\lambda)$. Hence the law of τ is a *mixed* distribution: an atom at 0 and a continuous exponential component of weight 0.8.

(a) Distribution function. For $x < 0$, $F_\tau(x) = 0$. For $x \geq 0$, by total probability,

$$F_\tau(x) = \mathbb{P}(\tau \leq x) = \mathbb{P}(\tau = 0) + \mathbb{P}(\tau \leq x \mid \tau > 0)\,\mathbb{P}(\tau > 0) = 0.2 + 0.8\big(1 - e^{-\lambda x}\big).$$

Equivalently, the density has an atom $\mathbb{P}(\tau = 0) = 0.2$ and for $x > 0$,

$$f_\tau(x) = 0.8\,\lambda e^{-\lambda x}.$$

(b) Mean and variance. Using the mixture structure (or integrating with f_τ),

$$\mathbb{E}[\tau] = 0 \cdot 0.2 + \int_0^\infty x\,(0.8\,\lambda e^{-\lambda x})\,dx = 0.8 \cdot \frac{1}{\lambda} = \frac{0.8}{0.05} = 16 \text{ hours}.$$

Since for an exponential variable $X \sim \mathrm{Exp}(\lambda)$ we have $\mathbb{E}[X^2] = 2/\lambda^2$, here

$$\mathbb{E}[\tau^2] = 0.8 \cdot \frac{2}{\lambda^2} = 0.8 \cdot \frac{2}{0.05^2} = 640, \qquad \mathrm{Var}(\tau) = \mathbb{E}[\tau^2] - \big(\mathbb{E}[\tau]\big)^2 = 640 - 16^2 = 384 \text{ hours}^2.$$

□

Problem 8. Let X have the standard Cauchy density $f_X(x) = \dfrac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.

(a) Show that $Y = \dfrac{1}{X}$ is also standard Cauchy.

(b) What is the expected value of X?

Solution. TBW.

(a) $Y = 1/X$ is Cauchy. Because $P(X = 0) = 0$, the map $g(x) = 1/x$ is invertible a.s. with inverse $g^{-1}(y) = 1/y$ and $\left|\dfrac{d}{dy} g^{-1}(y)\right| = \dfrac{1}{y^2}$. By change of variables,

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\tfrac{d}{dy} g^{-1}(y)\right| = \frac{1}{\pi(1 + (1/y)^2)} \cdot \frac{1}{y^2} = \frac{1}{\pi(1+y^2)}, \qquad y \in \mathbb{R},$$

so $Y \sim \mathrm{Cauchy}(0,1)$.

(Equivalent CDF check). With $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$, one gets $F_Y(y) = \frac{1}{2} + \frac{1}{\pi}\arctan y$ for all y, the Cauchy CDF.

(b) Expectation.

$$\int_{\mathbb{R}} |x|\, f_X(x)\,dx = \frac{2}{\pi}\int_1^\infty \frac{x}{1+x^2}\,dx = \frac{1}{\pi}\Big[\ln(1+x^2)\Big]_1^\infty = \infty.$$

Hence $X \notin L^1$ and the (Lebesgue) expectation $\mathbb{E}[X]$ *does not exist* (the symmetric improper integral equals 0 as a principal value, but this is not a finite expectation). □

Week 4 – Class 7

Multiple Random Variables

So far, we only worked with univariate models. Now: multivariate.

Definition. An n-dimensional random vector is a function from a sample space S into $\mathbb{R}^n$, n-dimensional Euclidian space.

Discrete Bivariate

Definition. Let (X, Y) be a discrete bivariate random vector. Then the function

$$f(x,y) = \mathbb{P}(X = x, Y = y), \quad (x,y) \in \mathbb{R}^2,$$

is called the joint probability mass function, or joint pmf, of (X, Y).

The joint pmf can be used to compute the probability of any event defined in terms of (X, Y). Let A be any subset of $\mathbb{R}^2$. Then

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

Remark. Because (X, Y) is discrete, $f(x, y)$ is nonzero at most at a countable number of points (x, y). Hence, this is a countable sum.

Expectations

Expectations work the same as with univariate random variables. Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) for (X, Y). Then $g(X, Y)$ is also a random variable, and

$$\mathbb{E}[g(X,Y)] \;=\; \sum_{(x,y)\in\mathbb{R}^2} g(x,y)f(x,y).$$

Properties of the Joint pmf

The joint pmf must satisfy certain properties:

- For any (x, y), $f(x, y) \geq 0$ because it is a probability.
- Since (X, Y) certainly takes values in $\mathbb{R}^2$,

$$\sum_{(x,y)\in\mathbb{R}^2} f(x,y) \;=\; \mathbb{P}\big((X,Y) \in \mathbb{R}^2\big) \;=\; 1.$$

Remark. We do not need to work with the underlying fundamental sample space S, which can be intractable. Instead, we work with the pmf.

Discrete Bivariate: Marginals

The variable X of the random vector (X, Y) is itself a random variable, with a pmf $f_X(x) = \mathbb{P}(X = x)$ (same for Y); we call this the *marginal pmf*.

Theorem 33 (4.1.6)**.** *Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = \mathbb{P}(X = x)$ and $f_Y(y) = \mathbb{P}(Y = y)$, are*

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \qquad \text{and} \qquad f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y).$$

Proof.

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}\big(X = x,\ -\infty < Y < \infty\big).$$

Let

$$A_x = \{(x', y') : \ x' = x,\ -\infty < y' < \infty\}.$$

Then

$$\mathbb{P}\big((X, Y) \in A_x\big) = \sum_{(x',y') \in A_x} f_{X,Y}(x', y') = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y).$$

An identical argument gives the expression for $f_Y(y)$. □

Continuous Bivariate

Definition (Joint pdf)**.** A function $f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ is called the joint probability density function (joint pdf) of a continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$,

$$\mathbb{P}\big((X, Y) \in A\big) = \iint_A f(x, y)\, dx\, dy.$$

Remark. Same as the univariate case, but now with double integrals.

If $g(x, y)$ is a real-valued function, then the expected value of $g(X, Y)$ is

$$\mathbb{E}\big[g(X, Y)\big] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)\, f(x, y)\, dx\, dy.$$

Continuous Bivariate: Marginals

The marginal probability densities of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \qquad -\infty < x < \infty,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \qquad -\infty < y < \infty.$$

As in the discrete case, any function $f(x, y)$ with $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ that integrates to 1, i.e.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1,$$

is the joint pdf of some continuous bivariate random vector (X,Y).

Continuous Bivariate: Joint CDF

The joint probability distribution of (X,Y) can be completely described with the joint cdf, defined by

$$F(x,y) = \mathbb{P}(X \le x, Y \le y), \qquad (x,y) \in \mathbb{R}^2.$$

We have

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\,dt\,ds,$$

and moreover

$$\frac{\partial^2}{\partial x\,\partial y}F(x,y) = f(x,y).$$

Conditional Distributions

Definition (4.2.1 Discrete case)**.** Let (X,Y) be a discrete bivariate random vector with joint pmf $f(x,y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $\mathbb{P}(X=x) = f_X(x) > 0$, the *conditional pmf of* Y *given* $X = x$ is the function of y denoted $f(y|x)$, defined by

$$f(y|x) = \mathbb{P}(Y = y \mid X = x) = \frac{f(x,y)}{f_X(x)}.$$

(and similarly for x given y).

Definition (4.2.3 Continuous case)**.** Let (X,Y) be a continuous bivariate random vector with joint pdf $f(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the *conditional pdf of* Y *given* $X = x$ is the function of y denoted $f(y|x)$, defined by

$$f(y|x) = \frac{f(x,y)}{f_X(x)}.$$

(and similarly for x given y).

Conditional Expectations

When we have conditional pmfs or pdfs, we can compute conditional expected values:

$$\mathbb{E}[g(Y) \mid X = x] = \sum_y g(y)\,f(y|x) \qquad \text{(discrete case)},$$

$$\mathbb{E}[g(Y) \mid X = x] = \int_{-\infty}^{\infty} g(y)\,f(y|x)\,dy \qquad \text{(continuous case)}.$$

Definition (4.2.5 Independence)**.** Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent* if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = f_X(x) f_Y(y).$$

If X and Y are independent, then the conditional distribution of Y given $X = x$ is

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

Lemma (4.2.7)**.** *Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for all $x, y \in \mathbb{R}$,*

$$f(x, y) = g(x) h(y).$$

Proof. Revise Carefully

($\Rightarrow$) Suppose X and Y are independent. By definition of independence,

$$f(x, y) = f_X(x) f_Y(y).$$

Hence $g(x) = f_X(x)$ and $h(y) = f_Y(y)$ satisfy the factorization.

($\Leftarrow$) Suppose instead that the joint distribution factorizes as

$$f(x, y) = g(x) h(y).$$

At this point $g(x)$ and $h(y)$ are just nonnegative functions, not necessarily probability distributions.

Compute the marginals:

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\, dy = \int_{\mathbb{R}} g(x) h(y)\, dy = g(x) \int_{\mathbb{R}} h(y)\, dy.$$

Define

$$c = \int_{\mathbb{R}} h(y)\, dy.$$

Then

$$f_X(x) = g(x) c.$$

Similarly,

$$f_Y(y) = \int_{\mathbb{R}} f(x, y)\, dx = \int_{\mathbb{R}} g(x) h(y)\, dx = h(y) \int_{\mathbb{R}} g(x)\, dx.$$

Define

$$d = \int_{\mathbb{R}} g(x)\, dx.$$

Then

$$f_Y(y) = h(y) d.$$

Now, because $f(x,y)$ is a valid pdf/pmf, we must have

$$1 = \iint_{\mathbb{R}^2} f(x,y)\,dxdy = \Big(\int_{\mathbb{R}} g(x)\,dx\Big)\Big(\int_{\mathbb{R}} h(y)\,dy\Big) = cd.$$

Therefore,

$$f_X(x)f_Y(y) = (g(x)c)(h(y)d) = g(x)h(y)\,(cd).$$

But since $cd = 1$, we get

$$f_X(x)f_Y(y) = g(x)h(y) = f(x,y).$$

Thus the joint distribution factorizes into the product of the marginals, which means X and Y are independent. □

Theorem 34 (4.2.10)**.** *Let X and Y be independent random variables.*

(a) For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$,

$$\mathbb{P}(X \in A,\, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B).$$

(b) If g is a function of x only and h a function of y only (with $\mathbb{E}|g(X)| < \infty$, $\mathbb{E}|h(Y)| < \infty$), then

$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)].$$

Proof. Revise Carefully

We write the proof for the continuous case (with joint pdf f, marginals f_X, f_Y). The discrete case is identical replacing integrals by sums.

Useful identity (for any Borel $A \subset \mathbb{R}$):

$$\mathbb{E}\big[\mathbf{1}_A(X)\big] = \int_{\mathbb{R}} \mathbf{1}_A(x) f_X(x)\,dx = \int_A f_X(x)\,dx = \mathbb{P}(X \in A). \tag{1}$$

(a) Using indicator functions and independence ($f(x,y) = f_X(x)f_Y(y)$),

$$\begin{aligned}
\mathbb{P}(X \in A,\, Y \in B) &= \mathbb{E}\big[\mathbf{1}_A(X)\mathbf{1}_B(Y)\big] \\
&= \iint_{\mathbb{R}^2} \mathbf{1}_A(x)\mathbf{1}_B(y) f(x,y)\,dx\,dy \\
&= \Big(\int_{\mathbb{R}} \mathbf{1}_A(x) f_X(x)\,dx\Big)\Big(\int_{\mathbb{R}} \mathbf{1}_B(y) f_Y(y)\,dy\Big) \\
&= \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B) \qquad \text{by (1) (and its } Y\text{-analogue).}
\end{aligned}$$

(b) By Fubini and independence,

$$\begin{aligned}\mathbb{E}\big[g(X)h(Y)\big] &= \iint_{\mathbb{R}^2} g(x)h(y)f(x,y)\,dx\,dy \\ &= \Big(\int_{\mathbb{R}} g(x)f_X(x)\,dx\Big)\Big(\int_{\mathbb{R}} h(y)f_Y(y)\,dy\Big) \\ &= \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)].\end{aligned}$$

Equivalently (and this is the missing step in the slide): by the Law of Iterated Expectations and independence,

$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}\big[g(X)\,\mathbb{E}[h(Y) \mid X]\big] = \mathbb{E}\big[g(X)\,\mathbb{E}[h(Y)]\big] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)],$$

since $Y \perp X$ implies $\mathbb{E}[h(Y) \mid X] = \mathbb{E}[h(Y)]$ a.s. □

Example (4.1.12 and 4.2.4)**.** Let $f(x,y) = e^{-y}$ for $0 < x < y < \infty$. We want to compute $\mathbb{P}(X+Y \geq 1)$.

Instead of integrating directly over the region $\{(x,y) : x+y \geq 1,\, 0 < x < y\}$, it is easier to use the complement:

$$\mathbb{P}(X+Y \geq 1) = 1 - \mathbb{P}(X+Y < 1).$$

The event $\{X+Y < 1\}$ corresponds to $0 < x < 1/2$ and $x < y < 1-x$. Thus

$$\mathbb{P}(X+Y < 1) = \int_0^{1/2} \int_x^{1-x} e^{-y}\,dy\,dx.$$

Evaluating the inner integral:

$$\int_x^{1-x} e^{-y}\,dy = e^{-x} - e^{-(1-x)}.$$

So

$$\mathbb{P}(X+Y < 1) = \int_0^{1/2} \Big(e^{-x} - e^{-(1-x)}\Big)\,dx.$$

Therefore,

$$\mathbb{P}(X+Y \geq 1) = 1 - \int_0^{1/2} \Big(e^{-x} - e^{-(1-x)}\Big)\,dx = 2e^{-1/2} - e^{-1}.$$

Now, let us compute the conditional distribution of Y given $X = x$. The marginal of X is

$$f_X(x) = \int_x^{\infty} e^{-y}\,dy = e^{-x}, \qquad x > 0.$$

Hence $X \sim \text{Exponential}(1)$.

The conditional pdf is

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \begin{cases} \dfrac{e^{-y}}{e^{-x}} = e^{-(y-x)}, & y > x, \\ 0, & y \le x. \end{cases}$$

Thus $Y|X = x \sim x + \text{Exponential}(1)$.

Finally, let us compute conditional expectation and variance.

$$\mathbb{E}[Y|X = x] = \int_x^\infty y\, e^{-(y-x)}\, dy.$$

Substitute $z = y - x$, $dz = dy$, $y = z + x$, lower limit $z = 0$:

$$\mathbb{E}[Y|X = x] = \int_0^\infty (z + x)e^{-z}\, dz = \int_0^\infty ze^{-z}\, dz + x \int_0^\infty e^{-z}\, dz = 1 + x.$$

For the variance, use

$$\mathbf{Var}(Y|X = x) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2.$$

We compute

$$\mathbb{E}[Y^2|X = x] = \int_x^\infty y^2 e^{-(y-x)}\, dy = \int_0^\infty (z + x)^2 e^{-z}\, dz.$$

Expanding:

$$\mathbb{E}[Y^2|X = x] = \int_0^\infty (z^2 + 2xz + x^2)e^{-z}\, dz = 2 + 2x + x^2.$$

Thus

$$\mathbf{Var}(Y|X = x) = (x^2 + 2x + 2) - (1 + x)^2 = 1.$$

Conclusion: The conditional distribution of $Y|X = x$ is exponential with mean $1 + x$ and variance 1, i.e. a shifted exponential.

Bivariate Transformations

Definition (Set-up)**.** Let (X, Y) be a bivariate random vector with known joint distribution. Define a transformed pair (U, V) by

$$U = g_1(X, Y), \qquad V = g_2(X, Y),$$

for given functions g_1, g_2. For any $B \subset \mathbb{R}^2$, write

$$A \;=\; \{(x, y) \in \mathbb{R}^2 :\; (g_1(x, y), g_2(x, y)) \in B\}.$$

Then $(U, V) \in B \iff (X, Y) \in A$, hence

$$\mathbb{P}\big((U, V) \in B\big) = \mathbb{P}\big((X, Y) \in A\big).$$

Remark (What this means)**.** All distributional information about (U, V) is inherited from (X, Y) via *preimages of sets*: probabilities for (U, V) over B equal probabilities for (X, Y) over the corresponding preimage A. This principle underlies both the discrete and continuous formulas below.

Discrete random vectors

Let $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ be the (countable) support of (X, Y) and

$$\mathcal{B} = \{(u, v) : \exists (x, y) \in \mathcal{A} \text{ s.t. } u = g_1(x, y),\ v = g_2(x, y)\}$$

the attainable set of (U, V). For $(u, v) \in \mathcal{B}$ define the preimage slice

$$A_{uv} = \{(x, y) \in \mathcal{A} : g_1(x, y) = u,\ g_2(x, y) = v\}.$$

Proposition (Joint pmf under a transformation)**.** *The joint pmf of (U, V) is*

$$f_{U,V}(u, v) = \mathbb{P}(U = u, V = v) = \sum_{(x,y) \in A_{uv}} f_{X,Y}(x, y), \qquad (u, v) \in \mathcal{B},$$

and $f_{U,V}(u, v) = 0$ for $(u, v) \notin \mathcal{B}$.

Proof. The sets $\{(X, Y) = (x, y)\}$ with $(x, y) \in A_{uv}$ are disjoint and their union is $\{U = u, V = v\}$. Add probabilities. □

Continuous random vectors

Let $A = \{(x, y) : f_{X,Y}(x, y) > 0\}$ and $B = \{(u, v) : \exists (x, y) \in A \text{ with } (u, v) = (g_1(x, y), g_2(x, y))\}$. Assume:

(i) $g = (g_1, g_2) : A \to B$ is one-to-one and onto B;

(ii) g is continuously differentiable on A with nonzero Jacobian determinant everywhere;

(iii) its inverse $h = (h_1, h_2) : B \to A$ is continuously differentiable.

Proposition (Change of variables for bivariate densities)**.** *On B the joint pdf of (U, V) is*

$$f_{U,V}(u, v) = f_{X,Y}\big(h_1(u, v), h_2(u, v)\big)\, \big|J(u, v)\big|, \qquad (u, v) \in B,$$

and $f_{U,V}(u,v) = 0$ *for* $(u,v) \notin B$*, where*

$$J(u,v) = \det\begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} \quad \textit{with} \quad x = h_1(u,v),\ y = h_2(u,v).$$

Equivalently,

$$f_{U,V}(u,v) = \left.\frac{f_{X,Y}(x,y)}{\left|\det\left(\frac{\partial(u,v)}{\partial(x,y)}\right)\right|}\right|_{(x,y)=h(u,v)}.$$

Intuition/derivation. Fix $(u_0,v_0) \in B$ and let $(x_0,y_0) = h(u_0,v_0)$. For a small rectangle $R_{uv} = [u_0, u_0+\Delta u] \times [v_0, v_0+\Delta v]$, its preimage under h is a small parallelogram R_{xy} around (x_0,y_0) whose area is approximately $\left|J(u_0,v_0)\right|\Delta u\,\Delta v$ (by the linear approximation of h). Hence

$$\mathbb{P}\big((U,V) \in R_{uv}\big) = \mathbb{P}\big((X,Y) \in R_{xy}\big) \approx f_{X,Y}(x_0,y_0)\,\left|J(u_0,v_0)\right|\,\Delta u\,\Delta v.$$

Divide by $\Delta u\,\Delta v$ and let $\Delta u, \Delta v \to 0$ to obtain the stated density. □

Remark (Support and zero density outside B)**.** By construction $f_{U,V}$ is supported on $B = g(A)$; if $(u,v) \notin B$, then no (x,y) maps to (u,v) and $f_{U,V}(u,v) = 0$.

Recall (Jacobian entries)**.** With $x = h_1(u,v)$ and $y = h_2(u,v)$,

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u,v)}{\partial u}, \quad \frac{\partial x}{\partial v} = \frac{\partial h_1(u,v)}{\partial v}, \quad \frac{\partial y}{\partial u} = \frac{\partial h_2(u,v)}{\partial u}, \quad \frac{\partial y}{\partial v} = \frac{\partial h_2(u,v)}{\partial v}.$$

Take the absolute value of the determinant.

Remark (Why inverse-Jacobian?)**.** The area element transforms as $dx\,dy = \left|\det(\partial(x,y)/\partial(u,v))\right|\,du\,dv$. Therefore we multiply by the *inverse* Jacobian (from (u,v) back to (x,y)). Using the forward Jacobian $\det(\partial(u,v)/\partial(x,y))$ is equivalent after inversion.

Remark. Checklist to apply the theorem:

1. Identify the support A of (X,Y) and define $B = g(A)$.
2. Verify one-to-one on A (otherwise, split into one-to-one branches).
3. Find the inverse map $h(u,v) = (x,y)$ explicitly.
4. Compute $J(u,v) = \det\big(\partial(x,y)/\partial(u,v)\big)$.
5. Write $f_{U,V}(u,v) = f_{X,Y}(h(u,v))\,|J(u,v)|$ on B and 0 otherwise.
6. (Sanity check) Verify $\iint_B f_{U,V}(u,v)\,du\,dv = 1$.

Example: Transformation to Polar Coordinates

Example (Uniform on the unit disk ⇒ polar coordinates). **Setup.** Let (X, Y) be uniform on the unit disk:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi}, & x^2 + y^2 \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Transformation. Define

$$R = \sqrt{X^2 + Y^2}, \qquad \Theta = \arctan\left(\frac{Y}{X}\right),$$

and note the inverse map

$$x = h_1(r,\theta) = r\cos\theta, \qquad y = h_2(r,\theta) = r\sin\theta.$$

Jacobian. Using $x = r\cos\theta$, $y = r\sin\theta$,

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \qquad J = \det\left(\frac{\partial(x,y)}{\partial(r,\theta)}\right) = r.$$

Support. Since $x^2 + y^2 \le 1$ iff $0 \le r \le 1$ and every point on the disk has a unique polar angle modulo 2π, we take

$$0 \le r \le 1, \qquad 0 \le \theta < 2\pi.$$

Joint pdf. By the change-of-variables formula,

$$f_{R,\Theta}(r,\theta) = f_{X,Y}\big(h_1(r,\theta), h_2(r,\theta)\big)\,\big|J\big| = \begin{cases} \dfrac{1}{\pi}r, & 0 \le r \le 1,\ 0 \le \theta < 2\pi, \\ 0, & \text{otherwise.} \end{cases}$$

Example. Let $U = X + Y$ and $V = X$ for a generic pair (X, Y) with joint pdf $f_{X,Y}$. The inverse map is $x = h_1(u,v) = v$, $y = h_2(u,v) = u - v$, with

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \quad \Rightarrow \quad \big|\det(\partial(x,y)/\partial(u,v))\big| = 1.$$

Hence

$$f_{U,V}(u,v) = f_{X,Y}\big(v,\, u - v\big) \times 1 = f_{X,Y}\big(v,\, u - v\big),$$

with support obtained by mapping the support of (X, Y) through $(u, v) = (x + y, x)$.

Bivariate Transformations and Mixtures

Theorem 35 (4.3.5 Independence under separate transformations)**.** *Let X and Y be independent random variables. Let $g_1(x)$ depend only on x and $g_2(y)$ only on y. Then the transformed variables*

$$U = g_1(X), \qquad V = g_2(Y)$$

are independent.

Continuous case. Let $M, N \subset \mathbb{R}$ and define

$$A_M = \{x : g_1(x) \in M\}, \qquad B_N = \{y : g_2(y) \in N\}.$$

Then

$$F_{U,V}(M, N) = \mathbb{P}(U \in M, V \in N) = \mathbb{P}\big(X \in A_M, Y \in B_N\big).$$

Since $X \perp Y$, this factorizes as

$$\mathbb{P}(X \in A_M)\, \mathbb{P}(Y \in B_N) = F_U(M)\, F_V(N).$$

Differentiating gives $f_{U,V}(u, v) = f_U(u) f_V(v)$, hence $U \perp V$. □

Remark. This result says that applying independent (measurable) transformations to independent variables preserves independence.

Non one-to-one transformations

Sometimes one is interested in a single transformed variable, say $U = g_1(X, Y)$ (e.g. XY or $X + Y$). To derive its distribution, we often introduce a convenient second variable $V = g_2(X, Y)$ such that the map $(X, Y) \mapsto (U, V)$ is one-to-one. We then compute the joint distribution of (U, V) and obtain the marginal of U.

If the transformation is not globally one-to-one, partition the support

$$A = \{(x, y) : f_{X,Y}(x, y) > 0\}$$

into subsets $\{A_i\}$ where (g_1, g_2) is one-to-one. For each i there is an inverse $(h_{1i}(u, v), h_{2i}(u, v))$ and a Jacobian J_i. Then

$$f_{U,V}(u, v) = \sum_{i=1}^{k} f_{X,Y}\big(h_{1i}(u, v), h_{2i}(u, v)\big)\, |J_i|.$$

Remark. This is the multivariate analogue of handling non-monotone transformations in the univariate case.

Hierarchical Models and Mixtures

Definition (Mixture distribution)**.** A random variable X is said to have a *mixture distribution* if its distribution depends on another random quantity Y which itself has a distribution. Equivalently, the parameter of X is random.

Example (Hierarchical model)**.** Let

$$X \mid Y \sim \text{Binomial}(Y, p), \qquad Y \sim \text{Poisson}(\lambda).$$

Here the distribution of X depends on the random parameter Y. The marginal distribution of X is therefore a mixture: averaging the binomial distribution over the Poisson distribution of Y.

Theorem 36 (4.4.3 Law of Iterated Expectations)**.** *If X and Y are two random variables (with the relevant expectations finite), then*

$$\mathbb{E}[X] \;=\; \mathbb{E}\big(\mathbb{E}[X \mid Y]\big).$$

Proof. Start with the definition:

$$\mathbb{E}[X] = \iint x\, f_{X,Y}(x,y)\, dx\, dy.$$

Factor the joint density as $f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y)$:

$$\mathbb{E}[X] = \int \left(\int x\, f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy.$$

The inner integral is by definition $\mathbb{E}[X \mid Y = y]$. Therefore

$$\mathbb{E}[X] = \int \mathbb{E}[X \mid Y = y]\, f_Y(y)\, dy = \mathbb{E}\big(\mathbb{E}[X \mid Y]\big).$$

□

Example (Binomial–Poisson mixture)**.** Let

$$X \mid Y \sim \text{Binomial}(Y, p), \qquad Y \sim \text{Poisson}(\lambda).$$

By the law of iterated expectations,

$$\mathbb{E}[X] = \mathbb{E}\big(\mathbb{E}[X \mid Y]\big) = \mathbb{E}[pY] = p\, \mathbb{E}[Y] = p\lambda.$$

Week 4 – Class 8

Theorem 37 (4.4.7 Law of Total Variance)**.** *For any random variables X, Y (with finite variances),*

$$\mathbf{Var}(X) = \mathbb{E}\big[\mathbf{Var}(X \mid Y)\big] + \mathbf{Var}\big(\mathbb{E}[X \mid Y]\big).$$

Proof. Improve explanation (slide 28 with proofs) Recall

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

Add and subtract $\mathbb{E}[X \mid Y]$ inside the square:

$$X - \mathbb{E}[X] = \underbrace{X - \mathbb{E}[X \mid Y]}_{\alpha} + \underbrace{\mathbb{E}[X \mid Y] - \mathbb{E}[X]}_{\beta}.$$

Then

$$(X - \mathbb{E}[X])^2 = \alpha^2 + 2\alpha\beta + \beta^2.$$

Take expectations:

$$\mathbf{Var}(X) = \mathbb{E}[\alpha^2] + 2\mathbb{E}[\alpha\beta] + \mathbb{E}[\beta^2].$$

Now:

- $\mathbb{E}[\alpha^2] = \mathbb{E}[\mathbf{Var}(X \mid Y)]$ by definition.
- $\mathbb{E}[\alpha\beta] = \mathbb{E}\big(\mathbb{E}[\alpha\beta \mid Y]\big) = \mathbb{E}\big(\beta\, \mathbb{E}[\alpha \mid Y]\big) = 0$ since $\mathbb{E}[\alpha \mid Y] = \mathbb{E}[X - \mathbb{E}[X \mid Y] \mid Y] = 0$.
- $\mathbb{E}[\beta^2] = \mathbf{Var}(\mathbb{E}[X \mid Y])$.

Thus

$$\mathbf{Var}(X) = \mathbb{E}[\mathbf{Var}(X \mid Y)] + \mathbf{Var}(\mathbb{E}[X \mid Y]).$$

□

Covariance and Correlation

Definition (Covariance)**.** The *covariance* between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big],$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Intuitively, covariance measures whether large (or small) values of X tend to be associated with large (or small) values of Y.

- If $\mathrm{Cov}(X, Y) > 0$, then X and Y tend to move together.
- If $\mathrm{Cov}(X, Y) < 0$, then X and Y move in opposite directions.
- If $\mathrm{Cov}(X, Y) = 0$, there is no *linear* association.

Definition (Correlation)**.** The *correlation coefficient* between X and Y is defined as

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Unlike covariance, correlation is scale–free: it is always bounded between -1 and 1.

- $\rho_{XY} = 1$: perfect positive linear relation.
- $\rho_{XY} = -1$: perfect negative linear relation.
- $\rho_{XY} = 0$: no linear relationship (but possibly nonlinear dependence).

Theorem 38. *For any random variables X and Y,*

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \mu_Y.$$

Theorem 39. *If X and Y are independent, then*

$$\operatorname{Cov}(X, Y) = 0 \quad \textit{and} \quad \rho_{XY} = 0.$$

Remark. Independence implies zero correlation, but the converse is not true: zero covariance (or correlation) does not imply independence. Covariance only captures *linear* dependence. For instance, two variables can have a strong nonlinear relationship (e.g. $Y = X^2$) but still satisfy $\rho_{XY} = 0$.

Geometric Interpretation

We can view covariance as an inner product in the Hilbert space of square–integrable random variables. If we define the centered variables

$$\tilde{X} = X - \mu_X, \quad \tilde{Y} = Y - \mu_Y,$$

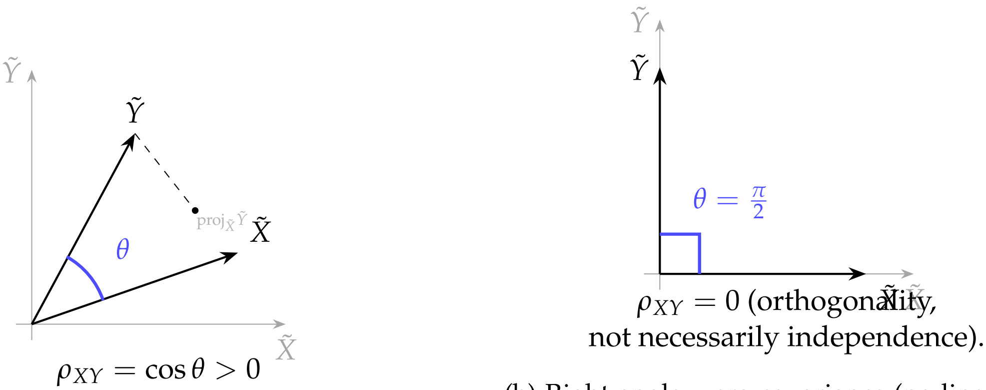
then

$$\rho_{XY} = \frac{\langle \tilde{X}, \tilde{Y} \rangle}{\|\tilde{X}\| \|\tilde{Y}\|} = \cos\theta.$$

Thus, correlation is literally the *cosine of the angle* between the two "vectors" $\tilde{X}$ and $\tilde{Y}$.

- A small (acute) angle means strong positive correlation.
- An obtuse angle means negative correlation.
- A right angle ($\theta = \pi/2$) means orthogonality: no linear dependence.

This explains the geometric view: two random variables are uncorrelated if their centered versions are orthogonal in this vector space. However, orthogonality does not preclude nonlinear dependence.



(a) Acute angle: positive covariance/correlation.

(b) Right angle: zero covariance (no linear relation).

Figure 10: Geometric view in L^2: $\rho_{XY} = \cos\theta$.

Theorem 40 (Variance of Linear Combinations). *If X and Y are any two random variables and a and b are any two constants, then*

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y) + 2ab\mathbf{Cov}(X, Y).$$

If X and Y are independent random variables, then

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y).$$

Proof. By definition,

$$\begin{aligned}
\mathbf{Var}(aX + bY) &= \mathbb{E}\left[\Big((aX + bY) - (a\mu_X + b\mu_Y)\Big)^2\right] \\
&= \mathbb{E}\left[(a(X - \mu_X) + b(Y - \mu_Y))^2\right] \\
&= \mathbb{E}\left[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)\right] \\
&= a^2\mathbb{E}[(X - \mu_X)^2] + b^2\mathbb{E}[(Y - \mu_Y)^2] + 2ab\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y) + 2ab\mathbf{Cov}(X, Y).
\end{aligned}$$

If X and Y are independent, then $\mathbf{Cov}(X, Y) = 0$ and the simplified formula follows. □

Definition (Bivariate Normal Distribution). Let $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$. The *bivariate normal distribution* with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ has joint density

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)}\left[\left(\tfrac{x - \mu_X}{\sigma_X}\right)^2 - 2\rho\left(\tfrac{x - \mu_X}{\sigma_X}\right)\left(\tfrac{y - \mu_Y}{\sigma_Y}\right) + \left(\tfrac{y - \mu_Y}{\sigma_Y}\right)^2\right]\right),$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Proposition (Properties of the Bivariate Normal)**.** *Some useful properties of the bivariate normal distribution are:*

(a) The marginal distribution of X is $N(\mu_X, \sigma_X^2)$.

(b) The marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$.

(c) The correlation between X and Y is $\rho_{XY} = \rho$.

(d) For any constants a and b, the linear combination $aX + bY$ is normally distributed:

$$aX + bY \sim N\big(a\mu_X + b\mu_Y,\, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y\big).$$

(e) All conditional distributions are also normal. For example,

$$Y \mid X = x \sim N\Big(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),\, \sigma_Y^2(1 - \rho^2)\Big).$$

Theorem 41 (Variance of Linear Combinations)**.** *If X and Y are any two random variables and a and b are any two constants, then*

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y) + 2ab\mathbf{Cov}(X, Y).$$

If X and Y are independent random variables, then

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y).$$

Proof. By definition,

$$\begin{aligned}
\mathbf{Var}(aX + bY) &= \mathbb{E}\left[\Big((aX + bY) - (a\mu_X + b\mu_Y)\Big)^2\right] \\
&= \mathbb{E}\left[(a(X - \mu_X) + b(Y - \mu_Y))^2\right] \\
&= \mathbb{E}\left[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)\right] \\
&= a^2\mathbb{E}[(X - \mu_X)^2] + b^2\mathbb{E}[(Y - \mu_Y)^2] + 2ab\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y) + 2ab\mathbf{Cov}(X, Y).
\end{aligned}$$

If X and Y are independent, then $\mathbf{Cov}(X, Y) = 0$ and the simplified formula follows. □

Definition (Bivariate Normal Distribution)**.** Let $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$. The *bivariate normal distribution* with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ has joint density

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)}\left[\left(\tfrac{x - \mu_X}{\sigma_X}\right)^2 - 2\rho\left(\tfrac{x - \mu_X}{\sigma_X}\right)\left(\tfrac{y - \mu_Y}{\sigma_Y}\right) + \left(\tfrac{y - \mu_Y}{\sigma_Y}\right)^2\right]\right),$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Proposition (Properties of the Bivariate Normal). *Some useful properties of the bivariate normal distribution are:*

(a) The marginal distribution of X is $N(\mu_X, \sigma_X^2)$.

(b) The marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$.

(c) The correlation between X and Y is $\rho_{XY} = \rho$.

(d) For any constants a and b, the linear combination $aX + bY$ is normally distributed:

$$aX + bY \sim N\big(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y\big).$$

(e) All conditional distributions are also normal. For example,

$$Y \mid X = x \sim N\Big(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),\ \sigma_Y^2(1 - \rho^2)\Big).$$

Bivariate Normal: marginal vs. joint normality

Revise this section.

Remark. *Marginal normality does not imply joint normality.*

Example. Let $X, Y \overset{\text{iid}}{\sim} N(0,1)$ and define

$$Z = \begin{cases} X, & XY > 0, \\ -X, & XY < 0, \end{cases} \qquad \text{(ignore } XY = 0\text{, which has probability 0).}$$

Then $X \sim N(0,1)$ and $Z \sim N(0,1)$ marginally, but (X, Z) is not jointly normal.

Why $Z \sim N(0,1)$. Write $S = \operatorname{sgn}(Y) \in \{-1, +1\}$. Since $Y \sim N(0,1)$ is symmetric, $\mathbb{P}(S = 1) = \mathbb{P}(S = -1) = \frac{1}{2}$, and S is independent of X. Note that

$$Z = X\,S.$$

Because S is an independent Rademacher variable, XS has the same distribution as X (a symmetric $N(0,1)$), hence $Z \sim N(0,1)$. □

Proposition. *The pair (X, Z) defined above is* not *jointly normal.*

Proof. Observe $Z = X$ on $\{Y > 0\}$ and $Z = -X$ on $\{Y < 0\}$. Therefore the support of (X, Z) is concentrated on the two lines

$$\{(x,z) : z = x\} \quad \text{and} \quad \{(x,z) : z = -x\},$$

a set of Lebesgue measure 0 in $\mathbb{R}^2$. A nondegenerate bivariate normal distribution has a strictly positive density on $\mathbb{R}^2$ (elliptical level sets). The only way a bivariate normal can be singular is when $Z = aX$ almost surely for a *fixed* constant a (degenerate Gaussian). Here $Z = X$ on $\{Y > 0\}$ and $Z = -X$ on $\{Y < 0\}$, so no fixed a satisfies $Z = aX$ a.s. Hence (X, Z) cannot be jointly normal. □

Multivariate Normal Distribution

Definition (Standard multivariate normal)**.** Let $Z = (Z_1, \ldots, Z_m)^\top$ have independent and identically distributed components $Z_i \sim N(0, 1)$. Then the joint pdf of Z is

$$f_Z(x_1, \ldots, x_m) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{x^\top x}{2}\right),$$

which is the pdf of the *multivariate standard normal* $N(0, I_m)$.

Proposition (Moments of the standard case)**.** *If* $Z \sim N(0, I_m)$*, then* $\mathbb{E}[Z] = 0$ *and* $\mathbf{Var}(Z) = I_m$.

Theorem 42 (Affine transformations of the standard normal)**.** *Let* $Z \sim N(0, I_m)$*, let* $\mu \in \mathbb{R}^q$ *and* B *be a* $q \times m$ *matrix. Define* $X = \mu + BZ$*. Then* X *has a multivariate normal distribution*

$$X \sim N(\mu, \Sigma), \qquad \Sigma = BB^\top.$$

Remark. Every $N(\mu, \Sigma)$ with Σ symmetric positive semidefinite can be written as in Theorem 42 for some B (e.g. a Cholesky factor of Σ).

Multivariate Distributions: basic facts

Definition (Joint pmf/pdf)**.** Let $X = (X_1, \ldots, X_n)$ be a random vector.

- If X is discrete, its joint pmf is $f(x) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and for any $A \subset \mathbb{R}^n$,
$$\mathbb{P}(X \in A) = \sum_{x \in A} f(x).$$
- If X is continuous, its joint pdf is a function $f(x_1, \ldots, x_n)$ such that for any Borel set $A \subset \mathbb{R}^n$,
$$\mathbb{P}(X \in A) = \int_A \cdots \int f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Proposition (Expectation of a function)**.** *Let* $g(x) = g(x_1, \ldots, x_n)$ *be a real–valued function. Then* $g(X)$ *is a random variable and its expectation is*

$$\mathbb{E}[g(X)] = \begin{cases} \displaystyle\int_{\mathbb{R}^n} g(x)\, f(x)\, dx, & \textit{continuous case,} \\ \displaystyle\sum_{x \in \mathbb{R}^n} g(x)\, f(x), & \textit{discrete case.} \end{cases}$$

Proposition (Marginalization)**.** *The marginal pdf/pmf of any subset of coordinates is obtained by integrating/summing over the remaining coordinates. In particular, for the first k coordinates* $(X_1, \ldots, X_k)$*:*

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \begin{cases} \displaystyle\int_{\mathbb{R}^{n-k}} f(x_1, \ldots, x_n)\, dx_{k+1} \cdots dx_n, & \textit{continuous}, \\ \displaystyle\sum_{(x_{k+1},\ldots,x_n)\in\mathbb{R}^{n-k}} f(x_1, \ldots, x_n), & \textit{discrete}. \end{cases}$$

Definition (Conditional distribution)**.** The conditional pdf or pmf of $(X_{k+1}, \ldots, X_n)$ given $(X_1 = x_1, \ldots, X_k = x_k)$ is defined by

$$f(x_{k+1}, \ldots, x_n \mid x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}.$$

That is, the *joint density* divided by the *marginal* of the conditioning variables.

Definition (Mutual independence)**.** Let $X_1, \ldots, X_n$ be random vectors with joint pdf or pmf $f(x_1, \ldots, x_n)$. Let $f_{X_i}(x_i)$ denote the marginal pdf or pmf of X_i. Then $X_1, \ldots, X_n$ are *mutually independent* if for every $(x_1, \ldots, x_n)$,

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

If the X_i's are one-dimensional, they are called mutually independent random variables.

Theorem 43 (Expectation factorization under independence)**.** *Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ depends only on x_i for $i = 1, \ldots, n$. Then*

$$\mathbb{E}[g_1(X_1) \cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdots \mathbb{E}[g_n(X_n)].$$

Remark. An intuitive way to see this result is to think of each X_i as an independent "observation". The product $\prod_i g_i(X_i)$ then separates cleanly into independent components, and expectations of products reduce to products of expectations. This mental picture can help when proving properties of independent random variables.

MGFs and Sums of Independent Random Variables

Definition (Moment generating function)**.** For a real r.v. X the mgf (when it exists on a neighborhood of 0) is

$$M_X(t) = \mathbb{E}(e^{tX}), \qquad t \in \mathbb{R}.$$

Theorem 44 (MGF of a sum)**.** *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs*

$M_{X_1}(t), \ldots, M_{X_n}(t)$. If $Z = X_1 + \cdots + X_n$, then

$$M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t).$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. with mgf $M_X(t)$, then

$$M_Z(t) = [M_X(t)]^n.$$

Proof. By definition and independence,

$$M_Z(t) = \mathbb{E}\left(e^{t(X_1+\cdots+X_n)}\right) = \mathbb{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} \mathbb{E}\left(e^{tX_i}\right) = \prod_{i=1}^{n} M_{X_i}(t).$$

The i.i.d. statement follows immediately. □

Example (Sum of exponentials gives Gamma)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, whose mgf is $M_X(t) = \dfrac{\lambda}{\lambda - t}$ for $t < \lambda$. By Theorem 44,

$$M_Z(t) = [M_X(t)]^n = \left(\frac{\lambda}{\lambda - t}\right)^n, \qquad t < \lambda.$$

This is the mgf of the Gamma distribution with shape n and rate λ; hence

$$Z = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

Corollary (Linear combinations of independent normals)**.** Let $X_1, \ldots, X_n$ be mutually independent with $X_i \sim N(\mu_i, \sigma_i^2)$. For fixed scalars $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, define

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i).$$

Then

$$Z \sim N\left(\sum_{i=1}^{n} (a_i \mu_i + b_i),\ \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Proof. The mgf of X_i is $M_{X_i}(t) = \exp(\mu_i t + \frac{1}{2}\sigma_i^2 t^2)$. Hence the mgf of $a_i X_i + b_i$ is $M_{a_i X_i + b_i}(t) = \exp\big((a_i \mu_i + b_i)t + \frac{1}{2} a_i^2 \sigma_i^2 t^2\big)$. By independence and Theorem 44,

$$M_Z(t) = \prod_{i=1}^{n} M_{a_i X_i + b_i}(t) = \exp\left(\Big(\sum_{i=1}^{n} (a_i \mu_i + b_i)\Big) t + \frac{1}{2}\Big(\sum_{i=1}^{n} a_i^2 \sigma_i^2\Big) t^2\right).$$

This is the mgf of $N\big(\sum_i (a_i \mu_i + b_i),\ \sum_i a_i^2 \sigma_i^2\big)$. □

Week 4 – Discussion

Problem 1. Let $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ be independent.

(a) Show that $X + Y \sim \text{Poisson}(\theta + \lambda)$.

(b) Show that the conditional distribution of X given $X + Y = n$ is $\text{Binomial}\big(n, \frac{\theta}{\theta+\lambda}\big)$. What is the distribution of $Y \mid X + Y = n$?

Solution.

(a) Done by the MGF argument above (or directly by convolution).

(b) For $n \in \mathbb{N}_0$ and $k = 0, \dots, n$,

$$\mathbb{P}(X = k \mid X + Y = n) = \frac{\mathbb{P}(X = k, Y = n-k)}{\mathbb{P}(X + Y = n)} = \frac{\dfrac{e^{-\theta}\theta^k}{k!} \cdot \dfrac{e^{-\lambda}\lambda^{n-k}}{(n-k)!}}{\dfrac{e^{-(\theta+\lambda)}(\theta+\lambda)^n}{n!}} = \binom{n}{k}\left(\frac{\theta}{\theta+\lambda}\right)^k \left(\frac{\lambda}{\theta+\lambda}\right)^{n-k}.$$

Thus

$$X \mid (X + Y = n) \sim \text{Binomial}\left(n, \frac{\theta}{\theta+\lambda}\right), \qquad Y \mid (X + Y = n) \sim \text{Binomial}\left(n, \frac{\lambda}{\theta+\lambda}\right).$$

Remark (Poisson splitting)**.** Equivalently, if $N \sim \text{Poisson}(\theta + \lambda)$ and, conditional on N, each of the N events is tagged "type X" with probability $p = \theta/(\theta + \lambda)$ independently, then $X \sim \text{Poisson}(\theta)$, $Y \sim \text{Poisson}(\lambda)$ and $X \perp Y$.

Problem 2.) Let X have the negative binomial distribution

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

where $0 < p < 1$ and $r \in \mathbb{N}$ ("number of failures before the r-th success").

(a) Compute the mgf $M_X(t)$.

(b) Define $Y = 2p\,X$. Show that, as $p \downarrow 0$,

$$\lim_{p \to 0} M_Y(t) = \Big(1 - 2t\Big)^{-r}, \qquad |t| < \tfrac{1}{2},$$

i.e. the mgf converges to that of a chi-squared random variable with $2r$ degrees of freedom.

Solution.

(a) MGF of X. Write $X = Y_1 + \cdots + Y_r$ where $Y_i \overset{iid}{\sim} \text{Geom}(p)$ with $\mathbb{P}(Y_i = k) = p(1-p)^k$, $k \geq 0$ (number of failures before one success). For t in a neighborhood of 0,

$$M_{Y_i}(t) = \mathbb{E}[e^{tY_i}] = p \sum_{k=0}^{\infty} \big((1-p)e^t\big)^k = \frac{p}{1-(1-p)e^t}, \qquad \text{whenever } |(1-p)e^t| < 1.$$

Independence then gives

$$\boxed{M_X(t) = \prod_{i=1}^{r} M_{Y_i}(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \qquad t < -\log(1-p) \text{ (in particular near 0).}}$$

(b) Scaling and limit to χ^2_{2r}. For $Y = 2p\,X$,

$$M_Y(t) = \mathbb{E}[e^{t(2pX)}] = M_X(2pt) = \left(\frac{p}{1-(1-p)e^{2pt}}\right)^r.$$

As $p \downarrow 0$, use $e^{2pt} = 1 + 2pt + o(p)$ to expand the denominator:

$$1-(1-p)e^{2pt} = 1 - \big(1-p\big)\big(1+2pt+o(p)\big) = p(1-2t) + o(p).$$

Hence

$$\frac{p}{1-(1-p)e^{2pt}} = \frac{p}{p(1-2t)+o(p)} \longrightarrow \frac{1}{1-2t} \quad \text{for } |t| < \tfrac{1}{2},$$

and therefore

$$\boxed{\lim_{p\to 0} M_Y(t) = \big(1-2t\big)^{-r}, \qquad |t| < \tfrac{1}{2}.}$$

This is precisely the mgf of a chi-squared distribution with $2r$ degrees of freedom, so $Y \underset{p\downarrow 0}{\Longrightarrow} \chi^2_{2r}$. $\square$

Week 5 – Discussion

Problem 1. Let $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ be independent.

(a) Show that $X + Y \sim \text{Poisson}(\theta + \lambda)$.

(b) Show that the conditional distribution of X given $X + Y = n$ is $\text{Binomial}\left(n, \dfrac{\theta}{\theta + \lambda}\right)$. What is the distribution of $Y \mid X + Y = n$?

Solution.

(a) Using MGFs, $M_{X+Y}(t) = M_X(t)M_Y(t) = \exp\{\theta(e^t - 1)\}\exp\{\lambda(e^t - 1)\} = \exp\{(\theta + \lambda)(e^t - 1)\}$, hence $X + Y \sim \text{Poisson}(\theta + \lambda)$.

(b) For $n \in \mathbb{N}_0$ and $k = 0, \ldots, n$,

$$\mathbb{P}(X = k \mid X + Y = n) = \frac{\mathbb{P}(X = k, Y = n - k)}{\mathbb{P}(X + Y = n)} = \binom{n}{k}\left(\frac{\theta}{\theta + \lambda}\right)^k\left(\frac{\lambda}{\theta + \lambda}\right)^{n-k}.$$

Thus $X \mid (X + Y = n) \sim \text{Binomial}\big(n, \frac{\theta}{\theta+\lambda}\big)$, and symmetrically $Y \mid (X + Y = n) \sim \text{Binomial}\big(n, \frac{\lambda}{\theta+\lambda}\big)$.

□

Problem 2. Let X and Y be independent with $X \sim \text{Gamma}(r, 1)$, $Y \sim \text{Gamma}(s, 1)$ (shape $r, s > 0$, unit scale). Define

$$Z_1 = X + Y, \qquad Z_2 = \frac{X}{X + Y}.$$

Show that Z_1 and Z_2 are independent, with $Z_1 \sim \text{Gamma}(r + s, 1)$ and $Z_2 \sim \text{Beta}(r, s)$.

Solution. Consider the bijection $(x, y) \mapsto (z_1, z_2)$ with inverse $x = z_1 z_2$, $y = z_1(1 - z_2)$, where $z_1 > 0$, $0 < z_2 < 1$. The Jacobian of the inverse is

$$J = \begin{pmatrix} \partial x/\partial z_1 & \partial x/\partial z_2 \\ \partial y/\partial z_1 & \partial y/\partial z_2 \end{pmatrix} = \begin{pmatrix} z_2 & z_1 \\ 1 - z_2 & -z_1 \end{pmatrix}, \qquad \left|\det J\right| = z_1.$$

The joint density of (X, Y) is $f_{X,Y}(x, y) = \dfrac{1}{\Gamma(r)\Gamma(s)}\, x^{r-1} y^{s-1} e^{-(x+y)}$ for $x, y > 0$. Hence, for $z_1 > 0$, $0 < z_2 < 1$,

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_{X,Y}\big(z_1 z_2, z_1(1 - z_2)\big)\, |\det J| \\ &= \frac{1}{\Gamma(r)\Gamma(s)}\,(z_1 z_2)^{r-1}\big(z_1(1 - z_2)\big)^{s-1} e^{-z_1}\, z_1 \\ &= \underbrace{\frac{1}{\Gamma(r + s)}\, z_1^{r+s-1} e^{-z_1}}_{\text{Gamma}(r+s,1)} \times \underbrace{\frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)}\, z_2^{r-1}(1 - z_2)^{s-1}}_{\text{Beta}(r,s)}. \end{aligned}$$

The factorization shows $Z_1 \perp Z_2$, with the stated marginal laws. □