



Discrete Random Variables

Expectation, variance, and the indicator method

Probability Theory • Lecture 3

Applied Data Analysis — Module I

September 2026

Roadmap & learning objectives

So far we assigned probabilities to *events*. Today we attach *numbers* to outcomes.

- Define a **random variable**, its **PMF** and **CDF**; prove the basic properties of the CDF.
- Handle **functions** of random variables and **joint** distributions; define independence of random variables.
- Define **expectation**; prove **LOTUS** and **linearity**; see when $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
- Define **variance** and **covariance**; prove the computational identities and the variance-of-a-sum formula.
- Master the **indicator method** — the single most useful trick of the course.
- Prove **Markov's** and **Chebyshev's** inequalities.
- ML tie-in: **risk is an expectation**, the empirical risk is a sample mean, and its variance shrinks like $1/n$.



Where we stand after Lectures 1-2

$$(\Omega, \mathcal{F}, \mathbb{P}) \text{ with (K1)-(K3); } \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}; \quad A \perp\!\!\!\perp B \iff \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B); \\ \mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

- Events are *sets*; probabilities are numbers attached to sets.
- Independence and conditioning let us build joint probabilities from simpler pieces.

The new question

A gambler cares about *how much* she wins, an analyst about *how many* users click, a model about *how large* its error is. We need numbers attached to outcomes — and a way to summarize their typical size and spread.





1.

Random variables

Motivating example — from outcomes to numbers

Roll two fair dice; $\Omega = \{1, \dots, 6\}^2$, 36 equally likely outcomes. Let S = the sum.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$S(\omega)$ for each $\omega = (i, j)$: a function on Ω .

- S assigns a number to every outcome:
 $S(3, 4) = 7$.
- “ $S = 7$ ” is the event
 $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, of probability $6/36$.
- Many outcomes, few values: S **compresses** Ω to what we care about.

$$\mathbb{P}(S = 7) = \frac{6}{36}, \quad \mathbb{P}(S = 2) = \frac{1}{36}, \quad \mathbb{P}(S \leq 4) = \frac{6}{36}.$$



Definition — random variable

Definition (Random variable)

A **random variable** on $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \rightarrow \mathbb{R}$ such that $\{X \leq x\} = \{\omega : X(\omega) \leq x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$. X is **discrete** if it takes values in a finite or countable set.

- The condition says that “ $X \leq x$ ” is an *event*, so $\mathbb{P}(X \leq x)$ makes sense. For countable Ω with $\mathcal{F} = 2^\Omega$ it is automatic.
- Notation: capitals X, Y, S for random variables, small x, y, s for values; $\mathbb{P}(X = x)$ abbreviates $\mathbb{P}(\{\omega : X(\omega) = x\})$.

Examples

Number of heads in 10 flips; the sum of two dice; the indicator $\mathbf{1}\{A\}$ of an event; the number of defective items in a batch; the number of clicks on an ad; the 0/1 error of a classifier.



The probability mass function

Definition (PMF)

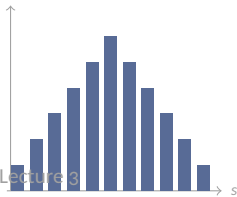
The **probability mass function** of a discrete X is $p_X(x) = \mathbb{P}(X = x)$, $x \in \mathbb{R}$.

Theorem

(i) $p_X(x) \geq 0$; (ii) $p_X(x) = 0$ except for countably many x ; (iii) $\sum_x p_X(x) = 1$; (iv) $\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x)$.

Proof.

(i) is (K1); (ii) holds since X is discrete. For (iii)–(iv): the events $\{X = x\}$ are disjoint and their union is Ω (resp. $\{X \in B\}$), so (K3) gives the sums. □



PMF of the sum of two dice:

$$p_5(s) = \frac{6 - |s - 7|}{36}, s = 2, \dots, 12 \text{ (heights shown } \times 36).$$



The cumulative distribution function

Definition (CDF)

$F_X(x) = \mathbb{P}(X \leq x)$. For discrete X : $F_X(x) = \sum_{t \leq x} p_X(t)$.

Theorem (Properties of the CDF)

(1) F_X is non-decreasing; (2) $F_X(x) \rightarrow 0$ as $x \rightarrow -\infty$ and $\rightarrow 1$ as $x \rightarrow +\infty$; (3) F_X is right-continuous; (4) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$ and $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$ — the jump at x .

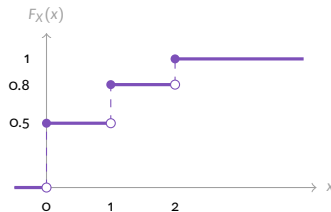
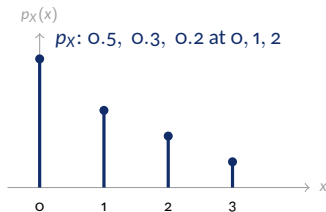
Proof.

(1) $x \leq y \Rightarrow \{X \leq x\} \subseteq \{X \leq y\}$; monotonicity. (2) $\{X \leq -n\} \downarrow \emptyset$, $\{X \leq n\} \uparrow \Omega$; continuity of $\mathbb{P}$ (Lecture 1). (3) $\{X \leq x + 1/n\} \downarrow \{X \leq x\}$. (4) $\{X \leq b\} = \{X \leq a\} \sqcup \{a < X \leq b\}$; and $\{X \leq x - 1/n\} \uparrow \{X < x\}$ gives $F_X(x^-) = \mathbb{P}(X < x)$. □

The CDF is defined for *every* random variable — it carries us to the continuous case in Lecture 6.



PMF and CDF side by side



- Jump sizes of F_X are exactly the PMF values:
 $F_X(1) - F_X(1^-) = 0.8 - 0.5 = 0.3 = p_X(1)$.
- Filled dot = value at the point (right-continuity); open dot = the left limit, not attained.
- Either function determines the other for a discrete random variable.

Functions of a random variable

If X is a random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$, then $Y = g(X)$ is a random variable with

$$p_Y(y) = \mathbb{P}(g(X) = y) = \sum_{x: g(x)=y} p_X(x).$$

Example

X takes values $-1, 0, 1, 2$ with probabilities $0.2, 0.3, 0.4, 0.1$. Let $Y = X^2$:

$$p_Y(0) = 0.3, \quad p_Y(1) = p_X(-1) + p_X(1) = 0.6, \quad p_Y(4) = 0.1.$$

Two values of X collapse into one value of Y : information is lost, probabilities are added.

- Common transformations: X^2 (squared error), $|X|$, $\max(X, 0)$ (option payoff, ReLU), $\mathbf{1}\{X > c\}$ (thresholding a score).
- Computing p_Y is always possible but often unnecessary — LOTUS lets us skip it.



Joint distributions of two random variables

Definition (Joint PMF, marginals)

The **joint PMF** of discrete X, Y is $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. The **marginals** are $p_X(x) = \sum_y p_{X,Y}(x, y)$ and $p_Y(y) = \sum_x p_{X,Y}(x, y)$.

$p_{X,Y}$	$Y=0$	$Y=1$	$Y=2$	p_X
$X=0$	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
$X=1$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
p_Y	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Two coins: X = first coin, Y = total number of heads. The margins of the table are the marginals (hence the name). Marginals do *not* determine the joint: many tables share the same marginals.

Proof of the marginal formula: $\{X = x\} = \bigsqcup_y \{X = x, Y = y\}$; apply (K3).



Independent random variables

Definition

X and Y are **independent** if $\{X \in A\}$ and $\{Y \in B\}$ are independent events for all $A, B \subseteq \mathbb{R}$. For discrete variables this is equivalent to

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \quad \text{for all } x, y.$$

Proof of the equivalence (sketch).

$(\Rightarrow)$ take $A = \{x\}$, $B = \{y\}$. $(\Leftarrow)$

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



- The joint table *factorizes* as an outer product of the margins. In the two-coin table it does not: $p_{X,Y}(0, 2) = 0 \neq \frac{1}{2} \cdot \frac{1}{4}$.
- $X_1, \dots, X_n$ are independent if $p_{X_1, \dots, X_n} = \prod_i p_{X_i}$; “i.i.d.” = independent and identically distributed.
- If $X \perp\!\!\!\perp Y$ then $g(X) \perp\!\!\!\perp h(Y)$ for any g, h .





2.

Expectation

Expectation — the definition

Definition (Expected value)

For a discrete X with PMF p_X , $\mathbb{E}[X] = \sum_x x p_X(x)$, provided $\sum_x |x| p_X(x) < \infty$.

- A **weighted average** of the values, weights = probabilities; the centre of mass of the PMF.
- Frequentist reading: the long-run average of many independent copies of X (the LLN, Lecture 13, makes this a theorem).
- Fair-price reading: the amount you should pay to play a game with payoff X .

Sum of two dice

$\mathbb{E}[S] = \sum_{s=2}^{12} s \cdot \frac{6-|s-7|}{36} = 7$ — by symmetry of the triangle around 7. We re-derive this in one line with linearity.



Why absolute convergence? The St. Petersburg game

A fair coin is flipped until the first head; if that happens on flip k you receive 2^k roubles.

$$\mathbb{P}(X = 2^k) = 2^{-k}, \quad \mathbb{E}[X] = \sum_{k \geq 1} 2^k \cdot 2^{-k} = \sum_{k \geq 1} 1 = +\infty.$$

- The expected payoff is infinite, yet nobody would pay 1,000 roubles to play: the median payoff is 2.
- For a variable taking both signs, rearranging a conditionally convergent series changes its sum — the “expectation” would depend on the order of summation. Hence the requirement $\sum |x|p_X(x) < \infty$.
- Heavy tails are not a curiosity: returns, insurance claims, and city sizes have distributions whose moments may not exist (Lecture 7).

Rule

When $\mathbb{E}[X]$ exists we say X is **integrable**. Infinite or undefined expectations are real possibilities, not pathologies.



LOTUS — the law of the unconscious statistician

Theorem (LOTUS)

For $Y = g(X)$, $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$ (whenever the series converges absolutely).

Proof.

$\mathbb{E}[Y] = \sum_y y p_Y(y) = \sum_y y \sum_{x: g(x)=y} p_X(x) = \sum_y \sum_{x: g(x)=y} y g(x) p_X(x) = \sum_x g(x) p_X(x)$, since every x belongs to exactly one group $\{x : g(x) = y\}$. $\square$

- “Unconscious”: we compute $\mathbb{E}[g(X)]$ without ever finding p_Y .
- The same proof gives $\mathbb{E}[g(X, Y)] = \sum_{x,y} g(x, y) p_{X,Y}(x, y)$.

Check with the earlier example

$\mathbb{E}[X^2] = 1 \cdot 0.2 + 0 + 1 \cdot 0.4 + 4 \cdot 0.1 = 1.0$; via p_Y : $0 \cdot 0.3 + 1 \cdot 0.6 + 4 \cdot 0.1 = 1.0$. Same number, less work.



Linearity of expectation

Theorem (Linearity)

For X, Y with finite expectations and constants a, b : $\mathbb{E}[aX + bY] = a \mathbb{E}[X] + b \mathbb{E}[Y]$. **No independence is required.**

Proof.

By two-variable LOTUS with $g(x, y) = ax + by$:

$$\mathbb{E}[aX + bY] = \sum_{x,y} (ax + by) p_{X,Y}(x, y) = a \sum_x x \sum_y p_{X,Y}(x, y) + b \sum_y y \sum_x p_{X,Y}(x, y),$$

and the inner sums are the marginals $p_X(x)$ and $p_Y(y)$. □

- By induction, $\mathbb{E}[\sum_i a_i X_i] = \sum_i a_i \mathbb{E}[X_i]$ for any finite family.
- The dependence structure disappears because we only ever sum out one variable at a time.

Two dice, in one line

$S = X_1 + X_2$, $\mathbb{E}[X_i] = 3.5$, so $\mathbb{E}[S] = 7$. No triangle needed.

Basic properties of expectation

Theorem

(1) $\mathbb{E}[c] = c$ and $\mathbb{E}[X + c] = \mathbb{E}[X] + c$; (2) **monotonicity**: $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$, in particular $X \geq 0 \Rightarrow \mathbb{E}[X] \geq 0$; (3) $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$; (4) if $X \geq 0$ and $\mathbb{E}[X] = 0$ then $\mathbb{P}(X = 0) = 1$.

Proof.

(1) The constant has PMF $p(c) = 1$; then linearity. (2) $Y - X \geq 0$, so $\mathbb{E}[Y - X] = \sum_z z p_{Y-X}(z) \geq 0$; then linearity. (3) $-|X| \leq X \leq |X|$ and monotonicity. (4) If $\mathbb{P}(X = x) > 0$ for some $x > 0$, then $\mathbb{E}[X] \geq x \mathbb{P}(X = x) > 0$. □

What linearity does *not* give

$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ in general: $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$ unless X is constant, $\mathbb{E}[1/X] \neq 1/\mathbb{E}[X]$, $\mathbb{E}[\max(X, Y)] \neq \max(\mathbb{E}[X], \mathbb{E}[Y])$. Jensen's inequality (Lecture 13) says which way it goes for convex g .



Expectation of a product

Theorem

If $X \perp\!\!\!\perp Y$ (with finite expectations), then $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$.

Proof.

LOTUS with $g(x, y) = xy$ and $p_{X,Y} = p_X p_Y$:

$$\mathbb{E}[XY] = \sum_{x,y} xy p_X(x) p_Y(y) = \left(\sum_x x p_X(x) \right) \left(\sum_y y p_Y(y) \right).$$



The converse is false

Let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Then $XY = X^3 = X$, so $\mathbb{E}[XY] = \mathbb{E}[X] = 0 = \mathbb{E}[X]\mathbb{E}[Y]$ — yet Y is a function of X , as dependent as it gets.

- More generally, $X \perp\!\!\!\perp Y \Rightarrow \mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \mathbb{E}[h(Y)]$ for all g, h — and this *does* characterize independence.
- Product rule needs independence; sum rule does not.





3.

Variance and covariance

Variance — the definition

Definition

Let $\mu = \mathbb{E}[X]$. The **variance** is $\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$, the **standard deviation** $\sigma_X = \sqrt{\text{Var}(X)}$; the k -th **moment** is $\mathbb{E}[X^k]$, the k -th **central moment** $\mathbb{E}[(X - \mu)^k]$.

- Variance is the expected squared distance from the mean: a measure of **spread**. It has the units of X^2 ; σ_X restores the units of X .
- $\text{Var}(X) \geq 0$, and $\text{Var}(X) = 0$ iff X is constant with probability 1.
- Why squared? Squares are differentiable, add up for independent sums, and make the mean the minimizer.

The minimizing property, proved

$\mathbb{E}[(X - c)^2] = \mathbb{E}[(X - \mu)^2] + 2(\mu - c)\mathbb{E}[X - \mu] + (\mu - c)^2 = \text{Var}(X) + (\mu - c)^2$, since $\mathbb{E}[X - \mu] = 0$. So $c = \mu$ is optimal.



Computing the variance

Theorem (Computational identities)

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad \text{Var}(aX + b) = a^2 \text{Var}(X).$$

Proof.

$\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2$ by linearity. For the affine rule: $aX + b$ has mean $a\mu + b$, so $\text{Var}(aX + b) = \mathbb{E}[a^2(X - \mu)^2] = a^2 \text{Var}(X)$. $\square$

- Shifting by b does not change spread; scaling by a scales the variance by a^2 .
- Corollary: $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ always — a first instance of Jensen.

Worked examples

$p_X(0) = 0.5, p_X(1) = 0.3, p_X(2) = 0.2$: $\mathbb{E}[X] = 0.7, \mathbb{E}[X^2] = 1.1, \text{Var}(X) = 0.61, \sigma_X \approx 0.78$.
One die: $\mathbb{E}[X] = 3.5, \mathbb{E}[X^2] = \frac{91}{6}, \text{Var}(X) = \frac{35}{12} \approx 2.92$.



Covariance

Definition

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Theorem (Properties)

(1) $\text{Cov}(X, X) = \text{Var}(X)$ and $\text{Cov}(X, Y) = \text{Cov}(Y, X)$; (2) **bilinearity**:
 $\text{Cov}(aX + bY, Z) = a \text{Cov}(X, Z) + b \text{Cov}(Y, Z)$, constants have zero covariance with everything; (3)
 $X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0$; the converse is **false**.

Proof.

(1) immediate. (2) Expand $\mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\mathbb{E}[Z]$ by linearity. (3) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for independent variables; counterexample: X uniform on $\{-1, 0, 1\}$, $Y = X^2$,
 $\text{Cov}(X, Y) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0$. □

Positive covariance: X and Y tend to be above their means together. Zero covariance = **uncorrelated**, which is weaker than independent.



Variance of a sum

Theorem

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y), \quad \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

If the X_i are pairwise uncorrelated (in particular independent), $\text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i)$.

Proof.

$\text{Var}(\sum_i X_i) = \text{Cov}(\sum_i X_i, \sum_j X_j) = \sum_{i,j} \text{Cov}(X_i, X_j)$ by bilinearity; separate the diagonal terms. $\square$

Two dice

$X_1 \perp\!\!\!\perp X_2$, so $\text{Var}(S) = 2 \cdot \frac{35}{12} \approx 5.83$, while $\text{Var}(2X_1) = 4 \cdot \frac{35}{12} \approx 11.7$: the sum of two *independent* dice is less spread than twice one die. Diversification in one formula.



Common error

Worked example — the mean of a sample

Let $X_1, \dots, X_n$ be i.i.d. with mean μ and variance σ^2 , and $\bar{X}_n = \frac{1}{n} \sum_i X_i$.

Theorem

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}, \quad \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}.$$

Proof.

Linearity: $\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_i \mu = \mu$. Independence and the affine rule:

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \text{Var}(\sum_i X_i) = \frac{1}{n^2} \cdot n\sigma^2.$$



- The average is **unbiased** for μ and its spread shrinks like $1/\sqrt{n}$: to halve the error, quadruple the sample.
- This is why averaging works — the engine behind the LLN (Lecture 13) and behind every “test accuracy $\pm$ ” you report.
- If the X_i are positively correlated, $\text{Var}(\bar{X}_n)$ has extra terms and does *not* vanish: correlated data are worth less.





4.

The indicator method

Indicator random variables

Definition

The **indicator** of an event A is $\mathbf{1}\{A\}(\omega) = 1$ if $\omega \in A$ and 0 otherwise.

Theorem

(1) $\mathbb{E}[\mathbf{1}\{A\}] = \mathbb{P}(A)$ and $\text{Var}(\mathbf{1}\{A\}) = \mathbb{P}(A)(1 - \mathbb{P}(A))$; (2) $\mathbf{1}\{A\}\mathbf{1}\{B\} = \mathbf{1}\{A \cap B\}$, $\mathbf{1}\{A^c\} = 1 - \mathbf{1}\{A\}$, $\mathbf{1}\{A \cup B\} = \mathbf{1}\{A\} + \mathbf{1}\{B\} - \mathbf{1}\{A \cap B\}$; (3) $\text{Cov}(\mathbf{1}\{A\}, \mathbf{1}\{B\}) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$, which is 0 iff $A \perp\!\!\!\perp B$.

Proof.

(1) $\mathbf{1}\{A\}$ takes value 1 with probability $\mathbb{P}(A)$; $\mathbf{1}\{A\}^2 = \mathbf{1}\{A\}$, so $\mathbb{E}[\mathbf{1}\{A\}^2] = \mathbb{P}(A)$. (2) Check both sides on $\omega \in A \cap B$, $\omega \in A \setminus B$, etc. (3) $\mathbb{E}[\mathbf{1}\{A\}\mathbf{1}\{B\}] = \mathbb{E}[\mathbf{1}\{A \cap B\}] = \mathbb{P}(A \cap B)$. □

Indicators translate **set algebra** into **arithmetic**, and **probabilities** into **expectations**. That is the whole trick.



The indicator method

If $X = \sum_{i=1}^n \mathbf{1}\{A_i\}$ counts how many of $A_1, \dots, A_n$ occur, then $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{P}(A_i)$ — **whether or not the A_i are independent.**

Proof.

Linearity of expectation plus $\mathbb{E}[\mathbf{1}\{A_i\}] = \mathbb{P}(A_i)$. □

The recipe

- 1 Recognize that X counts something.
- 2 Write X as a sum of indicators of the individual things being counted.
- 3 Compute each $\mathbb{P}(A_i)$ — usually easy, often all equal by symmetry.
- 4 Add. Ignore the dependence entirely.

Fixed points of a random permutation

Shuffle n cards: X = number of cards in their original position. $A_i = \{\pi(i) = i\}$, $\mathbb{P}(A_i) = \frac{1}{n}$, so

Indicator method — more examples

Birthday matches

n people; X = number of *pairs* sharing a birthday. For each of the $\binom{n}{2}$ pairs, $\mathbb{P}(\text{match}) = \frac{1}{365}$, so $\mathbb{E}[X] = \binom{n}{2} \frac{1}{365}$. At $n = 23$: $\mathbb{E}[X] = \frac{253}{365} \approx 0.69$ — consistent with $\mathbb{P}(X \geq 1) \approx 0.5$ from Lecture 1.

Distinct values (the coupon collector's cousin)

Draw n times uniformly from $\{1, \dots, m\}$; X = number of distinct values seen. With $A_j = \{\text{value } j \text{ appears}\}$, $\mathbb{P}(A_j) = 1 - (1 - \frac{1}{m})^n$, so $\mathbb{E}[X] = m[1 - (1 - \frac{1}{m})^n]$: the expected number of non-empty buckets when hashing n keys into m buckets.

Runs in coin flips

n flips; X = number of maximal runs. A new run starts at position 1 and at every $i \geq 2$ with $\omega_i \neq \omega_{i-1}$ (probability $\frac{1}{2}$), so $\mathbb{E}[X] = 1 + (n - 1) \cdot \frac{1}{2} = \frac{n+1}{2}$.

In all three cases the indicators are dependent. Nobody cares.



Variance with dependent indicators

Theorem

For $X = \sum_{i=1}^n \mathbf{1}\{A_i\}$,

$$\text{Var}(X) = \sum_i \mathbb{P}(A_i)(1 - \mathbb{P}(A_i)) + \sum_{i \neq j} [\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j)].$$

Proof.

The variance-of-a-sum formula with $\text{Var}(\mathbf{1}\{A_i\}) = \mathbb{P}(A_i)(1 - \mathbb{P}(A_i))$ and $\text{Cov}(\mathbf{1}\{A_i\}, \mathbf{1}\{A_j\}) = \mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j)$. □

Dependence enters *only* through the pairwise intersections $\mathbb{P}(A_i \cap A_j)$ — often as easy as the $\mathbb{P}(A_i)$ themselves.

Fixed points, continued

$\mathbb{P}(A_i) = \frac{1}{n}$ and $\mathbb{P}(A_i \cap A_j) = \frac{1}{n(n-1)}$ for $i \neq j$, hence

$$\text{Var}(X) = n \cdot \frac{1}{n} \left(1 - \frac{1}{n}\right) + n(n-1) \left[\frac{1}{n(n-1)} - \frac{1}{n^2}\right] = 1 - \frac{1}{n} + \frac{1}{n} = 1.$$

Inclusion-exclusion, proved with indicators

Theorem

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

Proof.

$\omega \notin \bigcup A_i$ iff $\omega \in A_i^c$ for all i , so

$$1 - \mathbf{1}\{\bigcup A_i\} = \prod_{i=1}^n (1 - \mathbf{1}\{A_i\}) = 1 - \sum_i \mathbf{1}\{A_i\} + \sum_{i < j} \mathbf{1}\{A_i\} \mathbf{1}\{A_j\} - \cdots$$

Use $\mathbf{1}\{A_i\} \mathbf{1}\{A_j\} = \mathbf{1}\{A_i \cap A_j\}$, rearrange, and take expectations — linearity turns every indicator into a probability. □

- Lecture 1 proved the case $n = 2$ by hand; indicators give all n at once, in three lines.
- Truncating the expansion yields the **Bonferroni inequalities**: the first sum alone gives the union bound $\mathbb{P}(\bigcup A_i) \leq \sum \mathbb{P}(A_i)$ — the tool behind generalization bounds in Lecture 13.





5.

Two inequalities

Markov's inequality

Theorem (Markov)

If $X \geq 0$ and $a > 0$, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

Proof.

For every ω : $a \mathbf{1}\{X \geq a\} \leq X$ (if $X(\omega) \geq a$ the left side is $a \leq X(\omega)$; otherwise it is $0 \leq X(\omega)$). Take expectations and use monotonicity: $a \mathbb{P}(X \geq a) \leq \mathbb{E}[X]$. $\square$

- One line — and it is the ancestor of every concentration inequality in the course.
- It is **tight**: for $X = a$ with probability p and 0 otherwise, both sides equal p . It is also often **weak**: for a die, $\mathbb{P}(X \geq 6) = \frac{1}{6}$ while Markov gives $\frac{3.5}{6} \approx 0.58$.

A useful reading

If average income is μ , at most a fraction $1/k$ of people earn $\geq k\mu$. No assumptions on the distribution at all.



Chebyshev's inequality

Theorem (Chebyshev)

If $\text{Var}(X) = \sigma^2 < \infty$ and $\mu = \mathbb{E}[X]$, then for every $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}, \quad \text{i.e.} \quad \mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof.

Apply Markov to the nonnegative variable $(X - \mu)^2$ with $a = t^2$:

$$\mathbb{P}((X - \mu)^2 \geq t^2) \leq \mathbb{E}[(X - \mu)^2]/t^2.$$



- At least 75% of the mass lies within 2σ of the mean, at least 89% within 3σ — for *any* distribution with a variance.
- Combine with $\text{Var}(\bar{X}_n) = \sigma^2/n$: $\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2) \rightarrow 0$. This is the weak law of large numbers — you have just proved it. Lecture 13 sharpens the rate.





6.

Machine learning: risk is an expectation

Risk, empirical risk, and why one estimates the other

Data $(X, Y) \sim \mathbb{P}$; predictor h ; loss $\ell(h(X), Y) \geq 0$. The **risk** and the **empirical risk** on an i.i.d. sample are

$$R(h) = \mathbb{E}[\ell(h(X), Y)], \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i).$$

Theorem

For a fixed h : $\mathbb{E}[\hat{R}_n(h)] = R(h)$ and $\text{Var}(\hat{R}_n(h)) = \frac{\text{Var}(\ell(h(X), Y))}{n}$.

Proof.

The losses $L_i = \ell(h(X_i), Y_i)$ are i.i.d. with mean $R(h)$; apply the sample-mean theorem. $\square$

Test error is an unbiased estimate of the true error, with standard deviation $O(1/\sqrt{n})$. Everything else in learning theory is about what changes when h is chosen using the same data (Lecture 13).



The 0/1 loss and its indicators

For classification, $\ell(h(X), Y) = \mathbf{1}\{h(X) \neq Y\}$, so

$$R(h) = \mathbb{P}(h(X) \neq Y), \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(X_i) \neq Y_i\} = \frac{\text{\#errors}}{n}.$$

- Accuracy = $1 - \hat{R}_n$ is an average of indicators — the indicator method again. Its variance is $\frac{R(1-R)}{n}$, so an accuracy of 0.90 on $n = 100$ test points has standard deviation 0.03: **two models at 0.90 and 0.92 on 100 points are indistinguishable.**
- Chebyshev: $\mathbb{P}(|\hat{R}_n - R| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$. For $\varepsilon = 0.01$ at confidence 95% you need $n \geq 50,000$ test points. (Lecture 13 improves this to $\approx 18,000$ via Hoeffding.)

Why surrogates

$\mathbf{1}\{h(X) \neq Y\}$ has zero gradient almost everywhere. We minimize the expectation of a *convex surrogate* — logistic, hinge, squared. The surrogate risk is still an expectation; today's mathematics applies unchanged.



Squared loss, bias, and variance — a preview

Predict a number Y by a constant c under squared loss: $R(c) = \mathbb{E}[(Y - c)^2]$.

Theorem

$R(c) = \text{Var}(Y) + (\mathbb{E}[Y] - c)^2$; hence the optimal constant predictor is $c^* = \mathbb{E}[Y]$ and the optimal risk is $\text{Var}(Y)$.

Proof.

This is the minimizing property of the mean proved earlier today. □

- **Variance** of the target is the floor below which no constant prediction can go; **bias** $(\mathbb{E}Y - c)^2$ is the penalty for a wrong centre.
- The full **bias-variance decomposition** (Lecture 11) replaces the constant by a learned $\hat{h}(X)$ and conditions on X ; the algebra is the same identity, applied twice.

Mean = best constant guess under squared loss; variance = its irreducible error.





7.

Wrap-up

Summary

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x); \quad \mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y; \quad \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2; \quad \text{Var}(X + Y) = \text{Var} X + \text{Var} Y + 2 \text{Cov}(X, Y).$$

- A random variable is a function on Ω ; its PMF and CDF encode all probabilities about it.
- Expectation is linear *without* independence; products factorize *with* independence — the converse fails.
- $\text{Var}(\bar{X}_n) = \sigma^2/n$: averaging independent measurements works, correlated ones work less.
- The indicator method: count $\rightarrow$ sum of indicators $\rightarrow$ sum of probabilities. Dependence enters the variance only through pairs.
- Markov and Chebyshev: tail bounds from a mean and a variance, no other assumptions.
- ML: risk is an expectation, empirical risk a sample mean; accuracy has error bars of size $\sqrt{R(1-R)/n}$.



Seminar problems & next lecture

- 1 X has PMF $p(-2) = 0.1$, $p(0) = 0.4$, $p(1) = 0.3$, $p(3) = 0.2$. Find F_X , $\mathbb{E}[X]$, $\text{Var}(X)$, and the PMF and mean of $Y = |X|$.
- 2 Two dice; $M = \max(X_1, X_2)$. Find the PMF of M , $\mathbb{E}[M]$, and $\text{Cov}(M, X_1)$.
- 3 Prove $\text{Var}(X) = \frac{1}{2}\mathbb{E}[(X - X')^2]$ for an independent copy X' of X .
- 4 n people throw their hats in the air and each picks one at random. Find the expected number and the variance of the number of people who get their own hat back. (You have already done this — say why.)
- 5 n balls into m boxes uniformly at random. Find the expected number of empty boxes and of boxes with exactly one ball.
- 6 Use Chebyshev to bound $\mathbb{P}(|S_{100} - 350| \geq 30)$ for the sum S_{100} of 100 dice; compare with simulation.
- 7 A classifier has true error 0.2. How many test points guarantee, via Chebyshev, that the test error is within 0.05 with probability ≥ 0.9 ?



Next: the essential discrete distributions — Bernoulli, binomial, geometric, Poisson.