

Session 4: Discrete Random Variables

Movement II: One Random Variable

Instructor: John Lipor

This is the first meeting of *Movement II: One Random Variable*. We restate the course slogan, as we will at the start of every movement:

A probability model answers three questions: What is uncertain? What is observed? How should observation change belief and action?

Movement I answered these for a target hiding among cells, where the uncertain thing was a *label*. Movement II keeps the whole belief-action loop and changes the objects inside it: from here on the uncertain thing is a *number*. The session thesis: *a random variable packages a numerical question about an uncertain world, and expectation is its single-number summary, linear no matter what.*

1 Launch

The search game so far asked *where*. Now ask *how much*. How many looks will tomorrow's search take before the target is found? How many mines will a single sweep of the minefield turn up? How many packets arrive at a relay in the next second? None of these is known in advance, yet each is a definite number once the world reveals itself. A quantity that is a number, but a number determined by the uncertain outcome, is a *random variable*. Today we build the three tools that tame such quantities: the distribution (what values, with what probabilities), expectation (the one-number summary), and the indicator method (which computes averages of complicated counts without ever writing down their distribution). By the end you will compute, in one line, how much ground a team of uncoordinated searchers covers, a count whose distribution we never write down.

2 Development

2.1 Random variables, PMF, and support

Definition 1 (Random variable). A **random variable** is a function $X : \Omega \rightarrow \mathbb{R}$ that assigns a number $X(\omega)$ to each outcome. Every set of values $B \subseteq \mathbb{R}$ then names an event $\{X \in B\} = \{\omega : X(\omega) \in B\}$, and X inherits its probabilities from $\mathbb{P}$: $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\})$.

(There is a technical condition, that each such set be a legitimate event, which measure theory handles; as promised in Session 1 we will never need it.) A random variable is *discrete* if it takes values in a finite or countable set. For discrete X the probabilities collapse to one value at a time.

Definition 2 (Probability mass function and support). *The **probability mass function** (PMF) of a discrete random variable X is $p_X(x) = \mathbb{P}(X = x)$. Its **support** is the set of values with $p_X(x) > 0$. The PMF is nonnegative and sums to one over the support: $\sum_x p_X(x) = 1$.*

Example: tossing a fair coin twice, let X count the heads. Then $\Omega = \{HH, HT, TH, TT\}$, and reading off the outcomes gives $p_X(0) = 1/4$, $p_X(1) = 1/2$, $p_X(2) = 1/4$. The event “at least one head” is $\{X \geq 1\}$, with $\mathbb{P}(X \geq 1) = p_X(1) + p_X(2) = 3/4$: once we have the PMF, probabilities of value-events are just sums.

2.2 The CDF: a running total

The PMF answers “what is the probability of exactly x ?” Often we want “at most x ,” and accumulating the PMF gives a function that means what it says at every real number.

Definition 3 (Cumulative distribution function). *The **cumulative distribution function** (CDF) of X is $F_X(x) = \mathbb{P}(X \leq x)$, defined for every $x \in \mathbb{R}$.*

For a discrete variable F_X is a running total of the PMF: it is flat between support points and jumps by $p_X(x_i)$ at each value x_i , producing a staircase (Figure 1). Three properties hold for any CDF and are worth reading off the staircase: F_X is nondecreasing; $F_X(x) \rightarrow 0$ as $x \rightarrow -\infty$ and $\rightarrow 1$ as $x \rightarrow +\infty$; and F_X is right-continuous, so the value *at* a jump is the higher one (the filled dot). The jump height recovers the PMF: $p_X(x_i) = F_X(x_i) - \lim_{x \uparrow x_i} F_X(x)$. The CDF is the object that will survive into Session 6, where continuous variables have no PMF but every variable has a CDF.

2.3 Expectation and LOTUS

Definition 4 (Expectation). *The **expectation** (expected value, mean) of a discrete random variable X is*

$$\mathbb{E}[X] = \sum_x x p_X(x), \quad (1)$$

the sum taken over the support.

Two readings, both correct and neither to be belabored. As a *long-run average*, $\mathbb{E}[X]$ is the value the sample mean of many independent copies of X settles toward. As a *center of mass*, $\mathbb{E}[X]$ is the balance point of the PMF regarded as weights placed along the number line. The mean of the coin-count above is $0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

To average a *function* of X we do not need the distribution of that function; we reweight the values of X themselves.

Theorem 1 (Law of the unconscious statistician (LOTUS)). *For any function $g : \mathbb{R} \rightarrow \mathbb{R}$,*

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x). \quad (2)$$

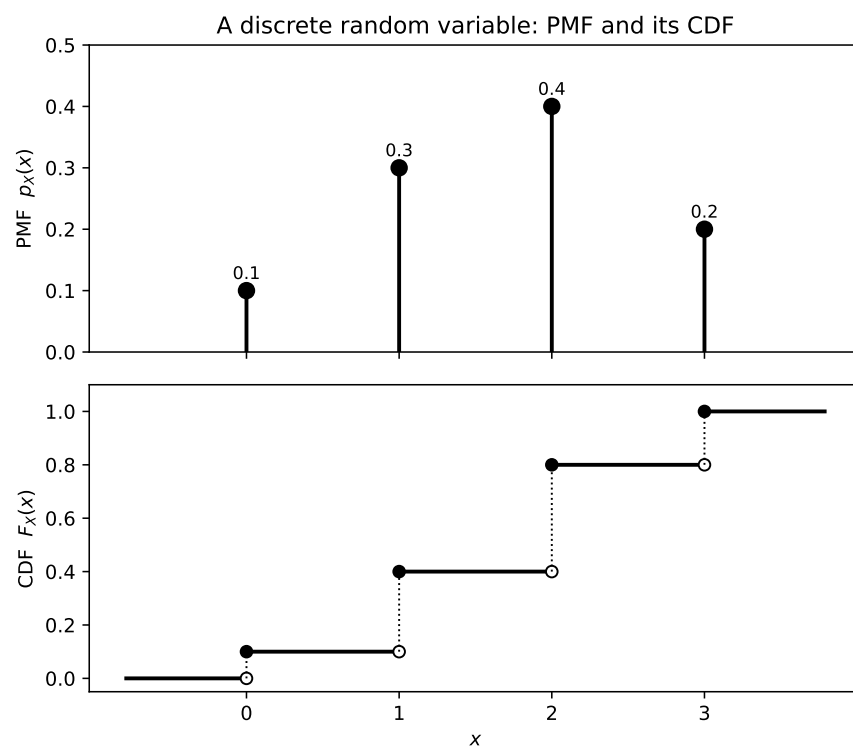


Figure 1: A discrete random variable's PMF (top) and its CDF (bottom). The CDF is the running total of the PMF; each jump equals the corresponding mass.

Proof. Group the outcomes of $g(X)$ by the value of X . By definition $\mathbb{E}[g(X)] = \sum_y y \mathbb{P}(g(X) = y)$, and $\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} p_X(x)$. Substituting and exchanging the order of the finite (or absolutely convergent) sums,

$$\mathbb{E}[g(X)] = \sum_y y \sum_{x:g(x)=y} p_X(x) = \sum_x g(x) p_X(x), \quad (3)$$

since for each x the outer value y equals $g(x)$. The point is that we never had to find the PMF of $g(X)$. $\square$

2.4 Linearity: the engine of the course

Theorem 2 (Linearity of expectation). *For random variables X, Y and constants a, b ,*

$$\mathbb{E}[aX + bY] = a \mathbb{E}[X] + b \mathbb{E}[Y]. \quad (4)$$

Proof. Work on the outcome space directly, where X and Y are ordinary functions of ω . Writing $\mathbb{E}[Z] = \sum_{\omega} Z(\omega) \mathbb{P}(\omega)$ for a discrete model,

$$\mathbb{E}[aX + bY] = \sum_{\omega} (aX(\omega) + bY(\omega)) \mathbb{P}(\omega) = a \sum_{\omega} X(\omega) \mathbb{P}(\omega) + b \sum_{\omega} Y(\omega) \mathbb{P}(\omega) = a \mathbb{E}[X] + b \mathbb{E}[Y]. \quad (5)$$

No assumption whatsoever was made about how X and Y relate. $\square$

Linearity needs no independence. Theorem 2 holds for *any* X and Y : independent or not, correlated or not, same experiment or different. This unconditional freedom is what makes the indicator method below so powerful. We will compute the mean of a complicated count by breaking it into a sum of simple pieces and adding their means, even when the pieces are tangled together, precisely because linearity never asks whether they are.

2.5 Indicators: counting by expectation

An **indicator** of an event A is the 0/1 random variable $\mathbb{1}\{A\}$ that equals 1 when A occurs and 0 otherwise. Its mean is just the probability of the event:

$$\mathbb{E}[\mathbb{1}\{A\}] = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A). \quad (6)$$

This tiny identity, married to linearity, is a counting machine. Whenever a random variable X *counts* how many of several events occur, write it as a sum of indicators, $X = \sum_i \mathbb{1}\{A_i\}$, and then

$$\mathbb{E}[X] = \sum_i \mathbb{E}[\mathbb{1}\{A_i\}] = \sum_i \mathbb{P}(A_i). \quad (7)$$

The expected count is the sum of the individual event probabilities, with no distribution of X ever computed and no independence required. The worked examples put this to work on a count whose pieces are tangled together.

2.6 Variance and standard deviation

Expectation locates the center; *variance* measures spread.

Definition 5 (Variance and standard deviation). *The **variance** of X is $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, and the **standard deviation** is $\sigma_X = \sqrt{\text{Var}(X)}$, carried in the same units as X .*

The two expressions agree by expanding the square and using linearity: $\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2$ with $\mu = \mathbb{E}[X]$. Variance is not linear; it responds to affine changes by the square of the scale:

$$\text{Var}(aX + b) = \mathbb{E}[(aX + b - a\mathbb{E}[X] - b)^2] = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2 \text{Var}(X). \quad (8)$$

The shift b drops out (moving the center does not change spread) and the scale a enters squared.

For a sum, the two-variable expansion shows what independence buys. For any X, Y ,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]), \quad (9)$$

the last bracket being the *cross term*. When X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, so the cross term vanishes and variances add: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. The same cancellation extends to any number of pairwise-independent terms.

Remark 1. *We are borrowing one fact on credit: that independence forces $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. It is intuitive (the average of a product of unrelated quantities is the product of the averages) and we will prove it properly in Session 8, once joint distributions give us the tools. Everything in this session's variance computations uses only the vanishing of that cross term.*

3 Worked examples

3.1 (a) Uncoordinated search: how much ground is covered?

Ten searchers set out over a field of $n = 20$ cells. They do not coordinate: each independently picks one cell uniformly at random and searches it. Let C be the number of *distinct* cells searched. Had they coordinated, C would be 10. How much coverage do they lose by not talking?

The PMF of C is a genuine chore: it requires counting the ways ten choices can land on exactly c distinct cells. The indicator method skips it, provided we choose the right pieces. Indicators on the *searchers* do not help, since whether searcher s reaches a new cell depends on everyone else's choices. Put the indicators on the *cells* instead: let $I_j = \mathbb{1}\{\text{cell } j \text{ is searched by someone}\}$, so that $C = \sum_{j=1}^{20} I_j$. Cell j is missed exactly when all ten searchers pick other cells; each does so with probability $19/20 = 0.95$, independently of the others, so

$$\mathbb{P}(I_j = 1) = 1 - (0.95)^{10} \approx 1 - 0.599 = 0.401. \quad (10)$$

By (6) and linearity,

$$\mathbb{E}[C] = \sum_{j=1}^{20} \mathbb{P}(I_j = 1) = 20(1 - (0.95)^{10}) \approx 8.03. \quad (11)$$

On average about two of the ten looks are wasted on cells someone else already searched.

Notice what linearity just ignored. The indicators are *not* independent: there are only ten searchers, so if cells 1 through 9 are all covered, at most one searcher is left for the other eleven cells. Linearity never asked. This is the freedom of Theorem 2 in action: a count whose pieces are tangled together, averaged in one line.

3.2 (b) The spread of the coverage count: cross terms that do not vanish

For the variance, expand as in (9), now with twenty terms:

$$\text{Var}(C) = \sum_{j=1}^{20} \text{Var}(I_j) + \sum_{j \neq l} \left(\mathbb{E}[I_j I_l] - \mathbb{E}[I_j] \mathbb{E}[I_l] \right). \quad (12)$$

Each indicator satisfies $I_j^2 = I_j$, so $\text{Var}(I_j) = 0.401 - 0.401^2 \approx 0.2403$. For the cross terms, the product $I_j I_l$ is itself an indicator, of the event that *both* cells are covered, so $\mathbb{E}[I_j I_l] = \mathbb{P}(j \text{ and } l \text{ both covered})$. Both cells are missed exactly when every searcher picks one of the other 18 cells, which each does with probability 0.90; inclusion–exclusion (Session 1) on the two “missed” events then gives

$$\mathbb{P}(j \text{ and } l \text{ both covered}) = 1 - 2(0.95)^{10} + (0.90)^{10} \approx 0.151. \quad (13)$$

Each cross term is therefore $0.151 - 0.401^2 \approx -0.00981$. It is negative because covering one cell uses up a searcher who might have covered the other. Each one is tiny, but there are $20 \times 19 = 380$ of them:

$$\text{Var}(C) \approx 20(0.2403) + 380(-0.00981) \approx 4.81 - 3.73 = 1.08, \quad \sigma_C \approx 1.04. \quad (14)$$

Had we treated the indicators as independent and dropped the cross terms, we would have reported a variance of 4.81, more than four times too large. Dependence left the mean untouched in (a), but it reshaped the spread: because the searchers form a fixed pool, the coverage count is far more predictable than twenty independent cells would be. Linearity of the mean needs no independence; additivity of variance does, and dependence enters the variance only through the cross terms.

3.3 (c) Planning with the average: why $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$

A search vehicle must cover one unit of distance, and its speed X depends on the sea state: 1 or 3 units per hour, equally likely. The time taken is $g(X) = 1/X$. A planner who averages first uses the mean speed $\mathbb{E}[X] = 2$ and budgets

$$g(\mathbb{E}[X]) = \frac{1}{\mathbb{E}[X]} = \frac{1}{2} = 0.5 \text{ hours}, \quad (15)$$

whereas the expected time, by LOTUS, is

$$\mathbb{E}[g(X)] = \frac{1}{2} \cdot \frac{1}{1} + \frac{1}{2} \cdot \frac{1}{3} = \frac{2}{3} \approx 0.667 \text{ hours}, \quad (16)$$

a third more than the plan: the slow case costs far more time than the fast case saves. The two disagree, and not by accident: $g(x) = 1/x$ is convex on the positive axis, and for convex g one always has $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. That inequality is *Jensen's inequality*; we name it and move on. The moral is operational: you may not push an expectation through a nonlinear function. Averaging and transforming do not commute unless g is affine, the one case where Theorem 2 applies.

4 Consolidation

Work these in class.

1. (*Indicators, tangled pieces.*) A faulty router delivers n packets in a uniformly random order (all $n!$ orders equally likely). Let X be the number of packets that arrive in their original position. Find $\mathbb{E}[X]$ without finding the PMF. Are your indicators independent?
Answer: $X = \sum_{i=1}^n \mathbb{1} \{\text{packet } i \text{ arrives in position } i\}$, and each indicator has probability $1/n$, so $\mathbb{E}[X] = n \cdot (1/n) = 1$ for every n . The indicators are dependent (if $n-1$ packets are in place, so is the last one), and linearity ignores that.
2. (*CDF.*) A discrete X has PMF $p_X(0), \dots, p_X(3) = (0.2, 0.5, 0.2, 0.1)$. Compute $F_X(1)$ and $\mathbb{P}(X \geq 2)$.
Answer: $F_X(1) = p_X(0) + p_X(1) = 0.7$; $\mathbb{P}(X \geq 2) = 1 - F_X(1) = 0.3$.
3. (*Conceptual.*) A relay's backlog X is 0 or 10 packets, each with probability $1/2$. Compute $\mathbb{E}[X]$. Is the expected value a backlog the relay can ever actually hold? What does this say about reading "expected" as "typical"?
Answer: $\mathbb{E}[X] = 5$, a value X never takes; the mean is a balance point, not a prediction of any single outcome, so "expected" is not the same as "typical."

5 Synthesis

A random variable is the course thesis pointed at a number: *what is uncertain* is now a numerical quantity, its distribution (PMF or, more durably, CDF) is the full model of it, and expectation compresses that model to one summary. The engine of the session was linearity, which holds with no strings attached, and its partner the indicator method, which turns any count into a sum of event probabilities and computed the expected coverage of ten uncoordinated searchers in a single line, even though the pieces of that count were dependent. Variance then measured the spread, and there, unlike for the mean, independence mattered: it is what makes the cross terms vanish so variances add, and in the coverage count, where it failed, the cross terms cut the variance by more than a factor of four. We also drew the boundary between averaging and transforming: $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ except when g is affine.

Next session takes the ad-hoc counts we built by hand and recognizes the recurring shapes among them, the handful of named distributions that model almost every sensor and arrival count, and it meets the first genuinely new tool of Movement II: the CDF method for the distribution of a maximum or a minimum.

Self-check questions

1. You know a random variable's PMF. In one sentence each, how do you get (i) its CDF and (ii) the probability of the event $\{X \geq 2\}$?

Answer: (i) $F_X(x)$ is the running total $\sum_{x_i \leq x} p_X(x_i)$; (ii) $\mathbb{P}(X \geq 2) = \sum_{x_i \geq 2} p_X(x_i) = 1 - F_X(1)$.

2. Why can the expected number of cells covered by uncoordinated searchers be computed without the distribution of that count, and why does treating the cell indicators as independent give the right mean but the wrong variance?

Answer: write the count as a sum of cell indicators and use linearity, which needs no independence; the variance, however, contains the cross terms $\mathbb{P}(\text{both covered}) - \mathbb{P}(\text{covered})^2$, which are negative here, so dropping them overstates the spread.

3. Give a one-line reason $\mathbb{E}[1/X]$ need not equal $1/\mathbb{E}[X]$, and name the general principle.

Answer: $1/x$ is nonlinear (convex), so averaging and transforming do not commute; Jensen's inequality gives $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ for positive X .