Chapter 9

Expected Value of a Discrete Random Variable

Lecture 9 | Devore 9e: 3.3

9.1 Leftovers on the cdf

Everything about X can be read off $F_X(x) = \mathbb{P}\{X \le x\}$, provided you are careful about which endpoints are included. Write

$$F_X(x^-) := \lim_{y \to x^-} F_X(y)$$

for the left limit (the value "just before" any jump). Then

$$\begin{aligned} \mathbb{P}\{X > x\} &= 1 - F_X(x), \\ \mathbb{P}\{X < x\} &= F_X(x^-), \\ \mathbb{P}\{X = x\} &= F_X(x) - F_X(x^-), \\ \mathbb{P}\{a < X \le b\} &= F_X(b) - F_X(a), \\ \mathbb{P}\{a \le X \le b\} &= F_X(b) - F_X(a^-). \end{aligned}$$

Derivation of the last one.

$$\{a \le X \le b\} = \{X \ge a\} \cap \{X \le b\} = \{X \le b\} \cap \{X < a\}^c = \{X \le b\} \setminus \{X < a\},$$

and $\{X < a\} \subset \{X \le b\}$, so the probabilities subtract: $\mathbb{P}\{a \le X \le b\} = \mathbb{P}\{X \le b\} - \mathbb{P}\{X < a\} = F_X(b) - F_X(a^-)$. □

Remark 9.1. For a discrete X the left limit is easy to evaluate: if X takes only the values $1, 2, 3, \ldots$ then $\mathbb{P}\{X < 3\} = \mathbb{P}\{X \le 2\} = F_X(2)$. If X can take the value 2.8 but nothing in $(2.8, 2.9]$, then $\mathbb{P}\{X < 2.9\} = \mathbb{P}\{X \le 2.8\} = F_X(2.8)$. In words: step back to the largest value X can actually attain.

9.2 Expected value

Example 9.2 (Where the definition comes from)**.** Among 100 students, the number registered for $0, 1, 2, 3, 4$ courses is $5, 10, 30, 30, 25$ respectively. On average, how many courses does a student

take?

$$\frac{\text{total courses}}{\text{total students}} = \frac{5\cdot 0 + 10\cdot 1 + 30\cdot 2 + 30\cdot 3 + 25\cdot 4}{100} = 0.05{\cdot}0{+}0.10{\cdot}1{+}0.30{\cdot}2{+}0.30{\cdot}3{+}0.25{\cdot}4 = 2.6.$$

Look at the middle expression: dividing through by 100 turned the counts into *proportions*, i.e. into a pmf. So the average is a **probability-weighted sum of the values**.

Definition 9.3 (Expected value)**.** Let X be a discrete random variable taking values in $D = \{x_1, x_2, \dots\}$ with $p_i = \mathbb{P}\{X = x_i\}$. The **expected value** (mean) of X is

$$\mathbb{E}[X] = \mu_X := \sum_{x_i \in D} x_i\, p_i = \sum_{x_i \in D} x_i\, \mathbb{P}\{X = x_i\},$$

provided the sum converges absolutely.

Remark 9.4. X is a *function*; $\mathbb{E}[X]$ is a *number*. Also, $\mathbb{E}[X]$ need not be a value X can take — no student registers for 2.6 courses.

Example 9.5 (Expectation generalizes the plain average)**.** Given numbers $a_1 \le a_2 \le \cdots \le a_n$, define X by $\mathbb{P}\{X = a_i\} = \frac{1}{n}$. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \cdot \frac{1}{n} = \frac{a_1 + \cdots + a_n}{n},$$

the ordinary mean. The general definition just allows unequal weights.

9.2.1 A worked series: the geometric mean

Example 9.6 (Shooting at a target)**.** $\mathbb{P}\{X = n\} = (1-\beta)^{n-1}\beta$ for $n \ge 1$. Then

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n(1-\beta)^{n-1}\beta = \beta \sum_{n=1}^{\infty} n(1-\beta)^{n-1}.$$

The trick is to recognize $n(1-\beta)^{n-1} = -\frac{d}{d\beta}(1-\beta)^n$:

$$\mathbb{E}[X] = -\beta \frac{d}{d\beta} \sum_{n=1}^{\infty} (1-\beta)^n = -\beta \frac{d}{d\beta}\left(\frac{1-\beta}{1-(1-\beta)}\right) = -\beta \frac{d}{d\beta}\left(\frac{1-\beta}{\beta}\right) = -\beta \frac{d}{d\beta}\left(\frac{1}{\beta} - 1\right) = -\beta\left(-\frac{1}{\beta^2}\right) = \boxed{\frac{1}{\beta}}.$$

(We used $\sum_{n\ge 1} a^n = \frac{a}{1-a}$ for $|a| < 1$, and differentiated term by term, which is legitimate inside the radius of convergence.)

Sanity check: if you hit 10% of the time you expect to need 10 shots. Good.

9.2.2 When the mean does not exist

Example 9.7 (Heavy tail)**.** Let X be the number of interviews needed to get a job, with

$$\mathbb{P}\{X = x\} = a \cdot \frac{1}{x^2}, \qquad x = 1, 2, 3, \dots$$

First find a. We need $\sum_{x\ge 1} a/x^2 = 1$, and $\sum_{x\ge 1} x^{-2} = \pi^2/6$, so $a = 6/\pi^2$. That is a perfectly legitimate pmf. But

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{6}{\pi^2}\frac{1}{x^2} = \frac{6}{\pi^2}\sum_{x=1}^{\infty} \frac{1}{x} = \infty.$$

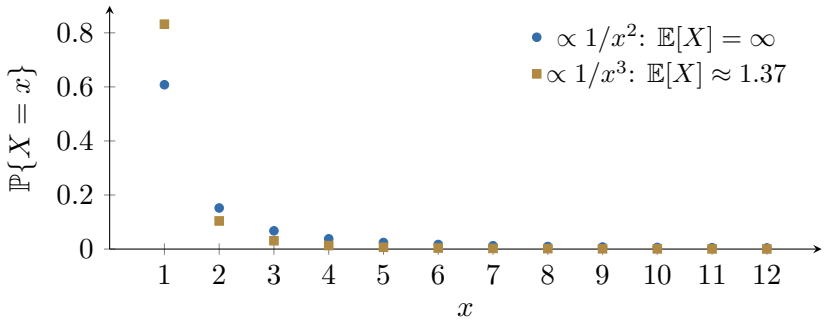
Remark 9.8. The mean is $+\infty$: the distribution has a **heavy tail**. Large values are individually unlikely but not unlikely *enough* to keep the weighted sum finite. This is not a pathology invented by mathematicians — insurance losses, file sizes, city populations, and income all behave this way, and for such data the sample mean is a badly behaved summary.

Example 9.9 (How much lighter must the tail be?)**.** Keep the same shape but decay one power faster: let

$$\mathbb{P}\{X = x\} = \frac{b}{x^3}, \qquad x = 1, 2, 3, \dots$$

Normalizing needs $\sum_{x\geq 1} x^{-3} = \zeta(3) \approx 1.2021$, so $b = 1/\zeta(3) \approx 0.832$. Now

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{b}{x^3} = b \sum_{x=1}^{\infty} \frac{1}{x^2} = b \cdot \frac{\pi^2}{6} \approx 0.832 \times 1.645 \approx 1.37 \ < \ \infty.$$



Remark 9.10. Both pmfs look like they decay to zero, and on a plot they are nearly indistinguishable past $x = 5$. Yet one has an infinite mean and the other does not. "Heavy tail" is therefore not something you can reliably eyeball — it is a statement about how fast $\sum x\, p(x)$ accumulates out in the tail, and the boundary sits exactly at the $1/x^2$ rate: for $p(x) \propto x^{-s}$ the sum $\sum x\, p(x)$ behaves like $\sum x^{1-s}$, which converges precisely when $s > 2$.

9.3 Functions of a random variable

If $X : S \to \mathbb{R}$ is a random variable and $h : \mathbb{R} \to \mathbb{R}$, then the composition $h(X) = h \circ X : S \to \mathbb{R}$ is again a random variable. Its expectation can be computed without ever finding the distribution of $h(X)$:

Theorem 9.11 (Law of the unconscious statistician)**.** *If X takes values in $D = \{x_1, x_2, \dots\}$ and $h : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[h(X)] = \sum_{x_i \in D} h(x_i)\, \mathbb{P}\{X = x_i\}.$$

Example 9.12. Shooting at a target, X = number of shots. Each bullet costs 5, and maintenance costs $\log X$. Total expense is $h(X) = 5X + \log X$:

X	1	2	3	$\cdots$
$h(X)$	$5 + \log 1$	$10 + \log 2$	$15 + \log 3$	$\cdots$
$\mathbb{P}$	β	$(1-\beta)\beta$	$(1-\beta)^2\beta$	$\cdots$

so $\mathbb{E}[h(X)] = \sum_{n\geq 1}(5n + \log n)(1-\beta)^{n-1}\beta$. Note we did *not* have to work out the pmf of $h(X)$ — that is exactly what the theorem buys us.

Corollary 9.13 (Linearity in the simplest case)**.** *For constants $a, b \in \mathbb{R}$,*

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b.$$

Proof.

$$\mathbb{E}[aX+b] = \sum_{x_i \in D}(ax_i + b)\mathbb{P}\{X = x_i\} = a\underbrace{\sum_{x_i\in D} x_i\mathbb{P}\{X = x_i\}}_{=\,\mathbb{E}[X]} + b\underbrace{\sum_{x_i\in D}\mathbb{P}\{X = x_i\}}_{=\,1} = a\mathbb{E}[X] + b. \quad \square$$

Remark 9.14. Linearity is what makes expectation usable. Note carefully that $\mathbb{E}[h(X)] \neq h(\mathbb{E}[X])$ in general — for instance $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$, and the gap between them is exactly the variance, which is next lecture.