1 Discrete Random Variables and Expectation

A **(real-valued) random variable** X on a sample space Ω is a measurable function $X : \Omega \to \mathbb{R}$, and a **discrete random variable** is one which may take on only a countable number of distinct values. "$X = a$" represents the set $\{s \in \Omega \mid X(s) = a\}$, and we denote the probability of that event by $\mathbf{P}(X = a) = \sum_{s\in\Omega:X(s)=a} \mathbf{P}(s)$.

Random variables $X_1, X_2, \cdots, X_n$ are **mutually independent** (simply called **independent** when $k = 2$) if and only if, for any subset $I \subseteq \{1, 2, \cdots, k\}$ and any values x_i $(i \in I)$, $\mathbf{P}(\bigcap_{i\in I}(X_i = x_i)) = \prod_{i\in I} \mathbf{P}(X_i = x_i)$.

The **expectation** of a discrete random variable X, denoted by $\mathbf{E}[X]$, is given by $\mathbf{E}[X] = \sum_i i \cdot \mathbf{P}(X = i)$. Note that the infinite series needs to be absolutely convergent (i.e. rearrangements do not change the value of the sum).

Theorem 1.1 (Linearity of expectation) For discrete random variables $X_1, X_2, \cdots, X_n$ with finite expectations and any contants $c_1, c_2, \cdots, c_n$, we have $\mathbf{E}[\sum_{i=1}^n c_iX_i] = \sum_{i=1}^n c_i\mathbf{E}[X_i]$.

Proof. Observe that we only need to prove the following two cases:

$$\begin{aligned}
\mathbf{E}[X + Y] &= \sum_i \sum_j (i + j) \cdot \mathbf{P}((X = i) \cap (Y = j)) \\
&= \sum_i i \sum_j \mathbf{P}((X = i) \cap (Y = j)) + \sum_j j \sum_i \mathbf{P}((X = i) \cap (Y = j)) = \mathbf{E}[X] + \mathbf{E}[Y], \\
\mathbf{E}[cX] &= \sum_i i \cdot \mathbf{P}(cX = j) = c \cdot \sum_j (j/c) \cdot \mathbf{P}(X = j/c) = c \cdot \sum_k k \cdot \mathbf{P}(X = k) = c \cdot \mathbf{E}[X].
\end{aligned}$$

When there are countably infinite variables, the situation becomes more subtle. We will discuss it later. ◀

Theorem 1.2 (Jensen's inequality) If f is a convex function, then $\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$.

Proof. Assume that f has a Taylor expansion. Let $\mu = \mathbf{E}[X]$. By Taylor's theorem, there is a value c such that

$$f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(c)(x - \mu)^2}{2} \geq f(\mu) + f'(\mu)(x - \mu)$$

Taking expectations of both sides

$$\mathbf{E}[f(X)] \geq \mathbf{E}[f(\mu) + f'(\mu)(X - \mu)] = \mathbf{E}[f(\mu)] + f'(\mu)(\mathbf{E}[X] - \mu) = f(\mu) = f(\mathbf{E}[X])$$

An alternative proof will be presented in Exercise 2.10. ◀

Define **conditional expectation** $\mathbf{E}[Y \mid Z = z] = \sum_y y \cdot \mathbf{P}(Y = y \mid Z = z)$ and $\mathbf{E}[Y \mid Z]$ as a random variable $f(Z)$ that takes on the value $\mathbf{E}[Y \mid Z = z]$ when $Z = z$.

Theorem 1.3 (Law of total expectation) For any random variables X and Y,

$$\mathbf{E}[X] = \sum_y \mathbf{P}(Y = y) \cdot \mathbf{E}[X \mid Y = y] = \mathbf{E}[\mathbf{E}[X \mid Y]].$$

A **Bernoulli** random variable X takes 1 with probability p and 0 with probability $1 - p$. A **binomial** random variable X with parameters n and p, denoted by $B(n, p)$, is defined by **probability distribution** $\mathbf{P}(X = k) = \binom{n}{k} \cdot p^k(1 - p)^{n-k}$, $n = 0, 1, \cdots, n$. Its expectation is np.

A **geometric** random variable X with parameter p is defined by probability distribution $\mathbf{P}(X = n) = (1-p)^{n-1}p$, $n = 1, 2, \cdots$. Its expectation is $1/p$. Geometric random variables are memoryless, that is, one ignores past failures as distribution does not change. Formally, we have the following statement.

Lemma 1.4 (Memorylessness) Let X be a geometric random variable with parameter p. Then, for $n > 0$,

$$\mathbf{P}(X = n + k \mid X > k) = \mathbf{P}(X = n).$$

Lemma 1.5 Let X be a discrete random variable that takes on only nonnegative integer values. Then,

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k \cdot \mathbf{P}(X = k) = \sum_{1\leq i\leq k} \mathbf{P}(X = k) = \sum_{i=1}^{\infty} \mathbf{P}(X \geq i)$$

Exercise 2.7 (a) By the memoryless property, we can ignore the case of $X > 1$ and $Y > 1$, thus $\mathbf{P}[X = Y] = \mathbf{P}[(X = 1) \cap (Y = 1)] \,/\, (1 - \mathbf{P}[(X > 1) \cap (Y > 1)])$. (b) Consider the first **trial**, and we can get an equation of $\mathbf{E}[\max(X, Y)]$. (c) Construct a **bernoulli trial** that success when there is at least one of two trials success. Its distribution of the first successful time provides the answer. (d) is the same as (a).

Exercise 2.14 (Negative binomial distribution) the k-th successful time. $\mathbf{P}(X = n) = \binom{n-1}{k-1}p^k(1 - p)^{n-k}$, $n \geq k$.

Exercise 2.16.b Break the sequence of flips up into disjoint blocks of $\lfloor \log_2 n - 2\log_2 \log_2 n \rfloor$ consecutive flips. For sufficiently large n, the probability is less than

$$\left(1 - 2^{\log_2 n - 2\log_2 \log_2 n}\right)^{\frac{n}{\log_2 n - 2\log_2 \log_2 n}} < \left(1 - \frac{n}{\log_2^2 n}\right)^{\frac{n}{\log_2^2 n} \cdot \log_2 n} < e^{-\ln n} = \frac{1}{n}.$$

Exercise 2.29 If $\{X_n\}$ is a sequence of random variable satisfying $X_n \to X$ **almost surely** (i.e. except possibly on an event of zero probability) then **(monotone convergence)** if $0 \le X_n \le X_{n+1}$ for all n almost surely, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$; **(dominated convergence)** if $|X_n| \le Y$ for all n almost surely and $\mathbf{E}[Y]$ is finite, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

Let $Z_n = \sum_{i=0}^n X_n$. We have $Z_n \to \sum_{i=0}^\infty X_n$ and $|Z_n| \le \sum_{i=0}^\infty |X_n|$ whose expectation is finite ($\mathbf{E}[\sum_{i=0}^\infty |X_n|] = \sum_{i=0}^\infty \mathbf{E}[|X_n|] < \infty$ is a consequence of monotone convergence). By dominated convergence, it follows that

$$\sum_{j=0}^{n} \mathbf{E}[X_j] = \mathbf{E}\left[\sum_{j=0}^{n} X_j\right] = \mathbf{E}[Z_n] \to \mathbf{E}[Z] = \mathbf{E}\left[\sum_{j=0}^{\infty} X_j\right], \qquad n \to \infty.$$

Exercise 2.32 For $i > m$, $\mathbf{P}(E_i) = \frac{1}{n} \cdot \frac{m}{i-1}$. Putting this all together, we get $\mathbf{P}(E) = \frac{m}{n} \sum_{j=m+1}^{n} \frac{1}{j-1}$. Then,

$$\frac{m}{n} \cdot \ln\left(\frac{n}{m}\right) = \frac{m}{n} \cdot \int_{m+1}^{n+1} \frac{\mathrm{d}x}{x-1} \le \mathbf{P}(E) \le \frac{m}{n} \cdot \int_{m}^{n} \frac{\mathrm{d}x}{x-1} = \frac{m}{n} \cdot \ln\left(\frac{n-1}{m-1}\right)$$

Note that $m(\ln n - \ln m)/n$ is maximized when $m = n/e$ and $\mathbf{P}(E) \ge 1/e$ for this choice of m.