DISCRETE RANDOM VARIABLES AND THEIR EXPECTATIONS

Contents

1 COMBINATORIAL PROBABILITY

In this section we will briefly review some combinatorial concepts, which come in handy when performing actual computations with discrete random variables. We start with two results from analysis:

1. Exponential function as a limit, for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

2. Stirling bounds on factorial

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

Definition 1. *Let Ω be a finite sample space. The discrete uniform probability space is $(\Omega, 2^{\Omega}, \mathbb{P})$, where*

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \quad \forall\, \omega \in \Omega.$$

Moreover, for any event $A \subset \Omega$, by finite additivity,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

Such assignments form the foundation of combinatorial probability wherein one is usually interested in counting the number of elements satisfying a particular criterion. This counting is often done following an iterative procedure. We collect some specific examples.

Example 1 (Permutations)**.** *A permutation of the numbers $1, \dots, n$ is an isomorphism $\pi \colon \{1, \dots, n\} \to \{1, \dots, n\}$. Proceeding iteratively, there are n choices for $\pi(1)$, $n-1$ choices for $\pi(2)$, $\cdots$, and 1 choice for $\pi(n)$. Therefore, there are $n \cdot (n-1) \cdots 2 \cdot 1 = n!$ possible permutations.*

Example 2 (Choices)**.** *Given a collection of n distinct objects there are $n \cdot (n-1) \cdots (n-(k-2)) \cdot (n-(k-1)) = n!/(n-k)!$ ways to select k of those objects in a particular order. Moreover, there are $n!/((n-k)!k!)$ ways to select k of those objects ignoring order, as $k!$ represents all possible ways to arrange k objects. This last expression $n!/((n-k)!k!)$ is denoted $\binom{n}{k}$ and referred to as n choose k.*

Example 3 (Birthday Paradox)**.** *Given n individuals, what is the probability that no two have the same birthday? Let $A =$ "All individuals in a group of size n have unique birthdays". Assume that an individual's birthday is independent of all other birthdays and occurs equally likely on any calendar day (non leap year), i.e. birthdays are independent and identically distribution uniformly on $\{1, \dots, 365\}$. A is the number of ways to uniquely select n birthdays, with $|A| = 365 \cdot 364 \cdots (365-(n-1))$, and the sample space is $\{1, \dots, 365\}^n$. Therefore, the resulting probability is*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{365!}{365^n (365-n)!}.$$

For $n = 23$, $\mathbb{P}(A) = .4927$, so it is more likely than two individuals will have the same birthday.

Example 4 (Mafia Game)**.** *Suppose that n members of the mafia are in one room and simultaneously shoot another mafia member uniformly at random (possibly themselves). Let $A =$ "Every member is shot". Let $\Omega = \{(\omega_1, \dots, \omega_n)\}$ be the set of assignments of mafia members to the people they shoot, i.e. $\omega_i \in$*

$\{1, \ldots, n\}$ *is the target of the* i*-th mafia member. The event* A *occurs for all* $\omega \in$ *that are permutations. Therefore, the corresponding probability is*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{n!}{n^n} \leq \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}}{n^n} \leq \left(\sqrt{2\pi ne}\right) e^{-n}.$$

Hence, for large n*, the event* A *is very unlikely.*

Example 5 (Mafia Survival)**.** *Under the setting of the Mafia Game, let* $B =$ *"The first mafia member survives". This time each shooter has* $(n-1)$ *admissible targets for event* B *to occur. Therefore, the corresponding probability is*

$$\Pr(B) = \frac{(n-1)^n}{n^n} = \left(1 - \frac{1}{n}\right) \xrightarrow{n} e^{-1}.$$

This result can be further generalized (see Section 2.1) to show that probability of the first mafia member dying from exactly 3 bullets is

$$\binom{n}{3} \frac{1}{n}\frac{1}{n}\frac{1}{n} \left(1 - \frac{1}{n}\right)^{n-3} \longrightarrow \frac{e^{-1}}{3!}.$$

Example 6 (Multinomial Coefficients)**.** *Similar to the binomial coefficient* $\binom{n}{k}$*, given* n *elements and* r *numbers* n_i*,* $i = 1, \ldots, r$*, with* $\sum_{i=1}^{r} n_i = n$*, the multinomial coefficient expresses the number of ways those* n *elements can be separated into* r *groups of size* n_i*. Proceeding iteratively, there are* $\binom{n}{n_1}$ *choices for the first group,* $\binom{n-n_1}{n_2}$ *choices for the second group,* $\cdots$*, and* $\binom{n_r}{n_r}$ *choices for the* r*-th group. In total this provides*

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n_r}{n_r} = \frac{n!}{n_1!n_2!\cdots n_r!} \triangleq \binom{n}{n_1, \ldots, n_r}$$

possible choices, where this last expression is the multinomial coefficient.

2 A FEW USEFUL RANDOM VARIABLES

Recall that a random variable $X : \Omega \to \mathbb{R}$ is called discrete if its range (i.e., the set of values that it can take) is a countable set. The PMF of X is a function $p_X : \mathbb{R} \to [0, 1]$, defined by $p_X(x) = \mathbb{P}(X = x)$, and completely determines the probability law of X.

The following are some important PMFs.

(a) **Discrete uniform** with parameters a and b, where a and b are integers with $a < b$. Here,

$$p_X(k) = 1/(b - a + 1), \qquad k = a, a + 1, \ldots, b,$$

and $p_X(k) = 0$, otherwise.[1]

(b) **Bernoulli** with parameter p, where $0 \leq p \leq 1$. Here, $p_X(0) = p$, $p_X(1) = 1 - p$.

(c) **Binomial** with parameters n and p, where $n \in \mathbb{N}$ and $p \in [0, 1]$. Here,

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

A binomial random variable with parameters n and p represents the number of heads observed in n independent tosses of a coin if the probability of heads at each toss is p.

(d) **Geometric** with parameter p, where $0 < p \leq 1$. Here,

$$p_X(k) = (1 - p)^{k-1} p, \qquad k = 1, 2, \ldots, .$$

A geometric random variable with parameter p represents the number of independent tosses of a coin until heads are observed for the first time, if the probability of heads at each toss is p.

(e) **Poisson** with parameter λ, where $\lambda > 0$. Here,

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots .$$

As will be seen shortly, a Poisson random variable can be thought of as a limiting case of a binomial random variable. Note that this is a legitimate PMF (i.e., it sums to one), because of the series expansion of the exponential function, $e^\lambda = \sum_{k=0}^{\infty} \lambda^k / k!$.

(f) **Power law** with parameter , where > 0. Here,

$$p_X(k) = \frac{1}{k} - \frac{1}{(k + 1)}, \qquad k = 1, 2, \ldots .$$

An equivalent but more intuitive way of specifying this PMF is in terms of the formula

$$\mathbb{P}(X \geq k) = \frac{1}{k}, \qquad k = 1, 2, \ldots .$$

[1] In the remaining examples, the qualification "$p_X(k) = 0$, otherwise," will be omitted for brevity.

Note that when α is small, the "tail" $\mathbb{P}(X \geq k)$ of the distribution decays slowly (slower than an exponential) as k increases, and in some sense such a distribution has "heavy" tails.

Notation: Let us use the abbreviations $\mathrm{dU}(a,b)$, $\mathrm{Ber}(p)$, $\mathrm{Bin}(n,p)$, $\mathrm{Geo}(p)$, $\mathrm{Pois}(\lambda)$, and $\mathrm{Pow}(\alpha)$ to refer the above defined PMFs. We will use notation such as $X \stackrel{d}{=} \mathrm{dU}(a,b)$ or $X \sim \mathrm{dU}(a,b)$ as a shorthand for the statement that X is a discrete random variable whose PMF is uniform on (a,b), and similarly for the other PMFs we defined. We will also use the notation $X \stackrel{d}{=} Y$ to indicate that two random variables have the same PMFs.

2.1 Poisson distribution as a limit of the binomial

To get a feel for the Poisson random variable, think of a binomial random variable with very small p and very large n. For example, consider the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss that results in a head when the word is misspelled), or the number of cars involved in accidents in a city on a given day (associate a car with a coin toss that results in a head when the car has an accident). Such random variables can be well modeled with a Poisson PMF.

More precisely, the Poisson PMF with parameter λ is a good approximation for a binomial PMF with parameters n and p, i.e.,

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

provided $\lambda = np$, n is large, and p is small. In this case, using the Poisson PMF may result in simpler models and calculations. For example, let $n = 100$ and $p = 0.01$. Then the probability of $k = 5$ successes in $n = 100$ trials is calculated using the binomial PMF as

$$\frac{100!}{95!\,5!} \cdot 0.01^5(1-0.01)^{95} = 0.00290.$$

Using the Poisson PMF with $\lambda = np = 100 \cdot 0.01 = 1$, this probability is approximated by

$$e^{-1}\frac{1}{5!} = 0.00306.$$

Proposition 1. (Binomial convergence to Poisson) *Let us fix some $\lambda > 0$, and suppose that $X_n \overset{d}{=} \mathrm{Bin}(n, \lambda/n)$, for every n. Let $X \overset{d}{=} \mathrm{Pois}(\lambda)$. Then, as $n \to \infty$, the PMF of X_n converges to the PMF of X, in the sense that $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$, for any $k \geq 0$.*

Proof: We have

$$\mathbb{P}(X_n = k) = \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \Big(1 - \frac{\lambda}{n}\Big)^{n-k}.$$

Fix k and let $n \to \infty$. We have, for $j = 1, \ldots, k$,

$$\frac{n-k+j}{n} \to 1, \quad \Big(1 - \frac{\lambda}{n}\Big)^{-k} \to 1, \quad \Big(1 - \frac{\lambda}{n}\Big)^{n} \to e^{-\lambda}.$$

Thus, for any fixed k, we obtain

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

as claimed. □

3 JOINT, MARGINAL, AND CONDITIONAL PMFS

In most applications, one typically deals with several random variables at once. In this section, we introduce a few concepts that are useful in such a context.

3.1 Marginal PMFs

Consider two discrete random variables X and Y associated with the same experiment. The probability law of each one of them is described by the corresponding PMF, p_X or p_Y, called a **marginal** PMF. However, the marginal PMFs do not provide any information on possible relations between these two random variables. For example, suppose that the PMF of X is symmetric around the origin. If we have either $Y = X$ or $Y = -X$, the PMF of Y remains the same, and fails to capture the specifics of the dependence between X and Y.

As another example let $X \overset{d}{=} \mathrm{Bin}(n, 1/2)$. Notice that then $Y = n - X$ also has $\mathrm{Bin}(n, 1/2)$ as a PMF. At the same time $X + Y = n$ and this is something that cannot be inferred from PMF alone.

3.2 Joint PMFs

The statistical properties of two random variables X and Y are captured by a function $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$, defined by

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y),$$

called the **joint PMF** of X and Y. We think of X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Namely, $X, Y : \Omega \to \mathbb{R}$. Then the event $A = \{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}$ is well defined. Then $\mathbb{P}(X = x, Y = y)$ is simply $\mathbb{P}(A)$. So that we can talk about the probability of this event we also need to ensure that the event is measurable.

Exercise 1. Suppose X and Y are two discrete random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Is it the case that for every $x, y \in \mathbb{R}$ the event $A = \{\omega \in \quad : X(\omega) = x, Y(\omega) = y\}$ is measurable with respect to $\mathcal{F}$? Either prove this or construct a counterexample.

From this point on we assume that the events A described above are measurable and will omit the measurability issues. Here and in the sequel, we will use the abbreviated notation $\mathbb{P}(X = x, Y = y)$ instead of the more precise notations $\mathbb{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbb{P}(X = x \text{ and } Y = x)$. More generally, the PMF of finitely many discrete random variables, $X_1, \ldots, X_n$ on the same probability space is defined by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Sometimes, we define a vector random variable X, by letting $X = (X_1, \ldots, X_n)$, in which case the joint PMF will be denoted simply as $p_X(x)$, where now the argument x is an n-dimensional vector.

The joint PMF of X and Y determines the probability of any event that can be specified in terms of the random variables X and Y. For example if A is the set of all pairs (x, y) that have a certain property, then

$$\mathbb{P}\big((X, Y) \in A\big) = \sum_{(x,y)\in A} p_{X,Y}(x, y).$$

In fact, we can calculate the marginal PMFs of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x, y), \qquad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

The formula for $p_X(x)$ can be verified using the calculation

$$p_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y p_{X,Y}(x, y),$$

where the second equality follows by noting that the event $\{X = x\}$ is the union of the countably many disjoint events $\{X = x, Y = y\}$, as y ranges over all the different values of Y. The formula for $p_Y(y)$ is verified similarly.

3.3 Conditional PMFs

Let X and Y be two discrete random variables, defined on the same probability space, with joint PMF $p_{X,Y}$. The **conditional PMF** of X given Y is a function $p_{X|Y}$, defined by

$$p_{X|Y}(x \,|\, y) = \mathbb{P}(X = x \,|\, Y = y), \qquad \text{if } \mathbb{P}(Y = y) > 0;$$

if $\mathbb{P}(Y = y) = 0$, the value of $p_{X|Y}(y \,|\, x)$ is left undefined. Using the definition of conditional probabilities, we obtain

$$p_{X|Y}(x \,|\, y) = \frac{p_{X,Y}(x, y)}{p_Y(y)},$$

whenever $p_Y(y) > 0$.

More generally, if we have random variables $X_1, \ldots, X_n$ ad $Y_1, \ldots, Y_m$, defined on the same probability space, we define a conditional PMF by letting

$$\begin{aligned} p_{X_1,\ldots,X_n \,|\, Y_1,\ldots,Y_m}&(x_1, \ldots, x_n \,|\, y_1, \ldots, y_m) \\ &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \,|\, Y_1 = y_1, \ldots, Y_m = y_m) \\ &= \frac{p_{X_1,\ldots,X_n,Y_1,\ldots,Y_m}(x_1, \ldots, x_n, y_1, \ldots, y_m)}{p_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m)}, \end{aligned}$$

whenever $p_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m) > 0$. Again, if we define $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$, the shorthand notation $p_{X|Y}(x \,|\, y)$ can be used.

Note that if $p_Y(y) > 0$, then $\sum_x p_{X|Y}(x \,|\, y) = 1$, where the sum is over all x in the range of the random variable X. Thus, the conditional PMF is essentially the same as an ordinary PMF, but with redefined probabilities that take into account the conditioning event $Y = y$. Visually, if we fix y, the conditional PMF $p_{X|Y}(x \,|\, y)$, viewed as a function of x is a "slice" of the joint PMF $p_{X,Y}$, renormalized so that its entries sum to one.

4 INDEPENDENCE OF RANDOM VARIABLES

We now define the important notion of independence of random variables. We start with a general definition that applies to all types of random variables, including discrete and continuous ones. We then specialize to the case of discrete random variables.

4.1 Independence of general random variables

Intuitively, two random variables are independent if any partial information on the realized value of one random variable does not change the distribution of the other. This notion is formalized in the following definition.

Definition 2. (Independence of random variables)

(a) Let $X_1, \ldots, X_n$ be random variables defined on the same probability space. We say that these random variables are independent if the events $X_1 \in B_1, \ldots, X_n \in B_n$ are independent for any Borel subsets $B_1, \ldots, B_n$ of the real line. Namely,

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n),$$

for any Borel subsets $B_1, \ldots, B_n$.

(b) Let $\{X_s \mid s \in S\}$ be a collection of random variables indexed by the elements of a (possibly infinite) index set S. We say that these random variables are independent if for every finite subset $\{s_1, \ldots, s_n\}$ of S, the random variables $X_{s_1}, \ldots, X_{s_n}$ are independent.

Verifying the independence of random variables using the above definition (which involves arbitrary Borel sets) is rather difficult. It turns out that one only needs to examine Borel sets of the form $(-\infty, x]$.

Proposition 2. *Suppose that for every n, every $x_1, \ldots, x_n$, and every finite subset $\{s_1, \ldots, s_n\}$ of S, the events $\{X_{s_i} \le x_i\}$, $i = 1, \ldots, n$, are independent. Then, the random variables X_s, $s \in S$, are independent.*

The proof is a simple application of Theorem 2 from Lecture 3 applied to a generating p-system $(-\infty, x]$.

Let us define the joint CDF of of the random variables $X_1, \ldots, X_n$ by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n).$$

In view of Proposition 2, independence of $X_1, \ldots, X_n$ is equivalent to the condition

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \qquad \forall\, x_1, \ldots, x_n.$$

Exercise 2. Consider a collection $\{A_s \mid s \in S\}$ of events, where S is a (possibly infinite) index set. Prove that the events A_s are independent if and only if the corresponding indicator functions I_{A_s}, $s \in S$, are independent random variables.

4.2 Independence of discrete random variables

For a finite number of discrete random variables, independence is equivalent to having a joint PMF which factors into a product of marginal PMFs.

Theorem 1. *Let X and Y be discrete random variables defined on the same probability space. The following are equivalent.*

(a) The random variables X and Y are independent.

(b) For any $x, y \in \mathbb{R}$, the events $\{X = x\}$ and $\{Y = y\}$ are independent.

(c) For any $x, y \in \mathbb{R}$, we have $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

(d) For any $x, y \in \mathbb{R}$ such that $p_Y(y) > 0$, we have $p_{X|Y}(x \mid y) = p_X(x)$.

Proof: The fact that (a) implies (b) is immediate from the definition of independence, since recall that the sets consisting of one point $\{x\}, \{y\}$ are Borel sets.

That (b) implies (c), and (c) implies (d) is also an immediate consequence of our definitions. Let us show that (d) implies (c). For the case when $p_Y(y) > 0$, we have $p_{X,Y}(x, y) = p_{X|Y}(x \mid y)p_Y(y) = p_X(x)p_Y(y)$. When $p_Y(y) = 0$, we have also $p_X(x)p_Y(y) = 0$. So in order to show the identity we need $p_{X,Y}(x, y) = 0$. But $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) \leq \mathbb{P}(Y = y) = 0$, and we have verified that both parts equal zero.

We complete the proof by verifying that (c) implies (a). Suppose that X and Y are independent, and let A, B, be two Borel subsets of the real line. We then

have

$$\begin{aligned}
\mathbb{P}(X \in A, Y \in B) &= \sum_{x \in A,\, y \in B} \mathbb{P}(X = x, Y = y) \\
&= \sum_{x \in A,\, y \in B} p_{X,Y}(x, y) \\
&= \sum_{x \in A,\, y \in B} p_X(x) p_Y(x) \\
&= \Big(\sum_{x \in A} p_X(x) \Big) \Big(\sum_{y \in B} p_Y(y) \Big) \\
&= \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B).
\end{aligned}$$

Since this is true for any Borel sets A and B, we conclude that X and Y are independent. □

We note that Theorem 1 generalizes to the case of multiple, but finitely many, random variables. The generalization of conditions (a)-(c) should be obvious. As for condition (d), it can be generalized to a few different forms, one of which is the following: given any subset S_0 of the random variables under consideration, the conditional joint PMF of the random variables X_s, $s \in S_0$, given the values of the remaining random variables, is the same as the unconditional joint PMF of the random variables X_s, $s \in S_0$, as long as we are conditioning on an event with positive probability.

We finally note that functions $g(X)$ and $h(Y)$ of two independent random variables X and Y must themselves be independent. This should be expected on intuitive grounds: If X is independent from Y, then the information provided by the value of $g(X)$ should not affect the distribution of Y, and consequently should not affect the distribution of $h(Y)$. Observe that when X and Y are discrete, then $g(X)$ and $h(Y)$ are random variables (the required measurability conditions are satisfied) even if the functions g and h are not measurable (why?).

Theorem 2. *Let X and Y be independent discrete random variables. Let g and h be some functions from $\mathbb{R}$ into itself. Then, the random variables $g(X)$ and $h(Y)$ are independent.*

The proof is left as an exercise.

4.3 Examples

Example. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with the same

parameter p. Then, the random variable X defined by $X = X_1 + \cdots + X_n$ is binomial with parameters n and p. To see this, consider n independent tosses of a coin in which every toss has probability p of resulting in a one, and let X_i be the result of the ith coin toss. Then, X is the number of ones observed in n independent tosses, and is therefore a binomial random variable.

Example. Let X and Y be independent binomial random variables with parameters (n, p) and (m, p), respectively. Then, the random variable Z, defined by $Z = X + Y$ is binomial with parameters $(n+m, p)$. To see this, consider $n+m$ independent tosses of a coin in which every toss has probability p of resulting in a one. Let X be the number of ones in the first n tosses, and let Y be the number of ones in the last m tosses. Then, Z is the number of ones in $n+m$ independent tosses, which is binomial with parameters $(n + m, p)$.

Example. Consider n independent tosses of a coin in which every toss has probability p of resulting in a one. Let X be the number of ones obtained, and let $Y = n - X$, which is the number of zeros. The random variables X and Y are not independent. For example, $\mathbb{P}(X = 0) = (1 - p)^n$ and $\mathbb{P}(Y = 0) = p^n$, but $\mathbb{P}(X = 0, Y = 0) = 0 \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0)$. Intuitively, knowing that there was a small number of heads gives us information that the number of tails must be large.

However, in sharp contrast to the intuition from the preceding example, we obtain independence when the number of coin tosses is itself random, with a Poisson distribution. More precisely, let N be a Poisson random variable with parameter λ. We assume that X has conditional PMF $p_{X|N}(\cdot \mid n)$ is binomial with parameters n and p (representing the number of ones observed in n coin tosses), and define $Y = N - X$, which represents the number of zeros obtained. We have the following surprising result. An intuitive justification will have to wait until we consider the Poisson process, later in this course. The proof is left as an exercise.

Theorem 3. (Splitting of a Poisson random variable) *The random variables X and Y are independent. Moreover, $X \overset{d}{=} \text{Pois}(\lambda p)$ and $Y \overset{d}{=} \text{Pois}\big(\lambda(1-p)\big)$.*

5 EXPECTED VALUES

5.1 Preliminaries: infinite sums

Consider a sequence $\{a_n\}$ of nonnegative real numbers and the infinite sum $\sum_{i=1}^{\infty} a_i$, defined as the limit, $\lim_{n\to\infty} \sum_{i=1}^{n} a_i$, of the partial sums. The infinite sum can be finite or infinite; in either case, it is well defined, as long as we allow the limit to be an extended real number. Furthermore, it can be verified that the value of the infinite sum is the same even if we reorder the elements of the sequence $\{a_n\}$ and carry out the summation according to this different order. Because the order of the summation does not matter, we can use the notation $\sum_{n\in\mathbb{N}} a_n$ for the infinite sum. More generally, if C is a countable set and $g : C \to [0, \infty)$ is a nonnegative function, we can use the notation $\sum_{x\in C} g(x)$, which is unambiguous even without specifying a particular order in which the values $g(x)$ are to be summed.

When we consider a sequence of nonpositive real numbers, the discussion remains the same, and infinite sums can be unambiguously defined. However, when we consider sequences that involve both positive and negative numbers, the situation is more complicated. In particular, the order at which the elements of the sequence are added can make a difference.

Example. Let $a_n = (-1)^n/n$. It can be verified that the limit $\lim_{n\to\infty} \sum_{i=1}^{n} a_n$ exists, and is finite, but that the elements of the sequence $\{a_n\}$ can be reordered to form a new sequence $\{b_n\}$ for which the limit of $\sum_{i=1}^{n} b_n$ does not exist.

In order to deal with the general case, we proceed as follows. Let S be a countable set, and consider a collection of real numbers a_s, $s \in S$. Let S_+ (respectively, S_-) be the set of indices s for which $a_s \geq 0$ (respectively, $a_s < 0$). Let $S_+ = \sum_{s\in S_+} a_s$ and $S_- = \sum_{s\in S_-} |a_s|$. We distinguish four cases.

(a) If both S_+ and S_- are finite (or equivalently, if $\sum_{s\in S} |a_s| < \infty$), we say that the sum $\sum_{s\in S} a_s$ is **absolutely convergent**, and is equal to $S_+ - S_-$. In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_{s_i} = S^+ - S_-.$$

(b) If $S_+ = \infty$ and $S_- < \infty$, the sum $\sum_{s\in S} a_s$ is not absolutely convergent; we define it to be equal to ∞. In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_{s_i} = \infty.$$

(c) If $S_+ < \infty$ and $S_- = \infty$, the sum $\sum_{s\in S} a_s$ is not absolutely convergent; we define it to be equal to $-\infty$. In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_{s_i} = -\infty.$$

(d) If $S_+ = \infty$ and $S_- = \infty$, the sum $\sum_{s\in S} a_n$ is left undefined. In fact, in this case, different arrangements of the elements of S in a sequence $\{s_n\}$ will result into different or even nonexistent values of the limit

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_{s_i}.$$

To summarize, we consider a countable sum to be well defined in cases (a)-(c), and call it absolutely convergent only in case (a).

We close by recording a related useful fact. If we have a doubly indexed family of nonnegative numbers a_{ij}, $i, j \in \mathbb{N}$, and if either (i) the numbers are nonnegative, or (ii) the sum $\sum_{(i,j)} a_{ij}$ is absolutely convergent, then

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij} = \sum_{(i,j)\in\mathbb{N}^2} a_{ij}. \tag{1}$$

More important, we stress that the first equality need not hold in the absence of conditions (i) or (ii) above.

5.2 Definition of the expectation

The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X. It is often desirable to summarize this information in a single representative number. This is accomplished by the **expectation** of X, which is a weighted (in proportion to probabilities) average of the possible values of X.

As motivation, suppose you spin a wheel of fortune many times. At each spin, one of the numbers $m_1, m_2, \ldots, m_n$ comes up with corresponding probability $p_1, p_2, \ldots, p_n$, and this is your monetary reward from that spin. What is the amount of money that you "expect" to get "per spin"? The terms "expect" and "per spin" are a little ambiguous, but here is a reasonable interpretation.

Suppose that you spin the wheel k times, and that k_i is the number of times that the outcome is m_i. Then, the total amount received is $m_1k_1 + m_2k_2 + \cdots + m_nk_n$. The amount received per spin is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k}.$$

If the number of spins k is very large, and if we are willing to interpret probabilities as relative frequencies, it is reasonable to anticipate that m_i comes up a fraction of times that is roughly equal to p_i:

$$\frac{k_i}{k} \approx p_i, \qquad i = 1, \ldots, n.$$

Thus, the amount of money per spin that you "expect" to receive is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k} \approx m_1p_1 + m_2p_2 + \cdots + m_np_n.$$

Motivated by this example, we introduce the following definition.

Definition 3. (Expectation) *We define the* **expected value** *(also called the* **expectation** *or the* **mean***) of a discrete random variable X, with PMF p_X, as*

$$\mathbb{E}[X] = \sum_x xp_X(x),$$

whenever the sum is well defined, and where the sum is taken over the countable set of values in the range of X.

Observe that $\mathbb{E}[X]$ is non-negative if X only takes non-negative values with positive probability. (Namely, $p_X(x) > 0$ implies $x \geq 0$).

5.3 Properties of the expectation

We start by pointing out an alternative formula for the expectation, and leave its proof as an exercise. In particular, if X can only take nonnegative integer values, then

$$\mathbb{E}[X] = \sum_{n \geq 0} \mathbb{P}(X > n). \tag{2}$$

Example. Using this formula, it is easy to give an example of a random variable for which the expected value is infinite. Consider $X \stackrel{d}{=} \mathrm{Pow}(\alpha)$, where $\alpha \leq 1$. Then, it can

be verified, using the fact $\sum_{n=1}^{\infty} 1/n = \infty$, that $\mathbb{E}[X] = \sum_{n \geq 0} \frac{1}{n} = \infty$. On the other hand, if $\alpha > 1$, then $\mathbb{E}[X] < \infty$.

Here is another useful fact, whose proof is again left as an exercise.

Proposition 3. *Given a discrete random variable X and a function $g : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}[g(X)] = \sum_{\{x \,|\, p_X(x) > 0\}} g(x) p_X(x). \tag{3}$$

More generally, this formula remains valid given a vector $X = (X_1, \ldots, X_n)$ of random variables with joint PMF $p_X = p_{X_1,\ldots,X_n}$, and a function $g : \mathbb{R}^n \to \mathbb{R}$.

For example, suppose that X is a discrete random variable and consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x^2$. Let $Y = g(X)$. In order to calculate the expectation $\mathbb{E}[Y]$ according to Definition 2, we need to first find the PMF of Y, and then use the formula $\mathbb{E}[Y] = \sum_y y p_Y(y)$. However, according to Proposition 3, we can work directly with the PMF of X, and write $\mathbb{E}[Y] = \mathbb{E}[X^2] = \sum_x x^2 p_X(x)$.

The quantity $\mathbb{E}[X^2]$ is called the **second moment** of X. More generally, if $r \in \mathbb{N}$, the quantity $\mathbb{E}[X^r]$ is called the rth moment of X. Furthermore, $\mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^r\big]$ is called the rth **central moment** of X. The second central moment, $\mathbb{E}[(X - \mathbb{E}[X])^2]$ is called the **variance** of X, and is denoted by $\mathrm{var}(X)$. The square root of the variance is called the **standard deviation** of X, and is often denoted by σ_X, or just σ. Note, that for every even r, the rth moment and the rth central moment are always nonnegative; in particular, the standard deviation is always well defined.

We continue with a few more important properties of expectations. In the sequel, notations such as $X \geq 0$ or $X = c$ mean that $X(\omega) \geq 0$ or $X(\omega) = c$, respectively, for all $\omega \in \Omega$. Similarly, a statement such as "$X \geq 0$, almost surely" or "$X \geq 0$, a.s.," means that $\mathbb{P}(X \geq 0) = 1$.

Proposition 4. *Let X and Y be discrete random variables defined on the same probability space.*

(a) If $X \geq 0$, a.s., then $\mathbb{E}[X] \geq 0$.

(b) If $X = c$, a.s., for some constant $c \in \mathbb{R}$, then $\mathbb{E}[X] = c$.

(c) **Linearity of expectation.** *For any $a, b \in \mathbb{R}$, we have $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ (as long as the sum $a\mathbb{E}[X] + b\mathbb{E}[Y]$ is well-defined).*

(d) If $\mathbb{E}[X]$ is finite, then $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

(e) For every $a \in \mathbb{R}$, we have $\mathrm{var}(aX) = a^2\mathrm{var}(X)$.

(f) If X and Y are independent and have finite expectations, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

(g) More generally, if $X_1, \ldots, X_n$ are independent and have finite expectations, then

$$\mathbb{E}\Big[\prod_{i=1}^{n} X_i\Big] = \prod_{i=1}^{n} \mathbb{E}[X_i],$$

and

$$\mathrm{var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{var}(X_i).$$

Remark: We emphasize that property (c) does not require independence.

Proof: We only give the proof for the case where all expectations involved are well defined and finite, and leave it to the reader to verify that the results extend to the case where all expectations involved are well defined but possibly infinite.

Parts (a) and (b) are immediate consequences of the definitions. For part (c), we use the second part of Proposition 3, and then Eq. (1), we obtain

$$\begin{aligned}
\mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by) p_{X,Y}(x, y) \\
&= \sum_{x} \Big(ax \sum_{y} p_{X,Y}(x, y) \Big) + \sum_{y} \Big(by \sum_{x} p_{X,Y}(x, y) \Big) \\
&= a \sum_{x} x p_X(x) + b \sum_{y} y p_Y(y) \\
&= a\mathbb{E}[X] + b\mathbb{E}[Y].
\end{aligned}$$

For part (d), we have

$$\begin{aligned}\mathrm{var}(X) &= \mathbb{E}\big[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2\big] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}\big[X\mathbb{E}[X]\big] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\end{aligned}$$

where the second equality made use of property (c).

Part (e) follows easily from (d) and (c). For part (f), we apply Proposition 3 and then use independence to obtain

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xyp_{X,Y}(x,y) \\ &= \sum_{x,y} xyp_X(x)p_Y(y) \\ &= \Big(\sum_x xp_X(x)\Big)\Big(\sum_y yp_Y(y)\Big) \\ &= \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

Furthermore, using property (d), we have

$$\begin{aligned}\mathrm{var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - (\mathbb{E}[X])^2 - (\mathbb{E}[Y])^2 - 2\mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Using the equality $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, the above expression becomes $\mathrm{var}(X) + \mathrm{var}(Y)$. The proof of part (g) is similar and is omitted. □

Remark: The equalities in part (f) need not hold in the absence of independence. For example, consider a random variable X that takes either value 1 or -1, with probability 1/2. Then, $\mathbb{E}[X] = 0$, but $\mathbb{E}[X^2] = 1$. If we let $Y = X$, we see that $\mathbb{E}[XY] = \mathbb{E}[X^2] = 1 \neq 0 = \big(\mathbb{E}[X]\big)^2$. Furthermore, $\mathrm{var}(X+Y) = \mathrm{var}(2X) = 4\mathrm{var}(X)$, while $\mathrm{var}(X) + \mathrm{var}(Y) = 2$.

Exercise 3. Show that $\mathrm{var}(X) = 0$ if and only if there exists a constant c such that $\mathbb{P}(X = c) = 1$.

6.436J / 15.085J Fundamentals of Probability

Fall 2018