Contents

1 Probability

1.1 Discrete Probability Spaces

In order to formalize the idea of a discrete probability space, we need the idea of a set being countable which, roughly, means it has size no larger than the size of the natural numbers.

Definition 1 (Finite). *A set S is **finite** if there is some $n \in \mathbb{N}$ such that $|S| = n$.*

Definition 2 (Countable). *A set S is **countable** if it is finite or $|S| = |\mathbb{N}|$.*

Likewise, we need the idea of a probability function.

Definition 3 (Probability Function). *A function $\Pr$ is a **probability function** if its domain is S, its codomain is $[0, 1]$ and $\sum_{\omega \in S} \Pr(\omega) = 1$.*

The following formalizes a discrete probability space.

Definition 4 (Discrete Probability Space). *A **discrete probability space** consists of a countable set S and a probability function $\Pr : S \to [0, 1]$.*

If $(S, \Pr)$ is a discrete probability space then we refer to S as the **sample space** and each $\omega \in S$ as an **outcome**.

The idea of an event allows us to formalize the idea of something happening in our probability space.

Definition 5 (Event). *An **event** in a discrete probability space $(S, \Pr)$ is any subset $E \subseteq S$.*

We extend the idea of the probability of an outcome to the probability of an event as follows.

Definition 6 (Event Probability). *Given discrete probability space $(S, \Pr)$, the probability of event $E \subseteq S$ is defined as $\Pr(E) = \sum_{\omega \in E} \Pr(\omega)$.*

1.2 Uniform Probability Spaces

One particularly nice kind of discrete probability space is one in which every outcome has the same probability.

Definition 7 (Uniform Probability Space). *A discrete probability space is **uniform** $(S, \Pr)$ if for any $\omega_1, \omega_2 \in S$, we have $\Pr(\omega_1) = \Pr(\omega_2)$.*

The following two facts show that understanding the probability of an outcome or event in such a probability space reduces to determining set cardinalities (i.e. combinatorics!).

Theorem. If discrete probability space $(S, \Pr)$ is uniform then for any outcome $\omega \in S$ we have $\Pr(\omega) = \frac{1}{|S|}$.

Theorem. If discrete probability space $(S, \Pr)$ is uniform then for any event $E \subseteq S$ we have $\Pr(E) = \frac{|E|}{|S|}$.

1.3 Probability Rules

Even when the discrete probability space is not uniform, the usual counting rules from combinatorics carry over. For example, the following are the analogues of the sum and subtraction rule from combinatorics but applied to probability.

Theorem (Probability Sum Rule)**.** If events $A_1, A_2, \ldots, A_k$ are *disjoint*, then

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_k) = \Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_k).$$

Theorem (Probability Subtraction Rule (Inclusion–Exclusion))**.** For events A and B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Likewise, there are many other probability rules which have analogues in combinatorics, such as the following.

Theorem (Difference Rule)**.** For events A and B,

$$\Pr(A \setminus B) = \Pr(A) - \Pr(A \cap B).$$

Theorem (Union Bound)**.** For events $A_1, \ldots, A_k$,

$$\Pr(A_1 \cup \cdots \cup A_k) \leq \Pr(A_1) + \cdots + \Pr(A_k).$$

Theorem (Monotonicity)**.** If $A \subseteq B$, then

$$\Pr(A) \leq \Pr(B).$$

Theorem (Complement Rule). If S is the sample space, then for any event A,

$$\Pr(S \setminus A) = 1 - \Pr(A).$$

1.4 Conditional Probability

Conditional probabilities allow us to formalize the idea that we should get a new probability if we know for sure that something happened. The following is the sense in which we can define a new probability for an outcome, assuming some event happened.

Definition 8 (Conditional Probability of an Outcome). *Given discrete probability space* $(S, \Pr)$ *and event* $E \subseteq S$*, the* ***probability of outcome*** $\omega \in E$ *conditioned on* E *is*

$$\Pr_E(\omega) = \frac{\Pr(\omega)}{\Pr(E)}.$$

These new outcome probabilities, in turn, give us a new discrete probability space, as described below.

Definition 9 (Result of Conditioning). *Given discrete probability space* $(S, \Pr)$ *and event* $E \subseteq S$*, we call* $(E, \Pr_E)$ *the* ***result of conditioning on*** E*.*

Theorem. Given discrete probability space $(S, \Pr)$ and event $E \subseteq S$, we have $(E, \Pr_E)$ is a discrete probability space.

Similar to the probability of an outcome, we can define the probability of an event E' conditioned on event E happening.

Definition 10 (Conditional Probability of an Event). *Given discrete probability space* $(S, \Pr)$ *and event* $E \subseteq S$*, the* ***probability of event*** $E' \subseteq S$ *conditioned on* E *is*

$$\Pr(E' \mid E) = \frac{\Pr(E' \cap E)}{\Pr(E)}.$$

The probability of E' conditioned on E is just the probability of E' in the discrete probability space we get when we condition on E. However, since the result of conditioning on E only has outcomes in E, we must take the intersection with E to make this make sense. This is formalized by the following theorem.

Theorem. Given discrete probability space $(S, \Pr)$ and events $E, E' \subseteq S$, we have

$$\Pr(E' \mid E) = \Pr_E(E' \cap E).$$

One nice application of conditional probability is it allows us to understand the "and" of two events (i.e. the intersection) as the probability that one happens times the probability that the other happens assuming the first one happened.

Theorem. Given events A and B we have

$$\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B).$$

Similarly, we can break up the probability of an event into the probability that it happens assuming each event in a partition happens as follows.

Theorem (Law of Total Probability)**.** Given partition $E_1, E_2, \ldots$ of sample space S and event $E \subseteq S$ we have

$$\Pr(E) = \sum_i \Pr(E \mid E_i) \cdot \Pr(E_i).$$

1.5 Independence (of Events)

Two events are independent if knowing one does not change the probability of the other.

Definition 11 (Independent Events)**.** *Event A is independent of event B if* $\Pr(B) = 0$ *or*

$$\Pr(A \mid B) = \Pr(A).$$

The following is a useful equivalent characterization of independent events.

Theorem. A is independent of B iff

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

Independence is symmetric.

Theorem. If A is independent of B then B is independent of A.

Thus, in general we will just talk about two independent events A and B (rather than, say, an event A that is independent of another event B).

Note that if A and B are disjoint, it is not necessarily the case that A and B are independent.

1.6 Random Variables

Random variables give a number for each outcome in our sample space.

Definition 12 (Random Variable)**.** *A **random variable** is a function $X : S \to \mathbb{R}$.*

Two important kinds of random variables are indicator and constant random variables.

Definition 13 (Indicator Random Variable)**.** *An **indicator random variable** is a function $X : S \to \{0, 1\}$.*

An indicator random variable always corresponds to an event and vice versa. In particular, indicator random variable X corresponds to the event $\{\omega \in S : X(\omega) = 1\}$. Event E corresponds to the indicator random variable X such that $X(\omega) = 1$ iff $\omega \in E$.

Definition 14 (Constant Random Variable)**.** *A **constant random variable** is a function $X : S \to \{c\}$ for some $c \in \mathbb{R}$.*

In a slight abuse of notation, we will often notate a constant random variable with the c which it always returns. Similarly, given any function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and random variables X and Y, we will notate the random variable that on ω returns $f(X(\omega), Y(\omega))$ with $f(X, Y)$. For example, $X + Y$ is the random variable which on $\omega \in S$ returns $X(\omega) + Y(\omega)$.

1.7 Expectation

The expected value (or just expectation) of a random variable is a probability-weighted average of its values. That is, if one value is far more likely to occur, we weight it higher in the average.

Definition 15 (Expectation)**.** *The **expectation** of random variable X is*

$$\mathbb{E}[X] = \sum_{\omega \in S} \Pr(\omega) \cdot X(\omega).$$

If a probability space is uniform, then the expectation really is just the average value output but the random variable across all outcomes.

Theorem. If $(S, \Pr)$ is uniform with S finite, then the expectation of random variable X is

$$\mathbb{E}[X] = \frac{\sum_{\omega \in S} X(\omega)}{|S|}.$$

Indicator and constant random variables also have nice expectations as follows.

Theorem. If X is an indicator random variable with corresponding event E then

$$\mathbb{E}[X] = \Pr(E).$$

Theorem. If c is a constant random variable then

$$\mathbb{E}[c] = c.$$

Given a predicate $P(x)$ and random variable X, we let $P(X)$ be the event $\{\omega \in S : P(X(\omega))\}$. For example, $X = i$ is the event consisting of all outcomes ω where $X(\omega) = i$. The following is a useful alternative form of the expectation.

Theorem. For any random variable X we have

$$\mathbb{E}[X] = \sum_i \Pr(X = i) \cdot i.$$

We can prove this using the fact that if X is a random variable, then $\{\, X = i : i \in \mathbb{R} \,\}$ is a partition of S.

1.8 Independence (of Random Variables)

Similar to how events can be independent, random variables can also be independent.

Definition 16 (Independent Random Variables)**.** *A random variable X is* ***independent*** *of a random variable Y if for all i, j we have that the event $X = i$ is independent of the event $Y = j$.*

1.9 Random Variable Distribution

The distribution of a random variable is the function which gives the probability it takes on each value.

Definition 17 (Distribution)**.** *The distribution of a random variable X is the function f_X which given i returns $f_X(i) = \Pr(X = i)$.*

f_X is a **Bernoulli distribution** if X is an indicator random variable. f_X is a **Binomial distribution** if X is the sum of independent indicator random variables (each with the same distributions).

1.10 Conditional Expectation

Using conditional probabilities, we can define conditional expectations which are, roughly, the expected value of the random variable assuming some event E happened.

Definition 18 (Conditional Expectation)**.** *Given event E, the **conditional expectation** of a random variable X is*

$$\mathbb{E}[X \mid E] \;=\; \sum_i \Pr(X = i \mid E) \cdot i.$$

The following shows that the conditional expectation of X conditioned on E really is just the expectation of X in the discrete probability space that results from conditioning on E.

Theorem. Given random variable X and event E we have

$$\mathbb{E}[X \mid E] = \sum_{\omega \in E} \Pr_E(\omega)\, X(\omega).$$

Similar to the Law of Total Probability, we can use conditional expectation to break up an expectation into conditional expectations.

Theorem (Law of Total Expectation)**.** Given random variable X, if $E_1, E_2, \ldots$ is a partition of S then

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid E_i] \cdot \Pr(E_i).$$

1.11 Linearity of Expectation

One of the most important facts about expectation is the fact that it is linear which is to say that it distributed over the addition of random variables and multiplication by reals.

Theorem (Linearity of Expectation)**.** Given random variables $X_1, X_2, \ldots$ and real numbers $a_1, a_2, \ldots$, we have

$$\mathbb{E}\left[\sum_i a_i \cdot X_i\right] = \sum_i a_i \cdot \mathbb{E}[X_i].$$

Importantly, in the above the random variables $X_1, X_2, \ldots$ do not have to be independent.

One nice byproduct of linearity of expectation is that it means that the expected number of events to happen in a collection is just the sum of their probabilities, as follows.

Theorem. Given events $E_1, E_2, \ldots$ with corresponding indicator random variables $I_1, I_2, \ldots$, we have

$$\mathbb{E}\left[\sum_i I_i\right] = \sum_i \Pr(E_i).$$

In general, expectation does not distribute over other things. One notable exception of this, however, is that if it distributes over the product of random variables *if they are independent.*

Theorem. Given independent RVs X, Y have

$$\mathbb{E}[X \cdot Y] \ = \ \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

1.12 Bayes' Rule

Bayes' rule gives a way of "flipping" a conditional probability.

Theorem (Bayes' Rule)**.** For any events A and B we have

$$\Pr(A \mid B) = \Pr(A) \cdot \frac{\Pr(B \mid A)}{\Pr(B)}.$$

Much of what makes Bayes' rule useful is that it allows us to update our probability of A, assuming we just learned that B happened.

1.13 Variance

Sometimes measuring the mean (expectation) of a random variable doesn't give us enough information: it can be helpful to know how much we expect the variable to *stray* from its average.

Definition 19 (Variance)**.** *Given a random variable X, the **variance** of X is*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Unpacking this from the inside out: $X - \mathbb{E}[X]$ is a random variable measuring the distance between X and its mean at each outcome. Averaging the square of this gives us a sense of, overall, how far X tends to be from its mean.

Variance is not linear but it satisfies the following nice properties.

Theorem. For any real number a, the variance satisfies

$$\mathrm{Var}(aX) = a^2\ \mathrm{Var}(X).$$

Theorem. If random variables X and Y are independent, then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Likewise, variance has a common and useful definition in terms of expectation.

Theorem. Variance can be expressed as

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

1.14 Concentration

Concentration is the study of how likely a random variable is to deviate from its expectation. Markov's inequality gives a coarse estimate of the probability that a random variable takes a value much larger than its mean. It requires that the random variable is nonnegative which is to say that it never outputs a negative value.

Theorem (Markov's Inequality)**.** Given a nonnegative random variable X with $\mathbb{E}[X] > 0$, for all $a > 0$ we have

$$\Pr(X \geq a \cdot \mathbb{E}[X]) \leq \frac{1}{a}$$

That is, the probability of X being more than a times its mean is at most $1/a$.

Chebyshev's inequality gives a more refined probability that a random variable deviates from its expectation but requires knowing the variance of the random variable.

Theorem (Chebyshev's Inequality)**.** Given RV X,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$