Discrete Probability: The Very Basics

Probability theory is used to analyze the likelihood of events where the *experiment* (or process, or measurement) in question involves chance or or is large or complex enough that it is overly difficult or impossible for us to analyze completely. Such an experiment could be anything: tossing a coin, counting the number of hairs on a mouse, measuring the distance from the Earth to the moon, pinning down the location of a proton at a given instant, and so on.

Definition. The undetermined *outcome* of an experiment is called a *random variable* and is usually denoted by an upper-case Roman letter from the end of the alphabet such as X or Y. The collection of all possible outcomes of the experiment (*i.e.* possible values of the random variable) is the *sample space* of the experiment, usually denoted by an upper-case Greek letter such as Ω or an upper-case Roman letter such as S.

Our textbook defaults to Ω to denote a sample space if only one is being considered, but the majority of other probability textbooks nowadays default to S. When talking about a generic sample space, we usually denote the individual outcomes by the corresponding lower-case letter, with subscripts if we need to tell them apart.

Examples

1. Toss a two-sided coin. If we ignore the very unlikely possibility that the coin ends up balanced on its edge, the possible outcomes are heads, usually abbreviated H, and tails, usually abbreviated T. The sample space for this experiment is thus $\Omega = \{H, T\}$.
2. Suppose we toss a two-sided coin three times in a row. There are eight possible outcomes (ignoring that pesky edge again :-), which makes the complete sample space be $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
3. Suppose we toss a coin until it comes up heads and then stop tossing. This time there are infinitely many outcomes: $\Omega = \{H, TH, TTH, TTTH, TTTTH, \dots\}$
4. Shuffle a standard 52-card deck[*] and then draw one card. There are 52 possible outcomes, namely the 52 cards, so the sample space is $\Omega = \{A\heartsuit, K\heartsuit, \dots, 3\spadesuit, 2\spadesuit\}$.
5. Roll two standard dice[†] and record the sum of the two faces that came up. In this experiment the sample space is $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.
6. Have an immortal monkey type at a computer keyboard until it hits the \ key or the end of time, whichever comes first. The sample space includes all finite sequences of characters on the keyboard in which there is one and only one \, at the very end. (Yes, this would include all the works of Shakespeare – with a backslash added at the end – as possible outcomes. :-)

[*] A standard 52-card deck has four *suits*: $\heartsuit$, $\diamondsuit$, $\clubsuit$, and $\spadesuit$. Each suit has one card of each of the thirteen *kinds*: A (ace), K (king), Q (queen), and J (jack), as well as cards numbered 10, 9, 8, 7, 6, 5, 4, 3, and 2.

[†] A standard die is a cube with the six faces numbered 1 through 6.

Definition. A random variable and its sample space are said to be *discrete* if the number of possible outcomes is either finite or countably infinite[‡]. All of the examples given above are examples of discrete sample spaces.

We still need to define how likelihood or probability is measured in an experiment. It is conventional in probability theory to measure probabilities with real numbers between 0 and 1. (0 means it certainly won't happen and 1 means it certainly will happen.) In some applications of probability other conventions may be used. For example, percentages are often used to express probabilities in statistics and odds are commonly used to express probabilities in gambling. Thus a probability of $\frac{1}{4} = 0.25$ corresponds to a percentage probability of 25% and odds of $1:3$. To get a probability from a percentage probability simply divide by 100, and odds of $r:s$ correspond to a probability of $\frac{r}{r+s}$.

Definition. Let X be a discrete random variable with sample space $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$. A *probability distribution function* for X is a function $m : \Omega \to \mathbb{R}$ (where $\mathbb{R}$ is the set of real numbers) such that:

1. $m(\omega) \geq 0$ for every outcome $\omega \in \Omega$.
2. $\displaystyle\sum_{\omega\in\Omega} m(\omega) = m(\omega_1) + m(\omega_2) + m(\omega_3) + \cdots = 1$

Probability distribution functions are often referred to simply as probability functions or, like our textbook, as distribution functions.

The definition boils down to saying that no outcome in our experiment can do worse than never happen (*i.e.* have probability 0) and that the experiment must have some outcome (*i.e.* the probabilities of all the outcomes put together is a certainty). Lets see what this could mean in our previous examples.

Examples Revisited

1. Toss a two-sided coin. If the coin is fair – *i.e.* H and T are equally likely outcomes – then the distribution function for this random variable must be $m(H) = m(T) = \frac{1}{2} = 0.5$. On the other hand if the coin were biased, with $m(H) = 0.6$ then we would have to have $m(T) = 0.4 = 1 - 0.6$ because the sum of the probabilities of all the outcome must be 1, and H and T are the only outcomes.
2. Suppose we toss a two-sided coin three times in a row. There are eight possible outcomes, with $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. If the coin is fair, these eight outcomes must be equally likely, so we would have $m(\omega) = \frac{1}{8} = 0.125$ for each outcome $\omega \in \Omega$.
3. Suppose we toss a coin until it comes up heads and then stop tossing, so the sample space is $\Omega = \{H, TH, TTH, TTTH, TTTTH, \dots\}$. If the coin is fair, then the distribution function is given by $m(H) = \frac{1}{2}$, $m(TH) = \frac{1}{4}$, $m(TTH) = \frac{1}{8}$, and so on. (Why?)

[‡] A set is countably infinite if it is infinite but can, in principle, be written out in an infinite list. For example, the positive integers are countably infinite; the usual way to list them is in order: 1, 2, 3, 4, ... Some sets, such the set of real numbers, are too large to be fully listed in such a way. This is one reason we will have to deal with continuous probability all too soon.

4. Shuffle a standard 52-card deck and then draw one card. The 52 cards are the possible outcomes, so the sample space is $\Omega = \{A\heartsuit, K\heartsuit, \dots, 3\spadesuit, 2\spadesuit\}$. Assuming you shuffled well and don't cheat, each card is as likely to be srawn as any other to be drawn. This means that the distribution function must be given by $m(\omega) = \frac{1}{52} \approx 0.01923$.
5. Roll two standard dice and record the sum of the two faces that came up, so the sample space is $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Assuming the dice are fair, so that each face is alikely to come up as any other when we roll one, we would have to have the distribution function be $m(2) = \frac{1}{36}$, $m(3) = \frac{2}{36}$, $m(4) = \frac{3}{36}$, $m(5) = \frac{4}{36}$, $m(6) = \frac{5}{36}$, $m(7) = \frac{6}{36}$, $m(8) = \frac{5}{36}$, $m(9) = \frac{4}{36}$, $m(10) = \frac{3}{36}$, $m(11) = \frac{2}{36}$, and $m(12) = \frac{1}{36}$. (Again, why?)
6. An immortal monkey types at a computer keyboard until it hits the \ key or the end of time, whichever comes first. The sample space includes all finite sequences of characters on the keyboard in which there is one and only one \, at the very end. Assuming the monkey is as likely to hit each key or symbol as any other key or symbol, … [It's a bit of an unholy mess, but in principle it works much like example 3 above. The key – cough, cough – information you still need is how many keys or symbols there are on the key board.]

In real life we are often not that interested in particular outcomes but in groups of them. For example, to have a (possibly barely) winning season a sports team must win more games than it loses. The particular sequence of wins and losses over the season (which would be an outcome of the season) doesn't matter too much; what matters that it is (or is not) one of the possible sequences of wins and losses in which there are more wins than losses. This leads us to the following definition:

Definition. Suppose X is a discrete random variable with sample space Ω and distribution function $m : \Omega \to \mathbb{R}$. An *event* $A \subseteq \Omega$ is a collection of outcomes, *i.e.* A a subset of the sample space. The *probability* of the event A, denoted by $P(A)$, is the likelihood that X will have a value in the event, namely:

$$P(A) = \sum_{\omega \in A} m(\omega) = \text{sum of the probabilities of all the outcomes in the event } A$$

We will usually use upper-case Roman letters from the beginning of the alphabet, A, B, C, and so on, for events.

Some Examples Revisited Again

1. Toss a two-sided coin. If the coin is fair – *i.e.* H and T are equally likely outcomes – then the distribution function for this random variable must be $m(H) = m(T) = \frac{1}{2} = 0.5$. The only possible events are $\emptyset = \{\}$, which has probability 0 (Why?), $\{H\}$, which has probability $\frac{1}{2} = 0.5$, $\{T\}$, which has probability $\frac{1}{2} = 0.5$, and $\{H, T\}$ which has probability 1.
2. Suppose we toss a two-sided coin three times in a row. There are eight possible outcomes, with $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. If the coin is fair, we have $m(\omega) = \frac{1}{8} = 0.125$ for each outcome $\omega \in \Omega$. Suppose B is the event

"Exactly two tails came up." Then $B = \{HTT, THT, TTH\}$ and the probability of B is:

$$P(B) = m(HTT) + m(THT) + m(TTH) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} = 0.375$$

3. Suppose we toss a coin until it comes up heads and then stop tossing, so the sample space is $\Omega = \{H, TH, TTH, TTTH, TTTTH, \dots\}$. If the coin is fair, then the distribution function is given by $m(H) = \frac{1}{2}$, $m(TH) = \frac{1}{4}$, $m(TTH) = \frac{1}{8}$, and so on. Suppose C is the event that the experiment ends after no more than four tosses. Then $C = \{H, TH, TTH, TTTH\}$ and the probability of C is:

$$P(C) = m(H) + m(TH) + m(TTH) + m(TTTH) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16} = 0.9375$$

4. Shuffle a standard 52-card deck and then draw one card. The 52 cards are the possible outcomes, so the sample space is $\Omega = \{A\heartsuit, K\heartsuit, \dots, 3\spadesuit, 2\spadesuit\}$. Assuming you shuffled well and don't cheat, each card is as likely to be srawn as any other to be drawn, so the distribution function is given by $m(\omega) = \frac{1}{52}$. Suppose A is the event that the card is a face card, *i.e.* a K, Q, or J. Then

$$A = \{K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit, Q\heartsuit, Q\diamondsuit, Q\clubsuit, Q\spadesuit, J\heartsuit, J\diamondsuit, J\clubsuit, J\spadesuit\}$$

and the probability of A is:

$$\begin{aligned} P(A) &= m(K\heartsuit) + m(K\diamondsuit) + m(K\clubsuit) + m(K\spadesuit) + m(Q\heartsuit) + m(Q\diamondsuit) \\ &\qquad + m(Q\clubsuit) + m(Q\spadesuit) + m(J\heartsuit) + m(J\diamondsuit) + m(J\clubsuit) + m(J\spadesuit) \\ &= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} \\ &= \frac{12}{52} = \frac{3}{13} \approx 0.23077 \end{aligned}$$

5. Roll two standard dice and record the sum of the two faces that came up, so the sample space is $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Assuming the dice are fair, wthe distribution function is given by $m(2) = \frac{1}{36}$, $m(3) = \frac{2}{36}$, $m(4) = \frac{3}{36}$, $m(5) = \frac{4}{36}$, $m(6) = \frac{5}{36}$, $m(7) = \frac{6}{36}$, $m(8) = \frac{5}{36}$, $m(9) = \frac{4}{36}$, $m(10) = \frac{3}{36}$, $m(11) = \frac{2}{36}$, and $m(12) = \frac{1}{36}$. Suppose B is the event that the sum you get is odd, *i.e.* $B = \{3, 5, 7, 9, 11\}$. Then the probability of B is:

$$P(B) = m(3) + m(5) + m(7) + m(9) + m(11) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36} = \frac{1}{2} = 0.5$$

The following facts about the probabilities of events are pretty easy to get from the definitions we've developed so far:

Theorem. Suppose X is a discrete random variable with sample space Ω and distribution function $m : \Omega \to \mathbb{R}$. Then the following are true:

1. $0 \leq P(A) \leq 1$ for every event $A \subseteq \Omega$, with $P(\emptyset) = 0$ and $P(\Omega) = 1$.
2. If event A contains the event B, *i.e.* $B \subseteq A$, then $P(B) \leq P(A)$.
3. If events A and B are *disjoint* – no outcome is on both A and B – then $P(A \cup B) = P(A) + P(B)$, where the *union* of A and B, $A \cup B$, is the event consisting of all the outcomes in A together with all the outcomes in B.
4. More generally, if $A \cap B$ is the event consisting of all the outcomes that are in both of the events A and B (*i.e.* the *intersection* of A and B), then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5. If $\bar{A}$ is the *complement* of event A, consisting of exactly those outcomes that are not in A, then $P\left(\bar{A}\right) = 1 - P(A)$.
6. More generally, if $A - B$ is the event consisting of all the outcomes in event A that are not in event B, then $P(A - B) = P(A) - P(A \cap B)$.
7. If A and B are any events, then $P(B) = P(B \cap A) + P\left(B \cap \bar{A}\right)$.

See §1.2 of the textbook for the proofs of these statements, as well as some additional variations on these statements. We'll be beating up on – er, using – these ideas in various applications and additional concepts coming soon.