Notes on Probability

Mark Schmidt

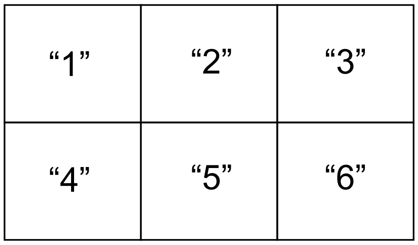
March 25, 2020

1 Probabilites

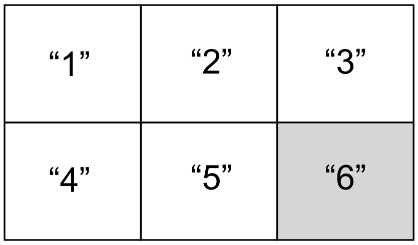
Consider an event A that may or may not happen. For example, if we roll a dice then we may or may not roll a 6. We use the notation $p(A)$ to denote the **probability** of event A happening, which is the likeliness that the event A will actually happen. Probabilities map from events A to a number between 0 and 1,

$$0 \leq p(A) \leq 1,$$

where a value of 0 means “definitely will not happen”, a value of 0.5 means that it happens half of the time, and a value of 1 means “definitely will happen”. It is helpful to think of probabilities as areas that divide up a geometric object. For example, we can represent the dice example with the following diagram:



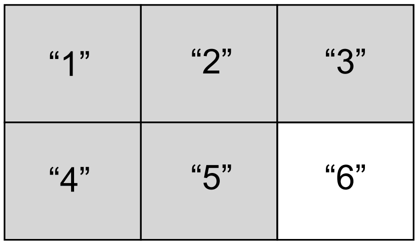
We have set up this figure so that the area associated with each event is proportional to its probability. In this case, each possible value of the dice takes up 1/6 of the area, so we have that $p(6) = 1/6$.



We can use $\neg A$ to represent the event that ‘A does not happen’, and its probability is given by

$$p(\neg A) = 1 - p(A).$$

Thus, the probability of *not* rolling a 6 is given by $1 - 1/6 = 5/6$. From the area figure, we see that all the events where rolling a 6 do not happen correspond to 5/6 of the total area.

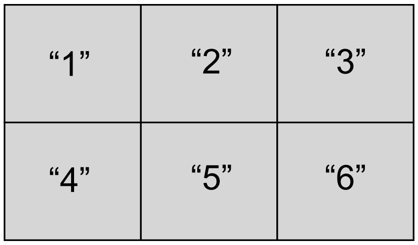


2 Random Variables

A **random variable** X is a variable that takes different values with certain probabilities. We can then consider probabilities of events involving the random variable, such as the event that $X = x$ for a specific value x. We usually use the notation $p(X = x)$ to denote the probability of the even that the random variable X takes the value x. In the dice example, X could be the value that we roll, and in that case we have $p(X = 6) = 1/6$. Often we will simply write $p(x)$ instead of $p(X = x)$, since the random variable is usually obvious from the context. Let's use $\mathcal{X}$ as the set of all possible values that the random variable X might take. In the dice example, this would be the set $\{1, 2, 3, 4, 5, 6\}$. Because the random variable must take some value, we have that the probabilities over all values must sum to one,

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

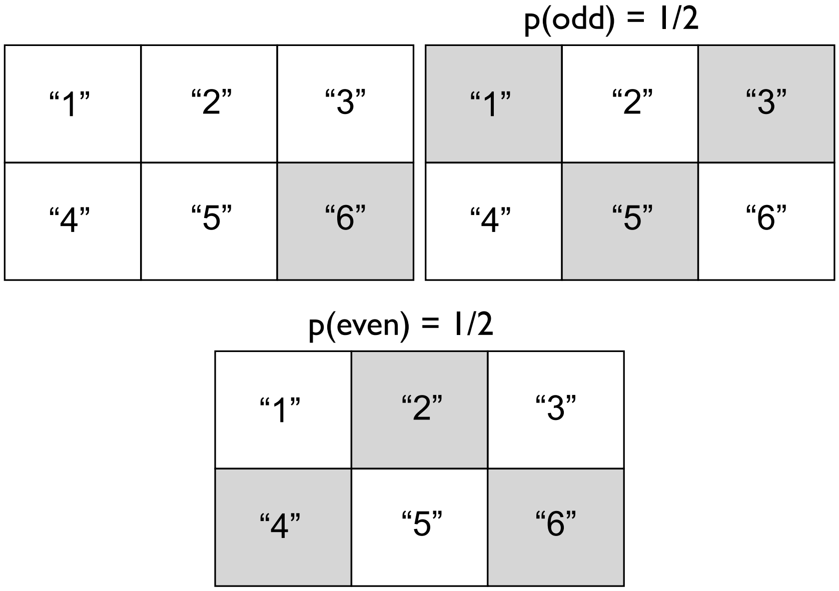
Geometrically, this just means that if we consider all events, that this includes the entire probability space:



In this note, we'll assume that random variables can only take a finite number of possible values. For continuous random variables, we replace sums like these with integrals.

3 Joint Probability

We are often interested in probabilities involving more than one event. For example, if we have two possible events A and B, we might want to know the probability that *both* of them happen. We use the notation $p(A, B)$ to denote the probability of both A and B happening, and we call this the **joint probability**. In terms of areas, this probability is given by the *intersection* of the areas of the two events. For example, consider the probability that we roll a 6 and we roll an odd number, $p(6, \text{odd})$. This probability is zero since the areas where this is true do not intersect.



On the other hand, $p(6, \text{even}) = 1/6$ since the intersection of rolling a 6 with rolling an even number is simply the area associated with rolling a 6. Note that the order of the arguments is not important, so $p(A, B) = p(B, A)$.

An important identity is the **marginalization rule**. This rule says that if we sum the joint probability $p(A, X = x)$ over all possible values x of a random variable X, then we obtain the probability of the event A,

$$p(A) = \sum_{x \in \mathcal{X}} p(A, X = x). \tag{1}$$

For example, the probability of rolling an even number is given by

$$p(\text{even}) = \sum_{i=1}^{6} p(i, \text{even}) = 0 + 1/6 + 0 + 1/6 + 0 + 1/6 = 1/2,$$

which corresponds to adding up all areas where the number is even. If we apply this marginalization rule twice, then we see that the joint probability summed over all values must be equal to one,

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) = \sum_{x \in \mathcal{X}} p(X = x) = 1.$$

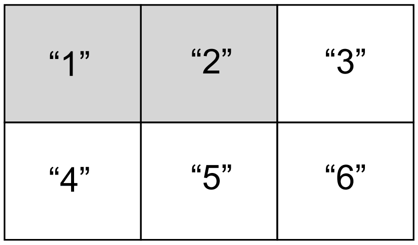
4 Union of Events

Instead of considering the probability of events A and B both occurring, we might instead be interested in the probability of *at least one* of them occurring. This is denoted by $p(A \cup B)$, and in terms of areas corresponds to the union of the areas associated with A and B. This union is given by

$$p(A \cup B) = p(A) + p(B) - p(A, B),$$

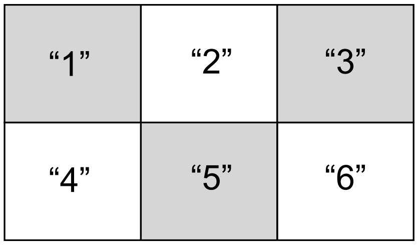
where the last term subtracts the common area that is counted in both $p(A)$ and $p(B)$. For example, the probability of rolling a 1 or a 2 is given by

$$p(1 \cup 2) = p(1) + p(2) - p(1, 2) = 1/6 + 1/6 - 0 = 1/3,$$



Simiarly, the probability of rolling a 1 or an odd number is given by

$$p(1 \cup \text{odd}) = p(1) + p(\text{odd}) - p(1, \text{odd}) = 1/6 + 1/2 - 1/6 = 1/2.$$



5 Conditional Probability

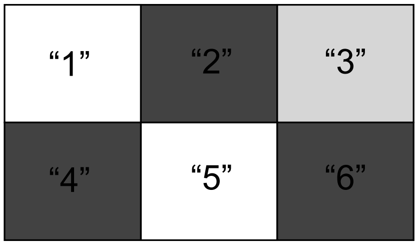
We are often interested in the probability of an event A, *given that we know an event B occurred.* This is called the **conditional probability** and it is denoted by $p(A \mid B)$. Viewed from the perspective of areas, this is the area of A restricted to the region where B happened, divided by the total area taken up by B. Mathematically, this gives

$$p(A \mid B) = \frac{p(A, B)}{p(B)}, \tag{2}$$

where we have $p(B) \neq 0$ since it happened. For example, the probability of rolling a 3 given that you rolled an odd number is given by

$$p(3 \mid \text{odd}) = \frac{p(3, \text{odd})}{p(\text{odd})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Geometrically, we remove the area associated with numbers that are not odd, and compute the area of where the event happened divided by the total area that is left:



Observe that conditional probabilities sum up to one when we sum over the left variable,

$$\sum_{x \in \mathcal{X}} p(x \mid B) = \sum_{x \in \mathcal{X}} \frac{p(x, B)}{p(B)} = \frac{1}{p(B)} \sum_{x \in \mathcal{X}} p(x, B) = \frac{p(B)}{p(B)} = 1,$$

where we have used the marginalization rule (1). If we sum over the conditioning variable B it does not need to sum up to one,

$$\sum_{x \in \mathcal{X}} p(A \mid x) \neq 1,$$

in general.

6 Product Rule and Bayes Rule

By re-arranging the conditional probability inequality, we obtain the **product rule**,

$$p(A,B) = p(A \mid B)p(B),$$

and similarly

$$p(A,B) = p(B \mid A)p(A).$$

This lets us express joint probabilities (which can be hard to deal with) in terms of conditional probabilities (which are often easier to deal with). It also gives a variation on the marginalization rule (1),

$$p(A) = \sum_{x \in \mathcal{X}} p(A, X = x) = \sum_{x \in \mathcal{X}} p(A \mid X = x)p(X = x).$$

By applying the product rule in the definition of conditional probability (2), we obtain **Bayes rule**,

$$p(A \mid B) = \frac{p(A,B)}{p(B)} = \frac{p(B \mid A)p(A)}{p(B)}.$$

This lets us express the conditional probability of A given B in terms of the reverse conditional probability (of B given A). We sometimes also write Bayes rule using the notation

$$p(A \mid B) \propto p(B \mid A)p(A),$$

where the '$\propto$' sign means that the values are equal up to a constant value that makes the conditional probabilities sum up to one over all values of A. Another form of Bayes rule that you often see comes from applying the marginalization rule (1) and then the product rule to $p(B)$,

$$p(x \mid B) = \frac{p(B \mid x)p(x)}{p(B)} = \frac{p(B \mid x)p(x)}{\sum_{x \in \mathcal{X}} p(B,x)} = \frac{p(B \mid x)p(x)}{\sum_{x \in \mathcal{X}} p(B \mid x)p(x)}.$$

7 Conditional Probabilities with More Than 2 Variables

We can also define conditional probabilities involving more than two events. We use the notation $p(A, B \mid C)$ to denote the probability of both A and B happening, given that C happened,

$$p(A, B \mid C) = \frac{p(A,B,C)}{p(C)}.$$

We similarly use the notation $p(A \mid B, C)$ to denote the probability that A happens, given that both B and C happened,

$$p(A \mid B, C) = \frac{p(A,B,C)}{p(B,C)}.$$

8 Conditional Probability Identities

If we use C to denote the event that we are conditioning on, then if we keep C the right side of the conditioning bar then all of the identities above generalize to conditional probabilities. Conceptually, we're just redefining the total probability area as the area where C happened. For example, the marginalization rule (1) is changed to

$$\sum_{x \in \mathcal{X}} p(A, x \mid C) = p(A \mid C),$$

the union of events is change to

$$p(A \cup B \mid C) = p(A \mid C) + p(B \mid C) - p(A, B \mid C),$$

the product rule is changed to

$$p(A, B \mid C) = p(A \mid B, C)p(B \mid C),$$

Bayes rule is changed to

$$p(A \mid B, C) = \frac{p(B \mid A, C)p(A \mid C)}{p(B \mid C)},$$

and so on.

9 Independence and Conditional Independence

We say that two events are **independent** if their joint probability equals the product of their individual probabilities,

$$p(A, B) = p(A)p(B).$$

In this case we use the notation $A \perp B$. Two random variables are independent if this is true for all values that the random variables can take.

By using the product rule, we see that two variables are independent iff

$$p(A)p(B) = p(A, B) = p(A \mid B)p(B),$$

or equivalently that

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

This means that knowing that B happened tells us nothing about the probability of A happening, and vice versa.

A generalization of independence is **conditional independence**, where we consider independence given that we know a third event C occurred,

$$p(A, B \mid C) = p(A \mid C)p(B \mid C),$$

and in this case we use the notation $A \perp B \mid C$. Conditional independence can be much weaker than marginal independence, and we often make use of it to model high-dimensional probability distributions.

10 Expectation

If we have a random variable X that can take values $x \in \mathcal{X}$, we define the **expectation** of X or $\mathbb{E}[X]$ by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(X = x)x,$$

where if X is continuous we again replace the sum with an integral. In the case of rolling a dice, we have $p(X = x) = 1/6$ for all x so we have

$$\mathbb{E}[X] = \sum_{x=1}^{6} p(X = x)x = \sum_{i=1}^{6} \frac{x}{6} = 3.5.$$

We can think of the expectation as a generalization of the notion of an *average*. If the probabilities are uniform then the expectation is the average over values of possible X, but the expectation can also give us the “average” value we expect to see if the probabilities are non-uniform. For example, if we flip a coin that lands heads ($x = 1$) 75% of the time and lands tails ($x = 0$) 25% of the time, then the expectation is

$$\mathbb{E}[X] = (0.75)(1) + (0.25)(0) = 0.75,$$

which is a weighted average of the possibilities that reflects their probabilities of actually happening.

We can also talk about the *expected value of a function* f that depends on a random variable,

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} p(X = x) f(X).$$

Because f depends on a random variable, we call f a random function. Because the expectation is just a sum, *expectation is a linear operator* meaning that if α and β are (non-random) scalar values while f and g are functions then we have

$$\mathbb{E}[\alpha f(X) + \beta g(X)] = \alpha \mathbb{E}[f(X)] + \beta \mathbb{E}[g(X)].$$

However, note that in general we may have $\mathbb{E}[f(X)g(X)] \neq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$. We define the *conditional expectation* of a random variable X given the value of another random variable $Y = y$ by

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} p(X = x \mid Y = y) f(X).$$

Notice that this is a function of Y but doesn't actually depend on X.

An important property of conditional expectations is that

$$\mathbb{E}[\mathbb{E}[X = x \mid Y = y]] = \mathbb{E}[X],$$

where the outer expectation is taken with respect to Y. This is called the *tower property*, *law of total expectation*, or the *iterated expectation* rule. It words it says that the expected value of X, averaged over all possible values that Y that we could condition on, is still the expected value of X.