Contents

CHAPTER 1

Random Variables and Expectations

1.1 RANDOM VARIABLES

Quite commonly, we would like to deal with numbers that are random. We can do so by linking numbers to the outcome of an experiment. We define a **random variable**:

Definition: *Discrete random variable*

Given a sample space Ω, a set of events $\mathcal{F}$, and a probability function P, and a countable set of of real numbers D, a discrete random variable is a function with domain Ω and range D.

This means that for any outcome ω there is a number $X(\omega)$. P will play an important role, but first we give some examples.

Example: *Numbers from coins*

We flip a coin. Whenever the coin comes up heads, we report 1; when it comes up tails, we report 0. This is a random variable.

Example: *Numbers from coins II*

We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable.

Example: *The number of pairs in a poker hand*

(from Stirzaker). We draw a hand of five cards. The number of pairs in this hand is a random variable, which takes the values $0, 1, 2$ (depending on which hand we draw)

A function of a discrete random variable is also a discrete random variable.

Example: *Parity of coin flips*

We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable. The parity of this number is also a random variable.

Associated with any value x of the random variable X is an event — the set of outcomes such that $X = x$, which we can write $\{X = x\}$; it is sometimes written as $\{\omega : X(\omega) = x\}$. The probability that X takes the value x is given by $P(\{X = x\})$. This is sometimes written as $P(X = x)$, and rather often written as $P(x)$.

Definition: *The probability distribution of a discrete random variable*

The probability distribution of a discrete random variable is the set of numbers $P(X = x)$ for each value x that X can take. The distribution takes the value 0 at all other numbers. Notice that this is non-negative.

Definition: *The cumulative distribution of a discrete random variable*

The cumulative distribution of a discrete random variable is the set of numbers $P(X <= x)$ for each value x that X can take. Notice that this is a non-decreasing function of x. Cumulative distributions are often written with an f, so that $f(x)$ might mean $P(X <= x)$.

Worked example 1.1 *Numbers from coins III*

We flip a biased coin 2 times. The flips are independent. The coin has $P(H) = p$, $P(T) = 1 - p$. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 2 bit random number, which is a random variable. What is the probability distribution and cumulative distribution of this random variable?

Solution: Probability distribution: $P(0) = (1 - p)^2$; $P(1) = (1 - p)p$; $P(2) = p(1 - p)$; $P(3) = p^2$. Cumulative distribution: $f(0) = (1 - p)^2$; $f(1) = (1 - p)$; $f(2) = p(1 - p) + (1 - p) = (1 - p^2)$; $f(3) = 1$.

Worked example 1.2 *Betting on coins*

One way to get a random variable is to think about the reward for a bet. We agree to play the following game. I flip a coin. The coin has $P(H) = p$, $P(T) = 1 - p$. If the coin comes up heads, you pay me $\$q$; if the coin comes up tails, I pay you $\$r$. The number of dollars that change hands is a random variable. What is its probability distribution?

Solution: We see this problem from my perspective. If the coin comes up heads, I get $\$q$; if it comes up tails, I get $-\$r$. So we have $P(X = q) = p$ and $P(X = -r) = (1 - p)$, and all other probabilities are zero.

1.1.1 Joint and Conditional Probability for Random Variables

All the concepts of probability that we described for events carry over to random variables. This is as it should be, because random variables are really just a way of getting numbers out of events. However, terminology and notation change a bit.

Assume we have two random variables X and Y. The probability that X takes the value x and Y takes the value y could be written as $P(\{X = x\} \cap \{Y = y\})$. It is more usual to write it as $P(x, y)$. You can think of this as a table of values, one for each possible pair of x and y values. This table is usually referred to as the **joint probability distribution** of the random variables. Nothing (except notation) has really changed here, but the change of notation is useful.

We will simplify notation further. Usually, we are interested in random variables, rather than potentially arbitrary outcomes or sets of outcomes. We will write $P(X)$ to denote the probability distribution of a random variable, and $P(x)$ or $P(X = x)$ to denote the probability that that random variable takes a particular value. This means that, for example, the rule we could write as

$$P(\{X = x\} \mid \{Y = y\})P(\{Y = y\}) = P(\{X = x\} \cap \{Y = y\})$$

will be written as

$$P(x|y)P(y) = P(x, y).$$

This yields **Bayes' rule**, which is important enough to appear in its own box.

Definition: *Bayes' rule*

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Random variables have another useful property. If $x_0 \neq x_1$, then the event $\{X = x_0\}$ must be disjoint from the event $\{X = x_1\}$. This means that

$$\sum_x P(x) = 1$$

and that, for any y,

$$\sum_x P(x|y) = 1$$

(if you're uncertain on either of these points, check them by writing them out in the language of events).

Now assume we have the joint probability distribution of two random variables, X and Y. Recall that we write $P(\{X = x\} \cap \{Y = y\})$ as $P(x, y)$. Now consider the sets of outcomes $\{Y = y\}$ for each different value of y. These sets must be disjoint, because y cannot take two values at the same time. Furthermore, each element of the set of outcomes $\{X = x\}$ must lie in one of the sets $\{Y = y\}$. So we have

$$\sum_y P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})$$

which is usually written as

$$\sum_y P(x, y) = P(x)$$

and is often referred to as the **marginal probability** of X.

Definition: *Independent random variables*

The random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$) are independent. This means that

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

which we can rewrite as

$$P(x, y) = p(x)p(y)$$

Worked example 1.3 *Sums and differences of dice*
You throw two dice. The number of spots on the first die is a random variable (call it X); so is the number of spots on the second die (Y). Now define $S = X + Y$ and $D = X - Y$.

- What is the probability distribution of S?
- What is the probability distribution of D?
- What is their joint probability distribution?
- Are X and Y independent?
- Are S and D independent?
- What is $P(S|D = 0)$?
- What is $P(D|S = 11)$?

Solution:

- S can have values in the range $2, \ldots, 12$. The probabilities are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$.
- D can have values in the range $-5, \ldots, 5$. The probabilities are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$.
- This is more interesting to display, because it's an 11x11 table. See Table 1.4.5.
- Yes
- No — one way to check this is to notice that the rank of the table, as a matrix, is 6, which means that it can't be the outer product of two vectors. Also, notice that if you know that (say) $S = 2$, you know the value of D precisely.
- You could work it out from the table, or by first principles. In this case, S can have values $2, 4, 6, 8, 10, 12$, and each value has probability $1/6$.
- You could work it out from the table, or by first principles. In this case, D can have values $1, -1$, and each value has probability $1/12$.

1.1.2 Just a Little Continuous Probability

Our random variables take values from a discrete set of numbers D. This makes the underlying machinery somewhat simpler to describe, and is often, but not always, enough for model building. Some phenomena are more naturally modelled as being continuous — for example, human height; human weight; the mass of a distant star; and so on. Giving a complete formal description of probability on a continuous space is surprisingly tricky, and would involve us in issues that do not arise much in practice.

$$\frac{1}{36} \times \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

TABLE 1.1: *A table of the joint probability distribution of S (vertical axis; scale $2, \ldots, 12$) and D (horizontal axis; scale $-5, \ldots, 5$) from example 3*

These issues are caused by two interrelated facts: real numbers have infinite precision; and you can't count real numbers. A continuous random variable is still a random variable, and comes with all the stuff that a random variable comes with. We will not speculate on what the underlying sample space is, nor on the underlying events. This can all be sorted out, but requires moderately heavy lifting that isn't particularly illuminating for us. The most interesting thing for us is specifying the probability distribution. Rather than talk about the probability that a real number takes a particular value (which we can't really do satisfactorily most of the time), we will instead talk about the probability that it lies in some interval. So we can specify a probability distribution for a continuous random variable by giving a set of (very small) intervals, and for each interval providing the probability that the random variable lies in this interval.

The easiest way to do this is to supply a **probability density function**. Let $p(x)$ be a probability density function for a continuous random variable X. We interpret this function by thinking in terms of small intervals. Assume that dx is an infinitesimally small interval. Then

$$p(x)dx = P(\{\text{event that } X \text{ takes a value in the range } [x, x+dx]\}).$$

This immediately implies some important facts about probability density functions. First, they are non-negative. Second, we must have that

$$P(\{\text{event that } X \text{ takes a value in the range } [-\infty, \infty]\}) = 1$$

which implies that if we sum $p(x)dx$ over all the infinitesimal intervals between $-\infty$ and ∞, we will get 1. This sum is, fairly clearly, an integral. So we have

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$

Notice also that for $a < b$

$$P(\{\text{event that } X \text{ takes a value in the range } [a, b]\}) = \int_a^b p(x)dx.$$

Probability density functions can be moderately strange functions. There is some (small!) voltage over the terminals of a warm resistor caused by noise (electrons moving around in the heat and banging into one another). This is a good example of a continuous random variable, and we can assume there is some probability density function for it, say $p(x)$. Now imagine I define a new random variable by the following procedure: I flip a coin; if it comes up heads, I report 0; if tails, I report the voltage over the resistor. This random variable has a probability $1/2$ of taking the value 0, and $1/2$ of taking a value from $p(x)$. Write this random variable's probability density function $q(x)$. Then $q(x)$ has the property that

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} q(x) dx = 1/2,$$

which means that $q(x)$ is displaying quite unusual behavior at $x = 0$. We will not need to deal with probability density functions that behave like this; but you should be aware of the possibility.

Every probability density function $p(x)$ has the property that $\int_{-\infty}^{\infty} p(x) dx = 1$; this is useful, because when we are trying to determine a probability density function, we can ignore a constant factor. So if $g(x)$ is a non-negative function that is proportional to the probability density function (often pdf) we are interested in, we can recover the pdf by computing

$$p(x) = \frac{1}{\int_{-\infty}^{\infty} g(x) dx} g(x).$$

This procedure is sometimes known as **normalizing**, and $\int_{-\infty}^{\infty} g(x) dx$ is the **normalizing constant**.

Notice that, while a pdf has to be non-negative, and it has to integrate to 1, it does *not* have to be smaller than one. The pdf doesn't represent the probability that a random variable takes a value — it wouldn't be useful if it did, because real numbers have infinite precision. Instead, you should think of $p(x)$ as being the limit of a ratio (which is why it's called a density):

$$\frac{\text{the probability that the random variable will lie in a small interval centered on } x}{\text{the length of the small interval centered on } x}$$

A ratio like this could be a lot larger than one, as long as it isn't larger than one for too many x (because the integral must be one).

Another good way to think about pdf's is as the limit of a histogram. Imagine you collect an arbitrarily large dataset of data items, each of which is independent. You build a histogram of that dataset, using arbitrarily narrow boxes. You scale the histogram so that the sum of the box areas is one. The result is a probability density function.

It is quite usual to write all pdf's as lower-case p's (and if one specifically wishes to refer to probability, to write an upper case P).

1.1.3 Expectations and Expected Values

Imagine we play the game of example 2 multiple times. Our frequency definition of probability means that in N games, we expect to see about pN heads and $(1-p)N$

tails. In turn, this means that my total income from these N games should be about $(pN)q - ((1-p)N)r$. The N in this expression is inconvenient; instead, we could say that for any single game, my income is

$$pq - (1-p)r.$$

This isn't the actual income from a single game (which would be either q or $-r$, depending on what the coin did). Instead, it's an estimate of what would happen over a large number of games, on a per-game basis. This is an example of an expected value.

Definition: *Expected value*

Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

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

This is sometimes written which is $\mathbb{E}_P[X]$, to clarify which distribution one has in mind

Notice that an expected value could take a value that the random variable doesn't take.

Example: *Betting on coins*

We agree to play the following game. I flip a fair coin ($P(H) = P(T) = 1/2$). If the coin comes up heads, you pay me 1\$; if the coin comes up tails, I pay you 1\$. The expected value of my income is 0\$, even though the random variable never takes that value.

Definition: *Expectation*

Assume we have a function f that maps a discrete random variable X into a set of numbers $\mathcal{D}_f$. Then $f(x)$ is a discrete random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f] = \sum_{u \in \mathcal{D}_f} uP(F = u) = \sum_{x \in \mathcal{D}} f(x)P(X = x)$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".

Expectations are linear, so that $\mathbb{E}[0] = 0$ and $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$. The expectation of a constant is that constant (or, in notation, $\mathbb{E}[k] = k$), because probabilities sum to 1. Because probabilities are non-negative, the expectation of a non-negative random variable must be non-negative.

1.1.4 Expectations for Continuous Random Variables

We can compute expectations for continuous random variables, too, though summing over all values now turns into an integral. This should be expected. Imagine you choose a set of closely spaced values for x which are x_i, and then think about x as a discrete random variable. The values are separated by steps of width Δx. Then the expected value of this discrete random variable is

$$\mathbb{E}[X] = \sum_i x_i P(X \in \text{interval centered on } x_i) = \sum_i x_i p(x_i)\Delta x$$

and, as the values get closer together and Δx gets smaller, the sum limits to an integral.

Definition: *Expected value*

Given a continuous random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \int_{x \in \mathcal{D}} x p(x) dx.$$

This is sometimes written which is $\mathbb{E}_p[X]$, to clarify which distribution one has in mind

The expected value of a continuous random variable could be a value that the random variable doesn't take, too. Notice one attractive feature of the $\mathbb{E}[X]$ notation; we don't need to make any commitment to whether X is a discrete random variable (where we would write a sum) or a continuous random variable (where we would write an integral). The reasoning by which we turned a sum into an integral works for functions of continuous random variables, too.

Definition: *Expectation*

Assume we have a function f that maps a continuous random variable X into a set of numbers $\mathcal{D}_f$. Then $f(x)$ is a continuous random variable, too, which we write F. The expected value of this random variable is

$$\mathbb{E}[f] = \int_{x \in \mathcal{D}} f(x) p(x) dx$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".

Again, for continuous random variables, expectations are linear, so that $\mathbb{E}[0] = 0$ and $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$. The expectation of a constant is that constant (or, in notation, $\mathbb{E}[k] = k$), because probabilities sum to 1. Because probabilities are non-negative, the expectation of a non-negative random variable must be non-negative.

1.1.5 Mean, Variance and Covariance

There are three very important expectations with special names.

Definition: *The mean or expected value*

The mean or expected value of a random variable X is

$$\mathbb{E}[X]$$

Definition: *The variance*

The variance of a random variable X is

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

Notice that

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

Definition: *The covariance*

The covariance of two random variables X and Y is

$$\mathsf{cov}\,(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Notice that

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

We also have $\mathsf{Var}[X] = \mathsf{cov}\,(X, X)$.

Now assume that we have a probability distribution $P(X)$ defined on some discrete set of numbers. There is some random variable that produced this probability distribution. This means that we could talk about the mean of a probability distribution P (rather than the mean of a random variable whose probability distribution is $P(X)$). It is quite usual to talk about the mean of a probability distribution. Furthermore, we could talk about the variance of a probability distribution P (rather than the variance of a random variable whose probability distribution is $P(X)$).

Worked example 1.4 *variance*

Can a random variable have $\mathbb{E}[X] > \sqrt{\mathbb{E}[X^2]}$?

Solution: No, because that would mean that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] < 0$. But this is the expected value of a non-negative quantity; it must be non-negative.

Worked example 1.5 *More variance*

We just saw that a random variable can't have $\mathbb{E}[X] > \sqrt{\mathbb{E}[X^2]}$. But I can easily have a random variable with large mean and small variance - isn't this a contradiction?

Solution: No, you're confused. Your question means you think that the variance of X is given by $\mathbb{E}\left[X^2\right]$; but actually $\mathsf{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$

Worked example 1.6 *Mean of a coin flip*

We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the mean of X? (i.e. $\mathbb{E}[X]$).

Solution: $\mathbb{E}[X] = \sum_{x \in D} x P(X = x) = 1p + 0(1-p) = p$

Useful facts: *Expectations*

1. $\mathbb{E}[0] = 0$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. $\mathbb{E}[kX] = k\mathbb{E}[X]$
4. $\mathbb{E}[1] = 1$
5. if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
6. if X and Y are independent, then $\mathsf{cov}\,(X, Y) = 0$.

All but 5 and 6 are obvious from the definition. For 5, recall that $\mathbb{E}[X] = \sum_{x \in D} xP(X = x)$, so that

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{(x,y)\in D_x \times D_y} xyP(X = x, Y = y) \\
&= \sum_{x \in D_x} \sum_{y \in D_y} (xyP(X = x, Y = y)) \\
&= \sum_{x \in D_x} \sum_{y \in D_y} (xyP(X = x)P(Y = y)) \\
&\quad \text{because } X \text{ and } Y \text{ are independent} \\
&= \left(\sum_{x \in D_x} xP(X = x)\right) \left(\sum_{y \in D_y} yP(Y = y)\right) \\
&\quad \text{because } X \text{ and } Y \text{ are independent} \\
&= (\mathbb{E}[X])(\mathbb{E}[Y]).
\end{aligned}$$

Notice that 5 is certainly not true when X and Y are not independent (try $Y = -X$). 6 follows from 5.

Useful facts: *Variance*

It is quite usual to write $\mathsf{Var}[X]$ for the variance of the random variable X.

1. $\mathsf{Var}[0] = 0$
2. $\mathsf{Var}[1] = 0$
3. $\mathsf{Var}[X] \geq 0$
4. $\mathsf{Var}[kX] = k^2\mathsf{Var}[X]$
5. if X and Y are independent, then $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$

1, 2, 3 are obvious. 4 follows because

$$\begin{aligned}
\mathsf{Var}[kX] &= \mathbb{E}\left[(kX - \mathbb{E}[kX])^2\right] && (1.1)\\
&= \mathbb{E}\left[(kX - k\mathbb{E}[X])^2\right] && (1.2)\\
&= \mathbb{E}\left[k^2(X - \mathbb{E}[X])^2\right] && (1.3)\\
&= k^2\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] && (1.4)\\
&= k^2\mathsf{Var}[X]. && (1.5)
\end{aligned}$$

5 follows because

$$\begin{aligned}
\mathsf{Var}[X+Y] &= \mathbb{E}\left[(X + Y - \mathbb{E}[X+Y])^2\right]\\
&= \mathbb{E}\left[(X + Y - \mathbb{E}[X] - \mathbb{E}[Y])^2\right]\\
&= \mathbb{E}\left[(X - \mathbb{E}[X])^2 + (Y - \mathbb{E}[Y])^2 + (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]\\
&= \mathsf{Var}[X] + \mathsf{Var}[Y] + \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\\
&= \mathsf{Var}[X] + \mathsf{Var}[Y] + \mathbb{E}[XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y]]\\
&= \mathsf{Var}[X] + \mathsf{Var}[Y] + \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\\
&= \mathsf{Var}[X] + \mathsf{Var}[Y] + \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y]\\
&\quad \text{because } X \text{ and } Y \text{ are independent}\\
&= \mathsf{Var}[X] + \mathsf{Var}[Y]
\end{aligned}$$

Worked example 1.7 *Variance of a coin flip*

We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the variance of X? (i.e. $\mathsf{Var}[X]$).

Solution: $\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = (1p - 0(1-p)) - p^2 = p(1-p)$

The variance of a random variable is often inconvenient, because its units are the square of the units of the random variable. Instead, we could use the **standard deviation**.

Definition: *Standard deviation*

The **standard deviation** of a random variable X is defined as

$$\mathsf{sd}\left(\{X\}\right) = \sqrt{\mathsf{Var}[X]}$$

You do need to be careful with standard deviations. If X and Y are independent random variables, then $\mathsf{Var}[X+Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$, but $\mathsf{sd}\left(\{X+Y\}\right) = \sqrt{\mathsf{sd}\left(\{X\}\right)^2 + \mathsf{sd}\left(\{Y\}\right)^2}$. One way to avoid getting mixed up is to remember that variances add, and derive expressions for standard deviations from that.

It is sometimes convenient when working with random variables to use **indicator functions**. This is a function that is one when some condition is true, and zero otherwise. The reason they are useful is that their expectated values have interesting properties.

Definition: *Indicator functions*

An indicator function for an event is a function that takes the value zero for values of X where the event does not occur, and one where the event occurs. For the event $\mathcal{E}$, we write

$$\mathbb{I}_{[\mathcal{E}]}(X)$$

for the relevant indicator function.

For example,

$$\mathbb{I}_{[\{|X|\} \leq a]}(X) = \begin{cases} 0 & \text{if } -a < X < a \\ 1 & \text{otherwise} \end{cases}$$

Indicator functions have one useful property.

$$\mathbb{E}\left[\mathbb{I}_{[\mathcal{E}]}\right] = P(\mathcal{E})$$

which you can establish by checking the definition of expectations.

1.1.6 Expectations from Simulation

One remarkably easy way to estimate expectations is by simulation. The reasoning is a straightforward extension of the way we used simulations to estimate probabilities. Recall we thought of a probability as a frequency. This means that for an event $\mathcal{E}$, $P(\mathcal{E}) = p$ if, in a very large number N of independent experiments, $\mathcal{E}$ occurs about pN times. We then argued that one could simulate independent experiments easily, the way to estimate p was to run a lot of experiments and count.

This argument works for expectations as well. Imagine we have a discrete random variable X that takes values in some domain D. Assume that we can easily produce indendent simulations, and that we wish to know $\mathbb{E}[f]$. We now run N independent simulations. The i'th simulation produces the value x_i for X. Then a very good estimate of $\mathbb{E}[f]$ is

$$\frac{\sum_{i=1}^{N} f(x_i)}{N}.$$

I will not prove this rigorously, but there is an easy argument that suggests the expression must be right. X takes values in the domain D. Choose some value x in that domain — this value should appear in our set of simulation outputs about $Np(x)$ times. This means that

$$\frac{\sum_{i=1}^{N} f(x_i)}{N} \approx \frac{\sum_{x \in D} f(x)(Np(x))}{N} = \mathbb{E}[f].$$

All the points I made about simulation in the previous chapter apply to this idea, too. The value that each block of N experiments estimates for $\mathbb{E}[f]$ is a random variable. You can get some idea of the accuracy of your estimate by running several blocks of experiments, then computing the mean and standard deviation of the results. The dataset that you produce will almost always look like a normal distribution. And sometimes it can be hard to produce independent simulations.

Worked example 1.8 *MTGDAF — How long before you can play a spell of cost 3?*

Assume you have a deck of 15 Lands, 15 Spells of cost 1, 14 Spells of cost 2, 10 Spells of cost 3, 2 Spells of cost 4, 2 Spells of cost 5, and 2 Spells of cost 6. What is the expected number of turns before you can play a spell of cost 3? Assume you always play a land if you can.

Solution: I get 6.3, with a standard deviation of 0.1. The problem is it can take quite a large number of turns to get three lands out. I used the code of listings 1.1 and 1.2

Listing 1.1: Matlab code used to estimate number of turns before you can play a spell of cost 3

```
simcards=[zeros(15, 1); ones(15, 1);...
    2*ones(14, 1);3*ones(10, 1); ...
    4*ones(2, 1); 5*ones(2, 1); 6*ones(2, 1)];
nsims=10;
ninsim=1000;
counts=zeros(nsims, 1);
for i=1:nsims
    for j=1:ninsim
        % draw a hand
        shuffle=randperm(60);
        hand=simcards(shuffle(1:7));
        %reorganize the hand
        cleanhand=zeros(7, 1);
        for k=1:7
            cleanhand(hand(k)+1)=cleanhand(hand(k)+1)+1;
            % ie count of lands, spells, by cost
        end
        landsontable=0;
        k=0; played3spell=0;
        while played3spell==0;
            [played3spell, landsontable, cleanhand]=...
                play3round(landsontable, cleanhand, shuffle, ...
                simcards, k+1);
            k=k+1;
        end
        counts(i)=counts(i)+k;
    end
   counts(i)=counts(i)/ninsim;
end
```

1.2 SOME PROBABILITY DISTRIBUTIONS

1.2.1 The Geometric Distribution

We have a biased coin. The probability it will land heads up, $P(\{H\})$ is given by p. We flip this coin until the first head appears. The number of flips required is a discrete random variable which takes integer values greater than or equal to one, which we shall call X.

To get n flips, we must have $n-1$ tails followed by 1 head. This event has probability $(1-p)^{(n-1)}p$. This means the probability distribution for this random variable is

$$P(\{X=n\}) = (1-p)^{(n-1)}p.$$

for $0 \le p \le 1$ and $n \ge 1$; for other n it is zero. This probability distribution is

Listing 1.2: Matlab code used to simulate a turn to estimate the number of turns before you can play a spell of cost 3

```
function [played3spell, landsontable, cleanhand]=...
    play3round(landsontable, cleanhand, shuffle, simcards, ...
    turn)
        % draw
ncard=simcards(shuffle(7+turn));
cleanhand(ncard+1)=cleanhand(ncard+1)+1;
% play land
if cleanhand(1)>0
     landsontable=landsontable+1;
     cleanhand(1)=cleanhand(1)-1;
end
played3spell=0;
if (landsontable>=3)&&(cleanhand(4)>0)
    played3spell=1;
end
```

known as the **geometric distribution**.

Worked example 1.9 *Geometric distribution*

Show that the geometric distribution is non-negative and sums to one (and so is a probability distribution).

Solution: Recall that for $0 < r < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}.$$

So

$$\begin{aligned}
\sum_{n=1}^{\infty} P(\{X = n\}) &= p\sum_{n=1}^{\infty}(1-p)^{(n-1)} \\
&= p\sum_{i=0}^{\infty}(1-p)^{i}\text{just reindexing the sum} \\
&= p\frac{1}{1-(1-p)} \\
&= 1
\end{aligned}$$

Useful facts: *The geometric distribution*

1. The mean of the geometric distribution is $\frac{1}{p}$.
2. The variance of the geometric distribution is $\frac{1-p}{p^2}$.

Proofs: *The geometric distribution - lemmas*

Three results are helpful. First,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}.$$

Second,

$$\begin{aligned}
\sum_{i=0}^{\infty} i r^i &= (\sum_{i=1}^{\infty} r^i) + r(\sum_{i=1}^{\infty} r^i) + r^2(\sum_{i=1}^{\infty} r^i) + \ldots \\
&= \frac{r}{1-r}\left(\sum_{i=0}^{\infty} r^i\right) \\
&= \frac{r}{(1-r)^2}.
\end{aligned}$$

Now write $\gamma = \sum_{i=0}^{\infty} r^i$. Third,

$$\begin{aligned}
\sum_{i=0}^{\infty} i^2 r^i &= (\gamma - 1) + 3r(\gamma - 1) + 5r^2(\gamma - 1) + \ldots \\
&= (\gamma - 1)\sum_{i=0}^{\infty}(2i+1)r^i \\
&= (\frac{r}{1-r})\left[\frac{2r}{(1-r)^2} + \frac{1}{1-r}\right] \\
&= \frac{r(1+r)}{(1-r)^3}
\end{aligned}$$

Proofs: *Geometric distribution - Mean*

The mean is

$$\begin{aligned}
\sum_{i=1}^{\infty} i(1-p)^{(i-1)}p &= \sum_{i=0}^{\infty}(i+1)(1-p)^i p \\
&= p\left(\sum_{i=0}^{\infty} i(1-p)^i + \sum_{i=0}^{\infty}(1-p)^i\right) \\
&= p\left(\frac{(1-p)}{p^2} + \frac{1}{p}\right) \\
&= \frac{1}{p}.
\end{aligned}$$

Proofs: *Geometric distribution - Variance*

The variance is $\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. We have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{i=1}^{\infty} i^2(1-p)^{(i-1)}p \\
&= \frac{p}{1-p}\sum_{i=0}^{\infty} i^2(1-p)^i \\
&= \frac{p}{1-p}\frac{(1-p)(2-p)}{p^3}
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 &= \frac{(2-p)}{p^2} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

1.2.2 The Binomial Probability Distribution

We interpreted the probability that a coin will come up heads as the fraction of times it will come up heads in an experiment that is repeated many times. But how many times will it come up heads in N flips? We can compute this in a straightforward way, by thinking about outcomes and independence. Assume we have a biased coin, so that $P(H) = p$ and $P(T) = 1 - p$. Now we flip it once. The probability distribution for the number of heads is $P_1(1) = p$ and $P_1(0) = (1 - p)$. Now flip it twice; then the probability distribution for the number of heads is $P_2(2) = pP_1(1) = p^2$; $P_2(1) = pP_1(0) + (1 - p)P_1(1)$; and $P_2(0) = (1 - p)P_1(0)$.

From this, we have that if we flip it N times, the probability distribution satisfies a recurrence relation. In particular $P_N(i) = pP_{N-1}(i-1) + (1-p)P_{N-1}(i)$.

TODO: Insert a figure of Pascal's triangle to drive home this analogy

Notice the similarity between this procedure and what happens when we expand $(x+y)^N$. We get $(x+y)^N = x(x+y)^{N-1} + y(x+y)^{N-1}$. For example, $(x+y)^2 = x(x+y) + y(x+y) = x^2 + xy + yx + y^2 = x^2 + 2xy + y^2$. The binomial theorem tells us that

$$(x+y)^N = \sum_{i=0}^{i=N} \left(\begin{array}{c} N \\ i \end{array} \right) y^N x^{(N-i)}$$

and by pattern matching, we obtain a probability distribution on the number of heads in N flips. In particular, the probability of i heads in N flips with a coin where $P(H) = p$ is

$$P_b(i; N, p) = \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{(N-i)}$$

(as long as $0 \le i \le N$; in any other case, the probability is zero). This is known as the **binomial distribution**.

Definition: *The Binomial distribution*

In N independent repetitions of an experiment with a binary outcome (ie heads or tails; 0 or 1; and so on) with $P(H) = p$ and $P(T) = 1-p$, the probability of observing a total of i H's and $N-i$ T's is

$$P_b(i; N, p) = \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{(N-i)}$$

(as long as $0 \le i \le N$; in any other case, the probability is zero).

Worked example 1.10 *The binomial distribution*

Write $P_b(i; N, p)$ for the binomial distribution that one observes i H's in N trials. Show that

$$\sum_{i=0}^{N} P_b(i; N, p) = 1$$

Solution:

$$\sum_{i=0}^{N} P_b(i; N, p) = (p + (1-p))^N = (1)^N = 1$$

by pattern matching to the binomial theorem

Definition: *Bernoulli random variable*

A Bernoulli random variable takes the value 1 with probability p and 0 with probability $1-p$. This is a model for a coin toss, among other things

Useful facts: *The binomial distribution*

1. The mean of $P_b(i; N, p)$ is Np.
2. The variance of $P_b(i; N, p)$ is $Np(1-p)$

Proofs: *The binomial distribution*

Notice that the number of heads in N coin tosses is can be obtained by adding the number of heads in each toss. This means that, if X has the binomial distribution $P_b(X; N, p)$, and Y has the binomial distribution $P_b(Y; 1, p)$, so we can get the mean easily by

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{j=1}^{N} Y\right] \\ &= \sum_{j=1}^{N} \mathbb{E}[Y] \\ &= N\mathbb{E}[Y] \\ &= Np.\end{aligned}$$

The variance is easy, too. Each coin toss is independent, so the variance of the sum of coin tosses is the sum of the variances. This gives

$$\begin{aligned}\mathsf{Var}[X] &= \mathsf{Var}\left[\sum_{j=1}^{N} Y\right] \\ &= N\mathsf{Var}[Y] \\ &= Np(1-p)\end{aligned}$$

1.2.3 Multinomial probabilities

The binomial distribution describes what happens when a coin is flipped multiple times. But we could toss a die multiple times too. Assume this die has k sides, and we toss it N times. The distribution of outcomes is known as the **multinomial distribution**.

We can guess the form of the multinomial distribution in rather a straightforward way. The die has k sides. We toss the die N times. This gives us a sequence of N numbers. Each toss of the die is independent. Assume that side 1 appears n_1 times, side 2 appears n_2 times, ... side k appears n_k times. Any single sequence with this property will appear with probability $p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$, because the tosses are independent. However, there are

$$\frac{N!}{n_1! n_2! ... n_k!}$$

such sequences. This means that the event "observing side 1 n_1 times, side 2 n_2 times, ... side k n_k times" has probability

$$\frac{N!}{n_1!n_2!...n_k!}p_1^{n_1}p_2^{n_2}\dots p_k^{n_k}$$

Worked example 1.11 *Dice*

I throw five fair dice. What is the probability of getting two 2's and three 3's?

Solution: $\frac{5!}{2!3!}(\frac{1}{6})^2(\frac{1}{6})^3$

1.2.4 The Discrete Uniform Distribution

If every value of a discrete random variable has the same probability, then the probability distribution is the discrete uniform distribution. We have seen this distribution before, numerous times. For example, I define a random variable by the number that shows face-up on the throw of a die. This has a uniform distribution. As another example, write the numbers 1-52 on the face of each card of a standard deck of playing cards. The number on the face of the first card drawn from a well-shuffled deck is a random variable with a uniform distribution.

Assume X and Y are discrete random variables with uniform distributions. Neither $X - Y$ nor $X + Y$ is uniform. We can see this by constructing the distribution. This is easiest to do with concrete ranges, etc. for these random variables. For concreteness, assume that both X and Y are integers in the range $1 - 100$. Write $S = X + Y$, $D = X - Y$. Then

$$P(S = k) = P(\cup_{u=1}^{u=100}\{\{X = k - u\} \cap \{Y = u\}\})$$

but the events $\{\{X = k - u\} \cap \{Y = u\}\}$ and $\{\{X = k - v\} \cap \{Y = v\}\}$ are disjoint if $u \neq v$. So we can write

$$P(S = k) = \sum_{u=1}^{u=100} P(\{\{X = k - u\} \cap \{Y = u\}\})$$

and the events $\{X = k - u\}$ and $\{Y = u\}$ are independent, so we can write

$$P(S = k) = \sum_{u=1}^{u=100} P(\{X = k - u\})P(\{Y = u\}).$$

Now, although $P(X)$ and $P(Y)$ are uniform, the shifting effect of the subtraction term in $\{X = k - u\}$ has very significant effects. For example, imagine $k = 2$; then there is only one non-zero term in the sum (i.e. $P(\{X = 1\})P(\{Y = 1\})$). But if $k = 3$, there are two (i.e. $P(\{X = 2\})P(\{Y = 1\})$ and $P(\{X = 1\})P(\{Y = 2\})$). And if $k = 100$, there are far more terms (which I'm not going to list here).

By a similar argument,

$$P(D = k) = \sum_{u=1}^{u=100} P(\{X = k + u\})P(\{Y = u\}).$$

Again, this isn't uniform; again, the shifting effect of the addition term in $\{X = k + u\}$ has very significant effects. For example, imagine $k = -99$; then there is only one non-zero term in the sum (i.e. $P(\{X = 1\})P(\{Y = 100\})$). But if $k = 98$, there are two (i.e. $P(\{X = 2\})P(\{Y = 100\})$ and $P(\{X = 1\})P(\{Y = 99\})$). And if $k = 0$, there are far more terms (which I'm not going to list here). Figure **??** shows a drawing of the distributions for these sums and differences.

One can construct expressions for the mean and variance of a discrete uniform distribution, but they're not usually much use (too many terms, not often used).

1.2.5 The Poisson Distribution

Assume we are interested in observations that occur in an interval of space (e.g. along some ruler) or of time (e.g. within a particular hour). We know these observations have two important properties. First, they occur with some fixed average rate. Second, an observation occurs independent of the interval since the last observation. Then the Poisson distribution is an appropriate model.

There are numerous such cases. For example, the marketing phone calls you receive during the day time are likely to be well modelled by a Poisson distribution. They come at some average rate — perhaps 5 a day right now, during an election year — and the probability of getting one clearly doesn't depend on the time since the last one arrived. As another example, mark the height of each dead insect on your car windscreen on a ruler; these heights should be well-modelled by a Poisson distribution. Classic examples include the number of Prussian soldiers killed by horse-kicks each year; the number of calls arriving at a call center each minute; the number of insurance claims occurring in a given time interval (outside of a special event like a hurricane, etc.).

If a random variable X has a Poisson distribution, then its probability distribution takes the form

$$P(\{X = k\}) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda > 0$ is a parameter often known as the **intensity** of the distribution.

Notice that this is a probability distribution, because it is non-negative and because

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$$

so that

$$\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = 1$$

Useful facts: *The Poisson Distribution*

1. The mean of a Poisson distribution with intensity λ is λ.
2. The variance of a Poisson distribution with intensity λ is λ (no, that's not an accidentally repeated line or typo).

1.2.6 The Continuous Uniform Distribution

Some continuous random variables have a natural upper bound and a natural lower bound but otherwise we know nothing about them. For example, imagine we are given a coin of unknown properties by someone who is known to be a skillful maker of unfair coins. The manufacturer makes no representations as to the behavior of the coin. The probability that this coin will come up heads is a random variable, about which we know nothing except that it has a lower bound of zero and an upper bound of one.

If we know nothing about a random variable apart from the fact that it has a lower and an upper bound, then a **uniform distribution** is a natural model. Write l for the lower bound and u for the upper bound. The probability density function for the uniform distribution is

$$p(x) = \begin{cases} 0 & x < l \\ 1/(u-l) & l \le x \le u \\ 0 & x > u \end{cases}$$

A continuous random variable whose probability distribution is the uniform distribution is often called a **uniform random variable**. The matlab function `rand` will produce independent samples from a continuous uniform distribution, with lower bound 0 and upper bound 1.

You can guess the expression for a sum of two continuous random variables from the expression for discrete random variables above. Write p_x for the probability density function of X and p_y for the probability density function of Y. Then the probability density function of $S = X + Y$ is

$$p(s) = \int_{-\infty}^{\infty} p_x(s-u)p_y(u)du.$$

Notice that the procedure we have applied to add two random variables could be applied to three, as well. So if we need to know the probability density function for $S_3 = X + Y + Z$, we have

$$p(s) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} p_x((s-v)-u)p_y(u)du \right) p_z(v)dv$$

and so on. It is a remarkable and deep fact, known as the **central limit theorem**, that adding many random variables produces a particular probability density function *whatever* the distributions of those random variables.

1.3 THE NORMAL DISTRIBUTION

Assume we flip a fair coin N times. The number of heads h follows the binomial distribution, so

$$P(h) = \begin{pmatrix} N \\ h \end{pmatrix} (1/2)^N$$

The mean of this distribution is $N/2$, and the variance is $N/4$. Now we reduce this to standard coordinates, as in chapter **??**, by subtracting the mean and dividing by the standard deviation. Our new variable is $x = \frac{2h-N}{\sqrt{N}}$, and we can go the other way by $h = (x + \sqrt{N})(\sqrt{N}/2)$. So the probability distribution for this new variable is

$$P(x) = \binom{N}{(x+\sqrt{N})(\sqrt{N}/2)} (1/2)^N.$$

Notice that N is very large, and we expect x to be relatively small — mostly in the range -3 to 3 or so. Write $u = (\sqrt{N} + x)(\sqrt{N}/2)$, $v = (\sqrt{N} - x)(\sqrt{N}/2)$. Now we have

$$\begin{aligned} P(x) &= \binom{N}{(x+\sqrt{N})(\sqrt{N}/2)} (1/2)^N \\ &= \left(\frac{N!}{u!v!}\right)(1/2)^N. \end{aligned}$$

Now we want a continuous function — a probability density — as N gets very large. This means we can ignore the $(1/2)^N$ term, because it isn't a function of x. We will have to choose a constant so that the probability density function normalizes to one anyhow, so there is no need to drag this term around. The important term is

$$\left(\frac{N!}{u!v!}\right).$$

By applying an approximation (appendix - this proof is informative, but by no manner of means essential) we get

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right).$$

This probability density function is the **standard normal distribution**. We started with a binomial distribution, but standardized the variables so that the mean was zero and the standard deviation was one. We then assumed there was a very large number of coin tosses, so large that that the distribution started to look like a continuous function. The function we get is quite familiar from work on histograms, etc. above, and looks like Figure 1.1. It should look familiar; it has the shape of the histogram of standard normal data, or at least the shape that the histogram of standard normal data aspires to. This argument explains why standard normal data is quite common.

There are two ways to evaluate the mean and standard deviation of this distribution. First, notice that we standardized the variables in the binomial distribution, and so we should have that the mean is zero and the standard deviation is one. The other is simply to compute the mean and standard deviation using integration (which yields the same answer!).

Any probability density function that is a standard normal distribution in standard coordinates is a **normal distribution**. Now write μ for the mean of a probability density function and σ for its standard deviation; we are saying that, if

$$\frac{x-\mu}{\sigma}$$

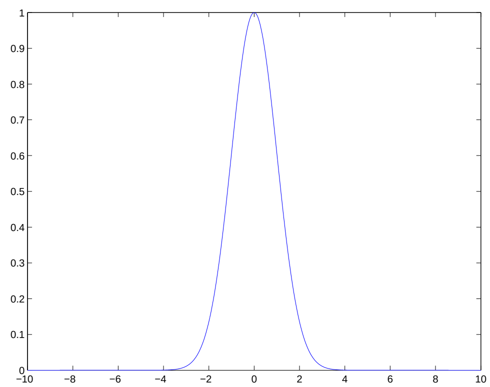

FIGURE 1.1: *A plot of the probability density function of the standard normal distribution. Notice how probability is concentrated around zero, and how there is relatively little probability density for numbers with large absolute values.*

has a standard normal distribution, then $p(x)$ is a normal distribution. We can work out the form of the probability density function in two steps: first, we notice that for any normal distribution, we must have

$$p(x) \propto \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

But, for this to be a probability density function, we must have $\int_{-\infty}^{\infty} p(x)dx = 1$. This yields the constant of proportionality, and we get

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

is the form of a normal distribution with mean μ and standard deviation σ.

Normal distributions are important for two reasons. We have seen a fairly tight argument that anything that behaves like a binomial distribution with a lot of trials — for example, the number of heads in many coin tosses; as another example, the percentage of times you win at a game of MTGDAF (as above) — should produce a normal distribution. Pretty much any experiment where you perform a simulation, then count to estimate a probability, should give you an answer that has a normal distribution.

The second reason I've hinted at above, but not shown in detail because it's a nuisance to prove. If you add together many random variables, each of pretty much any distribution, then the answer has a distribution close to the normal distribution. For these two reasons, we see normal distributions often.

Notice that it is quite usual to call normal distributions **gaussian distributions**.

A normal random variable tends to take values that are quite close to the mean, measured in standard deviation units. We can demonstrate this important fact by computing the probability that a standard normal random variable lies between u and v. We form

$$\int_u^v \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

It turns out that this integral can be evaluated relatively easily using a special function. The **error function** is defined by

$$\mathsf{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt$$

so that

$$\frac{1}{2}\mathsf{erf}\left(\left(\frac{x}{\sqrt{2}}\right)\right) = \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

Notice that $\mathsf{erf}(x)$ is an odd function (i.e. $\mathsf{erf}(-x) = \mathsf{erf}(x)$). From this (and tables for the error function, or Matlab) we get that, for a standard normal random variable

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^1 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.68$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^2 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.95$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^2 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.99.$$

These are very strong statements. They measure how often a standard normal random variable has values that are in the range $-1, 1$, $-2, 2$, and $-3, 3$ respectively. But these measurements apply to normal random variables if we recognize that they now measure how often the normal random variable is some number of standard deviations away from the mean. In particular, it is worth remembering that:

Useful facts: *Normal Random Variables*

- About 68% of the time, a normal random variable takes a value within one standard deviation of the mean.
- About 95% of the time, a normal random variable takes a value within one standard deviation of the mean.
- About 99% of the time, a normal random variable takes a value within one standard deviation of the mean.

1.4 USING EXPECTATIONS

1.4.1 Should you accept the bet?

We can't answer this as a moral question, but we can as a practical question, using expectations. The expected value of a random variable is estimate of what would happen over a large number of trials, on a per-trial basis. This means it tells us what will happen if we were to repeatedly accept a bet.

Worked example 1.12 *Red or Black?*
A roulette wheel has 36 numbers, 18 of which are red and 18 of which are black. Different wheels also have one, two, or even three zeros,which are colorless. A ball is thrown at the wheel when it is spinning, and it falls into a hole corresponding to one of the numbers (when the number is said to "come up"). The wheel is set up so that there is the same probability each number coming up. You can bet on whether a red number or a black number comes up. If you bet \$1 on red, and a red number comes up, you keep your stake and get \$1, otherwise you get $\$-1$ (i.e. the house keeps your bet).

- On a wheel with one zero, what is the expected value of a \$1 bet on red?
- On a wheel with two zeros, what is the expected value of a \$1 bet on red?
- On a wheel with three zeros, what is the expected value of a \$1 bet on red?

Solution: Write p_r for the probability a red number comes up. The expected value is $1 \times p_r + (-1)(1 - p_r)\$$ which is $2p_r - 1\$$.

- In this case, p_r = (number of red numbers)/(total number of numbers) = 18/37. So the expected value is $\$-1/37$ (you lose about 3 cents each time you bet).
- In this case, $p_r = 18/38$. So the expected value is $\$-2/38 = \$-1/19$ (you lose slightly more than five cents each time you bet).
- In this case, $p_r = 18/39$. So the expected value is $\$-3/39 = \$-1/13$ (you lose slightly less than 8 cents each time you bet).

Worked example 1.13 *Coin game*
In this game, P1 flips a fair coin and P2 calls "H" or "T". If P2 calls right, then P1 throws the coin into the river; otherwise, P1 keeps the coin. What is the expected value of this game to P1? and to P2?

Solution: To P2: P2 gets 0 if P2 calls right, and 0 if P2 calls wrong; these are the only cases, so the expected value is 0. To P1: P1 gets -1 if P2 calls right, and 0 if P1 calls wrong. The coin is fair, so the probability P2 calls right is 1/2. The expected value is $-1/2$. While I can't explain why people would play such a game, I've actually seen this done.

We call a bet fair when its expected value is zero. Taking an unfair bet is unwise, because that gets you into a situation equivalent to the gambler's ruin. Recall that, in that example, the gambler bet \$1 on (say) a biased coin flip. The gambler won \$1 with probability p, $\$-1$ with probability $1-p$. This means that the expected value of the bet is is $\$(2p-1)/2$. This is negative, because $p < 1/2$. Now imagine instead that the gambler bets on a fair coin; but the gambler gets $\$2p$ if it comes up heads, and $\$-1$ if it comes up tails. There is no difference between these bets, because the second bet has expected value $\$(2p-1)/2$ — the gambler loses money at the same rate

1.4.2 Odds and bookmaking — a cultural diversion

Gamblers sometimes use a terminology that is a bit different from ours. In particular, the term **odds** is important. The odds of an event are $a : b$ if the probability that the event occurs is $b/(a+b)$. Usually a and b are small integers.

This term comes from the following analysis. Assume you bet \$ 1 that a biased coin will come up heads. $P(H) = p$. If you win the bet, you get $\$k$ and your stake back. If you lose, you lose your stake. What choice of k makes the bet fair? (in the sense that the expected value is zero). The answer is straightforward. The value v of this coin flip to you is $\$-1$ if you lose, and k if you win, so

$$\mathbb{E}[v] = kp - 1(1-p)$$

which means that $k = 1/(1+p)$.

A bookmaker sets adds at which to accept bets from gamblers. The bookmaker does not wish to lose money at this business, and so must set odds which are potentially profitable. Doing so is not simple (bookmakers can lose catastrophically, and go out of business). In the simplest case, assume that the bookmaker knows the probability that a particular bet will win. Then the bookmaker could set odds of $1 : (1+p)$. In this case, the expected value of the bet is zero; this is fair, but not attractive business, so the bookmaker will set odds assuming that the probability is a bit higher than it really is.

In some cases, you can tell when you are dealing with a bookmaker who is likely to go out of business soon. Assume the bet is placed on a horse race, and that bets pay off only for the winning horse. Assume also that exactly one horse will win (i.e. the race is never scratched, there aren't any ties, etc.), and write the probability that the i'th horse will win as p_i. Then $\sum_{i \in \text{horses}} p_i$ must be 1. Now if the bookmaker's odds yield a set of probabilities that is less than 1, their business should fail, because there is at least one horse on which they are paying out too much. Bookmakers deal with this possibility by writing odds so that $\sum_{i \in \text{horses}} p_i$ is larger than one.

But this is not the only problem a bookmaker must deal with. The bookmaker doesn't actually know the probability that a particular horse will win, and must account for errors in this estimate. One way to do so is to collect as much information as possible (talk to grooms, jockeys, etc.). Another is to look at the pattern of bets that have been placed already. If the bookmaker and the gamblers agree on the probability that each horse will win, then there should be no expected advantage to choosing one horse over another — each should pay out slightly less than zero to the gambler (otherwise the bookmaker doesn't eat). But if the bookmaker has underestimated the probability that a particular horse will win, a gambler may get a positive expected payout by betting on that horse. This means that if one particular horse attracts a lot of money from bettors, it is wise for the bookmaker to offer less generous odds on that horse. There are two reasons: first, the bettors might know something the bookmaker doesn't, and they're signalling it; second, if the bets on this horse are very large and it wins, the bookmaker may not have enough capital left to pay out or to stay in business. All this means that real bookmaking is a complex, skilled business.

1.4.3 Ending a game early

Imagine two people are playing a game for a stake, but must stop early — who should get what percentage of the stake? One way to do this is to give each player what they put in at the start, but this is (mildly) unfair if one has an advantage over the other. The alternative is to give each player the expected value of the game at that state for that player. Sometimes one can compute that expectation quite easily.

Worked example 1.14 *Ending a game early*
(from Durrett), two players bet \$50 on the following game. They toss a fair coin. If it comes up heads, player H wins that toss; if tails, player T wins. The first player to reach 10 wins takes the stake. But one player is called away when the state is 8-7 (H-T) — how should the stake be divided?

Solution: The expectation for H is $\$50P(\{\text{H wins from 8-7}\}) + \$0P(\{\text{T wins from 8-7}\})$, so we need to compute $P(\{\text{H wins from 8-7}\})$. Similarly, the expectation for T is $\$50P(\{\text{T wins from 8-7}\}) + \$0P(\{\text{H wins from 8-7}\})$, so we need to compute $P(\{\text{T wins from 8-7}\})$; but $P(\{\text{T wins from 8-7}\}) = 1 - P(\{\text{H wins from 8-7}\})$. Now it is slightly easier to compute $P(\{\text{T wins from 8-7}\})$, because T can only win in two ways: 8-10 or 9-10. These are independent. For T to win 8-10, the next three flips must come up T, so that event has probability 1/8. For T to win 9-10, the next four flips must have one H in them, but the last flip may not be H (or else H wins); so the next four flips could be HTTT, THTT, or TTHT. The probability of this is 3/16. This means the total probability that T wins is 5/16. So T should get \$16.625 and H should get the rest (although they might have to flip for the odd half cent).

Example: *Ending a game early works out*

Imagine we have a game with two players, who are playing for a stake. There are no draws, the winner gets the whole stake, and the loser gets nothing. The game must end early. We decide to give each player the expected value of the game for that player, from that state. Show that the expected values add up to the value of the stake (i.e. there won't be too little or too much money in the stake.

1.4.4 Making a Decision

Imagine we have to choose an action. Once we have chosen, a sequence of random events occurs, and we get a reward with some probability. Which action should we choose? A good answer is to choose the action with the best expected outcome. In fact, choosing any other action is unwise, because if we encounter this situation repeatedly and make a choice that is even only slightly worse than the best, we could lose heavily (check the gambler's ruin example if you're uncertain about this).

This is a very common recipe, and it can be applied to many situations.

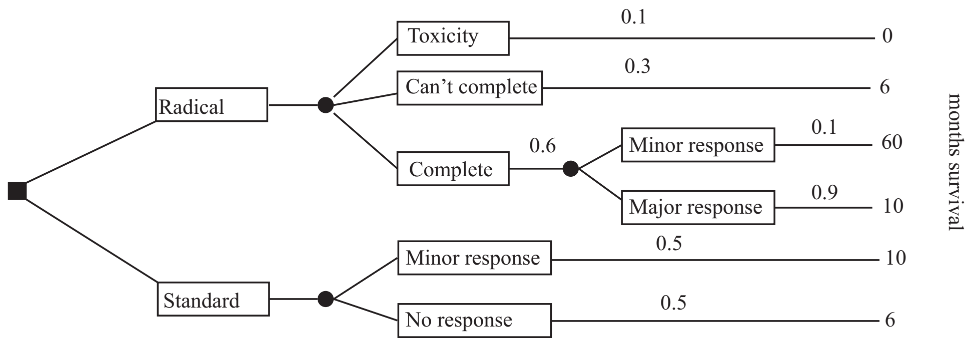

FIGURE 1.2: *A decision tree for the radical treatment example. Decision trees lay out potential outcomes from a decision or a series of decisions, with probabilities, so that one can compute expectations.*

Usually, but not always, the reward is in money, and we will compute with money rewards for the first few examples.

Worked example 1.15 *Vaccination*
It costs \$10 to be vaccinated against a common disease. If you have the vaccination, the probability you will get the disease is $1 - 1e - 7$. If you do not, the probability is 0.1. The disease is unpleasant; with probability 0.95, you will experience effects that cost you \$1000 (eg several days in bed), but with probability 0.05, you will experience effects that cost you \$1e6. Should you be vaccinated?

Solution: The expected cost of the disease is $0.95 \times \$1000 + 0.05 \times \$1e6 = \$50,950$. If you are vaccinated, your cost will be $\$10 + 1e - 7 \times \$50,950 = \$10.01$ (rounding to the nearest cent). If you are not, your cost is $\$5,095$. You should be vaccinated.

Sometimes it is hard to work with money. For example, in the case of a serious disease, choosing treatments often boils down to expected survival times.

Worked example 1.16 *Radical treatment*
(This example largely after Vickers, p97). Imagine you have a nasty disease. There are two kinds of treatment: standard, and radical. Radical treatment might kill you (with probability 0.1); might be so damaging that doctors stop (with probability 0.3); but otherwise you will complete the treatment. If you do complete radical treatment, there could be a major response (probability 0.1) or a minor response. If you follow standard treatment, there could be a major response (probability 0.5) or a minor response, but the outcomes are less good. All this is best summarized in a drawing called a **decision tree** (Figure 1.2). What gives the longest expected survival time?

Solution: In this case, expected survival time with radical treatment is $(0.1 \times 0 + 0.3 \times 6 + 0.6 \times (0.1 \times 60 + 0.9 \times 10)) = 10.8$ months; expected survival time without radican treatment is $0.5 \times 10 + 0.5 \times 6 = 8$ months.

Our reasoning about decision making has used expected values, which are like means. One feature of an expected value is that an unusual value that is very different from all others can have a large effect on the expected value (as in the example of the billionaire in the bar). We computed medians to avoid this effect. But avoiding this effect is not a good idea for decision making. Imagine the following bet: we roll two dice. If two sixes come up, you pay me $\$10,000$; otherwise, I pay you \$10. Notice the median value of this bet for you is \$10, but the mean is $\$-268$.

Working with money values is not always a good idea. For example, many people play state lotteries. The expected value of a \$1 bet on a state lottery is well below \$1 — why do they do it? One reason, harshly but memorably phrased, might be that state lotteries are just a tax on people not knowing how to do sums. This isn't a particularly good analysis, however. It seems to be the case that people value money in a way that doesn't depend linearly on the amount of money. So, for example, people may value a million dollars rather more than a million times the value they place on one dollar. If this is true, we need some other way to keep track of value; this is sometimes called **utility**. It turns out to be quite hard to know how people value things, and there is quite good evidence that (a) human utility is complicated and (b) it is difficult to explain human decision making in terms of expected utility.

Sometimes there is more than one decision. We can still do simple examples, though drawing a decision tree is now helpful, because it allows us to keep track of cases and avoid missing anything. For example, assume I wish to buy a cupboard. Two nearby towns have used furniture shops (usually called antique shops these days). One is further away than the other. If I go to town A, I will have time to look in two (of three) shops; if I go to town B, I will have time to look in one (of two) shops. I could lay out this sequence of decisions (which town to go to; which shop to visit when I get there) as Figure 1.3.

You should notice that this figure is missing a lot of information. What is the probability that I will find what I'm looking for in the shops? What is the value of finding it? What is the cost of going to each town? and so on. This information is not always easy to obtain. In fact, I might simply need to give my best subjective guess of these numbers. Furthermore, particularly if there are several decisions, computing the expected value of each possible sequence could get difficult. There are some kinds of model where one can compute expected values easily, but a good viable hypothesis about why people don't make optimal decisions is that optimal decisions are actually too hard to compute.

1.4.5 Two Inequalities

Mean and variance tell us quite a lot about a random variable, as two important inequalities show.

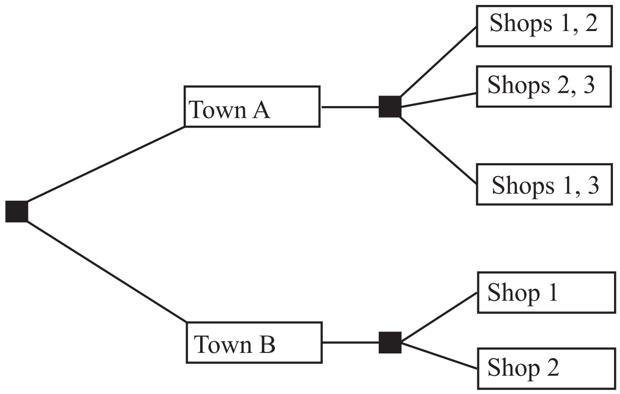

FIGURE 1.3: *The decision tree for the example of visiting furniture shops. Town A is nearer than town B, so if I go there I can choose to visit two of the three shops there; if I go to town B, I can visit only one of the two shops there. To decide what to do, I could fill in the probabilities and values of outcomes, compute the expected value of each pair of decisions, and choose the best. This could be tricky to do (where do I get the probabilities from?) but offers a rational and principled way to make the decision.*

Definition: *Markov's inequality*

Markov's inequality is

$$P(\{|X| \geq a\}) \leq \frac{\mathbb{E}[|X|]}{a}.$$

You should read this as indicating that a random variable is most unlikely to have an absolute value a lot larger than the absolute value of the mean. This should seem fairly intuitive from the definition of expectation. Recall that

$$\mathbb{E}[X] = \sum_{x \in D} xP(\{X = x\})$$

Assume that D contains only non-negative numbers (that absolute value). Then the only way to have a small value of $\mathbb{E}[X]$ is to be sure that, when x is large, $P(\{X = x\})$ is small. The proof is a rather more formal version of this observation, below.

Definition: *Chebyshev's inequality*

Chebyshev's inequality is

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

It is common to see this in another form, obtained by writing σ for the standard deviation of X, substituting $k\sigma$ for a, and rearranging

$$P(\{|X - \mathbb{E}[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

This means that the probability of a random variable taking a particular value must fall off rather fast as that value moves away from the mean, in units scaled to the variance. This probably doesn't seem intuitive from the definition of expectation. But think about it this way: values of a random variable that are many standard deviations above the mean must have low probability, otherwise the standard deviation would be bigger. The proof, again, is a rather more formal version of this observation, and appears below.

Proof of Markov's inequality: (from Wikipedia). Notice that, for $a > 0$,

$$a\mathbb{I}_{[\{|X| \leq a\}]}(X) \leq |X|$$

(because if $|X| < a$, the LHS is a; otherwise it is zero). Now we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X| \leq a\}]}\right] \leq \mathbb{E}[|X|]$$

but, because expectations are linear, we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X| \leq a\}]}\right] = a\mathbb{E}\left[\mathbb{I}_{[\{|X| \leq a\}]}\right] = aP(\{|X| \leq a\})$$

and so we have

$$aP(\{|X| \leq a\}) \leq |X|$$

and we get the inequality by division, which we can do because $a > 0$.

Proof of Chebyshev's inequality: Write U for the random variable $(X - \mathbb{E}[X])^2$. Markov's inequality gives us

$$P(\{|U| \geq w\}) \leq \frac{\mathbb{E}[|U|]}{w}$$

Now notice that, if $a^2 = w$,

$$P(\{|U| \geq w\}) = P(\{|X - \mathbb{E}[X]|\} \geq a)$$

so we have

$$P(\{|U| \geq w\}) = P(\{|X - \mathbb{E}[X]|\} \geq a) \leq \frac{\mathbb{E}[|U|]}{w} = \frac{\mathsf{Var}[X]}{a^2}$$

1.5 APPENDIX: THE NORMAL DISTRIBUTION FROM STIRLING'S APPROXIMATION

To deal with this term we need Stirling's approximation, which says that, for large N,

$$N! \approx \sqrt{2\pi}\sqrt{N}\left(\frac{N}{e}\right)^N.$$

Notice that $u+v=N$, substitute, and we have

$$\binom{N!}{u!v!} \approx \frac{1}{\sqrt{2\pi}}\sqrt{\frac{N}{uv}}\frac{N^N}{u^u v^v} = \frac{1}{\sqrt{2\pi}}\gamma.$$

Now

$$\log\gamma = \log N - \log u - \log v + N\log N - u\log u - v\log v.$$

We have $u = (1+x/\sqrt{N})(N/2)$ and $v = (1-x/\sqrt{N})(N/2)$. Recall from Taylor series that

$$\log a(1+x) \approx \log a + x$$

for small x. This means

$$\log u \approx \log N - \log 2 + \frac{x}{\sqrt{N}}$$

and

$$\log v \approx \log N - \log 2 - \frac{x}{\sqrt{N}}$$

and

$$\begin{aligned} u\log u &\approx (\frac{N}{2})(1+\frac{x}{\sqrt{N}})\left[\log\frac{N}{2}+\frac{x}{\sqrt{N}}\right] \\ &= (\frac{N}{2})(1+\frac{x}{\sqrt{N}})\left[\log\frac{N}{2}+\frac{x}{\sqrt{N}}\right] \\ &= (\frac{N}{2})\left[(\log N - \log 2) + \frac{x}{\sqrt{N}} + \frac{N}{2}\frac{x}{\sqrt{N}} + \frac{x^2}{N}\right] \end{aligned}$$

and

$$\begin{aligned} v\log v &\approx (\frac{N}{2})(1-\frac{x}{\sqrt{N}})\left[\log\frac{N}{2}-\frac{x}{\sqrt{N}}\right] \\ &= (\frac{N}{2})(1-\frac{x}{\sqrt{N}})\left[\log\frac{N}{2}-\frac{x}{\sqrt{N}}\right] \\ &= (\frac{N}{2})\left[(\log N - \log 2) - \frac{x}{\sqrt{N}} - \frac{N}{2}\frac{x}{\sqrt{N}} + \frac{x^2}{N}\right]. \end{aligned}$$

Substituting, and clearing terms, we get

$$\log\gamma \approx \frac{-x^2}{2}$$

meaning that

$$p(x) \propto \exp\left(\frac{-x^2}{2}\right).$$

We can get the constant of proportionality from integrating, to find

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right).$$

Notice what has happened in this blizzard of terms. We started with a binomial distribution, but standardized the variables so that the mean was zero and the standard deviation was one. We then assumed there was a very large number of coin tosses, so large that that the distribution started to look like a continuous function. The function we get is quite familiar from work on histograms, etc. above.