4 Random variables

Very frequently when we are applying statistics and probability we are dealing with data we want to quantify. Because we can add, subtract and multiply numerical data, applying probability to numerical data allows for some techniques and approaches that aren't possible for other data types. The mathematical construction used in this case is a **random variable**.

The expression 'random variable' is often used very loosely and in a way that can be confusing; however, it does have an exact meaning. A random variable X is a map from a sample space to a set of numerical values. For now we will deal with a discrete random variables, where those values are discrete. As an example consider the sample space for flipping a coin three times; the sample space will have eight elements, {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}. One random variable, say X, on this space is the number of heads. This random variable maps HHH to three and TTH to one.

Just like events have a probability, so do random variables, in fact it is nearly the same definition: the probability that $X = x$, $p(X = x)$, sometimes written $p(x)$, is the sum of the probabilities of all the outcomes with value x. For a sample space $\mathcal{S}$ with random variable X it is the probability of the event:

$$E_x = \{s \in \mathcal{S} | X(s) = x\} \tag{1}$$

When there are more than one random variable being considered we sometimes write $p_X(x)$ to mean the probability corresponding to the random variable X. Note that we use a small letter p for probabilities when we are talking about random variables, a big letter P is usually used when talking about events.

Going back to the coin flipping example, it is easy to see that

$$p(X = 1) = \frac{3}{8} \tag{2}$$

since there are three equally probable outcomes corresponding to $X = 1$. We can put all the possible values of the random variable in to a table

	0	1	2	3
p	1/8	3/8	3/8	1/8

A table like this is called a **probability distribution**.

The probabilities satisfy similar properties to the probabilities for events:

$$0 \leq p(x) \leq 1 \tag{3}$$

and

$$\sum_x p(x) = 1 \tag{4}$$

where the sum is over all possible values x with non-zero probabilities. More importantly, we interpret $p(x)$ as the frequency $X = x$, so if we choose random elements of the sample space

$$\frac{\text{number of times we get the value } x}{\text{number of samples we take}} \to p_X(x) \tag{5}$$

with the left hand side approaching the right hand side as the number of samples goes to infinity.

As another example, consider rolling two three-sided dice. I'll leave it to you to work out how to make a three-sided dice; having a more realistic number of dice faces makes the tables really big. Let X be the sum of the two numbers and Y be the larger of the two numbers. We can work out both distributions. For X we have

	2	3	4	5	6
p_X	1/9	2/9	1/3	2/9	1/9

and for Y

	1	2	3
p_Y	1/9	1/3	5/9

We won't discuss it much now but you could also work out a **joint distribution**, this is the distribution for pairs (X, Y):

Y \ X	2	3	4	5	6
1	1/9	0	0	0	0
2	0	2/9	1/9	0	0
3	0	0	2/9	2/9	1/9

so $p_{X,Y}(3, 2) = 2/9$ since it corresponds to the roll one and two along with the roll two and one. We can recover the original **marginal** distributions p_X and p_Y by summing out the variable we aren't interested in; clearly

$$p_Y(y = 2) = p_{X,Y}(2, 2) + p_{X,Y}(3, 2) + p_{X,Y}(4, 2) + p_{X,Y}(5, 2) + p_{X,Y}(6, 2) = \frac{2}{9} + \frac{1}{9} = \frac{1}{3} \quad (6)$$

We can also work out, again with the obvious notation, **conditional probabilities** $p_{X|Y}(x|y)$ so

	2	3	4	5	6
$p_{X\|Y}(x\|Y = 2)$	0	2/3	1/3	0	0

which is just

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \quad (7)$$

Expected values

So far we have introduced some new, and slightly confusing, notation without anything to show for it. However, there is an advantage; as we noted above, numerical values support arithmetic. This allows us to define the expectation value. If $g(x)$ is a function we define the **expected value** of $g(X)$ as

$$\langle g(X) \rangle = \sum_x p(x) g(x) \quad (8)$$

where the sum is over all values x with non-zero probability. The angle brackets $\langle \ldots \rangle$ are a common notation for expectation values; they lead to a very convenient notation called 'bra and ket' notation; we won't get to that here but we will use angle brackets for expectation values. Another common notation is to use a capital E, as in

$$E[g(X)] = \langle g(X) \rangle \quad (9)$$

The most obvious expected value is the **expected value** of X:

$$\langle X \rangle = \sum_x xp(x) \tag{10}$$

Clearly, if the $p(x)$ represent the frequencies, this is the average or **mean** value taken by X; it is often called μ. We are being careful to distinguish the expected value of X and the mean; there are occasions when this distinction matters, but generally it doesn't. Basically the expected value is defined no matter what the $p(x)$s represent. If, as they almost always do, they represent the frequencies, what we think of as the probability of getting $X = x$, then the expected value is equal to the mean. On the way to proving some theorem it might be useful to define another probability distribution on the same data, where the second distribution doesn't correspond to frequencies; this is why it can be useful to distinguish the mean and the expected value of X.

The expect value of X is also referred to as the **first moment**; the 'first' bit is because it is the expectation value for the first power of X.

As a simple example consider the probability distribution above for the number of heads when a coin is flipped three times:

$$\langle X \rangle = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3+6+3}{8} = \frac{3}{2} \tag{11}$$

Another common expected value is the **variance**:

$$V(X) = \langle (X - \mu)^2 \rangle \tag{12}$$

If $p(x)$ represents the frequencies, this is the square of the **standard deviation**: $V(X) = \sigma^2$. It is often interpreted as measuring how spread out the data is. For example, this distribution:

	0	1	2	3
p_X	1/8	3/8	3/8	1/8

and this distribution

	0	1	2	3
p_Y	1/16	7/16	7/16	1/16

both have expect valued 1.5:

$$\langle X \rangle = \langle Y \rangle = 1.5 \tag{13}$$

However,

$$V(X) = 0.75 \tag{14}$$

whereas

$$V(Y) = 0.5 \tag{15}$$

reflecting the way the second distribution is more tightly bunched up around the mean.

The expected value has nice properties that it inherits from the nice properties of the underlying arithmetic. First, scalar multiplication: for a constant c

$$\langle cg(X) \rangle = \sum_x cg(x)p(x) = c \sum_x g(x)p(x) = c\langle g(X) \rangle \tag{16}$$

Also, trivially:

$$\langle 1 \rangle = \sum_x p(x) = 1 \tag{17}$$

Finally, it is additive: if g_1 and g_2 are two different functions

$$\begin{aligned} \langle g_1(X) + g_2(X) \rangle &= \sum_x [g_1(x) + g_2(x)] p(x) \\ &= \sum_x g_1(x)p(x) + \sum_x g_2(x)p(x) = \langle g_1(X) \rangle + \langle g_2(X) \rangle \end{aligned} \tag{18}$$

We can use this to get another formula for the variance:

$$V(X) = \langle (X - \mu)^2 \rangle = \langle (X^2 - 2\mu X + \mu^2) \rangle \tag{19}$$

Now, using the additive property

$$V(X) = \langle X^2 \rangle - 2\mu \langle X \rangle + \langle \mu^2 \rangle \tag{20}$$

Finally, noting $\mu = \langle X \rangle$ and using the scalar multiplication property, this give

$$V(X) = \langle X^2 \rangle - \mu^2 \tag{21}$$

Incidentally $\langle X^2 \rangle$ is called the **second moment**, the variance is called the **second central moment**; the 'central' indicates that it is the second moment you get if you take away the mean first.

Summary

- A **random variable** is a map from sample space to a set of numerical values.
- The probability that $X = x$, $p(X = x)$, sometimes written $p(x)$, is the sum of the probabilities of all the outcomes with value x.
  1. $$0 \leq p(x) \leq 1 \tag{22}$$
  2. $$\sum_x p(x) = 1 \tag{23}$$
- A **probability distribution** is a table of probabilities for a random variable.
- For two random variables X and Y, the **joint distribution** is $p(x, y)$, the probability $X = x$ and $Y = y$; the **conditional distribution** of $X = x$ given $Y = y$ is $p(x|y)$ and the **marginal distribution** is
$$p(x) = \sum_y p(x, y) \tag{24}$$
- Is $g(x)$ is a function, the **expected value** is
$$\langle g(X) \rangle = \sum_x p(x)g(x) \tag{25}$$
- The **mean** is $\langle X \rangle$. It is often called μ.
- The **variance** is $\langle (X - \mu)^2 \rangle$. It is often called V or σ^2.
- The n**th moment**, often written μ_n, is $\langle X^n \rangle$ and the n**th central moment** is $\langle (X-\mu)^n \rangle$.
- Expected values have nice properties
  1. $\langle cg(X) \rangle = c\langle g(X) \rangle$
  2. $\langle 1 \rangle = 1$
  3. $\langle g_1(X) + g_2(X) \rangle = \langle g_1(X) \rangle + \langle g_2(X) \rangle$
- Using these nice properties it can be shown that $\sigma^2 = \langle X^2 \rangle - \mu^2$

Example question

Work out the variance for the result of rolling a die. **solution** So all the probabilities are 1/6 so the mean is

$$\mu = \frac{1+2+3+4+5+6}{6} = 3.5 \tag{26}$$

and the second moment is

$$\langle X^2 \rangle = \frac{1+4+9+16+25+36}{6} = 15.1667 \tag{27}$$

so

$$\sigma^2 = 15.1667 - 12.25 = 2.9167 \tag{28}$$