Chapter 6

Random Variables

6.1 Random Variables

In this chapter, we will put together the ideas we have developed for probability and descriptive statistics, to build tools that will help us understand statistics such as the average of a random sample.

The story begins with an *outcome space*, that is, the set of all possible outcomes of an experiment that involves chance. Standard notation for this space is Ω, the upper case Greek letter Omega. **For mathematical simplicity, we will assume that the outcome space Ω is finite.** Each outcome ω (that's lower case omega) is assigned a probability, and the total probability of all the outcomes is 1.

Example 1. You flip a fair coin three times, and record what face the coin landed on. What is a reasonable outcome space Ω for this experiment, and what is the probability distribution on that space ?

You can think of Ω as consisting of 8 equally likely outcomes, since the coin is not biased. Here's a table representing Ω and the probabilities of all the outcomes.

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

A *random variable* X is a real-valued function defined on Ω. That is, the domain of X is Ω and the range of X is the real line.

Note. We are going to restrict attention to random variables that have a finite number of values.

In Example 1, X could be the number of times the letter H appears in an outcome. You can think of X as the number of heads in three tosses of a coin.

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
$X(\omega)$	0	1	1	1	2	2	2	3

The probability function on Ω determines probabilities for X. For example, the chance that X is 1 is defined as follows:

$$P(X = 1) \;=\; P(\{\omega : X(\omega) = 1\}) \;=\; P(\text{TTH, THT, HTT}) \;=\; 1/8 + 1/8 + 1/8 \;=\; 3/8$$

6.2 Probability distribution

The *probability distribution* of X is a distribution on the range of X. It specifies all the possible values of X along with all their probabilities.

For random variables that have a finite number of possible values, the probability distribution is also known as the *probability mass function*.

For example, for X defined in the example above, the probability distribution is given by

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

All the probabilities in a distribution must add up to 1.

Note that the probabilities on the range of X are determined by the probabilities on the domain. In our example, had all 8 elements in the domain not been equally likely, the possible values of X would still have been the same but the probabilities might have been different.

Example 2: Uniform distribution A word is picked at random from the set science, computer. Let V be the number of distinct vowels in the word. What is the distribution of V?

Solution. $\Omega = \{\text{science, computer}\}$. The probabililty distribution on Ω assigns probability 1/2 to each element. Now $V(\text{science}) = 2$ and $V(\text{computer}) = 3$.

So the distribution of V puts probability 1/2 on each of the values 2 and 3. We say that V has the *uniform* distribution on $\{2, 3\}$.

Example 3: Binomial distribution. Suppose a coin is tossed n times. Let X be the number of heads. Find the distribution of X.

Solution. Each possible outcome of n tosses is a sequence of n H's and T's. You have seen in an earlier exercise that there are 2^n equally likely such sequences.

When you are finding a distribution, an excellent idea is to start with the possible values rather than the probabillities. The possible number of heads in n tosses is 0 through n.

For any fixed k in the range 0 through n, the number of sequences that contain exactly k H's is $\binom{n}{k}$. So the distribution of X is

$$P(X = k) = \frac{\binom{n}{k}}{2^n}, \quad k = 0, 1, \ldots, n$$

This is called the *binomial distribution with parameters n and* 1/2. You should convince yourself that the formula gives a sensible answer in the two edge cases $k = 0$ and $k = n$.

If the coin is unfair, it is still true that there are $\binom{n}{k}$ sequences with exactly k H's, but all 2^n sequences are no longer equally likely. To find the probability of each sequence, let p be the chance that the coin lands heads on a single toss, and make the natural assumption that the outcome of any set of tosses does not affect chances for any other set.

Let's work out an example under these assumptions before finding the general formula. Suppose $n = 3$ as in Example 1, and let $k = 2$. Then the sequences corresponding to 2 heads in 3 tosses are HHT, HTH, and THH. Each of these has probability $p^2(1-p)$, because there are two factors of p and one of $(1-p)$ in different order. So the chance of 2 heads in 3 tosses is $P(\text{HHT or HTH or THH}) = 3p^2(1-p)$.

Notice that the 3 in the answer above is the number of sequences that have 2 H's, that is, $3 = \binom{3}{2}$.

Now we are ready to extend the argument to the general case. Suppose the coin lands heads on a single toss with probability p. If X is the number of heads on n tosses, then the distribution of X is

$$P(X=k) \ = \ \binom{n}{k} p^k (1-p)^{n-k}, \quad k \ = \ 0,\ 1, \ldots,\ n$$

This is called the *binomial distribution with parameters n and p*. Notice that the fair coin is a special case:

$$P(X=k) \ = \ \frac{\binom{n}{k}}{2^n} \ = \ \binom{n}{k}(\frac{1}{2})^k(\frac{1}{2})^{n-k}$$

The name *binomial* comes from the fact that the probabilities in the binomial distribution are the terms in the *binomial expansion* $(a+b)^n$ for $a=p$ and $b=1-p$. This also shows why the sum of the terms is equal to 1.

The distribution can be used to find probabilities of events, as in the following examples.

The chance of getting between 45 and 55 heads (inclusive) in 100 tosses of a fair coin is

$$\sum_{k=45}^{55} \frac{\binom{n}{k}}{2^n}$$

The chance of getting fewer than 10 sixes in 12 rolls of a die is

$$\sum_{k=0}^{9} \binom{12}{k}(\frac{1}{6})^k(\frac{5}{6})^{12-k} \ = \ 1 - \sum_{k=10}^{12} \binom{12}{k}(\frac{1}{6})^k(\frac{5}{6})^{12-k}$$

The second form is simpler to calculate, as the sum only has three terms.

Example 4. A 5-card poker hand is dealt from a standard deck. What is the distribution of the number of queens in the hand?

Solution. Let's give the variable a name: Q for "queens". We will start by listing the possible values of Q. We are dealing 5 cards but there are only 4 aces in the deck. So the possible values of Q are $0,1,2,3,4$.

The total number of possible hands is the total number of subsets of 5 that can be chosen from among 52 cards. That's $\binom{52}{5}$, and they are all equally likely.

Now fix k in the range 0 through 4. To find $P(Q=k)$, we will need the number of hands that contain exactly k queens. These k queens can be chosen in $\binom{4}{k}$ ways. For each of these ways, there are $\binom{48}{5-k}$ ways of choosing the remaining $5-k$ cards in the hand. So the number of 5-card hands that contain exactly k queens is $\binom{4}{k}\binom{48}{5-k}$.

So the distribution of Q is given by

$$P(Q=k) \ = \ \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}, \ \ k \ = \ 0,1,2,3,4$$

The chance of three or more queens in a 5-card hand is

$$P(Q \geq 3) \ = \ \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} + \frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}}$$

6.3 Functions of Random Variables

We will frequently be interested in quantities that can be calculated based on a random variable. In other words, we will be interested in functions of random variables. A function of a random variable is also a random variable, because a function of a function on Ω is also a function on Ω.

Example 5. Let X have the binomial $(4, 1/2)$ distribution. Let Y be the absolute deviation of X from 2, that is, $Y = |X - 2|$. What is the distribution of Y?

Solution. The table below gives the distribution of X along with the possible value y of Y computed from each possible value x of X:

y	2	1	0	1	2
x	0	1	2	3	4
$P(X = x)$	1/16	4/16	6/16	4/16	1/16

Collect terms get the distribution of Y:

y	0	1	2
$P(Y = y)$	6/16	8/16	2/16

Example 6: Net gain on red at roulette. In Nevada roulette, the bet on "red" pays 1 to 1 and you have 18 in 38 chances of winning. This means that if you bet a dollar on "red" and the winning color is not red, you lose your dollar; if the winning color is red, then your net winnings are a dollar. Suppose the roulette wheel is spun 10 times, and you bet a dollar on red each time. What is the chance that you make money?

Solution. Let X be the number of times you win. Then X has the binomial distribution with parameters $n = 10$ and $p = 18/38$. But the question is about your net gain being positive overall. So let your net gain be G, and let's try and work out what G has to be for you to make money.

For any possible value x of X, the amount of money you make is the total of \$1 for each of the x bets that you win and -\$1 for each of the $10 - x$ bets that you lose. This corresponds to making g dollars overall, where

$$g = 1 \cdot x + (-1) \cdot (10 - x)$$

So

$$g = 2x - 10$$

So your random net gain G is the following function of the random number of bets X that you win:

$$G = 2X - 10$$

The question asks for $P(G > 0)$.

$$P(G > 0) = P(2X - 10 > 0) = P(X > 5) = \sum_{x=6}^{10} \binom{10}{x} \left(\frac{18}{38}\right)^x \left(\frac{20}{38}\right)^{10-x}$$

6.4 Expectation

The *expectation* of X, denoted $E(X)$, is the average of the possible values of X, weighted by their probabilities.

Remember that we are assuming that X has finitely many values. So the expectation of X is finite.

The expectation can be computed in two equivalent ways, one defined on the domain of X and one on the range:

$$E(X) \;=\; \sum_{\omega} X(\omega)P(\omega) \;=\; \sum_{x} xP(X=x)$$

Let X be the number of heads in three tosses of a coin. The natural outcome space Ω, along with the probabilities of all the outcomes ω, is given by:

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

X is a function from Ω to $\{0,1,2,3\}$, with distribution given by:

x	0	1	2	3
$P(X=x)$	1/8	3/8	3/8	1/8

So the first form of the calculation of the expectation of X is

$$E(X) = 0\cdot 1/8 + 1\cdot 1/8 + 1\cdot 1/8 + 1\cdot 1/8 + 2\cdot 1/8 + 2\cdot 1/8 + 2\cdot 1/8 + 3\cdot 1/8 \;=\; 1.5$$

The second form is based on the probability distribution of X:

$$E(X) \;=\; 0\cdot 1/8 + 1\cdot 3/8 + 2\cdot 3/8 + 3\cdot 1/8 \;=\; 1.5$$

Expectation is often denoted by μ, the lower case Greek letter mu. That's because expecations and averages are often called *means*.

The form

$$E(X) \;=\; \sum_{x} xP(X=x)$$

is most commonly used in calculations involving simple distributions. Go back and look at the section **Another Way to Calcuate the Average** at the end of the section on averages in Chapter 1. You will recognize that the formula there is analogous to the formula for $E(X)$ above.

This implies that $E(X)$ is just an ordinary average and thus has all the familiar properties of averages. For example, if the random variable X is a constant, that is, if there is a constant c such that $P(X=c)=1$, then $E(X)=c$.

Linear transformations. Also, expectation transforms linearly when the random variable is transformed linearly:

$$E(aX+b) \;=\; aE(X)+b$$

The expectation is the balance point of the histogram of the probability distribution. And so on. Expecations have the properties of averages.

Expectation of a function of X. Any function of X is also a random variable, and thus has an expectation. Usually, the easiest way to find the expectation of a function of X is to do the calculation using the distribution of X. That is, if the random variable $Y = g(X)$, then

$$E(Y) \ = \ E(g(X)) \ = \ \sum_x g(x)P(X = x)$$

So if X is the number of heads in three tosses of a coin, and $Y = |X - 1.5|$, then

$$\begin{aligned} E(Y) &= E(|X - 1.5|) \\ &= |0 - 1.5| \cdot (1/8) + |1 - 1.5| \cdot (3/8) + |2 - 1.5| \cdot (3/8) + |3 - 1.5| \cdot (1/8) \end{aligned}$$

As you will see below, we will sometimes need to find the expectation of the square of a random variable. By the "function rule" above,

$$E(X^2) \ = \ \sum_x x^2 P(X = x)$$

6.5 Standard Deviation and Bounds

How far away from its expectation is a random variable likely to be? To answer this question, we need a measure of deviation away from the expectation. We will develop one that is analogous to the standard deviation of a list of numbers.

Define the *deviation* of X to be the random variable $D = X - \mu$, where $\mu = E(X)$.

To find the rough size of D, suppose we calculate $E(D)$. Notice that D is a linear transformation of X. By our rule about linear tranformations,

$$E(D) \ = \ E(X - \mu) \ = \ E(X) - \mu \ = \ \mu - \mu \ = \ 0$$

The expected deviation is 0, no matter what the distribution of X is. This is the parallel to the result that the average of a list of deviations from average is 0 no matter what the list is.

As with the deviations of values in a list, the problem is cancellation: the negative deviations cancel out the positive ones. So we can't learn anything meaningful about the spread of the variable if we just calculate the expected deviation.

So, just as we did with the deviations from average of a list of numbers, we'll square the deviations to ensure that they don't cancel each other out. This leads us to the definition of the *variance* of X, denoted $Var(X)$.

$$Var(X) \ = \ E[D^2] \ = \ E[(X - \mu)^2]$$

To correct the units of measurement, we have to take the square root of the variance. The *standard deviation* of X is then

$$SD(X) \ = \ \sqrt{Var(X)} \ = \ \sqrt{E(D^2)} \ = \ \sqrt{E[(X - \mu)^2]}$$

$SD(X)$ is often denoted by σ, the lower case Greek letter sigma.

The standard deviation defined here can also be thought of as the ordinary SD of a list of numbers, computed using the distribution table of the list. As such, it is an ordinary SD and has all the properties of SDs that we discovered in Chapter 2.

For example, if a random variable X is a constant, then $SD(X) = 0$.

Linear transformations. Only the mutiplicative factor of a linear transformation affects the SD:

$$SD(aX + b) = |a|SD(X)$$

Thus if X is a random temperature in degrees Celsius, and Y the corresponding temperature in degrees Fahrenheit, then $Y = (9/5)X + 32$. So

$$E(Y) = (9/5)E(X) + 32 \qquad SD(Y) = (9/5)SD(X)$$

Another consequence is that a constant shift simply slides the probability distribution along and doesn't affect the SD. For any constant c,

$$SD(X + c) = SD(X)$$

6.6 Bounding Tail Probabilities

The tail bounds that we established for distributions of data work for probability distributions of random variables as well.

Markov's Inequality

Let X be a non-negative random variable. That is, assume that $P(X \geq 0) = 1$. Let $E(X) = \mu$. Then for any constant $c > 0$

$$P(X \geq c) \leq \frac{\mu}{c}$$

Chebychev's Inequality

Let X be a random variable that has $E(X) = \mu$ and $SD(X) = \sigma$. Let k be any positive constant. Then

$$P(X \text{ is outside the range } \mu \pm k\sigma) \leq \frac{1}{k^2}$$

More formally, for all $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

These are exactly the same as the bounds obtained earlier for distributions of data, and they are true for the same reasons.

If you would like to prove them afresh, you can follow the steps in the proof in Chapters 1 and 2 but write them in random variable notation.

Standard Units

Let X be a random variable with expecation μ_X and SD σ_X. The random variable Z defined by

$$Z = \frac{X - \mu_X}{\sigma_X}$$

is called X *in standard units*. Notice that Z is a linear transformation of X.

As you can see from the definition of Z, standard units measure the deviation of X relative to the SD of X. In other words, they measure *how many SDs above the expected value* the value of X is.

Conversion to standard units allows us to compare the distributions of random variables that have been measured on different scales.

Measuring a random variable in standard units is equivalent to setting the origin to be μ_X and measuring distances from the origin in units of SDs. By our results about linear transformations,

$$\mu_Z = E(Z) = \frac{E(X) - \mu_X}{\sigma_X} = 0$$

and

$$\sigma_Z = SD(Z) = \frac{SD(X)}{\sigma_X} = 1$$

We can re-write Chebychev's Inequality in terms of Z. For any $k > 0$,

$$P(|X - \mu_X| \geq k\sigma_X) = P(|\frac{X - \mu_X}{\sigma_X}| \geq k) = P(|Z| \geq k)$$

So if Z is X measured in standard units, Chebychev's Inequality becomes

$$P(|Z| \geq k) \leq \frac{1}{k^2}$$

The chance that a random variable is at least 4 in standard units (that is, at least 4 SDs away from its expected value) is 1/16, which is quite small. Thus, when any random variable is measured in standard units, the bulk of its probability distribution lies in the interval $(-4, 4)$.