Chapter 4

Discrete Random Variables

4.1 Probability mass function

Suppose we have three (distinguishable) dice (say coloured red, blue and green) and they are rolled together and the number facing upwards is recorded. Then the sample space of such an experiment is the set of all possible combinations: $\Omega = \{(d_R, d_B, d_G) : d_R, d_B, d_G = 1, 2, 3, 4, 5, 6\}$, of which there are $6 \times 6 \times 6$ distinct events. We recognise such an experiment as a random (unpredictable) process. Let us define a *discrete random variable* associated with the experiment, for example the total of the three numbers:

$$X = \text{sum of } \mathrm{d_R} \text{ and } \mathrm{d_B} \text{ and } \mathrm{d_G}. \tag{4.1}$$

Then the sample space for X is the set of events:

$$X \in \{3, 4, \ldots, 17, 18\} \quad . \tag{4.2}$$

For each of the values of X, we can assign a *probability mass function* (p.m.f) which we denote by the symbol $f_X(x)$:

$$f_X(x_i) \equiv P(X = x_i) \quad . \tag{4.3}$$

At this point, we do not specify the form of this function, but simply assert that it exists. Since all possible events are included, the total probability (that is one of the outcomes in the sample space is certain to occur) adds to one:

$$P(\Omega) = \sum_i f_X(x_i) = 1 \quad , \tag{4.4}$$

which is the expression for the law of total probability.

Example A (fair) coin is tossed twice and the outcome (ordered pair) of the first and second tosses is noted. In this process the event space is the following: $\Omega = \{HH, HT, TH, TT\}$.

Let X be the number of *heads* that occurs in this process. Then $X \in \{0, 1, 2\}$, and the corresponding *probability mass function*, $f_X(x)$, can be calculated. We have specified that the coin is fair and thus:

$$P(H) = P(T) \quad , \tag{4.5}$$

and since, by the law of total probability:

$$P(H) + P(T) = 1 \quad , \tag{4.6}$$

from these equations we deduce that; $P(H) = \frac{1}{2}$.

Given that the first and second tosses are independent events, we then have:

event	X	$f_X(x)$
HH	$X = 2$	$P(X = 2) = \frac{1}{4}$
HT, TH	$X = 1$	$P(X = 1) = \frac{1}{2}$
TT	$X = 0$	$P(X = 0) = \frac{1}{4}$

This probability mass can be tabulated (as above) and/or sketched on a graph (figure 4.1).

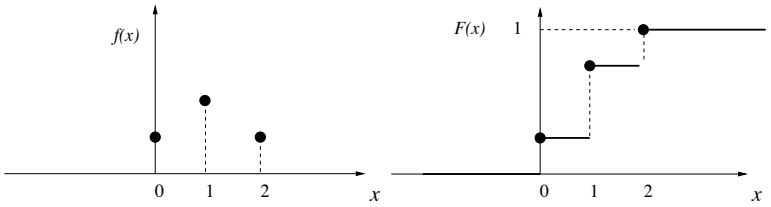

Figure 4.1: Left: Probability mass function: $f_X(x)$ for the number of HEADS after two tosses of a fair coin. Right: the corresponding probability distribution function, $F_X(x)$.

Of course, the law of total probability applies, so that when, $X = \{x_1, x_2, \ldots, x_n\}$ we note that

$$\sum_{i=1}^{n} f_X(x_i) = 1 \quad ,$$

that is, the probability of a certain event is 1.

4.2 Probability distribution function

Given a probability mass function, we can also define a corresponding *probability distribution function*, that we denote as:

$$F_X(x) = P(X \leq x).$$

Notice that $F_X(x)$ begins at $F = 0$ and rises (monotonically) in a series of steps terminating at $F = 1$.

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1 \qquad , \tag{4.7}$$

as illustrated in figure 4.1.

Since the probability mass is never negative, then the the distribution function is monotonically increasing. That is,

$$F(x) \leq F(y) \qquad , \qquad x \leq y \qquad . \tag{4.8}$$

Finally, the relation between the two functions may be expressed as

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \quad , \tag{4.9}$$

that is the distribution function increases in steps, with the height of each step being the corresponding mass at that point. Reciprocally, we write that:

$$f_X(x) = F_X(x) - \lim_{h \to 0+} F(x - h) \quad ,$$

that is the mass function can be estimated by the height of each step in the distribution function. The notation $\lim_{h \to 0+}$ means, h tends towards zero from the positive side.

We can introduce the notion of an *indicator function*, $\mathbb{I}(X)$, to give a numerical value to the statement *true or false*. It is defined as follows:

$$\mathbb{I}(x \in A) = \begin{cases} 1 & , \quad x \in A \\ 0 & , \quad x \notin A \end{cases} \tag{4.10}$$

This function is very useful in algebraic manipulations and, for example, it is useful is defining the probability distribution:

$$F_X(x) = \sum_i \mathbb{I}(x_i \leq x) \, f_X(x_i) \qquad . \tag{4.11}$$

The *median* of a distribution is the value of X such that it divides the probability distribution in half. That is, the median $\tilde{x}$ is the solution of the equation

$$F_X(\tilde{x}) = \frac{1}{2} \quad , \tag{4.12}$$

which can be written in the form:

$$\sum_{x_i \leq \tilde{x}} f_X(x_i) = \frac{1}{2} \quad . \tag{4.13}$$

The *mode* of a distribution is the value of x_i for which $f_X(x_i)$ is a maximum.

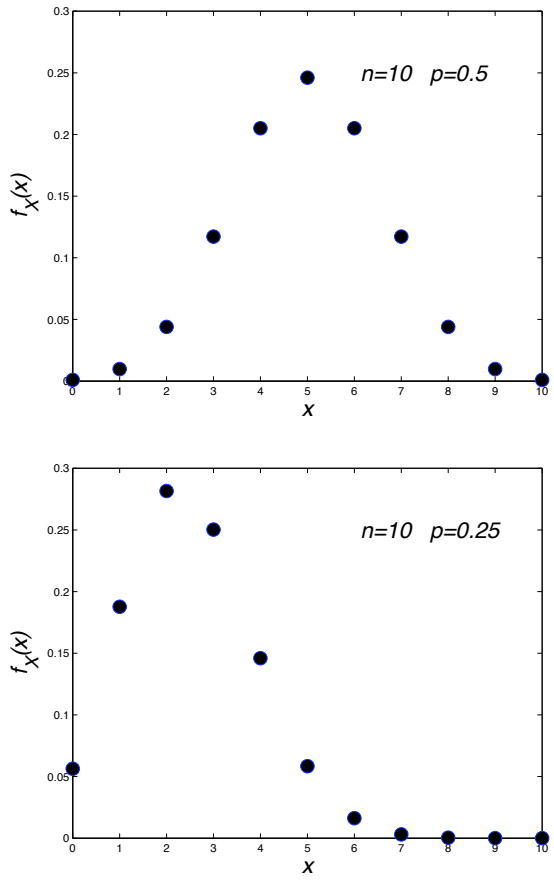

Figure 4.2: Examples of the probability mass function for the Binomial (n, p) distribution: $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$. Top: $n = 10, p = 0.5$ which is symmetric about the mode $x = 5$. Bottom $n = 10$, $p = 0.25$, which is asymmetric with a mode at $x = 2$.

4.3 Bernoulli variable

Consider a single coin toss, with the event space being the upward face of the coin. Then $\Omega = \{H, T\}$, suppose that $P(H) = p$, $P(T) = 1 - p$.
Let X be the number of heads that occur. Then the mass function is:

$$\begin{array}{ll} X(H) = 1 & X(T) = 0 \\ f_X(1) = p & f_X(0) = 1 - p \end{array}$$

with the corresponding distribution:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x \leq 1 \\ 1 & 1 \leq x \end{cases} \tag{4.14}$$

This is an example of a Bernoulli trial - a random process with only two outcomes possible. The term *dichotomous* variable is also used indicating there are exactly two (mutually exclusive) outcomes: the event (A) and its negation: (A^c).

The variable X is said to give a *Bernoulli distribution* (often abbreviated to Bernoulli (p)).

4.4 Combinatorics

Suppose I have n *distinct* objects: let's say a set of red ping-pong balls that are identical except they are numbered $1, 2, \ldots, n$ to make them distinct. If these are arranged in a row, how many *permutations* (different orderings) of the n objects is possible?

The answer is, the number of different orderings (permutations) of n distinct objects is:

$$\boxed{n \times (n-1) \times \cdots \times 2 \times 1 = n!} \quad . \tag{4.15}$$

Proof:
Suppose the number of permutations of n balls were M_n (which we don't yet know). But we can easily see that $M_1 = 1$ and $M_2 = 2$, for example. Then I bring in an extra ball, the $(n+1)$th ball, and this will give me M_{n+1} permutations. Consider a single arbitrary (random) permutation of the n balls lined up in a row. The new ball can be inserted at the start or at the end of this row, or squeezed in somewhere in the middle. Clearly there are $(n+1)$ possible places where it could go - all of these distinct. Since I have done this to an arbitrary n-ball permutation, I can do the same with any n-ball permutation. Each n-ball permutation will have $(n+1)$ new $(n+1)$-ball permutations.

So: $M_{n+1} = (n+1)M_n$. Then *reccurring* (that is repeating this formula), we have:

$$M_{n+1} = (n+1)M_n = (n+1)nM_{n-1} = (n+1)n(n-1)M_{n-2} = \cdots = (n+1)!M_1 = (n+1)!$$

This concludes our proof.

Now consider that we have these n distinct balls dropped into a bag and we choose one at a time at random (without replacement lining them up into a row of n balls. Then clearly there would be $n!$ different ordered sequences possible. That is, if we selected

If we replaced the balls after every selection, then in k selections we could have n^k possible orderings.

Now suppose that we have n (distinctly numbered) balls of which the first r balls are coloured red (and numbered $1, 2, \ldots, r$) and the remaining $n-r$ balls are white (and numbered $n-r+1, n-r+2, \ldots, n$). Consider those permutations (orderings) in which we are NOT interested in the ordering of the white balls. That is we consider the white balls *indistinguishable*. Then all permutations in which two white balls can be swapped will be the same for us, and if we have $n-r$ white balls there will be $(n-r)!$ ways of permutating them for *any* n-ball ordering. So there is a *fraction* of $(n-r)!$ repeats in our n-ball permutations.

Thus the number of distinct orderings of n objects of which r are distinguishable (and $n-r$ are indistinguishable) is:

$$\boxed{{}^nP_r = \frac{n!}{(n-r)!}} \quad . \tag{4.16}$$

We also see that ${}^nP_r$ is the number of distinct (different) ways in which r objects can be chosen from n objects (without replacement) but in which the ordering of the objects is important. Just consider an one-to-one correspondence between the objects and the numbered balls to see the analogy.

Consider an example of 4 balls, 2 of which are white (and identical) and the other two are red and numbered 1 and 2. Then the number of distinct sequences (orderings) is, according to (4.16), with $n = 4$ and $r = 2$: $4!/2! = 12$. We can see this explicitly by listing them:

$$12WW \quad 1W2W \quad 1WW2$$

where W denotes a white ball. Then swapping (permuting) 1 and 2 we get another 3 distinct permutations

$$21WW \quad 2W1W \quad 2WW1 \quad .$$

The last six are the arrangements:

$$W12W \quad W1W2 \quad , \quad W21W \quad W2W1 \quad , \quad WW12 \quad , \quad WW21 \quad .$$

So, in our bag with 4 balls, in which we did not pay attention to the number of the white balls we would have 12 sequences.

Finally suppose the ordering of the red balls was also of no interest to us. So we choose the balls from the bag (in sequence without replacement) until the bag is empty. Then the red balls would have $r!$ repetitions for every sequence of the white balls, and now this fraction of the permutations would not be considered as 'distinct'. So we have even fewer arrangements.

The number of distinct sequences (arrangements) of n objects of which r are distinct but in which ordering is not important is:

$$\boxed{{}^nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}} \quad . \tag{4.17}$$

The C is the familiar binomial coefficient and we read the symbol as 'n choose r '. That is, ${}^nC_r$ is the number of ways (sequences/arrangements) of choosing r (unordered) but distinct objects from n objects (without replacement).

So for the example of 4 balls, we have only:

$$\frac{4!}{2!2!} = 6$$

different combinations

$$RRWW \quad RWRW \quad RWWR$$
$$WRRW \quad WRWR \quad , \quad WWRR \quad .$$

4.5 Binomial distribution

Consider a sequence of coin tosses under identical conditions. On each toss the probabilities of HEADS and TAILS is as follows: $P(H) = p$, $P(T) = 1 - p \equiv q$, For a sequence of two tosses, each of which are independent events, we can calculate the probabilities of each outcome:

$$P(HH) = p.p = p^2 \quad , \quad P(HT) = P(TH) = p(1-p) = pq \quad , \quad P(TT) = q^2 \qquad .$$

In a sequence of n tosses, the probability that the first r tosses give HEADS , and all the subsequent $(n-r)$ tosses are TAILS is:

$$P(HHH\cdots HTT\cdots T) = p^r q^{n-r}.$$

Suppose we were simply interested in the probability of X HEADS in n tosses without regard to the ordering/sequence in which these occur. We can see that these sequences are just like the red and white balls discussed above. Then the number of possible choices of x objects from n, without regard to ordering, is the *binomial coefficient* (4.17):

$${}^nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad . \tag{4.18}$$

So if we think of the sequence of tosses, then the occurrence of a HEADS on the 10th toss, is like choosing the 10th object in an ordered row of n objects to be HEADS. Then our sequences of HEADS are just ways of picking x objects from a possible sequence of n.

Let X be the discrete random variable that equals the total number of HEADS in any ordering. Then the probability mass function for X is

$$P(X = x) = f_X(x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, \ldots, n \qquad . \tag{4.19}$$

X is said to have a *binomial probability distribution* with $0 \leq p \leq 1$ and $q \equiv 1 - p$. We note that, in accordance with the *binomial theorem*:

$$\sum_{x=0}^{n} f_X(x) = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} = (p+q)^n = (1)^n = 1 \quad . \tag{4.20}$$

Of courses this is as it should be: the total probability of all events should sum to 1. Some examples of the probability mass function for the binomial distribution are displayed in figure 4.2.

4.5.1 Example

The probability of HEADS occurring exactly once in 10 tosses of a coin for which $p = 0.4$ is:

$$f_X(x=1) = \left(\begin{array}{c} 10 \\ 1 \end{array} \right) p^1 q^{10-1} = 10(0.4)(0.6^9) \simeq 0.04 \quad .$$

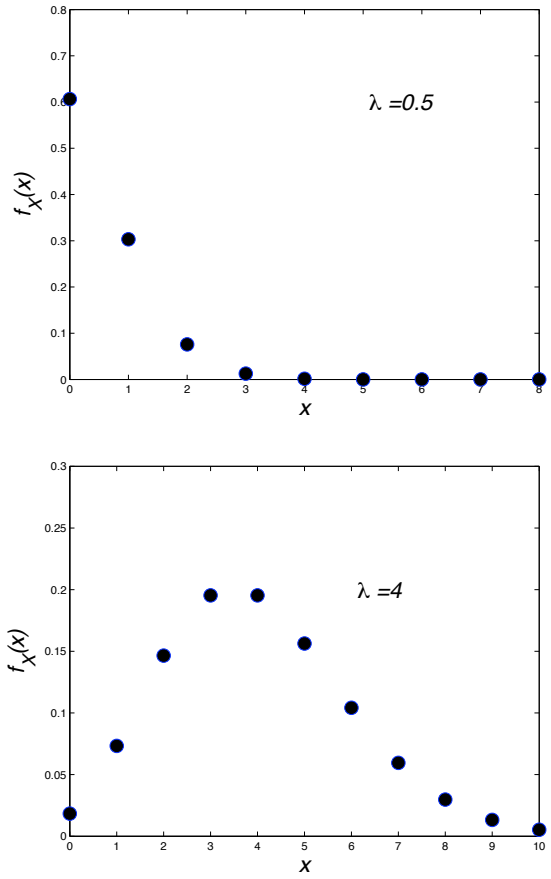

Figure 4.3: Examples of the probability mass function for the Poisson(λ) distribution: $f_X(x) = e^{-\lambda}\lambda^x/x!$. Top: $\lambda = 0.5$. Bottom: $\lambda = 4$.

4.6 Poisson distribution

A discrete random variable, often encountered, is the *Poisson distribution*. The random variable has an infinite range of values: $X = \{0, 1, 2, 3, \ldots\}$, and the distribution has a single parameter ($\lambda > 0$) that defines its shape (and hence all its properties). The probability mass function takes the form:

$$f_X(x) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad , \qquad x = 0, 1, 2, 3, \ldots \quad . \tag{4.21}$$

Examples of the Poisson distribution are shown in figure 4.3.

We can verify that the distribution is correctly *normalised* since:

$$\sum_{x=0}^{\infty} f_X(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{+\lambda} = 1 \quad . \tag{4.22}$$

The Poisson distribution also arises from a large number of Bernoulli trials in which the chance of success in each trial is very small. We know that in n Bernoulli trials, in which $0 \leq p \leq 1$ is the probability of success on each trial, the probability of x successes is the Binomial distribution (4.19) which can be written as:

$$P(X = x) = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} p^x (1-p)^{n-x} \quad , \quad x = 0, 1, 2, \ldots, n \quad . \tag{4.23}$$

Now suppose the success is vanishingly small, $p \to 0$, but that the trial is conducted many times, $n \to \infty$, such that their product tends to a finite value:

$$\lim_{p \to 0, n \to \infty} np = \lambda \quad . \tag{4.24}$$

We have in mind, perhaps, goals scored by a football team in which n could be the number of minutes in a game ~ 90 and p the probability of a goal in one minute intervals. Historically, the distribution was introduce in 1837 by Poisson, and a famous application was conducted in 1898 to a study of the number of soldiers accidentally killed in the Prussian army by kicks from the cavalry horses. Consider (4.22) and taking this limit of rare events ($p \to 0$) in a large number of trials ($n \to \infty$) we can make some *approximations*. For example, when n is very large, $n - 1 \approx n$, and in general (for finite x), $n - x \approx n$ so the numerator factor has the limit

$$\lim_{n \to \infty} n(n-1)(n-2)\cdots(n-x+1) = \lim_{n \to \infty} n^x \quad . \tag{4.25}$$

Therefore, writing $np = \lambda$, we have:

$$\lim_{p \to 0, n \to \infty} P(X = x) = \lim_{n \to \infty} \frac{\lambda^x}{x!} (1 - \lambda/n)^n \quad , \quad x = 0, 1, 2, \ldots, \tag{4.26}$$

where now, since $n \to \infty$, x can have values up to infinity. The Euler definition of the exponential function gives us:

$$\lim_{p \to 0, n \to \infty} P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad , \quad x = 0, 1, 2, \ldots . \tag{4.27}$$

which is the Poisson distribution. The fact that the Poisson distribution deals with rare events is one reason for its widespread use in the insurance industry, and some examples of this will be discussed later in the book.

4.7 Geometric distribution

Finally, the *geometric distribution* is defined for the discrete random variable, $X \in \{1, 2, 3, \ldots\}$, and by the parameter, $0 \leq p \leq 1$, with $q = 1 - p$.

$$f_X(x) = q^{x-1} p \quad , \quad x = 1, 2, 3, \ldots \quad . \tag{4.28}$$

This probability mass function is illustrated in figure 4.4.

Again, we can verify that the total probability sums to 1. Adding the terms, and changing the variable $x \to r$ where $r = x - 1$, we have:

$$\sum_x f_X(x) = \sum_{x=1}^{\infty} q^{x-1} p = p \sum_{r=0}^{\infty} q^r = \frac{p}{1-q} = \frac{p}{p} = 1 \qquad .$$

Here we have used the result for the infinite geometric series:

$$\sum_{r=0}^{\infty} x^r = \frac{1}{1-x} \qquad , \qquad |x| < 1 \qquad .$$

This distribution is yet another expression that can be related to Bernoulli trials. Let X denote the number of tosses required to get the *first* occurrence of 'heads', in which $0 \leq p \leq 1$, is the probability of heads for each toss. If the first heads arises on the xth toss, this means we had a sequence of $x-1$ tails preceding this event. The probability of a sequence of $x-1$ tails, followed by a heads is then given by the expression (4.28).

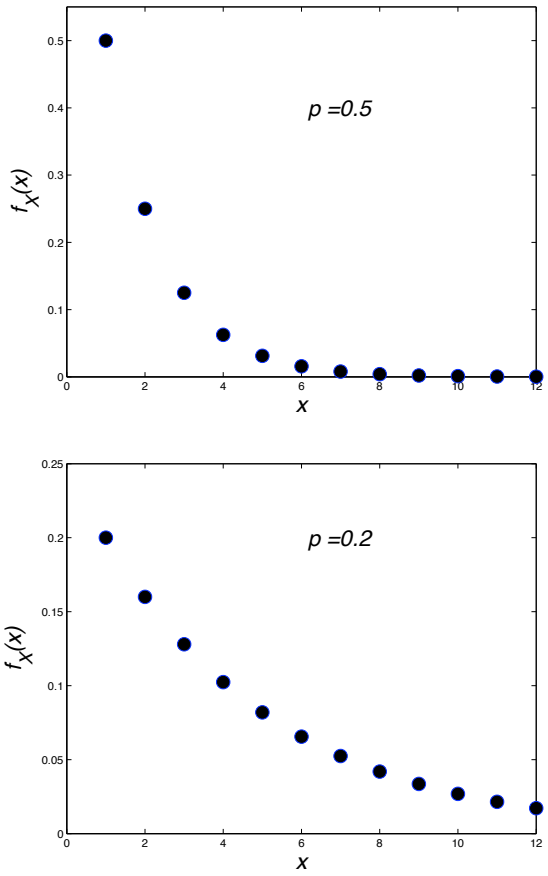

Figure 4.4: The geometric distribution illustrated: Geometric(p) distribution: $f_X(x) = (1-p)^{x-1}p$. Top: $p = 0.5$. Bottom: $p = 0.2$. Note that the probability always decreases with increasing x whatever the value of p. The larger the value of p the sharper the decrease.

4.8 Expectation (mean value, expected value)

The "expectation" of a discrete random variable X with probability mass function (pmf) $f_X(x)$

$$\mathbb{E}(X) = \sum_i x_i f_X(x_i) \quad . \tag{4.29}$$

where the sum is over the non-zero values of $f_X(x_i)$. The sum exists only if it is *uniformly convergent* (then the ordering is not important).

4.8.1 Expect the unexpected

Suppose we have a discrete variable X, such that $x = -2, -1, 1, 3$ and the probability mass for each of these values is $f_X(x) = \frac{1}{4}, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}$, respectively.

Then

$$\mathbb{E}(X) = \sum_i x_i f_X(x_i) = (-2)(\tfrac{1}{4}) + (-1)(\tfrac{1}{8}) + (1)(\tfrac{1}{4}) + (3)(\tfrac{3}{8}) = \tfrac{3}{4}$$

We recognise that the *expected value* does not, in general, equal any of the values of X. That is, the 'expected value' is not a value we ever 'expect' to get ! In that sense, it's a bit of a misnomer. By definition, the most likely outcome is the *mode* since it has the highest probability.

4.8.2 Expected values

Generalising the concept of expected value to any function of the random variable, we have the following identity:

$$\mathbb{E}\left(g(X)\right) = \sum_i g(x_i) f_X(x_i) \quad . \tag{4.30}$$

Note that, depending on the function, this sum might not exist! This could arise if the sum (or any of its terms) are divergent or infinite. For example, if:

$$f_X(x) = (6/\pi^2) \times \frac{1}{x^2} \quad , \quad x = 1, 2, 3, \ldots ,$$

then $\mathbb{E}\left(X\right)$ does not exist (or more properly we say it is *not defined* ! If the expectation does exist, it then follows that the expectation is a linear operator, in the sense that, for any constants a and b, we have:

$$\mathbb{E}\left(aX + b\right) = a\mathbb{E}\left(X\right) + b \qquad . \tag{4.31}$$

4.8.3 Expected value as 'best predictor'

Suppose there exists a discrete random variable, $X \in \{x_1, x_2, \ldots x_n\}$, with a known probability mass: $P(X = x_i) = f_X(x_i)$. Although it is impossible to predict the outcome, since the variable is random, one can make a prediction in the following sense.

Let us define the 'best predictor', a, of X as the value that minimizes the mean-square error. That is, a is such that:

$$g(a) \equiv \mathbb{E}\left((X - a)^2\right) \tag{4.32}$$

is a minimum.

Denoting $\mathbb{E}\left(X\right) = \mu$, we have,

$$\begin{aligned} g(a) &= \mathbb{E}\left((X - \mu + \mu - a)^2\right) \\ &= \mathbb{E}\left((X - \mu)^2\right) + 2(\mu - a)\mathbb{E}\left(X - \mu\right) + (\mu - a)^2 \\ &= \mathbb{E}\left((X - \mu)^2\right) + (\mu - a)^2 \end{aligned}$$

So the only dependence on the parameter a is in the second term. Since $(\mu - a)^2 \geq 0$, the value of a that minimizes this expression, $g(a)$, is simply: $a = \mu$. More formally, $g(a)$ has a minimum when:

$$\frac{d}{da} g(a) = 0 \qquad , \tag{4.33}$$

which leads to the same conclusion;

$$\frac{d}{da} g(a) = \mathbb{E}\left(2(X - a)\right) = 2\left(\mathbb{E}\left(X\right) - a\right) = 0 \qquad , \tag{4.34}$$

with solution:

$$a = \mu \qquad . \tag{4.35}$$

Thus, given any random variable (X), the best predictor, in the sense of minimizing the expected least-squares error, is the mean (expected) value: $a = \mathbb{E}\left(X\right)$.

As discussed already, the expected value, in general, will correspond to any event value (outcome). If this is the case then the 'best predictor' will never be correct. All we are asserting is that, the estimate (prediction) will outperform any choice of the set $\{x_1, x_2, \ldots x_n\}$ in being closest *on average*.

4.8.4 Median as predictor

There are alternatives ways of defining a *best estimator*, or what is termed a *measure of centrality* of a distribution. Suppose instead of (4.32) we chose the 'best predictor', b, of X as the value that minimizes absolute error defined as follows:

$$g(b) \equiv \mathbb{E}\left(|X - b|\right) \quad . \tag{4.36}$$

In this case we find b is the *median* of the distribution.

The proof is straightforward. Consider a random variable X with mass: $f_X(x)$ and distribution:

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \quad . \tag{4.37}$$

Then (4.36) can be written as:

$$g(b) = \sum_{x_i \leq b} (b - x_i) f_X(x_i) + \sum_{x_i > b} (x_i - b) f_X(x_i) \tag{4.38}$$

The value of b that minimizes this function is then, as before, the solution of $g'(b) = 0$ that is:

$$\sum_{x_i \leq b} f_X(x_i) - \sum_{x_i > b} f_X(x_i) = 0 \tag{4.39}$$

which can be written as:

$$F_X(b) - (1 - F_X(b)) = 0 \tag{4.40}$$

That is:

$$F_X(b) = \frac{1}{2} \quad , \tag{4.41}$$

which, referring to (4.12), is nothing other than the definition of the median, b, of the distribution,$F_X(x)$.

4.8.5 Predicting the lottery numbers

An observation by Galton in 1907 (reported in the journal *Nature* vol. 75, no. 1949, pp 450-451). *A weight-judging competition was carried out at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox, having been selected, competitors bought stamped and numbered cards for 6 pence each, on which to inscribe their respective names, addresses and estimates of what the ox would weigh after it had been slaughtered and 'dressed'.* You might be familiar with other versions of the competition such as guessing the number of sweets in a large glass jar.

Galton observed that the guesses were widely distributed (not quite a random normal distribution) but that the average value of the guesses was within 1% of the exact (correct) answer. He mischievously asserted that the *vox populi* (voice of the people) was remarkably accurate, and that this indicated the value of the democratic process. *'The result is, I think, more creditable to the trustworthiness of a democratic judgement than might have been expected'.*

So could one use a large sample of people to guess lottery numbers for example! Well guessing the weight of a cow is not quite the same as guessing a random variable. However, in guessing a random variable, as shown above, the average value is the best estimator. So does this translate to the sample mean as a best estimator for lottery numbers ?

Suppose we have M numbered balls, and we ask a group of N people to guess the outcome of the first ball to be drawn. The ball is chosen at random from the sample and has the value y. The N guesses, which are arrived at independently and may include repetition, are arranged in a ordered sequence: $\{x_1, x_2, x_3, \ldots, x_N\}$, where $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_N$. We make no assumption about how these guesses were obtained. In particular, we do *not* assume that they are identically distributed random variables. The sample mean is then:

$$\hat{x} \equiv \frac{1}{N}\left(x_1 + x_2 + x_3 + \cdots + x_N\right)$$

Now we choose, at *random*, an individual from the crowd. This person made the guess x_i, and the question we pose is the following. Is x_i (any individual's estimate) a better estimate of y than the sample

average $\hat{x}$ (the average of the guesses of the group) ? In mathematical terms, this is equivalent to the following proposition:

$$(x_i - y)^2 < (\hat{x} - y)^2 \qquad ? \tag{4.42}$$

Since y itself is a random variable, the inequality may or may not be satisfied; that is the individual may outperform the group. But we can ask whether, on *average*, this is the case. One logical choice for averaging would be with respect to the randomness of the choice of person - that is x_i. Another possibility would be to repeat the experiment over a number of weeks as our sample, and take the average result that way. Let us take expectations of the inequality (4.42), with respect to the random variable x_i, with respect to the sample of N people. Then:

$$\begin{aligned} \mathbb{E}\left((x_i - y)^2\right) &< \mathbb{E}\left((\hat{x} - y)^2\right) \\ (\hat{x} - y)^2 &< (\hat{x} - y)^2 \end{aligned}$$

which is clearly not true. Thus, on average (in the sense of choosing a person at random from the group), the guess of an individual will be worse than the sample average guess. So, in that sense, and only in that sense, can we assert, with mathematical justification, that two heads are better than one. That is, the crowd (large sample) is collectively superior to the individual in guessing numbers. Unfortunately, in the lottery you don't get any money for near misses.

Incidentally, as a game of chance, the lottery is strongly biased against any player playing for profit. For every pound wagered there is, on average, only 50p returned as prize money. Even the charities and organisations don't do that well, as a proportion of the amounts spent, with only 28p diverted towards good causes.