CHAPTER 5

Discrete Random Variables

This chapter is one of two chapters dealing with random variables. After introducing the notion of a random variable, we discuss discrete random variables: continuous random variables are left to the next chapter. Next on the menu we learn about calculating simple probabilities using a probability function. Several probability functions warrant special mention as they arise frequently in real-life situations. These are the probability functions for the so-called Geometric, Hypergeometric, Binomial and Poisson distributions. We focus on the physical assumptions underlying the application of these functions to real problems. Although we can use computers to calculate probabilities from these distributions, it is often convenient to use special tables, or even use approximate methods in which one probability function can be approximated quite closely by another.

The theoretical idea of "expected value" will be introduced. This leads to the notions of population mean and population standard deviation; population analogues of the sample mean and sample standard deviation that we met in Chapter 2. The population versions are derived for simple cases and just summarized for our four distributions mentioned above. Finally, the mean and standard deviations are obtained for $aX + b$ in terms of those for X.

5.1   Random Variables

At the beginning of Chapter 4 we talked about the idea of a random experiment. In statistical applications we often want to measure, or observe, different aspects or characteristics of the outcome of our experiment.

> A ***(random) variable*** is a type of measurement taken on the outcome of a random experiment.

We use upper-case letters X, Y, Z etc. to represent random variables. If our experiment consists of sampling an individual from some population we may be interested in measurements on the yearly income (X say), accommodation costs (Y say), blood pressure (Z), or some other characteristic of the individual chosen.

We use the term "measurement" loosely. We may have a variable $X =$ "marital status" with three categories "never married", "married", "previously married" with which we associate the numerical codes $1, 2$ and 3 respectively. Then the X-measurement of an individual who is currently married is $X = 2$. In Section 2.1 we distinguished between several types of random variables. In this chapter we concentrate on discrete random variables.

5.2 Probability Functions

5.2.1 Preliminaries

Example 4.2.1   Consider the experiment of tossing a coin twice and define the variable $X =$ "number of heads". A sample space for this experiment is given by $S = \{HH, HT, TH, TT\}$ and X can take values $0, 1$ and 2. Rather than write, "the X-measurement of TT is 0" we write $X(TT) = 0$. Similarly, $X(HT) = 1$, $X(TH) = 1$ and $X(HH) = 2$.

We use small letters x, y, z etc to represent possible values that the corresponding random variables X, Y, Z etc can take. The statement $X = x$ defines an event consisting of all outcomes with X-measurement equal to x. In Example 5.2.1, "$X = 1$" is the event $\{HT, TH\}$ while "$X = 2$" consists of $\{HH\}$. Thus we can assign probabilities to events of the form "$X = x$" as in Section 4.4.4 (alternative chapter on the web site) by adding the probabilities of all the outcomes which have X-measurement equal to x.

> The ***probability function*** for a discrete random variable X gives
> $$\mathrm{pr}(\boldsymbol{X = x})$$
> for every value x that X can take.

Where there is no possibility of confusion between X and some other variable, $\mathrm{pr}(X = x)$ is often abbreviated to $\mathbf{pr(\boldsymbol{x})}$. As with probability distributions, $0 \leq \mathrm{pr}(x) \leq 1$ and the values of $\mathrm{pr}(x)$ must add to one, i.e. $\sum_x \mathrm{pr}(x) = 1$. This provides a useful check on our calculations.

Example 5.2.2 Consider tossing a coin twice as in Example 5.2.1. If the coin is unbiased so that each of the four outcomes is equally likely we have $\text{pr}(0) = \text{pr}(TT) = \frac{1}{4}$, $\text{pr}(1) = \text{pr}(TH, HT) = \frac{2}{4}$ and $\text{pr}(2) = \text{pr}(HH) = \frac{1}{4}$. This probability function is conveniently represented as a table.

x	0	1	2
$\text{pr}(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The probabilities add to 1 as required. Values of x not represented in the table have probability zero.

Example 5.2.3 This is continuation of Example 4.4.6(c) (alternative chapter on the web site) in which a couple has children until they have at least one of each sex or a maximum of 3 children. The sample space and probability distribution can be represented as

Outcome	GGG	GGB	GB	BG	BBG	BBB
Probability	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Let X be the number of girls in the family. Then X takes values $0, 1, 2, 3$ with probability function

x	0	1	2	3
$\text{pr}(x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

The probabilities of the events "$X = 0$", "$X = 2$" and "$X = 3$" are easy to get as they correspond to single outcomes. However, $\text{pr}(X = 1) = \text{pr}(GB) + \text{pr}(BG) + \text{pr}(BBG) = \frac{1}{4} + \frac{1}{4} + \frac{1}{8} = \frac{5}{8}$. Note that $\sum_x \text{pr}(x) = 1$ as required.

Probability functions are best represented pictorially as a line graph. Fig. 5.2.1 contains a line graph of the probability function of Example 5.2.3.
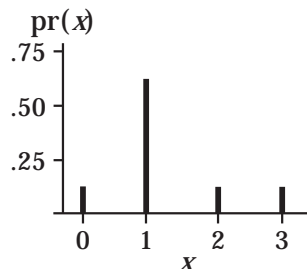

Figure 5.2.1 : Line graph of a probability function.

Example 5.2.4 Let us complicate the "tossing a coin twice" example (Examples 5.2.1 and 5.2.2) by allowing for a biased coin for which the probability of getting a "head" is p, say, where p is not necessarily $\frac{1}{2}$. As before, $\text{pr}(X = 0) =$

$\text{pr}(TT)$. By TT we really mean $T_1 \cap T_2$ or "tail on 1st toss" and "tail on 2nd toss".

Thus

$$\begin{aligned}
\text{pr}(TT) &= \text{pr}(T_1 \cap T_2) \\
&= \text{pr}(T_1) \times \text{pr}(T_2) \qquad \text{(as the tosses are independent)} \\
&= (1-p) \times (1-p) = (1-p)^2.
\end{aligned}$$

Similarly $\text{pr}(HT) = p(1-p)$, $\text{pr}(TH) = (1-p)p$ and $\text{pr}(HH) = p^2$. Thus $\text{pr}(X=0) = \text{pr}(TT) = (1-p)^2$, $\text{pr}(X=1) = \text{pr}(HT) + \text{pr}(TH) = 2p(1-p)$ and $\text{pr}(X=2) = \text{pr}(HH) = p^2$. A table is again a convenient representation of the probability function, namely:

x	0	1	2
$\text{pr}(x)$	$(1-p)^2$	$2p(1-p)$	p^2

Exercises on Section 5.2.1

1. Consider Example 4.4.6(b) (alternative chapter on the web site) and construct a table for the probability function of X, the number of girls in a 3-child family, where the probability of getting a girl is $\frac{1}{2}$.

2. Consider sampling 2 balls at random without replacement from a jar containing 2 black balls and 3 white balls. Let X be the number of black balls selected. Construct a table giving the probability function of X.

5.2.2   Skills in manipulating probabilities

We will often need to use a probability function to compute probabilities of events that contain more than one X value, most particularly those of the form $\text{pr}(X \geq a)$, e.g. $\text{pr}(X \geq 3)$, or of the form $\text{pr}(X > b)$ or $\text{pr}(X \leq c)$ or $\text{pr}(a \leq X \leq b)$. We will discuss the techniques involved while we still have our probability functions in the form of a simple table, rather than as a mathematical formula.

To find the probability of an event containing several X values, we simply add the probabilities of all the individual X values giving rise to that event.

Example 5.2.5   Suppose a random variable X has the following probability function,

x	1	3	4	7	9	10	14	18
$\text{pr}(X=x)$	0.11	0.07	0.13	0.28	0.18	0.05	0.12	0.06

then the probability that:

(a) X is at least 10 is

$$\text{pr}(X \geq 10) \;=\; \text{pr}(10) + \text{pr}(14) + \text{pr}(18) \;=\; 0.05 + 0.12 + 0.06 = 0.23;$$

(b) X is more than 10 is

$$\text{pr}(X > 10) \;=\; \text{pr}(X \geq 14) \;=\; \text{pr}(14) + \text{pr}(18) \;=\; 0.12 + 0.06 = 0.18;$$

(c) X is less than 4 is

$$\text{pr}(X < 4) \;=\; \text{pr}(X \leq 3) \;=\; \text{pr}(1) + \text{pr}(3) \;=\; 0.11 + 0.07 \;=\; 0.18;$$

(d) X is at least 4 and at most 9 is

$$\text{pr}(4 \leq X \leq 9) \;=\; \text{pr}(4) + \text{pr}(7) + \text{pr}(9) \;=\; 0.13 + 0.28 + 0.18 \;=\; 0.59;$$

(e) X is more than 3 and less than 10 is
$$\text{pr}(3 < X < 10) \;=\; \text{pr}(4 \leq X \leq 9) \;=\; 0.59 \text{ again.}$$

Theory   Suppose you want to evaluate the probability of an event A, but evaluating it involves many terms. If the complementary event, $\overline{A}$, involves fewer terms, it is helpful to use

$$\text{pr}(A) = 1 - \text{pr}(\overline{A}) \quad \text{or} \quad \text{pr}(A \text{ occurs}) = 1 - \text{pr}(A \text{ doesn't occur})$$

This is particularly valuable if complicated formulae have to be evaluated to get each individual value $\text{pr}(x)$.

Example 5.2.5 (cont.)   The probability that:
(f) X is at least 4 is

$$\text{pr}(X \geq 4) \;=\; 1 - \text{pr}(X \leq 3) \;=\; 1 - (0.11 + 0.07) \;=\; 0.82;$$

(g) X is at most 10 is

$$\text{pr}(X \leq 10) \;=\; 1 - \text{pr}(X \geq 14) \;=\; 1 - (0.12 + 0.06) \;=\; 0.82.$$

Exercises on Section 5.2.2

Suppose a discrete random variable X has probability function given by

x	3	5	7	8	9	10	12	16
$\text{pr}(X = x)$	0.08	0.10	0.16	0.25	0.20	0.03	0.13	0.05

[Note: The probabilities add to one.   Any value of x which is not in the table has probability zero.]

What is the probability that:
(a) $X > 9$?   (b) $X \geq 9$?   (c) $X < 3$?
(d) $5 \leq X \leq 9$?   (e) $4 < X < 11$?
(f) $X \geq 7$? Use the closer end of the distribution [cf. Example 5.2.5(f)].
(g) $X < 12$? Use the closer end of the distribution.

5.2.3 Using a formula to represent the probability function

Consider tossing a biased coin $\big($with $\mathrm{pr}(H) = p\big)$ until the first head appears. Then $S = \{H, TH, TTH, TTTH, \dots\}$. Let X be the total number of tosses performed. Then X can take values $1, 2, 3, \dots$ (an infinite number of values). We note, first of all, that

$$\mathrm{pr}(X = 1) = \mathrm{pr}(H) = p.$$

Then, using the notation of Example 5.2.4,

$$\begin{aligned}
\mathrm{pr}(X = 2) = \mathrm{pr}(TH) \quad &[\text{which means } \mathrm{pr}(T_1 \cap H_2)] \\
&= \mathrm{pr}(T_1)\mathrm{pr}(H_2) \quad \text{(as tosses are independent)} \\
&= (1 - p)p.
\end{aligned}$$

More generally,

$$\begin{aligned}
\mathrm{pr}(X = x) = \mathrm{pr}(\ \overbrace{TT \dots T}^{(x-1)\,\text{of them}}\ H) &= \mathrm{pr}(T_1 \cap T_2 \cap \dots \cap T_{x-1} \cap H_x) \\
&= \mathrm{pr}(T_1) \times \mathrm{pr}(T_2) \times \dots \times \mathrm{pr}(T_{x-1}) \times \mathrm{pr}(H_x) \quad \text{(independent tosses)} \\
&= (1 - p)^{x-1}p.
\end{aligned}$$

In this case the probability function is best represented as a mathematical function

$$\mathrm{pr}(X = x) = (1 - p)^{x-1}p, \qquad \text{for} \qquad x = 1, 2, 3 \dots\ .$$

This is called the *Geometric distribution*.[1] We write $X \sim \mathrm{Geometric}(p)$, where "$\sim$" is read "is distributed as".

> The ***Geometric distribution*** is the distribution of the number of tosses of a biased coin up to and including the first head.

The probability functions of most of the distributions we will use are, in fact, best presented as mathematical functions. The Geometric distribution is a good example of this as its probability function takes a simple form.

---

[1] The name Geometric distribution comes from the fact that the terms for $\mathrm{pr}(X = x)$ form a geometric series which can be shown to sum to 1.

Example 5.2.6   A *NZ Herald* data report quoted obstetrician Dr Freddie Graham as stating that the chances of a successful pregnancy resulting from implanting a frozen embryo are about 1 in 10. Suppose a couple who are desperate to have children will continue to try this procedure until the woman becomes pregnant. We will assume that the process is just like tossing a biased coin[2] until the first "head", with "heads" being analogous to "becoming pregnant". The probability of "becoming pregnant" at any "toss" is $p = 0.1$. Let X be the number of times the couple tries the procedure up to and including the successful attempt. Then X has a Geometric distribution.

(a) The probability of first becoming pregnant on the 4th try is

$$\text{pr}(X = 4) = 0.9^3 \times 0.1 = 0.0729.$$

(b) The probability of becoming pregnant before the 4th try is

$$\text{pr}(X \le 3) = \text{pr}(X = 1) + \text{pr}(X = 2) + \text{pr}(X = 3)$$
$$= 0.1 \ + \ 0.9 \times 0.1 \ + \ 0.9^2 \times 0.1 = 0.271.$$

(c) The probability that the successful attempt occurs either at the second, third or fourth attempt is

$$\text{pr}(2 \le X \le 4) = \text{pr}(X = 2) + \text{pr}(X = 3) + \text{pr}(X = 4)$$
$$= 0.9 \times 0.1 \ + \ 0.9^2 \times 0.1 \ + \ .9^3 \times 0.1 = 0.2439.$$

What we have seen in this example is an instance in which a trivial physical experiment, namely tossing a coin, provides a useful analogy (or model) for a situation of real interest. In the subsections to follow we will meet several simple physical models which have widespread practical applications. Each physical model has an associated probability distribution.

Exercises on Section 5.2.3

1. Suppose that 20% of items produced by a manufacturing production line are faulty and that a quality inspector is checking randomly sampled items. Let X be the number of items that are inspected up to and including the first faulty item. What is the probability that:

(a) the first item is faulty?

(b) the 4th item is the first faulty one?

(c) X is at least 4 but no more than 7?

(d) X is no more than 2?

(e) X is at least 3?

---

[2]We discuss further what such an assumption entails in Section 5.4.

2. In Example 5.2.6 a woman decides to give up trying if the first three attempts are unsuccessful. What is the probability of this happening? Another woman is interested in determining how many times t she should try in order that the probability of success before or at the tth try is at least $\frac{1}{2}$. Find t.

5.2.4 Using upper-tail probabilities

For the Binomial distribution (to follow), we give extensive tables of probabilities of the form $\mathbf{pr}(X \geq x)$ which we call upper-tail probabilities.

For the **_Geometric_** distribution, it can be shown that $\mathrm{pr}(X \geq x)$ has the particularly simple form

$$\mathrm{pr}(X \geq x) = (1-p)^{x-1}. \tag{1}$$

We will use this to learn to manipulate upper-tail probabilities to calculate probabilities of interest.

There are two important ideas we will need. Firstly, using the idea that $\mathrm{pr}(A \text{ occurs}) = 1 - \mathrm{pr}(A \text{ doesn't occur})$, we have in general

  (i) $\mathrm{pr}(X < x) = 1 - \mathrm{pr}(X \geq x)$.   Secondly

 (ii) $\mathrm{pr}(a \leq X < b) = \mathrm{pr}(X \geq a) - \mathrm{pr}(X \geq b)$.

We see, intuitively, that (ii) follows from the nonoverlapping intervals in Fig. 5.2.2. The interval from a up to but not including b is obtained from the interval from a upwards by removing the interval from b upwards.
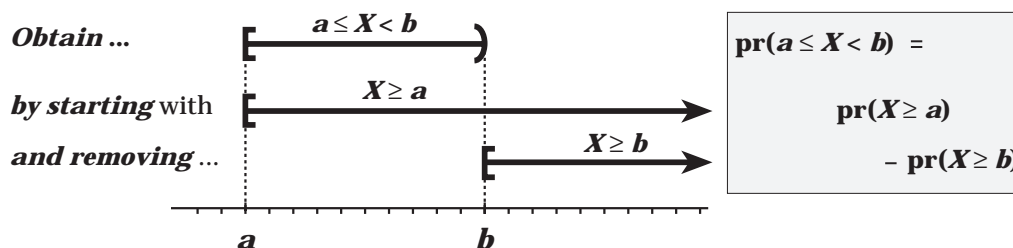

Figure 5.2.2 :   Probabilities of intervals from upper-tail probabilities.

The discrete random variables we concentrate on most take integer values $0, 1, 2, \ldots$. Therefore $\mathrm{pr}(X < x) = \mathrm{pr}(X \leq x-1)$, e.g. $\mathrm{pr}(X < 3) = \mathrm{pr}(x \leq 2)$. Similarly, $\mathrm{pr}(X > 3) = \mathrm{pr}(x \geq 4)$. For such random variables and integers a, b and x we have the more useful results

(iii) $\mathrm{pr}(X \leq x) = 1 - \mathrm{pr}(X \geq x+1)$,   and

(iv) $\mathrm{pr}(a \leq X \leq b) = \mathrm{pr}(X \geq a) - \mathrm{pr}(X \geq b+1)$.

These are not results one has to remember. Most people find it easy to "reinvent" them each time they want to use them, as in the following example.

Example 5.2.6 (cont.)   In Example 5.2.6, the probability of conceiving a child from a frozen embryo was $p = 0.1$ and X was the number of attempts up to and including the first successful one. We can use the above formula (1) to compute the following. The probability that:

(d) at least 2 attempts but no more than 4 are needed is

$$\mathrm{pr}(2 \leq X \leq 4) = \mathrm{pr}(X \geq 2) - \mathrm{pr}(X \geq 5) = 0.9^1 - 0.9^4 = 0.2439;$$

(e) fewer than 3 attempts are needed is

$$\mathrm{pr}(X < 3) \;=\; 1 - \mathrm{pr}(X \geq 3) \;=\; 1 - 0.9^2 \;=\; 0.19; \;\; \text{and}$$

(f) no more than 3 attempts are needed is

$$\mathrm{pr}(X \leq 3) \;=\; 1 - \mathrm{pr}(X \geq 4) \;=\; 1 - 0.9^3 \;=\; 0.271.$$

Exercises on Section 5.2.4

We return to Problem 1 of Exercises 5.2.3 in which a quality inspector is checking randomly sampled items from a manufacturing production line for which 20% of items produced are faulty. X is the number of items that are checked up to and including the first faulty item. What is the probability that:

(**a**) X is at least 4?
(**b**) the first 3 items are not faulty?
(**c**) no more than 4 are sampled before the first faulty one?
(**d**) X is at least 2 but no more than 7?
(**e**) X is more than 1 but less than 8?

Quiz for Section 5.2

1. You are given probabilities for the values taken by a random variable. How could you check that the probabilities come from probability function?

2. When is it easier to compute the probability of the complementary event rather than the event itself?

3. Describe a model which gives rise to the Geometric probability function. Give three other experiments which you think could be reasonably modeled by the Geometric distribution.

5.3   The Hypergeometric Distribution

5.3.1   Preliminaries

To cope with the formulae to follow we need to be familiar with two mathematical notations, namely $n!$ and $\binom{n}{k}$.

The n! or n-factorial [3] notation

By 4! we mean $4 \times 3 \times 2 \times 1 = 24$. Similarly $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. Generally $n!$, which we read as "n-factorial", is given by

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

An important special case is $0!$. $\boxed{0! = 1 \text{ by definition}}$

Verify the following: $3! = 6$, $6! = 720$, $9! = 362880$.

The $\binom{n}{k}$ or "n choose k" notation

This is defined by $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

e.g. $\binom{6}{2} = \frac{6!}{4! \times 2!} = 15$, $\qquad \binom{9}{4} = \frac{9!}{4! \times 5!} = 126$, $\qquad \binom{15}{8} = \frac{15!}{8! \times 7!} = 6435$.

(Check these values using your own calculator)

Special Cases: For any positive integer n

$$\binom{n}{0} = \binom{n}{n} = 1, \qquad \binom{n}{1} = \binom{n}{n-1} = n.$$

Given that the composition of a selection of objects is important and not the order of choice, it can be shown that the number of ways of choosing k individuals (or objects) from n is $\binom{n}{k}$, read as "n choose k". For example, if we take a simple random sample (without replacement) of 20 people from a population of 100 people, there are $\binom{100}{20}$ possible samples,[4] each equally likely to be chosen. In this situation we are only interested in the composition of the sample e.g. how many males, how many smokers etc., and not in the order of the selection.

> Ignoring order, there are $\binom{n}{k}$ ways of choosing k objects from n.

If your calculator does not have the factorial function the following identities are useful to speed up the calculation of $\binom{n}{k}$.

(i) $\binom{n}{k} = \frac{n}{k} \times \frac{n-1}{k-1} \times \ldots \times \frac{(n-k+1)}{1}$, e.g. $\binom{12}{3} = \frac{12 \times 11 \times 10}{3 \times 2 \times 1}$, $\binom{9}{2} = \frac{9 \times 8}{2 \times 1}$.

(ii) $\binom{n}{k} = \binom{n}{n-k}$ e.g. $\binom{12}{9} = \binom{12}{3} = 220$.

---

[3] A calculator which has any statistical capabilities almost always has a built-in factorial function. It can be shown that $n!$ is the number of ways of ordering n objects.

[4] Note that two samples are different if they don't have *exactly* the same members.

When using (i) the fewer terms you have to deal with the better, so use (ii) if $(n - k)$ is smaller than k. These techniques will also allow you to calculate $\binom{n}{k}$ in some situations where $n!$ is too big to be represented by your calculator.

Exercises on Section 5.3.1

Compute the following

(a) $\binom{9}{0}$ **(b)** $\binom{7}{1}$ **(c)** $\binom{15}{4}$ **(d)** $\binom{12}{9}$ **(e)** $\binom{11}{11}$ **(f)** $\binom{25}{23}$

5.3.2 The urn model and the Hypergeometric distribution

Consider a barrel or urn containing N balls of which M are black and the rest, namely $N - M$, are white. We take a simple random sample (i.e. without replacement)[5] of size n and measure X, the number of black balls in the sample. The Hypergeometric distribution is the distribution of X under this sampling scheme. We write[6] $X \sim \text{Hypergeometric}(N, M, n)$.
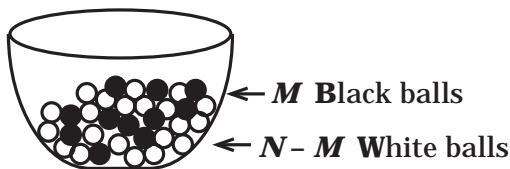
Sample
- **n** balls without replacement

Count **X** = # **Black** in sample

$X \sim$ Hypergeometric(N, M, n)

← **M Black balls**

← **$N - M$ White balls**

Figure 5.3.1 : The two-color urn model.

The Hypergeometric(N, M, n) distribution has probability function

$$\text{pr}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

for those values of x for which the probability function is defined.[7]

[**Justification:** Since balls of the same color are indistinguishable from one another, we can ignore the order in which balls of the same color occur. For this experiment, namely taking a sample of size n without replacement, an outcome corresponds to a possible sample. In the sample space there are $b = \binom{N}{n}$ possible samples (outcomes) which can be selected without regard to order, and each of these is equally likely (because of the random sampling). To find $\text{pr}(X = x)$, we need to

---

[5]Recall from Section 1.1.1 that a *simple random sample* is taken **without** replacement. This could be a accomplished by randomly mixing the balls in the urn and then either blindly scooping out n balls, or by blindly removing n balls one at a time without replacement.

[6]Recall that the "$\sim$" symbol is read as "is distributed as".

[7]If $n \leq M$, the number of black balls in the urn, and $n \leq N - M$, the number of white balls, x takes values $0, 1, 2 \dots, n$. However if $n > M$, the number of black balls in the sample must be no greater than the number in the urn, i.e. $X \leq M$. Similarly $(n - x)$, the number of white balls in the sample can be no greater than $(N - M)$, the number in the urn. Thus x ranges over those values for which $\binom{M}{x}$ and $\binom{N-M}{n-x}$ are defined. Mathematically we have $x = a, a + 1, \dots, b$, where $a = \max(0, n + M - N)$ and $b = \min(M, n)$.

know the number of outcomes making up the event $X = x$. There are $\binom{M}{x}$ ways of choosing the black balls without regard to order, and each sample of x black balls can be paired with any one of $\binom{N-M}{n-x}$ possible samples of $(n - x)$ white balls. Thus the total number of samples that consist of x black balls and $(n - x)$ white balls is $a = \binom{M}{x} \times \binom{N-M}{n-x}$. Since we have a sample space with equally likely outcomes, $\text{pr}(X = x)$ is the number of outcomes a giving rise to x divided by the total number of outcomes b, that is a/b.]

(a) If $N = 20$, $M = 12$, $n = 7$ then

$$\text{pr}(X = 0) = \frac{\binom{12}{0}\binom{8}{7}}{\binom{20}{7}} = \frac{1 \times 8}{77520} = 0.0001032,$$

$$\text{pr}(X = 3) = \frac{\binom{12}{3}\binom{8}{4}}{\binom{20}{7}} = \frac{220 \times 70}{77520} = 0.1987,$$

$$\text{pr}(X = 5) = \frac{\binom{12}{5}\binom{8}{2}}{\binom{20}{7}} = \frac{792 \times 28}{77520} = 0.2861,$$

(b) If $N = 12$, $M = 9$, $n = 4$ then

$$\text{pr}(X = 2) = \frac{\binom{9}{2}\binom{3}{2}}{\binom{12}{4}} = \frac{36 \times 3}{495} = 0.218.$$

Exercises on Section 5.3.2

Use the Hypergeometric probability formula to compute the following:
1. If $N = 17$, $M = 10$, $n = 6$, what is
   (**a**) $\text{pr}(X = 0)$?   (**b**) $\text{pr}(X = 3)$?   (**c**) $\text{pr}(X = 5)$?   (**d**) $\text{pr}(X = 6)$?
2. If $N = 14$, $M = 7$, $n = 5$, what is
   (**a**) $\text{pr}(X = 0)$?   (**b**) $\text{pr}(X = 1)$?   (**c**) $\text{pr}(X = 3)$?   (**d**) $\text{pr}(X = 5)$?

5.3.3 Applications of the Hypergeometric distribution

The two color urn model gives a physical analogy (or model) for any situation in which we take a simple random sample of size n (i.e. without replacement) from a finite population of size N and count X, the number of individuals (or objects) in the sample who have a characteristic of interest. With a sample survey, black balls and white balls may correspond variously to people who do (black balls) or don't (white balls) have leukemia, people who do or don't smoke, people who do or don't favor the death penalty, or people who will or won't vote for a particular political party. Here N is the size of the population, M is the number of individuals in the population with the characteristic of interest, while X measures the number with that characteristic in a sample of

size n. In all such cases the probability function governing the behavior of X is the Hypergeometric(N, M, n).

The reason for conducting surveys as above is to estimate M, or more often the proportion of "black" balls $p = M/N$, from an observed value of x. However, before we can do this we need to be able to calculate Hypergeometric probabilities. For most (but not all) practical applications of Hypergeometric sampling, the numbers involved are large and use of the Hypergeometric probability function is too difficult and time consuming. In the following sections we will learn a variety of ways of getting approximate answers when we want Hypergeometric probabilities. When the sample size n is reasonably small and $\frac{n}{N} < 0.1$, we get approximate answers using Binomial probabilities (see Section 5.4.2). When the sample sizes are large, the probabilities can be further approximated by Normal distribution probabilities (see web site material for Chapter 6 about Normal approximations). Many applications of the Hypergeometric with small samples relate to gambling games. We meet a few of these in the Review Exercises for this chapter. Besides the very important application to surveys, urn sampling (and the associated Hypergeometric distribution) provides the basic model for acceptance sampling in industry (see the Review Exercises), the capture-recapture techniques for estimating animal numbers in ecology (discussed in Example 2.12.2), and for sampling when auditing company accounts.

Example 5.3.1   Suppose a company fleet of 20 cars contains 7 cars that do not meet government exhaust emissions standards and are therefore releasing excessive pollution. Moreover, suppose that a traffic policeman randomly inspects 5 cars. The question we would like to ask is how many cars is he likely to find that exceed pollution control standards?

This is like sampling from an urn. The $N = 20$ "balls" in the urn correspond to the 20 cars, of which $M = 7$ are "black" (i.e. polluting). When $n = 5$ are sampled, the distribution of X, the number in the sample exceeding pollution control standards has a Hypergeometric$(N = 20, M = 7, n = 5)$ distribution. We can use this to calculate any probabilities of interest.

For example, the probability of no more than 2 polluting cars being selected is

$$\mathrm{pr}(X \le 2) = \mathrm{pr}(X = 0) + \mathrm{pr}(X = 1) + \mathrm{pr}(X = 2)$$
$$= 0.0830 + 0.3228 + 0.3874 = 0.7932.$$

Exercises on Section 5.3.3

1. Suppose that in a set of 20 accounts 7 contain errors. If an auditor samples 4 to inspect, what is the probability that:

(**a**) No errors are found?

(**b**) 4 errors are found?

(c) No more than 2 errors are found?

(d) At least 3 errors are found?

2. Suppose that as part of a survey, 7 houses are sampled at random from a street of 40 houses in which 5 contain families whose family income puts them below the poverty line. What is the probability that:

(a) None of the 5 families are sampled?

(b) 4 of them are sampled?

(c) No more than 2 are sampled?

(d) At least 3 are sampled?

Case Study 5.3.1 The Game of Lotto

Variants of the gambling game LOTTO are used by Governments in many countries and States to raise money for the Arts, charities, and sporting and other leisure activities. Variants of the game date back to the Han Dynasty in China over 2000 years ago (Morton, 1990). The basic form of the game is the same from place to place. Only small details change, depending on the size of the population involved. We describe the New Zealand version of the game and show you how to calculate the probabilities of obtaining the various prizes. By applying the same ideas, you will be able to work out the probabilities in your local game.

Our version is as follows. At the cost of 50 cents, a player purchases a "board" which allows him or her to choose 6 different numbers between 1 and 40.[8] For example, the player may choose 23, 15, 36, 5, 18 and 31. On the night of the Lotto draw, a sampling machine draws six balls at random without replacement from forty balls labeled 1 to 40. For example the machine may choose $25, 18, 33, 23, 12$, and 31. These six numbers are called the "winning numbers". The machine then chooses a 7th ball from the remaining 34 giving the so-called "bonus number" which is treated specially. Prizes are awarded according to how many of the winning numbers the player has picked. Some prizes also involve the bonus number, as described below.

Prize Type	Criterion
Division 1	All 6 winning numbers.
Division 2	5 of the winning numbers plus the bonus number.
Division 3	5 of the winning numbers but not the bonus number.
Division 4	4 of the winning numbers.
Division 5	3 of the winning numbers plus the bonus.

The Division 1 prize is the largest prize; the Division 5 prize is the smallest. The government makes its money by siphoning off 35% of the pool of money paid in

---

[8]This is the usual point of departure. Canada's Lotto 6/49 uses 1 to 49, whereas many US states use 1 to 54.

by the players, most of which goes in administrative expenses. The remainder, called the "prize pool", is redistributed back to the players as follows. Any Division 1 winners share 35% of the prize pool, Division 2 players share 5%, Division 3 winners share 12.5%, Division 4 share 27.5% and division 5 winners share 20%.

If we decide to play the game, what are the chances that a single board we have bought will bring a share in one of these prizes?

It helps to think about the winning numbers and the bonus separately and apply the two-color urn model to X, the number of matches between the player's numbers and the winning numbers. For the urn model we regard the player's numbers as determining the white and black balls. There are $N = 40$ balls in the urn of which $M = 6$ are thought of as being black; these are the 6 numbers the player has chosen. The remaining 34 balls, being the ones the player has not chosen, are thought of as the white balls in the urn. Now the sampling machine samples $n = 6$ balls at random without replacement. The distribution of X, the number of matches (black balls), is therefore Hypergeometric($N = 40, M = 6, n = 6$) so that

$$\text{pr}(X = x) = \frac{\binom{6}{x}\binom{34}{6-x}}{\binom{40}{6}}.$$

Now

$$\text{pr}(\text{Division 1 prize}) = \text{pr}(X = 6) = \frac{1}{\binom{40}{6}} = \frac{1}{3838380} = 2.605 \times 10^{-7},$$

$$\text{pr}(\text{Division 4 prize}) = \text{pr}(X = 4) = \frac{\binom{6}{4}\binom{34}{2}}{\binom{40}{6}} = \frac{8415}{3838380} = 0.0022.$$

The other three prizes involve thinking about the bonus number as well. The calculations involve conditional probability.

$$\text{pr}(\text{Division 2 prize}) = \text{pr}(X = 5 \cap \text{bonus}) = \text{pr}(X = 5)\text{pr}(\text{bonus} \mid X = 5).$$

If the number of matches is 5, then one of the player's balls is one of the 34 balls still in the sampling machine when it comes to choosing the bonus number. The chance that the machine picks the player's ball is therefore 1 in 34, i.e. $\text{pr}(\text{bonus} \mid X = 5) = 1/34$. Also from the Hypergeometric formula, $\text{pr}(X = 5) = 204/3838380$. Thus

$$\text{pr}(\text{Division 2 prize}) = \frac{204}{3838380} \times \frac{1}{34} = 1.563 \times 10^{-6}.$$

Arguing in the same way

$$\text{pr(Division 3 prize)} = \text{pr}(X = 5)\text{pr(no bonus} \mid X = 5)$$

$$= \text{pr}(X = 5) \times \frac{33}{34} = 5.158 \times 10^{-5}$$

and

$$\text{pr(Division 5 prize)} = \text{pr}(X = 3)\text{pr(bonus} \mid X = 3)$$

$$= pr(X = 3) \times \frac{3}{34} = 0.002751.$$

There are many other interesting aspects of the Lotto game and these form the basis of a number of the Review exercises.

Quiz for Section 5.3

1. Explain in words why $\binom{n}{k} = \binom{n}{n-k}$. (Section 5.3.1)

2. Can you describe how you might use a table of random numbers to simulate sampling from an urn model with $N = 100$, $M = 10$ and $n = 5$? (See Section 1.5.1)

3. A lake contains N fish of which M are tagged. A catch of n fish yield X tagged fish. We propose using the Hypergeometric distribution to model the distribution of X. What physical assumptions need to hold (e.g. fish do not lose their tags)? If the scientist who did the tagging had to rely on local fishermen to return the tags, what further assumptions are needed?

5.4   The Binomial Distribution

5.4.1   The "biased-coin-tossing" model

Suppose we have a biased or unbalanced coin for which the probability of obtaining a head is p $\big($i.e. $\text{pr}(H) = p\big)$. Suppose we make a fixed number of tosses, n say, and record the number of heads. X can take values $0, 1, 2, \ldots , n$. The probability distribution of X is called the ***Binomial distribution***. We write $\boldsymbol{X} \sim \text{Bin}(\boldsymbol{n},\ \boldsymbol{p})$.

> The distribution of the number of heads in $\boldsymbol{n}$ tosses of a biased coin is called the ***Binomial distribution***.
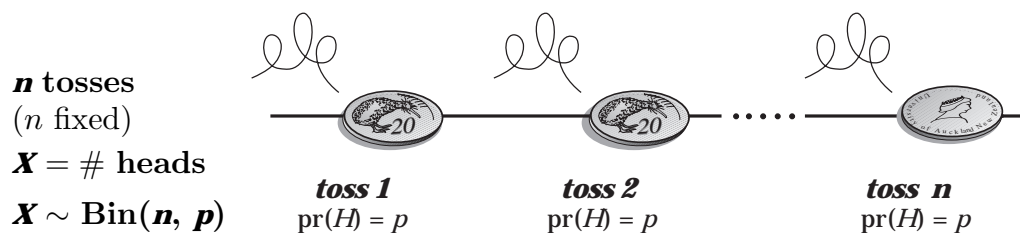

Figure 5.4.1 :   Biased coin model and the Binomial distribution.

The probability function for X, when $X \sim \text{Bin}(n, p)$, is

$$\text{pr}(\boldsymbol{X} = \boldsymbol{x}) = \binom{\boldsymbol{n}}{\boldsymbol{x}} \boldsymbol{p^x} (1 - \boldsymbol{p})^{\boldsymbol{n-x}} \quad \text{for} \quad \boldsymbol{x} = \boldsymbol{0, 1, 2, \ldots, n} .$$

[Justification:

The sample space for our coin tossing experiment consists of the list of all possible sequences of heads and tails of length n. To find $\text{pr}(X = x)$, we need to find the probability of each sequence (outcome) and then sum these probabilities over all such sequences which give rise to x heads. Now

$$\text{pr}(X = 0) = \text{pr}(\overbrace{TT \ldots T}^{n \text{ of them}}) = \text{pr}(T_1 \cap T_2 \cap \ldots \cap T_n) \quad \text{using the notation of Section 5.2.3}$$

$$= \text{pr}(T_1)\text{pr}(T_2) \ldots \text{pr}(T_n) \quad \text{as tosses are independent}$$

$$= (1 - p)^n,$$

and $\quad \text{pr}(X = 1) = \text{pr}(\{H \overbrace{T \ldots T}^{(n-1)}, THT \ldots T, \ldots, T \ldots TH\}),$

where each of these n outcomes consists of a sequence containing 1 head and $(n - 1)$ tails. Arguing as above, $\text{pr}(HT \ldots T) = \text{pr}(THT \ldots T) = \ldots = \text{pr}(T \ldots TH) = p(1 - p)^{n-1}$. Thus

$$\begin{aligned}\text{pr}(X = 1) &= \text{pr}(HT \ldots T) + \text{pr}(THT \ldots T) + \ldots + \text{pr}(T \ldots TH) \\ &= np(1 - p)^{n-1}.\end{aligned}$$

We now try the general case. The outcomes in the event "$X = x$" are the sequences containing x heads and $(n - x)$ tails, e.g. $\overbrace{HH \ldots H}^{x} \overbrace{T \ldots T}^{n-x}$ is one such outcome. Then arguing as for $\text{pr}(X = 1)$, any particular sequences of x heads and $(n - x)$ tails has probability $p^x(1 - p)^{n-x}$. There are $\binom{n}{x}$ such sequences or outcomes as $\binom{n}{x}$ is the number of ways of choosing where to put the x heads in the sequence (the remainder of the sequence is filled with tails). Thus adding $p^x(1 - p)^{n-x}$ for each of the $\binom{n}{x}$ sequences gives the formula $\binom{n}{x} p^x(1 - p)^{n-x}$.**]**

Calculating probabilities: Verify the following using both the formula and the Binomial distribution (individual terms) table in Appendix A2.

If $\qquad\qquad\qquad\qquad X \sim \text{Bin}(n = 5, p = 0.3),$

$$\text{pr}(2) \overset{\text{means}}{=} \text{pr}(X = 2) = \binom{5}{2}(.3)^2(.7)^3 = 0.3087,$$

$$\text{pr}(0) = \binom{5}{0}(0.3)^0(0.7)^5 = (0.7)^5 = 0.1681.$$

If $X \sim \text{Bin}(n = 9, p = 0.2)$ then $\text{pr}(0) = 0.1342$, $\text{pr}(1) = 0.3020$, and $\text{pr}(6) = 0.0028$.

Exercises on Section 5.4.1

The purpose of these exercises is to ensure that you can confidently obtain Binomial probabilities using the formula above and the tables in Appendices A2 and A2Cum.

Obtain the probabilities in questions 1 and 2 below using both the Binomial probability formula and Appendix A2.

1. If $X \sim \text{Binomial}(n = 10, p = 0.3)$, what is the probability that

    (a) $\text{pr}(X = 0)$?   **(b)** $\text{pr}(X = 3)$?   **(c)** $\text{pr}(X = 5)$?   **(d)** $\text{pr}(X = 10)$?

2. If $X \sim \text{Binomial}(n = 8, p = 0.6)$, what is the probability that

    (a) $\text{pr}(X = 0)$?   **(b)** $\text{pr}(X = 2)$?   **(c)** $\text{pr}(X = 6)$?   **(d)** $\text{pr}(X = 8)$?

Obtain the probabilities in questions 3 and 4 using the table in Appendix A2Cum.

3. (Continuation of question 1) Find

    (a) $\text{pr}(X \geq 3)$ **(b)** $\text{pr}(X \geq 5)$ **(c)** $\text{pr}(X \geq 7)$ **(d)** $\text{pr}(3 \leq X \leq 8)$

4. (Continuation of question 2) Find

    (a) $\text{pr}(X \geq 2)$ **(b)** $\text{pr}(X \geq 4)$ **(c)** $\text{pr}(X \geq 7)$ **(d)** $\text{pr}(4 \leq X \leq 7)$

5.4.2 Applications of the biased-coin model

Like the urn model, the biased-coin-tossing model is an excellent analogy for a wide range of practical problems. But we must think about the essential features of coin tossing before we can apply the model.

We must be able to view our experiment as a series of "tosses" or trials, where:

(1) each trial ("toss") has only 2 outcomes, success ("heads") or failure ("tails"),

(2) the probability of getting a success is the same, p say, for each trial, and

(3) the results of the trials are mutually independent.

In order to have a Binomial distribution, we also need

(4) X is the number of successes in a ***fixed*** number of trials.[9]

Examples 5.4.1

(a) Suppose we roll a standard die and only worry whether or not we get a particular outcome, say a six. Then we have a situation that is like

---

[9]Several other important distributions relate to the biased-coin-tossing model. Whereas the ***Binomial*** distribution is the distribution of the number of heads in a ***fixed*** number of tosses, the ***Geometric*** distribution (Section 5.2.3) is the distribution of the number of tosses up to and including the first head, and the ***Negative Binomial*** distribution is the number of tosses up until and including the kth (e.g. 4th) head.

tossing a biased coin. On any trial (roll of the die) there are only two possible outcomes, "success" (getting a six) and "failure" (not getting a six), which is condition (1). The probability of getting a success is the same as $\frac{1}{6}$ for every roll (condition (2)), and the results of each roll of the die are independent of one another (condition (3)). If we count the number of sixes in a *fixed number* of rolls of the die (condition (4)), that number has a Binomial distribution. For example, the number of sixes in 9 rolls of the die has a Binomial($n = 9, p = \frac{1}{6}$) distribution.

(b) In (a), if we rolled the die until we got the first six, we no longer have a Binomial distribution because we are no longer counting "the number of successes in a *fixed number* of trials". In fact the number of rolls till we get a six has a Geometric($p = \frac{1}{6}$) distribution (see Section 5.2.3).

(c) Suppose we inspect manufactured items such as transistors coming off a production line to monitor the quality of the product. We decide to sample an item every now and then and record whether it meets the production specifications. Suppose we sample 100 items this way and count X, the number that fail to meet specifications. When would this behave like tossing a coin? When would X have a Binomial($n = 100, p$) distribution for some value of p?

Condition (1) is met since each item meets the specifications ("success") or it doesn't ("failure"). For condition (2) to hold, the average rate at which the line produced defective items would have to be constant over time.[10] This would not be true if the machines had started to drift out of adjustment, or if the source of the raw materials was changing. For condition (3) to hold, the status of the current item i.e. whether or not the current item fails to meet specifications, cannot be affected by previously sampled items. In practice, this only happens if the items are sampled sufficiently far apart in the production sequence.

(d) Suppose we have a new headache remedy and we try it on 20 headache sufferers reporting regularly to a clinic. Let X be the number experiencing relief. When would X have a Binomial($n = 20, p$) distribution for some value of p?

Condition (1) is met provided we define clearly what is meant by relief: each person either experiences relief ("success") or doesn't ("failure"). For condition (2) to hold, the probability p of getting relief from a headache would have to be the same for everybody. Because so many things can cause headaches this will hardly ever be exactly true. (Fortunately the Binomial distribution still gives a good approximation in many situations even if p is only approximately constant). For condition (3) to hold, the people must be physically isolated so that they cannot compare notes and, because of the psychological nature of some headaches, change each other's chances of getting relief.

---

[10]In Quality Assurance jargon, the system would have to be ***in control***.

But the vast majority of practical applications of the biased-coin model are as an approximation to the two-color urn model, as we shall now see.

Relationship to the Hypergeometric distribution

Consider the urn in Fig. 5.3.1. If we sample our n balls one at a time at random *with* replacement then the biased-coin model applies exactly. A "trial" corresponds to sampling a ball. The two outcomes "success" and "failure" correspond to "black ball" and "white ball". Since each ball is randomly selected and then returned to the urn before the next one is drawn, successive selections are independent. Also, for each selection (trial), the probability of obtaining a success is constant at $p = \frac{M}{N}$, the proportion of black balls in the urn. Thus if X is the number of black balls in a sample of size n,

$$X \sim \text{Bin}\left(n, p = \frac{M}{N}\right).$$

However, with finite populations we do not sample *with* replacement. It is clearly inefficient to interview the same person twice, or measure the same object twice. Sampling from finite populations is done *without* replacement. However, if the sample size n is small compared with both M the number of "black balls" in the finite population and $(N-M)$, the number of "white balls", then urn sampling behaves very much like tossing a biased coin. Conditions of the biased-coin model (1) and (4) are still met as above. The probability of obtaining a "black ball" is always very close to $\frac{M}{N}$ and depends very little on the results of past drawings because taking out a comparatively few balls from the urn changes the proportion of black and white balls remaining very little.[11] Consequently, the probability of getting a black ball at any drawing depends very little on what has been taken out in the past. Thus when M and $N - M$ are large compared with the sample size n the biased-coin model is approximately valid and

$$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \simeq \binom{n}{x}p^x(1-p)^{n-x} \qquad \text{with} \qquad p = \frac{M}{N}.$$

If we fix p, the proportion of "black balls" in the finite population, and let the population size N tend to infinity i.e. get very big, the answers from the two formulae are indistinguishable, so that the Hypergeometric distribution is well approximated by the Binomial. However, when large values of N and M are involved, Binomial probabilities are much easier to calculate than Hypergeometric probabilities, e.g. most calculators cannot handle $N!$ for $N \geq 70$. Also, use of the Binomial requires only knowledge of the proportion p of "black balls". We need not know the population size N.

---

[11]If $N = 1000$, $M = 200$ and $n = 5$, then after 5 balls have been drawn the proportion of black balls will be $\frac{200-x}{1000-5}$, where x is the number of black balls selected. This will be close to $\frac{200}{1000}$ irrespective of the value of x.

The use of the Binomial as an approximation to the Hypergeometric distribution accounts for many, if not most, occasions in which the Binomial distribution is used in practice. The approximation works well enough for most practical purposes if the sample size is no bigger than about 10% of the population size ($\frac{n}{N} < 0.1$), and this is true even for large n.

Summary

If we take a sample of less than 10% from a large population in which a proportion p have a characteristic of interest, the distribution of X, the number in the sample with that characteristic, is approximately Binomial(n, p), where n is the sample size.

Examples 5.4.2

(a) Suppose we sample 100 transistors from a large consignment of transistors in which 5% are defective. This is the same problem as that described in Examples 5.4.1(c). However, there we looked at the problem from the point of view of the process producing the transistors. Under the Binomial assumptions, the number of defects X was Binomial($n = 100, p$), where p is the *probability* that the manufacturing process produced a defective transistor. If the transistors are independent of one another as far as being defective is concerned, and the probability of being defective remains constant, then it does not matter how we sample the process. However, suppose we now concentrate on the batch in hand and let p ($= 0.05$) be the *proportion* of defective transistors in the batch. Then, provided the 100 transistors are chosen at random, we can view the experiment as one of sampling from an urn model with black and white balls corresponding to defective and nondefective transistors, respectively, and 5% of them are black. The number of defectives in the random sample of 100 will have a Hypergeometric distribution. However, since the proportion sampled is small (we are told that we have a large consignment so we can assume that the portion sampled is less than 10%), we can use the Binomial approximation. Therefore using either approach, we see that X is approximately Binomial($n = 100, p = 0.05$). The main difference between the two approaches may be summed up as follows. For the Hypergeometric approach, the probability that a transistor is defective does not need to be constant, nor do successive transistors produced by the process need be independent. What is needed is that the sample is random. Also p now refers to a proportion rather than a probability.

(b) If 10% of the population are left-handed and we randomly sample 30, then X, the number of left-handers in the sample, is approximately a Binomial($n = 30, p = 0.1$) distribution.

(c) If 30% of a brand of wheel bearing will function adequately in continuous use for a year, and we test a sample of 20 bearings, then X, the number

of tested bearings which function adequately for a year is approximately Binomial($n = 20, p = 0.3$).

(d) We consider once again the experiment from Examples 5.4.1(d) of trying out a new headache remedy on 20 headache sufferers. Conceivably we could test the remedy on all headache sufferers and, if we did so, some proportion, p say, would experience relief. If we can assume that the 20 people are a random sample from the population of headache sufferers (which may not be true), then X, the number experiencing relief, will have a Hypergeometric distribution. Furthermore, since the sample of 20 can be expected to be less than 10% of the population of headache sufferers, we can use the Binomial approximation to the Hypergeometric. Hence X is approximately Binomial $(20, p)$ and we arrive at the same result as given in Examples 5.4.1(d). However, as with the transistor example, we concentrated there on the "process" of getting a headache, and p was a probability rather than a proportion.

Unfortunately p is unknown. If we observed $X = 16$ it would be nice to place some limits on the range of values p could plausibly take. The problem of estimating an unknown p is what typically arises in practice. We don't sample known populations! We will work towards solving such problems and finally obtain a systematic approach in Chapter 8.

Exercise on Section 5.4.2

Suppose that 10% of bearings being produced by a machine have to be scrapped because they do not conform to specifications of the buyer. What is the probability that in a batch of 10 bearings, at least 2 have to be scrapped?

Case Study 5.4.1   DNA fingerprinting

It is said that, with the exception of identical twins, triplets etc., no two individuals have identical DNA sequences. In 1985, Professor Alec Jeffries came up with a procedure (described below) that produces pictures or profiles from an individual's DNA that can then be compared with those of other individuals and coined the name DNA fingerprinting. These "fingerprints" are now being used in forensic medicine (e.g. for rape and murder cases) and in paternity testing. The FBI in the US has been bringing DNA evidence into court since December 1988. The first murder case in New Zealand to use DNA evidence was heard in 1990. In forensic medicine, genetic material used to obtain the profiles often comes in tiny quantities in the form of blood or semen stains. This causes severe practical problems such as how to extract the relevant substance and the aging of a stain. We are interested here in the application to paternity testing where proper blood samples can be taken.

Blood samples are taken from each of the three people involved, namely the mother, child and alleged father, and a DNA profile is obtained for each of them as follows. Enzymes are used to break up a person's DNA sequences into a unique collection of fragments. The fragments are placed on a sheet of jelly-like substance and exposed to an electric field which causes them to line

up in rows or bands according to their size and charge. The resulting pattern looks like a blurry supermarket bar-code (see Fig. 5.4.2).
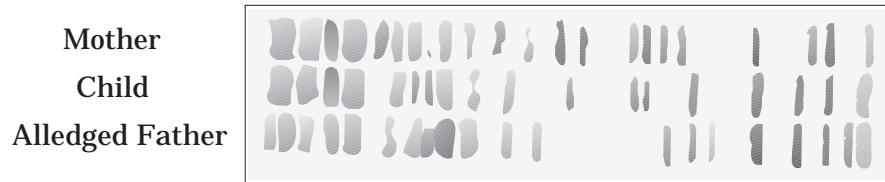

Figure 5.4.2 : DNA "fingerprints" for a mother, child and alleged father.

Under normal circumstances, each of the child's bands should match a corresponding band from either the father or the mother. To begin with we will look at two ways to rule the alleged father out of contention as the biological father.

If the child has bands which do not come from either the biological mother or the biological father, these bands must have come from genetic mutations which are rare. According to Auckland geneticists, mutations occur independently and the chance any particular band comes from a mutation is roughly

1 in 300. Suppose a child produces 30 bands (a fairly typical number) and let U be the number caused by mutations. On these figures, the distribution of U is Binomial($n = 30, p = 1/300$). If the alleged father is the real father, any bands that have not come from either parent must have come from biological mutation. Since mutations are rare, it is highly unlikely for a child to have a large number of mutations. Thus if there are too many unmatched bands we can say with reasonable confidence that the alleged father is not the father. But how many is too many? For a Binomial($n = 30, p = 1/300$) distribution $\text{pr}(U \geq 2) = 0.00454$, $\text{pr}(U \geq 3) = 0.000141$, $\text{pr}(U \geq 4) = 0.000003$. Two or more mutations occur for only about 1 case in every 200, three or more for 1 case in 10,000. Therefore if there are 2 or more unmatched bands we can be fairly sure he isn't the father.

Alternatively, if the alleged father is the biological father, the geneticists state that the probability of the child inheriting any particular band from him is about 0.5 (this is conservative on the low side as it ignores the chances that the mother and father have shared bands). Suppose the child has 30 bands and let V be the number which are shared by the father. On these figures, the distribution of V is Binomial($n = 30, p = 1/2$). Intuitively, it seems reasonable to decide that if the child has too few of the alleged father's bands, then the alleged father is not the real father. But how few is too few? For a Binomial($n = 30, p = 1/2$) distribution, $\text{pr}(V \leq 7) = 0.003$, $\text{pr}(V \leq 8) = 0.008$, $\text{pr}(V \leq 9) = 0.021$, and $\text{pr}(V \leq 10) = 0.049$. Therefore, when the child shares fewer than 9 bands with the alleged father it seems reasonable to rule him out as the actual father.

But are there situations in which we can say with reasonable confidence that the alleged father is the biological father? This time we will assume he is not

the father and try to rule that possibility out. If the alleged father is not the biological father, then the probability that the child will inherit any band which is the same as a particular band that the alleged father has is now stated as being about 0.25 (in a population without inbreeding). Suppose the child has 30 bands of which 16 are explained by the mother, leaving 14 unexplained bands. Let W be the number of those unexplained bands which are are shared by the alleged father. The distribution of W is Binomial($n = 14, p = 1/4$). If too many more than $\frac{1}{4}$ of the unexplained bands are shared by the alleged father, we can rule out the possibility that he is not the father. Again, how many is too many? For a Binomial($n = 14, p = 1/4$) distribution, $\text{pr}(W \geq 10) = 0.00034$, $\text{pr}(W \geq 11) = 0.00004$, $\text{pr}(W \geq 12) = 0.000003$. Therefore if there are 10 or more unexplained bands shared with the alleged father, we can be fairly sure he is the father.

There are some difficulties here that we have glossed over. Telling whether bands match or not is not always straight forward. The *New York Times* (29 January 1990) quotes Washington University molecular biologist Dr Philip Green as saying that "the number of possible bands is so great and the space between them is so close that you cannot with 100 percent certainty classify the bands". There is also another problem called band-shifting. Even worse, the New York Times talks about gross discrepancies between the results from different laboratories. This is more of a problem in forensic work where the amounts of material are too small to allow more than one profile to be made to validate the results. Finally, the one in four chance of a child sharing any particular band with another individual is too small if that man has a blood relationship with the child. If the alleged father is a full brother to the actual father that probability should be one chance in two.

Quiz for Section 5.4

1. A population of N people contains a proportion p of smokers. A random sample of n people is taken and X is the number of smokers in the sample. Explain why the distribution of X is Hypergeometric when the sample is taken without replacement, and Binomial when taken with replacement. (Section 5.4.2)

2. Give the four conditions needed for the outcome of a sequence of trials to be modeled by the Binomial distribution. (Section 5.4.2)

3. Describe an experiment in which the first three conditions for a Binomial distribution are satisfied but not the fourth. (Section 5.4.2)

4. The number of defective items in a batch of components can be modeled directly by the Binomial distribution, or indirectly (via the Hypergeometric distribution). Explain the difference between the two approaches. (Section 5.4.2)

5. In Case Study 5.4.1 it mentions that the probability of a child inheriting any particular band from the biological father is slightly greater than 0.5. What effect does this have on the probabilities $\text{pr}(V \leq 7)$ etc. given there? Will we go wrong if we work with the lower value of $p = 0.5$?

5.5   The Poisson Distribution

5.5.1   The Poisson probability function

A random variable X taking values $0, 1, 2 \ldots$ has a ***Poisson distribution*** if

$$\mathrm{pr}(\boldsymbol{X = x}) = \frac{\boldsymbol{e^{-\lambda}\lambda^x}}{\boldsymbol{x!}} \qquad \text{for} \qquad \boldsymbol{x = 0, \ 1, \ 2, \ \ldots} \quad .$$

We write $X \sim \mathrm{Poisson}(\lambda)$. [**Note** that $\mathrm{pr}(X = 0) = e^{-\lambda}$ as $\lambda^0 = 1$ and $0! = 1$.] For example, if $\lambda = 2$ we have

$$\mathrm{pr}(0) = e^{-2} = 0.135335, \quad \mathrm{pr}(3) = \frac{e^{-2} \times 2^3}{3!} = 0.180447.$$

As required for a probability function, it can be shown that the probabilities $\mathrm{pr}(X = x)$ all sum to 1.[12]

Learning to use the formula   Using the Poisson probability formula, verify the following: If $\lambda = 1$, $\mathrm{pr}(0) = 0.36788$, and $\mathrm{pr}(3) = 0.061313$.

The Poisson distribution is a good model for many processes as shown by the examples in Section 5.5.2. In most real applications, x will be bounded, e.g. it will be impossible for x to be bigger than 50, say. Although the Poisson distribution gives positive probabilities to all values of x going off to infinity, it is still useful in practice as the Poisson probabilities rapidly become extremely close to zero, e.g. If $\lambda = 1$, $\mathrm{pr}(X = 50) = 1.2 \times 10^{-65}$ and $\mathrm{pr}(X > 50) = 1.7 \times 10^{-11}$ (which is essentially zero for all practical purposes).

5.5.2   The Poisson process

Consider a type of event occurring randomly through time, say earthquakes. Let X be the number occurring in a unit interval of time. Then under the following conditions,[13] X can be shown mathematically to have a $\mathrm{Poisson}(\lambda)$ distribution.

(1) The events occur at a constant average rate of λ per unit time.

(2) Occurrences are independent of one another.

(3) More than one occurrence cannot happen at the same time.[14]

With earthquakes, condition (1) would not hold if there is an increasing or decreasing trend in the underlying levels of seismic activity. We would have to

---

[12]We use the fact that $e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^x}{x!} + \ldots$.

[13]The conditions here are somewhat oversimplified versions of the mathematical conditions necessary to formally (mathematically) establish the Poisson distribution.

[14]Technically, the probability of 2 or more occurrences in a time interval of length d tends to zero as d tends to zero.

be able to distinguish "primary" quakes from the aftershocks they cause and only count primary shakes otherwise condition (2) would be falsified. Condition (3) would probably be alright. If, instead, we were looking at car accidents, we couldn't count the number of damaged cars without falsifying (3) because most accidents involve collisions between cars and several cars are often damaged at the same instant. Instead we would have to count accidents as whole entities no matter how many vehicles or people were involved. It should be noted that except for the rate, the above three conditions required for a process to be Poisson do not depend on the unit of time. If λ is the rate per second then 60λ is the rate per minute. The choice of time unit will depend on the questions asked about the process.

The Poisson distribution[15] often provides a good description of many situations involving points randomly distributed in time or space,[16] e.g. numbers of microrganisms in a given volume of liquid, errors in sets of accounts, errors per page of a book manuscript, bad spots on a computer disk or a video tape, cosmic rays at a Geiger counter, telephone calls in a given time interval at an exchange, stars in space, mistakes in calculations, arrivals at a queue, faults over time in electronic equipment, weeds in a lawn and so on. In biology, a common question is whether a spatial pattern of where plants grow, or the location of bacterial colonies etc. is in fact random with a constant rate (and therefore Poisson). If the data does not support a randomness hypothesis, then in what ways is the pattern nonrandom? Do they tend to cluster (attraction) or be further apart than one would expect from randomness (repulsion). Case Study 5.5.1 gives an example of a situation where the data seems to be reasonably well explained by a Poisson model.

Case Study 5.5.1   Alpha particle[17] emissions.

In a 1910 study of the emission of alpha-particles from a Polonium source, Rutherford and Geiger counted the number of particles striking a screen in each of 2608 time intervals of length one eighth of a minute.
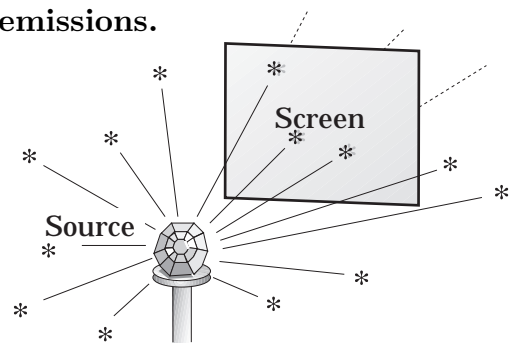
Their observed numbers of particles may have looked like this:
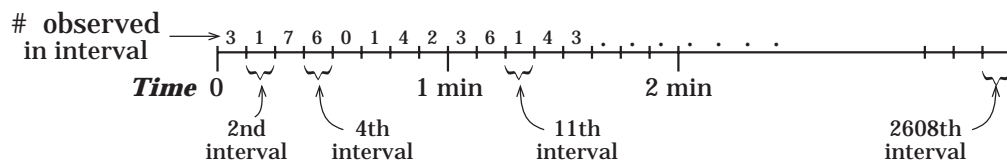
[15] Generalizations of the Poisson model allow for random distributions in which the average rate changes over time.

[16] For events occurring in space we have to make a mental translation of the conditions, e.g. λ is the average number of events per unit area or volume.

[17] A type of radioactive particle.

Rutherford and Geiger's observations are recorded in the repeated-data frequency table form of Section 2.5 giving the number of time intervals (out of the 2608) in which $0, 1, 2, 3$ etc particles had been observed. This data forms the first two columns of Table 5.5.1.

Could it be that the emission of alpha-particles occurs randomly in a way that obeys the conditions for a Poisson process? Let's try to find out. Let X be the number hitting the screen in a single time interval. If the process is random over time, $X \sim \text{Poisson}(\lambda)$ where λ is the underlying average number of particles striking per unit time. We don't know λ but will use the observed average number from the data as an estimate.[18] Since this is repeated data (Section 2.5.1), we use the repeated data formula for $\overline{x}$, namely

$$\overline{x} = \frac{\sum u_j f_j}{n} = \frac{0 \times 57 + 1 \times 203 + 2 \times 383 + \ldots + 10 \times 10 + 11 \times 6}{2608} = 3.870.$$

Let us therefore take $\lambda = 3.870$. Column 4 of Table 5.5.1 gives $\text{pr}(X = x)$ for $x = 0, 1, 2, \ldots, 10$, and using the complementary event,

$$\text{pr}(X \geq 11) = 1 - \text{pr}(X < 11) = 1 - \sum_{k=1}^{10} \text{pr}(X = k).$$

Table 5.5.1 : Rutherford and Geiger's Alpha-Particle Data

Number of Particles u_j	Observed Frequency f_j	Observed Proportion f_j/n	Poisson Probability $\text{pr}(X = u_j)$
0	57	0.022	0.021
1	203	0.078	0.081
2	383	0.147	0.156
3	525	0.201	0.201
4	532	0.204	0.195
5	408	0.156	0.151
6	273	0.105	0.097
7	139	0.053	0.054
8	45	0.017	0.026
9	27	0.010	0.011
10	10	0.004	0.004
11+	6	0.002	0.002

$$n = 2608$$

Column 2 gives the observed proportion of time intervals in which x particles hit the screen. The observed proportions appear fairly similar to the theoretical

---

[18]We have a problem here with the 11+ category. Some of the 6 observations there will be 11, some will be larger. We've treated them as though they are all exactly 11. This will make $\overline{x}$ a bit too small.

probabilities. Are they close enough for us to believe that the Poisson model is a good description? We will discuss some techniques for answering such questions in Chapter 11.

Before proceeding to look at some simple examples involving the use of the Poisson distribution, recall that the Poisson probability function with rate λ is

$$\mathrm{pr}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Example 5.5.1   While checking the galley proofs of the first four chapters of our last book, the authors found 1.6 printer's errors per page on average. We will assume the errors were occurring randomly according to a Poisson process. Let X be the number of errors on a single page. Then $X \sim \mathrm{Poisson}(\lambda = 1.6)$. We will use this information to calculate a large number of probabilities.

(**a**) The probability of finding no errors on any particular page is

$$\mathrm{pr}(X = 0) = e^{-1.6} = 0.2019.$$

(**b**) The probability of finding 2 errors on any particular page is

$$\mathrm{pr}(X = 2) = \frac{e^{-1.6}1.6^2}{2!} = 0.2584.$$

(**c**) The probability of no more than 2 errors on a page is

$$\mathrm{pr}(X \leq 2) = \mathrm{pr}(0) + \mathrm{pr}(1) + \mathrm{pr}(2)$$
$$= \frac{e^{-1.6}1.6^0}{0!} + \frac{e^{-1.6}1.6^1}{1!} + \frac{e^{-1.6}1.6^2}{2!}$$
$$= 0.2019 + 0.3230 + 0.2584 = 0.7833.$$

(**d**) The probability of more than 4 errors on a page is

$$\mathrm{pr}(X > 4) = \mathrm{pr}(5) + \mathrm{pr}(6) + \mathrm{pr}(7) + \mathrm{pr}(8) + \dots$$

so if we tried to calculate it in a straightforward fashion, there would be an infinite number of terms to add. However, if we use $\mathrm{pr}(A) = 1 - \mathrm{pr}(\overline{A})$ we get

$$\mathrm{pr}(X > 4) = 1 - \mathrm{pr}(X \leq 4) = 1 - [\mathrm{pr}(0) + \mathrm{pr}(1) + \mathrm{pr}(2) + \mathrm{pr}(3) + \mathrm{pr}(4)]$$
$$= 1 - (0.2019 + 0.3230 + 0.2584 + 0.1378 + 0.0551)$$
$$= 1.0 - 0.9762 = 0.0238.$$

(**e**) Let us now calculate the probability of getting a total of 5 errors on 3 consecutive pages.

Let Y be the number of errors in 3 pages. The only thing that has changed is that we are now looking for errors in bigger units of the manuscript so that the average number of events per unit we should use changes from 1.6 errors per page to $3 \times 1.6 = 4.8$ errors per 3 pages.

Thus
$$Y \sim \text{Poisson}(\lambda = 4.8)$$

and
$$\text{pr}(Y = 5) = \frac{e^{-4.8} 4.8^5}{5!} = 0.1747.$$

(f) What is the probability that in a block of 10 pages, exactly 3 pages have no errors?

There is quite a big change now. We are no longer counting events (errors) in a single block of material so we have left the territory of the Poisson distribution. What we have now is akin to making 10 tosses of a coin. It lands "heads" if the page contains no errors. Otherwise it lands "tails". The probability of landing "heads" (having no errors on the page) is given by (a), namely $\text{pr}(X = 0) = 0.2019$. Let W be the number of pages with no errors. Then

$$W \sim \text{Binomial}(n = 10, p = 0.2019)$$

and
$$\text{pr}(W = 3) = \binom{10}{3} (0.2019)^3 (0.7981)^7 = 0.2037.$$

(g) What is the probability that in 4 consecutive pages, there are no errors on the first and third pages, and one error on each of the other two?

Now none of our "brand-name" distributions work. We aren't counting events in a block of time or space, we're not counting the number of heads in a fixed number of tosses or a coin, and the situation is not analogous to sampling from an urn and counting the number of black balls.

We have to think about each page separately. Let X_i be the number of errors on the ith page. When we look at the number of errors in a single page, we are back to the Poisson distribution used in (a) to (c). Because errors are occurring randomly, what happens on one page will not influence what happens on any other page. In other words, pages are independent. The probability we want is

$$\text{pr}(X_1 = 0 \cap X_2 = 1 \cap X_3 = 0 \cap X_4 = 1)$$
$$= \text{pr}(X_1 = 0)\text{pr}(X_2 = 1)\text{pr}(X_3 = 0)\text{pr}(X_4 = 1),$$

as these events are independent.

For a single page, the probabilities of getting zero errors, $\text{pr}(X = 0)$, and one error, $\text{pr}(X = 1)$, can be read from the working of (b) above. These are 0.2019 and 0.3230 respectively. Thus

$$\text{pr}(X_1 = 0 \text{ and } X_2 = 1 \text{ and } X_3 = 0 \text{ and } X_4 = 1)$$
$$= 0.2019 \times 0.3230 \times 0.2019 \times 0.3230 = 0.0043. \quad \blacksquare$$

The exercises to follow are just intended to help you learn to use the Poisson distribution. No complications such as in (f) and (g) above are introduced. A block of exercises intended to help you to learn to distinguish between Poisson, Binomial and Hypergeometric situations is given in the Review Exercises for this chapter.

Exercises on Section 5.5.2

1. A recent segment of the *60 Minutes* program from CBS compared Los Angeles and New York City tap water with a range of premium bottled waters bought from supermarkets. The tap water, which was chlorinated and didn't sit still for long periods, had no detectable bacteria in it. But some of the bottled waters did![19] Suppose bacterial colonies are randomly distributed through water from your favorite brand at the average rate of 5 per liter.

(a) What is the probability that a liter bottle contains
   (i) 5 bacterial colonies?
   (ii) more than 3 but not as many as 7?

(b) What is the probability that a 100ml (0.1 of a liter) glass of bottled water contains
   (i) no bacterial colonies?   **(ii)** one colony?
   (iii) no more than 3?   **(iv)** at least 4?
   (v) more than 1 colony?

2. A Reuter's report carried by the *NZ Herald* (24 October 1990) claimed that Washington D.C. had become the murder capital of the US with a current murder rate running at 70 murders per hundred thousand people per year.[20]

(a) Assuming the murders follow a Poisson process,
   (i) what is the distribution of the number of murders in a district of 10,000 people in a month?
   (ii) What is the probability of more than 3 murders occurring in this suburb in a month?

(b) What practical problems can you foresee in applying an average yearly rate for the whole city to a particular district, say Georgetown? (It may help to pose this question in terms of your own city.)

(c) What practical problems can you foresee in applying a rate which is an average over a whole year to some particular month?

(d) Even if we could ignore the considerations in (b) and (c) there are other possible problems with applying a Poisson model to murders. Go through the conditions for a Poisson process and try to identify some of these other problems.

---

[19]They did remark that they had no reason to think that the bacteria they found were a health hazard.

[20]This is stated to be 32 times greater than the rate in some other (unstated) Western capitals.

[**Note:** The fact that the Poisson conditions are not obeyed does not imply that the Poisson distribution cannot describe murder-rate data well. If the assumptions are not too badly violated, the Poisson model could still work well. However, the question of how well it would work for any particular community could only be answered by inspecting the historical data.]

5.5.3 Poisson approximation to Binomial

Suppose $X \sim \text{Bin}(n = 1000, p = 0.006)$ and we wanted to calculate $\text{pr}(X = 20)$. It is rather difficult. For example, your calculator cannot compute 1000!.

Luckily, if n is large, p is small and np is moderate (e.g. $n \geq 100, np \leq 10$)

$$\binom{n}{x} p^x (1-p)^{n-x} \simeq \frac{e^{-\lambda} \lambda^x}{x!} \qquad \text{with} \qquad \lambda = np.$$

The Poisson probability is easier to calculate.[21] For $X \sim \text{Bin}(n = 1000, p = 0.006)$, $\text{pr}(X = 20) \simeq \frac{e^{-6} \times 6^{20}}{20!} = 3.725 \times 10^{-6}$.

Verify the following:

If $X \sim \text{Bin}(n = 200, p = .05)$, then $\lambda = 200 \times 0.05 = 10$, $\text{pr}(X = 4) \simeq 0.0189$, $\text{pr}(X = 9) \simeq 0.1251$. [True values are: 0.0174, 0.1277]

If $X \sim \text{Bin}(n = 400, p = 0.005)$, then $\lambda = 400 \times 0.005 = 2$, $\text{pr}(X = 5) \simeq 0.0361$. [True value is: 0.0359]

Exercises on Section 5.5.3

1. A chromosome mutation linked with color blindness occurs in one in every $10{,}000$ births on average. Approximately $20{,}000$ babies will be born in Auckland this year. What is the probability that

(**a**) none will have the mutation?

(**b**) at least one will have the mutation?

(**c**) no more than three will have the mutation?

2. Brain cancer is a rare disease. In any year there are about 3.1 cases per $100{,}000$ of population.[22] Suppose a small medical insurance company has $150{,}000$ people on their books. How many claims stemming from brain cancer should the company expect in any year? What is the probability of getting more than 2 claims related to brain cancer in a year?

Quiz for Section 5.5

1. The Poisson distribution is used as a probability model for the number of events X that occur in a given time interval. However the Poisson distribution allows X to take all values $0, 1, \ldots$ with nonzero probabilities, whereas in reality X is always bounded. Explain why this does not matter in practice. (Section 5.5.1)

---

[21] We only use the Poisson approximation if calculating the Binomial probability is difficult.
[22] US figures from *TIME* (24 December 1990, page 41).

2. Describe the three conditions required for X in Question 1 to have a Poisson distribution. (Section 5.5.2)

3. Is it possible for events to occur simultaneously in a Poisson process? (Section 5.5.2)

4. Consider each of the following situations. Determine whether or not they might be modeled by a Poisson distribution.
   (a) Counts per minute from a radioactive source.
   (b) Number of currants in a bun.
   (c) Plants per unit area, where new plants are obtained by a parent plant throwing out seeds.
   (d) Number of power failures in a week.
   (e) Pieces of fruit in a tin of fruit. (Section 5.5.2)

5. If X has a Poisson distribution, what "trick" do you use to evaluate $pr(X > x)$ for any value of x? (Section 5.5.2)

6. When can the Binomial distribution be approximated by the Poisson distribution? (Section 5.5.3)

5.6 Expected Values

5.6.1 Formula and terminology

Consider the data in Table 5.5.1 of Case Study 5.5.1. The data summarized there can be thought of as 2608 observations on the random event $X =$ "Number of particles counted in a single eighth of a minute time interval". We calculated the average (sample mean) of these 2608 observations using the repeated data formula of Section 2.5.1, namely

$$\overline{x} = \frac{\sum u_j f_j}{n} = \sum u_j \left(\frac{f_j}{n}\right),$$

where $X = u_j$ with frequency f_j. From the expression on the right hand side, we can see that each term of the sum is a possible number of particles, u_j, multiplied by the proportion of occasions on which u_j particles were observed. Now if we observed millions of time intervals we might expect the observed proportion of intervals in which u_j particles were observed to become very close to the true probability of obtaining u_j particles in a single time interval, namely $pr(X = u_j)$. Thus as n gets very large we might expect

$$\overline{x} = \sum u_j \frac{f_j}{n} \quad \text{to get close to} \quad \sum u_j pr(X = u_j).$$

This latter term is called the expected value of X, denoted $E(X)$. It is traditional to write the expected value formula using notation $x_1, x_2, \ldots$ and not $u_1, u_2, \ldots$. Hence if X takes values $x_1, x_2, \ldots$ then the ***expected value*** of X is

$$\mathbf{E}(\boldsymbol{X}) = \sum \boldsymbol{x_i} \, pr(\boldsymbol{X} = \boldsymbol{x_i}).$$

This is abbreviated to

$$\boxed{\mathbf{E}(\boldsymbol{X}) = \sum \boldsymbol{x} \, pr(\boldsymbol{X} = \boldsymbol{x}).}$$

As in the past we will abbreviate $\mathrm{pr}(X = x)$ as $\mathrm{pr}(x)$ where there is no possibility of ambiguity.

Example 5.6.1   In Example 5.2.3 we had a random variable with probability function

x	0	1	2	3
$\mathrm{pr}(x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Here

$$E(X) = \sum x\,\mathrm{pr}(x) = 0 \times \frac{1}{8} + 1 \times \frac{5}{8} + 2 \times \frac{1}{8} + 3 \times \frac{1}{8}$$
$$= 1.25.$$

Example 5.6.2   If $X \sim \mathrm{Binomial}(n = 3, p = 0.1)$ then X can take values $0, 1, 2, 3$ and

$$E(X) = \sum x\,\mathrm{pr}(x) = 0 \times \mathrm{pr}(0) \;+\; 1 \times \mathrm{pr}(1) \;+\; 2 \times \mathrm{pr}(2) \;+\; 3 \times \mathrm{pr}(3)$$
$$= 0 \;+\; 1 \times \binom{3}{1}(0.1)(0.9)^2 \;+\; 2 \times \binom{3}{2}(0.1)^2(0.9) \;+\; 3 \times \binom{3}{3}(0.1)^3(0.9)^0$$
$$= 0.3.$$

It is conventional to call $E(X)$ the *mean* of the distribution of X and denote it by the Greek symbol for "m", namely μ. If there are several random variables being discussed we use the name of the random variable as a subscript to μ so that it is clear which random variable is being discussed. If there is no possibility of ambiguity, the subscript on μ is often omitted.

$\mu_X = \mathbf{E(X)}$ is called the ***mean*** of the distribution of $\boldsymbol{X}$.

Suppose we consider a finite population of N individuals in which f_j individuals have X-measurement u_j, for $i = 1, 2, \ldots, k$. To make the illustration concrete, we will take our X-measurement to be the income for the previous year. Let X be the income on a single individual sampled at random from this population. Then $\mathrm{pr}(X = u_j) = \frac{f_j}{N}$ and $E(X) = \sum u_j \frac{f_j}{N}$ which is the ordinary (grouped) average of all the incomes across the N individuals in the whole population. Largely because $E(X)$ is the ordinary average (or mean) of all the values for a finite population, it is often called the *population mean*, and this name has stuck for $E(X)$ generally and not just when applied to finite populations.

$\mu_X = \mathbf{E(X)}$ is usually called the ***population mean***.

This is the terminology we will use. It is shorter than calling $\mu = E(X)$ the "mean of the distribution of X" and serves to distinguish it from the "sample mean", which is the ordinary average of a batch of numbers.

There is one more connection we wish to make. Just as $\overline{x}$ is the point where a dot plot or histogram of a batch of numbers balances (Fig. 2.4.1 in Section 2.4), μ_X is the point where the line graph of $pr(X = x)$ balances as in Fig. 5.6.1.

μ_X is the point where the line graph of $pr(X = x)$ balances.
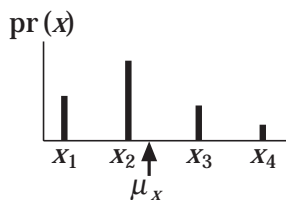
Figure 5.6.1 : The mean μ_X is the balance point.

Expected values tell us about long run average behavior in many repetitions of an experiment. They are obviously an important guide for activities that are often repeated. In business, one would often want to select a course of action which would give give maximum expected profits. But what about activities you will only perform once or twice? If you look at the expected returns on "investment", it is hard to see why people play state lotteries. The main purpose of such lotteries is to make money for government projects. NZ's Lotto game returns about 55 cents to the players for every dollar spent; so if you choose your Lotto numbers randomly, 55 cents is the expected prize money for an "investment" of \$1.00. In other words, if you played an enormous number of games, the amount of money you won would be about 55% of what you had paid out for tickets.[23] In the very long run, you are guaranteed to lose money. However, the main prizes are very large and the probabilities of winning them are tiny so that no-one ever plays enough games for this long-run averaging behavior to become reliable. Many people are prepared to essentially write off the cost of "playing" as being small enough that they hardly notice it, against the slim hope of winning a very large amount[24] that would make a big difference to their lives.[25]

---

[23]If you don't choose your numbers randomly things are more complicated. The payout depends upon how many people choose the same numbers as you and therefore share any prize. The expected returns would be better if you had some way of picking unpopular combinations of numbers.

[24]In New York State's Super Lotto in January 1991, 9 winners shared US\$90 million.

[25]This idea that profits and losses often cannot be expressed simply in terms of money is recognized by economists and decision theorists who work in terms of quantities they call **utilities**.

However you rationalize it though, not only are lottery players guaranteed to lose in the long run, the odds are that they will lose in the short run too. The Kansas State instant lottery of Review Exercises 4 Problem 19, returns about 47 cents in the dollar to the players. Wasserstein and Boyer [1990] worked out the chances of winning more than you lose after playing this game n times. The chances are best after 10 games at which point roughly one person in 6 will be ahead of the game. But after 100 games, only one person in 50 will have made a profit.

Exercises on Section 5.6.1

1. Suppose a random variable X has probability function

x	2	3	5	7
$\text{pr}(x)$	0.2	0.1	0.3	0.4

Find μ_X, the expected value of X.

2. Compute $\mu_X = \text{E}(X)$ where $X \sim \text{Binomial}(n=2, p=0.3)$.

5.6.2 Population standard deviation

We have discussed the idea of a mean of the distribution of X, $\mu = \text{E}(X)$. Now we want to talk be able to about the standard deviation of the distribution of X, which we will denote **sd(X)**. We use the same intuitive idea as for the standard deviation of a batch of numbers. The standard deviation is the square root of the average squared distance of X from the mean μ, but for distributions we use $\text{E}(.)$ to do the averaging.

The ***population standard deviation*** is

$$\text{sd}(X) = \sqrt{\text{E}[(X-\mu)^2]}$$

Just as with "population mean" and "sample mean", we use the word *population* in the term "population standard deviation" to distinguish it from "*sample* standard deviation", which is the standard deviation of a batch of numbers. The *population* standard deviation is the standard deviation of a distribution. It tells you about the spread of the distribution, or equivalently, how variable the random quantity is.

To compute $\text{sd}(X)$, we need to be able to compute $\text{E}[(X-\mu)^2]$. How are we to do this? For discrete distributions, just as we compute $\text{E}(X)$ as $\Sigma x_i \text{pr}(x_i)$, we calculate $\text{E}[(X-\mu)^2]$ using[26]

$$\text{E}[(X-\mu)^2] = \sum_i (x_i - \mu)^2 \text{pr}(x_i).$$

---

[26] An alternative formula which often simplifies the calculations is $\text{E}[(X-\mu)^2] = \text{E}(X^2) - \mu^2$.

[This is part of a general pattern, $E(X^2) = \Sigma x_i^2 \text{pr}(x_i)$, $E(e^X) = \Sigma e^{x_i} \text{pr}(x_i)$, or generally for any function $h(X)$ we compute $E[h(X)] = \Sigma h(x_i)\text{pr}(x_i)$.]

We now illustrate the use of the formula for $E[(X - \mu)^2]$.

Example 5.6.3 In Example 5.6.1, X has probability function

x	0	1	2	3
$\text{pr}(x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

and we calculated that $\mu = E(X) = 1.25$. Thus,

$$
\begin{aligned}
E[(X - \mu)^2] &= \sum (x_i - \mu)^2 \, \text{pr}(x_i) \\
&= (0 - 1.25)^2 \frac{1}{8} + (1 - 1.25)^2 \frac{5}{8} + (2 - 1.25)^2 \frac{1}{8} + (3 - 1.25)^2 \frac{1}{8} \\
&= 0.6875
\end{aligned}
$$

and $\text{sd}(X) = \sqrt{E[(X - \mu)^2]} = \sqrt{0.6875} = 0.8292$.

Example 5.6.4 If $X \sim \text{Binomial}(n = 3, p = 0.1)$ then X can take values $0, 1, 2, 3$ and we calculated in Example 5.6.2 that $\mu = E(X) = 0.3$. Then

$$
\begin{aligned}
E[(X - \mu)^2] &= \sum (x_i - \mu)^2 \, \text{pr}(x_i) \\
&= (0 - 0.3)^2 \text{pr}(0) + (1 - 0.3)^2 \text{pr}(1) + (2 - 0.3)^2 \text{pr}(2) \\
&\quad + (3 - 0.3)^2 \text{pr}(3) \\
&= 0.27.
\end{aligned}
$$

Thus, $\text{sd}(X) = \sqrt{E[(X - \mu)^2]} = \sqrt{0.27} = 0.5196$.

[Note that the formulae for $\text{pr}(0)$, $\text{pr}(1)$, etc. are given in Example 5.6.2.]

Exercises 5.6.2

1. Compute $\text{sd}(X)$ for Exercises 5.6.1, problem 1.

2. Compute $\text{sd}(X)$ for Exercises 5.6.1, problem 2.

5.6.3 Means and standard deviations of some common distributions

The Poisson distribution with parameter λ takes values $0, 1, 2, \ldots$ with probability function $\text{p}(x) = \frac{e^{-\lambda} \lambda^x}{x!}$. We can use the formulae of the previous subsections to calculate the mean $[\mu = E(X)]$ and the standard deviation of the Poisson distribution in the usual way. Thus

$$
\begin{aligned}
E(X) &= \Sigma_i x_i \text{pr}(x_i) \\
&= 0 \times \text{pr}(0) \ + \ 1 \times \text{pr}(1) \ + 2 \times \text{pr}(2) \ + \ldots \\
&= 0 \times \frac{e^{-\lambda} \lambda^0}{0!} \ + 1 \times \frac{e^{-\lambda} \lambda^1}{1!} \ + 2 \times \frac{e^{-\lambda} \lambda^2}{2!} \ + \ldots
\end{aligned}
$$

and it can be shown that this adds to λ. Thus, for the Poisson(λ) distribution, $E(X) = \lambda$.

The idea of calculating population (or distribution) means and standard deviations using these formulae is simple, but the algebraic manipulations necessary to obtain simple expressions for these quantities can be quite involved. The standard deviation of the Poisson is harder to find than the mean. Similarly, the Binomial calculations are more complicated than the Poisson. We do not want to go to that level of detail here and will just quote results, namely

$$\textbf{\textit{Hypergeometric}}(N, M, n): \quad \textbf{E}(\textbf{\textit{X}}) = \textbf{\textit{np}}, \quad \text{sd}(\textbf{\textit{X}}) = \sqrt{\textbf{\textit{np}}(\textbf{1}-\textbf{\textit{p}})\frac{\textbf{\textit{N}} - \textbf{\textit{n}}}{\textbf{\textit{N}}-\textbf{1}}} \,,$$

$$\text{where } p = \tfrac{M}{N} \,.$$

$$\textbf{\textit{Binomial}}(n, p): \qquad\qquad \textbf{E}(\textbf{\textit{X}}) = \textbf{\textit{np}}, \quad \text{sd}(\textbf{\textit{X}}) = \sqrt{\textbf{\textit{np}}(\textbf{1}-\textbf{\textit{p}})} \,,$$

$$\textbf{\textit{Poisson}}(\lambda): \qquad\qquad \textbf{E}(\textbf{\textit{X}}) = \lambda, \quad \text{sd}(\textbf{\textit{X}}) = \sqrt{\lambda} \,,$$

5.6.4 Mean and standard deviation of $aX + b$

Suppose that the length of a cellphone toll call (to the nearest minute) has a distribution with mean $E(X) = 1.5$ minutes and a standard deviation of $\text{sd}(X) = 1.1$ minutes. If the telephone company charges a fixed connection fee of 50 cents and then 40 cents per minute, what is the mean and standard deviation of the distribution of charges? Let Y be the charge.

We have $\qquad Cost = 40 \times Time + 50 \qquad$ i.e. $Y = 40X + 50,$

which is an example of $\qquad\qquad Y = aX + b.$

Means and standard deviations of distributions behave in exactly the same way, in this regard, as sample means and standard deviations (i.e. averages and sd's of batches of numbers).

$$\boxed{\textbf{E}(\textbf{\textit{a}}\,\textbf{\textit{X}} + \textbf{\textit{b}}) = \textbf{\textit{a}}\,\textbf{E}(\textbf{\textit{X}}) + \textbf{\textit{b}} \ \text{ and } \ \text{sd}(\textbf{\textit{a}}\,\textbf{\textit{X}} + \textbf{\textit{b}}) = |\textbf{\textit{a}}|\,\text{sd}(\textbf{\textit{X}}).}$$

We use $|a|$ with the standard deviation because, as a measure of variability, we want "sd" to be always positive and in some situations a may be negative. Application of these results to the motivating example above is straightforward. Intuitively, the expected cost (in cents) is 40 cents (per minute) times the expected time taken (in minutes) plus the fixed cost of 50 cents. Thus

$$E(40X + 50) = 40\,E(X) + 50 = 40 \times 1.5 + 50 = 110 \text{ cents.}$$

Also, a fixed cost of 50 cents should have no effect on the variability of charges and the variability of charges (in cents) should be 40 times the variability of times taken (in minutes). We therefore have

$$\text{sd}(40X + 50) = 40\,\text{sd}(X) = 40 \times 1.1 = 44\,\text{cents.}$$

It will be helpful to have some facility with using these results algebraically, so we will do a theoretical example. The particular example will be important in later chapters.

Example 5.6.5 If Y is the number of heads in n tosses of a biased coin, then $Y \sim \text{Binomial}(n, p)$. The proportion of heads is simply Y/n, which we denote by $\hat{P}$. What is $\text{E}(\hat{P})$ and $\text{sd}(\hat{P})$?

$$\text{E}(\hat{P}) = \text{E}(\frac{1}{n}Y) = \frac{1}{n}\text{E}(Y) = \frac{1}{n}np = p, \qquad \text{and}$$

$$\text{sd}(\hat{P}) = \text{sd}(\frac{1}{n}Y) = \frac{1}{n}\text{sd}(Y) = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}}.$$

Exercises on Section 5.6.4

Suppose $\text{E}(X) = 3$ and $\text{sd}(X) = 2$, compute $\text{E}(Y)$ and $\text{sd}(Y)$, where:

(**a**) $Y = 2X$ (**b**) $Y = 4 + X$ (**c**) $Y = X + 4$

(**d**) $Y = 3X + 2$ (**e**) $Y = 4 + 5X$ (**f**) $Y = -5X$

(**g**) $Y = 4 - 5X$.

Quiz for Section 5.6

1. Why is the expected value also called the population mean? (Section 5.6.1)
2. Explain how the idea of an expected value can be used to calculated the expected profit from a game of chance like LOTTO. (Section 5.6.1)
3. What is the effect on (a) the mean (b) the standard deviation of a random variable if we multiply the random variable by 3 and add 2? (Section 5.6.4)

5.7 Summary

5.7.1 General ideas

Random variable (r.v.)

A type of measurement made on the outcome of the random experiment.

Probability function

$\text{pr}(X = x)$ for every value X can take, abbreviated to $\text{pr}(x)$.

Expected Value   for a r.v. X, denoted $\mathrm{E}(X)$.

- Also called the ***population mean*** and denoted μ_X (abbrev. to μ).
- Is a measure of the long run average of X-values.

$$\mu_X = \mathrm{E}(X) \;\; = \sum x \mathrm{pr}(x) \qquad \text{(for a discrete r.v. } X\text{)}.$$

Standard deviation   for a r.v. X, denoted $\mathrm{sd}(X)$.

- Also called the ***population standard deviation*** and denoted σ_X.

$$(\sigma_X \text{ often abbrev. to } \sigma)$$

- Is a measure of the variability of X-values.

$$\sigma_X = \mathrm{sd}(X) = \sqrt{\mathrm{E}[(X-\mu)^2]}$$

$$\mathrm{E}[(X-\mu)^2] = \sum (x-\mu)^2 \mathrm{pr}(x) \qquad \text{(for a discrete r.v. } X\text{)}$$

Linear functions   For any constants a and b,

$$\mathrm{E}(aX+b) = a\mathrm{E}(X) + b \quad \text{and} \quad \mathrm{sd}(aX+b) = |a|\,\mathrm{sd}(X).$$

5.7.2 Summary of important discrete distributions

[Note: $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$ and $\binom{n}{x} = \frac{n!}{x!(n-x)!}$].

Hypergeometric(N, M, n) distribution

This is the distribution of the number of black balls in a sample of size n taken randomly without replacement from an urn containing N balls of which M are black.

$$\mathrm{pr}(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad \mu = np, \quad \sigma = \sqrt{np(1-p)\frac{N-n}{N-1}}, \;\; \text{where} \;\; p = \frac{M}{N}.$$

Binomial(n, p) distribution

This is the distribution of the number of heads in n tosses of a biased coin.

$$\mathrm{pr}(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad \mu = np, \quad \sigma = \sqrt{np(1-p)}.$$

$$\text{[Here } p = \mathrm{pr}(\text{Head}) \text{ on a single toss]}.$$

Poisson(λ) *distribution*

This is the distribution of the number of events occurring in a unit of time or space when events are occurring randomly at an average rate λ per unit.

$$\mathrm{pr}(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \mu = \lambda, \quad \sigma = \sqrt{\lambda}.$$

Review Exercises 4

1. Use the appropriate formulae to calculate the following:
   (a) $\mathrm{pr}(X = 4)$ when $X \sim \mathrm{Poisson}(\lambda = 2)$.
   (b) $\mathrm{pr}(X \le 5)$ when $X \sim \mathrm{Poisson}(p = 0.3)$.
   (c) $\mathrm{pr}(X = 2)$ when $X \sim \mathrm{Hypergeometric}(N = 6, M = 3, n = 4)$.
   (d) $\mathrm{pr}(X = 3)$ when $X \sim \mathrm{Binomial}(n = 10, p = 0.23)$.

2. [The main aim of this question is simply to give practice in identifying an appropriate model and distribution from a description of a sampling situation. You will be looking for situations that look like coin tossing (Binomial for a fixed number of tosses or Geometric for the number of tosses up to the first head), situations that look like urn sampling (Hypergeometric), and random events in time or space (where we are looking for the answer "Poisson($\lambda =$??)" although, in practice, we would still need to think hard about how good are the Poisson assumptions).]
   For each of the following situations, write down the name of the distribution of the random variable and identify the values of any parameters.

   (a) In the long run, 80% of Ace light bulbs last for 1000 hours of continuous operation. You need to have 20 lights in constant operation in your attic for a small business enterprise, so you buy a batch of 20 Ace bulbs. Let X_1 be the number of these bulbs that have to be replaced by the time 1000 hours are up.

   (b) In (a), because of the continual need for replacement bulbs, you buy a batch of 1000 cheap bulbs. Of these, 100 have disconnected filaments. You start off by using 20 bulbs (which we assume are randomly chosen). Let X_2 be the number with disconnected filaments.

   (c) Suppose that telephone calls come into the university switchboard randomly at a rate of 100 per hour. Let X_3 be the number of calls in a 1-hour period.

   (d) In (c), 60% of the callers know the extension number of the person they wish to call. Suppose 120 calls are received in a given hour. Let X_4 be the number of callers who know the extension number.

   (e) In (d), let X_5 be the number of calls taken up to and including the first call where the caller did *not* know the extension number.

(**f**) It so happened that of the 120 calls in (e), 70 callers knew the extension number and 50 did not. Assume calls go randomly to telephone operators. Suppose telephone operator A took 10 calls. Of the calls taken by operator A, let X_6 be the number made by callers who knew the extension number.

(**g**) Suppose heart attack victims come to a 200-bed hospital at a rate of 3 per week on average. Let X_7 be the number of heart attack victims admitted in one week.

(**h**) Continuing (g), let X_8 be the number of heart attack victims admitted in a month (4 weeks).

(**i**) Suppose 20 patients, of whom 9 had "flu", came to a doctor's surgery on a particular morning. The order of arrival was random as far as having flu was concerned. The doctor only had time to see 15 patients before lunch. Let X_9 be the number of flu patients seen before lunch.

(**j**) Suppose meteor showers are arriving randomly at the rate of 40 per hour. Let X_{10} be the number of showers arriving in a 15 minute period.

(**k**) A box of 15 fuses contains 8 defective fuses. Three are sampled at random to fill 3 positions in a circuit. Let X_{11} be the number of defective fuses installed.

(**l**) Let X_{12} be the number of ones in 12 rolls of a fair die.

(**m**) Let X_{13} be the number of rolls if you keep rolling until you get the first six.

(**n**) Let X_{14} be the number of double ones on 12 rolls of a pair of fair dice.

(**o**) A Scrabble set consists of 98 tiles of which 44 are vowels. The game begins by selecting 7 tiles at random. Let X_{15} be the number of vowels selected.

3. For each random variable below, write down the name of its distribution and identify the values of any parameter(s). Assume that the behavior and composition of queues is random.

(**a**) A hundred students are standing in a queue to get the Dean's approval. Ten of them are postgraduate students. An extra Dean's representative arrives and 50 of the students break off to make a queue in front of the new representative. Let X be the number of postgraduate students in the new queue.

(**b**) At the end of the day the Dean has 10 people left in the queue to process before going home. Suppose that this Dean approves 80% of the enrolling students she sees. Let Y be the number of students in this queue whom the Dean does not approve.

(**c**) On average, the cashier's computer that calculates tuition fees tuition will crash (stop working) unexpectedly once every eight hour working day on average. I must stand in line in front of the cashier for two hours to pay my fees. Let W be the number of times the computer crashes unexpectedly while I am in line.

(**d**) Suppose one of the cashiers feels like a coffee break. They decide to take it after the next part-time student they see. Part-time students make up 30% of the student population. Let V be the number of students that the cashier must see before taking a coffee break.

4. On 1988 figures (*NZ Herald*, 14 January 1989, page 9), there was one homicide every 6 days on average in NZ (population 3.2 million). What is a possible distribution of the number of homicides in NZ over a month (30 days)?

5. The following simple model is sometimes used as a first approximation for studying an animal population. A population of 420 small animals of a particular species is scattered over an area 2 km by 2 km.[27] This area is divided up into 400 square plots, each of size 100 m by 100 m. Twenty of these plots are selected at random. Using a team of observers, the total number of animals seen at a certain time on these 20 plots is recorded. It is assumed that the animals are independent of one another (i.e. no socializing!) and each animal moves randomly over the whole population area.

(**a**) What is the probability that a given animal is seen on the sampled area?

(**b**) Let X be the number of animals seen on these plots. Explain why we can use the Binomial model for X. Give the parameters of this distribution.

(**c**) Let W be the number of animals found on a single plot. State the distribution and parameters of W.

(**d**) Give two reasons why modeling the numbers of animals found on the sampled plots with a Binomial distribution may not be appropriate (i.e. which of our assumptions are not likely to be valid?)

6. A lake has 250 fish in it. A scientist catches 50 of the fish and tags them. The fish are then released. After they have moved around for a while the scientist catches 10 fish. Let X be the number of tagged fish in her sample.

(**a**) What assumptions would need to be satisfied if the Hypergeometric model is to be suitable for the distribution of X?

(**b**) Give the parameters of your Hypergeometric distribution and write down an expression for the probability of obtaining at least one tagged fish in the sample.

---

[27]1 km(kilometer)=1000 m(meters).

(**c**) What distribution could you use to approximate your Hypergeometric distribution? Use this distribution to evaluate the probability in (b).

7. Over 92% of the world's trade is carried by sea. There are nearly 80,000 merchant ships of over 500 tonnes operating around the world. Each year more than 120 of these ships are lost at sea. Assume that there are exactly 80,000 merchants ships at the start of the year and that 120 of these are lost during the year.

(**a**) A large shipping corporation operates a fleet of 160 merchant ships. Let L be the number of ships the company will lose in a year. What are the distribution and parameter(s) of L?

(**b**) What is the number of ships that the corporation expects to lose in a year?

(**c**) What assumptions were made in (a)? Do you think that these assumptions are reasonable here?

(**d**) What approximate distributions could we use for L? Justify your answer. Give the parameters for the approximating distribution.

8. According to a Toronto *Globe and Mail* report (7 August 1989, page A11), a study on $3,433$ women conducted by the Alan Guttmacher Institute in the US found that 11% of women using the cervical cap method of contraception can expect to become pregnant within the first year.[28] Suppose a family planning clinic fits 30 women with cervical caps in early January.

(**a**) What is the probability that none will become pregnant before the end of the year?

(**b**) What is the probability that no more than 2 will?

(**c**) What is the probability that none will be pregnant by the end of two years?

9. Homelessness has always existed in underdeveloped parts of the world. However, *TIME* magazine (6 December 1993) pointed out that homelessness had become increasingly prevalent in western nations, particularly during a recent recession. In Table 1 they give the estimated numbers of homeless for five countries.

Table 1 : Homeless People[a]

Country	Population	Homeless
Australia	$17,753,000$	$100,000$
France	$57,527,000$	$400,000$
Germany	$81,064,000$	$150,000$
India	$897,443,000$	$3,000,000$
USA	$258,328,000$	$2,000,000$

[a]Source: *TIME*, 6 December 1993.

---

[28]compared with 16% for the diaphragm, 14% whose partners use condoms, and 6% on the contraceptive pill.

(a) Using the priniciples introduced in Chapter 2, comment on the layout of Table 1. Can you suggest any improvements?

(b) Rank the countries from the highest to the lowest on percentage of homeless people. Comment on anything you find surprising.

(c) Suppose we take a random sample of 10 Australians. Let X be the number of homeless people in the sample.

   (i) What is the exact distribution of X? Give the parameter(s) of this distribution.

   (ii) What distribution can we use to approximate the distribution of X? Justify your answer.

   (iii) Find the probability that at least one person in the sample is homeless.

   (iv) Find the probability that no more than 2 people are homeless.

(d) Suppose we sample 10 people from the town of Alice Springs in Australia. Would the calculations in (c) give us useful information?

10. 18% of lots of condoms tested by the US Food and Drug Administration between 1987 and 1989 failed leakage tests. It's hard to know if this applies to individual condoms but let us suppose it does. In a packet of 12 condoms,

(a) what is the distribution of the number that would fail the tests?

(b) what is the probability none would fail the test?

[The failure rate in Canada was even worse with 40% of lots failing Health and Welfare Dept tests over this period!]

11. Suppose that experience has shown that only $\frac{1}{3}$ of all patients having a particular disease will recover if given the standard treatment. A new drug is to be tested on a group of 12 volunteers. If health regulations require that at least seven of these patients should recover before the new drug can be licensed, what is the probability the drug will be discredited even if it increases the individual recovery rate to $\frac{1}{2}$?

12. If a five card poker hand is dealt from a pack of 52 cards, what is the probability of being dealt

(a) **(i)** 3 aces?   **(ii)** 4 aces?

(b) **(i)** 3 of a kind?   **(ii)** 4 of a kind?

(c) a Royal flush in hearts (i.e. 10, Jack, Queen, King, Ace)?

(d) a Royal flush all in the same suit?

(e) Why can't the method applied in (b) be used for 2 of a kind?

13. A fishing fleet uses a spotter plane to locate schools of tuna. To a good approximation, the distribution of schools in the area is completely random with an average density of 1 school per $200,000$km^2 and the plane can search about $24,000$km^2 in a day.

(a) What is the probability that the plane will locate at least one school in 5 days of searching?

(**b**) To budget for fuel costs etc., the fleet management wants to know how many days searching would be required to be reasonably certain of locating at least one school of tuna. How many days should be allowed for in order to be 95% sure of locating at least one school?

14. In a nuclear reactor, the fission process is controlled by inserting into the radioactive core a number of special rods whose purpose is to absorb the neutrons emitted by the critical mass. The effect is to slow down the nuclear chain reaction. When functioning properly, these rods serve as the first-line defense against a disastrous core meltdown.

Suppose that a particular reactor has ten of these control rods (in "real life" there would probably be more than 100), each operating independently, and each having a 0.80 probability of being properly inserted in the event of an "incident". Furthermore, suppose that a meltdown will be prevented if at least half the rods perform satisfactorily. What is the probability that, upon demand, the system will fail?

15. Samuel Pepys, whose diaries so vividly chronicle life in seventeenth century England, was a friend of Sir Isaac Newton. His interest in gambling prompted him to ask Newton whether one is more likely to get:

(**a**) at least one 6 when six dice are rolled,

(**b**) at least two 6's when 12 dice are rolled, or

(**c**) at least three 6's when 18 dice are rolled? The pair exchanged several letters before Newton was able to convince Pepys that (a) was most probable. Compute these three probabilities.

16. Suppose X has a probability function given by the following table but **one** of the given probabilities is in error:

X	-3	0	1	3	8
$\mathrm{pr}(X = 3)$	0.23	-0.39	0.18	0.17	0.13

(**a**) Which one of the probabilities is in error? Give the correct value and use it in answering the following questions.

(**b**) What is the probability that X is at least 1?

(**c**) What is the probability that X is no more than 0?

(**d**) Calculate the expected value and standard deviation of X.

17. A standard die[29] has its faces painted different colors. Faces 3, 4, 6 are red, faces 2 and 5 are black and face 1 is white.

(**a**) Find the probability that when the die is rolled, a black or even numbered face shows uppermost.

A game is played by rolling the die once. If any of the red faces show uppermost, the player wins the dollar amount showing, while if a black face shows uppermost the player loses twice the amount showing. If a white face shows uppermost he wins or loses nothing.

---

[29]singular of dice.

(b) Find the probability function of the player's winnings X.

(c) Find the expected amount won. Would you play this game?

18. Almond Delight is a breakfast cereal sold in the US In 1988 the manufacturers ran a promotion with the come-on to the purchaser, "Instantly win up to 6 months of FREE UTILITIES" up to a value of $558. (Utilities are things like electricity, gas and water). A colleague bought a box of Almond Delight for $1.84 and sent us the box. The company had distributed 252,000 checks or vouchers amongst 3.78 million boxes of Almond Delight with at most one going into any box as shown in Table 2. If a check was found in a box the purchaser could redeem the prize by sending in a self addressed (and stamped!) envelope. (Assume the cost of stamps is 20 cents per letter.)

Table 2 :   Almond Delight Prizes

"Free utilities"	$ value	Number of such vouchers distributed
Light home for week	1.61	246,130
Wash and dry clothes for month	7.85	4,500
Hot water for 2 months	25.90	900
Water lawn all summer	34.29	360
Air condition home for 2 months	224.13	65
Free utilities for 6 months	558.00	45

Let X be the value of vouchers found in the packet of Almond Delight bought.

(a) Write down the probability function of X in table form.

(b) Calculate the expected value of X.

(c) Calculate the standard deviation of X.

(d) Would you buy Almond Delight if you found another brand which you liked equally well at $1.60? Why or why not? (Allow for the cost of 2 stamps. Assume envelopes are free.)

19. An instant lottery played in the US state of Kansas called "Lucky Numbers" costs $1.00 per ticket. The prizes and chances of winning them are:

Prize	$0	Free ticket	$3	$7	$11	$21	$2100
Prob	by subtraction	$\frac{1}{10}$	$\frac{1}{30}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{150}$	$\frac{1}{300000}$

We want to find the expected return on $1.00 spent on playing this game. Free tickets entitle the player to play again, so let us decide that every time we get a free ticket we will play again until we get a proper outcome. We get the relevant probabilities, which ignore free tickets, by dividing each of the probabilities above by 0.9. (Why?). Now work out the expected return on a $1.00 ticket. What is the standard deviation?

20. Suppose that, to date, a process manufacturing rivets has been "in control" and rivets which are outside the specification limits have been produced at the rate of 1 in every 100. The following (artificial) inspection process is operated. Periodically, a sample of 8 rivets is taken and if 2 or more of the 8 are defective, the production line is halted and adjusted. What is the probability that:

(a) production will be halted on the strength of the sample if the process has not changed?

(b) production will not be halted even though the process is now producing 2% of its rivets outside the specification limits?

21. Manufacturers whose product incorporates a component or part bought from an external supplier will often take a sample of each batch of parts they receive and send back the batch if there are "too many" defective parts in the sample. This is called "***acceptance sampling***".

Suppose an incoming shipment contains 100 parts and the manufacturer picks two parts at random (*without* replacement) for testing. The shipment is accepted if both sampled parts are satisfactory and is rejected if one or more is defective. Find the probability of rejecting the shipment if the number of defective parts in the shipment is

(a) 0 **(b)** 10 **(c)** 20 **(d)** 40 **(e)** 60 **(f)** 80.

(g) Graph the *acceptance* probability versus the percentage of defective parts. (Such graphs are called operating characteristic, or OC, curves. They show how sensitive the sampling rejection scheme is at detecting lapses in shipment quality.)

(h) Comment on the practical effectiveness or otherwise of this particular scheme. Can you think of a (very) simple way of improving it?

22. At a company's Christmas social function attended by the sister of a colleague of ours, ten bottles of champagne were raffled off to those present. There were about 50 people at the function. The 50 names were placed on cards in a box, the box was stirred and the name of the winner drawn. The winner's name was returned to the box for the next draw. The sister won three of the ten bottles of champagne. By the last one she was getting a little embarrassed. She definitely doubted the fairness of the draw. Even worse, she doesn't drink and so she didn't even appreciate the champagne!

Assuming a random draw:

(a) What is the distribution of X, the number of bottles won by a given person? Justify your answer and give the value(s) of the parameter(s).

(b) What is:
  (i) $\text{pr}(X = 0)$ and $\text{pr}(X \leq 2)$?
  (ii) the probability of winning 3 or more bottles?

(c) Do you think the names in the box were well stirred?

But perhaps it isn't as bad as all that. Sure, the chances of a particular person winning so many bottles are tiny, but after all, there were 50 people in the room. Maybe the chances of it happening to somebody in the room are quite reasonable.

(d) The events $E_i =$ "ith person wins 3 or more bottles" are not independent. Make an observation that establishes this.

(e) Even though the events are not independent, since $\text{pr}(E_i)$ is so small the independence assumption probably won't hurt us too much. Assuming independence, what is the probability that nobody wins 3 or more bottles? What is the probability that at least one person wins 3 or more bottles?

(f) Does this change your opinion in (c)?

23. In the "People versus Collins" case described in Case Study 4.7.1 in Section 4.7.3, we saw how the defendants were convicted largely on the basis of a probabilistic argument which showed that the chances that a randomly selected couple fitted the description of the couple involved were overwhelmingly small (1 in 12 million). At the appeal at the Supreme Court of California, the defense attacked the prosecution's case by attacking the independence assumption. Some of the characteristics listed clearly tend to go together such as mustaches and beards and, Negro men and interracial couples. However, their most devastating argument involved a conditional probability. The probability of finding a couple to fit the description was so small that it probably never entered the jurors' heads that there might be another couple fitting the description. The defense calculated the conditional probability that there were two ore more couples fitting the description given that at least one such couple existed. This probability was quite large even using the prosecution's 1 in 12 million figure. Reasonable doubt about whether the convicted couple were guilty had clearly been established. We will calculate some of these probabilities now.

Suppose there were n couples in Los Angeles (or maybe of the whole of California) and the chances of any given couple match the description is 1 in 12 million. Assuming independence of couples:

(a) What is the distribution of X, the number of couples who fit the description? Justify your answer.

(b) Write down formulae for

    (i) $\text{pr}(X = 0)$ **(ii)** $\text{pr}(X = 1)$ **(iii)** $\text{pr}(X \geq 2)$.

(c) Write down a formula for $\text{pr}(X \geq 2 \,|\, X \geq 1)$.

(d) Evaluate your probability in (c) when $n = 1$ million, 4 million and 10 million.

24. The Medicine column in *TIME* magazine (23 February 1987) carried a story about Bob Walters, a one-time quarterback with the San Francisco 49ers (a well known American football team). He had ALS (amyotrophic

laterial sclerosis), a rare fatal disease which involves a degeneration of the nerves. The disease has an incidence rate of roughly 1 new case per 50,000 people per year. Bob Walters had learned that 2 teammates from his 55 member strong 49ers squad had also contracted the disease. Both had since died. Researchers into rare diseases look for "clusters" of cases. We would expect clusters if the disease was to some extent contagious or caused by some localized environmental conditions. Walters and his doctors believed this was such a cluster and suspected a fertilizer used in the 49ers' practice field for about a decade in 1947. The fertilizer was known to have a high content of the heavy metal cadmium. The *TIME* story also describes 2 other clusters seen in the US (no time period is given). Do these "clusters" show that ALS does not strike randomly? Is it worth investigating the cases for common features to try and find a cause for ALS? We will do some calculations to try and shed a little bit of light on these questions.

Suppose ALS occurs randomly at the rate of 1 case per 50,000 people in a year. Let X be the number of cases in a single year in a group of size 55.

(a) What is the distribution of X? Justify your answer and give the value(s) of the parameter(s).

If you try to calculate $\text{pr}(X \geq 3)$, you will find it is almost vanishingly small, about 2×10^{-10}. However, if a large population like the US we might still expect some clusters even in the disease occurs randomly. The population of the US is about 250 million, so let us imagine that it is composed of about 4 million groups of 55 people ("training squads"). For 20 years (Walters and the others were from the 1964 squad) we have 80 million training squad-years. Suppose the chances of a single training squad becoming a "cluster" in any one year is 2×10^{-10} and "clusters" occur independently.

Let Y be the number of clusters noted over the 20 year period.

(b) What is the distribution of Y? Justify your answer and give the value(s) of the parameter(s).

(c) What distribution would you use to approximate the distribution of Y.

(d) What is the expected value of Y?

(e) Evaluate $\text{pr}(Y = 0)$, $\text{pr}(Y = 1)$, $\text{pr}(Y = 2)$, $\text{pr}(Y \geq 3)$.

I think at this stage you will agree that 3 clusters in the last 20 years is fairly strong evidence against ALS striking randomly. However there is another factor we haven't taken into account. Everybody belongs to a number of recognizable groups e.g. family, work, sports clubs, churches, neighborhoods etc. We will notice a "cluster" if several cases turn up in any one of those groups. Let us make a crude correction by multiplying the number of groups by 5 say[30] (to 400 million training squad-years). This

---

[30]This isn't quite correct because now the different groups are no longer independent as they are no longer composed of completely different people.

multiplies E(Y) in (d) by 5.

(**f**) What is E(Y) now after this modification?

(**g**) Do we still have strong evidence against the idea that ALT strikes randomly.

25. A March 1989 *TIME* article (13 March, Medicine Section) sounded warning bells about the fast growing in-vitro fertilization industry which caters for infertile couples who are desperate to have children. US in-vitro programs appear to charge in the vicinity of $7,000 per attempt. There is a lot of variability in the success rates both between clinics and within a clinic over time but we will use an average rate of 1 success in every 10 attempts.

Suppose that 4 attempts ($28,000) is the maximum a couple feels they are prepared to pay and will try until they are successful up to that maximum.

(**a**) Write down a probability function for the number of attempts made.

(**b**) Compute the expected number of attempts made. Also compute the standard deviation of the number of attempts made.

(**c**) What is the expected cost when embarking on this program.

(**d**) What is the probability of still being childless after paying $28,000?

The calculations you have made assume that the probability of success is always the same at 10% for every attempt. This will not be true. Quoted success rates will be averaged over both couples and attempts. Suppose the population of people attending the clinics is made up of three groups. Imagine that 30% is composed of those who will get pregnant comparatively easily, say pr(*success*) = 0.2 per attempt, 30% are average with pr(*success*) = 0.1 per attempt, and the third group of 40% have real difficulty and have pr(*success*) = 0.01.[31] After a couple are successful they drop out of the program. Now start with a large number, say 100,000 people and perform the following calculations.

(**e**) Calculate the number in each group you expect to conceive on the first attempt, and hence the number in each group who make a second attempt. [Note: It is possible to find the probabilities of conceiving on the first try, on the second given failure at the first try, etc. using the methods of Chapter 4. You might even like to try it. The method you were led through here is much simpler, though informal.]

(**f**) Find the number in each group getting pregnant on the second attempt, and hence the number in each group who make a third attempt. Repeat for the third and fourth attempts.

(**g**) Now find the proportions *of all couples who make the attempt* who conceive at each of the first, second, third and fourth attempts. Observe how these proportions decrease.

[The decrease is largely due to the fact that as the number of attempts increases, the "difficult" group makes up a bigger and bigger proportion of those making the attempt.]

---

[31]In practice there will be a continuum of difficulty to conceive, not several well defined groups.

26. Suppose you are interested in the proportion, p say, of defective items being produced by a manufacturing process. You suspect that p is fairly small. You inspect 50 items and find no defectives. Using zero as an estimate of p is not particularly useful. It would be more useful if we could put an upper limit on p, i.e. to say something like "we are fairly sure p is no bigger than 0.07". (Note: The actual number 0.07 here is irrelevant to the question.)

(a) What is the distribution of X, the number of defectives in a sample of 50?

(b) What assumptions are implicit in your answer to (a)? Write down some circumstances in which they would not be valid.

(c) Plot $\text{pr}(X = 0)$ versus p for $p = 0, 0.02, 0.04, 0.06, 0.08$ and sketch a smooth curve through the points. (This should show you the shape of the curve).

(d) Write down the expression for $\text{pr}(X = 0)$ in terms of p. We want to find the value of p for which $\text{pr}(X = 0) = 0.1$. Solve the equation for p if you can. Otherwise read it off your graph in (c).

[Note: **Your answer to (d) is a reasonable upper limit for** p **since any larger value is unlikely to give zero defectives (less than 1 chance in 10)**].

27. A 8 March 1989 report in the *Kitchener-Waterloo Record* (page B11) discussed a successful method of using DNA probes to sexing embryo calves developed by the Salk Institute Biotechnology and Industrial Associates in California. The method correctly identified the sex of all but one of 91 calves. Imitate the previous problem to calculate a likely upper limit for the failure rate p of the method by calculating the probability of getting at most 1 failure for various values of p and using an approximate graphical solution [cf. 26(d)].

28. Everyone seems to love to hate their postal service and NZ is no exception. In 1988 NZ Post set up a new mail system. Letters were either to go by "Standard Post" at a cost of 40 cents or "Fast Post" at a cost of 70 cents. Fast Post was promoted vigorously. According to NZ Post publicity, the chances of a Fast Post letter being delivered to the addressee the following day are "close to 100%". Very soon after the introduction of the new system, the *NZ Herald* conducted an experiment to check on the claims. Newspaper staff mailed 15 letters all over the country from the Otara Post Office in Auckland. According to the resulting story (9 June 1988), only 11 had reached their destination by the next day. This doesn't look very "close to 100%".

(a) On the grounds that we would have been even more disturbed by fewer than 11 letters arriving, determine the probability of 11 or fewer letters being delivered in a day if each had a 95% chance of doing so.

(b) What assumptions have you made.

(c) From your answer to (a), do you believe NZ Post's claims were valid or was the new system suffering from, what might be kindly referred to, as teething problems?

$\Big[$Note: You need not consider this in your answer, but there are different ways we might be inclined to interpret the 95%. Should it mean 95% of letters? This should be relatively easy for NZ Post to arrange if most of the traffic flows between main centers with single mail connections. Or should it mean 95% of destinations? This would give much more confidence to customers sending a letter somewhere other than a main center. The *Herald* sent its letters to 15 different cities and towns.$\Big]$

Note about the exercises: We end with a block of exercises built about the game of LOTTO which is described in Case Study 5.3.1 in Section 5.3.3.

29. In the game of Lotto, which set of six numbers $(1, 2, 3, 4, 5, 6)$ or $(25, 12, 33, 9, 27, 4)$ is more likely to come up? Give a reason for your answer.

30. (a) The minimum "investment" allowed in Lotto is \$2.00 which buys 4 boards. Provided none of the four sequences of numbers chosen are completely identical, what is the probability that one of the 4 boards wins a division 1 prize? Give a reason for your answer.

(b) Does this line of reasoning carry over to the probability of winning at least one prize using four boards? Why or why not?

(c) Someone (identity unknown) rang one of the authors about the possibility of covering all possible combinations of 39 of the 40 numbers. How many boards would they need? What would it cost?

(d) Another enquiry along the same lines was from someone representing a small syndicate who wanted to cover all combinations of the 14 numbers $2, 3, 4, 6, 7, 9, 10, 16, 21, 23, 28, 33, 34$. This type of thing can be done in many countries, even New Zealand. However, at the time of the enquiry all the boards have to be filled out by hand.
  (i) How many boards were there?
  (ii) How much did this exercise cost? (The people did it!)

(e) Give a method by which a player can increase the probability of getting at least one prize from the boards he or she is purchasing. $\Big[$Hint: Try with only 2 boards.$\Big]$

31. On 26 September 1987, the Division 1 prize of \$548,000 was shared by 13 people. On the 8th of October, after this had come to light, the *NZ Herald* ran a story which began as follows. "Mathematicians have been shaking their heads in disbelief and running their computers overtime for the last 10 days or so." One mathematician was quoted as saying that the odds were "billions of trillions" against such an event happening. Our favorite quotation came from a high school mathematics teacher who said, "It couldn't have happened in a thousand years. There is no mathematical explanation for it. If I cannot believe mathematics, what can I trust?" !!!

The prize money above corresponds to sales of about \$2.5 million which is about 10 million boards. Suppose 10 million boards were sold and all the 10 million sequences of 6 numbers have been chosen randomly (with replacement) from the available sequences.

(**a**) What is the distribution of X, the number of boards winning division 1 prizes? Justify your answer. What is the expected number of winning boards?

(**b**) What distribution could be used to give approximate values for $\text{pr}(X = x)$? Justify your answer.

(**c**) Write down an expression for the probability that X is at least 13 but you need not evaluate it. The probability is in fact about 3.7×10^{-6}.

(**d**) This appears to support the claims above. However what huge assumption has been made in making this calculation? Write down any ways in which you think human behavior is likely to differ from the assumed behavior. For each way, write down the qualitative effect this departure is likely to have on the estimate.

[Note: Morton [1990] stated that 14,697 entries in the 7 June 1986 New York State Lotto selected 8, 15, 22, 29, 36, 43 which was a main diagonal of the entry ballot paper.]

(**e**) If no-one wins a Division 1 prize, the allotted prize money is carried over to the next week. Under the assumptions above:

  (**i**) what is the probability that the division 1 prize is not struck in a given week?

  (**ii**) What is the probability that it is not struck for 3 weeks in a row?

32. The probability of obtaining a prize of some sort from a \$2.00 "investment" in Lotto is approximately 0.02. Suppose you decided to "invest" \$2.00 in Lotto every week for a year (52 weeks). Let X be the number of weeks in which you would win a prize.

(**a**) Name the distribution of X and give the parameter values.

(**b**) What is the mean (expected value) and standard deviation of X?

(**c**) What is the probability of winning no prizes in the whole year?

33. In any community when a Lotto game begins, the newspapers pick up on random quirks in the numbers that appear. After the 13th draw in NZ, a *NZ Herald* article asked "What has Lotto's Tulipe[32] got against the numbers 24 and 25?" These were the only numbers that had not appeared in the 13 draws. This looked strange enough for the journalist to write a short column about it. Recall that in each draw, 7 numbers are drawn without replacement from the numbers 1 to 40.

(**a**) What is the probability that 24 and 25 are not drawn on the first draw?

(**b**) What is the probability that neither is ever chosen in 13 draws?

The probability is about 0.12 which is not particularly unusual. It would happen in about one community in every 8 that started up a 40-ball Lotto game.

---

[32]the Lotto sampling machine

(c) The article went on to say "But reason says that 24 and 25 are bound to come up in the next few Lotto draws." Is this statement true? Why or why not? What is the probability that 24 and/or 25 come up in the next draw?

In (b) above, we have treated 24 and 25 as special. The journalist would probably have written exactly the same article if it was 12 and 36, or if any other pair of numbers hadn't been drawn. So the relevant probability is the probability that there are two numbers that have never appeared after 13 draws. The next few parts of the problem will give you some idea about how you might calculate the distribution of how many numbers appear after n Lotto draws.

Let X_1 count how many numbers have appeared after the first draw, X_2 count how many numbers have appeared after 2 draws,... etc. In parallel, let Y_i be the number of *new numbers* appearing on draw i. Then $X_i = X_{i-1} + Y_i$.

Note that 7 numbers appear on the first draw.

(d) What is the distribution of Y_2?

(e) What is the probability that $X_3{=}12$ given that $X_2 = 8$?

(f) More generally, what is the probability that $X_n = j + d$ given that up until and including the last draw, j numbers had appeared (i.e. $X_{n-1} = j$)?

Using the partition theorem,

$$\mathrm{pr}(X_3 = 12) = \Sigma_j \mathrm{pr}(X_3 = 12 \mid X_2 = j)\mathrm{pr}(X_2 = j),$$

so that we can use the distribution of X_2 and the conditional probabilities to build up the distribution of X_3. In the same way one can use the distribution of X_3 and the conditional probabilities $\mathrm{pr}(X_4 = i \mid X_3 = j)$ to obtain the distribution of X_4 and so on. The calculations are more complicated than we would expect from the reader and are tedious so we used a computer. After 13 games, the non-negligible values of $\mathrm{pr}(X_{13} = x)$, i.e. the probability that x of the numbers 1 to 40 have appeared by the end of the 13th draw is

x	33	34	35	36	37	38	39	40
$\mathrm{pr}(X_{13} = x)$	0.017	0.054	0.128	0.219	0.260	0.204	0.094	0.019

(g) What is the probability that 2 or more numbers still haven't appeared after 13 draws?

Appendix A2Cum   Binomial distribution (Upper tail probabilities)

The tabulated value is $\operatorname{pr}(X \geq x)$, where $X \sim \text{Binomial}(n, p)$.

[e.g. For $\;n = 5, \;\; p = 0.2, \;\; \operatorname{pr}(X \geq 2) = 0.263$
       For $\;n = 7, \;\; p = 0.3, \;\; \operatorname{pr}(X \geq 5) = 0.029$]

p

n x	.01	.05	.10	.15	.20	.25	.30	.35	.40	.50	.60	.65	.70	.75	.80	.85	.90	.95	.99
2 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.020	.097	.190	.278	.360	.438	.510	.577	.640	.750	.840	.877	.910	.938	.960	.978	.990	.997	1.00
2		.003	.010	.023	.040	.063	.090	.122	.160	.250	.360	.422	.490	.563	.640	.723	.810	.902	.980
3 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.030	.143	.271	.386	.488	.578	.657	.725	.784	.875	.936	.957	.973	.984	.992	.997	.999	1.00	1.00
2		.007	.028	.061	.104	.156	.216	.282	.352	.500	.648	.718	.784	.844	.896	.939	.972	.993	1.00
3			.001	.003	.008	.016	.027	.043	.064	.125	.216	.275	.343	.422	.512	.614	.729	.857	.970
4 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.039	.185	.344	.478	.590	.684	.760	.821	.870	.938	.974	.985	.992	.996	.998	.999	1.00	1.00	1.00
2	.001	.014	.052	.110	.181	.262	.348	.437	.525	.688	.821	.874	.916	.949	.973	.988	.996	1.00	1.00
3			.004	.012	.027	.051	.084	.126	.179	.313	.475	.563	.652	.738	.819	.890	.948	.986	.999
4				.001	.002	.004	.008	.015	.026	.063	.130	.179	.240	.316	.410	.522	.656	.815	.961
5 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.049	.226	.410	.556	.672	.763	.832	.884	.922	.969	.990	.995	.998	.999	1.00	1.00	1.00	1.00	1.00
2	.001	.023	.081	.165	.263	.367	.472	.572	.663	.813	.913	.946	.969	.984	.993	.998	1.00	1.00	1.00
3		.001	.009	.027	.058	.104	.163	.235	.317	.500	.683	.765	.837	.896	.942	.973	.991	.999	1.00
4				.002	.007	.016	.031	.054	.087	.188	.337	.428	.528	.633	.737	.835	.919	.977	.999
5						.001	.002	.005	.010	.031	.078	.116	.168	.237	.328	.444	.590	.774	.951
6 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.059	.265	.469	.623	.738	.822	.882	.925	.953	.984	.996	.998	.999	1.00	1.00	1.00	1.00	1.00	1.00
2	.001	.033	.114	.224	.345	.466	.580	.681	.767	.891	.959	.978	.989	.995	.998	1.00	1.00	1.00	1.00
3		.002	.016	.047	.099	.169	.256	.353	.456	.656	.821	.883	.930	.962	.983	.994	.999	1.00	1.00
4			.001	.006	.017	.038	.070	.117	.179	.344	.544	.647	.744	.831	.901	.953	.984	.998	1.00
5					.002	.005	.011	.022	.041	.109	.233	.319	.420	.534	.655	.776	.886	.967	.999
6							.001	.002	.004	.016	.047	.075	.118	.178	.262	.377	.531	.735	.941
7 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.068	.302	.522	.679	.790	.867	.918	.951	.972	.992	.998	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	.002	.044	.150	.283	.423	.555	.671	.766	.841	.938	.981	.991	.996	.999	1.00	1.00	1.00	1.00	1.00
3		.004	.026	.074	.148	.244	.353	.468	.580	.773	.904	.944	.971	.987	.995	.999	1.00	1.00	1.00
4			.003	.012	.033	.071	.126	.200	.290	.500	.710	.800	.874	.929	.967	.988	.997	1.00	1.00
5				.001	.005	.013	.029	.056	.096	.227	.420	.532	.647	.756	.852	.926	.974	.996	1.00
6					.001	.004	.009	.019	.063	.159	.234	.329	.445	.577	.717	.850	.956	.998	
7							.001	.002	.008	.028	.049	.082	.133	.210	.321	.478	.698	.932	

Appendix A2Cum   Binomial distribution (Uppertail probabilities) cont.

The tabulated value is $\mathrm{pr}(X \geq x)$, where $X \sim \mathrm{Binomial}(n, p)$.

p

n	x	.01	.05	.10	.15	.20	.25	.30	.35	.40	.50	.60	.65	.70	.75	.80	.85	.90	.95	.99
8	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.077	.337	.570	.728	.832	.900	.942	.968	.983	.996	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.003	.057	.187	.343	.497	.633	.745	.831	.894	.965	.991	.996	.999	1.00	1.00	1.00	1.00	1.00	1.00
	3		.006	.038	.105	.203	.321	.448	.572	.685	.855	.950	.975	.989	.996	.999	1.00	1.00	1.00	1.00
	4			.005	.021	.056	.114	.194	.294	.406	.637	.826	.894	.942	.973	.990	.997	1.00	1.00	1.00
	5				.003	.010	.027	.058	.106	.174	.363	.594	.706	.806	.886	.944	.979	.995	1.00	1.00
	6					.001	.004	.011	.025	.050	.145	.315	.428	.552	.679	.797	.895	.962	.994	1.00
	7							.001	.004	.009	.035	.106	.169	.255	.367	.503	.657	.813	.943	.997
	8									.001	.004	.017	.032	.058	.100	.168	.272	.430	.663	.923
9	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.086	.370	.613	.768	.866	.925	.960	.979	.990	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.003	.071	.225	.401	.564	.700	.804	.879	.929	.980	.996	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3		.008	.053	.141	.262	.399	.537	.663	.768	.910	.975	.989	.996	.999	1.00	1.00	1.00	1.00	1.00
	4		.001	.008	.034	.086	.166	.270	.391	.517	.746	.901	.946	.975	.990	.997	.999	1.00	1.00	1.00
	5			.001	.006	.020	.049	.099	.172	.267	.500	.733	.828	.901	.951	.980	.994	.999	1.00	1.00
	6				.001	.003	.010	.025	.054	.099	.254	.483	.609	.730	.834	.914	.966	.992	.999	1.00
	7						.001	.004	.011	.025	.090	.232	.337	.463	.601	.738	.859	.947	.992	1.00
	8								.001	.004	.020	.071	.121	.196	.300	.436	.599	.775	.929	.997
	9										.002	.010	.021	.040	.075	.134	.232	.387	.630	.914
10	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.096	.401	.651	.803	.893	.944	.972	.987	.994	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.004	.086	.264	.456	.624	.756	.851	.914	.954	.989	.998	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3		.012	.070	.180	.322	.474	.617	.738	.833	.945	.988	.995	.998	1.00	1.00	1.00	1.00	1.00	1.00
	4		.001	.013	.050	.121	.224	.350	.486	.618	.828	.945	.974	.989	.996	.999	1.00	1.00	1.00	1.00
	5			.002	.010	.033	.078	.150	.249	.367	.623	.834	.905	.953	.980	.994	.999	1.00	1.00	1.00
	6				.001	.006	.020	.047	.095	.166	.377	.633	.751	.850	.922	.967	.990	.998	1.00	1.00
	7					.001	.004	.011	.026	.055	.172	.382	.514	.650	.776	.879	.950	.987	.999	1.00
	8							.002	.005	.012	.055	.167	.262	.383	.526	.678	.820	.930	.988	1.00
	9								.001	.002	.011	.046	.086	.149	.244	.376	.544	.736	.914	.996
	10										.001	.006	.013	.028	.056	.107	.197	.349	.599	.904
11	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.105	.431	.686	.833	.914	.958	.980	.991	.996	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.005	.102	.303	.508	.678	.803	.887	.939	.970	.994	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3		.015	.090	.221	.383	.545	.687	.800	.881	.967	.994	.998	.999	1.00	1.00	1.00	1.00	1.00	1.00
	4		.002	.019	.069	.161	.287	.430	.574	.704	.887	.971	.988	.996	.999	1.00	1.00	1.00	1.00	1.00
	5			.003	.016	.050	.115	.210	.332	.467	.726	.901	.950	.978	.992	.998	1.00	1.00	1.00	1.00
	6				.003	.012	.034	.078	.149	.247	.500	.753	.851	.922	.966	.988	.997	1.00	1.00	1.00
	7					.002	.008	.022	.050	.099	.274	.533	.668	.790	.885	.950	.984	.997	1.00	1.00
	8						.001	.004	.012	.029	.113	.296	.426	.570	.713	.839	.931	.981	.998	1.00
	9							.001	.002	.006	.033	.119	.200	.313	.455	.617	.779	.910	.985	1.00
	10									.001	.006	.030	.061	.113	.197	.322	.492	.697	.898	.995
	11											.004	.009	.020	.042	.086	.167	.314	.569	.895

Appendix A2Cum   Binomial distribution (Uppertail probabilities) cont.

The tabulated value is $\mathrm{pr}(X \geq x)$, where $X \sim \mathrm{Binomial}(n, p)$.

p

n	x	.01	.05	.10	.15	.20	.25	.30	.35	.40	.50	.60	.65	.70	.75	.80	.85	.90	.95	.99
12	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.114	.460	.718	.858	.931	.968	.986	.994	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.006	.118	.341	.557	.725	.842	.915	.958	.980	.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3		.020	.111	.264	.442	.609	.747	.849	.917	.981	.997	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4		.002	.026	.092	.205	.351	.507	.653	.775	.927	.985	.994	.998	1.00	1.00	1.00	1.00	1.00	1.00
	5			.004	.024	.073	.158	.276	.417	.562	.806	.943	.974	.991	.997	.999	1.00	1.00	1.00	1.00
	6			.001	.005	.019	.054	.118	.213	.335	.613	.842	.915	.961	.986	.996	.999	1.00	1.00	1.00
	7				.001	.004	.014	.039	.085	.158	.387	.665	.787	.882	.946	.981	.995	.999	1.00	1.00
	8					.001	.003	.009	.026	.057	.194	.438	.583	.724	.842	.927	.976	.996	1.00	1.00
	9							.002	.006	.015	.073	.225	.347	.493	.649	.795	.908	.974	.998	1.00
	10								.001	.003	.019	.083	.151	.253	.391	.558	.736	.889	.980	1.00
	11										.003	.020	.042	.085	.158	.275	.443	.659	.882	.994
	12											.002	.006	.014	.032	.069	.142	.282	.540	.886
15	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.140	.537	.794	.913	.965	.987	.995	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.010	.171	.451	.681	.833	.920	.965	.986	.995	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3		.036	.184	.396	.602	.764	.873	.938	.973	.996	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4		.005	.056	.177	.352	.539	.703	.827	.909	.982	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5		.001	.013	.062	.164	.314	.485	.648	.783	.941	.991	.997	.999	1.00	1.00	1.00	1.00	1.00	1.00
	6			.002	.017	.061	.148	.278	.436	.597	.849	.966	.988	.996	.999	1.00	1.00	1.00	1.00	1.00
	7				.004	.018	.057	.131	.245	.390	.696	.905	.958	.985	.996	.999	1.00	1.00	1.00	1.00
	8				.001	.004	.017	.050	.113	.213	.500	.787	.887	.950	.983	.996	.999	1.00	1.00	1.00
	9					.001	.004	.015	.042	.095	.304	.610	.755	.869	.943	.982	.996	1.00	1.00	1.00
	10						.001	.004	.012	.034	.151	.403	.564	.722	.852	.939	.983	.998	1.00	1.00
	11							.001	.003	.009	.059	.217	.352	.515	.686	.836	.938	.987	.999	1.00
	12									.002	.018	.091	.173	.297	.461	.648	.823	.944	.995	1.00
	13										.004	.027	.062	.127	.236	.398	.604	.816	.964	1.00
	14											.005	.014	.035	.080	.167	.319	.549	.829	.990
	15												.002	.005	.013	.035	.087	.206	.463	.860
20	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	.182	.642	.878	.961	.988	.997	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	.017	.264	.608	.824	.931	.976	.992	.998	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	.001	.075	.323	.595	.794	.909	.965	.988	.996	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4		.016	.133	.352	.589	.775	.893	.956	.984	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5		.003	.043	.170	.370	.585	.762	.882	.949	.994	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	6			.011	.067	.196	.383	.584	.755	.874	.979	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	7			.002	.022	.087	.214	.392	.583	.750	.942	.994	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	8				.006	.032	.102	.228	.399	.584	.868	.979	.994	.999	1.00	1.00	1.00	1.00	1.00	1.00
	9				.001	.010	.041	.113	.238	.404	.748	.943	.980	.995	.999	1.00	1.00	1.00	1.00	1.00
	10					.003	.014	.048	.122	.245	.588	.872	.947	.983	.996	.999	1.00	1.00	1.00	1.00
	11					.001	.004	.017	.053	.128	.412	.755	.878	.952	.986	.997	1.00	1.00	1.00	1.00
	12						.001	.005	.020	.057	.252	.596	.762	.887	.959	.990	.999	1.00	1.00	1.00
	13							.001	.006	.021	.132	.416	.601	.772	.898	.968	.994	1.00	1.00	1.00
	14								.002	.006	.058	.250	.417	.608	.786	.913	.978	.998	1.00	1.00
	15									.002	.021	.126	.245	.416	.617	.804	.933	.989	1.00	1.00
	16										.006	.051	.118	.238	.415	.630	.830	.957	.997	1.00
	17										.001	.016	.044	.107	.225	.411	.648	.867	.984	1.00
	18											.004	.012	.035	.091	.206	.405	.677	.925	.999
	19											.001	.002	.008	.024	.069	.176	.392	.736	.983
	20													.001	.003	.012	.039	.122	.358	.818