Dr Isabella Gollini

Probability and Statistics

4 Discrete Probability Distributions

4.1 Random variables

Example

Consider the simple experiment of tossing a fair coin three times. There are 8 equally likely outcomes. In an obvious notation,

$$S = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}.$$

Let X denote the number of heads that occur in the three tosses. The only possible values that X can take are 0, 1, 2 and 3. Tabulated below is the value of X associated with each possible outcome.

outcome	hhh	hht	hth	thh	htt	tht	tth	ttt
X	3	2	2	2	1	1	1	0

We write

$$\begin{aligned} \{X = 0\} &= \{ttt\} \\ \{X = 1\} &= \{htt, tht, tth\} \\ \{X = 2\} &= \{hht, hth, thh\} \\ \{X = 3\} &= \{hhh\} \end{aligned}$$

and, counting the outcomes, deduce that

$$\begin{aligned} \Pr(X = 0) &= 1/8 \\ \Pr(X = 1) &= 3/8 \\ \Pr(X = 2) &= 3/8 \\ \Pr(X = 3) &= 1/8 \end{aligned}$$

The variable X as specified above is an example of what is known as a *random variable*, a number associated with the outcome of an experiment, which varies "randomly" from repetition to repetition of the experiment. Formally,

Definition

Given some sample space S, a *random variable* X is a function defined on the sample space.

In the present section we look at *discrete* random variables, where X maps S onto a finite or countable set, usually a subset of the non-negative integers.

4.2 Probability distributions

Let r be a particular value that a random variable X can take. Then $\{X = r\}$ is a simpler way of writing the event $\{a \in S : X(a) = r\}$. The probability of this event is written $\Pr(X = r)$, the probability that the random variable takes the value r.

As a further simplification, we write

$$\Pr(X = r) = p_r$$

for all values r that X can take. Then (p_r) is what is known as the *probability distribution* (or the *probability density function*) of the discrete random variable X. Now

$$\bigcup_r \{X = r\} = S,$$

where the union is over all values r that X can take. The events in this union are pairwise mutually exclusive. Hence from the probability axioms

$$\sum_r p_r = \sum_r \Pr(X = r) = \Pr\left(\bigcup_r \{X = r\}\right) = \Pr(S) = 1.$$

For simplicity of notation restricting attention to the common case where the values that X can take are the non-negative integers or a subset of them, we make the following definition.

Definition

A sequence (p_r) $(r = 0, 1, 2, \ldots)$ is a *discrete probability distribution* (or a *discrete probability density function*) if

1. $$p_r \geq 0 \qquad (r = 0, 1, 2, \ldots)$$

2. $$\sum_{r=0}^{\infty} p_r = 1\ .$$

In our earlier example of tossing a fair coin three times, we found that

$$p_0 = 1/8,\ p_1 = 3/8,\ p_2 = 3/8,\ p_3 = 1/8,$$

so that $\sum_{r=0}^{3} p_r = 1$. This is indeed a probability distribution.

4.3 The binomial distributions

Consider a simple trial with just two possible outcomes, which we shall refer to as "success" and "failure", respectively, and where the probability of success is p and the probability of failure is q, where $q = 1 - p$. We carry out n mutually independent repetitions of the trial and count up the number of successes. What is the probability distribution of the total number of successes in the n trials?

The following are some examples of situations where this scenario applies.

Coin tossing If a coin is tossed n times, "success" may be identified with "heads" and "failure" with "tails". What is the probability distribution of the total number of heads in the n tosses? If the coin is fair then $p = q = 1/2$. In Section 4.1 we considered the case $n = 3$.

Gaming If there are n independent plays of some game, where the probability of winning ("success") at each play is p and the probability of not winning ("failure") is q. What is the probability distribution of the total number of wins in the n trials?

Sampling with replacement Consider an urn that contains w white balls and b black balls, so that the proportion p of white balls in the urn is given by

$$p = \frac{w}{w+b}.$$

Suppose that the urn is shaken and a ball drawn at random from the urn. The colour of the ball is recorded and the ball is then returned to the urn. This process is repeated n times, as a result of which we have a random sample of size n "with replacement". At each step, independently of all other steps, the probability that a white ball is drawn is p. We may identify the drawing of a white ball with a "success" and the drawing of a black ball with a "failure". What is the probability distribution of the total number of white balls in the sample?

Sampling from a large population Suppose that we have a large population of individuals, a proportion p of whom have some specified characteristic. For example, the characteristic might be that an individual suffers from a particular disease or that a person is in full-time employment. A "random sample" of size n is drawn from the population. How many of the sample members have the specified characteristic? As another example, we might be considering a population of plants, a proportion p of which have white flowers, and the remainder have flowers that are not white. We may identify the occurrence of a plant with white flowers with a "success" and the occurrence of a plant with flowers of some other colour with a "failure". What is the probability distribution of the total number of plants with white flowers in a random sample of size n?

Quality control Consider large numbers of items coming off a production line, some of which are in some sense "defective" or "faulty". Let p denote the long-run proportion of items that are defective. From time to time a random sample of size n is drawn from the production line. We may identify the observation of a defective item with "success" and the observation of a non-defective item with "failure". What is the probability distribution of the total number of defective items in such a random sample of size n?

In the general case, there are 2^n different possible sequences of outcomes for the n trials. Consider a particular possible sequence of length n of successes and failures, such as, in an obvious notation,

$$sfffsfs\dots ffs.$$

Using the independence assumption, the probability of obtaining this particular sequence is

$$p^r q^{n-r},$$

where r is the total number of successes in the sequence. The total number of such sequences with exactly r successes is the number of ways of choosing r locations out of the total number of n locations in the sequence for where the successes occur, that is, the binomial coefficient

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}\ .$$

(An alternative notation is C_r^n.) Hence, since these sequences represent distinct outcomes of the sequence of trials, the probability of there being exactly r successes in the n trials is given by

$$\binom{n}{r} p^r q^{n-r}.$$

If X is the random variable that denotes the total number of successes in the n trials then X has what is known as the *binomial distribution* with parameters n and p, that is, the distribution specified by

$$p_r = \binom{n}{r} p^r q^{n-r} \qquad (r = 0, 1, 2, \dots n). \tag{1}$$

- A binomial distribution is specified for any positive integer value n and any p with $0 < p < 1$.
- The notation "$B(n, p)$ distribution" may be used for the binomial distribution with parameters n and p, and we may write $X \sim B(n, p)$.
- Note that q is defined by $q = 1 - p$.
- Using the binomial theorem,

$$\sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (p+q)^n = 1.$$

Hence the property of a probability distribution, $\sum_{r=0}^{n} p_r = 1$, is satisfied.

In the case of a sequence of tosses of a fair coin, for which $p = q = 1/2$, the probability of r heads in n tosses is given by

$$p_r = \frac{1}{2^n}\binom{n}{r} \qquad (r = 0, 1, 2, \dots n).$$

It is readily checked that this gives the same answer as we obtained earlier in the case $n = 3$.

4.4 Calculation of binomial probabilities

Firstly consider a random variable X with an arbitrary discrete probability distribution (p_r), so that $p_r = \Pr(X = r)$ $(r = 0, 1, 2, \ldots)$. The corresponding *cumulative distribution function* (or just *distribution function*) (F_r) $(r = 0, 1, 2, \ldots)$ is given by the cumulative probabilities F_r,

$$F_r = \Pr(X \leq r) = \sum_{i=0}^{r} p_i \qquad (r = 0, 1, 2, \ldots).$$

For any positive integers a and b with $a \leq b$,

$$\Pr(a \leq X \leq b) = \sum_{r=a}^{b} p_r.$$

Alternatively,

$$\Pr(a \leq X \leq b) = F_b - F_{a-1}.$$

In the binomial case, the (cumulative) distribution function is given by

$$F_r = \sum_{i=0}^{r} \binom{n}{i} p^i q^{n-i} \qquad (r = 0, 1, 2, \ldots n).$$

If, for given values of n and p, we wish to calculate binomial probabilities, we have a number of alternatives. We can do the calculations using the formula of Equation (1). For example, if $X \sim B(10, 0.2)$ then

$$\Pr(X = 3) = p_3 = \binom{10}{3}(0.2)^3(0.8)^7 = (120)(0.008)(0.2097152) = 0.201326592 = 0.201$$

to 3 decimal places.

4.4.1 Statistical Tables

However, it is tedious and unnecessary to do too many such calculations, especially if n is large. An alternative is to use statistical tables such as Table 1 of the *New Cambridge Statistical Tables.* This table lists values of the cumulative distribution function for $n = 2$ up to $n = 20$ and for values of p from 0.01 to 0.50 in steps of 0.01.

- The *New Cambridge Statistical Tables* will be available in the examination.
- Note that if $X \sim B(n, p)$ then, interchanging the roles of success and failure, $n - X \sim B(n, q)$, where $q = 1 - p$. So if $p > 0.5$ then we can work instead with the $B(n, q)$ distribution, where $q < 0.5$.

In our example where $X \sim B(10, 0.2)$, we find from the table for $n = 10$ and the row for $p = 0.20$ that $F_3 = 0.8791$ and $F_2 = 0.6778$. Hence

$$\Pr(X = 3) = p_3 = F_3 - F_2 = 0.8791 - 0.6778 = 0.2013 = 0.201$$

to 3 decimal places, which corroborates the answer obtained earlier by direct calculation.

4.4.2 Statistical Software

In the increasingly common situation where you have a statistical package or a spreadsheet package available, this provides an easier and more natural way of calculating binomial probabilities.

In Excel the `BINOMDIST` function may be used (Figures 1 and 2). For example, entering the formula `=BINOMDIST(3,10,0.2,TRUE)` yields the cumulative probability $F_3 = 0.879126118$ and entering the formula `=BINOMDIST(3,10,0.2,FALSE)` yields the probability $p_3 = 0.201326592$.

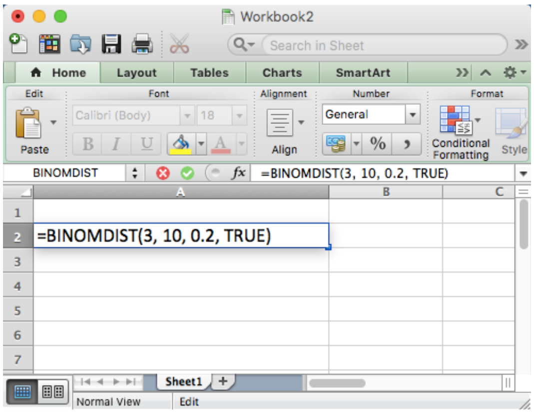

Figure 1: Cumulative distribution function F_r $(r = 3, n = 10, p = 0.2)$

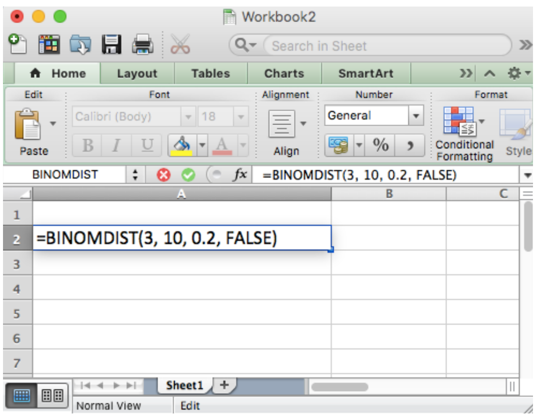

Figure 2: Probability density function p_r $(r = 3, n = 10, p = 0.2)$

To calculate binomial probabilities using `R`, you can use the function `pbinom` to get the cumulative distribution function or `dbinom` to calculated terms from the probability density function. The values you have to specify in the function are (in this order) r, n, and p.

Cumulative distribution function F_r $(r = 3, n = 10, p = 0.2)$:

```
pbinom(3, 10, 0.2)

## [1] 0.8791261
```

Probability density function p_r $(r = 3, n = 10, p = 0.2)$:

```
dbinom(3, 10, 0.2)

## [1] 0.2013266
```

Using `R` we can also specify a sequence of values for r:

```
r <- 0:10
n <- 10
p <- 0.2

Fr <- pbinom(r, n, p)
pr <- dbinom(r, n, p)
Tab <- cbind(r, Fr, pr)
Tab

##        r        Fr           pr
##  [1,]  0 0.1073742 0.1073741824
##  [2,]  1 0.3758096 0.2684354560
##  [3,]  2 0.6777995 0.3019898880
##  [4,]  3 0.8791261 0.2013265920
##  [5,]  4 0.9672065 0.0880803840
##  [6,]  5 0.9936306 0.0264241152
##  [7,]  6 0.9991356 0.0055050240
##  [8,]  7 0.9999221 0.0007864320
##  [9,]  8 0.9999958 0.0000737280
## [10,]  9 0.9999999 0.0000040960
## [11,] 10 1.0000000 0.0000001024
```

Figure 3 shows the probability density function and cumulative distribution function of $B(10, 0.2)$. Figure 3 shows the probability density function of other two binomial distributions with different values of p: $B(10, 0.5)$ and $B(10, 0.8)$. It is possible to notice that in all the three cases the density grows as r increases, up to a maximum that is around np, and then decrease. When $p = 0.5$ the distribution is symmetric around $np = \frac{10}{2}$. The distribution of $B(10, 0.2)$ is skewed right, and the pdf of $B(10, 0.8)$ is skewed left. It is also possible to notice that when $p = 0.8$ we obtain a plot that is the mirror image of Figure 3 $(p = 0.2)$.

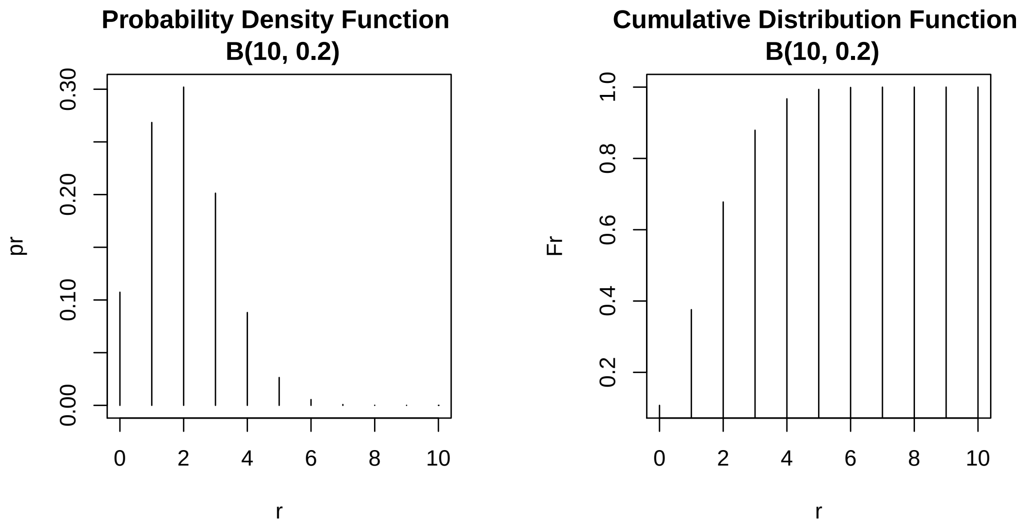

Figure 3: Probability Density Function (left) and Cumulative Distribution Function (right) of $B(10, 0.2)$

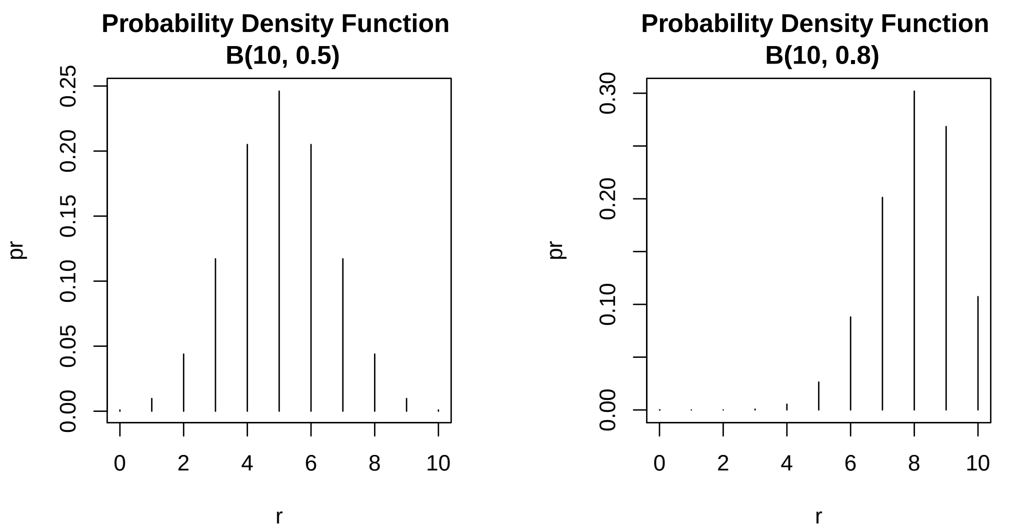

Figure 4: Probability Density Functions of $B(10, 0.5)$ (left) and $B(10, 0.8)$ (right)

If $X \sim B(10, 0.2)$ and we wish, for example, to evaluate $\Pr(1 \leq X \leq 3)$, we may do this by using the probability density function, so that

$$\Pr(1 \leq X \leq 3) = p_1 + p_2 + p_3 = 0.268435 + 0.301990 + 0.201327 = 0.771752 = 0.772$$

to 3 decimal places. Alternatively, we may use the cumulative distribution function, so that

$$\Pr(1 \leq X \leq 3) = F_3 - F_0 = 0.87913 - 0.10737 = 0.77176 = 0.772$$

to 3 decimal places.

Appendix

`R` code to reproduce Figure 3:

```
r <- 0:10
n <- 10
p <- 0.2

Fr <- pbinom(r, n, p)
pr <- dbinom(r, n, p)

plot(r, pr,
 type = "h",
 main = paste("Probability Density Function \n B(",n,", ",p,")",
              sep = ""))

plot(r, Fr,
 type = "h",
 main = paste("Cumulative Distribution Function \n B(",n,", ",p,")",
              sep = ""))
```

in order to get Figure 4, you only have to change the value of `p`.

4.5 Extra Exercises

Bolt Factory

In a bolt factory 20% of the bolts produced by a machine are defective. The bolts are sold in boxes containing 5 pieces. What is the probability that a box contains at most one defective bolt?

Airline Tickets

A regional airline uses small 37 seat aircraft. From previous records the airline knows that 30% of all those making reservations do not appear for the trip, so they are selling 44 tickets for a flight.

a. What is the probability that at least one passenger cannot take the flight?
b. What is the probability that the flight departs with between 2 and 4 empty seats?