Chapter 5

Discrete Distributions

References: Pruim 2.1.3, 2.3.1, Larsen & Marx 3.2, 3.3

Statistics and data science are concerned with numbers, but probability is focused on sample spaces and events. In this chapter we will discuss an important construct - random variables - which enables us to connect the two.

A **random variable** is a number which is obtained as or from the result of a random experiment. For example, say we flip a coin 3 times, then our sample space has $2^3 = 8$ possible outcomes:

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Typically, the particular sequence of heads or tails is of little interest; what does matter is the number of heads that result.

If we define

$$X = \text{number of heads in 3 tosses,}$$

then we have captured the essence of the problem. Note: in defining X, we have actually defined a mapping (a function) from the original sample space S to a set of numbers.

$$X(HHH) = 3, X(HHT) = 2, X(HTH) = 2, X(HTT) = 1$$

$$X(THH) = 2, X(THT) = 1, X(TTH) = 1, X(TTT) = 0$$

We are now ready to make a formal definition of a random variable.

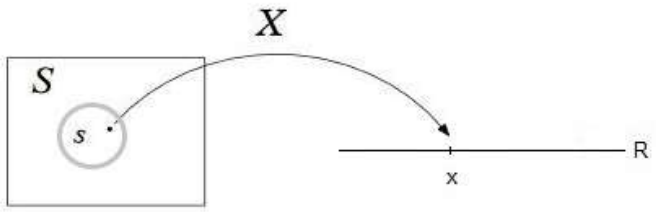
Definition 5.1. Let S be a sample space associated with a random experiment, $\mathbb{R}$ is the real line and let

$$X : S \to \mathbb{R}.$$

Then X is called a **random variable** and

$$P(X = x) = P(s \in S : X(s) = x),$$

where x is a possible value for X.



In the process of defining a random variable, we have also created a new sample space (the range of the random variable). Random variables often create a dramatically simpler sample space. They also allow us to describe certain kinds of events very succinctly. In the coin rolling example, we can now write $P(X = 2)$ instead of P(there are two heads and one tail).

Example 5.1. Independent trials consisting of the flipping of a coin having probability $\frac{1}{2}$ of coming up heads are continually performed until a head occurs. The sample space is

$$S = \{(H), (T, H), (T, T, H), (T, T, T, H) \ldots\}$$

Suppose we define the random variable X as the number of tosses until the first head appears.

a. What is the $range(X)$?

b. Express the following event in random variable notation and calculate its probability: at least three tosses must be made before the first head is observed.

. . . . . .

We will be concerned with two main types of random variables: discrete and

continuous. A **discrete** random variable is one that can only take on a finite or countably infinite set of values. A **continuous** random variable can take on all the values in an interval.

X, the number of heads in 3 flips of a coin is clearly a discrete random variable since $range(X) = \{0, 1, 2, 3\}$.

Let Y be the life length of a randomly selected bulb. Then Y is theoretically a continuous random variable since $range(Y) = [0, \infty)$. In practice, we may only measure Y to the nearest hour (or minute), but it is still more useful to model this situation by considering the underlying continuous random variable.

.....

Example 5.2. In each case, state whether the random variable is discrete or continuous.

a. The distance traveled by a football when thrown.

b. Toss a coin repeatedly until the first head appears and record the number of tails.

.....

We will turn our attention to describing the distribution of a random variable. We do this differently depending on whether the random variable is discrete or continuous. In this chapter, we will focus on the discrete case.

5.1 Probability Mass Functions

Recall that the distribution of a column of data in a dataset describes what values occur and with what frequency. What we need now is a way to do the same thing for a random variable. The difference is since the value of a random variable is based on the outcome of a random experiment, we will describe the possible values and the probabilities assigned to them. This is called a **probability distribution**.

One useful way to describe the probability distribution of a discrete random variable, especially one with a finite range, is by way of a table.

For example, for the random variable X which counts the number of heads in 3 tosses of a fair coin, we can produce the following table showing the possible values and their probabilities which were calculated assuming all 8 outcomes in the sample space are equally likely:

X	Number of heads in 3 tosses			
x	0	1	2	3
probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

This is an example of a **probability mass function** or PMF. As the name suggests, the PMF is associated with "point probabilities". Note that the table only shows the values which have positive probability.

Definition 5.2. For a discrete random variable X, we define the **Probability Mass Function (PMF)** $f(x)$ by:

$$f(x) = P(X = x), \ \forall \ x.$$

We will write f_X when we want to emphasize the random variable.

The PMF $f(x)$ of a discrete random variable can be positive for at most a countable number of values. That is, if X can take values x_1, x_2,x_3, ... then

$$\begin{aligned} f(x_i) &> 0, \ i = 1, 2, 3, ... \\ &= 0 \text{ otherwise.} \end{aligned}$$

Furthermore, since X must take one of the values x_i, we have

$$\sum_{i=1}^{\infty} f(x_i) = 1.$$

.....

Example 5.3. A store manager receives a shipment of 30 microwave ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens randomly without replacement from the shipment and tests them to see if they are defective. Let X denote the the number of defective ovens found.

Write the PMF of X in a tabular format.

.....

It is often instructive to present the PMF in a graphical format called a probability histogram. The vertical blocks of the histogram are each of width 1 and centered on the possible values of the random variable X. The height[1] of the blocks represents the probability that the random variable takes on a value covered by the base of the block.

```
# create data frame consisting of xi and f(xi)
library(tidyverse)
ovens <- data.frame(
  x = 0:4,
  f = c(0.4616, 0.4196, 0.1095, 0.0091, 0.0002)
)

ovens
##   x      f
## 1 0 0.4616
## 2 1 0.4196
```

[1] more accurately, the area which is equal to the height in the case of probability histograms

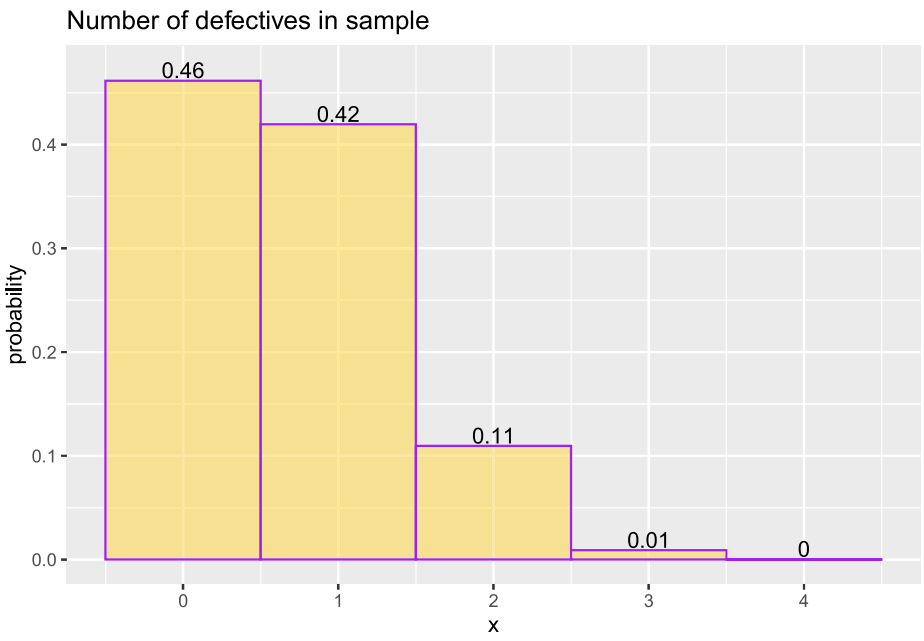
```
## 3 2 0.1095
## 4 3 0.0091
## 5 4 0.0002

# check that f sums to 1

ovens %>% summarise(sum(f))
##   sum(f)
## 1      1

# make probability histogram

ggplot(data = ovens, mapping = aes(x = x, y = f)) +
  geom_col(width = 1, alpha = 0.5, fill = "gold", color = "purple") +
  geom_text(mapping = aes(label = round(f, 2), y = f + 0.01)) +
  labs(
    x = "x",
    y = "probability",
    title = "Number of defectives in sample"
  )
```



For random variables with infinitely many possible values, a formula provides the most concise representation of the PMF. This is illustrated below for the random variable we saw in an earlier example.

Example 5.4. Let X denote the number of tosses until the first head when tossing a fair coin. Find the P.M.F. of X. You may assume the outcome on one toss is independent of the outcome on a different toss.

5.2 Cumulative Distribution Function

There is yet one more important way to describe the distribution of a discrete random variable, with a **cumulative distribution function**.

Definition 5.3. The **Cumulative Distribution Function (CDF)** F of a random variable X at a value x is defined by

$$F(x) = P(X \leq x) \quad \forall\, x.$$

Let's construct the CDF for the number of heads in 3 tosses of a fair coin from the PMF which was introduced at the beginning of the previous section and is

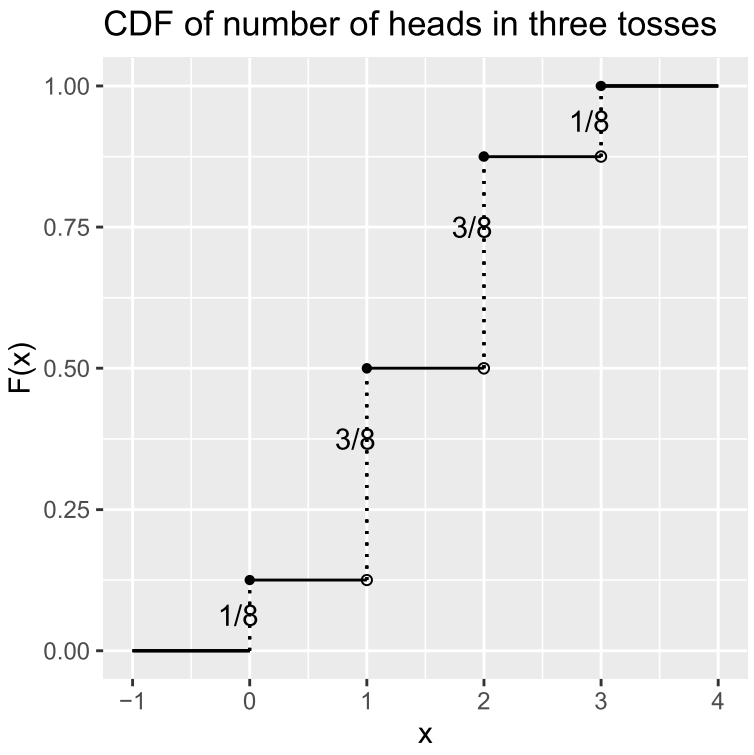
shown again below.

X;	Number of heads in 3 tosses			
x	0	1	2	3
probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The CDF F is a step function which jumps at each of the possible values x and the size of the jump is equal to $P(X = x)$.

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \leq x < 1, \\ \frac{1}{2} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases} \end{aligned}$$

This function is depicted graphically below.



Example 5.5. A discrete uniform random variable X has a PMF of the form

$$f(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n.$$

Find the CDF of X.

Both the CDF and the PMF completely describe the distribution of a random variable and effectively contain all the information about it. There is of course a connection between the PMF and CDF of a given random variable. To get the CDF from the PDF, we simply add up the probabilities for all possible values up to and including x. To get the PMF from the CDF, we look at how much the CDF has changed from the last jump.

5.3 Practice Problems

1. Find the value of c that makes the following a valid PMF:

$$f(x) = c/(x+1), \quad x = 0, 1, 2, 3$$

2. Two dice are rolled. You may assume each of the $6^2 = 36$ outcomes are equally likely. Write the PMF of the random variable "maximum of the two rolls" in a tabular format.

3. For a random variable X with the following CDF:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/2 & 0 \leq x < 1 \\ 3/5 & 1 \leq x < 2 \\ 4/5 & 2 \leq x < 3 \\ 9/10 & 3 \leq x < 3.5 \\ 1 & x \geq 3.5 \end{cases}$$

what is:

a. $f(1)$
b. $f(1.5)$
c. $f(3.5)$
d. $P(X \leq 3 | X > 1)$
e. $P(X \leq 3 | X \geq 1)$