Introduction to Statistics

Introduction, examples and definitions

Introduction

We begin the module with some basic data analysis. Since Statistics involves the collection and interpretation of data, we must first know how to understand, display and summarise large amounts of quantitative information, before undertaking a more sophisticated analysis.

Statistical analysis of quantitative data is important throughout the pure and social sciences. For example, during this module we will consider examples from Biology, Medicine, Agriculture, Economics, Business and Meteorology.

Examples

Survival of cancer patients: A cancer patient wants to know the probability that he will survive for at least 5 years. By collecting data on survival

rates of people in a similar situation, it is possible to obtain an empirical estimate of survival rates. We cannot know whether or not the patient will survive, or even know exactly what the *probability* of survival is. However, we can *estimate* the *proportion* of patients who survive from *data*.

Car maintenance: When buying a certain type of new car, it would be useful to know how much it is going to cost to run over the first three years from new. Of course, we cannot predict exactly what this will be — it will vary from car to car. However, collecting data from people who bought similar cars will give some idea of the *distribution* of costs across the *population* of car buyers, which in turn will provide information about the *likely* cost of running the car.

Definitions

The quantities measured in a study are called *random variables*, and a particular outcome is called an *observation*. Several observations are

collectively known as *data*. The collection of all possible outcomes is called the *population*.

In practice, we cannot usually observe the whole population. Instead we observe a sub-set of the population, known as a *sample*. In order to ensure that the sample we take is *representative* of the whole population, we usually take a *random sample* in which all members of the population are *equally likely* to be selected for inclusion in the sample. For example, if we are interested in conducting a survey of the amount of physical exercise undertaken by the general public, surveying people entering and leaving a gymnasium would provide a *biased* sample of the population, and the results obtained would *not* generalise to the population at large.

Variables are either *qualitative* or *quantitative*. Qualitative variables have non-numeric outcomes, with no natural ordering. For example, gender, disease status, and type of car are all qualitative variables. Quantitative variables have numeric outcomes. For example, survival time, height, age, number of children, and number of faults are all quantitative variables.

Quantitative variables can be *discrete* or *continuous*. Discrete random variables have outcomes which can take only a countable number of possible values. These possible values are usually taken to be integers, but don't have to be. For example, number of children and number of faults are discrete random variables which take only integer values, but your score in a quiz where "half" marks are awarded is a discrete quantitative random variable which can take on non-integer values. Continuous random variables can take any value over some continuous scale. For example, survival time and height are continuous random variables. Often, continuous random variables are rounded to the nearest integer, but the are still considered to be continuous variables if there is an underlying continuous scale. Age is a good example of this.

Data presentation

Introduction

A set of data on its own is very hard to interpret. There is lots of information contained in the data, but it is hard to see. We need ways of understanding important features of the data, and to summarise it in meaningful ways.

The use of *graphs* and *summary statistics* for understanding data is an important first step in the undertaking of any statistical analysis. For example, it is useful for understanding the main features of the data, for detecting *outliers*, and data which has been recorded incorrectly. Outliers are extreme observations which do not appear to be consistent with the rest of the data. The presence of outliers can seriously distort some of the more formal statistical techniques to be examined in the second semester, and so preliminary detection and correction or accommodation of such observations is crucial, before further analysis takes place.

Frequency tables

It is important to investigate the *shape* of the *distribution* of a random variable. This is most easily examined using *frequency tables* and *diagrams*. A frequency table shows a tally of the number of data observations in different *categories*.

For *qualitative* and *discrete quantitative* data, we often use all of the observed values as our categories. However, if there are a large number of different

observations, consecutive observations may be *grouped* together to form combined categories.

Example

For Example 1 (germinating seeds), we can construct the following frequency table.

No. germinating	85	86	87	88	89	90	91	92	93	94
Frequency	3	1	5	2	3	6	11	4	4	1

$n = 40$

Since we only have 10 categories, there is no need to amalgamate them.

For *continuous* data, the choice of categories is more arbitrary. We usually use 8 to 12 *non-overlapping consecutive intervals* of *equal width*. Fewer than this may be better for small sample sizes, and more for very large samples. The intervals must cover the entire observed range of values.

Example

For Example 2 (survival times), we have the following table.

Range			Frequency
0	—	39	11
40	—	79	4
80	—	119	1
120	—	159	1
160	—	199	2
200	—	240	1

$n = 20$

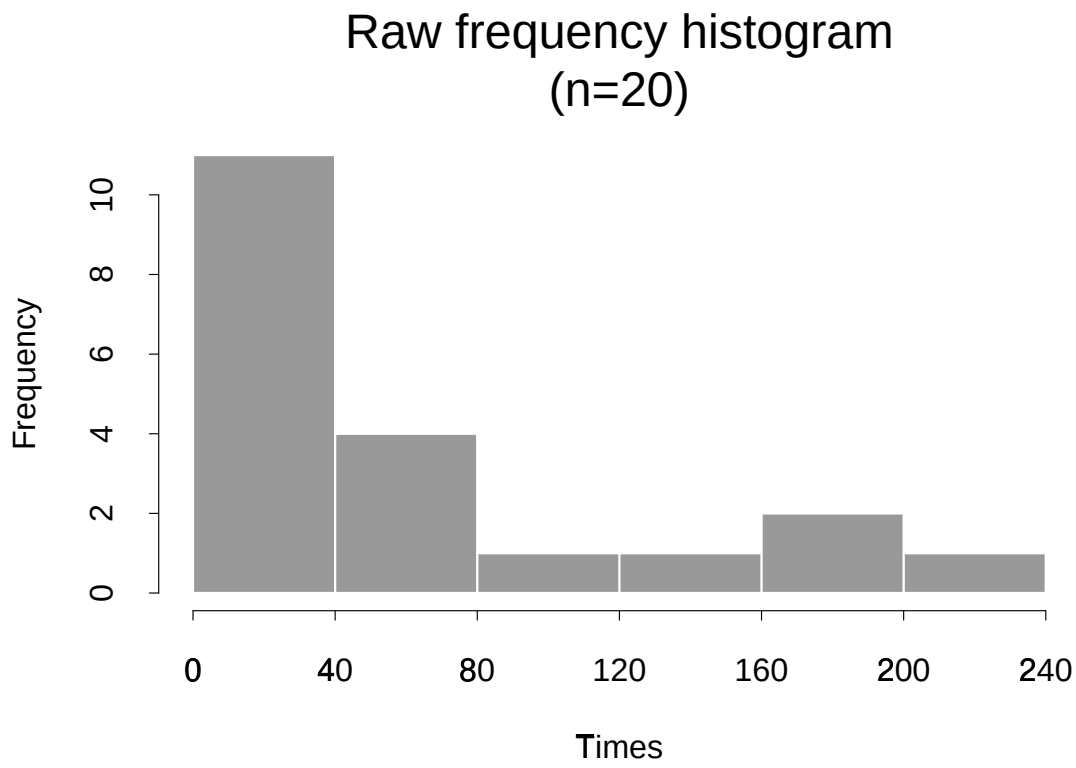
N.B. You should define the intervals to the same accuracy of the data. Thus, if the data is defined to the nearest integer, the intervals should be (as above). Alternatively, if the data is defined to one decimal place, so should the intervals. Also note that here the underlying data is continuous. Consequently, if the data has been *rounded* to the nearest integer, then the intervals are actually $0 - 39.5$, $39.5 - 79.5$, *etc.* It is important to include the sample size with the table.

Histograms

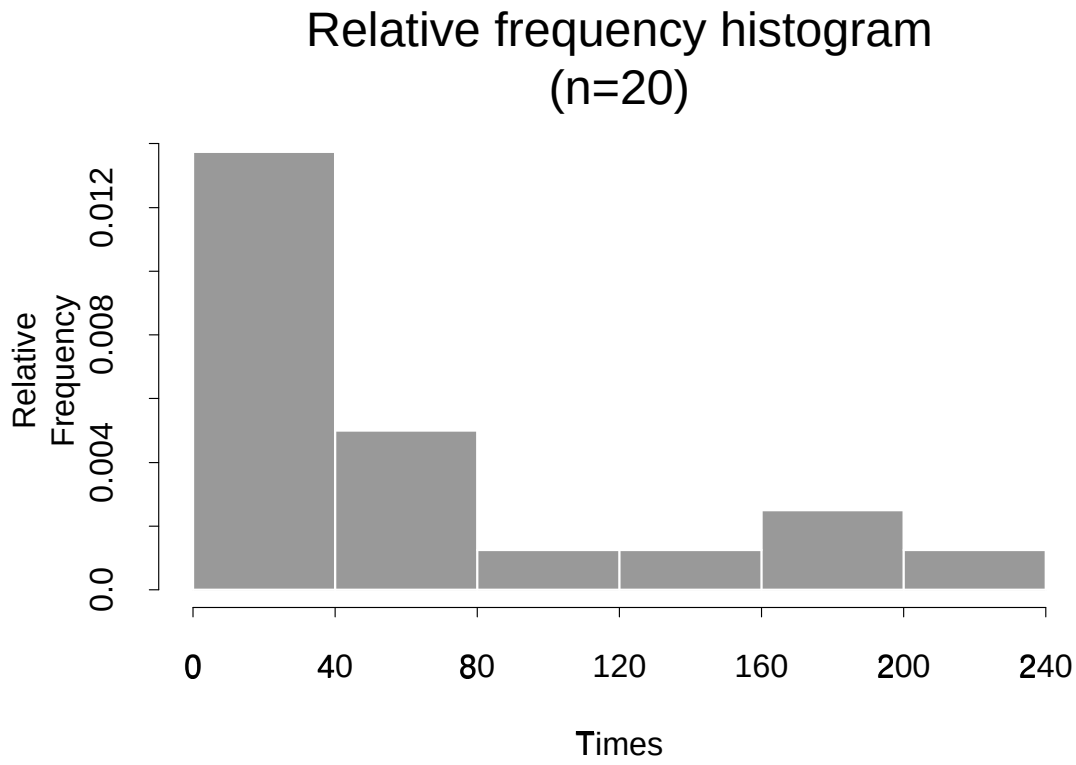
Once the frequency table has been constructed, pictorial representation can be considered. For most continuous data sets, the best diagram to use is a *histogram*. In this the classification intervals are represented to scale on the abscissa (x-axis) of a graph and rectangles are drawn on this base with their *areas* proportional to the frequencies. Hence the ordinate (y-axis) is frequency per unit class interval (or more commonly, *relative frequency* — see below). Note that the *heights* of the rectangles will be proportional to the frequencies if and only if class intervals of equal width are used.

Example

The histogram for Example 2 is as follows.



Note that here we have labelled the y-axis with the raw frequencies. This only makes sense when all of the intervals are the same width. Otherwise, we should label using relative frequencies, as follows.

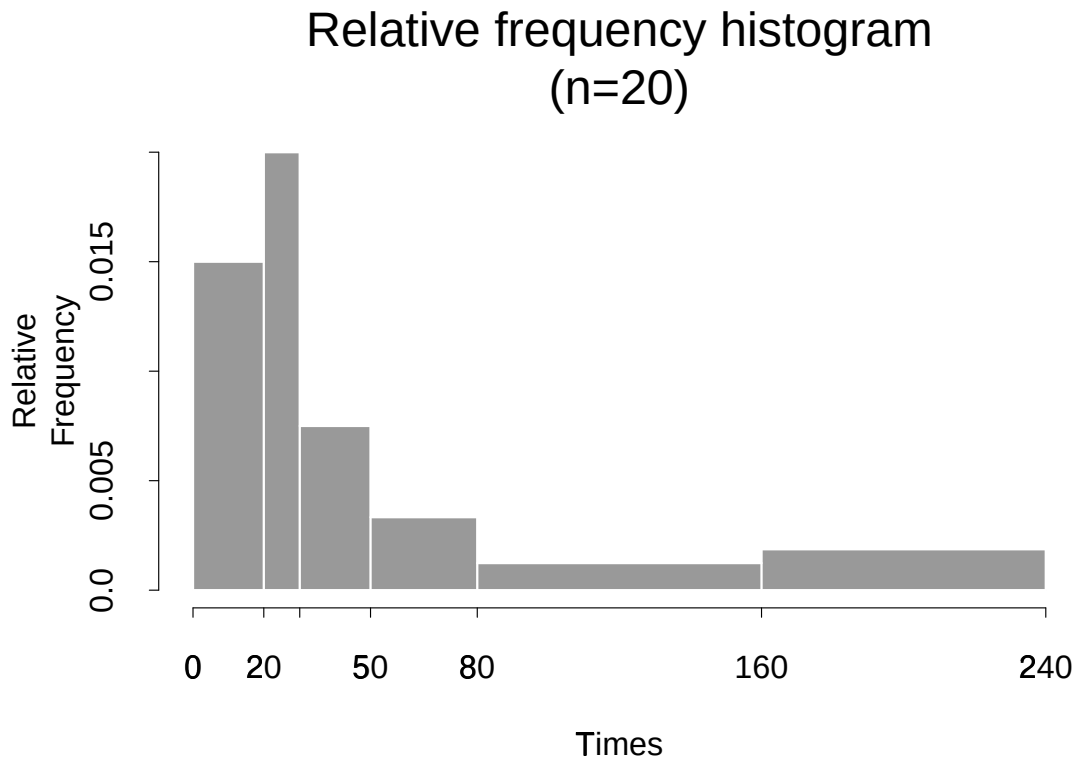


The y-axis values are chosen so that the area of each rectangle is the proportion of observations falling in that bin. Consider the first bin (0–39). The proportion of observations falling into this bin is 11/20 (from the frequency table). The area of our rectangle should, therefore, be 11/20. Since the rectangle has a base of 40, the height of the rectangle must be $11/(20 \times 40) = 0.014$. In general therefore, we calculate the bin height as

follows:

$$\text{Height} = \frac{\text{Frequency}}{n \times \text{BinWidth}}.$$

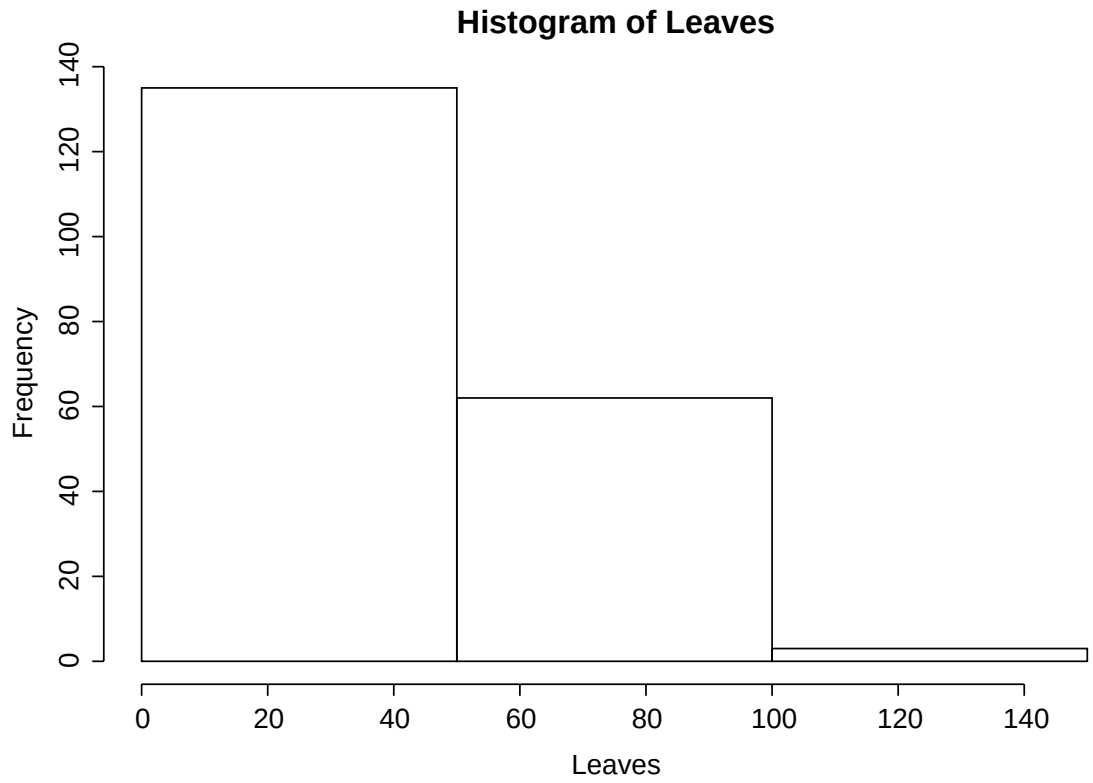
This method can be used when the interval widths are not the same, as shown below.

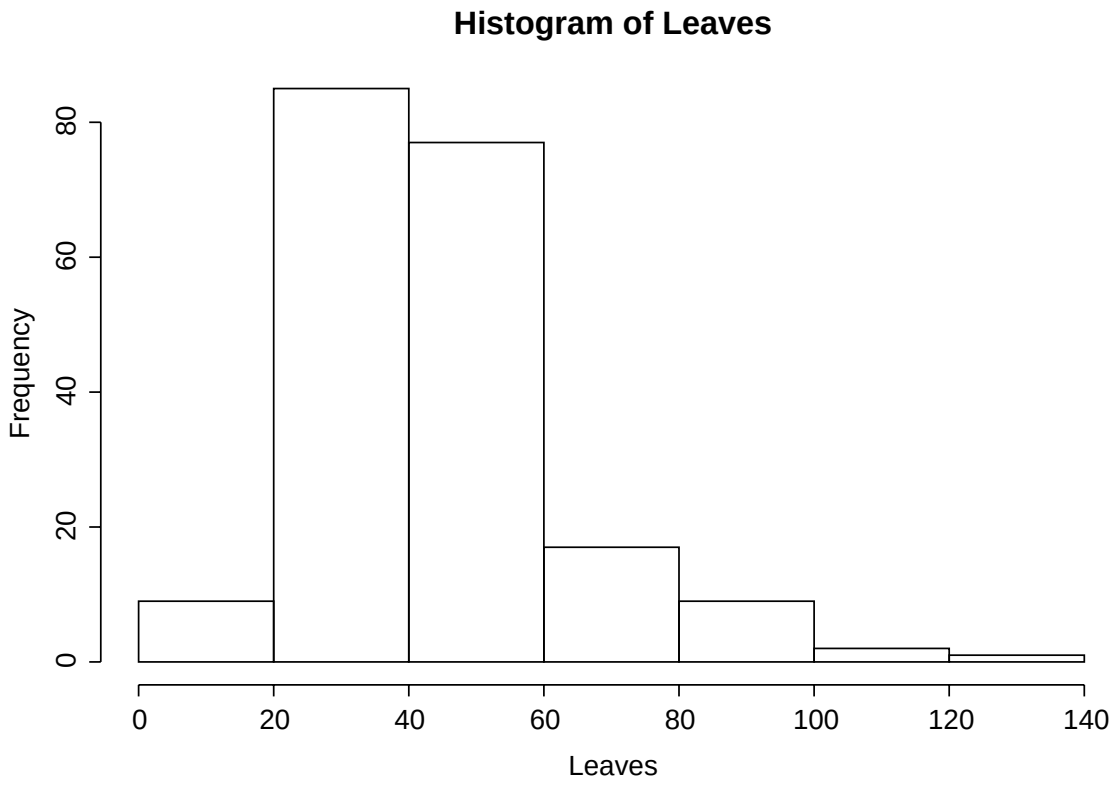


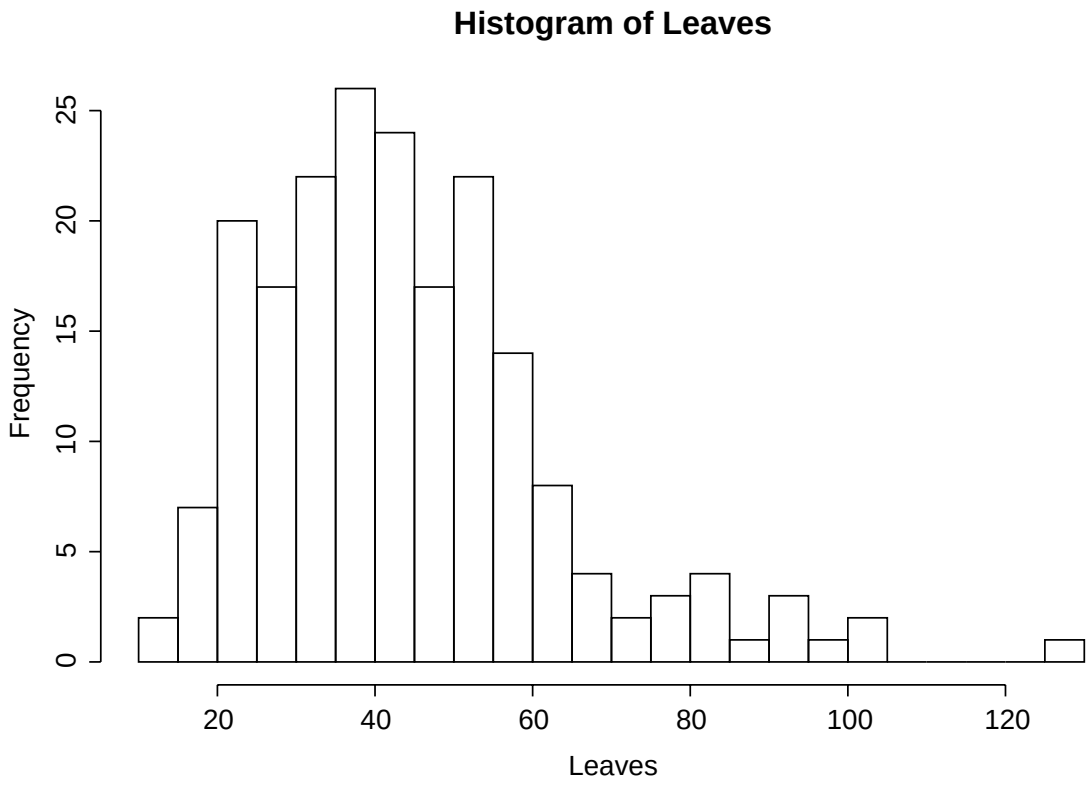
Note that when the y-axis is labelled with relative frequencies, the area under the histogram is always one. Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation.

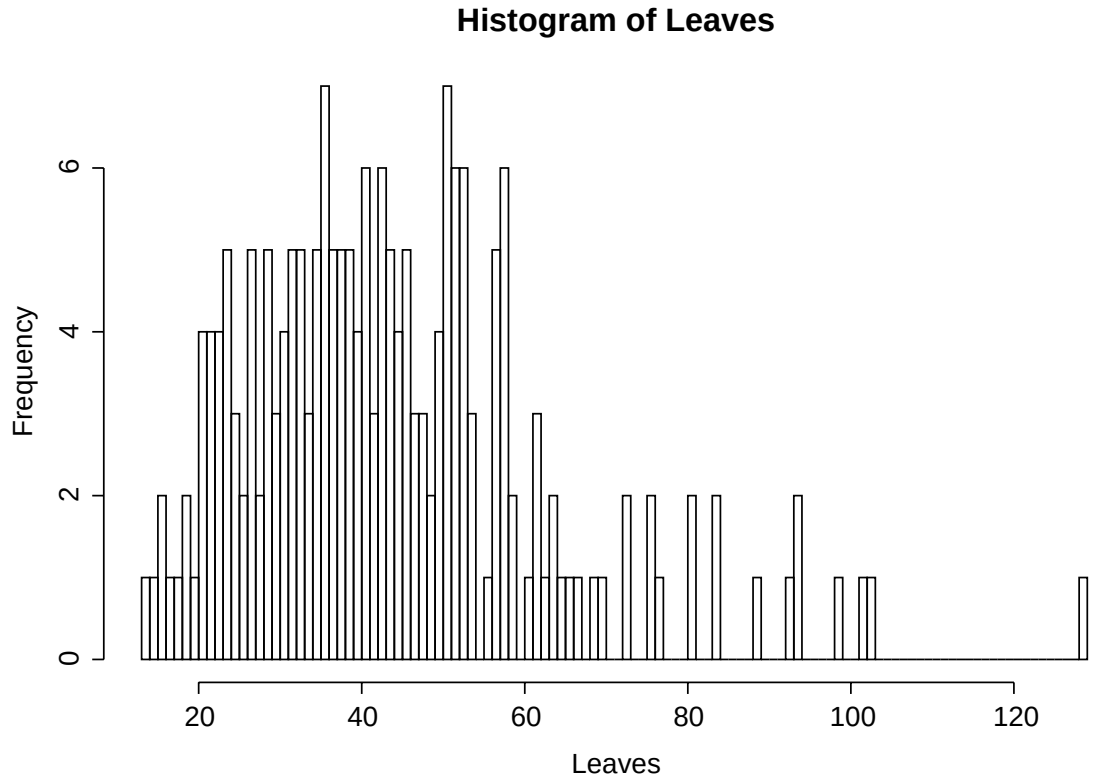
Example

Consider the leaf area data from Example 3. There follows some histograms of the data based on different bin widths. Which provides the best overview of the distribution of the data?




Histogram of Leaves
Frequency
0
20
40
60
80
0
20
40
60
80
100
120
140
Leaves


Histogram of Leaves
Frequency
0
5
10
15
20
25
20
40
60
80
100
120
Leaves



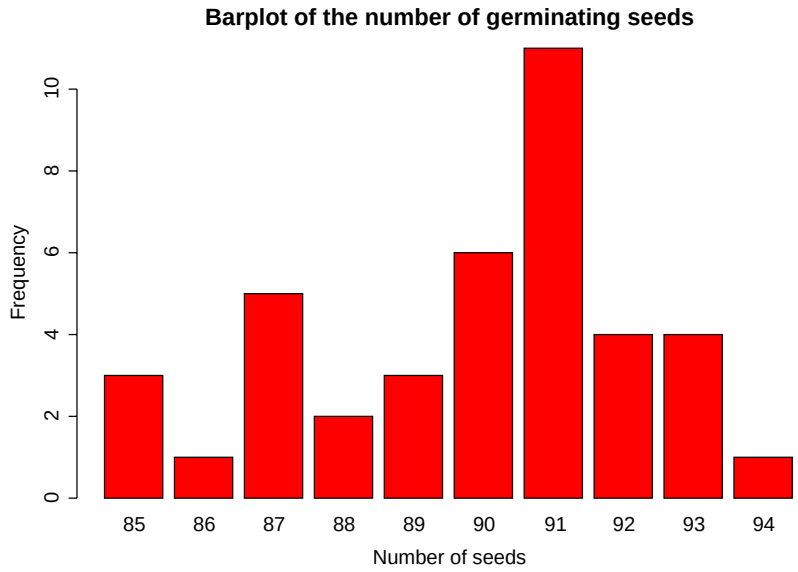
Bar charts and frequency polygons

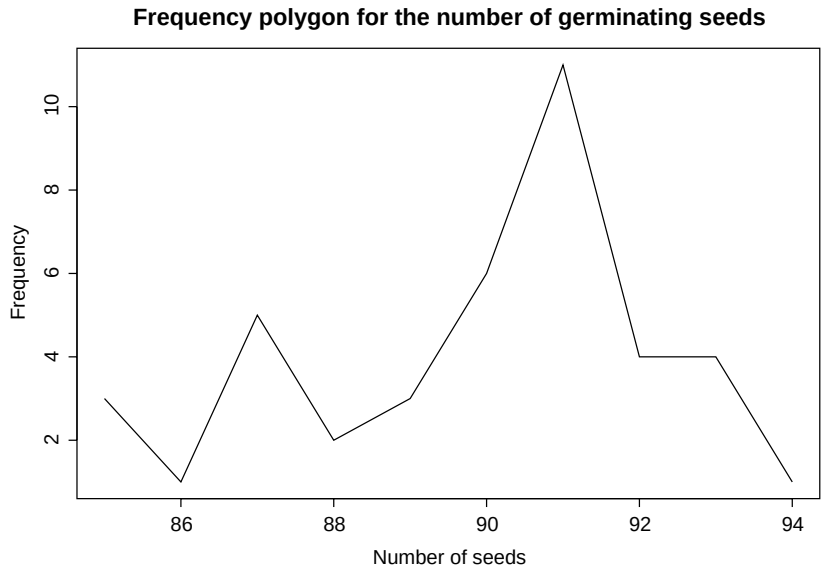
When the data are discrete and the frequencies refer to individual values, we display them graphically using a *bar chart* with heights of bars representing frequencies, or a *frequency polygon* in which only the tops of the bars are marked, and then these points are joined by straight lines. Bar charts are drawn with a gap between neighbouring bars so that they are easily

distinguished from histograms. Frequency polygons are particularly useful for comparing two or more sets of data.

Example

Consider again the number of germinating seeds from Example 1. Using the frequency table constructed earlier, we can construct a Bar Chart and Frequency Polygon as follows.





$$n = 40$$

A third method which is sometimes used for *qualitative* data is called a *pie chart*. Here, a circle is divided into sectors whose *areas*, and hence *angles* are proportional to the *frequencies* in the different categories. Pie charts should generally not be used for *quantitative* data – a bar chart or frequency polygon is almost always to be preferred.

Whatever the form of the graph, it should be clearly labelled on each axis and a fully descriptive title should be given, together with the number of observations on which the graph is based.

Stem-and-leaf plots

A good way to present both continuous and discrete data for sample sizes of less than 200 or so is to use a *stem-and-leaf plot*. This plot is similar to a bar chart or histogram, but contains more information. As with a histogram, we normally want 5–12 intervals of equal size which span the observations. However, for a stem-and-leaf plot, the widths of these intervals must be 0.2, 0.5 or 1.0 times a power of 10, and we are not free to choose the end-points of the bins. They are best explained in the context of an example.

Example

Recall again the seed germination Example 1. Since the data has a range of 9, an interval width of $2\ (= 0.2 \times 10^1)$ seems reasonable. To form the plot, draw a vertical line towards the left of the plotting area. On the left of this mark the interval boundaries in increasing order, noting only those digits that are common to all of the observations within the interval. This is called the *stem* of the plot. Next go through the observations one by one, noting down the next significant digit on the right-hand side of the corresponding stem.

8	5	5	5														
8	7	7	7	7	6	7											
8	8	9	8	9	9												
9	1	1	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1
9	3	2	2	2	3	3	3	2									
9	4																

For example, the first stem contains any values of 84 and 85, the second stem contains any values of 86 and 87, and so on. The digits to the right of the vertical line are known as the *leaves* of the plot, and each digit is known as a *leaf*.

Now re-draw the plot with all of the leaves corresponding to a particular stem ordered increasingly. At the top of the plot, mark the sample size, and at the bottom, mark the stem and leaf units. These are such that an observation corresponding to any leaf can be calculated as

$$\text{Observation} = \text{StemLabel} \times \text{StemUnits} + \text{LeafDigit} \times \text{LeafUnits}$$

to the nearest leaf unit.

$$n = 40$$

8	5 5 5
8	6 7 7 7 7 7
8	8 8 9 9 9
9	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
9	2 2 2 2 3 3 3 3
9	4

Stem Units = 10 seeds,

Leaf Units = 1 seed.

The main advantages of using a stem-and-leaf plot are that it shows the general shape of the data (like a bar chart or histogram), and that the all of the data can be recovered (to the nearest leaf unit). For example, we can see from the plot that there is only one value of 94, and three values of 89.

Example

A stem-and-leaf plot for the data from Example 5 is given below. An interval width of $0.5\ (= 0.5 \times 10^0)$ is used.

$$n = 24$$

1	5	9	9			
2	0	1	3	4	4	
2	6	7	7	8	8	9
3	0	1	2	2		
3	5	6	8	8	8	
4	1					

Stem Units = 1,

Leaf Units = 0.1.

Note that in this example, the data is given with more significant digits than can be displayed on the plot. Numbers should be “cut” rather than rounded to the nearest leaf unit in this case. For example, 1.97 is cut to 1.9, and then entered with a stem of 1 and a leaf of 9. It is *not* rounded to 2.0.

Summary

Using the plots described in this section, we can gain an empirical understanding of the important features of the distribution of the data.

- Is the distribution *symmetric* or *asymmetric* about its central value?

- Are there any *unusual* or *outlying* observations, which are much *larger* or *smaller* than the main body of observations?

- Is the data *multi-modal*? That is, are there *gaps* or *multiple peaks* in the distribution of the data? Two peaks may imply that there are two different groups represented by the data.

- By putting plots side by side with the *same scale*, we may compare the distributions of different groups.

Summary measures

Measures of location

In addition to the graphical techniques encountered so far, it is often useful to obtain quantitative summaries of certain aspects of the data. Most simple summary measurements can be divided into two types; firstly quantities which are “typical” of the data, and secondly, quantities which summarise the variability of the data. The former are known as *measures of location* and the latter as *measures of spread*. Suppose we have a sample of size n of *quantitative* data. We will denote the measurements by $x_1, x_2, \ldots x_n$.

Sample mean

This is the most important and widely used measure of location. The *sample mean* of a set of data is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This is the location measure often used when talking about the *average* of a set of observations. However, the term “average” should be avoided, as all measures of location are different kinds of averages of the data.

If we have *discrete* quantitative data, tabulated in a frequency table, then if the possible outcomes are $y_1, \ldots, y_k$, and these occur with frequencies $f_1, \ldots, f_k$, so that $\sum f_i = n$, then the sample mean is

$$\bar{x} = \frac{f_1 y_1 + \cdots + f_k y_k}{n} = \frac{1}{n} \sum_{i=1}^{k} f_i y_i = \frac{1}{\sum f_i} \sum_{i=1}^{k} f_i y_i.$$

For *continuous* data, the sample mean should be calculated from the original data if this is known. However, if it is tabulated in a frequency table, and the original data is *not* known, then the sample mean can be *estimated* by assuming that all observations in a given interval occurred at the mid-point of that interval. So, if the mid-points of the intervals are $m_1, \ldots, m_k$, and the corresponding frequencies are $f_1, \ldots, f_k$, then sample mean can be approximated using

$$\bar{x} \simeq \frac{1}{n} \sum_{i=1}^{k} f_i m_i.$$

Sample median

The *sample median* is the middle observation when the data are *ranked* in increasing order. We will denote the ranked observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. If

there are an even number of observations, there is no middle number, and so the median is defined to be the sample mean of the middle two observations.

$$\text{SampleMedian} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ odd}, \\ \\ \frac{1}{2}x_{(\frac{n}{2})} + \frac{1}{2}x_{(\frac{n}{2}+1)}, & n \text{ even}. \end{cases}$$

The sample median is sometimes used in preference to the sample mean, particularly when the data is asymmetric, or contains outliers. However, its mathematical properties are less easy to determine than those of the sample mean, making the sample mean preferable for formal statistical analysis. The ranking of data and calculation of the median is usually done with a stem-and-leaf plot when working by hand. Of course, for large amounts of data, the median is calculated with the aid of a computer.

Sample mode

The mode is the value which occurs with the greatest frequency. Consequently, it only really makes sense to calculate or use it with discrete data, or for continuous data with small grouping intervals and large sample

sizes. For discrete data with possible outcomes $y_1, \ldots, y_k$ occurring with frequencies $f_1, \ldots, f_k$, we may define the sample mode to be

$$\text{SampleMode} = \{y_k | f_k = \max_i \{f_i\}\}.$$

That is, the y_k whose corresponding f_k is largest.

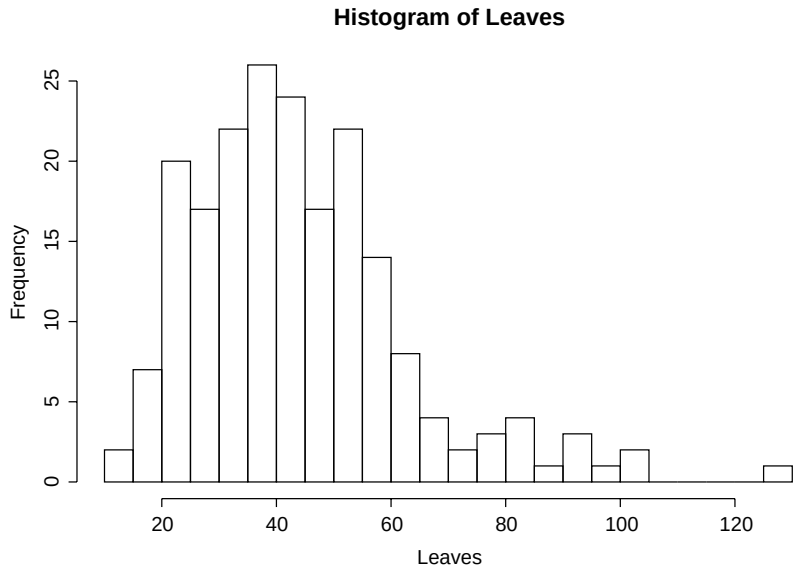
Summary of location measures

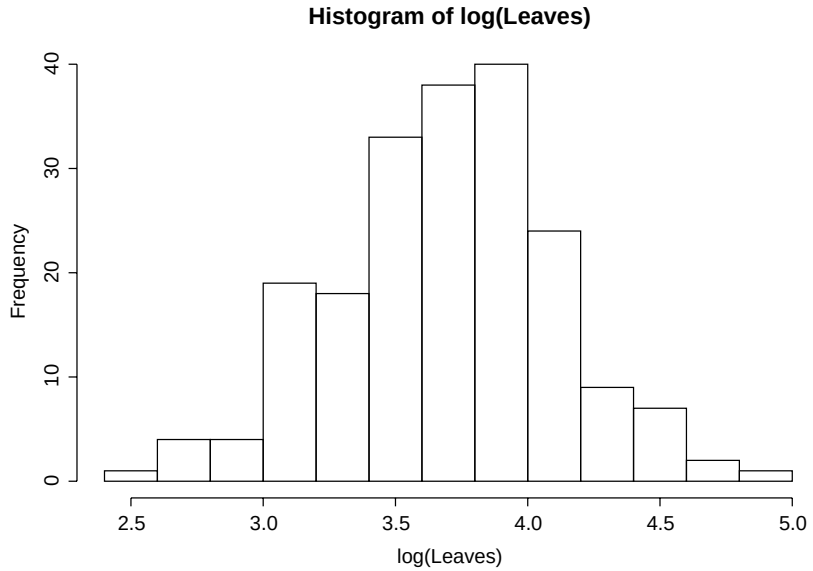
As we have already remarked, the sample mean is by far the most important measure of location, primarily due to its attractive mathematical properties (for example, the sample mean of the sum of two equal length columns of data is just the sum of the sample means of the two columns). When the distribution of the data is roughly symmetric, the three measures will be very close to each other anyway. However, if the distribution is very skewed, there may be a considerable difference, and all three measures could be useful in understanding the data. In particular, the sample median is a much more *robust* location estimator, much less sensitive than the sample mean to asymmetries and unusual values in the data.

In order to try and overcome some of the problems associated with skewed data, such data is often *transformed* in order to try and get a more symmetric distribution. If the data has a longer tail on the left (smaller values) it is known as *left-skewed* or *negatively skewed*. If the data has a longer tail on the right (larger values), then it is known as *right-skewed* or *positively skewed*. N.B. *This is the opposite to what many people expect these terms to mean*, as the "bulk" of the distribution is shifted in the opposite direction on automatically scaled plots. If the data is positively skewed, then we may take square roots or logs of the data. If it is negatively skewed, we may square or exponentiate it.

Example

One of the histograms for Example 3, the leaf area data, is repeated below. We can see that the long tail is on the right, and so this data is *positively* or *right* skewed. This is the case despite the fact that the bulk of the distribution is shifted to the left of the plot. If we look now at a histogram of the logs of this data, we see that it is much closer to being symmetric. The sample mean and median are much closer together (relatively) for the transformed data.





Measures of spread

Knowing the “typical value” of the data alone is not enough. We also need to know how “concentrated” or “spread out” it is. That is, we need to know something about the “variability” of the data. Measures of spread are a way of quantifying this idea numerically.

Range

This is the difference between the largest and smallest observation. So, for our ranked data, we have

$$\text{Range} = x_{(n)} - x_{(1)}$$

This measure can sometimes be useful for comparing the variability of samples of the same size, but it is not very robust, and is affected by sample size (the larger the sample, the bigger the range), so it is not a fixed characteristic of the population, and cannot be used to compare variability of different sized samples.

Mean absolute deviation (M.A.D.)

This is the average absolute deviation from the sample mean.

$$\text{M.A.D.} = \frac{|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

where

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0. \end{cases}$$

The M.A.D. statistic is easy to understand, and is often used by non-statisticians. However, there are strong theoretical and practical reasons for preferring the statistic known as the *variance*, or its square root, the *standard deviation*.

Sample variance and standard deviation

The *sample variance*, s^2 is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\}.$$

It is the average squared distance of the observations from their mean value. The second formula is easier to calculate with. The divisor is $n-1$ rather than

n in order to correct for the bias which occurs because we are measuring deviations from the sample mean rather than the “true” mean of the population we are sampling from — more on this in Semester 2.

For *discrete* data, we have

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(y_i - \bar{x})^2$$

and for continuous tabulated data we have

$$s^2 \simeq \frac{1}{n-1}\sum_{i=1}^{k} f_i(m_i - \bar{x})^2.$$

The *sample standard deviation*, s, is just the square root of the sample variance. It is preferred as a summary measure as it is in the units of the original data. However, it is often easier from a theoretical perspective to work with variances. Thus the two measures are complimentary.

Most calculators have more than one way of calculating the standard deviation of a set of data. Those with σ_n and σ_{n-1} keys give the sample

standard deviation by pressing the σ_{n-1} key, and those with σ and s keys give it by pressing the s key.

When calculating a summary statistic, such as the mean and standard deviation, it is useful to have some idea of the likely value in order to help spot arithmetic slips or mistakes entering data into a calculator or computer. The sample mean of a set of data should be close to fairly typical values, and $4s$ should cover the range of the bulk of your observations.

Quartiles and the interquartile range

Whereas the median has half of the data less than it, the *lower quartile* has a *quarter* of the data less than it, and the *upper quartile* has a quarter of the data above it. So the lower quartile is calculated as the $(n+1)/4^{th}$ smallest observation, and the upper quartile is calculated as the $3(n+1)/4^{th}$ smallest observation. Again, if this is not an integer, *linearly interpolate* between adjacent observations as necessary (examples below). There is no particularly compelling reason why $(n+1)/4$ is used to define the position of the lower quartile — $(n+2)/4$ and $(n+3)/4$ seem just as

reasonable. However, the definitions given are those used by Minitab, which seems as good a reason as any for using them!

Examples

Calculating lower quartiles

$n = 15$	LQ at $(15+1)/4 = 4$	LQ is $x_{(4)}$
$n = 16$	LQ at $(16+1)/4 = 4\frac{1}{4}$	LQ is $\frac{3}{4}x_{(4)} + \frac{1}{4}x_{(5)}$
$n = 17$	LQ at $(17+1)/4 = 4\frac{1}{2}$	LQ is $\frac{1}{2}x_{(4)} + \frac{1}{2}x_{(5)}$
$n = 18$	LQ at $(18+1)/4 = 4\frac{3}{4}$	LQ is $\frac{1}{4}x_{(4)} + \frac{3}{4}x_{(5)}$
$n = 19$	LQ at $(19+1)/4 = 5$	LQ is $x_{(5)}$

The *inter-quartile range* is the difference between the upper and lower quartiles, that is

$$\text{IQR} = \text{UQ} - \text{LQ}.$$

It measures the range of the middle 50% of the data. It is an alternative measure of spread to the standard deviation. It is of interest because it is much more robust than the standard deviation, and thus is often used to describe asymmetric distributions.

Coefficient of variation

A measure of spread that can be of interest is known as the *coefficient of variation*. This is the ratio of the standard deviation to the mean,

$$\text{Coefficient of variation} = \frac{s}{\bar{x}},$$

and thus has no units. The coefficient of variation does not change if the (linear) *scale*, but not the *location* of the data is changed. That is, if you take data $x_1, \ldots, x_n$ and transform it to new data, $y_1, \ldots, y_n$ using the mapping $y_i = \alpha x_i + \beta$, the coefficient of variation of $y_1, \ldots, y_n$ will be the same as the coefficient of variation of $x_1, \ldots, x_n$ if $\beta = 0$ and $\alpha > 0$, but not otherwise. So, the coefficient of variation would be the same for a set of length measurements whether they were measured in centimeters or inches (zero is the same on both scales). However, the coefficient of variation would be

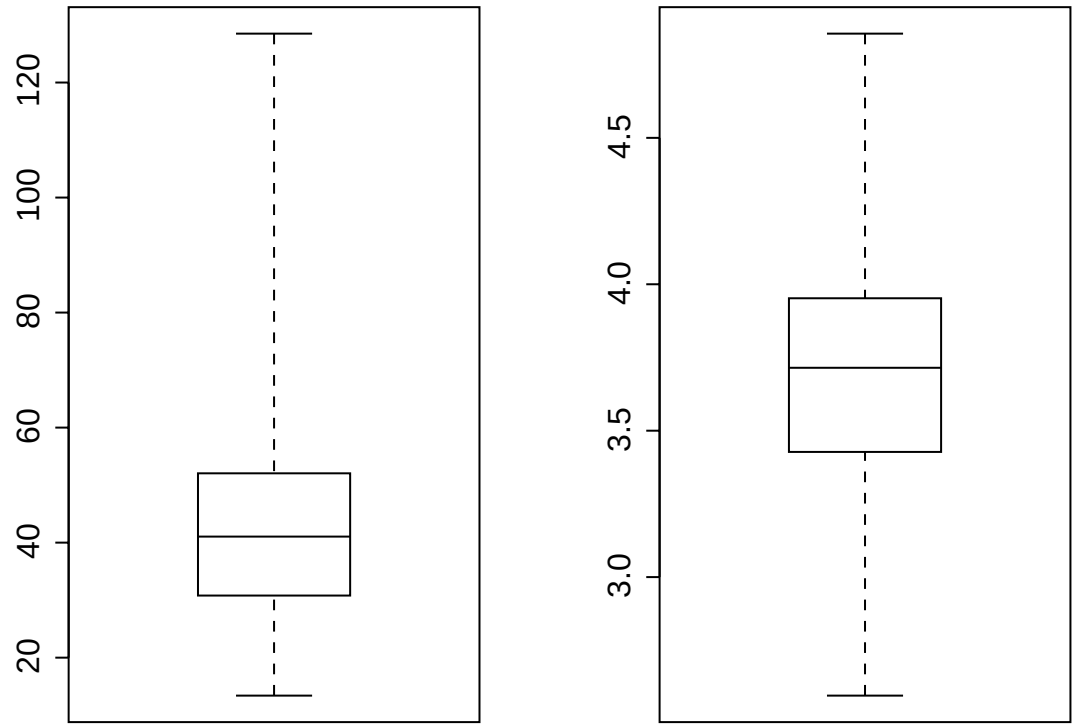
different for a set of temperature measurements made in Celsius and Fahrenheit (as the zero of the two scales is different).

Box-and-whisker plots

This is a useful graphical description of the main features of a set of observations. There are many variations on the box plot. The simplest form is constructed by drawing a rectangular *box* which stretches from the lower quartile to the upper quartile, and is divided in two at the median. From each end of the box, a line is drawn to the maximum and minimum observations. These lines are sometimes called *whiskers*, hence the name.

Example

Consider the data for Example 3, the leaf size data. Box plots for this data and the logs are given below. Notice how the asymmetry of the original distribution shows up very clearly on the left plot, and the symmetry of the distribution of the logs, on the right plot.



Box plots of raw and transformed leaf area data $(n = 200)$

Box-and-whisker plots are particularly useful for comparing several groups of observations. A box plot is constructed for each group and these are displayed on a common scale. At least 10 observations per group are required in order for the plot to be meaningful.

Introduction to Probability

Sample spaces, events and sets

Introduction

Probability is the language we use to model uncertainty. The data and examples we looked at in the last chapter were the *outcomes* of *scientific experiments*. However, those outcomes could have been different — many different kinds of *uncertainty* and *randomness* were part of the mechanism which led to the actual data we saw. If we are to develop a proper understanding of such experimental results, we need to be able to understand the randomness underlying them. In this chapter, we will look at the fundamentals of probability theory, which we can then use in the later chapters for modelling the outcomes of experiments, such as those discussed in the previous chapter.

Sample spaces

Probability theory is used as a model for situations for which the outcomes occur randomly. Generically, such situations are called *experiments*, and the set of all possible outcomes of the experiment is known as the *sample space* corresponding to an experiment. The sample space is usually denoted by S, and a generic element of the sample space (a possible outcome) is denoted by s. The sample space is chosen so that exactly one outcome will occur. The size of the sample space is *finite*, *countably infinite* or *uncountably infinite*.

Examples

Consider Example 1. The outcome of one of any replication is the number of germinating seeds. Since 100 seeds were monitored, the number germinating could be anything from 0 to 100. So, the sample space for the outcomes of this experiment is

$$S = \{0, 1, 2, \ldots, 100\}.$$

This is an example of a finite sample space. For Example 2, the survival time in weeks could be any non-negative integer. That is

$$S = \{0, 1, 2, \ldots\}.$$

This is an example of a countably infinite sample space. In practice, there is some upper limit to the number of weeks anyone can live, but since this upper limit is unknown, we include all non-negative integers in the sample space. For Example 3, leaf size could be any positive real number. That is

$$S = \mathbf{R}^+ \equiv (0, \infty).$$

This is an example of an uncountably infinite sample space. Although the leaf sizes were only measured to 1 decimal place, the actual leaf sizes vary continuously. For Example 4 (number of siblings), the sample space would be as for Example 2, and for Example 5 (plutonium measurements), the sample space would be the same as in Example 3.

Events

A *subset* of the sample space (a collection of possible outcomes) is known as an *event*. Events may be classified into four types:

- the *null event* is the empty subset of the sample space;

- an *atomic event* is a subset consisting of a single element of the sample space;

- a *compound event* is a subset consisting of more than one element of the sample space;

- the *sample space* itself is also an event.

Examples

Consider the sample space for Example 4 (number of siblings),

$$S = \{0, 1, 2, \dots\}$$

and the event *at most two siblings*,

$$E = \{0, 1, 2\}.$$

Now consider the event

$$F = \{1, 2, 3, \dots\}.$$

Here, F is the event *at least one sibling*.

The *union* of two events E and F is the event that at least one of E and F occurs. The union of the events can be obtained by forming the union of the sets. Thus, if G is the union of E and F, then we write

$$\begin{aligned} G &= E \cup F \\ &= \{0,1,2\} \cup \{1,2,3,\ldots\} \\ &= \{0,1,2,\ldots\} \\ &= S \end{aligned}$$

So the union of E and F is the whole sample space. That is, the events E and F together cover all possible outcomes of the experiment — at least one of E or F must occur.

The *intersection* of two events E and F is the event that both E and F occur. The intersection of two events can be obtained by forming the intersection of the sets. Thus, if H is the intersection of E and F, then

$$\begin{aligned} H &= E \cap F \\ &= \{0,1,2\} \cap \{1,2,3,\ldots\} \\ &= \{1,2\} \end{aligned}$$

So the intersection of E and F is the event *one or two siblings*.

The *complement* of an event, A, denoted A^c or $\bar{A}$, is the event that A does *not* occur, and hence consists of all those elements of the sample space that are not in A. Thus if $E = \{0,1,2\}$ and $F = \{1,2,\dots\}$,

$$E^c = \{3,4,5,\dots\}$$

and

$$F^c = \{0\}.$$

Two events A and B are *disjoint* or *mutually exclusive* if they cannot both occur. That is, their intersection in empty

$$A \cap B = \emptyset.$$

Note that for any event A, the events A and A^c are disjoint, and their union is the whole of the sample space:

$$A \cap A^c = \emptyset \quad \text{and} \quad A \cup A^c = S.$$

The event A is *true* if the outcome of the experiment, s, is contained in the event A; that is, if $s \in A$. We say that the event A *implies* the event B, and write

$A \Rightarrow B$, if the truth of B automatically follows from the truth of A. If A is a subset of B, then occurrence of A necessarily implies occurrence of the event B. That is

$$(A \subseteq B) \iff (A \cap B = A) \iff (A \Rightarrow B).$$

We can see already that to understand events, we must understand a little set theory.

Set theory

We already know about sets, complements of sets, and the union and intersection of two sets. In order to progress further we need to know the basic rules of set theory.

Commutative laws:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive laws:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

DeMorgan's laws:

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Disjoint union:

$$A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$$

and $A \cap B^c$, $A^c \cap B$ and $A \cap B$ are disjoint.

Venn diagrams can be useful for thinking about manipulating sets, but formal proofs of set-theoretic relationships should only rely on use of the above laws.

Probability axioms and simple counting problems

Probability axioms and simple properties

Now that we have a good mathematical framework for understanding events in terms of sets, we need a corresponding framework for understanding probabilities of events in terms of sets.

The real valued function $\mathrm{P}(\cdot)$ is a *probability measure* if it acts on subsets of S and obeys the following axioms:

I. $\mathrm{P}(S) = 1$.

II. If $A \subseteq S$ then $\mathrm{P}(A) \geq 0$.

III. If A and B are *disjoint* ($A \cap B = \emptyset$) then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)\,.$$

Repeated use of Axiom III gives the more general result that if $A_1, A_2, \ldots, A_n$ are mutually disjoint, then

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i)\,.$$

Indeed, we will assume further that the above result holds even if we have a *countably* infinite collection of disjoint events ($n = \infty$).

These axioms seem to fit well with our intuitive understanding of probability, but there are a few additional comments worth making.

1. Axiom I says that one of the possible outcomes must occur. A probability of 1 is assigned to the event "something occurs". This fits in exactly with our definition of sample space. Note however, that the implication does not go the other way! When dealing with infinite sample spaces, there are

often events of probability one which are not the sample space and events of probability zero which are not the empty set.

2. Axiom II simply states that we wish to work only with positive probabilities, because in some sense, probability measures the *size* of the set (event).

3. Axiom III says that probabilities "add up" — if we want to know the probability of *at most one sibling*, then this is the sum of the probabilities of *zero siblings* and *one sibling*. Allowing this result to hold for countably infinite unions is slightly controversial, but it makes the mathematics much easier, so we will assume it throughout!

These axioms are all we need to develop a theory of probability, but there are a collection of commonly used properties which follow directly from these axioms, and which we make extensive use of when carrying out probability calculations.

Property A: $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

Property B: $\mathrm{P}(\emptyset) = 0$.

Property C: If $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$.

Property D: (Addition Law) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Interpretations of probability

Somehow, we all have an intuitive feel for the notion of probability, and the axioms seem to capture its essence in a mathematical form. However, for probability theory to be anything other than an interesting piece of abstract pure mathematics, it must have an interpretation that in some way connects it to reality. If you wish only to study probability as a mathematical theory, then there is no need to have an interpretation. However, if you are to use probability theory as your foundation for a theory of statistical inference which makes probabilistic statements about the world around us, then there must be an interpretation of probability which makes some connection between the mathematical theory and reality.

Whilst there is (almost) unanimous agreement about the mathematics of probability, the axioms and their consequences, there is considerable disagreement about the interpretation of probability. The three most common interpretations are given below.

Classical interpretation

The classical interpretation of probability is based on the assumption of underlying equally likely events. That is, for any events under consideration, there is always a sample space which can be considered where all atomic events are equally likely. If this sample space is given, then the probability axioms may be deduced from set-theoretic considerations.

This interpretation is fine when it is obvious how to partition the sample space into equally likely events, and is in fact entirely compatible with the other two interpretations to be described in that case. The problem with this interpretation is that for many situations it is not at all obvious what the partition into equally likely events is. For example, consider the probability that it rains in Newcastle tomorrow. This is clearly a reasonable event to

consider, but it is not at all clear what sample space we should construct with equally likely outcomes. Consequently, the classical interpretation falls short of being a good interpretation for real-world problems. However, it provides a good starting point for a mathematical treatment of probability theory, and is the interpretation adopted by many mathematicians and theoreticians.

Frequentist interpretation

An interpretation of probability widely adopted by statisticians is the relative frequency interpretation. This interpretation makes a much stronger connection with reality than the previous one, and fits in well with traditional statistical methodology. Here probability only has meaning for events from experiments which could in principle be repeated arbitrarily many times under essentially identical conditions. Here, the probability of an event is simply the "long-run proportion" of times that the event occurs under many repetitions of the experiment. It is reasonable to suppose that this proportion will settle down to some limiting value eventually, which is the probability of the event. In such a situation, it is possible to derive the axioms of probability from

consideration of the long run frequencies of various events. The probability p, of an event E, is defined by

$$p = \lim_{n \to \infty} \frac{r}{n}$$

where r is the number of times E occurred in n repetitions of the experiment.

Unfortunately it is hard to make precise exactly why such a limiting frequency should exist. A bigger problem however, is that the interpretation only applies to outcomes of repeatable experiments, and there are many “one-off” events, such as “rain in Newcastle tomorrow”, that we would like to be able to attach probabilities to.

Subjective interpretation

This final common interpretation of probability is somewhat controversial, but does not suffer from the problems that the other interpretations do. It suggests that the association of probabilities to events is a personal (subjective) process, relating to your *degree of belief* in the likelihood of the event occurring. It is controversial because it accepts that *different* people will

assign *different* probabilities to the *same event*. Whilst in some sense it gives up on an objective notion of probability, it is in no sense arbitrary. It can be defined in a precise way, from which the axioms of probability may be derived as requirements of self-consistency.

A simple way to define *your* subjective probability that some event E will occur is as follows. Your probability is the number p such that you consider $\$p$ to be a *fair price* for a gamble which will pay you $\$1$ if E occurs and nothing otherwise.

So, if you consider 40p to be a fair price for a gamble which pays you $\$1$ if it rains in Newcastle tomorrow, then 0.4 is your subjective probability for the event. The subjective interpretation is sometimes known as the *degree of belief interpretation*, and is the interpretation of probability underlying the theory of *Bayesian Statistics* (MAS359, MAS368, MAS451) — a powerful theory of statistical inference named after Thomas Bayes, the 18th Century Presbyterian Minister who first proposed it. Consequently, this interpretation of probability is sometimes also known as the *Bayesian interpretation*.

Summary

Whilst the interpretation of probability is philosophically very important, all interpretations lead to the same set of axioms, from which the rest of probability theory is deduced. Consequently, for this module, it will be sufficient to adopt a fairly classical approach, taking the axioms as given, and investigating their consequences independently of the precise interpretation adopted.

Classical probability

Classical probability theory is concerned with carrying out probability calculations based on *equally likely outcomes*. That is, it is assumed that the sample space has been constructed in such a way that every subset of the sample space consisting of a single element has the same probability. If the sample space contains n possible outcomes ($\#S = n$), we must have for all $s \in S$,

$$\mathrm{P}(\{s\}) = \frac{1}{n}$$

and hence for all $E \subseteq S$

$$\mathrm{P}(E) = \frac{\#E}{n}.$$

More informally, we have

$$\mathrm{P}(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of outcomes}}.$$

Example

Suppose that a fair coin is thrown twice, and the results recorded. The sample space is

$$S = \{HH, HT, TH, TT\}.$$

Let us assume that each outcome is equally likely — that is, each outcome has a probability of $1/4$. Let A denote the event *head on the first toss*, and B denote the event *head on the second toss*. In terms of sets

$$A = \{HH, HT\},\ B = \{HH, TH\}.$$

So

$$\mathrm{P}(A) = \frac{\#A}{n} = \frac{2}{4} = \frac{1}{2}$$

and similarly $\mathrm{P}(B) = 1/2$. If we are interested in the event $C = A \cup B$ we can work out its probability using from the set definition as

$$\mathrm{P}(C) = \frac{\#C}{4} = \frac{\#(A \cup B)}{4} = \frac{\#\{HH, HT, TH\}}{4} = \frac{3}{4}$$

or by using the addition formula

$$\mathrm{P}(C) = \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = \frac{1}{2} + \frac{1}{2} - \mathrm{P}(A \cap B).$$

Now $A \cap B = \{HH\}$, which has probability 1/4, so

$$\mathrm{P}(C) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

In this simple example, it seems easier to work directly with the definition. However, in more complex problems, it is usually much easier to work out how many elements there are in an intersection than in a union, making the addition law very useful.

The multiplication principle

In the above example we saw that there were two distinct experiments — *first throw* and *second throw*. There were two equally likely outcomes for the first

throw and two equally likely outcomes for the second throw. This leads to a combined experiment with $2 \times 2 = 4$ possible outcomes. This is an example of the *multiplication principle*.

Multiplication principle

If there are p experiments and the first has n_1 equally likely outcomes, the second has n_2 equally likely outcomes, and so on until the pth experiment has n_p equally likely outcomes, then there are

$$n_1 \times n_2 \times \cdots n_p = \prod_{i=1}^{p} n_i$$

equally likely possible outcomes for the p experiments.

Example

A class of school children consists of 14 boys and 17 girls. The teacher wishes to pick one boy and one girl to star in the school play. By the multiplication principle, she can do this in $14 \times 17 = 238$ different ways.

Example

A die is thrown twice and the number on each throw is recorded. There are clearly 6 possible outcomes for the first throw and 6 for the second throw. By the multiplication principle, there are 36 possible outcomes for the two throws. If D is the event *a double-six*, then since there is only one possible outcome of the two throws which leads to a double-six, we must have $\mathrm{P}(D) = 1/36$.

Now let E be the event *six on the first throw* and F be the event *six on the second throw*. We know that $\mathrm{P}(E) = \mathrm{P}(F) = 1/6$. If we are interested in the event G, *at least one six*, then $G = E \cup F$, and using the addition law we have

$$\begin{aligned}
\mathrm{P}(G) &= \mathrm{P}(E \cup F) \\
&= \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cap F) \\
&= \frac{1}{6} + \frac{1}{6} - \mathrm{P}(D) \\
&= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} \\
&= \frac{11}{36}.
\end{aligned}$$

This is much easier than trying to count how many of the 36 possible outcomes correspond to G.

Permutations and combinations

Introduction

A repeated experiment often encountered is that of *repeated sampling* from a fixed collection of objects. If we are allowed duplicate objects in our selection, then the procedure is known as *sampling with replacement*, if we are not allowed duplicates, then the procedure is known as *sampling without replacement*.

Probabilists often like to think of repeated sampling in terms of drawing labelled balls from an urn (randomly picking numbered balls from a large rounded vase with a narrow neck). Sometimes the order in which the balls are drawn is important, in which case the set of draws made is referred to as a *permutation*, and sometimes the order does not matter (like the six main balls in the National Lottery), in which case set of draws is referred to as a

combination. We want a way of *counting* the number of possible permutations and combinations so that we can understand the probabilities of different kinds of drawings occurring.

Permutations

Suppose that we have a collection of n objects, $C = \{c_1, c_2, \ldots, c_n\}$. We want to make r selections from C. How many possible *ordered* selections can we make?

If we are sampling *with replacement*, then we have r experiments, and each has n possible (equally likely) outcomes, and so by the multiplication principle, there are

$$n \times n \times \cdots \times n = n^r$$

ways of doing this.

If we are sampling *without replacement*, then we have r experiments. The first experiment has n possible outcomes. The second experiment only has $n-1$

possible outcomes, as one object has already been selected. The third experiment has $n-2$ outcomes and so on until the rth experiment, which has $n-r+1$ possible outcomes. By the multiplication principle, the number of possible selections is

$$n \times (n-1) \times (n-2) \times \cdots \times (n-r+1) = \frac{n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1}{(n-r) \times (n-r-1) \times \cdots \times 3 \times 2 \times 1} = \frac{n!}{(n-r)!}.$$

This is a commonly encountered expression in combinatorics, and has its own notation. The number of ordered ways of selecting r objects from n is denoted P_r^n, where

$$\mathrm{P}_r^n = \frac{n!}{(n-r)!}.$$

We refer to P_r^n as the number of permutations of r out of n objects. If we are interested solely in the number of ways of arranging n objects, then this is clearly just

$$\mathrm{P}_n^n = n!$$

Example

A CD has 12 tracks on it, and these are to be played in random order. There are $12!$ ways of selecting them. There is only one such ordering corresponding to the ordering on the box, so the probability of the tracks being played in the order on the box is $1/12!$ As we will see later, this is considerably smaller than the probability of winning the National Lottery!

Suppose that you have time to listen to only 5 tracks before you go out. There are

$$\mathrm{P}_5^{12} = \frac{12!}{7!} = 12 \times 11 \times 10 \times 9 \times 8 = 95,040$$

ways they could be played. Again, only one of these will correspond to the first 5 tracks on the box (in the correct order), so the probability that the 5 played will be the first 5 on the box is $1/95040$.

Example

In a computer practical session containing 40 students, what is the probability that at least two students share a birthday?

First, let's make some simplifying assumptions. We will assume that there are 365 days in a year and that each day is equally likely to be a birthday.

Call the event we are interested in A. We will first calculate the probability of A^c, the probability that *no two* people have the same birthday, and calculate the probability we want using $\mathrm{P}(A) = 1 - \mathrm{P}(A^c)$. The number of ways 40 birthdays could occur is like sampling 40 objects from 365 *with* replacement, which is just 365^{40}. The number of ways we can have 40 *distinct* birthdays is like sampling 40 objects from 365 *without* replacement, P_{40}^{365}. So, the probability of all birthdays being distinct is

$$\mathrm{P}(A^c) = \frac{\mathrm{P}_{40}^{365}}{365^{40}} = \frac{365!}{325!365^{40}} \approx 0.1$$

and so

$$\mathrm{P}(A) = 1 - \mathrm{P}(A^c) \approx 0.9.$$

That is, there is a probability of 0.9 that we have a match. In fact, the fact that birthdays are *not* distributed uniformly over the year makes the probability of a match even higher!

Unless you have a very fancy calculator, you may have to expand the expression a bit, and give it to your calculator in manageable chunks. On the other hand, Maple loves expressions like this.

```
> 1-365!/(325!*365^40);
> evalf(");
```

will give the correct answer. However, Maple also knows about combinatoric functions. The following gives the same answer:

```
> with(combinat);
> 1-numbperm(365,40)/(365^40);
> evalf(");
```

Similarly, the probability that there will be a birthday match in a group of n people is

$$1 - \frac{\mathrm{P}_n^{365}}{365^n}.$$

We can define this as a Maple function, and evaluate it for some different values of n as follows.

```
> with(combinat);
> p := n -> evalf(1 - numbperm(365,n)/(365^n));
> p(10);
> p(20);
> p(22);
> p(23);
> p(40);
> p(50);
```

The probability of a match goes above 0.5 for $n = 23$. That is, you only need a group of 23 people in order to have a better than evens chance of a match. This is a somewhat counter-intuitive result, and the reason is that people think more intuitively about the probability that someone has the same birthday as *themselves*. This is an entirely different problem.

Suppose that you are one of a group of 40 students. What is the probability of B, where B is the event that at least one other person in the group has the same birthday as you?

Again, we will work out $\mathrm{P}(B^c)$ first, the probability that no-one has your birthday. Now, there are 365^{39} ways that the birthdays of the the other people can occur, and we allow each of them to have any birthday other than yours, so there are 364^{39} ways for this to occur. Hence we have

$$\mathrm{P}(B^c) = \frac{364^{39}}{365^{39}} \approx 0.9$$

and so

$$\mathrm{P}(B) = 1 - \frac{364^{39}}{365^{39}} \approx 0.1.$$

Here the probabilities are reversed — there is only a 10% chance that someone has the same birthday as you. Most people find this much more intuitively reasonable. So, how big a group of people would you need in order to have a better than evens chance of someone having the same birthday as you? The general formula for the probability of a match with n people is

$$\mathrm{P}(B) = 1 - \frac{364^{n-1}}{365^{n-1}} = 1 - \left(\frac{364}{365}\right)^{n-1},$$

and as long as you enter it into your calculator the way it is written on the right, it will be fine. We find that a group of size 254 is needed for the

probability to be greater than 0.5, and that a group of 800 or more is needed before you can be really confident that someone will have the same birthday as you. For a group of size 150 (the size of the lectures), the probability of a match is about 1/3.

This problem illustrates quite nicely the subtlety of probability questions, the need to define precisely the events you are interested in, and the fact that some probability questions have counter-intuitive answers.

Combinations

We now have a way of counting permutations, but often when selecting objects, all that matters is *which* objects were selected, not the order in which they were selected. Suppose that we have a collection of objects, $C = \{c_1, \ldots, c_n\}$ and that we wish to make r selections from this list of objects, *without replacement*, where the order does not matter. An unordered selection such as this is referred to as a *combination*. How many ways can this be done? Notice that this is equivalent to asking how many different subsets of C of size r there are.

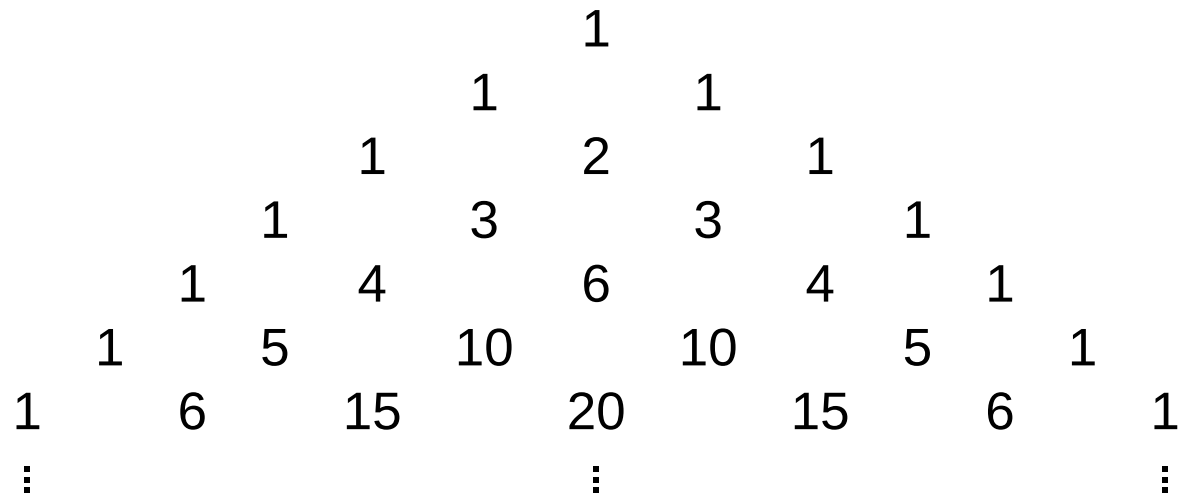
From the multiplication principle, we know that the number of *ordered* samples must be the number of *unordered* samples, multiplied by the number of orderings of each sample. So, the number of unordered samples is the number of ordered samples, divided by the number of orderings of each sample. That is, the number of unordered samples is

$$\begin{aligned}\frac{\text{number of ordered samples of size } r}{\text{number of orderings of samples of size } r} &= \frac{\mathrm{P}_r^n}{\mathrm{P}_r^r} \\ &= \frac{\mathrm{P}_r^n}{r!} \\ &= \frac{n!}{r!(n-r)!}\end{aligned}$$

Again, this is a very commonly found expression in combinatorics, so it has its own notation. In fact, there are two commonly used expressions for this quantity:

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

These numbers are known as the *binomial coefficients*. We will use the notation $\binom{n}{r}$ as this is slightly neater, and more commonly used. They can be found as the $(r+1)$th number on the $(n+1)$th row of *Pascal's triangle*:



Example

Returning to the CD with 12 tracks. You arrange for your CD player to play 5 tracks at random. How many different unordered selections of 5 tracks are there, and what is the probability that the 5 tracks played are your 5 favourite tracks (in any order)?

The number of ways of choosing 5 tracks from 12 is just $\binom{12}{5} = 792$. Since only one of these will correspond to your favourite five, the probability of getting your favourite five is $1/792 \approx 0.001$.

Example (National Lottery)

What is the probability of winning exactly $10 on the National Lottery?

In the UK National Lottery, there are 49 numbered balls, and six of these are selected at random. A seventh ball is also selected, but this is only relevant if you get exactly five numbers correct. The player selects six numbers before the draw is made, and after the draw, counts how many numbers are in common with those drawn. If the player has selected exactly three of the balls drawn, then the player wins $10. The order the balls are drawn in is irrelevant.

We are interested in the probability that exactly 3 of the 6 numbers we select are drawn. First we need to count the number of possible draws (the number of different sets of 6 numbers), and then how many of those draws correspond to getting exactly three numbers correct. The number of possible draws is the number of ways of choosing 6 objects from 49. This is

$$\binom{49}{6} = 13,983,816.$$

The number of drawings corresponding to getting exactly three right is calculated as follows. Regard the six numbers you have chosen as your “good” numbers. Then of the 49 balls to be drawn from, 6 correspond to your

“good” numbers, and 43 correspond to your “bad” numbers. We want to know how many ways there are of selecting 3 “good” numbers and 3 “bad” numbers. By the multiplication principle, this is the number of ways of choosing 3 from 6, multiplied by the number of ways of choosing 3 from 43. That is, there are

$$\binom{6}{3}\binom{43}{3} = 246,820$$

ways of choosing exactly 3 “good” numbers. So, the probability of getting exactly 3 numbers, and winning $10 is

$$\frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}} \approx 0.0177 \approx \frac{1}{57}.$$

Conditional probability and the multiplication rule

Conditional probability

We now have a way of understanding the probabilities of events, but so far we have no way of *modifying* those probabilities when certain events occur. For this, we need an extra axiom which can be justified under any of the interpretations of probability. The axiom defines the *conditional probability of* A *given* B, written $\mathrm{P}(A|B)$ as

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \quad \text{for } \mathrm{P}(B) > 0.$$

Note that we can only condition on events with positive probability.

Under the classical interpretation of probability, we can see that if we are told that B has occurred, then all outcomes in B are equally likely, and all outcomes not in B have zero probability — so B is the new sample space. The number of ways that A can occur is now just the number of ways $A \cap B$ can occur, and these are all equally likely. Consequently we have

$$\mathrm{P}(A|B) = \frac{\#(A \cap B)}{\#B} = \frac{\#(A \cap B)/\#S}{\#B/\#S} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Because conditional probabilities really just correspond to a new probability measure defined on a smaller sample space, they obey all of the properties of

“ordinary” probabilities. For example, we have

$$\begin{aligned} \mathrm{P}(B|B) &= 1 \\ \mathrm{P}(\emptyset|B) &= 0 \\ \mathrm{P}(A\cup C|B) &= \mathrm{P}(A|B) + \mathrm{P}(C|B), \quad \text{for } A\cap C = \emptyset \end{aligned}$$

and so on.

The definition of conditional probability simplifies when one event is a special case of the other. If $A \subseteq B$, then $A \cap B = A$ so

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)}{\mathrm{P}(B)}, \quad \text{for } A \subseteq B.$$

Example

A die is rolled and the number showing recorded. Given that the number rolled was even, what is the probability that it was a six?

Let E denote the event “even” and F denote the event “a six”. Clearly $F \subseteq E$, so

$$\mathrm{P}(F|E) = \frac{\mathrm{P}(F)}{\mathrm{P}(E)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

The multiplication rule

The formula for conditional probability is useful when we want to calculate $\mathrm{P}(A|B)$ from $\mathrm{P}(A \cap B)$ and $\mathrm{P}(B)$. However, more commonly we want to know $\mathrm{P}(A \cap B)$ and we know $\mathrm{P}(A|B)$ and $\mathrm{P}(B)$. A simple rearrangement gives us the multiplication rule.

$$\mathrm{P}(A \cap B) = \mathrm{P}(B) \times \mathrm{P}(A|B)$$

Example

Two cards are dealt from a deck of 52 cards. What is the probability that they are both Aces?

We now have three different ways of computing this probability. First, let's use conditional probability. Let A_1 be the event "first card an Ace" and A_2 be the event "second card an Ace". $\mathrm{P}(A_2|A_1)$ is the probability of a second Ace.

Given that the first card has been drawn and was an Ace, there are 51 cards left, 3 of which are Aces, so $\mathrm{P}(A_2|A_1) = 3/51$. So,

$$\begin{aligned}\mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_1) \times \mathrm{P}(A_2|A_1) \\ &= \frac{4}{52} \times \frac{3}{51} \\ &= \frac{1}{221}.\end{aligned}$$

Now let's compute it by counting ordered possibilities. There are P_2^{52} ways of choosing 2 cards from 52, and P_2^4 of those ways correspond to choosing 2 Aces from 4, so

$$\mathrm{P}(A_1 \cap A_2) = \frac{\mathrm{P}_2^4}{\mathrm{P}_2^{52}} = \frac{12}{2652} = \frac{1}{221}.$$

Now let's compute it by counting unordered possibilities. There are $\binom{52}{2}$ ways of choosing 2 cards from 52, and $\binom{4}{2}$ of those ways correspond to choosing 2 Aces from 4, so

$$\mathrm{P}(A_1 \cap A_2) = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{6}{1326} = \frac{1}{221}.$$

If possible, you should always try and calculate probabilities more than one way (as it is very easy to go wrong!). However, for counting problems where the order doesn't matter, counting the unordered possibilities using combinations will often be the only reasonable way, and for problems which don't correspond to a sampling experiment, using conditional probability will often be the only reasonable way.

The multiplication rule generalises to more than two events. For example, for three events we have

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\ \mathrm{P}(A_2|A_1)\ \mathrm{P}(A_3|A_1 \cap A_2)\,.$$

Independent events, partitions and Bayes Theorem

Independence

Recall the multiplication rule

$$\mathrm{P}(A \cap B) = \mathrm{P}(B)\ \mathrm{P}(A|B)\,.$$

For some events A and B, knowing that B has occurred will not alter the probability of A, so that $\mathrm{P}(A|B) = \mathrm{P}(A)$. When this is so, the multiplication rule becomes

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\ \mathrm{P}(B)\,,$$

and the events A and B are said to be *independent events*. Independence is a very important concept in probability theory, and is used a lot to build up complex events from simple ones. Do not confuse the independence of A and B with the exclusivity of A and B — they are entirely different concepts. If A and B both have positive probability, then they cannot be both independent and exclusive (exercise).

When it is clear that the occurrence of B can have no influence on A, we will *assume* independence in order to calculate $\mathrm{P}(A \cap B)$. However, if we can calculate $\mathrm{P}(A \cap B)$ directly, we can check the independence of A and B by seeing if it is true that

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\ \mathrm{P}(B)\,.$$

We can generalise independence to collections of events as follows. The set of events $A =$

$\{A_1, A_2, \ldots, A_n\}$ are *mutually independent events* if *for any subset*, $B \subseteq A$, $B = \{B_1, B_2, \ldots, B_r\}$, $r \leq n$ we have

$$\mathrm{P}(B_1 \cap \cdots \cap B_r) = \mathrm{P}(B_1) \times \cdots \times \mathrm{P}(B_r).$$

Note that mutual independence is much stronger that *pair-wise* independence, where we only require independence of subsets of size 2. That is, pair-wise independence *does not* imply mutual independence.

Example

A playing card is drawn from a pack. Let A be the event "an Ace is drawn" and let C be the event "a Club is drawn". Are the events A and C exclusive? Are they independent?

A and C are clearly not exclusive, since they can both happen — when the Ace of Clubs is drawn. Indeed, since this is the only way it can happen, we

know that $\mathrm{P}(A \cap C) = 1/52$. We also know that $\mathrm{P}(A) = 1/13$ and that $\mathrm{P}(C) = 1/4$. Now since

$$\begin{aligned} \mathrm{P}(A)\ \mathrm{P}(C) &= \frac{1}{13} \times \frac{1}{4} \\ &= \frac{1}{52} \\ &= \mathrm{P}(A \cap C) \end{aligned}$$

we know that A and C are independent. Of course, this is intuitively obvious — you are no more or less likely to think you have an Ace if someone tells you that you have a Club.

Partitions

A *partition* of a sample space is simply the decomposition of the sample space into a collection of mutually *exclusive* events with positive probability. That is, $\{B_1, \ldots, B_n\}$ form a *partition* of S if

- $S = B_1 \cup B_2 \cup \cdots \cup B_n = \bigcup_{i=1}^{n} B_i$,

- $B_i \cap B_j = \emptyset,\ \forall i \neq j,$

- $\mathrm{P}(B_i) > 0,\ \forall i.$

Example

A card is randomly drawn from the pack. The events $\{C, D, H, S\}$ (Club, Diamond, Heart, Spade) form a partition of the sample space, since one and only one will occur, and all can occur.

Theorem of total probability

Suppose that we have a partition $\{B_1, \ldots, B_n\}$ of a sample space, S. Suppose further that we have an event A. Then A can be written as the disjoint union

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_n),$$

and so the probability of A is given by

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}((A \cap B_1) \cup \cdots \cup (A \cap B_n)) \\
&= \mathrm{P}(A \cap B_1) + \cdots + \mathrm{P}(A \cap B_n), && \text{by Axiom III} \\
&= \mathrm{P}(A|B_1)\ \mathrm{P}(B_1) + \cdots + \mathrm{P}(A|B_n)\ \mathrm{P}(B_n), && \text{by the multiplication rule} \\
&= \sum_{i=1}^{n} \mathrm{P}(A|B_i)\ \mathrm{P}(B_i).
\end{aligned}$$

Example ("Craps")

Craps is a game played with a pair of dice. A player plays against a banker. The player throws the dice and notes the sum.

- If the sum is 7 or 11, the player wins, and the game ends (a *natural*).

- If the sum is 2, 3 or 12, the player looses and the game ends (a *crap*).

- If the sum is anything else, the sum is called the players *point*, and the player keeps throwing the dice until his sum is 7, in which case he looses, or he throws his *point* again, in which case he wins.

What is the probability that the player wins?

Bayes Theorem

From the multiplication rule, we know that

$$\mathrm{P}(A \cap B) = \mathrm{P}(B)\ \mathrm{P}(A|B)$$

and that

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\ \mathrm{P}(B|A)\,,$$

so clearly

$$\mathrm{P}(B)\ \mathrm{P}(A|B) = \mathrm{P}(A)\ \mathrm{P}(B|A)\,,$$

and so

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\ \mathrm{P}(A)}{\mathrm{P}(B)}.$$

This is known as *Bayes Theorem*, and is a very important result in probability, as it tells us how to "turn conditional probabilities around" — that is, it tells us how to work out $\mathrm{P}(A|B)$ from $\mathrm{P}(B|A)$, and this is often very useful.

Example

A clinic offers you a free test for a very rare, but hideous disease. The test they offer is very reliable. If you have the disease it has a 98% chance of giving a positive result, and if you don't have the disease, it has only a 1% chance of giving a positive result. You decide to take the test, and find that you test positive — what is the probability that you have the disease?

Let P be the event "test positive" and D be the event "you have the disease". We know that

$$\mathrm{P}(P|D) = 0.98 \text{ and that } \mathrm{P}(P|D^C) = 0.01.$$

We want to know $\mathrm{P}(D|P)$, so we use Bayes' Theorem.

$$\begin{aligned}\mathrm{P}(D|P) &= \frac{\mathrm{P}(P|D)\ \mathrm{P}(D)}{\mathrm{P}(P)} \\ &= \frac{\mathrm{P}(P|D)\ \mathrm{P}(D)}{\mathrm{P}(P|D)\ \mathrm{P}(D) + \mathrm{P}(P|D^c)\ \mathrm{P}(D^c)} \quad \text{(using the theorem of total probability)} \\ &= \frac{0.98\,\mathrm{P}(D)}{0.98\,\mathrm{P}(D) + 0.01(1 - \mathrm{P}(D))}.\end{aligned}$$

So we see that the probability you have the disease given the test result depends on the probability that you had the disease in the first place. This is a rare disease, affecting only one in ten thousand people, so that $\mathrm{P}(D) = 0.0001$. Substituting this in gives

$$\mathrm{P}(D|P) = \frac{0.98 \times 0.0001}{0.98 \times 0.0001 + 0.01 \times 0.9999} \simeq 0.01.$$

So, your probability of having the disease has increased from 1 in 10,000 to 1 in 100, but still isn't that much to get worried about! Note the *crucial* difference between $\mathrm{P}(P|D)$ and $\mathrm{P}(D|P)$.

Bayes Theorem for partitions

Another important thing to notice about the above example is the use of the theorem of total probability in order to expand the bottom line of Bayes Theorem. In fact, this is done so often that Bayes Theorem is often stated in this form.

Suppose that we have a partition $\{B_1, \ldots, B_n\}$ of a sample space S. Suppose further that we have an event A, with $\mathrm{P}(A) > 0$. Then, for each B_j, the probability of B_j given A is

$$\begin{aligned}
\mathrm{P}\big(B_j|A\big) &= \frac{\mathrm{P}\big(A|B_j\big)\ \mathrm{P}\big(B_j\big)}{\mathrm{P}(A)} \\
&= \frac{\mathrm{P}\big(A|B_j\big)\ \mathrm{P}\big(B_j\big)}{\mathrm{P}(A|B_1)\ \mathrm{P}(B_1) + \cdots + \mathrm{P}(A|B_n)\ \mathrm{P}(B_n)} \\
&= \frac{\mathrm{P}\big(A|B_j\big)\ \mathrm{P}\big(B_j\big)}{\sum_{i=1}^{n} \mathrm{P}(A|B_i)\ \mathrm{P}(B_i)}.
\end{aligned}$$

In particular, if the partition is simply $\{B, B^c\}$, then this simplifies to

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\ \mathrm{P}(B)}{\mathrm{P}(A|B)\ \mathrm{P}(B) + \mathrm{P}(A|B^c)\ \mathrm{P}(B^c)}.$$

Discrete Probability Models

Introduction, mass functions and distribution functions

Introduction

We now have a good understanding of basic probabilistic reasoning. We have seen how to relate events to sets, and how to calculate probabilities for events by working with the sets that represent them. So far, however, we haven't developed any special techniques for thinking about *random quantities*. *Discrete probability models* provide a framework for thinking about *discrete random quantities*, and *continuous probability models* (to be considered in the next chapter) form a framework for thinking about *continuous* random quantities.

Example

Consider the sample space for tossing a fair coin twice:

$$S = \{HH, HT, TH, TT\}.$$

These outcomes are equally likely. There are several random quantities we could associate with this experiment. For example, we could count the number of heads, or the number of tails.

Formally, a *random quantity* is a real valued function which acts on *elements* of the sample space (outcomes). That is, to each outcome, the random variable assigns a real number. Random quantities (sometimes known as *random variables*) are always denoted by upper case letters.

In our example, if we let X be the number of heads, we have

$$\begin{aligned} X(HH) &= 2, \\ X(HT) &= 1, \\ X(TH) &= 1, \\ X(TT) &= 0. \end{aligned}$$

The observed value of a random quantity is the number corresponding to the actual outcome. That is, if the outcome of an experiment is $s \in S$, then $X(s) \in \mathbf{R}$ is the observed value. This observed value is always denoted with a lower case letter — here x. Thus $X = x$ means that the observed value of the

random quantity, X is the number x. The set of possible observed values for X is

$$S_X = \{X(s) | s \in S\}.$$

For the above example we have

$$S_X = \{0, 1, 2\}.$$

Clearly here the values are not all equally likely.

Example

Roll one die and call the random number which is uppermost Y. The sample space for the *random quantity* Y is

$$S_Y = \{1, 2, 3, 4, 5, 6\}$$

and these outcomes are all equally likely. Now roll two dice and call their sum Z. The sample space for Z is

$$S_Z = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and these outcomes are *not* equally likely. However, we know the probabilities of the events corresponding to each of these outcomes, and we could display them in a table as follows.

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

This is essentially a tabulation of the *probability mass function* for the random quantity Z.

Probability mass functions (PMFs)

For any discrete random variable X, we define the *probability mass function* (PMF) to be the function which gives the probability of each $x \in S_X$. Clearly we have

$$\mathrm{P}(X = x) = \sum_{\{s \in S | X(s) = x\}} \mathrm{P}(\{s\}).$$

That is, the probability of getting a particular number is the sum of the probabilities of all those outcomes which have that number associated with

them. Also $\mathrm{P}(X = x) \geq 0$ for each $x \in S_X$, and $\mathrm{P}(X = x) = 0$ otherwise. The set of all pairs $\{(x, \mathrm{P}(X = x)) | x \in S_X\}$ is known as the *probability distribution* of X.

Example

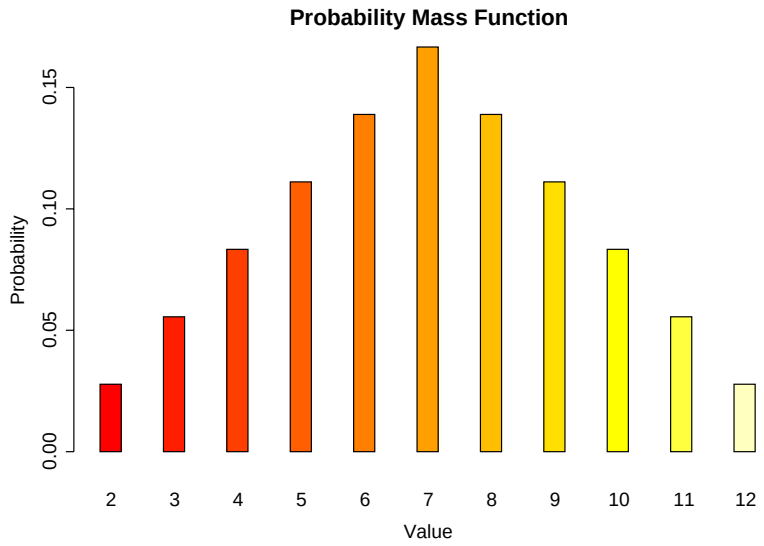
For the example above concerning the sum of two dice, the probability distribution is

$$\{(2, 1/36), (3, 2/36), (4, 3/36), (5, 4/36), (6, 5/36), (7, 6/36), \\ (8, 5/36), (9, 4/36), (10, 3/36), (11, 2/36), (12, 1/36)\}$$

and the probability mass function can be tabulated as

x	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and plotted graphically as follows.



Cumulative distribution functions (CDFs)

For any discrete random quantity, X, we clearly have

$$\sum_{x \in S_X} \mathrm{P}(X = x) = 1$$

as every outcome has some number associated with it. It can often be useful to know the probability that your random number is no greater than some particular value. With that in mind, we define the *cumulative distribution function*,

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{\{y \in S_X | y \leq x\}} \mathrm{P}(X = y)\,.$$

Example

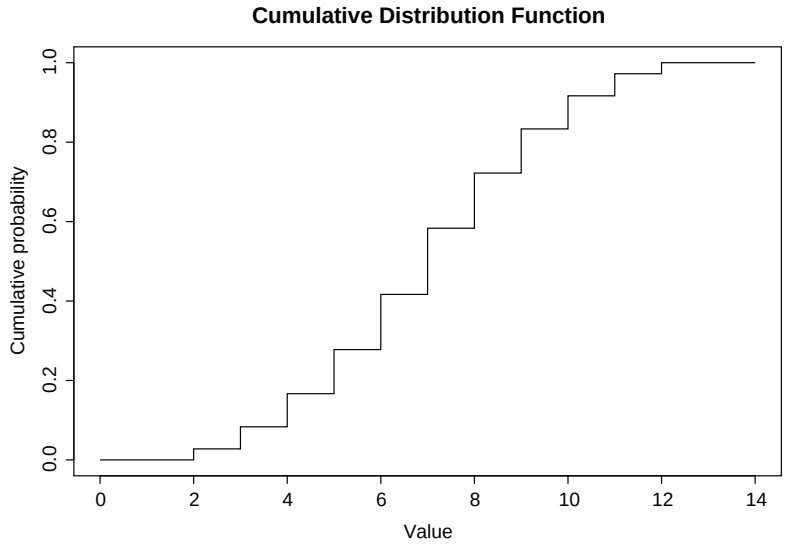
For the sum of two dice, the CDF can be tabulated for the outcomes as

x	2	3	4	5	6	7	8	9	10	11
$F_X(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36

but it is important to note that the CDF is defined *for all real numbers* — not just the possible values. In our example we have

$$\begin{aligned} F_X(-3) &= \mathrm{P}(X \leq -3) = 0, \\ F_X(4.5) &= \mathrm{P}(X \leq 4.5) = \mathrm{P}(X \leq 4) = 6/36, \\ F_X(25) &= \mathrm{P}(X \leq 25) = 1. \end{aligned}$$

We may plot the CDF for our example as follows.



It is clear that for any random variable X, for all $x \in \mathbf{R}$, $F_X(x) \in [0,1]$ and that $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to +\infty$.

Expectation and variance for discrete random quantities

Expectation

Just as it is useful to summarise data (Chapter 1), it is just as useful to be able to summarise the distribution of random quantities. The *location*

measure used to summarise random quantities is known as the *expectation* of the random quantity. It is the "centre of mass" of the probability distribution. The expectation of a discrete random quantity X, written $\mathrm{E}(X)$ is defined by

$$\mathrm{E}(X) = \sum_{x \in S_X} x \, \mathrm{P}(X = x) \, .$$

The expectation is often denoted by μ_X or even just μ. Note that the expectation is a known function of the probability distribution. It is *not* a random quantity, and in particular, it is *not* the sample mean of a set of data (random or otherwise). In fact, there is a *relationship* between the sample mean of a set of data and the expectation of the underlying probability distribution generating the data, but this is to be made precise in Semester 2.

Example

For the sum of two dice, X, we have

$$\mathrm{E}(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \cdots + 12 \times \frac{1}{36} = 7.$$

By looking at the symmetry of the mass function, it is clear that in some sense 7 is the "central" value of the probability distribution.

Variance

We now have a method for summarising the location of a given probability distribution, but we also need a summary for the *spread*. For a discrete random quantity X, the *variance* of X is defined by

$$\mathrm{Var}\,(X) = \sum_{x \in S_X} \left\{ (x - \mathrm{E}(X))^2 \, \mathrm{P}(X = x) \right\}.$$

The variance is often denoted σ_X^2, or even just σ^2. Again, this is a known function of the probability distribution. It is not random, and it is not the *sample* variance of a set of data. Again, the two are related in a way to be made precise later. The variance can be re-written as

$$\mathrm{Var}\,(X) = \sum_{x_i \in S_X} x_i^2 \, \mathrm{P}(X = x_i) - [\,\mathrm{E}(X)]^2,$$

and this expression is usually a bit easier to work with. We also define the *standard deviation* of a random quantity by

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}\,(X)},$$

and this is usually denoted by σ_X or just σ.

Example

For the sum of two dice, X, we have

$$\sum_{x_i \in S_X} x_i^2 \, \mathrm{P}(X = x_i) = 2^2 \times \frac{1}{36} + 3^2 \times \frac{2}{36} + 4^2 \times \frac{3}{36} + \cdots + 12^2 \times \frac{1}{36} = \frac{329}{6}$$

and so

$$\mathrm{Var}\,(X) = \frac{329}{6} - 7^2 = \frac{35}{6},$$

and

$$\mathrm{SD}(X) = \sqrt{\frac{35}{6}}.$$

Properties of expectation and variance

One of the reasons that expectation is widely used as a measure of location for probability distributions is the fact that it has many desirable mathematical properties which make it elegant and convenient to work with. Indeed, many

of the nice properties of expectation lead to corresponding nice properties for variance, which is one of the reasons why variance is widely used as a measure of spread.

Expectation of a function of a random quantity

Suppose that X is a discrete random quantity, and that Y is another random quantity that is a known function of X. That is, $Y = g(X)$ for some function $g(\cdot)$. What is the expectation of Y?

Example

Throw a die, and let X be the number showing. We have

$$S_X = \{1, 2, 3, 4, 5, 6\}$$

and each value is equally likely. Now suppose that we are actually interested in the square of the number showing. Define a new random quantity $Y = X^2$. Then

$$S_Y = \{1, 4, 9, 16, 25, 36\}$$

and clearly each of these values is equally likely. We therefore have

$$\mathrm{E}(Y) = 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + \cdots + 36 \times \frac{1}{6} = \frac{91}{6}.$$

The above example illustrates the more general result, that for $Y = g(X)$, we have

$$\mathrm{E}(Y) = \sum_{x \in S_X} g(x)\, \mathrm{P}(X = x)\,.$$

Note that in general $\mathrm{E}(g(X)) \neq g(\mathrm{E}(X))$. For the above example, $\mathrm{E}\left(X^2\right) = 91/6 \simeq 15.2$, and $\mathrm{E}(X)^2 = 3.5^2 = 12.25$.

We can use this more general notion of expectation in order to redefine variance purely in terms of expectation as follows:

$$\mathrm{Var}\,(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^2\right) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2\,.$$

Having said that $\mathrm{E}(g(X)) \neq g(\mathrm{E}(X))$ in general, it does in fact hold in the (very) special, but important case where $g(\cdot)$ is a *linear* function.

Expectation of a linear transformation

If we have a random quantity X, and a linear transformation, $Y = aX + b$, where a and b are known real constants, then we have that

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b.$$

We can show this as follows:

$$\begin{aligned}
\mathrm{E}(aX + b) &= \sum_{x \in S_X} (ax + b)\,\mathrm{P}(X = x) \\
&= \sum_{x \in S_X} ax\,\mathrm{P}(X = x) + \sum_{x \in S_X} b\,\mathrm{P}(X = x) \\
&= a \sum_{x \in S_X} x\,\mathrm{P}(X = x) + b \sum_{x \in S_X} \mathrm{P}(X = x) \\
&= a\,\mathrm{E}(X) + b.
\end{aligned}$$

Expectation of the sum of two random quantities

For two random quantities X and Y, the expectation of their sum is given by

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Note that this result is true irrespective of whether or not X and Y are independent. Let us see why. First,

$$S_{X+Y} = \{x+y|(x \in S_X) \cap (y \in S_Y)\},$$

and so

$$
\begin{aligned}
\mathrm{E}(X+Y) &= \sum_{(x+y)\in S_{X+Y}} (x+y)\,\mathrm{P}((X=x)\cap(Y=y)) \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} (x+y)\,\mathrm{P}((X=x)\cap(Y=y)) \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} x\,\mathrm{P}((X=x)\cap(Y=y)) \\
&\quad + \sum_{x\in S_X}\sum_{y\in S_Y} y\,\mathrm{P}((X=x)\cap(Y=y)) \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} x\,\mathrm{P}(X=x)\;\mathrm{P}(Y=y|X=x) \\
&\quad + \sum_{y\in S_Y}\sum_{x\in S_X} y\,\mathrm{P}(Y=y)\;\mathrm{P}(X=x|Y=y) \\
&= \sum_{x\in S_X} x\,\mathrm{P}(X=x)\sum_{y\in S_Y}\mathrm{P}(Y=y|X=x) \\
&\quad + \sum_{y\in S_Y} y\,\mathrm{P}(Y=y)\sum_{x\in S_X}\mathrm{P}(X=x|Y=y) \\
&= \sum_{x\in S_X} x\,\mathrm{P}(X=x) + \sum_{y\in S_Y} y\,\mathrm{P}(Y=y) \\
&= \mathrm{E}(X) + \mathrm{E}(Y)\,.
\end{aligned}
$$

Expectation of an independent product

If X and Y are *independent* random quantities, then

$$\mathrm{E}(XY) = \mathrm{E}(X)\ \mathrm{E}(Y).$$

To see why, note that

$$S_{XY} = \{xy | (x \in S_X) \cap (y \in S_Y)\},$$

and so

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_{xy \in S_{XY}} xy\, \mathrm{P}((X = x) \cap (Y = y)) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} xy\, \mathrm{P}(X = x)\ \mathrm{P}(Y = y) \\
&= \sum_{x \in S_X} x\, \mathrm{P}(X = x) \sum_{y \in S_Y} y\, \mathrm{P}(Y = y) \\
&= \mathrm{E}(X)\ \mathrm{E}(Y)
\end{aligned}$$

Note that here it is *vital* that X and Y are independent, or the result does not hold.

Variance of an independent sum

If X and Y are *independent* random quantities, then

$$\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)\,.$$

To see this, write

$$\begin{aligned}
\mathrm{Var}\,(X+Y) &= \mathrm{E}\left([X+Y]^2\right) - [\,\mathrm{E}(X+Y)]^2 \\
&= \mathrm{E}\left(X^2 + 2XY + Y^2\right) - [\,\mathrm{E}(X) + \mathrm{E}(Y)]^2 \\
&= \mathrm{E}\left(X^2\right) + 2\,\mathrm{E}(XY) + \mathrm{E}\left(Y^2\right) - \mathrm{E}(X)^2 - 2\,\mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(Y)^2 \\
&= \mathrm{E}\left(X^2\right) + 2\,\mathrm{E}(X)\,\mathrm{E}(Y) + \mathrm{E}\left(Y^2\right) - \mathrm{E}(X)^2 - 2\,\mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(Y)^2 \\
&= \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 + \mathrm{E}\left(Y^2\right) - \mathrm{E}(Y)^2 \\
&= \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)
\end{aligned}$$

Again, it is vital that X and Y are independent, or the result does not hold. Notice that this implies a slightly less attractive result for the standard deviation of the sum of two independent random quantities,

$$\mathrm{SD}(X+Y) = \sqrt{\mathrm{SD}(X)^2 + \mathrm{SD}(Y)^2},$$

which is why it is often more convenient to work with variances.

The binomial distribution

Introduction

Now that we have a good understanding of discrete random quantities and their properties, we can go on to look at a few of the standard families of discrete random variables. One of the most commonly encountered discrete distributions is the binomial distribution. This is the distribution of the number of “successes” in a series of independent “success”/“fail” trials. Before we look at this, we need to make sure we understand the case of a single trial.

Bernoulli random quantities

Suppose that we have an event E in which we are interested, and we write its sample space as

$$S = \{E, E^C\}.$$

We can associate a random quantity with this sample space, traditionally denoted I, as $I(E) = 1$, $I(E^C) = 0$. So, if $\mathrm{P}(E) = p$, we have

$$S_I = \{0, 1\},$$

and $\mathrm{P}(I=1)=p,\ \mathrm{P}(I=0)=1-p$. This random quantity, I is known as an *indicator variable*, and is often useful for constructing more complex random quantities. We write

$$I \sim Bern(p).$$

We can calculate its expectation and variance as follows.

$$\begin{aligned}
\mathrm{E}(I) &= 0\times(1-p)+1\times p = p \\
\mathrm{E}\left(I^2\right) &= 0^2\times(1-p)+1^2\times p = p \\
\mathrm{Var}\,(I) &= \mathrm{E}\left(I^2\right) - \mathrm{E}(I)^2 \\
&= p-p^2 = p(1-p)
\end{aligned}$$

With these results, we can now go on to understand the binomial distribution.

The binomial distribution

The binomial distribution is the distribution of the number of “successes” in a series of n independent “trials”, each of which results in a “success” (with

probability p) or a “failure” (with probability $1-p$). If the number of successes is X, we would write

$$X \sim B(n,p)$$

to indicate that X is a binomial random quantity based on n independent trials, each occurring with probability p.

Examples

1. Toss a fair coin 100 times and let X be the number of heads. Then $X \sim B(100, 0.5)$.

2. A certain kind of lizard lays 8 eggs, each of which will hatch independently with probability 0.7. Let Y denote the number of eggs which hatch. Then $Y \sim B(8, 0.7)$.

Let us now derive the probability mass function for $X \sim B(n,p)$. Clearly X can take on any value from 0 up to n, and no other. Therefore, we simply have to

calculate $\mathrm{P}(X = k)$ for $k = 0, 1, 2, \ldots, n$. The probability of k successes followed by $n - k$ failures is clearly $p^k(1-p)^{n-k}$. Indeed, this is the probability of *any* particular sequence involving k successes. There are $\binom{n}{k}$ such sequences, so by the multiplication principle, we have

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

Now, using the binomial theorem, we have

$$\sum_{k=0}^{n} \mathrm{P}(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + [1-p])^n = 1^n = 1,$$

and so this does define a valid probability distribution.

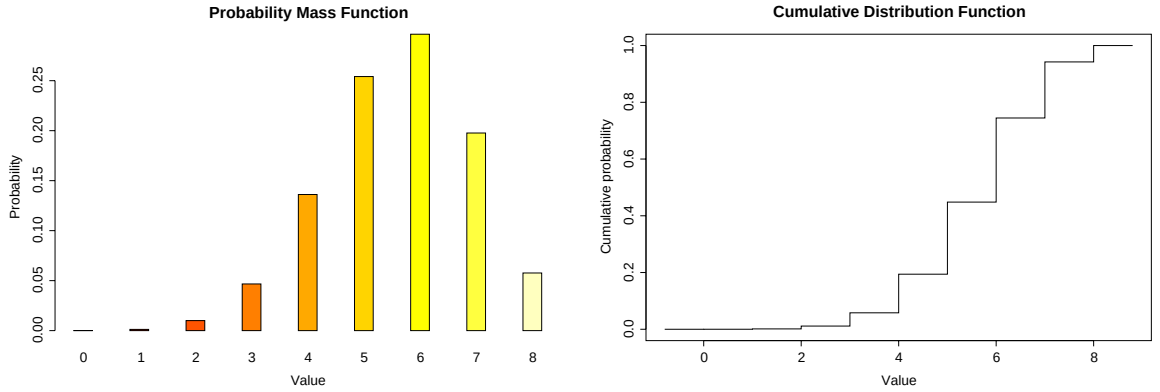
Examples

For the lizard eggs, $Y \sim B(8, 0.7)$ we have

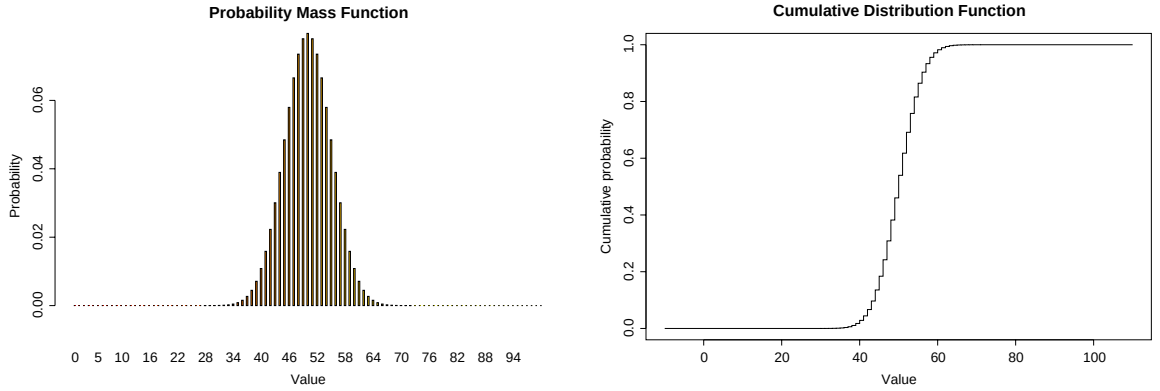
$$\mathrm{P}(Y = k) = \binom{8}{k} 0.7^k 0.3^{8-k}, \; k = 0, 1, 2, \ldots, 8.$$

We can therefore tabulate and plot the probability mass function and cumulative distribution function as follows.

k	0	1	2	3	4	5	6	7	8
$\mathrm{P}(Y=k)$	0.00	0.00	0.01	0.05	0.14	0.25	0.30	0.20	0.06
$F_Y(k)$	0.00	0.00	0.01	0.06	0.19	0.45	0.74	0.94	1.00



Similarly, the PMF and CDF for $X \sim B(100, 0.5)$ (number of heads from 100 coin tosses) can be plotted as follows.



Expectation and variance of a binomial random quantity

It is possible (but a little messy) to derive the expectation and variance of the binomial distribution directly from the PMF. However, we can deduce them rather more elegantly if we recognise the relationship between the binomial and Bernoulli distributions. If $X \sim B(n, p)$ then

$$X = \sum_{j=1}^{n} I_j$$

where $I_j \sim Bern(p),\ j = 1, 2, \ldots, n$, and the I_j are mutually independent. So we then have

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}\left(\sum_{j=1}^{n} I_j \right) \\ &= \sum_{j=1}^{n} \mathrm{E}(I_j) && \text{(expectation of a sum)} \\ &= \sum_{j=1}^{n} p \\ &= np \end{aligned}$$

and similarly,

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{Var}\left(\sum_{j=1}^{n} I_j\right) \\
&= \sum_{j=1}^{n} \operatorname{Var}\left(I_j\right) && \text{(variance of independent sum)} \\
&= \sum_{j=1}^{n} p(1-p) \\
&= np(1-p).
\end{aligned}$$

Examples

For the coin tosses, $X \sim B(100, 0.5)$,

$$\begin{aligned}
\mathrm{E}(X) &= np = 100 \times 0.5 = 50, \\
\operatorname{Var}(X) &= np(1-p) = 100 \times 0.5^2 = 25,
\end{aligned}$$

and so

$$\mathrm{SD}(X) = 5.$$

Similarly, for the lizard eggs, $Y \sim B(8, 0.7)$,

$$\begin{aligned} \mathrm{E}(Y) &= np = 8 \times 0.7 = 5.6, \\ \mathrm{Var}\,(Y) &= np(1-p) = 8 \times 0.7 \times 0.3 = 1.68 \end{aligned}$$

and so

$$\mathrm{SD}(Y) = 1.30.$$

The geometric distribution

PMF

The geometric distribution is the distribution of the number of independent Bernoulli trials until the first success is encountered. If X is the number of trials until a success is encountered, and each independent trial has probability p of being a success, we write

$$X \sim Geom(p).$$

Clearly X can take on any positive integer, so to deduce the PMF, we need to calculate $\mathrm{P}(X = k)$ for $k = 1, 2, 3, \ldots$. In order to have $X = k$, we must have an

ordered sequence of $k-1$ failures followed by one success. By the multiplication rule therefore,

$$\mathrm{P}(X=k) = (1-p)^{k-1}p, \qquad k = 1,2,3,\ldots.$$

CDF

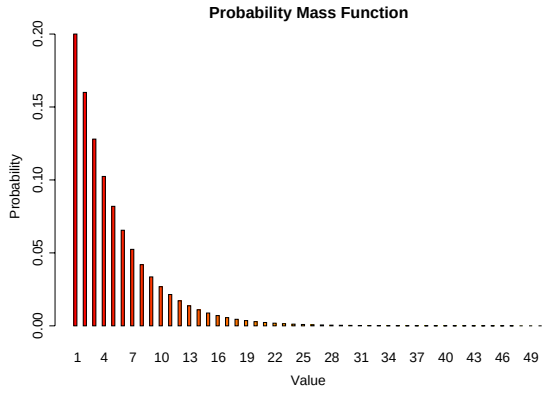
For the geometric distribution, it is possible to calculate an analytic form for the CDF as follows. If $X \sim Geom(p)$, then

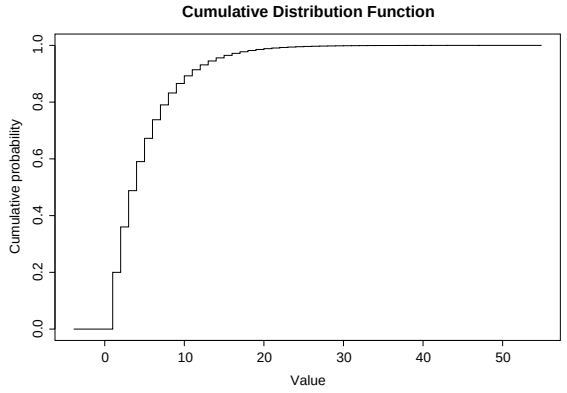
$$\begin{aligned} F_X(k) &= \mathrm{P}(X \leq k) \\ &= \sum_{j=1}^{k} (1-p)^{j-1}p \\ &= p\sum_{j=1}^{k} (1-p)^{j-1} \\ &= p \times \frac{1-(1-p)^k}{1-(1-p)} \qquad \text{(geometric series)} \\ &= 1-(1-p)^k. \end{aligned}$$

Consequently there is no need to tabulate the CDF of the geometric distribution. Also note that the CDF tends to one as k increases. This confirms that the PMF we defined does determine a valid probability distribution.

Example

Suppose that we are interested in playing a game where the probability of winning is 0.2 on any particular turn. If X is the number of turns until the first win, then $X \sim Geom(0.2)$. The PMF and CDF for X are plotted below.





Useful series in probability

Notice that we used the sum of a geometric series in the derivation of the CDF. There are many other series that crop up in the study of probability. A few of the more commonly encountered series are listed below.

$$\sum_{i=1}^{n} a^{i-1} = \frac{1-a^n}{1-a} \qquad (a > 0)$$

$$\sum_{i=1}^{\infty} a^{i-1} = \frac{1}{1-a} \qquad (0 < a < 1)$$

$$\sum_{i=1}^{\infty} i\,a^{i-1} = \frac{1}{(1-a)^2} \qquad (0 < a < 1)$$

$$\sum_{i=1}^{\infty} i^2 a^{i-1} = \frac{1+a}{(1-a)^3} \qquad (0 < a < 1)$$

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1).$$

We will use two of these in the derivation of the expectation and variance of the geometric distribution.

Expectation and variance of geometric random quantities

Suppose that $X \sim Geom(p)$. Then

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{i=1}^{\infty} i \, \mathrm{P}(X = i) \\
&= \sum_{i=1}^{\infty} i(1-p)^{i-1} p \\
&= p \sum_{i=1}^{\infty} i(1-p)^{i-1} \\
&= p \times \frac{1}{(1-[1-p])^2} \\
&= \frac{p}{p^2} \\
&= \frac{1}{p}.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{i=1}^{\infty} i^2 \, \mathrm{P}(X = i) \\
&= \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} p \\
&= p \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} \\
&= p \times \frac{1 + [1-p]}{(1-[1-p])^3} \\
&= p \times \frac{2-p}{p^3} \\
&= \frac{2-p}{p^2},
\end{aligned}
$$

and so

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}.\end{aligned}$$

Example

For $X \sim Geom(0.2)$ we have

$$\begin{aligned}\mathrm{E}(X) &= \frac{1}{p} = \frac{1}{0.2} = 5 \\ \mathrm{Var}\,(X) &= \frac{1-p}{p^2} = \frac{0.8}{0.2^2} = 20.\end{aligned}$$

The Poisson distribution

The Poisson distribution is a very important discrete probability distribution, which arises in many different contexts in probability and statistics. Typically,

Poisson random quantities are used in place of binomial random quantities in situations where n is large, p is small, and the expectation np is stable.

Example

Consider the number of calls made in a 1 minute interval to an Internet service provider (ISP). The ISP has thousands of subscribers, but each one will call with a very small probability. The ISP knows that on average 5 calls will be made in the interval. The actual number of calls will be a Poisson random variable, with mean 5.

A Poisson random variable, X with parameter λ is written as

$$X \sim P(\lambda)$$

Poisson as the limit of a binomial

Let $X \sim B(n, p)$. Put $\lambda = \mathrm{E}(X) = np$ and let n increase and p decrease so that λ remains constant.

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

Replacing p by λ/n gives

$$\begin{aligned}
\mathrm{P}(X=k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \\
&= \frac{\lambda^k}{k!}\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-k+1)}{n}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \\
&\to \frac{\lambda^k}{k!}\times 1\times 1\times 1\times\cdots\times 1\times\frac{e^{-\lambda}}{1}, \quad \text{as } n\to\infty \\
&= \frac{\lambda^k}{k!}e^{-\lambda}.
\end{aligned}$$

To see the limit, note that $(1-\lambda/n)^n \to e^{-\lambda}$ as n increases (compound interest formula).

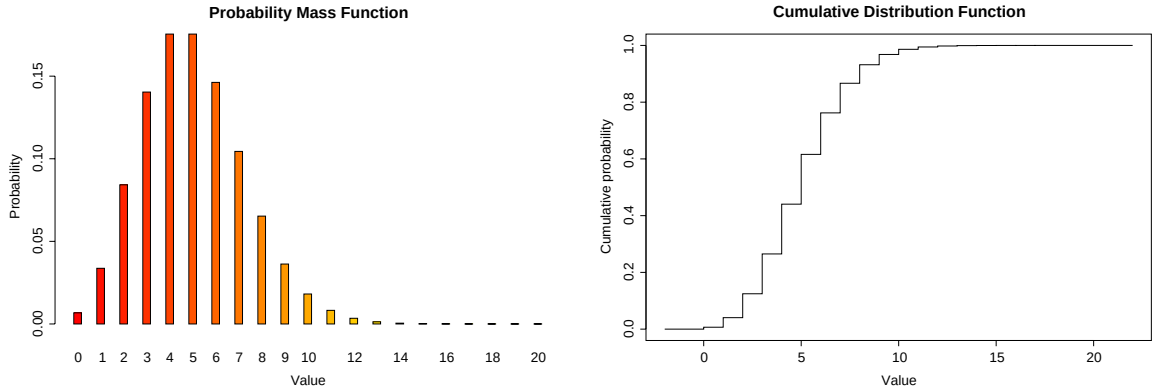
PMF

If $X \sim P(\lambda)$, then the PMF of X is

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, 3, \ldots .$$

Example

The PMF and CDF of $X \sim P(5)$ are given below.



Note that the CDF does seem to tend to 1 as n increases. However, we do need to verify that the PMF we have adopted for $X \sim P(\lambda)$ does indeed define

a valid probability distribution, by ensuring that the probabilities do sum to one:

$$\begin{aligned} \mathrm{P}(S_X) &= \sum_{k=0}^{\infty} \mathrm{P}(X = k) \\ &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1. \end{aligned}$$

Expectation and variance of Poisson

If $X \sim P(\lambda)$, we have

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{k=0}^{\infty} k\,\mathrm{P}(X=k) \\
&= \sum_{k=1}^{\infty} k\,\mathrm{P}(X=k) \\
&= \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\
&= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}e^{-\lambda} \\
&= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} \\
&= \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}e^{-\lambda} && \text{(putting } j=k-1\text{)} \\
&= \lambda \sum_{j=0}^{\infty} \mathrm{P}(X=j) \\
&= \lambda.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{k=0}^{\infty} k^2 \, \mathrm{P}(X=k) \\
&= \sum_{k=1}^{\infty} k^2 \, \mathrm{P}(X=k) \\
&= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
&= \lambda \sum_{j=0}^{\infty} (j+1) \frac{\lambda^j}{j!} e^{-\lambda} && \text{(putting } j = k-1\text{)} \\
&= \lambda \left[\sum_{j=0}^{\infty} j \frac{\lambda^j}{j!} e^{-\lambda} + \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} \right] \\
&= \lambda \left[\sum_{j=0}^{\infty} j \, \mathrm{P}(X=j) + \sum_{j=0}^{\infty} \mathrm{P}(X=j) \right] \\
&= \lambda [\, \mathrm{E}(X) + 1] \\
&= \lambda(\lambda+1) \\
&= \lambda^2 + \lambda.
\end{aligned}
$$

So,

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 \\ &= [\lambda^2 + \lambda] - \lambda^2 \\ &= \lambda.\end{aligned}$$

That is, the mean and variance are both λ.

Sum of Poisson random quantities

One of the particularly convenient properties of the Poisson distribution is that the sum of two independent Poisson random quantities is also a Poisson random quantity. If $X \sim P(\lambda)$ and $Y \sim P(\mu)$ and X and Y are independent, then $Z = X + Y \sim P(\lambda + \mu)$. Clearly this result extends to the sum of many independent Poisson random variables. The proof is straightforward, but is a little messy, and hence omitted from this course.

Example

Returning to the example of calls received by an ISP. The number of calls in 1 minute is $X \sim P(5)$. Suppose that the number of calls in the following minute is

$Y \sim P(5)$, and that Y is independent of X. Then, by the above result, $Z = X + Y$, the number of calls in the two minute period is Poisson with parameter 10. Extending this in the natural way, we see that the number of calls in t minutes is Poisson with parameter $5t$. This motivates the following definition.

The Poisson process

A sequence of timed observations is said to follow a *Poisson process* with *rate* λ if the number of observations, X, in any interval of length t is such that

$$X \sim P(\lambda t).$$

Example

For the ISP example, the sequence of incoming calls follow a Poisson process with rate 5 (per minute).

Continuous Probability Models

Introduction, PDF and CDF

Introduction

We now a have a fairly good understanding of discrete probability models, but as yet we haven't developed any techniques for handling continuous random quantities. These are random quantities with a sample space which is neither finite nor countably infinite. The sample space is usually taken to be the real line, or a part thereof. Continuous probability models are appropriate if the result of an experiment is a continuous *measurement*, rather than a *count* of a discrete set.

If X is a continuous random quantity with sample space S_X, then for any particular $a \in S_X$, we generally have that

$$\mathrm{P}(X = a) = 0.$$

This is because the sample space is so "large" and every possible outcome so "small" that the probability of any *particular* value is vanishingly small.

Therefore the probability mass function we defined for discrete random quantities is inappropriate for understanding continuous random quantities. In order to understand continuous random quantities, we need a little calculus.

The probability density function

If X is a *continuous* random quantity, then there exists a function $f_X(x)$, called the *probability density function* (PDF), which satisfies the following:

1. $f_X(x) \geq 0, \quad \forall x$;

2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$;

3. $\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx$ for any a and b.

Consequently we have

$$\begin{aligned} \mathrm{P}(x \leq X \leq x+\delta x) &= \int_x^{x+\delta x} f_X(y)\,dy \\ &\simeq f_X(x)\delta x, \qquad \text{(for small } \delta x\text{)} \\ \Rightarrow f_X(x) &\simeq \frac{\mathrm{P}(x \leq X \leq x+\delta x)}{\delta x} \end{aligned}$$

and so we may interpret the PDF as

$$f_X(x) = \lim_{\delta x \to 0} \frac{\mathrm{P}(x \leq X \leq x+\delta x)}{\delta x}.$$

Example

The manufacturer of a certain kind of light bulb claims that the lifetime of the bulb in hours, X can be modelled as a random quantity with PDF

$$f_X(x) = \begin{cases} 0, & x < 100 \\ \dfrac{c}{x^2}, & x \geq 100, \end{cases}$$

where c is a constant. What value must c take in order for this to define a valid PDF? What is the probability that the bulb lasts no longer than 150 hours?

Given that a bulb lasts longer than 150 hours, what is the probability that it lasts longer than 200 hours?

Notes

1. Remember that PDFs are *not* probabilities. For example, the density can take values greater than 1 in some regions as long as it still integrates to 1.

2. It is sometimes helpful to think of a PDF as the limit of a relative frequency histogram for many realisations of the random quantity, where the number of realisations is very large and the bin widths are very small.

3. Because $\mathrm{P}(X = a) = 0$, we have $\mathrm{P}(X \leq k) = \mathrm{P}(X < k)$ for continuous random quantities.

The distribution function

In the last chapter, we defined the cumulative *distribution function* of a random variable X to be

$$F_X(x) = \mathrm{P}(X \leq x), \quad \forall x.$$

This definition works just as well for continuous random quantities, and is one of the many reasons why the distribution function is so useful. For a discrete random quantity we had

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{\{y \in S_X | y \leq x\}} \mathrm{P}(X = y),$$

but for a continuous random quantity we have the continuous analogue

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) \\ &= \mathrm{P}(-\infty \leq X \leq x) \\ &= \int_{-\infty}^{x} f_X(z)\, dz. \end{aligned}$$

Just as in the discrete case, the distribution function is defined for all $x \in \mathbf{R}$, even if the sample space S_X is not the whole of the real line.

Properties

1. Since it represents a probability, $F_X(x) \in [0,1]$.

2. $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.

3. If $a < b$, then $F_X(a) \leq F_X(b)$. *ie.* $F_X(\cdot)$ is a non-decreasing function.

4. When X is continuous, $F_X(x)$ is *continuous*. Also, by the Fundamental Theorem of Calculus, we have

$$\boxed{\frac{d}{dx}F_X(x) = f_X(x),}$$

and so the *slope* of the CDF $F_X(x)$ is the PDF $f_X(x)$.

Example

For the light bulb lifetime, X, the distribution function is

$$F_X(x) = \begin{cases} 0, & x < 100 \\ 1 - \dfrac{100}{x}, & x \geq 100. \end{cases}$$

Median and quartiles

The *median* of a random quantity is the value m which is the "middle" of the distribution. That is, it is the value m such that

$$\mathrm{P}(X \leq m) = \mathrm{P}(X \geq m) = \frac{1}{2}.$$

Equivalently, it is the value, m such that

$$F_X(m) = 0.5.$$

Similarly, the *lower quartile* of a random quantity is the value l such that

$$F_X(l) = 0.25,$$

and the *upper quartile* is the value u such that

$$F_X(u) = 0.75.$$

Example

For the light bulb lifetime, X, what is the median, upper and lower quartile of the distribution?

Properties of continuous random quantities

Expectation and variance of continuous random quantities

The *expectation* or *mean* of a continuous random quantity X is given by

$$\boxed{\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx,}$$

which is just the continuous analogue of the corresponding formula for discrete random quantities. Similarly, the *variance* is given by

$$\begin{aligned}\mathrm{Var}\,(X) &= \int_{-\infty}^{\infty} [x - \mathrm{E}(X)]^2\, f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx - [\mathrm{E}(X)]^2.\end{aligned}$$

Note that the expectation of $g(X)$ is given by

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

and so the variance is just

$$\boxed{\mathrm{Var}\,(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^2\right) = \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2}$$

as in the discrete case. Note also that all of the properties of expectation and variance derived for discrete random quantities also hold true in the continuous case.

Example

Consider the random quantity X, with PDF

$$f_X(x) = \begin{cases} \frac{3}{4}(2x - x^2), & 0 < x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

Check that this is a valid PDF (it integrates to 1). Calculate the expectation and variance of X. Evaluate the distribution function. What is the median of this distribution?

PDF and CDF of a linear transformation

Let X be a continuous random quantity with PDF $f_X(x)$ and CDF $F_X(x)$, and let $Y = aX + b$ where $a > 0$. What is the PDF and CDF of Y? It turns out to be

easier to work out the CDF first:

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) \\ &= \mathrm{P}(aX + b \leq y) \\ &= \mathrm{P}\left(X \leq \frac{y-b}{a}\right) && \text{(since } a > 0\text{)} \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

So,

$$\boxed{F_Y(y) = F_X\left(\frac{y-b}{a}\right),}$$

and by differentiating both sides with respect to y we get

$$\boxed{f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).}$$

Example

For the light bulb lifetime, X, what is the density of $Y = X/24$, the lifetime of the bulb in days?

The uniform distribution

Now that we understand the basic properties of continuous random quantities, we can look at some of the important standard continuous probability models. The simplest of these is the uniform distribution.

The random quantity X has a *uniform distribution* over the range $[a, b]$, written

$$X \sim U(a, b)$$

if the PDF is given by

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

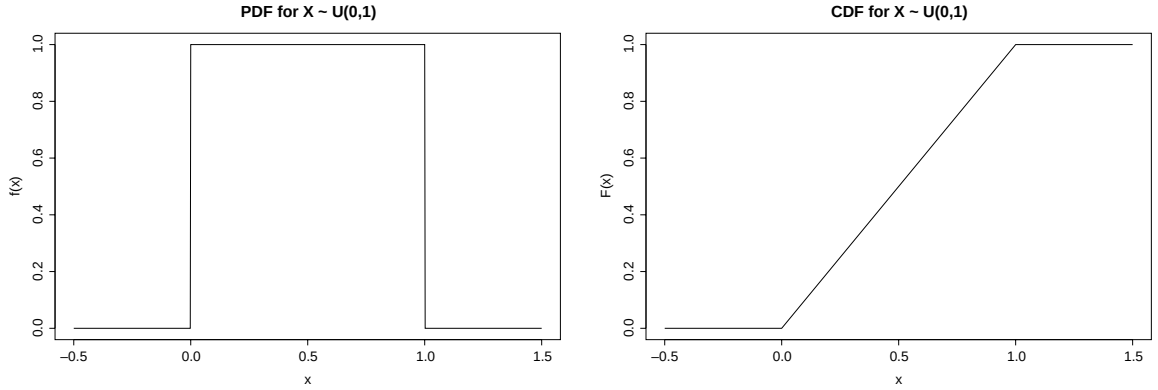
Thus if $x \in [a,b]$,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(y)\,dy \\ &= \int_{-\infty}^{a} f_X(y)\,dy + \int_{a}^{x} f_X(y)\,dy \\ &= 0 + \int_{a}^{x} \frac{1}{b-a} dy \\ &= \left[\frac{y}{b-a}\right]_{a}^{x} \\ &= \frac{x-a}{b-a}. \end{aligned}$$

Therefore,

$$F_X(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

We can plot the PDF and CDF in order to see the "shape" of the distribution. Below are plots for $X \sim U(0,1)$.



Clearly the lower quartile, median and upper quartile of the uniform distribution are

$$\frac{3}{4}a+\frac{1}{4}b, \quad \frac{a+b}{2}, \quad \frac{1}{4}a+\frac{3}{4}b,$$

respectively. The expectation of a uniform random quantity is

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \\
&= \int_{-\infty}^{a} x\, f_X(x)\, dx + \int_{a}^{b} x\, f_X(x)\, dx + \int_{b}^{\infty} x\, f_X(x)\, dx \\
&= 0 + \int_{a}^{b} \frac{x}{b-a} dx + 0 \\
&= \left[\frac{x^2}{2(b-a)} \right]_a^b \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{a+b}{2}.
\end{aligned}$$

We can also calculate the variance of X. First we calculate $\mathrm{E}\left(X^2\right)$ as follows:

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_a^b \frac{x^2}{b-a} \\
&= \left[\frac{x^3}{3(b-a)}\right]_a^b \\
&= \frac{b^3-a^3}{3(b-a)} \\
&= \frac{b^2+ab+a^2}{3}.
\end{aligned}$$

Now,

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left(X^2\right) - \operatorname{E}(X)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2}{12} \\ &= \frac{1}{12}[b^2-2ab+a^2] \\ &= \frac{(b-a)^2}{12}.\end{aligned}$$

The uniform distribution is rather too simple to realistically model actual experimental data, but is very useful for computer simulation, as random quantities from many different distributions can be obtained from $U(0,1)$ random quantities.

The exponential distribution

Definition and properties

The random variable X has an *exponential distribution* with parameter $\lambda > 0$, written

$$X \sim Exp(\lambda)$$

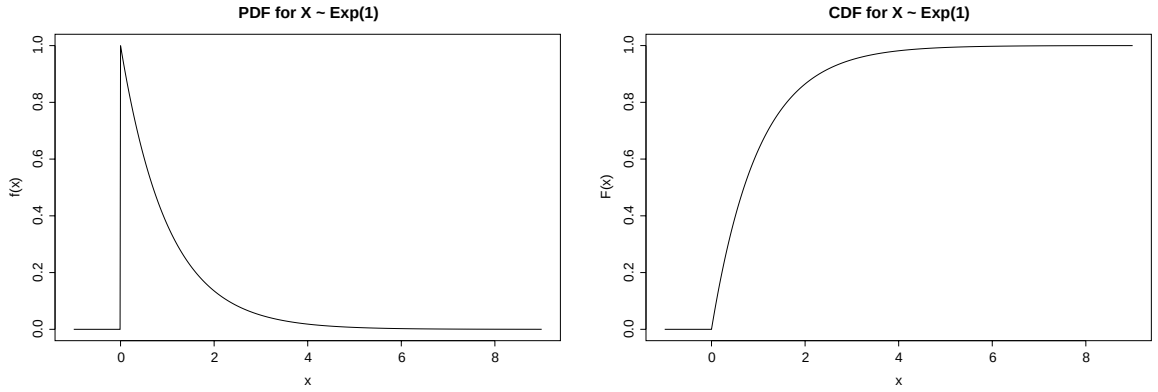
if it has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The distribution function, $F_X(x)$ is therefore given by

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

The PDF and CDF for an $Exp(1)$ are shown below.



The expectation of the exponential distribution is

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^\infty x\lambda e^{-\lambda x} dx \\
&= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx && \text{(by parts)} \\
&= 0 + \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^\infty \\
&= \frac{1}{\lambda}.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_0^\infty x^2\lambda e^{-\lambda x} dx \\
&= \frac{2}{\lambda^2},
\end{aligned}$$

and so

$$\begin{aligned}
\operatorname{Var}(X) &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}.
\end{aligned}$$

Note that this means the expectation and standard deviation are both $1/\lambda$.

Notes

1. As λ increases, the probability of small values of X increases and the mean decreases.

2. The median m is given by

$$m = \frac{\log 2}{\lambda} = \log 2 \, \mathrm{E}(X) < \mathrm{E}(X) \, .$$

3. The exponential distribution is often used to model lifetime and times between random events. The reasons are given below.

Relationship with the Poisson process

The exponential distribution with parameter λ is the time between events of a Poisson process with rate λ. Let X be the number of events in the interval

$(0,t)$. We have seen previously that $X \sim P(\lambda t)$. Let T be the time to the first event. Then

$$\begin{aligned} F_T(t) &= \mathrm{P}(T \leq t) \\ &= 1 - \mathrm{P}(T > t) \\ &= 1 - \mathrm{P}(X = 0) \\ &= 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \\ &= 1 - e^{-\lambda t}. \end{aligned}$$

This is the distribution function of an $Exp(\lambda)$ random quantity, and so $T \sim Exp(\lambda)$.

Example

Consider again the Poisson process for calls arriving at an ISP at rate 5 per minute. Let T be the time between two consecutive calls. Then we have

$$T \sim Exp(5)$$

and so $\mathrm{E}(T) = \mathrm{SD}(T) = 1/5$.

The memoryless property

If $X \sim Exp(\lambda)$, then

$$
\begin{aligned}
\mathrm{P}(X > (s+t) | X > t) &= \frac{\mathrm{P}([X > (s+t)] \cap [X > t])}{\mathrm{P}(X > t)} \\
&= \frac{\mathrm{P}(X > (s+t))}{\mathrm{P}(X > t)} \\
&= \frac{1 - \mathrm{P}(X \leq (s+t))}{1 - \mathrm{P}(X \leq t)} \\
&= \frac{1 - F_X(s+t)}{1 - F_X(t)} \\
&= \frac{1 - [1 - e^{-\lambda(s+t)}]}{1 - [1 - e^{-\lambda t}]} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\
&= e^{-\lambda s} \\
&= 1 - [1 - e^{-\lambda s}] \\
&= 1 - F_X(s) \\
&= 1 - \mathrm{P}(X \leq s) \\
&= \mathrm{P}(X > s) .
\end{aligned}
$$

So in the context of lifetimes, the probability of surviving a further time s, having survived time t is the same as the original probability of surviving a time s. This is called the “memoryless” property of the distribution. It is therefore the continuous analogue of the geometric distribution, which also has such a property.

The normal distribution

Definition and properties

A random quantity X has a normal distribution with parameters μ and σ^2, written

$$X \sim N(\mu, \sigma^2)$$

if it has probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty,$$

for $\sigma > 0$. Note that $f_X(x)$ is symmetric about $x = \mu$, and so (provided the density integrates to 1), the median of the distribution will be μ. Checking that

the density integrates to one requires the computation of a slightly tricky integral. However, it follows directly from the known "Gaussian" integral

$$\int_{-\infty}^{\infty} e^{-\alpha x^2}\,dx = \sqrt{\frac{\pi}{\alpha}}, \qquad \alpha > 0,$$

since then

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2} z^2\right\} dz \qquad \text{(putting } z = x-\mu\text{)} \\
&= \frac{1}{\sigma\sqrt{2\pi}} \sqrt{\frac{\pi}{1/2\sigma^2}} \\
&= \frac{1}{\sigma\sqrt{2\pi}} \sqrt{2\pi\sigma^2} \\
&= 1.
\end{aligned}$$

Now we know that the given PDF represents a valid density, we can calculate

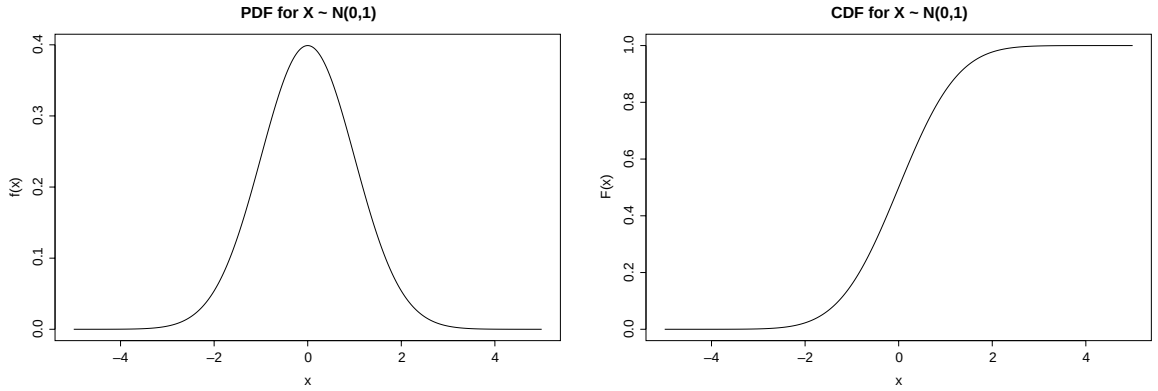
the expectation and variance of the normal distribution as follows:

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} dx \\
&= \mu
\end{aligned}$$

after a little algebra. Similarly,

$$\begin{aligned}
\mathrm{Var}\,(X) &= \int_{-\infty}^{\infty} (x-\mu)^2\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} dx \\
&= \sigma^2.
\end{aligned}$$

The PDF and CDF for a $N(0,1)$ are shown below.



The standard normal distribution

A standard normal random quantity is a normal random quantity with zero mean and variance equal to one. It is usually denoted Z, so that

$$Z \sim N(0,1).$$

Therefore, the density of Z, which is usually denoted $\phi(z)$, is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}, \qquad -\infty < z < \infty.$$

It is important to note that the PDF of the standard normal is symmetric about zero. The distribution function of a standard normal random quantity is denoted $\Phi(z)$, that is

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)dx.$$

There is no neat analytic expression for $\Phi(z)$, so tables of the CDF are used. Of course, we do know that $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$, as it is a distribution function. Also, because of the symmetry of the PDF about zero, it is clear that we must also have $\Phi(0) = 1/2$, and this can prove useful in calculations. The standard normal distribution is important because it is easy to transform any normal random quantity to a standard normal random quantity by means of a simple linear scaling. Consider $Z \sim N(0,1)$ and put

$$X = \mu + \sigma Z,$$

for $\sigma > 0$. Then $X \sim N(\mu, \sigma^2)$. To show this, we must show that the PDF of X is the PDF for a $N(\mu, \sigma^2)$ random quantity. Using the result for the PDF of a linear transformation, we have

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \end{aligned}$$

which is the PDF of a $N(\mu, \sigma^2)$ distribution. Conversely, if

$$X \sim N(\mu, \sigma^2)$$

then

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

Even more importantly, the distribution function of X is given by

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

and so the cumulative probabilities for any normal random quantity can be calculated using tables for the standard normal distribution.

Example

Suppose that $X \sim N(3, 2^2)$, and we are interested in the probability that X does not exceed 5. Then

$$\mathrm{P}(X < 5) = \Phi\left(\frac{5-3}{2}\right) = \Phi(1) = 0.84134.$$

Notes

The normal distribution is probably *the* most important probability distribution in statistics. In practice, many measured variables may be assumed to be approximately normal. For example, weights, heights, IQ scores, blood pressure measurements *etc.* are all usually assumed to follow a normal distribution.

The ubiquity of the normal distribution is due in part to the *Central Limit Theorem*. Essentially, this says that *sample means* and *sums* of independent random quantities are approximately normally distributed whatever the distribution of the original quantities, as long as the sample size is reasonably large — more on this in Semester 2.

Normal approximation of binomial and Poisson

Normal approximation of the binomial

We saw in the last chapter that $X \sim B(n, p)$ could be regarded as the sum of n independent Bernoulli random quantities

$$X = \sum_{k=1}^{n} I_k,$$

where $I_k \sim Bern(p)$. Then, because of the central limit theorem, this will be well approximated by a Normal distribution if n is large, and p is not too extreme (if p is very small or very large, a Poisson approximation will be more appropriate). A useful guide is that if

$$0.1 \leq p \leq 0.9 \quad \text{and} \quad n > \max\left[\frac{9(1-p)}{p}, \frac{9p}{1-p}\right]$$

then the binomial distribution may be adequately approximated by a normal distribution. It is important to understand exactly what is meant by this statement. No matter how large n is, the binomial will always be a discrete random quantity with a PMF, whereas the normal is a continuous random quantity with a PDF. These two distributions will always be qualitatively different. The similarity is measured in terms of the CDF, which has a consistent definition for both discrete and continuous random quantities. It is the CDF of the binomial which can be well approximated by a normal CDF. Fortunately, it is the CDF which matters for typical computations involving cumulative probabilities.

When the n and p of a binomial distribution are appropriate for approximation by a normal distribution, the approximation is done by matching expectation

and variance. That is

$$B(n,p) \simeq N(np, np[1-p]).$$

Example

Reconsider the number of heads X in 100 tosses of an unbiased coin. There $X \sim B(100, 0.5)$, which may be well approximated as

$$X \simeq N(50, 5^2).$$

So, using normal tables we find that $\mathrm{P}(40 \leq X \leq 60) \simeq 0.955$ and $\mathrm{P}(30 \leq X \leq 70) \simeq 1.000$, and these are consistent with the exact calculations we undertook earlier: 0.965 and 1.000 respectively.

Normal approximation of the Poisson

Since the Poisson is derived from the binomial, it is unsurprising that in certain circumstances, the Poisson distribution may also be approximated by the normal. It is generally considered appropriate to make the approximation

if the mean of the Poisson is bigger than 20. Again the approximation is done by matching mean and variance:

$$X \sim P(\lambda) \simeq N(\lambda, \lambda) \text{ for } \lambda > 20.$$

Example

Reconsider the Poisson process for calls arriving at an ISP at rate 5 per minute. Consider the number of calls X, received in 1 hour. We have

$$X \sim P(5 \times 60) = P(300) \simeq N(300, 300).$$

What is the approximate probability that the number of calls is between 280 and 310?