MATH1710 - Probability and Statistics 1

This module will introduce you to the basics of probability and statistics, which are essential for later modules. Although many of you will already have taken one or two courses which include similar topics, that is not true of the whole class. Even where there is overlap, we will cover the algebraic derivations and consider assumptions – there will be something new for everyone. Also, we will perform calculations and graphical tasks using the R coding language – a free software application – which is used throughout the undergraduate statistics teaching in Leeds and in the professional data science world.

The origins of probability and statistics are very different, and so perhaps it is a surprise that they are in fact so closely related. The study of probability first occurred in the 17th century, motivated by gambling, whereas statistics has its origins 5000 years earlier as a means of recording people and their possessions for tax purposes. Probability and statistics have, however, come a long way and are a critical aspect of the modern world.

The Chief Economist at Google, Hal Varian, said about statistics:

> "*The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it, is going to be a hugely important skill in the next decades . . . because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.* "

Probability and statistics are branches of mathematics which rigorously measure and describes uncertain (or random) systems and processes. Modern applications include: modelling hereditary disease in genetics, pension predictions in actuarial science, stock price movements in finance, epidemic modelling in public health.

The twin topics of probability and statistics can be artificially separated and taught in isolation – mirroring their origins – but by bringing them together the links between application and theory can be more clearly seen. This means that we will frequently switch from numerics and computing, to algebra and calculus.

Contact information:

Lecturer:	Dr Robert G. Aykroyd
Office:	Room 10.18 (School of Mathematics Satellite)
E-mail:	`r.g.aykroyd@leeds.ac.uk`
Webpages:	Note that all course material will be made available on `Minerva`.

Syllabus details:

Chapter 1: EDA in R Introduction to EDA. R and RStudio. Histogram and scatterplots. Mean, variance, correlation and five-number summary. Boxplots.

Chapter 2: Basic probability: Events. Probability as relative frequency. Law of Large Numbers. Subjective probability. Probability axioms. General addition rule.

Chapter 3: Conditional probability: Conditional probability and the multiplication rule. Independence. Total probability and Bayes' rule.

Chapter 4: Random variables: Expectations and variances. Functions of random variables. Probability generating functions. Joint, conditional and marginal distributions. Conditional expectation and conditional variance. Independence and correlation.

Chapter 5: Models for count data: Bernoulli trials. Binomial. Geometric. Poisson. Hypergeometric. Estimation of parameters. Sums of independent random variables.

Chapter 6: Models for measurement data: Uniform. Beta. Exponential. Normal. Estimation of parameters. Functions of random variables. Central limit theorem.

Chapter 7: Bayesian methods: Likelihood, prior and posterior distributions. Posterior mean and posterior variance.

Methods of teaching:
Lectures (22 hours) will be held on Mondays at 12–1 with a repeat at 3–4 and on Thursdays at 2–3 with a repeat at 4–5 — you must, however, attend the days/times as on your timetable. These will cover most of the required material and key admin.

Tutorials (5 hours) will be held in Weeks **2**, **4**, **6**, **8** and **10**. You are required to attend one tutorial in each of these weeks where you can get help from your Statistics Tutor to complete coursework and discuss the material covered in Lectures.

Assessment: 80% two-hour written examination at end of semester,
20% coursework (Tutorial and Computing Exercises).

Tutorial Exercises will be given out every two weeks.You should start these before your next Statistics Tutorial and hand-in your solutions a few days after your tutorial. Short Computing Exercises will be set every week, to reinforce the ideas seen in the Lectures. For some of these you will be required to hand-in your answers or a short report.

Late submission of assessed work is subject to a penalty and any work submitted after the solutions have been handed out will receive a mark of zero.

!!! In order to pass the module, you must pass the examination. !!!

Booklist: (Available from the University Library)

1. Scheaffer RL and Young LJ, Introduction to Probability and its Applications, 3rd Ed, Brooks/Cole Cengage Learning, 2010.
2. Stone JV, Bayes' Rule: A Tutorial Introduction, Sebtel Press, 2013.
3. Clarke GM and Cooke D, A Basic Course in Statistics, 5th Ed, Arnold, 2004.
4. Rice JA, Mathematical Statistics and Data Analysis, 3rd Ed, Thomson Press, 2007.
5. ‡Stirzaker DR, Elementary Probability, 2nd Ed, CUP, 2003.

‡ Available online via the Library website.

Contents

Preface

This module will introduce you to the basics of probability and statistics which are essential for later modules. Although many of you will already have taken one or two courses which include similar topics, that is not true of the whole class. Even where there is overlap, we will cover the algebraic derivations and consider assumption – there will be something new for everyone. Also, we will perform calculations, and other numerical tasks, using the R coding language – a free software package – which is used throughout the undergraduate teaching in Leeds and in the professional data science world.

The origins of probability and statistics are very different, and so perhaps it is a surprise that they are in fact so closely related. The study of probability first occurred in the 17th century, motivated my gambling, whereas statistics has its origins 5000 years earlier as a means of recording people and their possessions for tax purposes. Probability and statistics have, however, come a long way and are a critical aspect of the modern world.

The twin topics of probability and statistics can be artificially separated and taught in isolation – mirroring their origins – but by bringing them together the links between applications and theory can be more clearly seen. This means that we will frequently switch from numerics and computing to algebra and calculus.

Probability and statistics are branches of mathematics which rigorously measure and describes uncertain (or random) systems and processes. Modern applications include: modelling hereditary disease in genetics, pension predictions in actuarial science, stock price movements in finance, epidemic modelling in public health.

1 Exploratory Data Analysis in R

1.1 Introduction

Suppose we have performed a random experiment or applied some sampling technique to a population, then the collection of observations is called a *random sample* or *dataset* which we will write as a vector, for example $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ for a sample of n measurements.

In R we can easily define vectors and check their length:

```
> x=c(2.5,5,3)
> x
[1] 2.5 5.0 3.0
> length(x)
[1] 3
```

R-*note. To check details of an* R *command then you can use, for example,* `help(length)`.

If a dataset contains only a few values then we might see any interesting features by looking at the numbers themselves. Once we are faced with more than a handful of values, however, it is rare that any useful information can be gained by simply examining the individual values. Instead, we can look at a suitably chosen picture of the data and get a feel for the general structure. Here we shall look at a few simple graphical representations, but please note that these are only examples of what is available.

1.2 Frequency and relative frequency

For a dataset representing measurements, a simple approach is to divide the range of values into non-overlapping intervals, called classes, and to count how many values are in each class. Note that these classes are usually, but are not required to be, of equal width. (For discrete data, each distinct value might define a class.)

Consider the percentage return on Lloyds Banking Group shares for 24 consecutive months (Jan 2015-Dec 2016, source: shareprices.com):

```
 -2.7  7.1  -0.9  -1.1  13.4  -2.9  -2.4  -7.0  -2.9  -1.9  -1.0   0.2
-10.4 10.6  -6.0  -1.4   7.4 -24.9  -1.7  11.7  -8.1   5.0   1.1   8.0
```

```
> data = read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
MonthlyShareReturns.csv")
> data$returns
[1] -2.7 7.1 -0.9 -1.1 13.4 -2.9 -2.4 -7.0 -2.9 -1.9 -1.0 0.2 -10.4
[14] 10.6 -6.0 -1.4 7.4 -24.9 -1.7 11.7 -8.1 5.0 1.1 8.0
```

R-*note. To see details of the components of a variable use, for example,* `head(data)`.

The value `-24.9` is extreme — corresponding to June 2016 and the "Brexit" vote. Suppose we ignore this value as being unrepresentative – never to be repeated – and then choose to have 6 classes of equal width, fixing the boundaries at: -15, -10, -5, 0, 5, 10, 15. Loosely speaking, the first interval will contain all data values up to -10, the second the data values from -10 to 5, and so on. But in which interval would we place 5.0? We must be careful not to be ambiguous and so let us say that the first interval is up to and including -10, then from -10 up to and including -5 etc. In standard mathematical notation these can be written $(-15, -10]$, $(-10, -5]$ etc – where the shape of the bracket carries extra meaning – square brackets, [ ], mean including the endpoint value whereas parentheses, also known as round brackets, ( ), mean excluding the endpoint value.

It is now possible to consider each data value in turn and increment the count in the appropriate class leading to the following table:

	$(-15, -10]$	$(-10, -5]$	$(-5, -0]$	$(0, 5]$	$(5, 10]$	$(10, 15]$	Sum
Frequency	1	3	10	3	3	3	23
Rel. Freq.	0.04	0.13	0.43	0.13	0.13	0.13	1.0

Do you think this gives a good summary of the data? Why?

```
> which(data$returns == -24.9)
[1] 18
> returns.new = data$returns[-18]
> brks = c(-15,-10,-5,0,5,10,15)
> tmp = hist(returns.new, breaks=brks, plot=F)
> tmp$counts
[1] 1 3 10 3 3 3
> round(tmp$counts/sum(tmp$counts),2)
[1] 0.04 0.13 0.43 0.13 0.13 0.13
```

R-*note. Use the $ symbol to access components and* `head` *to see the available components.*

In general, suppose that we choose m classes. Then, the class midpoints can be written as a vector, for example $(x_1, x_2, \ldots, x_m)$, with corresponding class frequencies $(f_1, f_2, \ldots, f_m)$ — note that $n = f_1 + f_2 + \cdots + f_m$ or written as $n = \sum_{i=1}^{m} f_i$, that is n is the sum of the class frequencies.

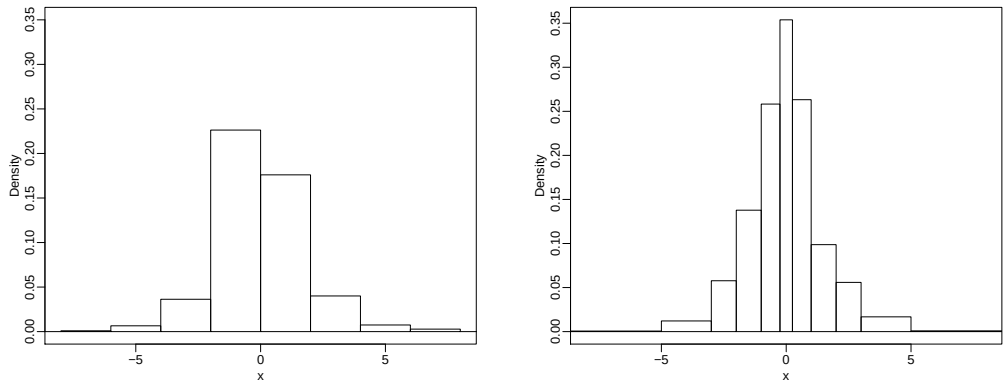
Rather than the frequencies, it is sometimes more useful to see what proportion of the values fall into each of the classes, for example to compare datasets of different sizes. The relative frequency is simply the class frequency divided by the sample size.

> If any of the statistical ideas in this Chapter are new, or you do not remember the details, then you will find good explanations in, for example, the first few chapters of the recommended textbook: Clarke GM and Cooke D, A Basic Course in Statistics. There are, of course, many good online resources which can also be useful to fill-in any gaps.

1.3 Histograms, time series plots and scatterplots

Although the frequency, or relative frequency, table helps to summarise a large dataset, it is even more useful to have a picture. The appropriate choice for continuous measurement data is the histogram (or the bar chart for discrete/categorical data).

Let us start by considering a bigger dataset using the **daily** percentage return for Lloyds Banking Group shares (Jan 2015-Dec 2016, source: shareprices.com). Figure 1 shows corresponding histograms using equal width intervals and unequal width intervals. Also, both horizontal and vertical scales are equal to allow direct comparison. Note that in each, the vertical scale is marked as `Density` which is the relative frequency within the interval divided by the interval width. This is called a density-scaled histogram and has the property that the total area of the bars equals one — more on this in a future week. Both show that most values are between -5 and $+5$, and that there is a big peak at zero.



(a) Equal intervals widths (b) Unequal interval widths

Figure 1: Density-scaled histograms of Lloyds Bank daily percentage returns.

Which of these histograms do you think gives the most accurate summary? Why?

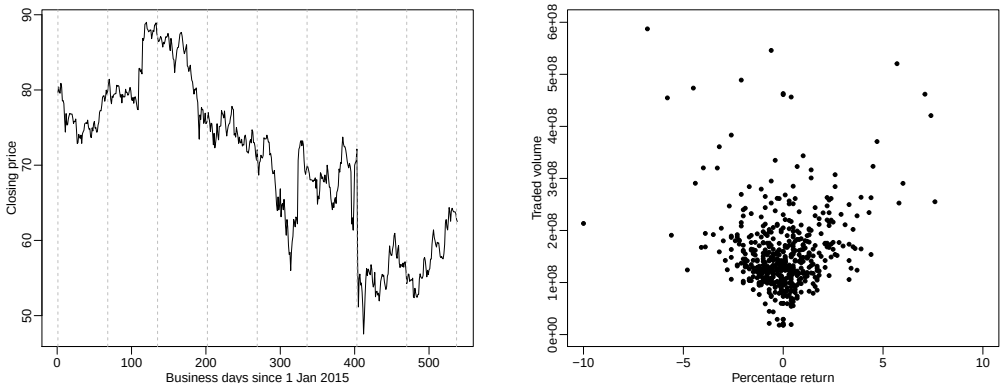
Note that with equal interval widths, the frequency histogram, relative frequency histogram and the density-scaled histogram would have exactly the same appearance – only the vertical scale would be changed. However, choosing the relative frequency or density-scaled histograms allows us to easily compare the shape of more than one sample even when the number of data points is different. Also, if we use unequal interval widths, then we must account for this by using the density-scaled histogram.

The appearance of the histogram depends on the number, and hence the width of the classes. As a general rule between 5 and 20 classes is reasonable — with the greater number of classes being used with larger datasets. A common starting point is to use the number of bars equal to the square-root of the size of the dataset, so 16 points leads to 4 bars; 25 data points to 5 bars, 100 data points to 10 bars, and so on.

```
> data=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
DailyShareData.csv")
> hist(data$daily.return, breaks=13, probability=T, ylim=c(0,0.35),
xlim=c(-8,8), xlab="x", ylab="Density", main="" )
[Output similar to Fig 1(a) above]
>
> brks = c(-25,-20,-5, -3, -2, -1, -0.25, 0.25, 1, 2, 3, 5, 20)
> hist(data$daily.return,breaks=brks, probability=T, xlim=c(-8,8),
xlab="x", ylab="Density", main="" )
[Output similar to Fig 1(b) above]
```

R-*note. You can choose, almost, any name for your variables but try to use something easy to remember and type.*

In many situations we collect data sequentially through time, or simultaneously in pairs, and hence pictures that reflect this structure are more appropriate. Figure 2(a) shows a time series plot of the closing price divided into quarter-year periods — note that in such plots, the time variable is always on the horizontal axis.



(a) Time series plot of closing price (b) Scatterplot of return and volume

Figure 2: Lloyds Bank daily closing price, and return against traded volume.

For this dataset, we might imagine some relationship between the daily return and the traded volume. Figure 2(b) shows a scatterplot of the percentage return against the traded volume for each day. It is clear that low-volume days have returns close to zero, but there is great variability in return on moderate and high volume days.

```
> plot(data$closing.price, type='l', xlab="Business days since 1 Jan
2015", ylab="Closing price")
[Output similar to Fig 2(a) above]
>
> plot(data$daily.return, data$daily.volume, xlim=c(-10,10), ylim=
c(1e6,6e8), pch=20, ylab="Traded volume", xlab="Percentage return")
[Output similar to Fig 2(b) above]
```

R-*note. Commands will have options, but defaults will be used if you don't specify values.*

1.4 Numerical summary statistics

It is usual to process a dataset to obtain a few important numerical features of the sample — as well as examining graphical summaries. The most often calculated are the sample mean and the sample variance (or equivalently mean and standard deviation). Here we consider two situations: (i) where we use the actual data values $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, as in the 24 consecutive monthly percentage returns, and (ii) with grouped data, as in the earlier frequency table, with class midpoints $(x_1, x_2, \ldots, x_m)$ and corresponding class frequencies $(f_1, f_2, \ldots, f_m)$ with $n = \sum_{i=1}^{m} f_i$.

The *sample mean* is denoted $\overline{x}$ and defined as:

$$\text{(i)} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{or} \qquad \text{(ii)} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{m} f_i x_i$$

and the *sample variance* is written s_x^2 and defined as:

$$\text{(i)} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{or} \qquad \text{(ii)} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{m} f_i(x_i - \overline{x})^2.$$

For computational purposes, it is more convenient to use alternative forms of the above variance definitions: (i) $s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$ or (ii) $s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{m} f_i x_i^2 - n\overline{x}^2\right)$.

The *sample standard deviation* of a random sample is written s_x and is defined as the (positive) square root of the sample variance. Note that it is usually more appropriate to quote mean and standard deviation of a sample as both statistics are then in the same units as the original measurements.

For the Lloyds Bank returns (with the outlier removed) we have: $\mathbf{x} = (-2.7, 7.1, -0.9, -1.1, 13.4, -2.9, -2.4, -7.0, -2.9, -1.9, -1.0, 0.2, -10.4, 10.6, -6.0, -1.4, 7.4, *, -1.7, 11.7, -8.1, 5.0, 1.1, 8.0)$ giving $\overline{x} = -14.1/23 = -0.61$ percent/day and $s_x^2 = (924.35 - 23 \times (-0.61)^2)/(23-1) = 41.63$ percent^2/day^2 and $s_x = \sqrt{41.63} = 6.45$ percent/day.

```
> X=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
MonthlyShareReturns.csv")
> returns.new = X$returns[-18]
> mean(returns.new)
[1] 0.6130435
> var(returns.new)
[1] 41.623
> sd(returns.new)
[1] 6.451589
```

As an exercise, supposing that we only had access to the frequency table data, then the class midpoints are $(-12.5, -7.5, -2.5, 2.5, 7.5, 12.5)$ with frequencies $(1, 3, 10, 3, 3, 3)$, giving a sample mean of 0.30 percent/day and standard deviation of 6.59 percent/day.

Whenever we measure more than one variable at the same time, such as with the price and traded volume in the Lloyds Bank dataset, we will be interested in quantifying any relationship — the most common measure is the correlation. Suppose we have n pairs of measurements $((x_1, y_1,), (x_2, y_2,), \ldots, (x_n, y_n))$, then the *sample correlation coefficient* is written r_{xy} and is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of x and y respectively, as above, and the sample covariance, s_{xy}, is defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Similarly to above, for computational purposes, it is more convenient to use the alternative form: $s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right)$.

Note that if small values in one variable are associated with small values in the second variable, and large values with large values, then there is a positive correlation. In contrast if small values of one variable are associated with large values of the other variable, then there is a negative correlation — this is less common.

```
> Y=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
DailyShareData.csv")
> cor(Y$daily.return, Y$daily.volume)
[1] -0.2646797
```

1.5 The 5-figure summary and boxplots

The 5-figure summary describes a dataset using five numerical summaries: the minimum, the lower (or first) quartile, the median, the upper (or third) quartile, and the maximum.

In R, with the full Lloyds Bank monthly percentage returns stored in vector `x`, the command `summary(x)` gives the output:

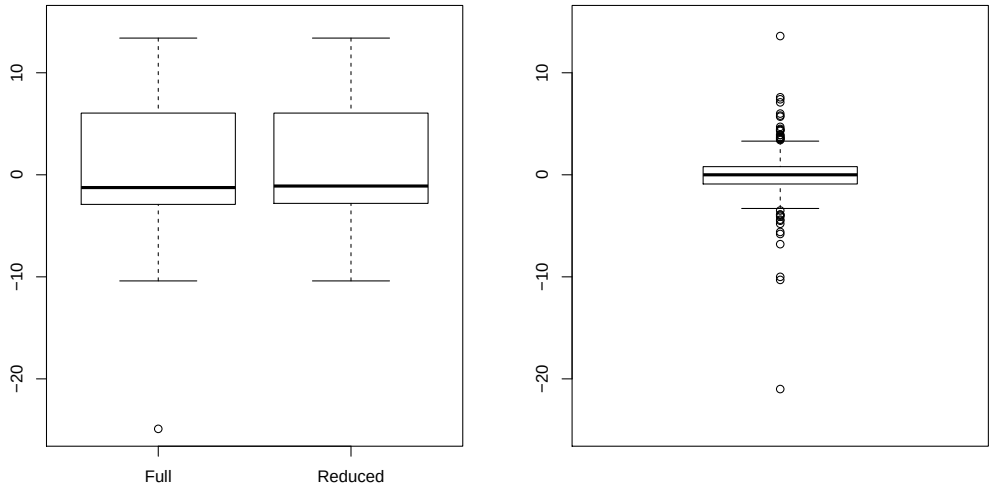
```
> x=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
MonthlyShareReturns.csv")
> summary(x$returns)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-24.900 -2.900 -1.250 -0.450 5.525 13.400
> summary(x$returns[-18])
Min. 1st Qu. Median Mean 3rd Qu. Max.
-10.400 -2.800 -1.100 0.613 6.050 13.400
```

Which of these do you think gives the best summary? Why?

The lower quartile, median and upper quartile are examples of *order statistics*. The lower quartile is the point in the data with 25% of values below and 75% above, the median has

50% below and 50% above, and the upper quartile has 75% below and 25% above. Hence these values divide the data into four, equal-sized, parts. It is important to note that there is no unique definition for all cases – you will see a discussion of this in Semester 2.

As before, a graph often conveys a clearer picture. Here the appropriate graphical summary is the boxplot. This gives a visual representation of the 5-figure summary and is very useful for checking symmetry and for comparing multiple datasets. Again, there are different definitions, although all give similar general information, and hence we must make sure we understand exactly the definitions used in any particular case.



(a) Monthly data (b) Daily data

Figure 3: Boxplots of Lloyds Bank percentage return.

In these, the box joins the lower quartile, $Q1$, and the upper quartile, $Q3$, and hence gives an interval containing 50% of the data – defining the interquartile range ($IQR = Q3 - Q1$). The thicker horizontal line, within the box, indicates the median. Here, the whiskers are also shown which extend out to the most extreme values within an interval from $Q1 - 1.5 \times IQR$ to $Q3 + 1.5 \times IQR$ – with any other data values shown individually.

```
> x=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
MonthlyShareReturns.csv")
> returns=x$returns
> boxplot(returns,returns[-18])
[Output similar to Fig 3(a) above]

> x=read.csv("http://www1.maths.leeds.ac.uk/~robert/MATH1710/
DailyShareData.csv")
> boxplot(x$daily.return)
[Output similar to Fig 3(b) above]
```

R-*note. We have now seen many* R *commands, you are not yet expected to understand them but you should try them out over the next few weeks.*

Installing R and RStudio - Windows version

In this module, we will perform calculations, and other numerical and graphical tasks, using the R coding language – a free software application – which is used throughout the undergraduate statistics teaching in Leeds and in the professional data science world. We will be using the RStudio interface to R which helps organise files and output, and in fact many people base all their work within RStudio and might never use R directly.

Although R and RStudio are installed on all University cluster PCs, it is likely that you will want to work on your own PC/laptop and hence this first exercise leads you through the stages of installing these two software applications. Please note that versions are available for Mac and Linux, but these instructions assume you are using Windows.

Please note that at the time of writing, the latest version of R is 3.6.1, and for RStudio it is 1.2.1335. Note, however, that there may be later versions once you come to install them – don't worry this is very unlikely to make any difference.

Start by downloading a copy of R from the internet:

1. Open the webpage: `cran.r-project.org`
2. Click on "Download R for Windows" (under Download and install R).
3. Click on "base".
4. Click on "Download R 3.6.1 for Windows" and, when prompted, select save the file `R-3.6.1-win.exe` to your computer – it takes about 1 to 2 minute depending on the speed of your connection.
5. Once downloaded, select Run. You may need to "allow the app to make changes to your device" by selecting "Yes". Then, follow the on-screen instructions.
6. At the end of the procedure, which takes about 1-2 minutes, click Finish.
7. You may have both R i386 3.6.1 and Rx64 3.6.1 icons, but these can be safely removed once you have completed the next step, as we will use R only through RStudio.
8. Double-click the Rx64 3.6.1 icon to see that R launches correctly, then File/Exit (choosing "No" when asked "Save workspace image?").
9. If you wish, you can now remove the R desktop icons.

Now download RStudio from the internet:

1. Open the webpage: `rstudio.com/products/rstudio/download`
2. Click on "RStudio 1.2.1335 - Windows 7+ (64-bit)" under the heading "Installers for Supported Platforms" near the bottom of the page, and, when prompted, select save the file `RStudio-1.2.1335.exe` to your computer – it takes less than 1 minute.
3. Once downloaded, select Run. You may need to "allow the app to make changes to your device" by selection "Yes". Then, follow the on-screen instructions.
4. At the end of the procedure, which takes about 1 minute, click Finish.
5. Double-click the icon to see that RStudio launches correctly – you may need to locate the app in your Start Menu – then File/Quit Session to exit.
6. If you wish, create a shortcut on your desktop or pin to the Start Menu.

Please note that you may also want to get a copy of Microsoft Office, in particular you might want to use Word and Excel during this module. As a University of Leeds student you can get a copy free of charge – please see online details or ask at the `IT Service Desk` on Level 10 in EC Stoner.

2 Basic Probability

Background

Probability (or probability theory) is a branch of mathematics which rigorously describes uncertain (or random) systems and processes. It has its roots in the 16th/17th century with the work of Cardano, Fermat & Pascal.

The French mathematician and astronomer, Pierre Simon, Marquis de Laplace said,

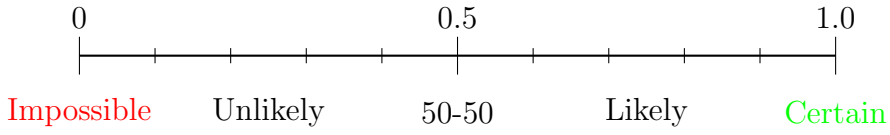
> "We see that the theory of probability is basically only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it.... "

Perhaps this is too extreme, but it is also an area of modern development and application such as: modelling heredity disease in genetics, pension calculations in actuarial science, stock pricing in finance, epidemic modelling in public health.

Put simply, probability measures the chance of some event occurring:

- probability 0 means the event is impossible,
- probability 1 means the event is certain.

The larger the probability, the more likely the event. The details, however, are much more complex as we will see.



2.1 Sample space and events

Let the sample space, Ω (the Greek letter capital "omega"), be the set of all possible outcomes of an experiment, and let ω (small "omega") be a single outcome, that is $\omega \in \Omega$. Then, let $|\Omega|$ (or $\#\Omega$) denote the number of possible outcomes.

An event, often denoted $A, B, C, \ldots$, is a set of outcomes of an experiment. The set can be empty, $A = \emptyset$, giving an impossible event, $Pr(\emptyset) = 0$, or can equal the sample space, $A = \Omega$, giving a certain event, $Pr(\Omega) = 1$. These extremes are not every interesting and so the event will usually be a non-empty, proper subset of the sample space.

Example 2.1: Experiment: Toss three standard coins.
$\Omega = \{(H,H,H),(H,H,T),(H,T,H),(T,H,H),(H,T,T),(T,H,T),(T,T,H),(T,T,T)\}$
with $|\Omega| = 8$.
Let $A = \{\text{There are at least two heads}\} = \{(H,H,H),(H,H,T),(H,T,H),(T,H,H)\}$
with $|A| = 4$ and so $Pr(A) = |A|/|\Omega| = 4/8 = 1/2$.

Example 2.2: Experiment: Roll two eight-sided dice.

$\Omega = \{(1,1),(1,2),\ldots,(8,8)\}$ with $|\Omega| = 64$.

Let $A = \{\text{The sum equals } 4\} = \{(1,3),(2,2),(3,1)\}$ with $|A| = 3$ and so $Pr(A) = |A|/|\Omega| = 3/64$.

Example 2.3: Experiment: Measure the height of a randomly selected student.

$\Omega = \mathbb{R}^+$ with $|\Omega| =$ "infinity"

Let $A = \{\text{Height more than 1.6m}\}$ with $|A| =$ "infinity"

Clearly the situation is different here, in that there are infinity many values to consider – more on this type of situation later in the module.

To assign a probability we might take a large number of students and consider the proportion with height greater than $1.6m$.

There are several ways to assign probability:

(i) counting — as in the first two examples above.

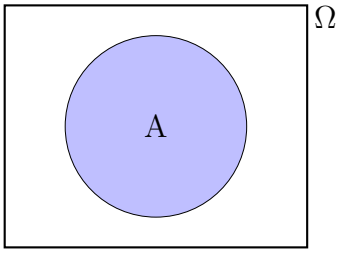
(ii) relative frequency, that is repeatedly performing an experiment under constant conditions — as in the third example.

(iii) subjectively — for when there is no argument of symmetry and where the experiment cannot be repeated. For example, England winning the next football world cup.

We will look more at these ideas later in the module.

The Venn diagram

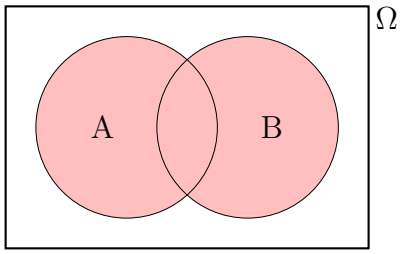
It is useful to show the relationships between events using Venn diagrams, such as the following.



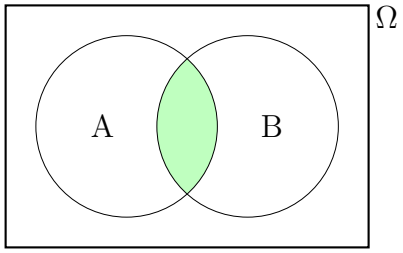
- Points represent outcomes, with
- the box representing the sample space, and
- the shaded region represents the outcomes in an event.

Operations with events

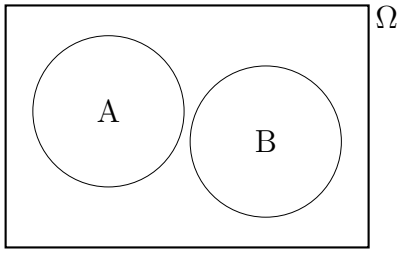
The union of A and B, written $A \cup B$ (say "A or B"), is the set of all outcomes belonging to at least one of the events A and B.



The intersection of A and B, written $A \cap B$ (say "A and B"), is the set of all outcomes belonging to both A and B.

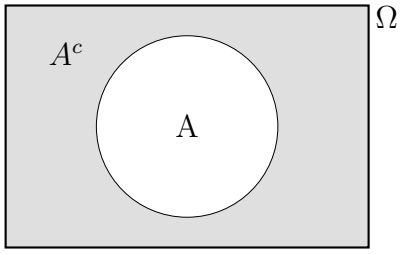


Events A and B are said to be mutually exclusive if they have no outcomes in common, then we write $A \cap B = \emptyset$, that is A and B cannot occur at the same time – we say that sets A and B are disjoint.



The complement of event A, written A^c (say "A complement") is the set of all outcomes which are not in A.

Note that $\Omega^c = \emptyset$ and $\emptyset^c = \Omega$, also $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$.



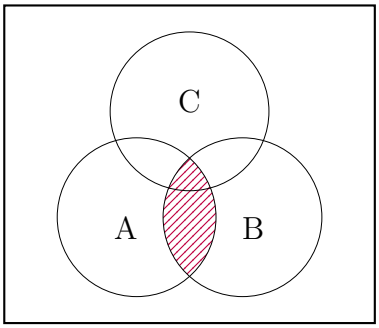
The operations of union and intersection can be further combined to give various set identities, for example:

$$\left.\begin{aligned} A \cup B &= B \cup A \\ A \cap B &= B \cap A \end{aligned}\right\} \text{Commutative laws}$$

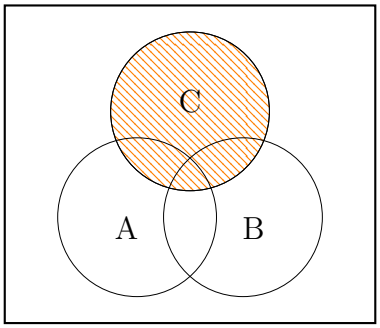
$$\left.\begin{aligned} A \cup (B \cup C) &= (A \cup B) \cup C \\ A \cap (B \cap C) &= (A \cap B) \cap C \end{aligned}\right\} \text{Associative laws}$$

$$\left.\begin{aligned} (A \cup B) \cap C &= (A \cap C) \cup (B \cap C) \\ (A \cap B) \cup C &= (A \cup C) \cap (B \cup C) \end{aligned}\right\} \text{Distributive laws}$$

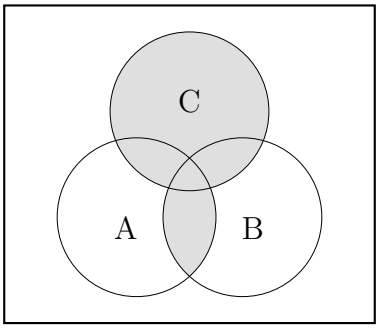
We can show, intuitively, that these laws are true by carefully constructing a set of Venn diagrams (but formal proof is more rigorous and powerful). As an example, consider the second distributive law. Drawing Venn diagrams of both left-hand side and right-hand side shows that the law is true. Starting with the left-hand side:



(a) $A \cap B$.

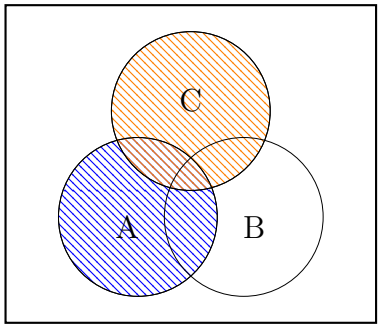


(b) C

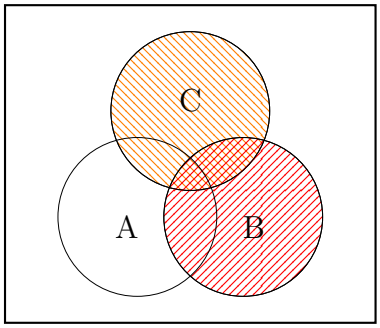


(c) $(A \cap B) \cup C$.

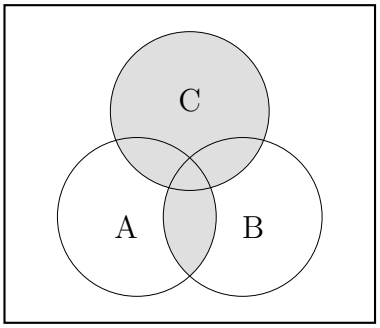
then the right-hand side:



(d) $A \cup C$.



(e) $B \cup C$



(f) $(A \cup C) \cap (B \cup C)$.

We see that the areas shaded in (c) and (f) are the same, and so we can claim that $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ is a true statement.

Now complete Worksheet 1 on Venn diagrams to check your understanding.

2.2 The axioms and basic rules of probability

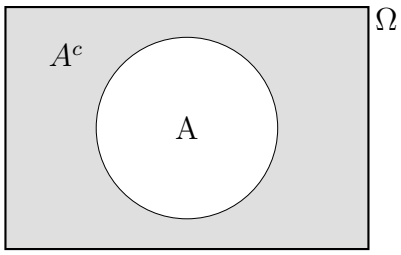
The (Kolmogorov) axioms of probability are:

- $Pr(A) \geq 0$ for any event A,
- $Pr(\Omega) = 1$ for any sample space Ω, and
- $Pr(A \cup B) = Pr(A) + Pr(B)$ for mutually exclusive events A and B (that is where $A \cap B) = \emptyset$).

Clearly, these are very basic statements, but they are sufficient to allow many complex rules to be derived.

Consider the proof of following (basic) rules:

(a) $Pr(A^c) = 1 - Pr(A)$.



Starting with the set relation

$$A \cup A^c = \Omega$$

then considering the probability of left and right

$$Pr(A \cup A^c) = Pr(\Omega)$$

and using K3 and K2

$$Pr(A) + Pr(A^c) = 1$$

leads to the required result

$$Pr(A^c) = 1 - Pr(A).$$

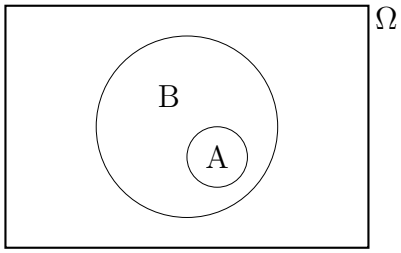
(b) $Pr(\emptyset) = 0$.

Start by noting that $\emptyset = \Omega^c$, then by result (a) above with $A = \Omega$ we have

$$Pr(\emptyset) = 1 - Pr(\Omega) \stackrel{\text{K2}}{=} 1 - 1 = 0, \quad \text{as required.}$$

Note that in the final step, K2 has been written over the equals sign to show that axiom K2 is needed.

(c) If $A \subseteq B$, then $Pr(A) \leq Pr(B)$.



Again, start with a set relation,

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

then using K3, since $A \cap (B \cap A^c) = \emptyset$, gives

$$Pr(B) = Pr(A) + Pr(B \cap A^c)$$

and since $Pr(B \cap A^c) \geq 0$ by K1 we get

$$Pr(B) \geq Pr(A), \quad \text{as required.}$$

A more important rule which can be derived from the axioms of probability is known as the addition rule for general events,

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B).$$

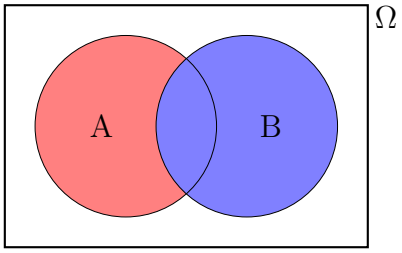
The key step is to realize that we can subdivide the events in the following two ways.

Starting with

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

then using K3, since $B \cap (A \cap B^c) = \emptyset$, gives

$$Pr(A \cup B) = Pr(B) + Pr(A \cap B^c). \qquad (1)$$

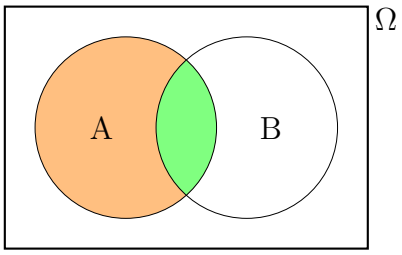


Also using the set relation,

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

so, using K3 with $(A \cap B) \cap (A \cap B^c) = \emptyset$, we get

$$Pr(A) = Pr(A \cap B) + Pr(A \cap B^c). \qquad (2)$$



Re-arranging (2) as $Pr(A \cap B^c) = Pr(A) - Pr(A \cap B)$ and substituting into (1) gives

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B), \quad \text{as required.}$$

If A and B are mutually exclusive, then $Pr(A \cap B) = 0$ and hence this result reduces to K3 – no contradiction. Hence K3 is refereed to as the addition rule for mutually exclusive events.

This result generalises to more that two events. For example consider three events, A, B and C, which gives the following result

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(B \cap C) - Pr(A \cap C) + Pr(A \cap B \cap C).$$

2.3 Assignment of probability

Classical probability for equally-likely events

The classical approach to assigning probability is to consider each member of the (finite) sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ to have equal probability. Then

$$Pr(\omega_i) = 1/N \quad \text{for } i = 1, 2, \ldots, N = |\Omega|.$$

Note that axioms K1 and K2 are clearly true. Now consider an event A, then using K3,

$$Pr(A) = \sum_{\omega_i \in A} Pr(\omega_i)$$

that is, the sum over all outcomes belonging to event A, which can be written

$$Pr(A) = |A|/|\Omega|$$

that is the number of outcomes in the event of interest divided by the number of events in the sample space – we saw examples of this earlier. So, in the classical approach, it is vital to be able to count the number of outcomes in events and in the sample space – often this involves permutations and combinations, a topic called combinatorics.

Probability as relative frequency and the Law of Large Numbers

Examples of truly equally-probable events are rare — the best examples are rolling dice and tossing coins. There is, however, a 1 in 6000 chance of a standard coin landing on its edge — this is less than 0.0002 and so perhaps can be ignored. More surprisingly, the expected 50-50, moves to 51-49 in favour of landing the same side up as started up. Is this enough to make an equal-probable assumption a bad approximation?

What about other events commonly thought of as being equally probable, such as birthdays or births of boy/girl babies? We will explore these situations later in the module.

Suppose that we have a simple situation but where an argument of symmetry, and hence equal probability, is not valid, such as a bent coin. If the coin were "fair" then p, the probability of `Heads`, is a half, $p = 1/2$, whereas if the coin is "biased", then $p \neq 1/2$.

Suppose we toss the coin n times and let S_n be the total number of `Heads`. So for the sequence: `Tails`, `Heads`, `Heads`, `Tails`, `Heads`, if we evaluated S_n after each toss, we would have $S_1 = 0$, $S_2 = 1$, $S_3 = 2$, $S_4 = 2$, $S_5 = 3$.

To estimate of the probability of `Heads` we can use the relative frequency of `Heads`, $\hat{p}_n = S_n/n$ and see this as depending on the number of tosses so far. Which, for the above sequence gives: $\hat{p}_1 = 0/1 = 0$, $\hat{p}_2 = 1/2$, $\hat{p}_3 = 2/3$, $\hat{p}_4 = 2/4 = 1/2$, $\hat{p}_5 = 3/5$.

The *Law of Large Numbers* says that

$$\hat{p}_n = \frac{S_n}{n} \quad \text{tends to} \quad p \qquad \text{as } n \text{ tends to } \infty.$$

That is to say, as the number of tosses grows the random *relative frequency* of `Heads` tends to a non-random value p, i.e. the *probability* of `Heads`.

A more technical explanation goes as follows. The convergence is understood in the sense that it becomes increasingly unlikely to observe even small deviations of the random quantity $\hat{p}_n$ from the value p, that is, for arbitrary small $a > 0$

$$Pr\left(|\hat{p}_n - p| > a\right) \to 0 \qquad \text{as } n \to \infty.$$

The Law of Large Numbers is also true in a stronger form, in that the convergence holds with probability 1. In other words, while successively calculating the frequencies S_n/n for a particular realisation of events, you may be sure that in the long run this random quantity will converge to p.

Subjective assignment of probability

The classical approach to assigning probability works well when we can exploit symmetry, but such cases are rare. Whereas the *frequentist*, also called *objectivist*, notion of probability is more general but relies on being able to endlessly repeat the experiment, and hence is idealistic and elusive. In particular, we may only have resources to repeat the experiment a few hundred times, or even only a few times, and it may not be possible to keep the conditions constant. Further, what can be said about similar, but no-identical, events or more importantly about events that are intrinsically unrepeatable, e.g. that England will win the Rugby World Cup in 2019?

An alternative view of probability is the *subjectivist*. For the subjectivist, probabilities only exist in the mind. Should I take the bus to work, or will I get there more quickly if I cycle? I can make a decision if I have some idea of the probability that I get to work more quickly when I cycle. If I randomly choose to cycle or go by bus to work every day, I will build up a fund of experience upon which to make the probabilistic assessment, which might depend on the weather, time of year, how I'm feeling today, etc. But even if this is my first day at this workplace, I can still make some kind of probabilistic assessment based on my general experience. A subjectivist can assign probabilities to events even though the experiment might never have been performed. What probability would you assign to England winning the Rugby World Cup in 2019?

Once you have completed the Essential directed reading on Combinatorics, then complete Worksheet 2 on Probability Notation and Worksheet 3 on More Probability to check your understanding.

Combinatorics

Basic definitions

The multiplication principle says that if an experiment has k stages with n_1 possible outcomes at the first stage, n_2 at the second, $\cdots$, and n_k at the k-th stage, then the total number of possible outcomes of the experiment is

$$|\Omega| = n_1 \times n_2 \times \cdots \times n_k.$$

This principle also allows us to breakdown events into stages and to count the number of outcomes in the event as a product of the outcomes of the stages.

Suppose we have n distinct objects, then the total number of ordered arrangements, or permutations, is

$$P_n = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = n!$$

and we say "n factorial".

Example 2.4: How many different arrangements of the letter a, b and c are possible? Here $n = 3$ and so there are $P_3 = 3! = 3 \times 2 \times 1 = 6$ permutations.

Suppose now that we only select r of the n objects and permute these. Then the number of permutations of r objects selected from n objects is

$${}^nP_r = n \times (n-1) \times \cdots (n-r+1) = \frac{n!}{(n-r)!}$$

and we say "n perm r" or "n-p-r".

Example 2.5: Suppose 20 people enter a 100m race. How many ways are there of distributing a Gold, Silver and Bronze? Here $n = 20$ and $r = 3$, and so we have ${}^{20}P_3 = 20!/(20-3)! = 20 \times 19 \times 18 = 6840$.

Suppose now that the ordering of the r selected objects is not important (only that we have selected them), then the total number of combinations of r objects selected from n objects is

$${}^nC_r = \frac{n!}{r!(n-r)!}$$

and we say "n choose r " or "n-c-r".

We can see that this is correct by imagining a two-stage procedure in which we first select the r objects from the n objects and then we permute these selected objects. Let the number of ways of completing the first stage be given the symbol ${}^nC_r$, and we know that once selected there are $r!$ ways of permuting the r objects. This gives a total number of ways of ${}^nC_r\, r!$ This two-stage process is equivalent to, in one stage, permuting r objects selected from n, and hence the number of ways must be equal. The number of ways of permuting r objects selected from n is ${}^nP_r = n!/(n-r)!$. Since these number must be equal

$$ {}^nP_r = \frac{n!}{(n-r)!} = {}^nC_r\, r! $$

which can be re-arranged to give

$$ {}^nC_r = \frac{n!}{r!(n-r)!} $$

as required.

Example 2.6: A football captain has to choose four other players to complete his team from a squad of 20. How many possible combinations are possible? With $n = 20$ and $r = 4$ we have

$$ \frac{20!}{4!(20-4)!} = \frac{20 \times 19 \times 18 \times 17}{4 \times 3 \times 2 \times 1} = 4845. $$

Suppose now that we are again interested in permutations of n objects, but that there are r of one type and $(n-r)$ of a second type. Then, the number of such permutations is

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

Notice that this is the same as ${}^nC_r$ above, even though it is describing a different situation, and we still say "n choose r". We shall see later in the module, that these also arise when dealing with the binomial distribution, and hence are sometimes called binomial coefficients.

We can show that this expression is correct by again considering the procedure divided into stages. In the first stage we permute the objects and let the number of distinguishable orderings be given the symbol $\binom{n}{r}$. If we now suppose that the objects are distinguishable, then there are $r!$ ways to rearrange the r objects and $(n-r)!$ ways of rearranging the $(n-r)$ objects. Hence the total number of rearrangements is $\binom{n}{r} r!\,(n-r)!$. This procedure is equivalent to the single-stage process of permuting all objects, assuming they are distinguishable, which can be done in $n!$ ways. Hence

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

which can be rearranged to give

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

as required.

If any of these ideas are new, or you would like further information, then take a look at the pages on "Counting Methods" in the recommended textbook: Clarke GM and Cooke D, A Basic Course in Statistics.

3 Conditional Probability and independence

3.1 Introduction

Sometimes we need to change probabilities in light of additional information.

Example 3.1: Consider data on the height of 89 first year students.

	<178cm	≥178cm	
Male	27	24	51
Female	33	5	38
	60	29	89

The probability that an individual is at least 178cm is $29/89 = 0.326$.
If we are now told that the individual is:

- Male, then the probability changes to $24/51 = 0.471$, and
- Female, then the probability changes to $5/38 = 0.132$.

Examples of events which might modify probabilities:

A patient has a certain illness	A test is negative
A political party wins an election	Opinion poll results
Your soccer team win	Play at home

3.2 Definitions

The conditional probability of event A occurring given (conditional on) event B having definitely occurred is defined by:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} \qquad [\text{provided } Pr(B) > 0]\,.$$

This is sometimes thought of as the "fourth axiom".

Example 3.1: (cont.) If $A = \{\text{Height at least 178cm}\}$ and $B = \{\text{Female}\}$, then, as before,

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{5/89}{38/89} = \frac{5}{38} = 0.132.$$

Notice that here $Pr(A|B) < Pr(A)$ (recall that $Pr(A) = 29/89 = 0.326$), and so we say that event B is unfavourable for event A.

If we already have the conditional probability of A given B, and the probability of B, then we can find the probability that both occur together using the multiplication rule which is a rearrangement of the definition of conditional probability

$$Pr(A \cap B) = Pr(A|B)Pr(B).$$

In some examples we are interested in the probabilities of intersections, but it is (much) easier to first evaluate conditional probabilities.

If we consider a sequence of n events, $A_1, A_2, \ldots, A_n$, then the generalisation is

$$Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = Pr(A_1)Pr(A_2|A_1)Pr(A_3|A_2 \cap A_1) \cdots Pr(A_n|A_{n-1} \cap \cdots \cap A_1).$$

Typically, these events may be related to successive stages of an experiment.

3.3 Independent events

This is a very important idea in general statistics, not just when considering events as here. For some pairs of events

$$Pr(A|B) > Pr(A) \quad \text{and then also} \quad Pr(B|A) > Pr(B) \qquad \begin{array}{l} A \text{ is favourable for } B \\ B \text{ is favourable for } A \end{array}$$

or

$$Pr(A|B) < Pr(A) \quad \text{and then also} \quad Pr(B|A) < Pr(B) \qquad \begin{array}{l} A \text{ is unfavourable for } B \\ B \text{ is unfavourable for } A. \end{array}$$

Sometimes, however, we have

$$Pr(A|B) = Pr(A) \quad \text{and then also} \quad Pr(B|A) = Pr(B).$$

in this case A and B are said to be independent. That is, knowing that one of these events has definitely occurred does not change the likelihood that the other will occur.

Definition: Two events are statistically independent if and only if

$$Pr(A \cap B) = Pr(A)Pr(B).$$

Comments:

1. This expression comes from a re-arrangement of the definition of conditional probability. It gives a single, symmetric equation which works even if $Pr(A)$ or $Pr(B)$ is zero.

2. Do not confuse independence with mutually exclusive! If events are mutually exclusive then $A \cap B = \emptyset$ and so $Pr(A \cap B) = 0$, whereas for independence $Pr(A \cap B) = Pr(A)Pr(B)$ which, in general, is not zero (only is $A = \emptyset$ or $B = \emptyset$, or both).

3. In a practical situation, that is when considering data, we can compare the relative frequencies and look for approximate equality.

4. If we know that the events are physically independent, then we can evaluate $Pr(A \cap B)$ using the product $Pr(A)Pr(B)$; or we can use this as a test for independence by evaluating $Pr(A \cap B)$and $Pr(A)Pr(B)$ separately and comparing the values.

5. Statistical independence concerns only "balance of probabilities" – knowing that B has occurred does not change the probability of A occurring. Physical independence, however, is a stronger situation requiring no change in the number of outcomes.

   For example, consider the roll of a fair six-sided die. Then, $\Omega = \{1, 2, , , 5, \}$. Let $A = \{2, , 6\}$, $B = \{1, 2\}$ and then $A \cap B = \{2\}$. So, $Pr(A) = 1/2$, $Pr(B) = 13/$ and $Pr(A \cap B) = 1/6$ and hence A and B are statistically independent.

Physical independence implies statistical independence, but statistical independence does not imply physical independence. If we use the above as a test, then if the condition holds we can only claim statistical independence, whereas if the condition does not hold then the events are neither statistically nor physically independent.

Comment:

Three events A, B and C, are pairwise independent is and only is

$$\begin{aligned} Pr(A \cap B) &= Pr(A)Pr(B) \\ Pr(B \cap C) &= Pr(B)Pr(C), \text{ and} \\ Pr(A \cap C) &= Pr(A)Pr(C). \end{aligned}$$

They are (completely) independent if and only if they are pairwise independent and

$$Pr(A \cap B \cap C) = Pr(A)Pr(B)Pr(C).$$

Now complete Worksheet 5 on conditional probability and independence to check your understanding.

3.4 Theorem of total probability and Bayes' theorem

Example 3.2: Suppose that there are two biased coins with probability of heads $p_1 = 0.3$ and $p_2 = 0.9$. In an experiment, a coin is selected at random and tossed once. What is the probability of getting {`Heads`}? Suppose that the outcome is {`Heads`}, then what is the probability that Coin 1 was selected?

Total probability formula

Suppose that we are interested in the probability of some event A, but that it is not easy to evaluate $Pr(A)$ directly.

Also, suppose that there are a set of events $B_1, B_2, \ldots, B_k$ which partition the sample space, and that we can easily find $Pr(A|B_1), \ldots, Pr(A|B_k)$ and $Pr(B_1), \ldots, Pr(B_k)$.

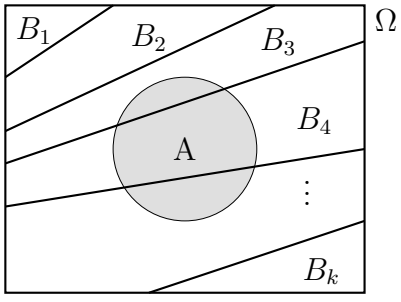
Then, we can evaluate the probability of A as

$$\begin{aligned} Pr(A) &= Pr(A|B_1)Pr(B_1) + Pr(A|B_2)Pr(B_2) + \cdots + Pr(A|B_k)Pr(B_k) \\ &= \sum_{j=1}^{k} Pr(A|B_j)Pr(B_j). \end{aligned}$$

For $B_1, \ldots, B_k$ to be a partition of Ω, these events must be: (i) mutually exclusive (disjoint sets), that is $B_i \cap B_j = \emptyset (i \neq j)$ and (ii) exhaustive for Ω, that is $B_1 \cup B_2 \cup \cdots \cup B_k = \Omega$.

Proof
We know that $A = A \cap \Omega$ and that $B_1 \cup \cdots \cup B_k = \Omega$ hence

$$\begin{aligned} A &= A \cap (B_1 \cup B_2 \cup \cdots \cup B_k) \\ &= (A \cap B_1) \cup \cdots \cup (A \cap B_k). \end{aligned}$$



Then

$$\begin{aligned} Pr(A) &= Pr\left((A \cap B_1) \cup \cdots \cup (A \cap B_k)\right) \\ &= Pr(A \cap B_1) + \cdots + Pr(A \cap B_k) \\ &= Pr(A|B_1)Pr(B_1) + \cdots + Pr(A|B_k)Pr(B_k) \\ &= \sum_{j=1}^{k} Pr(A|B_j)Pr(B_j). \end{aligned}$$

Example 3.3: Suppose a bag contains 3 red balls and 5 green balls. We remove two balls, what is the probability that the second ball is red?

Let $A = \{\text{Second ball is red}\}$ and $B_1 = \{\text{First ball is red}\}$, $B_2 = \{\text{First ball is green}\}$. Now, $Pr(B_1) = 3/8$ and $Pr(B_2) = 5/8$, but also $Pr(A|B_1) = 2/7$ and $Pr(A|B_2) = 3/7$.

Hence

$$Pr(A) = Pr(A|B_1)Pr(B_1) + Pr(A|B_2)Pr(B_2) = \frac{2}{7} \times \frac{3}{8} + \frac{3}{7} \times \frac{5}{8} = \frac{3}{8}.$$

Note that $Pr(\text{First ball is red}) = Pr(\text{Second ball is red})$.

If we want the probability that both balls are red, then we evaluate $Pr(A \cap B_1) = Pr(A|B_1)Pr(B_1) = (2/7) \times (3/8) = 3/28$.

Bayes' rule

Suppose we have a conditional probability, such as $Pr(A|B)$, but we are interested in the probability of the events conditioned the other way, that is $Pr(B|A)$, then

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)} \qquad \text{when } Pr(A) > 0.$$

Proof

By the definition of conditional probability

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)} = \frac{Pr(A \cap B)}{Pr(A)} \qquad \text{by symmetry}$$

then using the multiplication rule

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

In general, let $B_1, B_2, \ldots, B_k$ be a partition of Ω (as before), then

$$Pr(B_i|A) = \frac{Pr(A|B_i)Pr(B_i)}{Pr(A)}, \qquad i = 1, 2, \ldots, k.$$

Notice that $Pr(A)$ can be evaluated using the previous total probability formula, so that

$$Pr(B_i|A) = \frac{Pr(A|B_i)Pr(B_i)}{\sum_{j=1}^{k} Pr(A|B_j)Pr(B_j)}, \qquad i = 1, 2, \ldots, k.$$

Example 3.4: Suppose that a computer spam filter is 90% effective at flagging spam emails, but will also incorrectly flag 5% of safe emails. Suppose that 1 in 100 emails are spam.

Let $S = \{\text{Email is spam}\}$ and $F = \{\text{Flagged as spam}\}$, then $Pr(F|S) = 0.90$, $Pr(F|S^c) = 0.05$ and $Pr(S) = 0.01$.

What is the probability that a new email is flagged as spam?

$$Pr(F) = Pr(F|S)Pr(S) + Pr(F|S^c)Pr(S^c) = 0.90 \times 0.01 + 0.05 \times 0.99 = 0.0584.$$

Further, suppose that an email is flagged, then what is the probability that it is a spam email?

$$Pr(S|F) = \frac{Pr(F|S)Pr(S)}{Pr(F)} = \frac{0.90 \times 0.01}{0.0584} = 0.1538.$$

Note that the probability of an unflagged email being spam is

$$Pr(S|F^c) = \frac{Pr(F^c|S)Pr(S)}{Pr(F^c)} = \frac{(1 - Pr(F|S))Pr(S)}{1 - Pr(F)} = \frac{0.1 \times 0.01}{0.9406} = 0.0011.$$

Example 3.5: Suppose a fair die is rolled twice.

(i) Let A denote the event that the sum is 9, and
B denote the event that the second toss is an even value.

Now $A = \{(3,6),(4,5),(5,4),(6,3)\}$ and hence $Pr(A) = 4/36$, $Pr(B) = 1/2$, and $Pr(A \cap B) = 2/36$ since $A \cap B = \{(3,6),(5,4)\}$.

Then here $Pr(A)Pr(B) = (4/36) \times (1/2) = 2/36 = Pr(A \cap B)$ hence A and B are statistically independent.

(ii) Let C denote the event that both are odd, and D that their sum is 8.

We have $Pr(C) = 9/36$, $Pr(D) = 5/36$ and $Pr(C \cap D) = 2/36$.

So, since $Pr(C)Pr(D) = (9/36) \times (5/36) = 5/144 \neq 2/36 = Pr(C \cap D)$ then C and D are not independent.

Example 3.6: Consider the occurrence of colour blindness (CB) which is carried by the X-chromosome.

About 10% of X chromosomes are faulty and so males (who have one X chromosome) have a probability of 0.1 of being colour blind. Whereas females (with two X chromosomes) need two faulty chromosomes and so have a $(0.1)^2$ chance of being colour blind. In the UK population about 52% are male and 48% female.

So the overall probability of selecting a colour blind person is

$$\begin{aligned} Pr(CB) &= Pr(CB|Male)Pr(Male) + Pr(CB|Female)Pr(Female) \\ &= 0.1 \times 0.52 + (0.1)^2 \times 0.48 = 0.0568. \end{aligned}$$

Note that $\{Male\}$ and $\{Female\}$ are events which partition the sample space, and that $\{colour - blind\}$ is some other event.

Now suppose we have selected a person at random who is definitely colour blind. What can we say about the chances that they are Male or Female?

$$Pr(Male|CB)\frac{Pr(CB|Male)Pr(Male)}{Pr(CB)} = \frac{0.1 \times 0.52}{0.0568} = 0.915$$

and

$$Pr(Female|CB)\frac{Pr(CB|Female)Pr(Female)}{Pr(CB)} = \frac{(0.1)^2 \times 0.48}{0.0568} = 0.085.$$

So the colour blind person is 10 times more likely to be Male than Female.
Note that, of course $Pr(Male|CB) + Pr(Female|CB) = 1$ as the events $\{Male\}$ and $\{Female\}$ are complementary.

Example 3.7:
Suppose that two cards are chosen at random from a standard pack of 52 playing cards.

Let H_1 and H_2 denote the events that a heart is selected on the first and second choice respectively.

Now clearly, for the first selection we have,

$$Pr(H_1) = 13/52 \text{ and } Pr(H_1^c) = 39/52$$

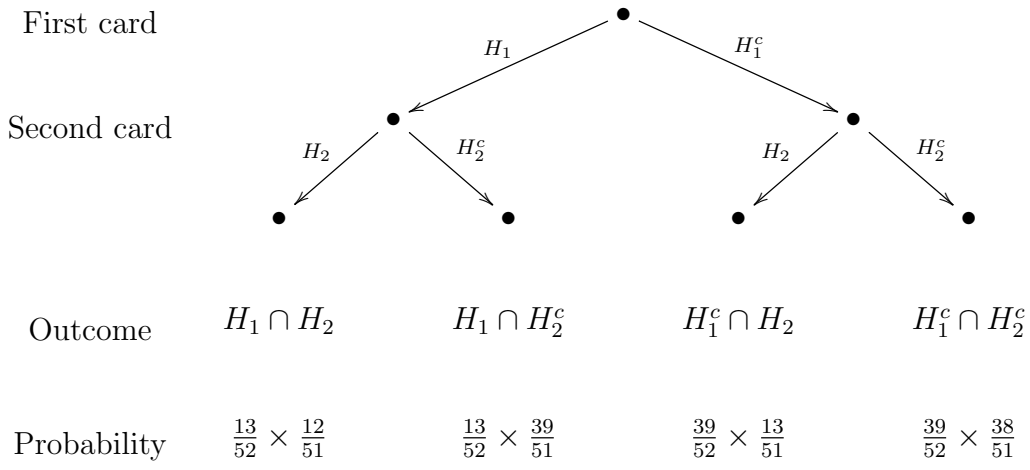
and for the second selection (given the first)

$$Pr(H_2|H_1) = 12/51 \text{ and } Pr(H_2|H_1^c) = 13/51$$
$$Pr(H_2^c|H_1) = 39/51 \text{ and } Pr(H_2^c|H_1^c) = 38/51.$$

Suppose we require $Pr(\text{both are hearts})$ that is $Pr(H_2 \cap H_1)$, then

$$Pr(H_2 \cap H_1) = Pr(H_2|H_1)Pr(H_1) = \frac{12}{51} \times \frac{13}{52} = \frac{1}{17}.$$

This type of approach can be illustrated using a tree diagram.



Outcome	$H_1 \cap H_2$	$H_1 \cap H_2^c$	$H_1^c \cap H_2$	$H_1^c \cap H_2^c$
Probability	$\frac{13}{52} \times \frac{12}{51}$	$\frac{13}{52} \times \frac{39}{51}$	$\frac{39}{52} \times \frac{13}{51}$	$\frac{39}{52} \times \frac{38}{51}$

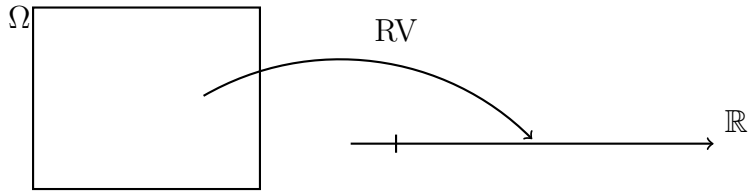
It is useful to check that these probabilities sum to 1.

4 Random variables

4.1 Basic definitions

Statistical models often describe the experimental outcome in terms of probability statements. In particular, a random variable is defined along with a distribution which might depend on a parameter. For example, let random variable X record the number of `Heads` when a coin is tossed. Then, $X = 1$ corresponds to `Heads`, with probability p, and $X = 0$ to `Tails`, with probability $1 - p$. In this example, p is the parameter of the model, which might be unknown. Tossing the coin many times and observing the outcomes would give us a better idea of the value of p. Hence, to learn about parameter values, we need to collect data. Then, taken together, our model and estimated parameter value give us an approximation to reality which can then be used to describe the current situation and to make predictions about the future.

Formally, a random variable is "a function which maps elements of the sample space onto the set of real numbers". Informally, it is a numerical value which summarizes a random experiment by measuring some property of interest.



It is common to use the letters X, Y, Z to denote random variables. The set of all possible values which can be taken by the random variable is called the range space.

A discrete random variable X, say, has a finite, or countably infinite, range space which is denoted $\Omega_X = \{x_1, x_2, \ldots\}$ and the corresponding probabilities are written as

$$Pr(\{X = x_i\}) = p_X(x_i) = p_i \qquad i = 1, 2, \ldots$$

The possible values and corresponding probabilities is called the probability mass function (pmf) and is often shown in a table. Note that the probability mass function must satisfy the Axioms of Probability, hence

$$0 \leq p_X(x_i) \leq 1 \qquad \text{for all } x_i \in \Omega_X$$

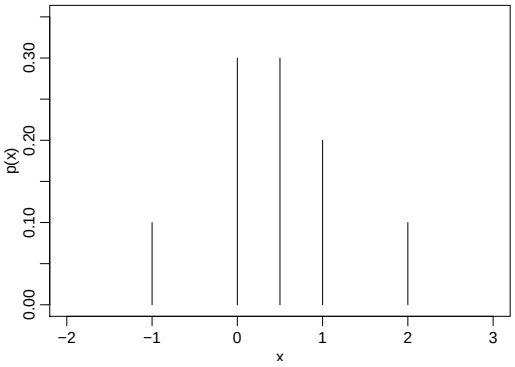
and

$$\sum_{x_i \in \Omega_X} p_X(x_i) = 1$$

where the sum is over all elements of the range space.

Example 4.1: The following is an example probability mass function, with $\Omega_X = \{-1, 0, 0.5, 1, 2\}$,

x	-1	0	0.5	1	2
$p_X(x)$	0.1	0.3	0.3	0.2	0.1



The graph is produced in R using the command `plot` with an option `type="h"`.

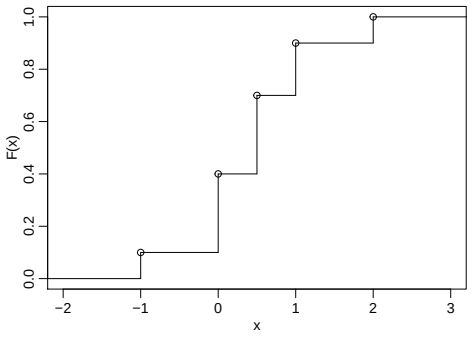
The cumulative distribution function (cdf) is defined as

$$F_X(x) = Pr(\{X \leq x\}) = \sum_{x_i \leq x} Pr(\{X = x_i\}).$$

Note that this function is defined for all real numbers, and not just at the possible values. This means that for discrete random variable the cdf is a *step function*.

Example 4.1: (Cont.) Consider the earlier probability mass function, which gives the following cumulative distribution function.

$$F_X(x) = \begin{cases} 0.0 & x < -1 \\ 0.1 & -1 \leq x < 0 \\ 0.4 & 0 \leq x < 0.5 \\ 0.7 & 0.5 \leq x < 1 \\ 0.9 & 1 \leq x < 2 \\ 1.0 & 2 \leq x \end{cases}$$



The graph is produced in R using the commands `cumsum`, `stepfun`, and `plot`.

4.2 Expected value and variance

Definition: The expectation, or expected value, of random variable X is defined as

$$E[X] = \sum_{i=1}^{N} x_i \, p_X(x_i).$$

This is a weighted average of the possible values. The quantity $E[X]$ is also called the "mean of X" and is sometimes denoted using the symbol μ (say "mu").

Example 4.1: (Cont.) Consider again the above random variable, then the expectation can be evaluated as

$$E[X] = (-1) \times 0.1 + 0 \times 0.3 + 0.5 \times 0.3 + 1 \times 0.2 + 2 \times 0.1 = 0.45.$$

Note that, although we call it the *expected value*, in fact for many random variables, such as the one above, the actual expected value can never occur. Instead we should think of it as the long-term average.

Properties of expectation

(E1) The expectation of a constant is the constant itself, if c is a constant then $E[c] = c$.

(E2) Expectation is a *linear operator*, that is if both a and b are constants, then

$$E[aX + b] = aE[X] + b.$$

A simple case of this is when $Y = 3 - 2X$, and then $E[Y] = E[3 - 2X] = 3 - 2E[X]$.

Proof: To see this let $Y = aX + b$, and then note that $\{Y = y_i\}$ where $y_i = ax_i + b$ and $\{X = x_i\}$ are equivalent events and hence $p_y(y_i) = p_X(x_i)$. Then

$$\begin{aligned} E[Y] &= \sum y_i p_Y(y_i) = \sum (ax_i + b) p_X(x_i) \\ &= a \sum x_i p_X(x_i) + b \sum p_X(x_i) = aE[X] + b \end{aligned}$$

where the last step uses the definition of the expectation of X and the second Axiom.

(E3) Consider a collection of random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$. Let $Y = \sum_{j=1}^{n} c_j X_j$, that is Y is a linear combination of the X_j, then

$$E[Y] = \sum_{j=1}^{n} c_j E[X_j].$$

A simple case of this is when $Y = X_1 - 2X_2$, and then $E[Y] = E[X_1 - 2X_2] = E[X_1] - 2E[X_2]$. Again, we see the linear properties of expectation.

Example 4.2:
In October the average monthly rainfall in Leeds is 2.3 inches with average minimum daily temperature of 11^oC and average maximum daily temperature of 18^oC.

Suppose we want these in cm (1 inch equal 2.5cm) and ^oF ($\mathtt{T}_F = \frac{9}{5} \times \mathtt{T}_C + 32$).

The expected rainfall is $2.3 \times 2.54 = 5.84$cm.
The expected minimum temperature is $\frac{9}{5} \times 11 + 32 = 51.8^oF$.
The expected maximum temperature is $\frac{9}{5} \times 18 + 32 = 64.4^oF$.

Estimation of parameters using the expectation Many models will contain unknown parameters and one aim of a statistical analysis is to use data to say something about likely values of the parameter — this process is called *estimation* or *inference*. One of the simplest cases is the unknown probability of `Heads` in a biased coin, but we will see many more examples later in the module. The basic idea is to use the data, $\underline{x} = (x_1, x_2, ..., x_n)$, to make a "good guess" at the numerical value of the parameter. Let the parameter be called θ (say "theta"), then an *estimate*, $\hat{\theta} = \hat{\theta}(\underline{x})$ is a numeric value which is a function of the data — different datasets lead to different estimates. The simplest approach to estimation is to choose the value of the parameter so that the theoretical mean is equal to the sample mean.

Example 4.3: Let random variable X represent the number of `Heads` when a coin is toss once with $Pr(\{X = 1\}) = p$ and $Pr(\{X = 0\}) = 1 - p$, and hence p is the unknown parameter. Now, the expectation is given by $E[X] = 1 \times p + 0 \times (1 - p) = p$.

Also, let $\underline{x} = (x_1, \ldots, x_n)$ be a corresponding dataset obtained by tossing the coin n times, with $\bar{x}$ being the sample mean. Hence, the estimate is simply given as $\hat{p} = \bar{x}$.

Suppose a sequence of 10 tosses yields $\underline{x} = (1, 0, 0, 1, 1, 1, 0, 1, 0, 1)$, then $\bar{x} = 0.6$ hence $\hat{p} = \bar{x} = 0.6$.

R commands: `x=c(1,0,0,1,1,1,0,1,0,1)` and then `phat = mean(x)`

Variance of a random variable

The mean (or expectation) gives a "typical" or "representative" value, $E[X] = \mu$ for random variable X, but it is also of interest to know about variation around the mean. We might imagine looking at the expected value of deviations of the random variable from the mean, but this is useless, as $E[X - \mu] = E[X] - \mu = 0$, hence, instead, we consider the expected squared deviation.

Definition: The variance of random variable X is defined as

$$Var(X) = E[(X - E[X])^2]$$

and the (positive) square-root is called the standard deviation, $SD(X) = \sqrt{Var(X)}$.

In practice we usually evaluate the variance using the equivalent expression

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

Proof: Starting with the definition and then multiplying the square

$$Var(X) = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2]$$

using the standard properties of expectation

$$= E[X^2] - 2E[X]E[X] + E[X]^2.$$

Finally, collecting terms together, gives

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

Example 4.1: (Cont.) Consider again the earlier random variable with $E[X] = 0.45$, then we need

$$E[X^2] = (-1)^2 \times 0.1 + 0^2 \times 0.3 + 0.5^2 \times 0.3 + 1^2 \times 0.2 + 2^2 \times 0.1 = 0.775$$

giving

$$Var(X) = E[X^2] - \{E[X]\}^2 = 0.775 - \{0.45\}^2 = 0.5725.$$

The first step can be easily calculated in R using `sum(xvals^2*probs)`, with `xvals` and `probs` as before.

Properties of variance

(V1) The variance of a constant is zero, $Var[c] = 0$.
(V2) For constants a and b, we have

$$Var[aX + b] = a^2 Var[X].$$

Proof Recall, with $Y = aX + b$, that $E[Y] = aE[X] + b$, then

$$\begin{aligned} Var(Y) &= E[(Y - E[Y])^2] \\ &= E\left[(\{aX + \not{b}\} - \{aE[X] + \not{b}\})^2\right] \\ &= E[(aX - aE[X])^2] = a^2 E[(X - E[X])^2] = a^2 Var(X). \end{aligned}$$

(V3) For (independent) random variables $X_1, \ldots, X_n$ and constants $c_1, \ldots, c_n$, and with $Y = \sum_{j=1}^{n} c_i X_i$, then

$$Var[Y] = \sum_{i=j}^{n} c_j^2 Var[X_j].$$

Notes:

- Variance is unaffected by an additive shift in the random variable, but a multiplicative scaling has a quadratic effect – compare to the linear properties of expectation.
- In property (V3) above, we see that an additional condition of independence was included. Two (or more) random variable are said to be independent if the value of one tells us nothing about the value of the other. If the random variables relate to physically separate experiments then we can assume independence, but otherwise we cannot.

Example 4.3: (cont.) For the eight-sided dice example earlier we saw that $E[X] = 14/3$ and $Var[X] = 47/9$.

Consider $Z = 6 - 3X$ then

$$E[Z] = E[6 - 3X] = 6 - 3E[X] = 6 - 3 \times \frac{14}{3} = -8$$

and

$$Var[Z] = Var[6 - 3X] = (-3)^2 Var[X] = 9 \times \frac{47}{9} = 47.$$

Now suppose that we also have a standard six-sided die which has $E[Y] = 7/2$ and $Var[Y] = 35/12$ Then, the variance of the sum of the two dice is

$$Var[X + Y] = Var[X] + Var[Y] = \frac{47}{9} + \frac{35}{12} = \frac{879}{108} = 8.14.$$

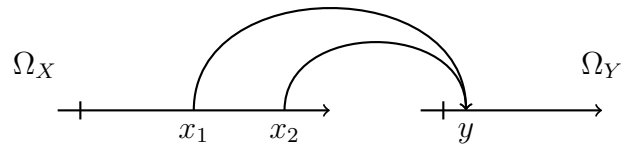
Note that here, X and Y are physically independent and so using this result for independent random variables is valid.

Functions of random variables

We have seen how to find the expectation and variance of linear functions, e.g. $E[aX+b] = aE[X]+b$. Now we shall see what to do for other general functions, $Y = g(X)$, such as s^X or e^{tX} (for constants s and t).

The function $y = g(x)$ maps points in the range space of X, Ω_X, to points in the range space of Y, Ω_Y. Hence, if $\Omega_X = \{x_1, x_2, \ldots\}$, then we can determine the range space of Y as the distinct elements of $\{g(x_1), g(x_2), \ldots\}$.

Note that as more than one element of Ω_X might map to the same point in Ω_Y, then Ω_Y cannot contain more elements than Ω_X. Hence if Ω_X is finite or countable infinite then so is Ω_Y which also means that if X is discrete then so is Y. If X is a continuous random variable, then usually Y will also be continuous but there is no guarantee, for example if $Y = \text{integer}(X)$. Hence, we should consider each case carefully.



We can then evaluate the probability mass function of Y by transferring probabilities using the idea of *equivalent events*.

Consider each of the elements of the range space of Y in turn; for element y from Ω_Y:

- if $\{Y = y\}$ implies that $\{X = x\}$, then
$$p_Y(y) = p_X(x)$$
- if $\{Y = y\}$ implies that $\{X \in (x_1, x_2, \ldots)\}$, then
$$p_Y(y) = p_X(x_1) + p_X(x_2) + \cdots = \sum_{x:g(x)=y} p_X(x).$$

Let us now check that the axioms are still valid:

Since $Pr(Y = y) = p_Y(y)$ is the sum of at least one $p_X(x)$ then since $p_X(x) > 0$ for all $x \in \Omega_X$, then $p_Y(y) > 0$ for all $y \in \Omega_Y$. Hence axiom (K1) is valid.

Also, starting with $Pr(\Omega_X) = \sum_x p_X(x) = 1$, then consider $Pr(\Omega_Y) = \sum_y p_Y(y) = \sum_y \sum_{x:g(x)=y} p_X(x) = \sum_x p_X(x) = 1$. Hence axiom (K2) is valid.

This can be illustrated as follows. Suppose that n_1 of the elements of Ω_X map to the first element of Ω_Y, n_2 map to the second, etc. until n_m map to the final element of Ω_Y. Then after possible reordering and relabelling we have

$x_1, \ldots, x_{n_1}$ map to y_1, and hence $g(x_1) = \cdots = g(x_{n_1}) = y_1$
$x_{n_1+1}, \ldots, x_{n_1+n_2}$ map to y_2, and hence $g(x_{n_1+1}) = \cdots = g(x_{n_1+n_2}) = y_2$
$\vdots$
$x_{n_1+\cdots+n_{m-1}}, \ldots, x_{n_1+\cdots+n_m}$ map to y_m, and so $g(x_{n_1+\cdots+n_{m-1}1}) = \cdots = g(x_{n_1+\cdots+n_m}) = y_m$

If we now want $p_Y(y)$ then this can be obtained by adding all the probabilities of the corresponding x values, that is all x such that $g(x) = y$ giving $p_Y(y) = \sum_{x:g(x)=y)} p_X(x)$, for example $p_Y(y_1) = \sum_{x:g(x)=y_1)} p_X(x) = p_X(x_1) + \cdots + p_X(x_{n_1})$.

Further,

$$\begin{aligned}\sum_y p_Y(y) &= p_Y(y1) + p_Y(y_2) + \cdots + p_Y(y_m) \\ &= p_X(x_1) + \cdots + p_X(x_{n_1}) + p_X(x_{n_1+1}) + \cdots + p_X(x_{n_1+n_2}) \\ &\quad + p_X(x_{n_1+\cdots+n_{m-1}}) + \cdots + p_X(x_{n_1+\cdots+n_m})\end{aligned}$$

Example 4.4: If X has probability mass function:

X	-1	0	1
$p_X(x)$	0.2	0.7	0.1

Suppose we are interested in $Y = X^2$, then clearly $\Omega_Y = \{0, 1\}$ and as $\{Y = 0\} = \{X = 0\}$ then $p_Y(0) = p_X(0) = 0.7$, but as $\{Y = 1\} = \{X = -1\} \cup \{X = 1\}$ hence $p_Y(1) = p_X(-1) + p_X(1) = 0.3$. Notice that $E[Y] = \sum_y y p_(y) = 0 \times 0.7 + 1 \times 0.3 = 0.3$.

The law of the unconscious statistician

A simple and automatic approach uses the following theorem: if $Y = g(X)$, then

$$E[g(X)] = \begin{cases} \sum g(x) p_X(x) & \text{if } X \text{ is discrete,} \\ \\ \int g(x) f_X(x) & \text{if } X \text{ is continuous.} \end{cases}$$

We have already seen examples of this with $E[X^2]$ and $E[X(X-1)]$, but there are many others. For example, $E[s^X]$ is known as the probability generating function and is particularly useful for deriving many theoretical results regarding discrete random variables.

Proof:

Starting with the definition of expectation of Y,

$$E[g(X)] = E[Y] = \sum_y y p_Y(y) = \sum_y y \sum_{x:g(x)=y} p_x(x) = \sum_y \sum_{x:g(x)=y} y p_x(x)$$

using $p_Y(y) = \sum_{x:g(x)=y} p_x(x)$ and replacing y by the equivalent numerical value $g(x)$

$$= \sum_y \sum_{x:g(x)=y} g(x) p_x(x)$$

and noting that the double sum can be replaced by a single sum

$$= \sum_x g(x) p_x(x).$$

Probability generating functions

We have already seen several discrete random variables (including Bernoulli, binomial, geometric, Poisson) and their corresponding probability mass functions. To help with the derivation of theoretical results, the same information can also, and sometimes more conveniently, be summarized by the probability generating functions (pgf). The pgf also has many special properties which make it more useful than the probability mass function.

For a discrete random variable, X, (taking only integer values) the probability generating function is defined as

$$G_X(s) = p_X(0)s^0 + p_X(1)s^1 + p_X(2)s^2 + \cdots = \sum_{x=0}^{\infty} s^x p_X(x)$$

that is

$$G_X(s) = E\left[s^X\right]$$

where s is an arbitrary variable. It is useful to note that each probability multiplies the corresponding power of s.

Example 4.5: Suppose that X described the outcome of the roll of a fair die, so $\Omega_X = \{1, 2, 3, 4, 5, 6\}$ and $p_X(x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$.

So for the probability generating function we have,

$$G_X(s) = \frac{1}{6}s^1 + \frac{1}{6}s^2 + \cdots + \frac{1}{6}s^6 = \frac{1}{6}\left(s + s^2 + \cdots + s^6\right).$$

If we want to find the expectation consider the following

$$G'_X(s) = \frac{dG_X(s)}{ds} = \frac{d}{ds}\sum_x p_X(x)s^x = \sum_x x p_X(x)s^{x-1}$$

and so,

$$G'_X(1) = \left.\frac{dG_X(s)}{ds}\right|_{s=1} = \sum_x x p_X(x) = E[X].$$

Example 4.5: (cont.) In this example we have

$$\frac{dG_X(s)}{ds} = \frac{1}{6}\left(1 + 2s + \cdots + 6s^5\right)$$

and with $s = 1$ we get

$$G'_X(1) = \frac{1}{6}\left(1 + 2 + \cdots + 6\right) = \frac{7}{2}$$

and so $E[X] = 7/2$.

Also,

$$G''_X(s) = \frac{dG^2_X(s)}{ds^2} = \frac{d^2}{ds^2}\sum_x p_X(x)s^x = \sum_x x(x-1)p_X(x)s^{x-2}$$

and so,

$$G''_X(1) = \left.\frac{d^2G_X(s)}{ds^2}\right|_{s=1} = \sum_x x(x-1)p_X(x) = E[X(X-1)].$$

Hence we can find the variance as

$$Var(X) = E[X^2] - \{E[X]\}^2 = E[X(X-1)] + E[X] - \{E[X]\}^2 = G''_X(1) + G'(1) - \{G'_X(1)\}^2$$

Example 4.5: (cont.) In this example we have

$$G''_X(1) = \frac{1}{6}\left(2 \times 1 + 3 \times 2s \cdots + 6 \times s^4\right)$$

and with $s = 1$ we get

$$G''_X(1) = \frac{1}{6}\left(2 + 6 + 12 + 20 + 30\right) = \frac{70}{6}.$$

Hence we have

$$Var(X) = \frac{70}{6} + \frac{7}{2} - \left\{\frac{7}{2}\right\}^2 = \frac{35}{12}.$$

Example 4.6: Suppose X follows a Bernoulli distribution, that is with $Pr(X = 1) = p$ and $Pr(X = 0) = 1 - p$.

Then

$$G_X(s) = (1-p)s^0 + ps^1 = 1 - p + ps$$

so

$$G'_X(s) = p \quad \text{hence} \quad G'_X(1) = p$$

and

$$G''_X(s) = 0 \quad \text{hence} \quad G''_X(1) = 0.$$

Hence we have

$$E[X] = p$$

and

$$Var(X) = G''_X(1) + G'_X(1) - \{G'_X(1)\}^2 = 0 + p - p^2 = p(1-p).$$

4.3 Joint distributions

In practice we are generally interested in analysing data from several random variables. For discrete variables, X and Y say, the joint probability mass function is defined as

$$p_{X,Y}(x_i, y_j) = Pr(X = x_i \text{ and } Y = y_j), \quad \text{for } (x_i, y_j) \in \Omega_{X,Y} = \Omega_X \times \Omega_Y$$

with $p_{X,Y}(x_i, y_j) \geq 0$ for all x_i and y_j and $\sum\sum p_{X,Y}(x_i, y_j) = 1$.

Further, we can define marginal probability mass functions by summing over one variable,

$$p_X(x_i) = \sum_{\text{all } y_j} p_{X,Y}(x_i, y_j) \quad \text{and} \quad p_Y(y_j) = \sum_{\text{all } x_i} p_{X,Y}(x_i, y_j).$$

Thus, to obtain a marginal pmf, we sum over the "unwanted" variable in the joint pmf.

Example 4.3: Consider the following table of probabilities corresponding to two discrete random variables X and Y.

		Values of X		
		$x_1 = 0$	$x_2 = 1$	$p_Y(y_j)$
Value of Y	$y_1 = 1$	0.30	0.27	0.57
	$y_2 = 2$	0.37	0.06	0.43
	$p_X(x_i)$	0.67	0.33	1.00

The joint probabilities are in the centre of the table with the marginal probabilities in the right-hand column and bottom row — the margins.

The next set of definitions are for the conditional probability mass functions

$$p_{X|Y}(x_i|y_j) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)} \quad \text{and} \quad p_{Y|X}(y_j|x_i) = \frac{p_{X,Y}(x_i, y_j)}{p_X(x_i)}.$$

These are essentially the same as the previous definitions of conditional probability.

Example 4.3: (cont) Examples of conditional probability mass functions are:

x_i	0	1
$p_{X\|Y}(x_i\|2)$	$\frac{0.37}{0.43} = 0.86$	$\frac{0.06}{0.43} = 0.14$

y_j	1	2
$p_{Y\|X}(y_j\|0)$	$\frac{0.30}{0.67} = 0.45$	$\frac{0.37}{0.67} = 0.55$

The two random variables are statistically independent if, and only if,

$$p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j), \qquad \text{for all } x_i \text{ and } y_j.$$

To check independence we need to check that this holds for every pair (x_i, y_j). To prove that X and Y are not independent, it is sufficient to find one counter-example.

We can evaluate (joint) expectations using the definition

$$E[g(X,Y)] = \sum\sum g(x_i, y_j)\; p_{X,Y}(x_i, y_j).$$

For example, and to use later, with $g(X,Y) = XY$ we have $E[XY] = \sum\sum x_i y_j\; p_{X,Y}(x_i, y_j)$.

We can also evaluate conditional expectations and variance using the definitions

$$\mathrm{E}[X|Y = y_j] = \sum\sum x_i\; p_{X|Y}(x_i|y_j) \quad \text{and} \quad \mathrm{Var}[X|Y = y_j] = E[(X - E[X])^2|Y = y_j].$$

There are similar definitions for $E[Y|X = x_i]$ and $\mathrm{Var}[Y|X = x_i]$. As elsewhere, an alternative formula for variance is $\mathrm{Var}[X|Y = y_j] = E[X^2|Y = y_j] - \{E[X|Y = y_j]\}^2$ where $E[X^2|Y = y_j] = \sum\sum x_i^2\; p_{X|Y}(x_i|y_j)$.

If we require marginal expectation and variances, then we can simply use the marginal probability mass function in the usual way.

To measure the degree of (linear) correlation between X and Y we can evaluate the correlation defined as

$$\mathrm{Cor}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}$$

where the covariance between X and Y is defined as

$$\mathrm{Cov}[X,Y] = E\left[(X - E[X])(Y - E[Y])\right].$$

As with the definitions of variance, it is often easier to perform calculations using the alternate formula, $\mathrm{Cov}[X,Y] = E[XY] - E[X]E[Y]$.

Suppose that we are interested in some linear function of the random variables, $Z = aX + bY + c$ with constants a, b and c. If X and Y are two general random variables, that is they are not known to be independent, then the formula for expectation is

$$E[Z] = E[aX + bY + c] = aE[X] + bE[Y] + c.$$

This is exactly the same situation as for independent random variables. To evaluate the variance, however, the following equation is needed

$$\mathrm{Var}[aX + bY + c] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\;\mathrm{Cov}[X,Y].$$

Note that, if X and Y are known to be independent, then it can be shown that $\mathrm{Cov}[X,Y] = 0$ and hence this reduces to the result seen earlier. Beware, however, that correlation zero does not mean the variables are independent.

Example 4.3: (cont) In this example, from the marginal distributions, we have $E[X] = 0.33$, $E[Y] = 1.43$, $\text{Var}[X] = 0.2211$ and $\text{Var}[Y] = 0.2451$.

It is clear that X and Y are not independent as, for example, $p_{X,Y}(0,1) = 0.30$ yet $p_X(0)p_Y(1) = 0.67 \times 0.57 = 0.3819$, that is they are not equal.

To find the covariance, and hence the correlation, we evaluate

$$\begin{aligned} E[XY] &= \sum\sum x_i y_j \; p_{X,Y}(x_i, y_j) \\ &= 0 \times 1 \times 0.30 + 0 \times 2 \times 0.37 + 1 \times 1 \times 0.27 + 1 \times 2 \times 0.06 = 0.39. \end{aligned}$$

Then, $\text{Cov}[X,Y] = E[XY] - E[X]E[Y] = -0.0819$ and hence

$$\text{Cor}[X,Y] = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{-0.0819}{\sqrt{0.2211 \times 0.2451}} = -0.3518.$$

5 Models for Count Data

Introduction

The world is now awash with data. Commercial and public-sector organisations increasingly depend on massive databases to refine their business operations. You may have thought, before starting this module, that statistics is all about calculating summary statistics such as means, standard deviations and correlations, and perhaps plotting pie charts and histograms of variables in a dataset. To be sure, these things are statistics, but they are not the subject of Statistics. Statistics is the art and science of learning about the real world through data and probability models.

Models can take many forms, but a key property is that they are approximations of reality. The famous statistician George Box said: "Essentially, all models are wrong, but some are useful" — though later this was re-expressed elsewhere as "all models are approximations".

5.1 Bernoulli trials and related distributions

We have discussed how random variables can be used to summarize the outcome of random experiments. We shall continue this by looking at distributions which arise from repeated experiments.

Suppose a single trial is performed which has just two possible outcomes, for example

- a tossed coin is `Heads` or `Tails`,
- a child is a boy or a girl,
- a learner passes their driving test or not.

Let random variable X describe this situation, taking value 1 if the event of interest occurs and value 0 if it does not occur, with

$$Pr(X = 1) = p \text{ and } Pr(X = 0) = 1 - p.$$

Note that, alternatively, we can write these in a single expression

$$Pr(X = x) = p_X(x) = p^x(1 - p)^{1-x}, \qquad x = 0, 1.$$

Here, X is called a Bernoulli random variable and such a random experiment is called a Bernoulli trial. Now suppose that we repeat this trial many times, but we keep all conditions constant. That is the trials are independent and the probability is fixed.

Example 5.1: Consider a Call Centre, and suppose we call five randomly chosen phone numbers and note whether the call was answered or not. Let A_i denote the event that the call was answered on the ith call, and let $Pr(A_i) = p$ for $i = 1, \ldots, 5$. It is assumed that calls are answered independently.
We might observe the sequence $A_1 A_2^c A_3 A_4 A_5^c$ (say) from the $2^5 = 32$ possible sequences. The probability of this particular sequence is

$$Pr(A_1 A_2^c A_3 A_4 A_5^c) = p \times (1-p) \times p \times p \times (1-p) = p^3(1-p)^2.$$

If we are only interested in the number of calls answered, then the above is only one of ten possible ways of getting three A's and two A^c's, and so

$$Pr(3A \text{ and } 2A^c) = 10 \times p^3(1-p)^2.$$

We can generalise this idea into the following result.

The binomial distribution

Let X be the number of "successes" in n repeated independent Bernoulli trials. Possible values for X are $0, 1, \ldots, n$ giving a range space $\Omega_X = \{0, 1, \ldots, n\}$. The probability mass function of X is

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n.$$

Note that $\binom{n}{x} = n!/(x!(n-x)!)$ arises by considering the number of permutations of n objects, of which x are of one type and the remaining $(n-x)$ are of another type (see Directed Reading on Combinatorics).

We say that X is a binomial random variable and can write $X \sim B(n, p)$.

The following conditions give rise to the binomial distribution.

- A sequence of n independent trials (n fixed).
- Two possible outcomes at each trial (say "success" and "failure").
- Fixed probability of "success" at each trail.
- The random variable counts the number of "successes".

Example 5.2: A fair die is rolled ten times, what is the probability that exactly four rolls give "6"?

Let X be the number of times "6" appears out of the ten rolls, then $X \sim B(n = 10, p = 1/6)$ and we require the probability

$$Pr(X = 4) = p_X(4) = \binom{10}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 = 0.0543.$$

This is easily done in R using the `dbinom` command, as `dbinom(4,10,1/6)`.

Unfortunately, it is not possible to produce a simple equation to give the cumulative distribution function. Instead, we would need to evaluate the individual probabilities, add them together and present the results as a table.

Example 5.2: (cont.)

First the probability mass function:

x	0	1	2	3	4	5	6	7	8	9	10
$p_X(x)$	0.16	0.32	0.29	0.16	0.05	0.01	0.00	0.00	0.00	0.00	0.00

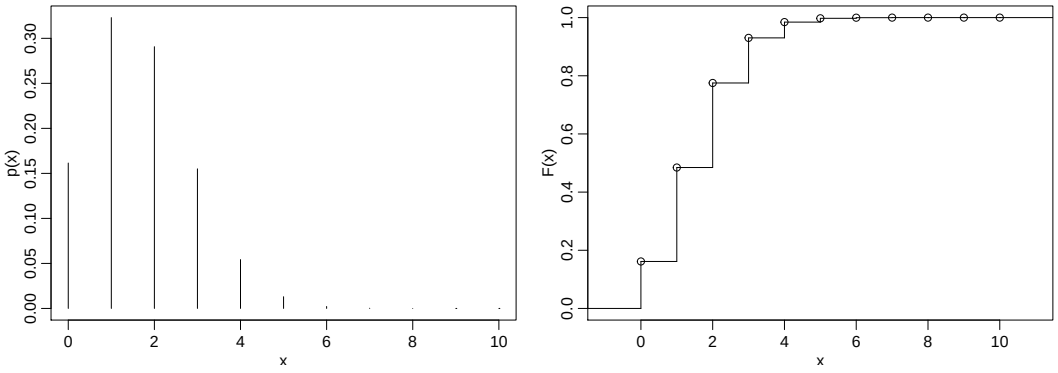
Then, the corresponding cumulative distribution function:

$F_X(x)$	0	0.16	0.48	0.77	0.93	0.98	1.00	$\cdots$
for $x \in$	$(-\infty, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$[5, 6)$	$\cdots$

These calculations use the commands: `round(dbinom(0:10,10,1/6), 2)` for the pmf and `round(cumsum(dbinom(0:10,10,1/6)), 2)` for the cdf.

Example 5.2: (cont.)

The probability mass function of $X \sim B(n = 10, p = 1/6)$ is shown below-left, and the corresponding cumulative distribution function on the right.



These graphs are produced in R using a combination of the commands `dbinom`, `cumsum`, `plot`, and `stepfun` – see R for MATH1710 Lesson 6.

Expectation and variance of the binomial: If $X \sim B(n, p)$ then

$$E[X] = np, \quad \text{and} \quad Var[X] = np(1-p).$$

Example 5.2: (cont.)

In the previous example, we had $X \sim B(10, 1/6)$ and so $E[X] = np = 10 \times \frac{1}{6} = \frac{5}{3}$ and $Var[X] = np(1-p) = 10 \times \frac{1}{6} \times \frac{5}{6} = \frac{50}{36} = \frac{25}{18}$.

Proof: Starting with the definition of expectation and using the probability mass function for a binomial distribution we have

$$E[X] = \sum_{x \in \Omega_X} x\, p_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

Then, the first term in the sum (with $x = 0$) is zero, also in the general term we have

$$x\binom{n}{x} = x\,\frac{n!}{x!(n-x)!} = x\,\frac{n(n-1)!}{x(x-1)!(n-x)!} = n\,\frac{(n-1)!}{(x-1)!(n-x)!} = n\binom{(n-1)}{(x-1)}$$

and so

$$E[X] = np \sum_{x=1}^{n} \binom{(n-1)}{(x-1)} p^{x-1}(1-p)^{n-x}.$$

Now, relabelling with $r = x - 1$, say, we get

$$E[X] = np \sum_{r=0}^{n-1} \binom{(n-1)}{r} p^r (1-p)^{(n-1)-r}.$$

which is the probability mass function of $B(n-1, p)$ so sums to 1, hence we get the required result

$$E[X] = np.$$

Next, let us consider the $E[X(X-1)]$ hence

$$E[X(X-1)] = \sum_{x \in \Omega_X} x(x-1)\, p_X(x) = \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x (1-p)^{n-x}.$$

Then, the first and second terms in the sum (with $x = 0$ and $x = 1$) are zero, also in the general term we have

$$\begin{aligned} x(x-1)\binom{n}{x} &= x(x-1)\, \frac{n!}{x!(n-x)!} = x(x-1)\, \frac{n(n-1)!}{x(x-1)(x-2)!(n-x)!} \\ &= n(n-1)\, \frac{(n-2)!}{(x-2)!(n-x)!} = n(n-1) \binom{(n-2)}{(x-2)} \end{aligned}$$

and so

$$E[X(X-1)] = n(n-1)p^2 \sum_{x=2}^{n} \binom{(n-2)}{(x-2)} p^{x-2} (1-p)^{n-x}.$$

Now, relabelling with $r = x - 2$, say, we get

$$E[X(X-1)] = n(n-1)p^2 \sum_{r=0}^{n-1} \binom{(n-1)}{r} p^r (1-p)^{(n-1)-r}.$$

which is the probability mass function of $B(n-2, p)$ so sums to 1, hence we get the result

$$E[X(X-1)] = n(n-1)p^2.$$

To obtain the required result note that $E[X(X-1)] = E[X^2] - E[X]$ hence $E[X^2] = E[X(X-1)] + E[X]$. Recalling that $E[X] = np$, we then get $E[X^2] = n(n-1)p^2 + np$ and, finally, this leads to $Var[X] = E[X^2] - \{E[X]\}^2 = n(n-1)p^2 + np - \{np\}^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$, as required.

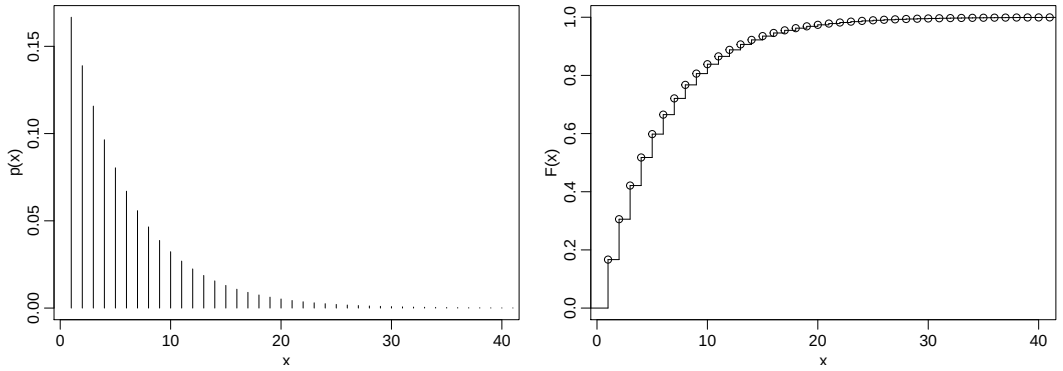
Geometric distribution

Suppose that we repeat independent Bernoulli trials until we get a "success". Let X be the number of failures before the first success. Possible values for X are $0, 1, \ldots$ (with no upper limit) giving $\Omega_X = \{0, 1, \ldots\}$ and its probability mass function

$$p_X(x) = Pr\,(x \text{ failures, then 1 success}) = (1-p)^x p, \qquad x = 0, 1, \ldots$$

This is called the geometric distributions and is denoted $X \sim Ge(p)$. The same conditions are valid as for the binomial, except that n is no longer fixed and the definition of the random variable has changed.

Example 5.3: Now suppose that a fair die is rolled until a six is shown and let X be the total number of tosses needed hence $X \sim Ge(p = 1/6)$. The probability mass function is shown below-left, and the corresponding cumulative distribution function on the right.



These graphs are produced in R using a combination of the commands `dgeom`, `cumsum`, `plot`, and `stepfun` – see R for MATH1710 Lesson 6.

Note that there are two versions of the geometric distribution. Compare to above the situation where we count the total number of trials, Y, which includes the failures but also include the final success – this is sometimes referred to as the shifted geometric. Here, possible values for Y are $1, 2, \ldots$ (with no upper limit) giving $\Omega_Y = \{1, 2, \ldots\}$ and its probability mass function

$$p_Y(y) = Pr\,((y-1) \text{ failures, then 1 success}) = (1-p)^{y-1} p, \qquad y = 1, 2, \ldots.$$

Note that $Y = X + 1$ and that $Pr(X = a) = Pr(Y = a + 1)$. Although here is no contradiction, we must be careful to ensure which version of the geometric is being used.

When evaluating cumulative probabilities, it is useful to recall the formula for the sum of a geometric series

$$S_n = 1 + r + \cdots + r^n = \sum_{k=0}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}$$

and when n is infinite, $S_\infty = 1/(1-r)$ (for $|r| < 1$). Further, when deriving the expectation and variance, we consider derivatives of this expression with respect to r giving

$$\sum_{k=1}^{\infty} k\, r^{k-1} = \frac{1}{(1-r)^2}, \quad \text{and} \quad \sum_{k=2}^{\infty} k(k-1)\, r^{k-2} = \frac{2}{(1-r)^3}, \qquad \text{for } |r| < 1.$$

Note that in these expressions the lower limit of summation has been adjusted to include only non-zero terms – but it would have been equally correct to leave $k = 0$.

Expectation and variance of the geometric: If $X \sim Ge(p)$ then

$$E[X] = (1-p)/p, \quad \text{and} \quad Var[X] = (1-p)/p^2.$$

Proof: Starting with the definition of expectation and using the probability mass function for a geometric distribution we have

$$E[X] = \sum_{x \in \Omega_X} x\, p_X(x) = \sum_{x=0}^{\infty} x\, (1-p)^x p = p \sum_{x=1}^{\infty} x\, (1-p)^x$$

then, replacing x by k and $1 - p$ by r and using one of the above results, we get

$$= p(1-p) \sum_{k=1}^{\infty} k\, r^{k-1} = p(1-p) \times \frac{1}{p^2} = \frac{1-p}{p}.$$

Next, let us consider the $E[X(X-1)]$ hence

$$E[X(X-1)] = \sum_{x \in \Omega_X} x(x-1)\, p_X(x) = \sum_{x=0}^{\infty} x(x-1)\, (1-p)^x p$$

$$= p(1-p)^2 \sum_{x=2}^{\infty} x(x-1)\, (1-p)^{x-2}$$

then, replacing x by k and $1 - p$ by r and using one of the above results, we get

$$= p(1-p)^2 \sum_{k=2}^{\infty} k(k-1)\, r^{k-2} = p(1-p)^2 \times \frac{2}{p^3} = \frac{2(1-p)^2}{p^2}.$$

To obtain the required result note that $E[X(X-1)] = E[X^2] - E[X]$ hence $E[X^2] =$

$E[X(X-1)]+E[X]$. Recalling that $E[X] = (1-p)/p$, we then get $E[X^2] = 2(1-p)^2/p^2 + (1-p)/p$ and, finally, this leads to $Var[X] = (1-p)/p^2$, as required.

Example 5.4: A fair coin is tossed until the first head is obtained. What is the probability that at least 3 tosses are needed?

Let X be the number of failures before the success, with $\Omega_X = \{0, 1, \ldots\}$, then we require $Pr(X \geq 2)$ where $X \sim Ge(1/2)$.

Rather than evaluating $Pr(X \geq 2) = Pr(X = 2) + Pr(X = 3) + \cdots$, instead consider the complementary event

$$\begin{aligned} Pr(X \geq 2) &= 1 - Pr(X < 2) = 1 - Pr(X = 0 \text{ or } X = 1) \\ &= 1 - \{Pr(X = 0) + Pr(X = 1)\} = 1 - \left\{ \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right) \right\} \\ &= 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

Further, the expectation and variance are:

$$E[X] = \frac{1 - 1/2}{1/2} = 1 \qquad \text{and} \qquad Var[X] = \frac{(1 - 1/2)}{(1/2)^2} = 2.$$

If, instead, we defined Y as the total number of tosses, with $\Omega_Y = \{1, 2 \ldots\}$, then we require $\Pr(Y \geq 3)$ which would give the same numerical value as above. In this situation, $\mathrm{E}[Y] = 2$ but the variance remains unchanged as $\mathrm{Var}[Y] = 2$. These can be verified from first principles, similar to the proof above, or by noting that $\mathrm{E}[Y] = \mathrm{E}[X + 1] = \mathrm{E}[X] + 1$ and $\mathrm{Var}[Y] = \mathrm{Var}[X + 1] = \mathrm{Var}[X]$.

5.2 Poisson distribution (the law of rare events)

Let X denote the number of events occurring (in some time interval or region in space) with known average rate, λ (say "lambda"), with probability mass function

$$Pr(X = x) = p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots,$$

then X is a Poisson random variable and we write $X \sim Po(\lambda)$. Note that here, the range space is (countably) infinite, $\Omega_X = \{0, 1, \ldots\}$.

As before, to evaluate the expectation and variance, knowing the relevant standard (Maclaurin) series result can be useful: $\sum_{k=0}^{\infty} x^k/k! = e^x$.

The distribution is named after French mathematician Siméon Poisson (1781-1840) and was made popular through application to the number of cavalrymen in the Prussian army killed by kicks from a horse – see, for example, Scheaffer and Young, 2010, pp 158-159.

Example 5.5: Suppose that the average number of students missing lectures due to flu virus in a week is $\lambda = 5.2$. What is the probability that in a randomly selected week there are fewer than 2 missing due to flu?

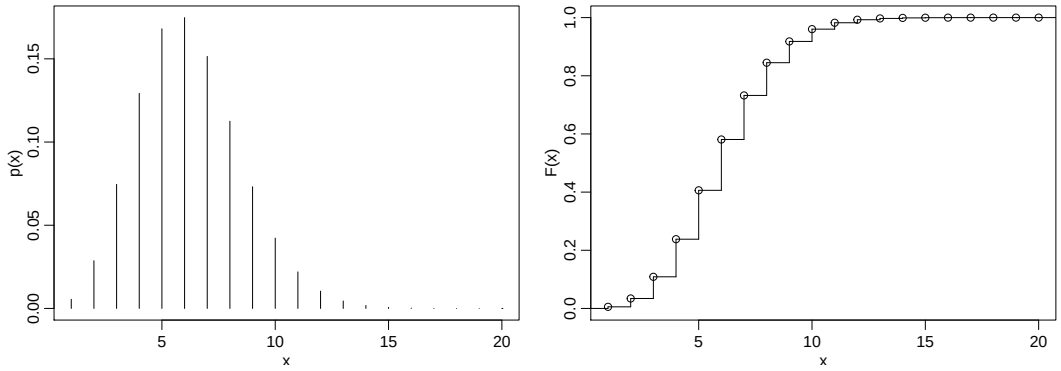
We have $X \sim Po(5.2)$ and we require $Pr(X < 2)$, so

$$\begin{aligned} Pr(X < 2) &= Pr(X = 0) + Pr(X = 1) \\ &= \frac{e^{5.2}(5.2)^0}{0!} + \frac{e^{5.2}(5.2)^1}{1!} = e^{5.2}\,(1 + 5.2) = 0.034. \end{aligned}$$

In this example you might have said that the binomial was a suitable model, with n the number of students and p the probability that a student gets flu. In fact, you are correct – see the next section. However, we may not know the values of n and p, but we might easily observe the average number missing.

Example 5.5: (cont.)

The probability mass function of a Poisson, $X \sim Po(\lambda = 5.2)$ is shown below-left, and the corresponding cumulative distribution function on the right.



These graphs are produced in R using a combination of the commands `dpois`, `cumsum`, `plot`, and `stepfun` – see R for MATH1710 Lesson 6.

Expectation and variance of the Poisson: If $X \sim Po(\lambda)$ then

$$E[X] = \lambda, \quad \text{and} \quad Var[X] = \lambda.$$

Proof: Starting with the definition of expectation and using the probability mass function for a geometric distribution we have

$$E[X] = \sum_{x \in \Omega_X} x\, p_X(x) = \sum_{x=0}^{\infty} x\, \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}$$

then, replacing $x - 1$ by k, we get

$$E[X] = \lambda \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = \lambda$$

since the sum is of a $Po(\lambda)$ over all possible values and hence equals 1.

Next, let us consider the $E[X(X-1)]$ hence

$$E[X(X-1)] = \sum_{x \in \Omega_X} x(x-1)\, p_X(x) = \sum_{x=0}^{\infty} x(x-1)\, \frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!}$$

then, replacing $x - 2$ by k, we get

$$E[X(X-1)] = \lambda^2 \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = \lambda^2$$

since the sum is of a $Po(\lambda)$ over all possible values and hence equals 1.

To obtain the required result note that $E[X(X-1)] = E[X^2] - E[X]$ hence $E[X^2] = E[X(X-1)] + E[X]$. Recalling that $E[X] = \lambda$, we then get $E[X^2] = \lambda^2 + \lambda$ and hence $Var[X] = E[X^2] - \{E[X]\}^2 = \lambda^2 + \lambda - \{\lambda\}^2 = \lambda$, as required.

Poisson approximation to the binomial

Suppose that $X \sim Po(\lambda)$ and $Y \sim B(n,p)$ with $\lambda = np$, then if p is small

$$Pr(Y = r) \to Pr(X = r), \quad \text{as } n \to \infty.$$

More usefully, we can say that for large n, and p small, that $Pr(Y = r) \approx Pr(X = r)$. In particular, the approximation is good if $n \geq 20$ and $p \leq 0.05$ and very good if $n \geq 100$ and $np \leq 10$.

Proof: See theoretical part of Example 4.10 in Stirzaker pp139-140.

We have, using $p = \lambda/n$,

$$Pr(Y = r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r}$$

which can be re-arranged to give

$$= \frac{n}{n} \times \frac{(n-1)}{n} \times \cdots \times \frac{(n-r+1)}{n} \, \frac{\lambda^r}{r!} \left(1 - \frac{\lambda}{n}\right)^{n-r}.$$

Now, $n \to \infty$ all the terms like $(n-r+1)/n \to 1$, and $\left(1 - \frac{\lambda}{n}\right)^{-r} \to 1$.

Now, what about the term $\left(1 - \frac{\lambda}{n}\right)^n$ as $n \to \infty$? Recall the binomial theorem: $(A+B)^n = \sum_{r=0}^{n} \binom{n}{r} A^r B^{n-r}$. Then with $A = -\lambda/n$ and $A = 1$ we have

$$\left(1 - \frac{\lambda}{n}\right)^n = \sum_{r=0}^{\infty} \binom{n}{r} \left(-\frac{\lambda}{n}\right)^r = \sum_{r=0}^{\infty} \frac{n}{n} \times \frac{(n-1)}{n} \times \cdots \times \frac{(n-r+1)}{n} \frac{(-\lambda)^r}{r!}$$

then following the same approach as above

$$= \sum_{r=0}^{\infty} \frac{(-\lambda)^r}{r!} = e^{-\lambda}.$$

Putting these various parts together gives

$$Pr(Y = r) = \frac{\lambda^r e^{-\lambda}}{r!} = Pr(X = r), \qquad r = 0, 1, \ldots$$

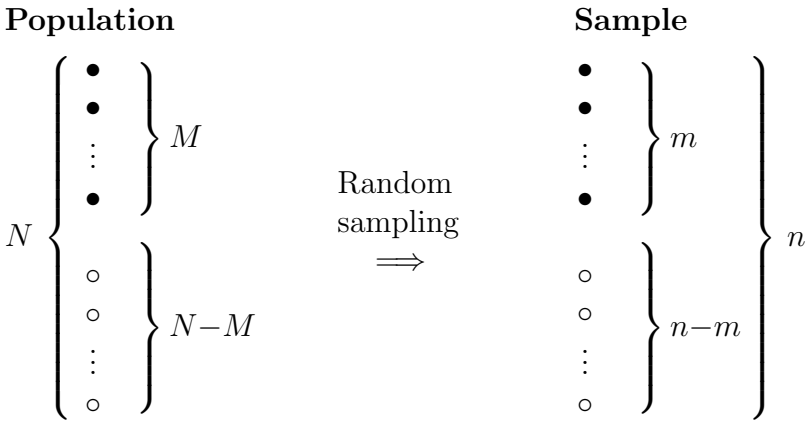
Poisson processes

Before moving on, it is worth noting another, and very important derivation of the Poisson distribution. Suppose that events occur *at random* throughout some period of time, or within some region, then the number of events occurring in the interval of time, or region, follows a Poisson distribution. As an example, suppose that on average an insurance company receives 120 claims per working day (12 hours), but that claims have equal chances of being made at any time through the day. Then, the actual number in a particular day follows a Poisson distribution with expected value 120, and that the number in a 1-hour period (say) follows a Poisson distribution with expected value 10. The process in time, or space, is called a Poisson process — you can see more of this in a Second Year module Markov Processes.

Now complete Worksheet 7 on standard discrete distributions to check your understanding.

Sampling from a finite population

Consider a bag of N balls, with M being black and the remaining $N - M$ being white. The experiment is to select a random sample of n balls from the bag of N. Let random variable X be the number of black balls in the sample. Suppose we want the probability of the event, A, that there are m black balls in the sample, that is $A = \{X = m\}$.



Before we can calculate the probability, we must know if the selected balls are returned to the population or not – this leads to: (I) sampling with replacement, and (II) sampling without replacement.

In (I) there are always the same N balls to choose from, and each is equally likely to be selected. So $|\Omega| = N \cdot N \cdots N = N^n$ and to calculate $|A|$ first consider a particular sequence of first m black and then $(n - m)$ white, once chosen these can be permuted to give $|A| = \binom{n}{m} M^m (N - M)^{n-m}$, hence

$$Pr(A) = \frac{|A|}{|\Omega|} = \frac{\binom{n}{m} M^m (N - M)^{n-m}}{N^n} = \binom{n}{m} \left(\frac{M}{N}\right)^m \left(1 - \frac{M}{N}\right)^{n-m}.$$

We can write $p = M/N$, which is the (unchanging) proportion of black balls in the population, and then

$$Pr(A) = \binom{n}{m} p^m (1 - p)^{n-m}.$$

This is the binomial probability formula seen earlier.

In (II), the selected ball is not replaced and so the number of balls changes at each stage.

Now, $|\Omega| = {}^N C_n$, as we are considering selecting n objects from N, and $|A| = {}^M C_m \, {}^{N-M} C_{n-m}$ as we require m from the M and $(n - m)$ from the $N - M$, hence

$$Pr(A) = \frac{{}^M C_m \, {}^{N-M} C_{n-m}}{{}^N C_n} = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}.$$

this is called the hyper-geometric probability formula.

If we let X be the number of black balls chosen then we can write $X \sim Hyp(n, M, N)$

and the probability mass function

$$p_X(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad \max(0, n+M-N), \ldots, \min(M, n)$$

and

$$E[X] = n\left(\frac{M}{N}\right) \qquad \text{and} \qquad Var(X) = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right).$$

Notice that, initially the proportion of black balls in the population is $p_0 = M/N$ and so

$$E[X] = np_0 \qquad \text{and} \qquad Var(X) = np_0(1-p_0)\left(\frac{N-n}{N-1}\right).$$

Comparing these to the mean and variance of the binomial we see that the expectation is equal, that is on average we obtain the same number of black balls, but that since $(N-n)/(N-1) < 1$, the variance of the hyper-geometric is smaller.

The value $(N-n)/(N-1)$ is called the finite population correction factor. Notice that for small sample sizes and large population sizes the factor is close to 1 (for example with $n = 10$ and $N = 1000$ then $(N-n)/(N-1) = 0.99$). In fact, when dealing with such situations we can use the binomial as a good approximation to the hyper-geometric – making calculations much easier.

Example 5.6: Consider sampling $n = 5$ students in a class of size $N = 50$ of whom $M = 28$ are male. What is the probability that all the sample are male?

Let $X = \{\text{The number of male students in the sample}\}$, then $X \sim Hyp(5, 28, 50)$ and we require

$$P_X(5) = \frac{\binom{28}{5}\binom{22}{0}}{\binom{50}{5}} = 0.04638562 = 0.0464 \text{ (4 dp)}.$$

If we had "forgotten" that we are sampling without replacement, then we would say that $X \sim B(n = 5, p = M/N = 0.56)$ and

$$p_X(5) = \binom{5}{5}(0.56)^5(1-0.56)^0 = 0.05507318 = 0.0551 \text{ (4 dp)}$$

which is considerably different to the correct value.

If we repeat these calculations for all maths students, with $N = 500$ and $M = 280$, then the exact hyper-geometric probability is $p_X(5) = 0.054$, whereas the binomial probability is unchanged at 0.055 – which is very close.

5.3 Additional Examples

Example 5.7: Suppose we roll an eight-sided die with sides labelled 1 to 8, where the even values are twice as likely as the odd numbers.

Let Y represent the outcome with probability mass function

Y	1	2	3	4	5	6	7	8
$p_Y(y)$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$

Now

$$E[Y] = \sum y\, p_Y(y)$$
$$= 1 \times \frac{1}{12} + 2 \times \frac{1}{6} + 3 \times \frac{1}{12} + 4 \times \frac{1}{6} + 5 \times \frac{1}{12} + 6 \times \frac{1}{6} + 7 \times \frac{1}{12} + 8 \times \frac{1}{6} = \frac{14}{3}.$$

Then, for the variance,

$$E[Y^2] = \sum y^2 p_Y(y)$$
$$= 1^2 \times \frac{1}{12} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{12} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{12} + 6^2 \times \frac{1}{6} + 7^2 \times \frac{1}{12} + 8^2 \times \frac{1}{6} = 27$$

giving

$$Var(Y) = E[Y^2] - \{E[Y]\}^2 = 27 - \left\{\frac{14}{3}\right\}^2 = \frac{47}{9}.$$

The first step can be easily calculated in R using `sum(yvals*probs)`, with `yvals=1:8` and `probs=c(1,2,1,2,1,2,1,2)/12`, and the second using `sum(yvals^2*probs)`.
Next consider $Z = 6 - 3X$ then

$$E[Z] = E[6 - 3Y] = 6 - 3E[Y] = 6 - 3 \times \frac{14}{3} = -8$$

and

$$Var[Y] = Var[6 - 3Y] = (-3)^2 Var[Y] = 9 \times \frac{47}{9} = 47.$$

Now suppose that we also have a standard six-sided die which has $E[X] = 7/2$ and $Var[X] = 35/12$ Then the expectation of the sum of the two dice is, $E[X + Y] = E[X] + E[Y] = \frac{14}{3} + \frac{7}{2} = \frac{49}{6}$ and the variance of the sum is

$$Var[X + Y] = Var[X] + Var[Y] = \frac{35}{12} + \frac{47}{9} = \frac{879}{108} = 8.14.$$

Note that here, X and Y are physically independent and so using this result for independent random variables is valid.

Important properties for sums of random variables

Suppose we have two independent random variables, X_1 and X_2, but are only interested in their sum, $Z = X_1 + X_2$, then

$$G_Z(s) = E[s^Z] = E[s^{X_1+X_2}] \overset{indep}{=} E[s^{X_1}]E[s^{X_2}] = G_{X_1}(s)G_{X_2}(s)$$

and the generalisation $Z = X_1 + X_2 + \cdots + X_n$ then

$$G_Z(s) = G_{X_1}(s)G_{X_2}(s)\cdots G_{X_n}(s).$$

Pgfs of standard distributions

Bernoulli	$1 - p + ps$
Binomial	$\{1 - p + ps\}^n$
Geometric	$ps/\left(1 - (1 - p)s\right)$
Poisson	$\exp\left(\lambda(s - 1)\right)$

Example 5.8: Suppose we are interested in the sum of n independent Bernoulli random variables, then

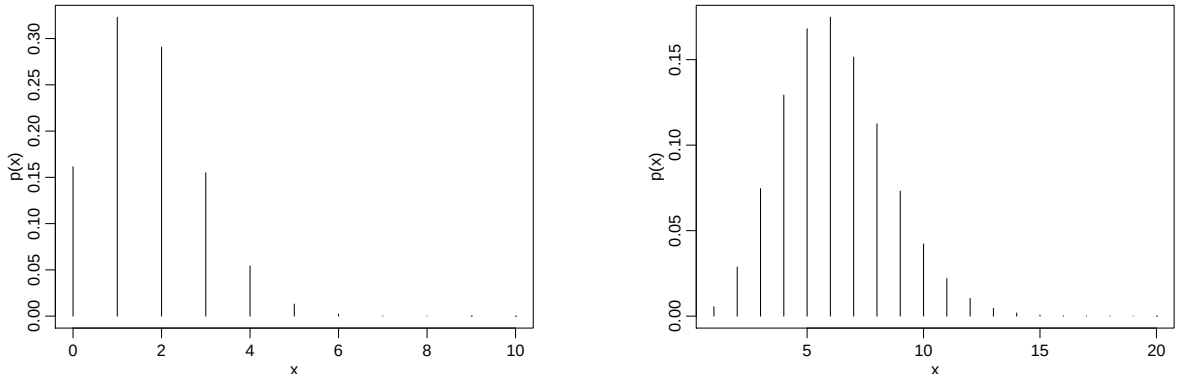
$$G_Z(s) = (1 - p + ps)\cdots(1 - p + ps) = (1 - p + ps)^n.$$

Although not yet derived, this is the probability generating function of the binomial, and so the sum of Bernoulli random variables is binomial, $X_1 + X_2 + \cdots + X_n \sim B(n, p)$.

6 Models for measurement data

6.1 Introduction

So far we have considered only random variables which have finite or countably infinite sample spaces, for example $\Omega = \{0, 1, \ldots, n\}$ for the binomial distribution or $\Omega = \{0, 1, \ldots\}$ for the Poisson distribution — that is, *discrete* random variables.



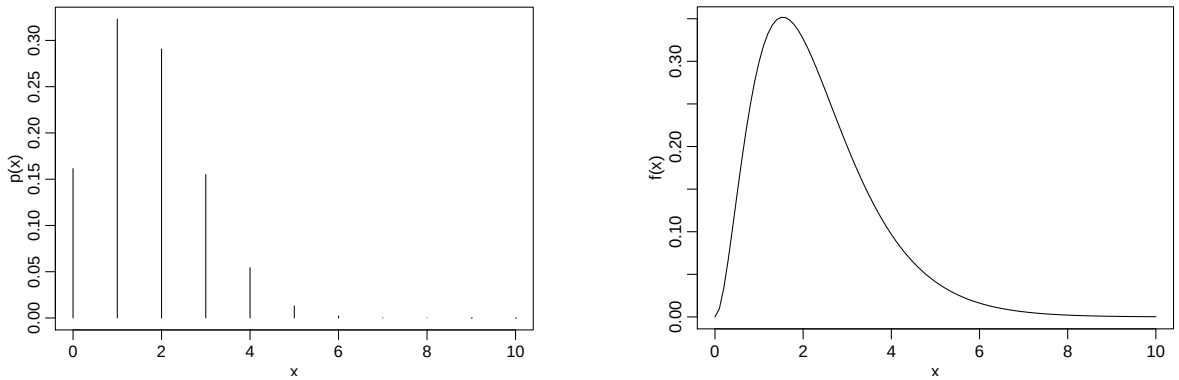
Many quantities of interest, however, can take values anywhere within some interval of the real line, for example:

- Proportion of the UK population currently unemployed: $\Omega = [0, 1]$.
- Time before the next shop customer arrives (min): $\Omega = [0, \infty)$.
- Total assets of a bank (£): $\Omega = (-\infty, \infty)$.

Such random quantities are called *continuous* random variables.

Consider the probability that a random variable X lies within a small interval, $[a, b]$, of width δx (with $\delta x \geq 0$) centred on the value x:

$$Pr(X \in [a, b]) = Pr\left(x - \frac{1}{2}\delta x \leq X \leq x + \frac{1}{2}\delta x\right).$$



If X is discrete, then we can choose δx to be sufficiently small that the interval contains just one element of the range space, x_i say. So

$$Pr(X \in [a, b]) = Pr(X = x_i), \quad \text{where } x_i \in [a, b].$$

For a continuous random variables this will never happen. Any interval, however small,

contains infinitely many possible values. But choosing $\delta x = 0$ gives

$$Pr(X \in [a, b]) = Pr(x \leq X \leq x) = Pr(X = x) = 0$$

as x is just one value out of an infinite number of values within even the smallest interval.

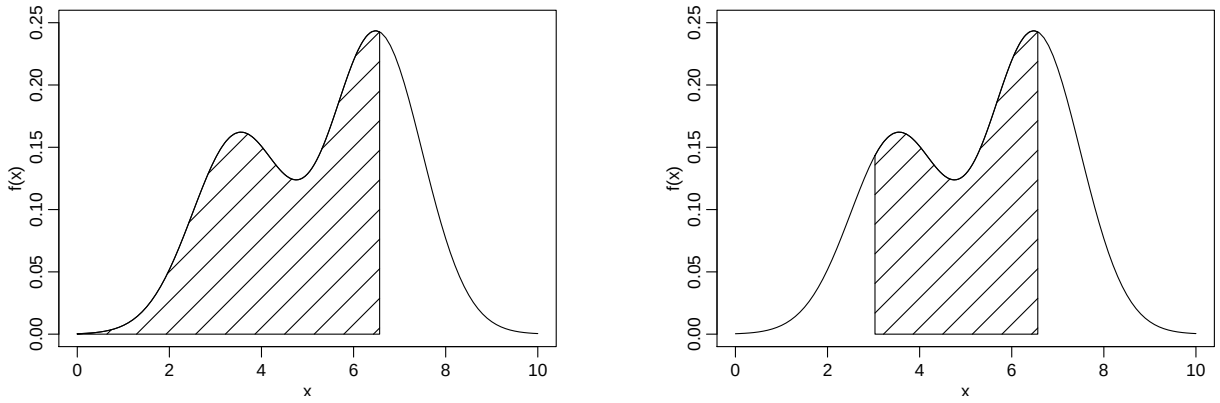
Probability density function (pdf)

For a continuous random variable X we define the probability density function, $f_X(x)$, through the following statement in terms of the cumulative distribution function, $F_X(\cdot)$:

$$F_X(b) = Pr(X \leq b) = \int_{-\infty}^{b} f_X(x)dx, \quad \text{for for any } b.$$

Note that for a continuous random variable we have $Pr(X = x) = 0$ for any x, hence

$$Pr(X \leq b) = Pr(X < b) + Pr(X = b) = Pr(X < b).$$



Alternatively, we can define the pdf through the following statement:

$$Pr(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx = F_X(b) - F_X(a), \quad \text{for any } a < b.$$

Note that

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

so given the cdf we can find the pdf by differentiating. Equally, given the pdf we can (in principle) find the cdf by integration.

Note also that, if $F_X(x)$ is continuous and $|\delta x|$ is small, we can approximate

$$Pr\left(x - \frac{1}{2}|\delta x| \leq X < x + \frac{1}{2}|\delta x|\right) \approx f(x)|\delta x|.$$

From the Axioms of Probability we have the following properties of the pdf:

K1 The pdf is always non-negative,

$$f_X(x) \geq 0, \quad \text{for any } x \in \Omega_X.$$

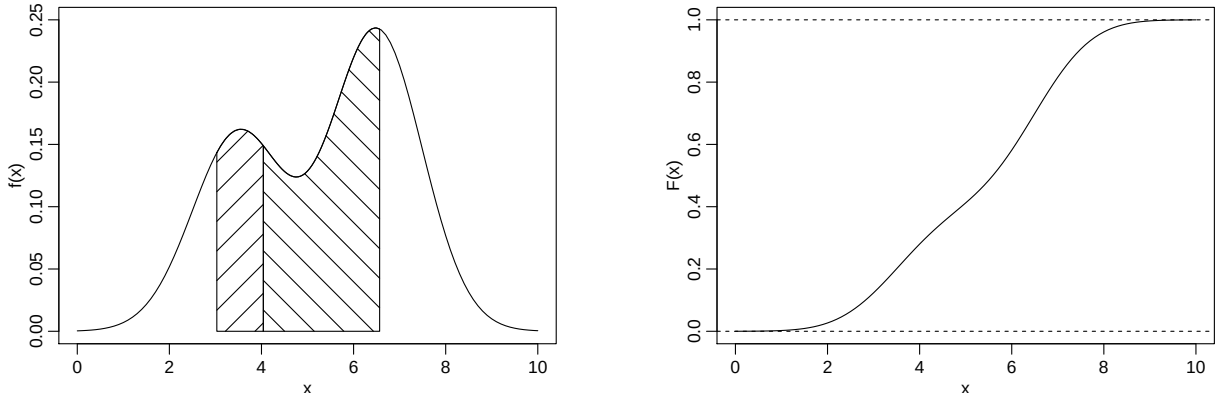
K2 The integral over the range space is equal to 1,

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

K3 For any non-overlapping subsets, S_1 and S_2, of the range space, Ω_X, then

$$\int_{S_1 \cup S_2} f_X(x)dx = \int_{S_1} f_X(x)dx + \int_{S_2} f_X(x)dx.$$

Which is the equivalent of the addition rule for mutually exclusive events.



We also have the following properties of the cdf:

C1 The cdf is bounded by 0 and 1,

$$0 \leq F_X(x) \leq 1, \quad \text{for any } x \in \Omega_X.$$

C2 The cdf is a non-decreasing function, that is

$$F_X(b) \geq F_X(a), \quad \text{where } b > a.$$

C3 The cdf has fixed points at the left and right,

$$F_X(-\infty) = 0 \quad \text{and} \quad F_X(\infty) = 1.$$

6.2 Expectation and variance of continuous random variables

To define the mean and variance of a continuous random variable, we generalise the definitions given previously for discrete random variables.

The expected value of a continuous random variable X is given by:

$$E[X] = \int_{-\infty}^{\infty} x \; f_X(x)dx.$$

Here we have simply replaced a summation by an integral. The definition of variance, in

terms of expectations, is unchanged:

$$Var[X] = E[(X-\mu)^2] = E[X^2] - \{E[X]\}^2$$

where $\mu = E[X]$. Then, we can evaluate $E[X^2]$ using the equation

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

We shall see examples of these later.

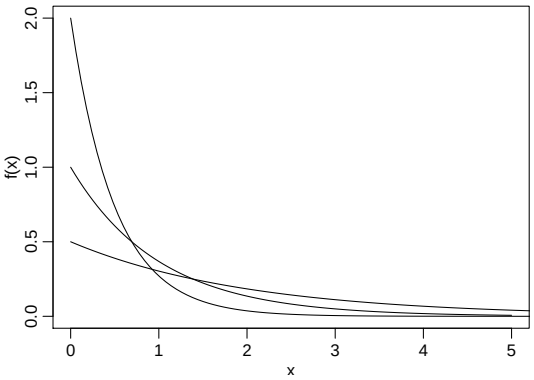
Now complete Worksheet 8 on basics properties of continuous random variables to check your understanding.

6.3 The exponential distribution

An exponential random variable, X, with rate λ is defined through its density,

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0; \quad 0 \text{ otherwise}$$

and is denoted $X \sim \text{Exp}(\lambda)$ — if $\lambda = 1$, this is called the *standard* exponential.
Clearly, $f_x(x) \geq 0$ for all x and it can be shown that $\int_{-\infty}^{\infty} f_X(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1$.



The cdf can be derived from the definition of the cdf and using the exponential pdf,

$$F_X(b) = Pr(X \leq b) = \int_{-\infty}^{b} f_X(x)dx = \int_0^b \lambda e^{-\lambda x} dx = \left[\frac{\lambda e^{-\lambda x}}{-\lambda}\right]_0^b = 1 - e^{-\lambda b}, \quad b \geq 0.$$

Note that in the derivation, the argument of the cdf has been changed to b to avoid any confusion between the variable of integration and the upper limit of the integral. Hence, with more usual symbols, the cdf is given by $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$; and 0 otherwise.

Example 6.1: Suppose that $\lambda = 1$, then what is $Pr(1 < X \leq 2)$?

$$Pr(1 < X \leq 2) = \int_1^2 \lambda e^{-\lambda x} dx = F_X(2) - F_X(1) = (1-e^{-2}) - (1-e^{-1}) = 0.2325.$$

The mean and variance can be obtained in one of two equivalent ways: either by direct integration, where $E[X]$ requires integration by parts, and $E[X^2]$ requires integration by parts twice, or by use of the gamma function.

Before looking at these results, note that the gamma function is defined as

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx,$$

with the properties that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ and hence, for integer n, $\Gamma(n) = (n-1)!$ Special cases are $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$ — see Essential Directed Reading for details.

Expectation and variance

From the definition of expectation we have

$$E[X] = \int_0^\infty x\ \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \int_0^\infty y e^{-y} dy = \frac{1}{\lambda}$$

where the first step uses the transformation $y = \lambda x$, hence $dy/dx = \lambda$, and the final step uses the fact that the integral gives $\Gamma(2) = 1! = 1$.

For the variance we need the following,

$$E[X^2] = \int_0^\infty x^2\ \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy = \frac{2}{\lambda^2}$$

using that the integral is $\Gamma(3) = 2! = 2$. Hence, the variance is

$$Var[X] = E[X^2] - \{E[X]\}^2 = \frac{2}{\lambda^2} - \left\{\frac{1}{\lambda}\right\}^2 = \frac{1}{\lambda^2}.$$

Exponential mean and variance of the exponential

Suppose we need to find the expectation and variance of the exponential distribution. In the lecture we will see how to use the *gamma function* to avoid performing the integration. The approach of re-arranging a sum or integral so that it matches a standard result is a useful technique, but it is not always the most intuitive. In previous mathematics courses you will have seen the method of *integration by parts* for tackling such integrals, but as we will see below, this is quite involved when we need to re-apply the formula multiple times. In this module, we only see a few of these types of calculations, and I do not mind what approach you use, but I recommend that you practice the approach used in the lecture as, in the long term, I believe it is much easier. However, below, is the solution using integration by parts for information.

Let X be an exponential random variable with rate parameter λ, that is $X \sim \exp(\lambda)$ with probability density function $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$ and zero otherwise. Then, for the expectation, starting with the general definition and then the specific form for the exponential

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\, dx.$$

We can integrate this by parts, recalling the usual formula

$$\int u\, v'\, dx = uv - \int u'\, v\, dx$$

with $u = x$, $v' = \lambda e^{-\lambda x}$ gives $u' = 1$, $v = \lambda e^{-\lambda x}/(-\lambda) = -e^{-\lambda x}$ and hence

$$E[X] = \Big[-xe^{-\lambda x} \Big]_0^{\infty} - \int_0^{\infty} \left(-e^{-\lambda x}\right)\, dx = \Big[-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \Big]_0^{\infty} = \frac{1}{\lambda}.$$

Before we can find the variance, we consider the expectation of the square,

$$E[X^2] = \int_{0}^{\infty} x^2 \cdot \lambda e^{-\lambda x}\, dx.$$

We will need to integrate by parts twice, starting with $u = x^2$, $v' = \lambda e^{-\lambda x}$ gives $u' = 2x$, $v = -e^{-\lambda x}$ and hence

$$E[X^2] = \Big[-x^2 e^{-\lambda x} \Big]_0^{\infty} - \int_0^{\infty} 2x(-e^{-\lambda x})\, dx.$$

Noting that the first term gives zero and the second can be rearranged, leads to the following

$$= \frac{2}{\lambda} \int_0^{\infty} x\lambda e^{-\lambda x} dx = \frac{2}{\lambda} \times \frac{1}{\lambda} = \frac{2}{\lambda^2}.$$

Where the integral is the same as for the expectation, and hence the required result.

Putting the two results together, gives the required result for the variance,

$$Var[X] = E[X^2] - \{E[X]\}^2 = \frac{2}{\lambda^2} - \left\{\frac{1}{\lambda}\right\}^2 = \frac{1}{\lambda^2}.$$

The Gamma Function

In lectures, for the exponential distribution, we will see that the gamma function can be used to avoid explicitly calculating some integrals. The approach of re-arranging an integral, so that it matches a standard result, is a useful technique but it is not always the most intuitive approach. In the long term it is, however, well work the effort!

Summary of key properties:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. $\Gamma(1) = 1$
3. $\Gamma(n) = (n - 1)!$

For this module you are expected to know and be able to use the above results, but you are not expected to know their derivation. The following is only for those who are inquisitive...

1. Start with the definition of the *gamma function* as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx \qquad \alpha > 0.$$

Now consider performing the integration by parts, recalling the usual formula

$$\int u\, v'\, dx = uv - \int u'\, v\, dx$$

with $u = x^{\alpha-1}$, $v' = e^{-x}$ gives $u' = (\alpha - 1)x^{\alpha-2}$, $v = -e^{-x}$ and hence

$$\Gamma(\alpha) = \Big[- x^{\alpha-1}\, e^{-x} \Big]_0^\infty + \int_0^\infty (\alpha - 1)x^{\alpha-2}\, e^{-x}\, dx.$$

Noting that the first term gives zero and the second can be rearranged, leads to the following

$$= (\alpha - 1) \int_0^\infty x^{\alpha-2}\, e^{-x}\, dx.$$

The integral is of the same form as in the above definition, except α has been replaced by $\alpha - 1$, and hence we get

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1).$$

2. Consider the definition, with α set to 1, then

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = \Big[- e^{-x} \Big]_0^\infty = 1.$$

3. Now, as a consequence of the general formula $\Gamma(\alpha-1) = (\alpha-2)\Gamma(\alpha-2)$ etc. and hence

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) \text{ etc.}$$

Suppose that we evaluate the function for positive integer values, $\alpha = n$ say, then

$$\Gamma(n) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)\cdots 2 \times \Gamma(1)$$

and using the above result for $\Gamma(1)$

$$\Gamma(n) = (n-1)(n-2)\cdots 2 \times 1 = (n-1)!$$

So, we see a relationship between the gamma function and the factorial.

Warning: the result

$$\Gamma(1/2) = \sqrt{\pi}$$

may seem simple, but its derivation is not. Proceed with caution!

From the definition, with $\alpha = 1/2$, we get

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty x^{-1/2}\, e^{-x}\, dx.$$

Consider the transformation, $x = t^2$, then $x^{-1/2} = 1/t$ and $dx/dt = 2t$, which then gives

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-t^2}\, dt.$$

But consider instead

$$\left\{\Gamma\left(\frac{1}{2}\right)\right\}^2 = 2\int_0^\infty e^{-s^2}\, ds \times 2\int_0^\infty e^{-t^2}\, dt = 4\int_0^\infty \int_0^\infty e^{-(s^2+t^2)} ds dt$$

transforming this to polar coordinates gives

$$= 4\int_0^\infty \int_0^{\pi/2} re^{-r^2} d\theta dr = 4\int_0^{\pi/2} 1\, d\theta \int_0^\infty re^{-r^2} dr = \pi\int_0^\infty 2re^{-r^2} dr = \pi\Big[-e^{-r^2}\Big]_0^\infty = \pi.$$

Hence, as required, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Now complete Worksheet 9 on more properties of continuous random variables to check your understanding.

6.4 The uniform and beta distributions

The simplest continuous distribution is the uniform which is defined via its density as:

$$f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b; \quad 0 \text{ otherwise}$$

and is denoted $U(a, b)$. The corresponding cumulative distribution function is

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

It can be shown that $E[X] = (a+b)/2$ and $Var[X] = (b-a)^2/12$.

Proof: Starting with the definition of expectation and using the pdf of the uniform distributions gives

$$E[X] = \int_{-\infty}^{\infty} x f_X(s)dx = \int_0^1 x \, \frac{1}{b-a} dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

Next,

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(s)dx = \int_0^1 x^2 \, \frac{1}{b-a} dx = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3-a^3}{3(b-a)}$$

$$= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

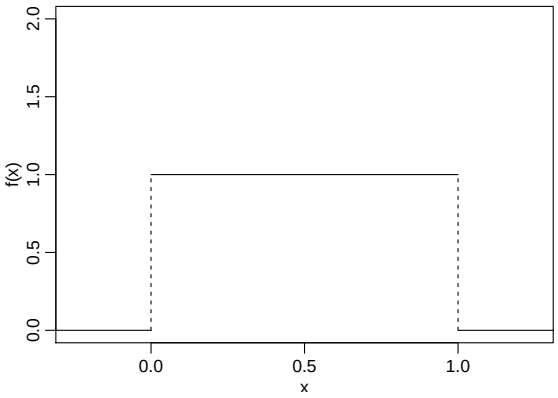
and hence

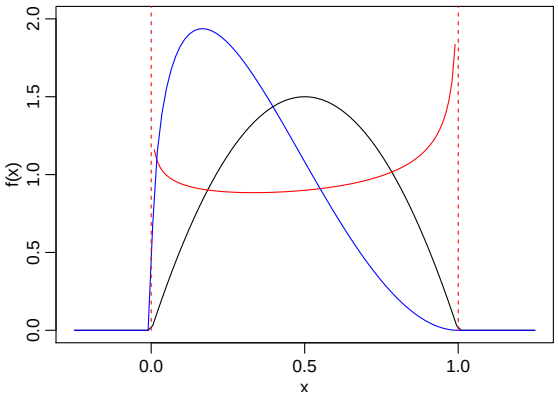
$$Var[X] = E[X^2] - \{E[X]\}^2 = \frac{b^2+ab+a^2}{3} - \left\{\frac{a+b}{2}\right\}^2$$

$$= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12}$$

$$= \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.$$

When $a = 0$ and $b = 1$, then X has a *standard uniform distribution* with pdf $f_X(x) = 1$ for $0 \leq x \leq 1$. Also, $F_X(x) = 0$ for $x < 0$, $F_X(x) = x$ for $0 \leq x \leq 1$, and $F_X(x) = 1$ for $x > 1$. Also, $E[X] = 1/2$ and $Var[X] = 1/12$.





The beta distribution with positive-valued parameters α and β, denoted Beta(α, β), is a generalization of the uniform distribution and is defined via its density:

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 \leq x \leq 1.$$

It can be shown that the mean is given by $E[X] = \alpha/(\alpha + \beta)$ and the variance by $Var[X] = \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$.

The role of the term $\mathrm{B}(\alpha, \beta)$ is merely to ensure that $\int f_X(x)dx = 1$ – it is a *constant of proportionality* or *normalizing constant*. That is, $\mathrm{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ but it can be shown also to be given by $\mathrm{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ where $\Gamma(\cdot)$ denotes the gamma function. The definition and properties of the gamma function are given in the Essential Directed Reading titled, *Exponential distribution and the gamma function*.

Notice that when $\alpha = 1$ and $\beta = 1$, then the pdf does not depend on x, that is it is a constant, and hence the beta distribution reduces to the uniform. Further, whenever $\alpha = \beta$, the pdf is symmetric and the expectation is $E[X] = 1/2$.

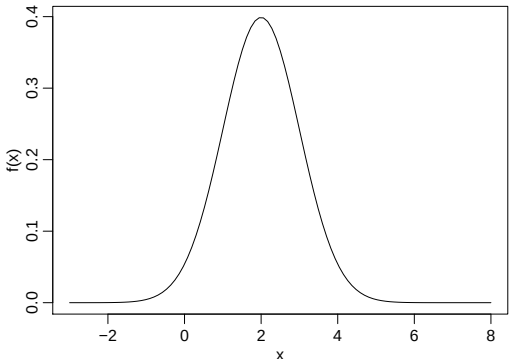
6.5 The normal (Gaussian) distribution

This is the most widely used distribution. In a few cases it has been proven to be the correct distribution, in some cases it has been proven to be an approximation and in many it is simply used as a "convenient model which seem to work well".

Let random variable X be normally distributed with parameters μ (say "mu") and σ^2 (say "sigma squared"), then we can write $X \sim N(\mu, \sigma^2)$.
The probability density function of X is

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

for $-\infty < x < \infty$.



If $\mu = 0$ and $\sigma^2 = 1$, then we obtain the standard normal — often this is denoted Z. Clearly, $Z \sim N(0, 1)$ has pdf

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

Although, clearly, $f_X(x) \geq 0$ (and $f_Z(z) \geq 0$) it is very difficult to show that the pdfs integrate to 1. Also, there is no equation for the cumulative distribution function — instead statistical tables, or a computer program such as R, are needed.

This distribution is so important that the pdf and cdf of the standard normal distribution have special notation

$$\phi(z) = f_Z(z) \quad \text{and} \quad \Phi(z) = F_Z(z) = Pr(Z \leq z).$$

Example 6.2: Suppose that $Z \sim N(0, 1)$ then $Pr(Z \leq 1) = 0.8413$ (from tables). From this we have, for example, $Pr(Z > 1) = 1 - Pr(Z \leq 1) = 0.1587$.

Properties of the normal:

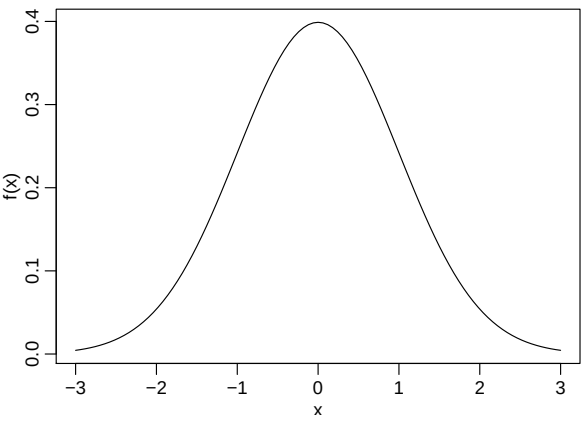
(N1) Symmetric about $x = \mu$, hence, for example with the standard normal $\phi(-z) = \phi(z)$ and $\Phi(-z) = 1 - \Phi(z)$.

(N2) Points of inflection in the pdf occur at $x = \mu - \sigma$ and $x = \mu + \sigma$.

(N3) If $X \sim N(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim N(0, 1)$ — this transformation is called standardization.

Because of this, $F_X(x) = Pr(X \leq x) = Pr(Z \leq (x - \mu)/\sigma) = \Phi((x - \mu)/\sigma)$.

Note that it can be shown that $E[X] = \mu$ and $Var[X] = \sigma^2$.

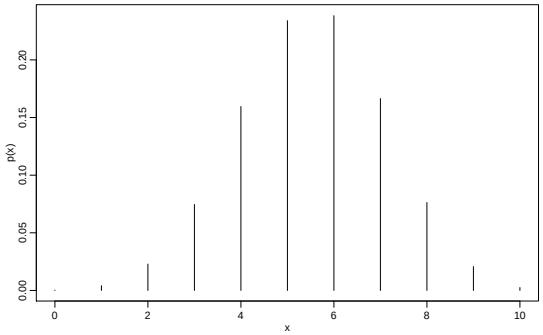


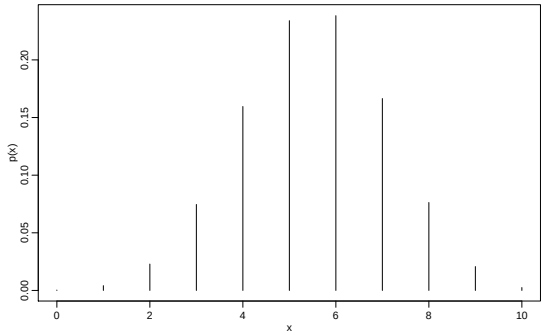
Normal approximations Here we note three important points regarding the use of the normal distribution.

1. Suppose $X \sim B(n,p)$ and that $Y \sim N(\mu, \sigma^2)$. Then, when n is large and p close to $1/2$, the binomial distribution can be approximated by $N(\mu = np, \sigma^2 = np(1-p))$.

   Note that X is a discrete random variable while Y is continuous. We can use a "continuity correction" to improve the accuracy of the approximation, so that

   $$Pr(X = x) \approx Pr(x - 1/2 \leq Y \leq x + 1/2)$$





   Note: We always "widen the interval" then using the continuity correction, so $Pr(X \leq x) \approx Pr(Y \leq x + 1/2)$ or $Pr(X \geq x) \approx Pr(Y \geq x - 1/2)$.

Example 6.3: let $X \sim B(n = 100, p = 1/2)$. What then is $Pr(X \leq 58)$?
Clearly,

$$Pr(X \leq 58) = \sum_{x=0}^{58} \binom{100}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{100-x}$$

which is lengthy to calculate exactly. Instead we use the normal distribution $Y \sim N(\mu = np = 50, \sigma^2 = np(1-p) = 25)$ and apply the continuity correction,

$$\begin{aligned} Pr(X \leq 58) \approx Pr(Y \leq 58.5) &= Pr(Z \leq \frac{58.5 - 50}{\sqrt{25}}) \\ &= Pr(Z \leq 1.7) = \Phi(1.7) = 0.9554. \end{aligned}$$

The exact value, using R, is $Pr(X \leq 58) = 0.9557$.

2. Another important use of the normal distribution model arises because of the "Central Limit Theorem", which states that,

   If $X_1, \ldots, X_n$ are independent observations from an arbitrary distribution with mean μ and finite variance σ^2, then as $n \to \infty$

   $$Z_n = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

where $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Clearly, for finite sample sizes then we can say that Z_n is approximately normal. Note that, very importantly, the there is no assumption that the $X_i's$ are normally distributed.

This result can also be converted to apply to $\bar{X}$ itself, or to the sum $X_1 + \cdots + X_n$.

3. Often we wish to evaluate $\Phi(z) = Pr(Z \leq x)$ for a value of z which is not in the statistical tables. If we have access to R then this is not a problem, but otherwise we would need to use linear interpolation to obtain an approximate value — see the details as part of the printed normal tables.

Now complete Worksheet 10 on basics of the normal distribution to check your understanding.

Mean and variance of the normal distribution

In lectures, we stated that the expectation of the standard normal distribution is zero and that the variance is 1, that is if $X \sim N(0,1)$, then $E[X] = 0$ and $Var[X] = 1$. Below is the proof of these results.

Recall that the probability density function for Z is as follows

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad \text{for } -\infty < z < \infty.$$

Then,

$$E[Z] = \int_{-\infty}^{\infty} z \, \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \left[-\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right]_{-\infty}^{\infty} = 0.$$

For the variance, we first need $E[Z^2]$, that is

$$E[Z^2] = \int_{-\infty}^{\infty} z^2 \, \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

which we can integrate by parts with $u = z$ and $dv/dz = ze^{-z^2/2}$, then $du/dz = 1$ and $v = -e^{-z^2/2}$ leading to

$$= \left[-z \, \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

In the first term it is not trivial to see the answer, but notice that the function is anti-symmetric, that is $h(z) = -h(-z)$, and hence will have answer zero. Also, the second term is the integral of the standard normal over its full rangre and hence equals one. Together these mean that

$$E[Z^2] = 1.$$

Hence,

$$Var[Z] = E[Z^2] + \{E[Z]\}^2 = 1 - 0^2 = 1.$$

General normal

Note that there is no need to "start again" for the general normal. Instead, consider the transformation $X = \sigma Z + \mu$, then $dx/dz = \sigma$ and the probability density of the general normal is transformed into that of the standard normal. Hence, $E[X] = E[\sigma Z + \mu] = \sigma E[Z] + \mu = \mu$, and $Var[X] = Var[\sigma Z + \mu] = \sigma^2 Var[Z] = \sigma^2$.

7 Bayesian Methods

7.1 Introduction

As motivation, and as a simple example to study, consider the tossing of a possibly biased coin with $Pr(\{\texttt{Heads}\}) = p$. Here we do not know p and hence we could consider it as a random variable and choose a suitable model. So, in an analysis we need to take into account the two sources of uncertainty: (i) that due to the unknown outcome of the coin tossing experiment, and (ii) that due to the unknown value of p.

- We know that p must be within the range $0 \leq p \leq 1$, but it might take any value within the range. This strongly suggests using a beta distribution, $P \sim \text{Beta}(\alpha, \beta)$ — with α and β specified appropriately.
- Suppose that we are given the value of p, then the experiment would correspond well to a sequence of Bernoulli trials. If we record the number of heads, X, in a fixed number of trials n, then $X|p \sim \text{B}(n, p)$, whereas if we record the number of trials, Y, until we get the first head then $Y|p \sim \text{Ge}(p)$.

In the terminology of Bayesian statistics, the distribution of p is called the prior distribution — as it is fixed before data is collected — and the conditional distribution of the experimental outcome given the value of p is called the likelihood.

Hence, in our coin tossing example, we now need to combine the prior distribution, $f(p)$, and the likelihood $l(x|p)$ — note that sometimes the symbol l is used instead of f. This can be done using Bayes' Theorem to produce what is called the posterior distribution, $f(p|x)$ — as it describes our knowledge of p after the experiment, that is

$$f(p|x) = l(x|p)f(p)/f(x)$$

where $f(x)$, a constant of proportionality, is known as the model evidence. In practice, its value is rarely needed as it is not a function of p — it can, however, be obtained by integration, $f(x) = \int f(x, p)\, dp = \int l(x|p)f(p)\, dp$, since p is continuous.

Note that when we have a general dataset of size N, that is $\underline{x} = (x_1, \ldots, x_N)$, then the notation for the posterior distribution changes to

$$f(p|\underline{x}) = l(\underline{x}|p)f(p)/f(\underline{x})$$

with $l(\underline{x}|p) = \prod_{i=1}^{N} f(x_i|p)$ assuming that the separate data values are independent.

Example 7.1: Suppose we have a coin, but we do not know if it is fair or not. We examine it and by appearance it looks like a regular coin.

It seems reasonable to toss the coin and estimate the probability of `Heads` using the relative frequency, that is if we obtain 5 heads out of 10 tosses, then $\hat{p} = 0.5$.

However, suppose we get 8 heads out of the 10 tosses, then would we really believe that $\hat{p} = 0.8$? Instead, might we simply say that we were "lucky"? Hence, it is very natural to involve our prior beliefs into the analysis and interpretation of experimental results.

7.2 The beta-binomial model

In this case our model is $X|p \sim \mathrm{B}(n,p)$ and $P \sim \mathrm{Beta}(\alpha,\beta)$ with

$$l(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{and} \quad f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

This gives a posterior distribution

$$f(p|x) \propto p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} = p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

where all terms not involving p have been ignored — these terms combine to give one big constant of proportionality.

The terms on the right should be "familiar" — compare it to the probability density for a beta random variable. Our function depends on p in exactly the same way as the beta distribution, and hence our posterior distribution is a beta distribution. This allows us to match the remaining constant terms to give

$$f(p|x) = \frac{1}{B(\alpha+x, \beta+n-x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

that is $P|x \sim \mathrm{Beta}(\alpha+x, \beta+n-x)$.

Recall the mean and variance of the $\mathrm{Beta}(a,b)$ distribution are

$$\frac{a}{a+b} \quad \text{and} \quad \frac{ab}{(a+b)^2(a+b+1)}.$$

This gives the mean and variance of the prior distribution as:

$$\frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

and the corresponding values for the posterior distribution as:

$$\frac{\alpha+x}{\alpha+\beta+n} \quad \text{and} \quad \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}.$$

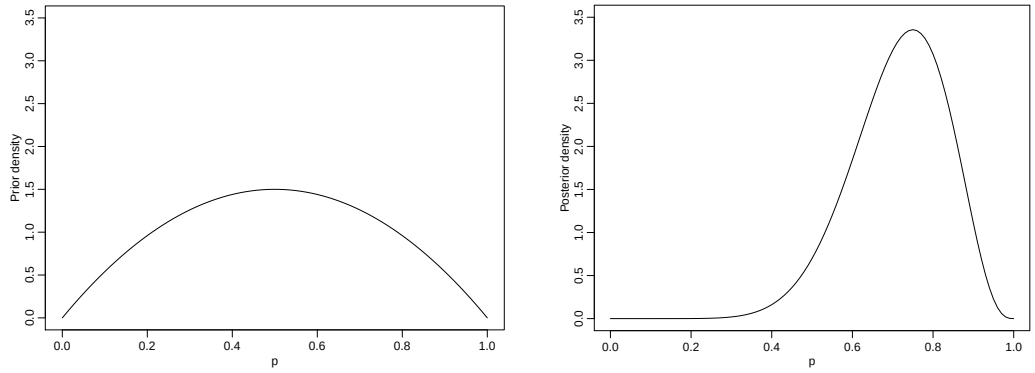
One very important question remains. How should you fix α and β? It is a subjective choice, and each of us might choose differently depending on the situation. We would just need to ensure that we choose fairly and to genuinely reflect our personal prior beliefs.

However, notice that for any (finite) choice of α and β, as $n \to \infty$ the mean tends to x/n, the sample proportion, and the variance to zero — this behaviour seems reasonable.

Example 7.2: Suppose that we believe that the prior mean should be 1/2 and the prior variance 1/20, then solving $\alpha/(\alpha+\beta) = 1/2$ and $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1) = 1/20$ gives the values $\alpha = \beta = 2$.

Suppose we toss a coin $n = 10$ times and observe $x = 8$ heads, then the posterior distribution is $P|x \sim \text{Beta}(\alpha + x = 10, \beta + n - x = 4)$ with posterior mean $(\alpha + x)/(\alpha + \beta + n) = 10/14 = 0.7143$ and posterior variance $(\alpha + x)(\beta + n - x)/(\alpha + \beta + n)^2(\alpha + \beta + n + 1) = 0.0136$.

Notice that the mean has increased from the prior belief of 0.5 towards the observed relative frequency of 0.8, and that the posterior variance is now less than the prior variance reflecting the inclusion of data (0.01 compared to 0.05).



(a) Prior distribution, Beta$(2, 2)$ (b) Posterior distribution, Beta$(10, 4)$

Example 6.1: Analysis of the beta-binomial model.

The approach can be applied sequentially where the posterior distribution from one stage can be used as the prior distribution at the next stage — with conjugate prior distributions this is very easy as we only need to update the parameter values.

7.3 The beta-geometric model

In this case our model is $X|p \sim \text{Ge}(p)$ and $P \sim \text{Beta}(\alpha, \beta)$ with probability functions

$$l(x|p) = p(1-p)^{x-1} \quad x = 1, 2, \ldots \quad \text{and} \quad f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad 0 \leq p \leq 1.$$

This gives a posterior distribution

$$f(p|x) \propto p(1-p)^{x-1} \times p^{\alpha-1}(1-p)^{\beta-1} = p^{\alpha+1-1}(1-p)^{\beta+x-1-1}$$

hence our posterior distribution must be a beta distribution. This allows us to match the remaining constant terms to give

$$f(p|x) = \frac{1}{B(\alpha+1, \beta+x-1)} p^{\alpha+1-1}(1-p)^{\beta+x-1-1}$$

that is $P|x \sim \text{Beta}(\alpha + 1, \beta + x - 1)$. Hence the prior mean and variance are:

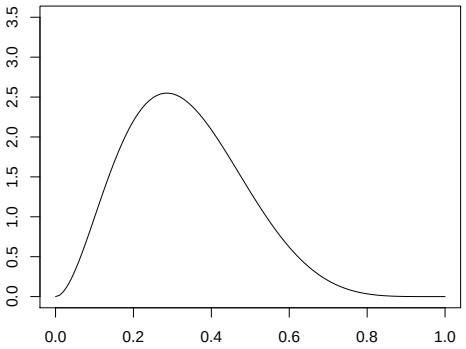
$$\frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

and the corresponding values for the posterior distribution are:

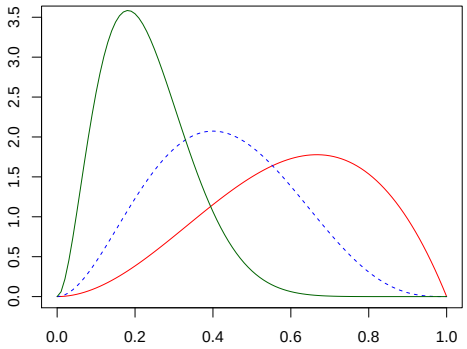
$$\frac{\alpha + 1}{\alpha + \beta + x} \quad \text{and} \quad \frac{(\alpha + 1)(\beta + x - 1)}{(\alpha + \beta + x)^2(\alpha + \beta + x + 1)}.$$

Example 7.3: Again we will use the prior parameter values $\alpha = 2$ and $\beta = 2$. Suppose the coin is tossed $x = 5$ times to obtain the first heads, that is we observe T, T, T, T, H, then the posterior distribution is $P|x \sim \text{Beta}(\alpha + 1 = 3, \beta + x - 1 = 6)$ with mean $(\alpha + 1)/(\alpha + \beta + x) = 3/9 = 1/3$ and variance $(\alpha + 1)(\beta + x - 1)/(\alpha + \beta + x)^2(\alpha + \beta + x + 1) = 18/810 = 0.0222$.

Notice that this time the mean has decreases from the prior belief of $1/2$, but again the posterior variance is less than the prior variance reflective the inclusion of data (0.02 compared to 0.05).



(a) Posterior distribution, Beta(3, 6)



(b) Posterior distributions with $x = 1$ (red), $x = 3$ (blue) and $x = 9$ (green)

Example 6.2: Analysis of the beta-geometric model.

Comments

1. In the above two cases the posterior distribution was from the same distribution family as the prior distribution — such priors are called conjugate — making the algebra much more straight forward. In other cases, however, this will not happen and it can even take complex numerical methods to solve the problem.

2. Note that the above are two special examples of Bayesian modelling, but similar approaches can be taken to model the rate parameter, λ, in the Poisson or exponential distributions or the mean and variance, μ and σ^2, in the normal distribution.

3. Although rare in the past, Bayesian methods are now part of main-stream statistics – every professional statistician should know about them. However, it is unusual for them to appear in first year statistics modules. Here we have only made a start, but you can choose later modules to go into greater depth.

7.4 The exponential-Poisson model

In this case our model is $X|\lambda \sim \text{Po}(\lambda)$ and $\lambda \sim \exp(\beta)$ with probability functions

$$l(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, \ldots \quad \text{and} \quad f(\lambda) = \beta e^{-,\beta\lambda} \quad \lambda > 0.$$

This gives a posterior distribution

$$f(\lambda|x) \propto \lambda^x e^{-\lambda} \times e^{-\beta\lambda} = \lambda^x e^{-\lambda(\beta+1)}.$$

which, unfortunately, is a distribution that we have not met — the gamma distribution. In later modules you will see this, but it turns out that the posterior density function is given by

$$f(\lambda|x) = \frac{(\beta+1)^{x+1}}{\Gamma(x+1)} \lambda^x e^{-\lambda(\beta+1)}$$

with $\lambda|x \sim \text{Gamma}(x+1, \beta+1)$ with mean and variance of the posterior distribution given by $(x+1)/(\beta+1)$ and $(x+1)/(\beta+1)^2$.

Note that the exponential is a special case of the gamma distribution and hence the gamma is the conjugate prior for the Poisson, and also it is conjugate for the exponential distribution.

Index