Quantifying Life

Dmitry Kondrashov

Table of contents

Preface

> What is a man, said Athos, who has no landscape? Nothing but mirrors and tides.
> – Anne Michaels, **Fugitive Pieces**

This is an online book to help biologists and biology-adjacent folks learn quantitative skills through the practice of programming in R. These skills can be roughly sorted into four types:

- Building models and understanding assumptions
- Writing code to perform computational tasks
- Performing mathematical analysis of models
- Working with data and using statistical tools

These skills interface, intertwine, and reinforce each other in the practice of biological research and are thus presented concurrently in this book, instead of being corralled into separate courses taught by different departments, like mathematics, statistics, and computer science. Here I combine ideas and skills from all of these disciplines into an educational narrative organized by increasing exposure to programming concepts.

A brief motivation of mathematical modeling

A mathematical model is a representation of some real object or phenomenon in terms of quantities (numbers). The goal of modeling is to create a description of the object in question that may be used to pose and answer questions about it, without doing hard experimental work. A good analogy for a mathematical model is a map of a geographic area: a map cannot record all of the complexity of the actual piece of land, because then the map would need to be size of the piece of land, and then it wouldn't be very useful! Maps, and mathematical models, need to sacrifice the details and provide a birds-eye view of reality in order to guide the traveler or the scientist. The representation of reality in the model must be simple enough to be useful, yet complex enough to capture the essential features of what it is trying to represent.

Mathematical modeling has long been essential in physics: for instance, it is well known that distance traveled by an object traveling at constant speed v is proportional to the time traveled (called t). This mathematical model can be expressed as an equation:

$$d = vt$$

Since the time of Newton, physicists have been very successful at using mathematics to describe the behavior of matter of all sizes, ranging from subatomic particles to galaxies. However, mathematical modeling is a new arrow in a biologist's quiver. Many biologists would argue that living systems are much more complex than either atoms or galaxies, since even a single cell is made up of a mind-boggling number of highly dynamic, interacting entities. That is true, but new advances in experimental biology are producing data that make quantitative methods indispensable for biology.

The advent of genetic sequencing in the 1970s and 80s has allowed us to determine the genomes of different species, and in the last few years next-generation sequencing has reduced sequencing costs for an individual human genome to a few thousand dollars. The resulting deluge of quantitative data has answered many outstanding questions, and also led to entirely new ones. We now understand that knowledge of genomic sequences is not enough for understanding how living things work, so the burgeoning field of systems biology investigates the interactions between genes, proteins, or other entities. The central question is to understand how a network of interactions between individual molecules can lead to large-scale results, such as the development of a fertilized egg into a complex organism. The human mind is not suited for making correct intuitive judgements about networks comprised of thousands of actors. Addressing questions of this complexity requires quantitative modeling.

Purpose of this book

This textbook is intended for a college-level course for biology and pre-medicine majors, or more established scientists interested in learning the applications of mathematical methods to biology. The book brings together concepts found in mathematics, computer science, and statistics courses to provide the student a collection of skills that are commonly used in biological research. The book has two overarching goals: one is to explain the quantitative language that often is a formidable barrier to understanding and critically evaluating research results in biological and medical sciences. The second is to teach students computational skills that they can use in their future research endeavors. The main premise of this approach is that computation is critical for understanding abstract mathematical ideas.

These goals are distinct from those of traditional mathematics courses that emphasize rigor and abstraction. I strongly believe that understanding of mathematical concepts is not contingent on being able to prove all of the underlying theorems. Instead, premature focus on abstraction obscures the ideas for most students; it is putting the theoretical cart before the experiential horse. I find that students can grasp deep concepts when they are allowed to experience them tangibly as numbers or pictures, and those with an abstract mindset can generalize and add rigor later. As I demonstrate in part 3 of the book, Markov chains can be explained without relying on the machinery of measure theory and stochastic processes, which require graduate level mathematical skills. The idea of a system randomly hopping between a few discrete states is far more accessible than sigma algebras and martingales. Of course, some abstraction is necessary when presenting mathematical ideas, and I provide correct definitions of terms and supply derivations when I find them to be illuminating. But I avoid rigorous proofs, and always favor understanding over mathematical precision.

The book is structured to facilitate learning computational skills. Over the course of the text students accumulate programming experience, progressing from assigning values to variables in the first chapter to solving nonlinear ODEs numerically by the end of the book. Learning to program for the first time is a challenging task, and I facilitate it by providing sample scripts for students to copy and modify to perform the requisite calculations. Programming requires careful, methodical thinking, which facilitates deeper understanding of the models being simulated. In my experience of teaching this course, students consistently report that learning basic scientific programming is a rewarding experience, which opens doors for them in future research and learning.

It is of course impossible to span the breadth of mathematics and computation used for modeling biological scenarios. This did not stop me from trying. The book is broad but selective, sticking to a few key concepts and examples which should provide enough of a basis for a student to go and explore a topic in more depth. For instance, I do not go through the usual menagerie of probability distributions in chapter 4, but only analyze the uniform and the binomial distributions. If one understands the concepts of distributions and their means and variances, it is not difficult to read up on the geometric or gamma distribution if one encounters it. Still, I omitted numerous topics and entire fields, some because they require greater mathematical sophistication, and others because they are too difficult for beginning programmers, e.g. sequence alignment and optimization algorithms. I hope that you do not end your quantitative journey with this book!

I take an even more selective approach to the biological topics that I present in every chapter. The book is not intended to teach biology, but I do introduce biological questions I find interesting, refer the reader to current research papers, and provide discussion questions for you to wrestle with. This requires a basic explanation of terms and ideas, so most chapters contain a broad brushstrokes summary of a biological field, e.g. measuring mutation rates, epidemiology modeling, hidden Markov models for gene structure, and limitations of medical testing. I hope the experts in these fields forgive my omitting the interesting details that they spend their lives investigating, and trust that I managed to get the basic ideas across without gross distortion.

Organization of the book

A course based on this textbook can be tailored to fit the quantitative needs of a biological sciences curriculum. At the University of Chicago the course I teach has replaced the last quarter of calculus as a first-year requirement for biology majors. This material could be used for a course without a calculus pre-requisite that a student takes before more rigorous statistics, mathematics, or computer science courses. It may also be taught as an upper-level elective course for students with greater maturity who may be ready to tackle the eigenvalues and differential equations chapters. My hope is that it may also prove useful for graduate students or established scientists who need an elementary but comprehensive introduction to the concepts they encounter in the literature or that they can use in their own research. Whatever path you traveled to get here, I wish you a fruitful journey through biomathematics and computation!

1 Arithmetic and variables

> You can add up the parts, but you won't have the sum;
> You can strike up the march, there is no drum.
> Every heart, every heart to love will come
> But like a refugee.
> –Leonard Cohen, *Anthem*

Mathematical modeling begins with a set of *assumptions*. In fact, one may say that a mathematical model is a bunch of assumptions translated into mathematics. These assumptions may be more or less reasonable, and they may come from different sources. For instance, many physical models are so well-established that we refer to them as laws; we are pretty sure they apply to molecules, cells, and organisms as well as to inanimate objects. Thus we may use physical laws as the foundation on which to build models of biological entities; these are often known as *first-principles* (theory-based) models. Other times we have experimental evidence which suggests a certain kind of relationship between quantities, perhaps we find that the amount of administered drug and the time until the drug is completely removed from the bloodstream are proportional to each other. This observation can be turned into an *empirical* (experiment-based) model. Yet another type of model assumption is not based on either theory or experiment, but simply on convenience: e.g. let us assume that the mutation rates in two different loci are independent, and see what the implications are. These are sometimes called *toy* or *cartoon* models. [**?**]

This leads to the question: how do you decide whether a model is good? It is surprisingly difficult to give a straightforward answer to this question. Of course, one major goal of a model is to capture some essential features of reality, so in most biological modeling studies you will see a comparison between experimental results and predictions of the model. But it is not enough for a model to be faithful to experimental data! Think of a simple example: suppose your experiment produced 5 data points as a function of time; it is possible to find a polynomial (of fourth degree) that passes exactly through all 5 points, by specifying the coefficients of its 5 terms. This is called *data fitting* and it has a large role to play in mathematical modeling of biology. However, I think you will agree that in this case we have learned very little: we just substituted 5 values in the data set with 5 values of the coefficients of the mathematical model. To heighten the absurdity, imagine a data set of 1001 points that you have modeled using a 1000-degree polynomial. This is an example of overfitting, or making the model agree with the data by making it overly complex.

Substituting a complicated model for a complicated real situation does not help understand it. One necessary ingredient of a useful model is *simplicity of assumptions*. Simplicity in modeling has at least two virtues: simple models can be grasped by our limited minds, and simple assumptions can be tested against evidence. A simple model that fails to reproduce experimental data can be more informative than a complex model that fits the data perfectly. If a simple model fails, you have learned that you are missing something in your assumptions; but a complex model can be right for the wrong reasons, like erroneous assumptions canceling each other, or it may contain needless assumptions. This is why good modeling is a difficult skill that balances simplicity of assumptions against fidelity to empirical data [**?**]. In this chapter you will learn how to do the following:

- distinguish variables and parameters in models
- describe the state space of a model
- perform arithmetic operations in R
- assign variables in R

1.1 Blood circulation and mathematical modeling

Galen was one of the great physicians of antiquity. He studied how the body works by performing experiments on humans and animals. Among other things, he was famous for a careful study of the heart and how blood traveled through the body. Galen observed that there were different types of blood: arterial blood that flowed out of the heart, which was bright red, and venous blood that flowed in the opposite direction, which was a darker color. This naturally led to questions: what is the difference between venous and arterial blood? where does each one come from and where does it go?

You, a reader of the 21st century, likely already know the answer: blood *circulates* through the body, bringing oxygen and nutrients to the tissues through the arteries, and returns back through the veins carrying carbon dioxide and waste products, as shown in figure **??**. Arterial blood contains a lot of oxygen while venous blood carries more carbon dioxide, but otherwise they are the same fluid. The heart does the physical work of pushing arterial blood out of the heart, to the tissues and organs, as well as pushing venous blood through the second circulatory loop that goes through the lungs, where it picks up oxygen and releases carbon dioxide, becoming arterial blood again. This may seem like a very natural picture to you, but it is far from easy to deduce by simple observation.

1.1.1 Galen's theory of blood

Galen came up with a different explanation based on the notion of *humors*, or fluids, that was fundamental to the Greek conception of the body. He proposed that the venous and arterial

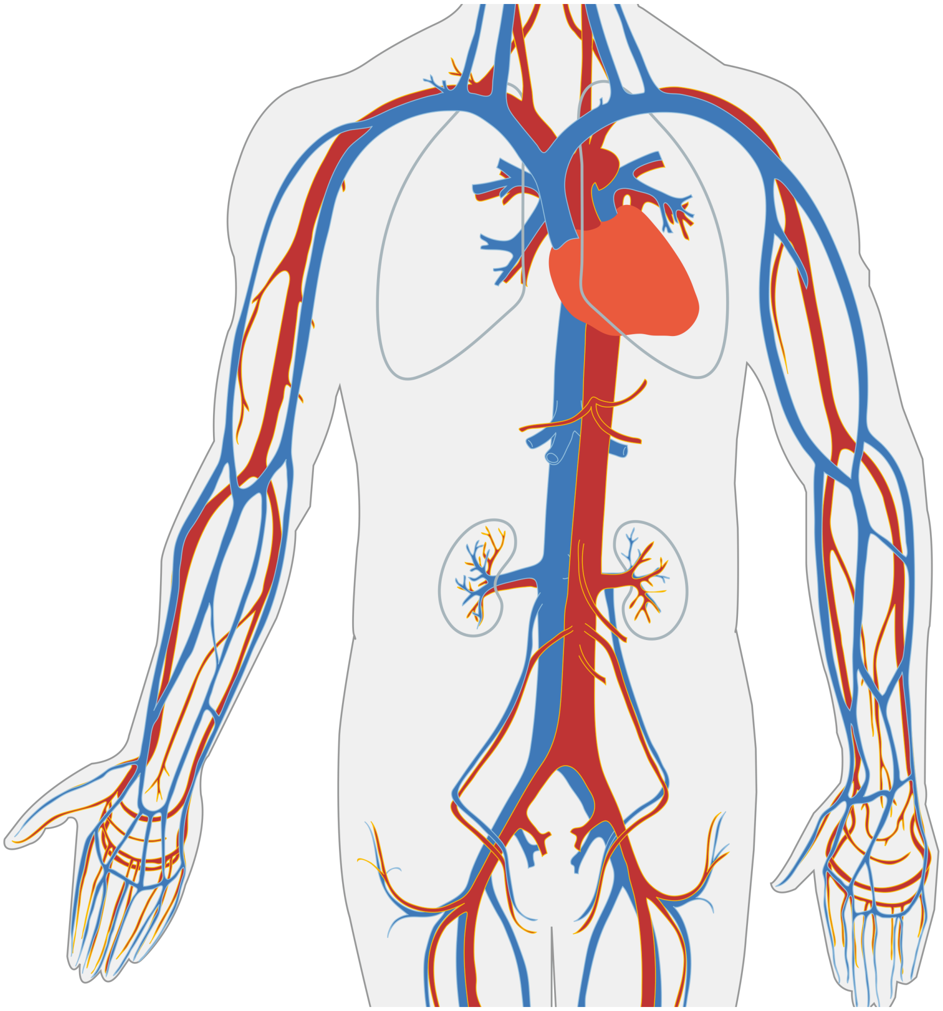


Figure 1.1: Human blood circulates throughout the body and returns to the heart, veins shown in blue and arteries in red. *Circulatory System en* by LadyofHats in public domain via Wikimedia Commons.

blood were different humors: venous blood, or *natural spirits*, was produced by the liver, while arterial blood, or *vital spirits*, was produced by the heart and carried by the arteries, as shown in figure **??**. The heart consisted of two halves, and it warmed the blood and pushed both the natural and vital spirits out to the organs; the two spirits could mix through pores in the septum separating its right and left halves. The vital and natural spirits were both consumed by the organs, and regenerated by the liver and the heart. The purpose of the lungs was to serve as bellows, cooling the blood after it was heated by the heart.

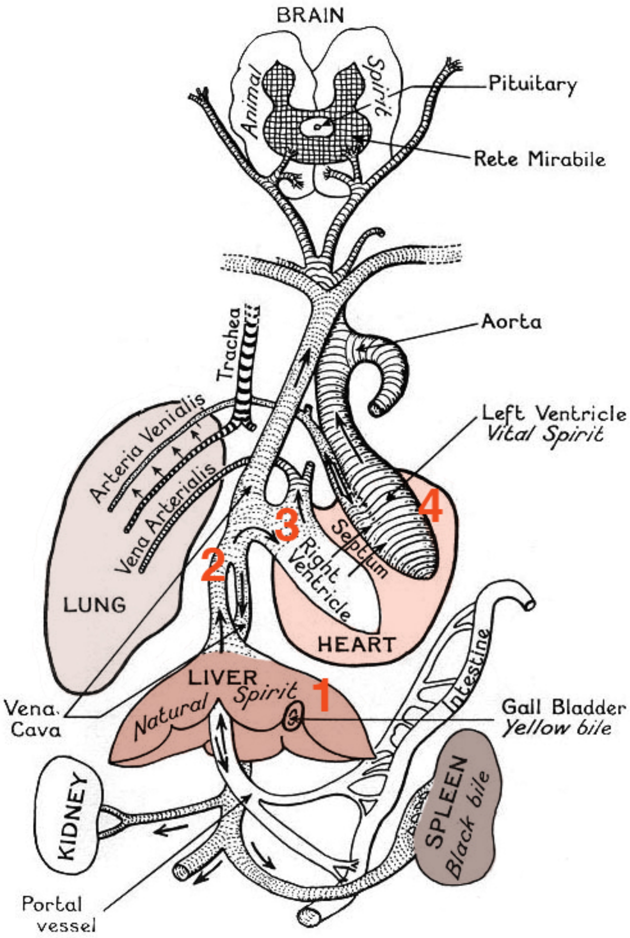


Figure 1.2: Illustration of Galen's conception of the blood system, showing different spirits traveling in one direction, but not circulating. Reproduced by permission of Barbara Becker.

Is this a good theory of how the heart, lungs, and blood work? Doctors in Europe thought so for over one thousand years! Galen's textbook on physiology was the standard for medical students through the 17th century. The theory seemed to make sense, and explain what was observable. Many great scientists and physicians, including Leonardo DaVinci and Avicenna, did not challenge the inaccuracies such as the porous septum in the heart, even though they could not see the pores themselves. It took both better observations and a quantitative testing of the hypothesis to challenge the orthodoxy.

1.1.2 Mathematical testing of the theory

William Harvey was born in England and studied medicine in Padua under the great physician Hieronymus Fabricius. He became famous, and would perform public demonstrations of physiology, using live animals for experiments that would not be approved today. He also studied the heart and the blood vessels, and measured the volume of the blood that can be contained in the human heart. He was quite accurate in estimating the correct volume, which we now know to be about 70 ml (1.5 oz). What is even more impressive is that he used this quantitative information to test Galen's theory.

Let us assume that all of the blood that is pumped out by the heart is consumed by the tissues, as Galen proposed; let us further assume that the heart beats at constant rate of 60 beats per minute, with a constant ejection volume of 70 ml. Then over the course of a day, the human body would consume about

$$\text{Volume} = 70 \text{ mL} \times 60 \text{ (beats per minute)} \times 60 \text{ (minutes per hour)} \times 24 \text{ (hours per day)}$$

or over 6,000 liters of blood! You may quibble over the exact numbers (some hearts beat faster or slower, some hearts may be larger or smaller) but the impact of the calculation remains the same: it is an absurd conclusion. Galen's theory would require the human being to consume and produce a quantity of fluid many times the volume of the human body (about 100 liters) in a day! This is a physical impossibility, so the only possible conclusion in that Galen's model is wrong.

This led Harvey to propose the model that we know today: that blood is not consumed by the tissues, but instead returns to the heart and is re-used again [**?**]. This is why we call the heart and blood vessels part of the circulatory system of the body. This model was controversial at the time - some people proclaimed they would "rather be wrong with Galen, than right with Harvey" - but eventually became accepted as the standard model. What is remarkable is that Harvey's argument, despite being grounded in empirical data, was strictly mathematical. He adopted the assumptions of Galen, made the calculations, and got a result which was inconsistent with reality. This is an excellent example of how mathematical modeling can be useful, because it can provide clear evidence against a wrong hypothesis.

1.2 Parameters and variables in models

Many biologists remain skeptical of mathematical modeling. The criticism can be summarized like this: a theoretical model either agrees with experiment, or it does not. In the former case, it is useless, because the data are already known; in the latter case, it is wrong! As I indicated above, the goal of mathematical modeling is not to reproduce experimental data; otherwise, indeed, it would only be of interest to theoreticians. The correct question to ask is, does a theoretical model help us understand the real thing? There are at least three ways in which a model can be useful:

- A model can help a scientist make sense of complex data, by testing whether a particular mechanism explains the observations. Thus, a model can help clarify our understanding by throwing away the non-essential features and focusing on the most important ones.
- A mathematical model makes predictions for situations that have not been observed. It is easy to change parameters in a mathematical model and calculate the effects. This can lead to new hypotheses that can be tested by experiments.
- Model predictions can lead to better experimental design. Instead of trying a whole bunch of conditions, the theoretical model can suggest which ones will produce big effects, and thus can save a lot of work for the lab scientist.

In order to make a useful model of a complex living system, you have to simplify it. Even if you are only interested in a part of it, for instance a cell or a single molecule, you have to make simplifying choices. A small protein has thousands of atoms, a cell consists of millions of molecules, which all interact with each other; keeping track mathematically of every single component is daunting if not impossible. To build a useful mathematical model one must choose a few quantities which describe the system sufficiently to answer the questions of interest. For instance, if the positions of a couple of atoms in the protein you are studying determine its activity, those positions would make natural quantities to include in your model. You will find more specific examples of models later in this chapter.

Once you have decided on the essential quantities to be included in the model, these are divided into *variables* and *parameters*. As suggested by the name, a variable typically varies over time and the model tracks the changes in its value, while parameters usually stay constant, or change more slowly. However, that is not always the case. The most important difference is that variables describe quantities **within the system** being modeled, while parameters usually refer to quantities which are controlled by something **outside the system**.

As you can see from this definition, the same quantity can be a variable or a parameter depending on the scope of the model. Let's go back to our example of modeling a protein: usually the activity (and the structure) of a protein is influenced by external conditions such as pH and temperature; these would be natural parameters for a model of the molecule. However, if we model an entire organism, the pH (e.g. of the blood plasma) and temperature are controlled by physiological processes within the organism, and thus these quantities will now be considered variables.

Perhaps the clearest way to differentiate between variables and parameters is to think about how you would present a data set visually. We will discuss plotting graphs of functions in chapter 2, and plotting data sets in chapter 3, but the reader has likely seen many such plots before. Consider which of the quantities you would to plot to describe the system you are modeling. If the quantity belongs on either axis, it is a variable, since it is important to describe how it changes. The rest of the quantities can be called parameters. Of course, depending on the question you ask, the same quantity may be plotted on an axis or not, which is why this classification is not absolute.

After we have specified the essential variables for your model, we can describe a complex and evolving biological system in terms of its *state*. This is a very general term, but it usually means the values of all the variables that you have chosen for the model, which are often called *state variables*. For instance, an ion channel can be described with the state variable of conformation, which may be in a open state or in a closed state. The range, or collection of all different states of the system is called the *state space* of the model. Below you will find examples of models of biological systems with diverse state spaces.

1.2.1 discrete state variables: genetics

There are genes which are present in a population as two different versions, called *alleles} - let us use letters A and B to label them. One may describe the genetic state of an individual based on which allele it carries. If this individual is haploid, e.g. a bacterium, then it only carries a single copy of the genome, and its state can be described by a single variable with the state space of A or B.

A diploid organism, like a human, possesses two copies of each gene (unless it is on one of the sex chromosomes, X or Y); each copy may be in either state A or B. This may seem to suggest that there are four different values in the genetic state space, but if the order of the copies does not matter (which is usually the case), then AB and BA are effectively the same, so the state space consists of three values: AA, BB, and AB.

1.2.2 discrete state variables: population

Consider the model of a population of individuals, with the variable of number of individuals (populations size) and parameters being the birth and death rates. The state space of this model is **all integers between 0 and infinity.**

Consider the model of a population of individuals who may get infected. Assume that the total number of individuals does not change (that is, there are no births and deaths) and that these individuals can be in one of two states: healthy or sick (in epidemiology these are called *susceptible* or *infectious*). There are typically two parameters in such models: the probability of infection and the probability of recovery. Since the total population is fixed at some number N, the space space of the model is all pairs of integers between 0 and N that add up to N.

1.2.3 continuous state variables: concentration

Suppose that a biological molecule is produced with a certain rate and degraded with a different rate, and we would like to describe the quantity of the molecule, usually expressed as concentration. The relevant variables here are concentration and time, and you will see those variables on the axes of many plots in biochemistry. Concentration is a ratio of the number of molecules and the volume, so the state space can be any positive real number (although practically there is a limit as to how many molecules can fit inside a given volume, but for simplicity we can ignore this).

Going even further, let us consider an entire cell, which contains a large number of different molecules. We can describe the state of a cell as the collection of all the molecular concentrations, with the parameters being the rates of all the reactions going on between those molecules. The state space for this model with N different molecules is N positive real numbers.

1.2.4 multiple variables in medicine

Doctors take medical history from patients and measure vital signs to get a picture of a patient's health. These can be all be thought of as variables in a model of a person that physicians construct. Some of these variables are discrete, for instance whether there is family history of hypertension, which has only two values: yes or no. Other variables are numbers with a range, such as weight and blood pressure. The state space of this model is a combination of *categorical* values (such as yes/no) and *numerical* values (within a reasonable range).

1.2.5 Discussion questions

Several biological models are indicated below. Based on what you know, divide the quantities into variables and parameters and describe the state space of the model. Note that there may be more than one correct interpretation

1. The volume of blood pumped by the heart over a certain amount of time, depending on the heart rate and the ejection volume.
2. The number of wolves in a national forest depending on the number of wolves in the previous year, the birth rate, the death rate, and the migration rate.
3. The fraction of hemes in hemoglobin (a transport protein in red blood cells) which are bound to oxygen depending on the partial pressure of oxygen and the binding cooperativity of hemoglobin.
4. The number of mutations that occur in a genome, depending on the mutation rate, the amount of time, and the length of the genome.

5. The concentration of a drug in the blood stream depending on the dose, time after administration, and the rate of metabolism (processing) of the drug.

6. Describing an outbreak of an infectious disease in a city in terms of the fractions of infected, healthy, and recovered people, depending on the rate of infection, rate of recovery, and the mortality rate of the disease.

1.3 First steps in R

A central goal of this book is to help you, the reader, gain experience with computation, which requires learning some programming (cool kids call it "coding"). Programming is a way of interacting with computers through a symbolic language, unlike the graphic user interfaces that we're all familiar with. Basically, programming allows you to make a computer do exactly what you want it to do.

There is a vast number of computer languages with distinct functionalities and personalities. Some are made to talk directly to the computer's "brain" (CPU and memory), e.g. Assembly, while others are better suited for human comprehension, e.g. python or Java. Programming in any language involves two parts: 1) writing a program (code) using the commands and the syntax for the language; 2) running the code by using a compiler or interpreter to translate the commands into machine language and then making the computer execute the actions. If your code has a mistake in it, the compiler or interpreter should catch it, and return an *error message* to you instead of executing the code. Sometimes, though, the code may pass muster with the interpreter/compiler, but it may still have a mistake (bug). This can be manifested in two different ways: either the code execution does not produce the result that you intended, or it hangs up or crashes the computer (the latter is hard to do with the kind of programming we will be doing). We will discuss errors and how to prevent and catch these bugs as you develop your programming skills.

In this course, our goal is to compute mathematical models and to analyze data, so we choose a language that is designed specially for these tasks, which is called R. To proceed, you'll need to download and install R, which is freely available here. In addition to downloading the language (which includes the interpreter that allows you to run R code on your computer) you will need to download a graphic interface for writing, editing, and running R code, called R Studio (coders call this an IDE, or an Integrated Developer Environment), which is also free and available here.

1.3.1 R Markdown and R Studio

In this course you will use R using R Studio and R Markdown documents, which are text files with the extension `.Rmd`. Markdown is a simple formatting syntax for creating reports in HTML, PDF, or Word format by incorporating text with code and its output. More details on using R Markdown are here. In fact, this whole book is written in R Markdown files and then compiled to produce the beautiful (I hope you agree) web book that you are reading.

If you open an Rmd file in R Studio, you will see a **Knit** button on top of the Editor window. Clicking it initiates the processing of the file into an output document (in HTML, PDF, or Word format) that includes the text as well as the output of any embedded R code chunks within the document. You can embed an R *code chunk* like this:

```r
print("Hello there!")
```

```
[1] "Hello there!"
```

To run the code inside a single R code chunk, click the green arrow in the top right of the chunk. This will produce an output, in this case the text "Hello there!". Inside the generated output file, for example the web book you may be reading, the output of code chunks is shown below the box with the R code and indicated by two hashtags.

You can make text **bold** or *italic* like so. You can also use mathematical notation called LaTeX, which you'll see used below to generate nice-looking equations. LaTeX commands are surrounded by dollar signs, for example \$e^x\$ generates e^x. Mathematical types love LaTeX, but you can use R Markdown without it.

1.3.2 numbers and arithmetic operations

When you get down to the brass tacks, all computation rests on performing *arithmetic operations*: addition, subtraction, multiplication, division, exponentiation, etc. The symbols used for arithmetic operations are what you'd expect: `+`, `-`, `*`, `/` are the four standard operations, and `^` is the symbol for exponentiation. For example, type `2^3` in any R code chunk and execute it:

```r
2^3
```

```
[1] 8
```

You see that R returns the result by printing it out on the screen. The number in square brackets [1] is not important for now; it is useful when the answer contains many numbers and has to be printed out on many rows. The second number is the result of the calculation.

For numbers that are either very large or very small, it's too cumbersome to write out all the digits, so R, like most computational platforms, uses the *scientific notation*. For instance, if you want to represent 1.4 billion, you type in the following command; note that 10 to the ninth power is represented as `e+09` and the prefix 1.4 is written without any multiplication sign:

```
1.4*10^9
```

```
[1] 1.4e+09
```

There are also certain numbers built into the R language, most notably π and e, which can be accessed as follows:

```
pi
```

```
[1] 3.141593
```

```
exp(1)
```

```
[1] 2.718282
```

The expression `exp()` is an example of a function, which we will discuss in section **??**; it returns the value of e raised to the power of the number in parenthesis, hence `exp(1)` returns e. Notice that although both numbers are irrational, and thus have infinitely many decimal digits, R only prints out a few of them. This doesn't mean that it doesn't have more digits in memory, but it only displays a limited number to avoid clutter. The number of digits to be displayed can be changed, for example to display 10 digits, type in `options(digits=10)`.

Computers are very good at computation, as their name suggests, but they have limitations. In order to manipulate numbers, they must be stored in computer memory, but computer memory is finite. There is a limit to the length of the number that is feasible to store on a computer. This has implications for both very large numbers and to very small numbers, which are close to zero, because both require many digits for storage.

All programming languages have an upper limit on the biggest number it will store and work with. If an arithmetic operation results in a number larger than that limit, the computer will call it an *overflow* error. Depending on the language, this may stop the execution of the program, or else produce a non-numerical value, such as `NaN` (not a number) or `Inf` (infinite). Do exercise **??** to investigate the limitations of R for large numbers.

On the other hand, very small numbers present their own challenges. As with very large numbers, a computer cannot store an arbitrary number of digits after the decimal (or binary) point. Therefore, there is also the smallest number that a programming language will accept and use, and storing a smaller number produces an *underflow* error. This will either cause the program execution to stop, or to return the value 0 instead of the correct answer. Do exercise **??** to investigate the limitations of R for small numbers.

This last fact demonstrates that all computer operations are imprecise, as they are limited by what's called the *machine precision*, which is illustrated in exercise **??**. For instance, two similar numbers, if they are within the machine precision of one another, will be considered the same by the computer. Modern computers have large memories, and their machine precision is very good, but sometimes this error presents a problem, e.g. when subtracting two numbers. A detailed discussion of machine error is beyond the scope of this text, but anyone performing computations must be aware of its inherent limitations.

1.3.3 R Coding Exercises

1. Calculate the value of π raised to the 10th power.
2. Use the scientific notation to multiply four billion by π.
3. Use the scientific notation with large exponents (e.g. 1e+100, 1e+500, etc.) to find out what happens when you give R a number that is too large for it to handle. Approximately at what order of magnitude does R produce an overflow error?
4. In the same fashion, find out what happens when you give R a number that is too small for it to handle. Approximately at what order of magnitude does R produce an underflow error?
5. How close can two numbers be before R thinks they are the same? Subtract two numbers which are close to each other, like 24 and 24.001, and keep making them closer to each other, until R returns a difference of zero. Report at what value of the actual difference this happens.

1.3.4 variable assignment

Variables in programming languages are used to store and access numerical or other information. After *assigning}* it a value for the first time (*initializing*), a variable name can be used to represent the value we assigned to it. Invoking the name of variable recalls the stored value from computer's memory. There are a few rules about naming variables: a name cannot be a number or an arithmetic operator like `+`, in fact it cannot contain symbols for operators or spaces inside the name, or else confusion would reign. Variable names may contain numbers, but not as the first character. When writing code it is good practice to give variables informative names, like *height* or *city_pop*.

The symbol '=' is used to assign a value to a variable in most programming languages, and can be used in R too. However, it is customary for R to use the symbols `<-` together to indicate assignment, like this:

```r
var1 <- 5
```

After this command the variable `var1` has the value 5, which you can see in the upper right frame in R Studio called *Environment*. In order to display the value of the variable as an output on the screen, use the special command `print()` (it's actually a function, which we will discuss in the next chapter). The following two commands show that the value of a variable can be changed after it has been initialized:

```
var1 <- 5
var1 <- 6
print(var1)
```

```
[1] 6
```

While seemingly contradictory, the commands are perfectly clear to the computer: first `var1` is assigned the value 5 and then it is assigned 6. After the second command, the first value is forgotten, so any operations that use the variable `var1` will be using the value of 6.

Entire expressions can be placed on the right hand side of an assignment command: they could be arithmetic or logical operations as well as functions, which we will discuss later on. For example, the following commands result in the value 6 being assigned to the variable `var2`:

```
var1 <- 5
var2 <- var1+1
print(var2)
```

```
[1] 6
```

Even more mind-blowing is that the same variable can be used on both sides of an assignment operator! The R interpreter first looks on the right hand side to evaluate the expression and then assigns the result to the variable name on the left hand side. So for instance, the following commands increase the value of `var1` by 1, and then assign the product of `var1` and `var2` to the variable `var2`:

```
var1 <- var1 + 1
print(var1)
```

```
[1] 6
```

```
var2 <- var1-1
print(var2)
```

```
[1] 5
```

```
var2 <- var1*var2
print(var2)
```

```
[1] 30
```

We have seen example of how to assign values to variables, so here is an example of how NOT to assign values, with the resulting error message:

```
var1 + 1 <- var1
```

The left-hand side of an assignment command should contain only the variable to which you are assigning a value, not an arithmetic expression to be performed.

1.3.5 R Coding Exercises

The following commands or scripts do not work as intended. Find the errors and correct them, then run them to make sure they do what they are intended to do:

1.3.6 Exercises:

The following R commands or short scripts contain errors; your job is to fix them so they runs as described. (Remove the # at the start of each line to "uncomment" the code first.)

1. Assign the value -10 to a variable

```
neg -> -10
```

2. Assign a variable the value 5 and then increase its value by 3:

```
2pac <- 5
2pac <- 2pac + 3
```

3. Assign the values 4 and 7 to two variables, then add them together and assign the sum to a new variable:

```
total <- part1 + part2
part1 <- 4
part2 <- 7
```

4. Add 5 and 3 and save it into variable my.number

```
5 + 3 <- my.number
```

5. Print the value of my.number on the screen:

```
print[my.number]
```

6. Replace the value of my.number with 5 times its current value

```
my.number <- 5my.number
```

7. Assign the values of 7 and 8 to variables a and b, respectively, multiply them and save the results in variable x

```
a<-7
b<-8
x<-ab
print(x)
```

9. Assign the value 42 to a variable, then increase it by 1

```
age <- 42
age + 1 <- age
```

10. Assign the value 10 to variable radius, then calculate the area of the circle with that radius using the formula $A = \pi r^2$:

```
r <- 10
area <- pir^2
```

2 Functions and their graphs

> Some fathers, if you ask them for the time of day, spit silver dollars.
> –Donald Barthelme, *The Dead Father*

Mathematical models describe how various quantities affect each other. In the last chapter we learned that these descriptions can be written down, often in the form of an equation. For instance, we can describe the total volume of blood pumped over a period of time as the product of stroke volume, the heart rate and the number of minutes, which can be written as an equation. The different quantities have their own meaning and roles, depending on what they stand for. To better describe how these quantities are related we use the deep idea of mathematical functions. In this chapter you will learn to do the following:

- use dimensional analysis to deduce the meaning of quantities in a model
- understand the concept of function, dependent and independent variables
- recognize basic functional forms and the shape of their graphs
- use R to plot functions
- understand basic models of reaction rates

2.1 Dimensions of quantities

What distinguishes a mathematical model from a mathematical equation is that the quantities involved have a real-world meaning. Each quantity represents a measurement, and associated with each one are the *units* of measurement. The number 173 is not enough to describe the height of a person - you are left to wonder 173 what? meters, centimeters, nanometers, light-years? Obviously, only centimeters make sense as a unit of measurement for human height; but if we were measuring the distance between two animals in a habitat, meters would be a reasonable unit, and it were the distance between molecules in a cell, we would use nanometers. Thus, any quantity in a mathematical model must have associated units, and any graphs of these quantities must be labeled accordingly.

In addition to units, each variable and parameter has a meaning, which is called the *dimension* of the quantity. For example, any measurement of length or distance has the same dimension, although the units may vary. The value of a quantity depends on the units of measurement, but its essential dimensionality does not. One can convert a measurement in meters to that in light-years or cubits, but one cannot convert a measurement in number of sheep to seconds - that conversion has no meaning.

Thus leads us to the fundamental rule of mathematical modeling: **terms that are added or subtracted must have the same dimension**. This gives mathematical modelers a useful tool called *dimensional analysis*, which involves replacing the quantities in an equation with their dimensions. This serves as a check that all dimensions match, as well as allowing to deduce the dimensions of any parameters for which the dimension was not specified. [**?**]

Example. As we saw in chapter 1, the relationship between the amount blood pumped by a heart in a certain amount of time is expressed in the following equation, where V_{tot} and V_s are the total volume and stroke volume, respectively, R is the heart rate, and t is the time:

$$V_{tot} = V_s R t$$

The dimension of a quantity X is denoted by $[X]$; for example, if t has the dimension of time, we write $[t] = time$. The dimension of volume is $[V_{tot}] = length^3$, the dimension of stroke volume is $[V_s] = volume/beat$ and the dimension of time t is time, so we can re-write the equation above in dimensional form:

$$length^3 = length^3/beat \times R \times time$$

Solving this equation for R, we find that it must have the dimensions of $[R] = beats/time$. It can be measured in beats per minute (typical for heart rate), or beats per second, beats per hour, etc. but the *dimensionality* of the quantity cannot be changed without making the model meaningless.

There are also *dimensionless* quantities, or pure numbers, which are not tied to a physical meaning at all. Fundamental mathematical constants, like π or e, are classic examples, as are some important quantities in physics, like the Reynolds number in fluid mechanics. [**?**] Quantities with a dimension can be made dimensionless by dividing them by another quantity with the same dimension and "canceling" the dimensions. For instance, we can express the height of a person as a fraction of the mean height of the population; then the height of a tall person will become a number greater than 1, and the height of a short one will become less than 1. This new dimensionless height does not have units of length - they have been divided out by the mean height. This is known as *rescaling* the quantity, by dividing it by a preferred scale. There is a fundamental difference between rescaling and changing the units of a quantity: when changing the units, e.g. from inches to centimeters, the dimension remains the same, but if one divides the quantity by a scale, it loses its dimension.

Example. The model for a population of bacteria that doubles every hour is described by the equation, where P_0 is initial number of bacteria and P is the population after t hours:

$$P = P_0 2^t$$

Let us define the quantity $R = P/P_0$, so we can say that population increased by a factor of R after t hours. This ratio is a dimensionless quantity because P and P_0 have the same dimension of bacterial population, which cancel out. The equation for R can be written as follows:

$$R = 2^t$$

According to dimensional analysis, both sides of the equation have to be dimensionless, so t must also be a dimensionless variable. This is surprising, because t indicates the number of hours the bacterial colony has been growing. This reveals the subtle fact that t is a rescaled variable obtained by dividing the elapsed time by the length of the reproductive cycle. Because of the assumption that the bacteria divide exactly once an hour, t counts the number of hours, but if they divided once a day, t would denote the number of days. So t doesn't have units or dimensions, but instead denotes the dimensionless number of cell divisions.

2.1.1 Exercises

For each biological model below determine the dimensions of the parameters, based on the given dimensions of the variables.

1. Model of number of mutations M as a function of time t:
$$M(t) = M_0 + \mu t$$

2. Model of molecular concentration C as a function of time t:
$$C(t) = C_0 e^{-kt}$$

3. Model of tree height H (length) as a function of age a (time):
$$H(a) = \frac{ba}{c+a}$$

4. Model of cooperative binding of ligands, with fraction of bound receptors θ as a function of ligand concentration L:
$$\theta(L) = \frac{L^n}{L^n + K_d}$$

5. Model of concentration of a gene product G (concentration) as a function of time t:
$$G(t) = G_m(1 - e^{-\alpha t})$$

6. Michaelis-Menten model of enzyme kinetics, v is reaction rate (1/time) and S is substrate concentration:
$$v(S) = \frac{v_{max} S}{K_m + S}$$

7. Logistic model of population growth, P is population size and time t:
$$P(t) = \frac{Ae^{kt}}{1 + B(e^{kt} - 1)}$$

2.2 Functions and their graphs

A relationship between two variables addresses the basic question: when one variable changes, how does this affect the other? An equation, like the examples in the last section, allows one to calculate the value of one variable based on the other variable and parameter values. In this section we seek to describe more broadly how two variables are related by using the mathematical concept of functions.

ℹ Definition

A function is a mathematical rule which has an input and an output. A function returns a well-defined output for every input, that is, for a given input value the function returns a unique output value.

In this abstract definition of a function it doesn't have to be written as an algebraic equation, it only has to return a unique output for any given input value. In mathematics we usually write them down in terms of algebraic expressions. As in mathematical models, you will see two different kinds of quantities in equations that define functions: variables and parameters. The input and the output of a function are usually variables, with the input called the *independent variable* and the output called the *dependent variable*.

The relationship between the input and the output can be graphically illustrated in a graph, which is a collection of paired values of the independent and dependent variable drawn as a curve in the plane. Although it shows how the two variables change relative to each other, parameters may change too, which results in a different graph of the function. While graphing calculators and computers can draw graphs for you, it is very helpful to have an intuitive understanding about how a function behaves, and how the behavior depends on the parameters. Here are the three questions to help picture the relationship (assume x is the independent variable and it is a nonnegative real number):

1. what is the value of the function at $x = 0$?
2. what does the function do when x becomes large ($x \to \infty$)?
3. what does the function do between the two extremes?

Below you will find examples of fundamental functions used in biological models with descriptions of how their parameters influence their graphs.

2.2.1 linear and exponential functions

The reader is probably familiar with linear and exponential functions from algebra courses. However, they are so commonly used that it is worth going over them to refresh your memory and perhaps to see them from another perspective.

ℹ Definition

A *linear function* $f(x)$ is one for which the difference in two function values is the same for a specific difference in the independent variable.

In mathematical terms, this can be written an equation for any two values of the independent variable x_1 and x_2 and a difference Δx:

$$f(x_1 + \Delta x) - f(x_1) = f(x_2 + \Delta x) - f(x_2)$$

The general form of the linear function is written as follows:

$$f(x) = ax + b \tag{2.1}$$

The function contains two parameters: the slope a and the y-intercept b. The graph of the linear function is a line (hence the name) and the slope a determines its steepness. A positive slope corresponds to the graph that increases as x increases, and a negative slope corresponds to a declining function. At $x = 0$, the function equals b, and as $x \to \infty$, the function approaches positive infinity if $a > 0$, and approaches negative infinity if $a < 0$.

ℹ Definition

An *exponential function* $f(x)$ is one for which the ratio of two function values is the same for a specific difference in the independent variable.

Mathematically speaking, this can be written as follows for any two values of the independent variable x_1 and x_2 and a difference Δx:

$$\frac{f(x_1 + \Delta x)}{f(x_1)} = \frac{f(x_2 + \Delta x)}{f(x_2)}$$

Exponential functions can be written using different symbolic forms, but they all have a constant base with the variable x in the exponent. I prefer to use the constant e (base of the natural logarithm) as the base of all the exponential functions, for reasons that will become apparent in chapter 15. This does not restrict the range of possible functions, because any exponential function can be expressed using base e, using a transformation: $a^x = e^{x \ln(a)}$. So let us agree to write exponential functions in the following form:

$$f(x) = ae^{rx} \tag{2.2}$$

The function contains two parameters: the *rate constant* r and the multiplicative constant a. The graph of the exponential function is a curve which crosses the y-axis at $y = a$ (plug in $x = 0$ to see that this is the case). As x increases, the behavior of the graph depends on the sign of the rate constant r. If $r > 0$, the function approaches infinity (positive if $a > 0$, negative if $a < 0$) as $x \to \infty$. If $r < 0$, the function decays at an ever-decreasing pace and asymptotically approaches zero as $x \to \infty$. Thus the graph of $f(x)$ is a curve either going to infinity or a curve asymptotically approaching 0, and the steepness of the growth or decay is determined by r.

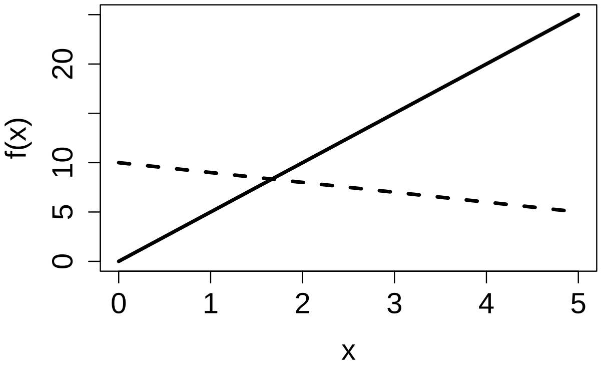


Figure 2.1: Plots of two linear functions (left) and two exponential functions (right). Can you identify which linear function has the positive slope and which one negative? Which exponential function has a positive rate constant and which one negative?

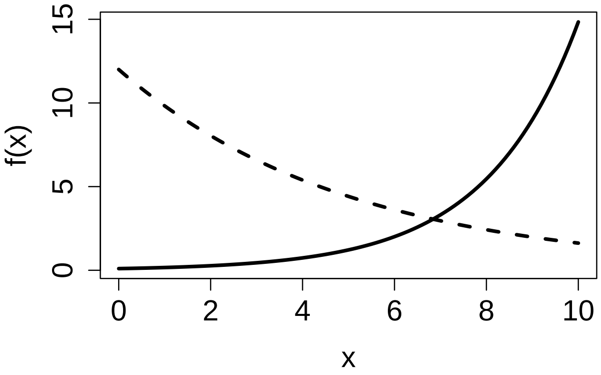


Figure 2.2: Plots of two linear functions (left) and two exponential functions (right). Can you identify which linear function has the positive slope and which one negative? Which exponential function has a positive rate constant and which one negative?

2.2.2 Exercises

Answer the questions below, some of which refer to the function graphs in figure @ref(ch2-funk1).

1. Which of the linear graphs in the first figure corresponds to $f(x) = 5x$ and which

corresponds to $f(x) = 10 - x$? State which parameter allows you to connect the function with its graph and explain why.

2. Which of the exponential graphs in the second figure corresponds to $f(x) = 0.1e^{0.5x}$ and which corresponds to $f(x) = 12e^{-0.2x}$? State which parameter allows you to connect the function with its graph and explain why.
3. Demonstrate algebraically that a linear function of the form given in equation **??** satisfies the property of linear functions from definition **??**.
4. Demonstrate algebraically that an exponential function of the form given in equation **??** satisfies the property of exponential functions from definition **??**.
5. Modify the exponential function by adding a constant term to it $f(x) = ae^{rx} + b$. What is is the value of this function at $x = 0$?
6. How does the function defined in the previous exercise, $f(x) = ae^{rx} + b$, how does it behave as $x \to \infty$ if $r > 0$?
7. How does the function $f(x) = ae^{rx} + b$ behave as $x \to \infty$ if $r < 0$?

2.2.3 rational and logistic functions

Let us now turn to more complex functions, made up of simpler components that we understand. Consider a ratio of two polynomials, called a rational function. The general form of such functions can be written down as follows, where ellipsis stands for terms with powers lower than n or m:

$$f(x) = \frac{a_0 + ... + a_n x^n}{b_0 + ... + b_m x^m} \tag{2.3}$$

The two polynomials may have different degrees (highest power of the terms, n and m), but they are usually the same in most biological examples. The reason is that if the numerator and the denominator are "unbalanced' ', one will inevitably overpower the other for large values of x, which would lead to the function either increasing without bound to infinity (if $n > m$) or decaying to zero (if $m > n$). There's nothing wrong with that, mathematically, but rational functions are most frequently used to model quantities that approach a nonzero asymptote for large values of the independent variable.

For this reason, let us assume $m = n$ and consider what happens as $x \to \infty$. All terms other than the highest-order terms become very small in comparison to x^n (this is something you can demonstrate to yourself using R), and thus both the numerator and the denominator approach the terms with power n. This can be written using the mathematical limit notation $\lim_{x\to\infty}$ which describes the value that a function approaches when the independent variable increases without bound:

$$\lim_{x\to\infty} \frac{a_0 + ... + a_n x^n}{b_0 + ... + b_n x^n} = \frac{a_n x^n}{b_n x^n} = \frac{a_n}{b_n}$$

Therefore, the function approaches the value of a_n/b_n as x grows.

Similarly, let us consider what happens when $x = 0$. Plugging this into the function results in all of the terms vanishing except for the constant terms, so

$$f(0) = \frac{a_0}{b_0}$$

Between 0 and infinity, the function either increases or decreases monotonically, depending on which value (a_n/b_n or a_0/b_0) is greater. Two examples of plots of rational functions are shown in figure **??**, which shows graphs increasing from 0 to 1. Depending on the degree of the polynomials in a rational function, it may increase more gradually (solid line) or more step-like (dashed line).

Example. The following model, called the Hill equation , describes the fraction of receptor molecules which are bound to a ligand, which is a chemical term for a free molecule that binds to another, typically larger, receptor molecule. θ is the fraction of receptors bound to a ligand, L denotes the ligand concentration, K_d is the dissociation constant, and n called the binding cooperativity or Hill coefficient:

$$\theta = \frac{L^n}{L^n + K_d}$$

The Hill equation is a rational function, and Figure **??** shows plots of the graphs of two such function in the right panel. This model is further explored in exercise 2.2.10.

Example. A common model of population over time is the logistic function. There are variations on how it is written down, but here is one general form:

$$f(x) = \frac{ae^{rx}}{b + e^{rx}} \tag{2.4}$$

The numerator and denominator both contain exponential functions with the same power. If $r > 0$ when $x \to \infty$, the denominator approaches e^{rx}, since it becomes much greater than b, and we can calculate:

$$\lim_{x\to\infty} = \frac{ae^{rx}}{e^{rx}} = a; \text{ if } r > 0$$

On the other hand, if $r < 0$, then the numerator approaches zero as $x \to \infty$, and so does the function

$$\lim_{x\to\infty} = \frac{0}{b} = 0; \text{ if } r < 0$$

Notice that switching the sign of r has the same effect as switching the sign of x, since they are multiplied. Which means that for positive r, if x is extended to negative infinity, the function approaches 0. This is illustrated in the second plot in Figure **??**, which shows two logistic functions increasing from 0 to a positive level, one with $a = 20$ (solid line) and the second with $a = 10$ (dashed line). The graph of logistic functions has a characteristic *sigmoidal* (S-shaped) shape, and its steepness is determined by the rate r: if r is small, the curve is soft, if r is large, the graph resembles a step function.

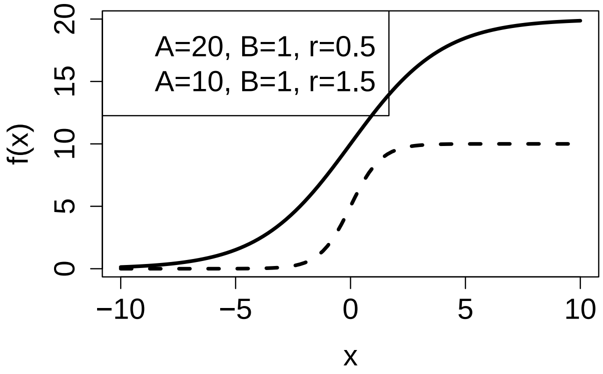


Figure 2.3: Examples of two graphs of logistic functions (left) and two Hill functions (right).

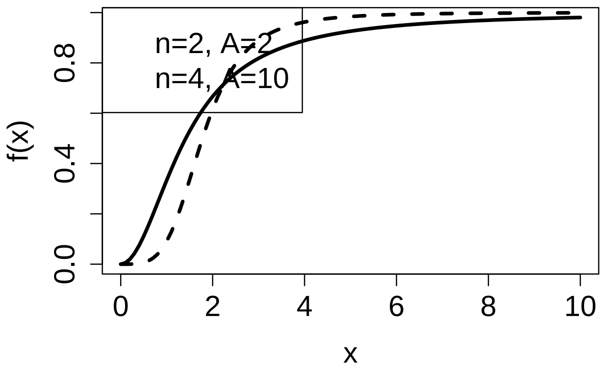


Figure 2.4: Examples of two graphs of logistic functions (left) and two Hill functions (right).

2.2.4 Exercises:

For each biological model below answer the following questions in terms of the parameters in the models, assuming all are nonnegative real numbers. 1) what is the value of the function when the independent variable is 0? 2) what value does the function approach when the independent variable goes to infinity? 3) verbally describe the behavior of the functions between 0 and infinity (e.g., function increases, decreases).

1. Model of number of mutations M as a function of time t:
$$M(t) = M_0 + \mu t$$

2. Model of molecular concentration C as a function of time t:
$$C(t) = C_0 e^{-kt}$$

3. Model of cooperative binding of ligands, with fraction of bound receptors θ as a function of ligand concentration L:
$$\theta = \frac{L^n}{L^n + K_d}$$

4. Model of tree height H (length) as a function of age a (time):
$$H(a) = \frac{ba}{c + a}$$

5. Model of concentration of a gene product G (concentration) as a function of time t:
$$G(t) = G_m(1 - e^{-\alpha t})$$

6. Michaelis-Menten model of enzyme kinetics, v is reaction rate (1/time) and S is substrate concentration:
$$v(S) = \frac{v_{max} S}{K_m + S}$$

7. Logistic model of population growth, P is population size and time t:
$$P(t) = \frac{Ae^{kt}}{1 + B(e^{kt} - 1)}$$

2.3 Vectors and plotting in R

2.3.1 writing scripts and calling functions

Programming means arranging a number of commands in a particular order to perform a task. Typing them one at a time into the command line is inefficient and error-prone. Instead, the commands are written into a file called a program or script (the name depends on the type of language; since R is a scripting language you will be writing scripts), which can be edited, saved, copied, etc. To open a new script file, in R Studio, go to File menu, and choose New R Script. This will open an editor window where you can type your commands. To save the script file (do this often!!), click the Save button (with the little floppy disk icon) or select Save from the File menu. You will also see small buttons at the top of the window that say `Run`, `Re-run`, and `Source`. The first two will run either the current line or a selected region of the script, while the `Source` button will run the entire file. Now that you know how to create a script, **you should never type your R code into the command line**, unless you're testing a single command to see what it does, or looking up help.

R comes equipped with many functions that correspond to standard mathematical functions. As we saw in section **??**, `exp()` is the exponential function that returns e raised to the power of the input value. Other common ones are: `sqrt()` returns the square root of the input value; `sin()` and `cos()` return the sine and the cosine of the input value, respectively. Note that all of these function names are followed by parentheses, which is a hallmark of a function (in R as well as in mathematics). This indicates that the input value has to go there, for example `exp(5)`. To compute the value of e^5, save it into a variable called `var1` and then print out the value on the screen, you can create the following script:

```
var1 <- exp(5)
print(var1)
```

```
[1] 148.4132
```

If you run the above code chunk in R Studio you will see two things happen: a variable named var1 appears in the Environment window (top right) with the value 148.41… and the same value is printed out in the command line window (bottom left).

The most important principle of the procedural brand of programming (which includes R) is this: the computer (that is, the compiler or interpreter) evaluates the commands from top to bottom, one at a time. The variables are used with the values that they are currently assigned. If one variable (`var1`) was assigned in terms of another (`var2`), and then `var2` is changed later, this does not change the value of `var2`. Here is an illustration of how this works:

```
var2 <- 20
var1 <- var2/20
print(var2)
```

```
[1] 20
```

```
var2 <- 10
print(var1)
```

```
[1] 1
```

Notice that `var1` doesn't change, because the R interpreter reads the commands one by one, and does not go back to re-evaluate the assignment for var1 after `var2` is changed. Learning to think in this methodical, literal manner is crucial for developing programming skills.

2.3.2 vector variables

Variables may contain more than a single number, they can also store a bunch of numbers, which is then called an array. When numbers in an array are organized as a single ordered list, this is called a *vector*. There are several ways of producing a vector of numbers in R.

2.3.2.1 c() function

The most direct method of making a vector is to put together several values by listing them inside the function `c()` and assigning the output to a variable, e.g. `my.vec`:

```
my.vec<-c(pi,45,912.8, 0)
print(my.vec)
```

```
[1]   3.141593  45.000000 912.800000   0.000000
```

This variable `my.vec` is now a vector variable that contains four different numbers. Each of those numbers can be accessed individually by referencing its position in the vector, called the *index*. In the R language the the index for the first number in a vector is 1, the index for the second number is 2, etc. The index is placed in square brackets after the vector name, as follows:

```
print(my.vec[1])
```

```
[1] 3.141593
```

```
print(my.vec[2])
```

```
[1] 45
```

```
print(my.vec[3])
```

```
[1] 912.8
```

```
print(my.vec[4])
```

```
[1] 0
```

2.3.2.2 the colon operator

Another way to generate a sequence of numbers in a particular order is to use the colon operator, which produces a vector of integers from the first number to the last, inclusive. Here are two examples:

```
my.vec1<-1:20
print(my.vec1)
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

```
my.vec2<-0:-20
print(my.vec2)
```

```
 [1]   0  -1  -2  -3  -4  -5  -6  -7  -8  -9 -10 -11 -12 -13 -14 -15 -16 -17 -18
[20] -19 -20
```

You can also access some but not all of the values stored in a vector simultaneously. To do this, enter a vector of positive integers inside the square brackets, either using the colon operator or using the `c()` function. Here are two examples, the first prints out the 4th through the 10th element of the vector `my.vec1`, while the second prints out the 1st, 5th, and 11th elements of the vector `my.vec2`:

```
print(my.vec1[4:10])
```

```
[1]  4  5  6  7  8  9 10
```

```
print(my.vec2[c(1,5,11)])
```

```
[1]   0  -4 -10
```

2.3.2.3 seq() function

If you want to generate a sequence of numbers with a constant difference other than 1, you're in luck: R provides a function called `seq()`. It takes three inputs: the starting value, the ending value, and the step (difference between successive elements). For example, to generate a list of numbers starting at 20 up to 50, with a step size of 3, type the first command; to obtain the same sequence in reverse, use the second command:

```
my.vec1<-seq(20,50,3)
print(my.vec1)
```

```
 [1] 20 23 26 29 32 35 38 41 44 47 50
```

```
my.vec2<-seq(50,20,-3)
print(my.vec2)
```

```
 [1] 50 47 44 41 38 35 32 29 26 23 20
```

2.3.2.4 rep() function

Sometimes you want to create a vector of repeated values. For example, you can create a variable with 20 zeros, you can use rep() like this:

```
zeros <- rep(0,20)
print(zeros)
```

```
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

You can repeat any value, say create a vector by repeating the number pi:

```
pies <- rep(pi,7)
print(pies)
```

```
[1] 3.141593 3.141593 3.141593 3.141593 3.141593 3.141593 3.141593
```

You can even repeat another vector, like the vector my.vec that was assigned above:

```
my.vecs <- rep(my.vec, 5)
print(my.vecs)
```

```
 [1]   3.141593  45.000000 912.800000   0.000000   3.141593  45.000000
 [7] 912.800000   0.000000   3.141593  45.000000 912.800000   0.000000
[13]   3.141593  45.000000 912.800000   0.000000   3.141593  45.000000
[19] 912.800000   0.000000
```

2.3.3 calculations with vector variables

```
NewVec <- 2*my.vec
print(NewVec)
```

```
[1]    6.283185   90.000000 1825.600000    0.000000
```

You can also perform calculations with multiple vector variables, but this requires extra care. R can perform any arithmetic operation with two vector variables, for instance adding two vectors results in a vector containing the sum of corresponding elements of the two vectors:

```
my.vec1<-1:5
my.vec2<-0:4
print(my.vec1)
```

```
[1] 1 2 3 4 5
```

```
print(my.vec2)
```

```
[1] 0 1 2 3 4
```

```
sum.vec<-my.vec1+my.vec2
print(sum.vec)
```

```
[1] 1 3 5 7 9
```

One needs to take care that the two vectors have the same number of elements (length). If you try to operate on (e.g. add) two vectors of different lengths, R will return a warning and the result will not be what you expect:

```
my.vec1<-1:2
my.vec2<-0:4
print(my.vec1)
```

```
[1] 1 2
```

```
print(my.vec2)
```

```
[1] 0 1 2 3 4
```

```
sum.vec<-my.vec1+my.vec2
```

```
Warning in my.vec1 + my.vec2: longer object length is not a multiple of shorter
object length
```

```
print(sum.vec)
```

```
[1] 1 3 3 5 5
```

2.3.4 Exercises

The following R commands or short scripts contain errors; your job is to fix them so they runs as described.

1. Assign a vector of three numbers to a variable:

```
date_num <- (3,8,16)
```

2. Assign a range of values to a vector variable and print out the third one:

```
the.vals <- 0:10
print[the.vals(3)]
```

3. Assign a range of values to a vector variable and print out the fourtieth and sixty-first values:

```
all.the.vals <- 0:100
print(all.the.vals[40,61])
```

4. Take the two vectors assigned above and assign their product to another vector:

```
product <- the.vals*all.the.vals
```

5. Create a vector vec1 of ten integers and print the second and the eighth elements:

```
vec1 <- 11:20
print(vec1[2:8])
```

6. Create a vector `vec1` and then multiply all of its elements by 20 and assign it to another vector:

```
vec1<-seq(-3,2,0.1)
vec2 <- 20vec1
```

7. Create a vector `vec1`, a vector `vec2` and print out all the elements of the first divided by the second:

```
vec1 <- 0:5
vec2 <- 3:8
print[vec1/vec2]
```

2.3.5 Plotting with vectors

```
curve(x^2, 0, 10, lwd = 3, xlab = "x", ylab = "quadratic",
    cex.axis = 1.5, cex.lab = 1.5)
curve(20 * exp(-0.5 * x), 0, 5, lwd = 3, xlab = "x",
    ylab = "exponential", cex.axis = 1.5, cex.lab = 1.5)
```

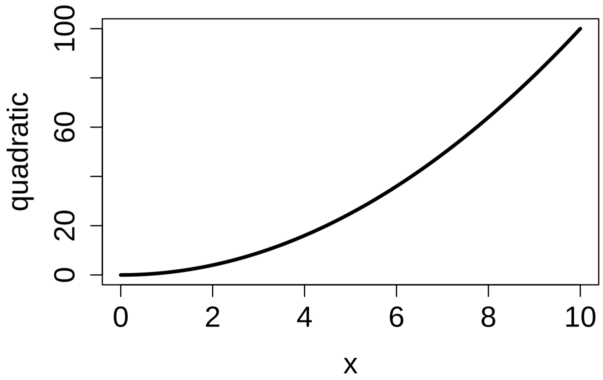


Figure 2.5: Two examples of plots using curve: quadratic ($y = x^2$) and exponential ($y = 20 * e^{-0.5x}$)

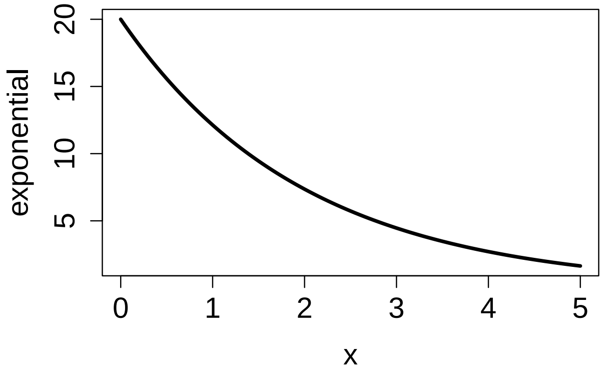


Figure 2.6: Two examples of plots using curve: quadratic ($y = x^2$) and exponential ($y = 20 * e^{-0.5x}$)

There are several ways of creating plots of mathematical functions or data R. If you want to plot a mathematical function, the simplest function is `curve()`. You can tell that this is a function, because it uses parentheses; the first input is an expression for the function, and the next two define the range of the independent variable over which to plot the graph. Two examples of plotting a quadratic function over the range from 0 and 5, and an exponential variation over the range of 0 to 10 are shown in figure **??**.

One can change the default look of the plot produced by curve by setting different options, which are optional inputs into the curve function, One is the line width `lwd` which can be increased from the default value of 1 to produce thicker curves, as demonstrated in the example above. One can add labels on the x and y axes with `xlab` and `ylab` options, respectively; note that these are strings of characters, and thus must be put in quotes to differentiate them from a variable name. There is one very important option not shown above: that of overlaying a curve on top of an existing plot, which is done by typing `add=TRUE`. This option takes logical (Boolean) values `TRUE` and `FALSE`, which must be typed in all caps and without quotes.

2.3.5.1 plot() function

In addition to curve, one can use the function `plot()` in R to create two dimensional graphs from two vector-valued variables of the same length, e.g. `plot(x,y)`. The first input variable corresponds to the *independent variable* (e.g. `x`), which is plotted on the x-axis, and the second variable corresponds to the *dependent variable* (e.g. `y`) which is plotted on the y-axis. In figure **??** you see graphs of exponential and logistic function plotted using `plot()`.

The following chunk creates a vector variable time, then calculates a new variable quad using time in a single operation:

```r
time <- 0:10
quad <- (time - 5)^2
print(time)
```

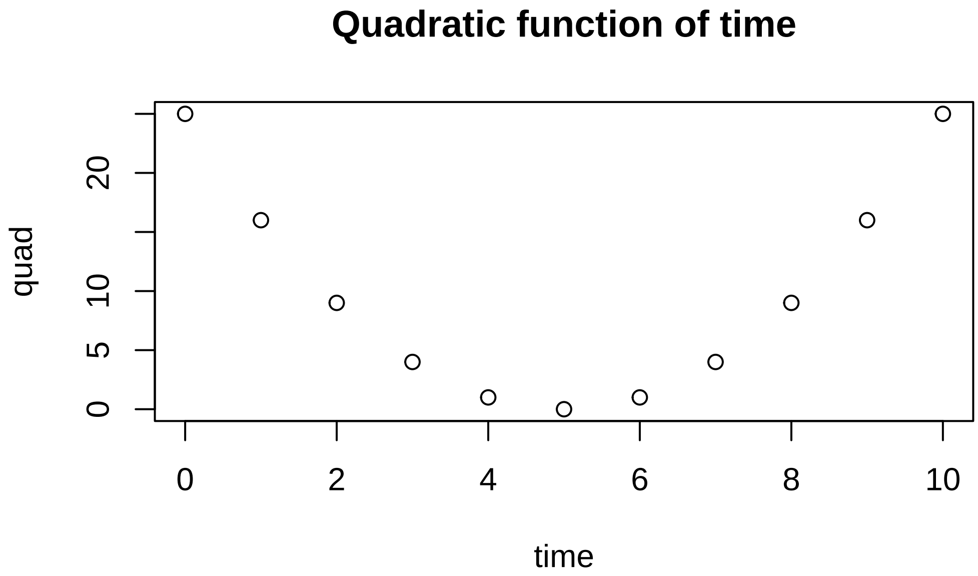
```
[1]  0  1  2  3  4  5  6  7  8  9 10
```

```r
print(quad)
```

```
[1] 25 16  9  4  1  0  1  4  9 16 25
```

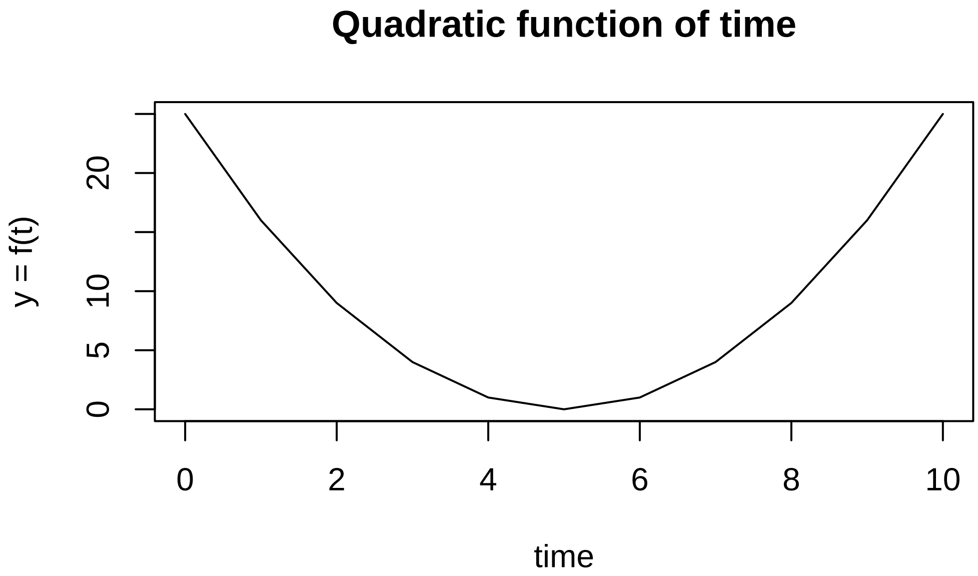
This chunk plots the two vector variables quad as a function of time, and adds a title to the plot

```r
plot(time, quad, main = "Quadratic function of time")
```



The default plot style in R uses circles to indicate each plotted point. To change it, you need to set the option `t` (type), for example, setting `t='l'` (the lowercase letter L) produces a continuous line connecting the individual data points.

```
plot(time, quad, main = "Quadratic function of time", type = "l", xlab = "time",
    ylab = "y = f(t)")
```



`plot()` is a versatile function that has many options function has many options which can be changed to determine the color, the style, and other attributes of the plot. For a full list type `help(plot)` in the console or type plot in the search bar of the Help pane in the bottom right window.

2.3.5.2 using lines() or points()

You may also want to plot multiple graphs on the same figure. The `plot()` function creates a new plot window, so if you want to add another plot on top of the first one, you have to use another function. There are two ones available: `lines()` which produces continuous curves connecting the points, and `points()` which plots individual symbols at every point.

Let us illustrate this by plotting two different exponential functions on one plot, and two different logistic functions on the second one, which were discussed in section **??**. When you've got multiple plots on the same figure, they need to be distinct and labeled. To distinguish them, below I use the option `col` to specify the color of the plot, and I add a legend describing the parameters of each plot to the figure **??**. The function has a lot of options, so if you want to understand the details, type `help(legend)` in the prompt or go to Help tab in the lower right frame of R Studio and type legend.

```r
x <- seq(0, 10, 0.5)
y <- 10 + 20 * exp(-0.5 * x)
plot(x, y, xlab = "x", ylab = "exponential", col = 1, lwd = 3)
y <- 10 + 20 * exp(-2 * x)
lines(x, y, col = 2, lwd = 3)
leg.txt = c("b=10,a=20,r=-0.5", "b=10,a=20,r=-2")
legend("topright", leg.txt, col = 1:2, pch = c(1, NA), lty = c(0, 1), lwd = 3)
x <- seq(-10, 10, 1)
y <- 20 * exp(0.5 * x)/(1 + exp(0.5 * x))
plot(x, y, xlab = "x", ylab = "logistic", col = 4, lwd = 3)
y <- 20 * exp(1.5 * x)/(1 + exp(1.5 * x))
lines(x, y, col = 2, lwd = 3)
leg.txt = c("a=20,b=1,r=0.5", "a=20,b=1,r=1.5")
legend("topleft", leg.txt, col = c(4, 2), pch = c(1, NA), lty = c(0, 1), lwd = 3)
```

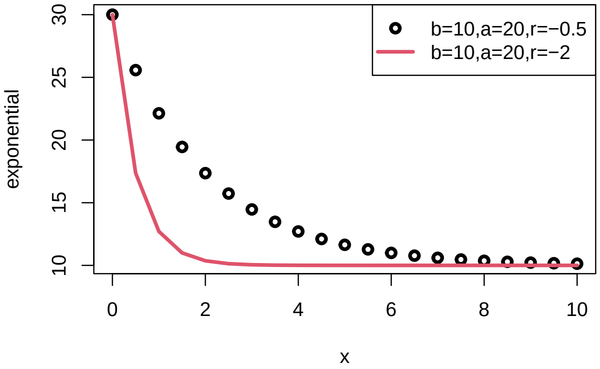


Figure 2.7: Overlaying multiple plots in R: two exponential functions of the form $y = b + ae^{rx}$ on the left, two logistic functions of the form $y = ae^{rx}/(b + e^{rx})$ on the right.

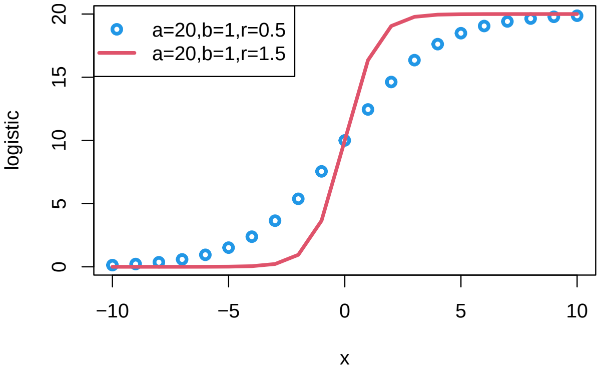


Figure 2.8: Overlaying multiple plots in R: two exponential functions of the form $y = b + ae^{rx}$ on the left, two logistic functions of the form $y = ae^{rx}/(b + e^{rx})$ on the right.

2.3.6 Exercises

The following R commands or short scripts contain errors; your job is to fix them so they runs as described.

1. Multiply a vector by a constant and add another constant and assign the result to a vector:

```
new.vals <- 5 + 8the.vals
```

2. Assign range to be a sequence of values from 0 to 100 with step of 0.1, and calculate the vector variable result as the square of the vector variable range:

```
range <- seq(0,0.1,100)
result <- square(range)
```

3. Plot result as a function of range:

```
plot(result, range)
```

4. Plot the graph of the function $f(x) = (45 - x)/(4x + 3)$ over the range of 0 to 100:

```
curve((45-x)/(4x+3), 0, 100)
```

5. Plot a quadratic function with specified coefficients a, b, c over a given range of independent variable x:

```
a<-10
b<- -15
c<- 5
y<-a*x^2+b*x+c
x<-seq(-0.5,2,0.01)
plot(x,y,type='l')
```

6. Overlay two different plots of the logistic function with different values of the parameter r:

```
time<-0:100
a<-1000
b<-50
r<-0.1
Population<-a*exp(r*time)/(b+exp(r*time))
plot(time,Population,type='l')
r<-10
lines(time,Population,col=2)
```

2.4 Rates of biochemical reactions

Living things are dynamic, they change with time, and much of mathematical modeling in biology is interested in describing these changes. Some quantities change fast and others slowly, and every dynamic quantity has a rate of change, or *rate* for short. Usually, the quantity that we want to track over time is the variable, and in order to describe how it changes we introduce a rate parameter. If we are describing changes over time, all rate parameters have dimensions with time in the denominator. As a simple example, the velocity of a physical object describes the change in distance over time, so its dimension is $[v] = length/time$.

On the most fundamental level, the work of life is performed by molecules. The protein hemoglobin transports oxygen in the red blood cells, while neurotransmitter molecules like serotonin carry signals between neurons. Enzymes catalyze reactions, like those involved in oxidizing sugar and making ATP, the energy currency of life. Various molecules bind to DNA to turn genes on and off, while myosin proteins walk along actin fibers to create muscle contractions.

In order to describe the activity of biological molecules, we must measure and quantify them. However, they are so small and so numerous that it is not usually practical to count individual molecules (although with modern experimental techniques it is sometimes possible). Instead, biologists describe their numbers using concentrations. Concentration has dimensions of number of molecules per volume, and the units are typically molarity, or moles ($\approx 6.022 * 10^{23}$ molecules) per liter. Using concentrations to describe molecule rests on the assumption that there are many molecules and they are well-mixed, or homogeneously distributed throughout the volume of interest.

Molecular reactions are essential for biology, whether they happen inside a bacterial cell or in the bloodstream of a human. *Reaction kinetics* refers to the description of the rates, or the speed, of chemical reactions. Different reactions occur with different rates, which may be dependent on the concentration of the reactant molecule. Consider a simple reaction of molecule A (called the substrate) turning into molecule B (called the product), which is usually written by chemists with an arrow:

$$A \xrightarrow{k} B$$

But how fast does the reaction take place? To write down a mathematical model, we need to define the quantities involved. First, we have the concentration of the molecule A, with dimensions of concentration. Second, we have the rate of reaction, let us call it v, which has dimension of concentration per time (just like velocity is length per time). How are the two quantities related?

2.4.1 Constant (zeroth-order) kinetics

In some circumstances, the reaction rate v does not depend on the concentration of the reactant molecule A. In that case, the relationship between the *rate constant* k and the actual rate v is:

$$v = k \tag{2.5}$$

Dimensional analysis insists that the dimension of k must be the dimension of v, or concentration/time. This is known as constant, or zero-order kinetics, and it is observed at concentrations of A when the reaction is at its maximum velocity: for example, ethanol metabolism by ethanol dehydrogenase in human liver cannot proceed any faster than about 1 drink per hour.

2.4.2 First-order kinetics

. In other conditions, it is easy to imagine that increasing the concentration of the reactant A will speed up the rate of the reaction. A simple relationship of this type is linear:

$$v = kA \tag{2.6}$$

In this case, the dimension of the rate constant k is 1/time. This is called first-order kinetics, and it usually describes reactions when the concentration of A is small, and there are plenty of free enzymes to catalyze more reactions.

2.4.3 Michaelis-Menten model of enzyme kinetics

However, if the concentration of the substrate molecule A is neither small nor large, we need to consider a more sophisticated model. An enzyme is a protein which catalyzes a biochemical reaction, and it works in two steps: first it binds the substrate, at which point it can still dissociate and float away, and then it actually catalyzes the reaction, which is usually practically irreversible (at least by this enzyme) and releases the product. The enzyme itself is not affected or spent, so it is free to catalyze more reactions. Let denote the substrate (reactant) molecule by A, the product molecule by B, the enzyme by E, and the complex of substrate and enzyme AE. The classic chemical scheme that describes these reactions is this:

$$A + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AE \xrightarrow{k_2} E + B$$

You could write three different kinetic equations for the three different arrows in that scheme. Michaelis and Menten used the simplifying assumptions that the binding and dissociation happens much faster than the catalytic reaction, and based on this they were able to write down an approximate, but extremely useful Michaelis-Menten model of an enzymatic reaction:

$$v = \frac{v_{max}A}{K_M + A} \tag{2.7}$$

Here v refers to the rate of the entire catalytic process, that is, the rate of production of B, rather than any intermediate step. Here the reaction rate depends both on the concentration of the substrate A and on the two constants v_{max}, called the maximum reaction rate, and the constant K_M, called the Michaelis constant. They both depend on the rate constants of the reaction, and v_{max} also depends on the concentration of the enzyme. The details of the derivation are beyond us for now, but you will see in the following exercises how this model behaves for different values of A.

3 Describing data sets

> Get your facts first, and then you can distort them as much as you please.
> – Rudyard Kipling, *An Interview with Mark Twain*

Science begins with experimental measurements, which are then verified by reproducing the results. But no experimental result is perfectly reproducible because all are subject to random noise, whether it is caused by unpredictable processes or is due to measurement error. Describing collections of numbers with noise is the first step to understanding the biological systems that are being measured. In this chapter you will learn to do the following:

- calculate means and medians of a data set
- calculate variances and standard deviations
- produce histograms and interpret them
- use R to plot and analyze data sets

3.1 Mutations and their rates

All Earth-based lifeforms receive an inheritance from their parent(s): a string of deoxyribonucleic acids (*DNA*) called the genetic sequence, or *genome* of an individual. The information to produce all the necessary components to build and run the organism is encoded in the sequence of the four different *nucleotides*: adenine, thymine, guanine, and cytosine (abbreviated as A, T, G, C). Different parts of the genome play different roles; some discrete chunks called *genes* contain the instructions to build *proteins*, the workhorses of biology. To make a protein from a gene, the information is transcribed from DNA into messenger ribonucleic acid (*mRNA*), which is then translated into a string of *amino acids* which constitute the protein. The genetic code determines the translation, using three nucleic acids in DNA and RNA to represent a single amino acid in a protein. Thus, a sequence of DNA results in a specific sequence of amino acids, which determine the structure and function of the protein.

The above processes involve copying and transferring information. As we know from experience, copying information inevitably means introducing errors. This is particularly important when passing information from parent to offspring, because then an entire organism has to develop and live based on a faulty blueprint. Changes introduced in the genome of an organism are called *mutations*, and they can be caused either by errors in copying DNA when making a new

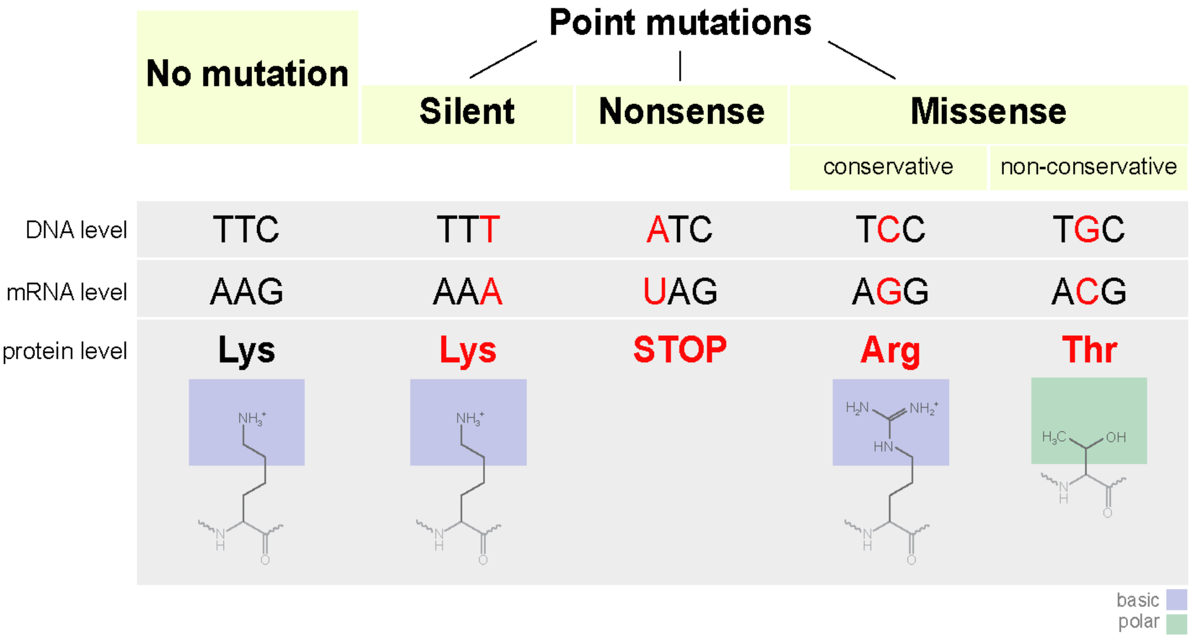


Figure 3.1: Different types of substitution point mutations are distinguished by their effects on the gene products; image by Jonsta247 in public domain via Wikimedia Commons.

cell (replication) or through damage to DNA through physical means (e.g. ionizing radiation) or chemical mechanisms (e.g. exogenous molecules that react with DNA). The simplest mutation involve a single nucleotide and are called *point mutations*. A nucleotide may be deleted, an extra nucleotide inserted, or a new one substituted instead: the three different types of substitution mutations are shown in figure ??. Large-scale mutations may involve whole chunks of the genome that are cut out and pasted in a different location, or copied and inserted in another position, but they are typically much more rare than point mutations.

Mutations can have different effects on the mutant organism, although acquisition of superpowers has not been observed. Usually, point mutations have either little observable effect or a negative effect on the health of the mutant. A classic example is *sickle-cell disease*, in which the molecules of the protein hemoglobin, responsible for carrying oxygen in the blood from the lungs to the tissues, tends to stick together and clump, resulting in sickle-shaped red blood cells. The disease is caused by a single substitution mutation in the gene that codes for one of the two components of hemoglobin, called β-globin. The substitution of a single nucleotide in the DNA sequence changes one amino acid in the protein from glutamate to valine, which causes the proteins to aggregate. This *missense*}* mutation (see figure ??) is carried by a fraction of the human population, and those who inherit the allele *allele* from both parents develop the painful and sometimes deadly disease. Such mutations that are present in some but not all of a population are called *polymorphisms*, to distinguish them from mutations that occurred in evolutionary lineages and differentiate species from each other.

One of the central questions of evolutionary biology is how frequently do mutations occur? Since mutations are generally undesirable, most living things have developed ways to minimize the frequency of errors in copying DNA, and to repair DNA damage. But although mutations are rare, they occur spontaneously in all organisms because molecular processes such as copying a DNA molecule are subject to random noise arising from thermal motion. So mutations are fundamentally a random process and we need to use *descriptive statistics* to analyze data with inherent randomness.

3.2 Describing data sets

3.2.1 central value of a data set

A data set is a collection of measurements. These measurements can come from many kinds of sources, and can represent all sorts of quantities. One big distinction is between numerical and categorical data sets. *Numerical* data sets contain numbers, either integers or real numbers. Some examples: number of individuals in a population, length, blood pressure, concentration. *Categorical* data sets may contain numbers, symbols, or words, limited to a discrete, usually small, number of values. The word categorical is used because this kind of data corresponds to categories or states of the subject of the experiment. Some examples: genomic classification of an individual on the basis of one locus (e.g. wild type or mutant), the state of an ion channel (open or closed), the stage of a cell in the cell cycle.

A data set contains more than one measurement, the number of them is called the size of the data set and is usually denoted by the letter n. To describe a data set numerically, one can use numbers called *statistics* (not to be confused with the branch of science of the same name). The most common statistics aim to describe the central value of the data set to represent a typical measurement. If you order all of the measurements from highest to lowest and then take the the middle value, you have found the *median* (if there is an even number of values, take the average between the middle two). Precisely half of the data values are less than the median and the other half are greater, so it represents the true "middle" value of the measurement. Note that the median can be calculated either for numerical or categorical data, as long as the categories can be ordered in some fashion.

The value that occurs most frequently in the data set is called its *mode*. For some data sets, particularly those which are symmetric, the mode coincides with the mean (see next paragraph) and the median, but for many others it is distinct. The mode is the most visual of the three statistics, as it can be picked out from the histogram plot of a data set (which is described in subsection 3.2.3) as the value corresponding to the maximum frequency. The mode can also be used for both categorical and numerical data.

The average or *mean* of a data set is the sum of all the values divided by the number of values. It is also called the *expected value* (particularly in the context of probability, which we will discuss later) because it allows to simply predict the sum of a large number of measurements with a given mean, by multiplying the mean by the number. The mean can be calculated only for a numerical data set, since we cannot add non-numerical values.

ℹ Definition

The *mean* of a data set X, also known as the average or the arithmetic mean is usually indicated with a bar over the variable symbol, and defined as the sum of the values divided by the number of values:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

The mean, unlike the median, is not the middle value of the data set, instead it represents the *center of mass* of the measured values [**?**]. Another way of thinking of the mean is as a **weighted sum of the values in the data set**. The weights represent the frequency of occurrence of each numeric value in the data set, which we will further discuss in subsection 3.2.3.

The mean is the most frequently used statistic, but it is not always interpreted correctly. Very commonly the mean is reported as the most representative value of a data set, but that is often misleading. Here are at least two situations in which the mean can be tricky: 1) data sets with a small number of discrete values; 2) data sets with outliers, or isolated numbers very far from the mean.

Examples of misleading means. Mean quantities for data sets with a few quantities are not the typical value, such as in the number of children born in a year per individual, also known as the birth rate. The birth rate per year in 2013 for both the United States and Russia is 1.3% per person, but you will have to look for a long time to find any individual who gave birth to 1.3% of a child. While this point may be obvious, it is often overlooked when interpreting mean values.

Outliers are another source of trouble for means. For example, a single individual (let's call him or her B.G.) with a wealth of $50 billion moves into a town of 1000 households with average wealth of $100,000. Although none of the original residents' assets have changed, the mean wealth of the town improves dramatically, as you can calculate in one of the exercises at the end of the chapter. One can site the improved per capita (per individual) in the town as evidence of economic growth, but that is obviously misleading. In cases with such dramatic outliers, the median is more informative as representation of a typical value of the data set.

3.2.2 Exercises

For the (small) data sets given below, calculate the mean and the median (by hand or using a calculator) and compare the two measures of the center.

1. Data set of the population of the city of Chicago (in millions) in the last 4 census years (2010, 2000, 1990, 1980): {2.7, 2.9, 2.8, 3.0}.
2. Data set of the numbers of the fish blacknose dace (*Rhinichthys atratulus*) collected in 6 different streams in the Rock Creek watershed in Maryland: {76, 102, 12, 55, 93, 98}.
3. Data set of tuberculosis incidence rates (per 100,000 people) in the 5 largest metropolitan areas in the US in 2012: {5.2, 6.6, 3.2, 5.5, 4.5}.
4. Data set of ages of mothers at birth for five individuals: {19, 20, 22, 32, 39}.
5. Data set of ages of fathers at birth for five individuals: {22, 23, 25, 36, 40}.
6. Data set of the number of new mutations found on maternal chromosomes for five individuals: {9, 10, 11, 26, 15}.
7. Data set of the number of new mutations found on paternal chromosomes for five individuals: {39, 43, 51, 53, 91}.
8. Consider the hypothetical town with 1000 households with mean and median wealth of $100,000 and one person with assets for $50 billion. Calculate the mean value of the combined data set, and compare it to the new median value.
9. Suppose you'd like to add a new observation to a data set; e.g. the 6-th largest metropolitan area (Philadelphia) to the tuberculosis incidence data set, which is 3.0. Calculate the mean of the 6-values data set, without using the 5 values in the original data set, but only using the mean of the 5-value data set and the new value. Generalize this to calculating the sample mean for any n-value data set, given the mean of the $n-1$ values, plus one new value.

3.2.3 spread of a data set

The center of a data set is obviously important, but so is the spread around the center. Sometimes the spread is caused by noise or error, for example in a data set of repeated measurements of the same variable under the same conditions. Other times the variance is due to real changes in the system, or due to inherent randomness of the system, and the size of the spread, as well as the shape of the histogram are important for understanding the mechanism. The simplest way to describe the spread of a numerical data set is to look at the difference between the maximum and minimum values, called the *range*. However, it is obviously influenced by outliers, since the extreme values are used. To describe the typical spread, we need to use all the values in the data set, and see how far each one is from the center, measured by the mean.

There is a problem with the naive approach: if we just add up all the differences of data values from the mean, the positives will cancel the negatives, and we'll get an artificially low spread. One way to correct this is to take the absolute value of the differences before adding them up. However, for somewhat deep mathematical reasons, the standard measure of spread uses not absolute values, but squares of the differences, and then divides that sum not by the number of data points n but by $n - 1$.

ℹ Definition

The *variance* of a data set X with n values is the sum of the squared differences of each value of the variable from the mean, divided by $n - 1$:

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X} - x_i)^2 \tag{3.2}$$

The variance is a sum of square differences, so its dimension is the square of the dimensions of the measurements in X. In order to obtain a measure of the spread comparable to the values of X, we take the square root of variance and call it the *standard deviation* of the data set X:

$$\sigma(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{X} - x_i)^2} \tag{3.3}$$

Just as the mean is a weighted average of all of the values in the data set, the variance is a weighted average of all the squared deviations of the data from the mean.

3.2.4 Exercises:

For the (small) data sets below, calculate the range, variance, and standard deviation (by hand or using a calculator). Compare the range and the standard deviation for each case: which one is larger? by how much?

1. Data set of the population of the city of Chicago (in millions) in the last 4 census years (2010, 2000, 1990, 1980): {2.7, 2.9, 2.8, 3.0}.
2. Data set of the numbers of the fish blacknose dace (*Rhinichthys atratulus*) collected in 6 different streams in the Rock Creek watershed in Maryland: {76, 102, 12, 55, 93, 98}.
3. Data set of tuberculosis incidence rates (per 100,000 people) in the 5 largest metropolitan areas in the US in 2012: {5.2, 6.6, 3.2, 5.5, 4.5}.
4. Data set of ages of mothers at birth for five individuals: {19, 20, 22, 32, 39}.
5. Data set of ages of fathers at birth for five individuals: {22, 23, 25, 36, 40}.
6. Data set of the number of new mutations found on maternal chromosomes for five individuals: {9, 10, 11, 26, 15}.

7. Data set of the number of new mutations found on paternal chromosomes for five individuals: {39, 43, 51, 53, 91}.
8. Consider the hypothetical town with 1000 households with mean and median wealth of \$100,000 and one person with assets for \$50 billion. Calculate the mean value of the combined data set, and compare it to the new median value.
9. (harder) Suppose that a data set has a fixed range (e.g. all values have to lie between 0 and 1). What is the greatest possible standard deviation for any data set within the range? Hint: think about how to place the points as far from the mean as possible. How do the data sets above relate to your prediction?}

3.2.5 describing data sets in graphs

Data sets can be presented visually to indicate the frequency of different values. This can be done in a number of ways, depending on the kind of data set. For a data set with only a few values, e.g. a categorical data set, a good way to represent it is with a pie chart. Each category is represented by a slice of the pie with the area of the same share of the pie as the fraction of the data set in the category. There is some evidence, however, that pie charts can be misleading to the eye, so R does not recommend using them.

For a numerical data set it is useful to plot the frequencies of a range of values, which is called a *histogram*. Its independent axis has the values of the data variable, and the dependent axis has the frequency of those values. If the data set consists of real numbers that range across an interval, that interval is divided into subintervals (usually of equal size), called bins, and the number of measurements in each bin is indicated on the y-axis. In order to be visually informative, there should be a reasonable number (usually no more than a few dozen, although it varies) of bins. The most frequent measurements are represented as the highest bars or points on the histogram. Histograms can denote either the counts of measurements in each bin, or to show the fraction of the total number of measurements in each bin. The only difference between those two kinds of histogram is the scale of the y-axis, and, confusingly, both can be called frequencies.

A histogram of the measured lengths of the bacterium *Bacillus subtilis* is shown in figure **??**. The data set was measured in increments in half a micron, with numbers varying between 1.5 and 4.5 microns. The histogram shows that the most common measurement (the mode) is 2 μm. Adding up all of the frequencies in the histogram tells us that there are approximately 200 total values in the data set. This allows us to find the median value by counting the frequencies of the first few bins until we get to 100 (the median point), which resides in the bin for 2.5 μm. It is a little bit more difficult to estimate the mean, but it should be clear that the center of mass of the histogram is also near 2.5 (it is actually 2.49). Finally, the hardest task is estimating the spread of the data set, such as the the standard deviation, based on the histogram. The range of the data set is $4.5 - 1.5 = 3$, so we know for sure that it is less than 1.5. The histogram shows that the deviations from the mean value of 2.5 range from 2

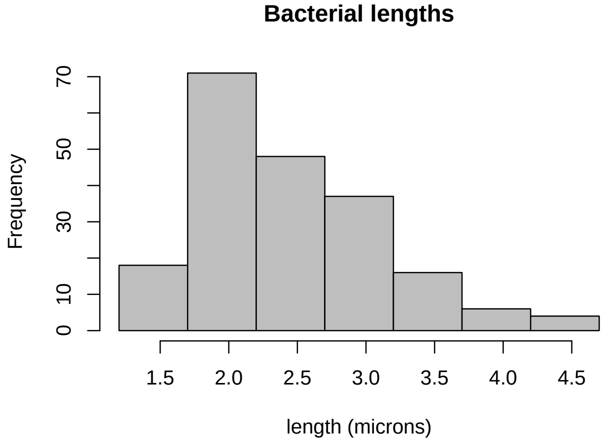


Figure 3.2: Length of bacteria Bacillus subtilis measured under the microscope as discrete values with step of 0.5; data from citep{watkins-intro-stats}

(rarely) to 0.5 (most prevalent). This should give you an idea that the weighted average of the deviations is less than 1. Indeed, the correct standard deviation is about 0.67.

There are different ways of plotting data sets that have more than one variable. For instance, a data set measured over time is called a time series. If the values are plotted with the corresponding times on the x-axis, then it is called a *time plot.* This is useful to show the changes of the values of your variable over time. If the data set doesn't undergo any significant changes over time, it makes more sense to represent it as a pie chart or histogram. More generally, one may plot two variables measured together on a single plot, which is called a *scatterplot.* We will explore such plots and the relationships between two measured variables in chapter 4.

3.2.6 Exercises

Answer the following questions, based on the histograms in figure **??** (mutation data) and in figure **??** (heart rate data).

1. How many people in the mutation data have fathers either younger than 20 or older than 40? How many have more than 80 new mutations?
2. Estimate the median and mean of the two variables in the mutation data set.
3. State the range of each data set, and estimate the standard deviation of the two variables in the mutation data set.
4. How many people in the heart rate data have heart rates greater than 80 bpm? How many have body temperature less that 97 F?

5. Estimate the median and mean of the two variables in the heart rate data set.
6. State the range of each data set, and estimate the standard deviation of the two variables in the heart rate data set.

```
my_data <- read.table("data/HR_temp.txt", header = TRUE)
hist(my_data$HR, col = "gray", main = "Heart rate data",
    xlab = "heart rate (bpm)")
hist(my_data$Temp, col = "gray", main = "Body temperature data",
    xlab = "temperature (F)")
```

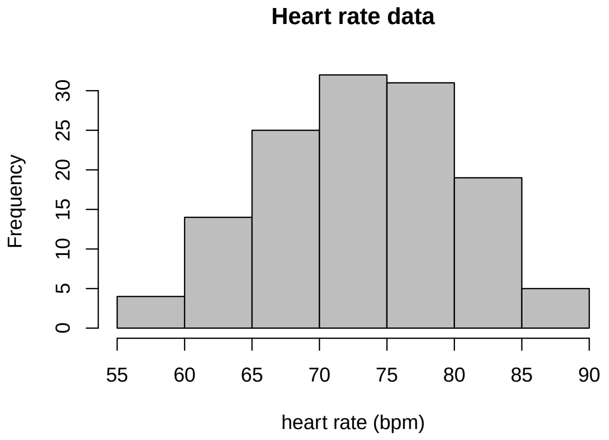


Figure 3.3: Histograms of heart rates and body temperatures

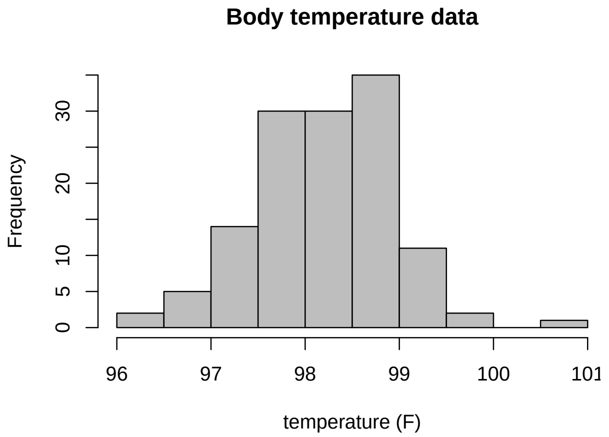


Figure 3.4: Histograms of heart rates and body temperatures

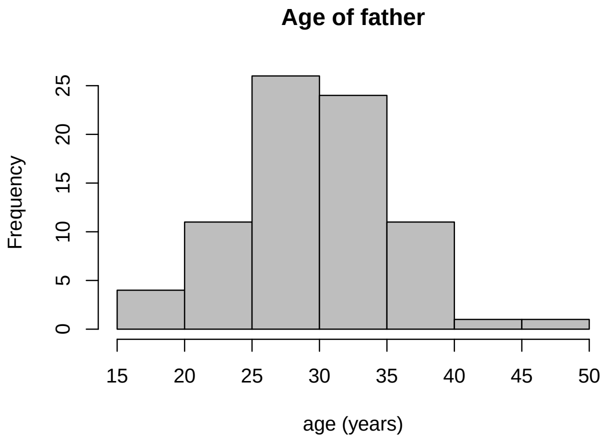


Figure 3.5: Histograms of paternal ages and the number of new mutations from 73 families; data from citep{kong_rate_2012}

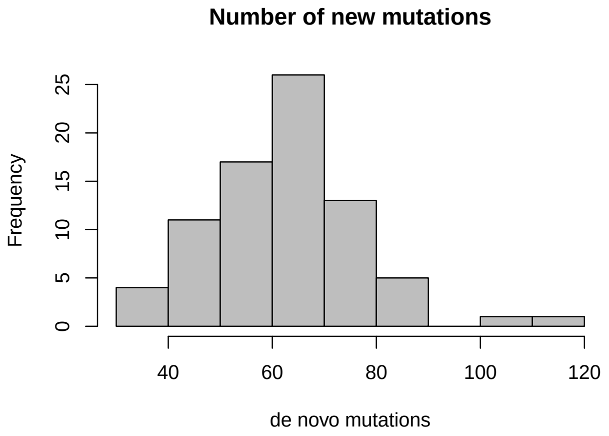


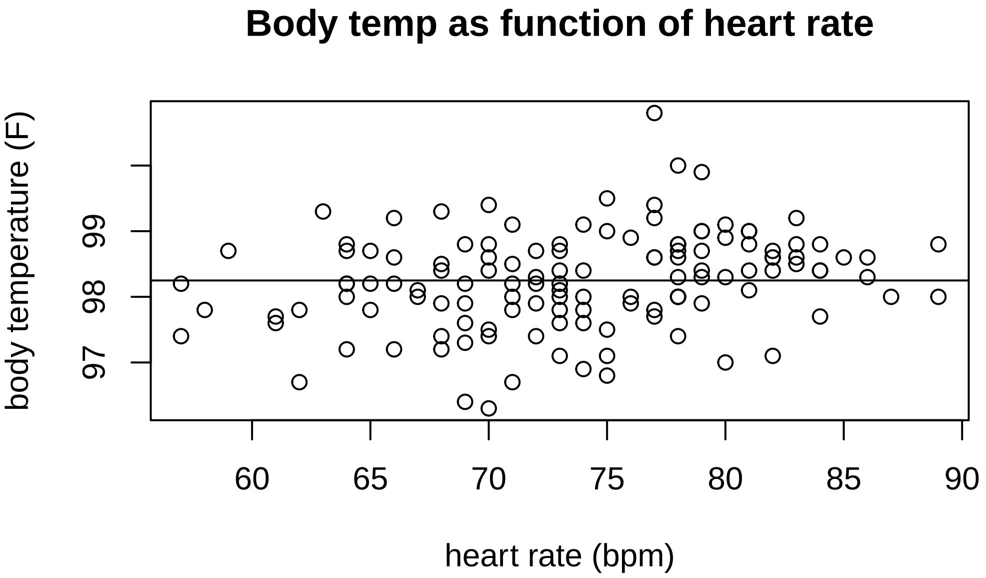
Figure 3.6: Histograms of paternal ages and the number of new mutations from 73 families; data from citep{kong_rate_2012}

3.3 Working with data in R

3.3.1 reading in data into data frames

One way to input data into R is to read in a text file, where several variables are stored in columns. For instance, the file `HR_temp.txt` contains three variables: body temperature (in Fahrenheit), sex (1 for male, 2 for female), and heart rate (in beats per minute). The values for the variables are arranged in columns, while first row of the file contains the names of the variables (Temp, Sex, and HR, respectively). Note that the data file has to be saved into the same folder as the .Rmd file week1.Rmd for this to work.

```
vitals <- read.table(file = "data/HR_temp.txt", header = TRUE)
plot(vitals$HR, vitals$Temp, main = "Body temp as function of heart rate",
    xlab = "heart rate (bpm)", ylab = "body temperature (F)")
mTemp <- mean(vitals$Temp)
sdTemp <- sd(vitals$Temp)
abline(mTemp, 0)
```



```
mean(vitals$HR)
```

```
[1] 73.76154
```

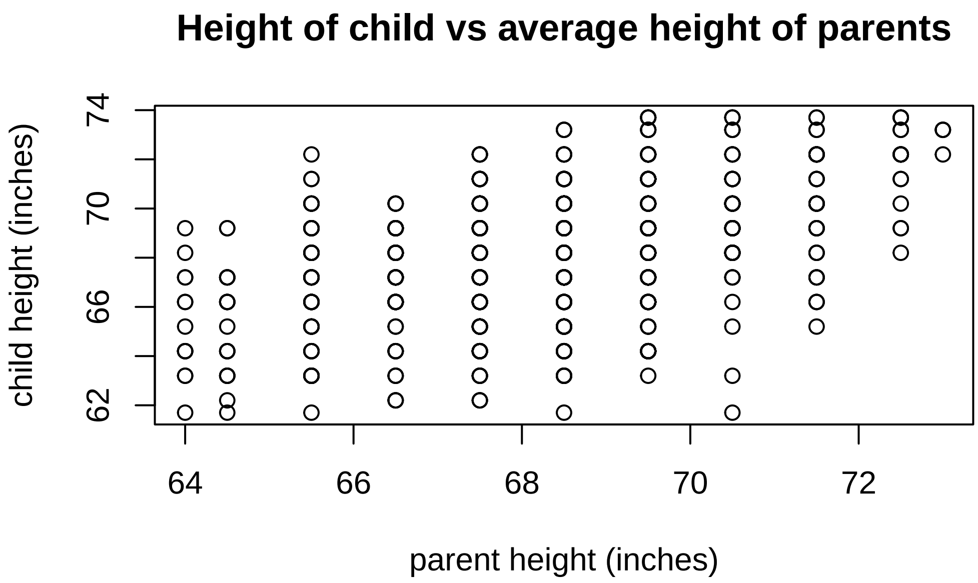
```
sd(vitals$HR)
```

```
[1] 7.062077
```

The R command read.table() reads this file and and puts it into a *data frame* called data. The three variables are stored inside the data frame, and can be accessed by appending the dollar sign and variable name to the data frame, so `data$HR` refers to only the heart rates, and `data$Temp` refers to the body temperatures. The plot shows the relationship of the two data variables, and the function `abline(98.6,0)` plots a line with the intercept 98.6a and slope 0 on top of the scatterplot.

You can also load data from a package, e.g. `HistData`, which contains many classic data sets. Got to the Packages tab in the lower right window in R Studio, click Install and type `HistData`. We will use the data set called `Galton` that contains the heights of parents (the mean of mother's and father's) and their children, in variables parent and child. The script below plots the two variables, with parent as the independent (explanatory) variable and child as the dependent (response) variable.

```
library(HistData)
plot(Galton$parent, Galton$child, main = "Height of child vs average height of parents",
    xlab = "parent height (inches)", ylab = "child height (inches)")
```



```
summary(Galton$parent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  64.00   67.50   68.50   68.31   69.50   73.00 
```

```r
summary(Galton$child)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  61.70   66.20   68.20   68.09   70.20   73.70 
```

3.3.2 descriptive statistics

One can also calculate basic descriptive statistics as follows:

```r
paste("The mean parental height is:", mean(Galton$parent))
```

```
[1] "The mean parental height is: 68.3081896551724"
```

```r
paste("The mean child height is:", mean(Galton$child))
```

```
[1] "The mean child height is: 68.0884698275862"
```

```r
paste("The standard deviation of parental height is:", sd(Galton$parent))
```

```
[1] "The standard deviation of parental height is: 1.78733340172202"
```

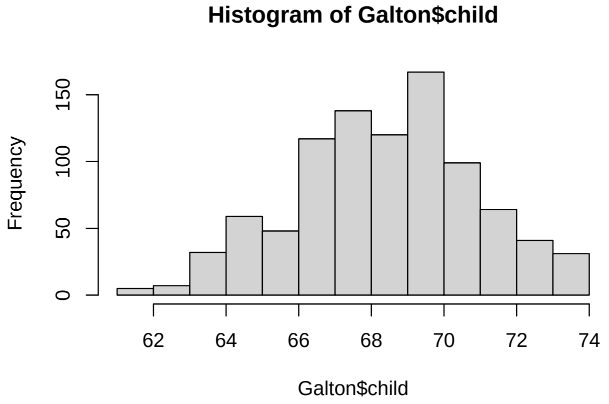
```r
paste("The standard deviation of  child height is:", sd(Galton$child))
```

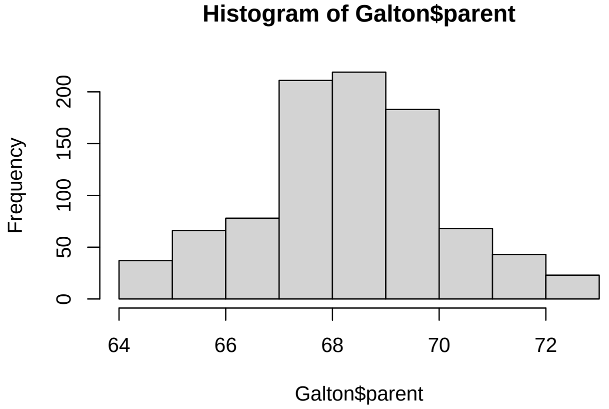
```
[1] "The standard deviation of  child height is: 2.51794136627677"
```

Why do you think the standard deviation of parental height is much smaller?

R has histogram function `hist()`, which does a passable job of representing the distribution of a variable such as child height or parent height. Compare the width of the two distributions and consider why they are different.

```r
hist(Galton$child)
hist(Galton$parent)
```





3.3.3 Exercises:

The following code chunks contain errors. Find and fix them so they work as intended.

1. Calculate the mean and standard deviation of the heart rates of the first 30 individuals in the data frame vitals:

```
mean(vitals$HR[30])
```

```
[1] 64
```

```
sd(vitals$HR[30])
```

```
[1] NA
```

2. Calculate the mean and standard deviation of the ratio of heart rates to body temperatures for the data set vitals:

```
mean(vitals$HR/Temp)
sd(vitals$HR/Temp)
```

3. Plot a scatterplot of the child heights as the response variable and the parent heights and the explanatory variable, and overlay the line y=x on top.

```
plot(parent, child, main = 'Height of child vs average height of parents', xlab= 'child he
abline(1,0)
```

4. Calculate the median of both parent and child heights:

```
median(Galton)
```

5. Plot the histogram of parent heights of the first half of the group:

```
hist(Galton$parent/2)
```

6. Plot the histogram for the ratio of parent and child heights for the entire data set and calculate its mean and variance:

```
hist(Galton$parent/child)
mean(Galton$parent/child)
sd(Galton$parent/child)
```

4 Random variables and distributions

> What is there then that can be taken as true? Perhaps only this one thing, that nothing at all is certain. –Rene Descartes

Mathematical models can be divided into *deterministic* and *stochastic* models. Deterministic models assume that the future can be perfectly predicted based on complete information of the past. Stochastic models instead assume that even perfect knowledge of the past does not allow one to predict the future with certainty.

Stochastic models may not sound very promising: after all, we want to make predictions, and randomness says that predictions are impossible! However, the word "random" in mathematics doesn't mean "completely unpredictable" or "without rules," as it does in common usage. It means that we can make probabilistic predictions, e.g. compute what fraction of molecules will diffuse from one place to another, or what fraction of genes mutate in one generation - we just can't make a definite prediction for each individual molecule or gene. Biological processes are so complex and are subject to so much environmental noise, that stochastic models are absolutely essential for our understanding of many living systems. Here is what you will learn to do in this chapter:

- define probability in terms of outcomes and events
- know what is a random variable and its distribution
- compute means and variances of distributions
- use the binomial distribution to model strings of binary trials
- generate random numbers in R

4.1 Random variables and distributions

4.1.1 definition of probability

In this section we will develop the terminology used in the mathematical study of randomness called probability. This begins with a *random experiment* which is a very broad term that can describe any natural or theoretical process whose outcome cannot be predicted with certainty. If the outcomes are numeric, they may be *discrete* (can be counted by integers) or *continuous* (corresponding to real numbers); they may also be *categorical*, meaning that they do not have a numeric meaning, like eye color. We will stick to experiments that have discrete outcomes in this chapter, but many important experiments produce continuous outcomes. The first step for studying a random process is to describe all of the outcomes it can produce:

> **i** Definition
>
> The collection of all possible outcomes of an experiment is called its *sample space* Ω. An *event* is a subset of the sample space, which means an event may contain one or more experimental outcomes.

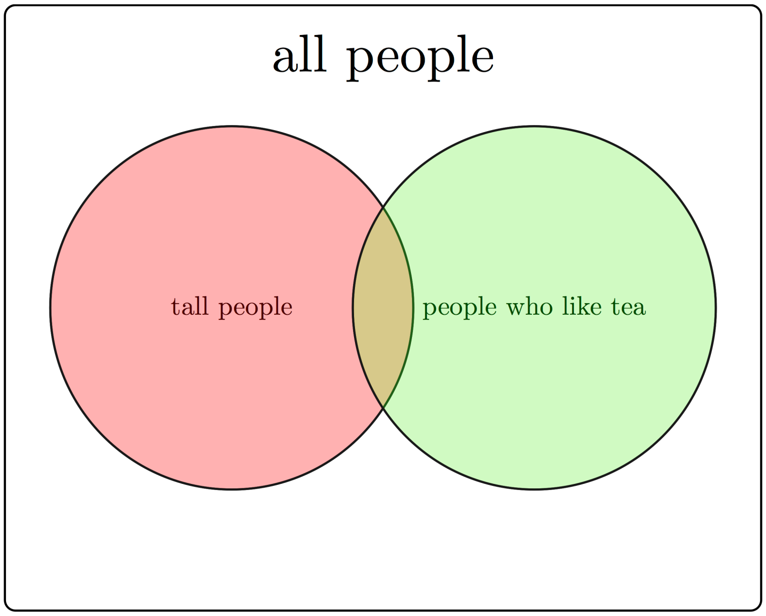


Figure 4.1: An illustration of the sample space of all people with two events: tall people and those who like tea.

Example. You can ask a person two questions: how tall are you (and classify them either as short or tall) and do you like tea (yes or no), and you've performed a random experiment. The randomness comes not from the answers (assuming the person doesn't randomly lie) but from the selection of the respondent. We will discuss randomly selecting a sample from a population in the next chapter. This random experiment has four outcomes: tall person who likes tea, tall person who does not like tea, short person who likes tea, and short person who does not like tea. This sample space and events is illustrated in figure **??** with a Venn diagram, which uses geometric shapes as representations of events as subsets of the entire sample space. These outcomes can be grouped into events by one of the responses: e.g. tall person (A) or person who doesn't like tea ($-B$).

Example. A random experiment with two outcomes, called a *Bernoulli trial* (after the famous

Swiss mathematician), can describe a variety of situations: a coin toss (heads or tails), a competition with two outcomes (win or loss), the allele of a gene (normal or mutant). The sample space for a single Bernoulli trial consists of just two outcomes: $\{H, T\}$ (for a coin toss). If the experiment is performed repeatedly, the sample space gets more complicated. For two Bernoulli trials there are four different outcomes $\{HH, HT, TH, TT\}$. One can define different events for this sample space: the event of getting two heads in two tosses contains one outcome: $\{HH\}$, the event of getting a single head contains two: $\{TH, HT\}$.

In order to to describe the composition of a sample space, we need to define the word *probability* [**?**]. While it is familiar to everyone from everyday usage, it is difficult to define without using other similar words, such as likelihood or plausibility, which are also in need of definition. It is accepted that something with a high probability happens often, while something with a low frequency is seldom observed. The other notion is that probability can range between 0 (meaning something that never occurs) and 1 (something that occurs every time). These notions lead to the commonly accepted definition:

> ℹ Definition
>
> The *probability* of an outcome or event in the sample space of a random experiment is the fraction of experiments with this outcome out of many repeated experiments.

This definition is at the heart of the *frequentist* view of probability, due to the underlying assumption that the experiment can be repeated as many times as necessary to observe the frequency of outcomes. There is an alternative view that focuses on what is previously known about the experiment (or about systems that produce that kind of experiment) that is called the *Bayesian* view:

> ℹ Definition
>
> The *probability* of an outcome or event in the sample space of a random experiment is the degree of *certainty* or *belief* that this outcome will occur based on prior experience.

We will investigate the Bayesian approach in chapter 12. Most of traditional probability and classical statistics is based on the frequentist view, as it grew out of attempts to understand games of chance, like cards and dice, which can be easily repeated, or simple experiments like those in agriculture, where many plots can be planted and observed. These easily repeatable simple experiments can be described with mathematical distributions that we will describe in this chapter. However, many contemporary research problems are not so easily repeated, and often require a Bayesian approach that does not yield to neat mathematical description and can be addressed using computation.

4.1.2 axioms of probability

One we have defined the probability of an outcome, one can calculate the probability of a collection of outcomes according to rules that ensure the results are self-consistent. These rules are called the axioms of probability:

> Definition
>
> The probability $P(A)$ of an event A in a sample space Ω is a number between 0 and 1, which obeys the following rules, called the *axioms of probability*:
>
> - $P(\Omega) = 1$
> - $P(\emptyset) = 0$
> - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Let us define some notation for sets: $A \cup B$ is called the *union* of two sets, which contains all outcomes that belong to either A or B, this is equivalent to the logical OR operator because it is true if either A or B is true. $A \cap B$ is called the *intersection* of two sets, which contains all outcomes that are in both A and B, this is is equivalent to the logical AND operator because it is true if both A and B are true. The $\emptyset$ denotes the empty set. Any event A has its *complement*, denoted $-A$, which contains all outcomes of Ω which are not in A.

Applying them to the sample space and events in figure **??**, the union of the two sets $A \cup B$ are all people who are either tall or like tea, the intersection of the two sets $A \cap B$ are all the tall people who like tea, and the intersection of the first set with the complement of the second $A \cup -B$ are all tall people who do not like tea.

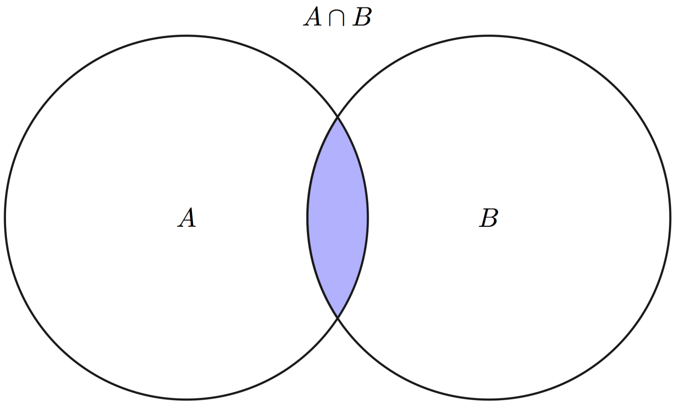


Figure 4.2: An illustration of the operation of intersection of sets A and B.

The first two axioms connect easily with our intuition about probability: the first axiom says that the probability of some outcome from the sample space occurring is 1, while the second says that the probability of nothing in the sample space occurring is 0. The intuition behind

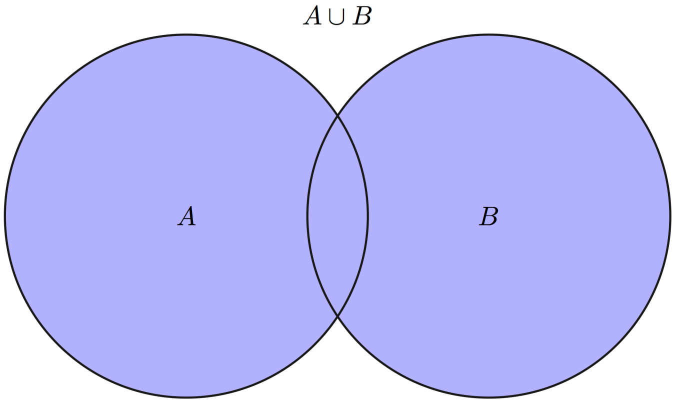


Figure 4.3: An illustration of the operation of the union of sets A and B.

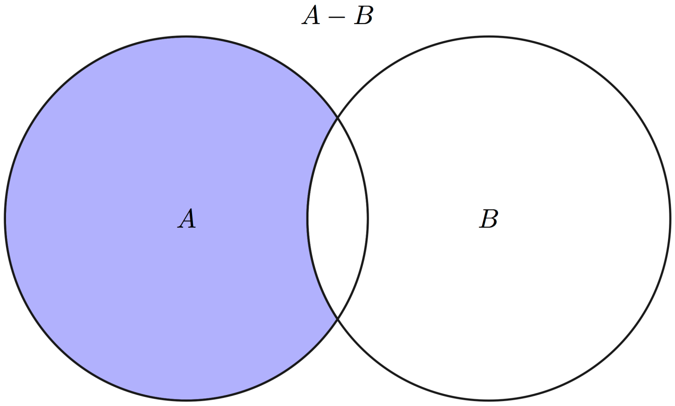


Figure 4.4: An illustration of the intersection of A with -B

axiom three is less transparent, but it can be see in a Venn diagram of two subsets A and B of the larger set Ω, as in figure **??**. Compare the size of the union of A and B and the sum of the sizes of sets A and B separately, and you will see that the intersection $A \cap B$ occurs in both A and B, but is only counted once in the union. This is why it needs to be subtracted from the sum of $P(A)$ and $P(B)$.

There are several useful rules that immediately follow from the axioms. First, if two events are mutually exclusive, meaning their intersection is empty ($A \cap B = \emptyset$), then the probability of either of them happening is the sum of their respective probabilities: $P(A \cup B) = P(A) + P(B)$ (from axiom 3). Further, since an event A and its complement $-A$ are mutually exclusive, their union is the entire sample space Ω: $P(A) + P(-A) = P(A \cup -A) = P(\Omega) = 1$, therefore $P(A) = 1 - P(-A)$.

Example. Assume one is using a fair coin, so the probability of a single head and a single tail is 1/2. The probability of getting two heads in a row is 1/4, because exactly half of those coins that come up heads once will come up heads again. In fact, the probability of getting any particular sequence of two coin toss results is 1/4. Here are some examples of what we can calculate:

- the probability of getting one head of out of two tosses is $1 - 1/4 - 1/4 = 1/2$ (by the complement rule).
- the probability of *not* getting two heads is $1 - 1/4 = 3/4$ (by the complement rule).

- the probability of getting either 0, 1, or 2 heads is 1 (by axiom 1).
- the probability of getting three heads is 0 (since this event is not in the sample space).

Example. Suppose one is testing people for a mutation which has the probability (prevalence) of 0.2 in the population, so for each person there are two possible outcomes: normal or mutant. The probability of drawing two mutants in a row is $0.2 * 0.2 = 0.04$ by the same argument as above; the probability of drawing two normal people is $0.8 * 0.8 = 0.64$. Based on this, we can calculate the following

- the probability of one mutant of out two people is $1 - 0.04 - 0.64 = 0.32$ (by the complement rule).
- the probability of not having two mutants is $1 - 0.04 = 0.96$ (by the complement rule).

- the probability of either 0, 1, or 2 mutants is 1 (by axiom 1).
- the probability of getting three mutants is 0 (since this event is not in the sample space).

Example (from Danny and Gaines Sarcastic fringeheads are a tropical ocea fish that engage in aggressive mouth-wrestling matches for their rocky residences. Let us treat each match as a stochastic experiment with two outcomes: win or loss. Then the sample space is equivalent to our coin-tossing experiment, e.g. for two matches the sample space is $\{WW, WL, LW, LL\}$. However, the probability distribution may different, for example if a particular fringehead wins 3/4 of its matches, then the probability distribution would be: $P(\{WW\}) = 9/16$, $P(\{LW\}) =$

$P(\{WL\}) = 3/16$, and \$ P({ LL }) = 1/16\$. Thus, the same sample space may have different probability distributions defined on it.

4.1.3 random variables

The outcomes of experiments may be expressed in numbers or words, but we generally need numbers in order to report and analyze results. One can describe this mathematically as a function (recall its definition form section **??**) that assigns numbers to random outcomes [**?**]. In practice, a random variable describes the measurement that one makes to describe the outcomes of a random experiment.

ℹ Definition

A *random variable* is a number or category associated to each outcome in a sample space Ω. This association has to follow the rules of a function as defined in chapter 2.

Example. Define the random variable to be the number of heads out of two coin tosses. This random variable will return numbers 0, 1, or 2, corresponding to different events. The random variable of the number of mutants out of two people (assuming there are only two outcomes, mutant and normal) has the same set of values. This random variable is a function on the sample space because it returns a unique value for each outcome.

Example. (Danny and Gaines) Suppose that our sarcastic fringehead, upon losing a wrestling match, has to search for another home for three hours. Then we can define the random variable of time wasted over two wrestling matches, which can be either 0, 3, or 6 hours, depending on the events defined above. Once again, this is a function because there is an unambiguous number associated with each outcome.

A random variable has a set of possible values, and each of those values may come up more or less frequently in an random experiment. The frequency of each measurement corresponds to the probability of the outcomes in the sample space that produce that particular value of the random variable. One can describe the behavior of the random variable in terms of the collection of the probabilities of its outcomes.

ℹ Definition

The probability of a random variable X taking some value a, written as $P(X = a)$, but usually simplified to $P(a)$ is the probability of the event corresponding to the value a of the random variable. This function $P(a)$ is called the *probability distribution* of the random variable X.

One important property of probability distribution functions for a discrete random variable is that all of its values have to add up to 1:

$$\sum_{i=1}^{N} P(a_i) = 1$$

The graph of a probability distribution function lies above zero because all probabilities are between 0 and 1. The graph of a probability distribution is very similar to a histogram, in that it represents the frequency of occurrence of each value of the random variable. A histogram of a variable from a data set can be thought is an approximation of the true probability distribution based on the sample. For a large sample size, the histogram approaches the graph of the probability distribution function, something which we will discuss in chapter 9.

Example. Assuming that each coin toss has probability 1/2 of resulting in heads, the probability distribution function for the number of heads out of two coin tosses is $P(0) = 1/4$; $P(1) = 1/2$; $P(2) = 1/4$ (as we computed in the example in the previous section). Note that the probabilities add up to 1, as they should.

Example. For the random variable of the number of mutants out of two people, for mutation prevalence of 0.2, the probability distribution function is $P(0) = 0.64$; $P(1) = 0.32$; $P(2) = 0.04$ (as we computed in the example in the previous section). Note that the probabilities add up to 1, as they should.

Example. For the time wasted by a fringehead, the distribution is $P(0) = 9/16$; $P(3) = 3/16$; $P(6) = 1/16$. Note that other values of the random variable have probability 0, because they correspond to the empty set in sample space.

4.1.4 expectation of random variables

> ℹ Definition
>
> The *expected value* (or mean) of a discrete random variable X with probability distribution $P(X)$ is defined as:
>
> $$E(X) = \mu_X = \sum_{i=1}^{N} a_i P(a_i)$$

This sum is over all values $\{a_i\}$ that the random variable X can take, multiplied by the probability of the random variable taking that value (meaning the probability of the event in sample space that corresponds to that value). This corresponds to the definition of the mean of a data set given in section **??**, if you consider $P(a_i)$ to be the number of times a_i occurs divided by the number of total measurements N. As in the case of the histogram and the distribution function, the mean of a sample for a large sample size N approaches the mean of the random variable, which we will discuss in more detail in the next chapter. Sometimes we will use the more concise $\mu_X = E(X)$ to represent the mean (expected) value. Here are some mathematical properties of the expectation:

- Expectation of a random variable which is always constant (c) is equal to c, since the probability of c is 1: $E(c) = cP(c) = c$
- Expectation of a constant multiple of a random variable is:

$$E(cX) = \sum_i cx_i P(x_i) = c \sum_i x_i P(x_i) = c\mu_X$$

- Expectation of a sum of two random variables is the sum of their expectations. This is a more complicated argument, so let us break it down. First, all possible values of the random variable $X + Y$ come from going through the possible values of X (a_i) and Y (b_i), and each combination of values has its own probability (called the joint probability distribution) $P(a_i, b_j)$:

$$E(X+Y) = \sum_i \sum_j (a_i + b_j) P(a_i, b_j)$$

We can split the sum into two terms by the distributive property of multiplication and then take out the values a_i and b_j out of the sum that they do not depend on:

$$E(X+Y) = \sum_i \sum_j a_i P(a_i, b_j) + \sum_i \sum_j b_j P(a_i, b_j) =$$
$$= \sum_i a_i \sum_j P(a_i, b_j) + \sum_j b_j \sum_i P(a_i, b_j)$$

The joint distributions added up over all values of one variable, become single-variable distributions, so this leaves us with two sums which are the two separate expected values:

$$E(X+Y) = \sum_i a_i P(a_i) + \sum_j b_j P(b_j) = E(X) + E(Y)$$

Example. The expected value of the number of heads out of two coin tosses can be calculated using the probability distribution function we found above:

$$E(X) = 0 \times P(0) + 1 \times P(1) + 2 \times P(2) = 0 + 1/2 + 2 \times 1/4 = 1$$

The expected number of heads out of 2 is 1, if each head comes up with probability 1/2, which I think you will find intuitive.

Example. The expected value of the number of mutants out of two people can be calculated using the probability distribution function we found above:

$$E(X) = 0 \times P(0) + 1 \times P(1) + 2 \times P(2) = 0 + 1 \times 0.32 + 2 \times 0.04 = 0.4$$

The expected number of mutants in a sample of two people is 0.4, which may seem a bit strange. Recall that mean or expected values do not have to coincide with values that are possible, as we discussed in section **??**, but are instead a weighted average of values, according to their frequencies or probabilities.

Example. Find the expected value of the number of wins out of two matches for a fringehead which has the probability of winning of 3/4.

$$E(X) = 0 \times 1/16 + 1 \times 6/16 + 2 \times 9/16 = 24/16 = 3/2$$

4.1.5 variance of random variables

Knowledge of the expected value says nothing about how the random variable actually varies: expectation does not distinguish between a random variable which is constant and one which can deviate far from the mean. In order to quantify this variation, one might be tempted to compute the mean differences from the mean value, but it does not work:

$$E(X - \mu_X) = \sum_i (x_i - \mu_x) P(x_i) = \sum_i x_i P(x_i) - \mu_x \sum_i P(x_i) = \mu_x - \mu_x = 0$$

The problem is, if we add up all the differences from the mean, the positive ones end up canceling the negative ones and the expected value of those deviations is exactly zero. This is why it makes sense to square the differences and add them up:

Definition

The *variance* of a discrete random variable X with probability distribution $P(x)$ is

$$Var(X) = E((X - \mu_X)^2) = \sum_{i=1}^{N} (x_i - \mu_x)^2 P(x_i)$$

One useful property of the variance is:

$$Var(X) = \sum_i (x_i^2 - 2x_i\mu_x + \mu_x^2) P(x_i) =$$

$$= \sum_i x_i^2 P(x_i) - 2\mu_x \sum_i x_i P(x_i) + \mu_x^2 \sum_i P(x_i) = E(X^2) - E(X)^2$$

So variance can be calculated as the difference between the expectation of the variable squared and the squared expectation. Note that the variance is given in units of the variable squared, so in order to measure the spread of the variable in the same units, we take the square root of the variance and call it the *standard deviation*:

$$\sigma_x = \sqrt{Var(X)}$$

While the expectation of a sum of random variables is the sum of their expectations, for any random variables, the same is not true for the variance. However, there is a special condition under which this is true. First, let us write the variance of a sum of two random variables X and Y:

$$Var(X+Y) = E\left[(X+Y) - (\mu_X + \mu_Y)\right]^2 =$$

$$= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 - 2(X-\mu_X)(Y-\mu_Y)] =$$
$$= E(X-\mu_X)^2 + E(Y-\mu_Y)^2 - 2E[(X-\mu_X)(Y-\mu_Y)] =$$

$$= Var(X) + Var(Y) - 2E[(X-\mu_X)(Y-\mu_Y)]$$

If you write out the last term as a sum, it is none other than the *covariance* of the two random variables X and Y, which we saw in the chapter on linear regression. So for any two random variables that have zero covariance, their variance is additive!

Example. The variance of the number of heads out of two coin tosses can be calculated using its probability distribution function and the expected value (1) from above:

$$Var(X) = (0-1)^2 \times P(0) + (1-1)^2 \times P(1) + (2-1)^2 \times P(2) = 1/4 + 0 + 1/4 = 1/2$$

Since the variance is 1/2, the standard deviation, or the expected distance from the mean value is $\sigma = \sqrt{1/2}$.

Example. The variance of the number of mutants out of two people can be calculated using its probability distribution function and the expected value (0.4) from above:

$$E(X) = (0-0.4)^2 \times P(0) + (1-0.4)^2 \times P(1) + (2-0.4)^2 \times P(2) =$$

$$= 0.4^2 \times 0.64 + 0.6^2 \times 0.32 + 1.6^2 \times 0.04 = 0.32$$

Since the variance is 0.32, the standard deviation, or the expected distance from the mean value is $\sigma = \sqrt{0.32}$.

Example. We have computed the expected value for the number of wins in two fringehead fights, so now let us find the variance and standard deviation. We already know the possible values of X, and the associated probabilities, so we calculate:

$$E(X^2) = 0^2 \times 1/16 + 1^2 \times 6/16 + 2^2 \times 9/16 = 42/16$$

Then the variance is:

$$Var(X) = E(X^2) - E(X)^2 = 42/12 - 9/4 = (42-27)/16 = 15/16$$

and the standard deviation is $\sigma = \sqrt{15}/4$ or just under 1.

4.1.6 Exercises

Calculate the expected values and variances of the following probability distributions, where the possible values of the random variable are in curly brackets, and the probability of each value is indicated as $P(x)$.

1. $X = \{0, 1\}$ and $P(0) = 0.1, P(1) = 0.9$.
2. $X = \{1, 2, 3\}$ and $P(1) = P(2) = P(3) = 1/3$.
3. $X = \{10, 15, 100\}$ and $P(10) = 0.5, P(15) = 0.3, P(100) = 0.2$.
4. $X = \{0, 1, 2, 3, 4\}$ and $P(0) = 1/8, P(1) = P(2) = P(3) = 1/4, P(4) = 1/8$.
5. $X = \{-1.5, -0.4, 0.3, 0.9\}$ and $P(-1.5) = 0.4, P(-0.4) = 0.2, P(0.3) = 0.35, P(0.9) = 0.05$.

4.2 Examples of distributions

4.2.1 uniform distribution

Perhaps the simplest random variable (besides a constant, which is not really random) is the *uniform random variable*, for which every outcome has equal probability. The distribution of a fair coin is uniform with two values, H or T, or 0 and 1, each with probability 1/2. More generally, a discrete uniform random variable has N outcomes and each one has probability $1/N$. This is what people often mean when they use the word random - an experiment where each outcome is equally likely.

We can calculate the expectation and variance of a uniform random variable U:

$$E(U) = \sum_{i=1}^{n} a_i P(a_i) = \frac{1}{n} \sum_{i=1}^{n} a_i$$

So the expected value is the mean of all the values of the uniform random variable.

Example. In the special case of the uniform distribution of $n + 1$ integers between 0 and n ($a_i = i$, for $i = 0, ..., n$), each value has probability $P = 1/(n + 1)$. The expected value is the average of the maximum and minimum values (using the fact that $\sum_{i=0}^{n} i = n(n + 1)/2$):

$$E(U) = \frac{n(n + 1)}{2(n + 1)} = \frac{n}{2}$$

Generalizing, for a random variable on integers between a and b, the expectation is

$$E(U) = \frac{a + b}{2}$$

We can also write down the expression for the variance of the discrete uniform distribution as follows:

$$Var(U) = E(U^2) - E(U)^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} a_i\right)^2$$

Example. In the special case of the uniform distribution of $n+1$ integers between 0 and n ($a_i = i$, for $i = 0, ..., n$), each value has probability $P = 1/(n+1)$. The variance can be calculated using the formula for the sum of squares: $\sum_{i=0}^{n} i^2 = n(n+1)(2n+1)/6$.

$$Var(U) = \frac{(n+1)(2n+1)n}{6(n+1)} - \frac{n^2}{4} = \frac{2n^2+n}{6} - \frac{n^2}{4} = \frac{n(n+2)}{12}$$

This can be generalize to a uniform random variable on integers between a and b (omitting the algebraic details) so the variance for that uniform random variable is:

$$Var(U) = \frac{(b-a+1)^2-1}{12} = \frac{(b-a)^2+2(b-a)}{12}$$

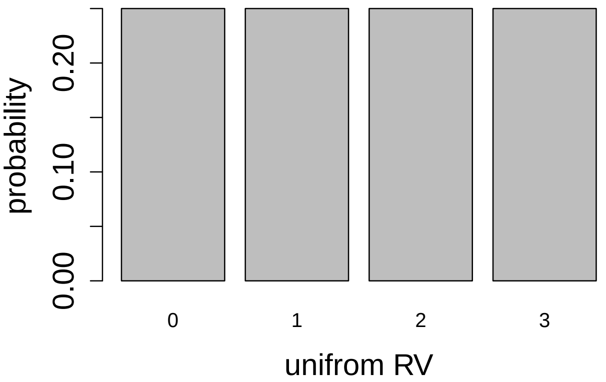


Figure 4.5: Two uniform random distributions with integer values with different ranges.

4.2.2 binomial distribution

We have introduced binary or Bernoulli trials in section **??**. Assume that the two values of the random variable X are 0 and 1, with probability $1-p$ and p, respectively. Then we can calculate the expectation and variance of a single Bernoulli trial:

$$E(X) = 0 \times (1-p) + 1 \times p = p$$
$$Var(X) = E(X^2) - E(X)^2 = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p)$$

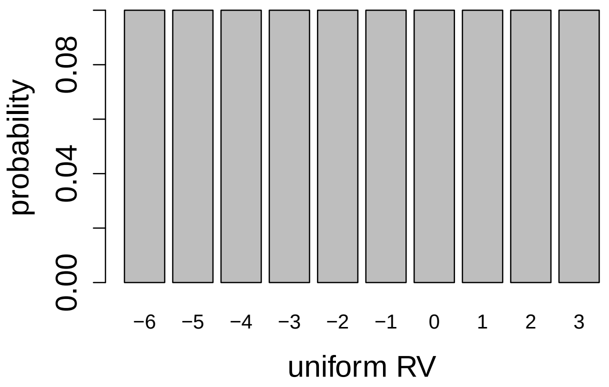


Figure 4.6: Two uniform random distributions with integer values with different ranges.

The first result is likely intuitive, but the second deserves a comment. Note that depending on the probability of 1, the variance, or the spread in outcomes of a Bernoulli trial is different. The highest variance occurs when $p = 1/2$, or equal probability of 0 or 1, but when p approaches 0 or 1, the variance approaches 0. Thus, as the probability approaches zero or one the random variable approaches a constant (either always 1 or 0); hence, no variance.

One can extend this scenario and ask what happens in a string of Bernoulli trials, for instance, in a string of 10 coin tosses, or in testing 20 randomly selected people for a mutation. The mathematical problem is to calculate the probability distribution of the number of success out of many trials. This is known as the binomial random variable, which is defined as the sum of n independent, identical Bernoulli random variables.

Definition

Given n independent Bernoulli trials X with the same probability of success p, the *binomial random variable* is defined as:

$$B = \sum_{i=1}^{n} X_i$$

where X_i is the random variable from the i-th Bernoulli trial, which takes values of 1 and 0.

In this definition I use the term independence without defining it properly, which will be done in chapter 10. Intuitively, independence between two Bernoulli trials (e.g. coin tosses) means that the outcome of one trial does not change the probability of the outcomes of any other trials. This amounts to the assumption that the probability of an outcome followed by another one is the product of the separate probabilities of the two outcomes. For example, if the two outcomes are wins and losses, then $P(\{WL\}) = P(W)P(L)$. This will be used below in the calculation of the variance of the binomial random variable.

To find the probability distribution of the binomial random variable, we need to define the event of k wins out of n trials. Consider the case of 4 trials. It is easy to find the event of 4 wins, as it is comprised only of the outcome $\{WWWW\}$. Then, $P(4) = p^4$, based on the independence assumption. The event of winning 3 times consists of four strings: $\{LWWW, WLWW, WWLW, WWWL\}$ so the probability of obtaining 3 wins is the sum of the four probabilities, each equal to \$ p^3(1-p)\$ from the independence assumption above, so $P(3) = 4p^3(1-p)$. The event of winning 2 times is even more cumbersome, and consists of six strings: $\{LLWW, WLLW, WWLL, WLWL, LWLW, LWWL\}$, so $P(2) = 6p^2(1-p)^2$ by the same reasoning.

Now imagine doing this to calculate 50 wins out of 100 trials. The counting gets ugly very fast. We need a general formula to help us count the number of ways of winning k times out of n trials. We denote this number $\binom{n}{k}$, also known as "n choose k" because it corresponds to the number of ways of choosing k distinct objects out of n without regard to order. The connection is as follows: let us label each trial from 1 to n. Then to construct a string with k wins, we need to specify which trials resulted in a win (the rest are of course losses). It does not matter in which order those wins are selected - it still results in the same string. Therefore the number of different strings of n binary trials with k successes is the same as the number of ways of selecting k different objects out of n different ones.

The number itself can be derived as follows: there are n possibilities for choosing the number of the first win, then $n-1$ possibilities for choosing the number of the second win, etc, and finally when choosing the k-th win there are $n-k+1$ possibilities (note that $k \leq n$, and if $n = k$ there is only one option left for the last choice.) Thus, the total number of such selections is: $n(n-1)...(n-k+1) = n!/(n-k)!$

But note that we overcounted, because we considered different strings of wins depending on the order in which a win was selected, even if the resulting strings are the same (example: $n = 4$ and $k = 4$ gives us 4! although there is only one string of 4 wins out of 4). In order to correct for the overcounting, we need to divide by the total number of ways of selecting the same string of k wins out of n. This is number of ways of rearranging k wins, or k! Thus, the number we seek is:

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

We can now calculate the general probability of winning k times out of n trials. First, each string of k wins and $n-k$ losses has the probability $p^k(1-p)^{n-k}$. Since we now know that the number of such strings is C_k^n, the probability is:

$$P(\text{k wins in n trials}) = P(B = k) = \binom{n}{k} p^k(1-p)^{n-k}$$

This is the probability distribution of the binomial random variable B.

The binomial random variable has much simpler formulas for the mean and the variance. First, we know that the mean of a sum of random variables is the sum of the means and the binomial random variable is a sum of n Bernoulli random variables X. Let us say X takes only the values of 0 and 1 with probabilities $1-p$ and p, so we can use the additive property of expected value to calculate $E(B)$:

$$E(B) = E\left[\sum_{i=1}^{n} X\right] = \sum_{i=1}^{n} E(X) = \sum_{i=1}^{n} p = np$$

This means that the expected number of heads/successes is the product of the probability of 1 head/success and the number of trials, e.g. if the probability of success is 0.3, then the expected number of successes out of 100 is 30.

Now let us calculate the variance, for which in general the same additive property is not true. But remember that in the section on variance above we showed that the variance of a sum of two random variables is the sum of their two separate variances as long as their covariance is zero. It turns out that for random variables that satisfy the product rule $P(x, y) = P(x)P(y)$ their covariance is 0:

$$E((X-\mu_X)(Y-\mu_Y)) = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)P(x_i, y_j) =$$
$$= \sum_i (x_i - \mu_X)P(x_i) \sum_j (y_j - \mu_Y)P(y_j)$$

We saw in section on variance above that the expected value of deviations from the mean is zero, which gives us:

$$E((X-\mu_X)(Y-\mu_Y)) = E(X-\mu_X)E(Y-\mu_Y) = 0$$

The demonstrates that for independent variables the variance of their sum is the sum of the variances and we can use this to compute the variance of the binomial random variable:

$$Var(B) = Var\left[\sum_{i=1}^{n} X\right] = \sum_{i=1}^{n} Var(X) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

For any given number of Bernoulli trials, the variance has a quadratic dependence on probability of success p: if $p = 1$ or $p = 0$, corresponding to all successes, or all failures, respectively, then the variance is zero, since there is no spread in the outcome. For a fair coin $p = 1/2$ the variance is highest. This can be seen in the plots of binomial random variables for $n = 2$, $n = 5$, and $n = 50$, shown in figures below.

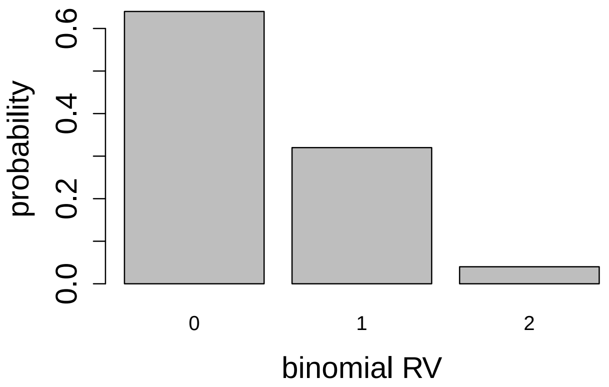


Figure 4.7: The binomial distribution for $n = 2$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

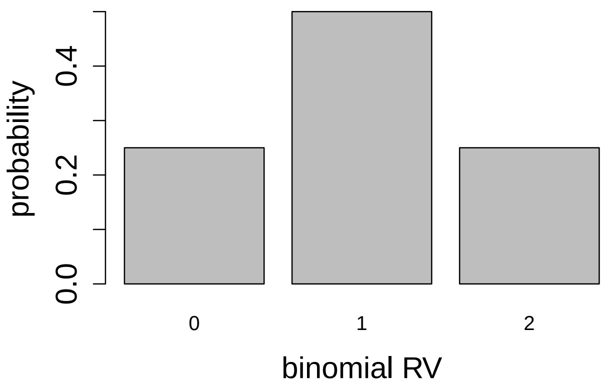


Figure 4.8: The binomial distribution for $n = 2$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

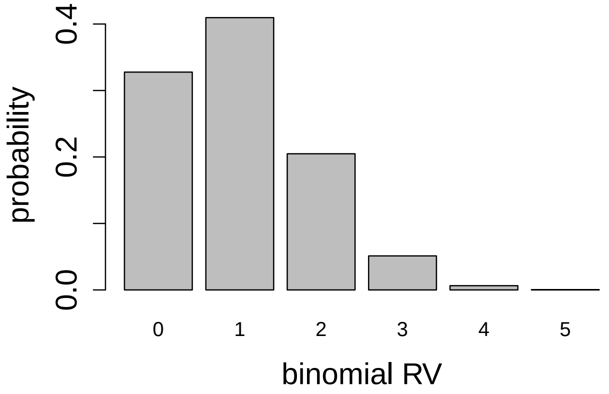


Figure 4.9: The binomial distribution for $n = 5$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

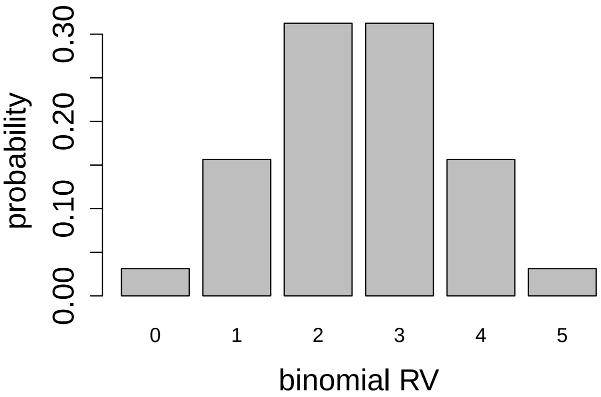


Figure 4.10: The binomial distribution for $n = 5$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

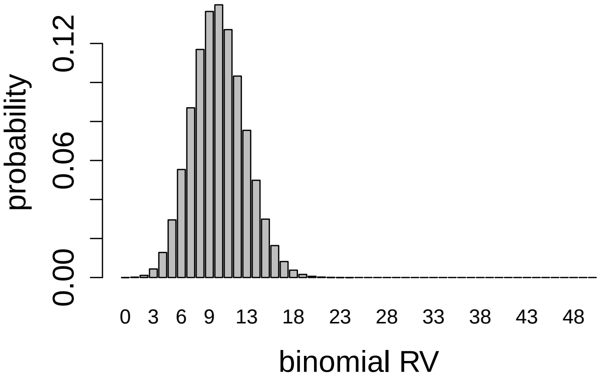


Figure 4.11: The binomial distribution for $n = 50$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

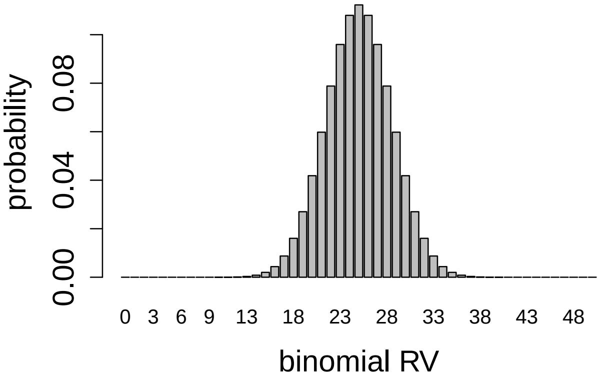


Figure 4.12: The binomial distribution for $n = 50$ and $p = 0.2$ and $p = 0.5$ (you should be able to tell which one is which!)

4.2.3 Exercises

Calculate the means and variances based on the plotted distributions using the definitions **??** and **??** and compare your calculations against equations **??** and **??** (for uniform random variables) and equations **??** and **??** (for binomial random variables)

1. Calculate the mean and the variance for the two uniform distributions plotted in figure @ref(fig:unif-dist).
2. Calculate the mean and the variance for the two binomial distributions plotted in figure @ref(fig:bin-dist-1).
3. Calculate the mean and the variance for the two binomial distributions plotted figure @ref(fig:bin-dist-2).
4. Estimate (approximately) the mean and the variance for the two binomial distributions plotted in figure @ref(fig:bin-dist-3).

4.2.4 testing for mutants

Suppose that you're screening people for a particular genetic abnormality. It is known from prior experience that about 5% of this population carry this mutation. You run your tests on a group of 20 people, and the results indicate that 3 of them are carriers. Clearly, this is higher than you expected - 3/20 is 15%, or 3 times higher than the estimate. One of your colleagues exclaims, What are the odds of this?

To answer this question, one must start by stating your assumptions. First, the people tested must be chosen from the same population, so we can assume a priori each had probability 5% of being a carrier. Second, the people must be selected without bias, that is, selection of one must be unlinked or independent of others. As a counter-example, if your selection included an entire biological family, that would be a biased selection - it may be that the whole family has the mutation, or maybe they don't, but either way probability is no longer determined on a person-by-person basis. If these assumptions are made, then one can calculate the probability of making a selection of 20 people that includes 3 carriers of the mutation, using the binomial distribution.

The formula for the binomial distribution in equation **??** provides the answer for any given number of mutants. For example, the probability of 3 people out of 20 being carriers for the mutation is:

$$P(3 \text{ out of } 20;\ p = 0.05) = \binom{20}{3} \times 0.05^3 \times 0.985^{17} =$$

$$= 1140 \times 0.05^3 \times 0.985^{17} \approx 0.0596$$

One may want to ask a different question: what is the probability that there are at least 3 mutants in the sample of 20 people? To most efficient way to calculate this it is to answer the complementary question first: what is the probability that there are fewer than 3 mutants out of 20 people? This corresponds to three values of the random variable: 0, 1, or 2. We can calculate the total probability by adding up the three separate probabilities, since they represent non-overlapping events (one can't have 1 and 2 mutants in a sample simultaneously):

$$P(B < 3;\ p = 0.05) = P(B = 0) + P(B = 1) + P(B = 2) =$$

$$= \binom{20}{2} \times 0.05^2 \times 0.985^{18} + \binom{20}{1} \times 0.05^1 \times 0.985^{19} + \binom{20}{0} \times 0.05^0 \times 0.985^{20} \approx$$

$$\approx 0.925$$

The answer to the original question is found by taking the complementary probability $1 - 0.925 = 0.075$. Thus the probability of finding at least 3 mutants in a sample of 20 with individual probability 0.0015 is approximately 0.075. The answer is close to the probability of having exactly 3 mutants because the probability of finding more than 3 mutants is very low.

4.3 Random number generators in R

Simulating randomness with a computer is not a simple task. Randomness is contrary to the nature of a computer, which is designed to perform operations exactly. However, there are algorithms that produce a string of numbers that are for all intents and purposes random: there is no obvious connection between one number and the next, and the values don't form any pattern. Such algorithms are called *random number generators*, although to be more precise they produce pseudo-random numbers. The reason is that they actually produce a perfectly predictable string of numbers, which eventually repeats itself, but with a humongous period. One can even produce the same random number, or the same string of random numbers, by specifying the seed for the random number generator. This is very useful if one wants to reproduce the results of a code that uses random numbers.

Of course, random variable are not all the same - they have different distributions. R has a number of functions for producing random numbers from different distributions. For example, to produce random numbers from a set of values with a uniform probability distribution, use the function `sample()`. For instance, the following command produces a random integer between 1 and 20. Repeating the same command produces a new random number, which (most likely) is not the same as the first. The first input argument (`1:20`) is the vector of values from which to draw the random number, and the second is the size of the sample:

```
x <- sample(1:20,1)
y <- sample(1:20,1)
print(x)
```

```
[1] 16
```

```r
print(y)
```

```
[1] 8
```

To generate 10 randomly chosen integers between 1 and 20, see the following two commands, which differ in setting the value of the option `replace`. The first command doesn't specify the value for replace, and by default it is set to FALSE, so the command draws numbers without replacing them (meaning that all the numbers in the sample are unique). In the second command `replace` is set to TRUE, so the numbers that were selected can be chosen again. In both cases, repeatedly running the command results in a different set of randomly chosen numbers, which you should investigate by copying the commands into R and running them yourself.

```r
x <- sample(1:20,10)
print(x)
```

```
 [1] 16  5 11  6  8 19  7 10 18  1
```

```r
y <- sample(1:20,10,replace=TRUE)
print(y)
```

```
 [1] 11  2  5  5 17 10  9  8  4  8
```

If you need to generate a random number from the binomial distribution, R has you covered. The command is `rbinom(s, n, p)` and it requires three input values: s is the number of observations (sample size), n is the number of binary trials in one observation, and p is the probability of success in one binary trial. The following two commands generate a single random number, the number of successes out of 20 trials with probability of success 0.2 and 0.6:

```r
x <- rbinom(1,20,0.2)
print(x)
```

```
[1] 3
```

```
y <- rbinom(1,20,0.6)
print(y)
```

```
[1] 17
```

To generate an entire sample of random numbers, change the first input parameter to 10. As you'd expect, the samples of 10 observations are (most likely) noticeably different: when the probability p is 0.2, the number of successes tend to be less than 6, while for probability 0.6, the numbers are usually greater than 10.

```
x <- rbinom(10,20,0.2)
print(x)
```

```
 [1] 8 6 3 5 1 3 2 1 1 3
```

```
y <- rbinom(10,20,0.6)
print(y)
```

```
 [1] 13 11 12 12 11 12 12 14 14 12
```

Notice that the range of possible values of this random variable is between 0 and 20, but unlike the uniform random numbers produced with the `sample()` function, the probability of obtaining different numbers are different, and depend on the parameter p. Calculation and plotting of the binomial distribution function can be accomplished with the command `dbinom(x,n,p)`, where x is the value of the random variable (between 0 and n), n is the number of trials, and p is the probability of success. For instance, the following script calculate the probability of obtaining 1 success out of 20 with probability $p = 0.2$:

```
n <- 20
p <- 0.2
print(dbinom(1,n,p))
```

```
[1] 0.05764608
```

The script above calculates the probabilities of all of the possible values of the random variable by substituting the vector of these values (e.g. 0 to 20) instead of the number 1, generating the probability distribution vector. This vector is plotted vs. the values of the random variable using the `barplot()` function, producing an aesthetically pleasing plot of the binomial distribution. The script plots two binomial probability distributions, both with $n = 20$, the first with $p = 0.2$ and the second with $p = 0.6$. Notice also the use of the axis labels in `barplot()` using the same options `xlab` and `ylab` as in `plot()` and use the `main` option to produce a title above each plot.

```
values.vec <- 0:n
prob.dist <- dbinom(values.vec,n,p)
barplot(prob.dist,names.arg=values.vec,xlab='binomial RV',ylab='probability',
main='binom dist with n=20 and p=0.2')
p<-0.6
prob.dist <- dbinom(values.vec,n,p)
barplot(prob.dist,names.arg=values.vec,xlab='binomial RV',ylab='probability',
main='binom dist with n=20 and p=0.6')
```

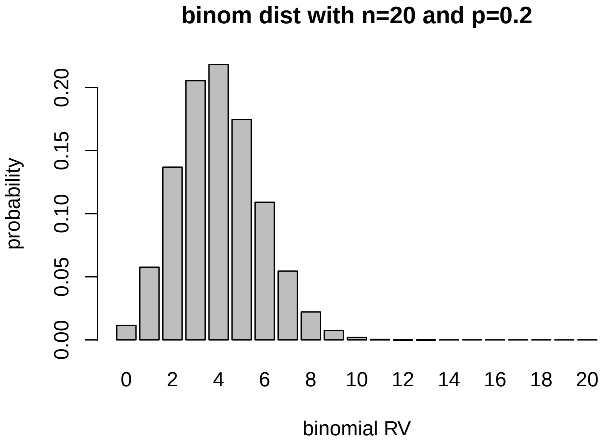


Figure 4.13: The binomial distribution for two different values of n and p produced using dbinom() function.

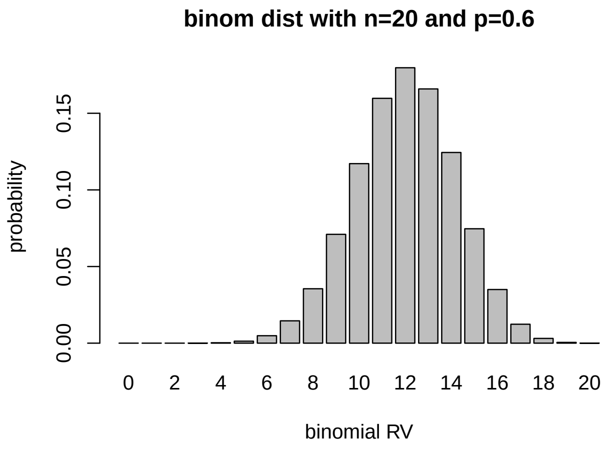


Figure 4.14: The binomial distribution for two different values of n and p produced using dbinom() function.

5 Linear regression

> The place in which I'll fit will not exist until I make it.
> –James Baldwin

In the last two chapters we learned to use data sets which fall into a few categories. We now turn to data which can be measured as a range of numerical values. We can ask a similar question of numerical data that we asked of categorical: how can we tell whether two variables are related? And if they are, what kind of relationship is it? This takes us into the realm of *data fitting*, raising two related questions: what is the best mathematical relationship to describe a data set? and what is the quality of the fit? You will learn to do the following in this chapter:

- define the quality of the fit between a line and a two-variable data set
- calculate the parameters for the best-fit line based on statistics of the data set
- use R to calculate and plot best-fit line for a data set
- understand the meaning of correlation and covariance
- understand the phenomenon of regression to the mean

5.1 Linear relationship between two variables

Although there is always error in any real data, there may be a relationship between the two variables that is not random: for example, when one goes up, the other one tends to go up as well. These relationships may be complicated, but in this chapter we will focus on the the simplest and most common type of relationship: linear, where a change in one variable is associated with a proportional change in the other, plus an added constant. This is expressed mathematically using the familiar equation for a linear function, with parameters slope (a) and intercept (b):

$$y = ax + b$$

Let us say you have measured some data for two variables, which we will call, unimaginatively, x and y. This data set consists of pairs of numbers: one for x, one for y, for example, the heart rate and body temperature of a person go together. They cannot be mixed up between different people, as the data will lose all meaning. We can denote this a list of n pairs of numbers: (x_i, y_i) (where i is an integer between 1 and n). Since this is a list of pairs of numbers, we can plot them as separate points in the plane using each x_i as the x-coordinate and each y_i as the y-coordinate. This is called a *scatterplot* of a two-variable data set. For example, two scatterplots of a data set of heart rate and body temperature are shown in figure **??**. In the first one, the body temperature is on the x-axis, which makes it the *explanatory* variable; in the second one, the body temperature is on the y-axis, which makes it the *response* variable.

```r
data <- read.table("data/HR_temp.txt", header = TRUE)
plot(data$Temp, data$HR, main = "heart rates vs. body temps",
    cex = 1.5, cex.axis = 1.5, cex.lab = 1.5)
plot(data$HR, data$Temp, main = "body temps vs. heart rates",
    cex = 1.5, cex.axis = 1.5, cex.lab = 1.5)
```

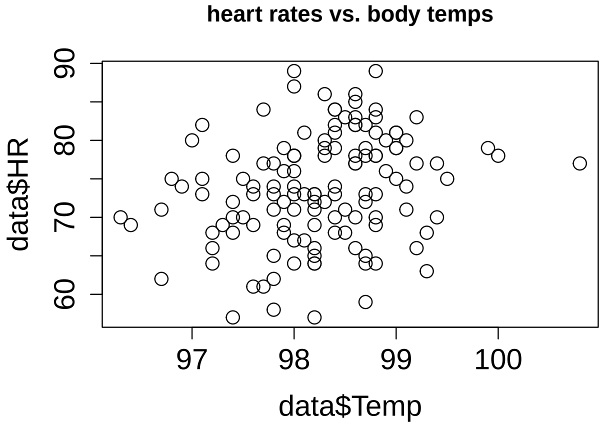


Figure 5.1: Scatterplot of heart rates and body temperatures: a) with heart rate as the explanatory variable; b) with body temperature as the explanatory variable.

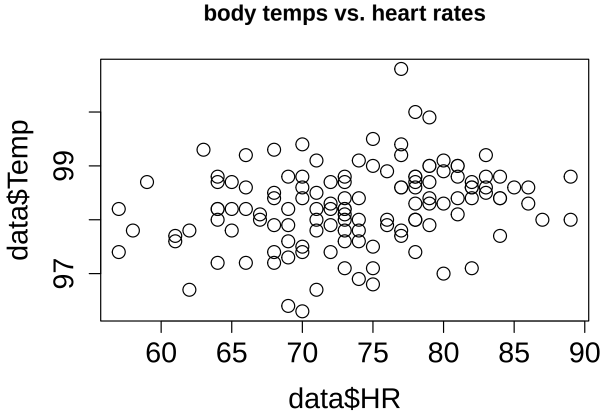


Figure 5.2: Scatterplot of heart rates and body temperatures: a) with heart rate as the explanatory variable; b) with body temperature as the explanatory variable.

5.2 Linear least-squares fitting

5.2.1 sum of squared errors

It is easy to find the best-fit line for a data set with only two points: its slope and intercept can be found by solving the two simultaneous linear equations, e.g. if the data set consists of $(3, 2.3), (6, 1.7)$, then finding the best fit values of a and b means solving the following two equations:

$$\begin{aligned} 3a + b &= 2.3 \\ 6a + b &= 1.7 \end{aligned}$$

These equations have a unique solution for each unknown: $a = -0.2$ and $b = 2.9$ (you can solve it using basic algebra).

However, a data set with two points is very small and cannot serve as a reasonable guide for finding a relationship between two variables. Let us add one more data point, to increase our sample size to three: $(3, 2.3), (6, 1.7), (9, 1.3)$. How do you find the best fit slope and intercept? **Bad idea:** take two points and find a line, that is the slope and the intercept, that passes through the two. It should be clear why this is a bad idea: we are arbitrarily ignoring some of the data, while perfectly fitting two points. So how do we use all the data? Let us write down the equations that a line with slope a and intercept b have to satisfy in order to fit our data points:

$$\begin{aligned} 3a + b &= 2.3 \quad (5.1) \\ 6a + b &= 1.7 \quad (5.2) \\ 9a + b &= 1.3 \quad (5.3) \end{aligned}$$

This system has no exact solution, since there are three equations and only two unknowns. We need to find a and b such that they are a *best fit* to the data, not the perfect solution. To do that, we need to define what we mean by the *goodness of fit*.

One simple way to asses how close the fit is to the data is to subtract the predicted values of y from the data, as follows: $e_i = y_i - (ax_i + b)$. The values e_i are called the *errors* or *residuals* of the linear fit. If the values predicted by the linear model $(ax_i + b)$ are close to the actual data y_i, then the error will be small. However, if we add it all up, the errors with opposite signs will cancel each other, giving the impression of a good fit simply if the deviations are symmetric.

A more reasonable approach is to take absolute values of the deviations before adding them up. This is called the total deviation, for n data points with a line fit:

$$TD = \sum_{i=1}^{n} |y_i - ax_i - b|$$

Mathematically, a better measure of total error is a sum of squared errors, which also has the advantage of adding up non-negative values, but is known as a better measure of the distance between the fit and the data (think of Euclidean distance, which is also a sum of squares) :

$$SSE = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Thus we have formulated the goal of fitting the best line to a two-variable data set, also known as linear regression: **find the values of slope and intercept that result in the lowest possible sum of squared errors**. There is a mathematical recipe which produces these values, which will be described in the next section. Any model begins with assumptions and in order for linear regression to be a faithful representation of a data set, the following must be true:

- the variables have a linear relationship
- all of the measurements are independent of each other
- there is no noise in the measurements of the explanatory variable
- the noise in the measurements of the response variable is normally distributed with mean 0 and identical standard deviation

The reasons why these assumptions are necessary for linear regression to work are beyond the scope of the text, and they are elucidated very well in the book *Numerical Recipes* [**?**]. However, it is important to be aware of them because if they are violated, the resulting linear fit may be meaningless. It's fairly clear that if the first assumption is violated, you are trying to impose a linear relationship on something that is actually curvy. The second assumption of independence is very important and often overlooked. The mathematical reasons for it have to do with properly measuring the goodness of fit, but intuitively it is because measurements that are linked can introduce a new relationship that has to do with the measurements, rather than the relationship between the variables. Violation of this assumption can seriously damage the reliability of the linear regression. The third assumption is often ignored, since usually the explanatory variable is also measured and thus has some noise. The reason for it is that the measure of goodness of fit is based only on the response variable, and there is no consideration of the noise in the explanatory variable. However, a reasonable amount of noise in the explanatory variable is not catastrophic for linear regression. Finally, the last assumption is due to the statistics of maximum-likelihood estimation of the slope and intercept, but again some deviation from perfect normality (bell-shaped distribution) of the noise, or slightly different variation in the noise is to be expected.

5.2.2 best-fit slope and intercept

Definition

The covariance of a data set of pairs of values (X, Y) is the sum of the products of the corresponding deviations from their respective means:

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

Intuitively, this means that if two variable tend to deviate in the same direction from their respective means, they have a positive covariance, and if they tend to deviate in opposite directions from their means, they have a negative covariance. In the intermediate case, if sometimes they deviate together and other times they deviate in opposition, the covariance is small or zero. For instance, the covariance between two independent random variables is zero, as we saw in section **??**.

It should come as no surprise that the slope of the linear regression depends on the covariance, that is, the degree to which the two variables deviate together from their means. If the covariance is positive, then for larger values of x the corresponding y values tend to be larger, which means the slope of the line is positive. Conversely, if the covariance is negative, so is the slope of the line. And if the two variables are independent, the slope has to be close to zero. The actual formula for the slope of the linear regression is [**?**]:

$$m = \frac{Cov(X, Y)}{Var(X)} \tag{5.4}$$

I will not provide a proof that this slope generates the minimal sum of squared errors, but that is indeed the case. To find the intercept of the linear regression, we make use of one other property of the best fit line: in order for it to minimize the SSE, it must pass through the point $(\bar{X}, \bar{Y})$. Again, I will not prove this, but note that the point of the two mean values is the central point of the "cloud" of points in the scatterplot, and if the line missed that central point, the deviations will be larger. Assuming that is the case, we have the following equation for the line: $\bar{Y} = a\bar{X} + b$, which we can solve for the intercept b:

$$b = \bar{Y} - \frac{Cov(X,Y)\bar{X}}{Var(X)} \tag{5.5}$$

5.2.3 Execises

Table 5.1: Body leanness (B) and heat loss rate (H) in boys; partial data set from [?]

B(m^2/kg)	H($^{\circ}C/min$)
7.0	0.103
5.0	0.091
3.6	0.014
3.3	0.024
2.4	0.031
2.1	0.006

Use the data set in table **??** to answer the following questions:

1. Compute the means and standard deviations of each variable.
2. Compute the covariance between the two variables.
3. Calculate the slope and intercept of the linear regression for the data with B as the explanatory variable.
4. Make a scatterplot of the data set with B as the explanatory variable and sketch the linear regression line with the parameters you computed.
5. Calculate the slope and intercept of the linear regression the data with H as the explanatory variable.
6. Make a scatterplot of the data set, with H as the explanatory variable and sketch the linear regression line with the parameters you computed.

5.2.4 correlation and goodness of fit

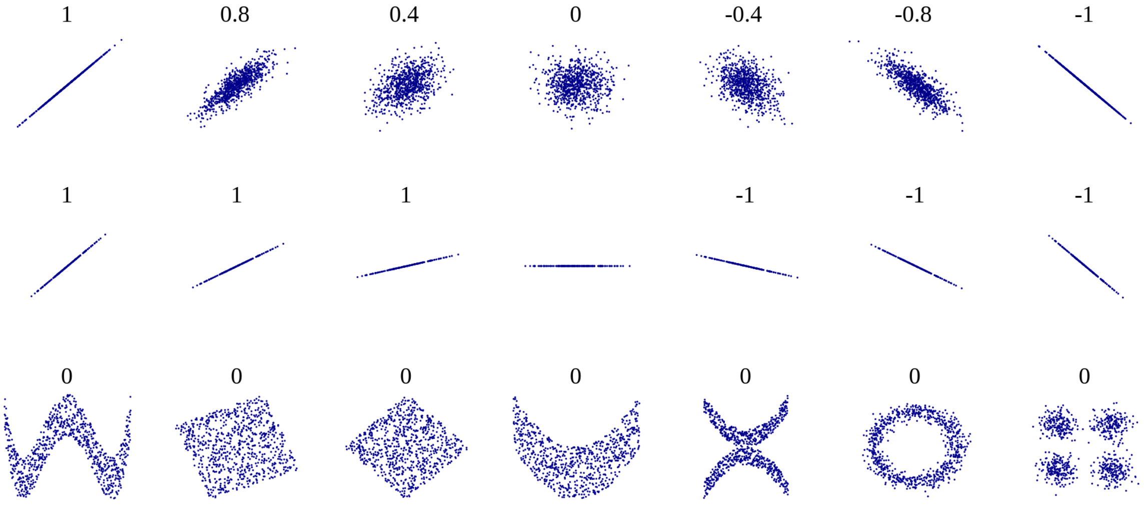
The correlation between two random variables is a measure of how much variation in one corresponds to variation in the other. If this sounds very similar to the description of covariance, it's because they are closely related. Essentially, correlation is a normalized covariance, restricted to lie between -1 and 1. Here is the definition:

Definition

The (linear) correlation of a dataset of pairs of data values (X,Y) is:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

If the two variables are identical, $X = Y$, then the covariance becomes its variance $Cov(X,Y) = Var(X)$ and the denominator also becomes the variance, and the correlation is 1. This is also true if X and Y are scalar multiples of each other, as you can see by plugging in $X = cY$ into the covariance formula. The opposite case if X and Y are diametrically opposite, $X = -cY$, which has the correlation coefficient of -1. All other cases fall in the middle, neither perfect correlation nor perfect anti-correlation. The special case if the two variables are independent, and thus their covariance is zero, has the correlation coefficient of 0.



This gives a connection between correlation and slope of linear regression:

$$a = r\frac{\sigma_Y}{\sigma_X} \tag{5.6}$$

Whenever linear regression is reported, one always sees the values of correlation r and squared correlation r^2 displayed. The reason for this is that r^2 has a very clear meaning of the **the fraction of the variance of the dependent variable** Y explained by the linear regression $Y = aX + b$. Let us unpack what this means.

According to the stated assumptions of linear regression, the response variable Y is assumed to be linear relationship with the explanatory variable X, but with independent additive noise (also normally distributed, but it doesn't play a role for this argument). Linear regression captures the linear relationship, and the remaining error (residuals) represent the noise. Thus, each value of Y can be written as $Y = R + \hat{Y}$ where R is the residual (noise) and the value predicted by the linear regression is $\hat{Y} = aX + b$. The assumption that R is independent of Y means that $Var(Y) = Var(\hat{Y}) + Var(R)$ because variance is additive for independent random variables, as we discussed in section **??**. By the same reasoning $Cov(X, \hat{Y} + R) = Cov(X, \hat{Y}) + Cov(X, R)$. These two covariances can be simplified further: $Cov(X, R) = 0$ because R is independent random noise. X and the predicted $\hat{Y}$ are perfectly correlated, so $Cov(X, \hat{Y}) = Cov(X, mX + b) = Var(X) = Var(\hat{Y})$. This leads to the derivation of the meaning of r^2:

$$\begin{aligned} r^2 = \frac{Cov(X,Y)^2}{Var(X)Var(Y)} = \frac{(Cov(X,\hat{Y}) + Cov(X,R))^2}{Var(X)Var(Y)} = \\ = \frac{Var(X)Var(\hat{Y})}{Var(X)Var(Y)} = \frac{Var(\hat{Y})}{Var(Y)} \end{aligned} \tag{5.7}$$

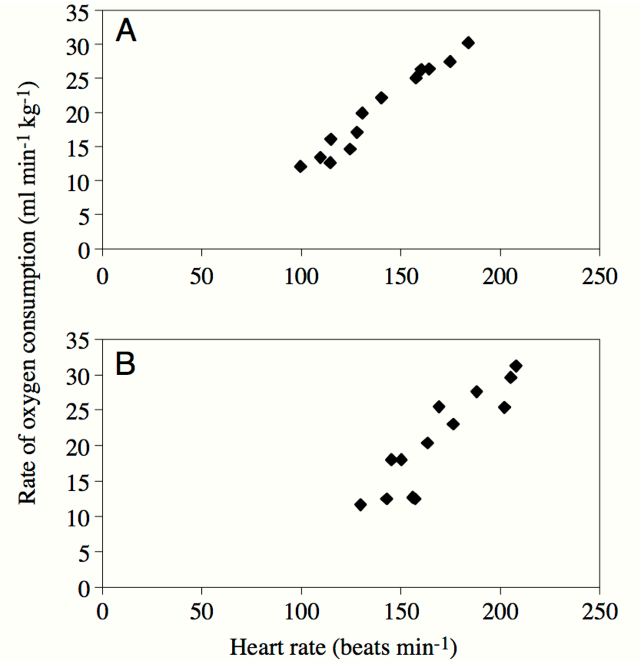
One should be cautious when interpreting results of a linear regression. First, just because there is no linear relationship does not mean that there is no other relationship. Figure **??** shows some examples of scatterplots and their corresponding correlation coefficients. What it shows is that while a formless blob of a scatterplot will certainly have zero correlation, so will other scatterplots in which there is a definite relationship (e.g. a circle, or a X-shape). The point is that **correlation is always a measure of the linear relationship between variables.**

The second caution is well known, as that is the danger of equating correlation with a causal relationship. There are numerous examples of scientists misinterpreting a coincidental correlation as meaningful, or deeming two variables that have a common source as causing one another. For example, one can look at the increase in automobile ownership in the last century and the concurrent improvement in longevity and conclude that automobiles are good for human health. It is well-documented, however, that a sedentary lifestyle and automobile exhaust do not make a person healthy. Instead, increased prosperity has increased both the purchasing power of individuals and enabled advances in medicine that have increase our lifespans. To summarize, one must be careful when interpreting correlation: a weak one does not mean there is no relationship, and a strong one does not mean that one variable causes the changes in the other.

There is another important measure of the quality of linear regression: the residual plot. The residuals are the differences between the predicted values of the response variable and the actual value from the data. As stated above, linear regression assumes that there is a linear relationship between the two variables, plus some uncorrelated noise added to the values of the response variable. If that were true, then the plot of the residuals would look like a vaguely spherical blob, with a mean value of 0 and no discernible trend (e.g. no increase of residual for larger x values). Visually assessing residual plots is an essential check on whether linear regression is a reasonable fit to the data in addition to the r^2 value.

5.2.5 Exercises

Figure **??** shows scatterplots of the rate of oxygen consumption (VO) and heart rate (HR) measured in two macaroni penguins running on a treadmill (really). The authors performed linear regression on the data and found the following parameters: $VO = 0.23HR - 11.62$ (penguin A) and $VO = 0.25HR - 20.93$ (penguin B). The datasets have the standard deviations: $\sigma_{VO} = 6.77$ and $\sigma_{HR} = 28.8$ (penguin A) and $\sigma_{VO} = 8.49$ and $\sigma_{HR} = 30.6$ (penguin B).



1. Find the dimensions and units of the slope and the intercept of the linear regression for this data (the units of HR and VO are on the plot).
2. Data set B has a larger slope than data set A. Does this mean the correlation is higher in data set B than in A? Explain.

3. Calculate the correlation coefficients for the linear regressions of the two penguins; explain how much variance is explained in each case.

4. Re-calculate the slopes of the two linear regressions if the explanatory and response variables were reversed. Does changing the order of variable affect the correlation?

5.3 Linear regression using R

We now have the tools to compute the parameters of the best-fit line, provided we can calculate the means, variances, and covariance of the two variable data set. Of course, the best way to do all this is to let a computer handle it. The function for calculating linear regression in R is `lm()`, which outputs a bunch of information to a variable called myfit in the script below. The slope, intercept, and other parameters can be printed out using the `summary()` function. In the script below you see a bunch of information, but we are concerned with the ones in the first column correspond to the best fit intercept (-166.2847) and the slope (2.4432). You can check that they correspond to our formulas by computing the covariance, the variances, and the means of the two variables:

```
my_data <- read.table("data/HR_temp.txt", header = TRUE)
myfit <- lm(HR ~ Temp, my_data)
summary(myfit)
```

```
Call:
lm(formula = HR ~ Temp, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6413  -4.6356   0.3247   4.8304  15.8474 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -166.2847    80.9123  -2.055  0.04190 * 
Temp           2.4432     0.8235   2.967  0.00359 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.858 on 128 degrees of freedom
Multiple R-squared:  0.06434,	Adjusted R-squared:  0.05703 
F-statistic: 8.802 on 1 and 128 DF,  p-value: 0.003591
```

```
a <- cov(my_data$HR, my_data$Temp)/var(my_data$Temp)
print(a)
```

```
[1] 2.443238
```

```
b <- mean(my_data$HR) - a * mean(my_data$Temp)
print(b)
```

```
[1] -166.2847
```

Here `Temp` and `HR` are the explanatory and response variables, respectively, and `my_data` is the name of the data frame they are stored in. The best fit parameters are stored in `myfit`, and the line can be plotted using `abline(myfit)`. The script below shows how to calculate a linear regression line and then plot it over a scatterplot in R, and the result is shown in figure **??**a.

```
plot(my_data$Temp, my_data$HR, main = "scatterplot and linear regression line",
    cex = 1.5, cex.axis = 1.5, cex.lab = 1.5)
abline(myfit)
HRresiduals <- resid(myfit)
plot(data$Temp, HRresiduals, main = "residuals plot",
    cex = 1.5, cex.axis = 1.5, cex.lab = 1.5)
abline(0, 0)
```

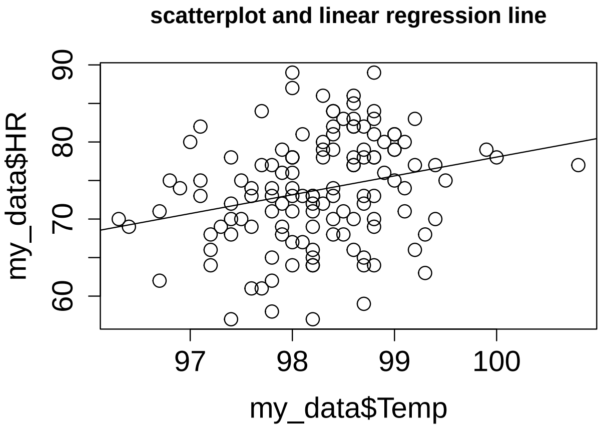


Figure 5.3: Linear regression for a data set of heart rates and body temperatures (a); and the residuals (b).

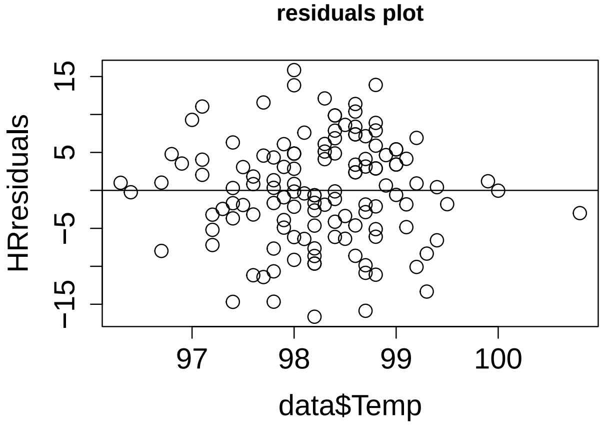


Figure 5.4: Linear regression for a data set of heart rates and body temperatures (a); and the residuals (b).

However, what does this mean about the quality of the fit? Just because we found a line to draw through a scatterplot does not mean that this line is meaningful. In fact, looking at the plot, there does not seem to be much of a relationship between the two variables. There are various statistical measures for the significance of linear regression, the most important one relies on the correlation between the two data sets. Look again at the summary statistics for the data set of heart rates and temperatures. There are several different statistics here, and the one that we care about is the r^2, which is reported here as 'Multiple R-squared'. This number tells us that the linear regression accounts for only about 6% of the total variance of the heart rate. In other words, there is no significant linear relationship in this data set.

As mentioned in section **??**, the other important check is plotting the residuals of the data set, after the linear fit is subtracted. You see the result in figure **??**b, showing that the residuals do not have any pronounced pattern. So it is reasonable to conclude that linear regression was a reasonable model to which to fit the data. The low correlation is because data seem to have little to no relationship, not because there is some complicated nonlinear relationship.

Here is an example of a linear regression performed and the line plotted over the basic R plot. Note that lm() uses the following syntax to indicate which variable is which: lm(Y ~ X) (where Y is the response variable and X is the explanatory variable.)

```
library(HistData)
myfit <- lm(child ~ parent, Galton)
summary(myfit)
```

```
Call:
lm(formula = child ~ parent, data = Galton)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-7.8050 -1.3661  0.0487  1.6339  5.9264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096 
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

```
print(paste("The best-fit slope is: ", myfit$coefficients[2]))
```

```
[1] "The best-fit slope is:  0.646290581993715"
```

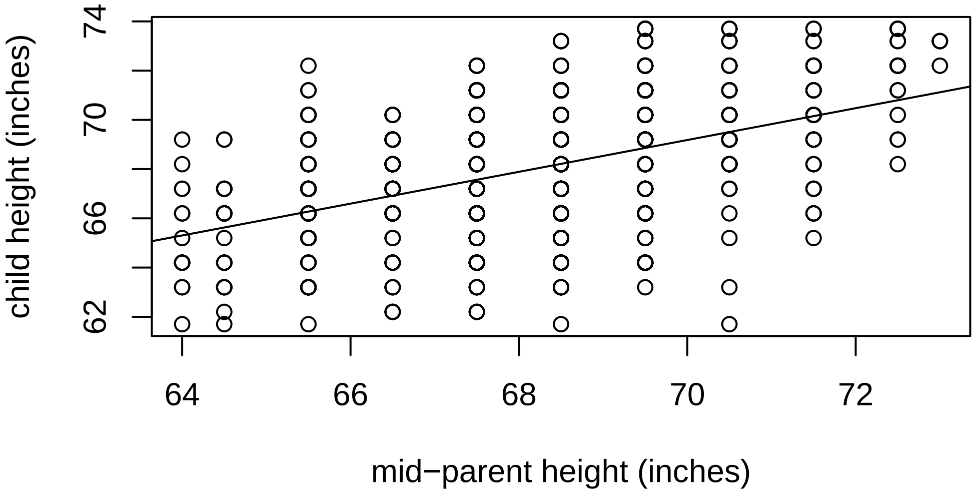
```
print(paste("The best-fit intercept is: ", myfit$coefficients[1]))
```

```
[1] "The best-fit intercept is:  23.9415301804086"
```

The summary outputs a whole bunch of information that is returned by the `lm()` function, as the object `myfit`. The most important are the intercept and slope, which may be printed out as shown above, and the R-squared parameter, also called the coefficient of determination. The value of R-squared is not accessible directly in `myfit`, but it is printed out in the summary (use multiple R-squared for our assignments.)

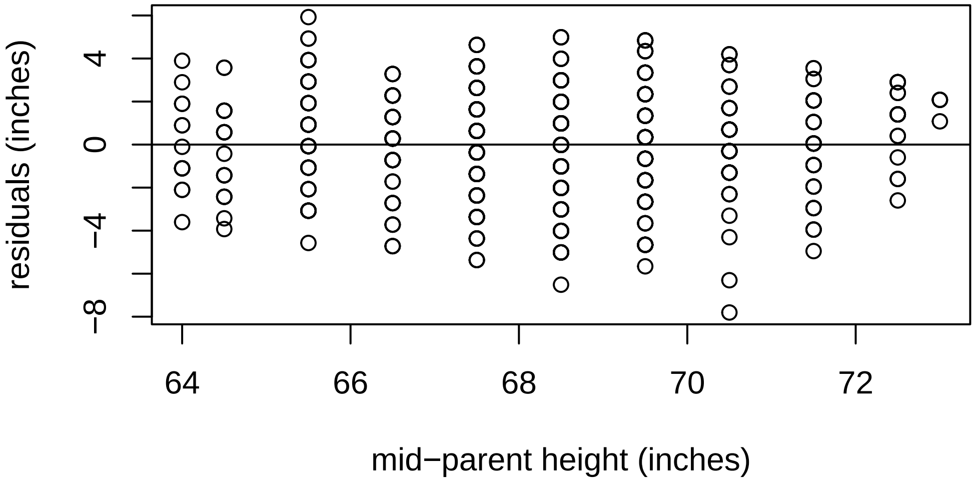
The actual best-fit line can be plotted as follows over a scatterplot of the data; notice that abline can take myfit as an input and use the slope and intercept:

```
#Overlay the best-fit line on the base R plot
plot(Galton$parent, Galton$child, xlab='mid-parent height (inches)', ylab='child height (i
abline(myfit)
```



After performing linear regression it is essential to check that the residuals obey the assumptions of linear regression. The residuals are the difference between the predicted response variable values and the actual values of the response variable, in this case the child height. The residuals are contained in the object myfit as variable residuals:

```
plot(Galton$parent, myfit$residuals, xlab='mid-parent height (inches)', ylab='residuals (i
abline(0,0)
```



It appears that the residuals meet the assumptions of being independent of measurement (shapeless scatterplot), are centered at zero, and look roughly normally distributed, although that can be checked more carefully using other tools.

5.3.1 Exercises:

1. Calculate descriptive statistics (mean and standard deviation) of the residuals from the linear regression above. What do you expect them to be, and how do they differ from the expectation? Using this calculation, check that the coefficient of determination really captures the fraction of total variance explained by linear regression

2. Perform linear regression on the Galton data set with the response and explanatory variables switched, and report which parameters changed and how.

3. Plot the residuals from your new linear regression and calculate and report their descriptive statistics. What do you expect them to be, and how do they differ from the expectation? Using this calculation, check that the coefficient of determination really captures the fraction of total variance explained by linear regression.

5.4 Regression to the mean

The phenomenon called regression to the mean is initially surprising. Francis Galton first discovered this by comparing the heights of parents and their offspring. Galton took a subset of parents who are taller than average and observed that their children were, on average, shorter than their parents. He also compared the heights of parents who are shorter than average, and found that their children were on average taller than their parents. This suggests the conclusion that in long run everyone will converge closer to the average height - hence “regression to mediocrity”, as Galton called it [**?**].

```
library("HistData")
myfit <- lm(child ~ parent, Galton)
plot(Galton$parent, Galton$child, xlab = "mid-parent height (inches)",
    ylab = "child height (inches)", cex = 1.5, cex.axis = 1.5,
    cex.lab = 1.5)
abline(myfit, lwd = 3, lty = 1)
abline(0, 1, lwd = 3, lty = 2, col = "red")
myfit <- lm(parent ~ child, Galton)
plot(Galton$child, Galton$parent, xlab = "child height (inches)",
    ylab = "mid-parent height (inches)", cex = 1.5,
    cex.axis = 1.5, cex.lab = 1.5)
abline(myfit, lwd = 3)
abline(0, 1, lwd = 3, lty = 2, col = "red")
```

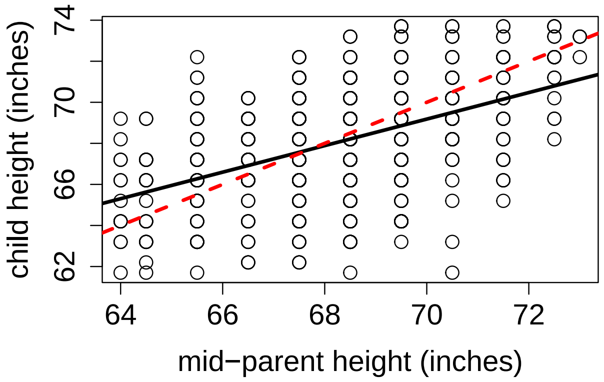


Figure 5.5: Galton data on heights of parents and of children as scatterplots (two versions with explanatory and response variables switched). The dotted red lines show the identity line y=x and the solid black line is the linear regression.

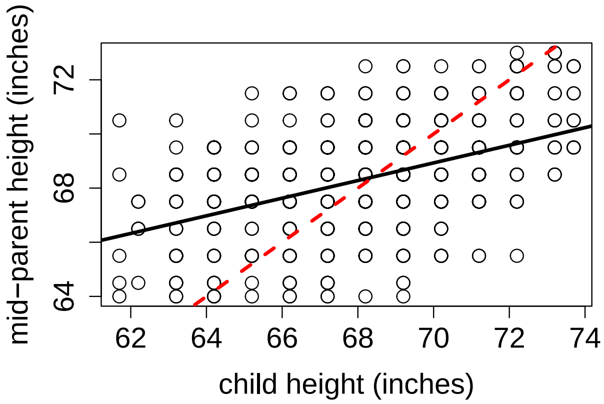


Figure 5.6: Galton data on heights of parents and of children as scatterplots (two versions with explanatory and response variables switched). The dotted red lines show the identity line y=x and the solid black line is the linear regression.

But that is not the case! The parents and children in Galton's experiment had a very similar mean and standard deviation. This appears to be a paradox, but it is easily explained using linear regression. Consider two identically distributed random variables (X, Y) with a positive correlation r. The slope of the linear regression is $m = r\sigma_Y/\sigma_X$ and since $\sigma_Y = \sigma_X$, the slope is simply r. Select a subset with values of X higher than $\bar{X}$, and consider the mean value of Y for that subset. If the slope $m < 1$ (the correlation is not perfect), then the mean value of Y for that subset is less than the mean value of X. Similarly, for a subset with values of X lower than $\bar{X}$, the mean value of Y for that subset is greater than the mean value of X, again as long as the slope is less than 1.

Figure **??** shows Galton's data set (available in R by installing the package 'HistData') along with the linear regression line and the identity like ($y = x$). If each child had exactly the same height as the parents, the scatterplot would lie on the identity line. Instead, the linear regression lines have slope less than 1 for both the plot with the parental heights as the explanatory variable and for the plot with the variables reversed. The correlation coefficient r does not depend on the order of the variables; so using the equation **??** we can see the difference in slopes is explained by the two data sets having different standard deviations, and reversing the explanatory and response variables results in reciprocation of the ratio of standard deviations. The children's heights have a higher standard deviation, which is likely an artifact of the experiment. In the data set the heights of the two parents were averaged to take them both into account, which substantially reduces the spread between male and female heights. To summarize, although the children of taller parents are shorter on average than their parents, and the children of shorter parents are taller than their parents, the overall standard deviation does not decrease from generation to generation.

5.4.1 Discussion questions

Please read the paper on measuring the rate of de novo mutations in humans and its relationship to paternal age [**?**].

1. What types of mutations were observed in the data set? What were the most and the least common?
2. The paper shows that both maternal and paternal age are positively correlated with offspring inheriting new mutations. What biological mechanism explains why paternal age is the dominant factor? What could explain the substantial correlation with maternal age?
3. Is linear regression the best representation of the relationship between paternal age and number of mutations? What other model did the authors use to fit the data, and how did it perform?
4. What do you make of the historical data of paternal ages the authors present at the end of the paper? Can you postulate a testable hypothesis based on this observation?

6 Independence

> Unconnected and free
> No relationship to anything.
> – They Might Be Giants, *Unrelated Thing*

In the first part of the book we learned how to describe data sets and probability distributions of random variables. So far we have not discussed how two or more variables may influence each other, and the next four chapters will be devoted to relationships between two variables. Many experiments in biology result in observations that naturally fall into a few categories, for example: sick or healthy patients, presence or absence of a mutation, etc. The resulting data sets are called *categorical.* Unlike numerical data sets that we will investigate later in chapters 8 and 9, they are not usually represented by numbers. Although it is possible, for instance, to denote mutants with the number 1 and wild type with 0, such designation does not add any value. Categorical variables require different tools for analysis than numerical ones; one cannot compute a linear regression between two categorical variables, because there is no meaningful way to place categories on axes. In this chapter you will learn the following:

- notion of conditional probability
- definition of independence for events and random variables
- produce a categorical data table
- compute a table of expected values based on independence

6.1 Contingency tables to summarize data

What kind of relationship can there be between categorical variables? It cannot be expressed in algebraic form, because without numeric values we cannot talk about a variable increasing or decreasing. Instead, the question is, does one variable being in a particular category have an effect on which category the second variable falls into? Let us say you want to know whether the age of the mother has an effect on the child having trisomy 21 (a.k.a. Down's syndrome), a genetic condition in which an embryo receives three chromosomes 21 instead of the normal two. The age of the mother is a numerical variable, but it can be classified into two categories: less than 35 and 35 or more years of age. The trisomy status of a fetus is clearly a binary, categorical variable: the fetus either has two chromosomes 21 or three.

The data are presented in a two-way or *contingency table*, which is a common way of presenting a data set with two categorical variables. The rows in such tables represent different categories of one variable and the columns represent the categories of the other, and the cells contain the data measurements of their overlaps. Table **??** shows a contingency table for the data set on Down's syndrome and maternal age, in which the rows represent the two categories of maternal age and the columns represent the presence or absence of the syndrome. Each internal cell (as opposed to the total counts on the margins) corresponds to the number of measurement where both variables fall into the specified category, for instance the number of fetuses with the syndrome and a mother under 35 is 28.

Maternal age	No DS	DS	Total
< 35	29,806	28	29,834
$>= 35$	8,135	64	8,199
Total	37,941	92	38,033

Contingency table for maternal age and incidence of Down's syndrome. Numbers represent counts of patients belonging to both categories in the row and the column. DS = Down's syndrome. From [**?**].

Once the data are organized into a contingency table, we can address the main question stated above: does the age of the mother have an effect on whether a fetus inherits three chromosomes 21? Perhaps the first approach that suggests itself is to compare the fraction of mothers carrying a fetus with DS for the two age categories. In this case, the fraction for the under-35 category is $28/29834 \approx 0.00094$, while for the 35-and-over category the fraction is $64/8199 \approx 0.0078$. The two fractions are different by almost a factor of 10, which suggests a real difference between the two categories. However, all data contain an element of randomness and a pinch of error, thus there needs to be quantifiable way of deciding what constitutes a real effect. But to determine if there is a relationship, we first have to define what it means to not have one.

6.2 Conditional probability

Let us return to the abstract description of probability introduced in section 8.1. There we used the notion of sample space and its subsets, called events, to describe collections of experimental outcomes. Suppose that you have some information about a random experiment that restricts the possible outcomes to a particular subset (event). In other words, you have ruled out some outcomes, so the only possible outcomes are those in the complementary set. This will affect the probability of other events in the sample space, because your information may have ruled out some of the outcomes in that event as well.

Definition

For two events A and B in a sample space Ω with a probability measure P, the probability of A given B, called the *conditional probability* is defined as:

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

$A\&B$ represents the intersection of events A and B, also known as A and B, the event that consists of all outcomes that are in both A and B. In words, given the knowledge that an event B occurs, the sample space is restricted to the subset B, which is why the denominator in the definition is $P(B)$. The num‘erator is all the outcomes we are interested in, which is A, but since we are now restricted to B, the numerator consists of all the elements of A which are also in B, or $A\&B$. The definition makes sense in two extreme cases: if $A = B$ and if A and B are mutually exclusive:

- $P(B|B) = P(B\&B)/P(B) = P(B)/P(B) = 1$ (probability of B given B is 1)
- if $P(A\&B) = 0$, then $P(A|B) = 0/P(B) = 0$ (if A and B are mutually exclusive, then probability of A given B is 0)

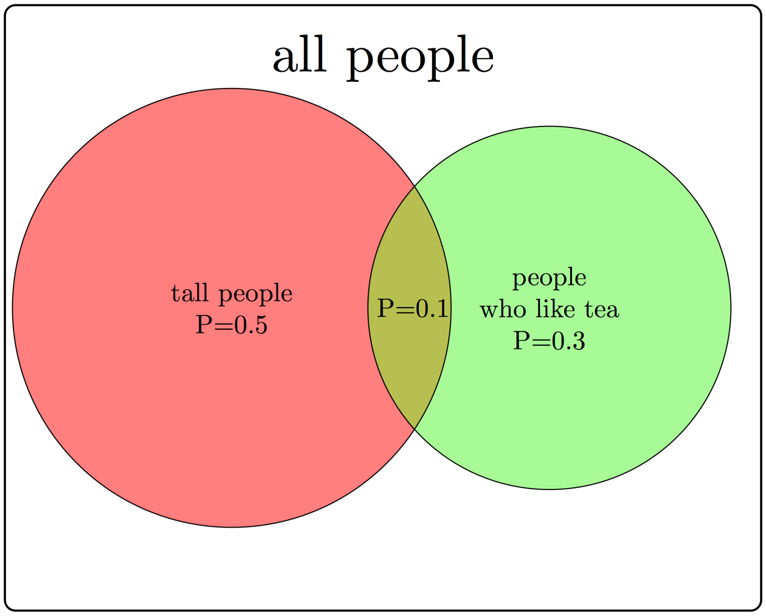


Figure 6.1: A Venn diagram of the sample space of all people with two events: tall people (A) and those who like tea (B) with probabilities of A, B and their intersection indicated.

There are some common misunderstandings about conditional probability, which are usually the result of discrepancies between everyday word usage and precise mathematical terminology.

First, the probability of A given B is not the same as probability of A and B. These concepts seem interchangeable because the statement "what are the odds of finding a tall person who likes tea?" is hard to distinguish from ''what are the odds that a person who is tall likes tea?" The difference in these concepts can be illustrated using a Venn diagram, shown in figure **??**. Based on the probabilities indicated there, the probability of randomly selecting a person who is both tall and likes tea is $P(A\&B) = 0.1$, while the probability that a tea drinker is tall is $P(A|B) = 0.1/0.3 = 1/3$, which are different values.

A similar misconception is to be cavalier about the order of conditionality. In general, $P(A|B) \neq P(B|A)$, except in special cases. Going back to the illustration in figure **??**, the probability that a tea drinker is tall $P(A|B) = 1/3$ is the different than the probability that a tall person is a tea drinker $P(B|A) = 0.1/0.5 = 0.2$. One must take care when interpreting written statements to carefully distinguish what is known *a priori* and what remains under investigation. In the statement $P(A|B)$, B represents what is known, and A represents what is still to be investigated.

Example. Let us return to the data set in the previous section. Data table **??** describes a sample space with four outcomes and several different events. One can calculate the probability of a fetus having Down's syndrome (event) based on the entire data set of 38,033 mothers, and 92 total cases of DS, so the probability is $92/38,033 \approx 0.0024$. Similarly, we can calculate the probability of a mother being above 35 as $8,199/38,033 \approx 0.256$.

Now we can calculate the conditional probability of a mother over 35 having a DS fetus, but first we have to be clear about what information is known and what is not. If the age of the mother is known to be over 35 (mature age or MA), then we calculate $P(DS|MA) = 64/8,199 \approx 0.008$. Notice that the denominator is restricted by the information that the mother is over 35, and thus only women in that category need to be considered for the calculation.

On the other hand, if we have the information that the fetus has DS, we can calculate the reversed conditional probability, what is the probability that a fetus with DS has a mother above age 35? $P(MA|DS) = 64/92 \approx 0.7$. Notice that is both calculations the numerators are the same, since they both are the intersection between the two events, but the denominators are different, because they depend on which event is given.

6.2.1 Exercises

In figure **??** there is a table of genotypes from the classic Mendelian experiment with genetics and color of pea flowers. The parents are both heterozygous, meaning each has a copy of the dominant (purple) allele B and the recessive (white) allele b. The possible genotypes of offspring are shown inside the square, and all four outcomes have equal probabilities. Based on this information, answer the following questions.

1. What is the probability of an offspring having purple flowers? white flowers?

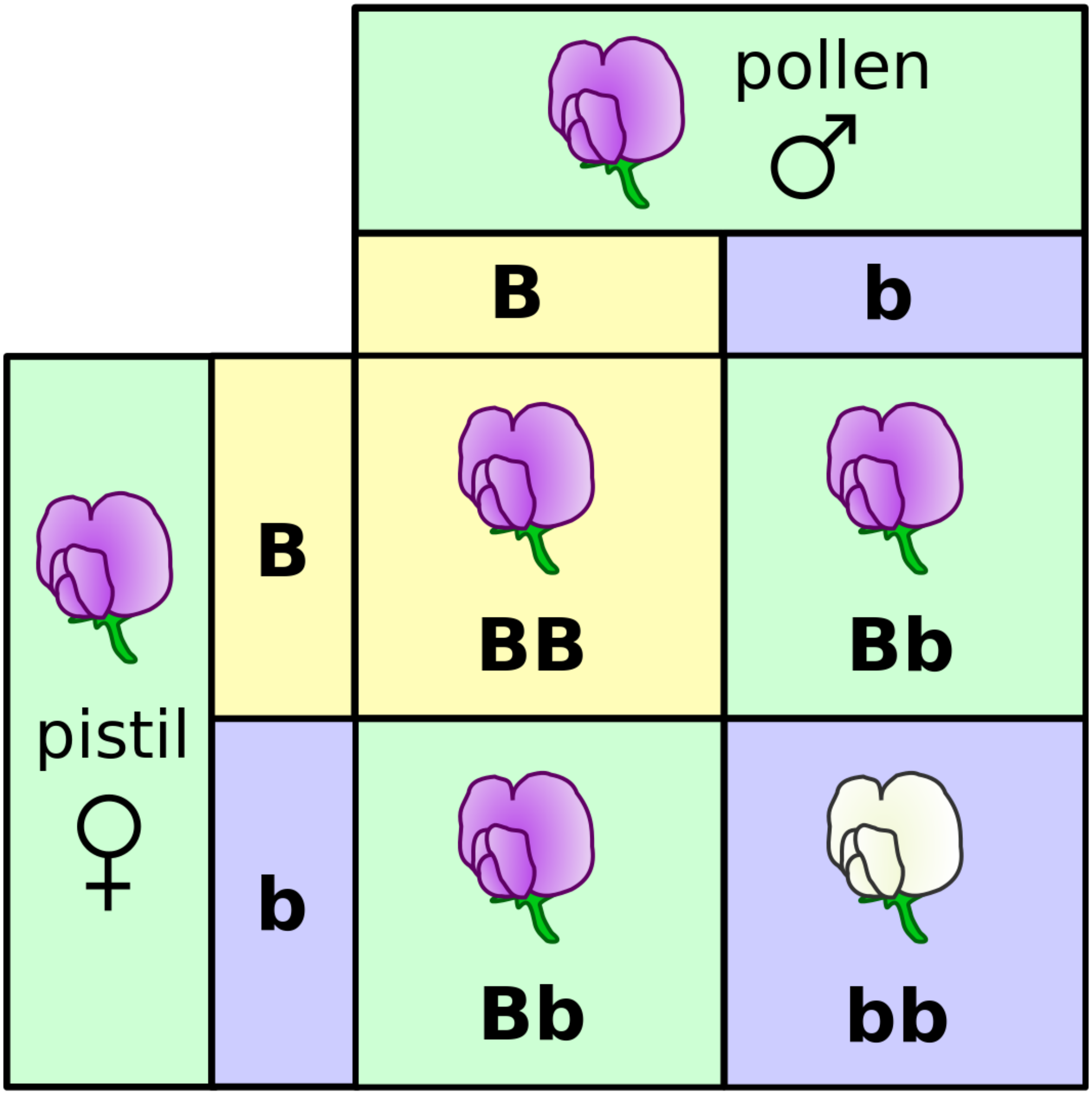


Figure 6.2: Punnet square of a cross of two heterozygous pea plants showing the possible genotypes and phenotypes of offspring (figure by Madprime in public domain via Wikimedia Commons.)

2. What is the probability of an offspring having genotype BB? genotype Bb? genotype bb?
3. What is the probability of an offspring having genotype BB, given that its flowers are purple?
4. What is the probability of an offspring having genotype Bb, given that its flowers are purple?
5. What is the probability of an offspring having genotype BB, given that its flowers are white?
6. What is the probability of an offspring having genotype Bb, given that its flowers are white?}
7. What is the probability of an offspring having purple flowers, given that its genotype is BB?

6.3 Independence of events

We first encountered the notion of independence in chapter 3, where two events were said to be independent if they did not affect each other. The mathematical definition uses the language of conditional probability to make this notion precise. It says that A and B are independent if given the knowledge of A, the probability of B remains the same, and vice versa.

> Definition
>
> Two events A and B are *independent* if $P(A|B) = P(A)$, or equivalently if $P(B|A) = P(B)$.

Independence is not a straightforward concept. It may be confused with mutual exclusivity, as one might surmise that if A and B have no overlap, then they are independent. That however, is false by definition, since $P(A|B)$ is 0 for two mutually exclusive events. The confusion stems from thinking that if A and B are non-overlapping, then they do not influence each other. But the notion of influence in this definition is about information; so of course if A and B are mutually exclusive, the knowledge that one of them occurs has an influence of the probability of the other one occurring.

A useful way to think about independence is in terms of fractions of outcomes. The probability of A is the fraction of outcomes out of the entire sample space which is in A, while the probability of A given B is the fraction of outcomes in B which are also in A. The definition of independence equates the two fractions, therefore, if A occupies $1/2$ of sample space, in order for A and B to be independent, events in A must constitute $1/2$ of the event B. In the illustration in figure **??**, the fraction of tall people is 0.5 of the sample space, but the fraction of tea-drinkers who are tall is $0.1/0.3 = 1/3$. Since the two fractions are different, A and B are not independent.

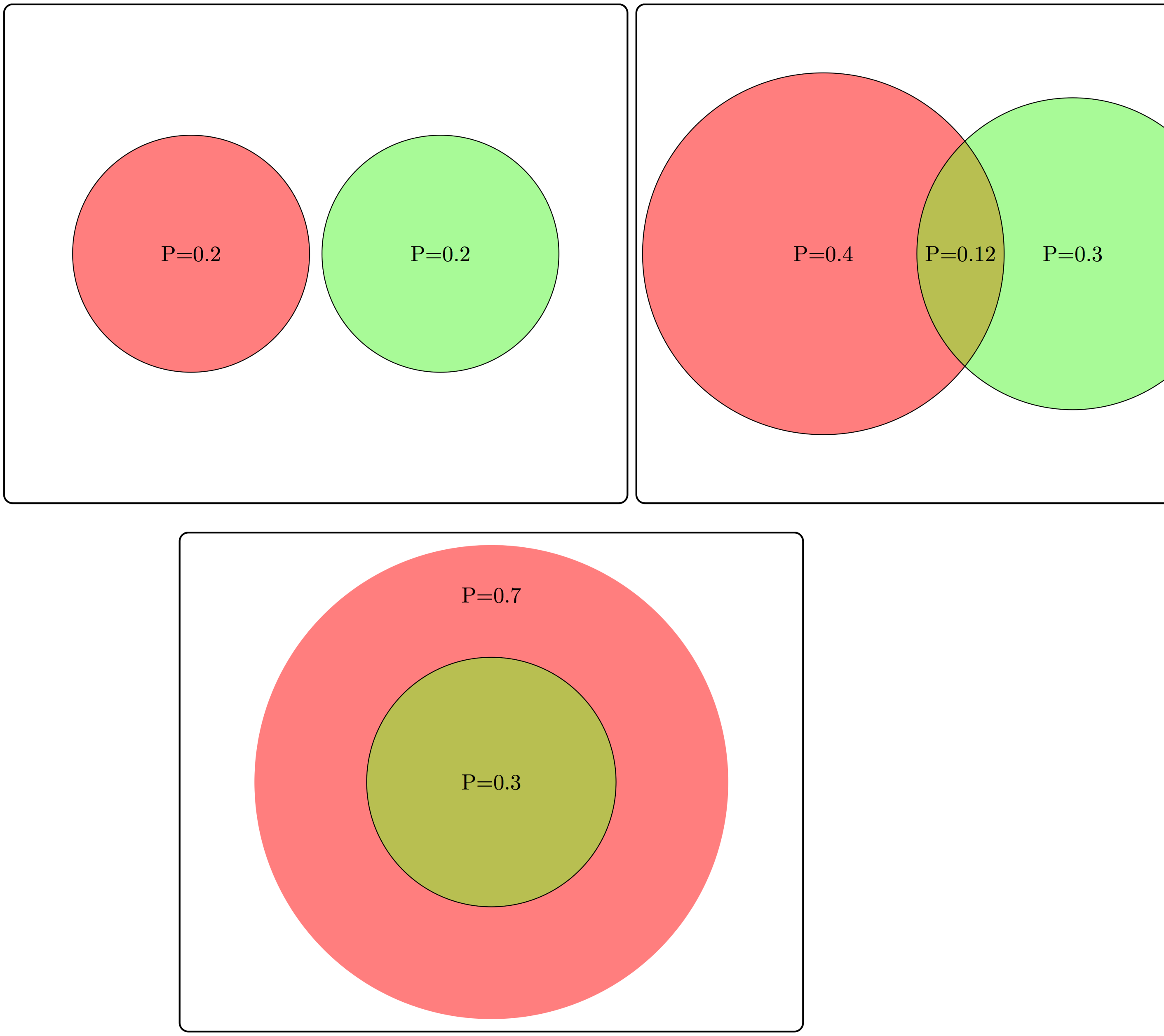


Figure 6.3: Illustration of two events inside a sample space where one is entirely contained in the other

6.3.1 Exercises

Consider three examples of events and their intersections in figure **??**.

1. Based on the two non-overlapping (mutually exclusive) events, calculate the conditional probability $P(A|B)$ and compare it with $P(A)$. Are A and B independent?
2. Based on the two partially overlapping events, calculate the conditional probability $P(A|B)$ and compare it with $P(A)$. Are A and B independent?
3. Based on the two completely overlapping events, calculate the conditional probability $P(A|B)$ and compare it with $P(A)$. Are A and B independent?

6.3.2 product rule

The definition of independence is abstract, but it has a direct consequence of great computational value. From the definition of conditional probability, $P(A|B) = P(A \cap B)/P(B)$, and if A and B are independent then $P(A|B)$ can be replaced with $P(A)$, leading to the expression $P(A) = P(A\&B)/P(B)$. Multiplying both sides by $P(B)$ gives us the formula called the *product rule*, which states that for two independent events the probability of both of them occuring is the product of their separate probabilities:

$$P(A\&B) = P(B)P(A)$$

The product rule is extremely useful for computing probability distributions of complicated random variables. Recall that the binomial distribution, which we saw in section **??** is based on a string of n Bernoulli trials which are independent of each other, which allows the calculation of the probability of a string of successes and failures, or heads/tails, etc. In practice, independence between processes is rarely true in the idealized mathematical sense. However, computing the probability of two random variables without independence is extremely difficult, so it is useful to make the independence assumption and then test it against the data. If it stands up, you have a good predictive model, and if it does not, you have learned that two processes are somehow linked, which is very useful.

6.4 Independence of variables

The product rule enables us to extend the notion of independence from events to variables. The concepts of independence is the same in both contexts, since the probability of a value x of a random variable X corresponds to the probability of the event that gets mapped to x by the variable. In order to make independence applicable to variables, the condition must hold true for all possible values of both random variables. That way, knowing the value of one variable has no effect on the probability of the other. In order to make it simpler to calculate, we will use the product rule as the equivalent condition for independence:

> Definition
>
> Two random variables X and Y are *independent* if for all possible values of X and Y it is true that
> $$P(X=a \& Y=b) = P(X=a)P(Y=b)$$

This allows us to address the question posed at the beginning of the chapter: how can one determine whether a data set has independent variables? The definition allows us to calculate what we would expect if the variables were independent. Given a data set in the form of a contingency table, such as table **??**, we can first calculate the probabilities of the two variables separately, and then from that predict the probabilities of the two variables together.

Example. Let us calculate the expected probabilities and frequencies of Down's syndrome in pregnant women in the two age categories. First, compute the probabilities of having Down's syndrome (and not having it), based on all the pregnancies in the data set: $P(DS) = 92/38033 \approx 0.002419$; the complementary probability is $P(no\ DS) = 1 - P(DS)$. Similarly, we can calculate the probability that a pregnant woman is 35 or over, based on the entire data set (let's denote this event MA for mature age). $P(MA) = 8199/38033 \approx 0.21558$; the complementary probability is $P(YA) = 1 - P(MA)$ (YA stands for young age).

These separate probabilities were calculated from the data, and now we can use them to calculate the predicted probabilities of different outcome, based on the assumption of independence. The probability of a mature-age woman having a pregnancy with Down's syndrome, based on the product rule is $P(MA\&DS) \approx 0.0024 \times 0.216 = 0.000518$. Similarly, we can calculate the probabilities of the other three outcomes: $P(YA & DS) 0.0019 $; $P(MA\&no\ DS) \approx 0.2156$; $P(YA\&no\ DS) \approx 0.782$.

These probabilities are the predictions based on the assumption that the two variables are independent. To compare the predictions with the data, we need to take one more step: convert the probabilities into counts, or frequencies of each occurrence. Since the probability is a fraction out of all outcomes, to generate the predicted frequency we need to multiply the probability by the total number of data points, in this case pregnant patients. The results of this calculation are seen in table **??** with expected frequencies shown instead of experimental observations.

Maternal age	No DS	DS	Total
< 35	29,761.8	72.2	29,834
>= 35	8,179.2	19.8	8,199
Total	37,941	92	38,033

Expected frequencies of Down's syndrome for two different age groups of mothers, assuming that age and Down's syndrome are independent.

Notice that expected frequencies do not need to be integers, because they are the result of prediction and not a data measurement. Now that we have a prediction, we can compare it with the measurements in the data table **??**. The numbers are substantially different, and we can see that the predicted frequency of Down's syndrome for women under 35 is larger than the frequencies for women at or above 35, due to the larger fraction of patients in the younger age group. We can calculate the differences between the *observed* and *expected* contingency tables to measure how much reality differs from the assumption of independence:

Maternal age	No DS	DS	Total
< 35	44.2	-44.2	29,834
>= 35	-44.2	44.2	8,199
Total	37,941	92	38,033

Differences between the observed frequencies of Down's syndrome for different maternal ages and the expected frequencies based on the assumption of independence.

The table of differences shows that the observed frequency of DS in the data set are higher than expected by 44.2 for women above 35 years of age and is lower than expected by the same number for women below age 35. This demonstrates that mathematically speaking, the two variables of age and DS are not independent.

However, real data is messy and subject to randomness of various provenance. First, there is sampling error that we explored in chapter 9, which means that samples from two perfectly independent variables can and will differ from expected frequencies. Second, measurement errors or environmental noise can contribute more randomness to the data. Thus, simply checking that observed frequencies are different from expected is not enough to conclude that the variables are not independent. We need a method to decide what scale of differences is enough to declare that there is an effect e.g. of maternal age on the likelihood of DS. To do this, we leave the cozy theoretical confines of probability and venture into the wild and treacherous world of statistics.

7 Hypothesis testing

> Sometimes I'm right and I can be wrong
> My own beliefs are in my song.
> – Sly and the Family Stone, *Everyday People*

This chapter introduces hypothesis testing and explains how to evaluate the results. This fundamentally involves two steps: stating the hypothesis and then making the binary decision whether to reject it or not. Although such a binary approach is necessarily reductive, there are many situations that make it necessary: deciding whether to approve a drug or start a treatment, for example. Much of the scientific method is based on hypothesis testing: scientists formulate an idea (hypothesis), then accumulate data that can challenge it, and if the data contradict the hypothesis, they discard it (the hypothesis, not the data!) No hypothesis in science is ever proven in an absolute sense, which is why it is fundamentally different from mathematics. A hypothesis that has survived many tests and was found to be consistent with all available observations becomes a theory, like the theory of gravity or of evolution. But unlike a theorem, a scientific theory is not certain, and if solid evidence were to surface that contradicts Newton's gravitational theory, it would be falsified and thrown out (again, the theory, not the evidence.)

In this chapter we will describe the framework of hypothesis testing and apply it to the specific task of deciding whether two variables are independent. After reading it you will know how to:

- Explain the difference between the truth of the hypothesis and a test result
- Describe four different outcomes of hypothesis testing
- Compute different hypothesis testing error rates
- Explain the meaning of p-value
- Use R to perform the chi-squared test

7.1 Terminology and quality measures

7.1.1 positives and negatives

In the classic statistical framework, the hypothesis to be tested is usually called the *null hypothesis*, which helpfully rhymes with dull, because it represents the lack of anything interesting, essentially the default state of the system. In order to reject the null hypothesis, the data has to be substantially different from what is expected as default. For instance, medical tests have the null hypothesis that the patient is normal/healthy, and only if the results are substantially different from normal the patient is considered ill. Another common example is the criminal justice system: a defendant on trial undergoes a binary test where the null hypothesis is innocence. Only if the prosecutor's evidence is strong, that is, shows guilt beyond a reasonable doubt, that the null hypothesis is rejected and the defendant found guilty.

Tests are binary, in that there are only two possible decisions: to reject the hypothesis or to not reject it. We can never truly accept a hypothesis as true, due to the impossibility of perfect knowledge of the world. The decision to reject a hypothesis is called a *positive* test result, which seems backwards, but remember that the default or null hypothesis is a lack of anything unusual or interesting, so if the data are different from default, it is called a positive result. The decision to not reject the null hypothesis is called a *negative* test result. You are probably familiar with this in a medical context: if you've ever been tested for a disease, you know that a negative result is good news!

7.1.2 types of errors

Hypothesis testing gives us a positive or negative result, but that does not mean that it is correct. Ideally, we want the test to reject a false null hypothesis, and not reject a true null hypothesis. These results are called, respectively, a *true positive* and a *true negative*. We can think of the hypothesis as a variable that can be either true or false, and of the test result as another variable than can be positive or negative. In the language of probability, the correct test results can be defined as follows:

> ℹ Definition
>
> For a hypothesis that can be either false (F) or true (T) and a test result that can be either positive (P) or negative (N), the probabilities of a *true positive* and *true negative* are:
>
> $$P(TP) = P(P\&F); \; P(TN) = P(N\&T)$$

However, hypothesis tests are not infallible, and they can make mistakes of two different types. A test that rejects a true null hypothesis makes a *type I error* or a *false positive* error, while a test that fails to reject a false null hypothesis makes a *type II error* or a false negative error. We can again define the probabilities of the two error types as the overlap of the events:

> ℹ Definition
>
> For a hypothesis that can be either false (F) or true (T) and a test result that can be

either positive (P) or negative (N), the two types of errors are:

$$P(FP) = P(P\&T);\ P(FN) = P(N\&F)$$

Test result	$H_O = F$	$H_O = T$
Positive	TP	FP
Negative	FN	TN

Table summarizing the four possible results of hypothesis testing, depending on the truth of null hypothesis H_0 and on the testing result.

7.1.3 test quality measures

Now that we have classified the four outcomes of hypothesis testing, we can define the measures of quality of a given hypothesis test. This aims to address a practical concern: how much can you trust a test result? One may answer this question by testing on data where the hypothesis is known to be either true or false. For example, if there is a "gold standard" method for determining the presence or absence of disease, one can use that information to measure the quality of a new test. By performing enough tests, we can measure the frequencies of the four testing outcomes and then measure the following two quality metrics:

ℹ Definition

The *sensitivity* (or power) of a test is the probability of obtaining a positive result, given a false hypothesis.

$$Sens = P(P|F) = \frac{P(TP)}{P(TP) + P(FN)}$$

The *specificity* of a test is the probability of obtaining the negative result, given a true hypothesis.

$$Spec = P(N|T) = \frac{P(TN)}{P(TN) + P(FP)}$$

Note that these are conditional probabilities, premised on knowing whether the hypothesis is actually true. On the other hand, there are two kinds of *error rates*:

ℹ Definition

The *type I error rate* or *false positive rate* is the probability of obtaining the positive result, given a true hypothesis (complementary to specificity):

$$FPR = \frac{FP}{TN + FP}$$

The *type II error rate* or *false negative rate* is the probability of obtaining the negative result, given a false hypothesis (complementary to sensitivity).

$$FNR = \frac{FN}{TP + FN}$$

Notice that knowledge of sensitivity and specificity determine the type I and type II error rates of a test since they are complementary events. Of course, it is desirable for a test to be both very sensitive (reject false null hypotheses, detect disease, convict guilty defendants) and very specific (not reject true null hypotheses, correctly identify healthy patients, acquit innocent defendants), but that is impossible in practice. In fact, making a test highly sensitive (e.g. diagnose every patient with a disease) will make it useless because of it lack of specificity, and vice versa. In statistics, as in life, tradeoffs are required.

7.1.4 Exercises

Test for TB	TB absent	TB present
Negative	1739	8
Positive	51	22

Data for TB testing using X-ray imaging

Table **??** shows the results of using X-ray imaging as a diagnostic test for tuberculosis in patients with known TB status. Use it to answer the questions below.

1. Calculate the marginal probabilities of the individual random variables, i.e. the probability of positive and negative X-ray test results, and of TB being present and absent.
2. Find the probability of positive result given that TB is absent (false positive rate) and the probability of a negative result given that TB is absent (specificity).
3. Find the probability of negative result given that TB is present (false negative rate) and the probability of a positive result given that TB is present (sensitivity).
4. Find the probability that a person who tests positive actually has TB (probability of TB present given a positive result).
5. Find the probability that a person who tests negative does not have TB (probability of no TB given a negative result).

6. Assuming the test result and the TB status are independent, calculate the expected probability of both TB being present and a positive X-ray test.}
7. Under the same assumption, calculate the expected probability of both TB being absent and a positive X-ray test.

7.1.5 rejecting the null hypothesis

Hypothesis testing is one of the most important applications of statistics. People often think of statistics as a collection of tests to be used for different hypotheses, which is too simplistic, but different tests do occupy a large fraction of statistics books. In this book we will only dip a toe into hypothesis testing, and will primarily approach it in a probabilistic (model-centered) way rather than from a statistical (data-centered) viewpoint. Probability allows us to calculate the sensitivity and specificity of a test for a given null hypothesis, provided the hypothesis is simple enough and the data are sampled correctly.

Example: testing whether a coin is fair. Suppose we want to know whether a coin is fair (has equal probabilities of heads and tails) based on a data set of several coin tosses. How much evidence do we need in order to reject the hypothesis of a fair coin with a small chance of making a type I error? What is the corresponding chance of making a type II error, not detecting an unfair coin?

Let us first consider a data set of two coin tosses. If one is heads and one is tails, it's obvious we have no evidence to reject the null hypothesis. But what if both times the coin landed heads? The probability of this happening for a fair coin is 1/4, which means that if you reject the null hypothesis based on the evidence, your probability of committing a type I error is 1/4. However, it is very difficult to answer the second question about making a type II error, because in order to do the calculation we need to know something about the probability of heads or tails. The hypothesis being false only means that the probability is not 1/2, but it could be anything between 0 and 1.

Let us see how this test fares for a larger sample size. Suppose we toss a coin n times, and if all n come up heads, then we reject the hypothesis that the coin is fair. A fair coin will come up all heads with probability $1/2^n$, so that is the rate of false positives for this test. For example, if a coin came up heads ten times in a row, there is only a 1/1024 probability that this is the result of a fair coin, so the probability of making a type I error is less than 0.1%. Is this careful enough? This question cannot be answered mathematically - it depends on your sense of acceptable risk of making a mistake. Notice that if you decide to use a very stringent criteria for rejecting a null hypothesis, you will necessarily end up not rejecting more false hypotheses. Such is the face of us mortals, dealing with imperfect information in an uncertain world.

This leads us to an important new idea: the probability that a given data set is produced from the model of the null hypothesis is called the *p-value* of a test. In the example of coin tosses we just studied, the p-value was $p = 1/2^n$. However, what if the data had 9 heads out of 10 tosses? The p-value then would be the probability of obtaining 9 or 10 heads out of 10. This is because to compute the probability of making a false positive error, we consider all cases that could have produced the result that is as different from expectation, or even further from expectation (in this case, 5 heads out of 10) than the data. [**?**].

ℹ Definition

For a data set D and a null hypothesis H_0, the *p-value* is defined as the probability of obtaining a result as far from expectation or farther than the data, given the null hypothesis.

The p-value is the most used, misused, and even abused quantity is statistics, so please think carefully about its definition. One reason this notion is frequently misused is because it is very tempting to conclude that the p-value is the probability of the null hypothesis being true, based on the data. That is not true! The definition has the opposite direction of conditionality - we assume that the null hypothesis is true, and based on that calculate the probability of obtaining the data. There is no way (according to classical frequentist statistics) of assigning a probability to the truth of a hypothesis, because it is not the result of an experiment. The simplest way to describe the p-value is that it is the likelihood of the hypothesis, based on the data set. This means that the smaller the p-value, the less likely the hypothesis, and one can be more certain about rejecting the hypothesis. Alternatively, the p-value represents the probability of making a type 1 error, or rejecting the correct null hypothesis for a particular data set. These two notions may seem to be in conflict, but they tell the same story: if the hypothesis is likely, the probability of making a type 1 error is high.

7.2 Chi-squared test

Now we are ready to address the question raised in the previous chapter of testing the independence hypothesis based on the table of observations and the calculated table of expected counts. In order to measure the difference between what is expected for a data table with two independent variables and the actual observations, we need to gather these differences into a single number. One can devise several ways of doing this, but the accepted measure is called the chi-squared statistic and it is defined as follows:

ℹ Definition

The chi-squared value for the independence test is calculated on the basis of a two-way table with m rows and n columns as the sum of the differences between the observed counts and the computed expected counts as follows:

$$\chi^2 = \sum_i \frac{(Observed(i) - Expected(i))^2}{Expected(i)}$$

The number of degrees of freedom of chi-squared is $df = (m-1)(n-1)$.

This number describes how far away the data is from what is expected for an independent data set. Therefore, the larger the chi squared statistic, the larger the differences between observed and expected frequency, and thus the null hypothesis of independence is less likely. However, simply obtaining the χ^2 is not enough to say whether the two variables are independent. We need to translate the chi-squared value into the language of probability, that is to ask, what is the probability of obtaining a data set with a particular χ^2 value, if those two variables were independent.

This question is answered using the *chi-squared probability distribution*, which describes the probability of the random variable χ^2. Like the normal distribution we saw in section **??** it is a continuous distribution, because χ^2 can take any (positive) real value. In another similarity, the χ^2 distribution has an even more complicated functional form than the normal distribution, so I do not present it here, because it is not enlightening. I will also not share the derivation of the mathematical form of the distribution, as it is far outside the goals of this text. In practice, nobody computes either the chi-squared statistic or its probability distribution function by hand, instead computers handle these chores. The chi-squared distribution has one key parameter, called the number of degrees of freedom, which was defined above. Depending on d.f. the distribution changes, specifically for more degrees of freedom the distribution moves to the right, that is, the chi-squared values tend to be larger.

The chi-squared distribution is used to determine the probability of obtaining a chi-squared statistic as at least as large as observed, based on the null hypothesis of independence. Figure **??** shows a plot of the chi-squared distribution, as well as the total probability to the right of an observed χ^2. This allows one to use it for the *chi-squared test* for independence between random variables, by comparing the p-value obtained from the distribution (by a computer) against a number called the *significance* level, which is decided by humans. The significance value α is a threshold that the test has to clear in order to reject the null hypothesis: if the p-value is less than α, the independence hypothesis is rejected, otherwise it stands, although one can never say that the independence hypothesis is accepted.

There is no mathematical or statistical method for determining the appropriate significance level, it is entirely up to the users to decide how much risk of rejecting a true null hypothesis they are willing to tolerate. If you choose 0.01, that means you want the likelihood of the hypothesis to be less than 1% percent in order to reject it. This is entirely arbitrary, and using a rigid significance level to decide whether a hypothesis is true can lead to major problems which we will discuss in the next chapter.

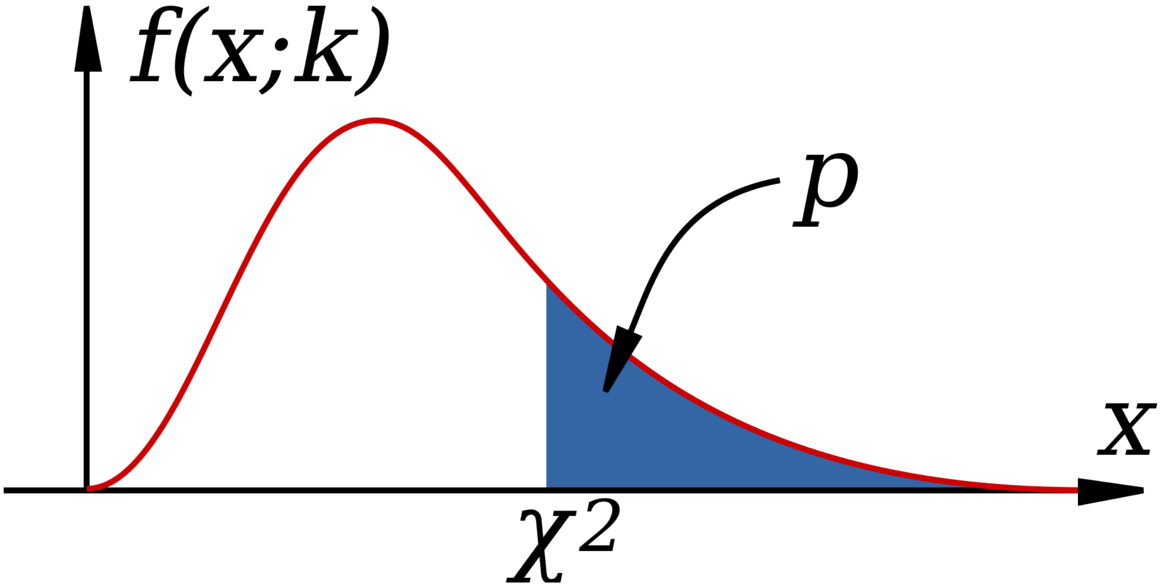


Figure 7.1: The chi-squared distribution is used to compute the p-value as the total probability of obtaining a χ^2 value at least as far from 0 as observed. (image by Inductiveload in public domain via Wikimedia Commons)

Like all mathematical models, the chi-squared distribution relies on a set of assumptions. If the assumptions are violated, then the probability distribution does not apply and the p-value does not reflect the actual likelihood of the hypothesis. Here are the assumptions:

- the data is from a simple random sample of the population
- the sample size is sufficiently large
- expected cell counts cannot be too small
- the observations are independent of each other

7.3 Hypothesis testing in R

R has many functions for different tests, including the chi-squared test. To use it, one first has to input a data set in the form of a two-way table, where each row represents the values of one random variable, and each column represents the values of the second random variable. The following script shows how to manually input a 2 by 2 contingency table into a matrix. In the matrix function, `ncol` stands for number of columns, and `nrow` for number of rows. Notice the order in which the numbers are put into the matrix: down the first column, then the second, etc. Type `help(matrix)` for more details. In order to access a specific element of the matrix, just like in vectors, R uses square brackets and two indices, first one for row, and second for column. Below are examples of accessing two elements of the matrix data defined above, and how to reference a particular element of the matrix.

```
data <- matrix(c(442,514,38,6),ncol=2,nrow=2)
print(data)
```

```
     [,1] [,2]
[1,]  442   38
[2,]  514    6
```

```
print(data[1,2])
```

```
[1] 38
```

```
print(data[2,1])
```

```
[1] 514
```

Based on a given data set, how likely is the hypothesis that the two random variables are independent? It is hard to do by hand (in the old days, you looked it up in a table of chi-squared values) but R will do it all for us: 1) calculate the expected counts, 2) compute the chi-squared value for the table, and 3) use the number of degrees of freedom and the chi-squared value to calculate the p-value of the independence hypothesis based on it. Use the `chisq.test()` function, and you will see output like this:

```
test.output <- chisq.test(data)
print(test.output)
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 25.555, df = 1, p-value = 4.3e-07
```

The results are the chi-squared values, the number of degrees of freedom (which depends on the number of rows and columns in the two-way table) and the p-value. The p-value is used to decide whether to reject the hypothesis, because it represents the likelihood of the hypothesis, given the data. In this case, the p-value is pretty small, so it seems relatively safe to reject the hypothesis of independence. To see the results of the hypothesis test, type `print(test.output)` and to access the p-value individually, use `test.output$p.value`.

Finally, we need to specify the significance level α for the hypothesis test. This refers to the probability of rejecting a true null hypothesis, by random chance. For instance, if you reject the hypothesis at $\alpha = 0.05$ significance, you're accepting a 5% chance that you falsely rejected a correct hypothesis. Note that it says nothing about failing to reject an incorrect hypothesis (also called the rate of false negatives.)

7.3.1 example with data

Let us return to the data presented in section **??**. We noted that the fraction of women in different age categories carrying fetuses with DS are different, but how certain are we that is not a fluke? To test the hypothesis of independence, we input the data into R and then run the chi-squared test:

```
data <- matrix(c(29806, 8135, 28, 64),ncol=2,nrow=2)
test.output <- chisq.test(data)
print(test.output)
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 122.86, df = 1, p-value < 2.2e-16
```

This tests of independence between the two variables of maternal age and DS status. The chi-squared parameter is about 122, reflecting the differences between expected and observed frequencies. This number us to calculate the p-value, which is very small (the number is actually caused by machine error). Therefore, the hypothesis can be rejected with a very small risk of making an error.

7.3.2 stop-and-frisk and race

The practice of New York Police Department dubbed "stop-and-frisk" gave police officers to power to stop, question, and search people on the street without a warrant. Since the practice commenced in the early 2000s, it has generated controversy for several reasons. First, the 4th amendment to the U.S. Constitution limits the power of the state to detain and search citizens, by mandating that officials first obtain a warrant based on "probable cause," while based on the Supreme Court interpretation, police are allowed to stop someone without a warrant provided "the officer has a reasonable suspicion supported by articulable facts" that the person may be engaged in criminal activity. Exactly what these conditions mean and whether officers in NYPD always had reasonable suspicions before stopping is a legal matter, rather than a statistical one, and you can read what federal judge Scheindlin ruled on this matter here [?].

The second issue raised by stop-and-frisk is whether it violates the principle of equal protection under the law enshrined in the 14th amendment of the Constitution. The idea that the law and its agents should treat people of different backgrounds the same, that people can be punished for their actions, but not for who they are, is deeply rooted in American law and culture. Critics of stop-and-frisk charge that officers disproportionately stop and search people of African-American and Hispanic background and therefore violate their constitutional rights to equal protection. As part of the trial, statistical evidence was introduced about the number of stops of New Yorkers of different racial backgrounds, how many of those stops resulted in the use of force, and how many uncovered evidence of criminal activity leading to an arrest. Let us analyze the data using our tools to address whether race and somebody being "stopped-and-frisked" are related.

The data in the summary of judge Scheindlin's decision is as follows: between 2004 to 2012, out of 4.4 million stops, 52% of the people stopped were black, 31% of the people stopped were Hispanic, and 10% of the people were white. The population of New York according to the 2010 census is approximately 23% black, 29% Hispanic, and 33% white. You may notice that the fractions are suggestive of a higher probability of stops of African-Americans, and lower probability of stops of white individuals, but we cannot use fractions to perform a chi-squared test, because actual counts are necessary to quantify the uncertainty in the testing.

Below I present data in the form of counts for only the calendar year 2011 [?], in the form of a contingency table with two variables: race/ethnicity and being stopped by police without a warrant. I have used the census population of New York (http://factfinder2.census.gov) and its breakdown by race (white only, black only, Hispanic, other). The data are presented in table ??, and then are input in R and run through a chi-squared independence test.

```r
data_mat <- matrix(c(61805, 2665172, 350743, 1527029, 223740,
2119718, 49436, 1201578),ncol=4,nrow=2)
rownames(data_mat) <- c('stopped', 'not stopped')
colnames(data_mat) <- c('White','Black', 'Hispanic', 'Other')
print(data_mat)
```

```
              White   Black Hispanic   Other
stopped       61805  350743   223740   49436
not stopped 2665172 1527029  2119718 1201578
```

```r
test.output <- chisq.test(data_mat)
print(test.output)
```

```
    Pearson's Chi-squared test

data:  data_mat
X-squared = 429039, df = 3, p-value < 2.2e-16
```

The results confirm what comparing the percentages suggested: the race of a person in NYC is not independent of whether or not they get stopped and frisked, with only a tiny probability that this disparity could have happened by chance. However, this is only the beginning of the analysis that experts performed for the court trial. Drawing conclusions about motives from the data is tricky, since two variables may be related without a causal connection. Defenders of the practice have argued that the racial disparities reflect differences in criminal activity. The data, however, show that only 6% of the stops result in arrests, and 6% more in court summons, so the vast majority of those stopped and frisked were not engaged in criminal activity.

8 Prior knowledge and Bayesian thinking

> Man can believe the impossible, but man can never believe the improbable.
> – Oscar Wilde, *The Decay of Lying*

In the last few chapters we defined the notion of independence and learned how to perform a statistical test to determine how likely a data set of two categorical variables is to have come from two independent random variables. This approach comes from the toolbox of classical "frequentist" statistics, which is taught in every statistics textbook in the world. It reduces statistical inference to a binary choice: reject or not reject the null hypothesis, based on the magic number called the p-value. However, this approach has deep problems, especially when applied mechanically and without understanding its limitations. Perhaps the most important limitation is that p-value based hypothesis testing does not incorporate any knowledge into its decision-making, aside from the given data. This may be reasonable at an early, exploratory stage of an experiment, but usually one has some prior knowledge about the likelihood of the hypothesis being tested. This knowledge cannot influence the data and the calculation of the p-value, of course, but it can have a dramatic effect on the interpretation, or inference one draws from the test. In this chapter you will learn to do the following:

- explain the effect or prior knowledge on interpretation of an experimental result
- calculate post-test probability based on prior probability, test result, and conditional probabilities of the test being in error
- use conditional statements to simulate random decisions with a given probability
- explain why conclusions based on binary p-value testing are frequently wrong

8.1 Prior knowledge

Suppose that a patient walks into a doctor's office, the doctor orders a pregnancy test, and the results indicate that the patient is pregnant. The doctor consults the published sensitivity and specificity values (that we defined in the last chapter) to discover, for instance, that 99% of positive pregnancy tests are correct. The doctor goes back to congratulate the patient with impending motherhood. Sound very reasonable, doesn't it? Would it still sound reasonable if the doctor knew that the patient does not have a uterus (whatever their gender may be)?

This is a slightly absurd example, of course, but it neatly illustrates the central point of this chapter: prior knowledge has an effect on the inference from a test, no matter how small or