Foundations of Biomedical Science: Quantitative Literacy: Theory and Problems

Julian Pakay

La Trobe eBureau

Melbourne

Contents

1

Acknowledgement of Country

La Trobe eBureau acknowledges that our publications are produced on the lands of many traditional custodians in Victoria. We recognise their ongoing connection to the land, their centuries of diverse knowledge systems, and value their unique contribution to the University and wider Australian society.

La Trobe University has campuses and undertakes teaching, learning and research activities in the traditional lands of the following people:

- Wurundjeri (Bundoora)
- Wurundjeri / Boonerwrung (City)
- Dja Dja Wurrung (Bendigo)
- Latji Latji / Barkindji (Mildura)
- Dhudhuroa / WayWurru (Wodonga)
- Yorta Yorta / Bangerang (Shepparton)
- Gadigal (Sydney)

2

Publisher Information

LA TROBE EBUREAU

La Trobe University, Melbourne, VIC 3086, Australia

https://library.latrobe.edu.au/ebureau/

Other information

Formatted by La Trobe eBureau
Copyediting by Adam Finlay
Cover design by Sebastian Kainey

Cover images adapted from: '1MBO' from NGL Viewer (AS Rose et al. (2018): web-based molecular graphics for large complexes. Bioinformatics doi:10.1093/bioinformatics/bty419), and RCSB PDB used under CC0; 'Relationship between mean and median under different skewness' by Diva Jain from Wikimedia Commons used under CC BY-SA 4.0; 'Glóbulos rojos' by Francisco Bengoa from Flickr used under CC BY-NC 2.0; 'Cross of a human artery' by Lord of Konrad from Wikimedia Commons used under CC0.

Icons used throughout this book: 'lightbulb' created by Maxim Kulikov from Noun Project, used under CC BY 3.0; 'Head' created by AFY Studio from Noun Project, used under CC BY 3.0; 'Colculator' created by rivercon from Noun Project, used under CC BY 3.0; 'Video' created by Adrien Coquet from Noun Project, used under CC BY 3.0; 'Owl' created by Teewara soontorn from Noun Project, used under CC BY 3.0; 'Search Pearson' created by Rank Sol from Noun Project, used under CC BY 3.0.

3

Accessibility Information

We believe that education must be available to everyone which means supporting the creation of free, open, and accessible educational resources. We are actively committed to increasing the accessibility and usability of the textbooks we produce.

Accessibility features of the web version of this resource

The web version of this resource has been designed with accessibility in mind by incorporating the following features:

- It has been optimized for people who use screen-reader technology.
  - all content can be navigated using a keyboard
  - links, headings, and tables are formatted to work with screen readers
  - images have alt tags
- Information is not conveyed by colour alone.

Other file formats available

In addition to the web version, this book is available in a number of file formats including PDF, EPUB (for eReaders), and various editable files. Choose from the selection of available file types from the 'Download this book' drop-down menu. This option appears below the book cover image on the eBook's landing page.

Third-Party Content

In some cases, our open text includes third-party content. In these cases, it is often not possible to ensure accessibility of this content.

Accessibility Improvements

While we strive to ensure that this resource is as accessible and usable as possible, we might not always get it right. We are always looking for ways to make our resources more accessible. If you have problems accessing this resource, please contact eBureau@latrobe.edu.au to let us know so we can fix the issue.

Category	Item	Status (Y / N)
Organising Content	Content is organised under headings and subheadings	Y
Organising Content	Headings and subheadings are used sequentially (e.g. Heading 1, Heading 2, etc.)	Y
Images	Images that convey information include Alternative Text (alt-text) descriptions of the image's content or function	Y
Images	Graphs, charts, and maps also include contextual or supporting details in the text surrounding the image	Y
Images	Images, diagrams, or charts do not rely only on colour to convey important information	Y
Images	Images that are purely decorative contain empty alternative text descriptions. (Descriptive text is unnecessary if the image doesn't convey contextual	Y

	content information)	
Tables	Tables include column headers, and row headers where appropriate	Y
Tables	Tables include a title or caption	Y
Tables	Tables do not have merged or split cells	Y
Tables	Tables have adequate cell padding	Y
Weblinks	The weblink is meaningful in context, and does not use generic text such as "click here" or "read more"	Y
Weblinks	Externals weblinks open in a new tab. Internal weblink do not open in a new tab.	Y
Weblinks	If a link will open or download a file (like a PDF or Excel file), a textual reference is included in the link information (e.g. '[PDF]').	N/A
Embedded Multimedia	A transcript has been made available for a multimedia resource that includes audio narration or instruction	Y
Embedded Multimedia	Captions of all speech content and relevant non-speech content are included in the multimedia resource that includes audio synchronized with a video presentation	Y
Embedded Multimedia	Audio descriptions of contextual visuals (graphs, charts, etc.) are included in the multimedia resource	N
Formulas	Formulas have been created using MathML	N, Formulas have been created using LaTeX.
Formulas	Formulas are images with alternative text descriptions, if MathML is not an option	N/A
Font Size	Font size is 12 point or higher for body text	Y
Font Size	Font size is 9 point for footnotes or endnotes	Y
Font Size	Font size can be zoomed to 200%	Y

About the Author

Julian Pakay is currently a senior lecturer in the Department of Biochemistry and Chemistry at La Trobe University, Melbourne, Australia.

Following his passion for biology, Julian completed his undergraduate studies and PhD in Biochemistry at the University of Western Australia where he investigated how some animals can control "metabolic time" and survive on stored fuels for extended periods. He continued in research at the Dunn Human Nutrition Unit in Cambridge, Geneva University, and Melbourne University working in diverse fields including bioenergetics, gene regulation and cellular signalling. He is now an education-focused academic, teaching biochemistry at all year levels from first year through to Masters. From his time in research Julian has seen how technological innovation has transformed biology into a more quantitative and predictive discipline. One of his major teaching goals is to help students attain proficiency in mathematics and quantitative literacy to help them navigate modern biology.

Acknowledgments

I

Chapter 1: Introduction to quantitative literacy

Quantitative literacy can be defined as the ability to interpret and communicate numbers and mathematical information throughout everyday life. Whether you are a scientist or not, you will be continuously confronted with claims based on quantitative data, sometimes regarding medical treatments and diagnostics but also in relation to the environment, politics, economics, and other aspects of life. Therefore, everybody needs to be able to interpret quantitative information to make informed decisions.

Quantitative literacy is especially important for students of biomedicine. Modern biomedicine is evidence based, which means fundamentally it is underpinned by quantitative data. This is becoming increasingly important as recent technological advances have led to biomedicine (and biology in general) becoming more 'data driven' and hence a more quantitative and predictive science. Think about the advances in DNA sequencing driving personalised medicine and then imagine the sorts of skills that will be in demand to navigate this kind of data in the future.

It is normal for many students to initially find aspects of mathematics and quantitative literacy difficult and therefore a source of anxiety. This resource, *Foundations of Biomedical Science*, and the problems within, are designed to help reduce this anxiety by targeting the skills you will need for later stages of your studies and beyond. These skills will help you interpret quantitative data and understand basic mathematical concepts and be able to apply them to authentic biomedical problems. However, your overarching goal here should be to habitually question any quantitative data you come across and to use the skills you will learn in this resource to make informed judgements about its veracity.

Why should you care about quantitative literacy?

Imagine the following scenario.

Before beginning a new job, you are sent for a mandatory health check. The health check is comprehensive and includes tests for some rare diseases. One of the diseases is found in 1 in 10,000 people but is incredibly deadly. Let's call it Disease X.

In your follow-up appointment the doctor tells you that you have tested positive for Disease X.

This is bad news, right?

Perhaps.

But first we need to know how accurate the test is.

The doctor tells you that the test is 99% accurate.

Okay, so now this is really bad news, right?

Well, maybe. But what does she mean by 'accurate'?

When we say a test is 99% accurate we mean this is the chance that if you have Disease X you will test positive, not that if you test positive there is a 99% chance that you have disease. You will learn later that using 'accuracy' to describe the effectiveness of a diagnostic test is a bad idea!

So, what are your chances of having Disease X?

Suppose 1 million people are tested for Disease X. Given that the prevalence is 1 in 10,000 people, then on average 100 out of the 1 million will have the disease. This means that out of the 1 million, 999,900 people do not have the disease. Since the test is only 99% accurate, 99 out of the 100 people with the disease will test positive, and 1 of the 100 with the disease will test negative but will still have the disease.

However, the accuracy of 99% also means that out of the 999,900 people that do not have the disease, 9,999 will test positive (1% of 999,900 = 0.01 x 999,900 = 9,999). Therefore, even though you tested positive you actually only have about a 1% chance (99 / (99 + 9,999) × 100) of actually having the disease.

Those odds suddenly look much better!

Having some quantitative skills will not just help with your grades – it will also provide you with some useful tools for questioning the truth and navigating life in general.

1

1.1 How to use this book

The lightbulb icon introduces a new idea or concept and provides some background information for the chapter. The boxed sections either provide a logical consequence of the concept or an interesting aside/example which will help explain the concepts to you.

The gear icon indicates sections that provide the rules or methods you will need to learn, as well as worked examples to help you put concepts to practical use and to solve the problems later in the chapter.
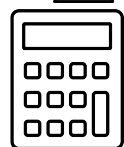
Each chapter contains some practice problems. There are worked solutions for all of these but resist the temptation to look at the answers first. Try to solve these practice problems on your own.

Where you see a play icon there is a link to a video demonstrating how to solve the problem. Again, always try to solve the problem on your own first!

The owl icon indicates a 'boffin' section. These are more advanced problems, examples and concepts that are designed to stretch your understanding. Some don't involve calculations but rather rely on thinking critically about the data. There are also worked solutions for these questions but have a go at solving them on your own first or try discussing them with friends or other students to see if you can come up with the answers between you.

The concepts taught here have real-life applications! Where you see this focus icon there is a link to a short interview with a professional which brings their career into focus. You will learn how they use maths and quantitative literacy in their everyday jobs, how they attained these skills and the strategies they employ to know if they are correct. The interviews solve the real problem stated in the text. Attempt to answer the problem on your own first and then watch the video to see the answer!

At the end of this chapter is a diagnostic test. Try this before you jump into the other chapters. It is designed to help you see where you may need to focus to improve your ability at tackling the various calculations you might be expected to know as a student of biomedical science or biology. The questions in the diagnostic are designed to be completed without a calculator and should take about 30 minutes. Attempt them seriously (under exam conditions!) and write out your full working. If you struggle with any of these problems do not worry as the content is covered in detail later in the book. Once you complete the diagnostic, check your answers. Full working is shown as well as directions to where the skill is covered in later chapters. Importantly, it is normal for many people to experience anxiety around maths or struggle with new concepts. Even though improving your quantitative literacy is empowering, perversely it can initially feel disempowering. My advice is to persevere and remain positive, but below are some useful strategies you can employ to help reduce maths anxiety.

2

1.2 Overcoming maths anxiety

Does the thought of having to do some maths make you feel nervous or anxious? If the answer is yes, you are not alone and most likely you are in the majority. In order to overcome this anxiety, it is important to understand that you were not born with it. There was a time when you didn't have it. Thinking about what contributes to anxiety and maybe what led to it in the first place can help us overcome it!

As students beginning our primary school education, we typically do not suffer from maths anxiety. In general, as very young students, we enthusiastically embrace learning new things, and early lessons in counting and basic arithmetic are not something we remember as traumatic.

However, the majority of adults report mild to moderate maths anxiety and a significant number of people (about 5% of adults) report severe maths anxiety.

Hart, S. A., & Ganley, C. M. (2019). The Nature of Math Anxiety in Adults: Prevalence and Correlates. Journal of Numerical Cognition, 5(2), 122-139. https://doi.org/10.5964/jnc.v5i2.195

It is unclear exactly what the contributing factors are for this increase in anxiety between early childhood and adulthood. However, it seems safe to posit that our anxiety tends to rise as we engage with more difficult content possibly coupled with poor teaching, become exposed to other people's negative attitudes to maths, or feel pressure to conform to social stereotypes. For example, there is a well-refuted but unfortunately pervasive myth in our society that females are not as good at maths as males. This myth has certainly contributed to the lower representation of females in certain STEM disciplines.

Wang, L. Mediation Relationships Among Gender, Spatial Ability, Math Anxiety, and Math

There is a strong negative correlation between maths anxiety and maths competency. This means that the more maths anxiety you have the more difficulty you will have with maths. So how can we overcome this? One way is to go back to the first principles of learning that you used when you were younger and free from maths anxiety.

1. Do what you understand first

Maths is incremental! This is why when we begin learning we start with very basic concepts and build from there. There is no shame in going back to the things you are more comfortable with first if the end goal is to build a base to tackle more challenging concepts later. This will help you build confidence and make it easier for you to visualise yourself succeeding. Try the easier problems first to help build your understanding of the concepts.

2. Understand the concepts

It is important to try to understand the 'why' of maths concepts rather than just memorising them. In a stressful situation (for instance an exam!) the first thing you will lose is your short-term memory. This will be a problem if you have only just memorised a set of rules. However, if you have a deeper conceptual understanding of the reason behind the rules you will fare much better. For example, a common problem is that many students are over-reliant on formulas. If you understand determination of molar concentration only as $C = n/V$, then it is possible you will misremember the equation and get your answers wrong. If you think about molar concentration in terms of the dimensions (C = concentration in moles per litre, n = number of moles and V = volume in L) then it is much more difficult to misremember or make a mistake when rearranging the equation. Understanding the concepts, in this case the units, will help you remember correctly.

A natural corollary to this is that you should be prepared to ask questions in order to understand. Remember always to ask for clear instruction, clear illustrations and examples!

3. Learn to self-check and troubleshoot errors

Learning to check your work is certainly something you would have been encouraged to do when you first began learning maths. There are three broad classes of errors you will make: conceptual errors, careless errors and computational errors. Conceptual errors can be avoided by understanding the concepts (as in the molar concentration example above). Careless errors result from things like copying the problem or a number down incorrectly, misinterpreting your bad handwriting, misreading the instructions, or pressing the wrong button on your calculator. Computational errors mean that a mistake has been made in the process of solving a problem stemming from an incorrect addition, subtraction, multiplication or division.

There are some strategies you can use to either avoid or troubleshoot careless and computational errors. One simple way is to take your time to read the problem carefully and or highlight important information. When solving the problem, carefully write down each step as this makes it harder to make a mistake and easier to check your working. This has the added benefit of that in an exam it will allow the person marking to follow your logic and give you partial credit if you have an incorrect answer but have clearly demonstrated an understanding of the underlying concept.

A lot of the operations you do are reversible so clearly articulating your steps also gives you an opportunity to work backwards to see if you get to the same starting point. Another very good way to troubleshoot errors is to mentally estimate the answer before you begin. This is known as sanity checking and is covered in a later chapter. Most of the time your estimate does not need to be that close for you to realise that you have made a mistake with your calculation. A dropped digit or inverted operation will often result in answers that are orders of magnitude incorrect!

Errors are going to occur, especially when you are initially learning, but in a professional context, they can have serious consequences. For example, a drug dosage calculation that is out by a factor of 10 could be fatal. Good professional practice involves taking steps to minimise error, a very effective way to do this is by writing out calculations and having another trained professional check them.

4. Identify the applications

It is difficult to find the motivation to tackle difficult maths or concepts which have previously vexed you. Your early explorations in maths were helped by obvious links to useful applications like being able to count or use money. It can help greatly to think about the applications for the maths you are learning. This book teaches you maths through a prism of biomedical science and ideally the relevant applications will be apparent to you. Look ahead to where you might need these skills in future study or employment and use this as motivation for persevering!

3

1.3 Diagnostic Test
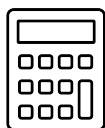
Have a go at this diagnostic test before diving into the rest of the book. Complete the questions without a calculator and give yourself about 30 minutes in total. Check your answers afterwards and reflect on what you might need to focus on. Don't worry if you are new to a concept! The underlying concepts and skills required to solve these problems are covered in later sections of this book.

Each answer is worth 1 mark, for a total of 20 marks.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=983#h5p-1

Side Bar – Bovine serum albumin (BSA) as the name implies is protein found in cow's blood. As it is abundant, easy and cheap to purify, small and stable it has found many uses in biochemical and clinical diagnostic laboratories. One such use is as a protein standard for determining unknown protein concentrations in biological extracts.
Side Bar – Square brackets, [ ], are often used as an abbreviation for the term concentration. e.g. a 0.1 molar hydrogen ion concentration can be referred to by $[H^+] = 0.1$ M.

Side Bar – Often concentrations are expressed as an amount per volume such as moles per litre or grams per millilitre. There are several valid ways to express concentrations, or in fact any rates or ratios which involve dividing one unit by another (e.g. kilometres per hour). It is worth becoming familiar with them. *Per* as in grams per millilitre is sometimes denoted with as a slash to indicate division (g/ml) or to indicate that ml is the denominator it can be expressed using a negative exponent (g.ml^{-1}).

4

1.4 Diagnostic test answers

1. $40 / (120 + 40) =$

   Complete the bracketed component first:

   $40/160 \times 6 =$

   Simplify by dividing the fraction by 10:

   $4/16 \times 6 =$

   $1/4 \times 6 = 1.5$

   Being able to do simple arithmetic is important for all calculations. A key skill is being able to quickly estimate the magnitude of an answer to avoid mistakes – see Chapter 3: Estimation (sanity checking). Being able to understand the order of operations is crucial for being able to solve for unknowns – see Chapter 10: Correlation, causation and confounding variables and Chapter 12: Growth and decay – Exponents and logarithms.

2. $(0.2/2) \times 500 =$

   Again, it helps to simplify, and recognising that 0.2 is 1/10th of 2 helps here:

   $(1/10) \times 500 = 50$

   Again, the key skill being taught here is to quickly estimate the magnitude of an answer – see Chapter 3: Estimation (sanity checking).

3. How many milligrams are in 1.25 g?

   The prefix milli denoted by 'm', in front of the base unit grams denoted by 'g', means one thousandth (1/1,000). That is, there are 1,000 mg in 1 g.

   Therefore, the answer is 1,000 mg × 1.25 = 1,250 mg

   If you are unfamiliar with the SI units and their prefixes, this is covered in Chapter 5: Scientific notation and SI units and Chapter 4: Biological scale. Being able to quickly interconvert between

units is essential to be able to work with concentrations and volumes of solutions – see Chapter 7: Solutions and concentrations and Chapter 8: Dilutions.

4. How many µl are in 0.68 ml?

The prefix micro denoted by 'µ', in front of the base unit litres denoted by 'l', means one millionth (10^{-6}). That is, there is 10^6 (1 million) µL in 1 L and 10^3 (1 thousand) µl in 1 ml.

Therefore, the answer is 1,000 µl × 0.68 = 680 µl

As in question 3 above, if you are unfamiliar with the SI units and their prefixes, this is covered in Chapter 5: Scientific notation and SI units. Being able to quick interconvert between units is essential to be able to work with concentrations and volumes of solutions – see Chapter 7: Solutions and concentrations and Chapter 8: Dilutions.

5. How many significant figures are in each of the following measured values?
   a. 50,000

   There is 1 significant figure as trailing zeros following a whole number are **not** considered significant unless the decimal is shown (e.g., 50,000.0) or it is referring to an exact number in which case there are infinite significant figures (e.g., 50,000 people) or the significant figure is denoted by an overbar (e.g., 50,00̄).

   a. 100.3

   There are 4 significant figures as all non-zero numbers are significant and any zeros between significant numbers are significant.

   Significant figures are important when thinking about measurements and how precise they are – see Chapter 2: Measurement uncertainty and significant figures.

6. Write 0.00361 in scientific notation.

3.61×10^{-3}

Quite often, numbers are too big or small to use conveniently. It is common practice to write these numbers as two numbers multiplied, almost always using exponents of base 10 – see Chapter 5: Scientific notation and SI units.

7. Write 1.35×10^{-2} in decimal notation.

0.0135

Being able to interconvert scientific notation and decimal notation is important, particularly when working with SI units and prefixes – see Chapter 5: Scientific notation and SI units.

8. How many moles are in 60 g of NaOH? (The molar mass of Na is 23 g/mol, of O is 16 g/mol and of H is 1 g/mol.)

First you need to calculate the molecular mass of 1 mole of NaOH:

23 + 16 + 1 = 40

Then determine the number of moles in 60 g:

60/40 = 1.5 moles

Understanding the concept of molecular weight (molecular mass) and the mole is essential to be able to understand and work with concentrations, which are often expressed as moles per litre – see Chapter 7: Solutions and concentrations.

9. Sunwise Lime Cordial is sold as a 10x concentrate. How much concentrate would you need to make 500 ml of 1x cordial?

We are diluting the cordial by a factor of 10 – that is, going from 10x to 1x – therefore we only need a 1/10th of the final volume of the undiluted (10x) cordial.

(1x/10x) × 500 ml =

(1/10) × 500 ml = 50 ml

This is essentially the same problem as in question 2 but this time just put into words. Being able to work with ratios and perform dilutions is covered in Chapter 8: Dilutions.

10. What is the final concentration of a solution where 40 ml of a 6 M solution is added to 120 ml of water?

This is the same problem as question 1 but this time put into words to make it more confusing!

We are adding 40 ml of a 6 M solution into a total volume of 40 ml + 120 ml, so the concentration should decrease by the amount of dilution.

= (40 ml / (120 ml + 40 ml)) × 6 M

= (40/160) × 6 M

= 4/16 × 6 M

= 1/4 × 6 M = 1.5 M

For help with problems similar to this, see Chapter 8: Dilutions.

11. If you had 1 l of a 2 M stock solution of NaCl but need 500 ml of a 0.2 M solution of NaCl, how much of your stock solution would you need to use to prepare your final solution?

You might recognise this problem as being similar to questions 2 and 9 but this time it has been made more difficult by including 'scientific' terms and by including some irrelevant information.

A good way to solve this is by using the formula $C_1V_1 = C_2V_2$ where C_1 refers to the initial

concentration and C_2 the final concentration while V_1 refers to the initial volume and V_2 to the final volume. The initial volume V_1 is what we need to determine.

$V_1 = ?$ ml

$C_1 = 2$ M

$V_2 = 500$ ml

$C_2 = 0.2$ M

So, you can now see that volume of the stock solution is not relevant to the question. (Unless there was not enough of it!)

$C_1V_1 = C_2V_2$

So, 2 M × V_1 = 500 ml × 0.2 M

It is important to have the same units for volume and concentration on each side of the equation here.

V_1 = (0.2 M / 2 M) × 500 ml = 50 ml

For help with problems like this, see Chapter 8: Dilutions.

12. Which of these is equivalent to $\log_2 3x^8$?

b. $8 \log_2 x$ as per the general rule $\log_c = n \log_c x$

This rearrangement is important for being able to solve for unknown exponents in growth and decay formulas – see Chapter 12: Growth and decay – Exponents and logarithms.

13. A DNA stock solution with a concentration of 100 µg/ml was diluted to make 3 other sequential solutions. That is, for each dilution, 1 ml of the previous solution was added to 4 ml of water for a final volume of 5 ml. What is the concentration of the final solution?

Each of these dilutions is a 1 in 5 dilution; that is, 1 ml is added into a total volume of (1 ml + 4 ml) 5 ml.

When the dilutions are carried out in series like this the dilution factors are multiplied:

1/5 × 1/5 × 1/5 × 100 µg/ml = 0.8 µg/ml

In this way, using serial dilutions, a large but accurate dilution can be made that does not require a massive volume of solvent – see Chapter 8: Dilutions.

14. For the dataset 10, 20, 25, 20, 20, 10, 25, 30, 20, calculate the:
    a. Mean: The mean is equal to the sum of the data divided by the number of data entries.Mean = sum of data/n = 180/9 = 20

b. Mode: Mode is the most common value in data set = 20
c. Range: Range is the absolute spread the data can be expressed as: 10–30 (or 20)

For working with descriptive statistics see Chapter 9: Medical diagnostics – Measurement, uncertainty and distributions.

15. There are 5 mice in a cage; 2 of them are known to have the *QLIT* mutation. What is the percentage chance that a researcher who picks 2 mice at random will choose both mice with the *QLIT* mutation?

There is a 2/5 chance of initially picking a *QLIT* mouse and then because of non-replacing selection there is a 1/4 chance of the second mouse being *QLIT*.

Therefore, chance = 1/5 × 1/4 = 2/20 = 1/10

Understanding probability is a vital skill and is especially important in being able to understand the sensitivity and specificity of medical diagnostic tests – see Chapter 10: Medical diagnostics – Sensitivity and specificity.

16. Which of the following correlation coefficients (r) represent the strongest relationship between two variables?

c. $r = –0.6$

The strength of the correlation between two variables can be estimated by the correlation coefficient, r. The closer r is to either $–1$ or 1 the stronger the correlation. That is, the absolute value of the correlation coefficient gives us the strength of the relationship.

For more information about bivariate analysis (the relationship between two variables) and correlation, see Chapter 11: Correlation, causation and confounding variables.

17. The rate of growth for a species of mesophilic bacteria was determined across a range of temperatures. The resulting graph is shown below.

Mesophilic bacterial growth
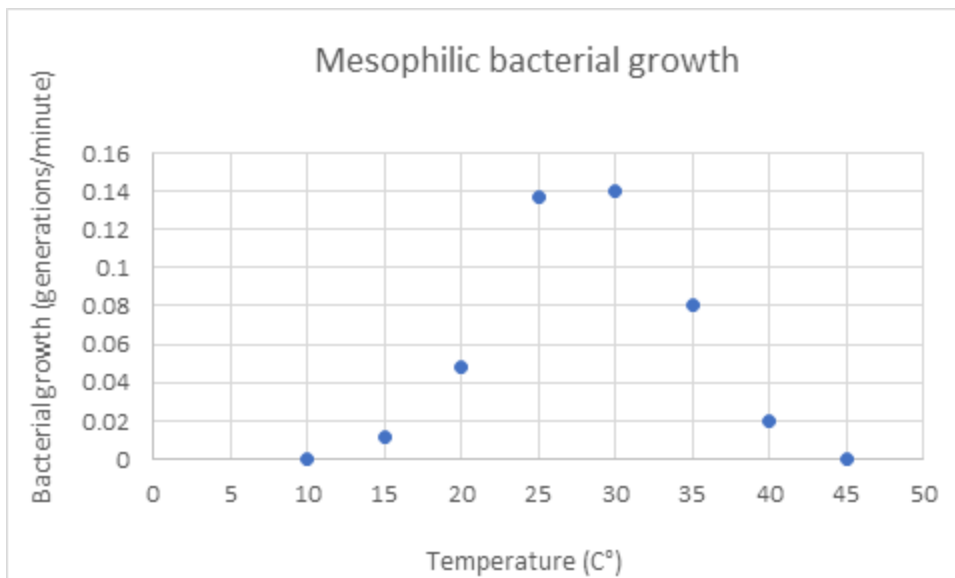
a. From the graph, what is the temperature range where bacterial growth is possible? 15 to 40°C
   Bacterial growth does not occur at 10°C or 45°C and you cannot extrapolate from 10–15°C and 40–45°C.
b. From the graph, what is the temperature range where would you expect maximal bacterial growth?
   Between 25 and 30°C.

18. A range of BSA concentrations were analysed using a colorimetric assay. The absorbance values at 750 nm were plotted against BSA concentrations, as shown in the graph below.

BSA standard curve
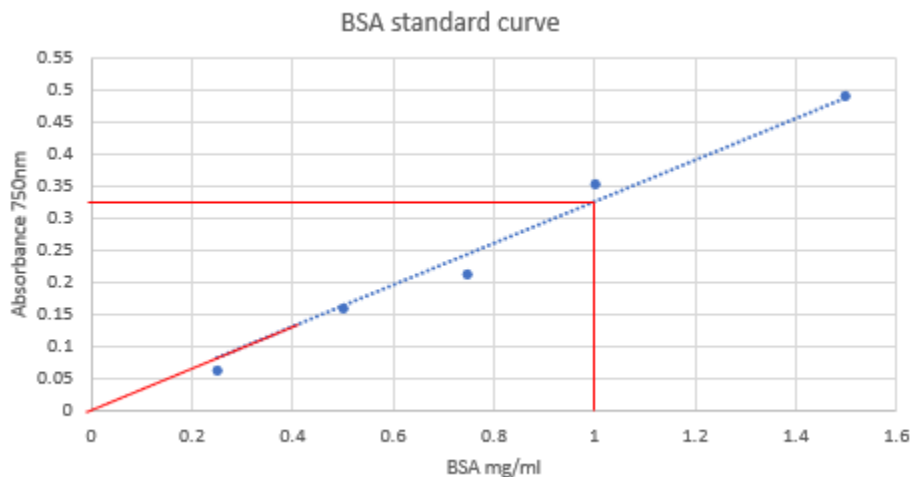
a. What would be the likely concentration for an unknown BSA solution which had an absorbance of 0.325 using the assay above?

The concentration corresponding to the absorbance value can be determined from the line of best fit. Draw a horizontal line from the absorbance value of 0.325 to the line of best fit. Where it intercepts, draw a vertical line down to the x-axis and read the concentration. In

this case the concentration is 1 mg/ml.

  a. What absorbance would you expect for a solution containing 0.1 mg/ml BSA using the assay above?

  Extrapolate the line of best fit down, to the point 0,0. You can see that the answer should be approximately 0.025 (note that absorbance does not have units).Just as a note, 0,0 is a valid point and can be included to calculate the line of best fit, but it is not strictly valid to force the line through 0,0. Occasionally you will see this and in general it is done in error or to ensure that nonsensical negative values are avoided.

Absorbance and spectrophotometry are covered in Chapter 7: Solutions and concentrations. More detail on linear regression is provided in Chapter 11: Correlation, causation and confounding variables.

19. Human blood plasma normally has an H^+ concentration of $10^{-7.4}$ M. What is the pH of human blood plasma? (Note that $pH = -\log_{10}[H^+]$ where $[H^+]$ is the H^+ concentration in M.)

This is solved by substituting the H^+ concentration into the equation:

$$pH = -\log_{10} 10^{-7.4}$$

The exponent can be brought to the front of the equation, changing the sign in the process:

$$pH = 7.4 \log_{10}$$

Since $\log_{10}$ of $10 = 1$

$$pH = 7.4$$

Further information about working with logarithms and exponents is provided in Chapter 13: Growth and decay – Exponents and logarithms.

20. What is in the equation below?

$$x^5 . x^{-2} = x^n$$

This is solved using the general rule

Therefore, $n = x^3$

Again, further information on the fundamental rules for working with logarithms and exponents is found in Chapter 13: Growth and decay – Exponents and logarithms.

II

Chapter 2: Measurement uncertainty and

significant figures

Imagine that you make a series of simple measurements on a group of people (e.g., patients in your care or subjects in your study). One of the measurements is height, which you record using a stadiometer (an instrument for measuring people's height) with 1 mm graduations. For a group of 3 people you measure their heights to be 167.8 cm, 178.2 cm and 175.8 cm and determine the mean. Your calculator will give you a result like this:

173.9333333333333

It is tempting to trust the number on the calculator, but is it right? How would you report this number?

Any measurement has a degree of uncertainty. This is because any tool or instrument we use for measuring has a limited capacity for precision; even our senses themselves have limitations. One way to describe the certainty we have about a measurement is by the number of digits we use when reporting the measurement. We call this the number of significant figures.

In the above example, the mean height should be reported as 173.9 cm. It would be misleading to use more decimal places because the stadiometer is only precise for 1 mm increments.

Determining the number of significant figures to use to express a final answer in a calculation can become complicated when you need to incorporate different measurements with different degrees of precision. Often you need to think about where the data comes from or how the measurement was made. The key rule to remember when reporting a final number, which was dependent on several measurements, is this:

Your answer should be reported in such a way that it reflects the reliability of the least precise measurement. That is, the number of significant figures in your answer will be limited to the measurement with the lowest number of significant figures.

Measurement precision

How **precise** is this 1 ml syringe?

Figure 2.1: A 1 ml Luer lock syringe showing 0.1 ml major increment markings each divided into 0.01 ml gradations.
Photo by Julian Pakay.
The major markings are at 0.1 ml increments and there are 9 markings between these creating 10 divisions. This means that, since 0.1 divided by 10 is 0.01, this syringe is accurate to measurements as small as 0.01 ml (1/100th of a ml)*.
Using this syringe, you could report that 0.45 ml was injected. However, it would be incorrect to report that 0.452 ml was injected as the syringe is not precise to 1/1,000th of a ml.
*In actual practice, for this syringe even an accuracy of 0.01 ml may be an overestimation. Most 1 ml syringes are rated at about 5% accuracy. Accuracy is often indicated on apparatus as /- followed by the percentage value of accuracy. For a 1 ml syringe rated with /- 5% accuracy, this means that we know, at best that measuring a 1 ml volume in that syringe will deliver somewhere between 0.95 and 1.05 ml.

2.1 Rules for significant figures

There are three rules for deciding how many significant figures there are in a number.

1. Non-zero digits are always significant

This rule should be self-explanatory. If you weigh out a chemical on an analytical balance and the mass is 1.38 g, then all these numbers are significant as a measurement has been made to their precise value.

2. Any zeros between two significant digits are significant

This time you weigh out a chemical on a balance and the mass is 203 mg. From rule 1, the numbers 2 (hundreds) and 3 (ones) are significant. However, the measurement also applies to the 0 (tens) as the value of the tens has been used to determine the value of the ones. Thus, the mass, 203 mg has 3 significant figures.

3. A final zero or trailing zeros in the decimal portion ONLY are significant

This is the more difficult rule. Imagine that you express the mass of the chemical from rule 1 in kg. The value would be 0.00138 kg.

The first zero (0.) is there to indicate the position of the decimal point – it could be left off if you chose to do so.

The next two zeros (.00) are only placeholders to put the decimal point in the correct position. If the number is written in scientific notation (1.38×10^{-3}) they disappear.

If the mass was measured at 1.00 g, then both these zeros are significant as a decision has been made on their value (i.e. that was the precise reading on the balance).

If there are trailing zeros in a whole number (e.g. 500) these zeros are NOT significant.

For example, 55,000 has only 2 significant figures.

But it is possible that 55,000 can be expressed as having more than 2 significant figures. If this is the case, you will be told (hopefully). There are a few ways to indicate if a trailing zero is significant:

- 55,0$\overline{0}$0 – an overbar (rare). This number has 4 significant figures.
- 55,000. – a decimal point (rare). This number has 5 significant figures.
- 5.5000×10^4 – scientific notation. This number has 5 significant figures.

Important

Exact numbers are numbers that we use for counting (e.g. 4 people, 6 pencils) and are considered to have an infinite number of significant figures, so you do not need to worry about these in your calculations (see the next section). For example, if you count the number of people in a room then the number has an infinite number of significant figures. If there are precisely 4 people, then there are 4.00000000 (etc.) people.

Defined numbers are also exact numbers. For example, there are 1,000 g in a kg. Or there are 1 carbon and 2 oxygen atoms in a molecule of carbon dioxide.

6

2.2 Calculations with significant figures

Remember the key rule! With calculations involving multiple measurements, your answer should be reported in such a way that it reflects the reliability of the least precise measurement. This means that the number of significant figures in your answer will be limited to the measurement with the lowest number of significant figures.

Multiplication / Division

If you multiply two numbers that have different numbers of significant figures, then the answer should have the same number of significant figures as the 'weaker' number.

For example, if you multiply 15.20 (4 significant figures) with 1.25 (3 significant figures) the answer would be limited to 19.0 (3 significant figures).

When you multiply two numbers, you more or less multiply the uncertainties. Therefore, it is the percentage by which you are uncertain that is important – the uncertainty in the number divided by the number itself. This is given roughly by the number of digits, regardless of their placement in terms of powers of ten. Hence the number of digits is what is important.

Addition / Subtraction

When adding and subtracting numbers, the rules of significant figures require that the number of places after the decimal point in the answer is less than or equal to the number of decimal places in every term in the sum. That is, limit the reported answer to the right-most column that all numbers have significant figures in common.

For example, if you were to add 1,992.123 and 34.12, note that the first number stops its significant figures in the thousandth's column, while the second number stops its significant figures in the hundredth's column. We therefore limit our answer to the hundredth's column.

1,992.123

+ 34.12

_____

2,026.243

2,026.24

This can be tricky to understand as it is not unusual for a sum to have more significant figures than the measurements added. <u>What is important here is to limit the answer to the right most column where all of numbers have a significant figure in common.</u>

If some of the numbers have no digits after the decimal point, use the same basic rule.

For example:

1,200

+ 85.88

_____

1,285.88

1,300

In this case, the answer has to be limited to the hundreds column as 1,200 only has two significant figures. Therefore, the answer is rounded to 1,300.

There is a difference in how addition/subtraction is handled compared to multiplication/division because when you add two or more numbers you add their uncertainties. If one of the numbers is smaller than the uncertainty of the other, it does not make much difference to the value (and hence, uncertainty) of the final result. Therefore, it is the location of the digits, not the number of digits that is important.

7

2.3 Practice problems
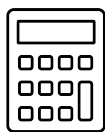
You may need to revise how to **round numbers** to complete some of these problems.

Why do we round numbers?

In general, there are two reasons:

1. To make a number simpler but keep its value close to what it was.
2. To round to the correct number of significant figures.

Rules for rounding

If the number to the right of the rounding digit is less than 5, then keep the rounding digit and change the rest of the digits right of the rounding digit to 0.

If the number to the right of the rounding digit is greater than or equal to 5, then add one to the rounding digit and change the rest of the digits right of the rounding digit to 0.

When rounding significant figures, the standard rules of rounding numbers apply, except those non-significant digits to the left of the decimal are replaced with zeros.

Example: 356 rounded to 2 significant digits is 360.

Note

These are 'rules of thumb'. There are a number of methods for rounding. For example, some statisticians will round to the nearest even number if the final digit is 5 to avoid bias. For example, 12.5 is rounded to 12 and so is 11.5. Ultimately, in practice you need to think about why you are rounding in the first place and apply a consistent method.

1. How many significant figures are in each of these measurements?
    a. 95.0°C
    b. 120°C
    c. 501 g
    d. 15 patients
    e. 0.450 ml
    f. 250 mM
    g. 2.38×10^{-3} g
    h. 0.00238 g
    i. 2,050 mmol
    j. 2.050 mmol
2. Rewrite the following answers so they are rounded to the correct precision (correct number of significant figures).

    a. 1.22 M × 1.3 l = 1.586 mol
    b. 500.00 g ÷ 125 ml = 4 g ml^{-1}
    c. 516.15 g − 0.005 g = 516.145 g
    d. 1.023 moles / 2.1 l = 0.4871 M
    e. 9.000 s − 0.5 s = 8.50 s
    f. The average height of 3 students with heights 170.3 cm, 167.23 cm and 171 cm

Solution to Practice Problem 2.2a.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=992

Solution to Practice Problem 2.2c

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=992

Significant figures and logarithms

Logarithmic values

For more information on logarithms see [Chapter 12](#).

when expressed as a decimal have two parts. The integer part is called the **characteristic** while the part to the right of the decimal is called the **mantissa**. So, for example $\log_{10} 15.0 = 1.176$. In this case 1 is the characteristic and 0.176 is the mantissa.

In terms of significant figures in logarithmic values only the numbers to the right of the decimal place (mantissa) count as significant. That is, the number of significant figures in the mantissa of a value expressed in scientific notation equals the number of significant figures to the right of the decimal in the logged value.

This is important for logarithmic quantities such as pH.

$$pH = -\log[H^+]$$

For example, pH = 8.85 has only 2 significant figures. It corresponds to a $[H^+]$ of 1.4×10^{-9} M (2 significant figures).

A pH of 7.6 corresponds to a concentration known to 1 significant figure.

Therefore, a pH given in a whole number such as pH = 4 corresponds to a concentration expressed only as a power of 10, in this case 10^{-4} M.

Significant figures in pH readings

In the case below, the pH reading of 7.29 involves only 2 significant figures. The leading digit in front of the decimal is not significant, only the numbers after the decimal are significant figures.

Figure 2.2: The digital display of a pH meter. Since pH is calculated using a logarithmic function, the value displayed has only 2 significant figures. Photo by Julian Pakay.

This is because the 7 indicates the magnitude of the number while the .29 indicates the number itself. This pH corresponds to a [H+] of 5.1×10^{-8} M, as $pH = -\log[H+]$

To see why this is so, take the log of 5.1×10^{-8}

That is, $\log (5.1 \times 10^{-8})$

The log of a product is equal to the sum of the logs of each multiplier.

Therefore, $\log (5.1 \times 10^{-8}) = \log (5.1) \log (10^{-8})$

$\log (5.1) = 0.71$ (2 significant figures, reflecting the uncertainty in the last digit of 5.1)

$\log (10^{-8}) = -8.000000 \ldots$ an infinite number of significant figures, as 10^{-8} is an exact number

Therefore, 0.71 (–8.000000 …) = –7.29

Remember, pH = –log[H+] so pH = 7.29

Here you are following the rules for addition of significant figures. That is, the number of places after the decimal point in the answer is less than or equal to the number of decimal places in every term in the sum.

8

2.4 Boffin questions

pH readings

1. What is the pH of a solution containing a $[H^+]$ of 2.56×10^{-5} M?

III

Chapter 3: Estimation (sanity checking)

There is a problem in the scenario below. What is the problem?

A patient is prescribed 3,000 micrograms (3,000 μg) of morphine over 24 hours. A nurse quickly calculates that the patient thus requires 12.5 μg per hour.

The calculation is clearly incorrect. To see this, round the 24 to 20, just to get a very approximate result using a calculation that is easy to do mentally: 3,000 ÷ 20 is equivalent to 300 ÷ 2, which is 150. So, we would expect an answer in the region of 150 μg. This suggests that the result 12.5 μg is wrong, and the calculation should be done again. The correct result is 125 μg. This is a potentially serious error in dosage!

This example highlights a very common problem. That is, when performing even very simple calculations, it is very easy to make an arithmetic error. The source of the error might be simply transcribing the number incorrectly, inverting a function (dividing instead of multiplying) or pressing the wrong button on your calculator. One of the best ways to check a calculation is to first estimate your answer.

9

3.1 The importance of estimations

Estimation is an important skill and used quite often in everyday life. In fact, a lot of the time estimation

is more important than performing precise calculations. In some situations, you can only make an estimate and often this requires a bit of experience to do well. For example, a paramedic may need to estimate the weight of a patient very quickly by eye (imagine a non-responsive accident victim at a roadside) to determine the dose of drug to administer to them.

However, when calculators or computers are being used, estimation is essential to be able to judge if the output is reasonable. You can estimate by rounding numbers to make them simpler, allowing you to quickly perform the calculation in your head. You can then get a sense of whether your precisely calculated answer is reasonable or not. Get into the habit of doing this – you will find it easier with practice.

Now try an estimation 'sanity check' on the next scenario.

You have 2 l of a 2 M stock solution of NaCl but need 500 ml of a 0.2 M solution of NaCl. How much of your stock solution would you need to use to prepare your final solution?

In the above problem, the answer must be less than 500 ml. Since you are making a dilution the volume of stock solution required must be less than the volume of the final solution! It is very easy to invert an operation and divide instead of multiply or vice versa resulting in a nonsensical answer.

Order of magnitude

If one amount is an order of magnitude larger than another, it is approximately 10 times larger. If it is two orders of magnitude larger, it is approximately 100 times larger. For example, an average mouse is around 20 g or 0.02 kg. An average human weighs around 70 kg. You could say that humans are approximately three orders of magnitude larger than mice.

10

3.2 Practice problems
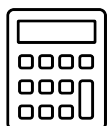
1. Round the following numbers:
   a. 37.7°C (to the nearest degree)
   b. 121 mM (to the nearest 10 mM)
   c. 505.3 g (to the nearest 100 g)
   d. 505.3 g (to the nearest 10 g)
   e. 14.5 g l^{-1} (to the nearest g l^{-1})
   f. 0.0245 g (to the nearest mg)
   g. 1.456 µg (to the nearest 100 ng)
2. Practice estimating the answers to the following calculations in your head. Then compare your answer to a precise calculation.

   a. $920 \times 27 =$
   b. $8{,}453 \div 53 =$
   c. $79 \times 91 =$

d. $1{,}205 \times 0.76 =$
e. $4{,}215 - 2{,}498 =$

Solution to Practice Problem 3.2a.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1002

Back of the envelope calculations

Sometimes you might hear these kind of calculations as 'back of the envelope' or 'back of the napkin' calculations. This is meant to emphasise that they are based on rough estimates. However, these kinds of calculations can be useful, particularly when assessing if an experimental approach may be feasible or as a sanity check when assessing data. Even though a rough estimate may be crude, it is often enough to know the relevant order of magnitude of a value.

To do these sorts of calculations you need to be confident with units, SI prefixes and working with indices. These are all incredibly useful skills.

For the next 'boffin' question, have a go at using the estimates to work out how many proteins are in a HeLa cell. HeLa cells are a human-derived cell line commonly used for preliminary research.

11

3.3 Boffin questions

 Estimating large numbers

1. How many proteins are in a HeLa cell?
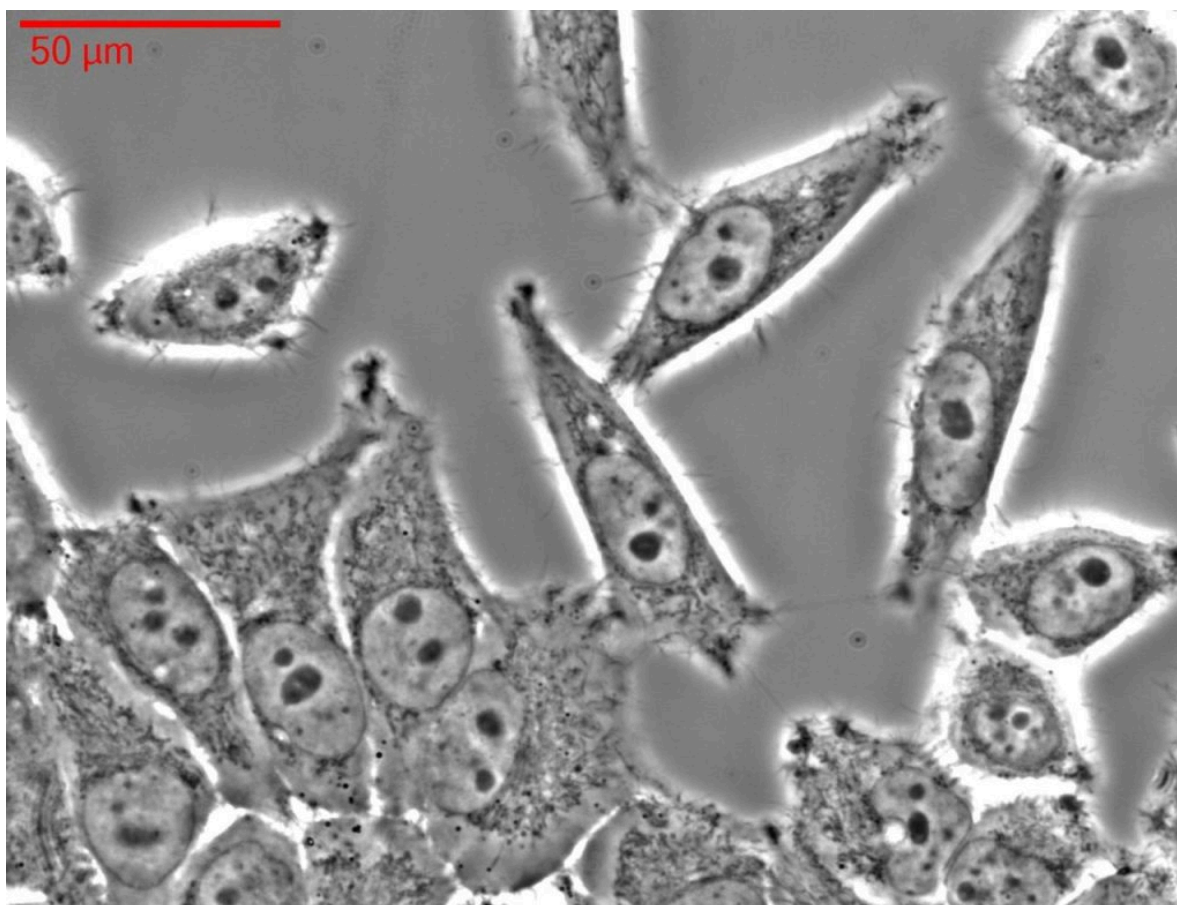
Figure 3.1: Phase contrast image of Hela cells in culture by Paul Anastasiadis, Eike Weiß from Fraunhofer Institute for Biomedical Technology, St. Ingbert, German ('HELA Cels' by Fraunhofer Institute for Biomedical Technology, St. Ingbert, Paul Anastasiadis, Eike Weiß from Wikimedia Commons used under CC BY-SA 3.0)

Use the following data to estimate the number of protein molecules in a HeLa cell:

- Protein mass per volume ≈ 0.3 g. ml^{-1}
- Mass of a typical amino acid ≈ 100 Da
- Average length of a protein ≈ 400 amino acids
- Volume of a HeLa cell ≈ 2,000 μm^3
- Avogadro's number = 6 × 10^{23}

Do not be afraid of the large numbers here! Work through this methodically and make sure you write down the units at each step.

Side Bar – The symbol, Da, refers to daltons (named after the English chemist and physicist, John Dalton) which is also known as the atomic mass unit. It is defined as 1/12th of the mass of a carbon-12 atom which is approximately equal to 1.660 x 10^{-27} kg. The mole is a unit of substance that was originally defined so that the mass of one mole of a substance, measured in grams, would be numerically equal to the average mass of one of its constituent particles, measured in daltons. So, carbon-12 (6

neutrons and 6 protons) is 12 Da and therefore the mass of one mole of carbon-12 is approximately 12 g. The molar weight of large molecules such as proteins or protein complexes can be expressed in kilodaltons (kDa) or megadaltons (MDa) as appropriate.

12

3.4 A focus on maths in clinical practice

Intravenous (IV) saline is used to replenish fluids or to deliver medications to patients. Normally the saline solution used is 0.9% (w/v) NaCl.

1. Why is this concentration of saline used?
2. What is the molar concentration of 0.9% (w/v) NaCl?
   (The molecular weight of NaCl is 58.44 g mol^{-1}.)
3. What is the osmolarity of 0.9% (w/v) NaCl?
   (Osmolarity refers to the molar concentration of particles of solute per litre of solution.)
4. If you needed to infuse 1 l of 0.9% (w/v) NaCl over 8 hours, what is the required flow rate in ml h^{-1}?

Maths in Clinical Practice

Yangama Jokwiro is a registered general nurse with extensive experience. He has worked in a variety of clinical settings but now focuses on teaching undergraduate nursing students. Watch an interview with Yangama to find out the answers to the questions above and learn how critical quantitative literacy is for clinicians to be able to effectively treat, monitor and care for their patients. Also learn how Yangama instils this literacy in his students!

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1006

Side Bar – Concentrations can be expressed as percentage weight per volume (e.g. 0.9% (w/v) NaCl). What this really means is the grams of solute in 100 ml of solution. That is, the percentage concentration tells us the parts of solute by mass per 100 parts by volume of solution. (see Chapter 7 Solutions and concentrations)

IV

Chapter 4: Biological scale

To understand biology, it helps to have some idea of the relative size of the components of biological systems. To get some idea have a look at the following graph. But just remember that these are typical sizes for these structures, and some will actually exhibit a range of sizes.

13

4.1 Linear and logarithmic scales

You will note that the scale on the graph is logarithmic. That is, each increment represents a 10-fold change in size (as we saw in Chapter 3, a 10-fold change in size is known as an order of magnitude). This kind of scale is useful to display widely different values – in this case the graph covers 11 orders of magnitude (or from 1 to 100 billion). This would be roughly the equivalent to the difference in the diameter of a grain of sand (0.05 mm) to the diameter of the Earth (12.7 km or 12,700,000,000 mm). On a linear scale the very small values would be bunched up together and be hard to distinguish. You will also note that rather than using the one unit, say metres (m), prefixes have been used to avoid either very large or very small numbers.

- The prefix m (milli) denotes 1/1,000th (10^{-3}) of the base unit – for example, there are 1,000 mm in a metre
- μ (micro) denotes 10^{-6} × the base unit
- n (nano) denotes 10^{-9} × the base unit

Some prefixes you will be familiar with while some you may need to memorise. It is important to be able to **quickly and confidently convert between these units**.

14

4.2 Practice problems
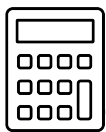
These practice problems are designed to get you thinking about biological scale. To answer some of these you may need to use the hints to look through a biology textbook. To do the calculations you may need to look at the section on SI units in Chapter 5.

1. Do you think that the protein myoglobin is bigger or smaller than the messenger RNA molecule which codes for it? Explain why or why not.
   Hint: Which is bigger, a nucleotide or an amino acid? How many nucleotides are in a codon? What is a codon?
2. How many orders of magnitude is the atomic radius bigger than the atomic nucleus?
3. The average bacillus ranges between 1 and 10 microns (μm) in length, and the mitochondria of both plant and animal cells measure in the same range. Do you think this is just a coincidence? Explain your reasoning.
   Hint: What is the proposed evolutionary origin of mitochondria? What is endosymbiosis?
4. If the distance between bases is approximately 0.3 nm, the number of base pairs (bp) in the

human genome is approximately 3 Gbp and most of our cells are diploid (have 2 copies of the genome), then how long is the DNA in a human cell?

Hint: The prefix n means nano and modifies the base unit by 10^{-9}. The prefix G means giga and modifies the base unit by 10^9.

5. How does your answer to question 4 compare to the size of a cell's nucleus? What are the implications of this? What is chromatin? What are chromosomes?

Solution to Practice Problem 4.4.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1011

15

4.3 Getting bigger or smaller
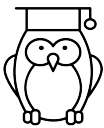
"… consider a giant man sixty feet high – about the height of Giant Pope and Giant Pagan in the illustrated *Pilgrim's Progress* of my childhood. These monsters were not only ten times as high as Christian [a normal human], but ten times as wide and ten times as thick, so that their total weight was a thousand times his, or about eighty to ninety tons. Unfortunately, the cross sections of their bones were only a hundred times those of Christian, so that every square inch of giant bone had to support ten times the weight borne by a square inch of human bone. As the human thighbone breaks under about ten times the human weight, Pope and Pagan would have broken their thighs every time they took a step."

The above is a quote by the geneticist J. B. S. Haldane in his 1928 essay *On being the right size*, discussing the giants 'Pope' and 'Pagan' who feature in the novel *Pilgrim's Progress* by John Bunyan (1678). Haldane's essay makes the point that structures in biology are ultimately constrained by the laws of physics. However, a common theme found in science fiction movies and speculative fiction is where people are made very small or an animal is made terrifyingly huge. Almost always, scientific feasibility is suspended for the sake of the story.

Consider the size of mammalian cells. Although there are a few exceptions (e.g. some specialised nerve cells and egg cells) almost all fall with a range of 5–50 μm. Even for animals which differ in size by several orders of magnitude the difference in size reflects the number of cells not the size of the cells.

So, what constrains almost all cells to this microscopic size?

Cells need to have constant interaction with their environment as dissolved gases and small molecules must be constantly absorbed and waste molecules must be excreted. These substances need to pass through the plasma membrane, which means that the internal regions of the cell are reliant on the cell surface. That is, what is important for the cell to function is a large enough surface-area-to-volume ratio. But as cells become larger their volume increases at a faster rate than their surface area.

Imagine a spherical cell with a diameter of 10 µm. Its surface area is calculated by $4\pi r^2$, so is approximately 300 µm^2. The volume is calculated by $4/3\pi r^3$, so is approximately 500 µm^3. Therefore, the surface-area-to-volume ratio is 0.6. However, if the diameter is increased to 100 µm (a 10-fold increase), the surface area is approximately 30,000 µm^2 (a 100-fold increase) and the volume is 500,000 µm^3 (a 1,000-fold increase). The surface-area-to-volume ratio has thus decreased to 0.06.

16

4.4 Boffin questions

Fantastic Voyage

In the American science fiction film *Fantastic Voyage* (20th Century Fox, 1966), a team of scientists are miniaturised along with a special submarine. The process of miniaturisation involves shrinking individual atoms. They are then injected into the bloodstream of an invaluable scientist to clear a blood clot and save his life. At one point in the film, the team travels through the inner ear where one team member, played by Raquel Welch, gets tangled in hair cell cilia. Obviously, the film is pure fantasy, but it is worth thinking about the implications of being shrunk to microscopic size.

1. In the film *Fantastic Voyage*, the hair cell cilia are about the width of Raquel Welch's arm. In reality hair cell cilia are about 0.2 µm in width. What is the wavelength range of visible light? What are the implications of this for her vision once miniaturised?
2. If the humans are miniaturised only by shrinking their atoms, then their mass stays the same while their volume decreases. By what factor does their density (mass/volume) change? One way to calculate this is to assume a starting mass of 1,000 kg m^{-3}. Volume will approximately decrease by (miniature height / normal height)^3. Assume they start at 1.8 m and decrease to 3 µm. What are the implications of this density change?

V

Chapter 5: Scientific notation and SI units

A criticism of the International System of Units (SI units) is they frequently involve numbers too big or too small to use conveniently. This is commonly seen in biology where the size of biological

components and their concentrations can be very small and the difference in the magnitude of different components can be huge (see **Chapter 4: Biological scale**). If a number needs to be expressed with a very long chain of zeros, it will take longer to record or write down, and there is a larger chance of a mistake.

One way to get around this is to use scientific notation, in which very large or small numbers are written as two numbers multiplied together (usually exponents of base 10). For example, 3,000,000 m could be written as 3×10^6 m and 0.000000005 mol could be rewritten as 5×10^{-9} mol.

Another common way to express larger or smaller numbers is to make the unit larger or smaller using prefixes. You will be familiar with some of these, such as kilo (k, 10^3), milli (m, 10^{-3}) and centi (c, 10^{-2}).

Most of the prefixes adjust the unit by increments of three decimal places. That is, they involve 1,000-fold increases or decreases in scale.

Become familiar with the prefixes in the table below. Learn their names and symbols so that you can easily convert between them.

Symbol	Name	Notation	Factor
T	tera	10^{12}	1,000,000,000,000
G	giga	10^{9}	1,000,000,000
M	mega	10^{6}	1,000,000
k	kilo	10^{3}	1,000
d	deci	10^{-1}	0.1
c	centi	10^{-2}	0.01
μ	micro	10^{-6}	0.000001
n	nano	10^{-9}	0.000000001
p	pico	10^{-12}	0.000000000001
f	femto	10^{-15}	0.000000000000001
a	atto	10^{-18}	0.000000000000000001

What are SI units?

SI stands for Système International d'Unités. This system, initially known as the metric system, was developed in France after the French Revolution and has now been adopted internationally. At first it consisted of three units – the metre, the kilogram and the second – but eventually expanded to include seven principle (basic) units. There are many more SI recognised units, but they are all derived from these basic seven units (e.g. the SI unit for force is the newton, N, which is derived from $kg.m.s^{-2}$)

Quantity	Unit	Symbol
Length	metre	m
Mass	kilogram	kg

Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Quantity of substance	mole	mol
Luminosity	candle	cd

Note

The litre is a non-SI unit which has been accepted as a measure for volume. Its symbol is either l or L. We've used L in this book as it can be easier to read, although it's conventional for prefixed units to use lowercase (e.g. ml or dl).

17

5.1 Converting between SI units

The following section provides a systematic approach you can use to convert between SI units. You will gain confidence with these kinds of conversions only through practice, and you will need to memorise the unit prefixes and what conversion factor they represent.

When converting to or from a base unit you need to determine the conversion factor. A way to do this is to use 1 for the prefixed unit and a power of 10 in front of the base unit.

Example 1

How many mg are in 1 g?

$1 \text{ mg} = 10^{-3} \text{ g}$

Multiply both sides by 1,000 (10^3)

Therefore, 1,000 mg = 1 g

Example 2

Convert 10 µg to g

$10 \text{ µg} = 10 \times 10^{-6} \text{ g}$

Example 3

What about converting from one prefix to another? You can do this as a two-step process by going through the base unit and using two conversion factors.

How many ng are in 1 kg?

Step 1: 1 ng = 10^{-9} g

Therefore, 10^9 ng = 1 g

Step 2: 1 kg = 10^3 g

Therefore, number of ng in 1 kg = $10^9 \times 10^3 = 10^{12}$ ng

Note that when multiplying numbers with the same base, add the exponents; the general rule is $a^b \times a^c = a^{b+c}$

Rules for using SI units

The SI is used because it is precise, but it has also developed some conventions to avoid ambiguity. Below are the rules you should follow when using SI units. However, you will come across many examples – often in popular media, but also unfortunately even in some textbooks or peer-reviewed publications – where these rules are broken due to ignorance or carelessness!

1. Never pluralise units. This can cause confusion, as kms does not mean kilometres but would be interpreted as km multiplied by s.
2. Never use a full stop after unit abbreviations unless it is the end of a sentence.
3. The unit (plus prefix) should be separated from the unit by a space (e.g. 10.3 µl).
4. Do not separate prefixes from the unit by a space. (e.g. 10 micrometres should written 10 µm not 10 µ m)
5. When you combine units, separate them by a space (e.g., 10 mg l^{-1}). This example means '10 milligrams per litre'. You can use –1 or use a slash (/) for per (e.g. 10 mg/l). Do not write 'per'.
6. Only use one prefix in front of a unit.
7. A unit named after a famous person, such the joule (named after the English physicist James Prescott Joule, 1818–1889), is written in lowercase but when abbreviated is written in uppercase (J).

18

5.2 Practice problems
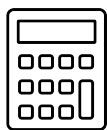
1. Convert the following decimal values to scientific notation:
   a. 123,000 m
   b. 299 l
   c. 0.0035 mol
2. Complete the following SI unit conversions:

a. 0.004 mol to mmol
b. 1.76 ml to μl
c. 0.0023 μm to nm
d. 20.3 μl to l
e. 0.12 fg to ng
f. 1.2 nmol to pmol
g. 1.2 mg μl^{-1} to g l^{-1}
h. 1.2 ng μl^{-1} to g l^{-1}

Solution to Practice Problem 5.2h.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1021

19

5.3 Boffin questions

How many cells are in your body?

1. You may have seen various estimates of the total number of cells in the human body in textbooks and online. A common estimate is 30,000,000,000,000 (30 trillion or 3×1013) cells. Have you ever thought about how these numbers are estimated? See if you can come up with a similar estimate for your own body.
   Hints: First, estimate your volume. You can assume 1 kg takes up about 1 l. Now convert your volume to m^3.
   Mammalian cells typically have a volume ranging from 10^3 to 10^4 μm^3.
   How many cells is that?
   Your answer should be a range based on the volume range. Does the range encompass the common estimate stated above?
2. How much of us is human? Based on Sender et al. (2016), what percentage of our body weight on average is composed of bacteria?

Side Bar – Take care when writing down units as it is easy to make a mistake especially when the difference in the symbols is slight. The goal should always be to avoid any ambiguity! For example, nm refers to nano metres or 10^{-9} metres while nM refers to nano molar or (10^{-9} moles per litre). Similarly, it is each to confuse nM with mM.

A more accurate estimate

Your estimate above is a great starting point to think about this problem but is likely to be very inaccurate due to the assumptions made regarding both the weight-to-volume ratio and range of volumes for mammalian cells.

Different cell types of different volumes will account for different proportions of the total cell population. Erythrocytes (red blood cells) are much smaller than most cells and there are typically about 5–6×10^{12} cells per litre of blood. With a blood volume of around 5 L on average that means there are around 2.5–3×10^{13} erythrocytes in an average human. This makes them the largest contributor to the overall human cell count, as shown in the next figure.
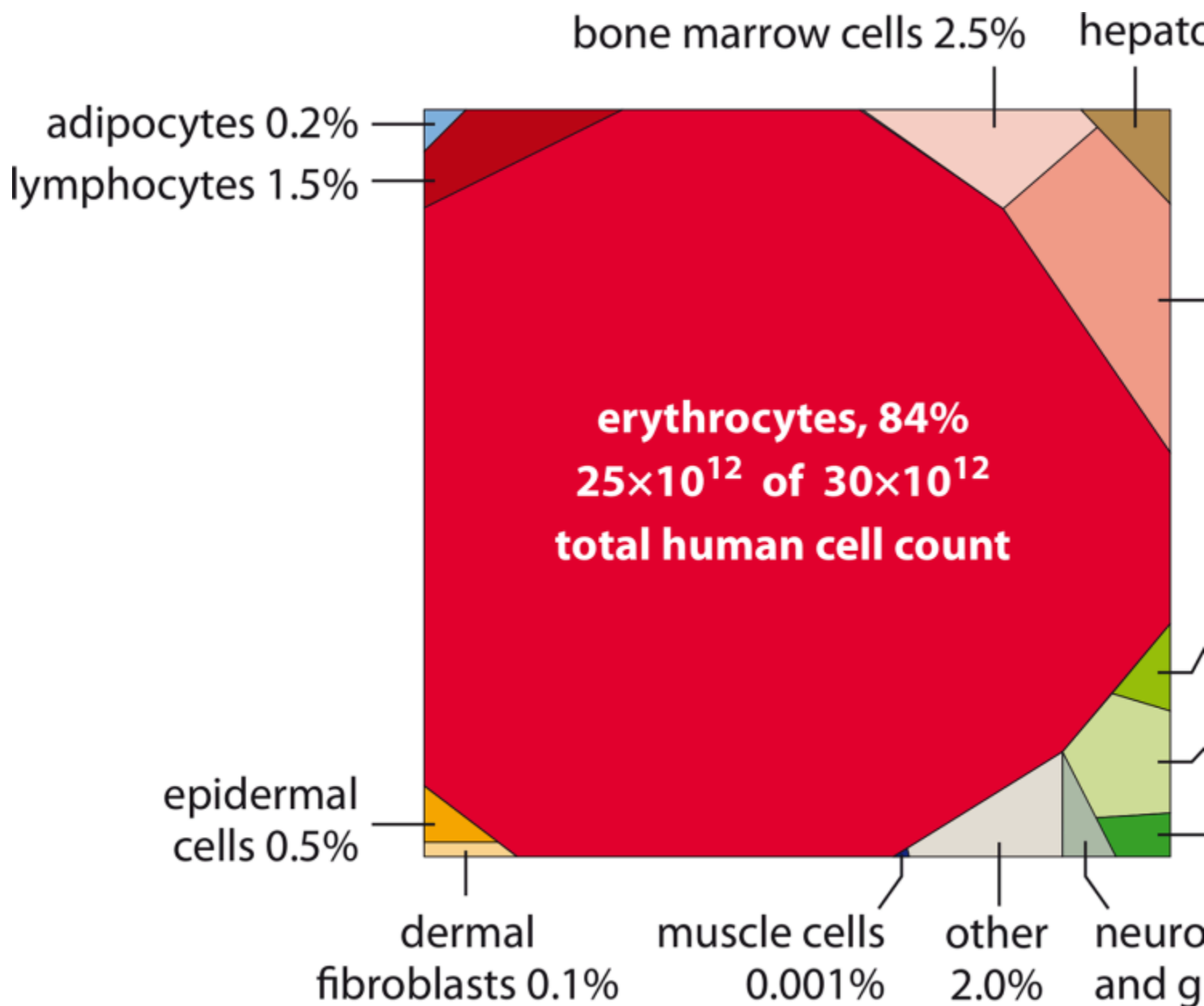
Figure 5.1: The relative contribution the major cell types to the total cell number in the human body ('The distributio type' by Ron Sender, Shai Fuchs and Ron Milo from Revised Estimates for the Number of Human and Bacteria Cel https://doi.org/10.1371/journal.pbio.1002533)

Sender et al. (2016) have attempted a more accurate estimate of the total cell number by estimating the relative contribution of each of the major cell types. It is worth looking at their paper to see how they made their calculations. They have also estimated the total number of bacteria in an average human, which has been the subject of a lot of controversy.

VI

Chapter 6: Blood composition

The primary function of blood is to deliver oxygen and nutrients to and remove waste from the body's cells. However, it is also responsible for transport of hormones, protection (clotting and immune defence), distribution of heat and maintenance of homeostasis (pH and water balance). To carry out these wide-ranging functions, blood has a necessarily complex composition.

Blood has two main components, plasma (blood fluid) and formed elements (cells and platelets). You quickly see these main components by performing a haemocrit test. In this test whole blood is centrifuged to pellet the formed elements at the bottom of the tube, leaving the plasma at the top of the tube. A haemocrit is normally conducted as part of a complete blood count and used to determine the volume percentage of erythrocytes in the blood. For a normal haemocrit the contribution of erythrocytes is about 45% but this varies depending on a number of factors including sex, age, pregnancy and even the altitude you live!
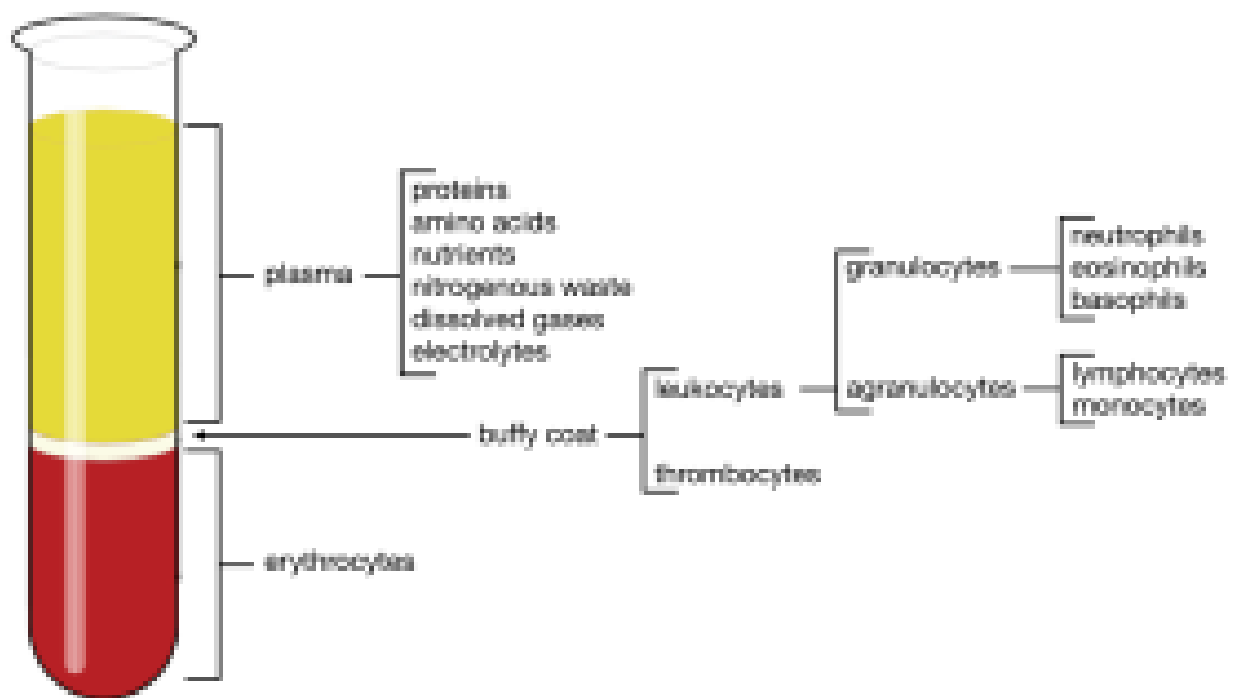
Erythrocytes are the dominant cell type; they comprise 99% of the cells in blood. In a haemocrit you will also see a thin layer of buff-coloured cells above the erythrocytes (buffy coat) which contains most of the leukocytes and thrombocytes (platelets).

How big are erythrocytes?

Erythrocytes are small cells with a diameter of 6–8 μm. They contain haemoglobin and their primary function is to distribute oxygen.
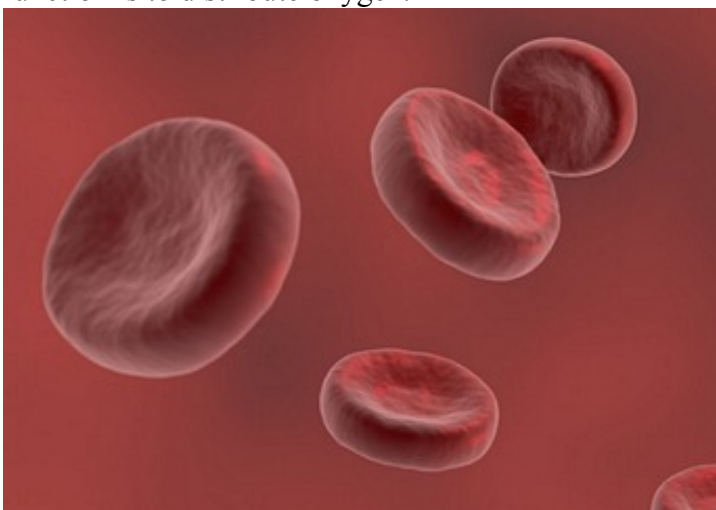
Figure 6.2: Red blood cells ('Red blood cell' by Crystal Blair from Pixabay used under Pixabay Licence)

Estimating cell size

We can estimate cell size by viewing cells under a microscope. The circle of light you see through the ocular lens is known as the field of view. You can determine the width of the field of view by using a stage micrometer, which is a microscope slide that has a scale etched into its surface. A haemocytometer (see the next section) can be used for crude estimates.

Once you have an idea of the width of the field of view you can estimate how many cells laid end to end it would take to equal the diameter of the field of view.
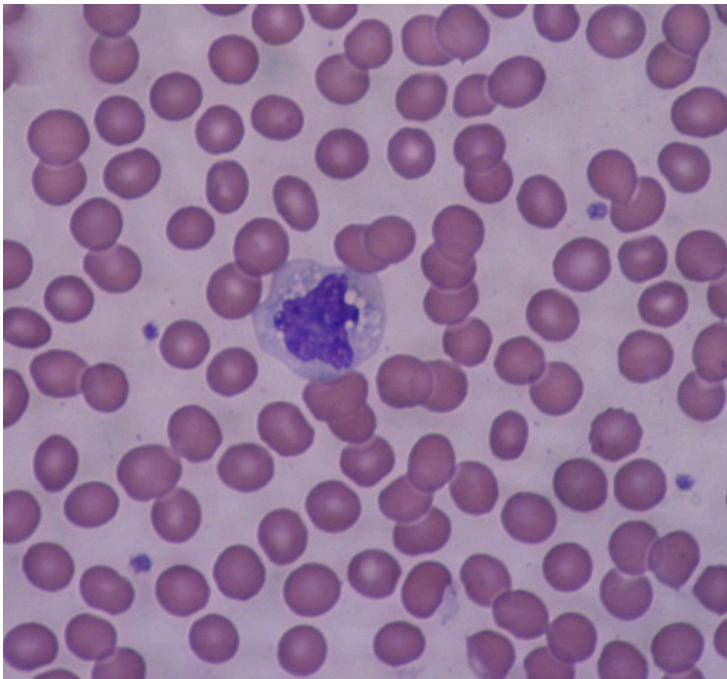
Figure 6.3: Monocyte ('Monocyte' by Bobjgalindo from
Wikimedia Commons used under CC BY-SA 4.0.)
Estimate the size of the monocyte (largest cell) in µm if the field of view is 0.2 mm wide.

20

6.1 Counting cells

The device shown in the next figure is used for accurately determining cell concentrations. It is known as a haemocytometer, since it was originally designed for performing blood cell counts. However, it is also used for cell culture, microbiology or any application where it is necessary to determine the number of cells per unit volume of a suspension.
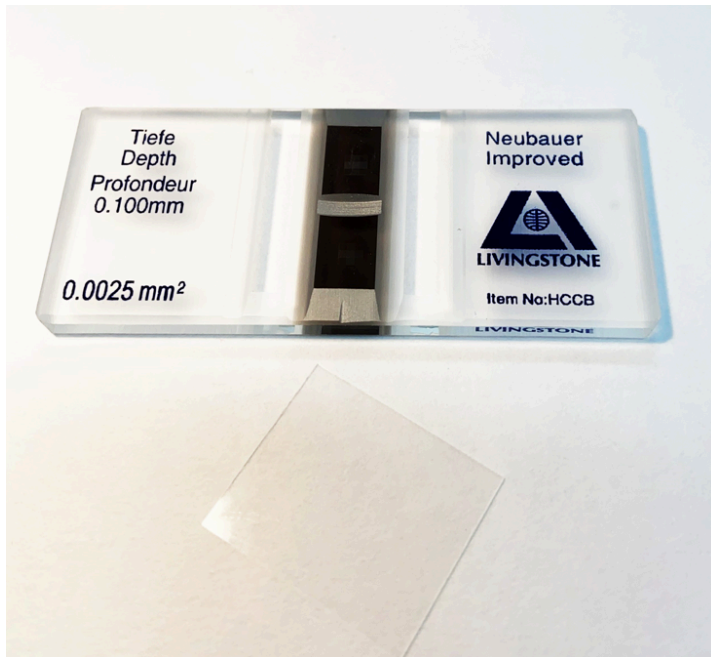
Figure 6.4: Haemocytomer used for counting cells.
Photo by Julian Pakay.

A haemocytometer is a modified and calibrated microscope slide which has two coverslip supports precisely raised above a cell counting surface inscribed with a precise grid, as shown in the next figure. This creates a defined volume for counting cells in suspension, which are pipetted into the counting chamber.
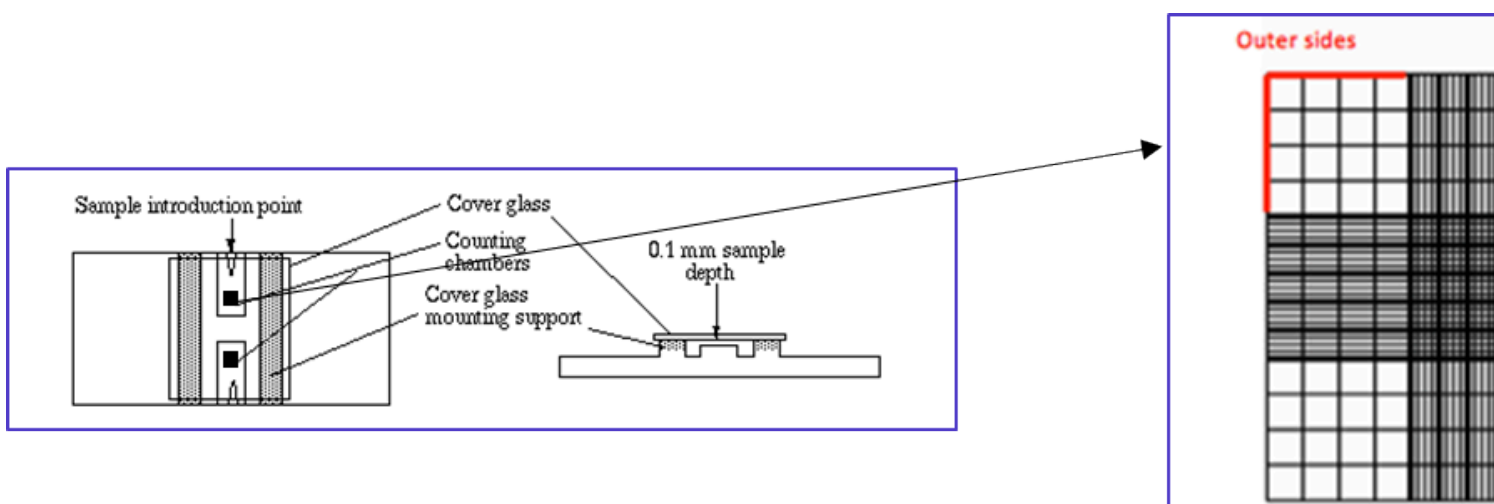
Figure 6.5: A haemocytometer counting chamber (Adapted from 'Using a Counting Chamber' created by David R. C University.)

Care is taken to dilute the cell suspension, so it is not so crowded it is impossible to count and also so that the cells are uniformly distributed. Often a dye is added to the cells which distinguishes dead cells from viable cells. The cell suspension is pipetted slowly into the counting chamber, which fills by capillary action. For most mammalian cell suspensions, the large corner squares (1 mm × 1 mm) are used for counting but smaller squares may be used for smaller cells such as erythrocytes or yeast. The cells are counted while viewing the haemocytometer under a microscope. Counting is done systematically to avoid bias. For instance, cells that overlap a line count as 'in' if they overlap the top or right outer line, and 'out' if they overlap the bottom or left outer line.

Calculating cell concentrations from a haemocytometer

A typical depth for a haemocytometer counting chamber is 0.1 mm. Therefore, the volume defined by one of the 1 mm outer squares is:

height x width x depth = volume

$1 \text{ mm} \times 1 \text{ mm} \times 0.1 \text{ mm} = 0.1 \text{ mm}^3$

(there are 1,000 mm^3 in 1 cm^3 and 1 cm^3 = 1 ml)

Therefore, 0.1 mm^3 = 1 ml/1,000 mm^3 x 0.1 mm^3 = 0.0001 ml = 0.1 µl
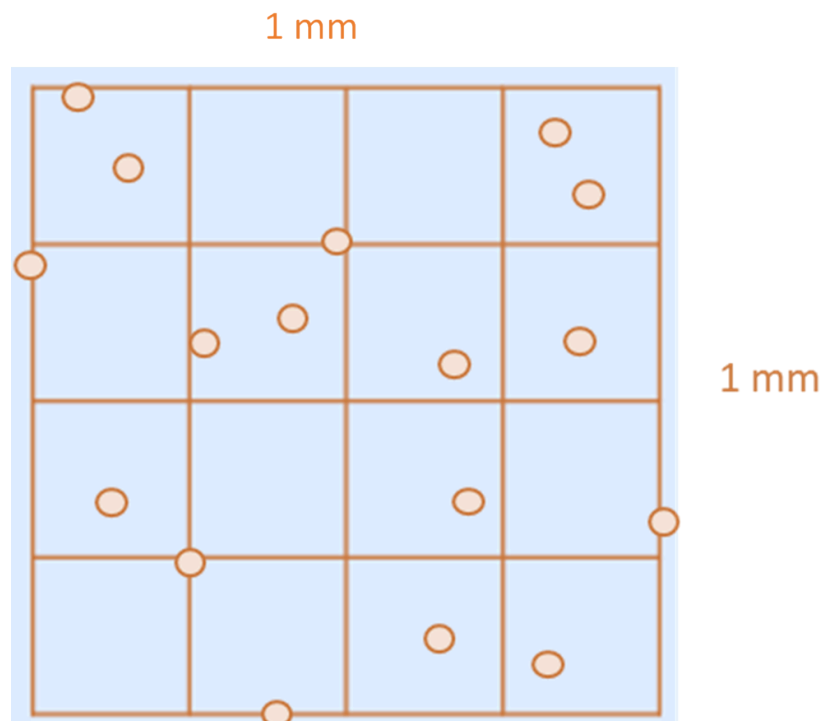
Figure 6.6: A diagram showing one of the outer corner squares from. The circles represent cells. Since some cells straddle the

outside border, a methodical approach is required to ensure consistent counting.

In the 1 mm × 1 mm square above, there are 15 cells (counting to avoid bias). Assuming this is representative of the cell suspension they were aliquoted from; this means that cell suspension contains:

15 cells per 0.1 µl

which is 150 cells per µl

which is 150,000 cells per ml.

21

6.2 Practice problems
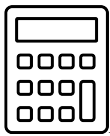
1. A researcher purified some platelets from human blood. She then added an aliquot of this suspension to a haemocytometer (Figure 6.7). Estimate the platelet concentration (cells ml^{-1}) of her suspension. The haemocytometer has a depth of 0.1 mm.
2. From your knowledge of erythrocytes, estimate the size (in µm) of the leukocyte in the micrograph (Figure 6.8). What type of leukocyte do you think it is?
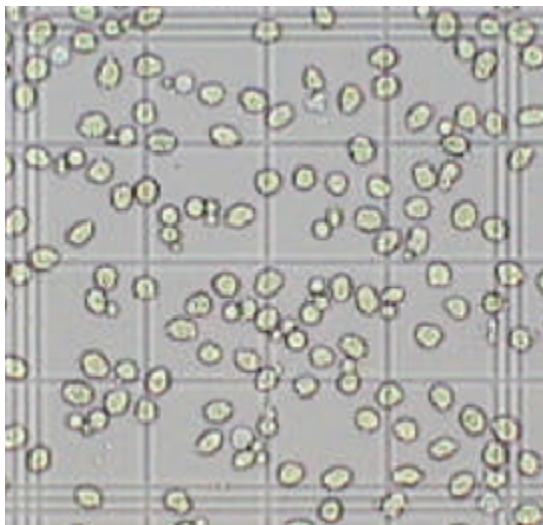
0.2 mm

Figure 6.7: A diagram showing one of the outer corner squares from a haemocytometer containing purified human platelets. The dimensions of the main square are 0.2 mm x 0.2 mm with a depth of 0.1 mm.
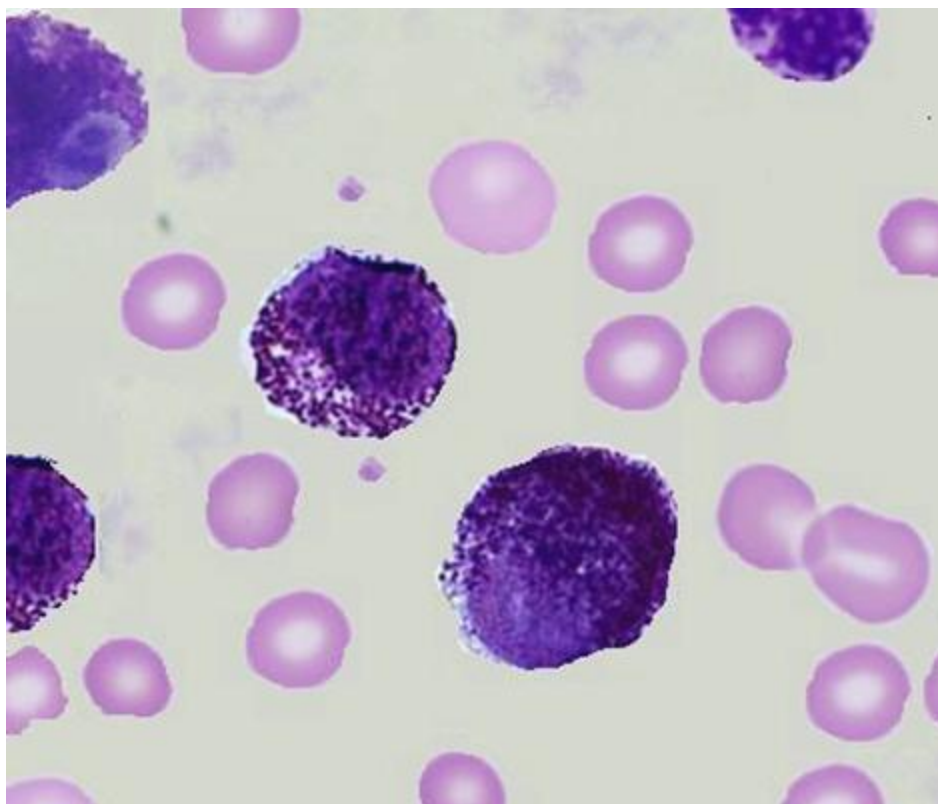
Figure 6.8: Unknown leukocyte ('Mast cell leukemia' by Ayman Qasrawi from Wikimedia Commons used under CC0)

VII

Chapter 7: Solutions and concentrations

A solution consists of at least two components: the solvent, which is the component in the largest quantity (water for aqueous solutions); and the solute (the component dissolved in the solvent). A true solution is a homogenous mixture of the solute(s) and solvent where the solute molecules are dispersed within the solvent.

Concentration refers to the amount of solute dissolved per unit volume of the solvent. That is, concentration quantitatively describes the ratio of solute to solvent. There are numerous ways to express concentration but most often you will encounter concentration expressed in one of the following ways.

- **Molarity:** expressed as M or mol l^{-1} and probably most used in chemistry and biochemistry. Molarity refers to the molar concentration, that is, the number of moles per litre. This is a particularly useful measure of concentration for determining the stoichiometry (relative amount of reactant and product) for a chemical reaction.
- **Percentage composition:** non-SI units that typically express the weight-to-volume percent (w/v % – percent of solute mass to solvent volume) or volume-to-volume percent (v/v % – percentage volume of solute to volume of solvent). These are commonly used for concentrated solutions.
- **Parts per million or parts per billion:** expressed as ppm or ppb. These are often used for very dilute solutions.

- **Mass per volume:** commonly used for drug solutions but also used in biochemistry to describe protein or nucleic acid solutions (expressed as g l^{-1} but the prefixes will vary depending on the concentration; e.g. μg ml^{-1}).

Molarity and the mole

The mole is the SI unit used to measure the amount of a substance. The technical definition is:

- the amount of substance which contains as many elementary entities are there are atoms in 0.012 kg of carbon-12 (^{12}C).

The number of particles in a mole is known as Avogadro's number

Amedeo Avogadro was the first person to clearly differentiate molecules and atoms, but the concept of the mole was only introduced after his death.

and is approximately equal to 6.02×10^{23}.

To determine the number of moles of a substance, you need to know its mass and molecular weight.

The molecular weight of a substance is the mass of 1 mol of a substance in g mol^{-1}.

Molecular weight or molecular mass is often interchangeably used to mean the mass of a single molecule in daltons (Da), which is the mass of molecules relative to 1/12th the mass of a carbon-12 atom. Note that although they are numerically identical, they are actually different units.

To determine the number of moles of a substance you need to divide the mass of the substance by its molecular weight:

$$\text{mol} = \frac{\text{mass (g)}}{\text{molecular weight}}$$

22

7.1 Calculating molar concentrations

To calculate concentration, you need to know the amount of a substance and the volume it is dissolved in.

Example 1

What is the molarity of a solution containing 0.30 mol of NaCl in 3.6 l?

Remember: $\text{molarity (M)} = \frac{\text{moles of solute (mol)}}{\text{volume of solution (l)}}$

$$\text{Molarity (M)} = \frac{0.30 \text{ mol}}{3.6 \text{ L}} = 0.08 M$$

Example 2

What is the molarity of a solution containing 210 g of NaCl in 3.6 l? The molecular weight of NaCl is 58.44 g mol^{-1}.

In this case, you first need to work out the number of moles from the mass:

Remember: $\text{number of moles (mol)} = \frac{\text{mass of substance (g)}}{\text{molecular weight}}$

Step 1: $\text{number of moles (mol)} = \frac{210}{58.44} = 3.6 \text{ mol}$

Step 2: $\text{molarity (M)} = \frac{3.6 \text{ mol}}{3.6 \text{ l}} = 1.0 \text{ M}$

Example 3

It can be useful to be able to do these calculations in reverse to work out the mass needed to make up a particular solution.

How would you prepare 200 ml of 0.25 M KNO$_3$ solution? The molecular weight of KNO$_3$ is 101.1 g mol^{-1}.

In this case you first need to work out the number of moles you need.

Remember: $\text{Molarity (M)} = \frac{\text{moles of solute (mol)}}{\text{volume of solution (L)}}$

You will need to rearrange the equation above to solve for moles of solute. It helps when rearranging equations to write out the units in full as it makes it much more obvious if you have made a mistake. Also remember to convert any volumes to litres.

$$\text{Molarity (mol l}^{-1}) = \frac{\text{moles of solute (mol)}}{\text{volume of solution (l)}}$$

Moles of solute mol = molarity (mol l^{-1}) × volume of solution (l)

Moles of KNO$_3$ = 0.25 mol l^{-1} × 0.2 L = 0.05 mol

Next you can work out the mass you need to weigh out.

Mass of substance (g) = number of moles (mol) × molecular weight (g mol^{-1})

Mass of KNO$_3$ (g) = 0.05 mol × 101.1 g mol^{-1} = 5.06 g

So, the full answer to question is: weigh out 5.06 g of KNO$_3$ and dissolve in water to a volume less than 200 ml. Adjust the final volume to 200 ml.

7.2 A focus on calculating mass from molarity and volume

If you had to prepare a 5 ml aliquot of 20 mM glucose for each of pair of students in a class of 120, how much glucose would you need to weigh out? The molecular weight of glucose is 180.156 g mol^{-1}.

Mass from Molarity and Volume

Vy Hoang is senior laboratory technician. Part of her work involves preparing reagents and materials for teaching laboratories. The problem above is a typical one she must solve on every day, and she would have solved many similar problems during her time in research. Watch an interview with Vy to see how this kind of problem is solved but also to see how to develop the quantitative mindset required to do routine laboratory work.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1037

7.3 Calculating other concentrations

To calculate concentration, you need to know the amount of a substance and the volume it is dissolved in.

Example 1: Mass per volume concentration

12 µg of protein is dissolved in 200 µl of water. What is the final concentration in mg ml^{-1}?

Remember: $\text{concentration} = \frac{\text{mass}}{\text{volume}}$

First convert the mass to mg and the volume to ml to get the answer in the correct form.

12 µg = 0.012 mg

200 µl = 0.2 ml

Example 2: Weight to volume percentage

What is the weight/volume percentage (w/v (%)) concentration of 250 ml of aqueous sodium chloride solution containing 25 g of NaCl?

In this case we assume that 1 ml is equivalent to 1 g.

w/v (%) = (mass solute ÷ volume of solution) × 100

Mass of solute (NaCl) = 25 g

Solution volume = 250 ml

w/v (%) = (25 ÷ 250) × 100 = 10%

Example 3: Volume to volume percentage

What is the percent by volume (v/v (%)) of a solution formed by mixing 15 ml of isopropanol with 60 ml of water?

v/v (%) = (volume of solute ÷ total volume of solution) × 100

Mass of solute (isopropanol) = 15 ml

Total volume of solution = 60 ml + 15 ml = 75 ml

v/v (%) = (15 ÷ 75) × 100 = 20%

Example 4: Calculating ppm and ppb

Parts per million (ppm) and parts per billion (ppb) are examples of expressing concentrations by mass. Volume/volume (v/v) and weight/volume (w/v) concentrations are sometimes expressed in ppm, especially if very dilute.

1 ppm is 1 part by weight, or volume, of solute in 1 million parts by weight, or volume, of solution.

$$\text{ppm} = \frac{\text{mass solute}}{\text{mass of solution}} \times 10^6$$

$$\text{ppb} = \frac{\text{mass solute}}{\text{mass of solution}} \times 10^9$$

What is the concentration of a solution, in parts per million, if 0.03 g of NaCl is dissolved in 1 l of water?

First convert the volume of solution to a mass:

1 l = 1 kg = 1,000 g

$$\text{ppm} = \frac{0.03}{1,000} \times 10^6 = 30 \text{ppm}$$

7.4 A focus on biopharmaceutical production

To ensure that the medications we take are safe, we make them comply with many regulations and an enormous degree of quality control. Biopharmaceuticals are pharmaceuticals which often must be purified from complex biological sources; they have very strict controls to ensure they do not contain harmful contaminants.

One class of monitored contaminants is endotoxins. Endotoxins are lipopolysaccharides found in the cell wall of gram-negative bacteria. They can elicit an immune response in humans and other animals which leads to inflammation, fever and in extreme cases anaphylaxis and death.

Drugs must often be tested in animal models. The US Food and Drug Administration (FDA) and the Australian Therapeutic Goods Administration (TGA) regulations stipulate that the dose limit of endotoxin does not exceed 5 endotoxin units (EU) per kg of body weight for rabbits.

If a drug 'x' needs to be injected at a dose of 0.1 ml kg^{-1} but contains 184 EU ml^{-1} is it safe? (The average weight of a rabbit is 5 kg.)

Biopharmaceutical Production

Adi Alhuwaider is a senior research assistant at St Vincent's Institute for Medical Research. He has extensive experience in research and in the production of biopharmaceuticals. Watch an interview with Adi to learn how to solve the problem above but also to see how important quantitative literacy is for working in drug development and production.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1041

7.5 Practice problems
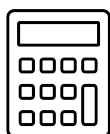
1. What is the molarity of a solution where 5 moles of Na_2CO_3 is dissolved in 2.5 l of water?
2. What is the molarity of 5 g of NaOH in 750 ml of solution? (The molecular weight of NaOH is

40 g mol^{-1}.)

3. How many moles of H_2SO_4 are in 10 l of a 2 M solution?

4. What weight (in g) of Na_2CO_3 is needed to make 750 ml of a 100 mM solution? (The molecular weight of Na_2CO_3 is 106 g mol^{-1}.)

5. A patient has a cholesterol count of 97 mg/dl. What is the molarity of cholesterol in this patient's blood? (The molecular weight of cholesterol is 386.64 g mol^{-1}.)

6. What is the mass of HCl present in 90 ml of a 3 M solution? (The molecular weight of HCl is 36.46 g mol^{-1}.)

7. Convert 50 mg of calcium carbonate, $CaCO_3$, into moles. (The molecular weight of $CaCO_3$ is 100.1 g mol^{-1}.)

8. You are given a 5 µM solution of a protein (molecular weight = 27,000 g mol^{-1}). What is the concentration in mg ml^{-1}?

9. You lyophilise 1 ml of a 50 nM solution of a 60 kDa protein dissolved in water. What is the expected weight of the dried protein? (Come back to this one if you find it difficult.)

10. You want to prepare 300 ml of a 60% (w/v) solution of sucrose. How much sucrose do you need to weigh out?

11. 2 l of an aqueous solution of KCl contains 45 g of KCl. What is the weight/volume percentage concentration of this solution?

12. A solution is made by adding 30 ml of benzene to 95 ml of toluene. What is the percent by volume of benzene?

13. How much ampicillin do you need to weigh out to make 150 ml of solution containing 50 µg ml^{-1} ampicillin?

14. If 0.025 g of $Pb(NO_3)_2$ is dissolved in 0.1 l of H_2O, what is the concentration of the resulting solution, in parts per million?

15. A 600 g sample of stream water has a concentration of 130 ppb dissolved nitrates. What is the mass of dissolved nitrates in this sample?

16. What is the total mass of solute in 1 l of water at a concentration of 5 ppm?

Solution to Practice Problem 7.2.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1043

Solution to Practice Problem 7.9.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1043

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1043

27

7.6 Boffin questions
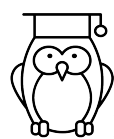
Determining an unknown concentration using spectrophotometry

Many molecules absorb light in the visible or ultraviolet portion of the electromagnetic spectrum. Spectrophotometry is a measure over what wavelengths and by what amount molecules absorb light. This technique uses a device known as a spectrophotometer which can pass light of a specific wavelength and known intensity through a solution and determine the amount of light transmitted. This can be used qualitatively to identify compounds as different molecules will have characteristic absorbance spectra. That is, they will absorb light at specific wavelengths. For example, proteins absorb light in the UV range with maxima at 220 nm due to absorbance by peptide bonds and at 280 nm due to aromatic amino acid side chains.

Absorbance can also be used quantitatively as the amount of light absorbed by a molecule in solution follows a linear relationship with its concentration. This relationship is described by Beer's law:

$$A = \varepsilon bc$$

where

A is the absorbance

b is the pathlength of light through the solution (cm)

c is the concentration of the analyte (M)

ε is the wavelength-dependent molar extinction coefficient or molar absorption coefficient $(M^{-1}\ cm^{-1})$.
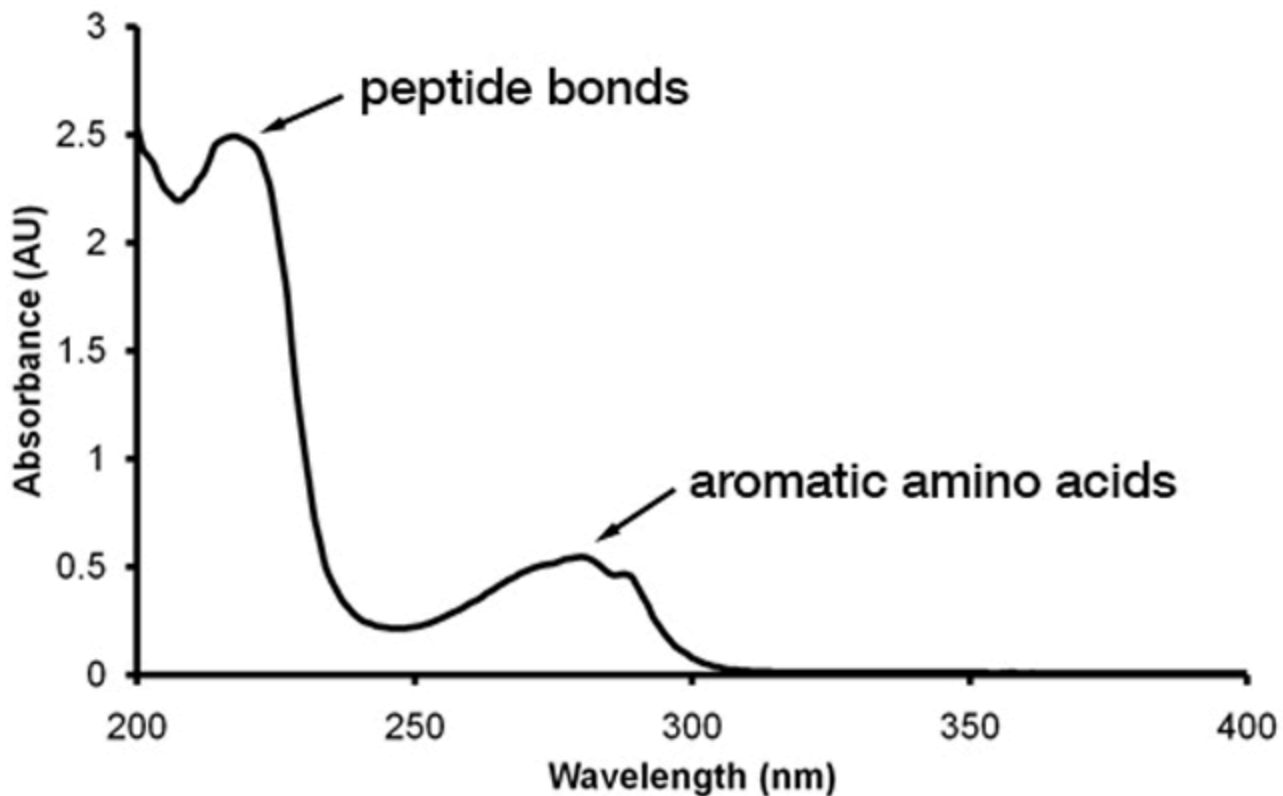
Figure 7.1: Characteristic UV absorbance spectrum for protein (Adapted from 'Absorption spectrum of a peptide in the near and far UV regions' from Nptel used under CC-BY-NC-SA)

Using Beer's law, you can calculate the concentration of a solution based on how much light it absorbs. There are two ways to do this.

1. If the molecule is well-characterised its molar extinction coefficient (ε) may be known and absorbance need only be measured over a defined pathlength to determine the concentration.
2. You can prepare a series of known concentrations of the same molecule that you are trying to determine the concentration of in your sample. Then you measure the absorbance of those solutions alongside that of the sample with unknown concentration. From this you can then work out the equation for the line of known concentrations with their corresponding absorbances (called a standard curve or calibration curve). Then you can use this equation to determine the concentration of the unknown sample from its absorbance.

How spectrophotometers work

First a collimator lens transmits a straight beam of light through a monochromator (prism) to split it into several component wavelengths (spectra). Then a wavelength selector transmits only the desired wavelengths, through the solution to be measured with a defined pathlength. A photometer then detects the number of photons transmitted and sends a signal which is converted to absorbance and displayed.
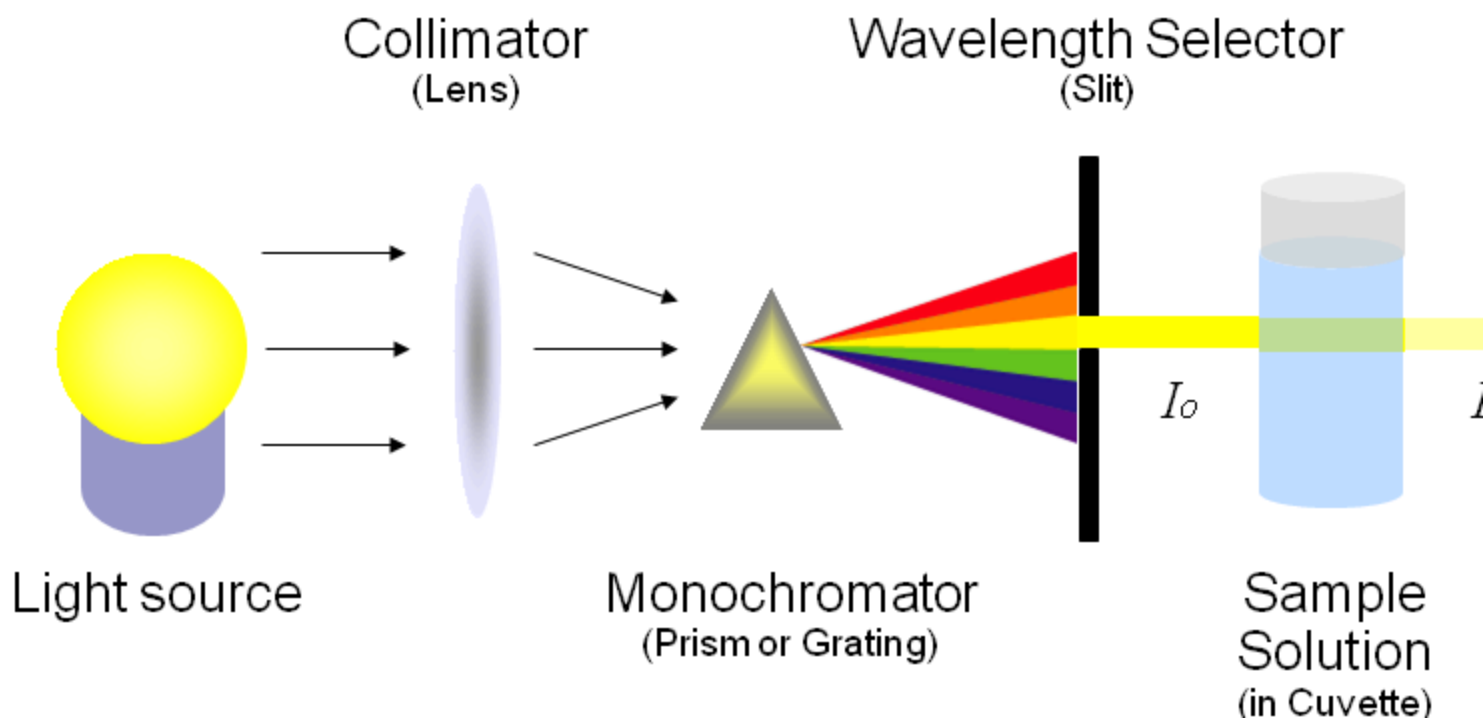
Figure 7.2: Basic structure of spectrophotometers ('Basic structure of spectrophotometers' illustrated by Heesung Sh
[BY-4.0.](#))

Determining an unknown concentration

1. The molar extinction coefficient of ATP is 15,400 M^{-1} cm^{-1} at 260 nm. A solution of purified ATP has an absorbance of 0.8 in a 1 cm cuvette. What is its concentration?
2. Below (Figure 7.3) is a standard curve for a spectrophotometric cholesterol assay, measured at 640 nm. What is the concentration of a sample with an absorbance of 0.865?
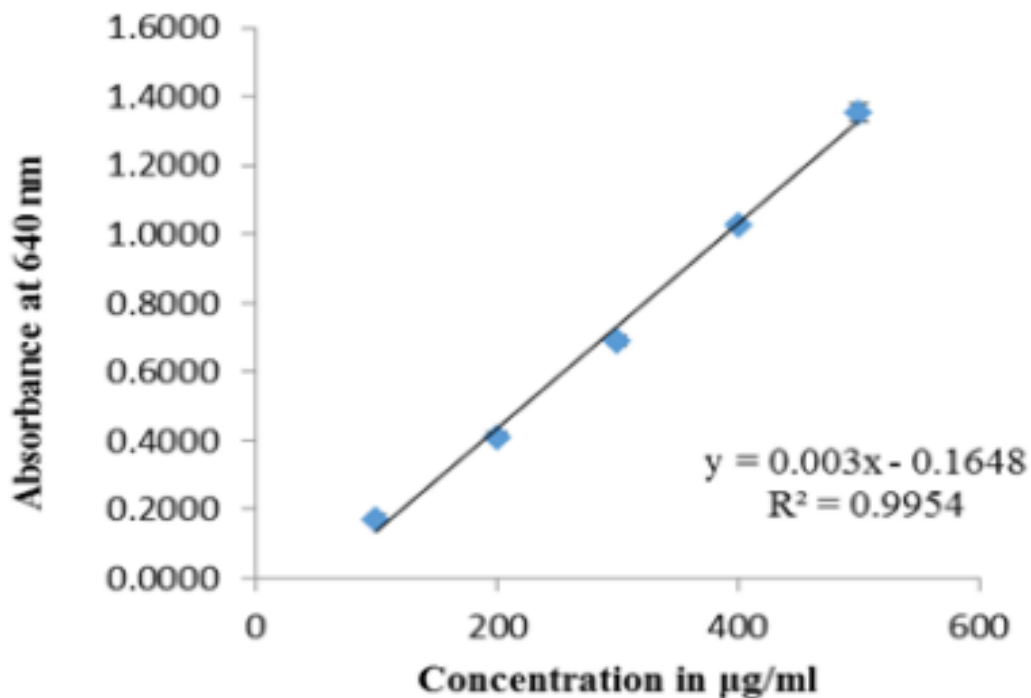
Figure 7.3: A standard curve for spectrophotometric determination of cholesterol where absorbance has been plotted against cholesterol concentration (µg/ml). A line of best fit has been plotted along with its linear regression equation

VIII

Chapter 8: Dilutions

Sometimes you will need to change the concentration of solution by changing the amount of solvent. This is often done when working in the laboratory or when dispensing drugs. Dilution is where additional solvent is added to a solution to decrease the concentration of the solute. It is important to remember that with dilution the amount of solute remains constant. It is only the concentration that changes.
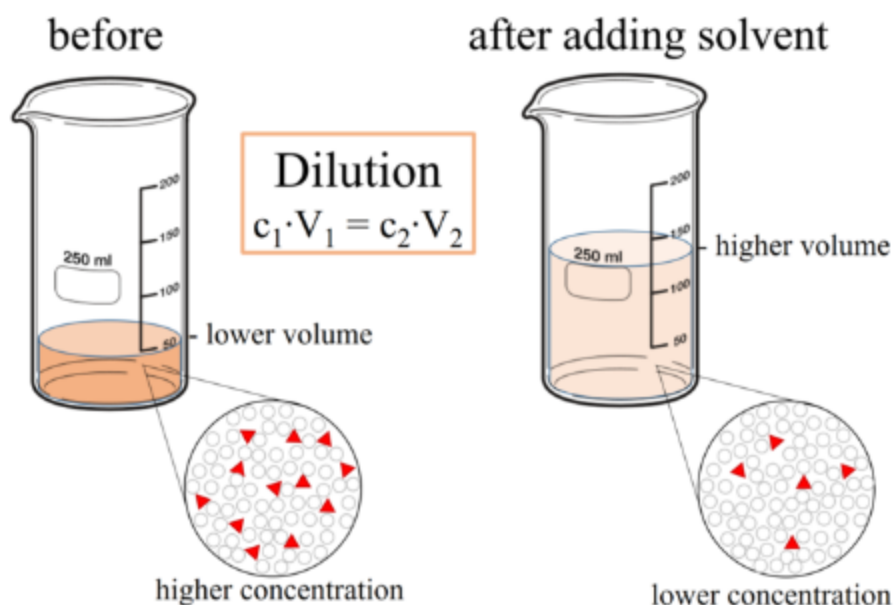
Figure 8.1: The two beakers contain an equal number of solute molecules; only the volume of solvent and hence concentration has changed ('Dilution' by Theislikerice from [Wikimedia Commons](#) used under [CC BY-SA 4.0](#))

The ability to accurately dilute is important as it allows many 'working' solutions of different concentrations to be made from a single concentrated stock solution. As the amount of solute remains the same following dilution the ratio between the concentration and volume remains constant before and after dilution. This can be exploited to work out a final concentration following dilution or the volume of additional solvent to add to achieve a target concentration.

You may come across the term 'dilution factor'. This refers to the change in volume of the solvent. For example, if you add 90 ml of water to 10 ml of an aqueous solution, you have increased the volume from 10 ml to 100 ml. This represents a 10-fold dilution (interchangeably referred to as a 1 in 10 dilution, 1/10 dilution or 10x dilution). Some concentrated stock solutions will be labelled with the suggested working dilution. For example, a commercial buffer may be labelled as 5x, which means it needs to be diluted five-fold for use. Often the dilution factor is simply stated as a number.To calculate the dilution factor, you need to determine the ratio of the volume of the initial (concentrated) solution (V_1) to the volume of the final (dilute) solution (V_2).

So, the dilution factor $= V_2 \div V_1$

Note that since this is a ratio it is important that both volumes are in the same units.

28

8.1 Serial dilutions

To make large dilutions, multiple serial dilutions of the initial stock are often made. In this case you can

determine the final dilution factor by multiplying the factors of each dilution. The final concentration can be determined by dividing the initial concentration by the dilution factor.
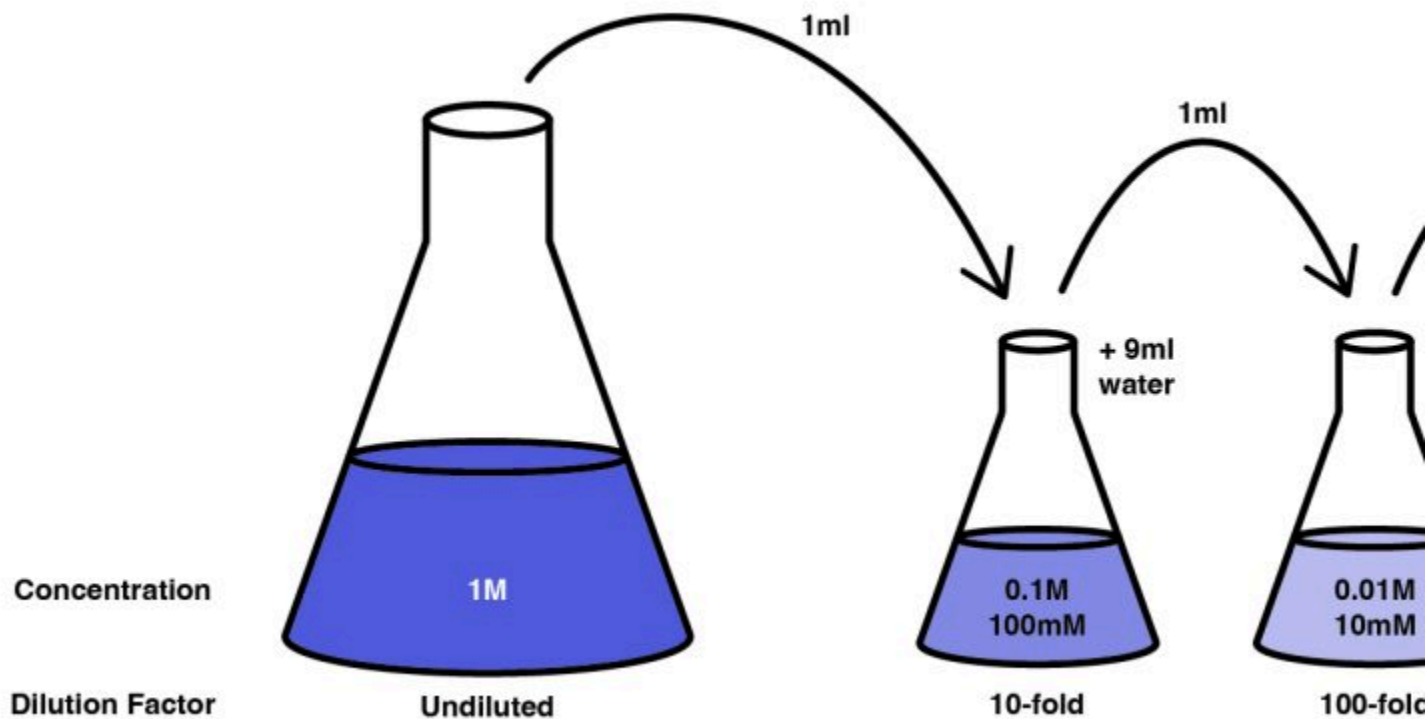
Figure 8.2: Each step involves a 10-fold dilution. The final dilution can be determined by multiplying the dilution fa

Concentration (verb)

As we saw in the previous chapter, concentration is the reverse process, where either solvent is removed or solute is added to increase the concentration of the solute.

How to concentrate solutions

There are several methods to reduce the volume of solvent. One of the simplest ways is by applying heat to **evaporate** some of the solvent. However, this method is not always appropriate. For example, proteins would be denatured by this process.

Other ways include removing the solvent by freeze drying **(lyophilisation)** or by **precipitating** the solute (in the case of proteins or nucleic acids), allowing the solvent to be removed before redissolving in a smaller volume.

A common way to concentrate macromolecules is to use a **semi-permeable membrane** which allows

solvent molecules to pass through but not the larger solute molecules. Such a semi-permeable barrier can be incorporated into a centrifuge tube where centrifugal force pushes the solution against the barrier, concentrating protein in the remaining solution.

Or the solution can be **dialysed** (placed in a sac made from the semi-permeable membrane) and the sac put into a solution of high concentration where again the solute molecules cannot pass through the membrane. In this case water will gradually move from the sac to equilibrate with the surrounding solution by a process called **osmosis**, decreasing the volume, and concentrating the solution in the sac.

29

8.2 Calculating dilutions

Remember that during dilution when additional solvent is added the amount of solute remains constant. It is only the concentration that changes. This means that the ratio between the concentration and volume remains constant before and after dilution. This can be exploited to work out a final concentration following dilution or the volume of additional solvent to add to achieve a target concentration. This ratio can be expressed as:

$$C_1 V_1 = C_2 V_2$$

where

C_1 = concentration of the original solution

V_1 = volume of the original solution

C_2 = concentration of the final solution

V_2 = volume of the final solution.

As this is a ratio, it does not matter what units you use for concentration or volume as long as you **use identical units on each side of the equation**.

Example 1

What volume of 2 M NaCl would you need to prepare 1 l of a 0.01 M solution?

When first doing this sort of calculation it helps to define the terms by writing out the equation in full. Here we are trying to determine what volume of the original solution (V_1) we need.

$C_1 = 2$ M

$V_1 = ?$

$C_2 = 0.01$ M

$V_2 = 1$ l

Since the units for concentration are the same on both sides, no conversion is required. Our V_2 is in litres so that is the unit our answer will be in.

$C_1 V_1 = C_2 V_2$

2 M $V_1 = 0.01$ M $\times$ 1 l

$V_1 = \frac{0.01\text{M}\times 1\text{l}}{2\text{M}}$

$V_1 = 0.005$ l $= 5$ ml

Example 2

What is the final concentration of a solution if 50 ml of a 1 M solution is added to 0.24 l?

Remember you need to total final volume here and you have to convert the volume to the same units on each side of the equation.

$C_1 = 1$ M

$V_1 = 50$ ml

$C_2 = ?$

$V_2 = 50$ ml + 240 ml = 290 ml

$C_2 = \frac{1\text{M}\times 50\text{ml}}{290\text{ml}}$

$C_2 = 0.17$ M = 170 mM

Example 3

Tube 1 contains a 2 M solution; 1 ml is removed to tube 2 containing 9 ml of water. Tube 2 is mixed, and 1 ml is removed from this tube and added to tube 3 containing 19 ml of water. Tube 3 is mixed, and 1 ml is removed from this tube and added to tube 4 containing 4 ml of water. What is the concentration of the solution in tube 4?

Remember for serial dilutions we multiply the dilution factors.

Dilution 1 = 1 / (1 + 9) = 1/10

Dilution 2 = 1 / (1 + 19) = 1/20

Dilution 3 = 1 /(1 + 4) = 1/5

Total dilution = 1/10 × 1/20 × 1/5 = 1/1,000

Concentration of tube 4 = 2 M × 1/1,000 = 0.002 M = 2 mM

8.3 A focus on diluting solutions

In molecular cloning, bacteria are often transformed (genetically modified) to be resistant to certain antibiotics. This allows you to grow the transformed bacteria in selective media (which contains the antibiotic) and prevent the growth of untransformed bacteria and any other microorganisms.

If you need to prepare 150 ml of Luria-Bertani (LB) agar medium containing 25 µg ml^{-1} ampicillin, how much 50 mg ml^{-1} ampicillin stock solution would you need to add?

Diluting Solutions

Khizar Iftikar is a graduating student studying biomedicine. He faced the problem above in his final year practical subject. Watch an interview with Khizar to find out how he solved it and how confidence in being able to do these kinds of calculations can help you in your later studies – and beyond!

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1053

8.4 Practice problems
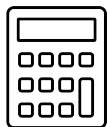
1. What volume of 2 M NaCl would you need to prepare 400 ml of a 0.3 M solution?
2. What volume of 2 M NaCl would you need to prepare 100 µl of an 800 mM solution?
3. What is the final concentration of a solution where 50 µl of a 0.5 M solution is added to 825 µl?
4. To make 100 ml of a 0.1 M solution of HCl how much concentrated HCL do you need? (Concentrated HCL is 11.6 M.)
5. If you have 250 ml of 1.25 M NaCl, how many ml of 0.25 M NaCl can you make?
6. You boil a 450 ml solution of 0.2 M KCl until the volume is 225 ml. What is the final concentration of KCl?
7. What is the dilution factor when 0.25 ml of aqueous solution is added to 38.75 ml of water?
8. You have 2 l of 0.25 M Na$_2$CO$_3$. What volume is required to make 100 ml of 0.1 M Na$_2$CO$_3$?
9. You have a BSA solution at 10 mg ml^{-1}. You remove an aliquot of 50 µl and mix it with 350 µl

of water. You then remove 25 µl of this dilution and add it to 1.25 ml of water. What is the BSA concentration of this solution? Give your answer in µg ml^{-1}.

10. You are making a PCR reaction with a final volume of 50 µl. How much 10x PCR buffer do you need to add to the reaction? (Hint: assume the final concentration should be 1x)

11. How much water must you add to a 1 ml protein sample with a concentration of 0.5 µg ml^{-1} to bring about a 10-fold dilution?

12. In performing a blood test for a crucial analyte, you dilute 100 µl of plasma with 300 µl of distilled water. The test still reads as needing further dilution as it is outside the range of your standard curve. You then take 100 µl of the first dilution and add 300 µl of distilled water. What number to you need to multiply your result by to determine the concentration of the analyte in the plasma?

13. You purify DNA from *E. coli* and obtain 100 µl of DNA at 0.75 mg ml^{-1}. You want to amplify a fragment of DNA by PCR. You need to add 50 pg of plasmid DNA to your PCR reaction.

    a. How much of the original solution do you need to add?
    b. Is this practical given the minimum volume you can accurately pipette is 1 µl?
    c. Work out a serial dilution scheme using 1 ml tubes. Your PCR reaction volume is 25 µl and we want to keep the added DNA to 10% or below of the total volume.

Solution to Practice Problem 8.3.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1055

Solution to Practice Problem 8.6.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1055

Solution to Practice Problem 8.9.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1055

Solution to Practice Problem 8.10.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1055

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1055

32

8.5 Homeopathy

Samuel Hahnemann was a German physician who established the system of medical treatment known as homeopathy. When testing an antimalarial drug, quinine, he observed that the drug produced a malaria-like fever. He concluded that the drug was effective because it produced a similar effect to the illness it was meant to cure. From this he developed the principle of homeopathic medicine, that is, 'like cures like'. The opposite concept is allopathy, which is the treatment of disease by conventional means, that is, with drugs having effects opposite to the symptoms.

Homeopathic dilution involves diluting the homeopathic preparation in alcohol or water. Practitioners believe that during the mixing process (known in homeopathy as succession) the preparation is 'activated' and successive dilutions (serial dilutions) increase the 'potency' of the preparation.

Dilutions of homeopathic products that are sold today usually range from 6X to 30X. This is homeopathy's system for measuring dilution, and it doesn't mean 1 part in 6 or 1 part in 30. X is the roman numeral representing 10. A 6X dilution means 1 part in 10^6, or 1 in 1 million. A 30X dilution means 1 part in 10^{30}, or 1 followed by 30 zeros. A few products are even marketed using the C scale, C being the roman numeral representing 100 – so 30C is 100^{30}!

- Do you think homeopathy works?

Extreme dilutions

Although it remains a popular alternative medicine, the British House of Commons Science and Technology Committee concluded in 2010 that homeopathic treatments fare no better than placebos. The most widely accepted method of testing the effectiveness of a new drug is the **randomised controlled trial**. In this method one group of patients, the **control** group, receives a placebo or standard treatment, and another group of patients receives the drug being tested. To avoid **bias** the trial is conducted **double-blind**. That is, neither the patient nor the practitioner making observations knows if the treatment is the trial drug or the placebo.

8.6 Boffin questions

More about homeopathy

1. In response to the Science and Technology Committee, the British Homeopathic Association concluded that 'the randomised control trial is not an appropriate test for the effectiveness of homeopathic treatments'. What are the implications of this for testing homeopathic treatments?
2. A popular homeopathic preparation is made from the seeds of *Aesculus hippocastanum*. The active compound is aescin and it makes up 30% by weight of the seeds. If a 10 mg ml^{-1} preparation of ground *Aesculus hippocastanum* seeds is diluted 18 times (homeopathy-wise) for a homeopathic preparation, how many molecules of aescin would you expect to find in a 1 ml aliquot? (The molecular weight of aescin is 1,131.26 g mol^{-1}.)

Side Bar – A placebo is a sham treatment or medication which is known to have no therapeutic value e.g. sugar pills. This is an important control in blind trials where one experimental group unknowingly receives the placebo instead of the treatment. There is a well-characterised psychological effect of receiving treatment which often results in patients perceiving a decrease in symptoms unrelated to the treatment itself and some illnesses will spontaneously resolve themselves in the absence of treatment. A change in the placebo group is known as the placebo response and the difference between no treatment and the placebo treatment is known as the placebo effect. It is important to be able to distinguish real therapeutic effects of a medical intervention from the placebo effect or spontaneous remission.

IX

Chapter 9: Medical diagnostics – Measurement, uncertainty and distributions

Blood tests and other clinical diagnostic tests are important for health practitioners to be able to make informed evaluations of patients. Collection and interpretation of this diagnostic data is central to evidence-based medicine. A diagnosis can be thought of as a prediction made by a health practitioner that can be supported by the presence of findings (i.e., certain diagnostic indicators) that occur, by definition, in patients with that diagnosis and not in others. If the diagnosis is to have some value to the patient, then the treatment, advice or intervention associated with the diagnosis should provide a benefit. On the other hand, for any diagnosis which is excluded, the patient should not lose any benefit by not being offered any treatment, advice or intervention.

For a diagnosis to be evidence based, there needs to be some sense of its reliability or certainty. That is, we must ask the question, what proportion of times is the diagnosis correct? To answer this, we need to understand types of data used for clinical diagnostics, how variability in data can be measured and the probability of correct diagnosis.

34

9.1 Types of data

To interpret and work with data you need to understand the different kinds of data. Certain statistical measurements can only be applied to specific data types. Broadly, data can be classified as qualitative (categorical) or quantitative (numerical) and within each of these types there are separate subgroups.
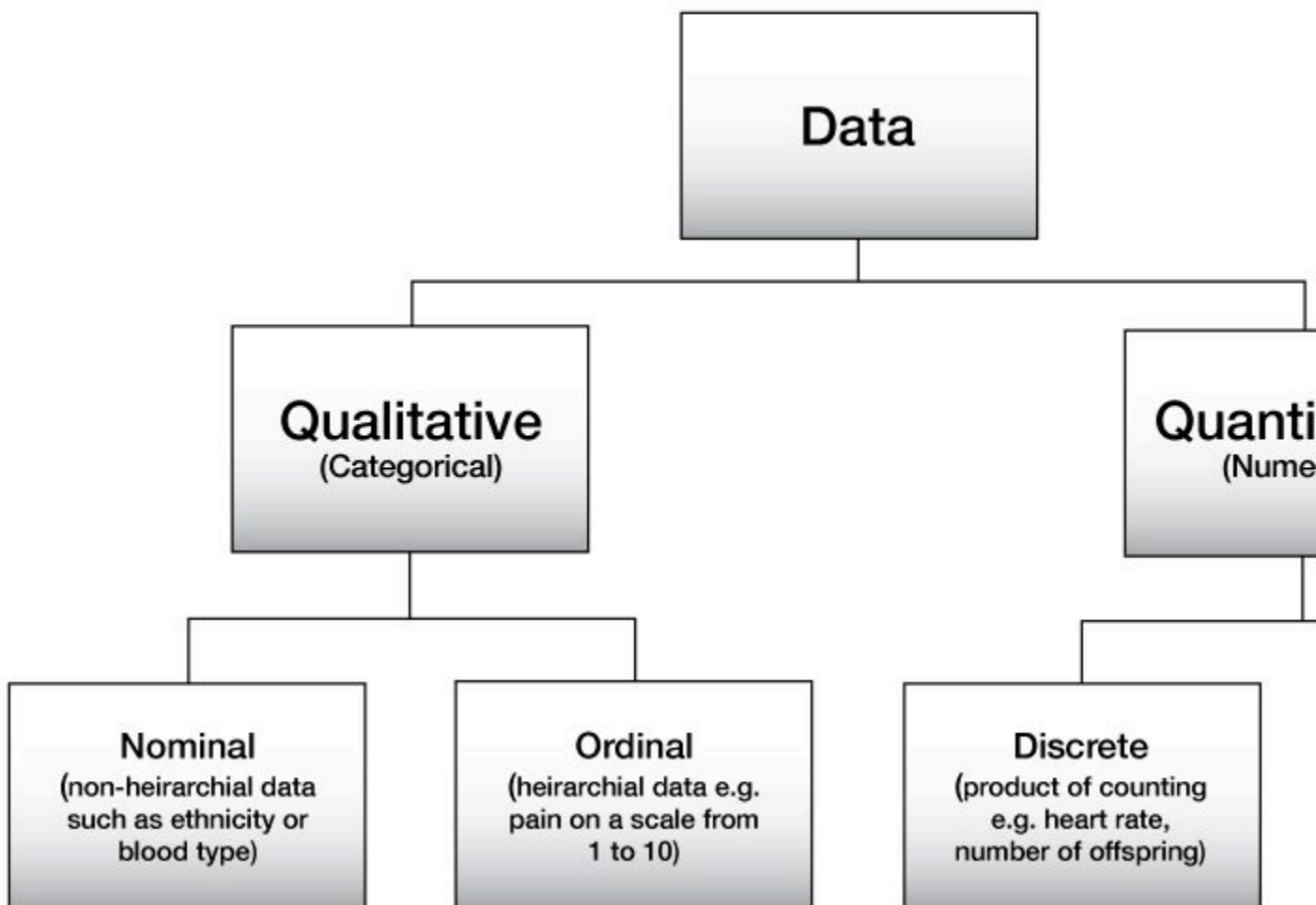
Figure 9.1: Categorizing the different forms of data.

Qualitative (categorical) data

Qualitative data includes such things as survey data. This kind of data can be further classified as nominal and ordinal. Nominal data are just labels and have no order; that is, if you assemble nominal data as a list you would not change the meaning of the values. Examples of nominal data could be patient ethnicity, country of origin or blood type. A further subtype of nominal data is binary data where the value can be one thing or another such as biological sex. Ordinal data on the other hand has a hierarchy and could include variables such as satisfaction ratings, socio-economic status, or perception of pain on a categorial scale.

Quantitative (numerical) data

Quantitative data involves numbers and can be grouped into discrete or continuous data. Discrete data can be counted, like heart rate in beats per minute or number of patients in a clinical trial. Continuous data is measured rather than counted. A characteristic of continuous data is that it can be divided (at least in theory) into infinitely finer parts. Examples of continuous data could be patient weight or serum sodium concentration.

35

9.2 Describing data

Since it is not possible to collect data from the entire population, from everybody, a subset or sample of the population must be made. Generally, the larger the sample the more representative it will be of the population. It is impossible to have zero bias in a sample from a population as no two samples will be exactly the same. However, researchers must take care to **ensure that there is minimal bias in sampling from the population**.

One of the common methods for organising population data is to construct a histogram or frequency distribution. A frequency distribution is an organised tabulation or graphical representation of the number of individuals in each category on the scale of measurement.

There are four important ways to describe frequency distributions:

- measures of central tendency (mean, median, mode)
- measures of dispersion (range, variance, standard deviation)
- the extent of symmetry/asymmetry (skewness)
- the flatness or 'peakedness' of the distribution (kurtosis).

The frequency distribution can be used statistically to help determine a reference range for the data or reference ranges can also be determined non-statistically. The non-statistical approach relies on reference intervals for an analyte being determined by a consensus of medical experts based on the results of clinical outcome studies. (see **Determining what is normal**)

Sometimes other factors can influence the frequency distribution of an analyte and then the reference range used for a patient must also consider that factor (e.g. age, sex).

36

9.3 Using data to make a diagnosis

Blood tests and other diagnostics are important for health practitioners to be able to make informed evaluations of patients. Collection of this data is central to evidence-based medicine.

So, imagine you have a blood test.

	Time:14:44		Units	Ref Range
	------------	------------	----------	-----------
SERUM/PLASMA				
Sodium	**128L**		mmol/L	135-145
Potassium	**6.7H**		mmol/L	3.5-5.5
Chloride	99		mmol/L	95-110
HCO3	**15L**		mmol/L	22-30
Creatinine	**260H**		µmol/L	50-90
eGFR	18			SEE-BELOW
Urea	**31.2H**		mmol/L	2.5-8.3
Glucose			mmol/L	3.3-7.7
Calcium	**1.87L**		mmol/L	2.10-2.60
Phosphate	**2.3H**		mmol/L	0.8-1.5
Magnesium	**1.45H**		mmol/L	0.70-1.30
Albumin	**31L**		g/L	35-50
AP			IU/L	<120
GGT			IU/L	<50
ALT			IU/L	<55
AST			IU/L	<40
Bili Total			µmol/L	<19
T. Protein			g/L	60-82
LD	**678H**		IU/L	210-420
Urate	**1.78H**		mmol/L	0.15-0.40
Lactate	0.5		mmol/L	0.2-1.8

Figure 9.2: An example of the typical data output from a whole blood exam. The values for the different analytes are provided and are also annotated either "L" for being lower than the corresponding reference range or "H" for being higher than the corresponding reference range. Note the units of concentration are also provided given. IU/L indicates international units per litre. An international unit is an arbitrary amount of a substance agreed upon by scientists and doctors

How does your clinician interpret this data? They will look at the concentration of the analytes on the left and compare them to the range of values or threshold on the right, which represent the 'normal' values. Any analyte outside the range is potentially indicative of a problem or abnormality and will help form a diagnosis. There are two important questions to consider here:

- **How do we know if a measurement is normal?** How are normal ranges of data and threshold determined? Data collected from patients will exhibit variability from patient to patient, so it is important to understand the amount of variability there is in this data across the population to know if a measure is likely to be normal or abnormal.
- **How reliable are the measurements?** How do we know the data collected for an individual patient is reliable?

37

9.4 Determining what is normal

Reference ranges provide the values to which a health practitioner compares the test results to determine health status. Values which fall outside the reference range for a particular test are considered abnormal. For example, in the blood test above any sodium concentration below 135 mM or above 145 mM is outside the reference range. For some analytes, the reference range is defined as 'less than' or 'greater than' a certain value.

Reference ranges

Reference ranges (intervals) for some analytes are determined by a consensus of medical experts based on clinical outcome studies. For example, the American Diabetes Association has decreed (based on clinical studies) a fasting plasma glucose concentration of < 100 mg/dl as non-diabetic, 100–125 mg/dl as prediabetic and ≥ 126 mg/dl as diabetic.

However, most reference ranges are determined statistically based on data collected from a number of people and observing what appears to be 'typical' for them. Therefore, the sample should demographically match the population whose tests will be compared to the resultant reference range.

Let's look at an example of how a reference range might be established. One diagnostic test for pulmonary function is forced expiratory volume (in litres) in 1 second (called FEV1). This is the volume of air you can force out of your lungs in 1 second and is measured by forcefully breathing into a mouthpiece connected to a spirometer (an instrument which records the volume of breath).

The data below represent FEV1 tests from a number of individuals, in this case 57 male, biomedical science students, ranked from lowest to highest. As these measurements have been obtained from 57 different individuals, the sample size is 57.

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43

| 3.04 | 3.39 | 3.57 | 3.75 | 4.08 | 4.20 | 4.47 | 4.70 | 5.00 |

| 3.10 | 3.42 | 3.60 | 3.78 | 4.10 | 4.30 | 4.47 | 4.71 | 5.10 |

| 3.10 | 3.48 | 3.60 | 3.83 | 4.14 | 4.30 | 4.50 | 4.78 | 5.10 |

Even when data is organised, like this is, it can be difficult to interpret. If we want to use this data to determine a reference range than we need to understand the variation in the dataset and understand the central position of the data (measure of central tendency).

There are three ways to think about the central tendency or 'average' of a dataset. The most common way is by determining the mean of the data. This is the sum of all the values divided by the sample size.

$$\text{Mean (sample)} = \bar{x} = \frac{(\sum x_i)}{n}$$

where

$\bar{x}$ = mean of the sample

$\sum$ = summation (add up)

x_i = all of the values

n = sample size.

In the case of the FEV1 data, the mean is 4.06.

Rather than simply listing the data as above it can help to display the data points graphically.

Forced Expiratory Volume (litres)
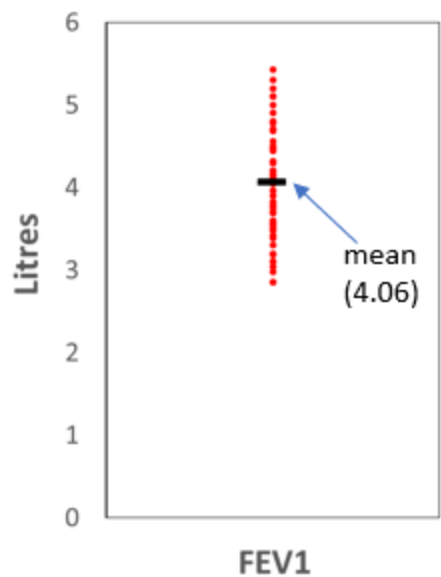
Figure 9.3: A plot showing all 57 individual FEV1 (forced expiratory volume in 1 second) data points. This type of plot is useful for showing the spread of the data and the mean has been indicated but it is difficult to see the other measures of central tendency, median and mode.

While this is somewhat useful it does not clearly indicate the other measures of central tendency: the median and mode. The median is the middle observation if the data are arranged in order. In the FEV1 data, the median data point is 4.10. The mode is the data point which occurs most often, in this case 4.47 occurs three times in the data.

These other measures of central tendency are useful particularly when compared to the mean as they provide an insight into how the data is distributed. For example, is the data spread evenly on either side of the mean or is most of the data bigger or smaller than the mean?

A common method for organising this kind of data to help answer these questions is to construct a histogram or frequency distribution table. A frequency distribution is an organised tabulation of the number of individuals in each category on the scale of measurement.

The following table shows the possible ranges of values for FEV1 in the left column and the number of data points in that range in the right column.

Range	Frequency
2-2.5	0

2.5-3	3
3-3.5	10
3.5-4	13
4-4.5	17
4.5-5	9
5-5.5	5
5.5-6	0

To make it easier to interpret, this table can then be graphed as a histogram.
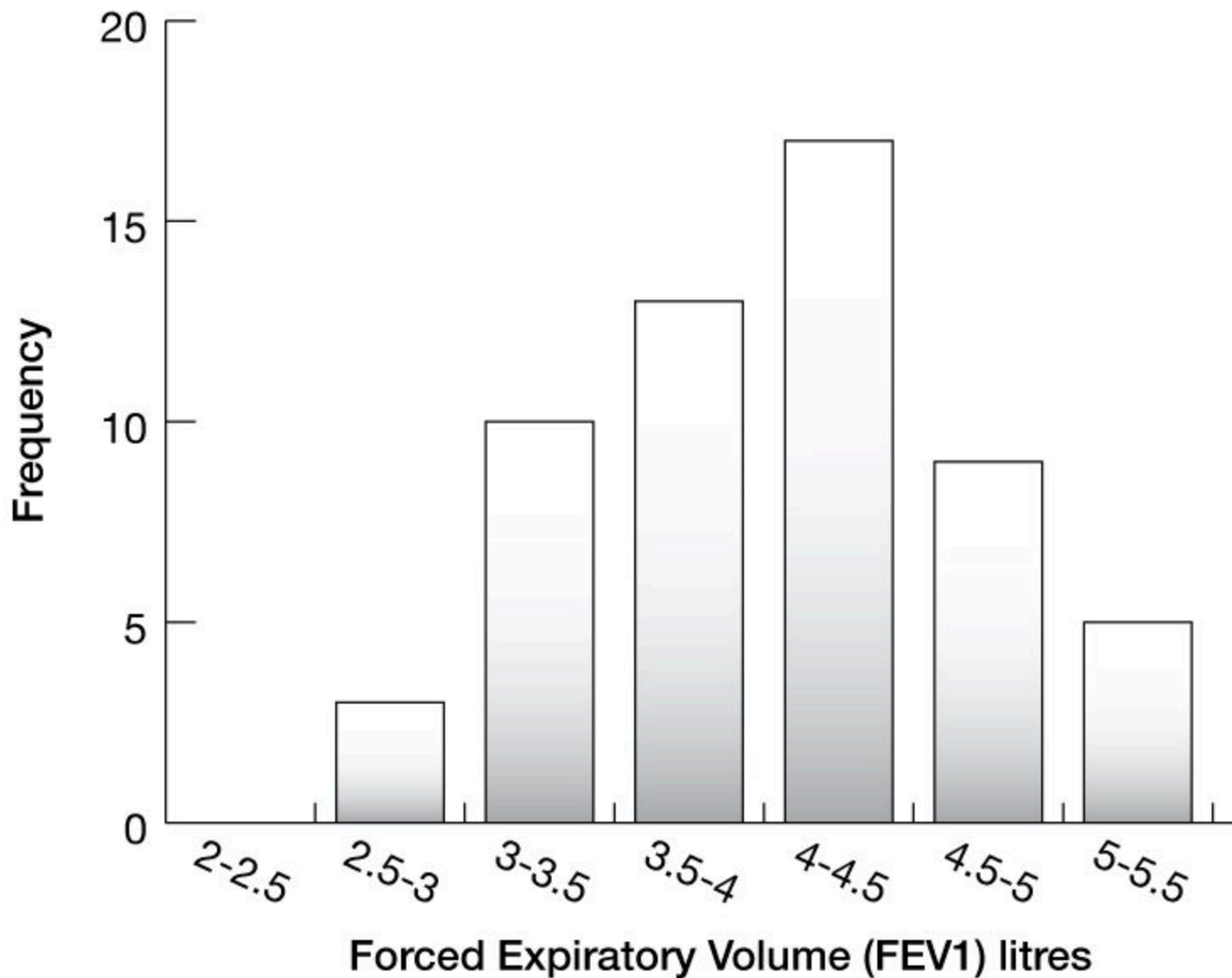
Figure 9.4: A histogram showing the distribution of the FEV1 (forced expiratory volume in 1 second) data. This type
useful for determining whether or not the data conforms to a normal distribution.

We can now observe some critical features of this data. The mean (4.06) is very close to the median
(4.10) and importantly both are close to the mode, as most measurements fall into the 4–4.5 category.
You can also see that roughly 50% of the data points fall on either side of the mean and that there is a
single peak (mode).

These are the characteristics of a normal distribution. To understand how this distribution can be used to
create reference ranges we first need to understand how to describe variations in data.

Describing data variation

The simplest way to look at variation is to look at the range (the minimum and maximum values). In this
case the range of the data is 2.85–5.43 l. But it is important to note that the range does not provide any

information as to how the data is arranged between these extreme values. The range is also likely to increase as sample size increases and can be sensitive to outliers.

The interquartile range is another measure of variability but based on dividing a dataset into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The interquartile range is the difference between the upper quartile and lower quartile. In the set of data above this is 3.54–4.50. The interquartile range indicates the spread of the middle 50% of the data and has an advantage over the range in that it is less sensitive than range to sample size and outliers. It still has a disadvantage in that it does not use all of the information in the data and quite often the tail ends of the data are important.

Variance – often denoted as σ^2 (population) or s^2 (sample) – is a more useful measure of variation as it uses all of the data points. Variance is calculated as the 'average' squared deviation from the mean. That is, the sum of every measurement's deviation from the mean divided by the number of measurements.

Variance (population) $= \sigma^2 = \frac{\sum (x-\mu)^2}{N}$

where

   N = population size

   μ = population mean

   x = each measure.

Variance (sample) $= s^2 = \frac{\sum (x-\tilde{x})^2}{n-1}$

where

   n = sample size

   $\bar{x}$ = sample mean

   x = each measure.

As most datasets involve a sample of the population rather than the entire population the sample variance will be used much more frequently. Even though variance uses all of the data it is measured in the original units squared and is therefore considerably affected by extreme values or outliers.

One thing you might be curious about is why square the deviations from the mean. The reason is that if you just add up the differences from the mean then the negatives will cancel the positives. A reasonable solution to this would be to use the absolute values of the differences (i.e. ignore the sign) but the problem here is that the averaging effect will still diminish the effect of extreme values.

The best way to solve this problem is to take the square root of the variance. This value is known as the standard deviation.

Standard deviation (population) $= \sigma = \sqrt{\sigma^2}$

Standard deviation (sample) = s = $\sqrt{s^2}$

The standard deviation still provides a sense of how the data is spread out from the mean but unlike variance is in the same units as the measurement. Simply put, if the standard deviation is small, the data values are close to the mean value. If the standard deviation is high, the data values are widely spread out from the mean value.

Standard error of the mean

Standard deviation is useful when we want to indicate how widely scattered the measurements are. Though you will frequently see the **standard error of the mean** used to report variability in a set of data, a standard error is really a measure of how precise an estimate is. So, the standard error of the mean is an estimate of how precise the mean is.

A formal definition of the standard error of the mean is that it is the standard deviation of many sample means of the same sample size drawn from the same population.

But most of the time we only calculate one mean for a set of data, not multiple means. Therefore, unlike the standard deviation of the measurements, the standard error of the mean is an estimate rather than a measurement. We refer to it as an inferential rather than a descriptive statistic.

The standard error of the mean is calculated by the following formula:

Standard error of the mean = $\frac{s}{\sqrt{n}}$

where

      n = sample size

      s = sample standard deviation.

This formula shows us that as n, the sample size, increases, the standard error of the mean decreases. This makes sense – the more data we have, the more precise will be our estimate of the mean. The standard error of the mean gets larger as the data exhibits more variability and standard deviation increases.

A real point of confusion is the nomenclature around the standard error of the mean. It is commonly (though imprecisely) shortened to just standard error. It is also abbreviated to SEM or sometimes SE. Standard deviation and standard error of the mean are often incorrectly interchanged.

The standard error is most useful as a means of calculating a confidence interval. For a large sample, a 95% confidence interval is obtained as the values 1.96 times the standard error of the mean either side of the mean.

Frequency distributions

As we saw earlier, there are four important ways to describe frequency distributions: measures of central tendency (mean, median, mode); measures of dispersion (range, variance, standard deviation); symmetry/asymmetry (skewness); and the flatness or 'peakedness' of the distribution (kurtosis).

In a normal distribution, the mean, median and mode are all approximately the same. That is, there is symmetry about the centre with about 50% of values less than the mean and 50% greater than the mean. Importantly there is only one peak (mode). Data where the mean is greater than the mode and median is positively skewed and data where the mean is less than the median and mode is negatively skewed. Some data which follow a non-normal distribution may be bimodal or even multi-modal, which means it contains more than one peak.
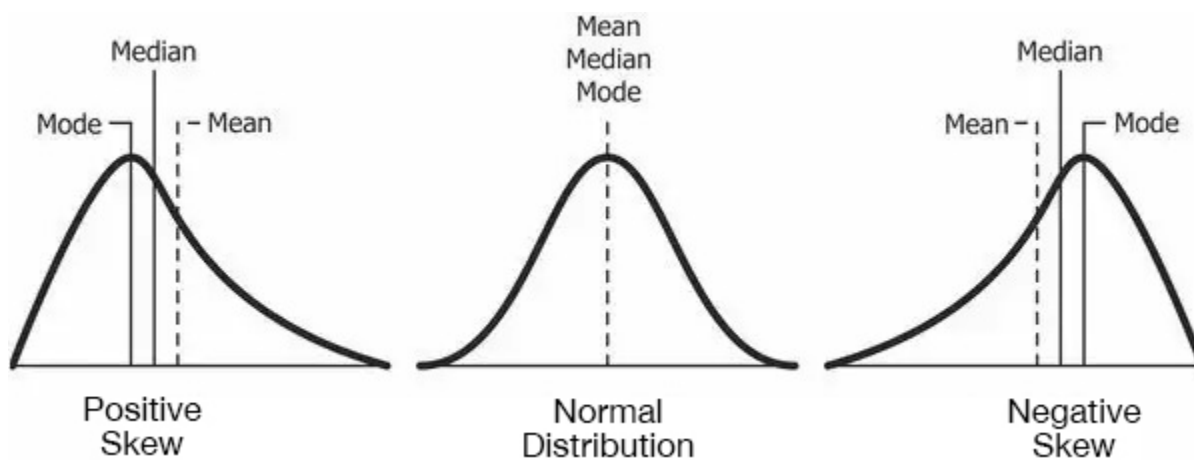
Figure 9.5: Positive Skew, Normal Distribution, Negative Skew (Adapted from 'Relationship between mean and median under different skewness' by Diva Jain from Wikimedia Commons used under CC BY-SA 4.0)

A normal distribution is a common probability distribution. It is often referred to as a 'bell curve' and many datasets follow this distribution. In fact, as you increase the number of samples taken from a population, the more the resulting distribution will resemble a normal distribution.

This is a tenet of the central limit theorem, which states that **given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population**.

What this means in practice is that you might perform a study where, for instance, you measure the FEV1 from a sample of individuals from the general population, and calculate the mean of that one sample. Now, imagine that you repeat the study many times and collect the same sample size for each. You then calculate the mean for each of these samples and plot them on a histogram. Given enough studies, the histogram will display a normal distribution of these sample means. This also means that the sample mean will start to approximate the population mean as sample size increases. So, in this way, we can use a sample distribution to make inferences about the population it is taken from.

But how large does the sample size have to be for the data to approximate a normal distribution? It

depends on the shape of the variable's distribution in the underlying population. The more the population distribution differs from being normal, the larger the sample size must be. Fortunately, many parameters will have an underlying normal distribution in the population and if not, care can be taken to sample from a normally distributed subset. In the case of FEV1 it would make sense to sample from either males or females as sampling from the entire population would result in a non-normal bimodal distribution. A rule of thumb agreed on by statisticians is that a sample size of 30 is sufficient for most variables.

So, the frequency distribution can be used to help determine a reference range for the data. That is, we can use the distribution to decide whether a particular measure can be considered 'abnormal'.

One property of normally distributed data is that 95% of values are approximately within two standard deviations of the mean. This can be best understood by looking at a standard normal distribution with a mean of 0 and a standard deviation of 1.
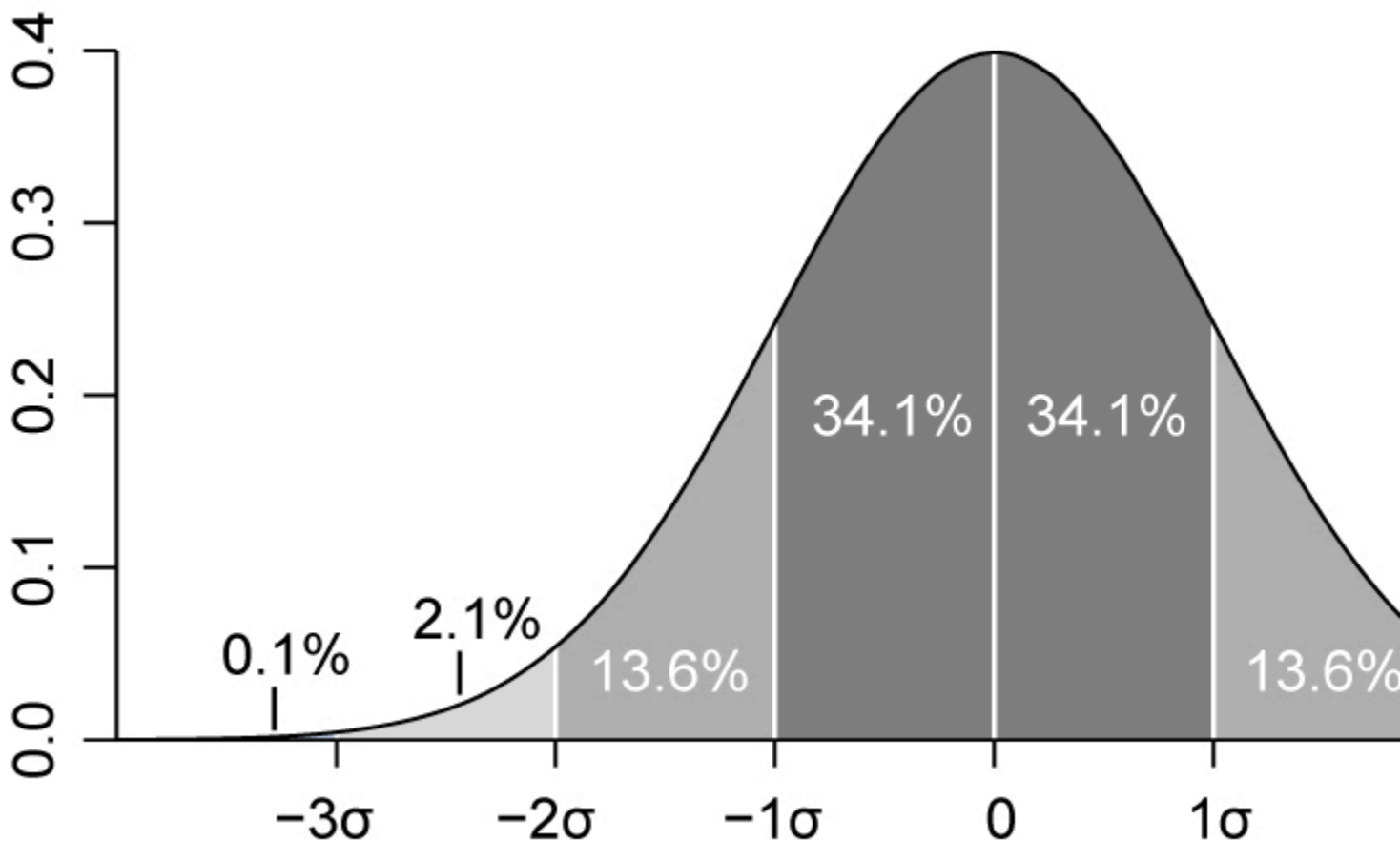
Figure 9.6: Standard deviation from the mean (Adapted from 'Standard deviation diagram' by M. W. Toews from W 2.5)

If you look at the area under the curve, the portion of the curve covered by the region 2 standard deviations either side of the mean is about 95%. You will also note that 68% of the curve is covered by

the region 1 standard deviation either side of the mean and over 99% of the curve is covered by the region 3 standard deviations either side of the mean. This is useful to know because it tells us that any value is:

- likely to be within 1 standard deviation of the mean (68 out of 100)
- very likely to be within 2 standard deviations of the mean (95 out of 100)
- almost certain to be within 3 standard deviations of the mean (99 out of 100).

This value of ±2 standard deviations covering 95% of the standard normal distribution either side of the mean is known as the Z score. Note that ±2 is an approximation, and the correct number is really ±1.96.

One way to state this is:

$$p\,(-1.96 < Z < 1.96) = 0.95$$

That is, there is a 95% probability that a standard normal variable, Z, will fall between −1.96 and 1.96. This is known as the 95% confidence interval. But real sample data will not be the same as the standard normal distribution (mean of 0 and a standard deviation of 1), so how can we calculate this for other data?

We can use the following formula:

$$95\% \text{ confidence interval} = \bar{x} \pm 1.96\,\frac{s}{\sqrt{n}}$$

where

n = sample size

$\bar{x}$ = sample mean

s = sample standard deviation.

This works when the sample size is greater than 30 and the underlying population distribution is normal.

You will notice that this formula uses the standard error of the mean $\frac{s}{\sqrt{n}}$. This is because when we are calculating confidence intervals, we are really trying to indicate the uncertainty around the estimate of the mean.

Confidence intervals for smaller sample sizes

For sample sizes less than 30 ($n < 30$), the central limit theorem does not apply. So, another distribution, the t distribution, is used instead. The t distribution is like the standard normal distribution but takes a slightly different shape depending on the sample size. So, instead of 'Z' values, there are 't' values for confidence intervals, which are larger for smaller samples, producing larger margins of error, because small samples are less precise.

The confidence interval estimate works in a similar way:

$$\text{Confidence interval} = \bar{x} \pm t\,\frac{\sigma}{\sqrt{n}}$$

where

n = sample size

$\bar{x}$ = sample mean

s = sample standard deviation

t = t-value.

We can look up the appropriate t-value (from a t-distribution table) based on the desired confidence interval. t-values are listed by degrees of freedom which are equal to n − 1 (see the next boffin section for more on this). Just as with large samples, the t distribution assumes that the outcome of interest is approximately normally distributed.

Example: 95% confidence intervals

Calculate the 95% confidence interval for serum creatine levels based on a study of 11 participants. The sample mean was 72 µM and the standard deviation was 8.5.

The first step is to look up the appropriate t-value. Since we are looking for a 95% confidence interval, this means that the tails outside this confidence interval will be 5% of the total area under the curve. Therefore, each tail will be 2.5%. This value, converted into a decimal, is known as α. In this case α = 0.025.

$\alpha/2$

$1 - \alpha$

$-t_{n-1}(\alpha/2)$

0
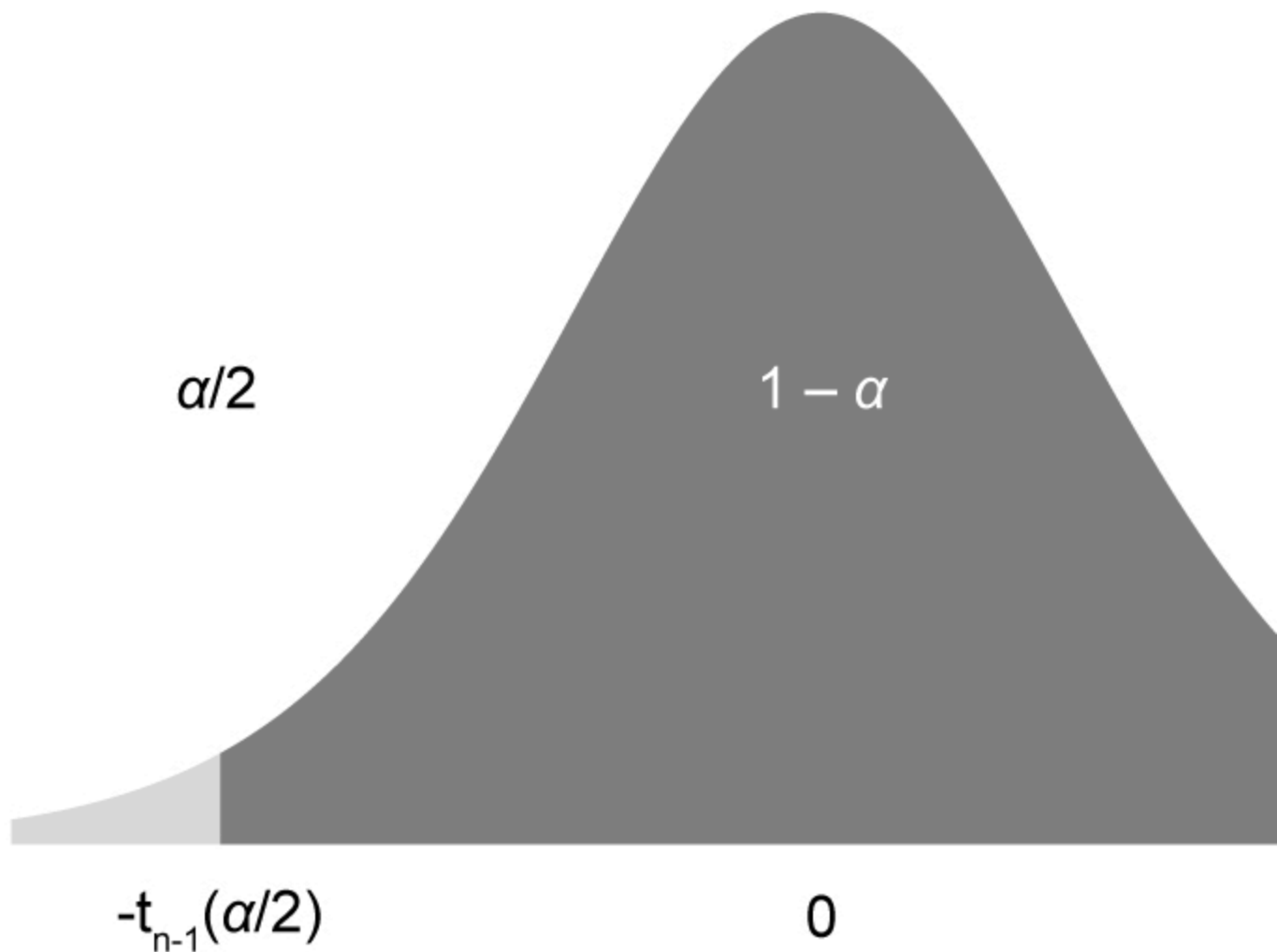
Figure 9.7: Confidence interval ('Confidence interval' by ARAKI Satoru from <u>Wikimedia Commons</u> used under <u>CC</u>

The next step is to determine the degrees of freedom (df). Since $n = 11$, the degrees of freedom are equal to 10 ($n - 1$).

You can then look up the appropriate t-value on a t-distribution table.

α / df	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073

Figure 9.8: Determining the t-value from a t-distribution table.

By cross-referencing α (0.025) and the degrees of freedom (10) we can find the t-value of 2.228 (Figure 9.8).

This can now be substituted into the equation:

$$\text{95\% confidence interval} = \bar{x} \pm t\,\frac{\sigma}{\sqrt{n}}$$

$$\text{95\% confidence interval} = 72 \pm 2.228\,\frac{8.5}{\sqrt{11}}$$

$$\text{95\% confidence interval} = 66.3,\ 77.7$$

Reference ranges and false positives

In biomedicine a 95% confidence interval is most often used for determining a reference interval, in turn to decide whether a diagnostic result is atypical. But remember, 95% is an arbitrary value and given that the primary goal of diagnostic testing is to differentiate 'healthy' from 'non-healthy' individuals, other confidence levels can be selected.

One consequence of using this statistical approach for determining reference ranges for diagnostic testing is that even in a 'healthy' population, a single test result with a 95% reference range or confidence interval is outside the reference range in 5% (or 1 in 20) of cases. These cases are called

false positives. The proportion of false positives may be even greater when the patient population is not closely matched to the control subjects with respect to age, sex, ethnic group, and other factors.

Why are degrees of freedom n – 1?

Degrees of freedom indicates how much independent information goes into a parameter estimate. This can be difficult to understand but a useful analogy is this:

Imagine a series of numbers 5, 4, 3, x where we know the mean is 5.

This means that x must equal 8. As you can see, the last number, x, has no freedom to vary. It is not an independent piece of information because it cannot be any other value. Estimating a parameter such as the mean imposes a constraint on the freedom to vary. The last value and the mean are entirely dependent on each other. So, when calculating the distribution of the t statistic, which is dependent on the sample mean, the degrees of freedom (df) are equal to the sample size minus 1 (n-1).

Sample bias

Sometimes other factors can influence the frequency distribution of an analyte and then the reference range used for a patient must consider that factor (e.g. age, sex). A clear indicator that there might be some sort of sample bias is a non-normal distribution.

With our example of FEV1 measurements in male biomedical students, a negative skew may result if we included in our sample group another 10 students who were all endurance athletes who had atypically high pulmonary capacity. Therefore, care needs to be taken to avoid sample bias when selecting individuals from a population. **The resulting reference range will only be applicable to the cohort used to generate it.**

A bimodal distribution might indicate some other factor that needs to be considered. For example, if we measured FEV1 for 57 male and 57 female biomedical students, it is likely we would end up with a bimodal distribution just due to inherent physiological differences between males and females. It would be more appropriate to separate the data by sex and derive male- and female-specific reference ranges.

38

9.5 Measurement reliability

Diagnostic tests must be accurate and precise. Accuracy represents how closely the test provides a value which is close to the true value for the patient. Precision tells us how repeatable a test result is. That is, if

we repeat the test, under the same conditions using the same samples, we should get similar results.

For example, in the blood test above, how do we know that the value obtained for the measurement of sodium concentration in serum is actually the concentration of sodium in the patient's serum? And if we were to take blood tests multiple times from the same patient how close would the measurements for serum concentration be to each other?

For determining accuracy, how could we know what the true value is? One way this is done in biomedical diagnostics is to compare the results of the test with results obtained using another method (or methods). Quite often these other methods are 'gold standard' techniques which are known to be highly accurate and precise for a certain measurement but are not widely used because they are less convenient or more expensive. The other alternative to estimate the accuracy of a test is to use a known standard in the measurement. For example, in a blood test where the concentration of a particular analyte is being measured then a measurement could also be made of serum which has had a known amount of the analyte added. In this way a standard has been produced to provide a theoretically true value.

How accurate a measurement is can be described by the percentage error of the measurement. This is given by the formula:

$$Percentage\ error\ (\%) = \frac{measured\ value\ -\ true\ value}{true\ value} \times 100$$

You will notice from this formula that it is possible to get a negative error depending on the magnitude of the measured value. This can be useful to know but quite often percentage error is calculated and reported using the absolute value:

$$Percentage\ error\ (\%) = \left|\frac{measured\ value\ -\ true\ value}{true\ value}\right| \times 100$$

Also, it is worth noting that you will occasionally see the same information expressed as relative percentage accuracy:

$$Relative\ accuracy\ (\%) = \left|\frac{true\ value\ -\ (true\ value\ -\ measured\ value)}{true\ value}\right| \times 100$$

Precision provides us with an indication of how repeatable or reliable our measurement is. If we only take one measurement, the reliability or certainty surrounding that measurement is zero – we have no idea how a precise the measurement is. So, in order to determine precision, we need to take multiple measurements and then need a way of expressing how close the measured values are to each other.

This is known as variability and there are several ways of expressing variability (see the earlier section on describing variations in data). The most useful measure of variability in most cases is the standard deviation. So, the magnitude of the standard deviation for a set of repeated measures provides an indication of the precision of the measurement. The higher the standard deviation the greater the variability of the data and therefore the less precise the measurement. Once we have an estimate of the precision of a particular measurement, we have an estimate of how reliable our data is.

What if we want to compare the precision of two different datasets, perhaps from two different measurement methods? It is not reasonable to simply compare standard deviations as the magnitude of the standard deviation changes with the mean. (Recall that standard deviation is calculated by determining the square root of the variance which in turn is calculated by determining the average of the squared differences between the measurements from the mean.) Therefore, we need a way to normalise

standard deviation against the mean to be able to make comparisons. The way to do this is to use the coefficient of variation.

The coefficient of variation is calculated using the following formula:

$$cv = \frac{s}{\bar{x}} \times 100$$

where

cv = coefficient of variation

s = sample standard deviation

$\bar{x}$ = mean.

With the coefficient of variation you can compare the precision of two diagnostic tests or compare the precision of the high and low ranges for a particular diagnostic test.

For example, there are two methods for determining glucose concentrations in plasma. One uses the enzyme hexokinase and the other uses the glucose oxidase.

Analyte: Glucose	Analyte: Glucose
Method: Hexokinase	Method: Glucose oxidase
Standard deviation = 4.8	Standard deviation = 4.2
Mean = 120	Mean = 110

Which of these methods is more precise?

Hexokinase:

$$cv = \frac{4.8}{120} \times 100 = 4.0\%$$

Glucose oxidase:

$$cv = \frac{4.2}{110} \times 100 = 3.8\%$$

The glucose oxidase assay is more precise.

9.6 Practice problems
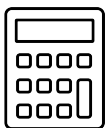
1. Determine the mean, mode and range of the following dataset:
   20, 20, 25, 19, 17, 18, 17, 22, 23, 17, 23

2. As part of a diagnostic test for levels of serum cortisol two standards were tested, a high concentration standard and a low concentration standard. The following results were obtained.

	High	**Low**
Mean	1.005	0.104
Standard deviation	0.051	0.006

   Is the assay more precise at the lower control or the higher standard concentration?

3. Pretend you are in charge of determining a reference interval for the general population for a new discovered blood analyte. Abnormally high or low levels of this analyte could be an indicator of kidney dysfunction. How would you go about selecting individuals to test? What factors would you need to consider in selecting them?

▶

Solution to Practice Problem 9.2.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1074

What do we do if measurements are inaccurate or imprecise?

Before this can be answered we need to consider, what is an acceptable level of accuracy and precision? There is no simple answer as this must be defined for each variable or analyte being measured. In a biomedical context this is normally based on medical significance.

A rule of thumb for many diagnostic assays is that the precision should be equal to or less than half of the 'within subject' biological variation. Ultimately the value of any medical diagnostic procedure or test is determined by how well it discriminates between the two conditions of interest (health and disease; two stages of a disease etc.)

Since it is sometimes difficult to know the true value for something being measured, precision is often used a proxy for accuracy. That is, a number of measurements are made under the same conditions that are in 'good' agreement with one another and therefore it is assumed that the measurements are accurate. This is not a correct assumption as it is possible for a measurement to be precise but not accurate. It is also possible for a measurement to be accurate but not precise. Ideally, we want a diagnostic test to be both accurate and precise.

Poor precision for a measurement usually results from poor technique and is associated with 'random errors'. That is, the error (the deviation from the mean) has a random sign and varying magnitude. For example, the technician performing a particular assay uses an automatic pipette which has a precision limitation that introduces random errors. This kind of error can be detected by examining the variability of the results and can be reduced by averaging over many repetitions.

In general, poor accuracy is associated with 'systematic errors'. This kind of error will have a reproducible sign and magnitude. For example, the automatic pipette used in an assay is incorrectly calibrated and consistently dispenses a higher volume than expected. Systematic errors are often more difficult to detect and require either the use of known standards or verification by a different method.

40

9.7 A focus on understanding standard curves

Standard curves (sometimes called calibration curves) are incredibly important in research and medical diagnostics for quantifying an analyte. How do they work and how do we know if the data we get when using them is accurate and precise?

Professor Robyn Murphy

Professor Robyn Murphy is an accomplished researcher who focuses on various aspects of skeletal muscle biochemistry in health and disease. In addition, she is a Deputy Dean and Associate Dean of Learning and Teaching with a passion for science and quantitative literacy education. Watch an interview with Robyn to learn why standard curves are so important and how we know if they are accurate and precise. You will also hear about strategies you can employ to reduce anxiety around and become more proficient in mathematics.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1077

41

9.8 Boffin questions
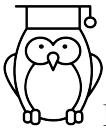
Do ACE inhibitors increase susceptibility to COVID-19?

An angiotensin-converting enzyme, or ACE, is a central component of the renin–angiotensin system. This enzyme controls blood pressure by regulating the volume of fluids in the body. It converts the hormone angiotensin I to the active vasoconstrictor angiotensin II. People with hypertension are prescribed ACE inhibitors. These medications reduce blood pressure by encouraging the blood vessels to relax and open.

One side effect of ACE inhibitors, at least reported in animal models, is that they increase the expression of angiotensin-converting enzyme 2 (ACE2). SARS-CoV-2, the coronavirus which causes COVID-19, enters cells by binding a viral spike protein to ACE2. Therefore, there is concern that patients taking ACE inhibitors may be more susceptible to contracting COVID-19 and more susceptible to worse outcomes.

In the UK, COVID-19 infections were monitored by a COVID-19 symptom tracker app. The app allowed members of the public to contribute to research through self-reporting data including demographics, conditions, medications, symptoms and COVID-19 test results. Researchers observed that people reporting ACE inhibitor use were twice as likely to have a COVID-19 infection based on symptoms, even after adjusting for differences in age, body mass index, sex, diabetes and heart disease.

- Speculate on why this data could be misleading or incorrect.

X

Chapter 10: Medical diagnostics – Sensitivity and specificity

Aside from performing a test, we also need to know if it is diagnostically valid. That is, is it sufficiently sensitive and specific to be clinically useful?

In the perfect diagnostic test for a disease:

- all the people who have a positive test result really would be ill (true positives)
- there would be no positive test results in people who are not ill (false positives)
- all the people with a negative test result would not be ill (true negatives)
- people who are ill would not have a negative test result (false negatives).

Imagine a screening test for early cancer detection. There are 10 people without symptoms. The 2 people highlighted in orange have cancer.

Figure 10.1: The two "orange" individuals have cancer and the eight "blue" individuals do not

With a perfect diagnostic test, all the blue figures would test negative, and all the orange figures would test positive.
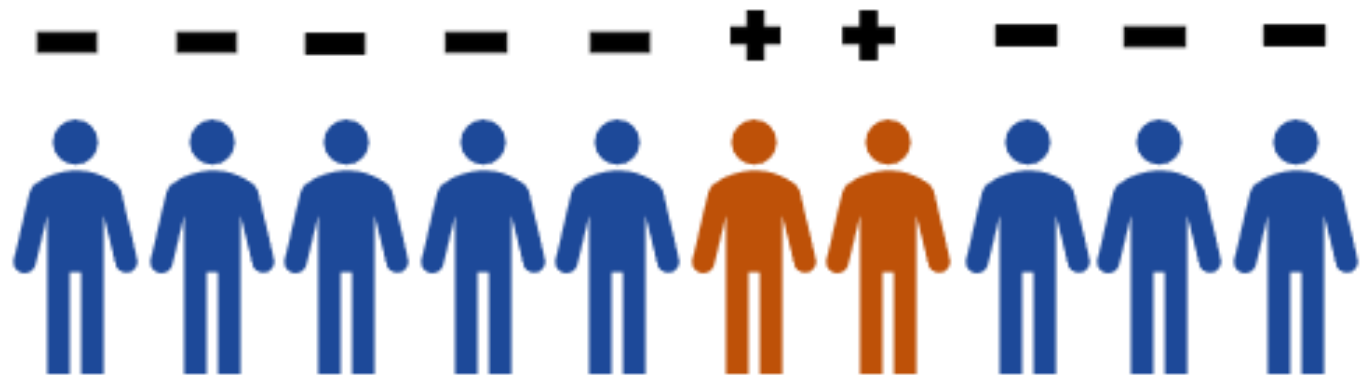
Figure 10.2: A perfect test. The signs indicate the results for a cancer diagnostic. Only individuals with cancer test positive and those without test negative.

However, most tests are not perfect. In the following image the test has given a positive result in only 1 of the people who has cancer and there is a false alarm in 2 of the people who do not have cancer! In this test there are 2 false positives and 1 false negative.
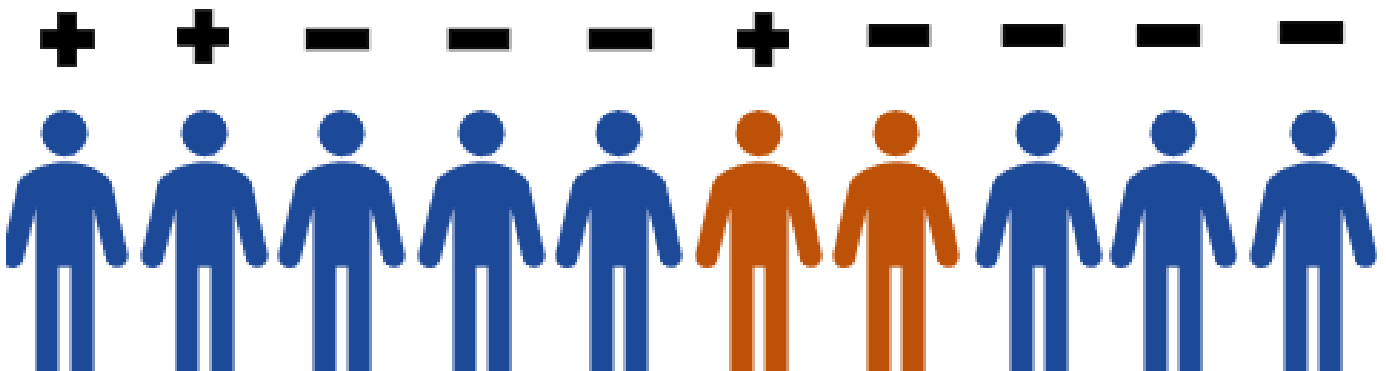
Figure 10.3: An imperfect test. One individual with cancer has tested negative (a false negative) and two without cancer have tested positive (false positives).

10.1 Sensitivity and specificity

The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease. A test with 100% sensitivity correctly identifies all patients with the disease. A test with 80% sensitivity detects 80% of patients with the disease (true positives) but 20% with the disease go undetected (false negatives).The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease. Therefore, a test with 100% specificity correctly identifies all patients without the disease. A test with 80% specificity correctly reports that 80% of patients without the disease test negative (true negatives) but 20% of patients without the disease are incorrectly identified and test positive (false positives).No clinical test is perfect and in general there is an inverse relationship between sensitivity and specificity. Also, quite often there is an overlap between health and disease for the parameter being measured. This makes it difficult to decide where a reference interval for a test should be. For example, in the figure below, setting the reference limit to point A will mean that the test will have good sensitivity but poor specificity, while setting it at point B will mean the specificity is better but sensitivity is poor.
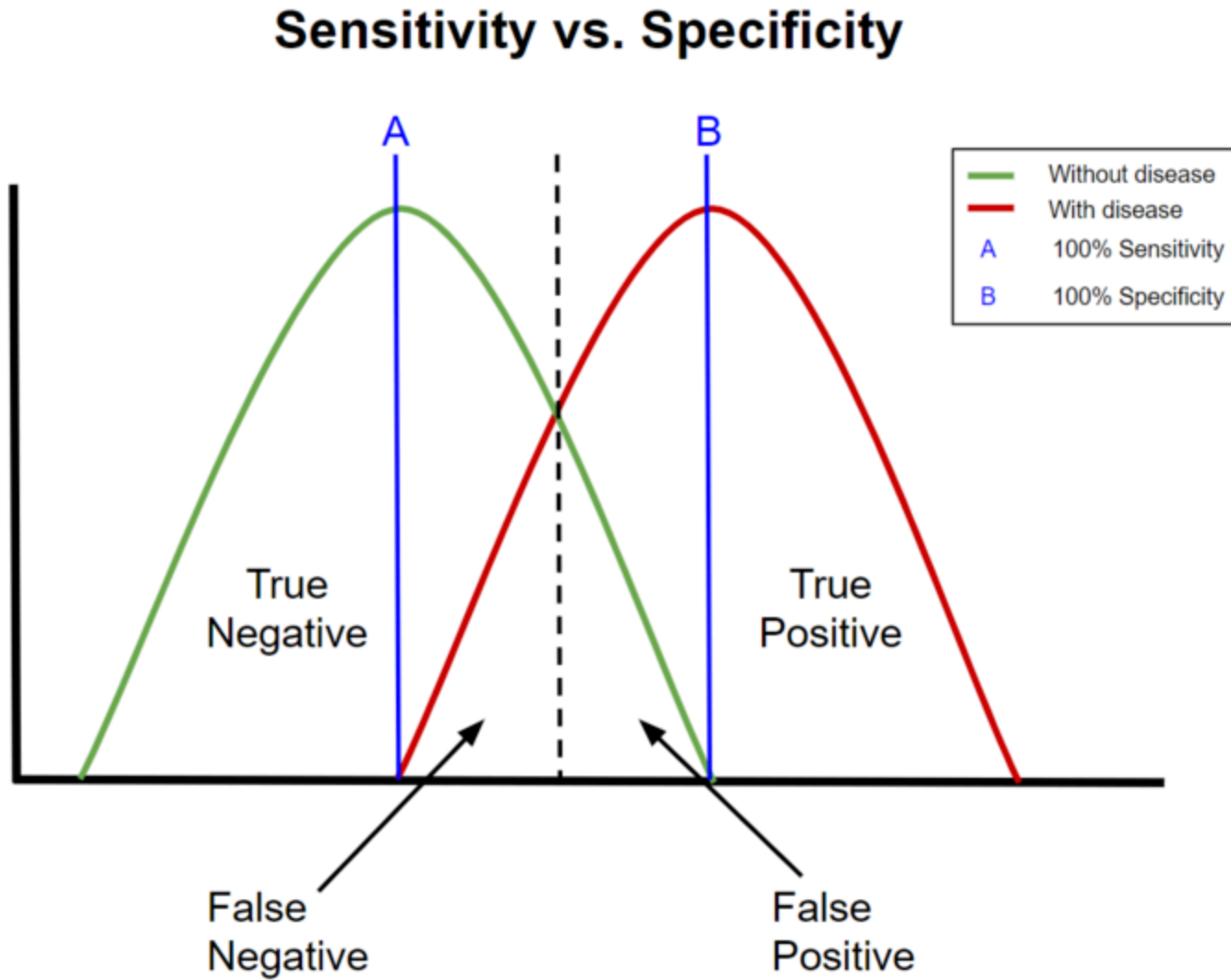
Figure 10.4: Sensitivity vs. Specificity (Adapted from 'Specificity vs Sensitivity Graph' by Blue64701 from Wikimedia Commons used under CC BY-SA 4.0.)

10.2 A focus on medical diagnostics

Imagine a disease with a prevalence of 1 in 1,000. The diagnostic test to detect it has a sensitivity and specificity of 90%. What is the chance that if you test positive, you actually have the disease?

Medical Diagnostics

Julian Pakay is a biochemist and science educator. He has a keen interest in promoting quantitative literacy. Watch an [interview](#) with Julian to learn why we need to be careful interpreting data around diagnostic tests and to see a solution to the question above. You will also learn some tips to minimise mistakes when performing calculations.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1084

We can calculate the sensitivity of a diagnostic test using the formula:

Sensitivity = (true positives / all diseased) × 100

or

Sensitivity = true positives / (true positives + false negatives) × 100

And we can calculate the specificity of a diagnostic test using the formula:

Specificity= (true negatives / all non-diseased) × 100

or

Specificity = true negatives / (true negatives + false positives) × 100

It is important to remember that if a diagnostic test for a disease has a sensitivity and specificity of 99%, and you test positive, the probability that you actually have the disease is not 99%! In fact, the rarer the disease, the lower the probability that a positive result indicates that you have it.

Diagnostic test 'accuracy'

You may see the term 'accuracy' used in the popular media to describe diagnostic tests. This term can be very misleading and should be avoided. For example, imagine a disease which has a high prevalence, for example, 10 out of 100 people suffer from it. A diagnostic test that fails to detect the disease at all would still be correct 90% of the time! It would correctly be negative for the 90% of people who do not have the disease. Therefore, it would be 90% accurate even though useless.In biomedical literature 'accuracy' is usually defined as all the accurate results (i.e. the sum of the true positive and true negatives) divided by the sum of all test results. This means that a diagnostic test for a disease with a very low prevalence could have a high accuracy even if it fails to detect any patients with the disease!So, it is much better to qualify diagnostic tests with sensitivity and specificity but even this can be misleading.

Example 1

A disease has a prevalence of 0.1%. This means that 1 in 1,000 people have it. So, from 100,000 people you would expect 100 people to have it. There is a diagnostic test for the disease with a sensitivity and

specificity of 99%. If you test positive for this disease what is the chance that you have it?

The best way to answer this question is by constructing a table of test results versus actually having the disease or not. In this case we test 100,000 people. We know that the prevalence is 0.1% so we expect that 100 of them will be sick. This leaves 99,900 as healthy.

	Sick	Healthy	Totals
Test positive			
Test negative			
Totals	100	99,900	100,000

If the test has 99% sensitivity, then true positives should make up 99% of the individuals with the disease. That is, 99 should be sick and test positive, which leaves 1 person who is sick but tests negative (false negative).

	Sick	Healthy	Totals
Test positive	99		
Test negative	1		
Totals	100	99,900	100,000

If the test has 99% specificity, then true negatives should make up 99% of all the individuals without the disease. That is, 99% of the 99,900 should be sick and test negative. This is equal to 98,901, which leaves 999 people who are healthy but test positive (false positives).

	Sick	Healthy	Totals
Test positive	99	999	1,098
Test negative	1	98,901	98,902
Totals	100	99,900	100,000

Therefore, if you test positive for this disease, you have a 99/1,098 or 9% chance of having the disease.

In the case above everyone was screened for the disease regardless of having symptoms or not, but most people only get tested for a disease if they have symptoms, are known to be at risk of having the disease or otherwise exhibit an indicator of the disease. In these cases, positive results

will be more likely to be true.

Example 2

A breast cancer diagnostic test is conducted by biopsy. About 30% of women tested have breast cancer. The false positive rate is 2% and the false negative rate is 14%. What is the chance that an individual testing positive has breast cancer?

The way to solve this is again to create a table and put in some theoretical numbers.

If 100,00 people are tested and we expect 30% to have breast cancer then we will have 30,000 with breast cancer, leaving 70,000 without. As the false positive rate is 2%, we expect that 1,400 of the 70,000 without breast cancer will test positive. As the false negative rate is 14%, we expect that 4,200 of the 30,000 with breast cancer will test negative.

	Have breast cancer	**Don't have breast cancer**	**Totals**
Test positive	25,800	1,400	27,200
Test negative	4,200	68,600	72,800
Totals	30,000	70,000	100,000

In this case, the probability that an individual testing positive has breast cancer:

= 25,800/27,200 = 95%

44

10.3 Diagnostic test 'accuracy'

You may see the term 'accuracy' used in the popular media to describe diagnostic tests. This term can be very misleading and should be avoided. For example, imagine a disease which has a high prevalence, for example, 10 out of 100 people suffer from it. A diagnostic test that fails to detect the disease at all would still be correct 90% of the time! It would correctly be negative for the 90% of people who do not have the disease. Therefore, it would be 90% accurate even though useless.In biomedical literature 'accuracy' is usually defined as all the accurate results (i.e. the sum of the true positive and true negatives) divided by the sum of all test results. This means that a diagnostic test for a disease with a very low prevalence could have a high accuracy even if it fails to detect any patients with the disease!So, it is much better to qualify diagnostic tests with sensitivity and specificity but even this can be misleading.

Example 1

A disease has a prevalence of 0.1%. This means that 1 in 1,000 people have it. So, from 100,000 people you would expect 100 people to have it. There is a diagnostic test for the disease with a sensitivity and specificity of 99%. If you test positive for this disease what is the chance that you have it?

The best way to answer this question is by constructing a table of test results versus actually having the disease or not. In this case we test 100,000 people. We know that the prevalence is 0.1% so we expect that 100 of them will be sick. This leaves 99,900 as healthy.

	Sick	Healthy	Totals
Test positive			
Test negative			
Totals	100	99,900	100,000

If the test has 99% sensitivity, then true positives should make up 99% of the individuals with the disease. That is, 99 should be sick and test positive, which leaves 1 person who is sick but tests negative (false negative).

	Sick	Healthy	Totals
Test positive	99		
Test negative	1		
Totals	100	99,900	100,000

If the test has 99% specificity, then true negatives should make up 99% of all the individuals without the disease. That is, 99% of the 99,900 should be sick and test negative. This is equal to 98,901, which leaves 999 people who are healthy but test positive (false positives).

	Sick	Healthy	Totals
Test positive	99	999	1,098
Test negative	1	98,901	98,902
Totals	100	99,900	100,000

Therefore, if you test positive for this disease, you have a 99/1,098 or 9% chance of having the disease.

In the case above everyone was screened for the disease regardless of having symptoms or not, but most people only get tested for a disease if they have symptoms, are known to be at risk of having the disease or otherwise exhibit an indicator of the disease. In these cases, positive results will be more likely to be true.

Example 2

A breast cancer diagnostic test is conducted by biopsy. About 30% of women tested have breast cancer. The false positive rate is 2% and the false negative rate is 14%. What is the chance that an individual testing positive has breast cancer?

The way to solve this is again to create a table and put in some theoretical numbers.

If 100,00 people are tested and we expect 30% to have breast cancer then we will have 30,000 with breast cancer, leaving 70,000 without. As the false positive rate is 2%, we expect that 1,400 of the 70,000 without breast cancer will test positive. As the false negative rate is 14%, we expect that 4,200 of the 30,000 with breast cancer will test negative.

	Have breast cancer	Don't have breast cancer	Totals
Test positive	25,800	1,400	27,200
Test negative	4,200	68,600	72,800
Totals	30,000	70,000	100,000

In this case, the probability that an individual testing positive has breast cancer:

= 25,800/27,200 = 95%

10.4 Sensitivity and specificity are inversely related

A perfect diagnostic test is one that has both high sensitivity and specificity such that the test parameter perfectly distinguishes diseased from healthy patients. However, some diagnostic tests can be sensitive without being specific, or vice versa. With diagnostic tests it is often possible to shift the threshold (cut-off value) used to decide whether a test result is positive or negative to optimise either sensitivity or specificity. In most cases there will be no perfect threshold as there is an inverse relationship between sensitivity and specificity.

Setting thresholds

For example, in the diagnostic test described in the next figure, the threshold could be set at position A. This would ensure that all healthy patients are deemed negative (no false positives, so high specificity) but it does mean that a large proportion of diseased patients will test negative (high false negatives, so low sensitivity). If the threshold was set at position B, then all diseased patients would test positive (no false negatives, so high sensitivity) but many healthy patients would test positive (high false positives, so low specificity).
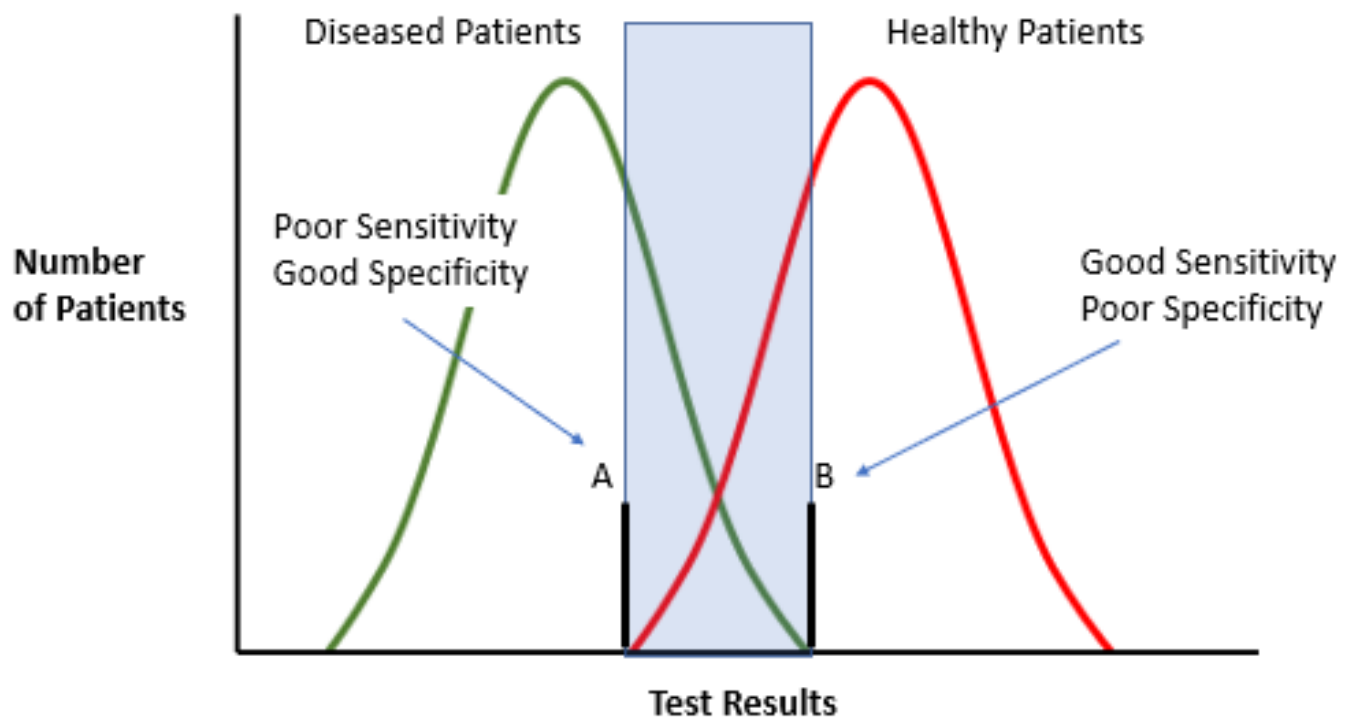
Figure 10.5: Sensitivity vs. Specificity (Adapted from 'Specificity vs Sensitivity Graph' by Blue64701 from Wikimedia Commons used under CC BY-SA 4.0.).

Adjusting the cut-off threshold for a diagnostic test either increases specificity at the expense of sensitivity or increases sensitivity at the expense of specificity.The choice of threshold will depend on the nature of the disease. If you were screening for a very serious disease such as a cancer where early detection could prevent fatality, it makes sense to set the threshold so the test has high sensitivity (few false negatives). In a mass screening test for a less serious condition such as monitoring cholesterol levels or for one where early detection is not critical, it may make sense to have a higher specificity to not overburden the healthcare system.

10.5 Practice problems
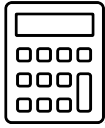
1. A total of 1,500 children had a rapid strep test (RST) done by a standardised culture technique. Of the 1,500 children, 1,338 have a negative RST and 162 have a positive RST. In addition, a backup throat culture (gold standard test) was done on all children. Of those children with a negative RST, 1,302 have a negative throat culture. In the group with a positive RST, 159 have a positive throat culture. Complete the table below and then calculate the sensitivity and specificity of the RST (provide answers to one decimal place).

	Target disorder (strep tonsillitis)		
	Present	**Absent**	**Totals**
Diagnostic test result Positive RST			
Negative RST			
Total			

Sensitivity =
Specificity =

1. A diagnostic test is 92% sensitive and 94% specific. A test group comprises 500 people known to have the disease and 500 people known to be free of the disease.
   a. How many of the known positives will test positive?
   b. How many of the known negatives will test negative?
2. A screening test for a particular disease has a sensitivity of 96% and a specificity of 92%. You plan to screen a population in which the prevalence of the disease is 0.2%.
   a. What is the chance that if you test positive you actually have the disease?
   b. What percentage of the known negatives will test negative?
3. For the diagnostic test below (Figure 10.6) for COVID-19, would you advise shifting the threshold (A) to the left, or right? Explain your answer.
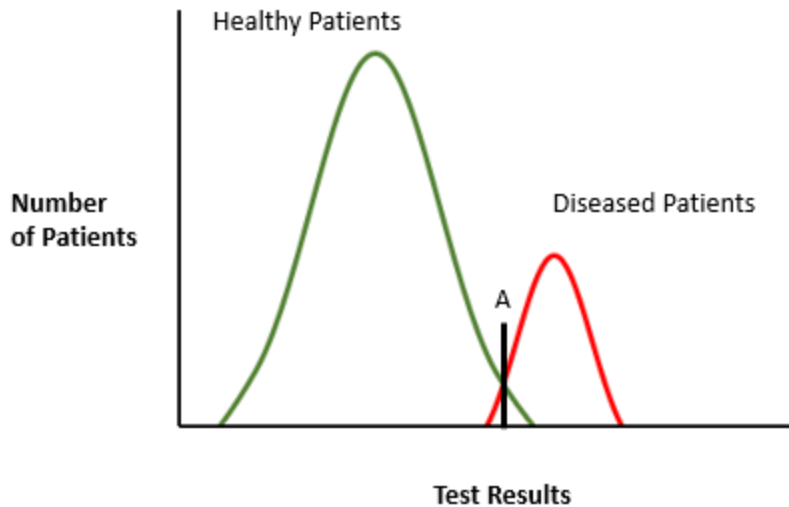
Figure 10.6: Setting the threshold (A) for a diagnostic test for COVID-19.

Solution to Practice Problem 10.3.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1091

XI

Chapter 11: Correlation, causation and confounding variables

When there is a set of data containing two variables that might be related, we refer to this data as **bivariate**. Often the goal of a statistical analysis is to determine the nature of the relationship between these two variables. For example, cancer epidemiologists may want to investigate the relationship between smoking and lung cancer. One way to do this is to perform a bivariate analysis by looking at both the incidence of cancer and the number of cigarettes smoked by individuals.
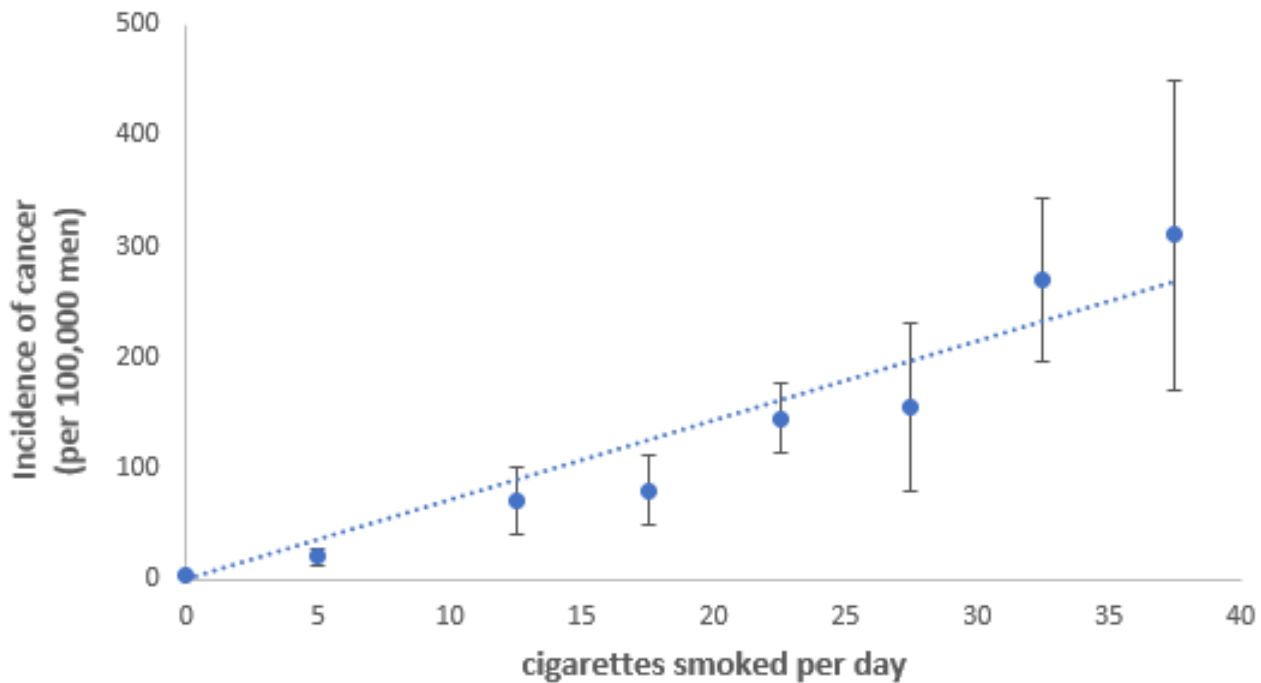
Figure 11.1: Incidence of cancer in men increases with the number of cigarettes smoked per day. (Data adapted from New Scientist, October 22, 1994, Volume 144)

You can see from the graph that as the number of cigarettes smoked per day increases so does the incidence of cancer. We can say that these two variables are correlated. Correlation (sometimes called dependence) is any statistical relationship between two variables.

Furthermore, we can say that there is a positive correlation between with the number of cigarettes smoked per day and the incidence of cancer – positive because as one variable increases so does the other. Variables can also be negatively correlated. That is, as one variable increases the other decreases.

One vital point about correlation is that just because there is a correlation between two variables, it does not necessarily follow that there is a causal relationship between them.

We can explain correlations in four broad ways:

1. The correlation is occurring purely by chance and the variables are not in fact correlated.
2. It could be that variable x causes variable y.
3. It could be that variable y causes variable x.
4. The correlation is real but there is a third variable which causes x and y.

If we think specifically about the case of smoking and cancer, we can rule out that the correlation is occurring purely by chance simply due to the sample size of these kinds of studies and the strength of the correlation (see below).So the question is now, why are these two variables correlated? It seems unlikely that getting cancer would increase your desire to smoke! A more plausible explanation is that smoking causes cancer but that still leaves the possibility of a third variable (or confounding variable) which is responsible for the relationship. This is the argument or alternative hypothesis put forward by tobacco companies. That is, that smoking itself does not cause cancer but rather there is another variable which is associated with smoking that is actually causing the cancer. It might be that smokers also tend to consume more alcohol and it is the increased alcohol consumption (the confounding variable) that

causes the increased incidence of cancer.

Determining causation can be difficult and time-consuming. It often requires rigorous experiments using large datasets. Ideally, such experiments lead to functional data, whereby manipulating the cause (independent variable) changes the effect (dependent variable). Tobacco companies exploit this difficulty in establishing causal links between associated variables to deflect criticism on the effects on health of smoking. However, in this specific case, some testable predictions have been used to further demonstrate the link. For example, stopping smoking or smoking filtered versus unfiltered cigarettes both should and do decrease the incidence of cancer.

47

11.1 Investigating the relationship between two variables

When investigating the relationship between two variables the first step is to display the data graphically on a scatter plot. This allows you to see if the two variables appear correlated. Often you will be able to see if the relationship between the two variables appears linear.
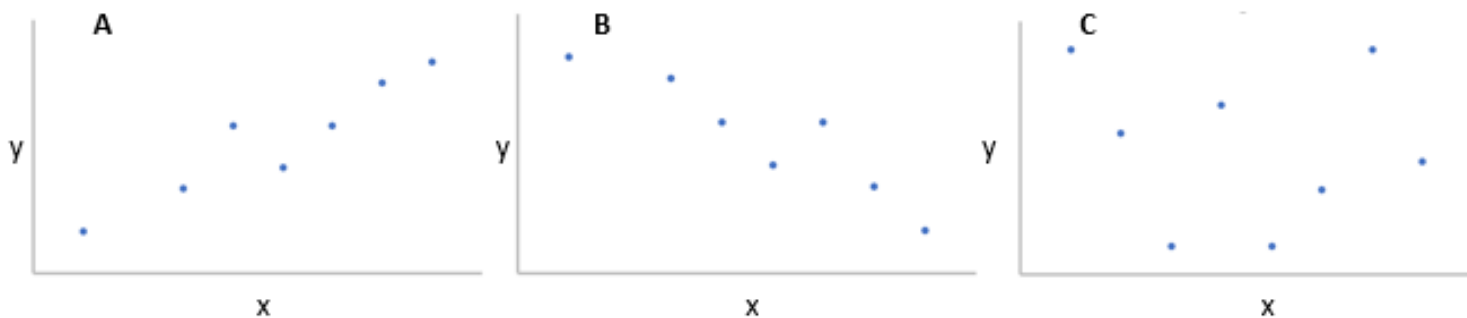
Figure 11.2: Different examples of correlation between data points. A represents a positive correlation between x and B represents a negative correlation and C no correlation.

A represents a positive correlation between x and y, B represents a negative correlation and C no correlation. For the three datasets plotted above we can clearly see that:

- for A, there is a positive correlation between x and y
- for B, there is a negative correlation between x and y
- for C, there is no correlation between x and y.

After this the most important techniques for investigating the relationship between two variables are determining the correlation coefficient and performing a linear regression analysis. The correlation coefficient quantifies the strength of the linear relationship between a pair of variables and the direction

of the correlation, whereas regression expresses the relationship in the form of an equation, which is useful in being able to make predictions regarding the data. If a curved line is needed to express the relationship between the variables, correlation and regression analysis can still be undertaken but require more complicated measures outside the scope of this resource. However, spreadsheet programs like Excel can easily perform these analyses for you.

11.2 Determining the correlation coefficient

The strength of a correlation can be estimated using the correlation coefficient. It is sometimes called Pearson's correlation coefficient after its Karl Pearson. This coefficient is denoted by r and is a measure of linear association. The correlation coefficient (r) can vary between -1 and 1.

An r of 1 indicates a perfect positive correlation; that is, all the data points fall on a straight line and y increases as x increases. An r of -1 indicates a perfect negative correlation; that is, all the data points fall on a straight line and y decreases as x increases.

An r of 0 indicates there is no correlation; that is, points appear to be random in their association between x and y, and even if we know the values of one variable we cannot conclude anything about the values of the other variable.

So, strength of correlation can be estimated by r. The closer r is to either -1 or 1 the stronger the correlation. That is, the absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. It is worth noting that a perfect correlation ($r = +1$ or $r = -1$) is generally not seen in biological systems and mostly occurs in theoretical models only.

For a general 'rule of thumb' for interpreting and describing correlation coefficients:

- .90 to 1.00 ($-.90$ to -1.00) is a very strong positive (negative) correlation
- .70 to .90 ($-.70$ to $-.90$) is a strong positive (negative) correlation
- .50 to .70 ($-.50$ to $-.70$) is a moderate positive (negative) correlation
- .30 to .50 ($-.30$ to $-.50$) is a low or weak positive (negative) correlation
- .00 to .30 (.00 to $-.30$) is a negligible or very weak positive (negative) correlation.

11.3 Calculating the correlation coefficient (r)

The correlation coefficient (r) is defined as how close a set of data points associate with a line (linear regression line) based on those points. You calculate it by taking the ratio of the covariance of the two variables normalised to the square root of their variances. Practically, you are unlikely to calculate this value by hand, especially for a large dataset, but it can be of value to follow a calculation to understand how this coefficient is derived. The formula is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2}\sqrt{n \sum y^2 - (\sum y)^2}}$$

So, to calculate r for the data for graph A in the previous figure, first calculate the mean.

x	y
1	2
3	4
4	7
5	5
6	7
7	9
8	10
Mean 4.86	6.29

Now you can calculate the deviation from the mean for each of the values followed by the square of each.

x	y	x − x_{mean}	y − y_{mean}	$(x - x_{mean})^2$	$(y - y_{mean})^2$
1	2	−3.9	−4.3	14.9	18.4
3	4	−1.9	−2.3	3.4	5.2
4	7	−0.9	0.7	0.7	0.5
5	5	0.1	−1.3	0.0	1.7
6	7	1.1	0.7	1.3	0.5
7	9	2.1	2.7	4.6	7.4
8	10	3.1	3.7	9.9	13.8
$\overline{x}$ = 4.9	$\overline{x}$ = 6.3	$\sum$ = 0	$\sum$ = 0	$\sum$ = 34.9	$\sum$ = 47.4

Now you calculate the sum of the cross product of these deviation scores (e.g. for the first line this is −3.9 × −4.3).

x	y	x − x_{mean}	y − y_{mean}	$(x - x_{mean})^2$	$(y - y_{mean})^2$	Cross product
1	2	−3.9	−4.3	14.9	18.4	16.5
3	4	−1.9	−2.3	3.4	5.2	4.2
4	7	−0.9	0.7	0.7	0.5	−0.6
5	5	0.1	−1.3	0.0	1.7	−0.2
6	7	1.1	0.7	1.3	0.5	0.8
7	9	2.1	2.7	4.6	7.4	5.8

8	10	3.1	3.7	9.9	13.8	11.7
$\overline{x} = 4.9$	$\overline{x} = 6.3$	$\sum = 0$	$\sum = 0$	$\sum = 34.9$	$\sum = 47.4$	$\sum = 38.3$

We can now put these values into the equation:

$$r = \frac{38.3}{(\sqrt{34.9})(\sqrt{47.4})}$$

$$r = 0.94$$

A very strong positive correlation!

Note

To calculate the correlation coefficient, the data for both variables should be continuous or on an interval scale (you cannot calculate the correlation coefficient for categorical data). Also, the data for at least one variable should be normally distributed and should have a linear relationship – which can be read from looking at a scatter plot of the data.

Calculating linear regression

We can also use these values to help calculate the line of regression (line of best fit for the data points).A straight line can be described using the equation:

$$y = a + bx$$

where a = the intercept and b = the slope of the line.

The slope can be determined by the following equation:

$$b = \frac{\sum(x - \overline{x})\sum(y - \overline{y})}{\sum(x - \overline{x})^2}$$

Therefore, in this example:

$$b = \frac{38.3}{34.9}$$

$$b = 1.09$$

We can then calculate the intercepts by substituting the mean values for x and y into our linear regression equation:

$$y = a + bx$$

$$6.3 = a + 1.09 \times 4.9$$

$$a = 0.96$$

So now our equation is:

$$y = 0.96 + 1.09x$$

And we could represent the data using the following graph.
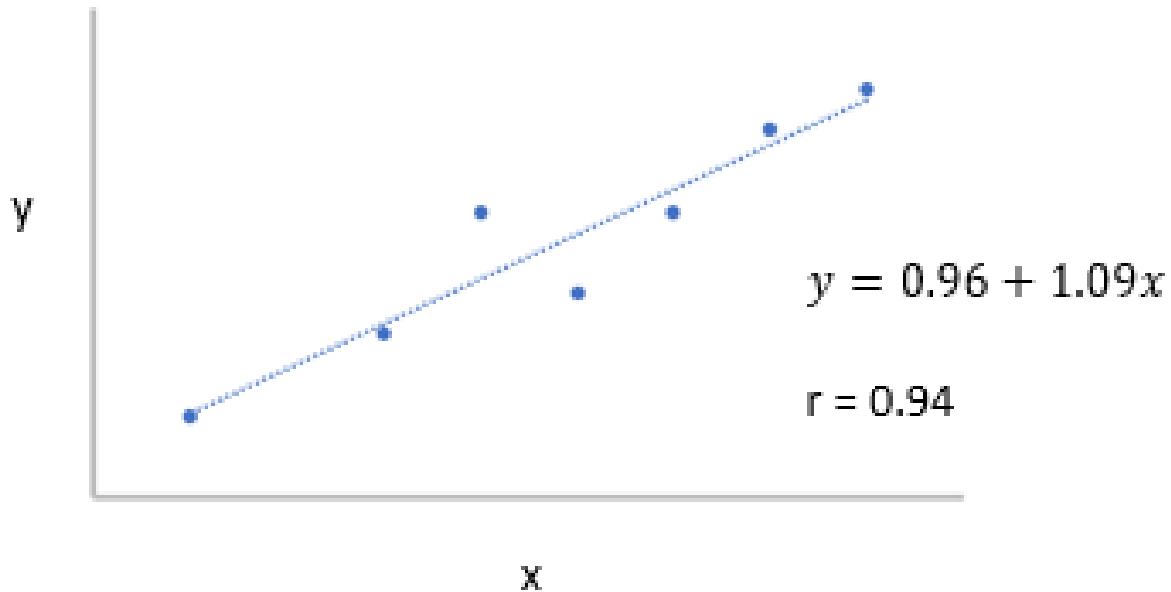
Figure 11.3: A linear regression equation is shown on the graph and its function has been plotted. This makes it easier to visualise how the actual data points deviate from this. The correlation coefficient has also been provided indicated a strong positive correlation.

50

11.4 Explanatory power of correlations (R2)

One way to determine the explanatory power of a linear regression equation is by determining the coefficient of determination, R^2 – that is, the square of the correlation coefficient. R^2 measures the percentage of variation in the dependent variable that can be attributed to the independent variable. That is, it is a goodness of fit measure for linear regression models. R^2 is always a value between 0 and 1. A dataset with a low R^2 value will have data points spread wide from the regression line; a dataset with a high R^2 value will have data points clustered close to the regression line.

Example: Correlation and causation

Researchers found a correlation between the latitude people live in and mortality rate due to skin cancer.

A scatter plot revealed that the relationship appeared to be linear, and the correlation coefficient of the data was 0.71 (a high positive correlation). What percentage of the mortality rate due to skin cancer can be attributed to latitude?

$R^2 = (0.71)^2 = 0.50$. Therefore, 50% of the mortality rate due to skin cancer can be attributed to latitude. Or in other words, 50% of the mortality rate due to skin cancer is due to factor(s) other than latitude. Also, take note from this example that seemingly high values of r (e.g. $r \approx 0.7$) explain only about 50% of the variability in the response variable.

It is important to use the correct notation for the correlation coefficient (use r or R) and the coefficient of determination (use r^2 or R^2). Using these incorrectly is likely to lead to misleading data interpretations.

51

11.5 Misleading regression models

There are some things to note about R^2. First, it does not always indicate how well a regression model fits the data. It is possible that a good model has a low R^2 value, and a biased model has a high R^2 value.You need to examine scatter plots to see how well a regression model fits the data and a good way to check is to plot the residuals. Residuals are measured values minus the predicted value.Look at the next figures, again examining the relationship between number of cigarettes smoked per day and the incidence of cancer, in this case lung cancer. In both scatter plots, A and B, there is a strong positive correlation, and the data appears to fit the linear regression line well.
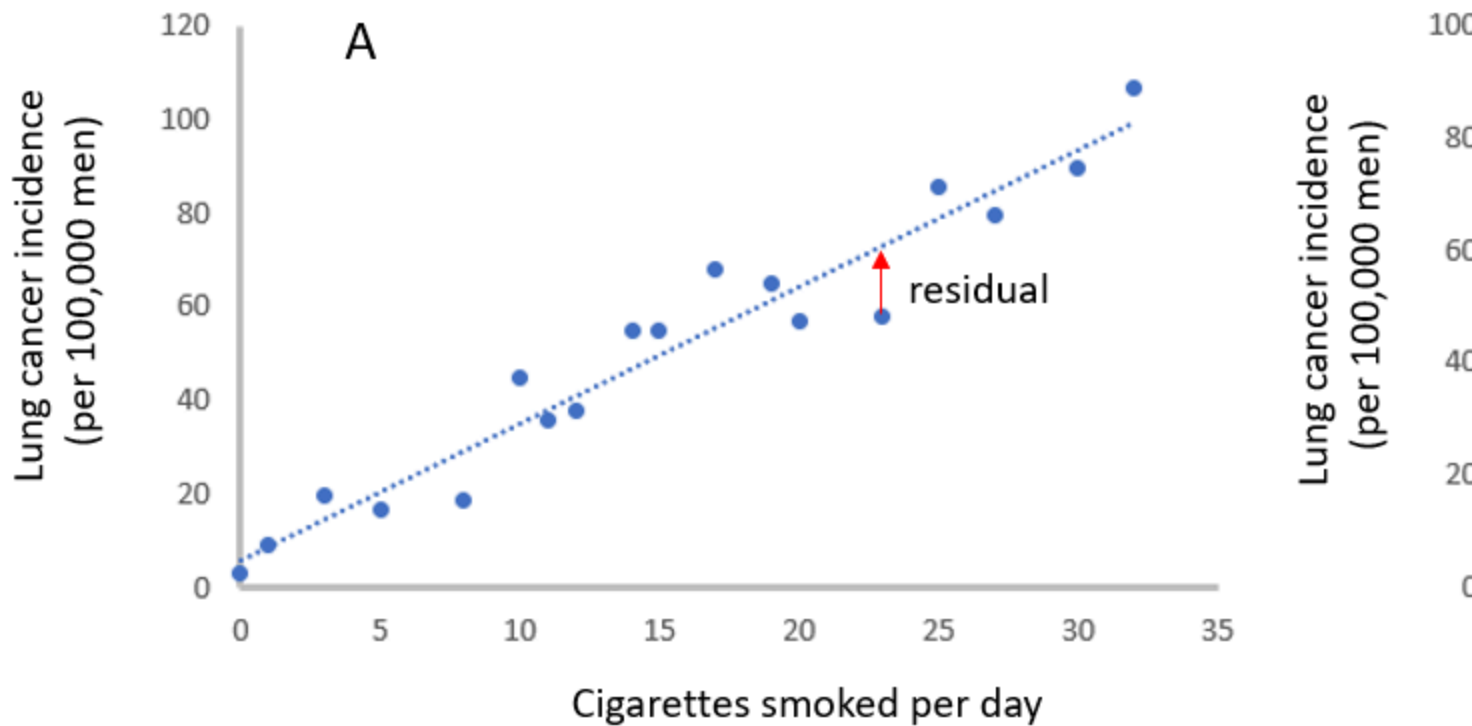
Figure 11.4: The value of plotting data. Even though both sets of data have a strong positive correlation, in B lung ca smoking, indicating there is a bias in the model.

However, you may notice that in plot B the incidence of lung cancer consistently falls below the regression line for the high values of cigarettes smoked. This indicates a possible bias in the model. To test for biases, the values of the residuals (measured value – predicted value) can be calculated and plotted. The next figure shows residual plots for the datasets A and B above.
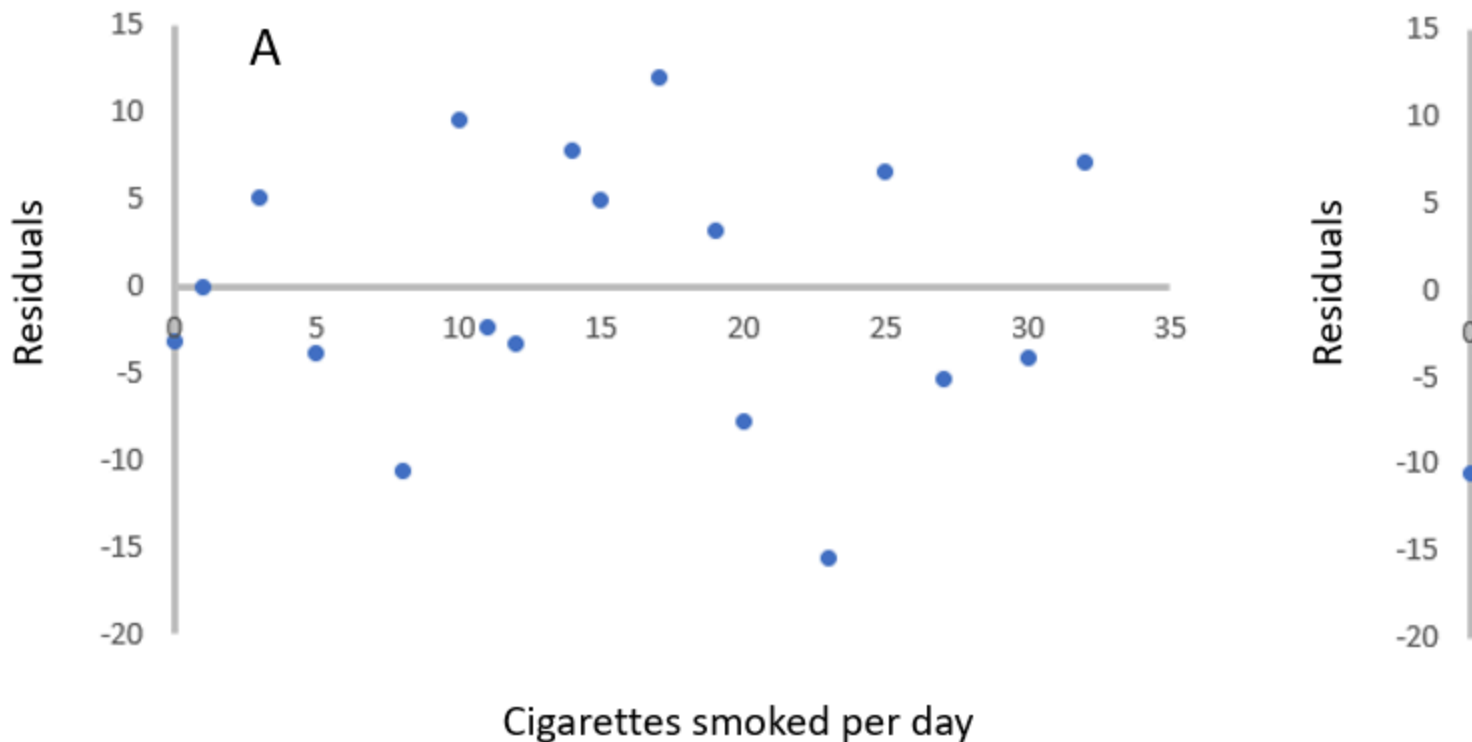
Figure 11.5: Plotting the residuals (measured value – predicted value) is a useful way to determine if there is bias (a

If the regression model is unbiased the residual plot should reveal that the residual values are randomly scattered above and below the regression line (plot A). However, in the residual plot data for B you can see that at the extreme low and high ends the residuals are consistently below the regression line and in the middle range of cigarettes smoked per day are consistently above the regression line. A curved rather than a linear model would fit dataset B much better. It is always crucial to plot the data!

52

11.6 Practice problems
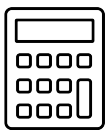
1. Which of the following correlation coefficients (r) indicates the strongest correlation?
   a. $r = 0.45$
   b. $r = -0.5$
   c. $r = 0.3$
   d. $r = -0.2$
2. Is it correct to calculate the Pearson's correlation coefficient for blood type and the duration of hospitalisation following COVID-19 infection expressed by the number of days? Explain your answer.
3. The degree of late gadolinium enhancement (a measure of cardiac injury) was found to correlate to circulating troponin I levels ($r = 0.52$). Is it correct to say that the degree of late gadolinium enhancement explains 52% of the variation in troponin I levels?

4. In recent research, a correlation was found between the latitude people live in the US and the mortality rate due to skin cancer. A scatter plot revealed that the relationship appeared to be linear, and the correlation coefficient of the data was −0.71. The researchers concluded that the major contributing factor to skin cancer in the population is the amount of sunlight people receive. Why might this conclusion be incorrect? See if you can propose some confounding variables to explain this correlation.

5. A colorimetric spectrophotometric assay was conducted to determine protein concentrations. A series of bovine serum albumin (BSA) standards were measured in the assay and the absorbance values determined.

BSA (mg/ml)	Absorbance
0	0.03
0.1	0.03
0.2	0.1
0.3	0.09
0.4	0.12
0.5	0.16
0.6	0.15
0.8	0.26
0.9	0.24
1	0.31

    a. Plot this data on a graph.
    b. What is the correlation coefficient for this data?
    c. What is the linear regression equation for this data?
    d. What is the concentration of a solution with an absorbance of 0.135?
    e. The following additional BSA standards were measured.

BSA (mg/ml)	Absorbance
1.4	0.37
1.5	0.36
1.6	0.37

Add these data points to your graph and describe what they show. What are the implications of these additional data points?

Football and COVID-19 – A correlation?

Figure 11.6: Image by Capri23auto from [Pixabay](#) used under [Pixabay License](#).

The COVID-19 pandemic spawned an unprecedented number of scientific publications. It was a huge challenge for clinicians and biomedical scientists to navigate this literature. Many variables were shown to correlate with COVID-19 susceptibility or severity including blood group, vitamin D levels, air quality, socio-economic status, population density and more. For many of these variables a plausible hypothesis for a causal link to COVID-19 could be put forward. However, testing that causal link is much harder. It is difficult to design controlled experiments to demonstrate cause and effect as it is not possible in most cases to manipulate the independent variable (e.g. air quality) and randomly allocate individuals to different test groups to see the effect on the dependent variable (COVID-19 susceptibility or severity). For many of the correlating variables, we know there is no direct causal relationship but identifying the likely variable to explain the correlation is not straightforward.

A paper published in the journal *Clinical Microbiology and Infection* (Ayoub et al., 2021) provided a humorous but cautionary example of one correlation. The authors discovered that there was a strong positive correlation between the number of COVID-19 infections in a country and its global ranking in the Fédération Internationale de Football Association (FIFA).

Each point in the next graph represents an individual country's global COVID-19 ranking (based on the total number of COVID-19 infections per million people, as of July 2021) plotted against its men's FIFA football ranking (based on the importance, outcomes and relative strengths of opposition in

international matches over the previous four years, as of July 2021).
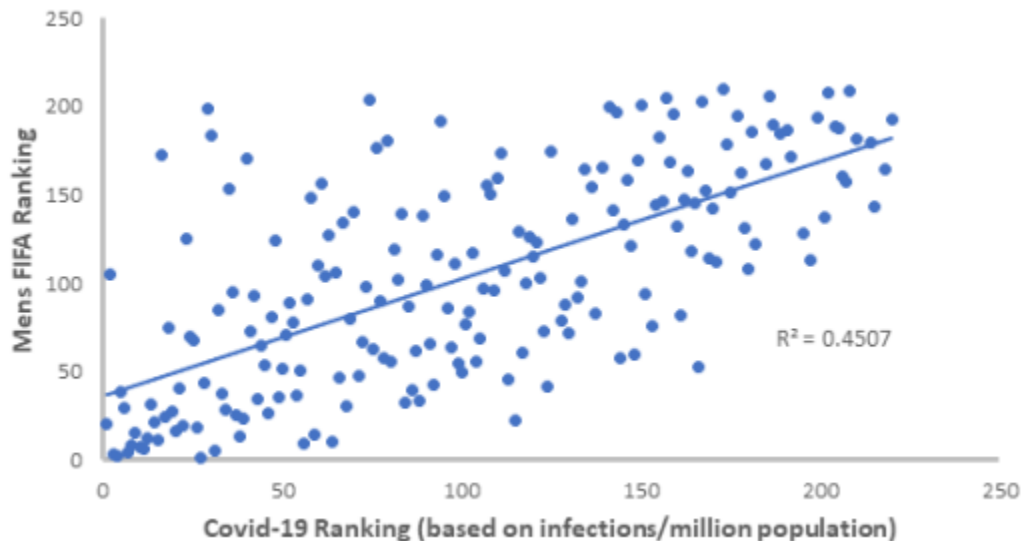
Figure 11.7: A positive correlation between COVID-19 infection rates and FIFA ranking (Ayoub et al., 2021)

What does this correlation mean? Does it mean that people who are good footballers are at increased risk of catching COVID-19 or at increased risk of spreading it? Possibly it means that COVID-19 makes you a better footballer – this seems unlikely!

However, a good biomedical scientist should never ignore a correlation. While there may not be a direct link, we can think about any confounding or explanatory variables that link football and COVID-19 infection.

53

11.7 Boffin questions
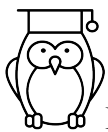
Football and COVID-19

- Speculate on what you think the explanatory variable(s) might be in this case.

Chapter 12: Growth and decay – Exponents and logarithms

In biology (and every other discipline in science) you will often have to deal with numbers which vary in magnitude to an extraordinary degree. Often the only way to graphically represent such data is to use logarithmic or exponential scales and you will find that you often need to interpret this kind of data. Related to this, there are many important measurements which use an exponential scale (e.g. pH). Moreover, many relationships between variables are best described by exponential functions. For example, in a growing bacterial or cancerous cell population, the number of cells can increase exponentially over time or the clearance of a drug from your body can decrease exponentially over time. Being able to work with logarithms and exponents to describe such relationships is fundamental to understanding them and allows us to make accurate predictions.

The spread of infectious diseases can often also be described by exponential functions and these are certain to be a part of any mathematical modelling of their epidemiology (studying the distribution, risk factors and spread in the community). It is of vital importance in managing infectious disease that we are able to understand how it is likely to spread in the community, to know the risk factors involved and to predict the effects of any interventions we might implement.

54

12.1 A focus on infectious disease modelling

Modelling Infectious Disease

Joel Miller is a mathematician who models infectious disease. He has worked at various public institutions and advises policymakers regarding the management of infectious diseases. Watch an interview with Joel to learn how the spread of infectious disease is modelled and the difficulty in incorporating human behaviour into models. (Especially when predictions made by the models affect human behaviour!) You will also learn some tips on how to apply your skills in mathematics to real-world problems.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1110

55

12.2 Dealing with large changes in magnitude

Think about the large range in scale of the components of biological systems from small molecules right through to large multicellular organisms. For example. *Escherichia coli* is a typical gram-negative, cylindrical bacterium about 1–2 μm long, with radius about 0.5 μm. That is, $1\text{–}2 \times 10^{-6}$ m long with a

radius of 0.5×10^{-6} m. The average size of a globular protein in the cytosol of a bacterium is approximately 300 amino acids long, which equates to a molecular weight of around 33,000 Da or 33 kDa. A protein of this size is 6 nm in diameter, that is, 6×10^{-9} m. A typical mammalian cell has a diameter about 10 times the length of an *E.coli* cell. Making a graphical comparison of components that vary so greatly with respect to biological size is very difficult to do using a linear scale.

The data in the next figure are the same as those taken from the previous chapter on biological scale. Notice that when representing relative sizes using a linear scale, the resultant graph has most of the components bunched up close to the origin (plot A) making it difficult to interpret. However, if the scale is transformed to a log scale, then it is much easier to see the relative size differences (plot B). You just need to be aware that each increment along the x-axis in this case represents a 10-fold change from the previous.
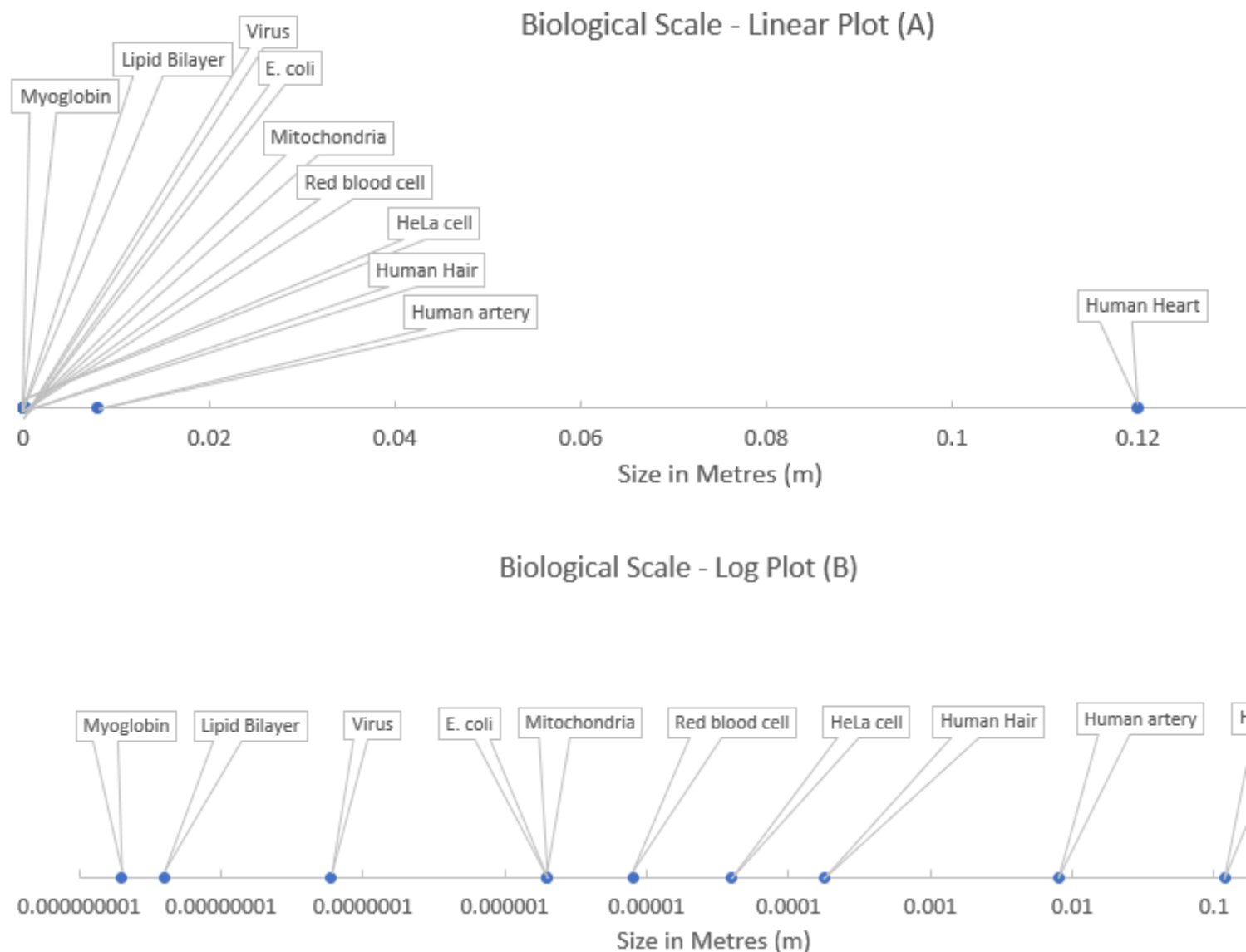

Figure 12.1: It is impossible to resolve various biological components when they are plotted by size on a linear scale

sizes vary by many orders of magnitude. A logarithmic scale (B) is far more useful for displaying this data.

Exponential measurements

Certain important measurements rely on a logarithmic scale. For example, the pH scale is logarithmic, which means that the difference in 1 pH unit is a difference of 10 times the concentration of hydrogen ions in solution. Similarly, in spectrophotometry, the amount of light absorbed by a solution is the logarithm of the light transmitted through a solution (see the section on determining an unknown concentration using spectrophotometry in Chapter 7). Understanding this is important as it means that as an absorbance reading increases there is exponentially less light being transmitted though the sample and therefore much more chance of an error in the spectrophotometer's reading.

Rates of growth and decay

For most biological systems the amount of growth in a population is directly proportional to the size of the population. To understand this, it is useful to think about how a population of bacteria grows. Bacterial cells divide by binary fission so if growth is 100% efficient (each new cell divides into two daughter cells each with the same potential to divide) then the population will double each generation, as shown in the next figure. This kind of growth is known as exponential and can be described by an exponential function where the variable of time is an exponent (i.e. the power to which the base amount is raised).
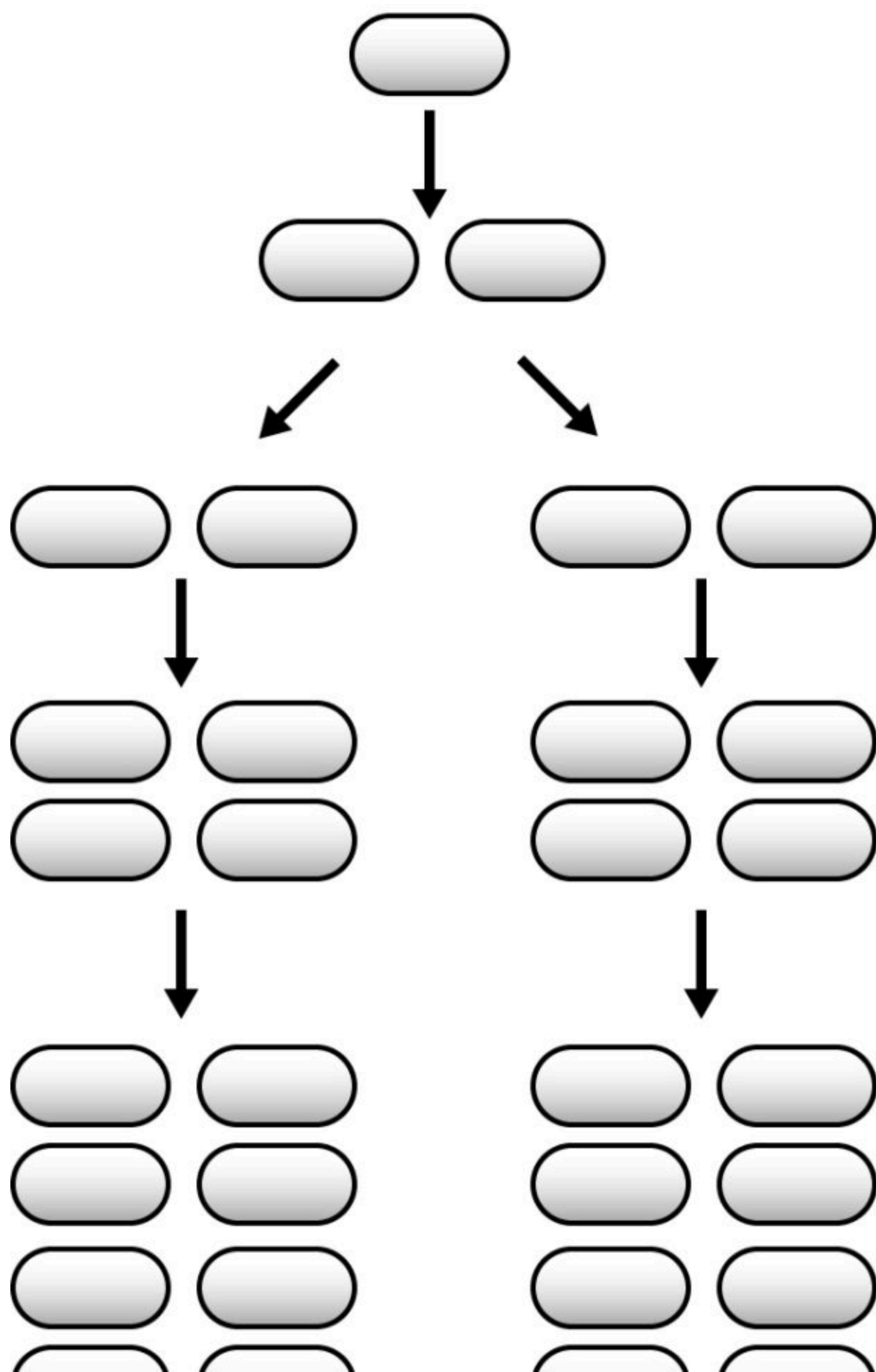
Figure 12.2: Cells displaying exponential growth. Diagram adapted from Cronodon used under CC BY-NC.

The opposite of exponential growth is exponential decay, which occurs when variables decrease proportionally as a function of their current value. Examples of this are radioactive decay, chemical reactions where the rate depends on the concentration of the reactants, and clearance of drugs from the body.

56

12.3 Anatomy of an exponential function

One confusing aspect in all this is that exponential functions may look similar to functions like $y = x^2$. But they are quite different in that the variable x is instead in the superscript position (i.e. is the exponent). For example, in $y=4^x$, the variable x is the exponent, and 4 is a constant and the base. The difficulty in making predictions based on exponential functions is often due to having to solve for an unknown exponent.

$$64 = 4^x$$

Power

Answer

Base

Why is Euler's number so important?

Euler's number, usually just shown as e, is a very important number in mathematics and appears in many functions that describe natural phenomena. It is worth understanding its origins. e is approximated by 2.718 but is an irrational number (i.e. it cannot be written as a simple fraction) and the base of the natural logarithms. e was discovered by Swiss mathematician Leonhard Euler (1707–1783), who was investigating compound interest. Compound interest is where interest earnt on money is added to the principal and interest earnt in the next increment is calculated on this compounded amount.

Imagine you deposit $1 in the bank at an annual interest rate of 2%, compounded each year. After five years the interest earnt would be equal to:

$$(1 + 0.02)^{\,5}$$

If this interest was instead calculated and compounded each quarter, it would be equal to:

$$(1 + 0.02/4)^{4 \times 5}$$

You can see that you will earn more money if the interest is calculated and compounded with more increments. But as you increase the number of increments does the amount increase indefinitely? You could describe this by a more general equation:

$$\$ \lim_{n \to \infty} (1 + 1/n)^n$$

The answer here is no. As you increase the number of increments (n), the total will start to converge on the value of e (2.718). This equation for compound interest also appears when modelling many natural processes – you can think of population growth as similar to compound interest.

Euler's number

Euler's number *e* becomes very useful where we need to model a relationship in which a constant change in the independent variable gives the same proportional change (i.e. percentage increase or decrease) in the dependent variable. For example, because the growth rate of a population of cells in vitro is proportional to the size of the population then the number of cells at any given time can be modelled by an exponential function such as:

$$y = Ae^{bx}$$

where

> y = the total number of cells at time *x*
>
> *A* = the initial number of cells
>
> *e* = Euler's number
>
> *b* = the growth constant
>
> *x* = time.

e and the natural logarithm

So, *e* is used in 'natural' exponential functions, as the following chart shows.
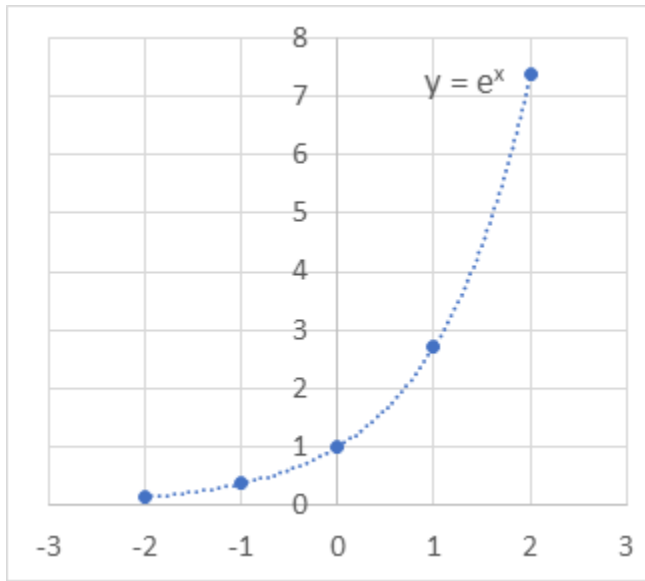
Figure 12.3: A plot of y = e^x. The slope of the curve e^x at any point is equal to the value of e^x. This property of e makes it useful for problems related to growth or decay, where the rate of change is determined by the present value of the number being measured.

One of the amazing things about this function of e is that the slope of the line is equal to the value. That is, if you calculate the slope of the curve e^x at any point it is equal to the value of e^x.

So, when x = 0, the value of e^x = 1, and the slope = 1

And when x = 1, the value of e^x = e, and the slope = e

These are important rules to remember when solving exponential functions.

So, a function like e^x is all about describing growth and is useful for determining the amount of growth or total number of things after x units of time.

The natural logarithm is the inverse of e^x. Because of the inverse nature of logarithms, they are very useful for determining the value of an unknown exponent. For example, they allow us to insert the amount of growth into an equation and then work out the amount of time it would take to get there.

What are logarithms?

The simplest way to explain logarithms is that they are functions which allow us to determine how many of a number we need to multiply to get another number. For example, how many 4s do you need to multiply to get 64? $4 \times 4 \times 4 = 64$, so the answer in this case is 3. So, the logarithm is 3. This is written as $\log_4(64) = 3$.

$4 \times 4 \times 4 = 64$ or $4^3 = 64$ is the equivalent of $\log_4(64) = 3$

In this function, 4 is the base of the logarithm. It is the same as the base of the equivalent exponential function.

For the natural logarithm the base is e. It is most often written in the form ln rather than $\log_e$ and represents how many times e needs to be multiplied to achieve the desired number. For example:

ln $(20.086) = \log_e (20.086) \approx 3$, since $2.71828^3 \approx 20.086$

A logarithm with base 10 is also known as the common logarithm $\log_{10}(x)$ and is most often written as $\log(x)$. So, if a log is used and the base is not specified it is safe to assume the base is 10. The common logarithm is used frequently in science and engineering.

$\log_{10}10 = 1$, $\log_{10}100 = 2$, $\log_{10}1{,}000 = 3$, $\log_{10}10{,}000 = 4$ and so on.

Working with exponents

In order to solve problems involving exponential functions it is worth examining their properties. These properties will provide you with some easy tools for simplifying exponential equations.

Multiplying like bases

When multiplying exponents with like bases we can simply add the bases. For example:

$$x^2 . x^3$$

is the equivalent of

$$x^2 . x^3 = x.x.x.x.x$$

$$x.x.x.x.x = x^5$$

which leads to general rule of

$$x^m . x^n = x^{m+n}$$

Dividing like bases

Dividing like bases is similar to multiplying like bases except instead of adding the exponents you subtract them. For example:

$$\frac{y^5}{y^2} = \frac{y.y.y.y.y}{y.y} = \frac{y.y.y}{1} = y^3$$

which leads to general rule of

$$y^m y^n = y^{m-n}$$

Negative powers

Negative exponents may initially seem more difficult to understand. However, a negative exponent just indicates how many times to divide by the base number. For example:

$$4^{-3} = 1/4^3 = 1(4 \times 4 \times 4) = 0.015625$$

This means that one way of handling negative exponents is to take the reciprocal of the exponent. When you do this, you change the sign of the exponent from minus to plus or from plus to minus, which leads to general rule of:

$$y^{-p} = \frac{1}{y^p}$$

Zero powers

Any number raised to the power of zero is equal to 1. How does this work? Remember the rule about subtracting the exponents when dividing like bases:

$$\frac{y^3}{y^3} = \frac{y \cdot y \cdot y}{y \cdot y \cdot y} = \frac{1}{1}$$

which leads to general rule of

$$y^0 = 1$$

Working with logs

An important property of logarithms and exponents is that they are the inverse of one another. That is, if you apply the log to a number and then apply the exponent, providing you use the same base you will get back to the same number. This property can be written as:

$$\log_y(y^x) = x$$

and the inverse is

$$y^{\log_y(x)} = x$$

This property can be important for solving exponential functions. For example:

What is x in $\log_4(x) = 5$?

Take the exponent of both sides to determine x:

$$4^{\log_4(x)} = 4^5$$

Since $4^{\log_4(x)} = x$

then $x = 4^5 = 1,024$

This leads to the general rule that the log equation can be transformed into its exponential equivalent in this way:

$$\log_b(M) = N \leftrightarrows M = b^N$$

This also means that the natural log of e is equal to 1:

$\ln(e) = 1$

Following on from this we get the following properties of logarithms:

$$\log_3 abc$$

$$= \log_3 a + \log_3 b + \log_3 ac$$

and

$$\log_2 \frac{x}{y} = \log_2 x - \log_2 y$$

This then leads to another important property, which is useful for solving for unknown exponents.

$$\log_c x^3$$

$$= \log_c x.x.x$$

$$= \log_c x + \log_c x + \log_c x$$

$$= 3 \log_c x$$

This further leads to a very important general rule:

$$\log_c x^n = n \log_c x$$

A useful way to further understand logarithms is to use the analogy of growth over time.

If we imagine a population growing over time then the equation $\ln(y) = x$ could be thought of as, how long (x) would it take to get a growth (y) in the population?

If you have the equation $\ln(1) = x$, then x must be equal to 0.

That is, it takes zero time to reach 1 times the current population.

Therefore, the logarithm of 1 always equals 0.

This also makes sense for logarithms of fractional values.

$\ln(0.5) = x$ could be thought of as, how long (x) would it take to get half the current amount in the population?

To get less than we started with means we must go in reverse.

So, $\ln(0.5) = -0.693$

In terms of time, it would take negative time to have half our current value.

Can you take the logarithm of a negative number? For example:

$\ln(-1) = x$

Using the population growth analogy, the question would be how long would it take to get a negative population? You can't have a negative population, so the answer is undefined.

Therefore, $\log(\text{anynegativenumber}) = \text{undefined}$

Putting exponents and logs to use

Can we put these properties of logarithms and exponents to use in solving real-world problems?

A strain of E. coli *bacteria has a doubling rate of 20 minutes in LB broth at 37 °C. If there are 1,000* E. coli *bacteria that are allowed to grow under ideal conditions, how long will it take to reach 1 million bacteria?*

We first need to think about how fast the bacteria are growing.

At time zero there are 1,000 bacteria and there is a doubling every 20 minutes (3 times per hour).

Time (minutes)	0	20	40	60	80	120
# Bacteria		1,000	2,000	4,000	8,000	16,000	32,000

This is an exponential rate of growth where the amount of growth in the bacterial population is directly proportional to the size of the population. The first thing to do is define an exponential equation which describes this rate of growth.

Recall the general form of the exponential growth equation:

$$y = Ae^{bx}$$

where

y = the total number of cells at time x

A = the initial number of cells

e = Euler's number

b = the growth constant

x = time.

In this case, however, we can be more defined:

$$B = (1,000)2^{3t}$$

where

B = number of bacteria

1,000 = starting number of bacteria (multiplied by 2 since they double)

t = time (hours) (multiplied by 3 since they double 3 times per hour).

You can check if the equation works by looking at the table of bacterial growth above.

So, when does the number of bacteria equal 1 million?

$$1,000,000 = (1,000)2^{3t}$$

What we need to do is isolate the exponent t, to be able to solve it. First simplify the equation by dividing both sides by 1,000.

$$\frac{1,000,000}{1,000} = \frac{(1,000)2^{3t}}{1,000}$$

$$1,000 = 2^{3t}$$

A way to isolate t is to take the log of both sides. In this case the common log (base 10) is applied to both sides. This is easy to work with if we have numbers which are multiples of 10.

$$\log 1,000 = \log 2^{3t}$$

$\log 1,000 = 3t \log 2$ (remember the general rule, $\log_c x^n = n \log_c x$)

$$t = \frac{\log 1,000}{3 \log 2}$$

$$t = \frac{3}{3 \times 0.3}$$

t = 3.32 hours (3 hours, 19 minutes, 12 seconds)

Exponential decay

Although we use exponential functions to describe growth, we can also use them to model quantities which rapidly fall toward zero without ever reaching zero. A classic example is modelling the half-life of radioactive isotopes, but exponential functions can also be very useful for modelling the half-life or clearance of drugs from the body, wound healing or even cooling.

An example of an exponential decay function is $y = 2^{-x}$. This is a continuously falling curve where the rate of falling slows as x becomes larger.
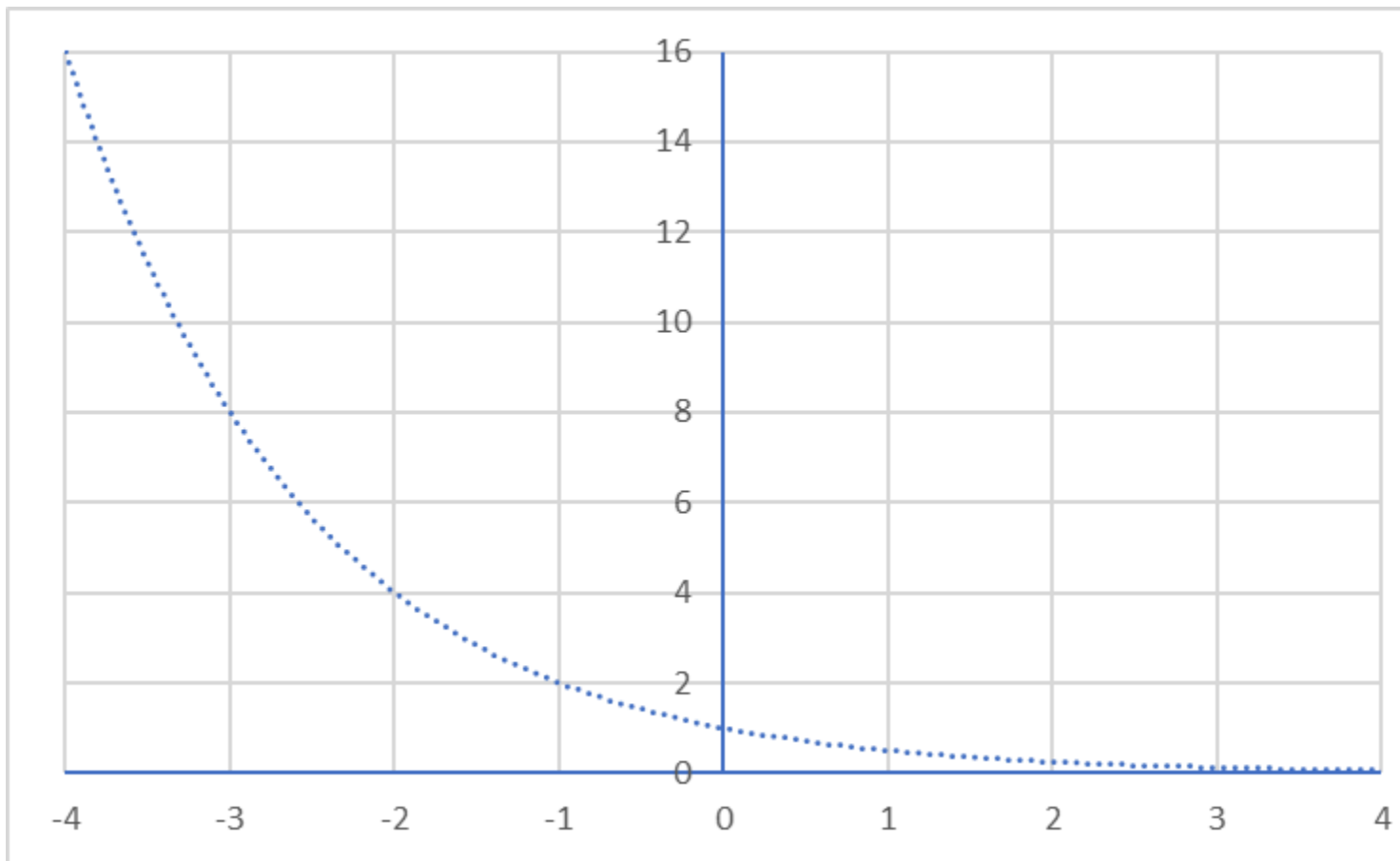
Figure 12.4: An example of an exponential decay function. Note the negative exponent – as x becomes larger, the rate of decrease slows.

The same behaviour is exhibited by functions where the base of the function is a positive number less than 1.

Logarithmic plots

As mentioned earlier logarithmic plots can be useful to display data where there are large changes in magnitude. The pH scale is a good example of this. pH is a calculated as the negative base 10 logarithm

of the molar hydrogen ion concentration of a solution:

$$pH = -\log_{10}[H^+]$$

However, logarithms also have another important function when displaying exponential data graphically. When you log transform a power function the result is a straight line. Consider the function:

$$y = b^x$$

If you take the log of both sides:

$$\log_{10} y = \log_{10} b^x$$

you can transform the equation to

$$\log_{10} y = x \log_{10} b$$

and then further simplify to

$$y = xb$$

which now has the form of straight line.

So, often a logarithmic scale for plotting exponential data as a straight line is easier to work with and a good test of whether the data is exhibiting exponential growth or decay.

Example: Plotting data

The biological effects of drugs are not linear with dose and quite often more closely resemble an exponential function. Imagine increasing the dose of a drug that stimulates cardiac activity. For any drug there will be a lower limit (or threshold) which produces no biological effect (increased heart rate). Above this, the biological effect will start rapidly rising as the dose increases. However, there is a limited capacity for the heart to continually contract faster, so therefore there is an upper limit.

If we plot dose response curves on an arithmetic scale, they will be difficult to interpret as the points on the lower portion of the scale will be plotted close to one another. On a semi-logarithmic scale, as in the following chart, this is not the case, and it is much easier to determine parameters (such as EC_{50} – the concentration of a drug that gives half-maximal response) that in turn are useful for comparing the effectiveness of similar drugs.
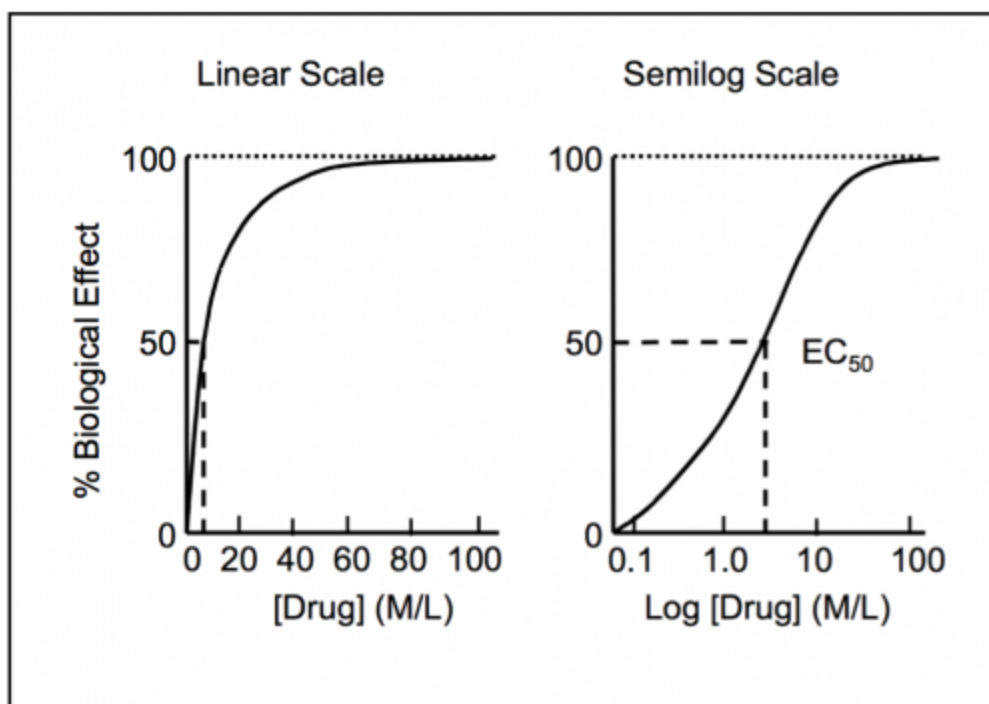
Figure 12.5: Using a semi-log scale (right-hand plot) for a dose response curve makes it much easier to determine the EC50 – the concentration of a drug that gives half-maximal response compared to using a linear scale (left-hand plot). (From TUSOM | Pharmwiki used under CC BY-NC.)

57

12.4 Practice problems
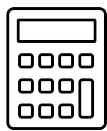
1. $\dfrac{x^9}{x^3} = ?$

2. Rewrite in logarithmic form.

3. Rewrite $10^3 = 1{,}000$ in logarithmic form.

4. Rewrite $\log_x y = z$ in exponential form.

5. Rewrite $\log_3 1 = 0$ in exponential form.

6. $\log_5 5^p = ?$

7. True or false: $\log\left(\dfrac{x}{y^2}\right) = \log x - \log y^2$

8. True or false: $\log(a - b) = \log a - \log b$

9. True or false: $\log x^t = \log a - \log b$

10. True or false: $(\ln x)^t = t \ln x$

11. Solve for x: $\ln x = -3$

12. What is x, when $7^x = 287$

13. Calculate $\ln \dfrac{1}{\sqrt{e}}$

14. What is the approximate hydrogen ion concentration [H$^+$] of blood with pH 7.4?
15. A bacterial population with a starting population of 50 doubles every 20 minutes under ideal conditions. Which equation best describes the growth of the bacterial population?
    a. number of bacteria $\left(\text{at time } t \right) = 50(2)^{3t}$, where t = hours
    b. number of bacteria $\left(\text{at time } t \right) = 2^{t}$, where t = hours
    c. number of bacteria $\left(\text{at time } t \right) = 50^{t}$, where t = hours
    d. number of bacteria $\left(\text{at time } t \right) = 3(50)^{t}$, where t = hours
16. A certain type of bacteria, given a favourable growth medium, doubles in population every 8 hours. Given there were approximately 1,000 bacteria to start with, how many bacteria will there be in 2.5 days?
17. The mad scientist Dr Herbert West accidently causes a zombie outbreak on a small Pacific island. The zombie outbreak can be modelled by the following exponential equation:
$$Z(t) = \frac{50,000}{\left(2 + e^{10-t}\right)}$$
where Z is the number of zombies after t days.
    a. After 5 days, approximately how many zombies are there?
    b. What is the maximum number of people that can turn into zombies?
18. Electrocardiograms (ECGs) are most often used by clinicians to identify potential arrhythmias. The QT interval on an ECG is the time interval representing ventricular depolarisation and subsequent repolarisation (i.e. the duration of activation and recovery of the heart). The RR interval on an ECG is the interval used to calculate the heart rate. The QT interval varies with heart rate (i.e. RR interval) and can be modelled using the following function:
QT$= 425 - 676 \cdot e^{-0.0037 \cdot \text{RR}}$
where RR is the RR interval from an ECG (all units are in milliseconds).
What is the QT interval for an individual with an RR interval of 850 ms? Answer to the nearest whole millisecond.
19. Using the ECG formula for the QT interval
QT$= 425 - 676 \cdot e^{-0.0037 \cdot \text{RR}}$
determine the RR interval for a patient with a QT interval of 330 ms. Answer to the nearest whole millisecond.
20. The area of a wound decreases exponentially with time. The area A of a wound after t days can be modelled by $A = A_0 e^{-0.05t}$ where latex] A_{0}[/latex] is the initial wound area. If the initial wound area is 4 cm^2, how many days until the wound is 50% of its initial size?

▶

Solution to Practice Problem 12.14.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1118

Solution to Practice Problem 12.16.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1118

Solution to Practice Problem 12.20.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://oercollective.caul.edu.au/foundations-of-biomedical-science/?p=1118

58

12.5 Boffin questions

Exponential bias

In the famous fable known as the Wheat and Chessboard problem, the ancient Indian mathematician Sessa (alternatively known as Sissa) invented the game of chess. His king upon seeing the game was so impressed he asked Sessa to name his reward. Sessa asked for a single grain of wheat (in some versions of the story it is a grain of rice) to be placed in the first square of the chessboard, two in the second, four in the third and so on – doubling each time. The king agrees, thinking it a very small reward. The king's treasurer, who was much less arithmetically challenged, was incredibly dismayed with this agreement!

1. Can you work out the amount of wheat Sessa would be owed?

The king in the fable exhibits what is known as exponential bias. Many people underestimate how fast an exponential function can increase in value. Epidemiologists are often frustrated by this when trying to introduce control measures to limit the spread of contagions. The most effective way to stem the spread of a pandemic such as COVID-19 is social distancing, but the introduction of such measures is hampered if a sizeable part of the population fails to see the need as they focus on the initial low numbers of infections.

2. The following data on the number of COVID-19 cases in Canada during the early stages of the pandemic in 2020 are presented in the next figure in two different ways.

Total number of COVID-19 cases in Canada
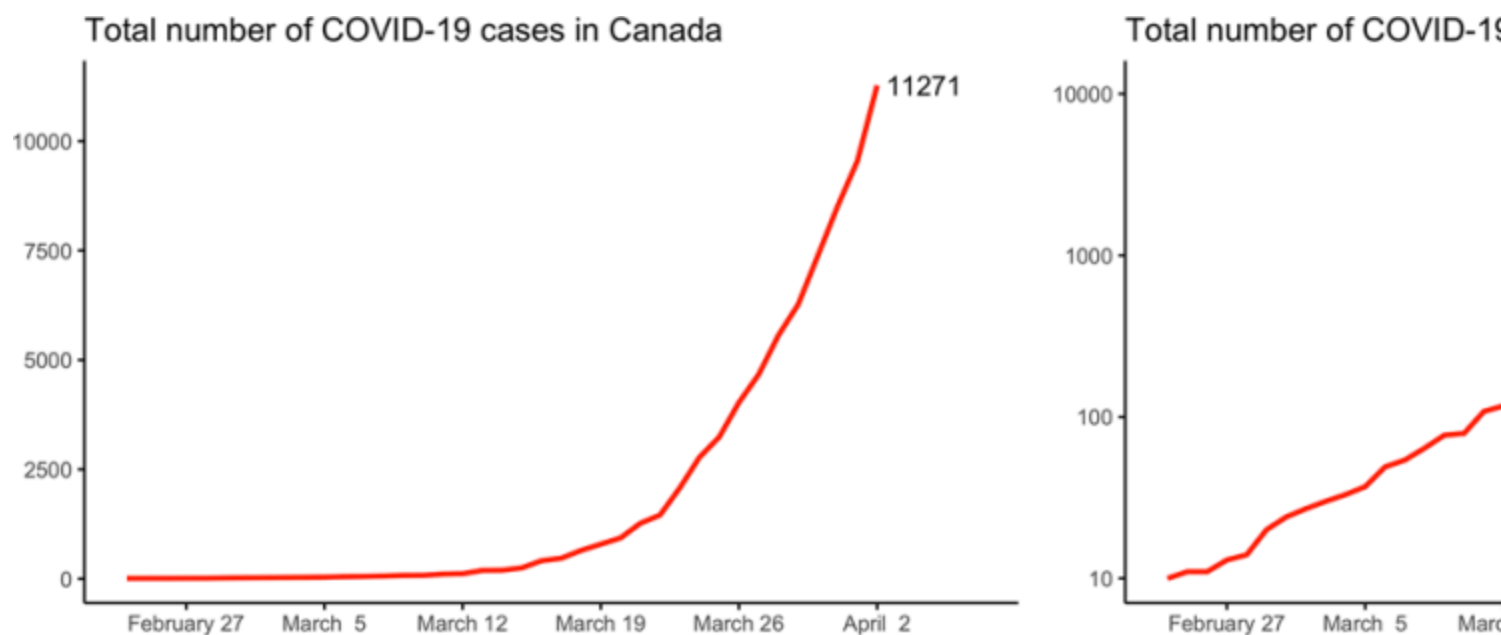
Source: Johns Hopkins University (CSSE)

Figure 12.6: Total number of COVID-19 cases in Canada until 2020 ('Two Time-Series Plots Showing the Cumulat
Up to April 2, 2020' by Sevi, S. et. al. from Canadian Journal of Political Science, 53(2), 385-390 DOI: https://doi.o
under CC BY.)

Which presentation do you think would be more effective at convincing the Canadian public that
social distancing measures were required?

3. Imagine that in a particular country a COVID-10 outbreak occurs. Initial testing shows that 23
   people have been infected so far. Modelling predicts that the number of infected people will
   grow by 28% per day. The country's government wants to push back the moment when
   1,000,000 people are infected as much as possible so as to not overwhelm the healthcare sector.
   Therefore, they wish to adopt increased handwashing to reduce infection rates. Modelling
   suggests that with these measures the number of infected people would grow at 21% per day.
   a. How much time would the country gain with these measures until 1,000,000 people are
      infected?
   b. How many infections could be avoided in the following 30 days with these measures?

What is the unit for pH?

The short answer to this is that pH is dimensionless; that is, it has no units. You may recall that pH is
defined as the negative log10 of the hydrogen ion concentration expressed in mol L^{-1}.

$$pH = -\log_{10}[H^+]$$

So, why does pH have no units? You can also think of the equation as:

$$\mathrm{pH} = -\log_{10} \mathrm{H}^+ \mathrm{mol.} \ \mathrm{L}^{-1}$$

This could be expanded as the sum of the logs of the different units (see the section above on working with logs), which does not make sense.Instead, before taking the logarithm, we have to divide the hydrogen ion concentration by the unit of concentration (M or mol L^{-1}) in order to obtain a dimensionless or unitless number. What we are really doing is this:

$$\mathrm{pH} = -\log_{10} \left(\frac{\mathrm{H}^+ \mathrm{mol.} \ \mathrm{L}^{-1}}{1 \mathrm{mol.} \ \mathrm{L}^{-1}} \right)$$

But you will almost never see the equation for determining pH written like this in a textbook!

Logarithms are what is known in mathematics as a **transcendental** function. A transcendental function is one which cannot be expressed as an algebraic function. Examples of transcendental functions include the exponential function and its logarithmic inverse as well as trigonometric functions such as sin.

XIII

Chapter 13: Further reading and bibliography

Useful books

Cann, A. J. (2013). *Maths from scratch for biologists*. Wiley.

> A useful and reasonably comprehensive problem-based book which start with basic manipulation of numbers, algebra and works towards a basic understanding of descriptive and inferential statistics. Though there is not an emphasis on the biomedical context, this book links, when possible, mathematical concepts to biological examples.

Foster, P. C. (1998). *Easy mathematics for biologists*. CRC Press.

> This is another useful problem-based book which covers many of the topics included here albeit from a non-biomedical perspective, the notable exceptions being the sections related to cell biology, clinical diagnostics and descriptive statistics. The book also includes some problems to help with fractions, decimals, percentages and ratios, which is assumed understanding here.

Bowers, D. (2019). *Medical statistics from scratch: An introduction for health professionals*. Wiley.

> This book does not assume any prior knowledge of statistics. It begins with a primer on variables and types of data, includes a chapter on how to interpret graphical information and builds towards the theory behind experimental design and more complex statistical analysis. An accessible book that tries to avoid jargon and complex mathematical proofs and instead teaches using current examples from the medical literature.

Triola, M., Triola, M., & Roy, J. (2017). *Biostatistics for the biological and health sciences* (2nd edn). Pearson.

This text will help reinforce the descriptive statistics presented here as well as help you to understand probability distributions. However, the text builds on this to describe how to use statistics to analyse authentic data using interesting examples from biology, health sciences and everyday life.

Milo, R., & Phillips R. (2015). *Cell biology by the numbers*. Garland Science.

This book is based on an online resource, Bionumbers (see below), developed by one of the authors, and investigates cell biology through a quantitative prism. It does this in Socratic fashion by asking questions and answering them through small mathematical vignettes. Reading this will help you understand many key concepts in cell biology and give you a lot of confidence to tackle back of the envelope calculations.

La Barbera, M. (2015). *It's alive! The science of B-movie monsters*. University of Chicago Press.

A fun exploration of the scientific plausibility of movie monsters which surprisingly contains a lot of basic biology and mathematics (particularly when discussing issues of scale, e.g. giant insects!). A useful book for learning how think 'outside the box'.

Useful websites

Bionumbers

This vast, searchable database is an incredible resource that allows you to look up numbers related to biology and in particular molecular/cell biology. Every entry has a reference, often with comments about the methods used to obtain the number or notes provided by the authors. Contains numbers from the practical, such as the number of protein molecules per cell, to the esoteric such as the rate of fingernail growth, and everything in-between!

Cell size and scale

A visual demonstration of biological scale. This site will help you appreciate the relative sizes of various cell types and macromolecules as well as learn unit prefixes.

Prepare for university – Mathematical biology

From Cambridge University, a problem-based resource on quantitative biology to help prepare students starting university. A challenging and interesting set of problems which will definitely get you thinking.

Biomath

The Biomath page for the University of Arizona Biology Project contains a series of quantitative biology problems as well as links to some mathematics tutorials.

The maths of COVID-19

A collection of articles from *Plus Magazine* about the mathematics behind COVID-19. Mathematics has

played an enormous role in fighting COVID-19 and the articles here are written in an accessible and informative style.

Useful journal articles

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *British Medical Journal, 331*, 903. https://doi.org/10.1136/bmj.331.7521.903

> An excellent article explaining the difference between standard deviation and standard error and when to use them.

Ayoub F., Sato, T., & Sakuraba A. (2021). Football and COVID-19 risk: correlation is not causation. *Clinical Microbiology and Infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, *27*(2), 291–292. https://doi.org/10.1016/j.cmi.2020.08.034

> An excellent article explaining that observational studies which demonstrate correlations are only useful for hypothesis forming and closer examination of a situation is required to establish true causation between two variables.

Bar-On, Y. M., Flamholz, A., Phillips, R., & Milo, R. (2020). SARS-CoV-2 (COVID-19) by the numbers. *eLife*, *9,* e57309. https://doi.org/10.7554/eLife.57309

> An infographic showing the key numbers from peer-reviewed literature which describe the biology of the virus and the characteristics of infection of a single human host.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. Journal of Cell Biology, 177(1), 7–11. https://doi.org/10.1083/jcb.200611141

> A short publication explaining how to use and interpret the error bars you see in graphical data.

Sender, R., Fuchs, S, & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology*, *14*(8), e1002533. https://doi.org/10.1371/journal.pbio.1002533

> An excellent example of how to use the literature to make informed estimates in quantitative biology.

Swift, A., Heale, R., & Twycross, A. (2020). What are sensitivity and specificity? *Evidence-Based Nursing*, *23*(1), 2–4. http://dx.doi.org/10.1136/ebnurs-2019-103225

> An explanation of how to describe the validity of diagnostic tests. It contains a detailed sample calculation for sensitivity and specificity with a helpful visual interpretation.

Appendix: Answers to problems

Chapter 2: Uncertainty in measurement and significant figures

1.   a. 95.0°C – 3 significant figures
           b. 120°C – 2 significant figures
           c. 501 g – 3 significant figures
           d. 15 patients – An exact number, so infinite
           e. 0.450 ml – 3 significant figures
           f. 250 mM – 3 significant figures
           g. 2.38×10^{-3} g – 3 significant figures
           h. 0.00238 g – 3 significant figures
           i. 2,050 mmol – 3 significant figures
           j. 2.050 mmol – 4 significant figures

2.   a. 1.22 M × 1.3 l = 1.586 mol – 1.6 mol
           b. 500.00 g ÷ 125 ml = 4 g ml^{-1} – 4.00 g ml^{-1}
           c. 516.15 g − 0.005 g = 516.145 g – 516.14 g
           d. 1.023 moles / 2.1 l = 0.4871 M – 0.49 M
           e. 9.000 s − 0.5 s = 8.50 s – 8.5 s
           f. The average height of 3 students with heights 170.3 cm, 167.23 cm and 171 cm
              (170.3 + 167.23 + 171) / 3 = 169.51 cm 170 cm

BOFFIN QUESTIONS: pH readings

pH = –log[H$^+$]

pH = –log[2.56×10^{-5}]

pH = –(log[2.56] + log[10^{-5}]) (note that 2.56 has 3 significant figures)

pH = –(0.408 + –5) = 4.592

Following the rules for addition of significant figures:

0.408 has 3 significant figures, reflecting the uncertainty in the last digit of 2.56

while log 10^{-5} = –5.0000 … has an infinite number of significant figures as 10^{-5} is an exact number.

The rules of addition of significant figures state the number of places after a decimal point is less than or equal to the number of decimal places in every term in the sum.

So, the final answer is a pH of 4.592

Chapter 3: Estimation (sanity checking)

1.   a. 37.7°C (to the nearest degree) – 38°C
            b. 121 mM (to the nearest 10 mM) – 120 mM
            c. 505.3 g (to the nearest 100 g) – 500 g
            d. 505.3 g (to the nearest 10 g) – 510 g
            e. 14.5 g l^{-1} (to the nearest g l^{-1}) – 14.0 g l^{-1}
            f. 0.0245 g (to the nearest mg) – 24 mg
            g. 1.456 µg (to the nearest 100 ng) – 1,500 ng

2.   a. $920 \times 27 - (900 \times 30 = 27{,}000)$
            b. $8{,}453 \div 53 - (8{,}450 \div 50 = 169)$
            c. $79 \times 91 - (80 \times 90 = 7{,}200)$
            d. $1{,}205 \times 0.76 - (1{,}200 \times 0.75 = 900)$
            e. $4{,}215 - 2{,}498 - (4{,}200 - 2{,}500 = 1{,}700)$

BOFFIN QUESTIONS: Estimating large numbers

First, we need to work out the amount of protein in a HeLa cell.

We know the protein mass per volume is approximately 0.3 g/ml

We also know the volume of a HeLa cell is approximately 2,000 µm^3

We need to convert µm^3 into ml:

1 ml is 1 cm^3

and 1 µm is 1×10^{-4} cm

therefore 1 µm^3 is 1×10^{-4} cm $\times 1 \times 10^{-4}$ cm $\times 1 \times 10^{-4}$ cm $= 1 \times 10^{-12}$ cm^3

So, 2,000 µm^3 $= 2 \times 10^{-9}$ ml

So now we can determine the amount of protein in one cell:

0.3 g/ml $\times 2 \times 10^{-9}$ ml $= 6 \times 10^{-10}$ g

We can work out the average molecular weight of HeLa cell proteins by multiplying the typical mass of an amino acid (110 Da) by the average length of HeLa proteins (400 amino acids) to give 44,000 Da (or g/mol).

Next we can calculate the number of moles of proteins per cell by dividing the mass of protein by the average molecular weight:

6×10^{-10} g / 44,000 g/mol $= 1.4 \times 10^{-14}$ mol

Finally, we can multiply this by Avogadro's number (the number of molecules per mole) to

determine the number of proteins:

$$1.4 \times 10^{-14} \text{ mol} \times 6 \times 10^{23} \text{ molecules/mol} = 8 \times 10^9 \text{ protein molecules}$$

Chapter 4: Biological scale

1. The mRNA molecule is certainly much larger! Proteins are made up of chains of amino acids. The average molecular weight of an amino acid is 110 Da while the average molecular weight of a ribonucleotide monophosphate (the monomer unit in an RNA molecule) is 327 Da. However, each amino acid in a protein is coded for by a triplet of nucleotides (each triplet is known as a codon). So, each nucleotide is 3 times bigger than an amino acid and there are at least 3 times more of them than amino acids in the corresponding protein, which is close to an order of magnitude (10-fold) difference in molecular weight.

   If we think in spatial terms the difference in size between proteins and mRNA can be even greater. Proteins typically take on a more compact folded structure while RNA molecules tend to be more diffuse and linear, punctuated with secondary structures in the form of hairpins, stem-loops and pseudoknots.

2. The atomic radius is approximately 0.1 nm while the atomic nucleus is around 1 fm.

   0.1 nm is 0.1×10^{-9} m (1×10^{-10} m) and 1 fm is 1×10^{-15} m

   This means that there is a difference of $1 \times 10^{-10} / 1 \times 10^{-15} = 100,000$ or five orders of magnitude between them!

3. The scientific consensus is that this is not a coincidence! The endosymbiotic theory proposes that eukaryotic cells are the result of separate prokaryotic cells which joined together in a symbiotic union. At some point a free-living bacterium was engulfed by another cell and became the mitochondrion. The engulfed bacterium benefited from being in a protected nutrient-rich environment while the host cell benefited from the production of chemical energy by the symbiont. This endosymbiotic event occurred very early in the eukaryotic lineage as all eukaryotes contain mitochondria.

   The evidence that supports this does not solely rely on size! Like bacteria, mitochondria have their own circular DNA, their own membranes, reproduce by fission, and some remnant protein components which are highly similar to those in bacteria.

4. The length of DNA in a human cell is therefore 6 Gbp $\times$ 0.3 nm bp^{-1}

   That is, $6 \times 10^9 \times 0.3 \times 10^{-9}$ m $= 1.8$ m!

5. This 1.8 m length of DNA must fit inside the cell's nucleus, a space typically 5 μm across. This implies there must be a very efficient way of packing the DNA. The DNA is indeed packed into a series of higher order structures to form chromosomes. DNA initially bonds with proteins called histones to form chromatin. The DNA winds 1.65 times around a core nucleosome structure of 8 histone molecules. Theses nucleosomes fold to form a 30 nm wide fibre. This fibre then forms loops an average of 300 nm long. The loops are then compressed and folded to

form a 250 nm wide fibre. This fibre is then coiled again to form the chromatids of the chromosome. The packing ratio from naked DNA to chromosomes is in the order of 10,000:1!

BOFFIN QUESTIONS: Fantastic Voyage

1. Based on the width of a human arm, we might generously estimate that Raquel's eyes (pupils) are about 1/10th the width of the hair cell cilia. That would make them 0.02 μm in diameter, which is 20 nm. The human eye can detect wavelengths of light from 380 to 700 nm. This means that Raquel's eyes, being much smaller than the wavelength of visible light, will not allow her to see at all!

This is not the only scientific fault with this film. However, if you can suspend disbelief and get past some of the outdated special effects it is still a lot of fun!

2. To solve this, you first need to convert both heights to same units:

$3 \ \mu m = 3 \times 10^{-6} \ m$

Therefore, the volume decrease will be $(3 \times 10^{-6}/1.8)^3 = 4.6 \times 10^{-18}$

If their starting mass stays the same, then their density would be $2.16 \times 10^{20} \ kg \ m^{-3}$

This is far denser than the Earth's core! If there was a way to shrink people like this they would sink through the floor, bedrock and all the way through the Earth's mantle!

Chapter 5 Scientific notation and SI units

1.   a. $123,000 \ m - 1.23 \times 10^5 \ m$
      b. $299 \ l - 2.99 \times 10^2 \ l$
      c. $0.0035 \ mol - 3.5 \times 10^{-3} \ mol$
2.   a. 0.004 mol to mmol – 4 mmol
      b. 1.76 ml to μl – 1,760 μl
      c. 0.0023 μm to nm – 2.3 nm
      d. $20.3 \ \mu l$ to $l - 2.03 \times 10^{-5} \ l$
      e. $0.12 \ fg$ to $ng - 1.2 \times 10^{-7} \ ng$
      f. 1.2 nmol to pmol – 1,200 pmol
      g. $1.2 \ mg \ \mu l^{-1}$ to $g \ l^{-1} - 1.2 \ mg \ \mu l^{-1} = 0.0012 \ g \ \mu l^{-1} = 1,200 \ g \ l^{-1}$
      h. $1.2 \ ng \ \mu l^{-1}$ to $g \ l^{-1} - 0.0012 \ g \ l^{-1}$

BOFFIN QUESTIONS: How many cells are in your body?

1. To solve this, you need the volume of your body to be in the same units as the volume of a cell. In this case the volume of the body will be converted into μm^3.

Assuming the mass of the body is 70 kg, 1 kg takes up about 1 l so the volume is 70 l.

1 L is 0.001 m^3 so the total volume is 0.07 m^3

1 m = 1 × 10^6 μm

Therefore, 1 m^3 = 1 × 10^{18} μm^3

The volume of the body is 0.07 × 1 × 10^{18} μm^3 = 7 × 10^{16} μm^3

Now the number of cells can be calculated:

7 × 10^{16} μm^3 / 1 × 10^3 μm^3 cell^{-1} = 7 × 10^{13} cells

7 × 10^{16} μm^3 / 1 × 10^4 μm^3 cell^{-1} = 7 × 10^{12} cells

The answer is there are between 7 × 10^{12} and 7 × 10^{13} cells in a 70 kg body. So the answer is between 7 trillion and 70 trillion cells.

2. It is worth reading the paper by Sender et al. (2016) to see how they approach calculating an estimate regarding the total weight of bacteria in the human body.

They use a 70 kg 'reference man' to make their estimate. They obtain estimates of the total number of bacteria from human tissues and conclude that compared to the colon the total number of bacteria is negligible. So, the focus of their calculation is the volume of the colon and the concentration of bacteria within.

They estimate that the volume of the colon is 0.4 l (corresponding to 400 g) and the number of bacteria per gram of wet stool is 0.9 × 10^{11}. This gives them an estimate of 3.6 × 10^{13} bacteria in the colon and considering that the contribution to the total number of bacteria from other organs is at around 10^{12}, they use 3.8 × 10^{13} as their estimate for the number of bacteria across the whole body for their reference man.

Since bacteria make up about 50% of the mass of the content of the colon, an estimate of their total weight is 200 g. This number is consistent with the accepted estimate of the wet weight of a bacterium being 5 pg.

Total weight of bacteria = 3.8 × 10^{13} bacteria × 5 pg per bacterium

Total weight of bacteria = 3.8 × 10^{13} bacteria × 5 × 10^{-12} g per bacterium

Total weight of bacteria = 190 g (approximately 200 g)

This total mass of bacteria mass represents about 0.3% of the overall body weight.

Percentage by weight of bacteria in a human is (0.2 kg / 70 kg) × 100 = 0.3%

Chapter 6: Composition of blood

1. In order to estimate the platelet concentration of the suspension you need to determine the number of platelets present in a defined region of the haemocytometer, then calculate the volume that region encompasses.

   Cells in the central 16 squares are counted. But what to do with cells overlapping the outside borders of those 16 squares? The best way to accurately count is to use a consistent approach such as counting only those cells on the border for the top and right borders and excluding those on the bottom and left borders. Normally for an accurate determination of cell numbers, multiple counts are made, and an average is taken.

   The cell count obtained is 102 platelets. (If your estimate is slightly different do not be too concerned – there will be an operator error with this kind of estimate.)

   The volume of the 16 central squares is 0.2 mm × 0.2 mm × 0.1 mm = 0.004 mm^3

   We need to now convert that volume to ml. Often it is easier to start with the base units and work towards your values:

   $1 \text{ m}^3 = 1,000 \text{ l}$

   $0.001 \text{ m}^3 = 1 \text{ l}$

   $1 \times 10^{-6} \text{ m}^3 = 1 \text{ ml}$

   Remember that units of volume are the cubes of units of length:

   $1 \text{ m}^3 = 1 \times 10^9 \text{ mm}^3$

   So, from above, $1 \text{ ml} = 1 \times 10^{-6} \times 1 \times 10^9 \text{ mm}^3 = 1,000 \text{ mm}^3$

   Therefore, the volume in ml of the 16 central squares of the haemocytometer is:

   $0.004/1,000 = 4 \times 10^{-6} \text{ ml}$

   So now we can calculate the platelet concentration:

   $102 \text{ platelets} / 4 \times 10^{-6} \text{ ml} = 2.55 \times 10^7 \text{ platelets/ml}$

2. The leukocyte in the micrograph is roughly 1.5 times the diameter of the erythrocytes. Given that erythrocytes are typically around 6–8 µm, this would make the cell around 9–12 µm in diameter. This size and the secretory granules contained within the cell are consistent with a type of granulocyte known as a mast cell.

Chapter 7: Solutions and concentrations

1. Molarity (M) refers to concentration in moles per litre. In this case you can use the equation:

$C = n/V$

where

C = molar concentration (M)

n = number of moles

V = volume (l).

Therefore, $C = 5$ mol $/ 2.5$ l $= 0.2$ M

2. This is a two-step problem. First you need to determine the number of moles using the mass of NaOH and the molecular weight:

Number of moles = mass/molecular weight = 5 g/40 g mol^{-1} = 0.125 mol

Then you can use the number of moles and the volume to determine the molarity. The volume needs to be converted from ml to L:

750 ml = 0.75 l

$C = 0.125$ mol/0.75 l $= 0.17$ M $= 170$ mM

3. $C = n/V$

Therefore, M $= n / 10$ l

$n = 2$ M $\times 10$ l $= 20$ mol

4. This is a two-step problem. First you need to determine the moles of Na2CO3 using the molarity and the volume. Remember that the volume should be converted to l and the concentration should be converted to M.

$C = 100$ mM $= 0.1$ M

$V = 750$ ml $= 0.75$ l

$C = n/V$

M = n / 0.75 L

$n = 0.075$ mol

Then you can use the number of moles and the molecular weight to work out the mass required:

Mass = number of moles × molecular weight = 0.075 mol × 106 g mol^{-1} = 7.95 g

5. In this problem we first need to work out the mass of cholesterol in 1 l.

   There is 0.1 l in 1 dl

   Therefore, the cholesterol concentration (mg l^{-1}) is 10 × 97 mg = 970 mg l^{-1}

   Next, we can use the mass and the molecular weight to determine the number of moles per litre (molarity). Remember that the mass must first be converted to g.

   Moles = 0.970 g / 386.64 g mol^{-1} = 0.0025 mol

   Therefore, the molarity is 0.0025 M = 2.5 mM

6. Moles of HCl in 90 ml of a 3 M solution = 0.09 l × 3.0 M = 0.27 mol

   Mass = number of moles × molecular weight = 0.27 mol × 36.46 g mol^{-1} = 9.8 g

7. First you need to convert mg to g:

   50 mg = 0.05 g

   Then, use the mass and molecular weight to determine the number of moles:

   Number of moles = mass / molecular weight = 0.05 g / 100.1 g mol^{-1} = 5 × 10^{-4} mol = 0.5 mmol

8. First you need to determine the mass present in 1 l:

   5 μM = 5 × 10^{-6} M = 5 × 10^{-6} mol l^{-1}

   Mass per litre = moles per litre × molecular weight

   Mass per litre = 5 × 10^{-6} mol l^{-1} × 27,000 g mol^{-1} = 0.135 g l^{-1}

   1 g l^{-1} = 1 mg ml^{-1}

   So the final answer is 0.135 mg ml^{-1}

9. First you need to determine the number of moles present in 1 ml. You need to convert the volume to l and the molarity to M:

   1 ml = 0.001 l

   50 nM = 5 × 10^{-8} M

   n = CV = 0.001 L × 5 × 10^{-8} M = 5 × 10^{-11} mol

Once you have determined the number of moles you can use this and the molecular weight to determine the mass:

The molecular weight is 60 kDa, which = 60,000 Da, which = 60,000 g mol^{-1}

Mass = number of moles × molecular weight = 5×10^{-11} mol × 60,000 g mol^{-1} = 3×10^{-6} g

3×10^{-6} g = 3×10^{-3} mg = 3 μg

10. Assume 1 ml is the equivalent of 1 g

Therefore, the mass required is 300 × 60/100 = 180 g

11. There are 45 g in 2 l

Assume 1 ml is the equivalent of 1 g, so convert the volume from l to ml

Therefore, the weight/volume percentage concentration is:

45/2,000 × 100 = 2.25%

12. For this question assume that the volumes are completely additive.

Volume of benzene = 30 ml

Total volume = 30 ml + 95 ml = 125 ml

The percent volume of benzene = 30/125 × 100 = 24%

13. Mass = concentration × volume

Amount of ampicillin = 50 μg ml^{-1} × 150 ml = 7,500 μg = 7.5 mg

14. Parts per million $\text{(ppm)} = \frac{\text{mass solute}}{\text{mass of solution}} \times 10^6$

Assume 1 ml = 1 g

So, 0.1 l = 100 ml = 100 g

ppm = 0.025 g / 100 g × 10^6 = 250

15. $\text{ppb} = \frac{\text{mass solute}}{\text{mass of solution}} \times 10^9$

Substituting into the equation above:

$130 = \frac{\text{mass of dissolved nitrates}}{600g} \times 10^9$

Mass of dissolved nitrates = $(600 \times 130)/10^9 = 7.8 \times 10^{-5}$ g = 0.078 mg = 78 µg

16. ppm = $\frac{\text{mass solute}}{\text{mass of solution}} \times 10^6$

Convert the volume to mass:

1 l = 1,000 ml = 1,000 g

Substituting this into the equation above:

$5 = \frac{\text{mass solute}}{1,000} \times 10^6$

Mass of solute = $(5 \times 1,000)/10^6 = 0.005$ g = 5 mg

BOFFIN QUESTIONS: Determining an unknown concentration

1. To determine the concentration of the ATP solution, substitute the values into the Beer's law equation ($A = \varepsilon bc$) and solve for c:

$0.8 = 15,400 \text{ M}^{-1} \text{ cm}^{-1} \times 1 \text{ cm} \times c$

$c = 0.8 / 15,400 \text{ M}^{-1} \text{ cm}^{-1} \times 1 \text{ cm}$

$c = 5.2 \times 10^{-5} \text{ M} = 0.052 \text{ mM} = 5.2 \text{ µM}$

2. To calculate the unknown cholesterol concentration, you can use the standard curve's equation, substituting the absorbance for y and solving for x (µg ml^{-1} cholesterol):

$y = 0.003x - 0.1648$

$0.865 = 0.003x - 0.1648$

$x = (0.865 + 0.1648)/0.003 = 343 \text{µg ml}^{-1}\text{cholesterol}$

Chapter 8: Dilutions

1. To solve this, use the equation:

$C_1V_1 = C_2V_2$

In this case you need to determine the volume of the initial solution required (V_1):

$2 \text{ M} \times V_1 = 0.3 \text{ M} \times 400 \text{ ml}$

$V_1 = (0.3 \text{ M} \times 400 \text{ ml}) / 2 \text{ M} = 60 \text{ ml}$

2. To solve this, use the equation:

$$C_1V_1 = C_2V_2$$

However, you first need to convert the concentrations to the same units:

800 mM = 0.8 M

$2\text{ M} \times V_1 = 100 \text{ μl} \times 0.8\text{ M}$

$V_1 = (100 \text{ μl} \times 0.8\text{ M}) / 2\text{ M} = 40 \text{ μl}$

3. Use the equation $C_1V_1 = C_2V_2$ solving for C_2

You need to also calculate the final volume (V_2) in this case as 50 μl + 825 μl = 875 μl

$0.5\text{ M} \times 50 \text{ μl} = C_2 \times 875 \text{ μl}$

$C_2 = 0.029\text{ M} = 30$ mM

4. Use the equation $C_1V_1 = C_2V_2$ solving for V_1

$11.6\text{ M} \times V_1 = 100 \text{ ml} \times 0.1$

$V_1 = 0.86 \text{ ml} = 860 \text{ μl}$

5. Use the equation $C_1V_1 = C_2V_2$ solving for V_2

$250 \text{ ml} \times 1.25 = 0.25 \times V_2$

$V_2 = 1{,}250 \text{ ml}$

6. Use the equation $C_1V_1 = C_2V_2$ solving for C_2

$450 \text{ ml} \times 0.2\text{ M} = C_2 \times 225 \text{ ml}$

$C_2 = 0.4\text{ M}$

7. The dilution factor is the total volume (V2) / initial volume (V1). Note that since this is a ratio it is important that both volumes are in the same units.

Dilution factor = (38.75 ml + 0.25 ml) / 0.25 ml = 156 (or a 156-fold dilution)

8. Use the equation $C_1V_1 = C_2V_2$ solving for V_1

The fact that there is 2 L of 0.25 M Na_2CO_3 is not relevant to the problem unless it is an insufficient amount to prepare the final solution!

$0.25\text{ M} \times V_1 = 100 \text{ ml} \times 0.1\text{ M}$

$V_1 = 40$ ml

9. This is a serial dilution as two consecutive dilutions have been made. To determine the final concentration the two separate dilution factors can be determined and multiplied to provide the overall dilution. Then the initial concentration can be divided by this factor to determine the final concentration. Remember to convert volumes to the same units when calculating the dilution factors.

1st dilution factor = (50 µl + 350 µl) / 50 µl = 8

2nd dilution factor = (25 µl + 1,250 µl) / 25 µl = 51

Total dilution factor = 8 × 51 = 408

Final BSA concentration = 10 mg ml^{-1}/408 = 0.025 mg ml^{-1} = 25 µg ml^{-1}

10. Use the equation $C_1V_1 = C_2V_2$ solving for V_1

Note that the concentration 10x is a suggested dilution so you can assume the final concentration should be 1x.

$$10x \times V_1 = 1x \times 50\text{µl}$$

$$V_1 = 5\text{µl}$$

11. In this case, the stated concentration is not relevant to the problem – only the dilution factor matters.

Dilution factor = $V_2 \div V_1$

10 = (1 ml + x) / 1 ml

x = 9 ml

That is, 9 ml of water needs to be added to 1 ml to achieve a 10-fold dilution.

12. There are two serial dilutions made here. To find the overall dilution factor, which requires working back to the original concentration of the analyte, you need to multiply the two individual dilution factors.

Dilution factor = (100 µl + 300 µl) / 100 µl = 4

Since this was done twice the overall dilution factor is 4 × 4 = 16

13. The first part of the question can be reworded:

Since there is 0.75 mg of DNA in 1 ml, what volume will contain 50 pg?

Since you are working with ratios the units must be the same. So, you can convert pg to mg or mg

to pg:

0.75 mg $= 750$ µg $= 750{,}000$ ng $= 750{,}000{,}000$ pg $= 7.5 \times 10^8$ pg

Volume containing 50 pg $= (50 \text{ pg} / 7.5 \times 10^8 \text{ pg}) \times 1$ ml $= 6.67 \times 10^{-8}$ ml

6.67×10^{-8} ml $= 6.67 \times 10^{-5}$ µl

This is far too small a volume to pipette! The DNA solution first needs to be diluted.

To determine the dilution factor required divide the original concentration by the final concentration.

From above, the original concentration was 7.5×10^8 pg ml^{-1} $= 7.5 \times 10^5$ pg µl^{-1}

The final concentration required is 50 pg in 2.5 µl (i.e. 10% of 25 µl) $= 20$ pg µl^{-1}

The dilution factor required is 7.5×10^5 pg µl^{-1}/20 pg µl^{-1} $= 37{,}500$

You could achieve this dilution by adding 1 µl of the original DNA to 37.5 ml of water or preferably do this by a serial dilution. There are infinite ways to solve this. Here is one possible solution which minimises volumes:

1. Dilute the original solution by adding 100 µl to 900 µl of water (dilution = 10)
2. Dilute the previous solution 10-fold by adding 100 µl to 900 µl of water (overall dilution = 100)
3. Dilute the previous solution 10-fold by adding 100 µl to 900 µl of water (overall dilution = 1,000)
4. Dilute the previous solution 10-fold by adding 100 µl to 900 µl of water (overall dilution = 10,000)
5. Dilute the previous solution 3.75-fold by adding 100 µl to 275 µl of water (overall dilution = 37,500)

BOFFIN QUESTIONS: Homeopathy

1. Without randomised control trials it is very difficult to test the effectiveness of homeopathy as this is the best way to isolate the variable of receiving the homeopathic treatment. The current scientific consensus is that when randomised control trials have been conducted, they have shown that the effectiveness of homeopathic treatment is indistinguishable from the placebo

   Ernst, E. (2002), A systematic review of systematic reviews of homeopathy. British Journal of Clinical Pharmacology, 54: 577-582. https://doi.org/10.1046/j.1365-2125.2002.01699.x

   .
2. First, you need to determine the concentration of aescin in the original preparation.

Since aescin is 30% by weight of the seeds, 10 mg ml^{-1} of ground Aesculus hippocastanum seeds

should contain 3 mg ml^{-1} aescin.

From this and the molecular weight of aescin you can determine its molar concentration:

3 mg ml^{-1} = 3 g l^{-1}

Number of moles = mass / molecular weight

Number of moles = 3 g / 1131.26 g mol^{-1} = 2.652 × 10^{-3} mol

Therefore, the preparation is 2.652 × 10^{-3} M

This preparation is diluted 18 times, which is actually 1 × 10^{18}

Therefore, the concentration of the homeopathic treatment is:

2.652 × 10^{-3} M / 1 × 10^{18} = 2.652 × 10^{-21} M

To determine the number of molecules of aescin in a 1 ml aliquot, determine the number of moles found in 1 ml and then multiply by Avogadro's number (6.022 × 1023):

2.652 × 10^{-21} M = 2.652 × 10^{-21} mol L^{-1} = 2.652 × 10^{-24} mol ml^{-1}

Number of molecules per ml = 2.652 × 10^{-24} mol × 6.022 × 10^{23} = 1.6

At this concentration the homeopathic preparation approaches a Poisson distribution as to whether it contains a single molecule of aescin!

Side Bar – A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times within a given interval of time or space and often used to describe the distribution of rare events in a large population.

Chapter 9: Medical diagnostics – Measurement, uncertainty and distributions

1.
   a. Mean = 20.1
   b. Mode = 17
   c. Range = 8 (25–17)
2. To compare you need to determine the coefficient of variation for each standard. Use the formula:

$$cv = \frac{s}{x} \times 100$$

$cv_{(high)}$ = 0.051/1.005 × 100 = 5.07%

$cv_{(low)} = 0.006/0.104 \times 100 = 5.77\%$

The assay is more precise at the higher standard concentration.

3. Since this is a reference interval for the general population a major consideration is to have a large enough sample to provide a reasonable estimate of the distribution of the analyte. Also, the subjects being tested (the reference population) should be as similar as possible to the population (in this case the general population) for which the test will be applied, except for the presence of disease. The major factors normally considered are age and sex but there could be racial or environment factors such as socio-economic status that could be taken in consideration along with common health factors such as obesity or diabetes.

BOFFIN QUESTIONS: Do ACE inhibitors increase susceptibility to COVID-19?

The key to understanding why this data might be flawed is the self-reporting aspect. People who have hypertension and are prescribed ACE inhibitors are likely to monitor their health more fastidiously than the general population. They are less likely to ignore symptoms and, being in an already high-risk group, more likely to get tested for COVID-19 infection. This highlights a common flaw in observational or self-reporting studies.

Despite the flaws in these early studies, this correlation did lead some clinicians to initially advise patients to stop taking prescribed ACE inhibitors. However, more rigorous clinical studies demonstrated that there was no evidence of ACE inhibitors increasing the chances of developing COVID-19 or affecting its severity. In fact, patients are more at risk of developing complications from COVID-19 without these medications due to increased hypertension.

Chapter 10: Medical diagnostics – Sensitivity and specificity
   1.

		Target disorder (strep tonsillitis)		
		Present	Absent	Totals
Diagnostic test result	Positive RST	159	3	162
	Negative RST	36	1,302	1,338
Total		195	1,305	1,500

Sensitivity = true positives / (true positives + false negatives) × 100

Sensitivity = 159 / (159 + 36) × 100 = 81.5%

Specificity = true negatives / (true negatives + false positives) × 100

Specificity = 1,302 / (1,302 + 3) × 100 = 99.8%

   2.   a. Sensitivity = true positives / (true positives + false negatives) × 100
            92 = x/500 × 100
            x = 460
            That is, 460 of the positives would test positive.

b. Specificity = true negatives / (true negatives + false positives) × 100

94 = x/500 × 100

x = 470

That is, 470 of the negatives would test negative.

3.

| | Disease | | |
	Present	Absent	Totals
Diagnostic test result Positive	192	7,984	8,176
Negative	8	91,816	91,824
Total	200	99,800	100,000

Sensitivity = true positives / (true positives + false negatives) × 100

96 = x/200 × 100

True positives = 192

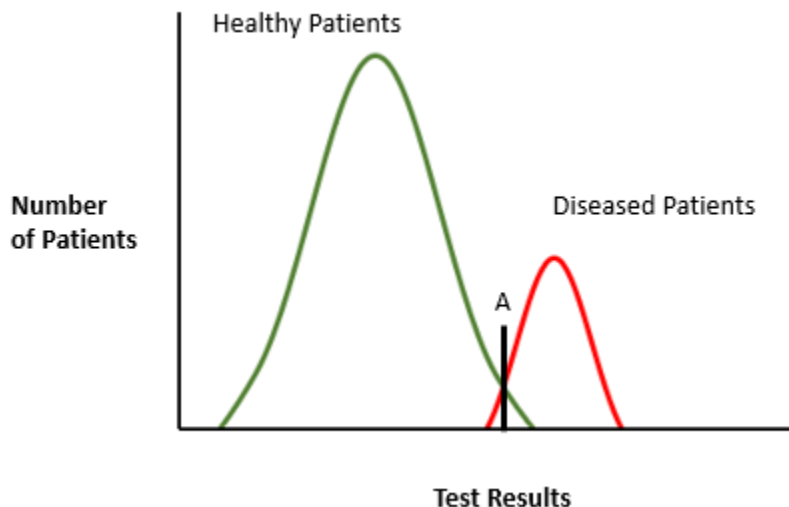Specificity = true negatives / (true negatives + false positives) × 100

92 = x/99,800 × 100

True negatives = 91, 816

a. 192/8,176 × 100 = 2.35%
b. 91,816/99,800 × 100 = 92% (note the definition of specificity)

4.

Since COVID-19 is a serious illness with the added problem of being highly contagious, it would be better to shift the threshold to the left. That is, increase sensitivity but reduce specificity. It would result in an increase in false positives but ensure that all diseased individuals can be quarantined. Follow-up diagnostic tests would be required to clear the false positives.

Chapter 11: Correlation, causation and confounding variables

1. The answer is (b) r = −0.5.
2. No, this is not correct. Blood type is categorical data and cannot be used to determine correlation. Both variables need to be continuous or on an interval scale. One variable should also be normally distributed, and the data should follow a linear relationship.
3. No, to determine how much the degree of late gadolinium enhancement explains the variation in troponin I levels you need to use the coefficient of determination, R^2. In this case it is $(0.52)2 = 0.27$. So, it would be correct to say that based on the data, the degree of late gadolinium enhancement explains 27% of the variation in troponin I levels. It is important to remember not to confuse the correlation coefficient (r or R) with the coefficient of determination (r^2 or R^2).
4. While latitude will directly correspond to the amount of sunlight and it is plausible to suggest that the amount of sunlight might explain the incidence of skin cancer there could be many cofounding variables which account for this correlation. For example, there could be demographic differences in the people living at the different latitudes. Different ethnicities of populations at different latitudes may have different genetic profiles which either protect them from or render them more susceptible to skin cancer. Likewise, there could be differences in diet at different latitudes. These could be potentially controlled for by careful analysis of the data.
5.   a. For the plot, BSA concentration is the independent variable (set by the experimenter) and by convention is plotted on the x-axis, while absorbance is the dependent (measured) variable and by convention is plotted on the y-axis.
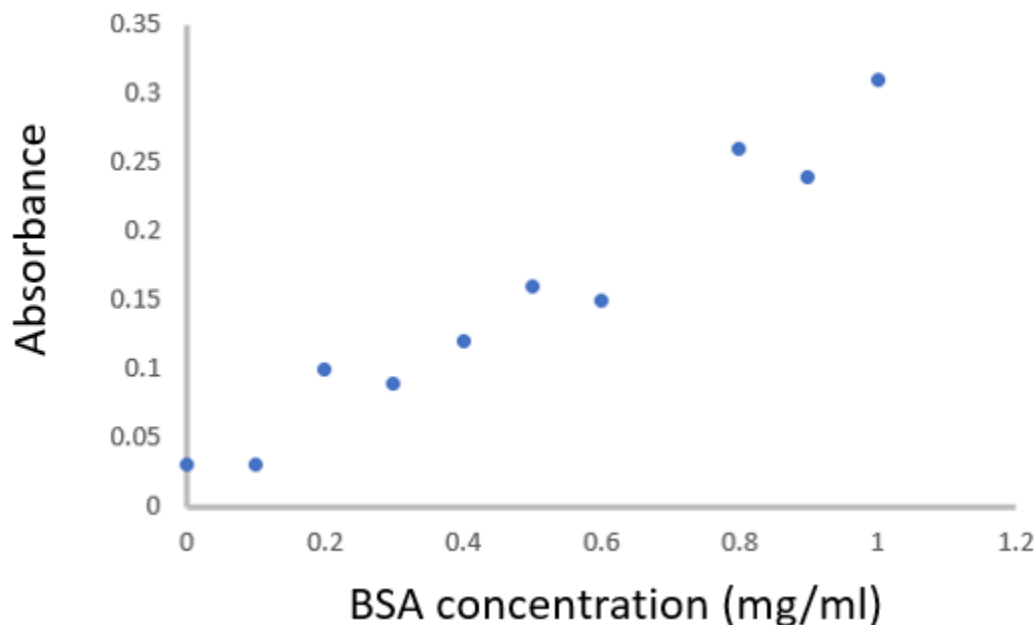
b. To determine the correlation coefficient (r), you could use Excel or another program to calculate this for you. Your answer should be close to 0.98 – a very strong positive correlation! Remember not to confuse the correlation coefficient (r) and the coefficient of determination (R^2). You can also calculate the correlation coefficient manually.

x	y	x − xmean	y − ymean	$(x - x_{mean})^2$	$(y - y_{mean})^2$	Cross product
0	0.03	−0.48	−0.119	0.2304	0.014161	0.05712
0.1	0.03	−0.38	−0.119	0.1444	0.014161	0.04522
0.2	0.1	−0.28	−0.049	0.0784	0.002401	0.01372
0.3	0.09	−0.18	−0.059	0.0324	0.003481	0.01062
0.4	0.12	−0.08	−0.029	0.0064	0.000841	0.00232
0.5	0.16	0.02	0.011	0.0004	0.000121	0.00022
0.6	0.15	0.12	0.001	0.0144	0.000001	0.00012
0.8	0.26	0.32	0.111	0.1024	0.012321	0.03552
0.9	0.24	0.42	0.091	0.1764	0.008281	0.03822
1	0.31	0.52	0.161	0.2704	0.025921	0.08372
$\bar{x} = 0.48$ $\bar{x} = 0.149$		$\Sigma = 0$	$\Sigma = 0$	$\Sigma = 1.056$	$\Sigma = 0.08169$	$\Sigma = 0.2868$

We can now put these values into the correlation coefficient equation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2}\sqrt{n \sum y^2 - (\sum y)^2}}$$

$$r = \frac{0.2868}{(\sqrt{1.056})(\sqrt{0.08169})}$$

$$r = 0.976$$

c. You can use Excel or another program to calculate this for you. When using Excel, you can fit a trendline to the data. You should choose a linear trendline and the equation of the line should be close to:

y = 0.2716x + 0.0186

That is, absorbance = 0.2716 (BSA concentration in mg/ml) + 0.0186

When fitting a trendline you will be given the option of forcing the line through the origin (0,0). While 0,0 is a valid data point you should not force the line through this point. Rather, you want a line of best fit.

d. Now that you have an equation for the trendline or line of best fit you can use this to calculate the concentrations of unknown solutions. For a solution with an absorbance of 0.135 the expected concentration would be:

Absorbance = 0.2716 (BSA concentration in mg/ml) + 0.0186

0.135 = 0.2716 (BSA concentration in mg/ml) + 0.0186

0.135 − 0.0186 = 0.2716 (BSA concentration in mg/ml)

0.1164/0.2716 = BSA concentration in mg/ml

BSA concentration = 0.43 mg/ml

 e. The additional data points illustrate an important consideration for standard curves, which is that it is not valid to extrapolate beyond the range of the standards used to create the standard curve. For example, there may be a strong linear correlation between the dependent and independent variable, but this will not necessarily hold true when the value of the independent variable increases. In this case the additional data points with higher BSA concentrations do not appear to fit the trendline. Absorbance plateaus around 0.36–0.37 at BSA concentrations of 1.4 mg/ml or above.
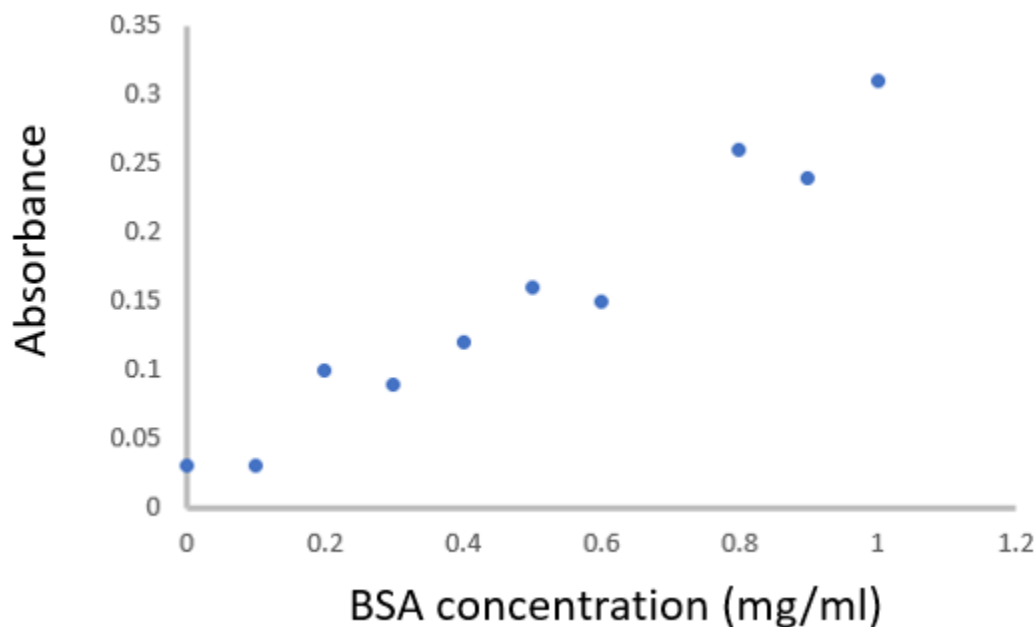
Practically this means that if you used the original trendline to estimate the BSA concentration of a solution with an absorbance of 0.37 there would be a danger of greatly underestimating the actual BSA concentration.

BOFFIN QUESTIONS: Football and COVID-19 – A correlation?

It is unlikely that being good at football leads to an increased risk of catching or spreading COVID-19. It is also not plausible that COVID-19 makes you a better footballer and, besides, the FIFA ranking data predates the emergence of COVID-19. However, this does not mean that the data is necessarily spurious and correlations, while not always directly linked to causation, are worth investigating.

So how could these variables be linked? There are many possibilities:

- ◦ Where football is popular there are more mass gatherings due to football matches, which contributes to the spread of COVID-19.
- ◦ A nation's football ranking may be linked to economic status and a better football ranking correlates with increased travel or increased social activity, which leads to increased spreading.
- ◦ Conversely, many poorer nations may not be as good at football and as they are poorer they are less able to test for COVID-19, so the infection rate is under-reported.

To establish a causal link, all of these variables would need to be investigated.

Chapter 12: Growth and decay – Exponents and logarithms

1. To solve this, you can use the general rule for dividing exponents:

$$y^m y^n = y^{m-n}$$

Therefore, the answer is $x^{9-3} = x^6$

2. To convert between logarithmic and exponential forms you can use the general rule:

$$log_b(M) = N M = b^N$$

In this case, $x = 3^9$ is converted to $\log_3 x = 9$

3. $\log_{10} 1,000 = 3$
4. $x^z = y$
5. $3^0 = 1$
   This is an important rule – any number raised to the power of 0 is equal to 1.
6. You can solve this using the general rule:

$$\log_y(y^x) = x$$

or by performing the conversion below (a useful operation when solving for unknown exponents).

$$\log_c x^n = n \log_c x$$

$$\log_5 5^p = p \log_5 5 = p < spanstyle = "font - size : 14pt" >$$

7. True
8. False; in this case the answer is $\log a / \log b < spanstyle = "font - size : 14pt" >$
9. True, following the general rule:
   $$\log_c x^n = n \log_c x$$
10. False; think about how $\log_{10} 10^2$ differs from $(\log_{10} 10)^2$

$$\log_{10} 10^2 = 2 \log_{10} 10 = 2$$

but

$$(\log_{10} 10)^2 = (\log_{10} 10) \times (\log_{10} 10) = 1$$

11. To solve for x, use the general rule to get the equation into an exponential form:

$$\log_b(M) = N \text{can be converted to} M = b^N$$

$$\ln x = -3$$

can be written as:

$$\log_e x = -3$$

$$x = 0.0498$$

12. To solve for an unknown exponent, take the log of both sides to isolate the exponent:

$$7x = 287$$

$$\ln 7^x = \ln 287$$

$$x \ln 7 = \ln 287$$

$$x = \ln 287/\ln 7$$

$$x = 5.66/1.96 = 2.91$$

13. You can just go ahead and plug this into a calculator or first simplify using the general rule:

$$\log_2 \frac{x}{y} = log_2 x - \log_2 y$$

So

$$\ln \frac{1}{\sqrt{e}} = \ln 1 - \ln \sqrt{e} = 0 - 0.5 = -0.5$$

14. There is a logarithmic relationship between hydrogen ion concentration and pH where:

$$\text{pH} = -\log_{10}[\text{H}^+]$$

$$[\text{H}^+]$$

is the hydrogen ion concentration in mol L^{-1}

To solve for $[\text{H}^+]$ given the pH, convert the logarithmic equation into an exponential equation:

$$[\text{H}^+] = 10^{-\text{pH}}$$

Therefore, at pH 7.4, $[\text{H}^+]$ is:

$$10^{-7.4} = 3.98 \times 10^{-8} \text{mol} L^{-1}$$

15. The answer is (a), in which 50 represents the starting number of bacteria. This is multiplied by 2 since they are doubling with each time increment. t is equal to time in hours and this needs to be multiplied by 3 since they double every 20 minutes (3 times per hour). The other equations are incorrect as they either do not include the starting population or fail to incorporate the doubling at each time increment.

16. You can use this information to write an equation for the number of bacteria at time d (days of 24 hours):

$$\text{Number of bacteria (at time} d) = 1,000(2)^{3d}$$

Since the population doubles every 8 hours then this will occur 3 times per day.

$$\text{Number of bacteria (at time} d) = 1,000(2)^{3(2.5)} \; 180,000$$

17. a.

$$Z(t) = \frac{50,000}{(2+e^{10-5})}$$

$$Z(5) = \frac{50,000}{(2+e^5)}$$

$$Z(5) = \frac{50,000}{(2+148.41)}$$

Zombies at 5 days $= 332$

b.

As t approaches infinity, e^{10-t} approaches 0, so the equation becomes 50,000/2 = 25,000

So, 25,000 is the maximum number of people that can turn into zombies.

18. By substituting 900 into the equation for RR, the answer is 396 ms.
19. Substitute 330 for QT:

$$330 = 425 - 676e^{-0.0037.RR}$$

$$676e^{-0.0037.RR} = 95$$

$$e^{-0.0037.RR} = 95/676$$

To solve for RR take the natural logarithm of both side of the question:

$$\ln(e^{-0.0037.RR}) = \ln\left(\frac{95}{676}\right)$$

$$\text{RR} = \frac{\ln\frac{95}{676}}{-0.0037} \; 503$$

Therefore, a QT of 330 ms corresponds to RR of approximately 530 ms

20. $A = A_0 e^{-0.05t}$

$2 = 4e^{-0.05t}$

$0.5 = e^{-0.05t}$

$\ln 0.5 = \ln e^{-0.05t}$

$\ln 0.5 = -0.05t \ln e$

$\frac{-0.693}{-0.05 \ln e} = t$

$t = \frac{-0.693}{-0.05}$

$t = 13.9 \text{days}$

BOFFIN QUESTIONS: Exponential bias

1. Let's use rice instead of wheat. One grain of rice is placed in the first square. This will double for each square. Since there are 64 squares in total there are 63 doublings.

   Grains of rice $= 1 \times 2^{63} = 9{,}223{,}372{,}036{,}854{,}775{,}808$ grains!

   The average weight of 1 grain of rice (from a sample of short, medium and long grains) is 21 mg. Therefore, the total weight of rice is:

   0.021 g $\times 9{,}223{,}372{,}036{,}854{,}775{,}808 = 193{,}690{,}812{,}773{,}950{,}292$ g $= 193{,}690{,}812{,}774$ tonnes

   To put this in perspective current worldwide annual production is around 740,000,000 tonnes!

1. The answer here is subjective. Many people may not realise this is the same data but with either a linear y-axis (graph A) or an exponential y-axis (graph B). If only data for the initial period was available, then graph B might be more effective to convince the general population to practice social distancing as this graph clearly shows the increase over time while in graph A the number of infections appears static. However, when looking at the entire time course the exponential nature of the increase in infection over time in graph A is probably more obvious to a casual observer.
2. This question highlights the fact that a small change to a parameter in an exponential function can produce a very large effect.
    a. First you need to develop an equation to model the growth in infections.

       Number of people initially infected $= 23$

       No handwashing growth rate expressed as a decimal $= 0.28$

       Handwashing growth rate expressed as a decimal $= 0.23$

Final number of people infected = 1,000,000

Time (days) = t

No handwashing

$1,000,000 = 23(1 + 0.28)^t$

$43,478.3 = 1.28^t$

$\ln 43,478.3 = \ln 1.28^t$

$t = \frac{\ln 43,478.3}{\ln 1.28}$

$t = 43.3 \text{days}$

Handwashing

$1,000,000 = 23(1 + 0.23)^t$

$43,478.3 = 1.23^t$

$\ln 43,478.3 = \ln 1.23^t$

$t = \frac{\ln 43,478.3}{\ln 1.23^t}$

$t = 51.5 \text{days}$

Handwashing buys 8.2 days!

b. The same formula can be used for this question but in this case we know the time period (30 days) and need to work out the number of infections.

No handwashing

$\text{Infections} = 23(1 + 0.28)^{30}$

$\text{Infections} = 37,846$

Handwashing

$\text{Infections} = 23(1 + 0.23)^{30}$

$\text{Infections} = 11,452$

Therefore, 26,394 infections are avoided over this period!

Versioning History

This page provides a record of changes made to this textbook. Each set of edits is acknowledged with a 0.01 increase in the version number. The exported files for this book reflect the most recent version.

If you find an error, please contact eBureau@latrobe.edu.au

Version	Date	Change	Details
1.00	March 2023	Published first edition	

Review Statement

La Trobe eBureau open publications rely on mechanisms to ensure that they are high quality, and meet the needs of all students and educators. This takes the form of both editing and double peer review.

Editing

This publication has been reviewed by an [IPED accredited editor](#) to improve the clarity, consistency, organization structure flow, and any grammatical errors.

Peer review

Two rounds of peer review were completed for this publication in December 2022 by:

- Jodie Young, La Trobe University
- Brandon Cheong, Australian Catholic University

The peer review was structured around considerations of the intended audience of the book, and examined the comprehensiveness, accuracy, and relevance of content, as well as longevity and cultural relevance.

Changes suggested by the editor and reviewers were incorporated by the author in consultation with the publisher.