

Quantitative Human Physiology

Jeffrey A. Walker

2019-04-01

Contents

Directions	5
1 Blood	7
1.1 How to solve an estimation problem	7
1.2 Problem set	8
2 Immune	9
2.1 Combinations	9
2.2 Problem set	10
3 Cardiovascular	11
3.1 Background	11
3.2 Problems	11
4 Respiratory – Why we need Hemoglobin	15
4.1 How much O ₂ can dissolve into blood?	15
5 Renal	17
5.1 A back of the envelope calculation of GFR	17
5.2 Using renal clearance to measure GFR in an individual	17
6 Fluid and electrolyte balance	19
7 Digestion, Nutrition, and Metabolism	21
7.1 Estimating Causal Effects	21
7.2 Simulation 1	22
7.3 Simulation 2	23
7.4 Questions	25
8 Metabolism	27
9 Physiological Genetics	29
10 Cancer	31

Directions

1. Create a Google Sheets file that you will share with me. Name the sheet “last name BIO 223”, for example “Walker BIO 223”
2. Create one sheet for each assignment. There are 10 assignments so your document should have 10 sheets.
3. Name each sheet using the assignment name. For example “1. Blood”, “2. Immune”, etc.

Do not create ten separate Google Sheets files. Do not spread a single assignment over multiple sheets.

Chapter 1

Blood

Goals: back-of-the-envelope estimation, scale, google search

Many complex problems in biology can be broken down into a series of smaller problems and a common smaller problem is the **estimation** of some number, such as the number of bacteria per cell. Estimation problems range from **back-of-the-envelope estimations** that are imprecise but useful in that they give one a general sense of the **magnitude** of a phenomenon to more precisely modeled estimates that are used for making **decisions under uncertainty**. Back-of-the-envelope estimations are called that because most can, literally, be done with a pencil and the back of an envelope. They can be done with pencil because the computations uses rounded instead of exact numbers like 10 or 300 that are easily multiplied/divided. In this module, you will compute some back-of-the-envelope estimations.

A problem like “how many bacteria can colonize a cell” depends on the distribution of the sizes of the bacteria, the size of the cell, and how packed the cell is with its own molecules and organelles. Here, I simply want to get you started on addressing a problem like this with very simple models of the problem. Along the way, solving the problem should give you a sense of scale of what it is like to be a bacterium or a virus living in a cellular world.

1.1 How to solve an estimation problem

I’ll solve an analogous problem: How many beach balls can fit in a barn? If a barn has Volume V_{barn} and a beach ball has volume V_{ball} then the number of balls that could fit into the barn would be approximately, $N_{balls} = \frac{V_{barn}}{V_{ball}}$. To solve this, I need to **parameterize the model** by assigning numbers to these variables. And, the answer is dependent on what numbers I choose for the size of the beach balls and the shape and size of the barn and how filled the barn is with hay (or furniture or horses or whatever) – that is how much of the volume of the barn is available for beach balls. A back-of-the-envelope calculation simply uses a *reasonable* value for the parameters. So, here are my numbers.

1. the barn is a typical vermont barn. I have a sense of what “typical” is because I live in New England and see barns every day. But what if I were a martian, and had never seen a barn? Then I would need to find this information from a reliable source. So, to find “typical”, I used a google search and found what looks like a reputable source that says a typical hay barn is 30 feet wide by 40 feet long. I estimated wall height from the figure as half the width and I used a 12/12 pitch for the roof, so the peak is centered and 15 feet (half of the width) high.

Again – if you don’t have to look up information to parameterize your model, don’t!

2. I used a big beach ball of 2 feet in diameter (because big beach balls are fun). I didn’t need to look this up!

	A	B	C	D	E	F	G	H	I	J	K
1		W (ft)	L (ft)	H (ft)	volume (ft ³)	ball diameter (ft)	volume of cube enclosing ball (ft ³)	max number of balls	volume fraction already filled	number of balls	
2	main room	30	40	20	24000	2	8	3000	0.1	2700	
3	attic	30	40	15	9000	2	8	1125	0.1	1012.5	
4									Total	3712.5	
5											
6	Source for barn size: "The most often utilized size is 30 feet wide and 40 feet long." http://www.thebarnpeople.com/the-barn-people-barn-inventory.html										

Figure 1.1: Estimation of maximum number of beach balls that could colonize a barn. Cells in red are computed.

3. 10% of the barn is filled with hay.

I use Google Sheets to compute the number of balls for each room (the main room and the attic) and then add these. Here is my sheet

My column labels include the units of the measure. **Do not add units to the measure itself** because this makes the format of the cell “text” instead of “number” and you cannot refer to the cell in an equation. I also cite the source of the parameterization below the table (I cite the source for the size of the barn. The size of the beach ball I just made up).

1.2 Problem set

Do these on the same sheet. Name the sheet “1. Blood”.

1. How many red blood cells in a drop of blood? Note, I don’t want you to look up how many RBCs are in a drop, I want you to estimate it using a back-of-the-envelope estimation. You don’t need to look up the volume of a drop of water if you are able to use available information in your head to derive a reasonable volume for a drop of water.

For the next three questions, assume the host cell is “empty”, that is, it contains no organelles or molecules that take up space.

2. How many bacteria could colonize a red blood cell?
3. How many bacteria could colonize a macrophage?
4. How many virus particles could colonize a red blood cell?

– These should all be on the same google sheet.

– **Do not hardcode parameters**, that is, if a virus is 30 feet wide do not put “30” in an equation but instead make your equations reference the cell with this information.

– You may need to google search information to parameterize the model, such as, how big a virus is. Part of the goal of this is for you to develop your skills finding reliable information using a google search. There is variation in virus size or cell size so use something in the middle or “typical”. Again – these are back-of-the-envelope estimates so you don’t need to be very precise – in fact all of these problems could be computed by most working biologists without looking up any information. We all have a pretty good sense for how big a virus, a bacterium, a blood cell, and a drop of water is. But you can look up this information because you are at the beginning of your biology career.

– Cite a webpage giving the source of the information, as I’ve done for the barns. There is no “right” or “wrong” place to get this information, only more or less reliable. I’m not grading you on where you get it, but I want to see where you get it! And all I want for a citation is the webpage, this is not a formal citation that you put in a scientific paper.

Chapter 2

Immune

Goals: combinations

How many kinds of antibody can a human make by V(D)J recombination?

An individual human produces many different antibody proteins, where “different” is amino acid sequence. How is this possible given that there are only a few “antibody” genes? Part of the answer is V(D)J recombination. An antibody is constructed from two pairs of polypeptides. Each pair consists of a light chain and a heavy chain. Each chain has a “variable” region and a “constant” region. The heavy chain is constructed from one gene (located on chromosome 14) while the light chain is constructed from two genes: the light chain locus λ (“lambda”) located on chromosome 22 and the light chain locus κ (“kappa”) located on chromosome 2. The variable region of both light chain loci is composed of a V part and a J part. The variable region of the heavy chain locus is composed of V, D, and J parts. A V, D, or J part consists of multiple copies of the exon that will be spliced into the mRNA but each of these copies has a slightly different nucleotide sequence and some of the copies do not produce functional mRNA.

To make the heavy chain mRNA for the antibody

1. Choose one of the copies of the V region of the heavy chain locus.
2. Choose one of the copies of the D region of the heavy chain locus.
3. Choose one of the copies of the J region of the heavy chain locus.

combine with the C (constant) region to make the heavy chain mRNA

To make the light chain mRNA for the antibody

1. Choose one of the copies of the V region of one (either λ or κ) light chain locus.
2. Choose one of the copies of the J region of the same light chain locus.

combine with the C region to make the light chain mRNA.

Finally, combine the light and heavy chains (these are actually translated independently and then joined into the protein but the math is the same).

So an antibody is a combination of combinations. It is a combination (light + heavy combined) of combinations (V, J, and D combined)

2.1 Combinations

If there are n_1 elements in set 1 and n_2 elements in set 2, how many combinations of 1 element of each set are there? Answer: $n_1 \times n_2$

	A	B	C	D
1				
2		lambda	kappa	heavy
3	V	3	4	5
4	D			2
5	J	2	2	3
6				
7	combinations	6	8	30
8				
9	light chain combinations		14	
10	heavy chain combinations		30	
11	possible antibody mRNA types		420	

Figure 2.1: How many kinds of antibodies

In the table below, I use this math to compute the number of antibodies that could be made using only V(D)J recombination. The

2.2 Problem set

Do these on the same sheet. Name the sheet “2. Immune”

1. (Ken, Jeff, David, and Doug) is the set of male faculty members in the Biology department. (Chris, Terry, Rachel1, Rachel2, and Rachel3) is the set of female faculty members in the Biology department. If the biology department has a square dance, how many combinations of male-female partners could there be? Write all of these out to confirm (write this in a column of your google sheet)
2. Figure 4.3 in this online textbook is a table containing the number of copies of each of the gene segments. Use this table to compute the number of different antibodies that can be synthesized using V(D)J recombination alone.

Chapter 3

Cardiovascular

3.1 Background

This exercise explores equations 12-1 and 12-2 from Vander’s Physiology.

Regulation of blood flow is critical to increase or decrease delivery of blood to organs as they need more or less blood. Blood flow can be modeled with the equation for fluid flow used in almost any system (rivers, wind, etc)

$$F = \frac{\Delta P}{R} \quad (12.1) \quad (3.1)$$

1. F is flow
2. P is pressure. Here, and almost everywhere you’ll see it, Δ (the greek letter “delta”) means “change”, so ΔP (“delta p”) is a *difference* in pressure between two points in space. Here this is two points along the length of a blood vessel. It is the difference in pressure that is driving the blood to flow.
3. R is the resistance to flow due to friction. Friction sucks kinetic energy from moving objects (the lost kinetic energy is transformed to heat).

Resistance is an important concept in understanding human physiology. Resistance can be modeled using the Poiseuille equation

$$R = \frac{8L\eta}{\pi r^4} \quad (12.2) \quad (3.2)$$

1. L is the length of section of blood vessel
2. η (the greek letter “eta”) is the viscosity of the fluid (or more specifically, the dynamic viscosity)
3. r is the radius of the lumen of the vessel.

3.2 Problems

Create a sheet named “3. cardiovascular”

1. Create a table like that below. Write down the units of each of the terms. There is no “right” answer, because units can be written different ways. For example I could write the units of volume as L (“liter”) or gallon or L^3 (here “L” is length). Write down a definition of each term. Write a formal definition and then add your own interpretation of that definition. For example, Wikipedia defines density as the “mass per unit volume” which I’ll interpret as “the amount of matter in given amount of space”,

which doesn't quite capture the nuances but is helpful for understanding. Also notice that wikipedia's definition of density here is an equation expressed as words, this can help with thinking about units.

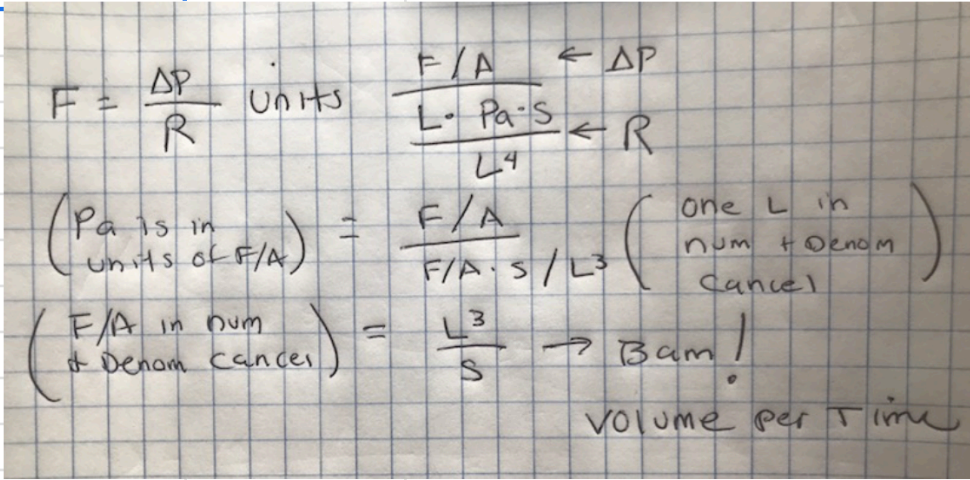
	A	B	C
1		units	definition
2	Flow		
3	pressure		
4	resistance		
5	length		
6	viscosity		
7	lumen radius		
8	stress		

2. The typical units of viscosity, $\text{Pa} \cdot \text{s}$ is not very intuitive. Wikipedia gives a nice way to think about viscosity:

Viscosity is the material property which relates the viscous stresses in a material to the rate of change of a deformation (the strain rate)

Using your knowledge of stress and strain from last semester, how would you express this in units? To answer this 1) write "Viscosity...relates the viscous stresses in a material to the rate of change of a deformation" as an equation, and then 2) determine the units from this equation. Show how the units expressed this way equals $\text{Pa} \cdot \text{s}$. Do this with pencil and paper, snap a photo, and insert it below the table in your Google sheet.

Here is an example to follow: While the equation for flow is useful for understanding how variation in pressure and resistance cause variation in flow, if I use the equation to define flow, I would get something like "the change in pressure of the fluid per unit resistance", which isn't very helpful in thinking about flow. Flow is "the volume of fluid that moves past a point per unit time". So how do I get from the equation to this definition? I worked this out, snapped a photo, re-sized the image to 800 pixels wide, then inserted the image in my google sheet.

	A	B	C
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			

3. The radius of the lumen of an arteriole leading into a capillary increases 50%. What is the change in blood flow to the capillary? Use the google sheet to show your work, including *all* calculations.

Chapter 4

Respiratory – Why we need Hemoglobin

We need hemoglobin because our blood cannot carry enough dissolved O₂ to support our cell activity. That's the short answer. Let's explore a quantitative answer.

4.1 How much O₂ can dissolve into blood?

We can model what is going on in the alveoli using a beaker filled with water. Diffusion of gas molecules from the air into the water (“going into solution”) or the reverse (“coming out of solution”) is not simply a function of the concentration gradient of the gas between the air and the water because *gasses have different solubilities in solution*.

Equilibrium (when an equal amount of gas is going into and out of solution) is modeled by the following equation

$$c_{O_2} = h_{O_2} P_{O_2} \tag{4.1}$$

- c_{O_2} is the concentration of O₂ in the water
- P_{O_2} is the partial pressure of O₂ in air (so a measure of concentration)
- h_{O_2} is Henry's solubility coefficient (or “constant of proportionality”).

Scientists in different fields have different ways of expressing the relationship between c_{O_2} and P_{O_2} so you may land on a web page or a textbook that expresses the constant in something like $\frac{1}{c_{O_2}}$ or even a dimensionless constant. I like this way for physiology because it pumps our intuition about how \bar{P}_{O_2} controls our dissolved O₂ levels.

This equation tells us how much O₂ will dissolve in the water at different partial pressures of O₂ in the air, or, switching back to the lung, how much O₂ will dissolve in the blood plasma given the partial pressure of O₂ in the alveolar air.

Below is a table of P_{O_2} of alveolar air and the resulting concentration of dissolved O₂ at equilibrium.

P _{O2} (mmHg)	c _{O2} (mL O ₂ /dL blood)
20	0.066
30	0.091
40	0.137
50	0.156
60	0.195
70	0.220
80	0.257
90	0.281
100	0.307
110	0.349

1. Transfer the data into your spreadsheet.
2. Plot c_{O_2} (y-axis) against P_{O_2} (x-axis)
3. Compute the slope and in the cell next to the computation, write the units.
4. What is this slope? (the concept not the value)
5. What is c_{O_2} in healthy arterial blood entering an organ (use $P_{O_2} = 97$ mmHg)?
6. How much dissolved O₂ is ejected from our left ventricle each minute? Again, use $P_{O_2} = 97$ mmHg
7. How much O₂ do our tissues need each minute? For this, you need to look up resting O₂ consumption, which is usually in units of mL O₂ per min per Kg. From this, you use the mass of a person to compute their O₂ consumption per minute.
8. Compare the dissolved O₂ sent by the left ventricle to the O₂ required at rest? Do we send enough dissolved O₂ to our tissues?

Chapter 5

Renal

5.1 A back of the envelope calculation of GFR

1. Estimate GFR using a back-of-the-envelope calculation. The calculation is

$\text{GFR} = \text{cardiac output} \times \text{renal fraction} \times \text{plasma fraction} \times \text{filtration fraction}$

Look up reasonable values for the four variables to parameterize this equation. Do the computation in your Google Sheet. Insert the units of GFR in the adjacent cell.

5.2 Using renal clearance to measure GFR in an individual

$$C_s = \frac{\dot{M}_s}{P_s} \quad (5.1)$$

where C_s is the clearance of solute s , \dot{M}_s (“m dot”) is the mass of s excreted in the urine per unit time, and P_s is the plasma concentration of s .

2. What are the units of C_s ? These are the units of what kind of measure (for example Force per Area are the units of a pressure)?
3. Remember that a dot over a variable is a first derivative; here we assume that this is constant and so $\dot{M}_s = \frac{\Delta \text{Mass}}{\Delta \text{Time}}$. What are the units of \dot{M}_s ? This kind of measure is “kinda like” the kind of measure in #2. Google around to see what we call \dot{M}_s .
4. The clearance of a solute is useful in pharmacology but we can also use the concept to measure the GFR in a person. This is done using a solute s that is filtered but no amount is either 1) secreted into the nephron, or 2) is not reabsorbed from the nephron). Inulin is an example. We could give a person some inulin and then measure the urine concentration of inulin (U_{in}), the volume of urine generated per time (\dot{V}), and the plasma concentration of inulin (P_{in}) to compute the GFR

$$\text{GFR} = \frac{U_{in} \dot{V}}{P_{in}} \quad (5.2)$$

(Note that I use \dot{V} and not V to make it crystal clear that this is a measure of the volume of urine produced per time not simply a volume).

Using this information, compute the GFR for a person in which 1) inulin was given continuously to generate a constant plasma concentration of 1.0 mg/dL. 1.6 L of urine was collected over a 10 hour period. The urinary concentration of inulin was 462 mg/L.

Chapter 6

Fluid and electrolyte balance

Chapter 7

Digestion, Nutrition, and Metabolism

7.1 Estimating Causal Effects

Think about headlines in human health, performance and disease: red wine decreases colon cancer, or coffee increases dementia, or oxygenated water increases marathon performance. These “conclusions” constantly seem to be flipping. To understand why, it is important to think about the problem with math. The *mathematical* way to think about these assertions is $X \rightarrow Y$, or “X causes Y”. Importantly, if a scientist says something like “X causes Y”, this does not mean that X is the only cause of Y – other things may also cause Y. For example: vegetarian diet \rightarrow low blood cholesterol AND running \rightarrow low blood cholesterol AND statins \rightarrow low blood cholesterol.

Most importantly “cause” is not binary (causes v. doesn’t cause) but has some magnitude (tivially small, or small, or big, or huge). Here, we use the greek letter β (“beta”) to indicate effect size.

We are going to use Google Sheets to create fake data that were generated by a known causal process (known β), and then use a **statistical model** to estimate the causal process (estimate β) from the fake data. The statistical model is **regression**, which is the principle statistical method used in the biological sciences to estimate causal effects. We are purposefully using abstract notation (X and Y) instead of meaningful variables (dietary cholesterol and atherosclerotic plaque development) because it is good to be able to think abstractly.

	A	B	C	D	E	F	G	H
1		True Value	Estimate			True Value	Estimate	
2	beta_0	0			beta_0	0		
3	beta_1	0.5	0.472		beta_1	0.5	-0.004	
4	E[b1]	0.5			beta_2	-0.7		
5					r	0.7	0.701	
6					E[b_1]	0.01		
7								
8	sd	0.9828666119	0.9735327391			1.0066371	0.9910731407	1.032272965
9	ID	X	Y		Z	X1	X2	Y
10	1	1.441519473	0.7105884214		-2.957506448	-1.450754737	-2.42224558	0.1065029495
11	2	-1.067223794	-0.3866287056		-1.024169967	-1.095180579	-1.44544354	0.1770615012
12	3	1.690301322	0.8807498814		1.655578157	1.308331782	1.883125131	-1.010955082
13	4	-0.2457007644	0.2254569788		0.2663960621	-1.002803094	-0.5660069822	0.8461727216
14	5	-1.306222918	-0.8542635306		0.2086526157	-0.2589033115	0.7362888197	0.1166237296
15	6	-0.2527736106	-0.5132589204		0.9056618838	-0.1066822945	1.550229258	-1.604065218

7.2 Simulation 1

Open your Google spreadsheet and

7.2.1 Step 1. Set up the parameters

1. In column A, cells 2-4, insert “beta_0”, “beta_1”, “E[b1]” (see figure above)
2. In row 1, columns B and C, insert “True Value”, “Estimate”
3. In B2, insert a number (it doesn’t matter)
4. In B3, insert 0.5 (this is the true generating effect of X on Y)
5. In B4, insert =B3 (this is the expected value of the generating effect of X on Y given the statistical model)

7.2.2 Step 2. Generate fake data

1. In row 9, columns A-C, insert “ID”, “X”, “Y”
2. In A10 insert “1”
3. In B10 insert =normsinv(rand())
4. In C10 insert =\$B\$2 + \$B\$3*B10 + sqrt(1-\$B\$3^2)*normsinv(rand())
5. In A11 insert =A10 + 1
6. Highlight cells B10 and C10. Click on the handle on the lower right corner of the box and drag down 1 row. Your formulas from row 10 should now be in row 11.
7. Highlight cells A11, B11, C11. Click on the handle on the lower right corner of the box and drag down and down and down until you get to row 1000. You should have copied all three formulas all the way down.

What is step 2 doing? It is creating fake data. The value is caused by three things, the value in Cell B2, the product of B3 and X, and a random number. The value in B3 is the contribution of X to Y or how “X causes Y” or the “causal effect of X on Y”. If B3 is 0 then there is no causal effect. If B3 is 1 or -1, then the random component is zero.

You have just created fake data with a known generating mechanism! But it is imperative to check the equations you entered don’t have bugs. If the equations were entered correctly, the standard deviation of the X and Y columns should both be one. Check this

7.2.3 Step 3. Fake data check

1. In A8, insert “sd”
2. In B8, insert =stdev(B10:B1000)
3. Copy B8 and paste in C8.

These numbers should be close to 1.0 (something is probably wrong if it is less than 0.95 or more than 1.05). Refresh the spread sheet by typing command-R (Mac) or control-R (Windows)

7.2.4 Step 4. Does a statistical model recover the known effect?

1. In C3, insert =slope(C10:C1000, B10:B1000)
2. In C3, round to three places after the decimal

This is the slope of the regression (the statistical model) of Y on X. It is the **estimate** of the causal effect. The number should be very close to the true value.

This slope is the **regression coefficient** b_1 . The cell labeled “E[b1]” is the “expectation of b_1 ” or the expected value of b_1 . Your estimate of β_1 should also be very close to $E(b_1)$ since $E(b_1)$ is equal to the true generating effect (β_1).

7.2.5 What you did

7.2.5.1 ... in a nutshell

you generated Y using a “data generating” mechanism and then using the available data (X and Y), you used a statistical analysis to see if you could recover this data generating mechanism. The data generating mechanism is the set of two coefficients β_0 and β_1 .

7.2.5.2 the data generating mechanism in a little more detail

The fake data are two variables, X and Y . Y is caused by three things:

$$y_i = \beta_0 + \beta_1 x_i + \sigma_i \quad (7.1)$$

the subscript is the “ i th” individual (if ID=7 then $i=7$). The three components generating y_i are

1. β_0 is “the intercept”; it is common to all i
2. $\beta_1 x_i$ is the product of the effect (β_1) and an individual's value of x . β_1 is the same for all i but the product is unique to each i .
3. σ_i is “the error”; this is the random variation due to other factors that “cause” Y but are unique to each i . That is, these factors are **not correlated** with X .

7.2.6 The model you fit is

$$y_i = b_0 + b_1 x_i + e_i \quad (7.2)$$

1. b_0 is the intercept
2. b_1 is the slope
3. e_i is the residual (the difference between the modeled value and the actual value)

Notice that the statistical model is the same as the generating model. It is not at all surprising that the statistical model “recovers” the data generating mechanism (or the “true values”). **The problem in science is, we don’t know the data generating model so we don’t know the correct statistical model.** This will hopefully make more sense in the next exercise.

7.3 Simulation 2

7.3.1 Step 5. Set up the parameters

1. In column E, rows 2-6, insert the labels “beta_0”, “beta_1”, “beta_2”, “r”, “E(b_1)”
2. In row 1, columns F and G, insert the labels “True Value”, “Estimate”
3. In F2, insert a number (it doesn’t matter) (this is the baseline value of generating model)
4. In F3, insert 0.5 (this is the true generating effect of X_1 on Y)
5. In F4, insert -0.7 (this is the true generating effect of X_2 on Y)
6. In F5, insert 0.7 (this is the true correlation between X_1 and X_2)

7.3.2 Step 6. Generate fake data

1. In row 9, columns E-H, insert “Z”, “X1”, “X2”, “Y”
2. In E10 insert `=normsinv(rand())`
3. In F10 insert `=sqrt(F5)*$E10 + sqrt(1-$F$5)*normsinv(rand())`
4. In G10, copy the equation from F10 and insert into G10
5. In H10, insert `=F2 + F3*F10 + F4*G10 + sqrt(1-F3^2 - F4^2 - 2*F3*F4*F5)*normsinv(rand())`
6. Highlight cells E10 through H10. Click on the handle on the lower right corner of the box and drag down and down and down until you get to row 1000. You should have copied all four formulas all the way down.

What is step 6 doing? Like Step 2 in Simulation 1 above, it is creating fake data. But here the Y value is caused by five things:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_i \quad (7.3)$$

1. β_0 is “the intercept”; it is common to all i
2. $\beta_1 x_{1i}$ is the product of the effect (β_1) and an individual's value of x_1 . β_1 is the same for all i but the product is unique to each i . This is the causal or generating effect of X_1 on Y
3. $\beta_2 x_{2i}$ is the product of the effect (β_2) and an individual's value of x_2 . β_2 is the same for all i but the product is unique to each i . This is the causal or generating effect of X_2 on Y
4. σ_i is “the error”; this is the random variation due to other factors that “cause” Y but are unique to each i . That is, these factors are **not correlated** with X .

what is the 5th cause of Y ?

5. r – the correlation between X_1 and X_2 . A correlation is a measure of association and is always between -1 and 1

7.3.3 Step 7. Fake data check

1. Check the standard deviation of X_1 , X_2 , and Y as in Step 3 above. All of these should be close to 1.0
2. insert `=correl(F10:F1000, G10:G1000)` in G5. This should be close to the true correlation in F5 (The starting correlation is 0.7, so the estimate should be 0.67-0.73)

7.3.4 Step 8. Does a statistical model recover the known effect?

1. In G3, insert `=slope(H10:H1000, F10:F1000)`
2. In G3, round to three places after the decimal
3. In F6, insert `=F3 + F4*F5`

As in Step 4 in Simulation 1 above, this is the slope of the regression (the statistical model) of Y on X . It is the **estimate** of the causal effect. The number will not be very close to the true generating value (β_1), at least using the default values specified in Step 5. But it should be close to $E(b_1)$ (the expected value of b_1), given the statistical model. But, unlike simulation 1, $E(b_1)$ is not similar to β_1 , the true generating effect. Huh?

1. $E(b_1)$ should not equal the true value of β_1 (at least using default values in Step 5), unlike in Simulation 1.
2. Your estimate of β_1 should be very close to $E(b_1)$ but not to β_1

7.3.5 What's going on is the whole point of this exercise

You have measured Y and X_1 but have not measured X_2 . Because you haven't measured X_2 , it is *not* in your statistical model, so your statistical model is just like that in Simulation 1.

$$y_i = b_0 + b_1 x_{1i} + e_i \quad (7.4)$$

But the generating model for Y is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_i \quad (7.5)$$

That is, your statistical model has an **omitted causal variable** (X_2) and your estimate of the effect of X_1 is **biased**. This kind of bias is called **omitted variable bias**. The true effect of X_1 on Y is β_1 but you are actually estimating $E(b_1)$ with the regression coefficient! Researchers often think their result will get closer to the truth as the sample size increases but if an causal effect is missing from a statistical model, the estimate of the effects of the factors in the model will never ever get closer to the truth - instead it gets closer to the wrong thing (the biased expectation of the effect given the model).

7.4 Questions

7.4.1 Simulation 1

1. Given the default parameters specified above, what is the estimated effect of X_1 on Y in Simulation 1?
2. What is the true effect of X_1 on Y in Simulation 1?
3. If you increase your sample size, will the estimated effect of X_1 on Y move toward the true effect of X_1 on Y in Simulation 1?

7.4.2 Simulation 2

4. Given the default parameters specified above, what is the estimated effect of X_1 on Y in Simulation 2?
5. What is the true effect of X_1 on Y in Simulation 2?
6. If you increase your sample size, will the estimated effect of X_1 on Y move toward the true effect of X_1 on Y in Simulation 2?
7. If you did a study of X_1 on Y and the true generating model of Y is that in Simulation 2, what would you conclude about the effect of X_1 on Y ?

7.4.3 Simulation 2 with new parameters

Redo Simulation 2 with the parameters: $\beta_1 = 0.0$. Leave $\beta_2 = -0.7$ and $r = 0.7$.

8. What is the estimated effect of X_1 on Y ?
9. What is the true effect of X_1 on Y in Simulation 2?
10. If you did a study of X_1 on Y and the true generating model of Y is these new parameters in Simulation 2, what would you conclude about the effect of X_1 on Y ?

Chapter 8

Metabolism

Chapter 9

Physiological Genetics

Chapter 10

Cancer