Math review for new biostatistics students

Patrick Breheny

Contents

Chapter 1

Introduction

This guide is intended as a review of fundamental math concepts for students who will be starting an MS or PhD program in biostatistics. More specifically, its intended audience is new students at the University of Iowa, but the material here is quite general and I would expect it to be useful to any new student in a biostatistics program regardless of where it is.

Why a math review? Is math the most important skill for a statistician? Not necessarily. However, in our experience, a shaky/rusty foundation in math is the thing most likely to lead to problems in the first year of graduate school. When you encounter new statistical concepts, instructors will introduce and explain them. But the mathematical techniques this guide covers, they will assume you already know.

This guide focuses in particular on two areas of mathematics, and for different reasons. Calculus, because it is a big topic – students often take Calc I, Calc II, and Calc III. That's a lot of material and it's not clear what needs to be reviewed and what can be skipped. Also matrix algebra, because it tends not to be taught very well at the undergraduate level. Perhaps more accurately, courses tend to focus on old-fashioned topics like inverting matrices by hand and not on the kinds of manipulations that one uses in statistics.

In principle, any idea from math *could* come up and be helpful to a statistician. In reality, however, certain ideas come up far more often than others, and this guide focuses on topics of greatest relevance. A good example is trigonometry: this almost never comes up in statistics. There is really no need to spend any time whatsoever reviewing it prior to stating a graduate program in statistics. On the other hand, properties of exponents and logarithms come up *constantly*. You need to know every property, because you will use them more or less every day, and if you don't know them, you will be constantly making errors on all of your homework and tests.

Finally, the focus here is really on the math – as noted above, we expect to teach you the statistics once you get here. However, to help make connections, I will occasionally point out the relevance of certain concepts to the field of biostatistics. If you've taken statistics in the past and terms like "independent events" and "regression model" mean something to you, great. If not, however, don't worry about it. You can appreciate the connection later once you learn about these ideas in graduate school.

To reiterate: this guide is intended to be a concise *review* of main ideas in calculus. If a section is unfamiliar or confusing, it would probably be a good idea to read the corresponding section of your calculus textbook, which will have a lot more examples, explanations, graphs, etc. Also, while we may add exercises and solutions in the future, there aren't any right now. Obviously, exercises are extremely helpful, especially if you feel you are rusty on a particular section. We recommend either finding problems from your calculus book or purchasing a book such as Schaum's Outline of Calclus.

Finally, if you spot any mistakes or typos, please let me know!

Chapter 2

Calculus

In this chapter, we will review/collect a large number of results that you should know and be familiar with from calculus. I'm not going to prove them or provide a bunch of details and explanations and graphs, so if anything strikes you as unfamiliar or you want more details, please consult your calculus textbook.

Calculus is important to statistics for lots of reasons, but I would like to point out three major ones before we begin the review.

Finding most likely solutions

A statistical analysis typically begins with some sort of model for how the data (which I'll call d) depends on an unknown parameter (which I'll call θ). We observe the data, but what's θ? To estimate it, we typically create some function (which I'll call f) that is large when θ is in agreement with the data we've seen. To find the "best" value of θ, we can take the derivative of f to find the optimal value. Note that this is actually a partial derivative, since f would be a function of both θ and d.

Probability and density

For continuous quantities such as height, the distribution of likely values is specified in terms of a probability density f. Calculating the probability from a probability density involves integration. For example, if we wanted to know the probability that a person's height was between 63 and 66 inches:

$$\int_{63}^{66} f(x)\,dx.$$

Independent observations

Suppose we are interested in the probability of events $A_1, A_2, \ldots, A_n$. If those events are independent, this is given by

$$P(A_1)P(A_2)\cdots P(A_n) = \prod_{i=1}^{n} P(A_i).$$

However, it is almost *always* easier to deal with this kind of quantity after taking the log:

$$\log\left(\prod_{i=1}^{n} P(A_i)\right) = \sum_{i=1}^{n} \log P(A_i).$$

To see why, go ahead and multiply a bunch of probabilities together and see how useful the result is to work with. The same trick can be used with dependent terms as well, although the results are messier.

It is hard to overstate how often you will do this. This isn't some occasional trick – this is standard operating procedure, so it is critical that you know the properties of exponents and logs extremely well.

2.1 Functions

The concept of a function is not difficult or foreign, but since it's the most important concept in all of mathematics, it's worth reviewing and knowing the formal definition.

Definition: Given two sets, A and B, a *function* (or *map*) is a rule that assigns, to each element in A, exactly one element from B. The set A is called the *domain* of the function and the set B is called the *range*. This is represented by the mathematical notation $f : A \to B$.

Commentary

A few remarks on this definition and its implications:

1. This is an *extremely* general definition. A and B could sets of numbers, but they could collections of numbers... A could be a set numbers and B could consist of intervals, with each interval being itself an infinite collection of numbers. Or A and B might not involve numbers at all. They can be *anything*. The only restriction is that given the same input $x \in A$, we always get the same output $f(x) \in B$.
2. Sometimes domains are obvious from context and not explicitly specified, but it's an important part of the function. For example, consider the function $f(x) = \sqrt{x}$. This is not a function that works for all numbers – in particular, it doesn't work for negative numbers. The domain, then, is the set of non-negative numbers $\{x : x \geq 0\}$. Functions don't have to be defined everywhere, they just need to work on their domain[1].
3. Keep in mind that a function needs to be defined for *every* element in its domain. This can get complicated, especially if your function is the integral of another function (as probability functions are). It's tempting to say, "The domain of my function is 'any set of numbers'. You enter a set, it returns a value." However, this is dangerous – a devious troublemaker could say, "Oh? How about the set of transcendental numbers?" Do you really want to be responsible for defining the value of your function for such complicated sets? Sometimes you need to limit the domain to make defining the function easier. Keep this in mind when you encounter things like "sigma algebras", typically one of the most bewildering concepts to grasp for first-year students.

If this seems very abstract, don't worry too much about it – for the purposes of this review, domain and range will almost always be sets of single numbers, but it's worth keeping an open mind about what functions can represent, since at various points in your education you may encounter other kinds of functions, especially functions that map vectors or matrices to numbers (or to other vectors or matrices).

Inverse functions

Recall that for a given input $x \in A$, the function must always return the exact same element of B. The converse, however, is not true: there may be lots of elements of A that all get mapped to the same element of B. For example, in statistics one often encounters "indicator functions" that can have various types of things as input but always return a 0 or 1 as output (i.e., the range of an indicator function is the set $\{0, 1\}$).

Now, if it **is** the case that whenever $x_1 \neq x_2$, we have $f(x_1) \neq f(x_2)$, then this is a special class of function called a *1:1 function*. Such functions are important because they have *inverses*: there exists a function f^{-1} such that whenever $f(a) = b$, we have $f^{-1}(b) = a$. A function has an inverse if and only if it is 1:1. This is important to be aware of, since there are a number of important results involving inverses, but be aware that

[1] We *could* extend the domain of the function to include negative numbers, but then the range would have to include complex numbers.

not all functions have inverses. For example, $f(x) = x^2$ does not have an inverse: $f^{-1}(4)$ could be either 2 or -2[2].

2.2 Limits and continuity

Limits

Definition: We say that the **limit** of a function $f(x)$, as x approaches a, is L if we can make the values of $f(x)$ get as close as we want to L by taking x sufficiently close to a (but not equal to a)[3]. Mathematically, we can express this idea as

$$\lim_{x \to a} f(x) = L.$$

For example, if $f(x) = x^2$, then it is the case that

$$\lim_{x \to \sqrt{5}} f(x) = 5.$$

Suppose we set x equal to 2.236 (this is close to $\sqrt{5}$ but not equal). Then $f(x) = 4.999696$, which is close to 5. There is no value of x other than $\sqrt{5}$ such that $f(x) = 5$, but we can get as close as we want by moving x closer to $\sqrt{5}$. For example, if 4.999696 isn't close enough to satisfy us and someone demands that we be within 0.000000001 of 5, we can always accomplish that by simply moving x closer to $\sqrt{5}$.

Infinite limit: A variation on this idea is to say that the limit is infinite:

$$\lim_{x \to a} f(x) = \infty.$$

This means that as x gets closer to a, $f(x)$ keeps getting bigger, with no bound. For example, we can make $1/x^2$ be as large as we want by moving x closer to 0, so $\lim_{x \to 0} 1/x^2 = \infty$ (limits of $-\infty$ are defined similarly).

One-sided limit: Sometimes, different things happen if we approach a from the left or right. We say that the **left-hand limit** of $f(x)$ as x approaches a "from the left" is L if $f(x)$ we can make the values of $f(x)$ as close to L as we want by moving x closer to a, but only considering points such that $x < a$. We denote this by

$$\lim_{x \to a^-} f(x) = L.$$

Right-hand limits are defined similarly. For example $\lim_{x \to 0^-} 1/x = -\infty$, whereas $\lim_{x \to 0^+} 1/x = \infty$.

The limit of $f(x)$ as $x \to a$ is L if and only both the left and the right-hand limits are also L.

Calculating limits

Limit laws: The following laws are helpful for calculating limits. In what follows, let

$$s = \lim_{x \to a} f(x) \tag{2.1}$$

$$t = \lim_{x \to a} g(x); \tag{2.2}$$

it is critical that these limits exist, or none of the results below necessarily hold.

[2]The astute reader may note that $f(x) = x^2$ *could* be 1:1 if I restrict its domain to be, say, $(0, \infty)$.

[3]This section covers limits and continuity from a conceptual standpoint. For a variety of technical reasons, the definition given here isn't actually satisfactory, and a more rigorous definition is required; see the chapter on analysis.

$$\lim_{x\to a}\{f(x)+g(x)\} = s+t \tag{2.3}$$

$$\lim_{x\to a}\{f(x)-g(x)\} = s-t \tag{2.4}$$

$$\lim_{x\to a}\{cf(x)+g(x)\} = cs \text{ where } c \text{ is a constant} \tag{2.5}$$

$$\lim_{x\to a}\{f(x)g(x)\} = st \tag{2.6}$$

$$\lim_{x\to a}\frac{f(x)}{g(x)} = \frac{s}{t} \text{ if } t \neq 0 \tag{2.7}$$

$$\lim_{x\to a}\{f(x)^n\} = s^n \tag{2.8}$$

Continuity

You may have noticed that with limits, the value of $f(x)$ at a is irrelevant. For example, if $f(x) = x^2$ everywhere except $x = 2$, where $f(2) = -10$, it would still be the case that $\lim_{x\to 2} f(x) = 2$. In fact, $f(x)$ wouldn't even need to be *defined* at 2 for this to work. If we add the requirement that $f(a)$ has to equal its limit, we end up with continuity.

Definition: A function f is **continuous at** a if

$$\lim_{x\to a} f(x) = f(a).$$

Note that this requires three things:

1. $f(a)$ is defined
2. $\lim_{x\to a} f(x)$ exists
3. These two things are equal

Expanding on this definition, we say that a function f is **continuous on an interval** if f is continuous at every number in the interval. We say that f is **continuous** if f is continuous at every point in its domain.

One-sided continuity: A function f is **continuous from the left at** a if

$$\lim_{x\to a^-} f(x) = f(a).$$

For example, consider the function

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

In this case, $f(x)$ is continuous from the right at 0, but not from the left at 0 (since $\lim_{x\to 0^-} = 0$, but $f(0) = 1$).

Continuity laws: The property of continuity behaves similarly to the limit laws above. If $f(x)$ and $g(x)$ are continuous at a, then the following functions are also continuous at a:

- $f(x) + g(x)$
- $f(x) - g(x)$
- $cf(x)$, where c is a constant
- $f(x)g(x)$
- $f(x)/g(x)$ if $g(a) \neq 0$

Composition: Finally, suppose that g is continuous at a and that f is continuous at $g(a)$. Then $f(g(x))$ is continuous at a. In words, a continuous function of a continuous function is continuous. The function $h(x) = f(g(x))$ is known as the *composition* of f and g.

2.3 Derivatives

Definition

The slope of a straight line is straightforward: $\Delta y/\Delta x$. For a curved line, however, we will get different answers depending on the range over which we calculate these changes. Nevertheless, we can calculate the limit of this slope over shorter and shorter ranges. This is known as the derivative of the function.

Definition: The **derivative of a function** f **at** a, denoted $f'(a)$, is

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

if this limit exists.

If the limit exists, f is said to be **differentiable at** a. If a function is not continuous at a, it is not possible for it to be differentiable at a. The converse, however, is not true. For example, the function $f(x) = |x|$ is continuous everywhere, and differentiable just about everywhere, but not differentiable at zero since the limit from the left is -1 and the limit from the right is 1.

Expanding on this pointwise definition, we can define a whole function, $f'(x)$. This function is known as the **derivative of** f.

Formulas

It is hard to overstate the importance of knowing the following formulas; you will use them constantly. Below, we assume that f and g are differentiable, and that c is a constant.

$$c' = 0 \tag{2.9}$$
$$(x^n)' = nx^{n-1} \tag{2.10}$$
$$(cf)' = cf' \tag{2.11}$$
$$(f+g)' = f' + g' \tag{2.12}$$
$$(f-g)' = f' - g' \tag{2.13}$$
$$(fg)' = fg' + gf' \text{ (product rule)} \tag{2.14}$$
$$\left(\frac{f}{g}\right)' = \frac{gf' - fg'}{g^2} \text{ (quotient rule)} \tag{2.15}$$
$$\tag{2.16}$$

These basic rules can be combined into all sorts of additional rules with the *chain rule*, which states that if the derivatives $g'(x)$ and $f'(g(x))$ both exist, then the derivative of $f(g(x))$ exists, and its derivative is $f'(g(x))g'(x)$. The rule is often expressed in Leibniz notation:

$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx}.$$

The section on logarthim and exponential functions provides additional important differentiation formulas.

Higher derivatives

Since $f'(x)$ is itself a function, we can also take *its* derivative. This is called the **second derivative** of f, and is denoted $f''(x)$.

Third derivatives, fourth derivatives, and so on are defined similarly.

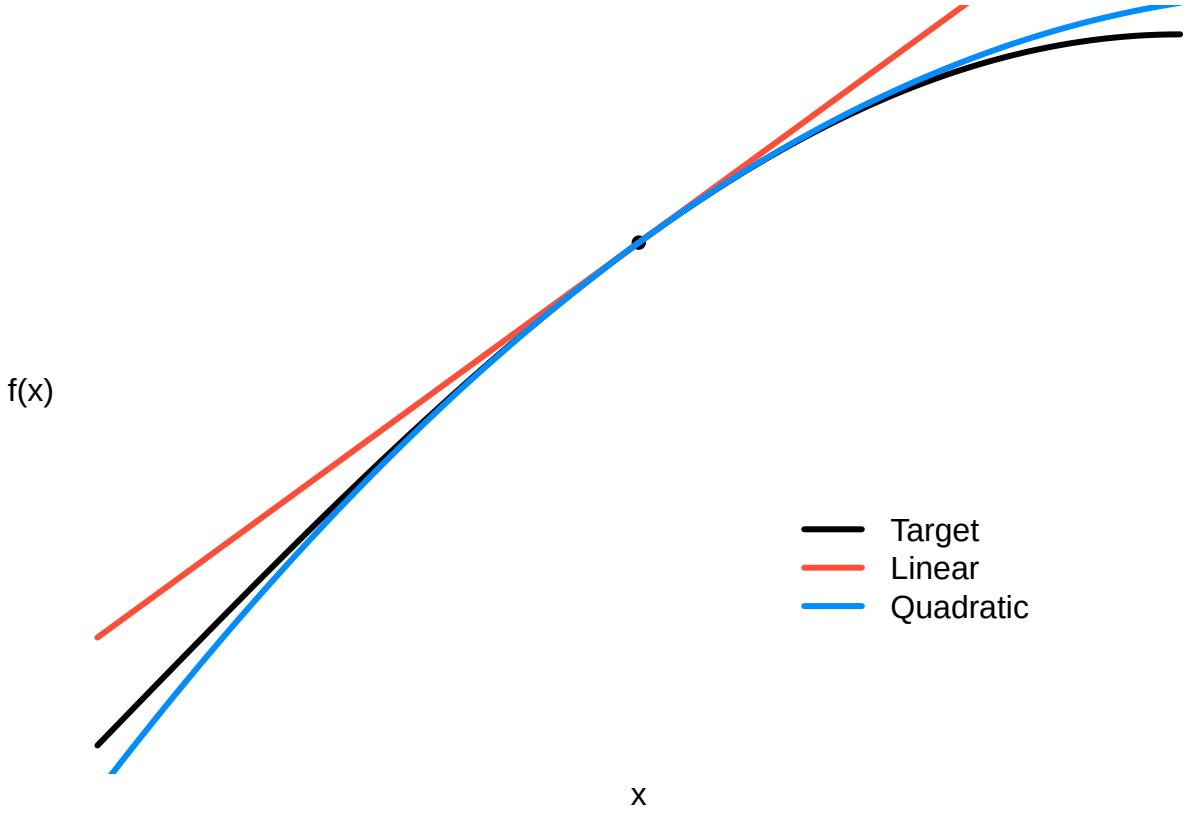
An important application of higher derivatives is to approximate functions. The **linear approximation** of f at a is given by

$$f(x) \approx f(a) + f'(a)(x - a).$$

The **quadratic approximation** of f at a is given by

$$f(x) \approx f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2.$$

To see these approximations in action, here's a figure:



Note that (a) both approximations are very good close to a, which is denoted by the black dot and (b) the quadratic approximation is more accurate than the linear approximation. Both of these observations are true broadly speaking; they are not particular to this example.

2.4 Optimization

Terminology

The most useful thing about derivatives is that they enable us to find the maximum and minimum values of a function. As noted earlier, this arises constantly in statistics. First, some terminology (below, f is a function with domain D):

- **absolute maximum**: The point c is an absolute maximum of f if $f(c) \geq f(x)$ for all x in D.
- **maximum value**: The maximum value of f is $f(c)$, where c is an absolute maximum of f.
- **local maximum**: The point c is a local maximum (or relative maximum) of f if there is an interval I containing c such that $f(c) \geq f(x)$ for all x in I.

Absolute minimum, minimum value, and local minimum are defined similarly. Finally, a point c is an **extreme value** if c is either an absolute maximum or an absolute minimum, while c is a **local extremum** if c is a local maximum or local minimum.

Derivatives and extreme values

What does this have to do with derivatives? The following result is so important, you should memorize it word for word and never forget it.

> If f has a local extremum at c, and if $f'(c)$ exists, then $f'(c) = 0$.

A point c satisfying $f'(c) = 0$ is called a *critical point* of f. Practically speaking, this means that if we want to maximize or minimize a function, we just need to find its critical points. However, we do need to be aware of a few caveats:

1. The derivative has to exist. For example, we cannot minimize $f(x) = |x|$ with derivatives, because the minimum occurs at 0 and f is not differentiable at 0.
2. The converse of the above statement is not true. It is true that if c is a local extremum (and f differentiable), then c is a critical point. However, c can be a critical point without being an extremum. For example, 0 is a critical point of $f(x) = x^3$, but it is not a local minimum or maximum.
3. If we find a critical point c, even if it is an extremum, we don't know whether it minimizes or maximizes f.
4. The function f might not have any critical points.

More information about caveats 3 and 4 is given below.

Don't led these caveats obscure the main result, though – this is arguably the most useful thing in all of calculus.

Monotonicity and convexity

Monotone functions: If f is differentiable, why wouldn't it have any critical points (#4 above)? The most likely answer is that it is monotone. A function f is called **increasing** if $f(x_1) < f(x_2)$ for all $x_1 < x_2$ and **decreasing** if $f(x_1) > f(x_2)$ for all $x_1 < x_2$. A function that is either increasing or decreasing is called **monotone**.

For a differentiable function, whether it is monotone or not is related to its derivative:

- If $f'(x) > 0$ for all x, then f is increasing.
- If $f'(x) < 0$ for all x, then f is decreasing.

So there you have it. If f is differentiable, there are three possibilities: it is always going up ($f'(x) > 0$), always going down ($f'(x) < 0$), or sometimes going up and sometimes going down, in which case it will cross zero and have a critical point (due to the intermediate value theorem.)

** Tests for min/max:** Often, it is obvious whether a critical point c is a minimum or maximum. However, if you're not sure, you can do one of two things:

1. Plug in a number less than c, then greater than c. If f' changes from negative to positive, c is a local maximum. If it changes from positive to negative, c is a local minimum. If it does not change sign, c is not a local extremum. This is known as the "first derivative test".
2. Take the second derivative at c (assuming it exists). If $f''(c) > 0$, then c is a local minimum. If $f''(c) < 0$, then c is a local maximum. This is known as the "second derivative test". Note that if $f''(c) = 0$, the test is inconclusive – c could be a local max, a local min, or neither.

Convexity and concavity: If a function is always curving upwards or downwards, then no tests are needed and no distinctions between local and global extrema are necessary. To define this formally, imagine drawing a tangent line to a function f at every point in its domain. If f always lies above the tangent line, it is said to be **convex** (curving upwards). If f always lies below the tangent line, it is **concave** (curving downwards). With respect to optimization,

- If f is convex, then any critical point is a global minimum.
- If f is concave, then any critical point is a global maximum.

Some textbooks / math classes refer to these as "concave up" and "concave down", but you should learn concave/convex since it is far more common in the statistics, mathematics, and optimization literature.

Optimization is an enormous subject with giant textbooks devoted to it, so obviously this isn't the whole story. However, taking the derivative and setting it equal to zero truly is the main idea, and solves a huge range of optimization problems.

2.5 Integration

There are two core questions with which calculus is concerned. One is generalizing the idea of slope to nonlinear functions. The other is how to calculate the total contribution of some entity, where the contribution at any given instant is given by a function. As with slopes, this is trivial if the function is linear and becomes much harder when the function is nonlinear. For example, if someone burns 700 calories/hr while exercising, and they exercise for half an hour, then they burn 350 calories. But what if their exercise intensity varies over time, with $f(t)$ describing the rate at time t (in minutes)? In this case we would have to add the contributions:

$$f(0)\frac{1}{60} + f(1)\frac{1}{60} + \dots.$$

However, this still doesn't really answer the question, as it assumes $f(t)$ is constant over the first minute, then allowed to change, then constant again for the next minute, and so on. We could get a more accurate answer by summing up these contributions at each second, and still more accurate by summing over each nanosecond, and so on. The limit of this process is known as the "integral", which we define below.

As noted earlier, this comes up constantly in statistics when calculating probabilities and expected values.

Definition

Let the interval $[a, b]$ be partitioned as follows:

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b,$$

let x_i^* be any point in $[x_{i-1}, x_i]$, $\Delta x_i = x_i - x_{i-1}$, and $m = \max\{\Delta x_1, \Delta x_2, \dots, \Delta x_n\}$. Then the **integral of f from a to b** is

$$\int_a^b f(x)\,dx = \lim_{m\to 0}\sum_{i=1}^{n} f(x_i^*)\Delta x_i$$

if this limit exists. If the limit does exist, then f is said to be **integrable** over $[a, b]$.

Relating this definition to our example above, m represents the time resolution and $\lim_{m\to 0}$ represents moving from minutes to seconds to nanoseconds and so on.

The above definition assumes that $a < b$; if $a > b$ the integral is defined as

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

How can we know if a function is integrable?

> If f is either continuous or monotonic on $[a, b]$, then f is integrable on $[a, b]$.

If f is jumping up and down discontinuously, then anything can happen – f may or may not be integrable, and we would need a deep dive into the theory of integration to really answer this question. Thankfully, as a first-year graduate student, you will only ever need to integrate continuous functions.

Properties of integrals

If all of the following integrals exist, then they obey these rules:

$$\int_a^b c\,dx = c(b-a) \tag{2.17}$$

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx \tag{2.18}$$

$$\int_a^b \big\{f(x)+g(x)\big\}\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx \tag{2.19}$$

$$\int_a^b \big\{f(x)-g(x)\big\}\,dx = \int_a^b f(x)\,dx - \int_a^b g(x)\,dx \tag{2.20}$$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \tag{2.21}$$

If we further suppose that $a < b$, then we also have

- If $f(x) \geq 0$ for all x, then $\int_a^b f(x)\,dx \geq 0$.
- If $f(x) \geq g(x)$ for all x, then $\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx$.
- If $m \leq f(x) \leq M$ for all x, then

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a)$$

- $\left|\int_a^b f(x)\,dx\right| = \int_a^b \big|f(x)\big|\,dx$

Finally, if $a = b$, then

$$\int_a^b f(x)\,dx = 0.$$

Fundamental theorem of calculus

Somewhat remarkably, the two branches of calculus (differentiation and integration) are closely related. This relationship is known as the **fundamental theorem of calculus**:

> If f is continuous on $[a,b]$, then
>
> $$g(x) = \int_a^x f(t)\,dt$$
>
> is continuous and differentiable and $g'(x) = f(x)$.

In other words, if we integrate a function, then differentiate the result, we get back to the original function. The same is true if we start with differentiation:

If f is continuous on $[a, b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

where F is any function that satisfies $F' = f$.

Functions satisfying $F' = f$ are particularly important, and discussed below.

Antiderivatives

A function F is called an **antiderivative** (or a **primitive**) of f if $F'(x) = f(x)$ for all x. This can also be written

$$\int f(x)\,dx = F(x).$$

This "equation" means the same thing, that $F'(x) = f(x)$ for all x. However, it is not truly an equation since there are an infinite number of functions F that satisfy $F'(x) = f(x)$ for all x. For example, both of the following are correct:

$$\int 2x\,dx = x^2 \tag{2.22}$$

$$\int 2x\,dx = x^2 + 5. \tag{2.23}$$

$$\tag{2.24}$$

This is potentially confusing because the left hand side is the same in each case, but the right hand side is different – hence the scare quotes around "equation". Some people prefer to write

$$\int 2x\,dx = x^2 + C$$

to emphasize this point. Whether you do this or not is up to you, but either way, it is critical to understand the distinction between $\int_a^b f(x)\,dx$ and $\int f(x)\,dx$. The first quantity (with the integration limits) is a known as a **definite integral**, and it is a number. The second quantity (without the limits) is known as an **indefinite integral**, and it is *not* a number – it is a function (or more precisely, a collection of an infinite number of functions)[4].

So what's the point of antiderivatives/primitives/indefinite integrals? If we have one, we can easily calculate (definite) integrals. For example,

$$\int_1^4 2x\,dx = F(4) - F(1) \tag{2.25}$$

$$= 4^2 - 1^2 \tag{2.26}$$

$$= 15. \tag{2.27}$$

[4]Students sometimes have a tendency to think of the indefinite integral as the "true" integral and the definite integral as an application of it. This is completely backwards. The indefinite integral is actually a statement about derivatives. We don't even need to define the concept of an integral in order to say that $\int 2x\,dx = x^2$. It is only once the integral has been defined and the fundamental theorem of calculus has been proven that indefinite integrals have any purpose. The "definite" integral defines the concept of the integral; the only reason we add the "definite" modifier to distinguish them from indefinite integrals.

Note that I get the same answer regardless of which antiderivative I use (i.e., it's not important to find the collection of all antiderivates... any antiderivative is fine).

In other words, we can integrate any function f if can find an antiderivative of it. How do we find these antiderivatives? Unfortunately, this is often challenging and sometimes impossible. However, there are several techniques for doing this, which will be discussed in a later section.

2.6 Logarithm and exponential

Exponential definition

The exponential function is $f(x) = a^x$; the *base* a must be a positive real number but the *exponent* x can be any real number. The precise definition doesn't come up often, but here it is in case you ever need it (defining all these cases is necessary in order to ensure that the resulting function is continuous):

1. If x is a positive integer n, then $a^n = a \cdot a \cdots a \quad (n \text{ times})$
2. If $x = 0$, then $a^0 = 1$
3. If x is a negative integer, then $a^{-n} = \frac{1}{a^n}$
4. If x is a rational number p/q, with $q > 0$, then $a^{p/q} = \sqrt[q]{a^p}$
5. If x is an irrational number, then it's defined as the limit of a^r, where r is a sequence of rational numbers whose limit is x.

Note that we would run into trouble at step 4 if we tried to allow negative bases.

Exponential rules

$$a^{x+y} = a^x a^y \tag{2.28}$$
$$a^{x-y} = \frac{a^x}{a^y} \tag{2.29}$$
$$(a^x)^y = a^{xy} \tag{2.30}$$
$$(ab)^x = a^x b^x \tag{2.31}$$

Exponential limits

$$\lim_{x\to\infty} a^x = \infty \quad \text{if } a > 1 \tag{2.32}$$
$$\lim_{x\to-\infty} a^x = 0 \quad \text{if } a > 1 \tag{2.33}$$
$$\lim_{x\to\infty} a^x = 0 \quad \text{if } 0 < a < 1 \tag{2.34}$$
$$\lim_{x\to-\infty} a^x = \infty \quad \text{if } 0 < a < 1 \tag{2.35}$$
$$\lim_{h\to 0} \frac{e^h - 1}{h} = 1 \tag{2.36}$$
$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e \tag{2.37}$$

Exponential derivatives and integrals

$$\frac{d}{dx}e^x = e^x \tag{2.38}$$

$$\frac{d}{dx}e^u = e^x\frac{du}{dx} \tag{2.39}$$

$$\int e^x\,dx = e^x \tag{2.40}$$

$$\frac{d}{dx}a^x = a^x\log(a) \tag{2.41}$$

$$\int a^x\,dx = \frac{a^x}{\log a} \quad (a \neq 1) \tag{2.42}$$

Note that the last two results use the logarithmic function, which we haven't actually introduced yet (see below).

Logarithm definition

The logarithmic function with base a is defined as the function satisfying

$$\log_a x = y \iff a^y = x$$

If we leave off the base, it is assumed to be base e, the "natural logarithm":

$$\log x = \log_e x$$

in other words,

$$\log x = y \iff e^y = x;$$

the notation $\ln x$ can also be used for this. In some disciplines, when we leave off the base, one assumes the base is 10; statistics is **not** one of those disciplines. Note that

$$\log(e^x) = x \tag{2.43}$$

$$e^{\log x} = x \tag{2.44}$$

$$\log e = 1. \tag{2.45}$$

Logarithm rules

$$\log_a(xy) = \log_a x + \log_a y \tag{2.46}$$

$$\log_a \frac{x}{y} = \log_a x - \log_a y \tag{2.47}$$

$$\log_a(x^y) = y\log_a x \tag{2.48}$$

$$\log_a x = \frac{\log x}{\log a} \tag{2.49}$$

Logarithm limits

If $a > 1$, then

$$\lim_{x\to\infty} \log_a x = \infty \tag{2.50}$$

$$\lim_{x\to 0^+} \log_a x = -\infty \tag{2.51}$$

$$\tag{2.52}$$

Logarithm derivatives and integrals

$$\frac{d}{dx} \log x = x^{-1} \tag{2.53}$$

$$\frac{d}{dx} \log u = u^{-1} \frac{du}{dx} \tag{2.54}$$

$$\int \frac{1}{x}\, dx = \log|x| \tag{2.55}$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \log a} \tag{2.56}$$

2.7 Improper integrals

In statistics, it is very common to encounter integrals that look like this:

$$\int_0^\infty f(x)\, dx.$$

This expression is a little confusing because if the integration region is infinite, then our earlier definition of the integral no longer works. What the expression means is that we're taking the limit of the definite integrals (all of which are well-defined) as the region gets larger and larger:

$$\int_0^\infty f(x)\, dx = \lim_{b\to\infty} \int_a^b f(x)\, dx.$$

For example,

$$\int_0^b e^{-x}\, dx = 1 - e^{-b};$$

(if you don't follow this derivation, see here and here). Since $\lim_{b\to\infty} e^{-b} = 0$ (see here), we have

$$\int_0^\infty e^{-x}\, dx = 1.$$

Integrals with an infinite bound (either upper or lower) are known as **improper integrals**. There are two other kinds of improper integrals.

Both bounds are infinite: One might expect that $\int_{-\infty}^{\infty} f(x)\, dx$ would be defined as the limit of $\int_{-a}^{a} f(x)\, dx$ as $a \to \infty$. But you'd be wrong! The actual definition is:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx,$$

provided that both improper integrals exist. To see why, suppose we wanted to calculate $\int_{-\infty}^{\infty} x\,dx$. This integral does not exist, even though $\lim_{a\to\infty}\int_{-a}^{a} x\,dx = 0$. The problem with saying that $\int_{-\infty}^{\infty} x\,dx$ equals 0 is that it depends entirely on how fast the upper and lower bounds are going to infinity. For example, $\lim_{a\to\infty}\int_{-a}^{2a} x\,dx = \infty$. Without more information on exactly how fast the upper and lower bounds are going to infinity, $\int_{-\infty}^{\infty} x\,dx$ could equal anything.

Unbounded functions: For technical reasons, we also run into problems when f is unbounded. Suppose we're interested in integrating $f(x) = 1/\sqrt{x}$, for example. At $x = 0$, $f(x)$ is undefined. Even if we were to define it, f wouldn't be continuous or monotone no matter what we chose, which causes problems with integration. As you might guess, however, we can extend our definition of the integral to include 0 as a lower bound by taking the limit as the bound goes to zero (if the limit exists):

$$\begin{aligned}
\int_0^1 x^{-1/2}\,dx &= \lim_{a\to 0}\int_a^1 x^{-1/2}\,dx && (2.57)\\
&= \lim_{a\to 0} 2\sqrt{x}\Big|_a^1 && (2.58)\\
&= 2\sqrt{1} - \lim_{a\to 0} 2\sqrt{a} && (2.59)\\
&= 2 && (2.60)
\end{aligned}$$

Note that in this case, if we failed to realize that f was not bounded over the integration region and blindly plugged in 0 anyway, it wouldn't make a difference – we'd get the same answer. However, this is not always true and if you ever run into a situation where this arises, it's important to know the proper definition.

2.8 Integration techniques

Every time we compute a derivative, we get a formula for integration. For example,

$$f(x) = \log(x^2+2) \implies f'(x) = \frac{2x}{x^2+2}.$$

This is great news if we are ever faced with the problem of calculating

$$\int_a^b \frac{2x}{x^2+2}\,dx,$$

but if we need to calculate $\int_a^b f(x)\,dx$, how can we reverse engineer a function $F(x)$ so that its derivative is $f(x)$?

Unfortunately, this task is sometimes easy, sometimes hard, and sometimes impossible (and you have no way of knowing in advance which situation you are in). One could create a huge table of integral formulas by taking derivatives of various things. Providing such a table is beyond the scope of this review, but such tables exist online and are useful resources to be aware of.

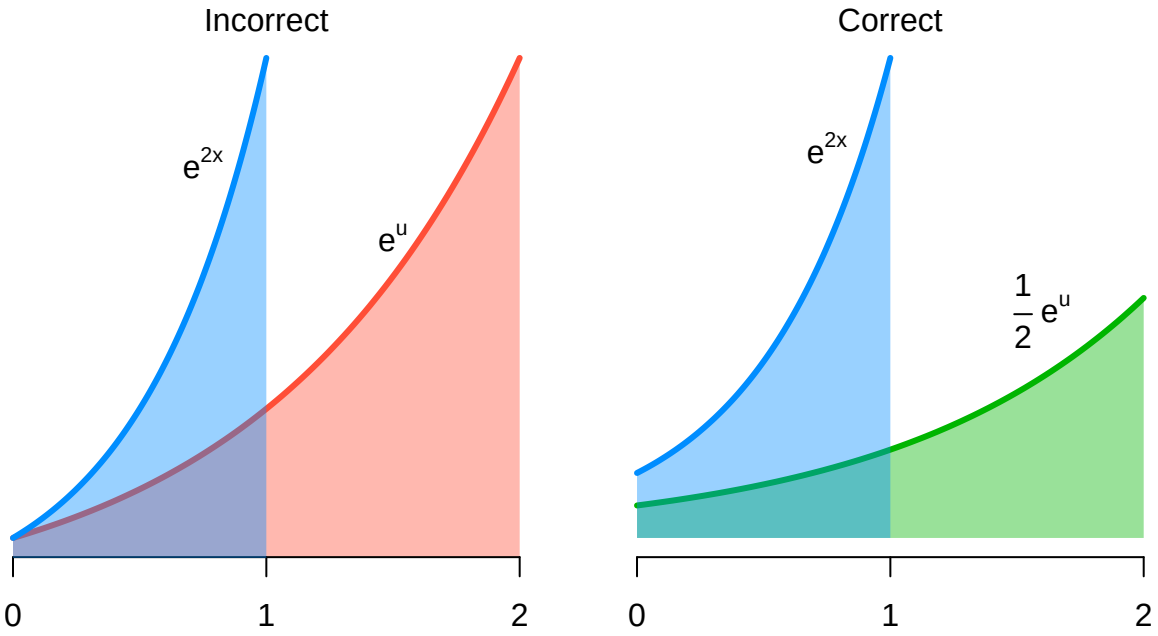
Even with such a table, however, there are a few useful integration techniques to be familiar with. Among other things, (a) it may be faster to use one of these techniques than looking up an integral (b) you might not have access to such a table at the moment, and (c) the form that appears in the table might be slightly different than what you need, and you might have to use one of these techniques in combination with the table to compute the integral.

Substitution

By far the most important technique to be aware of is substitution. For example, we know that $\int e^x\,dx = e^x$, but what if we have to find $\int e^{2x}\,dx$? Is it e^{2x}? The answer (and this is extremely important to understand, because it comes up in statistics all the time) is that no, it isn't. We can check this easily using the chain rule: the derivative of e^{2x} is $2e^{2x}$, so $\int e^{2x}\,dx \neq e^{2x}$.

In this case, it's also fairly clear what we need to do in order to fix the problem: $\int e^{2x}\,dx$ must be $\frac{1}{2}e^{2x}$: there must be a $1/2$ present to cancel the 2 that comes from the chain rule.

Conceptually, letting $u = 2x$, we can visualize what's going on here as follows. Each unit of u covers twice as much ground as a unit of x. If we don't do something to correct for this, we're going to artificially inflate the area under the curve integral (i.e., the integral). This is what's going on in the red region below, which clearly has greater area than the blue region (the integral we're trying to calculate).



However, if we compensate for this – we're stretching x out by a factor of 2, so we need to shrink the value of the function by a factor of 2 to preserve the correct area – we get the green region, which has the same area as the original blue region.

To formalize this thinking into a procedure, if $u = g(x)$, then (this works for any differentiable function g)

1. $du = g'(x)\,dx$
2. Substitute u for $g(x)$ and $1/g'(x)\,du$ for dx
3. Take the integral
4. Substitute $g(x)$ back for u

If we are calculating a definite integral, then instead of step 4, we can transform the limits of integration a and b to $g(a)$ and $g(b)$; this is usually preferable.

As practice, use this procedure to calculate

$$\int x(x^2-1)^5\,dx.$$

You should get $\frac{1}{12}(x^2-1)^6$.

Integration by parts

Just as the chain rule gave us substitution, the product rule gives us a formula called integration by parts, which is usually written in the form:

$$\int u\,dv = uv - \int v\,du.$$

As an example of integration by parts in action, suppose we want to integrate $\int \log x\,dx$. We can write this as

$$u = \log x \qquad dv = dx \tag{2.61}$$

$$du = \frac{1}{x}\,dx \qquad v = x \tag{2.62}$$

Thus,

$$\int \log x\,dx = x\log x - \int x\frac{dx}{x} \tag{2.63}$$

$$= x\log x - \int dx \tag{2.64}$$

$$= x\log x - x \tag{2.65}$$

As practice, use this procedure to calculate

$$\int xe^x\,dx.$$

You should get $xe^x - e^x$.

Kernel trick

The above techniques are useful, but in statistics it is often the case that you can avoid them entirely and calculate the answer much faster using something I will call the "kernel trick" (I am not aware of this idea having an official name).

For example, suppose we need to calculate

$$\int_0^\infty e^{-5x}\,dx.$$

Sure, we *can* use substitution, but most statisticians will find it easier to recognize that this is very similar to the exponential distribution, which (like all distributions) integrates to 1:

$$\int_0^\infty \lambda e^{-\lambda x}\,dx = 1 \text{ for all } \lambda > 0.$$

Applying this shortcut:

$$\int_0^\infty e^{-5x}\,dx = \frac{1}{5}\int_0^\infty 5e^{-5x}\,dx \tag{2.66}$$

$$= \frac{1}{5} \tag{2.67}$$

The *kernel* of a distribution is the part that has the variable we're integrating over. This is the only part that needs to match in order for the trick to work: we can always manipulate the constants as we did above.

As another example, suppose we need to find

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx.$$

This is actually impossible to solve using any of the integration techniques above – there is no elementary form form for its antiderivative. However, it has the kernel of a normal distribution:

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

Letting $\mu = 0$ and $\sigma = 1/\sqrt{2}$, we get

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}/\sqrt{2}}\exp\left\{-\frac{1}{2}\left(\frac{x}{1/\sqrt{2}}\right)^2\right\} \tag{2.68}$$

$$= \sqrt{\pi} \tag{2.69}$$

This may seem complicated at first, but I cannot emphasize enough how important it is to learn this. As a statistician you will become very familiar with these distributions and this will get easier and easier. Every fall, in a ritual as constant as the turning of the leaves, first-year graduate students labor away, trying to solve integrals using elaborate integration by parts techniques, and a professor or older graduate student will look at what they are doing and solve it in seconds using this trick.

As practice, use this procedure to calculate

$$\int_0^{\infty} x^2 e^{-x}\,dx$$

by using the kernel trick with respect to the gamma distribution, which has density function

$$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}.$$

You should get $\Gamma(3)$, which is 2: $\Gamma(\alpha) = (\alpha-1)!$ if α is an integer.

2.9 Sequences and series

Finite series

Some important finite series are sum i, sum i2, sum i3, sum i4 (p 261)

2.10 Partial derivatives

As problems get more complicated, there are almost always multiple variables involved. Suppose we have the function

$$f(x,y) = x^2 y$$

and we are told to find its derivative. This is an ambiguous request, and could mean several different things. For a function of several variables, a **partial derivative of f with respect to** x means to calculate the ordinary derivative treating f as a function of x alone with all the other variables held constant.

For the example above, taking the derivative with respect to x would yield $2xy$, whereas taking the derivative with respect to y, would yield x^2.

To give a further example, as well as introduce notation, for the function $f(x, y, z) = \frac{x^3 z}{y}$, the partial derivatives with respect to x, y, and z would be written as:

$$\frac{\partial}{\partial x}(\frac{x^3 z}{y}) = \frac{3x^2 z}{y}$$

$$\frac{\partial}{\partial y}(\frac{x^3 z}{y}) = \frac{-x^3 z}{y^2}$$

$$\frac{\partial}{\partial z}(\frac{x^3 z}{y}) = \frac{x^3}{y}$$

Notice how for each of the partial derivative operations, two of the input variables are left constant whereas the derivative operation is performed on one "aspect" of the function. For $\frac{\partial}{\partial x}$, the $\frac{z}{y}$ term of the function was treated as constant and the derivative operation was performed on x^3. For $\frac{\partial}{\partial y}$, the $x^3 z$ term of the function was treated as constant and the derivative operation was performed on $\frac{1}{y}$. For $\frac{\partial}{\partial z}$, the $\frac{x^3}{y}$ term of the function was treated as constant and the derivative operation was performed on z. Notation-wise, the symbol ∂ is used to represent a partial derivative whereas d is used to represent a total derivative.

Conceptually speaking, a partial derivative details the instantaneous rate of change in a function when one of its input variables is changed while keeping all other input variables constant. For instance, for each of the results above, it is conveyed that the instantaneous rate of change in the function $f(x, y, z) = \frac{x^3 z}{y}$ is $\frac{3x^2 z}{y}$, $\frac{-x^3 z}{y^2}$, $\frac{x^3}{y}$ when x, y, or z is changed respectively (while keeping the other two variables constant).

Partial derivatives can also be taken to the second order (similar to total derivatives) as well as "mixed". For second order partial derivatives, they are written and calculated similarly to total derivatives (just with keeping other input variables constant):

$$\frac{\partial^2}{\partial x^2}(\frac{x^3 z}{y}) = \frac{\partial}{\partial x}(\frac{\partial}{\partial x}(\frac{x^3 z}{y})) \tag{2.70}$$

$$= \frac{\partial}{\partial x}(\frac{3x^2 z}{y}) \tag{2.71}$$

$$= \frac{6xz}{y} \tag{2.72}$$

$$\frac{\partial^2}{\partial y^2}(\frac{x^3 z}{y}) = \frac{\partial}{\partial y}(\frac{\partial}{\partial y}(\frac{x^3 z}{y})) \tag{2.73}$$

$$= \frac{\partial}{\partial y}(\frac{-x^3 z}{y^2}) \tag{2.74}$$

$$= \frac{2x^3 z}{y^3} \tag{2.75}$$

$$\frac{\partial^2}{\partial z^2}(\frac{x^3z}{y}) = \frac{\partial}{\partial z}(\frac{\partial}{\partial z}(\frac{x^3z}{y})) \tag{2.76}$$

$$= \frac{\partial}{\partial z}(\frac{x^3}{y}) \qquad = 0 \tag{2.77}$$

By "mixed" partial derivatives, I mean partial derivatives with respect to different input variables that are taken in succession. For instance, if I write $\frac{\partial^2}{\partial x \partial y}$, this is equivalent to writing $\frac{\partial}{\partial x}\frac{\partial}{\partial y}$ and states that the partial derivative with respect to y is taken first and then the result from that operation is then used to take the partial derivative with respect to x. A couple examples using the same $f(x, y, z) = \frac{x^3z}{y}$ are given below:

$$\frac{\partial^2}{\partial x \partial y}(\frac{x^3z}{y}) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}(\frac{x^3z}{y}) \tag{2.78}$$

$$= \frac{\partial}{\partial x}(\frac{-x^3z}{y^2}) \tag{2.79}$$

$$= \frac{-3x^2z}{y^2} \tag{2.80}$$

$$\frac{\partial^2}{\partial z \partial x}(\frac{x^3z}{y}) = \frac{\partial}{\partial z}\frac{\partial}{\partial x}(\frac{x^3z}{y}) \tag{2.81}$$

$$= \frac{\partial}{\partial z}(\frac{3x^2z}{y}) \tag{2.82}$$

$$= \frac{3x^2}{y} \tag{2.83}$$

Since we are on the subject of partial derivatives, it is also worth briefly discussing the gradient operation as this will be one of the more common operations involving partial derivatives in statistics. The gradient, symbolized by ∇, when applied to a function made up of n input variables (i.e. $f(x_1, ..., x_n)$), results in a vector where each element is the partial derivative of $f(x_1, ..., x_n)$ with respect to one of the input variables.

$\nabla f(x_1, ..., x_n) = [\frac{\partial}{\partial x_1} f(x_1, ..., x_n), \frac{\partial}{\partial x_2} f(x_1, ..., x_n), ..., \frac{\partial}{\partial x_n} f(x_1, ..., x_n)]$

The gradient will be described in much more detail in the first year or two of the program but getting a little exposure to it now is important. Conceptually, the gradient represents, on a high level, the direction and rate of fastest increase for a function made up of multiple input variables.

2.11 Multiple integrals

Similar to partial derivatives, motivating the idea of multiple integrals can be done simply by considering trying to take the integral of the following function:

$f(x, y) = x^2y$

Whereas previous discussions of integration had involved only a singular input variable, functions with multiple input variables require specification of how you are performing integration (are you integrating with respect to x or respect to y or respect to both?). For instance, integrating the function with respect to x (while treating y as constant) would yield $\frac{x^3y}{3}$ whereas integrating with respect to y (while treating x as constant) would yield $\frac{x^2y^2}{2}$. However, it is also entirely possible to integrate the function $f(x, y)$ with respect to x and y. Performing this operation would be written as follows:

$\int\int f(x, y)dxdy = \int(\int f(x, y)dx)dy$

In this case, we will be integrating the function with respect to x first and then that result will be immediately integrated again with respect to y. The order of integrating with respect to x and then y is conveyed by how dx is written before dy. The parentheses around the inner integral should give some further intuition about how the dx and dy line up with the inner and outer integrals to convey order.

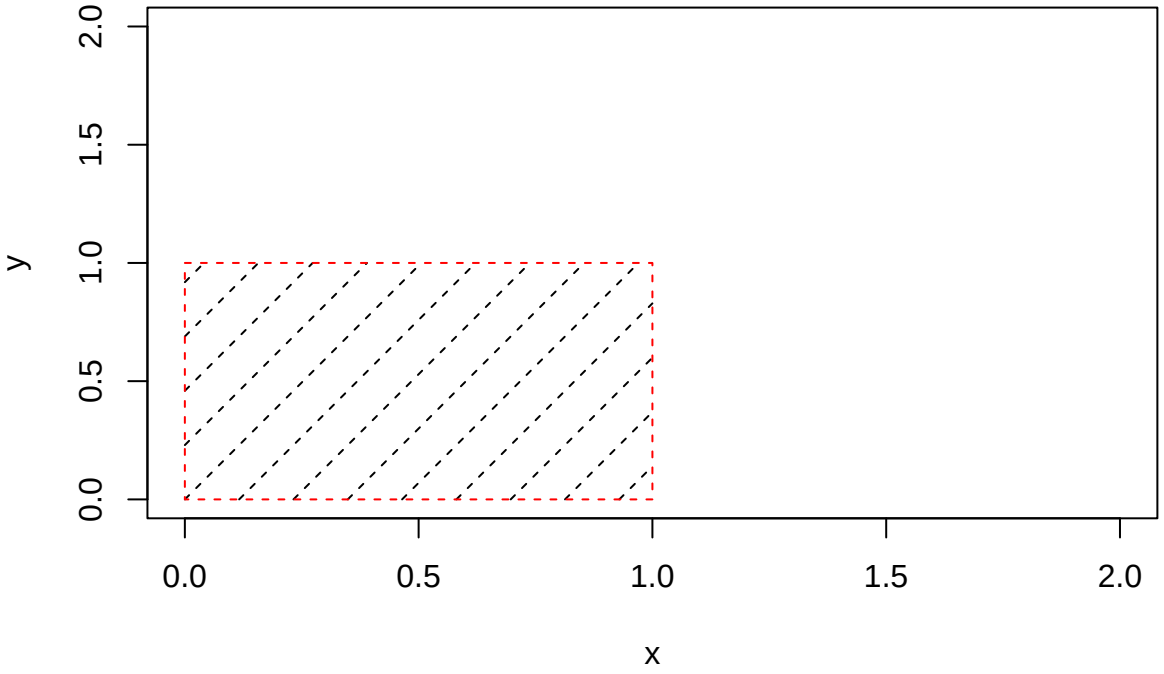
Performing the operation completely is done as follows:

$\int\int f(x,y)dxdy = \int(\int f(x,y)dx)dy = \int(\int x^2ydx)dy = \int \frac{x^3y}{3}dy = \frac{x^3y^2}{6}$

To extend this example to the situation of a definite integral, if I wanted to integrate $f(x,y) = x^2y$ across the regions $0 \leq x \leq 1$ and $0 \leq y \leq 1$, the operation would look like this:

$\int_0^1\int_0^1 f(x,y)dxdy = \int_0^1(\int_0^1 f(x,y)dx)dy = \int_0^1(\int_0^1 x^2ydx)dy = \int_0^1(\frac{(1)^3y}{3} - \frac{(0)^3y}{3})dy = \int_0^1 \frac{y}{3}dy = \frac{1^2}{6} - \frac{0^2}{6} = \frac{1}{6}$

Conceptually, this value of $\frac{1}{6}$ represents the volume encompassed within the function x^2y across the rectangular region $0 \leq x \leq 1$ and $0 \leq y \leq 1$. I say rectangular region because if I consider all the possible (x,y) pairs that satisfy the bounds $0 \leq x \leq 1$ and $0 \leq y \leq 1$, the resultant region is a rectangle.



This idea of taking integrals of functions made up of multiple input variables with respect to all those input variables is known as **multiple integrals**. Multiple integrals can be done over any dimension depending on how many input variables a function may have. For instance, the multiple integral of $f(x,y)$ is a double integral since the integration would be across two dimensions (x and y). However, the multiple integral of $f(x,y,z)$ is a triple integral since the integration would be across three dimensions (x, y, and z). To give a general formula for a multiple integral, for a function of n input variables (i.e. $f(x_1,...,x_n)$), the multiple integral would be written as:

$\int ... \int f(x_1,...,x_n)dx_1...dx_n$

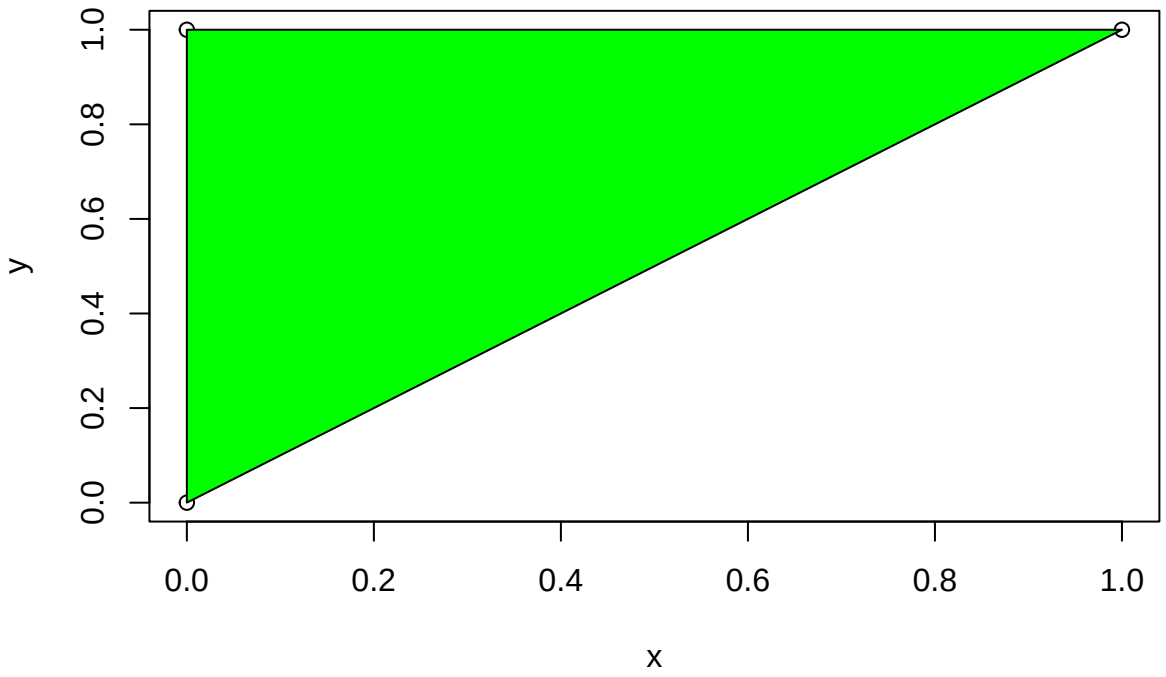
To give one more in-depth example of a multiple integral, consider integrating $f(x,y,z) = \frac{x^3z}{y}$ across x, then y and then z (indefinite integral). The operation sequence would be as follows:

$\int\int\int f(x,y,z)dxdydz = \int\int(\int \frac{x^3z}{y}dx)dydz = \int(\int \frac{x^4z}{4y}dy)dz = \int \frac{x^4zln(y)}{4}dz = \frac{x^4z^2ln(y)}{8}$

Since we are on the topic of multiple integrals, it is important to discuss the possibility of non-rectangular regions of integration. To make this easier to explain, this explanation will only be in the context of double integrals (i.e. $f(x,y)$). Earlier I talked about the rectangular region resultant when integrating $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Interestingly, any double integral integrated over a region like $a \leq x \leq b$ and $c \leq y \leq d$ is

going to be a rectangular region as all the (x,y) points that satisfy the bounds will be located inside the rectangle with vertices (a,c), (a,d), (b,c), and (b,d). However, it is possible for a region of integration to not be rectangular and this will arise when there is an inequality governing how x and y relate to each other.

For instance, what would be the multiple integral of $f(x, y) = x^2y$ across the region $0 \leq x \leq y \leq 1$. To give an idea of what the region of integration looks like, let's plot the area of (x,y) points that satisfy $0 \leq x \leq y \leq 1$.

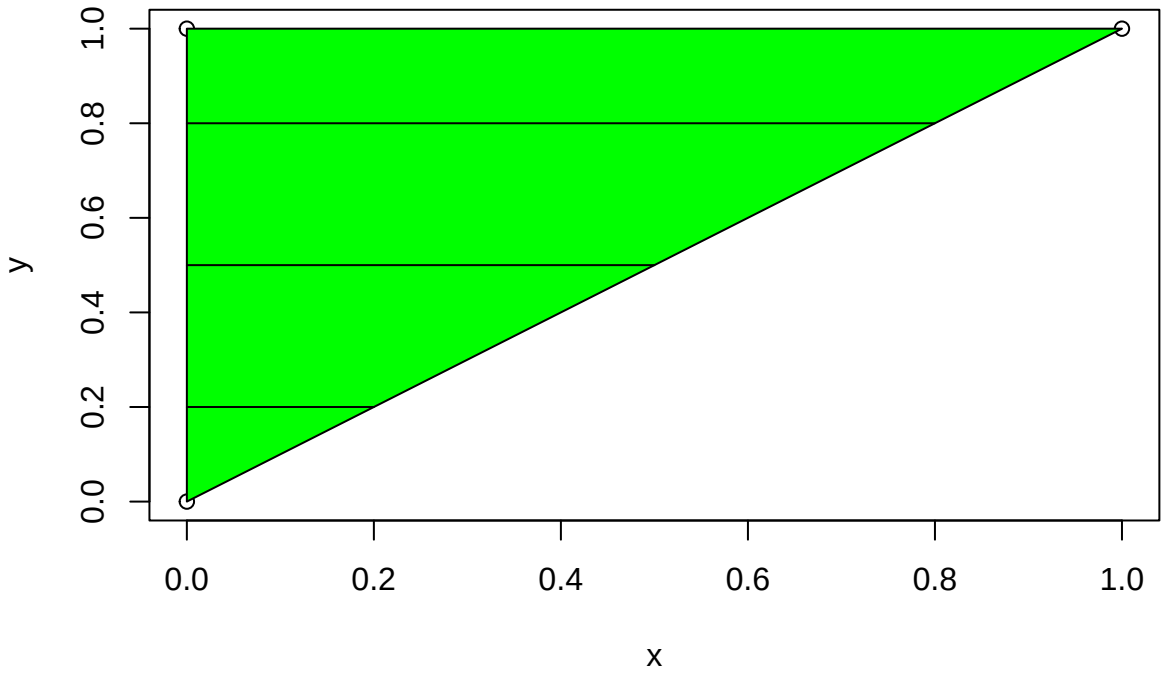


Notice that any (x,y) point in the green region satisfies the inequality $0 \leq x \leq y \leq 1$. At x = 0.4 for instance, only points where y is greater than or equal to 0.4 would allow for the point to be in the green region. This region, as opposed to the rectangular region shown earlier, is actually a triangle. The inequality governing x to always be less than or equal to y forces the bottom half triangle that would make the region rectangular to be "cut-off".

Regarding actually performing the double integral, the general formula is still applicable. We will start with integrating with respect to x and then with respect to y.

$$\int \int f(x, y)dxdy = \int (\int f(x, y)dx)dy$$

As soon as we figure out the bounds for the two integrals, we will be good to go. Since we are starting with respect to x, let's figure out the appropriate x bounds. The green region plotted above is immensely helpful in understanding what the appropriate upper and lower bounds are for x. To determine the lower and upper bounds of the inner integral with respect to x, we need to identify at any given value of y, what values of x are within the green region. For instance, at given values of y = 0.2, 0.5, and 0.8, I can draw straight lines to determine that $0 \leq x \leq 0.2$, $0 \leq x \leq 0.5$, and $0 \leq x \leq 0.8$ are valid x values for the given values of y respectively.



Extrapolating to the general case for any given value y, the region $0 \leq x \leq y$ consists of all possible x values that will fall in the green region. Thus, the bounds for the inner integral will be 0 and y.

$\int \int_0^y f(x,y)dxdy = \int (\int_0^y f(x,y)dx)dy$

Regarding the outer integral, since the inner integral goes over all the valid x values for any given y value, integrating y over the entire range of 0 to 1 will encompass the entire triangular region. Think about it as if you drew a line segment of valid x values for every y value in the range of 0 to 1 (y = 0.001, y = 0.002, ..., y = 0.1, ..., y = 0.5, ..., y = 1), all the line segments together would form the solid triangle.

$\int_0^1 \int_0^y f(x,y)dxdy = \int_0^1 (\int_0^y f(x,y)dx)dy$

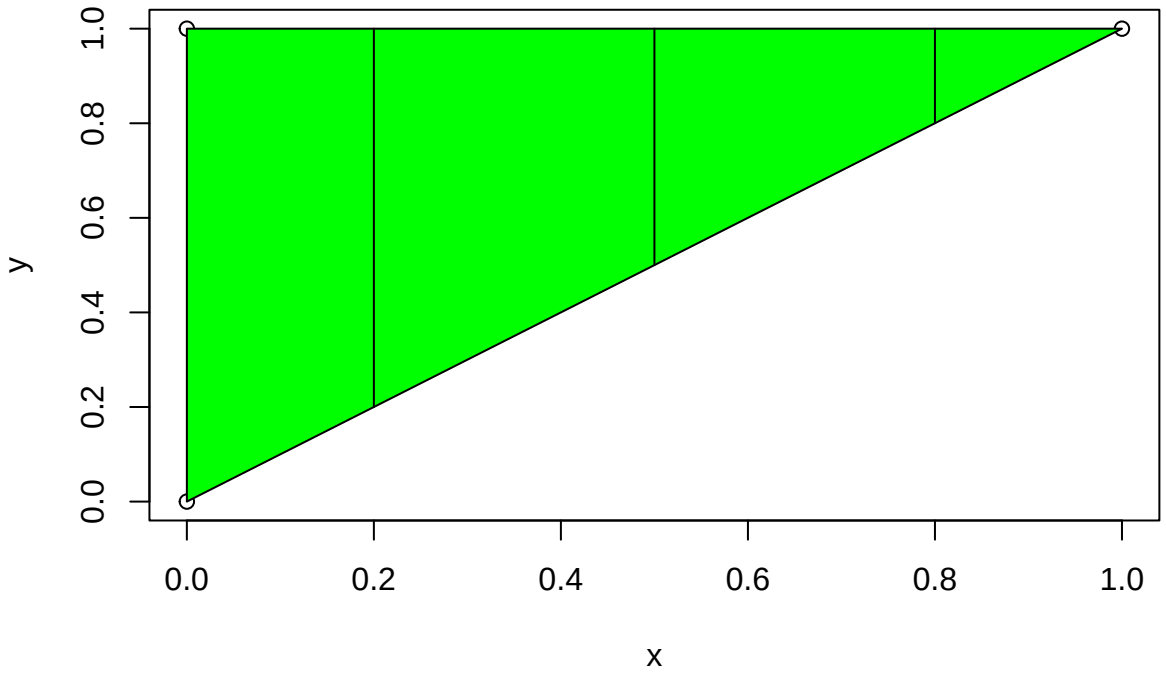
Now all that's left is to do the math:

$\int_0^1 \int_0^y f(x,y)dxdy = \int_0^1 (\int_0^y x^2 ydx)dy = \int_0^1 \frac{y^3 y}{3}dy = \frac{1^5}{15} = \frac{1}{15}$

Interestingly, it is entirely possible to start with integrating with respect to y and then with respect to x and yield the same answer. In this case, the general formula would be modified to account for the changed order:

$\int \int f(x,y)dydx = \int (\int f(x,y)dy)dx$

Using the same process as before but starting with respect to y, let's figure out the appropriate y bounds. To determine the lower and upper bounds of the inner integral with respect to y, we need to identify at any given value of x, what values of y are within the green region. For instance, at given values of x = 0.2, 0.5, and 0.8, I can draw straight lines to determine that $0.2 \leq y \leq 1$, $0.5 \leq y \leq 1$, and $0.8 \leq y \leq 1$ are valid y values for the given values of x respectively.



Extrapolating to the general case for any given value x, the region $x \leq y \leq 1$ consists of all possible y values that will fall in the green region. Thus, the bounds for the inner integral will be x and 1.

$\int \int_x^1 f(x,y)dydx = \int (\int_x^1 f(x,y)dy)dx$

Regarding the outer integral, since the inner integral goes over all the valid y values for any given x value, integrating x over the entire range of 0 to 1 will encompass the entire triangular region. Think about it as if you drew a line segment of valid y values for every x value in the range of 0 to 1 (x = 0.001, x = 0.002, ..., x = 0.1, ..., x = 0.5, ..., x = 1), all the line segments together would form the solid triangle.

$\int_0^1 \int_1^x f(x,y)dydx = \int_0^1 (\int_x^1 f(x,y)dy)dx$

Now all that's left is to do the math:

$\int_0^1 \int_x^1 f(x,y)dydx = \int_0^1 (\int_x^1 x^2 ydy)dx = \int_0^1 (\frac{1^2x^2}{2} - \frac{x^2x^2}{2})dx = \int_0^1 (\frac{x^2}{2} - \frac{x^4}{2})dx = \frac{1^3}{6} - \frac{1^5}{10} = \frac{1}{6} - \frac{1}{10} = \frac{1}{15}$

I bring up this idea of deciding whether to integrate with respect to x or y first because it is possible that, while both routes will lead to the same answer, one route might be much easier to calculate compared to the other. It is important to practice this concept of integrating across non-rectangular regions in order to develop intuition on how these problems can be visualized and deciding on appropriate integral bounds. Drawing the region of integration like what was done in the practice problems is also a valuable tool that can help in solving these types of problems.

Chapter 3

Matrix algebra

modeling, relevance to statistics

3.1 Definitions and conventions

A *matrix* is a collection of numbers arranged in a rectangular array of *rows* and *columns*, such as

$$\begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix}$$

A matrix with r rows and c columns is said to be an $r \times c$ matrix (e.g., the matrix above is a 3×2 matrix).

In the case where a matrix has just a single row or column, it is said to be a *vector*, such as

$$\begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

Conventionally, vectors and matrices are denoted in lower- and upper-case boldface, respectively (e.g., x is a scalar, $\mathbf{x}$ is a vector, and $\mathbf{X}$ is a matrix). In addition, vectors are taken to be *column vectors* – i.e., a vector of n numbers is an $n \times 1$ matrix, not a $1 \times n$ matrix.

The ijth element of a matrix $\mathbf{M}$ is denoted by M_{ij} or $(\mathbf{M})_{ij}$.

For example, letting $\mathbf{M}$ denote the above matrix, $M_{11} = 3$, $(\mathbf{M})_{32} = 2$, and so on. Similarly, the jth element of a vector $\mathbf{v}$ is denoted v_j; e.g., letting $\mathbf{v}$ denote the above vector, $v_1 = 3$.

3.2 Basic operations

Transposition

It is often useful to switch the rows and columns of a matrix around. The resulting matrix is called the *transpose* of the original matrix, and denoted with a superscript $^\top$ or an apostrophe $'$:

$$\mathbf{M} = \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix} \qquad \mathbf{M}^\top = \begin{bmatrix} 3 & 4 & -1 \\ 2 & -1 & 2 \end{bmatrix}$$

Note that $M_{ij} = M_{ji}^\top$, and that if $\mathbf{M}$ is an $r \times c$ matrix, $\mathbf{M}^\top$ is a $c \times r$ matrix.

Addition

There are two kinds of addition operations for matrices. The first is *scalar addition*:

$$\mathbf{M} + 2 = \begin{bmatrix} 3+2 & 2+2 \\ 4+2 & -1+2 \\ -1+2 & 2+2 \end{bmatrix} = \begin{bmatrix} 5 & 4 \\ 6 & 1 \\ 1 & 4 \end{bmatrix}$$

The other kind is *matrix addition*:

$$\mathbf{M} + \mathbf{M} = \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix} + \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 6 & 4 \\ 8 & -2 \\ -2 & 4 \end{bmatrix}$$

Formally, $(\mathbf{A} + \mathbf{B})\_ij = A_{ij} + B_{ij}$.

Note that only matrices of the same dimension can be added to each other – there is no such thing as adding a 4×5 matrix to a 2×9 matrix.

Multiplication

There are also two common kinds of multiplication for matrices. The first is *scalar multiplication*:

$$4\mathbf{M} = 4 \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 12 & 8 \\ 16 & -4 \\ -4 & 8 \end{bmatrix}$$

Formally, $(c\mathbf{M})\_ij = cM_{ij}$.

The other kind is *matrix multiplication*. The product of two matrices, $\mathbf{AB}$, is defined by multiplying all of $\mathbf{A}$'s rows by $\mathbf{B}$'s columns in the following manner:

$$(\mathbf{AB})\_ik = \sum_j A_{ij} B_{jk}$$

$$\begin{bmatrix} 1 & 2 & 1 \\ 4 & -1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 0 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 12 & 9 \end{bmatrix}$$

Note that matrix multiplication is only defined if the number of columns of $\mathbf{A}$ matches the number of rows of $\mathbf{B}$, and that if $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is an $n \times p$ matrix, then $\mathbf{AB}$ is an $m \times p$ matrix.

The following elementary algebra rules carry over to matrix algebra:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \tag{3.1}$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \tag{3.2}$$

$$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B} \tag{3.3}$$

One important exception, however, is that $\mathbf{AB} \neq \mathbf{BA}$; the order of matrix multiplication matters, and we must remember to, for instance, "left multiply" both sides of an equation by a matrix $\mathbf{M}$ to preserve equality.

Inner and outer products

Suppose $\mathbf{u}$ and $\mathbf{v}$ are two $n \times 1$ vectors. We can't multiply them in the sense defined above, $\mathbf{uv}$, because the number of columns of $\mathbf{u}$, 1, doesn't match the number of rows of $\mathbf{v}$, n. However, there are two ways in which vectors of the same dimension can be multiplied.

The first is called the *inner product* (also, the "cross product"):

$$\mathbf{u}^\top \mathbf{v} = \sum_j u_j v_j \tag{3.4}$$

$$\begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 6 - 2 = 4. \tag{3.5}$$

Note that when we multiply matrices, the element $(\mathbf{AB})_{ij}$ is equal to the inner product of the ith row of $\mathbf{A}$ and the jth column of B.

The second way of multiplying two vectors is called the *outer product*:

$$(\mathbf{uv}^\top)_{ij} = u_i v_j \tag{3.6}$$

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \end{bmatrix} = \begin{bmatrix} 6 & -3 \\ 4 & -2 \end{bmatrix} \tag{3.7}$$

Note that the inner product returns a scalar number, while the outer product returns an $n \times n$ matrix.

3.3 Special matrices

In the special case where a matrix has the same numbers of rows and columns, it is said to be *square*. If $\mathbf{A}^\top = \mathbf{A}$, the matrix is said to be *symmetric*.

$$\text{Symmetric: } \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \qquad \text{Not symmetric: } \begin{bmatrix} 3 & 2 \\ 0 & -1 \end{bmatrix}$$

Note that a matrix cannot be symmetric unless it is square.

The elements A_{ii} of a matrix are called its *diagonal entries*; a matrix for which $A_{ij} = 0$ for all $i \neq j$ is said to be a *diagonal matrix*:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Consider in particular the following diagonal matrix:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that this matrix has the interesting property that $(\mathbf{AI})\_ij = A_{ij}$ for all i, j; in other words, $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$. Because of this property, $\mathbf{I}$ is referred to as the *identity matrix*.

Some other notations which are commonly used are $\mathbf{1}$, the vector (or matrix) of 1s, and $\mathbf{0}$, the vector (or matrix) of zeros:

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \qquad \mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

The dimensions of these matrices is sometimes explicitly specified, as in $\mathbf{0}_{2\times 2}$, $\mathbf{I}_{5\times 5}$, or $\mathbf{1}_{4\times 1}$. Other times it is obvious from context what the dimensions must be.

Finally, the vector $\mathbf{e}_j$ is also useful: it has element $e_j = 1$ and $e_k = 0$ for all other elements:

$$\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

This is useful for selecting a single element of a vector: $\mathbf{u}^\top \mathbf{e}_3 = u_3$.

3.4 Inversion and related concepts

Suppose $\mathbf{Ax} = \mathbf{B}$ and we want to solve for $\mathbf{x}$ ... can we "divide" by $\mathbf{A}$? The answer is: "sort of". There is no such thing as matrix division, but we can multiply both sides by the *inverse* of $\mathbf{A}$. If a matrix $\mathbf{A}^{-1}$ satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, then $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$. If we know what $\mathbf{A}^{-1}$ is, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{B}$. Note that $\mathbf{x}$ is *not* equal to $\mathbf{BA}^{-1}$; we need to *left* multiply by the inverse and order or multiplication matters.

If two vectors $\mathbf{u}$ and $\mathbf{v}$ satisfy $\mathbf{u}^\top \mathbf{v} = 0$, they are said to be *orthogonal* to each other. If all the columns and rows of a matrix $\mathbf{A}$ are orthogonal to each other and satisfy $\mathbf{a}^\top \mathbf{a} = 1$, then $\mathbf{A}$ (transposed) can serve as its own inverse: $\mathbf{A}^\top \mathbf{A} = \mathbf{AA}^\top = \mathbf{I}$. In this case, the matrix $\mathbf{A}$ is said to be an *orthogonal matrix*. If a matrix $\mathbf{X}$ is not square, then it is possible that $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$ but $\mathbf{XX}^\top \neq \mathbf{I}$; in this case, the matrix is said to be *column orthogonal*, although in statistics it is common to refer to these matrices as orthogonal also. A somewhat related definition is that a matrix is said to be *idempotent* if $\mathbf{AA} = \mathbf{A}$.

Does every matrix have one and only one inverse? If a matrix has an inverse, it is said to be *invertible* – all invertible matrices have exactly one, unique inverse. However, not every matrix is invertible. For example, there are no values of a, b, c, and d that satisfy

$$\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Why doesn't this matrix have an inverse? There are four equations and four unknowns, but some of those equations contradict each other. The term for this situation is *linear dependence*. If you have a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, then you can form new vectors from *linear combinations* of the old vectors: $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$. A collection of vectors is said to be *linearly independent* if none of them can be written as a linear combination of the others; if it can, then they are linearly dependent. This is the key to whether a matrix is invertible or not: a matrix $\mathbf{A}$ is invertible if and only if its columns (or rows) are linearly independent. Note that the columns of our earlier matrix were not linearly independent, since $2(2 \quad 1) = (4 \quad 2)$.

The *rank* of a matrix is the number of linearly independent columns (or rows) it has; if they're all linearly independent, then the matrix is said to be of *full rank*.

Additional helpful identities:

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top \tag{3.8}$$

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top \tag{3.9}$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{3.10}$$

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top \tag{3.11}$$

Chapter 4

Analysis

The material in chapters 2 and 3 is intended as review for *incoming* graduate students to prepare them for courses they will take their first year in the program. For students in our PhD program, there is an additional sequence of courses (BIOS 7110 and BIOS 7250) that covers the mathematical foundations of statistics in greater depth. For this material, an understanding of analysis is important.

Analysis is concerned with the same topics as calculus, but calculus focuses on tools from a user perspective ("how do I calculate a derivative?") whereas analysis focuses on theoretical properties (e.g., proving theorems about derivatives and differentiability). So, the table of contents here might appear similar to chapter 2, but the focus is quite different.

NOTE: I am in the process of migrating material here from this pdf version, so this page is something of a placeholder for now.

REMINDER: If you are an incoming first-year student, you really don't need to worry about this material yet! Just focus on chapters 2 and 3.