Math review for new biostatistics students

Patrick Breheny and Eldon Sorensen

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 are helpful 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, there are very few exercises and solutions provided in this review. 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 Calculus.

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 $f : A \mapsto B$.

Commentary

A few remarks on this definition and its implications:

1. This is an *extremely* general definition. A and B could be sets of numbers, but not necessarily: A could be a set of 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

$$s = \lim_{x \to a} f(x)$$
$$t = \lim_{x \to a} g(x);$$

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.

$$\lim_{x\to a}\{f(x)+g(x)\} = s+t$$
$$\lim_{x\to a}\{f(x)-g(x)\} = s-t$$
$$\lim_{x\to a}\{cf(x)+g(x)\} = cs+t \text{ where } c \text{ is a constant}$$
$$\lim_{x\to a}\{f(x)g(x)\} = st$$
$$\lim_{x\to a}\frac{f(x)}{g(x)} = \frac{s}{t} \text{ if } t \neq 0$$
$$\lim_{x\to a}\{f(x)^n\} = s^n$$

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 $\lim_{x\to 2} f(x) = 4$. 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.

$$\begin{aligned}
c' &= 0 \\
(x^n)' &= nx^{n-1} \\
(cf)' &= cf' \\
(f+g)' &= f' + g' \\
(f-g)' &= f' - g' \\
(fg)' &= fg' + gf' \text{ (product rule)} \\
\left(\frac{f}{g}\right)' &= \frac{gf' - fg'}{g^2} \text{ (quotient rule)}
\end{aligned}$$

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.

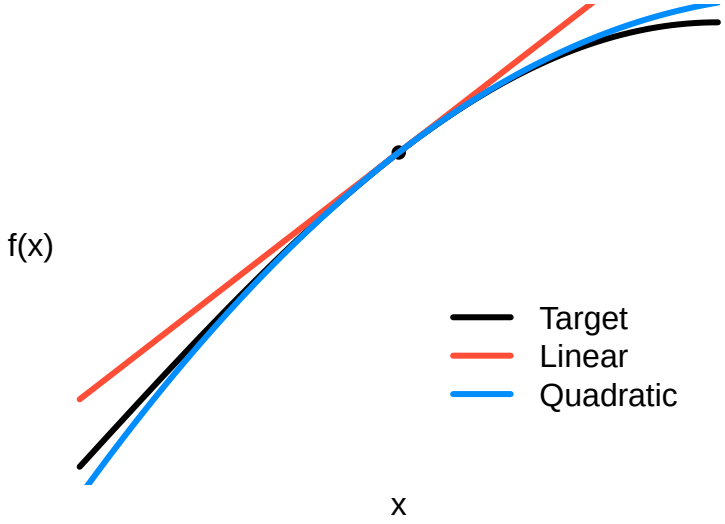
An important application of higher derivatives is to approximate functions. The **linear approximation** of f at a is given by

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

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 minimum. If it changes from positive to negative, c is a local maximum. 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:

$$\int_a^b c\,dx = c(b-a)$$
$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx$$
$$\int_a^b \big\{f(x)+g(x)\big\}\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$
$$\int_a^b \big\{f(x)-g(x)\big\}\,dx = \int_a^b f(x)\,dx - \int_a^b g(x)\,dx$$
$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

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

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

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

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:

$$\int 2x\,dx = x^2$$
$$\int 2x\,dx = x^2 + 5.$$

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,

$$\begin{aligned}\int_1^4 2x\,dx &= F(4) - F(1)\\ &= 4^2 - 1^2\\ &= 15.\end{aligned}$$

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.

[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.

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

$$\begin{aligned} a^{x+y} &= a^x a^y \\ a^{x-y} &= \frac{a^x}{a^y} \\ (a^x)^y &= a^{xy} \\ (ab)^x &= a^x b^x \end{aligned}$$

Exponential limits

$$\begin{aligned} \lim_{x\to\infty} a^x &= \infty \quad \text{if } a > 1 \\ \lim_{x\to-\infty} a^x &= 0 \quad \text{if } a > 1 \\ \lim_{x\to\infty} a^x &= 0 \quad \text{if } 0 < a < 1 \\ \lim_{x\to-\infty} a^x &= \infty \quad \text{if } 0 < a < 1 \\ \lim_{h\to 0} \frac{e^h - 1}{h} &= 1 \\ \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n &= e \end{aligned}$$

Exponential derivatives and integrals

$$\begin{aligned} \frac{d}{dx} e^x &= e^x \\ \frac{d}{dx} e^u &= e^x \frac{du}{dx} \\ \int e^x \, dx &= e^x \\ \frac{d}{dx} a^x &= a^x \log(a) \\ \int a^x \, dx &= \frac{a^x}{\log a} \quad (a \neq 1) \end{aligned}$$

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

$$\begin{aligned} \log(e^x) &= x \\ e^{\log x} &= x \\ \log e &= 1. \end{aligned}$$

Logarithm rules

$$\begin{aligned} \log_a(xy) &= \log_a x + \log_a y \\ \log_a \frac{x}{y} &= \log_a x - \log_a y \\ \log_a(x^y) &= y \log_a x \\ \log_a x &= \frac{\log x}{\log a} \end{aligned}$$

Logarithm limits

If $a > 1$, then

$$\begin{aligned} \lim_{x \to \infty} \log_a x &= \infty \\ \lim_{x \to 0^+} \log_a x &= -\infty \end{aligned}$$

Logarithm derivatives and integrals

$$\begin{aligned} \frac{d}{dx} \log x &= x^{-1} \\ \frac{d}{dx} \log u &= u^{-1} \frac{du}{dx} \\ \int \frac{1}{x}\, dx &= \log |x| \\ \frac{d}{dx} \log_a x &= \frac{1}{x \log a} \end{aligned}$$

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):

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

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.

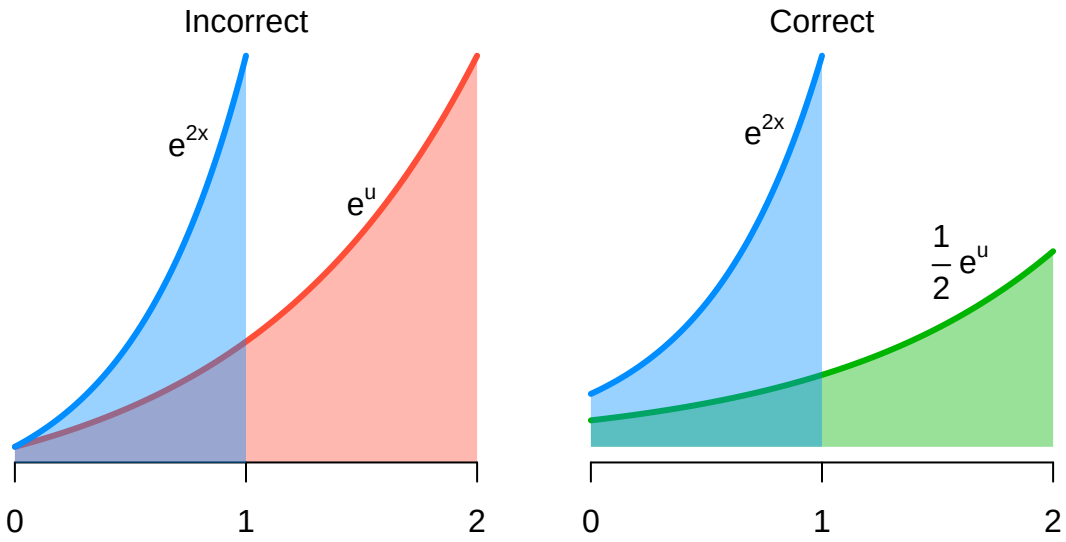
Even with such a table, however, there are a few useful integration techniques to be familiar with. Among other things, (a) it may be faster to use one of these techniques than looking up an integral (b) you might not have access to such a table at the moment, and (c) the form that appears in the table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

$$u = \log x \qquad dv = dx$$
$$du = \frac{1}{x}\,dx \qquad v = x$$

Thus,

$$\begin{aligned}\int \log x \, dx &= x \log x - \int x \frac{dx}{x} \\ &= x \log x - \int dx \\ &= x \log x - x\end{aligned}$$

As practice, use this procedure to calculate

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

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:

$$\begin{aligned}\int_0^\infty e^{-5x} \, dx &= \frac{1}{5} \int_0^\infty 5e^{-5x} \, dx \\ &= \frac{1}{5}\end{aligned}$$

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

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

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

A *sequence* is an ordered list of objects (usually numbers): $x_1, x_2, \ldots, x_n$. A *series* is a special type of sequence that arises from the cumulative sum of another sequence: $x_1 + x_2 + \cdots + x_n$, denoted $\sum_{i=1}^n x_i$. The limiting behavior of sequences and series (what happens to them as $n \to \infty$) is very important to statistics as we are often concerned with what happens to an estimate x_n as we collect more data (n here representing the number of observations). In particular, does x_n get closer and closer to the quantity we are trying to estimate?

The study of infinite sequences and series is a big topic; this review is not exhaustive but covers the core definitions and commonly arising results that come up often in statistics.

Finite series

Most of this section centers on infinite sequences and series, but here are some important finite series to be familiar with (you should know the first three by heart, and know where to look up the rest of them when you need them):

$$\sum_{i=1}^{n} 1 = n$$
$$\sum_{i=1}^{n} c = nc$$
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$
$$\sum_{i=1}^{n} i^3 = \left\{\frac{n(n+1)}{2}\right\}^2$$
$$\sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

Infinite sequences

The limit of a sequence is very similar to the limit of a function that we have previously encountered. A sequence a_n has limit a if a_n gets closer and closer to a (i.e., the difference $|a_n - a|$ gets smaller and smaller) as n goes to infinity. This is denoted $a_n \to a$[5].

As with functions, $\mathbf{a}_n \to \infty$ means that a_n just keeps getting bigger, with no bound.

If the limit exists, we say that the sequence **converges** (or is **convergent**). Otherwise, we say the sequence **diverges** (or is **divergent**). If $\mathbf{a}_n \to \infty$, people often say that the "limit is infinity", although keep in mind that if this happens, the sequence diverges (it just diverges in a particular way).

All of the limit laws we discussed earlier for functions are equally valid when stated in terms of sequences. For example, if $a_n \to a$ and $b_n \to b$, we have

$$\lim_{n\to\infty}\{a_n b_n\} = ab,$$

just like we did for functions.

Techniques: It is often unclear what the limit of a ratio is – both numerator and denominator could be going to infinity, or both going to zero. Two techniques to remember are dividing by the largest power and L'Hôpital's rule.

To illustrate the first:

$$\frac{n^2-5}{n^3+n+3} = \frac{n^{-1}-5n^{-3}}{1+n^{-2}+3n^{-3}}$$
$$\to 0$$

To illustrate the second:

[5]As before, we're skipping some technical details here because you shouldn't really need them for the first year. The topic is covered more rigorously here, which is intended for second-year students.

$$\begin{aligned}\lim_{n\to\infty} \frac{\log n}{n} &= \lim_{n\to\infty} \frac{n^{-1}}{1} \\ &= 0\end{aligned}$$

Recall that L'Hôpital's rule only holds if the ratio of the derivatives converges and the original fraction is indeterminate.

Some special sequences: It is useful to know the limits of some sequences, because they come up frequently:

$$r^n \to \begin{cases} 0 & \text{if } |r| < 1 \\ 1 & \text{if } r = 1 \\ \text{diverges otherwise} \end{cases}$$

$$n^r \to \begin{cases} 0 & \text{if } r < 0 \\ 1 & \text{if } r = 0 \\ \text{diverges otherwise} \end{cases}$$

$$r^{1/n} \to \begin{cases} 1 & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ \text{undefined otherwise} \end{cases}$$

$$\frac{n^a}{(1+r)^n} \to 0 \text{ for all } a \text{ if } r > 0$$

Monotone convergence theorem: Another way of establishing that a sequence converges is the **monotone convergence theorem**, which states that every bounded, monotonic sequence converges.

Infinite series

Similarly, we say that the series $\sum_{i=1}^{\infty} a_i$ **converges** if the sequence of partial sums $s_n = \sum_{i=1}^{n} a_i$ converges. If $s_n \to s$, then s is the called the **sum** of the series. Otherwise, the series is said to **diverge**.

Note that if the series $\sum_{n=1}^{\infty} a_n$ converges, then $a_n \to 0$. The converse, however, is not true. For example, the following is known as the **harmonic series:**

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \cdots = \infty.$$

Note that series obey some of the limit rules, but not all of them. For example, the following are true if $\sum_{n=1}^{\infty} a_n = a$ and $\sum_{n=1}^{\infty} b_n = b$:

$$\begin{aligned}\sum_{n=1}^{\infty} ca_n &= ca \\ \sum_{n=1}^{\infty} (a_n + b_n) &= a + b \\ \sum_{n=1}^{\infty} (a_n - b_n) &= a - b\end{aligned}$$

However, not all of the limit laws carry over. In particular, if we see the quotient

$$\sum_{n=1}^{\infty} \frac{a_n}{b_n},$$

we *cannot* conclude that its limit is a/b. This is a common mistake, so take some time to realize *why* this is false.

There are two important series to be aware of:

$$\sum_{n=1}^{\infty} r^{n-1} \to \begin{cases} \frac{1}{1-r} & \text{if } |r| < 1 \\ \text{diverges otherwise} \end{cases} \qquad \text{(geometric series)}$$

$$\sum_{n=1}^{\infty} n^r \to \begin{cases} \text{converges} & \text{if } r < -1 \\ \text{diverges otherwise} \end{cases}$$

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. This is denoted $\partial f/\partial x$.

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 $f(x, y, z) = x^3 z/y$, the partial derivatives with respect to x, y, and z would be written as:

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{x^3 z}{y}\right) &= \frac{3x^2 z}{y} \\ \frac{\partial}{\partial y}\left(\frac{x^3 z}{y}\right) &= \frac{-x^3 z}{y^2} \\ \frac{\partial}{\partial z}\left(\frac{x^3 z}{y}\right) &= \frac{x^3}{y} \end{aligned}$$

In each partial derivative, two of the input variables are treated as constants. For example, when we take $\frac{\partial}{\partial x}$, the z/y portion of the function is treated as a constant, and so on. Conceptually, a partial derivative represents the instantaneous rate of change in a function when only one of its input variables is changed while keeping all other input variables constant. Thus, f is changing at the rate $3(2)^2(4)/3$ as we increase x from, say, 2 to 2.001 when $x = 2$, $y = 3$, and $z = 4$.

Higher orders

Partial derivatives can also be taken to the second order, such as

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\left(\frac{x^3z}{y}\right) &= \frac{\partial}{\partial x}(\frac{\partial}{\partial x}(\frac{x^3z}{y})) \\ &= \frac{\partial}{\partial x}(\frac{3x^2z}{y}) \\ &= \frac{6xz}{y}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial z^2}\left(\frac{x^3z}{y}\right) &= \frac{\partial}{\partial z}(\frac{\partial}{\partial z}(\frac{x^3z}{y})) \\ &= \frac{\partial}{\partial z}(\frac{x^3}{y}) \\ &= 0.\end{aligned}$$

Higher-order partial derivatives can also be "mixed", meaning that we take a derivative first with respect to one variable, then with respect to a different variable. In terms of mathematical symbols, $\partial^2 f/\partial x\partial y$ is the same as $\partial/\partial x(\partial f/\partial y)$. A couple examples using the same $f(x,y,z) = x^3z/y$ are given below:

$$\begin{aligned}\frac{\partial^2}{\partial x\partial y}\left(\frac{x^3z}{y}\right) &= \frac{\partial}{\partial x}\frac{\partial}{\partial y}(\frac{x^3z}{y}) \\ &= \frac{\partial}{\partial x}(\frac{-x^3z}{y^2}) \\ &= \frac{-3x^2z}{y^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial z\partial x}\left(\frac{x^3z}{y}\right) &= \frac{\partial}{\partial z}\frac{\partial}{\partial x}(\frac{x^3z}{y}) \\ &= \frac{\partial}{\partial z}(\frac{3x^2z}{y}) \\ &= \frac{3x^2}{y}\end{aligned}$$

It is worth noting (and a bit surprising) that the order is irrelevant, a result known as Clairaut's Theorem. If $\partial^2 f/\partial x\partial y$ and $\partial^2 f/\partial y\partial x$ are both continuous, then

$$\frac{\partial^2 f}{\partial x\partial y} = \frac{\partial^2 f}{\partial y\partial x}.$$

Partial derivatives have many uses in statistics, but the main ideas above are all you really need to know for the first year. Later years will have an increasing focus on multivariable mathematical statistics.

Gradients

A closely related idea to the partial derivative is that of the **gradient**, which simply collects all of the partial derivatives of a function into a vector, symbolized by $\nabla f(\mathbf{x})$, where $\mathbf{x}$ represents the vector $(x_1, x_2, \ldots, x_n)$:

$$\nabla f(x_1, ..., x_n) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Again, gradients will play a larger role in later years than they do in the first year, but familiarity with the basic idea now will help. One reason (among many) that gradients are useful is that in multiple dimensions, there are infinitely many ways we could change the input: we could increase x a little while decreasing y at the same rate but change z three times as fast, and so on. With the gradient, we can calculate these changes by taking the inner product of the gradient and direction: $\nabla f(\mathbf{x})^\top \mathbf{d}$, where $\mathbf{d}$ is the direction.

2.11 Multiple integrals

In statistics, we also (quite often!) need to integrate functions involving multiple variables. Considering trying to integrate the following function:

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

Similar to partial derivatives, we could integrate with respect to x (treating y as a constant) or integrate with respect to y (treating x as constant). This should be relatively straightforward for you to get $x^3y/3$ and $x^2y^2/2$ in these two scenarios.

However, we often need to integrate with respect to *both* x and y. Thankfully, this can be done one variable at a time:

$$\begin{aligned} \int\int f(x, y)\, dx\, dy &= \int \left\{ \int f(x, y)\, dx \right\} dy \\ &= \int \left\{ \int f(x, y)\, dy \right\} dx. \end{aligned}$$

As we saw with second-order partial derivatives, it turns out that if f is continuous, the order of integration doesn't matter (this is known as Fubini's theorem); more on this later.

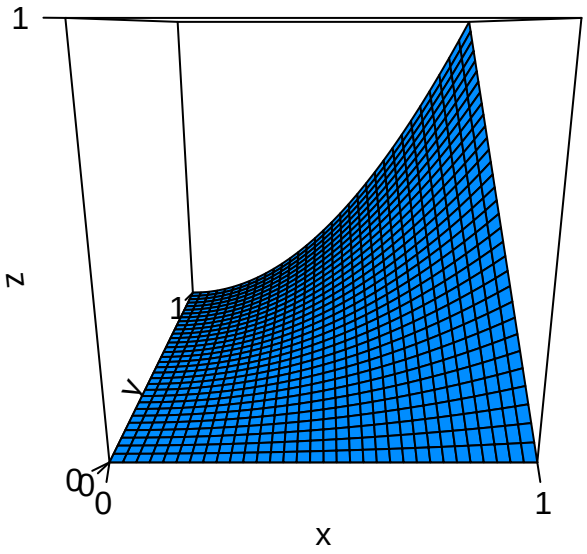
Let's say we've decided to integrate with respect to x first, then y (as we indicate when we write $dx\, dy$):

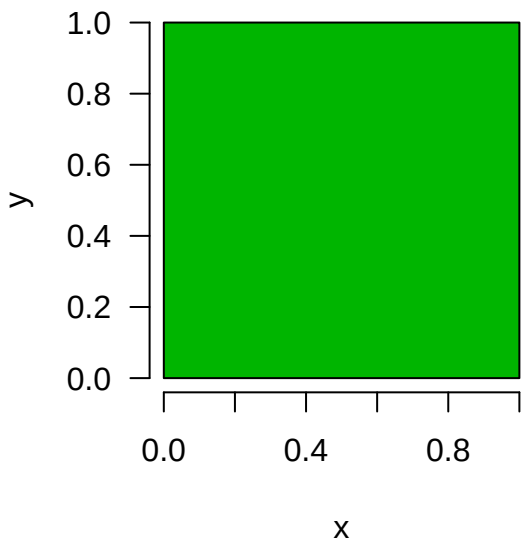
$$\begin{aligned} \int\int x^2 y\, dx\, dy &= \int \left\{ \frac{x^3 y}{3} \right\} dy \\ &= \frac{x^3 y^2}{6} \end{aligned}$$

Suppose now that we want to integrate $f(x, y) = x^2 y$ across the rectangular region $0 \leq x \leq 1$ and $0 \leq y \leq 1$ (i.e., to take the definite integral). This would go as follows:

$$\begin{aligned}
\int_0^1 \int_0^1 x^2 y \, dx \, dy &= \int_0^1 \left\{ \frac{x^3 y}{3} \Big|_{x=0}^{1} \right\} dy \\
&= \int_0^1 \left\{ \frac{(1)^3 y}{3} - \frac{(0)^3 y}{3} \right\} dy \\
&= \int_0^1 \frac{y}{3} \, dy \\
&= \frac{1^2}{6} - \frac{0^2}{6} \\
&= \frac{1}{6}
\end{aligned}$$

Conceptually, $\frac{1}{6}$ represents the volume encompassed within the function $x^2 y$ across the rectangular region $0 \leq x \leq 1$ and $0 \leq y \leq 1$; this volume and the rectangular x, y region are depicted below:





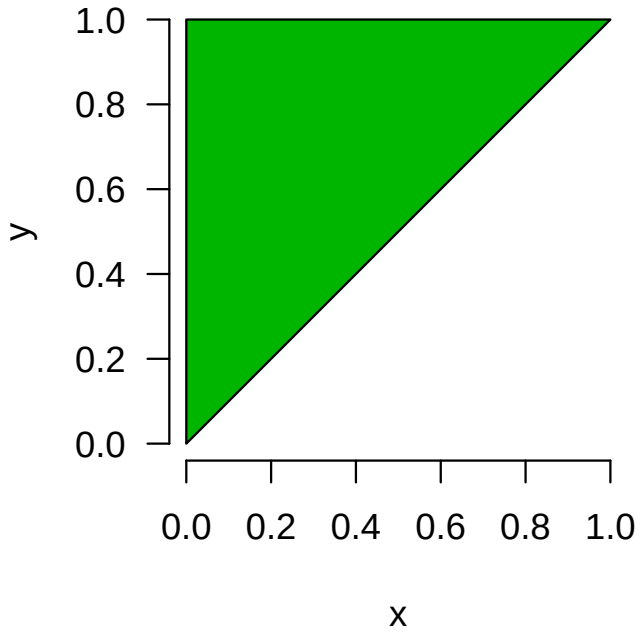
Multiple integrals can taken over any number of input variables:

$$\int \cdots \int f(x_1, \ldots, x_n) \, dx_1 \ldots \, dx_n$$

As practice, integrate $f(x, y, z) = x^3 z / y$ across x, then y and then z (a triple integral). You should get $x^4 z^2 \log(y)/8$.

Non-rectangular regions

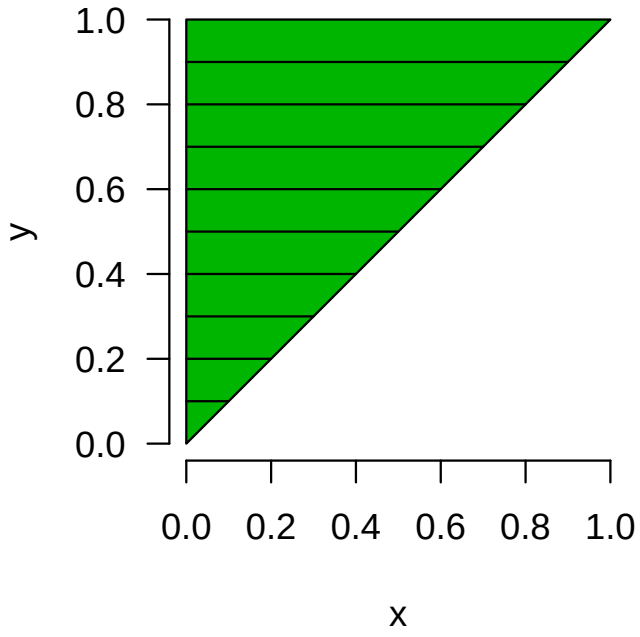
A final complication to be aware of with multiple integrals is that the region over which we are integrating may not be rectangular. For example, let's consider our same function as before, $f(x, y) = x^2 y$, but now our region of integration will be $0 \leq x \leq y \leq 1$:



Note that all the points in the green triangle (and only the points in the green triangle) satisfy $0 \leq x \leq y \leq 1$. At $x = 0.4$ for instance, only points where y is greater than or equal to 0.4 and less than 1 are shaded green. In comparison with the earlier rectangular region, note that the lower right half is no longer included.

The procedure to integrate $\int \int f(x, y)\, dx\, dy$ is the same as before, but we need to pay closer attention to the bounds. Since we're starting with respect to x, let's figure out the appropriate x bounds. Visualizing the green region plotted above is immensely helpful in determining the upper and lower bounds for x (you don't necessarily need to use a computer, but you should always draw the region of integration if it is not rectangular). To determine the lower and upper bounds of the inner integral with respect to x, we need to answer the question: at any given value of y, what values of x are within the green region? The answer is: $0 \leq x \leq y$. For instance, at $y = 0.2$, x can range from 0 to 0.2.

Regarding the range of y (the outer integral), we need to include the entire range of 0 to 1 in order to capture the entire triangular region. To visualize this, imagine drawing a line over the valid x values, and repeating this process for every y value in the range of 0 to 1. A depiction of what this would look like is provided below (hopefully it is clear that the collection of lines would yield the triangle):

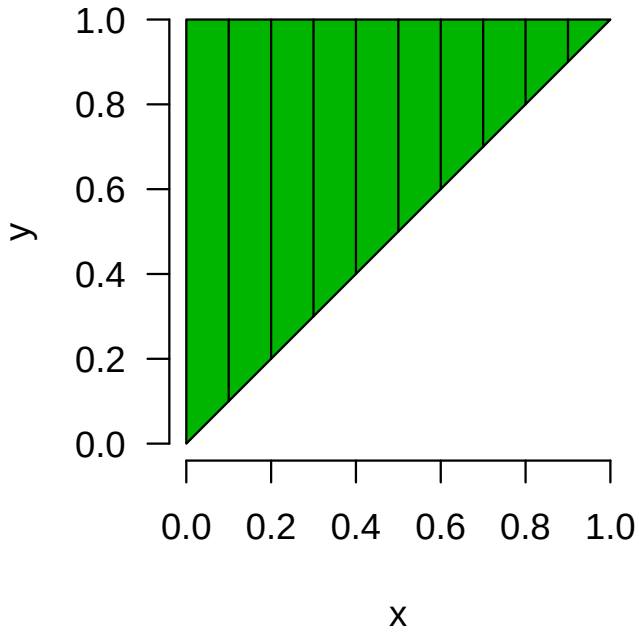


Evaluating the integral, we therefore have:

$$\begin{aligned}
\int_0^1 \int_0^y x^2 y \, dx \, dy &= \int_0^1 \left\{ \frac{x^3 y}{3} \Big|_{x=0}^{y} \right\} dy \\
&= \int_0^1 \left\{ \frac{(y)^3 y}{3} - \frac{(0)^3 y}{3} \right\} dy \\
&= \int_0^1 \frac{y^4}{3} \, dy \\
&= \frac{1^5}{15} - \frac{0^5}{15} \\
&= \frac{1}{15}
\end{aligned}$$

Intuitively, this makes sense – our region is smaller, so the volume should be smaller (1/15 instead of 1/6).

As we mentioned earlier, the order of integration doesn't matter. As practice, integrate $x^2 y$ over the region $0 \leq x \leq y \leq 1$, but this time start by integrating over y. Obviously, you should still get 1/15. Hint: to help visualize things, you should draw a figure that looks like this:



If you want even more practice, trying integrating x^2y over the region $0 \leq y \leq x \leq 1$. You should get $1/10$ (i.e., $1/6 - 1/15$).

Note that while either order gives the same answer, they do not necessarily involve the same amount of work. It is often the case that one route is much easier to calculate than the other, so keep this in mind if one of the integrals becomes difficult to calculate. And remember to draw those regions of integration (really can't stress that enough)!

Chapter 3

Matrix algebra

In this chapter, we will review a number of definitions and results that you should be familiar with from a course on linear algebra. This review will not focus at all on the theory of linear algebra, only on the practical matter of manipulating matrices – hence the title "matrix algebra", which is intended to reflect a focus on algebraic manipulation, not the abstract characterization of linear functions and transformations.

Matrices come up constantly in statistics. Before we begin the review, I'll point out two reasons for this.

Representing data

The data we collect in studies typically has the form of a matrix: for every subject, we collect information on a number of variables (age, weight, blood pressure, etc.). By convention, we write this in matrix form where every subject is assigned a row of the matrix and each variable is assigned a column. For a study with n subjects and p variables, we can represent the data we have collected in the form of a $n \times p$ matrix.

Multivariable manipulations

As we saw in the previous chapter, calculus is important for finding optimal solutions. When multiple variables are present, however, we need to simultaneously solve over all the variables to find the maximum or minimum. For example, if there are two unknown parameters θ_1 and θ_2, and we want to find the most likely values of θ_1 and θ_2 based on data we've collected, we can't just find the most likely value of θ_1 and then find the most likely value of θ_2 – the most likely value of θ_1 is going to *depend* on what θ_2 is. If ten parameters are involved, this process is going to seem hopeless unless we can manipulate and solve all ten equations simultaneously.

Thankfully, manipulating and solving multiple equations at the same time is what matrix algebra is all about.

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:

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

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"):

$$\mathbf{u}^\top \mathbf{v} = \sum_j u_j v_j$$
$$\begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 6 - 2 = 4.$$

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*:

$$(\mathbf{uv}^\top)_{ij} = u_i v_j$$
$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \end{bmatrix} = \begin{bmatrix} 6 & -3 \\ 4 & -2 \end{bmatrix}$$

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 the order of 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:

$$\begin{aligned} (\mathbf{A} + \mathbf{B})^\top &= \mathbf{A}^\top + \mathbf{B}^\top \\ (\mathbf{AB})^\top &= \mathbf{B}^\top \mathbf{A}^\top \\ (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\ (\mathbf{A}^\top)^{-1} &= (\mathbf{A}^{-1})^\top \end{aligned}$$

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.

Furthermore, constructing abstract proofs involves a rather different set of skills than deriving results in calculus or linear algebra, so it's also important to cover techniques and terminology that arise in constructing proofs. This is especially important to read if you have never taken a course in which you were asked to construct mathematical proofs.

If you've never had a course in real analysis, or it's been a while and you've forgotten, this document should be useful if you're thinking about taking Likelihood Theory. **REMINDER: If you are an incoming first-year student, you don't need to worry about this material yet! Just focus on chapters 2 and 3.**

Before we begin, a note on style. Most textbooks and papers provide proofs in an unstructured paragraph style. For the purposes of learning how to prove things, however, I recommend a more structured approach based on making consecutive explicit statements with explicit justifications. There are several reasons for this:

1. When you finish a structured proof, it is very clear exactly which conditions were required, why they were required, and what supporting theorems or results were used. When you're learning in a course, this is extremely valuable as it will be more clear to you how everything is connected.
2. The other major reason is that it's much harder to make a mistake in a structured proof. This doesn't make the proof easier, it just means that in an unstructured proof, one can easily skip steps without realizing it. We'll see some examples of this later.
3. It's also very beneficial from a grading and feedback perspective, as it makes it much clearer to the person grading the proof whether you understand all the pieces or not.

4.1 There exists and For all

Many proofs in theoretical statistics involve convergence. A loose definition was provided earlier: as we collect more and more data, we can be increasingly sure that certain things will happen. We'll discuss probabilistic convergence in the course, but before that, it will help to be familiar with the rigorous definition of (deterministic) convergence. Here's the definition; we'll come back to it later.

Definition. *A sequence $x_1, x_2, \ldots$ of real numbers is said to* ***converge*** *if there exists a real number x such that for all $\epsilon > 0$, there is an integer N such that $|x_n - x| < \epsilon$ for all $n > N$. In this case, we say that x_n converges to x, or that x is the limit of x_n, and write $x_n \to x$ or $\lim x_n = x$ (these mean the same thing). If no such number exists, the sequence is said to diverge.*

To make this definition concrete, let's consider a specific sequence: $e^{-1}, e^{-2}, \ldots$. Intuitively, these numbers are getting closer to zero, so it seems right to say the sequence converges to zero. But does it satisfy the definition? Notice that there are two specific clauses that come up in the above definition, and they occur constantly in mathematical proofs and theorems:

- "there exists": There is at least one number satisfying this condition. Denoted $\exists$.
- "for all": This has to be true in every possible case, no exceptions. Denoted $\forall$.

So, when we say that "there exists a real number x", we're only requiring these conditions to be satisfied for the limit point; in the case of our example $x_n = e^{-n}$, we only need to show that it happens at $x = 0$. Maybe this condition is met at other points, maybe it isn't.

On the other hand, when we say "for all $\epsilon > 0$", the rest of the condition has to true for *every single positive real number*. Maybe we can find some ϵ values for which the condition is met, but it if doesn't hold for all of them, it isn't good enough.

For example, suppose we consider the point $x = 0.001$: does x_n converge to 0.001? Well, if $\epsilon = 0.002$, then yes, we can find a number N such that $|x_n - x| < \epsilon$ for all $n > N$. For example, $N = 6$ works:

$$\begin{aligned} |e^{-6} - 0.001| &= 0.0015 < 0.002 \\ |e^{-7} - 0.001| &= 0.000088 < 0.002 \\ &\ldots \end{aligned}$$

However, this isn't enough – we need to be able to find a suitable N for *all* positive ϵ. If $\epsilon = 0.0005$, for example, we would never be able find an N that works. Therefore, x_n does not converge to 0.001.

One final important point to note here is that the order of these clauses matters. For every ϵ, there exists a suitable N. We choose ϵ first, and can then find an N that works for that ϵ – we don't need to find an N that works for all ϵ (and in fact, there typically isn't such a N). If $\epsilon = 0.001$, then $N = 7$ works. If $\epsilon = 0.00000000000000000000000001$, we are going to have to choose a much larger N (but we can *always* find such an N).

Finding N is fairly simple in this case; we can simply take the log of both sides:

$$\begin{aligned} e^{-n} &< \epsilon \\ -n &< \log \epsilon \\ n &> -\log \epsilon. \end{aligned}$$

So, $N = \lceil -\log \epsilon \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function (i.e., round up the number inside to the next integer). To make things even more concrete, one could think of $N(\epsilon)$ as a function:

```
n <- function(eps) {ceiling(-log(eps))}
n(0.001)
## [1] 7
n(0.0000001)
## [1] 17
n(0.00000000000000001)
## [1] 40
```

4.2 Structured proofs

That was a lot of preliminary stuff, but it's very important to understand definitions before you move on to proofs. Now, let's prove that $e^{-n} \to 0$. What, specifically, do we need to prove? That for all $\epsilon > 0$, there is an integer N such that $|e^{-n}| < \epsilon$ for all $n > N$. So, we need to start with "Let $\epsilon > 0$" (this is how **a lot** of proofs start):

Proof. Let $\epsilon > 0$, and let $N = \lceil -\log \epsilon \rceil$. Then for all $n > N$,

$$\begin{aligned}
|e^{-n}| &= e^{-n} && e^{-n} \text{ is always positive} \\
&< e^{-N} && e^{-n} \text{ is strictly decreasing} \\
&= \exp\{-\lceil -\log \epsilon \rceil\} && \text{Definition of } N \\
&\leq e^{\log \epsilon} && e^{-n} \text{ decreasing} \\
&= \epsilon
\end{aligned}$$

Thus, $\mathbf{x}_n \to 0$. □

The left column is a series of statements or claims; this is the main logic of what's happening in the proof. Chaining all the lines together, we have $|e^{-n}| < \epsilon$. Thus, we've done what we needed to do: given any positive ϵ, I can find always find an N that meets the requirement. Thus, $e^{-n} \to 0$ by the definition of convergence.

The right column provides the justification for each step. For example, in the fourth line we claimed that $\exp\{-\lceil -\log \epsilon \rceil\} \leq e^{\log \epsilon}$. How do we know that this is true? Well, $\lceil x \rceil \geq x$ since the operation involves rounding up, and e^{-x} is a decreasing function of x. So by replacing the argument to e^{-x} (i.e., $\lceil -\log \epsilon \rceil$) with something smaller (i.e., $-\log \epsilon$), the result must be larger (or equal, since $-\log \epsilon$ could be an integer already).

How far to go with these justification depends on what the reader of the proof would likely consider obvious, and is therefore a judgment call. For example, we could include a proof of the fact that e^{-x} is a decreasing function of x, but in my judgment this is going off on a bit of a tangent that distracts from the main point of the proof. In a similar fashion, I didn't even provide a justification for the final step, but to be really thorough, I could have added that exp() and log() are inverse functions. In general, for the purposes of a course, you should err on the side of being very thorough and justifying every step. In a paper or thesis, however, going into this level of detail for every proof is probably unnecessary as the audience (other people with PhDs in statistics or biostatistics) will probably find many of the steps obvious and not requiring justification.

One final comment: as the above proof should hopefully make clear, a final proof does not describe one's thought process in terms of how you arrived at the result. Clearly I didn't know what I should set N equal to until I'd worked through the math to solve for n in the previous section. If you're reading the proof for the first time, the line "let $N = \lceil -\log \epsilon \rceil$" is going to seem mysterious; where did *that* come from? Keep in mind that we're writing a proof, not a novel of self-discovery. The point is to construct an iron-clad, rigorous argument, not to communicate our thoughts and feelings and realizations. It might not be immediately obvious *why* I am letting $N = \lceil -\log \epsilon \rceil$, but it should be completely obvious that I *can* set $N = \lceil -\log \epsilon \rceil$. One can certainly write about the thought process and intuition behind the proof, but the place for this is outside the formal mathematical proof – we don't want to mix formal logic with informal intuition.

4.3 Convergence

Let's go through a few more proofs involving convergence (refer back to the definition at needed). For example, you might find the "there exists a real number x" clause in the definition of convergence unsatisfying, as it leaves open the possibility that many such numbers exist and a sequence might converge to lots of different things. However, this is in fact not possible.

Theorem. *If $x_n \to a$ and $x_n \to b$, then a=b.*

Proof. Let $\epsilon > 0$.

$$\begin{array}{lll} \exists N_a : n > N_a \implies |x_n - a| < \frac{\epsilon}{2} & & x_n \to a \\ \exists N_b : n > N_b \implies |x_n - b| < \frac{\epsilon}{2} & & x_n \to b \end{array}$$

Thus, for all $n > N = \max(N_a, N_b)$, we have

$$\begin{array}{lll} |a - b| = |a - x_n + x_n - b| & & \\ \quad \leq |x_n - a| + |x_n - b| & & \text{Triangle inequality} \\ \quad < \epsilon & & \end{array}$$

$\therefore a = b$ □

In the above proof, the symbol $\implies$ means "implies" and the symbol $\therefore$ means "therefore" (writing out the words is fine too, I'm just introducing the symbols because I occasionally use them in class). This proof introduces a standard technique that comes up often: since $x_n \to a$, I can get x_n as close to a as I want. I could find an n such that $|x_n - a| < \epsilon$, but why stop there? I can keep going and get $|x_n - a| < \epsilon/2$, which ends up making the rest of the proof more clear. I could keep going even further and get $|x_n - a| < \epsilon/1000000$ if I wanted, but this isn't necessary for what I'm trying to show.

Summarizing the above proof, if $x_n \to a$ and $x_n \to b$, then for any positive number ϵ, it must be the case that $|a - b| < \epsilon$. In other words, they have to be equal ($a - b = 0$). Here, the triangle inequality states that

$$|x + y| \leq |x| + |y|.$$

Applied to line three, this gives

$$\begin{array}{c} |a - x_n + x_n - b| \leq |a - x_n| + |x_n - b| \\ |a - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2}. \end{array}$$

Hopefully the logic here is clear and all the steps of the proof make sense (if not, feel free to come by my office). You may very well be thinking "I understand the proof, but I could never come up with that!" This is normal; don't worry about it. Any technique that you've never seen before is going to seem incredible and clever, but the more exposure to proofs you have, the more you will recognize all of the above steps as fairly standard and you will definitely be able to recognize when you need them in the future.

Theorem. *If $x_n \to x$, then the sequence x_n is bounded.*

Proof.

$$\begin{array}{lll} \exists N : n > N \implies |x_n - x| < 1 & & x_n \to x \\ \exists r = \max\{1, |x_1 - x|, \ldots, |x_N - x|\} & & \text{Maximum of finite set exists} \end{array}$$

Therefore, $\{x_n\}_{n=1}^{\infty}$ is bounded (above by $x + r$, below by $x - r$). □

This is a fairly simple proof, but it illustrates a few important points. First, the choice of "1" in the first line is completely arbitrary; I could have chosen any number. Second, and more importantly, the heart of this proof is the second line, where we establish that there is a number r that bounds $|x_n - x|$ and therefore also bounds x_n. However, it is important to recognize that this is a *claim* (of existence), and it requires a justification. It is very easy in an unstructured proof to just say "let r by the maximum of $\{1, |x_1 - x|, \ldots, |x_N - x|\}$".

But how do you know that this maximum exists? If the set were infinite, we wouldn't know this. For example, we saw earlier that $e^{-n} \to 0$, so by our proof, $e^{-1}, e^{-2}, \ldots$ is bounded. However, consider the set $\{\ldots, e^{-(-2)}, e^{-(-1)}, e^{-0}, e^{-1}, e^{-2}, \ldots\}$. This sequence converges as $n \to \infty$, but because there is an infinite collection of numbers leading up to x_N, our proof above doesn't work – we don't know that this set has a maximum (and indeed, it doesn't have a maximum, and the set isn't bounded).

This is a common way in which it is easy to skip steps in an unstructured proof. Of course, one could still write "Let $r = \max\{1, |x_1 - x|, \ldots, |x_N - x|\}$" in a structured proof, but it would be immediately clear that the right hand column is empty and that this statement has not been justified. Now, sometimes it's fine to say "let" without a justification: writing "Let $\epsilon > 0$" and justifying it with "positive numbers exist" is pedantic. Going back to an earlier proof, one could argue that letting $N = \lceil -\log \epsilon \rceil$ should be justified in the sense that this only exists if $\epsilon > 0$ (which it is), and furthermore we should be more careful in the definition: $N = \max\{1, \lceil -\log \epsilon \rceil\}$ is more technically sound (if ϵ is very large, N could be negative according to our original definition). It never hurts to think about these things, but at the same time, it's something of a judgment call and I felt that going into these justifications was a distraction from the main idea. On the other hand, the "maximum of a finite set" justification is quite important because it's really the main idea of the proof.

Now would be a good time to try proving things on your own. Here are three theorems to start with.

Theorem. *Suppose $x_n \to x$ and $y_n \to y$. Then $x_n + y_n \to x + y$.*

Theorem. *Suppose $x_n \to x$ and $y_n \to y$. Then $x_n y_n \to xy$.*

Theorem. *Suppose $x_n \to x$, with $x_n \neq 0$ for all n and $x \neq 0$. Then $1/x_n \to 1/x$.*

For the third theorem, is $x_n \neq 0$ actually required? If we know that $x \neq 0$, can we still have $x_n = 0$?

See here for solutions, although you should definitely try proving them on your own first before looking at the solutions.

4.4 Continuity

Another major concept in analysis is that of continuity (again, we came across this earlier, but it's worth revisiting this concept not that we've formally defined limits and convergence). Suppose we have a function f and a convergent sequence $x_n \to x$. Do we know that as $f(x_n) \to f(x)$? The answer is that no, this doesn't always happen. Only some functions have this property, and those functions are said to be "continuous". Below is the formal definition.

Definition. *A function f is said to be **continuous at the point** x_0 if for every $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(x_0)| < \epsilon$ for all $x : |x - x_0| < \delta$. If f is continuous at every point in its domain, then the entire function f is said to be **continuous**.*

This is similar to the concept of convergent sequences, except now instead of a countable sequence of points $x_1, x_2, \ldots$, we are concerned with all the points in the "neighborhood" $\{x : |x - x_0| < \delta\}$; that is, the points near x_0. Many of the techniques that we encountered earlier with convergence are very similar to the techniques one uses with continuity. These techniques are often referred to as "delta-epsilon" techniques. For example, the techniques used in the following proof should look fairly familiar by now.

Theorem. *Suppose $x_n \to x_0$ and f is continuous at x_0. Then $f(x_n) \to f(x_0)$.*

Proof. Let $\epsilon > 0$.

$$\begin{aligned} &\exists \delta : |x - x_0| < \delta \implies |f(x) - f(x_0)| < \epsilon && f \text{ continuous at } x_0 \\ &\exists N : n > N \implies |x_n - x_0| < \delta && x_n \to x_0 \end{aligned}$$

Therefore, $n > N \implies |f(x_n) - f(x_0)| < \epsilon$. □

One important thing to note here is that the order of these steps is important. Students often switch the order of the first two lines in this proof, but this makes no sense. The claim that $\exists N : n > N \implies |x_n - x_0| < \delta$ is meaningless if δ hasn't been defined yet. This isn't just semantics: if you were trying to determine how large n had to be in order to ensure that $f(x_n)$ is within a certain tolerance of $f(x_0)$, you couldn't start by finding N. Without using continuity first, you'd have no idea how close x_n must be to x_0 in order to ensure that $f(x_n)$ is within ϵ of $f(x_0)$.

It is worth noting that we can actually make the above theorem into an "if and only if" statement, and thus, an equivalent definition of continuity, but we would have to add the condition that $f(x_n) \to f(x)$ for all sequences $x_n \to x$. For example, a function could satisfy $f(x_n) \to f(x_0)$ for increasing sequences $x_n \nearrow x_0$ but not for decreasing sequences $x_n \searrow x_0$; such functions are not continuous at x_0.

Here are some additional proof exercises related to continuity for you to practice with. Note that the sum and product proofs are very similar to the corresponding proofs for sequences; however, they are still useful exercises if you've never done delta-epsilon proofs before.

Theorem. *Let the functions f and g be continuous at x_0. Then $h = f + g$ is continuous at x_0.*

Theorem. *Let the functions f and g be continuous at x_0. Then $h = f \cdot g$ is continuous at x_0.*

Theorem. *Let the function f be continuous at x_0 and the function g be continuous at $f(x_0)$. Then $h(x) = g(f(x))$ is continuous at x_0.*

Exercise. *Write an R function `n(eps)` that returns the smallest N for which $n > N \implies |f(x_n) - f(x_0)| < \epsilon$ for $x_n = 2^{1/n}$ and $f(x) = e^x$.*

See here for solutions, although you should definitely try proving them on your own first before looking at the solutions.

4.5 Solutions

Convergence

Theorem. *Suppose $x_n \to x$ and $y_n \to y$. Then $x_n + y_n \to x + y$.*

Proof. Let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists N_x : n > N_a \implies |x_n - x| < \frac{\epsilon}{2} & x_n \to x \\ ② & \exists N_y : n > N_y \implies |y_n - y| < \frac{\epsilon}{2} & y_n \to y \end{array}$$

Thus, for all $n > N = \max(N_x, N_y)$, we have

$$\begin{array}{ll} |x_n + y_n - (x + y)| \le |x_n - x| + |y_n - y| & \text{Triangle inequality} \\ < \epsilon & ①, ② \end{array}$$

□

Theorem. *Suppose $x_n \to x$ and $y_n \to y$. Then $x_n y_n \to xy$.*

Proof. First, let's establish an identity:

$$\begin{aligned} x_n y_n - xy &= x_n y_n - x_n y + x_n y - xy \\ &= x_n(y_n - y) + y(x_n - x) \\ &= (x_n - x + x)(y_n - y) + y(x_n - x) \\ &= (x_n - x)(y_n - y) + x(y_n - y) + y(x_n - x) \end{aligned}$$

Now, let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists N_x : n > N_x \implies |x_n - x| < \min\left(\frac{\sqrt{\epsilon}}{3}, \frac{\epsilon}{3|y|}\right) & x_n \to x \\ ② & \exists N_y : n > N_y \implies |y_n - y| < \min\left(\frac{\sqrt{\epsilon}}{3}, \frac{\epsilon}{3|x|}\right) & y_n \to y \end{array}$$

Thus, for all $n > N = \max(N_x, N_y)$, we have

$$\begin{array}{ll} |x_n y_n - xy| = |(x_n - x)(y_n - y) + x(y_n - y) + y(x_n - x)| & \text{Identity above} \\ \quad \leq |x_n - x||y_n - y| + |x||y_n - y| + |y||x_n - x| & \text{Triangle inequality} \\ \quad < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} & ①, ② \\ \quad = \epsilon & \end{array}$$

In the construction of N above, note that we are assuming $x, y \neq 0$. If either is zero, the second term in the sum can simply be omitted, as the corresponding term below is zero. □

Theorem. *Suppose $x_n \to x$, with $x_n \neq 0$ for all n and $x \neq 0$. Then $1/x_n \to 1/x$.*

Proof.

First, let us note that $|a - b| < \frac{1}{2}|b| \implies |a| > \frac{1}{2}b$. This is fairly obvious when you think about it; to prove it, we can break the claim up into cases:

- $b > 0$ and $a > b$: $a > b > b/2$
- $b > 0$ and $b > a$: $b - a < \frac{1}{2}b$, so $a > \frac{1}{2}b$

The cases where $b < 0$ follow the same reasoning. Now, let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists N_1 : n > N_1 \implies |x_n - x| < \frac{1}{2}|x|^2\epsilon & x_n \to x \\ ② & \exists N_2 : n > N_2 \implies |x_n - x| < \frac{1}{2}|x| & x_n \to x \\ ③ & \text{so that } |x_n| > \frac{1}{2}|x| & ②, \text{see above} \end{array}$$

Thus, for all $n > N = \max(N_1, N_2)$, we have

$$\begin{array}{lll} \left|\frac{1}{x_n} - \frac{1}{x}\right| = \left|\frac{x - x_n}{x_n x}\right| & & \\ \quad \leq \frac{2}{|x|^2}|x_n - x| & ③ & \\ \quad < \epsilon & ① & \end{array}$$

Note that in this third theorem, the requirement that $x_n \neq 0$ is unnecessary. As we see from ③, if $x_n \to x$ and $x \neq 0$, then there is an N such that $x_n \neq 0$ for all $n > N$.

□

Continuity

The first two theorems are essentially the same as their sequence counterparts, but the differences are worth paying attention to.

Theorem. *Let the functions f and g be continuous at x_0. Then $h = f + g$ is continuous at x_0.*

Proof. Let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists \delta_f : |x - x_0| < \delta_f \implies |f(x) - f(x_0)| < \frac{\epsilon}{2} & f \text{ continuous at } x_0 \\ ② & \exists \delta_g : |x - x_0| < \delta_g \implies |g(x) - g(x_0)| < \frac{\epsilon}{2} & g \text{ continuous at } x_0 \end{array}$$

Thus, for all $x : |x - x_0| < \delta = \min(\delta_f, \delta_g)$, we have

$$\begin{array}{lll} |h(x) - h(x_0)| &= |f(x) + g(x) - f(x_0) - g(x_0)| & \text{Def } h \\ &\leq |f(x) - f(x_0)| + |g(x) - g(x_0)| & \text{Triangle inequality} \\ &\leq \epsilon & ①, ② \end{array}$$

□

Theorem. *Let the functions f and g be continuous at x_0. Then $h = f \cdot g$ is continuous at x_0.*

Proof. Let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists \delta_f : |x - x_0| < \delta_f \implies |f(x) - f(x_0)| < \dfrac{\sqrt{\epsilon}}{3} + \dfrac{\epsilon}{3|g(x_0)|} & f \text{ continuous at } x_0 \\ ② & \exists \delta_g : |x - x_0| < \delta_g \implies |g(x) - g(x_0)| < \dfrac{\sqrt{\epsilon}}{3} + \dfrac{\epsilon}{3|f(x_0)|} & g \text{ continuous at } x_0 \end{array}$$

Thus, for all $x : |x - x_0| < \delta = \min(\delta_f, \delta_g)$, we have

$$\begin{array}{lll} |h(x) - h(x_0)| &= |f(x)g(x) - f(x_0)g(x_0)| & \text{Def } h \\ &\leq |\{f(x) - f(x_0)\}\{g(x) - g(x_0)\}| & \\ &\quad + |f(x_0)\{g(x) - g(x_0)\}| & \\ &\quad + |g(x_0)\{f(x) - f(x_0)\}| & \text{See earlier proof} \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} & ①, ② \\ &= \epsilon & \end{array}$$

□

Theorem. *Let the function f be continuous at x_0 and the function g be continuous at $f(x_0)$. Then $h(x) = g(f(x))$ is continuous at x_0.*

Proof. Let $\epsilon > 0$.

$$\begin{array}{lll} ① & \exists \eta : |y - f(x_0)| < \eta \implies |g(y) - g(f(x_0))| < \epsilon & g \text{ continuous at } f(x_0) \\ ② & \exists \delta : |x - x_0| < \delta \implies |f(x) - f(x_0)| < \eta & f \text{ continuous at } x_0 \end{array}$$

Thus, for all $x : |x - x_0| < \delta$, we have

$$\begin{array}{lll} |h(x) - h(x_0)| &= |g(f(x)) - g(f(x_0))| & \text{Def } h \\ &< \epsilon & ② \implies ① \end{array}$$

□

Exercise: Write an R function `n(eps)` that returns the smallest N for which $n > N \implies |f(x_n) - f(x_0)| < \epsilon$ for $x_n = 2^{1/n}$ and $f(x) = e^x$.

Conceptually, this is a three-part process:

1. Determine what x_n is converging to. Here, $x_n \to 1$.
2. Determine the largest value of delta that satisfies $e^{1+\delta} - e^1 < \epsilon$.
3. Determine the smallest value of N such that $2^{1/n} - 1 < \delta$.

```r
n <- function(eps) {
  delta <- log(eps + exp(1)) - 1
  ceiling(1/log2(1+delta))
}
```

Let's test this out and make sure it works:

```r
n(0.01)
## [1] 190
exp(2^(1/189)) - exp(1)  ## 189 not good enough
## [1] 0.01000582
exp(2^(1/190)) - exp(1)  ## 190 within 0.01
## [1] 0.009952969
```