Math review for new biostatistics students

Patrick Breheny

Contents

Chapter 1

Introduction

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

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

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

In principle, any idea from math *could* come up and be helpful to a statistician. In reality, however, certain ideas come up far more often than others, and this guide focuses on topics of greatest relevance. A good example is trigonometry: this almost never comes up in statistics. There is really no need to spent 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.

At present, this guide is very much a work in progress – in places, not much more than an outline. However, our hope is to continue adding to it and making it more useful over the years. In particular, we realize that exercises and solutions would be very helpful, and we have not yet added any.

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*.

Commentary

A few remarks on this definition and its implications:

1. This is an *extremely* general definition. A and B could be single numbers, but they could collections of numbers... A could consist of sets of 7 numbers and B could consist of intervals, which are themselves infinite collections 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. As a footnote, we *could* extend the domain of the function to include negative numbers, but then the range would have to include complex numbers.

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 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. As a footnote, the astute reader will point out that $f(x) = x^2$ could be 1:1 if I change its domain.

2.2 Limits and continuity

2.3 Derivatives

2.4 Optimization

2.5 Integration

2.6 Log and exp

Exp definition

The exponential function a^x actually has a pretty complicated definition:

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.

Exp rules

$$a^{x+y} = a^x a^y \tag{2.1}$$
$$a^{x-y} = \frac{a^x}{a^y} \tag{2.2}$$
$$(a^x)^y = a^{xy} \tag{2.3}$$
$$(ab)^x = a^x b^x \tag{2.4}$$

Exp limits

$$\lim_{x\to\infty} a^x = \infty \quad \text{if } a > 1 \tag{2.5}$$
$$\lim_{x\to-\infty} a^x = 0 \quad \text{if } a > 1 \tag{2.6}$$
$$\lim_{x\to\infty} a^x = 0 \quad \text{if } 0 < a < 1 \tag{2.7}$$
$$\lim_{x\to-\infty} a^x = \infty \quad \text{if } 0 < a < 1 \tag{2.8}$$
$$\lim_{h\to 0} \frac{e^h - 1}{h} = 1 \tag{2.9}$$
$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e \tag{2.10}$$

Exp derivatives and integrals

$$\frac{d}{dx}e^x = e^x \tag{2.11}$$
$$\frac{d}{dx}e^u = e^x \frac{du}{dx} \tag{2.12}$$
$$\int e^x\, dx = e^x \tag{2.13}$$
$$\frac{d}{dx}a^x = a^x \log(a) \tag{2.14}$$
$$\int a^x\, dx = \frac{a^x}{\log a} \quad (a \neq 1) \tag{2.15}$$

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

Log definition

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

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

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

$$\log x = \log_e x$$

in other words,

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

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

$$\log(e^x) = x \tag{2.16}$$
$$e^{\log x} = x \tag{2.17}$$
$$\log e = 1. \tag{2.18}$$

Log rules

$$\log_a(xy) = \log_a x + \log_a y \tag{2.19}$$

$$\log_a \frac{x}{y} = \log_a x - \log_a y \tag{2.20}$$

$$\log_a(x^y) = y \log_a x \tag{2.21}$$

$$\log_a x = \frac{\log x}{\log a} \tag{2.22}$$

Log limits

If $a > 1$, then

$$\lim_{x\to\infty} \log_a x = \infty \tag{2.23}$$

$$\lim_{x\to 0^+} \log_a x = -\infty \tag{2.24}$$

$$\tag{2.25}$$

Log derivatives and integrals

$$\frac{d}{dx} \log x = x^{-1} \tag{2.26}$$

$$\frac{d}{dx} \log u = u^{-1} \frac{du}{dx} \tag{2.27}$$

$$\int \frac{1}{x}\, dx = \log |x| \tag{2.28}$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \log a} \tag{2.29}$$

2.7 Integration techniques

Substitution + Jacobian

Integration by parts

Kernel trick

2.8 Sequences and series

2.9 Partial derivatives

2.10 Multiple integrals

Non-rectangular boundaries

Chapter 3

Matrix algebra

modeling, relevance to statistics

3.1 Definitions and conventions

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

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

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

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

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

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

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

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

3.2 Basic operations

Transposition

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

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

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

Addition

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

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

The other kind is *matrix addition*:

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

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

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

Multiplication

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

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

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

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

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

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

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

The following elementary algebra rules carry over to matrix algebra:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \tag{3.1}$$
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \tag{3.2}$$
$$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B} \tag{3.3}$$

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

Inner and outer products

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

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

$$\mathbf{u}^\top \mathbf{v} = \sum_j u_j v_j \tag{3.4}$$
$$\begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 6 - 2 = 4. \tag{3.5}$$

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

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

$$(\mathbf{uv}^\top)_{ij} = u_i v_j \tag{3.6}$$
$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \end{bmatrix} = \begin{bmatrix} 6 & -3 \\ 4 & -2 \end{bmatrix} \tag{3.7}$$

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

3.3 Special matrices

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

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

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

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

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

Consider in particular the following diagonal matrix:

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

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

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

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

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

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

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

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

3.4 Inversion and related concepts

Suppose $\mathbf{A}\mathbf{x} = \mathbf{B}$ 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 $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, 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 $\mathbf{B}\mathbf{A}^{-1}$; we need to *left* multiply by the inverse and order or multiplication matters.

If two vectors $\mathbf{u}$ and $\mathbf{v}$ satisfy $\mathbf{u}^\top \mathbf{v} = 0$, they are said to be *orthogonal* to each other. If all the columns and rows of a matrix $\mathbf{A}$ are orthogonal to each other and satisfy $\mathbf{a}^\top \mathbf{a} = 1$, then $\mathbf{A}$ (transposed) can serve as its own inverse: $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$. 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 $\mathbf{X}\mathbf{X}^\top \neq \mathbf{I}$; 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 $\mathbf{A}\mathbf{A} = \mathbf{A}$.

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

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

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

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

Additional helpful identities:

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

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

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

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