Vectors, inequalities, proofs

Patrick Breheny

Introduction

- Before we get to likelihood theory, we are going to spend the first part of this course reviewing/extending/deepening our knowledge of mathematical and statistical tools
- In particular, lower-level analysis and mathematical statistics courses often focus on single-variable results
- In practice, however, statistics is almost always a multivariate pursuit
- Thus, one of the things we will focus on in this review is covering results you may have seen for single variables in terms of vectors

Asymptotic theory

- A large amount (but not all) of statistical theory is based on asymptotic, or large sample, arguments
- Exact theoretical results are often very complicated and difficult to obtain, but we can typically simplify the problem greatly by considering what happens as $n \to \infty$
- A core idea here from analysis is that of a convergent sequence: x_n converges to x if, as n gets larger, x_n gets closer and closer to x
- We'll discuss this more next week, but first, we need to take a step back and define what it means for x_n to be "close" to x

Norms: Introduction

- Throughout this course, we need to be able to measure the distance between two vectors, or equivalently, the size of a vector; such a measurement is called a *norm*
- This is straightforward for scalars: the distance from a to b is $|a - b|$
- Vectors are more complicated; as we will see, there are many ways of measuring the size of a vector
- In order to be a meaningful measure of size, however, there are certain conditions any norm must satisfy

Norm: Definition

- **Definition:** A *norm* is a function $\|\cdot\| : \mathbb{R}^d \mapsto \mathbb{R}$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
  - $\|\mathbf{x}\| \geq 0$, with $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$ (positivity)
  - $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$ for any $a \in \mathbb{R}$ (homogeneity)
  - $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)
- The triangle inequality is also sometimes expressed as

$$\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|,$$

or

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}),$$

where $d(\mathbf{x}, \mathbf{y})$ quantifies the distance between $\mathbf{x}$ and $\mathbf{y}$

Reverse triangle inequality

- A related inequality:
- **Theorem (reverse triangle inequality):** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|$$

- **Corollary:** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned} \|\mathbf{x}\| - \|\mathbf{y}\| &\leq \|\mathbf{x} + \mathbf{y}\| \\ \|\mathbf{y}\| - \|\mathbf{x}\| &\leq \|\mathbf{x} + \mathbf{y}\| \\ \|\mathbf{y}\| - \|\mathbf{x}\| &\leq \|\mathbf{x} - \mathbf{y}\| \end{aligned}$$

Examples of norms

- By far the most common norm is the Euclidean (L_2) norm:

$$\|\mathbf{x}\|_2 = \sqrt{\sum\nolimits_i x_i^2}$$

- However, there are many other norms; for example, the Manhattan (L_1) norm:

$$\|\mathbf{x}\|_1 = \sum\nolimits_i |x_i|$$

- Both Euclidean and Manhattan norms are members of the L_p family of norms: for $p \geq 1$,

$$\|\mathbf{x}\|_p = \left(\sum\nolimits_i |x_i|^p\right)^{1/p}$$

Examples of norms (cont'd)

- Another norm worth knowing about is the L_∞ norm:
$$\|\mathbf{x}\|_\infty = \max_i |x_i|,$$
which is the limit of the family of L_p norms as $p \to \infty$
- One last "norm" worth mentioning is the L_0 norm:
$$\|\mathbf{x}\|_0 = \sum_i 1\{x_i \neq 0\};$$
be careful, however: this is not a proper norm! (why not?)

Inner products

- The Euclidean norm can also be thought of in terms of something called an *inner product*:

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i$$

- Inner products come up all the time in statistics and mathematics, and can also be written using angle bracket notation: $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$
- Two critical things to remember:
  - The inner product $\mathbf{x}^\top \mathbf{y}$ takes two vectors and returns a scalar
  - Writing $\mathbf{x}^2$ is meaningless — never do this — because there are two ways to multiply a vector $\mathbf{x}$ with itself

Outer products

- When we change which vector is transposed, instead of getting something simpler (a scalar), we get something more complicated (a matrix)
- The operation $\mathbf{x}\mathbf{y}^\top$ results in an $n \times n$ matrix where the element in row i, column j is $x_i y_j$
- This is known as *outer product*
- We will see inner and outer products all the time in this course, so this needs to be something you understand instantly in order to read equations and formulas that will appear (we will talk more about matrix algebra in a future lecture)

Matrix norms

- There are also matrix norms, although we will not work with these as often
- In addition to the three requirements listed earlier, matrix norms must also satisfy a requirement of *submultiplicativity*:

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|;$$

  unlike the other requirements, this only applies to $n \times n$ matrices
- The simplest matrix norm is the *Frobenius* norm

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

Spectral norm

- Another common matrix norm is the *spectral norm*:

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}},$$

  where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{A}^\top\mathbf{A}$
- There are many other matrix norms

Cauchy-Schwarz

- There are several important inequalities involving norms that you should be aware of; the most important is the Cauchy-Schwarz inequality, arguably the most useful inequality in all of mathematics
- **Theorem (Cauchy-Schwarz):** For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
$$\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2,$$
where equality holds only if $\mathbf{x} = a\mathbf{y}$ for some scalar a
- Note: the above is *the* Cauchy-Schwarz inequality, but in statistics, its probabilistic version goes by the same name:
$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$
for random variables X and Y, with equality iff $X = aY$

Hölder's inequality

- The Cauchy-Schwarz inequality is actually a special case of Hölder's inequality:
- **Theorem (Hölder):** For $1/p + 1/q = 1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q,$$

  again with exact equality iff $\mathbf{x} = a\mathbf{y}$ for some scalar a (unless p or q is exactly 1)
- Probabilistic analogue:

$$\mathbb{E}|XY| \leq \sqrt[p]{\mathbb{E}|X|^p}\,\sqrt[q]{\mathbb{E}|Y|^q}$$

Jensen's inequality

- Another extremely important inequality is Jensen's inequality; surely you've seen it before, but perhaps not in vector form
- **Theorem (Jensen):** For $\mathbf{a}, \mathbf{x} \in \mathbb{R}^d$ with $a_i > 0$ for all i, if g is a convex function, then

$$g\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i g(x_i)}{\sum_i a_i}$$

- Probabilistic analog:

$$g(\mathbb{E}X) \leq \mathbb{E}g(X)$$

- The inequalities are reversed if g is concave

Relationships between norms

- Getting back to the different norms, there are many important relationships between norms that are often useful to know
- **Theorem:** For all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$$

- Obvious, but useful:

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq d\|\mathbf{x}\|_\infty$$
$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{d}\|\mathbf{x}\|_\infty$$

Equivalence of norms

- The relationships on the previous slide suggest the following statement, which is in fact always true: for any two norms a and b, there exist constants c_1 and c_2 such that

$$\|\mathbf{x}\|_a \leq c_1 \|\mathbf{x}\|_b \leq c_2 \|\mathbf{x}\|_a$$

- This result is known as the *equivalence of norms* and means that we can often generalize results for any one norm to all norms
- For example, we will often encounter results that look like:

$$A = B + \|\mathbf{r}\|$$

  and show that $\|\mathbf{r}\| \to 0$, so $A \to B$

Equivalence of norms (cont'd)

- By the equivalence of norms, if, say, $\|\mathbf{r}\|_1 \to 0$, then $\|\mathbf{r}\|_2 \to 0$ and so on for all norms (except not the L_0 "norm"!)
- In this course, we will almost always be working with the Euclidean norm, so much so that I will often write $\|\mathbf{x}\|$ to mean the Euclidean norm and not even bother with the subscript
- That said, it is important to note that with these relationships, we can always derive corollaries that extend results to other norms

Equivalence of matrix norms

- Like vector norms, matrix norms are also equivalent
- For example,

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{r}\|\mathbf{A}\|_2,$$

where r is the rank of $\mathbf{A}$

Inclusion

- Many of the inequalities we will work with in this class involve random variables, so it is worth covering a simple but important point that doesn't seem to have a name (that I know of) but I will call the *principle of inclusion*
- **Theorem (Inclusion):** Suppose $a < b$ and Y is a random variable. Then

$$\mathbb{P}\{Y < a\} \leq \mathbb{P}\{Y < b\}$$
$$\mathbb{P}\{b < Y\} \leq \mathbb{P}\{a < Y\}$$

- These results follow from the monotonicity of probability: given sets A and B, if $A \subset B$, then $P(A) \leq P(B)$

Intuition

- In words, if the "big side" gets bigger or the "small side" gets smaller, the probability goes up
- For example, combining this with the triangle inequality:

$$\mathbb{P}\{X < |a+b|\} \leq \mathbb{P}\{X < |a| + |b|\}$$

  (the "big side" got bigger, so the probability went up)
- This is not hard to understand, but also a common source of mistakes, because it's easy to wind up with the inequality going the wrong way:

$$\mathbb{P}\{|a+b| < X\} \leq \mathbb{P}\{|a| + |b| < X\}$$

Inclusion and convergence

- Inclusion will come up frequently as we study convergence, as it is a very common proof technique
- Suppose now we have sequences of sets A_n and B_n, and for every n, $A_n \subseteq B_n$
    - If $P(A_n) \to 1$, then $P(B_n) \to 1$
    - If $P(B_n) \to 0$, then $P(A_n) \to 0$

Union bound

- Another useful technique is the *union bound*, also known as Boole's inequality
- **Theorem (Union bound):** For a countable set of events $A_1, A_2, A_3, \ldots,$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

- This idea can also be stated in terms of intersections: For a *finite* set of events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \ge \sum_{i=1}^{n} \mathbb{P}(A_i) - n + 1.$$

Proofs

- We will spend a considerable amount of time in this course proving results, both in class and on homework and exams
- It is therefore worth discussing some general principles for writing mathematical proofs, especially for students who have not previously taken a proof-based course
- I'll illustrate this with a simple example involving convergence, the logical structure of which demonstrates several important ideas that recur throughout mathematical statistics

Convergence

- **Definition:** A sequence of scalars x_n converges to x, written $x_n \to x$, if, for every $\epsilon > 0$, there exists N such that for all integers $n > N$, we have $|x_n - x| < \epsilon$.
- The order is important: Given an arbitrary $\epsilon > 0$, we must find an N, possibly depending on ϵ, such that the desired property holds for every $n > N$
- This is the definition for a sequence of scalars; we will discuss convergence of vectors next week

Proof

Let us prove that $1/n \to 0$.

1. Let $\epsilon > 0$.
2. Choose N such that $N > 1/\epsilon$.
3. Let $n > N$. Then

$$\begin{aligned} \left|\frac{1}{n} - 0\right| &= \frac{1}{n} \\ &< \frac{1}{N} && \text{because } n > N \\ &< \epsilon && \text{because } N > 1/\epsilon. \end{aligned}$$

- The inequalities above use the fact that for positive numbers, $a < b \implies a^{-1} > b^{-1}$

Order and dependence

- The steps of a proof must form a sequential chain of reasoning — each statement must follow from assumptions, definitions, or results established earlier in the proof
- For example, the following order would not be valid:
    1. Let $N > 1/\epsilon$.
    2. Let $\epsilon > 0$.
- The first line uses ϵ before it has been defined — if you were writing code, your program would crash here
- More fundamentally, this order misses an essential feature of convergence: there is no single N that works for every ϵ. The value of N depends on ϵ. We must *first* take an arbitrary $\epsilon > 0$ and *then* choose an appropriate N for that ϵ.

Arbitrary quantities: “for every”

- The statement “Let $\epsilon > 0$” puts the idea “for every $\epsilon > 0$” into practice
  - We fix ϵ, but we assume nothing about it except that it is positive
  - Because the argument works without making any additional assumptions about ϵ, it works for every positive ϵ
- The phrase “for every” also applies to n. Once N has been chosen, the inequality must hold for every sequence element after N:

$$x_1, x_2, \ldots, x_{N-1}, x_N, \color{red}{x_{N+1}, x_{N+2}, x_{N+3}, \ldots}.$$

It is not enough to verify the inequality for one particular value of n

Chosen quantities: "there exists"

- The role of N is different
- The definition says that *there exists* an N with the required property
- We therefore do not need the result to hold for every possible N — we only need to identify one choice that works
- Such a choice is usually not unique: here we picked $N > 1/\epsilon$, but we could have picked $N > 2/\epsilon$ or $N = 10 + 1/\epsilon$ or lots of other things
- The distinction between "for every" ($\forall$) and "there exists" ($\exists$) is fundamental in analysis and statistical theory

Defining a quantity fixes its value

- When you define a symbol, this fixes its value
- This is obvious if you say something like "let $x = 0.5$", but it also applies to "let $\epsilon > 0$"
- In the latter case, we don't know exactly what numerical value ϵ has, but it *remains fixed* at that value for the rest of the proof
- For similar reasons, you should not redefine a symbol partway through a proof — if you need two different ϵ values, use ϵ_1 and ϵ_2

Discovery versus presentation

- Students often ask: how did you know to choose $N > 1/\epsilon$?
- Usually, we *don't* know that in advance.
- When you're figuring out a proof, you might work backward from the desired conclusion, try several possibilities, and revise unsuccessful attempts
- This exploratory reasoning is part of finding an approach that works, but the final proof should present the resulting argument in a direct logical order — it's a mathematical argument, not a journal of self-discovery

Chains of equalities and inequalities

- Finally, when several mathematical statements form a sequence, it is usually clearest to display them as a single chain: $a < b = c < d = e$
- This makes the conclusion $a < e$ immediately visible and shows how each step is connected to the next.
- The same information is huch harder to follow if presented as an unordered collection: $a < b, d > c, e = d, b = c$
- Chains are especially helpful for inequalities: keeping all inequality signs oriented in the same direction makes the logic easier to follow and provides a useful check that your argument is sensible