Talking and Writing about Mathematics

Charles J. Geyer

School of Statistics
Minnesota Center for Philosophy of Science
University of Minnesota

September 8, 2022

Slides for this talk:
http://users.stat.umn.edu/~geyer/math.pdf

Me

I have been thinking about writing and mathematics (including statistics) for more than 40 years.

I have been thinking about talks about mathematics (including statistics) for more than 30 years.

I started off the one time I taught Stat 8913 (literature seminar) with a talk about how to give talks.

That talk (https://www.stat.umn.edu/geyer/8913/talks.pdf) had a little bit of advice specifically about mathematics. But way too little. Hence this talk.

8913 Talk about Talks

The points about math (at least vaguely) from that talk were

- The slides should have nothing that doesn't help the audience understand the main point (or points).
- Your talk should *not* look like . . . textbooks and lectures in classes . . . *way* too detailed.
- Never define notation until just before you use it.
- Use the least possible mathematical notation.

But most of the talk was about other issues.

Moreover — as the 8913 students pointed out — these generalities don't say *how to do it*.

Great Books About Writing

Simple & Direct by Jacques Barzun

On Writing Well by William Zinsser

The Sense of Style by Steven Pinker

The essay by Paul Halmos in How to Write Mathematics

Great Books About Writing (cont.)

All of these books say good writing is not easy. You need lots of thought and lots of practice, which takes years.

So you will not be able to instantly apply anything I say in this talk.

But you will eventually if you keep working at it.

Math is Hard

Math is Hard.

Writing about math is even harder.

Math can be correct but still "almost maximally confusing" (Halmos's description of Birkhoff's proof of the ergodic theorem, the first proof published).

Often it takes a lot of reworking of a subject, perhaps by many authors, before a clear version of a subject emerges.

The same will happen to you. It will take a lot of rewriting to get any of your mathematics clear.

Or any writing, mathematical or not.

Math is Hard: Disclaimer

A criticism of my practice talk from my sister (Professor, Department of Ecology, Evolution, and Behavior, University of Minnesota, member of the National Academy of Sciences) was that some students — the perfectionists — will take this as even more pressure and even more reason not to be able to do anything unless it is perfect (which no writing ever is).

So I am *not* saying that.

Her husband said that his father (an English professor) always said the only way to learn how to write is to write.

And my father (a business executive with a PhD in chemistry) always said first get something down on paper. Just thinking about it goes nowhere.

Less is More

Less is more.

Don't fuss. Except all mathematics is fuss, so don't fuss any more than is absolutely necessary.

You need symbols to write math, but most writers overdo it. Use fewer symbols, fewer formulas, shorter formulas.

Less is More (cont.)

The formula

$$\theta = a + M\beta$$

is not helped by

$$\theta \in \mathbb{R}^n$$
$$a \in \mathbb{R}^n$$
$$M \in \mathbb{R}^{n \times p}$$
$$\beta \in \mathbb{R}^p$$

It is enough to say that M is a matrix — or perhaps a linear operator — and the other things are vectors.

That the equation is supposed to make sense (dimensions are OK) usually goes without saying.

Zinsser Brackets

From the chapter titled “Clutter” in *On Writing Well*

> *Is there any way to recognize clutter at a glance? Here's a device my students at Yale found helpful. I would put brackets around any component in a piece of writing that wasn't doing useful work. Often just one word got bracketed . . . . Often my brackets surrounded the little qualifiers that weaken any sentence they inhabit . . . . Sometimes my brackets surrounded an entire sentence . . . . Most first drafts can be cut by 50 percent. They are swollen with words and phrases that do no new work.*

The same goes for symbols and formulas that do no new work — or no work not absolutely necessary.

Zinsser Brackets (cont.)

Geyer and Thompson (1992) was accepted as a JRSS-B discussion paper but the editor for discussion papers asked us to cut the length.

So we painfully and laboriously applied Zinsser's method. For each word, phrase, sentence, paragraph, symbol, or formula, if it was not pulling its weight, then out it went.

The length was cut 25% with no change in content.

The writing was clearer, stronger, better.

But it was hard work, and work that we wouldn't have bothered with unless asked.

Zinsser Brackets (cont.)

Still Geyer and Thompson (1992) is not good writing.

It lacks focus. It covers too many things that could have gone in separate papers. But I and my thesis advisor did not yet see the big picture yet on many of the topics in this paper.

Later papers

- Geyer (1992)
- Geyer (1994, JRSS-B)
- Geyer (1994, Annals)
- Geyer (2009)

are much clearer but required a lot more work and thought.

Motivation Before Notation

If chapter one is “preliminaries” or first slide is “establishing notation”, then you know it’s bad.

Don’t bury the lede.

Your first job is to interest readers in your material. If you don’t do that, you are *boring!*

95% of statistics papers and talks *are* boring. You can do better if you try.

Motivation Before Notation (cont.)

In a talk, this is especially important.

No one will memorize your “established notation” on the fly.

They won’t remember the meaning when you use it two slides later.

Introduce notation as it is needed.

And most of it isn’t needed.

Replace Formulas by Concepts Expressed in Words

My thesis (1990) is partly about exponential families of distributions.

It defines them with lots of formulas and an incredible amount of unnecessary measure-theoretic nonsense. And goes on for several pages. In this it more or less follows the authoritative literature (Barndorff-Nielsen, 1978; Brown, 1986).

Geyer (2009) and later writings of mine are much simpler. A statistical model is an exponential family if it has log likelihood of the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta)$$

where y is a vector statistic, θ a vector parameter, and $\langle \cdot , \cdot \rangle$ is a bilinear form placing vector spaces in duality.

Replace Formulas by Concepts Expressed in Words (cont.)

It can go without saying that c is a real-valued function on the parameter space.

It can go without saying that the parameter space is a nonempty subset of some vector space.

Usually I say that we can take $\langle \cdot , \cdot \rangle$ to be defined by

$$\langle y, \theta \rangle = \sum_i y_i \theta_i$$

But that is unnecessary additional fuss. Vector spaces and their duals should be concepts known to anyone who has had linear algebra.

Replace Formulas by Concepts Expressed in Words (cont.)

No explicit measure theory. All you need to know to understand the definition is what a log likelihood is.

This defines the *same mathematical concept* as all the measure theoretic woof.

If you know how to define *log likelihood* measure-theoretically, then you know how to prove we haven't changed the concept.

Replace Formulas by Concepts Expressed in Words (cont.)

How did this characterization of exponential families evolve?

I noticed that most of the reasoning in proofs about exponential families *started with the log likelihood*. Hence we could make that the star of the show.

But you have to be looking for this sort of thing. If you don't look, then you won't see.

Replace Formulas by Concepts Expressed in Words (cont.)

My Stat 5421 notes on exponential families contains the following:

Theorem

Suppose we are working with a regular full exponential family. If the canonical parameterization is identifiable, any algorithm that always goes uphill if it can on the log likelihood defined by [the formula four slides back] converges to the unique MLE if the observed value of the canonical statistic vector y is a possible value of the mean value parameter vector.

A *very important theorem* and *no formulas* — but many concepts.

There aren't any formulas in the proof either because it uses sledgehammer theorems from convex analysis.

The Math is the Hero, not You

Do you say “I propose” and “I prove” or “We propose” and “We prove”?

Neither!

The math is the star of the show, the hero of the story. We don’t need you to appear at all.

The Math is the Hero, not You (cont.)

This is a Zinsser point.

He once wrote a *New Yorker* series about two jazz musicians.

He thought his first submission was some of his best work.

He thought he really captured how they performed and his own reactions.

The *New Yorker* editor told him to cut all of his own reactions.

He was really loath to do that, but when he did he realized that the article got much better. He wasn't getting in the way of the reader's reactions to the musicians.

The Math is the Hero, not You (cont.)

So instead of trying to impress us, let the material speak for itself.

Even when the characters in the story are mathematical objects rather than jazz musicians.

Try to make your math come alive.

The Curse of Knowledge

> *Monads . . . come with a curse. The monadic curse is that, once someone learns what monads are and how to use them, they lose the ability to explain it to other people.*
>
> — *Douglas Crockford*

Pinker in *The Sense of Style* says this phenomenon is ubiquitous.

When you learn anything, you forget what you were like before you learned it. And you cannot imagine what it is like not to know it.

Thus you don't even see a need to explain it. You know it. Doesn't everybody?

The Curse of Knowledge (cont.)

This is one of the main reasons for bad teaching.

It is why my first year teaching evaluations were so bad that I had Joe Eaton and Glen Meeden assigned to mentor me until I got at least some rudimentary idea of what the students *didn't know* (and I had to teach).

Many teachers suffer from the same. You are not there to show off.

You are there to bring the students from where they are to somewhere more advanced.

The Curse of Knowledge (cont.)

Everybody is affected by the curse of knowledge.

If you are aware of it, then you can fight it.

You won't be entirely successful, but you can do better.

The Curse of Knowledge (cont.)

This is one of the main reasons for bad talks.

In the median statistics talk, no one in the audience really understands what the speaker is on about.

Why would anyone *want* to do that?

It's the curse of knowledge. They are blissfully unaware of it.

The Curse of Knowledge (cont.)

This is one of the main reasons for bad writing.

Pinker says the only real cure is to get comments from readers and actually fix the problems they complain about.

And he does not mean comments from referees.

He means friendly readers who are asked to be harsh about the *writing* (not mathematical correctness) — who are asked to complain about everything unclear.

Similarly for talks. Practice talks with harsh critics are the only remedy.

Care at the Beginning

Sung and Geyer (2007) is a typical *Annals* paper.

The proofs are hard to read because they skip many details.

I keep the first submitted version of that paper on-line for no reason other than it has clear proofs that are much longer than the published proofs.

This is the only *Annals* paper I have been on that had no mistakes in the first submission. This is because we wrote it with very clear proofs.

The only comment we got from the referees was that the proofs were way too long. Hence we complied with the referees request (and usual practice) and made the proofs unreadable.

Care at the Beginning (cont.)

There is a lesson there.

Do not write papers like those you see published. First get them correct. Then get them accepted. Then make them less readable if the editors and referees insist.

The Big Picture

Understanding the big picture — where the work you are explaining lies in the landscape of all statistics — is the key to good explanation. It can be used to organize your work.

The book *Finite-Dimensional Vector Spaces* by Paul Halmos is a great example.

It is a linear algebra book written as if it were specializing functional analysis to the finite-dimensional case. This organizes his whole presentation.

He emphasizes the concepts and uses the proofs that generalize to the infinite-dimensional case.

Readers who don't read prefaces don't even notice. But even if they don't, they do get the benefit of what is generally recognized to be the clearest presentation of the subject ever.

The Big Picture (cont.)

Markov chain theory organizes the subject of Markov chain Monte Carlo.

But it wasn't always that way.

We had a more than a decade of confusion about what the foundations of the subject were before they were clarified.

Even the math was not available until the late 1980's and early 1990's and was not widely understood by (some) statisticians until years later.

Invented in 1953, understood a bit before 2000.

The Big Picture (cont.)

The *right* concepts organize subjects.

Sometimes it can take decades or centuries of work by many authors to find the right concepts.

In new areas, the right concepts probably have not yet been found.

Use these concepts when you can. Look for better concepts.

You know you have found a good concept when it simplifies theorems and explanations.

It's Not All Equally Important

In bad mathematical writing, the reader is given no guidance.

It's all important. Pay the closest possible attention to *everything*.

Organization? We don't need no stinking organization!.

In good mathematical writing, the author thinks carefully about what is more important and what is less important and carefully guides the reader.