

A table of narratives and distributions

Ben Klemens

9 December 2014

Remember your first probability class? At some point, you covered how, if you take sets of independent and identically distributed (iid) draws from a pool of items, it can be proven that the means of those draws will be Normally distributed.

But if you're like the great majority of the college-educated population, you never took a probability class. You took a statistics class, where the first few weeks covered probability, and the next few weeks covered other, much more structured models. For example, the derivation of the linear regression formula is based on assumptions regarding an affine relation between variables and the minimization of an objective that is sensible, but just as sensible as many other possible objectives.

Another way to put it is that the Normal Distribution assumptions are bottom-up, with statements about each draw that lead to an overall shape, while the linear regression is top-down, assuming an overall shape and deriving item-level information like the individual error terms from the global shape.

There are lots of bottom-up models with sound microfoundations, each a story of the form *if each observation experienced this process, then it can be proven that you will observe a distribution like this*: Polya urns, Poisson processes, orthogonal combinations of the above. In fact, I'm making a list¹.

Maybe you read the many posts on this blog [this post (entry #148), et seq] about writing and using functions to transform well-defined models into other well-defined models. A chain of such transformations can lead to an increasingly nuanced description of a certain situation. But you have to start the chain somewhere, and so I started compiling this list.

I've been kicking around the idea of teaching a probability-focused stats class (a colleague who runs a department floated the idea), and the list of narrative/distribution pairs linked above would be the core of the first week or two. You may have some ideas of where you'd take it from here; me, I'd probably have the students code up some examples to confirm that convergence to the named distribution occurs, which leads to discussion of fitting data to closed-form distributions and testing claims about the parameters; and then start building more complex models from these basic models, which would lead to more theoretical issues like decomposing joint distributions into conditional parts, and estimation issues like Markov Chains. Every model along the way would have a plausible micro-story underlying it.

¹<http://b-k.github.io/narratives-distributions>

This post is mostly just to let you know that the list of narrative/distribution pairs² mentioned above exists for your reference. But it's also a call for your contributions, if your favorite field of modeling includes distributions I haven't yet mentioned, or if you otherwise see the utility in expanding the text further.

I've tried to make it easy to add new narrative/distribution pairs. The project is hosted on GitHub³, which makes collaboration pretty easy. You don't really have to know a lot about git (it's been on my to-do list to post the git chapter of *21st Century C* on here, but I'm lazy). If you have a GitHub account and fork a copy of the repository underlying the narrative/distribution list, you can edit it in your web browser; just look for the pencil icon.

Technical The formatting looks stellar on paper and in the web browser, if I may say so, including getting the math and the citations right. There isn't a common language that targets both screen and text for this sort of application, so I invented one, intended to be as simple as possible:

```
Items(  
• Write section headers like Section(Title here)  
• Write emphasized text like em(something important)  
• Write citation tags like Citep(fay:herriot)  
• and itemized lists like this.  
)
```

See the tech guide⁴ associated with the project for the full overview.

[Pandoc didn't work OK for me, and Markdown gets really difficult when you have technical documents. When things like ++ and * are syntactically relevant, mentioning C++ will throw everything off, and the formatting of $y = a * b * c$ will be all but a crash.

On the back end, for those who are interested, these formatting functions are m4 macros that expand to LaTeX or HTML as needed. I wrote the first draft of the macros for this very blog, around June 2013, when I got tired of all the workarounds I used to get LaTeX2HTML to behave, and started entrusting my math rendering to MathJax⁵. The makefiles prep the documents and send them to LaTeX and BibTeX for processing, which means that you'll need to clone the repository to a box with make and LaTeX installed to compile the PDF and HTML.

But the internals are no matter. This is a document that could, with further contributions from you, become a very useful reference for somebody working with probability models—and not just students, because, let's all admit it, working practitioners don't remember all of these models. It is implemented using a simple back-end that could be cloned off and used for generating collaborative technical documents of equal (or better!) quality in any subject.

²<http://b-k.github.io/narratives-distributions>

³<https://github.com/b-k/narratives-distributions>

⁴https://github.com/b-k/narratives-distributions/blob/master/tech/tech_guide

⁵<http://mathjax.org>