How (not) to introduce newcomers to Bayesian analysis

Robert L Grant, BayesCamp Ltd, 2018 ©

Thanks to Rasmus Bååth for some comments and encouragement.

When I decided to start my own business doing training in Bayesian methods, I read a lot of other people's introductions to the subject. I wanted to see how others approached the subject, and I wanted to steal the best ideas. I looked at books, videos, websites, blogs… and I'm still going, because they keep coming out and some are buried away in obscure places. Although there are some absolutely outstanding exemplars, the very beginning never quite satisfies me. I mean, the way that the idea of Bayesian statistics is introduced to the reader / listener / whatever.

In this post, I'll set out what I like and don't like about introductions to Bayes, and I'll explain how I do it as I go along.

First, I have to be clear about my intended audience; teaching a room of doctors would be different to a room of maths grad students. Not necessarily better or worse, easier or harder, just different. I aim at people who think about problems quantitatively, but not mathematicians or statisticians. I want to help everyone else who falls between the cracks. They might be healthcare professionals, marketing analysts or machine learning folk who want to get stronger at stats.

So, for starters, my introduction is not very mathematical. It's not that I don't appreciate the importance of mathematical ability if you want to be a theoretical statistician, it's just that my audience don't intend to become theoretical statisticians. Plenty of visual aids help here, and I think that the flipchart or whiteboard is a much more useful tool than slides, because it is interactive and allows students to come up and try out things (for instance, after a small group activity). I like to prepare several pages on the flipchart ahead of time so we can just skip through from one concept to another while it's fresh in their minds without them being distracted, thinking stuff like, "I wonder if that pen is going to hold out… the ink is looking thin." An example of this is showing a regression line in variable space (X on the horizontal axis and Y on the vertical), then flipping the page to parameter space (beta1 on the horizontal and beta0 on the vertical) to show it as a point.

Later, when we get into software, there is huge value in demonstrating how to code something up and look at the results via a projector. A Jupyter notebook is a great way of doing this because you can quickly go back and tweak something and see its effect, although I feel uneasy about getting my learners to spend time gaining familiarity with a tool that few of them will use in earnest. It's not considered cool, but I still think WinBUGS is a neat way of walking through reading in the data, checking the model, running a few hundred warmup iterations, then going for it. Of course, this is another tool that learners probably won't use in

years to come, but there's no reason why you can't do that same sequence of steps in R+Stan, for example.

I don't like:

Bayes theorem at the outset. Why do people do this? It shows us two things, the historical connection (who cares?) and the principle of reversing conditional probability by multiplying likelihood and prior. But that's not what we do in practice, we simulate, so why not show students that. They can also work out quickly that the formula you would really use is not that simple; there's also a normalising constant and a denominator that has to be integrated. It also confuses them that this theorem can be used for reversing-probability reasons that are not "Bayesian", like the classic examples with medical test results. I prefer to introduce the value and flexibility of multiplying conditional probabilities together with a practical example and an explanation more like a particle filter (although I wouldn't use that term), because that's closer to the simulation that follows.

Philosophical distinctions about the meaning of probability or randomness (but see below). This is important for clever students but even then will only interest them once they are getting comfortable with the analysis. We should make our learners good at applying the methods first, then they can reflect on theory and finally history.

History. Nobody gives a monkey's whether Jaynes and Keynes were the inspiration for a Gilbert and Sullivan operetta, or Peirce abducted Neyman's cat.

Contrived examples. Oh, I like old, well-worn datasets, like irises or the Titanic (more on that another time), but tossing a coin ten times? Come on. (As a self-defensive footnote, I have used coin-tossing with success when introducing the concept of hypothesis testing, but that's a very different goal and hence a different metaphor for the mental processes and quantification at work. I got that coin-tossing exercise from Beth Chance and Nathan Tintle at ICOTS9, and I think it rocks.)

A lot of maths: theorem-proof-lemma format for example, or matrix algebra when the learners would get the idea faster from talking about what happens to one observation, one parameter, one iteration at a time. Mathematicians get a habit to set out the most general possible exposition at the beginning, in the most general possible terms. But you can't fully grasp it until later, when the fine details have sunk in. I think it's better to have carefully chosen examples that illustrate one principle at a time, then gradually accumulate them. Debugging where someone lost the thread is much simpler. And we don't need to see the proof unless we are studying how to prove similar things in future. You, the teacher, need to do the maths, but keep it out of sight.

Analytical solutions and conjugacy. Yawn. It's not the 50s. Don't waste your learners' time.

1-d density plots of prior, likelihood and posterior that hardly overlap. You all know the sort. They are mathematically correct but unrealistic. That prior is a BAD prior, and your students can see that. The likelihood is out in low-prior region and that should give you pause for thought in real life. Don't drag your students out of real life critical thinking and into some abstract ritual!

I do like:

Simulation at the outset, shortcut formulas later (à la GAISE). Note that calling asymptotics "shortcut formulas" gives students the right attitude to conceptualising their analyses in a grounded and critical way; it's not intended to disparage the value of solid statistical theory.

Prior and posterior predictive checking, where you use your model before it sees any data, and after it has "learnt" from the data, to generate new phony data. Take a look at the phony data and see if they look anything like the real ones. Where the prior does not include the data, you've got problems. Likewise, when the posterior does n't look like the data in some way. These are intuitive ways of doing an open-ended check on your model.

A focus on computation, even if it's not specific to Bayes (floating point accuracy, digital rounding error, or setting RNG seeds are all good examples).

Approximate Bayesian Computation (ABC) as an inroads to thinking about: (1) simulation as a way of combining probability densities and/or likelihoods, (2) letting the computer try different values of the parameter and seeing how it matches the data, and (3) the need for a sensible prior to guide the computer away from no-hope regions as well as problems like no overlap in logistic regression. However, this will work well for people who have already thought about random number generators, less well for those who haven't. Because we are going to simulate, we need to introduce RNGs and distributions anyway, and throwing ABC on top of that might just be too much. It all depends on the audience.

Tilde notation, like x ~ norm(10,3). Rasmus Bååth and I both like this for introductory teaching and neither of us know what to call it. Let's go with "tilde probability notation". It's much easier to write once you know a few common distributions, and you can pile them up thus:

```
mu ~ unif(0.5, 10)
y[i] ~ poisson(mu)
```

That's a little univariate Poisson model. But it easily extends into models that are quite painful to read in algebra.

```
mu ~ unif(0.5, 10)
sigma ~ unif(0.1, 5)
y[i] ~ poisson(mu)
x_mean[i] = ((y[i] > 5) * 5) + ((y[i] <= 5) * y[i])
x_measure_error[i] = ((y[i] > 5) * 0.001) + ((y[i] <= 5) * sigma)
x[i] ~ round(norm(x_mean[i], x_measure_error[i]))
```

That's a model for Poisson-distributed data that are censored and heaped at 5 and also have some measurement error below that point. Pretty advanced, pretty quickly. Also, once you get familiarity with this, you can use it in Stan or BUGS or JAGS.

Emphasising the communication advantage, for example, "our analysis shows that there's an 81% chance that return on investment will be over $1m within 5 years" (check out Frank Harrell's blog for some medical examples). Who doesn't love that? Perverts, that's who. Or ultra-frequentists, though I'm not sure they exist any more. So that just leaves perverts.

Emphasising flexibility — we can get beyond simple models quickly and without a lot of jiggery-pokery like frequentists, who have to juggle with REML and E-M and profile likelihoods and goodness knows what all, just to avoid sins of the tongue.

Grounding everything we demonstrate in real-life research needs, like the communication and flexibility above.

Showing data space and parameter space and flipping between them

Getting quickly into models that are complex enough to be of real-life value. Students know when they are being shown some dumbed-down stuff that they could never use in vocational settings. This is a challenge of course, but you're not being paid to dodge challenges.

I honestly don't know what to think about:

Emphasising ways of thinking about data, models, truth, etc (à la McElreath). It's extremely important, to be sure, but I don't know if it helps to hear it early on. I am too immersed in the subject to be able to judge.

Bayes as the only true approach to probability for adherents of religions who contend that everything is predetermined or decided by supernatural force (this would include most Muslims and Calvinists). In essence, if mortal humans know nothing for certain, including the immutability of parameters or the infinite replicability of your experiments, then it follows that data, latent variables, parameters and hyperparameters all move in ways you cannot fully understand, and so are subject to the same mathematics of probability. Is this a helpful assertion? Or not. I tend to think it best not to get involved in such matters, especially as I don't believe it.

Emphasising networks early on, like David Barber's (otherwise great) book does. I suppose some people work with those models and that decision-theoretic application, and need it. I just don't know how it fits into everyone else's learning curve.

Just what is it that you want to do?

Now that I am free to design my own educational products in Bayes, I find that actually, it always has to be tailored to the audience, unless it's a quick overview. So, I set myself up to provide face-to-face training and coaching. I might have the odd quick overview as an online course, but the really in-depth stuff has to involve discussion, reflection and interaction, not just with me but with the other learners too. Of course, that means it's not fair on people in far-flung places, but I can't reach everyone.

I do training (a group of learners with clear learning outcomes at the outset) and coaching (one person and me, talking about their career and goals, where I mostly ask questions and there are no outcomes at the outset).

I think any training session should avoid getting bogged down in long sessions of chalk 'n' talk, but inevitably there has to be some of that. So, I keep them to 30 minutes long if I can, and alternate them with some small group activity. You could have several small group activities over the course of the day. I think it's a good idea to put the groups together so that they are diverse, and that requires finding out a little about learners' work experience, qualifications, etc before we begin. That mimics the diverse composition of data science teams in this day and age, and I tell people that from the outset so that they know they should respect and listen in the group to get the most out of the learning experience.

What about this maths avoidance? It strikes me as odd that there are many introductory textbooks and courses for statistics that play down the maths, on the basis that the learners are going to operate a computer and construct a model in code, not by matrix algebra and calculus. Yet, this doesn't happen for Bayesian statistics. This is perhaps down to two historical factors. Firstly, Bayes was an advanced subject so only people who already had degrees in statistics or mathematics would encounter it. Secondly, Bayesians spent decades being mocked and sidelined, and responded by foregrounding their mathematical rigour in the hope of beating their critics. But nowadays, we want all sorts of people who analyse data to think about using Bayes from the beginning of their careers, so we should offer them the same option. If you want the maths, there are plenty of options for you, but I would like to offer something a little different, a little more inclusive.