‘Educating your beliefs’ versus ‘Testing your Hypotheses’

This is the handout for our session about Bayesian issues.— B. Mazur

1. ‘Bayesian intertwining’

Here are a few words about the Bayesian point of view. But first a disclaimer: I know very little statistics; I'm a total outsider to this field[1] and especially to the extended conversation—and the somewhat sharp disagreements—that Bayesians and Frequentists have.

But I do want to make sense of the issue for myself, insofar as it reflects on our understanding of the interplay between *hypothesis, data, conclusions*, those basic building blocks of empirical investigation. Now I have just listed those “building blocks” in the standard (expected) order of procedure of an empirical investigation.

Set-up and Hypotheses ⟶ Data Collecting ⟶ Processing Data and Conclusion

The Bayesian viewpoint will methodically intertwine the first two steps.

In our discussion last week, we saw a hint of another kind of intertwining possible in the interpretive work of Art History. One might imagine that the standard order of procedure in front of a work of art—say the Arnolfini double portrait—is to gather evidence of all sorts about it, and then to use this mass of evidence to develop an interpretation or viewpoint of it or perhaps a deeper emotional connection with it. But the suggestion was made that, at times, it is (at least) *possible*—and perhaps necessary–to read back from the arrived-at interpretation to its supporting evidence; in that the depth of the interpretation—or lack of depth—might lead us to reassess the value of various pieces of evidence[2].

Beyond *Bayesian intertwining* there are other essential contrasts between the Bayesian's and the Frequentist's work; their methods are not the same, and they even seem to have slightly different *primary goals*. We'll get to that, eventually.

[1] I learned about the Bayesian point of view largely from a phone conversation this week with Susan Holmes, who is a statistician at Stanford.

[2] The suggestion was that we have something akin to the Dworkinian moral normative attitude towards law, in that one might value more the kind of evidence that leads to the best (deepest) interpretation of a work of art.

2. Prior information and the Birthday problem

To introduce ourselves to this 'Bayesian intertwining' (taking as a **black box**—at least at first—some of the mathematical procedures involved) let's revisit a famous problem: the birthday problem. You have a class of fifth graders in an elementary school. Suppose there are 23 students in the class. What is the probability that two of them have the same birthday? Or, to seem more mathematical, suppose there are n students. What is the answer as a function of n?

Here is the simple naive analysis of this problem. We assume, of course, that the probability of anyone having a birthday at any specific day, e.g., April 22, is $1/365$ (ignoring the leap year issue). Think of the teacher marking off—successively— on a calendar the birthdays of each student. We are going to gauge the possibility that in his class of n students there are no two birthdays on the same calendar day. The first student's birthday is duly marked. We can't possibly have a concurrence of birthdays (call it a *hit*) at this point, there being only one mark. So we can record "1" as the probability that we didn't get a *hit* at least so far[3].

As for the second student, the probability of him or her not having a birthday on the same day as student #1—i.e., that there not be a *hit*— is

$$1 - \frac{1}{365} \quad = \quad \frac{364}{365}.$$

Given this situation, and passing to the third student, in order for there not to be a hit, his or her birthday has to avoid two days, so that probability is

$$1 - \frac{2}{365} \quad = \quad \frac{363}{365}.$$

Putting the two probabilities together we get that–so far in our count—the probability that there isn't a hit with these three students is

$$(1 - \frac{1}{365})(1 - \frac{2}{365}) \quad = \quad (\frac{364}{365}) \cdot (\frac{363}{365}).$$

Working up (by mathematical induction) the probability that there's no hit, with n students is then:

$$(1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{n-1}{365}),$$

[3] We are going to write probabilities as numbers between 0 and 1. So if the probability of an event is $\frac{1}{2}$ that's the same as saying that it is *even odds of it happening or not happening* or that 50% *of the time it happens*, or one sometimes simply says that there's is a 50/50 chance of it occurring.

which when $n = 23$ is close to $\frac{1}{2}$. That is, for a class of 23 students the chances are 50/50 that there's a concurrence of birthdays—given this analysis.

My Bayesian friend Susan Holmes tells me that she has actually tried this out a number of times in real live classes, and discovered that the odds seem to be much better than 50/50 for 23 students; you even seem to get 50/50 with classes of as low as 16 students.

There is something too naive in the analysis above, says Susan. We should, at least, make the following (initial) correction to our setting-up of the problem. We said above:

> We assume, of course, that the probability of anyone having a birthday at any specific day, e.g., April 22, is 1/365

We actually *know* stuff about the structure of our problem that we haven't really registered in making that assumption.

For example, it is a class of fifth-graders so, chances are, they were all (or mostly) born in the same year. In particular, the years of their birth all (or mostly) had the same weekends and weekdays. In the era of possible c-sections and induced births—given that doctors and hospital staff would prefer to work on weekdays rather than weekends—one might imagine that the probability of being born on a weekday is somewhat skewed. We also know more that might make us think that fixing 1/365 at the rate is too naive.

Perhaps then, instead of sticking to the probability $p = 1/365$ per day hypothesis, allow a bit of freedom and a priori allow that there are different probabilities

$$p_1, p_2, p_3, \cdots, p_{365}$$

for each day of the year[4], about which we can make very very rough guesses. But let us not write this in stone yet. Make a mildly educated guess of these p_i; e.g., if "i" is a weekend, then p_i is slightly less than 1/365; if a weekday, slightly more. This initial guess we'll call a **Prior**. From the prior we can deduce—essentially as we did above with the "1/365"—all the expected odds and whatever statistics one wants. But we have hardly gotten our best answer!

For, we now consider whatever **Data** we've actually accumulated by sampling classes of sixth-graders. What the Bayesians have developed is a *somewhat uniform* way of actually it using the accumulated data to "correct" the prior, so that it is more in tune with the data. This corrected-prior is called (naturally) the **Posterior**.

The movement here is as follows:

$$\text{Prior} \quad \xrightarrow{\textbf{Data}} \quad \text{Posterior.}$$

In this protocol, the **Data** enjoys a slight augmentation of its principal role: no matter what else it will be used for, it will be used to educate the Prior!

[4] these summing to 1

Starting anew with this **Posterior** as if it were a more educated-guess than the original **Prior** we can deduce—essentially as we did above with the “constant 1/365” or the prior **Prior** —all the expected odds and whatever statistics one wants.

This is a preliminary move in the Bayesian direction, but we aren't quite there yet. Another–and better–way of viewing this move (reflecting our most up-to-date version of belief about the set-up) is that the initial values

$$p_1, p_2, p_3, \cdots, p_{365}$$

should not be taken as hard unchangeable numbers but rather are to be viewed as “random variables” in their own right, and subject to their own distribution, which we are bent on determining, given enough **Data**. The grand function of the data is to educate the prior.

The **black box**—so far—is that I have not yet said anything about the mathematical procedure Bayesians use to feed back (as an afterburner) information obtained by the Data into the prior assumptions, in order to effect the “education” of these prior assumptions and thereby produce the Posterior. For the moment—in this discussion—it is more important for me simply to say that *whatever this procedure is* it is, in fact, a *predetermined procedure.*

3. Predesignation versus the self-corrective nature of inductive reasoning

Now you might well worry that this Bayesian ploy is like curve-fitting various hypotheses[5] to the data—a kind of hypothesis-fishing expedition, if you want. You keep changing the entire format of the problem, based on accumulating data. The Bayesians have, as I understand it, a claim: that any two 'reasonable' priors, when “corrected” by enough data will give very close posteriors. That is, the initial rough-hewn nature of the prior will iron out with enough data. Their motto:

> *Enough data swamps the prior.*

I've been playing around with another formulation of that motto:

> *Any data-set is, in fact, a ‘data point’ giving us information about the probability distribution of priors.*

In contrast, there is a motto that captures the sentiment of a Frequentist:

[5] I want to use the word *hypothesis* loosely, for the moment; that is, the way we generally use the word; and not in the specific manner that statisticians use it.

> *Fix hypotheses. This determines a probability distribution to be expected in the data. Compute data. If your hypotheses are good,* **in the limit** *the data should conform to that probability distribution.*

About the above, one of the early great theorizers in this subject (and specifically regarding probability, randomness, and the law of large numbers) was Jacob Bernoulli. He *also* was a theologian preaching a specifically Swiss version of Calvinism. You see the problem here! There is a strict vein of *predetermined* destiny or fatalism in his theology, someone who is the father of the theory of randomness. How does he reconcile these two opposites? Elegantly, is the answer! He concludes[6] his treatise *Ars Conjectandi*, commenting on his law of large numbers, this way:

> Whence at last this remarkable result is seen to follow, that if the observations of all events were continued for the whole of eternity (with the probability finally transformed into perfect certainty) then everything in the world would be observed to happen in fixed ratios and with a constant law of alternation. Thus in even the most accidental and fortuitous we would be bound to acknowledge a certain quasi necessity and, so to speak, fatality. I do not know whether or not Plato already wished to assert this result in his dogma of the universal return of things to their former positions [apokatastasis], in which he predicted that after the unrolling of innumerable centuries everything would return to its original state.

Apokatastasis is a theological term, referring to a return to a state before the fall (of Adam and Eve)[7].

At this point we might connect the above discussion with C.S. Peirce's 1883 paper " A Theory of Probable Inference" as mentioned in the Len O'Neill reading. O'Neill points out the fundamental distinction that Peirce makes between *statistical deduction* and *statistical induction* the first being thought of as reasoning from an entire population to a sample, and the second being reasoning from sample to population. As O'Neill says: in the first it is a matter of long run frequency (i.e, the Frequentist's motto) whereas the second is related to a Peircean conception of *the self-corrective nature of inductive reasoning* (and this sounds like the Bayesian protocol).

[6] It is, in fact, the conclusion of the *posthumously* published treatise (1713) but it isn't clear to me whether or not he had meant to keep working on the manuscript.

[7] In the class discussion, Noah suggested that Calvinists might be perfectly at home with random processes leading to firm limiting fatalism, in that the fates of souls—in Calvinist dogma—are *randomly assigned* and not according to any of their virtues; i.e., to misquote someone else: "goodness had nothing to do with it."

Peirce dwells on the issue of *predesignation* in the Frequentist's context (i.e., you fix a model and then collect evidence for or against it; you don't start changing the model midstream in view of the incoming evidence). There is a curious type of *meta-predesignation* in the Bayesian context, in that the manner in which you change the model, given incoming evidence, is pre-designated. We'll take a brief look at this.

4. Priors as 'Meta-probabilities'

Suppose you are a cancer specialist studying a specific kind of cancer and want to know if there is a gender difference: do more men than women get this type of cancer? Or more women than men?

Now suppose I asked you (cancer specialist) to make some kind of guess—when considering groups of people that get this cancer—about the proportion of men-to-women that get it. You might tabulate this as a probability P that a random choice of person in this group is male. So P is a number between 0 and 1. You might actually give me a number if you are very confident, but more likely, for a spread of possible values of P, you'll give me an estimate of greater or lesser levels of confidence you have that this P is indeed the sought-for-probability. Taking the question I asked more systematically, you might interpret it as follows:

> As P ranges through all of its possible values, from 0 (no males get it) to 1 (only males get it) tell me (your guess of) the probability that P is the ratio $\frac{M}{M+W}$ where M is the number of men and W the number of women in the group? In effect, draw me a graph telling your probability-estimate for each of the P's in the range between 0 and 1.

Your initial guess, and initial graph, is the Prior (I privately call it the *meta-probability*). It *will* be educated by the data accumulating.

Let's imagine that you say "I have no idea! This probability P could–as far as I know–equally likely be any number between 0 and 1." If so, and if you had to draw a graph illustrating this noncommittal view, you'd draw the graph of a horizontal line over the interval $[0, 1]$. Or, you might have some reason to believe that P is close to $1/2$ but no really firm reason to believe this and you might have no idea whether gender differences enter at all. Then the graph describing your sense of the likelihood of the values of P would be humped symmetrically about $P = 1/2$. Or if you are essentially certain that it is $1/2$ you might draw it to be symmetrically spiked at $P = 1/2$.

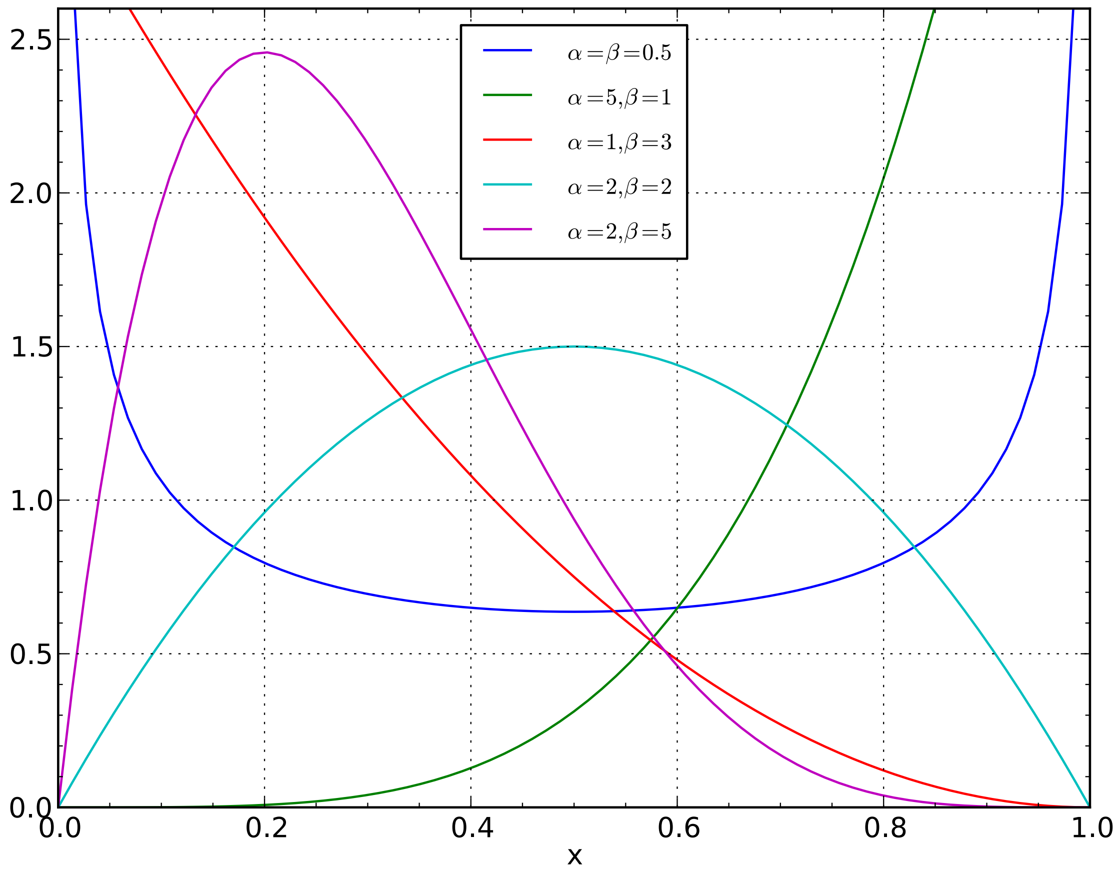
What you are drawing is–in a sense–a *meta-probability density* since you are giving a portrait of your sense of how probable you think each value between 0 and 1 might be the actual probability-that men-get-this-type-of-cancer. Your portrait is the graph of some probability density function $f(t)$.

There are theoretical reasons to suggest, for some such problems, that you would do well to be drawing the graphs of a specific well-known family called **beta-distributions**. These beta-distributions come as a two parameter family[8] $\beta_{a,b}(t)$. That is, fix any two positive numbers a, b (these numbers a, b are called the *shape parameters* of the beta-distribution) and you get such a graph.

Here are some general ground-rules for choosing these βs: shape parameters that are equal give distributions symmetric about $1/2$; i.e., you choose such a β if you expect that gender plays no role in the probability of contracting this cancer. Choosing $a > b$ means that you are skewing things to the left; i.e., you believe that men get this type of cancer less frequently than women; choosing $b > a$ means the reverse. The larger these parameters, the sharper the peak of the curve; i.e., the more "sure" you are that the probability occurs at the peak.

Choose parameters, say, $a = 2, b = 5$; or, say, $a = 2, b = 2$ and you have probability distributions $\beta_{2,5}(t)$, or $\beta_{2,2}(t)$, these being the blue and the magenta graphs in the figure below.

[8] These are distributions $t^{a-1}(1-t)^{b-1}dt$ normalized to have integral equal to 1 over the unit interval.



5. Back to our three steps

(1) (**Choosing the Prior**) Now, Bayesian cancer doctor that you are, when you start doing your statistics, choose a Prior. For this type of question you might do well, as I said, to choose some beta-distribution. If you imagine that there might be a gender bias here, but have no idea in which direction, you might choose one that is symmetric about $t = 1/2$ (which, as it turns out, means that you'd be taking shape parameters a equal to b). But size up the situation as best as you can, taking into account everything that you think is important to the problem and come up with a choice of a Prior. Let us say that your Prior is $\beta_{a,b}(t)$.

(2) (**The Data**) Suppose you now get a data sample of 100 people with cancer—perhaps the result of some specific study of some particular population, and suppose that 60 of these cancer victims are men (so 40 are women).

(3) (**Passing to the Posterior**) The beauty of the family of beta-distributions is that when you appropriately *educate* a beta-distribution (the Prior) with new data, the new distribution (the Posterior) is again a beta-distribution. The only thing is that the shape parameters may change; say, from (a, b) to a new pair of numbers (a', b'):

$$\beta_{a,b}(t) \quad \xrightarrow{\text{new data}} \quad \beta_{a',b'}(t)$$

I'm told that this change can be very easily computed. That is, in this example problem, the a', b' will depend on hardly more than the original a, b, the percentage of men with cancer, and the size of the study.

6. A numerical example and a question

For this example I'm normalizing things so the numbers work simply so we don't get bogged down in mere arithmetic. Imagine that your Prior is $\beta_{20,20}$ and you test a sample population (of just the right size for the normalizations to work out as I'm going to assume they do below) and in that population Men/ Women cancer ratio is 60/40. The Posterior is then (I'm told) $\beta_{20+60,20+40}$. And if you compute (based on that Posterior) the probability that men get this type of cancer more than women, that probability is:

$$0.955\ldots$$

If you did the analogous thing with the Prior $\beta_{10,10}$, getting, as Posterior, $\beta_{10+60,10+40}$ you'd compute (based on that Posterior) the probability that men get this type of cancer more than women to be:

$$0.966\ldots$$

Question: Why is it *reasonable* that the second estimate of probability of gender-difference be bigger than the first?