#Paradoxes of probability theory

← Back to Chapters

Comments.

Here is a mathematical take: Since no contradiction in mainstream mathematics has been fount yet, there are no actual paradoxes. There are two types of things that get called "paradoxes": 1) unintuitive results (a la Banach-Tarski theorem) and 2) fallacious arguments, resulting in incorrect conclusions.

The "paradoxes" of the first type are useful in order refine one's intuition: they shed light on the framework on is operating in (in the case of Banach-Tarski, roughly speaking, mesaure theory), elucidating this framework's strengths or weaknesses (in the case of Banach-Tarski, restriction to measurable sets). One is then free to reject the framework or to keep it – augmented with appropriate warning labels about the range of its validity.

The "paradoxes" of the second type (like inappropriately "summing" infinite sequences to prove $0 = 1$) are useful in highlighting non-obvious errors and, hopefully, preventing one from committing similar errors later. Thus, one learns not to do unjustified things. One then may also learn, separately, how to do things in a more appropriate way. In the specific example of "summing infinite sequences" we say that one learns mathematical analysis – a more appropriate and contradiction-free (as far as we know) framework. Failure to use analysis properly does not constitute an indictment of methods of analysis – an ironic parallel with the statements Jaynes makes about critics of Bayeseanism.

Comments on 15.3 – 15.6.

What Jaynes is arguing, in effect, is that finitely-additive axiomatization is too weak, as manifest by the fact that it allows pathologies like "nonconglomerability". This means that some other axiomatization is needed; the mainstream mathematicians of today seem to agree, as it is the countably-additive axioms that are universally taught to students.

It is a curious linguistic phenomenon: who is "more careful" 1) the mathematician who works with weaker axioms, so as to not put in more assumptions than they deem justified OR 2) the mathematician works with more restrictive axioms, so as to avoid "pathological" examples allowed by the more permissive axioms. I'll leave the judgement up to linguist (though Jaynes seems to have a particular position).

Indeed, "nonconglomerability" is a phenomenon that does not appear in mainstream probability theory – the paper of Kadane, Schervish and Seidenfeld, to which this section addresses itself, is explisitely set in the context of finitely-additive, but not countably additive, probability. In fact, one of the points

of the KSS paper is to advocate for finitely-additive measures (despite their "pathologies" like nonconglomerability) as good fit for Bayesean framework (in particular for their ability to handle improper non-informative priors).

One way of reading section 15.3 is to see it as an argument against both the finitely-additive and countably-additive approaches, pointing out that some alternative, third way of thinking about probability is needed, perhaps more directly based on the "taking the limit in the end" idea. Coming up with a rigorous (which is to say, fully specified) version of such an alternative theory is an interesting, but unsolved and probably difficult question.

Exercise 15.1

For a quicker approximate result see in the end.

An exact calculation is as follows:

Let $N(l, a)$ be the number of sequences of results producing a specific record x of length l in with a annihilations (that is, in $n = l + 2a$ tosses). We will find a recursion relation for $N(l, a)$ (which will also establish that this number depends on x only via l, and so is well-defined; compare Jaybes in 15.5 "n and $y = y(n)$ are sufficient statistics" – his y is what we called l). Now, clearly,

$$N(l, 0) = 1.$$

Observe that $N(1, a)$ is independent of the one character $x = \alpha$ – given any sequence generating some other $\hat{x} = \hat{\alpha}$ we can relabel the characters to get a sequence generating α, and vice versa, establishing a bijection between the sequences generating α and $\hat{\alpha}$.

(In group theoretic language, we are studying paths in the Cayley graph of the free group on 2 generators e, μ; we will call all four of the $e = e^+, e^{-1} = e^-, \mu = \mu^+, \mu^{-1} = \mu^-$ "generators". The groups admits an automorphism sending any generator α to any other generator $\hat{\alpha}$; this automorphism produces an automorphism of the Cayley graph and hence bijections on paths from the identity to α and paths from identity to $\hat{\alpha}$.)

Now, observe that

$$N(0, a) = 4N(1, a - 1),$$

because the way to obtain empty word in the end is to start with some symbol α (4 options), and then produce a word that is equal to α^{-1}, which one can do in $N(1, a - 1)$ ways.

Finally, for all $l \geq 1$ one we have

$$N(l, a) = N(l - 1, a) + 3N(l + 1, a - 1),$$

because the first toss either produces the first symbol in x – after which one has to generate the rest of the symbols of x with a annihilations; or it produces one of 3 other symbols, after which one has to produce the inverse of that symbol and then x, with $a-1$ annihilations (one annihilation being taken up by cancelling the first tossed symbol with its inverse).

Thus we have our recurrence relation and boundary conditions:

$$N(l,0) = 1$$
$$N(0,a) = 4N(1,a-1)$$
$$N(l,a) = N(l-1,a) + 3N(l+1,a-1)$$

The problem asks for $N(20,10)$. A small dynamic programming script gives

$$38192689856872 \approx 0.38 \times 10^{14}.$$

The recurrence relation above confirms that (again, as per Jaynes in section 15.5) "it is a standard textbook random walk problem".

The removal of the reflection at 0 gives the following approximation: there are $\binom{40}{10}3^{30}$ total paths of length 40 with exactly 10 annihilations (aka "steps to the left"). They are distributed between 4×3^{19} final records. Thus the number per record is

$$\binom{40}{10} \times 3^{11}/4 = 37540129888404 \approx 0.38 \times 10^{14},$$

which is in good agreement with the exact result.

Comments 15.7

Mathematically, one says these days that there is no procedure for conditioning on an event of probability zero (aka a circle on a sphere). There are rigorous procedures for defining conditional expectations with respect to a random variable (e.g. latitude or longitude), or, more generally, a "sigma subalgebra", see Wiki and this paper of Chang and Pollard. Then one can define conditional probability of an event as conditional expectation of its characteristic function.

Alternatively, conditional probability can be defined in some more restrictive context via "disintegration". See section 4 of Terry Tao's blog post and, again, Chang and Pollard's paper.

Comments on 15.8

Some of Jaynes's points seem to be: 1) proper priors are incompatible with 15.60 removing the paradox, but not in the way DSZ thought 2) improper uninformative priors do not suffer from the paradox 3) for improper informative priors the paradox is resolved by observing that B_1 and B_2 start with different information.

It seems that 3) is somewhat doubtful. See Kevin Van Horn's page and his alternative resolution of the paradox via approximation by proper priors (and attributing the "paradox" to improper handling of infinities and non-uniform convergence). His basic point is echoed by a paper of Wallstrom, and from the perspective of disintegration by Examples 11 and 12 in Chang and Pollard's paper.

Exercise 15.2

Without checking convergence (!) we wirte:

$$p(\xi|z) \propto \int d\eta \;\; p(z|\eta\xi)\pi(\eta,\xi)$$

$$= \int d\eta \int dy \;\; p(z,y|\eta,\xi)\pi(\eta,\xi)$$

by 15.59 the inner integral is $p(z|\xi)$ so

$$= \int d\eta \;\; p(z|\xi)\pi(\eta,\xi) = p(z|\xi) \int d\eta \;\; \pi(\eta,\xi) = p(z|\xi)\pi(\xi)$$

which after normalization is 15.61.

Exercise 15.3

First of all, note that in the change-point problem it is $s = \frac{1}{\eta}$ which is the scale parameter. However, if distribution of the scale is $p(s)ds = s^{-1}ds$ then for the inverse scale $\eta(s) = \frac{1}{s}$ we have $|d\eta| = |\frac{1}{s^2}||ds|$ then $p(s)ds = s^{-1}s^2 d\eta = \frac{1}{\eta}d\eta$ (and conversely). So scale being distributed via s^{-1} and inverse scale being distributed via η^{-1} is equivalent.

Now, to the exercise itself. Let $u = \frac{y}{\eta}$. Then, assuming y is 1D, $dy = \eta du$

$$\int dy \;\; p(z,y|\eta,\xi) = \int dy \;\; \frac{1}{\eta} h(z,\xi,u) = \int du \;\; h(z,\xi,u)$$

is independent of η, so indeed 15.59 holds.

Then 15.58 becomes

$$p(\xi|x) \propto \int d\eta \;\; h(z, \xi, y/\eta) \frac{1}{\eta} \pi(\eta, \xi)$$

while 15.61 is

$$p(\xi|x) \propto \pi(\xi)\pi(z|\xi) = \pi(\xi) \int du \;\; h(z, \xi, u)$$

Now if we assume $\pi(\eta, \xi) \propto \pi(\xi) \times \frac{1}{\eta}$ then the two match up:

Put $\frac{y}{\eta} = v$, so $d\eta = -\eta^2 dv$. Then (noting that limit reversal in the integral kills the minus sign) we get from 15.58

$$p(\xi|x) \propto \int d\eta \;\; h(z, \xi, y/\eta) \frac{1}{\eta} \pi(\eta, \xi)$$

$$\propto \int dv \;\; h(z, \xi, v)\pi(\xi) = \pi(\xi) \int du \;\; h(z, \xi, u)$$

matching 15.61.