

The A Priori Problem of Observed Probabilities

Nassim Nicholas Taleb

Life is not a laboratory in which we are supplied probabilities. Nor is it an exercise in textbooks on statistics. Nor is it an urn. Nor is it a casino where the state authorities monitor and enforce some probabilistic transparency.

Alas we do not observe probabilities; we estimate them from observations and samples. This discussion presents the principal problem of empirical probabilistic knowledge and discovery. The central problem is as follows. Without a strong, necessarily normative, *a priori* specification of the underlying reality of a process, epistemic confidence is inversely proportional to the consequences of the knowledge at hand. The more events matter, the worse our empirical knowledge about their properties. Large deviations, the one that can have the most meaningful effect on the total properties, are far more error-prone in *empirical* evaluation than regular ones.

This note will also present the epistemological difference between the non-scalable and scalable distributions –as well as more minor issues often bundled under the archaic designation “Knightian” uncertainty.

THE TELESCOPE PROBLEM

Assume that there are series of discrete “events” i of magnitude λ_i (taken as departures from a “norm”, a “mean”, or a center) and with probability π_i (i can also be a “slice”, a discretizing interval for a continuous distribution, in which case we assume equality in the slices). Assume further that λ_i can take values on the real line, with no *known* upper bounds or lower bounds, between minus infinity and infinity. (*I am not assuming that an upper or lower bound does not exist, only that we do not know where it is.*) Assume that you are observing samples in a finite set, and deriving properties that are general, applying outside the sample set (you could also be deriving a probability distribution of future events from past events). You observe a set of values λ_i and make an inference about their possible frequency outside such set.

Let me make it clear: You are not sampling within an urn, the composition of which you know. You are making statement outside of a sample set. This means that you are now subjected to the problem of induction.

Now consider the moment contribution of every state of the world i , with $M_{i,m} = \pi_i \lambda_i^m$ the moment contribution of order m . The total moment of order n would be the summation of M over all possible values of i .

$$M_m = \sum \pi_i \lambda_i^m$$

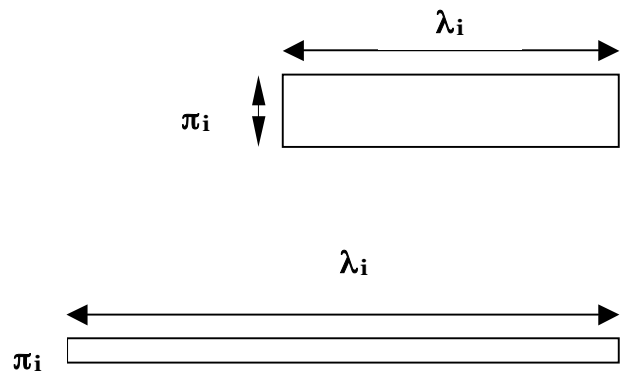
Let us turn to the estimation of probabilities. How does the sample reveals the values π_i ? Assume that you get them by simply taking the sample of size n , and adding up the values of observations for a given λ , here n_λ/n , where n_λ are the number of observations of the event of magnitude λ .

Let us call the “true” probabilities π_i^* , the probabilities that can be obtained by having full knowledge of the generating process. We do not observe them, but they are the ones that, in a world where they can be designed, the data is sampled from them.

The constraint that all π_i add up to 1 is not sufficient in many situations. So the central problem is as follows.

- **The small probability estimation error:** The smaller the π_i^* , the larger we need the sample n in order to be to be satisfied with our inference, and, for a given sample size, the higher the error in using π_i as an estimator of π^* .
- **The telescope problem:** if the $|\lambda|$ is some decreasing function of π , for $|\lambda|$ large enough, then *the smaller the probability, the more consequential* the impact of the error on the total moment. Effectively this is the case with unimodal distributions.

So the errors in the smaller π are multiplied by a larger λ . The pair $\pi_i \lambda_i$ is a rectangle that gets thinner as π becomes smaller, but its surface is more stochastic.



The Telescope Problem: Smaller probability π multiplies larger deviation λ , with the error in the

estimation of the product $\pi\lambda$ getting larger as π gets smaller.

SAINT PETERSBURG COMPOUNDED

But things can get worse for the "rectangle". Let us now consider the *shape* of the probability decrease, how the π need to drop as $|\lambda|$ rises. It is a fact that *the π cannot decrease too slowly*. The first intuition of the problem is in the well-known Saint Petersburg paradox (i.e., for n between 1 and infinity, a payoff λ_i of 2^i with probability $\pi_i = 1/2^i$), showing that no matter how low the π_i , the increase in $|\lambda|$ can be such that the moment contributions of all M_i are equal. Simply, the situation in which $\lambda_i = \frac{1}{\pi_i} K$, where K is a constant, makes the computation of all moments >1 impossible.

Generalization of the Saint Petersburg problem to higher moments m . So more generally, for the moment M_m to exist we need (in a unimodal distribution) the slices of $M_{i,m}$ to decrease as a certain rate, i.e.,

$$\lambda_i < \left(\frac{1}{\pi_i} K \right)^{1/m}$$

So the problem is that it is hard for a probability distribution to do the job *for all moments m* unless π falls faster than λ^m for all m . For that to happen, the only typical form is one with *a characteristic scale*, like $\pi_i = e^{-a\lambda_i}$. It makes the product $\pi_i \lambda_i^m$ decrease for larger values of λ . What I call non-nonscalable is, simply, the situation in which we do not have that exponential decline. See Note 1 for details.

The other solution is to give up on moments.

THE A PRIORI PROBLEM

Let us now look at the error rate in the estimation of π , and the difficulty that for higher orders of M , there is a compounding effect of a higher error it multiplies larger and larger values of λ . This is clearly intractable.

This can be solved, of course, with an arbitrary function that decreases the moment contributions M_i of the λ as these become larger. In other words, by assuming *a priori* a certain class of distributions.

Accordingly, we need one of the following three solutions:

Solution 1: A metaprobabilistic framework: allowing us to estimate the errors of the observed π_i some measure of the difference between probabilities and perfect information. In other words, we would have $\pi_{i,j}^*$ for every probability i , making every moment contribution $M_{i,m}$ stochastic. But here again there is the

risk of regress as the metaprobability needs to be checked as well for its own error –we need a model of error measurement for that.

Risk v/s Uncertainty: Note the difference between the so-called "Knightian" risk and uncertainty can be expressed as follows: risk is normatively set with unitary metaprobabilities $\pi_{i,j}^* = 1$ for all j (or no metaprobability). This makes the difference between the two entirely normative, not empirical. In other words, the difference does not exist in an environment where one cannot accept epistemological certainties about probabilities without a priori.

Solution 2: Assuming beforehand a probability distribution: If distributions are popular, it is because they allow normatively to make inferences about probabilities by analogy with other probabilities in place of the metaprobabilistic framework. Here the estimation error concerns some parameters of the probability distribution, but not the probabilities themselves.

This option is problematic because there is no justification for the derivation of probability distributions *internally from the data*, i.e. empirically, causing another infinite regress argument. We need data for probability distribution and a probability distribution to know how much data we need. So we cannot do a meta-probability distribution. This is where it becomes normative.

Note that there are situations in which one can put a subordination: you sample between two probability distributions, say two Gaussians. The

Solution 3: Truncation: of some values of the λ allowing the integration and the finiteness of M_m . This is not done by assumption, but, rather, in eliminating the sensitivity of the variable above a certain amount. (Note one aspect of Saint-Petersburg: The use of utility of payoff introduced a soft truncation. But you can set the game in a way to truncate "organically".)

CONCLUSION

There are no ways to deal with unbounded payoffs probabilistically without making assumptions, and assuming that these assumptions are not subjected to probabilistic judgment. This is the tragedy of probabilistic reasoning in modern domains, for which we cannot tolerate *a priori* probabilities.

MATHEMATICAL POINTS

Note 1: Assume that $\pi(x)$ is a continuous function. Assume x in the positive domain and p monotonic in each domain "for x large enough". We need $M[n]$, i.e., $\pi(\lambda) \lambda^n$ to be a decreasing function of λ for all n . To satisfy the strict negativity of $M'[n]$ the derivative of the "moment slice" with respect to λ , we have

$$\lambda^{n-1} (n \pi(\lambda) + \lambda \pi'(\lambda)) < 0$$

which fails for n at infinity with any scalable distribution with exponent α ; as we $\pi(\lambda) = K \lambda^{-\alpha}$

$$(n-\alpha) \lambda^{n-\alpha-1} < 0$$

Here since we are in the positive domain ($\lambda > 0$ and "large")

$$n > \alpha$$

Now, on the other hand take $\pi(x) = K e^{c x}$. We need

$$e^{cn} \lambda^{n-1} (n+c\lambda) < 0$$

$n > -c \lambda$, so c needs to be negative scaling constant.

Note 2: Assume the rosy case that you have *a priori* knowledge of the distribution. If the π are not of the form $\pi_i = e^{-a\lambda_i}$, then the increase in n makes the contribution of $\text{Max} [\pi_i \lambda_i^n] / M_m$, the moment contribution of the extremal variable to the total moments drop extremely slowly as n becomes larger (of the order of $1/6$ for a finite variance distribution). Simply, the distribution of the maximum will then have a Frechet form $\sim x^{-u}$. Accordingly, even if you know the distribution, the sampling error in the tails remains huge. Too huge for any application.