

# **On the Very Unfortunate Problem of Not Observing Probability Distributions**

**Nassim Nicholas Taleb and Avital Pilpel<sup>1</sup>**

SECOND DRAFT OCTOBER 2003

Cannot be quoted without permission

---

<sup>1</sup> We thank participants at the American Association of Artificial Intelligence Symposium on Chance Discovery in Cape Cod in November 2002, Stanford University in March 2003, the Italian Institute of Risk Studies in April 2003, and the ICBI Derivatives conference in Barcelona in May 2003.

A severe problem with risk bearing is when one does not have the faintest idea about the risks incurred. A more severe problem is when one does not have the faintest idea about the risks incurred yet thinks he has a precise idea of them. Simply, one needs a probability distribution to be able to compute the risks and assess the likelihood of some events.

These probability distributions are not directly observable, which makes any risk calculation suspicious since it hinges on knowledge about these distributions. Do we have enough data? If the distribution is, say, the traditional bell-shaped Gaussian, then yes, we may say that we have sufficient data. But if the distribution is not from such well-bred family, then we do not have enough data. But how do we know which distribution we have on our hands? Well, *from the data itself*. If one needs a probability distribution to gauge knowledge about the future behavior of the distribution from its past results, and if, at the same time, one needs the past to derive a probability distribution in the first place, then we are facing a severe regress loop—a problem of self reference akin to that of Epimenides the Cretan saying whether the Cretans are liars or not liars. And this self-reference problem is only the beginning.

What is a probability distribution? Mathematically, it is a function with various properties over a domain of “possible outcomes”,  $X$ , which assigns values to (some) subsets of  $X$ . A probability distribution describes a general property of a system: a die is a *fair die* if the probability distribution assigned to it gives the expected values that satisfy the requirements from the expectation operator. It is not that different, essentially, than describing mathematically other properties of the system (such as describing its mass by assigning it a numerical value of two kilograms).

The probability function is (usually, although not always—footnote to Poincaré) derived from specific instances from the system’s past: the tosses of the die in the past might justify the conclusion that, in fact, the die has the property of being fair, and thus correctly described by the probability function above.

Typically with time series one uses the past for sample, and generates attributes of the future based on what was observed in the past. Very elegant perhaps, very rapid shortcut maybe, but certainly dependent on the following: that the properties of the future resemble those of the past, that the observed properties in the past are sufficient, and that one has an idea on how large a sample of the past one needs to observe to infer properties about the future.

But there are worst news. Some distributions change all the time, so no matter how large the data, definite attributes about the risks of a given event cannot be inferred. Either the properties

are slippery, or they are unstable, or they become unstable because we tend to act upon them and cause some changes in them.

Then what is all such fuss about “scientific risk management” in the social sciences with plenty of equations, plenty of data, and neither any adequate empirical validity (these methods regularly fail to estimate the risks that matter) or any intellectual one (the argument above). Are we missing something?

An example. Consider the statement "it is a ten sigma event", which is frequently heard in connection with an operator in a stochastic environment who, facing an unforeseen adverse rare event, rationalizes it by attributing the event to a realization of a random process whose moments are well known by him, rather than considering the possibility that he used the wrong probability distribution.

Risk management in these sciences (particularly Economics) is plagued by the following central problem: *one does not observe probability distributions, merely the outcome of random generators*. Much of the sophistication in the growing science of risk measurement (since Markowitz 1952) has gone into the mathematical and econometric details of the process, rather than realizing that the hazards of using the wrong probability distribution will carry more effect than those that can be displayed by the distribution itself. This recalls the story of the drunkard looking for his lost keys at night under the lantern, because "that is where the light is". One example is the blowup of the hedge fund Long Term Capital Management in Greenwich, Connecticut<sup>2</sup>. The partners explained the blowup as the result of "ten sigma event", which should take place once per lifetime of the universe. Perhaps it would be more convincing to consider that, rather, they used the wrong distribution.

It is important to focus on catastrophic events for this discussion, because they are the ones that cause the more effect –so no matter how low their probability (assuming it is as low as operators seem to believe) the effect on the expectation will be high. We shall call such catastrophic events *black swan events*. Karl Popper remarked<sup>3</sup> that when it comes to generalizations like “all swans are white”, it is enough for *one* black swan to exist for this conclusion to be false. Furthermore, before you find the black swan, *no amount of information* about white swans – whether you observed one, 100, or 1,000,000 of them – could help you to determine whether or not the generalization “all swans are white” is true or not. We claim that risk bearing agents are in the same situations. Not only can they not tell before the fact whether a catastrophic event will happen, but *no amount of information about the past behavior of the*

---

<sup>2</sup> Lowenstein 2000.

<sup>3</sup> We use “remarked” not “noticed”—Aristotle already “noticed” this fact; it’s what he did with the fact that’s important.

*market* will allow them to limit their ignorance – say, by assigning meaningful probabilities to the “black swan” event. The only thing they can honestly say about catastrophic events before the fact is: “it might happen”. And, if it *does* indeed happen, then it can *completely destroy our previous conclusions about the expectation operator*, just like finding a black swan does to the hypothesis “all swans are white”. But by then, it’s too late.

Obviously, mathematical statistics is unequipped to answer questions about whether or not such catastrophic events will happen: it *assumes* the outcomes of the process we observe is governed by a probability distribution of a certain sort (usually, a Gaussian curve.) It tells us nothing about why to prefer this type of “well behaved” distributions to those who have “catastrophic” distributions, or what to do if we suspect the probability distributions might change on us unexpectedly.

This leads us to consider epistemology. By epistemology we mean the problem of the theory of knowledge, although in a more applied sense than problems currently dealt with in the discipline: *what can we know about the future, given the past?* We claim that there are good philosophical and scientific reasons to believe that, in Economics and social sciences, one *cannot* exclude the possibility of future “black swan events”.

### **THREE TYPES OF DECISION MAKING AND THE PROBLEM OF RISK MANAGEMENT**

Suppose one wants to know whether or not  $\underline{P}$  is the case for some proposition  $\underline{P}$  – “The current president of the United States is George W. Bush, Jr.”; “The next coin I will toss will land ‘heads’”; “There are advanced life forms on a planet orbiting the star Tau Ceti”.

In the first case, one can become *certain* of the truth-value of the proposition if one has the right data: who is the president of the United States. If one has to choose one’s actions based on the truth (or falsity) of this proposition – whether it is appropriate, for example, to greet Mr. Bush as “Mr. President” – one is in a state of *decision making under certainty*. In the second case, one cannot find out the truth-value of the proposition, but one can find out the *probability* of it being true. There is – in practice - no way to tell whether or not the coin will land “heads” or “tails” on its next toss, but under certain conditions one can conclude that  $p(\text{‘heads’}) = p(\text{‘tails’}) = 0.5$ . If one has to choose one’s actions based on the truth (or falsity) of this proposition – for example, whether or not to accept a bet with 1:3 odds that the coin will land

“heads” – one is in a state of *decision making under risk*.

In the third case, not only can one not find out the truth-value of the proposition, but one cannot give it any meaningful probability. It is not only that one doesn't know whether advanced life exist on Tau Ceti; one does not have any information that would enable one to even estimate its probability. If one must make a decision based on whether or not such life exists, it is a case of *decision making under uncertainty*<sup>4</sup>.

More relevant to economics, is the case when one needs to make decisions based whether or not future social, economical, or political events occur – say, whether or not a war breaks out. Many thinkers believed that, where the future depends not only on the physical universe but on human actions, there are no laws – even probabilistic laws – that determine the outcome; one is always “under uncertainty”<sup>5</sup>. As Keynes(1937) says:

By “uncertain” knowledge... I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty... The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention... About these matters, there is no scientific basis on which to form any calculable probability whatever. We simply do not know!<sup>6</sup>

Certainty, risk, and uncertainty differ not merely in the probabilities (or range of probabilities) one assigns to P, but in the strategies one must use to make a decision under these different conditions. Traditionally, in the “certainty” case, one chooses the outcome with the highest

---

<sup>4</sup>For the first distinction between risk and uncertainty see Knight (1921) for what became known as "Knightian risk" and "Knightian uncertainty". See Keynes (1937).

<sup>5</sup>Queasiness about the issue of uncertainty, especially in the case of such future events, had lead Ramsey (1931), DeFinetti(1937), and Savage(1954) to develop a “personalistic” or “subjective” view of probability, independent of any objective chance or lack thereof.

<sup>6</sup>“We simply do not know” is not necessarily a pessimistic claim. [Add about the “essential unknowledge” theory of free will and choice—mention Shackle, but go back to the origins! It goes back to the question of an omniscient God vs. free will, of course. Indeed, Shackle(1955) bases his entire economic theory on this “essential unknowledge” – that is, uncertainty - of the future. It is this “unknowledge” that allows for effective human choice and free will: for the human ability to *create* a specific future out of “unknowledge” by its efforts. – in other words, for free will. If the rules of probability applied to the future in the same way they apply to games of chance, says Shackle, all that one could do would be to passively *predict* which of several already-determined futures would occur.

utility. In the “risk” case, one chooses the outcome with the highest *expected* utility<sup>7</sup>. In the (completely) “uncertain” case, many strategies have been proposed. The most famous one is the minmax strategy (von Neumann and Morgenstern, 1944; Wald, 1950), but others exist as well, such as Savage’s “minmax of Regret” or “Horowitz’s alpha”. These strategies require bounded distributions. In the event of the distributions being unbounded the literature provides no meaningful answer.

### THE CENTRAL PROBLEM OF RISK BEARING

Using a decision-making strategy relevant to decision under risk in situations that are best described as cases of uncertainty, will lead to grief. If a shadowy man in a street corner offers me to play a game of three-card Monte, I will quickly lose everything if I consider the game a risk situation with  $p(\text{winning}) = 1/3$ . I should also consider the possibility that the game is rigged and my actual chances of winning are closer to zero. Being uncertain where in the range  $[0, 1/3]$  does my real chance of winning lies should lead one to the (minmax) uncertainty strategy, and reject the bet.

We claim that the practice of risk management (defined as the monitoring of the possibility and magnitude of adverse outcomes) subjects agents to just such mistakes. We argue below that, for various reasons, risk managers cannot rule out “catastrophic events”. We then show that this ever-present possibility of black swan events means that, in most situations, Risk managers are *essentially uncertain* of the future in the Knightian sense: where no meaningful probability can be assigned to possible future results.

Worse, it means that in many cases no known lower (or upper) bound can even be assigned to the range of outcomes;

---

<sup>7</sup>These ideas seem almost tautological today, but this of course is not so. It took von Neumann and Morgenstern(1944), with their rigorous mathematical treatment, to convince the world that one can assign a meaningful expected-utility function to the different options when one makes choices under risk or uncertainty, and that maximizing this expected utility (as opposed to some other parameter) is the “rational” thing to do. The idea of “expected utility” *per se* is already in Bernoulli(1738) and Cramer(1728), but for a variety of reasons its importance was not clearly recognized at the time.

Worst of all., it means that, while it is often the case that sampling or other actions can reduce the uncertainty in many situations, risk managers often face situations where no amount of information will help narrow this uncertainty.

The general problem of risk management is that, due to essential properties of the generators risk managers are dealing with, they are dealing with a situation of essential uncertainty, and not of risk.

To put the same point slightly more formally: risk managers look at collection of state spaces<sup>8</sup> that have a cumulative probability that exceeds a given arbitrary number. That implies that a *generator* of a certain *general type* (e.g., known probability distribution: Normal, Binomial, Poisson, etc. or mere histogram of frequencies) determines the occurrences. This generator has specific *parameters* (e.g. a specific mean, standard deviation, and higher-level moments) that – together with the information about its general type – determine the values of its distribution. Once the risk manager settles on the distribution, he can calculate the “risk” – e.g., the probability - of certain states of the world in which he is interested.

In almost all important cases, whether in the “hard” or “soft” sciences, the generator is hidden., There is *no* independent way to find out the parameters – e.g. the mean, standard deviation, etc. - of the generator except for trying to *infer it from the past behavior* of the generator. On the other hand, in order to give any estimate of these parameters in the first place, one must first *assume* that the generator in question *is* of a certain general type: that it is a Normal generator, or a Poisson generator, etc. The agent needs to provide a joint estimation of the generator and the parameters.

Under some circumstances, one is justified in assuming that the generator is of a certain general type and that the estimation of parameters from past behavior is reliable. This is the situation, for example, in the case of a repeated coin toss as one can observe the nature of the generator and assess the boundedness of its payoffs.

Under other circumstances, one might be justified in assuming that the generator is of a certain general type, but *not* be justified in using the past data to tell us anything reliable about the moments of the generator, no matter how much data one has.

Under even more troubling circumstances, one might have no justification not only for guessing the generator’s parameters, but also in guessing what *general type* of generator one is dealing with. In that case, naturally, it is meaningless to assign any values to the parameters of the generator, since we don’t know what parameters to look for in the first place.

---

<sup>8</sup>By "state-space" is meant the foundational Arrow-Debreu state-space framework in neoclassical economics.

We claim that most situations risk managers deal with are just such “bad” cases where one cannot figure out the general type of generator solely from the data, or at least give worthwhile estimate of its parameters. This means that any relation between the risks they calculate for “black swan” events, and the *actual* risks of such events, may be purely coincidental. We are in uncertainty: we cannot tell not only whether or not  $\underline{X}$  will happen, but not even give any reliable estimate of what  $p(\underline{X})$  is. The cardinal sin risk managers commit is to “force” the square peg of uncertainty into the round hole of risk, by becoming convinced without justification both of the generator type and of the generator parameters.

In the remainder of this paper we present the problem in the “Gedanken” format. Then we examine the optimal policy (if one exists) in the presence of uncertainty attending the generator.

#### Four “Gedanken” Monte Carlo Experiments

Let us introduce an invisible generator of a stochastic process. Associated with a probability space it produces observable outcomes. What can these outcomes reveal to us about the generator – and, in turn, about the future outcomes? What – if anything – do they tell us about its mean, variance, and higher order moments, or how likely the results in the future are to match the past?

The answer depends, of course, on the properties of the generator. As said above, Mother Nature failed to endow us with the ability to observe the generator--doubly so in the case of social science generators (particularly in economics).

Let us consider four cases. In all of these cases we observe mere *realizations* while the generator is operated from behind a veil. Assume that the draws are generated by a Monte Carlo generation by a person who refuses to reveal the program, but would offer samples of the series.

Table 1: The Four *Gedanken* Experiments

Gedanken	Probability Space	Selected Process	Effect	Comments
1	Bounded	Bernouilli	Fast convergence	"Easiest" case

2	Unbounded	Gaussian (General)	Semi-fast convergence	"Easy" case
3	Unbounded	Gaussian (mixed)	Slow convergence	Problems with solutions
4	Unbounded	Stable Pareto-Levy	No convergence	No known solutions

### THE “REGULAR” CASE, TYPE I: DICE AND CARDS

The simplest kind of random process (or “chance setups” as they are sometimes called) is when all possible realizations of the process are bounded. A trivial case is the one of tossing a die. The probability space only allows discrete outcomes between 1 and 6, inclusive.

The effect of having the wrong moments of the distribution is benign. First, note that the generator is *bounded*: the outcome cannot be more than 6 or less than 1. One cannot be off by more than a finite amount in estimating the mean, and similarly by some finite amount when estimating the other moments (although, to be sure, it might become a relatively large amount for higher-level moments) (note the difference between unbounded and infinite. As long as the moments exist, one *must* be off by only a finite amount, no matter what one guesses. The point is that the finite amount is *unbounded* by anything *a priori*. Give examples in literature—original one, preferably, E.).

Second, the bounded-ness of the generator means that there are *no extreme events*. There are no rare, low-probability events that any short run of the generator is unlikely to cover, but yet have a significant effect on the value of the true moments. There are certainly *no “black swan” events*—no outcomes whose result could destroy our previous estimates of the generator’s moments no matter how much previous data we have. That is,  $E(\underline{X}_n)$  (the observed mean) is likely to be close to  $E(\underline{X})$  (the “real”) mean since there is no rare, 1-in-1,000,000 chance of the die landing “1,000,000” – which would raise  $E(\underline{X})$  from the  $E(\underline{X})$  of a “regular” die almost by 1 but will not be in the observed outcomes  $x_1, x_2, \dots, x_n$  unless one is extremely lucky.

### THE “REGULAR” CASE, TYPE II: NORMAL DISTRIBUTION

A more complicated case is the situation where the probability space is unbounded. Consider the normal distribution with density function  $f_2$ . In this case, there is a certain  $>0$  probability for the outcome to be arbitrarily high or low; for it to be  $>M$  or  $<m$  for arbitrary  $M, m \in \mathbb{R}$ .

However, as  $M$  increases and  $m$  decreases, the probability of the outcome to be  $>M$  or  $<m$  becomes very small very quickly.

Although the outcomes are unbounded the epistemic value of the parameters identification is simplified by the “compactness” argument used in economics by Samuelson<sup>9</sup>.

A compact distribution, short for “distribution with compact support”, has the following mathematical property: the moments  $M[n]$  become exponentially smaller in relation to the second moment<sup>10</sup>.

But there is another twist to the Gaussian distribution. It has the beautiful property that it can be *entirely characterized by its first two moments*<sup>11</sup>. All moments from  $n = \{3, 4, \dots, \infty\}$  are merely a multiple of  $M[1]$  and  $M[2]$ . Thus, knowledge of the mean and variance of the distribution would be sufficient to derive higher moments. We will return to this point a little later. (Note tangentially that the Gaussian distribution would be the maximum entropy distribution conditional on the knowledge of the mean and the variance.)

From this point on—consider Levi more? Also induction more? These are the things that we need to add...

---

<sup>9</sup>See Samuelson (1952).

<sup>10</sup> A Noncentral moment is defined as  $M[n] \equiv \int_{\Omega} x^n \phi(x) dx$ .

<sup>11</sup> Take a particle  $W$  in a two dimensional space  $W(t)$ . It moves in random increments  $\Delta W$  over laps of time  $\Delta t$ . At times  $t+\Delta t$ , we have  $W(t+\Delta t) = W + \Delta W + \frac{1}{2} \Delta W^2 + \frac{1}{6} \Delta W^3 + \frac{1}{24} \Delta W^4 + \dots$  Now taking expectations on both sides:  $E[W(t+\Delta t)] = W + M[1] + \frac{M[2]}{2} + \frac{M[3]}{6} + \frac{M[4]}{24}$ , etc. Since odd moments are 0 and even moments are a multiple of the second moment, by stopping the Taylor expansion at  $M[2]$  one is capturing most of the information available by the system.

Another intuition: as the Gaussian density function for a random variable  $x$  is written as a scaling of  $e^{-\frac{(x-m)^2}{2\sigma^2}}$ , we can see that the density wanes very rapidly as  $x$  increases, as we can see in the tapering of the tail of the Gaussian. The interesting implication is as follows: Using basic Bayes' Theorem, we can compute the conditional probability that, given that  $(x-m)$  exceeds a

given  $2\sigma$ , that it falls under  $3\sigma$  and  $4\sigma$  becomes  $\frac{\int_2^3 \phi\left(\frac{x-m}{\sigma}\right) dx}{\int_{-\infty}^2 \phi\left(\frac{x-m}{\sigma}\right) dx} = 94\%$  and

$\frac{\int_2^4 \phi\left(\frac{x-m}{\sigma}\right) dx}{\int_{-\infty}^2 \phi\left(\frac{x-m}{\sigma}\right) dx} = 99.8\%$  respectively.

### THE "SEMI-PESSIMISTIC" CASE, TYPE III: "WEIRD" DISTRIBUTION WITH EXISTING MOMENTS.

Consider another case of unbounded distribution: this time, a linear combination of a "regular" distribution with a "weird" one, with very small probabilities of a very large outcome.

For the sake of concreteness assume that one is sampling from two Gaussian distributions. We have  $\pi_1$  probability of sampling from a normal  $N_1$  with mean  $\mu_1$  and standard deviation  $\sigma_1$  and  $\pi_2 = 1 - \pi_1$  probability of sampling from a normal  $N_2$  with mean  $\mu_2$  and standard deviation  $\sigma_2$ .

Assume that  $N_1$  is the "normal" regime as  $\pi_1$  is high and  $N_2$  the "rare" regime where  $\pi_2$  is low. Assume further that  $|\mu_1| \ll |\mu_2|$ , and  $|\sigma_1| \ll |\sigma_2|$ . The density function  $f_3$  of this distribution is a linear combination of the density functions  $N_1$  and  $N_2$ . Its moment-generating function,  $M_3$ , is also the weighted average of the moment generating functions  $M_1$  and  $M_2$ , of the "regular" and "weird" normal distributions, respectively, according to the well-known theorem in Feller (1971)<sup>12</sup>. This in turn means that the moments themselves ( $\mu_3, \sigma_3, \dots$ ) are a linear combination of the moments of the two normal distributions

---

<sup>12</sup> The mean  $m = \pi_1 \mu_1 + \pi_2 \mu_2$  and the standard deviation  $\sigma = \sqrt{\pi_1(m_1^2 + \sigma_1^2) + \pi_2(m_2^2 + \sigma_2^2) - (\pi_1 \mu_1 + \pi_2 \mu_2)^2}$ .

While the properties of this generator and the outcomes expected of it are much less “stable” (in a sense to be explained later) than either of the previous cases, it is at least the case that the mean, variance, and higher moments exist for this generator, Moreover, this distribution over time settles to a Gaussian distribution, albeit at an unknown rate

This, however, is not much a of a consolation when  $\sigma_2$  or  $\mu_2$  are very large compared to  $\sigma_1$  and  $\mu_1$ , as assumed here. It takes a sample size in inverse proportion to  $\pi_2$  to begin to reach the true moments: When  $\pi_2$  is very small, say 1/1000, it takes at least 1000 observations to start seeing the contribution of  $\sigma_2$  and  $m_2$  to the total moments.

### **THE “PESSIMISTIC CASE: NO FIXED GENERATOR**

Consider now a case where the generator itself is not fixed, but changes continuously over time in an unpredictable way; where the outcome  $x_1$  is the result of a generator  $G_1$  at time  $t_1$ , outcome  $x_2$  that of generator  $G_2$  at later time  $t_2$ , and so on. In this case, there is of course no single density function, moment-generating function, or moment can be assigned to the changing generator.

Equivalently, we can say that the outcome behaves as if it is produced by a generator which has no moments – no definite mean, infinite variance, and so on. One such generator is the one with moment-generating function  $M_4$  and density function  $f_4$  – the Pareto-Levy distribution<sup>13</sup> which is a special case of the stable distribution "L" Stable (for Levy-stable).

### **THE DIFFERENCES BETWEEN THE GENERATORS**

Suppose now that we observe the outcomes  $x_1, x_2, x_3 \dots x_n$  of generators of type (1)-(4) above, from the bound dice-throwing to the Pareto-Levy distribution. What could we infer from that data in each case? To figure this out, there are two steps: first, we need to do is figure out the

---

<sup>13</sup>See Samorodnitsky and Taqqu(1994). It is interesting that the Pareto-Levy distribution is only known by its characteristic function, not its density which cannot be expressed in closed form mathematically, but only as a numerical inversion of the Fourier transform.

*mathematical* relation between the observed moments ( $E(X_n)$ ,  $\text{Var}(X_n)$ , etc.) and the actual moments of the generator. Then, we need to see what epistemology tells us about the significance of these relations to our ability to *know* the actual moments.

### THE FIRST AND SECOND CASES.

In the first and second case, the moments of the generator (e.g.,  $E_1(X)$ ,  $\text{Var}_1(X)$ ,  $E_2(X)$ ,  $\text{Var}_2(X)$ , and higher-level moments) can be quickly inferred from the observation of the actual outcomes.

For example, the observed first moment – the observed mean  $E(X_n) = (x_1+x_2+\dots+x_n)/n$  – quickly converges to the actual mean  $E_1(X)$  or  $E_2(X)$  as  $n$  increases. Same with the observed variance of the sample  $\{x_1\dots x_n\}$ ,  $\text{Var}(X_n)$ , converging to  $\text{Var}_1(X)$  or  $\text{Var}_2(X)$ . The same is also true with higher-level moments.

Let us illustrate this point—the fast convergence of the observed moments to the actual moments—by considering the first moment, or the mean. In the first case (the dice), the outcomes are bounded, so that we know that  $\min(X) < x < \max(X)$  for sure. In the second case (the Normal distribution) the outcomes are not bounded, but their probability decreases drastically as they vary from the mean.

That is,  $p_i(x) \cdot x \rightarrow 0$  quickly as  $x$  increases to extreme values both in the case of the first and the second generator (that is, for  $i=1,2$ ). In the first case this is due to the fact that  $p_1(x)=0$  for  $x < \min(X)$  or  $x > \max(X)$ ; in the second, because  $p_2(x)$  decreases much faster than the deviation of  $x$  from the mean.

This means that the effect of extreme values on the mean of the generator,  $E_i(X) = \sum_x x \cdot p_i(x)$ , is negligible in both the bounded case ( $i=1$ ) and the Normal case ( $i=2$ ). That is,  $\sum_x x \cdot p_i(x) \sim \sum_x$  not an extreme value  $x \cdot p_i(x)$  for both generators.

Consider now the data we actually observe. Even if the low-probability extreme values of the generator (if such exist) are *not* observed at all in the outcomes  $x_1, x_2 \dots x_n$ , the “experimental”  $E(X_n) = (x_1+x_2+\dots+x_n)/n$  is *still* converging towards  $\sum_x$  not an extreme value  $x \cdot p_i(x)$ . This, as we said, will not differ much from the actual  $E_1(X)$  or  $E_2(X)$ . One does not, in other words, need to wait until a rare extreme event occurs, even if the possibility of such events exists, in order to get a reasonable estimate of the real  $E_1(X)$  or  $E_2(X)$  from the experimental  $E(X_n)$ .

For similar reasons,  $\text{Var}(X_n)$  will converge quickly to  $\text{Var}_1(X)$  or  $\text{Var}_2(X)$ , and the same for higher-level moments, even if  $x_1, x_2, \dots x_n$  does not include any of the extreme values that could occur – if any.

### THE “SEMI-PESSIMISTIC” CASE

Suppose now that the generator which generated our data – outcomes  $x_1, x_2, \dots, x_n$  – is of the third type, the “semi-pessimistic” case of a linear combination between a Normal and Poisson distribution.

In this case, the extreme values of the generator are *not* negligible for the calculations of the generator’s moment. That is since, while  $p_3(x) \rightarrow 0$  as  $x$  deviates greatly from the mean, it does not do so “fast enough” to make extreme values negligible. That is,  $p_3(x) \cdot x$  does not  $\rightarrow 0$  as  $x$  becomes extreme.

In such situations,  $E_3(X) = \sum_x p_3(x) \cdot x \neq \sum_{x \text{ not extreme value}} p_3(x) \cdot x$ . Therefore, as long as the rare extreme events do not occur, the “experimental”  $E(X_n)$  is converging towards  $\sum_{x \text{ not extreme value}} p_3(x) \cdot x$  - which might be very different from  $E_3(X) = \sum_x p_3(x) \cdot x$ .

In other words, the rare, extreme events need to *actually occur* before  $E(X_n)$  will be close to  $E_3(X)$  (if then). And similarly for  $\text{Var}(X_n)$  vs.  $\text{Var}_3(X)$  and the higher-level moments.

This is seen by the fact that in such generators, the conversion is much slower.

Furthermore, until extreme “black swan” results actually occur, the observed outcomes of the second (Normal) generator would be *indistinguishable* from the results of the third (Normal + Poisson) generator. We shall consider the implications of this later.

### THE “PESSIMISTIC” CASE

In the “pessimistic” case, things can be intractable. It is not that it takes time for the experimental moments  $E(X_n)$ ,  $\text{Var}(X_n)$ , etc. to converge to the “true”  $E_4(X)$ ,  $\text{Var}_4(X)$ , etc. In this case, these moments simply do not exist. This means, of course, that no amount of observation whatsoever will give us  $E(X_n)$ ,  $\text{Var}(X_n)$ , or higher-level moments that are close to the “true” values of the moments, since no true values exist.

### THE PROBLEM OF INDUCTIVE INFERENCE AND ITS RELATION TO THE MATHEMATICAL RELATIONS DISCUSSED ABOVE

So far, we have just *described four generators* and saw the mathematical relation they imply

between the value of the estimated moments and the actual moments (if they exist).

We now need to see how these properties affect the original question we considered: namely, under what circumstances can we use the data of the previous outcomes of the generator to establish the type of the generator and its parameters, and thus be able to predict the risk of future outcomes.

It should be emphasized that while these two problems – the *mathematical relation* between the generator's true moments and the observed moments, on the one hand, and the ability to *predict the future outcomes* of the generator are closely related, they are by no means identical. The first one is a purely mathematical problem. The second is an epistemological problem.

One can never conclude much about the future *solely* from a small specific set of outcomes, our “experimental data”. In the modern literature<sup>14</sup>, a corpus of knowledge, suggesting availability of background information is always imperative.

For example, one cannot tell, from a million observations of a coin toss *alone*, that the coin has a certain probability of landing “heads” on the next toss. There is nothing “in the data” itself that excludes, for example, the possibility that the coin will land neither “heads” nor “tails” the next time, but will explode like a nuclear bomb. Despite the close *mathematical* relation between the observed and actual moments, unless we have the right “background information”, we will not be able to make any *epistemological* conclusion from the data to the future behavior of the generator. The reason such outcomes as “will explode like a nuclear bomb” are excluded is that, in most case, we *have* the right kind of “background information” to exclude it – e.g., our knowledge of physics.

On the other hand, even if the generator is of the “pessimistic” Pareto-Levy type above, the lack of *mathematical* relation between the observed moments and the real moments might not – in theory! – exclude one from making an *epistemological* conclusion about the future outcomes of the generator. If by some miracle, for example, we have an access to an angel that whispers in our ear the next outcome of the generator before it occurs, then part of our “background information” simply *includes* the generator's outcome, and we could tell what the outcomes would be.

However, such cases are usually of a fantastic nature—in most cases we deal with, as seen below, the mathematical information is necessary, but *not* sufficient, to reach the epistemological conclusions we are interested in.

---

<sup>14</sup>See for example Levi(1980), Kyburg(1974).

## THE IMPLIED BACKGROUND INFORMATION AND OUR CLAIMS

As we said we are interested here in the epistemological problem given a *specific type of background information*, which is the situation in practice when risk managers need to “show their stuff”. We assume that the background information is such that:

- 1) Outcomes are created by some random generator;
- 2) That this random generator will continue to produce them in the future;
- 3) One does not have any independent way to estimate either the type of generator or its parameters except from the data of the previous outcomes, and that furthermore
- 4) The generator can be any one of the four different types of exclusive and exhaustive generators discussed above.

The first three assumptions about risk are not controversial. The fourth one is.

Our epistemological question is: *if* the background information is as above, *what if anything* can we conclude about the moments of the generator (and, hence, about its future behavior) from 1) the observed past behavior of the generator, and 2) this background information? Our practical question is: *when is it* the case that, indeed, the generator can be of all four types, or at least of the “pessimistic” type, type 3 or 4?

We claim that:

- 1) If the generator *can be* or type 3 or 4 (“semi-pessimistic” or “pessimistic”), that is enough to *invalidate* our ability to conclude much from its past behavior to its future behavior; in particular, it makes it impossible for us to assign *any specific probability* to future outcomes, which makes the situation one of uncertainty, as claimed in the introduction above.
- 2) It is precisely in situations dealt with by risk managers where the generator *can be* of type 3 or 4.

## THE PROBLEM OF INDUCTIVE INFERENCE: THE FIRST PART

Let us begin, then, with the first part of the problem, the “if-then” part: namely, under what circumstances we can (or cannot) say something about the moments of the the generator *if* we know (or do not know) the background something about what the generator is, or what type it could be.

There are two possibilities. It might be that certain information about the moments is a *deductive consequence* of what I already know about it. For example, if I know that a generator’s outcomes are bound between the values *a* and *b*, I know that the first moment is also so bound. This is not a matter of choice or decision: to be logically consistent, I *must* accept all such consequences the background information implies about the moments<sup>15</sup>.

More complicated is the case of *induction*. Even when (as we always assume) all the deductive consequences of the background information are known, it might be that no specific value for the moments emerges. In that case, we are not forced to settle on a specific value for them. Nevertheless, we might conclude that under the circumstances, we are *inductively justified* in assigning the mean of the generator a certain value (say, “3.5” in the “fair die” case), and similarly for higher moments.

We discuss induction more specifically below, in a separate part. But before we begin this section, a short summary is necessary.

As Peirce showed, this is really a epistemic *decision problem*. I am given background information about the generator (“it looks like a die of some sort is tossed”) and the previous outcomes (“the outcomes were 4, 4, 3, 2, 1”). I need to decide whether adding a new conclusion about the generator’s moments to my beliefs based on this data is justified (say, “the die is a *fair die*”, or more formally “the die’s first moment is 3.5”).

To solve the decision problem, as in all decisions problems, one needs to consider the *goal* (or goals) one tries to achieve, and the *options* one can choose from. To choose correctly means to choose the option that best achieves one’s goals. Decision-making goals can be anything from winning a nuclear war to choosing a good restaurant. The goals of inductive inference is (as James showed, below) to *seek new information* while at the same time *avoiding error*. Similarly, the available options can be anything from launching a Trident II missile to driving to the restaurant. In inductive inference, the options are *adding new claims* to one’s beliefs—in this case, claims about the value of a random generator’s moments<sup>16</sup>.

---

<sup>15</sup> See also the distinction between “doxastic commitment” and “doxastic performance” in the section about induction and deduction, below.

<sup>16</sup> Ref Levi & others.

These two goals are in tension: the more information I accept, the more likely it is that one will mistakenly include error. The question is, what new claims give me the most information for the least risk if I add them. The result of the inductive inference—the solution of the decision problem—is *adding to one's beliefs the claim that best balances these goals*. Adding this claim is the inductive inference justified under the circumstances.

Note that the null claim—“adds no new information”—is always available. If the optimal option is the null option, it means that the justified inductive inference is *no* inference. In our case it would mean that we are not justified in concluding anything about the generator's moments from our background information and past outcomes. As we shall see, this is often the case.

Note, further, that mere high probability, e.g. low risk of error, is *not* itself good enough for acceptance. Consider a lottery with a million tickets: the probability of each ticket winning is 1/1,000,000; but if we accept that this low probability, in itself, is enough to conclude that ticket  $n$  will *not* win, we reach the absurd conclusion that *no* ticket will win.

In what follows, we need to formalize and quantify the decision situation faced by the agent. For this we use the system developed by Levi. Other formalizations of epistemic decision-making in inquiry exist; in fact, one of the authors (Pilpel) is investigating the differences between these systems. But in the cases of risk management described below, all of them will recommend the same (pessimistic) conclusion.

## TYPE #1 AND #2 GENERATORS

Suppose that an angel came to us and told us the following: “the phenomena which you measured so far, with results  $x_1, x_2, \dots, x_n$ , is produced by a generator which is bound (type 1 above) between  $a$  and  $b$ , or which gives a normal distribution (type 2 above). However, I will not tell you what the mean, variance, or higher moments are; this you need to figure out from that data.” Could we do it? The answer is positive. Based on the mathematical analysis above, even if we do not know what the moments of the generator are, we still know that as  $n$  increases,  $E(X_n)$  quickly converges to the actual  $E(X)$  (and the same for higher-level moment, of course). We know that this is the case even when the generator itself is capable of producing extreme results, e.g., in the case of the normal distribution, as long as we have not observed them yet, as we presume is the case: we are dealing here, not with what kind of predictions we

can make if a rare event *already happened*, but with what to do if such an event has *not yet* happened.

But things are more “positive” than that. It is not only the case that we know that the observed moments—say the first one,  $E(X_n)$ —quickly converge to the real moments,  $E(X)$  in this case. In both the bounded and normal distribution case, we also have a good *numerical* idea of how likely the observed  $E(X_n)$  is to be more than a certain distance away from the real moment,  $E(X)$ .

This means that the agent dealing with estimating  $E(X)$  can tell, pretty quickly, just what the maximum probability is that  $E(X)$  and  $E(X_n)$  being more than a certain  $\epsilon$  away from each other—and that, for every given  $\epsilon$ , this probability becomes small quickly.

As we said above, induction is a matter of a *decision problem* concerning the *risk of error* versus the *gain in informational value* of adding something to one’s beliefs. In this case, the problem is whether or not the agent is justified in accepting that  $E(X)$  is *within some distance,  $\epsilon$ , from  $E(X_n)$* . The question is whether the risk of error incurred when accepting this statement is worth the informational value added to the agent’s beliefs when accepting it.

It turns out that this is the case. Due to the fact that the agent has a good idea of the risk of error incurred by adding this hypothesis to their belief, and the fact that the risk of error becomes small rather quickly, in most situations the agent would be likely to conclude that, indeed, the information gained is worth the small risk of making a mistake, requiring only a relatively small number of observations  $x_1, x_2 \dots x_n$  to accept the claim that  $E(X)$  is within a small distance  $\epsilon$  from the observed  $E(X_n)$ .

The answer is positive.

### **TYPE #3 GENERATORS – PART 1.**

The problem is that in most cases, the agent does *not* know that the generator is of type I or type II. The agent so *assumes*, but for no better reason than the fact that it is easy to reach seemingly “exact” results with such an assumption.

Suppose, for example, that so far the daily change in a stock’s price have been limited to the range between 0 and 10 points. Is there any reason to suspect that it will not move 1000 points one way or the other in the future? If we *knew* the generator that was producing the stock’s movements was normal, perhaps. But often we do not know it.

Suppose, however, that an angel told us: “the phenomena you are observing is generated by a generator of type #3. It is a combination of a “regular” Normal distribution and a Poisson distribution which distribution that gives us very large results with very low probabilities. I will not tell you what the mean, variance, or other moments of this generator are, however. You will have to figure them out from the data.”

What could we say about the mean, variance, and higher moments of this generator by looking at the data? Very little indeed – at least as long as no catastrophic “black swan” event *had* in fact occurred.

The reason is that in the case of such a distribution, most of the value of the moments comes from the rare and improbable “black swan” events that are due to the extreme Normal Poisson distribution, and not the regular and non-catastrophic events that are due to the Normal distribution. As long as no such catastrophic events occurs, then, we only know a “negative” point: that the observed moments  $E(X_n)$ ,  $\text{Var}(X_n)$ , etc. are not close to the actual moments  $E(X)$ ,  $\text{Var}(X)$ , etc. But that is all we know, *no matter how much (non-catastrophic) data we have*. We cannot say anything about what the size of the difference is until we actually observe such catastrophic events.

Let us put this in more formal epistemic form. As in the case of the first and second generators, let us presume that John wishes to evaluate what  $E(X)$  is based on the observed  $E(X_n)$ . (The same argument works, *mutatis mutandis*, for higher-level moments, as always.) As in the case of the first and second generators, the question is whether or not the agent is inductively justified in concluding that  $E(X)$  and  $E(X_n)$  are within some distance,  $\varepsilon$ , from each other.

Once more, this is a decision problem: is the risk of error in accepting this hypothesis worth the gain in informational value? In this case, the answer is usually negative, as the appendix shows. Before a catastrophic event occurs, the agent *doesn't know* what the risk of  $E(X)$  being more than  $\varepsilon$  away from  $E(X_n)$  is; and, in such a situation, the inductive expansion (addition of information, e.g., coming to believe that  $E(X_n)$  is within  $\varepsilon$  of  $E(X)$ ) is not justified.

In conclusion, even if we *know* that a certain generator is a type 3 (Normal + Poisson, or Normal + “wild” Normal with large mean and variance), before a catastrophic event occurs we cannot say anything about the difference between the observed  $E(X_n)$  and actual  $E(X)$  (and similarly for higher level moments). This, as the appendix shows in detail, remains the case *no matter how many observations occur*, as long as we know that no extreme event had yet occurred. Estimating the generator’s moments before an extreme event occurred is impossible

no matter how much data we have.

In conclusion: even if we *know* that a certain generator is a type 3 (Normal + Poisson) distribution, before a catastrophic event occurs we cannot say anything about the difference between the observed  $E(X_n)$  and  $E(X)$ , the observed  $\text{Var}(X_n)$  and  $\text{Var}(X)$ , or any other observed moment and the “real” one. Before such an event occurs, extrapolating from past data to future behavior of such a system is *worthless*.

### TYPE #4 GENERATORS – PART 1

Things are even worse with type 4 generators, for obvious reasons. If an angel tells us that a certain generator is a type 4 one (Pareto-Levy), we know that no relation between the observed moments  $E(X_n)$ ,  $\text{Var}(X_n)$ , etc. and the “real” moments of the generator exist – for the very good reason that there are *no such moments*.

Therefore, if we are observing a type 4 generator, no amount of observational data will tell us anything. Both type 3 and type 4 generators are “bad news”: before a catastrophic event occurs (in case 3) or even after (case 4), extrapolating from past data to future behavior of such a system is *worthless*.

### TYPE #3 AND #4 GENERATORS – PART 2

But things are even worse than that. We have just seen that if we know that the generator is of type 1 or type 2, we can rely on the observed moments to be close to the “real” moments. We also showed that if we know that the generator is of type 3 or type 4, the observed moments (at least before a catastrophic “black swan” event occurs) are worthless in finding the values of the real moments.

But all these scenarios assume that we know what type the generator is. Suppose we *don't* know what it is, and want to see if the data helps us figure this out? In that case, the mathematical equality between the observed and actual moments, *even if it holds* (even if the generator, that is, is in fact of type #1 or #2), might not be enough to reach any epistemological conclusions about the similarity of the past to the future. The mathematical equality is necessary, but not sufficient.

Consider the following situation. Suppose an angel tells you that a certain generator is either type 2 (Normal) *or* type 3 distribution (a mixed combination of Normal and Poisson). Consider the data  $x_1, x_2, \dots, x_n$ . As long as no catastrophic “Poisson event” had actually occurred, the data would be *indistinguishable* between type 2 and type 3 generators, since all the outcomes of

the type 3 generator would still be due to the “Normal” part of its distribution. We will not be able to tell due to anything in the data whether it is one or the other.

More generally, suppose that an angel tells us that a certain outcome *might* be due to a generator of type 3 or 4, as well as a type 1 or 2 generators. Does any amount of data tell us anything about whether or not this is true, before a “black swan” event happens? No, since until a low-probability catastrophe actually occurs, *if* the generator is in fact of type 3 or 4, the data would look *indistinguishable* from that of a generator of type 1 or 2, as we’ve just seen..

So if we *don’t know* that the generator is *not* type 3 or 4, then our data is *just as worthless* in assessing the future behavior of the generator as if we knew that it is type 3 or 4. This is not because  $E(X_n)$ ,  $\text{Var}(X_n)$  and so on *must* be far from the “real”  $E(X)$ ,  $\text{Var}(X)$ , etc. (if they exist), but because we can never tell from the data *whether* they are or not before a catastrophe happens.

And if we don’t know the moments, *ipso facto* we don’t know anything about the probabilities of the generator’s outcomes, which depend for their calculation on these moments. We cannot tell anything about the risk of any future outcome. We are in a situation of *decision making under uncertainty*.

In summary, for the epistemic inductive inference from the past outcomes to the future ones to be worthless, we need not know that the generator is of the “dangerous” type: it need not be the case that  $E(X) \neq E(X_n)$  (or the same for the other moments). It is enough *not* to know that it is not of that type. In such a situation, a “black swan” could surprise it at any moment – and we wouldn’t be able to tell whether it would happen or not until after the fact. The mathematical equality  $E(X) = E(X_n)$  is of no use to us if we cannot *know in advance* that it holds before a catastrophic event occurs.

### **COULD SUCH GENERATORS EXIST?**

This entire discussion would have remained completely theoretical if it was not the case that the situations risk managers deal with *could* involve the “bad” types of the generators – that is, unless epistemic assumption #4 above holds.

We have seen above that many economists dismiss the possibility of assumption #4. we claim that, unfortunately, in economical situations generators of this type *can* occur. Physical

systems (to borrow from Mandelbrot) must be of the “benign” type – type 1 or 2, or, more specifically, of type 1 (a “bounded” generator). The laws of physics bound their values – specifically, the amount of energy in the system, the entropy of the system, and other such physical characteristics cannot move beyond a certain range.

In physical and social systems, therefore, it is often the case that we can tell in advance, due to external, purely deductive reasons, that the “generator” must be bounded and therefore (relatively) benign; we can therefore use the past data for inductive inference about the future, as we seen above.

In many *financial* systems, however, this is not the case. There are potential events in many such systems that would cause losses (or gains) that are, in theory, *unbounded*. To convince oneself of this, one need only look at a simple “option”: the possibility exists of losing an infinite amount of money combined with the fact that such probability may remain unknown by us.

This is not to say, of course, that death is somehow “better” than losing a lot of money, or that gaining or losing an infinite (or very, very, large) amount of money is physically possible. The point is, rather, that in the case of a physical system one knows that one can describe the system with a bounded (or, at worse, a compact-supported) generator, while if we look at a financial system this cannot be promised. (Remove this paragraph, perhaps? Or give more references?)

### **THE RECOMMENDED STRATEGY IN SUCH SITUATIONS, AND “LONG-TERM CAPITAL” REVISITED**

The conclusion of this epistemological excursion is as follows: In such situations, then, we are in an essentially “uncertain” situation.

If we must make decisions in such a situation, our best bet is to use a strategy suited to “uncertainty”. Minmax (or similar strategies) will not work, because of unboundedness. (references to the strategies of uncertainty—perhaps again?) “Forcing” oneself to use a specific probability value will lead to grief: it is useless to protect oneself against the risk of a certain outcomes when you really have no reason to give it *any* specific probability.

Note in particular that the well-known device of taking “safety margins” will not work. Suppose that one is willing to take a one-in-a-million risk of bankruptcy, but – in order to “hedge” one’s bets – only makes trades that (according to his or her calculations) have a one-in-a-trillion chance of going so badly as to lead into bankruptcy. Will taking such a ludicrous “safety margin” – a factor of 1,000,000 – help the risk manager avoid bankruptcy in such situations?

The answer is no. Taking such “safety measures” is a reasonable device if one *knows* that the generator is of one of the “benign” types, e.g. type 1 or 2, and therefore one *knows* that one *is* justified in making assumptions about the probabilities of events happening in the future using the observed parameters as approximations for the actual parameters of the generator, but might not be completely sure about the *exact values* the parameters should have. In other words, this would work in cases where one knows one can safely describe the situation as one of decision making under risk, although one is not sure exactly what risk.

In a situation where the generator might be of type 3 or 4, however, one doesn’t simply have a *vague* idea of what the risk is; one has *no* idea what it is, and cannot assign *any* value to it. Taking only “trillion-to-1” bets against bankruptcy is worthless in such a situation since the assessment of the risk of a certain trade *as* trillion-to-1 is worthless in the first place. There is no ‘there’ there: the calculated “million to one safety margin” doesn’t correspond to anything in reality.

We have no real base to give credence to this estimation; the relaxing number “a trillion to 1” has only psychological significance in such a situation – as the occurrence of the “impossible”  $10\text{-}\sigma$  event in the case of “Long Term Capital” shows. It is not as if a  $10\text{-}\sigma$  event actually occurred. Rather, the belief that it *is* a  $10\text{-}\sigma$  event was based on the *unjustified conclusion* that the generator involved is of the benign type in the first place.

Therefore, the risk managers did not consider the possibility of the generator being of the third or fourth type, where events that *would be*  $10\text{-}\sigma$  events *if* the generator were of the benign type, actually occur far more frequently.

Our only recourse in such situations is Popper’s solution: to wait for the “black swan”, and make sure that we are not destroyed by it.

## SUMMARY

In this paper, we have tried to show the essential problem of risk management is forcing situations of decision making under uncertainty into the straightjacket of decision making under risk.

We showed this in a few steps: four steps:

First, we showed that certain random generators have a “bad” relation between their observed moments and their actual moments. This is a purely mathematical issue.

Second, we have shown if one’s background information satisfies certain conditions, then *if* such generators are not ruled out, the mere possibility that they are the generator one is dealing with sabotages any attempt to assign specific values to the “real” moments of the generator, due to the “black swan” problem – the possibility of rare extreme events which have a large influence on the moments. This is an epistemological issue.

Third, this forces us to conclude we that we are in a situation of *decision making under uncertainty*. This is a decision-theoretic matter.

Fourth, we showed that, in fact, the situation risk managers deal with are precisely those where such generators cannot be ruled out. This is a scientific issue: it has to do with the different nature of physical and economic systems.

Fifth, closely related to the third issue, , we showed that common “avoidance” procedures – taking only what seems like “very low” risks – will not work, since the implicitly assume the situation is one of decision making under risk in the first place. Even “usually” procedures for decision making under uncertainty – minmax, minmax of regret, etc. – will not work, since the “bad” generators are not bound.

Finally, we show that in such situation, the only thing we can do is protect ourselves against the black swan –and recognize that we may not know much about it. This is the only strategy applicable to this situation.

## **Appendix: Solving the Inductive Problems for the Generator: The Technical Details**

In the previous chapter, we have claimed that the correct inductive inference about the generator's moments—in particular, of the first moment,  $E(X)$ —changes depending what we know about the type of the generator.

In some cases, it is reasonable to give an estimate of it; in others, no matter how much information we have about previous outcomes, this cannot be done.

It is now the time to fill in this debt—to show how, indeed, induction works in these cases.

### **TYPE 1 GENERATORS: FORMAL TREATMENT**

Let us put things more formally, using Levi's notation (Levi, 1980, and also below). To simplify things, let us fix  $a$  and  $b$  as 1 and 6, and first consider a bounded generator (Type 1) with a finite number of outcomes—say a tossed die with outcomes  $\{1, 2, 3, 4, 5, 6\}$ . John, at time  $t_0$ , has to make a decision about the properties of this random generator. What can we say about this situation, epistemically?

### **THE CORPUS OF KNOWLEDGE: BACKGROUND INFORMATION AND EXPERIMENTAL DATA**

First of all, John has a *corpus of knowledge* (or belief),  $K_{\text{John},t_0}$ . It includes the following information:

- 1) *Background information* John knows about the generator. As the angel said to John,  $K_{\text{John},t_0}$  includes:

The outcomes of the dice throws are governed by a random generator defined by a probability function  $X: \{1,2,3,4,5,6\} \rightarrow [0,1]$ .

The outcomes of the generators are always one of the set  $\{1,2,3,4,5,6\}$ .

The generator's mean ( $E(X)$ ), variance ( $\text{Var}(X)$ ), and higher-level moments are fixed, both in

the past and in the future.

John knows the laws of statistics, methods of statistical inference, and so on, e.g., the central limit theorem, etc.

John's corpus of knowledge  $K_{\text{John},t_0}$  also includes the *outcomes of the previous trials* up to time  $t_0$ :

The first toss of the die had outcome  $x_1 \in \{1,2,3,4,5,6\}$ .

The second toss of the die had outcome  $x_2 \in \{1,2,3,4,5,6\}$ .

...

....

n) The nth toss of the die (the last one before time  $t_0$ ) was  $x_n \in \{1,2,3,4,5,6\}$ .

We also assume something else of significant importance: that  $n$  is large enough for us to use the *normal approximation* for  $E(X_n)$ . We shall see the importance of this later.

The result of 1-n above and John's knowledge of statistics is that, of course, John has *estimates* of the first, second, and higher moments in his corpus:

The estimated first moment of  $X$  given the first  $n$  tosses ( $E(X_n)$ ) is  $(\sum_i x_i)/n$ . Note that this itself is a random variable, dependant on both the properties of  $X$  and on  $n$ .

The estimated second moment given the first  $n$  tosses (Estimated variance, or  $\text{Var}(X_n)$ ) is the square of the sample's standard error, or  $(\sum_i (x_i - E(X_n))^2)/(n-1)$ .

... and so on for higher-level moments.

Finally, John's corpus of belief includes (by definition, as seen below) *all the deductive consequences* of the above information. In particular, that  $1 \leq E(X) \leq 6$ ,  $0 \leq \text{Var}(X) \leq 25$  ( $= (6-1)^2$ ) (actually less, but we can afford to be generous here), etc.

However, John's corpus does *not* limit where  $E(X)$  can be *deductively* any more than that. It is not logically follow from the outcomes and the background information that  $E(X)$  is more specific than  $[1,6]$ .

John is engaged, at time  $t_0$ , in solving a decision problem: given the information above in  $K_{\text{John},t_0}$ , can he give a reliable estimate of the moments of the generator  $X$ —and thus, of its future behavior? To simplify, once more, we shall consider only the case of John estimating the

first moment,  $E(X)$ .

### THE DECISION PROBLEM: THE OPTIONS

To repeat, giving a reliable estimate of  $E(X)$  is another name for saying that John is *justified to infer* that  $E(X)$  is of a certain value—that it is a legitimate inductive inference. This is a decision problem; we need to first consider *what options* for inductive inference exist—that is, between what estimates of  $E(X)$  John *can* choose; then, to decide which one (if any) of those John *should* choose.

What are the options available? This depends both on what is deductively excluded by  $K_{\text{John},t0}$  and the goals that interest John. In this case, we know that:

- 1)  $K_{\text{John},t0} \vdash 1 \leq E(X) \leq 6$ . Whatever value John chooses as his estimate of  $E(X)$ , it must be between 1 and 6 on pain of logical inconsistency.

From the statistics in  $K_{\text{John},t0}$  one knows that the estimate “ $E(X)=E(X_n)$ ” is the only one that is free from an built-in bias.

Now, we *can* limit the number of options John considers accepting or rejecting to a finite number (even to two). For a fixed  $\epsilon_0$ , we can consider the two options as whether  $|E(X)-E(X_n)| < \epsilon_0$  or not ( $H_0$ ). On this view, there are four options altogether: to accept that  $E(X)$  is at most  $\epsilon_0$  from the observed  $E(X_n)$ , to accept that  $E(X)$  is  $\epsilon_0$  or more from the observed  $E(X_n)$ , to accept both (which means that John decides to add information to  $K_{\text{John},t0}$  that makes his beliefs inconsistent, by adding  $H_0 \wedge \sim H_0$ ) and to accept neither (that is, to add nothing to  $K_{\text{John},t0}$ , by “adding” the tautology  $H_0 \vee \sim H_0$ )

However, there is no need to *a priori* limit the number of possible options. There is a natural set of potential basic options, mutually exclusive and exhaustive (as they must be—see Levi, 1980), that are the most specific possible: namely the set  $\{U_x =_{\text{def}} “E(X)=x” \mid 1 \leq x \leq 6\}$ .

In this case, John has a total number of  $2^8$  options: those that are defined by any sort of (measurable) subset of  $[1,6]$ . For example, John might decide that the strongest claim that he accepts is that  $E(X)$  is between  $\frac{1}{2}$  and 1 or between 4 and 5; that is, John accepts the infinite

disjunction  $(\bigvee_{0.5 < j < 1} U_j) \vee (\bigvee_{4 < j < 5} U_j)$  as true, but does not accept anything more specific. Note that the previous “basic” option  $H_0$  is a non-basic one, the disjunction  $\bigvee_{E(X_n) - \varepsilon < j < E(X_n) + \varepsilon} U_j$ .

In particular, John still has the weakest option—accept only the disjunction  $\bigvee_{1 \leq j \leq 6} U_j$ , that is, that  $1 \leq E(X) \leq 6$ , which is already in  $K_{\text{John}, t_0}$  and therefore a null addition; and there is a single strongest option—accepting the disjunction  $\bigvee_{\emptyset}$ , that is, to accept that *none* of the basic hypotheses  $U_j$  are true. This means to accept that  $E(X) \notin [1, 6]$ , in contradiction with information already in  $K_{\text{John}, t_0}$  that it is; that is, the strongest option is to add a contradiction.

## THE DECISION PROBLEM: RISK OF ERROR

The next issue to consider in the decision problem is the risk of error by accepting any of the options, and, in particular, the basic options. The *risk of error*, from the agent’s point of view, is the *probability that it is wrong*.

Since we are dealing with the infinite case, we must deal not with probability itself (for every basic option,  $p(U_j) = p(E(X) = \text{exactly } j)$  is 0), but with the *density function*,  $f$ , which in turn determines the probability for any measurable set. Can John estimate this density function? The laws of statistics tell us that John can do so.

The calculations themselves can be found in any statistics textbook. Here is a short sketch: for a “large enough”  $n$  ( $n > 30$ ), the random variable  $X_n^* =_{\text{def}} (\sum_{i=1}^n X_i) / n$  behaves roughly like a normal variable (due to the central limit theorem) with mean  $E(X)$  and standard deviation of  $\sigma_X / \sqrt{n}$ . We do not know what  $\sigma_X$  itself is, of course (the generator’s moments are hidden from us) but, since  $\sigma_X$  is bounded from above—if by nothing else, then by  $\text{sqr}((6-1)^2) = 5$ , in this case—there is a known upper limit to the standard deviation of  $X_n^*$  is for every  $n$ . So, for every  $n$ , the laws of statistics tell John that he can assume that  $X_n^*$ ’s density function is roughly that of a normal random variable with a mean  $E(X)$  and (maximal) standard deviation of (in this case)  $5 / (\sqrt{n})$ .

(For smaller  $n$ , one needs to use Gossett’s “Student- $t$ ” distribution, but we can assume  $n$  is large enough. Also that the normal approximation of  $X_n^*$  is unbounded—it can go to, say,  $-1000$  or  $+1,000,000$ —while the “real”  $X_n^*$  is the observed average of  $n$  die tosses, and must be bound between 1 and 6; but, again, for a “large enough”  $n$  the tails will be so close to 0 as to make no difference. Finally, one can estimate  $\sigma_X$  by using  $s = \text{sqr}[(\sum_{j=1}^n (X_j - E(X_n))^2) / (n-1)]$ , the sample’s standard error, which would usually be much smaller than 5; but we can afford to take the “worse case scenario” here.)

What, then, are the *allowable* probability functions,  $Q_{\text{John}, t_0}$ , John can consider (for a given  $n$ ) as possibly representing the actual density function of the probability of the real  $E(X)$  being at a

certain point around the observed  $E(X_n)$ ? It is a family of normal distributions with mean  $E(X_n)$  and maximal variance  $5/\sqrt{n}$ . So, the density functions are:

*Allowable density functions* for John at time  $t_0$ :  $Q_{\text{John},t_0} = \{f_v \equiv N(E(X_n), v) \mid 0 < v < 5/\sqrt{n}\}$ .

Note that the agent can use the laws of statistics to reach conclusions about the probabilities partially because the original random variable  $X$  describing the generator does not change with time. Therefore, the risk of error John takes (given a fixed density function  $f$  and  $n$ ) if John accepts the infinite disjunction  $(\bigvee_{0.5 < j < 1} U_j) \vee (\bigvee_{4 < j < 5} U_j)$  as true (that is, adds it to  $K_{\text{John},t_0}$ ) but does not accept anything more specific, is  $1 - (\int_{[0.5,1]} f(x) dx + \int_{[3,4]} f(x) dx)$ , that is, 1-the probability of it being the case that  $E(X)$  is in that range.

## THE DECISION PROBLEM: INFORMATIONAL VALUE

Now we come to *informational value*. What informational value should be assigned to every  $U_i$ ?

According to it (Levi 1980) the informational value of an hypothesis,  $\text{Cont}(H)$ , is inversely correlated with a probability function,  $M(H)$ : the higher the “probability” of an hypothesis, the less information it carries. This must be so, if we want certain basic properties of information value to hold: say, that the informational value of a tautology is the minimal possible one, or that the  $\text{Cont}(A \vee B) \leq \text{Cont}(A)$ ,  $\text{Cont}(B) \leq \text{Cont}(A \wedge B)$ .

Note that:

$M$  is *not* the same as the probability function that the agent assigns to the hypothesis being true, unlike what Popper (1950) and others believed. On the view advocated by Peirce, James, and Levi, informational value is not merely a way to say that something is improbable; probability and informational value are distinct characteristics.

The inverse proportion between  $M(H)$  and  $\text{Cont}(H)$  can take several forms: say,  $\text{Cont}(H) =_{\text{def}} 1/M(H)$ ,  $\text{Cont}(H) =_{\text{def}} -(\log(M(H)))$ , etc. Levi prefers the simple  $\text{Cont}(H) = 1 - M(H)$ , for reasons not crucial to this discussion (for the record, in this way his version of information content mimics in certain respects Savage’s “degrees of surprise”, see Savage (1953), Levi (1980).)

There is here a natural suggestion: that every  $U_i$  have an *equal* informational value: it is precisely as informative, or as specific, to say that  $E(X)$  is 0.453 as it is to say that it is 0.991. That means that the M-function, as well, must be “the same” for every  $U_i$ . Since we are dealing with the infinite case, any M-function would give probability 0 to every  $U_i$ , so we need to look at the density function: we wish the density function  $m$  of the M-function to be the constant one. In this case, we have  $m=0.2$  over  $[1,6]$ .

On this view, the informational value of every hypotheses  $H$  is  $1-M(H)$ , that is,  $1-0.2*(H$ 's measure). For example, if John accepts the infinite disjunction  $H=(\bigvee_{0.5<j<1}U_j)\vee(\bigvee_{4<j<5}U_j)$  as true (that is, adds it to  $K_{John,t0}$ ) but does not accept anything more specific, John gains informational value of  $Cont(H) = 1-M(H) = 1-(\int_{[0.5,1]}0.2dx+\int_{[3,4]}0.2dx)$ .

To illustrate, here is a graph of the  $m$ -function and a few of the potential density functions:

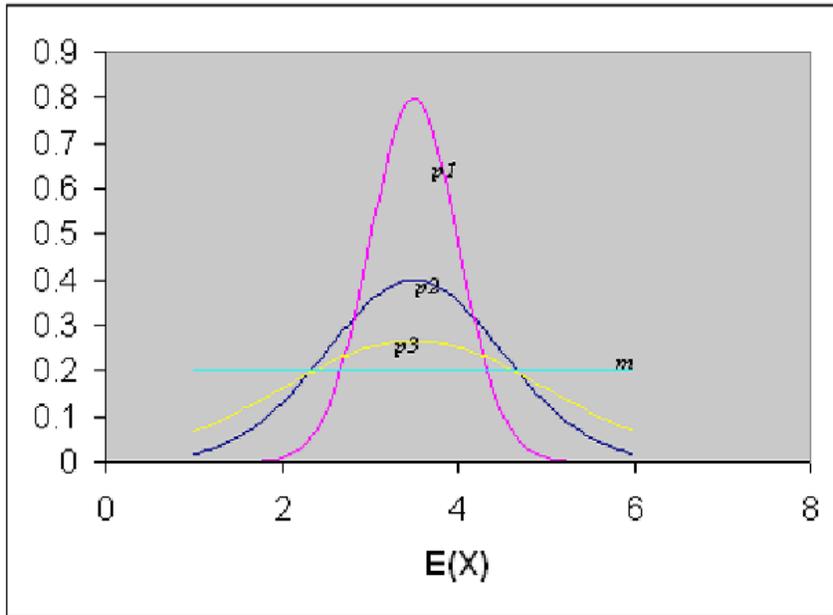


Figure 1: The agent's  $m$ - and  $p$ -functions

## THE DECISION PROBLEM: THE OPTIMAL INDUCTIVE STRATEGY

## THE DECISION PROBLEM: STAGE 1: THE FORMULAS

As always, we follow Levi(1980). Levi recommends to accept an hypothesis where the information value (defined by  $\text{Cont}(H)$ , etc.) is big enough to justify the risk of error (defined by  $p(H)$ , etc.)

How does one determine what is a “small enough” risk of error or a “large enough” informational value? Levi (1980) concludes that the way to go is as follows:

*Rejection Rule:* if  $U_i$  is a basic option,  $p(U_i)$  is the credal probability (e.g., the probability the agent assigns to  $U_i$  being true) of  $U_i$ , and  $M(U_i)$  the probability function determining its informational value  $\text{Cont}(U_i) =_{\text{def}} 1 - M(U_i)$ , the agent should *reject*  $U_i$  (e.g., *add*  $\sim U_i$  to their corpus of knowledge) if and only if  $p(U_i) < qM(U_i)$ , where  $0 < q < 1$  is the agent's “boldness index”.

Let us consider this for a moment. To *accept* hypothesis  $U_i$  is the same thing as *rejecting*  $\sim U_i$ ; and  $\text{Cont}(U_i) = 1 - M(U_i) = M(\sim U_i)$ . For a fixed  $p$  function and fixed  $q$ , the higher the informational value  $\text{Cont}(U_i)$ , the higher  $M(\sim U_i)$ , and the more likely that  $\sim U_i$  will be rejected—that is,  $U_i$  accepted. That is, the higher the informational value of  $U_i$ , then—*ceteris paribus*—the more likely it is to be accepted.

Similarly, On the other hand, for a fixed  $\text{Cont}(U_i)$  and  $q$ , the higher  $p(U_i)$ , the lower  $p(\sim U_i) = 1 - p(U_i)$ . This means that it is more likely that  $p(\sim U_i)$  will be lower than  $qM(\sim U_i)$ ; that is,  $\sim U_i$  will be rejected, or  $U_i$  accepted. The more probable  $U_i$ , the more likely it is (*ceteris paribus*) to be accepted.

Now, what is  $q$ ? This depends on the agent and the situation. For a fixed  $p$  and  $M$  functions, the higher  $q$  is, the more options are rejected, and the smaller (and more specific) the number of remaining options. The agent is therefore bolder in accepting the risk of error for information. The lower  $q$  is, the less options are rejected, and the larger (and less specific) the number of remaining options.

There is no *a priori* reason to fix  $q$  at a specific number. However, as Levi shows,  $q$  should never be 0 (let alone below), since this would mean the agent might hesitate and not accept options even if they carry *no* risk of error (e.g., they have probability =1). And  $q$  should never be 1 (or above), since that would mean the agent might accept to their beliefs options that carry a risk of error *for sure* (e.g., have probability = 0).

In the infinite case, as in here, one cannot use the probability functions themselves, since for

every basic option  $U_j$ ,  $p(U_j) = M(U_j) = 0$ , and therefore for every  $q$  the inequality does not hold (it is  $0 < 0$ ). The natural extrapolation (see also Levi, 1980, esp. 5.10, 5.11) is, in this case, to consider the density functions: to reject  $U_j$  for  $1 \leq j \leq 6$  if and only if  $f(j) < qm(j)$ , that is, if and only if  $f(j) < 0.2q$ .

This means that, for a specific  $q$  and  $f$ , there is a “cutoff point  $\varepsilon_0$ , where  $f(E(X_n) - \varepsilon_0) = f(E(X_n) + \varepsilon_0) = 0.2q$ ; John should reject the tails re the value of  $f$  is below  $0.2q$ , that is, the agent adds the information that the value of  $E(X)$  is between  $E(X_n) - \varepsilon_0$  to  $E(X_n) + \varepsilon_0$  to  $K_{\text{John}, t_0}$ .

## THE DECISION PROBLEM, STAGE 2: E-ADMISSIBILITY AND SUSPENDING JUDGMENT

Things, however are not that simple, for two reasons: first, John has more than one possible density function, and they do not always give the same recommendation. Second, once it is decided by John what he should add to his belief given a specific density function, the question is: which one of those to actually recommend?

An option that is recommended by a specific probability function the agent considers legitimate is called by Levi an *E-admissible* option. In this case, the set of E-admissible options for John are:

{Add to  $K_{\text{John}, t_0}$  that  $(E(X_n) - \varepsilon(f) \leq E(X) \leq E(X_n) + \varepsilon(f))$  for every  $f \in Q_{\text{John}, t_0}$ ,  $\varepsilon(f) =_{\text{def}}$  distance from  $E(X_n)$  where  $f(\varepsilon(f)) = 0.2q$ }

It is easy to see that this set is a set of stronger and stronger options, depending on what the variance of the allowable density function is, since the set of density functions is the normal density functions with mean  $E(X_n)$  and standard deviation from 0 to  $5/\sqrt{n}$ , as said above. This means that if  $f$  is a “spread out” function (with a relatively high variance),  $\varepsilon(f)$  is relatively large and John only accepts, given that  $f$ , that the true value of  $E(X)$  is between relatively far apart  $E(X_n) - \varepsilon(f)$  and  $E(X_n) + \varepsilon(f)$ . if  $f$  is a “concentrated” function (with a low variance),  $\varepsilon(f)$  is correspondingly smaller and John accepts a stronger claim—that the real value of  $E(X)$  is within a narrower range.

So much for the E-admissible options. Which one to choose? Levi suggests (Levi, 1980) a *rule for ties*:

*Rule for ties*: If an agent has two E-admissible options  $E_1$  and  $E_2$ , and it is reasonable to *suspend judgment* between them (accept  $E_1 \vee E_2$ )—that is, in particular, that  $E_1 \vee E_2$  is itself E-admissible—then one should choose the E-admissible  $E_1 \vee E_2$  over either the E-admissible  $E_1$  or the E-admissible  $E_2$ .

In this case, all the possible options are arranged by logical strength from the weakest (accept only that  $E(X)$  is between  $E(X_n)-\varepsilon$  to  $E(X_n)+\varepsilon$  when  $\varepsilon$  is when the density function  $N(E(X_n), 5/\sqrt{n})=0.2q$ ) to the strongest (accept that  $E(X_n)=E(X)$  exactly; that is, to consider the limit case where the normal distribution has variance 0). Of any two options, one implies the other, so that their disjunction is simply the weaker option. The rule for tie tells us to take the total disjunction—in this case, the *weakest* possibility. So, in sum, John accepts that:

**John's Acceptance, stage 1:** Adds to  $K_{John,t0}$  the fact that  $E(X_n)$  is between  $E(X_n)-\varepsilon$  and  $E(X_n)+\varepsilon$  when  $\varepsilon$  is when the density function  $N(E(X_n), 5/\sqrt{n})=0.2q$ .

### THE DECISION PROBLEM, STAGE 3: ITERATION

However, we are still not done. Now that John accepted certain claims to be true, says Levi, John needs to *iterate* the inductive inference. John's new  $K$ ,  $K_{John,t1}$ , is the deductive closure of  $K_{John,t0}$  and the disjunction  $\bigvee_{x|E(X_n)-\varepsilon \leq x \leq E(X_n)+\varepsilon} ("E(X)=x")$ . Or, in Levi's symbolism, John *expanded* his corpus to a larger one, holding more beliefs. In Levi's symbolism, if  $H_1 =_{\text{def}} \bigvee_{x|E(X_n)-\varepsilon \leq x \leq E(X_n)+\varepsilon} ("E(X)=x")$ :

$$K_{John,t1} = K_{John,t0} \overset{+}{H_1}.$$

John's probability functions, in  $Q_{John,t0}$ , also change: also change: they are now the set of *conditional* probabilities, given that John added the disjunction that  $E(X)$  is between  $E(X_n)-\varepsilon$  and  $E(X_n)+\varepsilon$  to his beliefs. (Levi calls this the *conditionalization commitment*. See Levi, 1980.) That is, John's new probability functions at time  $t_1$ , is:

$$Q_{John,t1} = \{p \mid p(x) = q(x|H_1), \text{ for every } q \in Q_{John,t0}\}$$

The informational value function also changes. It becomes determined by the conditional, new M-function, which is 0 outside  $[E(X_n) - \varepsilon_0, E(X_n) + \varepsilon_0]$  and  $1/2\varepsilon_0$  inside this interval.

John now has a *stage 2 decision problem*: which, if any, of the options  $U_x = "E(X)=x"$ , for  $x \in [E(X_n) - \varepsilon_0, E(X_n) + \varepsilon_0]$ , with these new probability and content functions, should he reject?

As before, one does the calculations and sees that one should reject just those  $U_x$ 's where the weakest conditional density function,  $N(E(X_n), 5/\sqrt{n})$  given that  $x$  is between  $E(X_n) - \varepsilon$  and  $E(X_n) + \varepsilon$ , is below  $qm(x)$ —that is,  $q(1/2\varepsilon_0)$ .

Possibly some more hypotheses will get rejected. If there are some, then John needs to yet again add more information to his beliefs—add to  $K_{John,t1}$  the fact that  $E(X)$  is *not* farther away from  $E(X_n)$  than some  $\varepsilon'$ , ( $0 < \varepsilon' < \varepsilon$ ). Then, John needs once more iterate—conditionalize  $Q_{John,t2}$  based on  $Q_{John,t1}$  given the new rejections, make  $M$  defined by the new  $m \equiv 1/2\varepsilon'$ , and so on.

This process continues indefinitely. John solves a series of decision problems given  $K_{John,t0}, K_{John,t1}, K_{John,t2}, \dots$  each saying that  $E(X)$  is at most  $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \dots$  away from  $E(X_n)$ , with the conditional  $Q_{John,t0}, Q_{John,t1}, Q_{John,t2}, \dots$ , each based on the previous one and the new information added, with the new  $m$  density function being  $1/5$  (the original one),  $1/2\varepsilon, 1/2\varepsilon', 1/2\varepsilon'', 1/2\varepsilon''' \dots$ , etc.

It can be shown that eventually—and perhaps even the first time—John will reach a certain  $K_{John,t*}, Q_{John,t*}$ , with the strongest claim in  $K_{John,t*}$  being that  $E(X)$  is at most  $0 < \varepsilon^*$  away from  $E(X_n)$ ,  $m$  being  $1/2\varepsilon^*$ , where *the recommendation is not to reject any more hypotheses*. John, as it were, rejected all the he could reasonably reject.

## JOHN'S FINAL DECISION

The final recommendation—the strongest—is:

**John's Acceptance, stage 1:** Adds to  $K_{John,t0}$  the fact that  $E(X_n)$  is between  $E(X_n) - \varepsilon^*$  and  $E(X_n) + \varepsilon^*$  when  $0 < \varepsilon^* \leq \varepsilon$ ,  $\varepsilon$  being the value where John's original density function, the (unconditional)  $N(E(X_n), 5/\sqrt{n}) = 0.2q$ .

## DISCUSSION

The result of the inductive decision problem is "John's acceptance", above. That is, induction recommends that John, in this situation, and for a given  $n$  and  $q$ , accept that  $E(X)$  is in the range described by "John's Final Decision".

In practice, this means two things:

Unless  $q$  is very small, then for any  $n$  that is not too small (say,  $n \approx 30$  or so, or higher, as we assume) the range that John accepts as possible value for  $E(X)$  is relatively small, even if one uses the maximum possible estimation of  $X_n^*$ 's standard deviation, that is,  $5/\sqrt{n}$ .

If one uses the standard estimation of  $X_n^*$ 's standard deviation (the standard error), then  $\epsilon$ , even after only one iteration, will be even smaller, since the weakest (most spread out) density function John considers in the first case will be  $N(E(X_n), s/\sqrt{n})$ , with  $s$  the standard error, not  $N(E(X_n), 5/\sqrt{n})$ , and  $s < 5$ —and thus  $N(E(X_n), s/\sqrt{n})$  would reach  $0.2q$  faster (closer to  $E(X_n)$ ).

Successive iterations of the decision problem might lead the agent to reject even more hypotheses, eventually settling on the claim that  $E(X)$  is in  $[E(X_n) - \epsilon^*, E(X_n) + \epsilon^*]$ , with  $0 < \epsilon^* \leq \epsilon$ .

(2) and (3), in this case, are almost unnecessary, however. For a reasonably large  $n$ —one large enough to use the normal approximation for  $X_n^*$ —even doing *only one iteration* of the decision process and using the *maximal possible size* of  $X_n^*$ 's standard deviation would usually significantly limit what is accepted.

In short, So when one has a type 1 generator, John *can* tell, pretty quickly, quite a bit about the value of the generator's moment,  $E(X)$ . John is *justified in inductively accepting* that it is within a range,  $\epsilon$ , that is small to begin with in most circumstances (as 1 above says) and gets smaller quickly as the number of observations increases.

Note, also, an important point. First, obviously information about the previous outcomes of the generator is essential for the agent to reach the conclusion. But the law of statistics could only be used *because* we have background information about the type of generator we have here—a “well-behaved” one.

## TYPE 2 GENERATORS: NORMAL DISTRIBUTION

### BACKGROUND INFORMATION

Type 2 generators are similar to type 1 generators, as we shall see, with a few difference. Again, to fix the discussion, let us presume that the generator is normal, with (actual) moments  $E(X)$ ,  $\text{Var}(X)$ , etc. As before, let us assume that John is trying to estimate the first moment, or

what  $E(X)$  is.

The background information is very similar to the one with the case of the bounded distribution, of course with the change that John knows that the generator is normal, not bounded. In particular, John knows that  $E(X)$  and  $\text{Var}(X)$  are fixed and will remain so in the future, and the laws of statistics. John also knows, due to these laws, that the (same) estimates ( $E(X_n)$ ,  $\text{Var}(X_n)$ ) are the only ones of the generator's moments that do not have a built-in bias.

When it comes to the data, John knows what the past outcomes ( $x_1, \dots, x_n$ ) of the generator were. As before, let us consider the first moment  $E(X)$  and John's estimation of it.

### **DIFFERENCES FROM BOUNDED DISTRIBUTION—AND WHY IT DOESN'T MATTER IN THIS CASE**

There are two things that can ruin it for John. In the bounded case, there were *no extreme events*, first, and  $\sigma_X$  was *bounded from above by a known quantity*. In the normal case, it *could* be that an extreme event would be observed in  $x_1, \dots, x_n$ , and significantly “throw off”  $E(X_n)$ . Or, if  $\sigma_X$  is extremely large, it might take a very large  $n$  to get  $E(X_n)$  to converge to  $E(X)$ . In both cases, even for a large  $n$ ,  $E(X_n)$  could still be significantly different from  $E(X)$ .

Consider, however, what we are trying to achieve in the first place. We are *not* claiming that *all* “well behaved” generators—e.g., all normal distributions—can be easily “worked on” in practice, no matter what their properties or what the outcomes in the past happened to be. If the normal distribution has a very large variance, it will indeed take a lot of time for that pattern to emerge. If an extreme  $10\text{-}\sigma$  event *dis* in fact occur, the estimate  $E(X_n)$  will be off from  $E(X)$  for a while.

But such occurrences are *observable*: John will see them occurring in the outcomes, and know to be care in reaching conclusions about the future. Our problem is not with the “bad” generators (large  $\sigma_X$ ) or “bad” outcomes ( $10\text{-}\sigma$  events) that wear their “badness” on their sleeves, that is, in the outcomes already observed. We are concerned here with exactly the opposite: what we can say about a generator when it is assumed that the outcomes *do* look good—that is, when  $\sigma_X$  is small and no extreme events occurred in the past.

So we can assume that the outcomes *do* “look good”: that  $\sigma_X$  is relatively small and that no  $10\text{-}\sigma$  events occurred. We want to know: *given these outcomes, what can the agent deduce, if anything, about the qualities of the generator?* In this case, quite a lot.

As above, the random variable  $X_n^*$  behaves normally, with random variable  $X_n^* =_{\text{def}} (\sum_{j=1}^n x_j)/n$  behaves roughly like a normal variable (due to the central limit theorem) with mean  $E(X)$  and standard deviation of  $\sigma_X/\sqrt{n}$ . We do not know what  $\sigma_X$  itself is, of course (the generator's

moments are hidden from us) but we can estimate it, since one can estimate  $\sigma_X$  by using  $s = \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1))]$ , the sample's standard error. This means that we can assume (presuming, again, that  $n$  is "large enough" as above) that the density function of  $X_n^*$  is  $N(E(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})$ . This is a Normal distribution that, as  $n$  increases, becomes more and more "centralized" since its  $\sigma \rightarrow 0$  as fast as  $1/\sqrt{n}$ .

In this case, then, John has *one* probability function that determines how probable it is that the actual  $E(X)$  is within a certain range of the observed  $E(X_n)$ :

*John's density function,  $f$ :  $N(E_n(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})$ .*

(Of course, we could have used this technique to minimize the set of allowable probability functions in the bounded case, as well. But we deliberately did not, to show that even if we *do* allow many probability functions that create a "worst-case scenario" in the bounded situation, John can *still* tell us much about the generator's moments. It also gave us a way to illustrate the rule for ties and E-admissability and to the iteration process, which will be important later on.)

## INFORMATIONAL VALUE: SOME COMPLICATIONS

Assigning an M-function (and therefore an informational value function) is a bit more complicated this time.  $M$ 's density function cannot be the constant function  $m$  whose integral over the possible range— $(-\infty, +\infty)$ —is equally to 1, since there is no such function (the integral is 0 for  $m=0$  and diverges otherwise). There is, simply put, no way for an agent to assign "equal informational value" to " $E(X)=x$ " for every  $x \in \mathbb{R}$  and still have the informational value be based on a probability function.

What, then, should  $M$  be? There are several possibilities. The one we use—due to our concern with "extreme events"—is as follows. Consider some large  $L_0$ , and the range  $[E(X_n)-L_0, E(X_n)+L_0]$ . There is an infinite number of hypotheses of the value of  $E(X)$  within this range (namely,  $U_x =_{\text{def}} "E(X)=x"$  for every  $x \in [E(X_n)-L_0, E(X_n)+L_0]$ , and two additional hypothesis:  $U =_{\text{def}} "E(X) < E(X_n) - L_0"$ , and  $U^+ =_{\text{def}} "E(X_n) + L_0 < E(X)"$ . The M-function that determines the content function will give both of these hypotheses some the hypothesis  $U^*$  some probability,

$p^-$  and  $p^+$ ; we can assume they are the same,  $p_0$ .

We can be careful and assume that, first,  $L_0$  is large (relative to the standard error of the sample,  $s$ )—say,  $10s$  in length; the reason is that we want these hypotheses to represent extreme possible values of  $E(X)$ . We also assume that  $U^-$  and  $U^+$  are very informative—that is, that  $p_0$  is very small. Within  $[E(X_n)-L_0, E(X_n)+L_0]$ , we assume that  $M$  is determined by the usual, fixed density function  $m$ ; only this time its integral of the  $2L_0$  interval isn't 1, but  $1-sp_0$ . So John's  $M$ -function is defined as:

$$M(U^+) = M(U^-) = p_0; m \equiv (1-2p_0)/2L_0 \text{ over the range } [E(X_n)-L_0, E(X_n)+L_0].$$

### THE DECISION PROBLEM

As usual, the agent should *reject an hypothesis  $U$  if and only if  $p(U) < qM(U)$* —or, in the case of point hypotheses, use the density functions of  $p$  and  $M$ , respectively: *reject the hypothesis  $U$  if and only if  $f(U) < qm(U)$* . On this view, we have:

Reject  $U^-$  if and only if  $p(U^-) < qM(U^-)$ : reject  $U^-$  if and only if  $\int_{-\infty}^{E(X_n)-L_0} [N(E_n(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})] dz < qp_0$ .

Reject  $U^+$  if and only if  $p(U^+) < qM(U^+)$ : reject  $U^+$  if and only if  $\int_{E(X_n)+L_0}^{+\infty} [N(E_n(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})] dz < qp_0$ .

Reject  $U_x$  for  $x \in [E(X_n)-L_0, E(X_n)+L_0]$  if and only if  $f(U_x) < qm(U_x)$ , that is, if and only if the value of the normal curve,  $N(E_n(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})] < q(1-2p_0)/2L_0$ .

Let us consider the possibilities. Suppose as above that  $L_0$  is large and that  $p_0$  is small. Nevertheless, unless  $p_0$  or  $q$  are very small indeed, the  $\int_{-\infty}^{E(X_n)-L_0} [N(E_n(X), \text{sqr}[(\sum_{j=1}^n (x_j - E(X_n))^2 / (n-1)] / \sqrt{n})] dz$  is going to be far smaller than  $p_0q$ , since it is the “tail end” of a normal distribution that is many standard deviations away from the mean. So both  $U^-$  and  $U^+$  will be rejected.

Now consider the middle case (3). What we have here is precisely the same situation as in the “bounded” case—with the small difference that the  $m$ -function is somewhat smaller than the  $m$ -function in the bounded case over the same range, since  $m \equiv (1-sp_0)/2L_0$  and not simply  $1/2L_0$ , for  $m$  must account for the possibility of  $U^-$  and  $U^+$ .

We know how to solve this problem. In fact, it is even easier, since we have a fixed probability

function and not a set of such functions. Following the exact same steps as in the bounded case, we get that, after the first iteration:

**John's first step:** John should reject  $U^-$ , reject  $U^+$ , and those hypotheses " $E(X)=x$ " in the range  $[E(X_n)-L_0, E(X_n)+L_0]$  such that  $f(x) < qm(x)$ , or  $N(E_n(X), s/\sqrt{n}) < q(1-2p_0)/2L_0$ , when  $s$  is the standard error, that is,  $\text{sqr}[(\sum_{j=1 \text{ to } n} (x_j - E(X_n))^2 / (n-1))]$ . In other words, John should *accept into*  $K_{John, 10}$  the claim that  $E(X) \in [E(X_n) - \varepsilon, E(X_n) + \varepsilon]$ , when  $\varepsilon$  is where the density function  $N(E_n(X), s/\sqrt{n}) = q(1-2p_0)/2L_0$ .

As before, even in this first step, if  $n$  is large enough for  $X_n^*$  to use the normal approximation in the first place,  $\varepsilon$  will be small. And, in addition, for the same reasons as above, it might be that further iterations will allow John to reject even more hypotheses, and accept:

**John's Final Inductive Conclusion:** John should accept that  $E(X) \in [E(X_n) - \varepsilon^*, E(X_n) + \varepsilon^*]$ , when  $0 < \varepsilon^* \leq \varepsilon$ ,  $\varepsilon$  being the value where the (original) density function of the probabilities,  $N(E_n(X), s/\sqrt{n}) = q(1-2p_0)/2L_0$ .

We see, then, that even if the generator is unbounded, John can usually justifiably conclude that its  $E(X)$  is within a narrow range, as long as the number of observations is large enough to apply the usual laws of statistics (e.g., the assumption that  $X_n^*$  is normal). The mere fact that the generator's moment  $E(X)$  could be any value, including a very large one, does not require John to take that possibility seriously. And the same, as before, holds *mutatis mutandis* for higher-level moments of the generator.

### TYPE 3 GENERATORS: NORMAL + RARE "WEIRD" DISTRIBUTION

In this case, we are dealing with the case of a generator that is a combination of two other generators. The first is a "regular" normal generator with low  $E(X)$  and  $\text{Var}(X)$ . The second, a "wild" one with high  $E(X)$  and  $\text{Var}(X)$ ; what prevents the second from swamping the first is

that it only occurs rarely.

Once more, consider what John knows. From the background information, John knows what the outcome of the generator so far has been. John also knows the laws of statistics. Furthermore, John knows that the generator is of the form  $X = (1-p)X' + pX''$ , where  $E(X') < E(X'')$  and  $p < 1$ . But John does *not* know what  $p$  is, or what  $E(X')$ ,  $E(X'')$  are.

In addition, John knows that *no extreme events occurred*. That is, John knows that all the outcomes so far have been from  $X'$ . Can John estimate  $E(X)$ ?

The answer is negative. To estimate  $E(X)$ , the agent needs to do two things: 1) estimate  $p$ , given that no events from  $X''$  occurred, and 2) estimate  $E(X'')$ . While  $p$  *can* be estimated, in fact, the fact that we have no information about about  $E(X'')$  precludes more deliberate information.

How does John estimate  $p$ ? Let us ignore the values of the outcomes and consider a simplification: the outcome is either due to generator  $X'$  (with probability  $1-p$ ) or due to generator  $X''$  (with probability  $p$ ). To help out John, and simplify the calculation we will assume that he *knows* (by psychic means, perhaps) whether an outcome is from  $X'$  or  $X''$ . The question is: what is  $p$ ?

### STEP 1: EVALUATING $p$

John, here, has an obvious set of options ( $p$  from 0 to 1), with an obvious M-function (namely,  $m \equiv 1$ ). John has a set of outcomes of length  $n$  which we know produces the  $p$  event exactly 0 times. For every  $p$ , this means that the probability of this occurring is  $(1-p)^n$ .

Now, when do we reject a hypothesis? We reject the hypothesis  $U_x$  (" $p = x$ ") if and only if  $q(U_x) < qM(U_x)$ , or, in this case,  $(1-x)^n < q$ ; that is, John will fail to reject only such  $x$ 's such that  $(1-x)^n \geq q$ , or that  $1-x \geq q^{1/n}$ , or  $-x \geq q^{1/n} - 1$ , or  $x \leq 1 - q^{1/n}$ . That is, John accepts that the real  $p$  is in the range  $(0, 1 - q^{1/n}]$ ; as  $n$  increases, and  $q^{1/n} \rightarrow 1$  (since  $0 < q < 1$ ), this range becomes smaller and smaller.

### STEP 2: EVALUATING $E(X'')$

So far so good. However, John has no information at all about  $E(X'')$ , and therefore cannot limit  $E(X)$  in any way, even with this information about  $p$ .

The problem is this. Consider evaluating  $E(X'')$  given the outcomes,  $E(X_n)$ —or, more precisely,

$E(X_n')$ . First, what are the options John has? John is interested in is as before. John is interested in *whether or not the real  $E(X)$  ( $= (1-p)E(X') + pE(X)$ ) is close, or not close, to the observed  $E(X_n)$  ( $= E(X'_n)$ )*. This means that John can use the same options as before: namely, for a given  $L_0$  which is large in relation to the standard error of the sample, John has  $U^- = "E(X'') < E(X_n) - L_0"$ ;  $U^+ = "E(X'') > E(X_n) + L_0"$ , and  $U_x = "E(X'') = x"$  for  $E(X_n) - L_0 \leq x \leq E(X_n) + L_0$ .

John is interested in is as before. John is interested in *whether or not the real  $E(X'')$  is close, or not close, to the observed  $E(X_n)$* . This means that John can use the same options as before: namely, for a given  $L_0$  which is large in relation to the standard error of the sample, John has  $U^- = "E(X'') < E(X_n) - L_0"$ ;  $U^+ =$  have  $M(U^+) = M(U) = p_0$ ;  $m \equiv (1 - 2p_0) / 2L_0$  over the range  $[E(X_n) - L_0, E(X_n) + L_0]$ .

Consider, however, what the allowable probability functions about  $E(X)$  being in any range are. But John has *no* data at all—no observations—about  $X''$ , only about  $X'$ . So there is no way to evaluate  $E(X'')$ . To put it differently, since there are no observations, *any* probability density function from  $-\infty$  to  $+\infty$  is in John's  $Q_{\text{John}, t_0}$ . This, of course, is always the case when one has literally no observations of the parameter.

Consider now the situation. For any given density function  $f$ ,  $U_x$  in  $[E(X_n) - L_0, E(X_n) + L_0]$  will be rejected if and only if the density function  $f(x) < q(1 - 2p_0) / 2L_0$ ; for  $U^-$  and  $U^+$ , if and only if  $\int_{-\infty}^{E(X_n) - L_0} f(z) dz < qp_0$  or  $\int_{E(X_n) + L_0}^{+\infty} f(z) dz < qp_0$ , respectively.

But since *all* probability functions, all  $f$ 's, that is, are allowed, for *every one* of the hypotheses,  $U^-$  and  $U^+$  included, there are some probability functions that recommend rejecting it and some that recommend accepting it. In particular, there is always some probability functions (for example,  $f \equiv$  the M-function itself!) that will recommend rejecting *no* hypothesis.

What to do? We can use Levi's rule of ties. Since *every* possible strategy from rejecting no hypothesis to rejecting all but one (it is impossible to reject all of them, as seen above, since that means adding an inconsistency to  $K_{\text{John}, t_0}$ , which is never recommended, see Levi, 1980 about "deliberate inductive inference", Ch. 5), that is, they are all E-admissible, the rule of ties recommends using the disjunction of all of them—the hypothesis "reject nothing"—as long as it is "reasonable" (e.g., itself at least E-admissible.) This is the case, as we've just seen.

Finally, there is the case of iteration. But in this case, since nothing is rejected, there is no iteration—the first action ("add nothing") is the final one that is recommended to John. There is no reason to conditionalize the probability functions or M, since nothing is added to  $K_{\text{John}, t_0}$  in

the first place.

So the recommended strategy is:

***John's Recommended Inductive Inference for  $E(X'')$*** : Remain in complete suspense about  $E(X'')$ ; accept nothing stronger than " $E(X'') \in \mathbb{R}$ ".

### **STEP 3: EVALUATING $E(X) = (1-p)E(X') + pE(X'')$**

Now John is finally ready to evaluate  $E(X)$  itself. Could, perhaps, the fact that at least  $p$  can be bounded by the agent be of use? The answer is negative. For if there is no information at all about  $E(X'')$ , then there is similarly no information about  $(1-p)E(X') + pE(X'')$ .

The reason is that the evaluating of  $E(X'')$  is undounded—it can be anything as far as John is concerned—so that the fact that it is multiplied by a small  $p$  is of no consequence. John cannot exclude the possibility that  $E(X'') = 1,000,000p$ , or  $10^{100}p$ , for that matter.

To put it somewhat more formally, consider any probability function  $g$  which supposedly gives us the definition of how  $E(X) = (1-p)E(X') + pE(X'')$  is distributed around  $\mathbb{R}$ . It is easy to find some other probability function,  $g''$ , such that if  $g''$  is the distribution of  $E(X'')$  in  $\mathbb{R}$ , then  $g$  is that of  $E(X)$ . The fact that  $p$  is small doesn't mean that  $E(X)$  must be small; if  $g$  (say) says that the likelihood of the average of  $E(X)$  is distributed around 1,000,000, just choose a  $g''$  where the likelihood is that  $E(X'')$  is distributed around  $1,000,000/p$ .

So John's possible functions for the likelihood of  $E(X)$  being anywhere in  $\mathbb{R}$  is still *all of the possible probability functions*. And for the same reasons as above:

***John's Recommended Inductive Inference for  $E(X)$*** : Remain in complete suspense about  $E(X)$ ; accept nothing stronger than " $E(X) \in \mathbb{R}$ ".

In conclusion: even if we *know* that a certain generator is a type 3 distribution, before a catastrophic event occurs we cannot say anything about the difference between the observed  $E(X_n)$  and  $E(X)$ , the observed  $\text{Var}(X_n)$  and  $\text{Var}(X)$ , or any other observed moment and the "real" one. Before such an event occurs, extrapolating from past data to future behavior of such

a system is *worthless*.

Here we see that the mathematical information is *necessary* for reaching the epistemological conclusion. To conclude that the future is like the past we must know that the mathematical equality  $E(X) \sim E(X_n)$  (and the same with other moments) will hold. If we know that this mathematical relations does not hold, then naturally we cannot make any epistemological conclusion about the future based on the past in that case.

