THE FUNDAMENTAL DIFFICULTIES OF PROBABILITY THEORY FACING NATURAL PHENOMENA

ZULFIKAR MOINUDDIN AHMED

1. Introduction

I was just reading the introductory chapter of Michel Loeve's Probability Theory. The intuition he provides of probability are based on the assumptions that probability theory shall be mapped to repeated trials where one counts occurrence of an event A, and marks n_A/n where n is the number of trials and n_A the count of occurrence of A as the probability of A.

Now I will specialise to the actual data I am examining, which is survey data of mostly categorical variable measurements of people around the world. In this case *event* concept does not give us the best feel for the issues that matter for scientific models. Here what is going on is that a highly complex system – the human race system – is providing us with a high-dimensional measurement of more than 400 measurements per person. Furthermore, the joint probability of the variables is unknown. This is the most significant distance between probability theory and science, that when confronted with measurements of a complex system, we actually have no control over what is independent and what is not.

And this is the difference between science and mathematics. In mathematics, issues of independence and dependence are matters of axioms or in exemplary cases proven. In data from nature, there is no control of what variables are independent from others. *In fact that is the central problem of science: any clarity regarding whether variables are independent or dependent is a scientific result of import.*

In science there is no choice of examples to illustrate any deep theorem. In science, if any deep theorem has any value, we celebrate because we are never going to be confident enough that any of the axioms of any mathematical field has any value faced with nature.

2. The Delicate Problem of Interpretation

So I go around the world and survey everyone with the final question of most importance, and the answer is ordered categorical. I ask a single question only. I record their answers, and it is a range of values from $\{1, 2, \ldots, 10\}$. It's the final question that will be the be all and end all of all questions, so I can't reveal to you what it is. It's my secret, ok? I have special knowledge and know the questions that would stump even the Oracle of Delphi and so on, so I don't reveal the question.

But I use symbols, so I refer to the question as $\alpha\omega$.

We consider $\alpha\omega$ as a variable over the Human Race, i.e.

$$\alpha\omega : H_0 \to \{1, 2, \ldots, 10\}$$

Date: May 17, 2021.

So what sort of thing is $\alpha\omega$? Well it's not actually anything but a deterministic function if I ask everyone in the entire world.

This is useful, because I am contemplating producing a solid scientific theory.

Let us then do the following set of dubious maneuvers:

- We only sample a random set of 100,000 people out of 7.8 billion. But we pretend as though the statistics came from 7.8 billion and we asked $\alpha\omega$ to everyone in the world.
- Pretend that we actually sampled the real interval $[0, 10]$
- Embed the humans into a Hilbert space H and assume $\alpha\omega$ is smooth on H

After these steps, we massage the frequency table of $\alpha\omega$ to a smooth function on $\mathbf{R}$ and get a strong fit of distribution $g(\theta) = GHD(\lambda, \mu, \sigma, \gamma, \bar{\alpha})$. Then we assume that there exists a probability P_H on H such that

$$\alpha\omega_*(P_H) = g(\theta)$$

And that is the setup we would like for our scientific model.

3. The Probability Viewpoint

The Probability Viewpoint in this case is not exactly the same as the one in the last section. It is, instead, something like this. We consider $\Omega = H_0$ as the set of all humans. Then we consider the probability space $(\Omega, \mathcal{F}, Q)$ as the probability space where now Q is the uniform probability over all humans. Then we consider the random variable $\alpha\omega : \Omega \to \{1, 2, \ldots, 10\}$. Then we consider the sequence of independent random variables $X_1, \ldots, X_k, \ldots$ for $1 \leq k < \infty$. Then we are interested in using the uniform probability to sample people from H and obtain a sequence of random choice ω_j from Ω and evaluate $\alpha\omega_j(\omega_j)$ Then we assume that the frequencies of values of $\alpha\omega(\omega_j)$.

We don't care here to query the full population but are interested in sampling from a subpopulation.

4. Final Expected Outcome

We should be able to produce a Human Nature model with around 3000 parameters. I could 500 variables, 5 parameters each plus 500 parametes for eigenvalues of the covariance part.

Then we will have a 500 dimensional Generalised Hyperbolic Levy Process model for all Social Science variables using World Values Survey.

After than you can use the same parameters to set up appropriate integro-differential equations for time evolution. The Levy process is latent, but is the sample path of a heat equation with generator containing and the pure jump integral operator that comes from Levy-Khinchine representation. Now you have all Social Science variables being projected into the future by solving exact equations.

5. The 1713 Law of Large Numbers

Michel Louve records the 1713 Law of Large Numbers for Bernoulli random variables as

$$P(|\frac{S_n}{n} - p| \geq \varepsilon) \to 0$$

as $n \uparrow \infty$. We can deduce the case of finite values frequencies converging to the true probabilities from this by induction.

An interesting problem here is to attempt to bound the probability of $P(|S_n/n - p| \geq \varepsilon)$. Chebyshev inequality gives

$$P(|S_n/n - p| \geq \varepsilon) \leq \varepsilon^2 E(S_n^2/n^2) \leq \frac{1}{n}\varepsilon^{-2}$$

6. We Need Probabilistic Point of View

We need both the probabilistic view and the analytic view for a good scientific model.

Now let us find some sense assuming that it is the Chebyshev inequality that will give us inference about the frequency. The World Values Survey has $O(10^4)$ data points but for particular curves, such as with ethnicity, many classes have sample size between 1100 to 6000. What inference can we actually get from Chebyshev in this case?

Can we ask for accuracy to 1% in the frequency table? No. We'll get $\varepsilon^2 = 10^{-4}$ and that needs $N = 10000$. We want both small p-value, i.e.

$$P(|S_n/n - p| \geq \varepsilon) \leq 0.05$$

and we want ε small.

$$N^{-1}\varepsilon^{-2} \leq 0.05$$

and so

$$20N^{-1/2} \leq \varepsilon$$

With $N = 1000$ we have

$$\varepsilon \geq 0.6$$

Since we would like something sensible in accuracy, we give up high probability and we can guarantee $p = 0.25$ with $N = 6000$ that accuracy can be

$$\varepsilon = 0.051$$

Direct application of Chebyshev for ensuring frequency is the true mean with $N = 6000$ will not guarantee anything like $\varepsilon = 0.001$ with p-value 0.01.

7. Delicate Issues in Social Science not Present in Physics or Chemistry

By its very nature, the ambitions of social scientists of what they will address in Nature is substantially greater than most physics and chemistry measurements. Measurements of extremely complex systems such as the human race systems seeking nontrivial variables that are products of a great deal of unknown complexity is the norm in social science. Noise is naturally much higher in social science measurements than in the precision measurements of physicists on much simpler quantities. In this situation, issues of statistical inference cannot be considered secondary, because there are no comparable highly established theories that have been tested to 10 decimal figures of precision. If the sample size issues are not resolved to ensure some *guarantee* that our means S_n/n will be certain to within some precision to the totally mysterious and unknown truth, by any method that can provide such guarantees, then social scientists are taking a grave risk of introducing *uncontrolled noise* in their models. And this will lead to great ideas producing mediocre gossip rather than illumination of deep features of complex nature. This is a shame, because noise is nightmare in science and can bring defeat from the jaws of victory of great scientific insights.

Now I think that there is quite a bit of anarchy and confusion in issues of sample size in social sciences. The rule of thumbs and heuristics for choice are good for some things but in fact I would say Chebyshev Inequality is one of the most potent tools to decide sample size because it is robust with minor assumptions.

It is important to understand that inference is not about visible data but guarantees about whether the various manipulations of data are referring to Nature 'out there' or merely telling us about variations in measurement. It is a pragmatic choice to understand noise as dispersion from mean. But in measurement there is uncontrolled noise if the sample size is insufficiently high leading to soggy unknown actual truth and instead measurements with high noise of unknown origin.

Social scientists cannot risk the possibility of low signal to noise from inappropriate choice of sample size determinations. Normal sample sizes are inappropriate because data are rarely normal.

Use proper guarantees for sample sizes from Weak Law such as Chebyshev inequality rather than overoptimistic bounds from normal theory.

8. Science is Hard Because Theories are not the World

In Science, our theories are *about the World*, i.e. Nature, that is external to our concerns. We want to produce theories about Nature and we want to be right *about the World*. This is a spectacularly bold undertaking, because we just have theories and some measurements that we believe represent some features that are valid about the World.

The bridge between theories and the World is not automatic. This is the fruit of Statistical Inference, and the idea here is that after some measurements are recorded, they are not automatically truth. The nontrivial miracle is that tools like Chebyshev Inequality give us confidence that even though we do not know the true mean we still can bound the difference between the *sample mean* and the *true mean* by Chebyshev Inequality provided our random variables do have second moments. This is necessary for it allows us to guarantee that we are indeed able to declare things about the world with confidence.

9. A Return to Quetelet and Lexis

I just realised that some of my impulses today, unknown to myself, was pioneered already by Adolphe Quetelet (17–1874). The following is from [1]

"The new science of probability and statistics was mainly used in astronomy at the time, where it was essential to account for measurement errors around means. This was done using the method of least squares. Quetelet was among the first to apply statistics to social science, planning what he called "social physics". He was keenly aware of the overwhelming complexity of social phenomena, and the many variables that needed measurement. His goal was to understand the statistical laws underlying such phenomena as crime rates, marriage rates or suicide rates. He wanted to explain the values of these variables by other social factors. These ideas were rather controversial among other scientists at the time who held that it contradicted the concept of freedom of choice."

Some years ago, I, too, after the success of Four-Sphere Theory turned to Social Science with similar ideas. I am pleased to learn about Adolphe Quetelet, as I had not studied his works before.

I, Zulf, am in possession today of vastly superior understanding of probability theory than Quetelet, who died in 1874, could access. The key figure whose works allowed this possibility is Paul Levy. For it was Paul Levy's great genius that allowed the world to go far beyond the constraints of Gaussian to processes associated with infinitely divisible distributions on the real line. His works are from 1920s and 1930s. And of course the discovery of what I will consider the *most important Levy Processes in the history of human race* which are Generalised Hyperbolic Processes, and this is the great work of O. Barndorff-Nielsen from the mid 1970s. The discovery of these Generalised Hyperbolic Processes will settle, in my view all classical questions of actual Noise in Nature and quell confusion of two centuries in every field of Science.

References

[1] https://en.wikipedia.org/wiki/Adolphe_Quetelet#Social_physics