Open Science Framework DOI:10.31219/osf.io/***

What is a probability model? [draft]

P.G.L. Porta Mana
Western Norway University of Applied Sciences <pgl@portamana.org>

22 May 2020; updated 28 September 2021

An operational*** definition of probability model is given, and some of its consequences discussed. Some definitions of statistical model used in the literature turn out not to be models according to the present definition; in particular, their probability conditional on data cannot be calculated.

1 ***

****

OLD SNIPPETS
****

2 Would you like to know which one is true, eventually?

The word *model* seems to have been taking the place of *hypothesis* since around the 1960s. This replacement does not seem to be connected with the shift from frequentist to Bayesian methods.

Does the replacement of “hypothesis” by “model” indicate a shift in concepts in all probability theory and statistics? Or are these two words simply equivalent?

Many hodiernal authors indeed seem to use them interchangeably. An example is Kass & Raftery’s famous review[1]. Initially they define a statistical model as something that represents the probability of the data according to a hypothesis (p. 773). Eventually they use the two terms interchangeably, for example saying at times “the hypothesis H_k”, at times “the model H_k”. ***

“It’s just a matter of terminology”, some may say. But I want to argue that that there are three deeper important problems, and they need to be solved.

The first problem is pedagogical. Many students of probability and statistics have uncertainties and misconceptions about what a model is. Model of what?

[1] **kassetal1995**.

This document is designed for screen reading and two-up printing on A4 or Letter paper

The second problem is an *inadvertent* mismatch between hypotheses declared to be under analysis, and hypotheses that are actually analysed. That is, some authors state that they will compare a particular set of hypotheses, but an analysis of their mathematics reveals that they are unintentionally comparing a different set.

The third problem is the *verifiability* of a hypothesis or model. I mean the following.

There's a box with three balls, which can be blue or red. You have four hypotheses about their colours: H_0 : no blue balls, H_1 : one blue ball, two red balls, H_2 : two blue balls, one red, H_3 : all blue balls. Each of these hypotheses, granted additional assumptions about the drawing procedure, leads to a probability for the colour in the first draw and in all four draws.

You can assign a probability to each hypothesis. As draws are made, your probability for each hypothesis changes, and so does the probability for each colour in the next draw.

Once all three balls are drawn you see which hypothesis is true (some were proven false along the way). This is reflected in their probabilities conditional on all data: one hypothesis gets unit probability the rest zero.

*** example with composite hypotheses

*** example with models. ... Something peculiar happens: we have collected *all possible* data, but none of the models has reached unit probability. You are still uncertain about the models. This leads to some questions:

- What are your models about? they cannot be about the colour composition, because now you know that perfectly and yet the truth of the models is still unsettled.
- What kind of data would you need to settle their truth?
- Are your models really relevant for this urn problem?

There's a closed box. You have three hypotheses: the box contains one, two, or three balls. By shaking the box a little and listening to the rattle you get the impression that there should be three balls. Maybe two. Unlikely to be just one. You may express your beliefs with a probability distribution.

But how do you *verify* how many balls there are? If possible you could open and look; or X-ray the box; or weigh it (assuming you knew the separate weights of box and balls). The electromagnetic or weigh data would tell you which hypothesis is true. You verify that there are two balls.

The verification of other hypotheses, often entertained in science, is not so straightforward. It would require an infinite amount of data or time.

Someone states that, during the whole history of a specific coin, it will come up heads 51% of the times it will be tossed, and tails 49% of the times. Another person says 50%/50%. How do you verify these hypotheses? You would need to monitor the coin, probably bequeathing this scientific task to several future generations (the oldest coin existing today is about 2 600 years old[2]), amassing a very long sequence of data. Or imagine ancient Greeks making hypotheses about the Earth's precise distance from the Sun. The data necessary to verify their hypotheses could not be gathered then, owing to lack of technology. But they were some centuries later[3]. In cases like these we say that the hypotheses are verifiable *in principle*. Scientific hypotheses are often of this kind.

There are hypotheses that cannot be verified even with infinite data or futuristic technologies. Their verification hinges on dubious notions, such as multiple copies of this universe with slightly different initial conditions. It is debatable whether we can speak of "hypotheses" and "verification" in this case.

The three cases above are not sharply distinct. There is an increasing difficulty with the kind or quantity of data necessary to verify a set of hypotheses.

I believe that whenever we propose a set of hypotheses or models we should always point out what kind of data would be necessary for their verifiability. Such a requirement is in line with today's emphasis on reproducibility and the abandonment of "significance" in favour of statistical thinking[4]

🧩 "the null must be nested within the alternative" (end of p. 776)

I have never seen a paper in which a probability model is *refuted*. Sure, there was "hypothesis rejection at some significance level"

***comparing two frequency-priors using exch. data is like using one data point only.

***[5] Raftery: non-testable models ***[6] is the "hot hand" example really testable?

[2] https://rg.ancients.info/lion/article.html

[3] **goldstein1985**.

[4] **asa2019**.

[5] **rafteryetal1989**.

[6] **kassetal1995**.

3 Discussion

The shift from "hypothesis" to "model" seems to reflect the gradual abandonment of trying to understanding a phenomenon to simply trying to fit to it some equation pulled out of a hat. This is also reflected in the defacing of the verb *explain*: many authors say "this distribution explains the data" (or the variance of the data, or similar) when in reality it just *fits* the data – it does not explain them.

****

4 What is a model?

*** model must be about decidable statement. Composite models are not
[7]

A ***probability model*** *is a combination of knowledge and hypotheses that allows You to assign numerical probabilities to a specific set of statements and to all their logical combinations* (1)

Such assignment is usually done by specifying the joint probability distribution over the *constituents* (or *basic conjunctions*, or *fundamental products*) of the set of statements, that is, conjunction of all of these in which some, none, or all are negated[8]. In scientific inferences such constituents statements are typically of the form "$X = x$" where X is a measurable quantity.

It is implicit in this definition that the set of statements includes *all* statements involved in Your inference. If Your inference is about given data , then the model must have assigned the probability $\mathrm{p}(\land \mid)$.

This definition also contains an implicit distinction between *model* and *hypothesis*. A scientific hypothesis by itself is often not sufficient to assign probabilities for statements about the quantities it involves. Measurements of such quantities, for example, often require additional knowledge or hypotheses about the measurement process.

For the moment let us neglect this distinction between *model* and *hypothesis* and use the two words interchangeably, as often done in the statistics literature; replace "hypotheses" with "conjectures" in the definition above. In view of that definition, some problems discussed

[7] **hailperin2011**.

[8] **hailperin1996**; **hailperin2011**.

in the statistics literature appear in a new light, and some disappear altogether.

Let consider the following specific, common situation: our statements are $_1 \coloneqq$ ı$X_1 = x_1$ȷ, $_2 \coloneqq$ ı$X_2 = x_2$ȷ, and so on. They concern a set of one, two, or more similar measurements.

4.1 Parametric models and model comparison

4.2 Composite hypotheses

4.3 Nested models

5 ***

The probability calculus, just like the truth calculus, proceeds purely syntactically rather than semantically. That is, if I tell you that H and $H \Rightarrow D$ are true, you can conclude that D is true; similarly, if I tell you that $\mathrm{p}(H \mid) = p$ and $\mathrm{p}(H \Rightarrow D \mid) = q$, you can conclude (try it as an exercise) that $\mathrm{p}(H \; D \mid) = p + q - 1$. And in either case you don't need to know what H and D are about – they could be about Donald Duck or parallel universes. Don't we too often abuse of this syntactical property? We often say "under model H our belief about the value of quantity x is expressed by such and such distribution $(x \mid H) = f(x)$", without explaining what θ really is and why it leads to f. Aren't terms such as "model" and "hypothesis", as often used in probability and statistics, convenient and respectable-looking carpets under which we can sweep the fact that we don't quite know what we're speaking about? The need to look under the carpet arises, though, the moment we have to specify our pre-data belief, the prior, about the mysterious H.

And yet again, semantics can very well be a by-product of syntax, or the distinction between the two be a chimera[9]. Such important matters are unfortunately rarely discussed in probability and statistics.

***

Only in extremely rare cases does the set of hypotheses $\{H_h\}$ encompass and reflect the extremely complex and fuzzy hypotheses lying in the backs of our minds. The background knowledge is therefore only a simplified picture of our actual knowledge. That's why or the hypotheses $\{H_h\}$ are often called *models*. "A theory cannot duplicate nature, for if it did so in all respects, it would be isomorphic to nature itself and hence useless, a mere repetition of all the complexity which nature presents to us, that very complexity we frame theories to penetrate and set aside. If a theory were not simpler than the phenomena it was designed to model, it would serve no purpose. Like a portrait, it can represent only a part of the subject it pictures. This part it exaggerates, if only because it leaves

[9] **wittgenstein1945_t1999**; **girard2001**; **girard2003**.

out the rest. Its simplicity is its virtue, provided the aspect it portrays be that which we wish to study" (Truesdell[10]).

***

A good portion of recent literature seems to focus on the relation between a hypothesis and *future* data rather than the collected data; but I think it's equally important to see how it relates to the data we already have. This tension between future and past data stems, in my opinion, from a misunderstanding about which model or hypothesis we're actually interested in. I hope to discuss this topic more at length in another work; but let me give a brief sketch of this misunderstanding.

As mentioned in the introduction, H sometimes denotes some unspecified knowledge that leads to a specific probabilistic model $\mathrm{p}(\dots \mid H)$ (see the remark about syntax and semantics on p. 5). We must remember that the conjunction of H and some data D defines a new and *different* probabilistic model $H' \coloneqq D\,H$. In the context of exchangeable models[11] there's an important distinction between two kinds of models:

(a) models that lead to the *same* belief about new data, independently of the data already observed:

$$\mathrm{p}(D' \mid D\,H) = \mathrm{p}(D' \mid H) \qquad (\text{if } D \nRightarrow D'); \tag{2}$$

these are usually called "simple" hypotheses or models[12], but I prefer to call them *non-learning* models, because when we assume such a model, observed data become irrelevant for our beliefs about unobserved data (although observed data are of course relevant for our belief about the model);

(b) models that lead to different beliefs about new data, depending on the data already observed:

$$\mathrm{p}(D' \mid D\,H) \ne \mathrm{p}(D' \mid H) \qquad (\text{if } D \nRightarrow D'); \tag{3}$$

these are usually called "composite" hypotheses or models,[12]but I prefer to call them *learning* models, for reasons analogous to the ones in the previous point.

Geometrically, learning models constitute a convex set, of which the non-learning models are the extreme points.[13] In fact, a learning model can

[10] **truesdelletal1980**.

[11] **bernardoetal1994_r2000dawid2013**; **definetti1937**; **kingman1978**; **kochetal1982diaconisetal1980**.

[12] **bernardoetal1994_r2000**.

[13] **lauritzen1974**; **lauritzen1982_r1988**; **lauritzen1984**; **barndorffnielsen1978_r2014**; **dynkin1978dawid2013**.

be considered as a state of uncertainty about which specific non-learning model holds, within a certain set of non-learning models. Our belief about new data is therefore give, by the theorem of total probability, by a mixture of our beliefs conditional on the contemplated non-learning models entertained; the weights of this mixture being our distribution of belief among these models themselves.

From this point of view the question "how well does the *learning* model forecast new data?" appears somewhat strange. Because it's the *non-learning* models that, so to speak, forecast new data, while the learning model only expresses our beliefs in those models. And it's the *observed* data that help us sharpen such beliefs (future data can't help us: we don't have them) – precisely by quantifying how much the competing individual non-learning models have been able to forecast the observed data so far. And this is exactly what the likelihood or log-likelihood (**??**) does, whereas the log-score (**??**) only does so incompletely.

Learning models have the important, factual property that *their conjunction with observed data tends to a non-learning model* as the number of observed data increases. Geometrically, as data accumulate, our model moves in the convex set mentioned above, approaching an extreme point.[14] I say "factual" because this is not a mathematical theorem: we may imagine and do encounter sequences of data for which no convergence occurs;[15] but in practice this usually makes us replace our simplified set of hypotheses with a less simplified one and start anew. There's a sort of Bayes theorem operating on a large hypothesis space which we keep on the back of our minds, of which the space represented on paper is just a necessarily simplified portrait or caricature (see Truesdell's quote in § **??**). Compare Jaynes's[16] discussion about the "resurrection of dead hypotheses".

The hypotheses typically investigated in the sciences are non-learning models.

the degrees of belief about data given some hypothesis only *after* some data have been collected, but I think it's equally important to see how it relates to the data we already have. (compare for example[17], although their motivation stems from o

If we see the symbol H as just a placeholder for a probabilistic model, rather than an empirical hypothesis (see the remark about syntax and

[14] **barronetal1986hjort1986**. [15] **bruno1964**; **berk1966**; **berk1970diaconis1988**. [16] **jaynes1994_r2003**. [17] **ohagan1995**; **bergeretal1996**; **bergeretal1998**.

semantics on p. 5), we must remember that the conjunction of H and some data D simply defines a new probabilistic model $H' \coloneqq H\,D$.

***

This approximation is reasonable if the amount of data is large with respect to the dimension of the space of a single datum, because *** (ref to geisser, stone, gelfandetal)

***

*** remark that $D\,H$ is a *new* probability model: which of the two are we assessing? Connection with learning vs non-learning models which I hope to take up in another work.

***[18] show the approximation only valid if number of data large enough – not very interesting situation in today's problems.

*** problems calculation with time-relevant hypotheses

"we cannot give a universal rule for them beyond the common-sense one, that if anybody does not know what his suggested value is, or whether there is one, he does not know what question he is asking and consequently does not know what his answer means"[19].

[18] **bernardoetal1994**. [19] **jeffreys1939_r1983**.