Statistical Inference Without Frequentist Justifications

Jan Sprenger

November 29, 2008

Abstract

Statistical inference is often justified by long-run properties of the sampling distributions, such as the repeated sampling rationale. These are *frequentist justifications* of statistical inference. I argue, in line with existing philosophical literature, but against a widespread image in empirical science, that these justifications are flawed. Then I propose a novel interpretation of probability in statistics, the *artefactual interpretation.* I believe that this interpretation is able to bridge the gap between statistical probability calculations and rational decisions on the basis of observed data. The artefactual interpretation is able to justify statistical inference without making any assumptions about probability in the material world.

1 Frequentist Statistics and Frequentists Justifications

In modern science, inductive inference often amounts to statistical inference. Statistical techniques have steadily conquered terrain over the last decades and extended their scope of application to more and more disciplines. Explanations and predictions, in high-level as well as in low-level sciences, are nowadays fueled by statistical models. However, this development did not occur because scientists believe the underlying systems to be irreducibly stochastic. This might sometimes be the case, but certainly not in general. Rather, even traditionally "deterministic" sciences (such as several branches of physics, psychology and economics) use statistics to model uncertainty, noise, imperfect measurement etc. A wide spectrum of techniques can be used to draw valid conclusions from data: Hypothesis tests help scientists to see which of two competing hypotheses is better supported by the data. Confidence intervals narrow down the set of values of an unknown model parameter which is compatible with the observations. And so on.

The classical methodology to answering these questions is *frequentist inference* (cf. Cox 2006). For reasons that will soon become obvious, I believe the term "frequentism" to be a misnomer. Rather, as pointed out by Mayo (1996), that school of statistical inference is characterized by a focus on the probability of making an error in the inference to a certain hypothesis or in setting up a confidence interval – hence, the name *error statistics.* A statistical procedure is good if and only if the two probabilities of committing an error – accepting a hypothesis when it is false, rejecting it when it is true – is low. For instance, assume that you want to test whether in a culture of 10000 cells, less than 5% have been infected with a certain virus. That is your working hypothesis. To perform the test, you draw a sample of 100 cells. Then you formulate a decision rule whether or not to accept that hypothesis, dependent on how many infected cells are in your sample. You calculate the *error probabilities* of that rule – i.e. the probability that you accept the hypothesis when more than 5% of all cells are infected, and the probability that you reject it when less than 5% are infected. The lower these probabilities, the more powerful and the better your test. Finally, you look at the data and come up with a conclusion according to your decision rule.

This example can easily be transferred to other applications of statistics in science. The rationale behind these procedures is that the more sensitive a hypothesis test which suggests a certain conclusion, the more substantiated the conclusion which the test yields (Neyman and Pearson 1967, Mayo 1996). The crucial question – the one this paper tries to answer – is why probabilistic properties of such a test should affect rational decisions and actions (i.e. to base the next experiment on hypothesis H rather than on $\neg H$). What precisely does it mean that a test has error probabilities of, say, 0.01 and 0.034 (for the two types of error)? And why are these values relevant for our decision to accept or to reject our working hypothesis in a specific, concrete problem?

An answer is suggested by the popular *repeated sampling rationale* of statistical inference. Statistical data are considered as a (small) sample out of a (large) population, and in principle, this sampling process could be replicated. Thus, the error probabilities of a test are interpreted as *relative frequencies* with which the testing procedure makes an error if the sampling procedure were repeated. For instance, if the probability of erroneous rejection is 0.01, then, if this test were repeated very often, we would erroneously reject our hypothesis only in about 1% of all cases where it is true. In particular, if the test were repeated infinitely often, the rate of erroneous rejection would almost surely settle at 0.01, due to the Strong Law of Large Numbers. This is supposed to justify our confidence in a particular inference:

> "We intend, of course, that this long-run behavior is some assurance that with our particular data currently under analysis sound conclusions are drawn."[1]

We call this a *frequentist justification* of statistical inference – our actual inference is supported by the long-run properties of the procedure(s) we use in our inference. In other words, we can confidently endorse the result of our test because, if the test were repeated, we would rarely go wrong. In a similar vein, Neyman and Pearson, the founding fathers of the error-statistical approach, write:

> "[...] we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false."[2]

However, the frequentist justification of inductive inference is not limited to frequentist, error-statistical inference – it can be applied in Bayesian statistics

[1] Cox 2006, 8.

[2] Neyman and Pearson 1967, 142.

as well. Bayesians aim at posterior probabilities of a hypothesis H conditional on evidence E which they calculate by means of Bayes's Theorem:

$$P(H|E) \quad = \quad \frac{P(H)\,P(E|H)}{P(H)\,P(E|H) + P(\neg H)\,P(E|\neg H)}. \tag{1}$$

The values of $P(H)$ and $P(\neg H)$ are, on a Bayesian account, standardly interpreted as the subjective degrees of belief in the truth of hypothesis H. However, the other probabilities that occur – the so-called *likelihoods* $P(E|H)$ and $P(E|\neg H)$ – fulfil the same role as error probabilities above: they are probabilities of certain events under specific hypotheses (H is true vs. H is false). Thus we can assign them a repeated sampling interpretation as well: if we draw a lot of samples out of a H-distributed population, the actual evidence will occur in a fraction of cases roughly equal to $P(E|H)$. What remains to argue is that these repeated sampling properties bear on the actual inference we make.

Although an answer seems obvious ("just set your credence to the computed probability"), it is hard to *argue* for this step without invoking circularity. Why should the non-subjective, statistical property $P(E|H)$ affect our rational credences? Howson and Urbach (1993, chapter 9), in their philosophical monograph on Bayesian inference, make recourse to frequentist justifications: They (correctly) observe that the value of $P(E|H)$ makes an assertion about the frequency of observing E in a long, potentially infinite run of scenarios where H is true. That said, they give an argument why our credences in a single case should match those probabilities, i.e. they give a frequentist justification of inference in a single case. Hence, frequentist justifications do not only arise in frequentist statistics, but in all varieties of probabilistic inference. For this reason, I find the term "frequentist statistics" misleading.[3] However, are frequentist justifications justified at all?

2 Against Frequentist Justifications

The frequentist justification of statistical inference has been under pressure from various sides, some in the philosophy of statistics, others in the philosophy of probability. Objective (as opposed to purely subjective) probabilities have often been interpreted as limiting relative frequencies (von Mises 1928), i.e. the meaning of a probability statement such as $P(E|H) = x$ was *identified* with the fact that in an infinite run of trials where H was the case, the limiting frequency of E would approach $x \in [0, 1]$. This account has been popular among scientists and philosophers alike because metaphysical assumptions about the nature of probability are avoided.

However, Albert (2005) points out that a frequentist, repeated sampling interpretation of objective probability (and in particular, statistical probability) cannot support rational credences in single cases, i.e. our degree of belief that we have actually committed an error. The limiting relative frequency of, say, erroneous rejection does not impose any constraints on the expectedness of erroneous rejection in a *finite* set of tests; and arguments to the contrary (such as Howson and Urbach 1993) can be shown to be fallacious. Any result in a finite

[3] However, from a historical point of view it is certainly true that frequentists justifications emerged from the error-statistical school of statistical inference (in particular from Jerzy Neyman), and this explains why frequentist statistics got their name.

set of tests is compatible with any limiting frequency because in the limit, the "weight" of the finite segment will be zero.[4] In other words, whether a test erroneously rejects a hypothesis under test is logically independent of the limiting frequency of rejection in the infinite run. Therefore, limiting relative frequencies cannot ground confidence in our inferences, e.g. following acceptance/rejection rules in a particular case.

Albert intends to make a point against the connection between frequentist justifications based on a repeated sampling interpretation of probability and rational credences, but his argument transfers to a broader sort of frequentist justifications as well. No matter how we understand probability, the Strong Laws of Large Numbers guarantee for any interpretation of probability that the limiting relative frequency almost surely converges to the true probability value. This explains why frequentist justifications used to be so attractive.

That point is not to be confused with the (false) claim that probability theory does not tell us anything about finite samples. For instance, with increasing sample size the sensitivity of a hypothesis test increases, too. This is not contentious. Rather, Albert shows that, if all we assume about statistical probabilities is that they determine limiting relative frequencies, we cannot make meaningful assertions about finite samples. The point of frequentist justifications was to circumvent a difficult task: namely, to describe the link between mathematical, statistical probabilities and rational credences and decisions. The lesson of the above criticism is that we cannot avoid to address that question.

In other words, we need a *bridge principle* connecting objective statistical probabilities and rational beliefs and decisions. Standardly, the *Principal Principle* (Lewis 1981) has played this role. I do not want to bother with the technicalities of the various formulations of the Principal Principle, but it can be seen as the attempt to capture the idea that

- if we know that the objective probability of E, given information H, is q
- if we do not have any "inadmissible" information that affects these probabilities

then our rational credence that E, conditional on H, should be equal to q.

Obviously, the Principal Principle serves as the desired bridge principle. The likelihoods $P(E|H)$ and $P(E|\neg H)$ are probabilities of certain observations under fixed distributions H or $\neg H$, and so are the error probabilities $P(reject H_0|H_0)$ and $P(reject H_1|H_1)$ since the decision whether or not to reject depends exclusively on the observations. Thus, both types of probabilities are *objective probabilities* and the Principal Principle applies. However, Strevens (1999) has argued that it is difficult to justify the Principal Principle in a non-circular way; and claiming that it is "evidently true" seems to beg the question against those who don't believe in bridge principles between objective and subjective probability. Therefore, invoking the Principal Principle is not a convincing way to justify statistical inference. Rather, we should address the following two questions:

1. What kind of objective probabilities are statistical probabilities – and more general, probabilities in scientific modeling?

[4] Strevens (1999) makes the same point with respect to "long-run propensity" interpretations.

2. Why do these probabilities affect rational decisions?

3 Artefactual Probabilities

At the end of his (2005), Albert mentions that due to the failure of repeated sampling interpretations to account for the normative force of statistical inference, we might be pressed to understand objective probabilities as *propensities*, i.e. irreducible causal tendencies to bring about a certain result. For instance, a Pb^{210} atom might have an irreducible tendency of 1/2 to decay in the next hour. However, this interpretation comes with several drawbacks that have been discussed extensively in the literature.[5] For our problem, it is most salient that propensity interpretations take probabilities as features of the world, e.g. as a causal power or irreducible tendency to bring about a certain event. However, it is questionable that practicing statisticians need accept such a strong, realist interpretation of objective probability. Maybe the world has no stochastic features at all. Even in such circumstances, our statistical inference ought to be meaningful, and if all objective probabilities have to be propensities, this desideratum might be lost. Instead, we require a *minimal* interpretation of objective probabilities – an interpretation that vindicates their normative force for probabilistic inference, without making too many metaphysical commitments.

To my mind, the lack of distinction between *objective* and *ontic* probability (cf. Rosenthal 2004) is the culprit for the lack of an account of the normative force of statistical probability. Ontic theories understand probabilities, or at least certain types of probabilities, as features of the world, such as mass, charge or entropy. Probabilities become properties of a physical system or a certain experimental arrangement. Of course, all ontic accounts are objective: there is a fact of the matter about the true value of an ontic probability; thus, in case of disagreement at least one side is wrong. However, objective probabilities need not be ontic – there might be probabilities whose value is unanimously accepted but which are not grounded in the material world. I believe that statistical probabilities are of that type.

My account of statistical probabilities is the *artefactual interpretation* (cf. Gillies 2000, 179). It is motivated by the very idea of statistical modeling. When we are modeling a physical system, we are not so presumptuous to believe that our (probabilistic) model will capture all aspects of the system, and that inconsistent observations are merely the result of measurement inaccuracies. Such a complete stochastic model is rather the exception than the rule – it might occur in some fundamental branches of science, but certainly not in complex disciplines as in economics, psychology, or geophysics. In those science, we know very well that most of the time, there is no "correct" statistical model, or at least, we will not be able to find it. So we should not understand our statistical model as outright guesses about the true model. Rather, we *idealize* the target system into a mathematical model and hope that we approximately caught the interesting and fundamental properties of the system. This will help us to understand the system's dynamics and governing mechanisms as well as to make reasonable predictions. But in making inferences about the model, we are well aware of the imperfections of the model. In particular, conclusions about the model transfer to the target system only *cum grano salis*.

[5] See Eagle 2004 for a comprehensive critique.

Therefore I contend that, when we use probability as a scientific modeling tool, we do not take it literally, e.g. as the belief that the studied population is fundamentally, irreducibly Normally or exponentially distributed. (Note that these two distributions are themselves mathematically convenient idealizations of discrete distributions!) Rather we reason like this: "OK, let's assume for the sake of the argument that the random variables $(X_1, \ldots, X_n)$ are independent and all Normally distributed (either mean value $\mu \leq 0$ or $\mu > 0$), and let's see how far we get using that assumption."[6] In other words, we act *as if* the suggested distributions – call them again H and $\neg H$ – were the only two possible models of the underlying system. We imagine a world in which some events are more expected than others, and we take this to be the meaning of $P(\cdot|H)$ respectively $P(\cdot|\neg H)$. Of course, reality might be quite unlike that imagined world. Still, we can compare two of such models with the help of real data. Therefore I call the probabilities of observations which are calculated in such an imagined probability model *artefactual probabilities* – they are the offsprings of our mathematical artefact, the probability model.[7]

I believe that artefactual probabilities capture the way working statisticians and empirical scientists reason about probability. For them, (objective) probability is mainly a modeling tool, and it should not be taken to mean more than that.[8] But can statisticians really dismiss the metaphysics of probability so easily? When an error statistician rejects a null hypothesis in favor of an alternative, isn't she committed to accepting the claims the alternative makes? Aren't we suspect to the realist argument that when a theory works well, we should have an account of the quantities it posits? Are theories where chances figure not similar to theories where electrons or quarks figure? I don't think so. Probabilistic models are often used although it is crystal-clear that they are not literally true, but rather a refined prediction tool. They are just the only way to make sense of the apparently messy, biased and untidy data that we collect, apt to model sampling error even if the sampling process was not genuinely random.

Some more explanations might help. Clearly, artefactual probabilities qualify as objective: typical statistical assertions as

> (%) "If a coin is fair, the chance of getting two 'heads' in two independent and identically distributed tosses is 1/4."

are objective and not open to subjective disagreement. Rather, it is part of the *meaning* of a fair coin that, if two i.i.d. tosses are performed, the chances of observing two times "heads" are 1/4. Artefactual probability statements are conditionals where the antecedens specifies a certain distribution and the consequens gives the probability of a particular event under that distribution.

[6] In that example, $P(\cdot|H)$ and $P(\cdot|\neg H)$ are not uniquely determined because H and $\neg H$ are composite hypotheses ($\mu \leq 0$ vs. $\mu > 0$). Bayesians rescue the objectivity of these probabilities by assigning a prior distribution over μ, and frequentists have their own sophisticated techniques to solve the problem; however, to explain them would go beyond the scope of this paper.

[7] It might be argued that the artefactual interpretation, instead of giving a new account, rather argues that interpretation questions are irrelevant to the problem of justifying statistical inferences. But I believe that the above explanations help us to understand what happens when scientists use probability as a modeling tool.

[8] Still, frequentist conceptions are sometimes encountered, but this phenomenon might be due to the fact that in the history of modern statistics, quite a lot of influential figures (such as Jerzy Neyman and Richard von Mises) have been arguing for a frequentist justification or interpretation of statistics.

Therefore they avoid the conceptual confusion of frequentism as well as the metaphysical hazzles of propensity accounts. Note that statistical inference builds on conditional statements such as (%) and does not require giving meaning to unconditional assertions about probabilistic mechanics in the real world, such as

> (&) "That particular coin is fair.".

Propensity accounts have, historically, been motivated by the need to explicate such sentences – e.g. what it means that a particular, material coin is fair. I do not doubt that this is a challenging semantical and ontological question and deserves serious philosophical scrutiny. I just believe that science does not need to bother with this business – for matters of inductive inferences, conditional assertions such as (%) are fully sufficient, *regardless of the preferred school of statistical inference.* Bayesians and error statisticians can, for this matter, happily agree – the only objective probabilities which they need are of type (%), not of type (&). Therefore statisticians can restrict themselves to artefactual probabilities.

It remains to argue why artefactual probabilities are normatively compelling. Recall that Strevens (1999) pointed out the problem of *justifying* bridge principles between objective and subjective probability, such as the Principal Principle. Part of his argument was the observation that it is impossible to establish a logical or semantic connection between ontic probability and subjective degrees of belief: The ones are in the material world, the others in our head. If one accepts that all objective accounts of probability have to be ontic (unless they are beset with other difficulties), then the sceptical conclusion apparently follows. However, as argued above, there is no need to accept this identification. In particular, we have developed an objective, non-ontic account of conditional probability assertions in statistics. These conditional accounts make it possible to defend the transfer of objective probability to rational expectations on semantic grounds: When scientists work with artefactual probabilities, they use probability as a model for reasoning about uncertainty. They decide to act *as if* $P(\cdot|H)$ (or $P(\cdot|\neg H)$) really gave the expectations for the observations which we make. In other words, the act of statistical modeling is free of ontological elements. All that we assume is that in this imagined model, some events are more expected than others. Our inductive inference builds on considerations that the observed event is more surprising under one distribution than under another, as witnessed by the key role of likelihood ratios both in frequentist statistics (Neyman-Pearson Lemma) and Bayesian inference. To repeat, when building and analyzing a statistical model we firstly abstract from a target system that is too complex to model in every nuance, and secondly, we interpret the different distributions in the model as making assertions about the expectedness of certain events. Models can then be compared each other in terms of how expected they render the real data, and the more expected our actual findings, the better the model, ceteris paribus. Thus, scepticism with respect to bridge principles between objective and subjective probability does not concern us: *the act of statistical modeling itself creates the crucial link to rational expectations.*

The reader might find that the account which I sketched resembles Gillies's own (2000) account of artefactual probabilities and Sober's (2008) No-Theory Theory (NTT) of objective probability. But there are subtle, and important, points of disagreement. Gillies uses the term “artefactual probability” as a subclass of objective probability and distinguishes it from “fully objective” probabilities, as the probability that an uranium atom disintegrates in a given time interval. But he also claims that artefactual probabilities “can be considered as existing in the material world”[9]. This is certainly nothing I would subscribe to because it is, to my mind, the very point of artefactual probabilities that they avoid reference to the material world and that they are mere artefactual objects, byproducts of scientific modeling. – Sober, on the other hand, likens probability to intrinsic properties of physical objects, such as mass and charge, and claims that like the aforementioned concepts, probability cannot be reduced to anything else. Certainly, by considering conditional probability as an objective relation between pairs of propositions, he is quite close to parts of my own position. But the imaginative, artefactual character of probability in scientific modeling is missing in his analysis, and he goes on with an argument for the *reality* of certain probabilities in the world. Again, I am not interested in correspondence relations between probability and the material world.

4 Conclusion

The failure of frequentist justifications of statistical inference – justifications that draw on the long-run properties of sampling distributions – triggers the question how to explicate the link between mathematical probability calculations and sound scientific inference. This task arises for Bayesian statisticians and error-statisticians alike. Certainly, statisticians need to link conditional probabilities – probabilities of evidence given a hypothesis – to their rational beliefs. I have argued that this connection can be established by adopting the *artefactual interpretation* of objective probability. Thereby we conceive statistical probabilities as rational expectations in imagined worlds. This gives an account of inductive reasoning in statistics that avoids metaphysical commitments and is thus close to the practice of empirical science. Moreover, scepticism with respect to bridge principles between subjective and objective probability, such as the Principal Principle, does not apply when the proposed artefactual interpretation is adopted.

References

[1] Albert, Max (2005): “Should Bayesians Bet Where Frequentists Fear to Tread?”, *Philosophy of Science* **72**, 584-593.

[2] Cox, David (2006): *Principles of Statistical Inference.* Cambridge: Cambridge University Press.

[3] Eagle, Antony (2004): “Twenty-one arguments against propensity analyses of probability”, *Erkenntnis* **60**, 371-416.

[4] Gillies, Donald (2000): *Philosophical Theories of Probability.* London: Routledge.

[9] Gillies 2000, 179.

[5] Howson, Colin and Peter Urbach (1993): *Scientific Reasoning: The Bayesian Approach.* Second Edition, La Salle: Open Court.

[6] Lewis, David (1980): “A Subjectivist’s Guide to Objective Chance”, in: Richard C. Jeffrey (ed.) *Studies in Inductive Logic and Probability*, Vol II., Berkeley and Los Angeles: University of California Press.

[7] Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.

[8] Mises, Richard von (1928): *Wahrscheinlichkeit, Statistik und Wahrheit.* Wien: Springer.

[9] Neyman, Jerzy and Egon Pearson (1967): “On the problem of the most efficient tests of statistical hypotheses”, in (iid.): *Joint statistical papers*, Berkeley: University of California Press. Originally published in 1933.

[10] Rosenthal, Jacob (2004): *Wahrscheinlichkeiten als Tendenzen.* Paderborn: mentis.

[11] Sober, Elliott (2008): “Evolutionary Theory and the Reality of Macro Probabilities”, in: Ellery Eells and James Fetzer (eds.), *Probability in Science*, La Salle: Open Court.

[12] Strevens, Michael (1999): “Objective Probability as a Guide to the World”, *Philosophical Studies* **95**, 243-275.