The Trouble with Objectivity and Rationality in Statistics[1]

Revised essay submitted for the course MNSES9100

Emil Aas Stoltenberg[2]

Department of Mathematics, University of Oslo

January 19, 2018

1 Introduction

Anyone acquainted with modern empirical science recognises its reliance on statistical tools and methods. Statistics is used in fields ranging from sociology and economics, to biology and physics. Political decisions are made on the basis of the likely value of a coefficient in a statistical model; the team of CERN-researchers that discovered the Higgs boson used statistical tools familiar to anyone with an introductory course in hypothesis testing.

The importance of statistics stems from many factors, one of which is its claim to objectivity. In this essay I will explain how this strive for objectivity might be at odds with another scientific desideratum, namely that of rationality, and vice versa. The point is that in order to make inductive inferences in a rational manner, one ought to (A) use methods that do not lead to contradictions, and (B), use methods that minimise the distance between ones beliefs and the true state of nature. (A) is a problem in the foundations of statistics, a problem which Hacking (1965, p. 1) defines as

> [...] to state a set of principles which entail the validity of all correct statistical inference, and which do not imply that any fallacious inference is valid.

(B) is a problem of *decision theory*, which is the theory of how to make, evaluate and compare decisions. What I hope this essay makes clear, is that presently, Bayesian statistics which satisfies (A) and (B) is not regarded as objective, and frequentist statistics, which is regarded as objective, does not satisfy (A) nor (B).

Before we plunge deeper into these problems, we first have to agree on what a statistical model is. This topic is treated in Section 2. In sections 2.1

[1] I have written about these things before in the student journal *Filosofisk supplement*, and some sections and examples in the current essay are adopted from that text. See Stoltenberg (2017).

[2] emilas@math.uio.no

and 2.2 I define what is meant by the terms 'objectivity' and 'rationality', as used in this essay. Section 3 provides an outline of the methods associated with the dominating school of statistics, known as *frequentism*.

In the spirit of problem (A), Section 4 argues for two principles that should be obeyed when making inductive inferences about the true state of nature, and examples intended to make these two principles intuitively obvious are presented. In this section it is also shown that the frequentist approach to statistics, which is often regarded as an objective approach, does not obey these two principles. An approach to statistics that is in accordance with the likelihood principle is the approach based on Bayes' theorem. Section 5 contains a brief introduction to Bayesian statistics, and explains why this approach does obey the likelihood principle. This section also provides examples of the challenges posed by the subjective element of Bayesian analyses. Section 6 summarises the arguments leading to the seeming incompatibility of rationality and objectivity. A tentative, and perhaps somewhat dispiriting, solution to this problem is presented.

Two more things are worth mentioning before we start. First, in this essay I only enter into problems in the foundations of probability as far as these are of direct relevance for the foundations of statistics. Second, this essay is about normative issues. It is concerned with how human beings *ought* to behave, and not how people actually do behave.

2 Models, objectivity and rationality

It is decided that a coin is to be tossed 12 times, and the results are recorded,

$$T\,T\,T\,T\,H\,T\,H\,T\,T\,T\,H\,H\,T. \tag{1}$$

The tosses are independent,[3] and the probability of the coin showing heads in a single toss is a number between 0 and 1. As an example, what we call a fair coin has probability 1/2 of showing heads in a single toss. We do not know whether the coin flipped above is fair or not, and we may therefore denote the probability of this coin showing heads by the symbol θ. Apart from the fact that θ is a number between 0 and 1, it is an unknown quantity. On the other hand, if we magically knew the value of θ, the probability of obtaining 4 heads in 12 tosses, as in the sequence above, could be computed via the formula,

$$\Pr(\#\text{heads} = x \,|\, \theta) = \binom{n}{x} \theta^x \,(1-\theta)^{n-x}, \tag{2}$$

[3]The independence assumption is not as innocent as it may appear, see O'Neill (2009).

where $\Pr(\cdot)$ denotes the probability of the event inside the parentheses, the vertical bar '|' reads 'given that', n is the number of tosses, and x a generic element of the set of possible outcomes,

$$\{0, 1, 2, \cdots, n\}.$$

The formula in (2), combined with the fact that $0 \leq \theta \leq 1$, is an example of a statistical model. A prototypical statistical task in this example is to make an inference about θ based on the 12 coins tosses above.

In Section 3 I get back to this example, and outline how a frequentist statistician goes about testing the hypothesis of a coin being fair. But first, I need to clarify what I mean by objectivity and rationality.

2.1 Objectivity

The question in this section is not whether there exists an objective truth out there. It does. The question is rather how a synthetic proposition[4] comes to gain the status of being objective. Drawing on Sprenger (2017), I distinguish between two, not mutually exclusive, forms of objectivity of particular relevance for the topic of this essay.

Procedural objectivity: Experiments, data gathering and coding of data are carried out according to strict protocols and procedures. The point here is that two researchers who carry out the same experiment, gather the data in the same manner, and follow the same procedure of coding, obtain data with the same characteristics. What has become known as the 'replicability crisis' in science, where scientists are unable to reproduce the results of scientific studies, may in part be caused by a lack of procedural objectivity.

Concordant objectivity: Members of a research community agree on the correctness of a model for some natural phenomenon. This type of objectivity is purely factual (it is a fact that the researchers do agree), and does not concern the way the given research community reaches this consensus.

As an example consider the discovery of the Higgs boson in 2012. The researchers at CERN reasoned that the experimental results they obtained from the Large Hadron Collider deviated from what one would expect under the assumption that *the Higgs boson does not exist*, to such an extent that they stopped believing in this hypothesis (Sprenger, 2016). To my knowledge no one has questioned this finding, which must mean that the large majority

[4] I rely on the distinction between *analytic* and *synthetic* propositions as defined by Ayer. According to Ayer (1952, p. 78) 'a proposition is analytic when its validity depends solely on the definition of the symbols it contains, and synthetic when its validity is determined by the facts of experience.'

of (if not all) physicists agree that the model of the universe that did not include the Higgs boson (the model that was rejected), was indeed the correct one for such a universe.

2.2 Rationality and subjectivity

In order not to get bogged down in details,[5] I adopt I.J. Good's (1952) principle of rationality.

> *Principle of rational behaviour*: The recommandation always to behave so as to maximize expected utility.

Elster (2010b, p. 30) defines rationality as consisting of three operations of optimisation, all three of which, I argue, are encapsulated by Good's principle.[6] The reason for including them here is that they might be more intuitive and do not require the notion of expected utility. They are:

(i) *Instrumental rationality*: Choosing the action that best realises one's preferences, given one's beliefs about the world;

(ii) *Epistemic rationality*: The art of achieving accurate beliefs about reality, given the information at hand;

(iii) *Optimal acquisition of information*: Investing, if necessary, in the collection of more information, until the cost of acquiring more information equals the expected profit of having more information.

When it comes to these three aspects of rationality, there is some philosophical debate about whether epistemic rationality is just one form of instrumental rationality, in which knowledge and truth are goals in themselves. Kelly (2003) and Yudkowsky (2009) defend such a view. Yet another debate, with particular importance for economics and social science, concerns the *explicative* power of the theory of rationality summarised in (i), (ii) and (iii). Elster (2010a,b) provides a critical contribution to this debate.

In this essay these debates will not concern us. The latter debate because we are only interested in rationality as a *normative* theory, not an explicative one. As for the former debate, I argue that Good's principle is violated if one or more of (i), (ii) or (iii) do not hold. It does then not matter whether or not (ii) is included in (i).

[5] The details I think of here involve foundational constructions of probability and decision theory, as given in for example Savage (1972), Bernardo and Smith (1994), and in Berger (1985).

[6] I.J. Good (1967) argued that (iii) follows from his principle of rationality.

With Good's principle at hand, we can determine whether a statistical method deserves the epithet rational or not. In the remainder of this section I argue, by way of example, that using only the empirical average to make a guess about an unknown probability, can fail to be rational.

Let us now formalise the notion of epistemic rationality in the context of coin tossing. In the sentence 'accurate beliefs about reality', the word 'reality' refers to the unknown parameter θ, which is the probability of heads in one toss. By 'beliefs' we will understand a procedure for making a decision about θ based on data, and denote it by δ, or $\delta(\text{data})$ to stress when it is a function of the data. An example of such a procedure is: 'take the average number of heads',[7] that is

$$\delta = \frac{\#\text{heads}}{\#\text{tosses}}. \tag{3}$$

According to the definition, a rational agent seeks to make δ as accurate for θ as possible. By this we understand that the distance between δ and θ should be minimised, and by 'distance' we think of some function of δ and θ that takes the value 0 when our beliefs are in complete correspondence with reality, that is $\delta = \theta$, and takes positive values if this is not the case. In the field of statistical decision theory, such a function is called a *loss function* and is typically denoted by L. A very common loss function is the squared error loss function,[8]

$$L(\delta, \theta) = (\delta - \theta)^2.$$

Since we decide on the decision procedure before the experiment is conducted, $L(\delta, \theta)$ depends on the data and is random. Interest is therefore in the *expected* loss function, called the *risk* function, which is the hypothetical average of the loss function in infinitely many trials. We write

$$R_\delta(\theta) = \mathrm{E}\, L(\delta(\text{data}), \theta),$$

for this function, where E stands for 'expectation' or hypothetical average. The risk $R_\delta(\theta)$, is a function of reality θ, with the decision procedure δ held fixed. Note that the risk function is just the negative of a utility function, so minimisation of $R_\delta(\theta)$ is just what Good's Principle of Rationality prescribes.

Since we aspire to rationality, the challenge is to choose a decision procedure δ that minimises the risk. We should not expect to find a decision

[7] In light of how late in history human beings understood that averages had something going for them (see Stigler (2016) for a historical account), it is interesting to note that today averages are so familiar to us that one often confounds them with the truth itself.

[8] That this loss function is common and mathematically convenient, does not mean that one should stop thinking about the choice of loss function. The loss function ought to be suited to the particular problem at hand. A symmetric loss function might for example be a very unfortunate choice for a Dutch dike engineer.

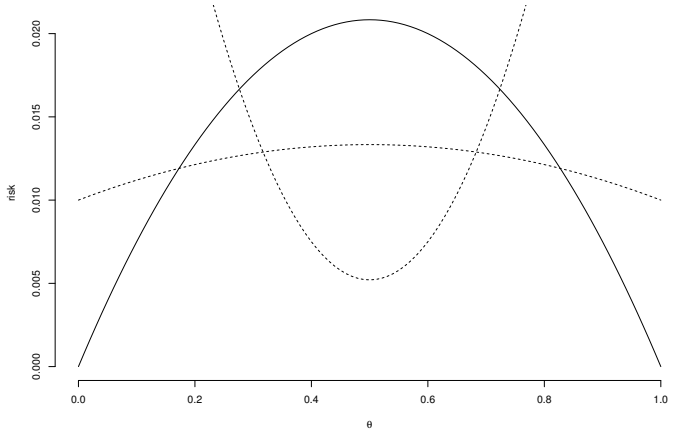

Figure 1: Three risk functions.

procedure that performs better than all other decision procedures *no matter what reality is*, that would be too much to ask. Rather, we should be happy if we can find a decision procedure that performs well for values of θ that we find likely.

In Figure 1 I have plotted the risk function (the solid line) of the decision procedure given in (3) (i.e. the average) along with the risk functions of two other decision procedures. More on these shortly.

We see that on average the distance between δ and θ is very small if the truth happens to be close to 0 or 1, and increasing as θ approaches $1/2$ (from both sides), with a maximum in $\theta = 1/2$. This is worrying! No more than a primitive understanding of the physical laws tells us, just by looking at a coin, that the probability of heads ought to be somewhere close to $1/2$. Therefore it is discomforting that our decision procedure is the furthest away from reality in the area where we have strong reasons to believe that reality is located.

To remedy this we might try to incorporate our hunch about θ into the decision procedure. One way of doing this is by letting the decision procedure be a weighted combination of such a prior hunch and the data. Consider

$$\delta = w\,\frac{1}{2} + (1 - w)\,\frac{\#\text{heads}}{\#\text{tosses}}, \tag{4}$$

where the weight w is a number between 0 and 1. By choosing w close to 1, we put a lot of weight on our hunch, while values of w closer to zero corresponds to less confidence in our guess. The two dotted lines in Figure 1 are the risk

functions of the decision procedure in (4) with $w = 1/5$ and $w = 4/5$, for the flatter and the downward pointing curve, respectively. The plot shows clearly that there is potentially a lot to gain by incorporating our rudimentary knowledge of physical laws into the decision procedure. This is seen by noting that the two dotted risk functions both lie below the solid curve in a region around $\theta = 1/2$. More importantly, according to the above definition of rationality, incorporating a prior hunch into our decision procedure seems to be the rational thing to do.

The Bayesian approach to statistics, which provides a formal framework for how to incorporate such prior hunches into decision procedures, and for updating beliefs in light of data, is presented in Section 5.

In this section we have seen an example where the empirical average, formed by the numerical data at hand, ought to be supplemented by subjective beliefs, that is more vague data, to get closer to reality. Since an objective analysis is often seen as one that is based solely on hard undisputed numerical data, it is troubling to see that relying exclusively on such data might not be the rational thing to do.

In the next section I give a brief presentation of the frequentist approach to statistics. This presentation provides the necessary background for understanding why the frequentist approach violates the principles presented in Section 4.

3 Frequentism and p-values

Scientific findings based on the proper use of frequentist statistical methods are often regarded as objective (Berger and Berry, 1988). Without going too deeply into the reasons for this (reasons which are mathematical, historical and sociological), it here suffices to say, somewhat loosely, that when a model supposed to generate the observed data is chosen, it is up to the data to decide the rest. 'The rest' refers to the likely values of the parameters in the model, confidence intervals for these, and so on.

This means that as long as the data are collected in a routine manner, yielding a sample satisfying the requirements of procedural objectivity, and the model supposed to generate the data is agreed upon by a community of researchers (i.e. concordant objectivity), the subsequent inferences are regarded as objective. Importantly, no vague data, such as the subjective hunches invoked in the example of Section 2.2, ever enters the analysis (at least not in a transparent manner, see Section 5.2).

Frequentist statistics is, as the name suggests, closely related to the view that probability expresses hypothetical long run frequencies. If we toss a coin

infinitely many times the share of heads will converge towards a number, and it is this number that we call the probability of heads. On this view, the probability of heads is a property of a coin, just as weight and circumference are properties of a coin. And similarly to how we would use a fine tuned kitchen weighing machine to ascertain the weight of a coin, we use a finite number of coin tosses to measure the property 'probability of heads' of a given coin.

This view of probability has implications for what kind of probability statements that are meaningful. Since the tendency of the coin of showing heads in a single toss is a constant quantity, it is, on the frequentist interpretation of probability, meaningless to claim that this quantity lies in a certain interval, between 1/3 and 2/3 for example, with a given probability. Either the quantity 'probability of heads' is in this interval, or it is not, the probability is 1 or 0, and nothing in between. In other words, the frequentist view of probability entails that we cannot make probability statements about the tendency of the coin to show heads. What we *can* make probability statements about is the tendency of the coin to show heads in 4 of 12 tosses, under a given assumption about the coin, for example that we are dealing with a fair coin.

Say we want to ascertain whether or not the 12 coin tosses in (1) stem from a coin that shows heads and tails with equal probability, that is a fair coin. The hypothesis we put to test, and under which the subsequent probabilities are computed, is called a null-hypothesis. So in this case our null-hypothesis is that the coin is fair. A very common way of assessing the evidence in data against a null-hypothesis is by a p-value. *A p-value is the probability of observing what we actually observed, or something even less in favour of the null-hypothesis, under the assumption that the null-hypothesis is true.* If this probability is below a predefined threshold, often 5 percent, we stop believing in the truth of the null-hypothesis.

The important thing to notice is that due to its reliance on the frequentist interpretation of probability, the probability given by the p-value concerns the outcome and possible outcomes(!) of an experiment, under a given assumption about a property of the coin, and not this property as such.

These features of frequentist statistics should be kept in mind when the likelihood principle, and two principles equivalent to it, are presented in the next section. After having presented and argued for the likelihood principle, I move on to explain why the frequentist approach to statistics violates this principle in Section 4.4.

4 Sufficiency, conditionality and likelihood

In the Introduction I claimed that when making inductive inferences one ought to use methods that do not lead to contradictions. The contradictions I have in mind, are possible contradictions between an inferential procedure and some foundational principles that every inductive inference should obey. Contrary to what is the case in deductive logic, where most logicians and mathematicians agree on the axiomatic basis provided by Zermelo-Fraenkel set theory (Suppes, 1972, ch. 1), there is no agreed upon principled foundation for inductive logic.[9] The lack of a foundation is, however, not due to a lack of trying, and in the following three sections I present what I deem to be the most successful attempt at providing a principled foundation for statistics.

In Section 4.1 and 4.2 I present two principles that are supposed to be so intuitively obvious that any inferential procedure that does not obey them should be dubious. The acceptance of these two principles forces one to accept at third principle, the likelihood principle, which is presented in Section 4.3. This principle has far ranging consequences for statistics, in particular it entails that the frequentist approach to statistics is not permissible.

4.1 The sufficiency principle

Look back at the sequence of coin tosses in (1). We are to use these tosses for making inferences about θ, the unknown probability of heads in a single toss. Of the twelve tosses, four came up heads and eight came up tails. The average number of heads is then 1/3. Suppose that this was all you knew - thus instead of knowing that the first toss came up tails, the second tails,..., the fifth heads, and so on, you only knew the number 1/3 - would you then have less, equally much or more information about the underlying θ? That you should have more information is clearly not the case, after all, from the entire sequence you can compute the average. Claiming that you have less information is equivalent to saying that you would have reached a different conclusion about θ with a sequence of tosses containing 4 heads and 8 tails *in a different order* than in (1). This would be strange (the independence assumption is crucial here). The conclusion is therefore that the full sequence

[9] I follow the Stanford Encyclopedia of Philosophy (Hawthorne, 2017), who defines a system of inductive logic as any system of inductive inference that obeys what is known as the *Criterion of Adequacy*: *As evidence accumulates, the degree to which the collection of true evidence statements comes to support a hypothesis, as measured by the logic, should tend to indicate that false hypotheses are probably false and that true hypotheses are probably true.*

in (1) and the average 1/3 contain the same amount of information about θ.

Summary statistics with this property are called *sufficient statistics*. Such statistics summarise the data in a manner that preserves all the information the data contain about the parameters. Formally, a sufficient statistic $S = S(x)$ is any function of the data x, such that the distribution of the data given S does not depend on θ.

In the coin tossing example it can be shown that the average number of heads have this property. We are now ready for the sufficiency principle.

Sufficiency principle: *Two observations x and y which are such that $S(x) = S(y)$, must lead to the same inference about θ.*

Consider the two sequences of independent coin tosses,

$$x = (T\,T\,T\,T\,H\,T\,H\,T\,T\,T\,H\,H\,T),$$
$$y = (T\,H\,T\,T\,T\,H\,T\,T\,T\,H\,T\,T\,H).$$

The average number of heads is 1/3 in both sequences, so with the notation used above, $S(x) = S(y) = 1/3$. According to the sufficiency principle the observations x and y should lead to the same conclusion.

4.2 The conditionality principle

You wake up early one morning and feel ill, thinking you might have a fever. A bit dazed you start searching around for your newly acquired thermometer, but you can't find it, give up, and settle for your old thermometer that is still in the drawer you put it in last winter. Being a biology student, you are well aware that no temperature measurement device is completely accurate, there is uncertainty attached to the number glimmering on the little screen. When making up your mind about whether to keep or reject your null-hypothesis of not being ill, should you take into account the fact that with a certain probability, you could have found the new and more accurate thermometer? Certainly not, is the obvious answer. After all, why would you let your decision - 'I'm in good shape' or 'I should stay in bed', be influenced by a temperature measurement that was never taken. This intuition is the essence of the following principle.

Conditionality principle: *If two experiments with the potential of yielding an inference about θ can be conducted, and one of the two experiments is randomly selected, then the resulting inference about θ should only depend on the selected experiment.*

4.3 The likelihood principle

The likelihood principle is not quite as intuitive as the two preceding ones. It is therefore of fundamental importance that it is equivalent with them. That is, *if you accept the sufficiency principle and the conditionality principle, then you must accept the likelihood principle* (Birnbaum, 1962; Berger and Wolpert, 1988; Robert, 2007). The reason for the likelihood principle not being quite as intuitive as the two preceding principles, is that it requires the notion of a likelihood function. We must start with this.

'Eirik Jensen would never have run the risk for such a small amount of money.'[10] The person uttering this sentence makes a case for Jensen not being corrupt (running the risk), because he thinks the probability of the data (a relatively small amount of Norwegian kroners) is low, *given that* Jensen is corrupt. So, under the hypothesis that Jensen is indeed corrupt, the data have a low probability, hence the hypothesis seems unlikely. This *conditional* probability can be expressed as

$$\Pr(\text{data} \,|\, \text{hypothesis}).$$

Notice that the person above argues for Jensen's innocence by considering the conditional probability as a function of the hypotheses. The argument is that

$$\Pr(\text{small money} \,|\, \text{corrupt}) < P(\text{small money} \,|\, \text{not corrupt}),$$

where it is the hypothesis that ranges over two competing hypotheses, while the datum (small money) is held constant. Generally, the function being used in this argument may therefore be written,

$$L(\text{hypothesis}) = \Pr(\text{data} \,|\, \text{hypothesis}), \tag{5}$$

where 'hypothesis' ranges over a set of mutually exclusive hypotheses. A conditional probability viewed as a function of the hypotheses (in which case it is no longer a probability in the strict sense of the word), is called a *likelihood function*. The likelihood principle states that,

Likelihood principle: *All the information about a hypothesis H obtained from an experiment is contained in the likelihood function $L(H)$. Two likelihood functions $L_1(H)$ and $L_2(H)$ that are proportional contain the same information abouth H.*

Some pharmacology students[11] claim that they have developed a new medicine that is superior to the old one. It is known that patients with some

[10] Jensen is a former Norwegian policeman. In 2017, he was convicted of corruption and drug trafficking, and sentenced to the maximum penalty of 21 years imprisonment.

[11] This example is originally due to Lindley and Phillips (1976), and appears in various forms in books on Bayesian statistics and the likelihood principle.

kind of cancer have a 70 percent chance of survival after treatment with the old medicine. The claim of the pharmacology students is based on an experiment where the new drug was administered to 10 patients, of which only one was not cured. This, the pharmacologists claim, gives a p-value of 0.04, leading to rejection of the null-hypothesis of the new drug being equally good or worse than the old drug. Pfixer, the company producing the old drug, disagrees with the students, claiming that there is no evidence to the effect that the new drug is superior to the old one. In fact, in a radio debate Pfixer's spokesperson tells the student representative to redo her introductory statistics course, because the correct p-value is 0.15, not 0.04 as the students claim.

Who is right? In the radio debate the student retorts that it is the people at Pfixer who ought to redo their introductory courses. The reason for this, she continues, is that with limited resources the students first tested the drug on one person, who was cured. So was the second, the third,..., and the ninth. To avoid saying that 100 percent of their test patients were cured, which would appear too good to be true, the students had decided to keep on recruiting new patients until they found one who was not cured by their drug. This was the tenth person. This experiment gives the likelihood function

$$L_{\mathrm{g}}(\theta) = \Pr(\#\text{cured until failure} = 9 \,|\, \theta) = \theta^9\,(1-\theta).$$

Under the null-hypothesis of $\theta = 0.7$, the probability of observing what the students observed, or even more cured patients, is then

$$(0.7)^9\,(1-0.7) + (0.7)^{10}\,(1-0.7) + (0.7)^{11}\,(1-0.7) + \cdots = 0.04. \tag{6}$$

Pfixer, on the other hand, based their claim on the supposition that the students had sampled 10 patients, administered the drug to them, and counted the number of cured people. This, would give the likelihood

$$L_{\mathrm{b}}(\theta) = \Pr(\#\text{cured} = 9 \,|\, \theta) = \binom{10}{9}\,\theta^9(1-\theta),$$

and a probability of an observation equally or more in favour of the students' drug, compared to the result actually obtained, of

$$10 \times (0.7)^9\,(1-0.7) + (0.7)^{10} = 0.15. \tag{7}$$

With a confidence level of 5 percent, the students rightly reject the null-hypothesis, while with the same confidence level, Pfixer just as correctly, does *not* reject the null-hypothesis. This means that the same data, 1 out of 10 which is 1 out of 10 however you might look at it, lead to conflicting

conclusions. Intuitively, this does not seem right, how can the same average lead to different conclusions? This intuition is in accordance with the sufficiency principle, because the average happens to be a sufficient statistic in both experiments.

As a consequence, this example shows that the classical frequentist methods used by both the students and by Pfixer are in conflict with the likelihood principle. So, to answer the question raised above - Who is right? - if one adopts the likelihood principle, both the students and Pfixer are equally wrong.

To see that the frequentist procedures are in conflict with the likelihood principle, notice that the ratio of the likelihoods is equal to 10 for all values of θ, $L_{\mathrm{b}}(\theta)/L_{\mathrm{g}}(\theta) = 10$, that is, they are proportional. With two proportional likelihoods at hand, the likelihood principle commands that the same inferences should be made about θ.

4.4 Summing over possible worlds

To see what it is about the p-values 0.04 and 0.15 that is problematic with respect to the likelihood principle, look back at equations (6) and (7) and notice how they are computed (or just read the definition of a p-value). Consider (7), restated here

$$\Pr(\#\text{cured} = 9) + \Pr(\#\text{cured} = 10) = 10 \times (0.7)^9 \, (1 - 0.7) + (0.7)^{10} = 0.15.$$

This is a sum of the data actually obtained (#cured = 9), and of a *more extreme* observation (#cured = 10), that was *not observed*. This means that data that could have been observed, but were in fact not, contribute to our conclusion. The same is true for the pharmacology students. In other words, the students use hypothetical data as evidence against the null-hypothesis, while Pfixer use hypothetical data as evidence in favour of the null-hypothesis. Imagine the defence lawyer of Eirik Jensen using a hypothetical scenario where Jensen received even less money than what he provably did receive, as evidence against the hypothesis that the former policeman is corrupt. It would not hold up.

For those scientists and statisticians who take the likelihood principle seriously, Bayesian statistical methods are often regarded as the solution. In the next section we will see why, while Section 5.1 problematises the subjective element of Bayesian statistics.

5 Bayesian statistics

Bayesian statistics is associated with the view that a probability expresses a persons subjective belief in a proposition. Consider the proposition 'I think there is a 10 percent chance that I will experience the singularity.'[12] This is a proposition that demands a non-trivial leap of thought if one holds the frequentist interpretation of probability. For this proposition to be meaningful on a frequentist account of probability, one has to imagine that we are living in one of infinitely many comparable universes, and that in 10 percent of these, the singularity does occur.[13]

Or what about a woman who claims to be 60 percent certain of being pregnant? This woman is either pregnant or not pregnant, she cannot be anything in between. On a strict frequentist account of probability, it is not easy to tell what the 60 percent the woman in question attaches to the proposition 'I'm pregnant', really means.

To rid oneself of these qualms, one must reject the strict frequentist account of probability, and accept that probabilities express something subjective.[14]

The woman who is 60 percent sure of being pregnant decides to take a pregnancy test. The test tells her that she is pregnant, but a pregnancy test is not always correct, so the woman wonders what probability she should ascribe to the proposition 'I'm pregnant', in view of the the positive test. Bayes' theorem tells her how this updating of beliefs should be carried out. Let H_0 denote the hypothesis 'I'm pregnant', and H_1 its negation. Prior to taking the test, the woman thought that $\Pr(H_0) = 0.6$ (called a *prior* probability for natural reasons). The probability she is interested in after taking the test is the so called *posterior* probability, written $\Pr(H_0 \,|\, \text{data})$, that is the probability of H_0 *given* data. Bayes' theorem gives that

$$\Pr(H_0 \,|\, \text{data}) = \frac{\Pr(\text{data} \,|\, H_0)\Pr(H_0)}{\Pr(\text{data} \,|\, H_0)\Pr(H_0) + \Pr(\text{data} \,|\, H_1)\Pr(H_1)}. \tag{8}$$

Since the datum in question is a positive test, the probability $\Pr(\text{data} \,|\, H_0)$ is the probability of the test telling the woman that she is pregnant when she is pregnant, while $\Pr(\text{data} \,|\, H_1)$ is the probability of the test telling the

[12] See Bostrom (2014) for more on the singularity.

[13] For some physicists and philosophers this might not be that much to swallow? See for example Deutsch (1997).

[14] Or one might claim that some probabilities are objective (frequentist) and some are subjective. For such a view, see Schweder and Hjort (2016) who distinguish between *aleatory* and *epistemic* probability. For Bayesians holding such a view, see Gelman and Robert (2013).

woman that she is pregnant when she is in fact not. The denominator is then the total probability of positive test.[15]

Now, notice that with the notation for the likelihood function introduced above, (i.e. $L(H) = \Pr(\text{data} \mid H)$), Bayes' theorem in (8) may be written

$$\Pr(H_0 \mid \text{data}) = \frac{L(H_0)\Pr(H_0)}{L(H_0)\Pr(H_0) + L(H_1)\Pr(H_1)}, \tag{9}$$

which makes it clear that the posterior probability $\Pr(H_0 \mid \text{data})$ only depends on the data actually observed, and not on data that could have been observed, which was the case with the p-values. Furthermore, suppose that we have two likelihood functions L_{g} and L_{b}, which are such that $L_{\text{g}}(H) = cL_{\text{b}}(H)$, for all hypotheses H and some constant c. Then it is easy to see that the posterior probability will be unaffected by whether we base the analysis on the one or the other likelihood function (the constant c appears in nominator and denominator of (9)).

The upshot of this is that if the pharmacology students and Pfixer had used Bayesian methods, they would - under one important condition! - reach the same conclusion about the effectiveness of the new medicine. That the pharmacology students used the likelihood function $L_{\text{g}}(\theta)$ and Pfixer the likelihood function $L_{\text{b}}(\theta)$ would not matter because these are proportional, as we saw $L_{\text{b}}(\theta) = 10L_{\text{g}}(\theta)$.

5.1 Subjectivity and the difficult prior distribution

The 'important condition' alluded to in the previous paragraph is that the pharmacology students and Pfixer use the same prior distribution. Without the same prior, they will arrive at different conclusions about the effectiveness of the new drug. This means that if pharmacology students and Pfixer are to agree on the conclusions, they must hold the same beliefs about the effectiveness of the new drug, *before any experiment has been carried out.* That's a tall order.

If the two parties fail to agree on a prior distribution, then Pfixer may accuse the students of having tarnished their analysis by letting their bias in favour of the new medicine sneak into their conclusions. This is subjective and unscientific. Similarly, the students may accuse Pfixer of giving the potential superiority of the new drug vanishingly little prior weight, simply to avoid competition on the market. Whom to believe?

[15] In medicine the probabilities $\Pr(\text{data} \mid H_0)$ and $1 - \Pr(\text{data} \mid H_1)$ are called the *sensitivity* and *specificity* of a test, respectively. These numbers are found in the user manual for a test.

	cooled	non-cooled
good	59	57
death/disability	19	22
total	78	79

Table 1: Data from Laptook et al. (2017), as given in STK4021 (2017).

This example highlights the difficulty of using Bayesian methods in scientific analyses. It simply seems wrong to let subjective prior opinions influence the conclusions. The example above is made up, but the difficulty is by all means real.

Take for example the article by Laptook et al. (2017) which recently appeared in JAMA. This article reports on a clinical trial investigating the effect of hypothermia administered between 6 and 24 hours after birth on death and disability from hypoxicischemic encephalopathy.[16] In some rare cases, newborns are deprived of oxygen to the brain due to a difficult birth. Cooling these kids down to a body temperature of 33 degrees Celsius just after birth, can save their life, with no later mental or motoric impairment. The controversy revolves around how long after birth this cooling action is still beneficial. Laptook et al. study the effect of cooling when it is initiated inside the time window 6 hours to 24 hours after birth, as opposed to the time window 0 to 6 hours after birth, which has been the recommandation so far (this summary is taken from STK4021 (2017)).

Due to the rarity of hypoxicischemic encephalopathy, and the fact that the cooling had to have taken place in the time window 6 to 24 hours after birth, or not at all, for a child to be eligible for the study, Laptook et al. were only able to recruit a limited number of kids. The data is given in Table 1. As can be seen from Table 1, the possible effect of cooling comes down to assessing the difference between

$$\frac{19}{78} = 0.244 \quad \text{and} \quad \frac{22}{79} = 0.278.$$

I think most people would agree that this difference is not dramatic. Now, the reason for this difference being so small might be because there is in fact no difference, or because the sample is rather limited. The challenge of a limited sample is what led Laptook et al. to use Bayesian methods to analyse their data. In the *Guide to Statistics and Methods* section of the same issue of

[16] I was made aware of this article by Nils Lid Hjort who has written a blog post on the matter (Hjort, 2017), contributed to a discussion (Walløe et al., 2017), and made an exam exercise about it (STK4021, 2017).

JAMA, Quintana et al. (2017) comment on Laptook et al., and write that due the limitations of a small sample, Laptook et al.

> [...] used a Bayesian analysis of the treatment effect to ensure that a clinically useful result would be obtained even if traditional approaches for defining statistical significance were impractical.

To a disapproving reader this sounds like 'since there is no difference in the data, we supplanted the data with our prior beliefs to make a difference'. How many such disapproving readers there are, I don't know, the point is that the objectivity of the finding - which Laptook et al. report as in favour of late cooling - might be questioned.[17]

5.2 Making Bayes objective

So what does the Bayesian retort when confronted by the frequentist with her inclination towards subjectivity? Somewhat simplified, there are two typical answers. One line of defense is to admit that, yes of course, Bayesian analyses are subjective, but so are all frequentist analyses as well, the frequentist only cover the subjectivity up better (Berger and Berry, 1988). There is definitively something to this line of defense. It is not as if all steps of a scientific process are completely objective, and then a prior distribution comes and destroys all this. On the contrary, every step in a research project - from building the model to collecting the data - is full of subjective choices that have to be made. Now, as briefly discussed in Section 2.1 the objectivity of this process may stem from the fact that it is carried out in a routine way, agreed upon by a community of researchers, and according to strict protocols. But if it is this that makes a finding objective, one might well imagine a community of researchers agreeing on the correct prior distribution for a given problem. Such a consensus on a prior distribution would make it qualify as concordantly objective (see Section 2.1).

Another response to the accusation of excessive Bayesian subjectivity, is attempts at making Bayesian analyses objective (see (Berger, 2006) for a review). Objective Bayesian analyses have traditionally been understood as analyses where the prior distribution does not favour certain values of the parameters at the expense of others.

One might for example imagine that the pharmacology students and Pfixer could agree on carrying out a Bayesian analyses where the prior dis-

[17] The quote above raises further questions. Why is it that a finding of no difference does not qualify as a 'clinically useful result'? And what is 'traditional approaches for defining statistical significance were impractical' supposed to mean? A p-value can always be defined, whether or not it indicates a significant finding is another matter.

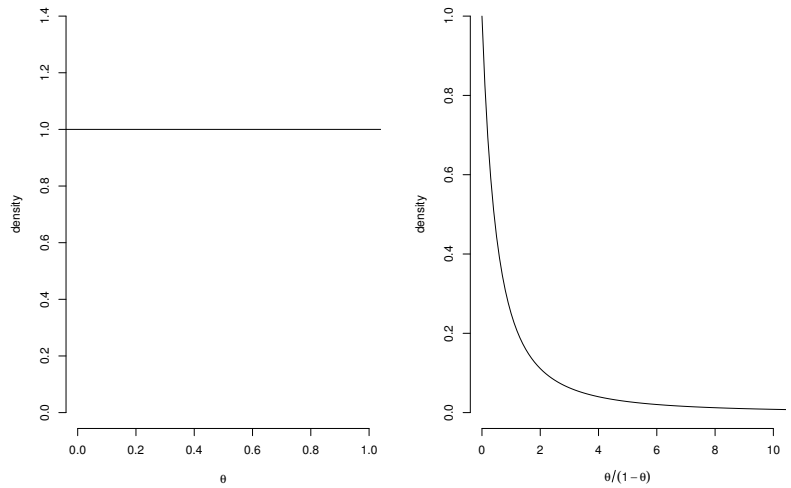

Figure 2: A uniform prior distribution for θ (left panel), and the resulting prior distribution for the odds $\theta/(1-\theta)$ (right panel).

tribution ascribed equal weight to all possible values of θ (recall that θ was the probability of an ill patient being cured by treatment, and that the old drug had $\theta = 0.7$). In this example, this would mean that the prior probability was smeared out equally over the interval from 0 to 1. Such a prior distribution is displayed on the left in Figure 2.

So far so good. Now, consider the much used measure in medicine and epidemiology, namely the odds. The odds of an event is the probability of the event divided by the probability of it not occurring, so in our example, the odds of an ill patient being cured by treatment is $\theta/(1-\theta)$. Since we have prior beliefs about θ, these prior beliefs obviously translate into prior beliefs about the odds. The problem is that the objective prior for θ, does not translate into an objective prior for the odds. In fact, the prior for θ that ascribes equal probability to all values of θ, translates into a prior for the odds which favours certain values of the odds above others. The resulting prior probability density for the odds is given on the right in Figure 2.

Figure 2 shows clearly how most of the probability mass is tilted in favour of small values of the odds, while larger values (those wished for by the Pharmacology students) receive far less prior weight. In summary, a prior that is objective for a parameter, does not translate into an objective prior for functions of that parameter. This is problematic, because it is clearly unreasonable to claim complete objectivity with regards to θ, while being

heavily biased when it comes to $\theta/(1-\theta)$. This is merely one of the challenges in the branch of statistics known as Objective Bayes.

6 Rationality vs. objectivity?

The argument made in this essay is that a system of inductive inference qualifies as rational if it obeys the definition of rationality given in Section 2.2, and does not contradict certain basic principles. The basic principles argued for in this essay are the sufficiency and the conditionality principle, which together imply the likelihood principle. In Section 4.3 I indicated why the dominant school of statistics, frequentism, is in conflict with the likelihood principle. In Section 2.2 it was also shown that incorporating prior beliefs into our decision procedures can bring us closer to reality, and I claimed that procedures that do not incorporate prior beliefs are in many cases irrational.

The strive for rationality led to the consideration of Bayesian methods. Bayesian inference is in accordance with the likelihood principle, and incorporates prior beliefs into decision procedures in a manner prescribed by the laws of probability. In Section 5.1 it was, however, shown that Bayesian methods may fail to qualify as scientifically objective.

It therefore appears as if the strive for rationality drives us to adopt methods that are scientifically dubious (Bayes), while the strive for objectivity drives us to accept methods (frequentism) that do not qualify as rational. This conflict has been recognised by many statisticians, and a general attitude seems to be that in certain situations Bayes is right, while in others one ought to be a frequentist ('In business I am a Bayesian', while in science 'I am a frequentist', for example). This is a dissatisfying state of affairs.

Many possible solutions to this conflict have been proposed, one of which is Objective Bayes. One major challenge of this approach has already been mentioned, another no lesser challenge is that the objective Bayesian often needs to work with prior (and even posterior) distributions that are not probabilities in the classical sense.[18]

Other inferential systems meant to resolve this conflict are the Error-Statistical account of Deborah Mayo (Mayo and Spanos, 2010), and methods based on the so-called Law of the Likelihood (Royall, 2000; Gandenberger, 2013). It is difficult (at least for me, at present) to tell how promising these approaches are, but so far they have not been adopted by the statistical community at large, and the statistical research on these topics is limited.

[18] Taraldsen and Lindqvist (2010) provide an introduction to a theory of probability intendended to tackle this problem.

The perhaps dispiriting solution I alluded to in the introduction is a Bayesian one. As hinted at in the text, I think that there is nothing in the way of Bayesian methods being applied in scientific analyses with the same claim to objectivity as frequentist methods.

This would require that the community of researchers agrees on what prior distributions are the correct ones for given problems, and that the updating of beliefs in a field of study is carried out in a routine and agreed upon way. That is, the community of researchers must agree on how, when a new study is conducted, one is to use the present pool of knowledge to inform the priors used in the new study.

This might appear as an insurmontable task, but when one thinks about the amount of tacit knowledge shared by a community of researchers when performing all other parts of a study, as well as when interpreting the results of new study, it might not be impossible.

References

Ayer, A. J. (1952). *Language, Truth & Logic*. Dover Publications, Inc., New York.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Berger, J. O. (2006). The Case For Objective Bayesian Analysis. *Bayesian Analysis*, 1:385–402.

Berger, J. O. and Berry, D. A. (1988). Statistical Analysis and the Illusion of Objectivity. *American Scientist*, 76:159–165.

Berger, J. O. and Wolpert, R. L. (1988). The Likelihood Principle. *Lecture notes-Monograph series*, 6:iii–199.

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory*. John Wiley & Sons.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, 57:269–306.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press, Oxford.

Deutsch, D. (1997). *The Fabric of Reality. The Science of Parallel Universes*. Allen Lane, New York.

Elster, J. (2010a). *Le Désintéressement,. Traité critique de l'homme économique I.* Éditions du Seuil, Paris.

Elster, J. (2010b). *L'irrationalité. Traité critique de l'homme économique II.* Éditions du Seuil, Paris.

Gandenberger, G. (2013). Why Likelihoodism is Not a Frequentist Methodology. *Greg Gandenbergers blog*, 18 June 2013. Available: `http://gandenberger.org/2013/06/18/likelihoodism-not-frequentist/` [Last accessed: 19 December 2017].

Gelman, A. and Robert, C. P. (2013). "Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference. *The American Statistician*, 67:1–5.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

Good, I. J. (1967). On the Principle of Total Evidence. *The British Journal for the Philosophy of Science*, 17:319–321.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, Cambridge.

Hawthorne, J. (2017). Inductive logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, `https://plato.stanford.edu/archives/spr2017/entries/logic-inductive/`, spring 2017 edition.

Hjort, N. L. (2017). Cooling of Newborns and the Difference Between 0.244 and 0.278. Online; accessed 17-December-2017, `http://www.mn.uio.no/math/english/research/projects/focustat/the-focustat-blog%21/`.

Kelly, T. (2003). Epistemic Rationality as Instrumental Rationality: A Critique. *Philosophy and Phenomenological Research*, 66:612–640.

Laptook, A., Shankaran, S., Tyson, J., and et al. (2017). Effect of Therapeutic Hypothermia Initiated After 6 Hours of Age on Death or Disability Among Newborns With Hypoxic-Ischemic Encephalopathy: A Randomized Clinical Trial. *Journal of the American Medical Association*, 318(16):1550–1560.

Lindley, D. V. and Phillips, L. (1976). Inference for a Bernoulli process (A Bayesian View). *The American Statistician*, 30:112–119.

Mayo, D. G. and Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press.

O'Neill, B. (2009). Exchangeability, Correlation, and Bayes' Effect. *International Statistical Review*, 77:241–250.

Quintana, M., Viele, K., and Lewis, R. J. (2017). Bayesian Analysis: Using Prior Information to Interpret the Results of Clinical Trials. *Journal of the American Medical Association*, 318(16):1605–1606.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, Berlin.

Royall, R. (2000). On the Probability of Observing Misleading Statistical Evidence. *Journal of the American Statistical Association*, 95:760–768.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press.

Sprenger, J. (2016). Bayesianism vs. frequentism in statistical inference. In Hájek, A. and Hitchcock, C., editors, *The Oxford Handbook of Probability and Philosophy*, pages 382–405. Oxford University Press.

Sprenger, J. (2017). The Objectivity of Subjective Bayesianism. PhilSci Archive, `http://philsci-archive.pitt.edu/13199/`.

Stigler, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

STK4021 (2017). Exam Project STK4021 - Applied Bayesian Analysis and Numerical Methods, University of Oslo. Exam written by Nils Lid Hjort. Online; accessed 17-December-2017, `http://www.uio.no/studier/emner/matnat/math/STK4021/h17/exam_stk4021_2017prosjekt.pdf`.

Stoltenberg, E. A. (2017). Frekventisme, Bayes og Likelihoodprinsippet. *Filosofisk supplement*, 13(2):58–65.

Suppes, P. (1972). *Axiomatic Set Theory*. Dover Publications, Inc., New York.

Taraldsen, G. and Lindqvist, B. H. (2010). Improper Priors Are Not Improper. *The American Statistician*, 64:154–158.

Walløe, L., Hjort, N. L., and Thoresen, M. (2017). Important data on effects of late hypothermia. Online; accessed 17-December-2017, http://www.mn.uio.no/math/english/research/projects/focustat/the-focustat-blog!/comments-to-laptook-jama.docx.

Yudkowsky, E. (2009). What Do We Mean By 'Rationality'? Online; accessed 18-January-2018, http://lesswrong.com/lw/31/what_do_we_mean_by_rationality/.