Classical vs. Bayesian statistics

Eric Johannesson

Department of Philosophy

Stockholm University

johannesson.eric@gmail.com

Forthcoming in *Philosophy of Science*

Abstract

In statistics, there are two main paradigms: *classical* and *Bayesian* statistics. The purpose of this paper is to investigate the extent to which classicists and Bayesians can (in some suitable sense of the word) *agree*. My conclusion is that, in certain situations, they can't. The upshot is that, if we assume that the classicist isn't allowed to have a higher degree of belief (credence) in a null hypothesis after he has rejected it than before, then (in certain situations), either he has to have trivial or incoherent credences to begin with, or fail to update his credences by conditionalization.

1 Introduction

In statistics, there are two main paradigms: *classical* and *Bayesian* statistics. A notorious problem with the Bayesian approach is the choice of *prior credences.* Due to Bertrand-style paradoxes, there doesn't seem to be any privileged way of choosing them. Classical statistics is, in a sense, an attempt to factor them out. A problematic consequence of this approach is its limited applicability. Another problem is how to interpret classical notions (such as *significance levels*) in terms of degrees of evidential support.[1] Setting these issues aside, the purpose of this paper is rather to investigate the extent to which classicists and Bayesians can (in some suitable sense of the word) *agree* in situations where classical statistics is applicable. More precisely: is there always a non-trivial prior credence distribution such that, for any possible observation, a Bayesian with that credence will agree with the classicist whether the observation speaks in favor of rejecting the hypothesis or not? The answer, as we will see, is no. In section 2 and 3, I present the classical and Bayesian frameworks, respectively. In section 4, I define what I take to be a reasonable notion of *agreement*, and construct a case in which agreement is impossible. In section 5, I present a more general diagnosis of the conflict between classical statistics and the Bayesian theory of rational degrees of belief.

2 Classical statistics

In classical statistics, hypotheses about the distribution of various properties in a population are tested by observing random samples of it. On the assumption that the

[1] Cf. Lindley (1957), Berger and Sellke (1987) and Casella and Berger (1987).

sample is random, the conditional probability $P(E|H)$ for every possible observation E and hypothesis H is usually derived as a matter of combinatorics. Given a particular hypothesis H_0 (a so called *null hypothesis*, which is the hypothesis to be tested), a rejection region for H_0 has to be selected, which is the set of outcomes leading to the rejection of H_0. Formally, if Ω is a non-empty set of possible outcomes (a sample space), Σ is a σ-algebra on Ω (an event space), and Θ is a non-empty set of hypotheses, there's a function $P : \Sigma \times \Theta \to [0, 1]$ such that, for each hypothesis $H \in \Theta$, the function $P(\cdot|H) : \Sigma \to [0, 1]$ is a probability function. A rejection region for a hypothesis $H_0 \in \Theta$ is an event $R \in \Sigma$ such that H_0 is rejected just in case an event $E \in \Sigma$ is observed such that $E \subseteq R$. The question is, what should this region look like?

Saying, for instance, that one should reject a hypothesis whenever an observation is made whose probability given that hypothesis is below a certain threshold will not do, since it means rejecting the hypothesis that a coin is fair given any sufficiently long sequence of observed tosses. Likewise, it means rejecting the hypothesis that the length of the male population is somewhat normally distributed around an average of 1.85 meters, after having observed a sample with exactly that average (in general, it's highly unlikely for the average of the sample to exactly coincide with the average of the population). There are essentially two classical approaches to this problem, one by Fisher and the other by Neyman and Pearson. I will look at each of them in turn.

2.1 Fisher

Instead of rejecting a hypothesis H just in case an event E is observed such that $P(E|H)$ is sufficiently low (which would be absurd), the general idea proposed by Fisher

(1925) is to reject H just in case $P(E \textit{ or something more extreme}|H)$ is sufficiently low. Formally, for a given hypothesis H, let $\prec_H$ be a strict partial order on Ω such that, for any $E \in \Sigma$, $\{x \in \Omega : y \prec_H x \text{ for some } y \in E\} \in \Sigma$. The intended interpretation of $x \prec_H y$ is that y *is more extreme than* x *with respect to* H. For any event $E \in \Sigma$, the so called *p-value* of E with respect to H and $\prec_H$ can then be defined as follows:

$$p(E|H, \prec_H) =_{df} P(E \cup \{x \in \Omega : y \prec_H x \text{ for some } y \in E\}|H) \tag{1}$$

Thus, the p-value of E with respect to H and $\prec_H$ is to be understood as *the probability of observing* E *or something more extreme under the assumption that* H *is true.* The lower the p-value, the more reason one has to reject H, according to Fisher. For instance, consider again the null hypothesis that the length of the male population is normally distributed around an average of 1.85 meter. Assuming that the measured average of a sample is to be considered more extreme the more it differs from 1.85, the p-value of measuring an average of exactly 1.85 meters is 1. Such a high p-value will not warrant the rejection of the null hypothesis.

But what makes an outcome more extreme than some other outcome with respect to given hypothesis H? In general, it would be unwise to simply stipulate that $x \prec_H y$ just in case $P(\{y\}|H) < P(\{x\}|H)$. For one thing, it might be the case that $\{x\}, \{y\} \notin \Sigma$, in which case the probabilities in question are undefined. Or it might be the case that $P(\{x\}|H) = 0$ for all $x \in \Omega$. That typically happens when Ω is the set $\mathbb{R}$ of real numbers, Σ is the set of Lebesgue measurable subsets of $\mathbb{R}$ and $P(\cdot|H) : \Sigma \to [0, 1]$ is defined as the integral of a continuous probability density function $f : \mathbb{R} \to \mathbb{R}$. But it's equally unwise to stipulate in such cases that $x \prec_H y$ just in case $f(y) < f(x)$, as the

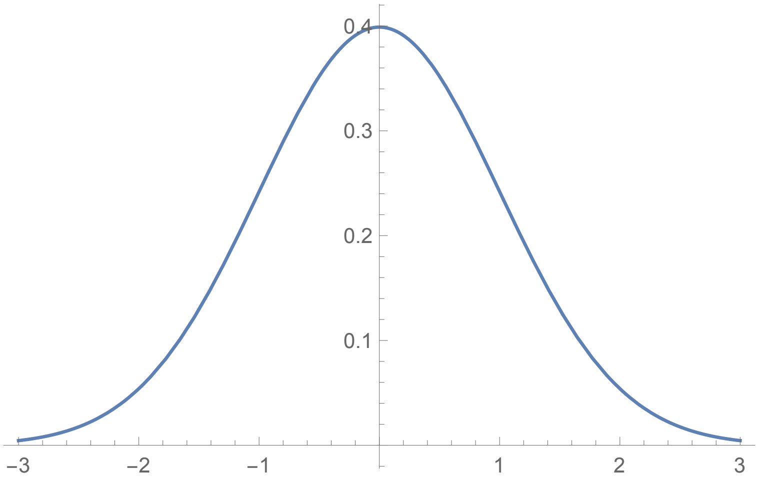

Figure 1: Expected distribution on the measured mean according to the null hypothesis, given by a probability density function $f : \mathbb{R} \to \mathbb{R}$.

following example illustrates:

Example 2.1 (The shooting range)**.** Suppose you want to test whether the aim on a rifle is straight or not (with respect to the horizontal direction). You go to a shooting range and fix the rifle aimed at a particular target. Your plan is to fire several bullets at the target, measure the horizontal distance in decimeters between the target and the point of impact for each bullet, and calculate the mean. Your null hypothesis H is that the aim is straight. Suppose that, on the assumption that the null hypothesis is true, due to random influences (in particular the shape of the bullet leaving the pipe), the sampled mean can be expected to be normally distributed around 0 with variance 1, as depicted in Figure 1. Any statistics textbook will tell you that an outcome y is to be considered more extreme than an outcome x relative to the null hypothesis just in case the absolute distance between y and the mean 0 is greater than the absolute difference between x and 0. In other words, for any $x, y \in \mathbb{R}$, $x \prec_H y$ just in case $|x| < |y|$. One

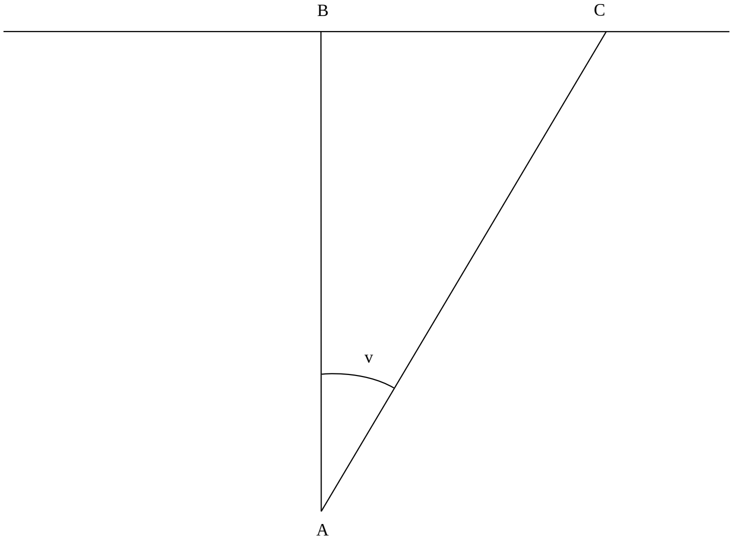

Figure 2: The measured mean represented by the angle $v \in (-90, 90)$.

might be tempted to infer this from the shape of the probability density function f in Figure 1 alone, since $f(y) < f(x)$ just in case $|x| < |y|$. To see why that would be unwise, let's look at a different but equivalent representation of the same situation, depicted in Figure 2. Let A be the point 1 decimeter from the target in the direction of the rifle, let B be the target, and let C be a point to the left or right of the target indicating the measured mean deviation, and consider the angle v between AB and AC, ranging in the open interval between -90 to 90 degrees. With the same random influences as before, the expected distribution on this angle is given by the probability density function g depicted in Figure 3. But here it's no longer the case that, for any angles $u, v \in (-90, 90)$, $g(v) < g(u)$ just in case $|u| < |v|$.

In conclusion, the relation of being a more extreme outcome with respect to a particular hypothesis cannot be inferred from the probability distribution associated

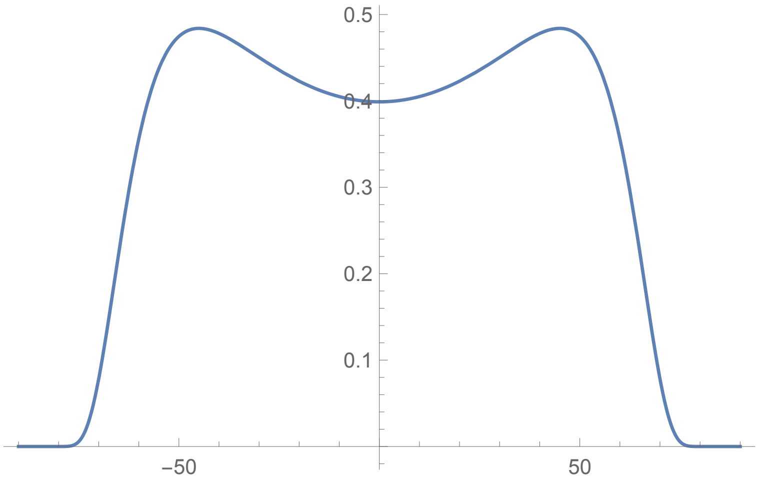

Figure 3: Expected distribution on the angle corresponding to the measured mean according to the null hypothesis, given by a probability density function $g : (-90, 90) \to \mathbb{R}$.

with the hypothesis in any obvious way. Although it's intuitively clear enough in practice what relation to use, it does make Fisher's theory less appealing from a philosophical point of view.

2.2 Neyman and Pearson

A philosophically more sophisticated theory is offered by Neyman and Pearson (1933a,b). The basic intuition is that, whether a hypothesis should be rejected depends on whether there are better alternatives, i.e. whether there are plausible enough alternative hypotheses according to which the probability of the observation is higher. More specifically, determining an adequate rejection region for the null hypothesis involves two notions: type 1 and type 2 errors. To commit a type 1 error is to reject a true null hypothesis. To commit a type 2 error is to not reject a false null hypothesis. Naturally, one does not want to reject a true null hypothesis. Hence, for each proposed

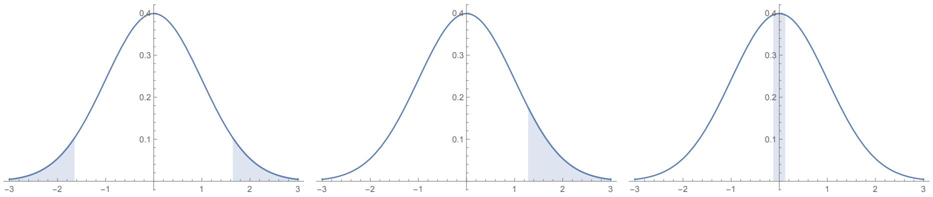

Figure 4: Three possible rejection regions of size 0.1 for the null hypothesis, R_1, R_2 and R_3.

rejection region $R \in \Sigma$, one may ask: what is the probability of rejecting the null hypothesis, on the assumption that it's true? The answer, of course, is $P(R|H_0)$. For some suitably small number $\alpha \in [0, 1]$ (e.g. $\alpha = 0.05$), it's therefore reasonable to require that $P(R|H_0) \leq \alpha$. However, if all we care about is minimizing the risk of making a type 1 error, then we should let $R = \emptyset$ and never reject any null hypothesis. In general, however, even if we require that $P(R|H_0) = \alpha$, this condition does not single out a unique rejection region. As illustrated by Barnett (1999, pp. 167-168), if $\alpha = 0.1$ and the null hypothesis yields a normal distribution with mean 0 and variance 1 over the sample space $\mathbb{R}$, the three regions depicted in Figure 4 all meet said requirement.

The solution, according to Neyman and Pearson, is to consider the probability of making a type 2 error. Relative to some alternative hypothesis $H \in \Theta$, the probability of rejecting the null hypothesis on the assumption that H is true (and hence H_0 is false) is given by $P(R|H)$. In order to minimize the risk of making a type 2 error, we want this probability to be as high as possible. Consider Figure 5. Suppose the alternative hypothesis H_1 yields a normal distribution with mean 1 and variance 1. Then $P(R_3|H_1) < P(R_1|H_1) < P(R_2|H_1)$. Hence, with H_1 considered as the alternative hypothesis, region R_2 is clearly the better choice. One can even show that, for every alternative region R such that $P(R|H_0) \leq \alpha$ and $P(R - R_2|H_1) \neq 0$, we have

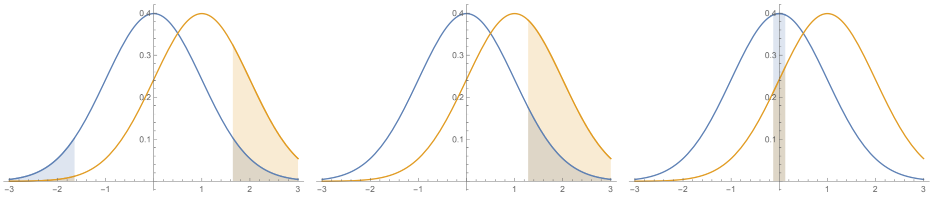

Figure 5: R_2 is a most powerful test of size 0.1 for H_0 against H_1.

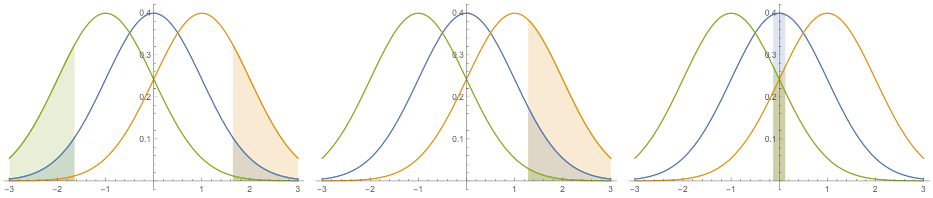

Figure 6: R_2 is not a most powerful test for H_0 against H_{-1} (R_1 is more powerful).

$P(R|H_1) < P(R_2|H_1)$. In this sense, R_2 dominates every other rejection region. It is a so called *most powerful test of size* α for H_0 against the alternative H_1. As a matter of fact, R_2 is a most powerful test of size α against *every* alternative hypothesis yielding a normal distribution with mean above zero and variance 1. Against this particular set of alternative hypotheses, R_2 is a so called *uniformly most powerful test of size* α for H_0. Intuitively, R_2 thereby makes for a natural non-arbitrary choice of rejection region. Indeed, it's the one Neyman and Pearson recommend. From a Bayesian point of view, they even show that such a rejection region is *optimal independently of prior probabilities* in the sense that, whatever the prior probabilities on hypotheses are, the probability of making a type 2 error can only increase by switching to another rejection region of the same size (Neyman and Pearson, 1933b).

Sometimes, however, no uniformly most powerful test exists. For instance, in the presence of an alternative hypothesis H_{-1} yielding a normal distribution with mean -1 and variance 1, R_2 is dominated by R_1, as can be seen in Figure 6. Hence, against any set of alternative hypotheses including both H_1 and H_{-1}, no uniformly most powerful test exists for H_0. This fact, of which Neyman and Pearson were aware, limits the applicability of their method. The lack of a uniformly most powerful tests in certain situations is a problem for classical statistics, but not one I will address in this paper.

3 Bayesian statistics

Compared with its classical counterparts, Bayesian statistics is straightforward. Basically, it falls out from the more general Bayesian theory of rational degrees of belief (rational credences), comprised of the following two postulates:

1. Rational credences are *coherent* (in the sense of satisfying the laws of probability).

2. Rational credences are updated by conditionalizing on evidence.

However, the theory only tells us how to rationally adjust our prior credences in relation to new evidence. It does not tell us what our prior credences should be. Hence, it does not tell us in absolute terms what our posterior credences should be in relation to new evidence. Unlike classical statistics, it doesn't tell us whether a particular hypothesis should be rejected or not.

Initially, this may lead one to suspect that classical statistics harbors some implicit assumptions about what kind of prior credences one may have. On this picture, a classicist would just be a Bayesian with a particular kind of prior credences. As we shall see, however, matters are much worse. We we will show that, in a certain rather natural sense, there are situations where classical and Bayesian statistics are incompatible.

I will now present what I take to be the essence of Bayesian statistics. For technical reasons, one usually distinguishes between the case where the set of hypotheses is countable (the discrete case) and when it's continuum-sized (the continuous case). For our purposes, it's enough to consider the discrete case. Thus, let Θ be a countable non-empty set of hypotheses, each of which is associated with a probability function $P(\cdot|H) : \Sigma \to [0,1]$ on an event space Σ (just as in classical statistics), and let $\mathrm{Cr} : \mathcal{P}(\Theta) \to [0,1]$ be a probability function representing the prior credence in each hypothesis. The idea is to extend this function to the mixed domain $\Sigma \times \mathcal{P}(\Theta)$ by something like *the principal principle* (Lewis, 1980). Hence, we assume that, for each event $E \in \Sigma$ and hypothesis $H \in \Theta$,

$$\mathrm{Cr}(E\&H) =_{\mathit{df}} P(E|H)\mathrm{Cr}(H) \tag{2}$$

and, for each $\Delta \subseteq \Theta$,

$$\mathrm{Cr}(E\&\Delta) =_{\mathit{df}} \sum_{H\in\Delta} \mathrm{Cr}(E\&H) \tag{3}$$

For the sake of legibility, we write $\mathrm{Cr}(E\&H)$ instead of $\mathrm{Cr}(\langle E, \{H\}\rangle)$, and $\mathrm{Cr}(H)$ instead of $\mathrm{Cr}(\{H\})$. For the same reason, we write $\mathrm{Cr}(E\&\Delta)$ instead of $\mathrm{Cr}(\langle E, \Delta\rangle)$.

Now, for any observation $E \in \Sigma$ such that $\mathrm{Cr}(E\&\Theta) \neq 0$, the posterior credence Cr_E is then given for each $H \in \Theta$ by conditionalizing on E:

$$\mathrm{Cr}_E(H) =_{\mathit{df}} \frac{\mathrm{Cr}(E\&H)}{\mathrm{Cr}(E\&\Theta)} =_{\mathit{df}} \frac{P(E|H)\mathrm{Cr}(H)}{\sum_{H\in\Theta} P(E|H)\mathrm{Cr}(H)} \tag{4}$$

The following fact is easy to establish, and will be used in the next section:

Fact 3.1. *Provided that $0 < Cr(H_0) < 1$, we have $Cr(H_0) < Cr_E(H_0)$ if, for all*

$H \in \Theta - \{H_0\}$, $P(E|H) < P(E|H_0)$. Likewise, we have $Cr_E(H_0) \leq Cr(H_0)$ if, for all $H \in \Theta - \{H_0\}$, $P(E|H_0) \leq P(E|H)$.

4 Agreement

In order to compare classical and Bayesian statistics, let's assume that we have a situation where there is a uniformly most powerful test for the null hypothesis against the alternatives. We shall focus on the Neyman-Pearson version of classical statistics, but also say something about Fisher's version as we go along. Now, a classical statistician is not in the business of assigning rational credences to hypotheses. One may still wonder: are there any credences he might have that would (in some suitable sense of the word) *agree* with his statistical methods? In the discrete case, the following definition of agreement seems natural:

Definition 4.1 (Agreement)**.** Let Θ be a countable non-empty set of hypotheses, each of which is associated with a probability function $P(\cdot|H) : \Sigma \to [0, 1]$ on an event space Σ, and let $\mathrm{Cr} : \mathcal{P}(\Theta) \to [0, 1]$ be a probability function representing the initial credences of the Bayesian. Assume that the null hypothesis $H_0 \in \Theta$ has uniformly most powerful test $R \in \Sigma$ of size $\alpha \in [0, 1]$. We then say that the Bayesian and the classicist *agree* (with respect to α) just in case there's no event $E \in \Sigma$ such that $E \subseteq R$ but $\mathrm{Cr}(H_0) < \mathrm{Cr}_E(H_0)$.

In plain English: the Bayesian and the classicist agree just in case there's no observation prompting the classicist to reject the null hypothesis while prompting the Bayesian to increase his credence in it.

Observe that, by (4), if $\mathrm{Cr}(H_0) = 1$ or $\mathrm{Cr}(H_0) = 0$, then the Bayesian and the classicist will agree trivially, since $\mathrm{Cr}(H_0) = \mathrm{Cr}_E(H_0)$ for all $E \in \Sigma$ for which Cr_E is defined. They will also agree trivially when $\alpha = 0$, since that means $P(E|H_0) = 0$ for any $E \subseteq R$. What we're interested is of course whether they can agree *non-trivially.* By Fact 3.1, it follows that non-trivial agreement is guaranteed in the situation depicted in Figure 5, with rejection region R_2 for H_0 against H_1. However, for any given $\alpha > 0$, there are situations where the Bayesian and the classicist *cannot* agree non-trivially. Here's such an example:

Example 4.1. Suppose you want to test the null hypothesis (H_0) that 1/3 of all members of a population have property F against the alternative hypothesis (H_1) that 2/3 have it. You decide to perform a random sample of 50 individuals and count the number of F:s in that sample. The expected number of F:s according to each hypothesis is given by a binomial distribution, depicted in Figure 7. A most powerful test for the null hypothesis of size 0.05 against the alternative is an event where the sample contains anything between 23 and 50 F:s. Let E be the event that the sample contains 23 F:s. Thus, using a test of size 0.05, the event E should lead to a rejection of the null hypothesis. However, since $P(E|H_1) < P(E|H_0)$, it follows by Fact 3.1 that non-trivial agreement is impossible. Whatever his prior credences are (assuming that $0 < \mathrm{Cr}(H_0) < 1$), the Bayesian will deem the null hypothesis more likely after observing E than before.

Clearly, for any $\alpha > 0$, a similar situation may arise if we make the sample size n large enough. Likewise, in the case of Fisher's theory of p-values, assuming that $x \prec_{H_0} y$ just in case $P(\{y\}|H_0) < P(\{x\}|H_0)$ for any $x, y \in \{0, 1, ..., n\}$, we can for any $\alpha > 0$ find

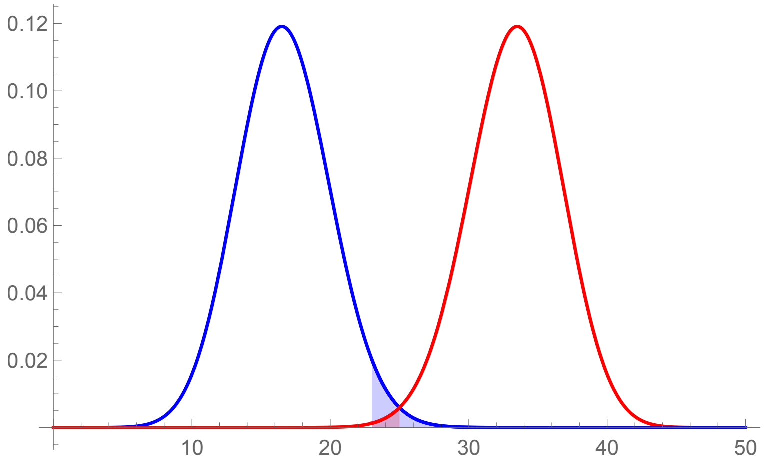

Figure 7: The expected number of F:s in a random sample of 50 individuals according to the null hypothesis (left) and the alternative hypothesis (right). A most powerful test of size 0.05 for the null hypothesis against the alternative is indicated by the blue area. In this case, Bayesians and classicists cannot agree.

a large enough sample size n for which an event E exists whose p-value is below α, although the event in question will prompt the Bayesian to increase his credence in H_0. This is essentially just a version of Lindley's paradox (Lindley, 1957). In the original paradox, however, a null hypothesis H_0 is tested against a set of alternatives $\{H_\mu : \mu > 0\}$ (or, as statisticians like to put it, a precise null hypothesis is tested against an *imprecise* or *composite* alternative). In order to generate the original paradox (i.e. in order to derive the conclusion that the Bayesian will increase his credence in the rejected null hypothesis), one has to assume that the Bayesian assigns a non-zero prior credence to it. But since the set of hypotheses is uncountable, this assumption is not very plausible, and it's possible to obtain agreement by simply denying it. In that case, the result is perhaps less problematic for the classicist. According to Spanos (2013), it merely reveals that powerful tests are more sensitive, which is why we can use them to detect small changes in a population. Moreover, as argued by Sprenger (2013), the

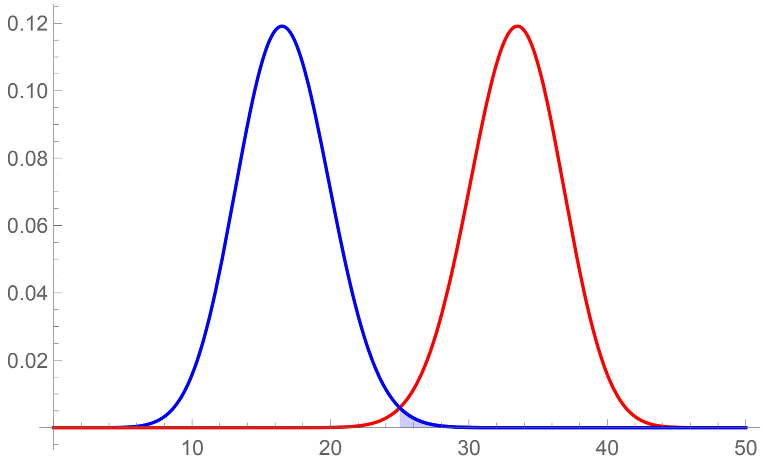

Figure 8: A most powerful test of size 0.02 for the null hypothesis (left) against the alternative (right) is indicated by the blue area. In this case, Bayesians and classicist must agree.

original paradox may also be a problem for the Bayesian. But in the example at hand, where a precise null hypothesis is tested against an equally precise alternative, the result is more clearly problematic only for the classicist. Intuitively, there's no sense in which the observation of 23 F:s favors the alternative H_1 over H_0.

The example indicates that, from a classical point of view, it may be unwise to settle for any particular test size in advance (e.g. $\alpha = 0.05$), regardless of how small it is. Indeed, for any particular sample size n, there is a most powerful test of *some* size (e.g. $\alpha = 0.02$ when $n = 50$) for H_0 against H_1 with respect to which Bayesians and classicists *can* and, in fact, *must* agree. Such a situation is depicted in Figure 8. According to Mayo and Spanos (2006, p. 345, footnote 21), some statisticians have even suggested that the significance level should be adjusted as a function of the sample size.[2] As a

[2] That's not the solution the authors themselves recommend, however. Although they don't provide any examples, a suggestion of this sort can indeed be found in Pérez and

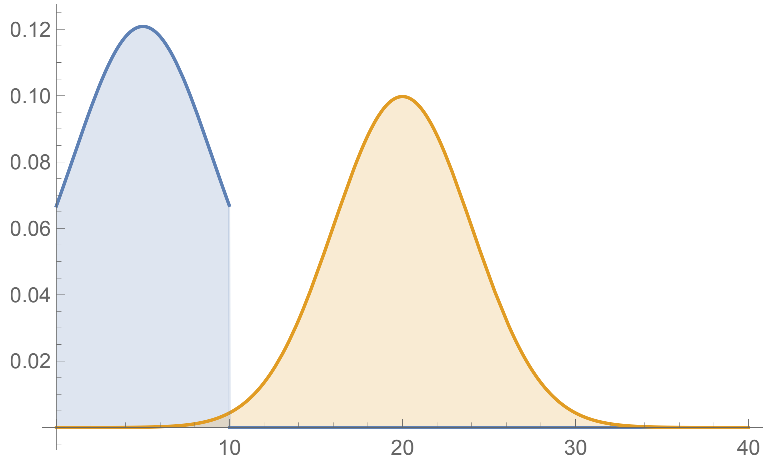

Figure 9: The distribution of length in a population according to the null hypothesis (left) and the alternative (right).

general solution, however, that suggestion is problematic. As witnessed by the following example, it's possible to construct a situation where there's no $\alpha > 0$ with respect to which Bayesians and classicists can agree:

Example 4.2. Let the null hypothesis H_0 be that the length of individuals in a certain population ranges between 0 and 10, and that it's (approximately) normally distributed with mean 5 and some variance v. Let the alternative hypothesis H_1 be that it ranges between 0 and 40, and that it's (approximately) normally distributed with mean 20 and the same variance v. the situation is depicted in Figure 9. According to the null hypothesis, the average length of a random sample will range between 0 and 10 and be (approximately) normally distributed with mean 5 and some variance $v' \leq v$. According to the alternative, it will range between 0 and 40 and be (approximately) normally distributed with mean 20 and the same variance v'. The situation is depicted in Figure 10. Clearly, for any $\alpha > 0$, a most powerful test of size α of the null hypothesis against

Pericchi (2014).

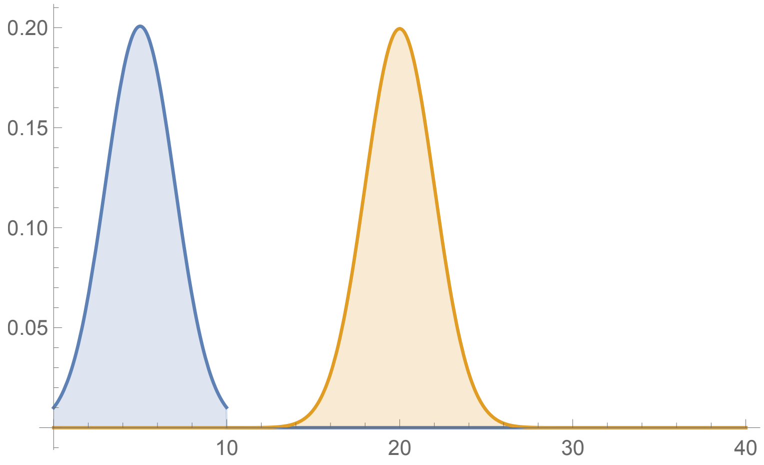

Figure 10: The distribution of average length of a random sample according to the null hypothesis (left) and the alternative (right).

the alternative is given by the rejection region $[a, 40]$, for some $a < 10$. However, for each such region R, there's an event $E \subseteq R$ such that $P(E|H_0) > P(E|H_1)$, namely $E = [a, 10]$. Hence, by Fact 3.1, for each $\alpha > 0$, non-trivial agreement is impossible.

The general features making non-trivial agreement impossible in this example are the following:

The null hypothesis is associated with a probability density function f on a sample space $[p, q] \cup [q, r] \subseteq \mathbb{R}$, where $p < q < r$, such that

(a) $f(x) = 0$ for all $x \in (q, r]$, and

(b) each alternative hypothesis is associated with a probability density function g on the sample space such that

   i. $f(x) > g(x)$ for all $x \in [p, q]$, and

   ii. $f(x)/g(x)$ is strictly decreasing on $[p, q]$.

Features (a) and (b)-ii guarantee that, for each $\alpha > 0$, a uniformly most powerful test of that size is given by the interval $[a, r]$, for some $a < q$. Feature (b)-i guarantees that the event $[a, q] \subseteq [a, r]$ is more probable according to the null hypothesis than according to any alternative. By Fact 3.1, this event will therefore prompt the Bayesian to increase his his credence in the null.

In summary, the example describes a situation where, for each $\alpha > 0$, there's a most powerful test of size α of the null hypothesis against the alternative, and for which there's an event E such that

1. The null hypothesis is rejected by E according to the Neyman-Pearson theory,
2. The p-value of E is smaller than α according to Fisher's theory[3], and
3. Conditionalizing on E will increase one's credence in the null hypothesis given any non-trivial assignment of prior credences to the null hypothesis and its alternative.

The upshot is that, if we assume that a classical statistician isn't allowed to increase his credence in the null hypothesis after he has rejected it, he either has to have trivial or incoherent credences to begin with, or fail to update his credences by conditionalization.

Some classicists might respond that although an event with a sufficiently low p-value always indicates *some* discrepancy from the null, it doesn't indicate the large discrepancy represented by the alternative in this case. The problem with this response is that, under the assumptions in the case at hand, any discrepancy from the null entails the alternative. So anything indicating the former indicates the latter. Alternatively, they might argue that the testing of a precise null hypothesis against an equally precise

[3] Assuming that $x \prec_{H_0} y$ just in case $|x - 5| < |y - 5|$, for all $x, y \in [0, 40]$.

alternative is “artifical” or “illegitimate”.[4] If so, I fail to see in what sense. Surely it’s possible to know that either of two such hypotheses obtain, and to gather evidence for or against them by means of random sampling. A general theory of statistical inference should be able to explain how and why.

5 A diagnosis

Example 4.1 and 4.2 illustrate (as does Lindley’s paradox) a basic conflict between classical and Bayesian statistics. The conflict is that the former (in the case of Neyman-Pearson, where $R \subseteq \Omega$ is a uniformly most powerful rejection region) equates the evidential value of a piece of information $E \subseteq \Omega$ to the evidential value of $E \cup R$, or (in the case of Fisher) equates it to the evidential value of E *or something more extreme*. Their evidential values are equated in the sense that they license exactly the same classical statistical inferences, respectively: E is sufficient for rejecting the null hypothesis just in case $E \cup R$ is, and the p-value of E is the same as the p-value of E *or something more extreme*. But E is generally not the same information as $E \cup R$ or E *or something more extreme*. For the Bayesian, they generally don’t have the same evidential value. The point is a familiar one. Jeffreys (1980) put it this way:[5]

> I have always considered the arguments for the use of [p-values] absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable [under that hypothesis]; that is, that it has not predicted something that has not

[4] Cf. Mayo and Spanos (2006, 2011).

[5] I’m grateful to Uwe Saint-Mont for pointing this out to me.

happened.

Indeed, as the examples show, the null hypothesis may assign an arbitrarily low probability to *E or something more extreme*, and yet assign a higher probability to E than any alternative hypothesis. By Fact 3.1, the Bayesian will then increase his credence in the null hypothesis (given any non-trivial prior credences) although it's been rejected by the classicist.

6 Conclusion

Suppose that a classical statistician isn't allowed to have a higher degree of belief (credence) in a null hypothesis after he has rejected it than before. We have shown that, in certain situations, either he has to have trivial or incoherent credences to begin with, or fail to update his credences by conditionalization. Remember that, in the trivial case, he is either absolutely certain that the null hypothesis is true, or absolutely certain that it's false. For obvious reasons, that case is irrelevant in this context. In the non-trivial case, there are plenty of good arguments to the effect that the classical statistician is irrational. I'm thinking primarily of synchronic and diachronic Dutch book arguments, which I shall not rehearse here. At any rate, there's a conflict between classical statistics and the Bayesian theory of rational degrees of belief.

References

Barnett, V. (1999). *Comparative Statistical Inference.* Probability and statistics. Wiley.

Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association 82*(397), 112–122.

Casella, G. and R. L. Berger (1987). Reconciling bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association 82*(397), 106–111.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Studies in Bayesian econometrics: Essays in honor of Harold Jeffreys*, pp. 451–453. North-Holland.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, pp. 83–132. University of California Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*(1/2), 187–192.

Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science 57*(2), 323–357.

Mayo, D. G. and A. Spanos (2011). Error statistics. In P. S. Bandyopadhyay and M. R. Forster (Eds.), *Philosophy of Statistics*, Volume 7 of *Handbook of the Philosophy of Science*, pp. 153 – 198. Amsterdam: North-Holland.

Neyman, J. and E. S. Pearson (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231*(694-706), 289–337.

Neyman, J. and E. S. Pearson (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society 29*(4), 492–510.

Pérez, M.-E. and L. R. Pericchi (2014). Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics & probability letters 85*, 20–24.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science 80*(1), 73–93.

Sprenger, J. (2013). Testing a precise null hypothesis: The case of Lindley's paradox. *Philosophy of Science 80*(5), 733–744.