# Bayesianism vs. Frequentism in Statistical Inference

Jan Sprenger*

May 15, 2014

## Contents

**Motivation: The Discovery of the Higgs Particle** **2**

**1 Bayesian Inference** **4**

**2 Frequentism: Principles and Significance Tests** **5**

**3 Frequentism: p-values** **8**
3.1 The Base Rate Fallacy . . . . . 11
3.2 p-values and Bayesian measures of evidence . . . . . 12
3.3 Are most published research findings false? . . . . . 14
3.4 Confidence intervals . . . . . 15

**4 The Likelihood and Stopping Rule Principles** **16**

**5 Discussion: The Search for Objectivity** **18**

**Conclusion** **20**

*Contact information: Tilburg Center for Logic, General Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

## Motivation: The Discovery of the Higgs Particle

Bayesianism and frequentism are the two grand schools of statistical inference, divided by fundamentally different philosophical assumptions and mathematical methods. In a nutshell, **Bayesian inference** is interested in the credibility of a hypothesis given a body of evidence whereas frequentists focus on the reliability of the procedures that generate their conclusions. More exactly, a **frequentist inference** is valid if in the long run, the underlying procedure rarely leads to a wrong conclusion. This line of reasoning also dominates scientific practice. To understand the role of Bayesianism in statistical inference, and to evaluate its prospects for improving scientific reasoning, we have to appreciate the principles, advantages and drawbacks of the frequentist school of statistics. Clarifying them and contrasting them to the principles of Bayesian inference is the purpose of the article.

For understanding the difference in scope between Bayesianism and frequentism, Royall's (1997, 6) distinction between three main questions in statistical analysis is helpful:

1. What should we *believe*?
2. What should we *do*?
3. When do data count as *evidence* for a hypothesis?

Typically, no statistical school treats all three questions on a par. Bayesians focus on the first question—rational belief—because for them, scientific hypotheses are an object of personal, subjective uncertainty. They are concerned with the question of how data should change our degree of belief in a uncertain hypothesis. Consequently, Bayesians answer the second and third question—what are rational decisions and good measures of evidence?—within a formal model of rational belief provided by the probability calculus. Frequentists, on the other hand, are united in rejecting the use of subjective degree of belief in science. For them, the first question ("what should we believe?") is not a properly scientific one. Still, there is no agreement on which of the other two questions is more fundamental: The famous statisticians Jerzy Neyman and Egon Pearson emphasize the role of decision theory for statistical inference and construct scientific hypothesis tests as reliable decision procedures. Others, such as Ronald A. Fisher, stress the relevance of (post-experimental) evidential assessments and refuse to combine statistical inference with decision theory.

I begin this article with a recent example how the divergences between Bayesianism and frequentism can have marked impacts on the assessment of scientific findings. On 4 July, 2012, the CERN (European Center for

Nuclear Research) at Geneva surprised the public by the announcement to have discovered the *Higgs boson*—a particle in the Standard Model of modern physics, which had been searched for since 1964. Since the discovery of Higgs boson proved the existence of a particular mechanism for breaking the electroweak symmetry, the discovery was of extreme importance for particle physics.

In the statistical analysis, the researchers at CERN reasoned in frequentist spirit: under the assumption that the Higgs boson did not exist, the experimental results deviated more than five standard deviations from the expected value. Since such an extreme result would occur by chance only once in two million times, the statisticians concluded that the Higgs boson had indeed been discovered.

This analysis sparked a vivid debate between Bayesian and frequentist statisticians. The well-known Bayesian statistician Tony O'Hagan sent an email to the newsletter of the International Society for Bayesian Analysis (ISBA) where the entire statistical analysis was heavily attacked:

> We know from a Bayesian perspective that this [a frequentist evidence standard, J.S.] only makes sense if (a) the existence of the Higgs boson [...] has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. Neither seems to be the case [...]. Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is? (O'Hagan 2012)

O'Hagan's message prompted a vivid exchange in the ISBA forum, with prominent statisticians and particle physicists taking part in the discussion. In the first place, the debate concerned a specific standard of evidence, but since this notion depends on the chosen statistical framework, it quickly developed into a general dispute about the merits of Bayesian and frequentist statistics. Thus, the discovery of the Higgs particle exemplifies how the interpretation of a fundamental scientific result depends on methodological issues about statistical inference. Such cases are not limited to particle physics: they occur in every branch of science where statistical methods are used, and they include issues as applied as the admission process for medical drugs.

In this contribution, we focus on the interpretation of statistical evidence. This is not only the most contested ground between Bayesians and frequentists, but also utterly relevant for statisticians, experimenters, and scientific policy advisors. The article is structured as follows: Section 1 summarizes the principles of Bayesian inference. Sections 2 explains the

principles of frequentist procedures while Section 3 deals with the controversy between Bayesians and frequentists. Section 4 deals with the Likelihood and the Stopping Rules Principle while Section 5 discusses objections to subjective Bayesianism and intermediate positions between Bayesianism and frequentism.

# 1 Bayesian Inference

The basic assumption of Bayesian inference is to interpret probability as rational degree of belief. That is, an agent's system of degrees of belief is represented by a probability function $p(\cdot)$, and $p(H)$ quantifies his or her degree of belief that hypothesis $H$ is true. That degrees of belief should satisfy the probability calculus can be defended in various ways: either by pointing to consequences of actions guided by non-probabilistic degrees of belief (these are the Dutch Book Theorems, cf. Kemeny 1955), or by pointing to the resulting loss in accuracy (Joyce 2003; Leitgeb and Pettigrew 2010a,b).

Probabilistic degrees of belief are changed in the light of incoming information. The degree of belief in hypothesis $H$ after learning evidence $E$ is expressed by the conditional probability of $H$ given $E$, $p(H|E)$:

**Bayesian Conditionalization:** The rational degree of belief in a proposition $H$ after learning $E$ is the conditional probability of $H$ given $E$: $p_{\text{new}}(H) = p(H|E)$.

$p(H)$ and $p(H|E)$ are called the **prior probability** and **posterior probability** of $H$. They can be related by means of **Bayes' Theorem:**

$$p_{\text{new}}(H) := p(H|E) = p(H)\frac{p(E|H)}{p(E)} \tag{1}$$

The terms $p(E|H)$ and $p(E|\neg H)$ are called the **likelihoods** of $H$ and $\neg H$ on $E$, that is, the probability of the observed evidence $E$ under a specific hypothesis, in this case $H$ or $\neg H$.

The label "Bayesian inference" is usually associated with the following principles:

- The representation of subjective degrees of belief in terms of probabilities.
- The use of Bayesian Conditionalization for rationally revising one's degrees of belief.
- The use of the posterior probability distribution for assessing evidence, accepting hypotheses and making decisions.

However, not all Bayesians agree with these principles. In an objective Bayesian framework (e.g., Jeffreys 1939; Bernardo 2012), prior probabilities need not express subjective uncertainty; their assignment may also be guided by convenient mathematical properties, such as transformation invariance. Conditionalization is rejected by Bayesians who accept the Principle of Maximum Entropy as a guide to determining rational beliefs, and according to which rational degrees of belief should be the most middling and equivocating ones that are compatible with the empirical evidence (Williamson 2010). See Section 5 for more details on these approaches. Finally, some statisticians put more emphasis on Bayesian measures of evidence (e.g., the Bayes factor, see Kass and Raftery 1995; Goodman 1999b) than on the shape of the posterior distribution. To simplify things, we focus on the classical subjectivist position in Bayesian statistics that is built on the conjunction of the above principles.[1]

## 2 Frequentism: Principles and Significance Tests

In the 19th century, probability theory gradually extended its scope from games of chance to questions of data analysis in science, industry and public administration. This was perhaps the birth hour of modern inferential statistics. It emerged as a tool for handling complex datasets and for quantifying lack of precision in prediction and measurement. Due to the prevalent ideal that science should strive for certainty and provide an objective, impartial view of reality, Bayesian inference was eschewed by many founding fathers of modern statistics. Subjective degree of belief was, after all, thought to be quite different from objective evidence. The great statistician Ronald A. Fisher (1935, 6–7) even spoke of "mere psychological tendencies, theorems respecting which are useless for scientific purposes". Fisher clarified that he did not believe Bayesian reasoning to be logically invalid, but that there is rarely any reliable information on which a non-arbitrary prior probability distribution could be based.

But how should one infer from data to theory if the road via Bayes' Theorem is blocked? Here, the main innovation of frequentist statistics pops in: belief updating is replaced by **hypothesis testing**. In their groundbreaking 1933 paper, the British statisticians Jerzy Neyman and Egon Pearson connected statistical inference to rational decision-making by developing a genuinely frequentist approach to hypothesis testing: decisions about the "acceptance" or "rejection" of a scientific hypothesis should be made as to minimize the relative frequency of wrong decisions

[1] See Howson and Urbach (2006) for a philosophical introduction, and Bernardo and Smith (1994) for a more mathematically oriented treatment.

in a hypothetical series of repetitions of the test. That is, the adequacy of a conclusion would not be secured by its high posterior probability, but by the fact that it emerged from a reliable decision rule.

An example may illustrate their approach. Suppose that we must decide whether a medical drug is going to make it to the next phase in a cumbersome and expensive drug admission process. Of course, we would not like to admit a drug that is no better than existing treatments in terms of efficacy, side effects, costs, and so on. On the other hand, we do not want to erroneously eliminate a superior drug from the admission process. These are the two possible kinds of errors, commonly called **type I** and **type II error**.

The standard procedure for such tests consists in choosing a default or **null hypothesis** $H_0$. Often, this hypothesis states that the experimental intervention has no effect on the target variable. By contrast, the **alternative** $H_1$ states that such an effect is present. While a type I error corresponds to erroneous rejection of the null hypothesis, a type II error stands for erroneous acceptance of the null. Conventionally, acceptable type I error rates are set at a **level** of 5%, 1% or 0.1%, although Neyman and Pearson insist that these levels have no special meaning, and that striking the balance between type I and type II error rates is a highly context-sensitive endeavor. We may then be rational in following the test procedure because of its favorable long-run properties:

> [...] 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. (Neyman and Pearson 1967, 142)

While Neyman and Pearson engaged in the project of finding tests with optimal properties, another grandfather of frequentist statistics, the eminent geneticist and statistician Ronald A. Fisher, violently opposed the entire behavioral, decision-theoretic approach to statistical inference. According to Fisher, determining an acceptable type I error rate implies an implicit assessment of the severity of an error, thereby imposing a decision-theoretic utility structure on the experiment in question. Fisher argued, however, that

> in the field of pure research no assessment of the cost of wrong conclusions [...] can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence. (Fisher 1935, 25–26)

Two arguments are implied here. First, that quantifying the overall scientific utility of a statistical inference is a pretty impossible task. The ultimate consequences of an acceptance or rejection of the tested hypothesis are beyond the epistemic horizon of any scientist. Hence, decision-theoretic models are not apt for describing scientific inference. Second, statistical hypothesis tests should state the *evidence* for or against the tested hypothesis and not be obscured by the practical consequences of working with a particular hypothesis. In other words, while Neyman-Pearson tests may be helpful in industrial quality control and other applied contexts, they do not investigate the *truth* of scientific hypotheses and are no suitable tool for (pure) scientific research.

As an alternative, Fisher (1956) invented the paradigm of **significance tests**. For Fisher, the purpose of statistical analysis consisted in assessing the relation of a (null) hypothesis to a body of observed data. That hypothesis usually stands for the absence of an interesting phenomenon, e.g., no causal relationship between two variables, no observable difference between two treatments, etc. In remarkable agreement with Popper's falsificationist methodology, Fisher states that the only purpose of an experiment is to "give the facts a chance of disproving the null hypothesis" (Fisher 1925, 16) and that failure to reject a hypothesis does not conclude positive evidence for the tested (null) hypothesis. But unlike Popper (1934/59), Fisher aims at experimental and statistical *demonstrations* of a phenomenon. Thus, he needs a criterion for when an effect is real and not an experimental fabrication. As such a criterion, he suggests the incompatibility of the null hypothesis with the observed data, as measured by their improbability under the null:

> "either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true." (Fisher 1956, 39)

This basic scheme of inference, called **Fisher's Disjunction** by Hacking (1965), stands at the heart of significance testing. The occurrence of such an exceptionally rare chance may have both epistemological and practical consequences: first, the null hypothesis is rendered "objectively incredible" (Spielman 1974, 214), second, the null should be treated as if it were false. Notably, Fisher's approach is essentially asymmetric: while a "rejection" strongly discredits the null hypothesis, an "acceptance" just means that the facts have failed to disprove the null. By contrast, Neyman-Pearson tests are essentially symmetric in the interpretation of the outcome. They give an explicit role to alternative hypotheses whereas Fisher's approach is focused on the null hypothesis only.

The scheme of inference inherent in Fisher's Disjunction, a sort of probabilistic modus tollens (see also Gillies 1971), has been criticized fre-

quently. Hacking (1965, 81–82) has pointed out that the explication of the term "exceptionally rare chance" inevitably leads into trouble. If it means that the observed event must be exceptionally unlikely compared to other events, some statistical hypotheses could never be tested. For instance, a uniform distribution over a finite set of events assigns equal probability to all observations. How should we test—and possibly reject—such a hypothesis in Fisher's framework?

To expand on this point, imagine that we test the hypothesis that a particular coin yields independent and identically distributed outcomes with equal probability of heads and tails. Compare now two series of outcomes: 'HTTHTTTHHH' and 'HHHHHHHHHH'. The probability of both events under the null is the same, namely $(1/2)^{10} = 1/1024$. Still, the second series, but not the first, seems to strongly speak against the null. Why is this the case? Implicitly, we have specified *the way in which the data are exceptional*: we are interested in the propensity $\theta$ of the coin to come up tails, rather than in questioning the independence between the tosses or another tacit assumption of the test. Therefore we can restrict our attention to the observed number of tails $T$, and indeed, $\{T = 0\}$ is indeed a much less likely event than $\{T = 5\}$ (cf. Royall 1997, ch. 3).

It seems that we cannot apply significance tests without an implicit specification of alternative hypotheses; here: that the coin is biased toward heads or tails. Spielman (1974) further presses this point in an extended logical analysis of significance testing: inferring from an unlikely result to the presence of a significant effect *presupposes* that the observed result is much more likely under an implicitly conceived alternative than under the null. Indeed, modern frequentist approaches, such as Mayo's (1996) error-statistical theory, take this into account by explicitly setting up statistical inference in a contrastive way. That is, testing always occurs with respect to a direction of departure from the tested hypothesis. This insight is important when we compare Bayesian and frequentist measures of evidence.

## 3 Frequentism: p-values

Significance testing in the Fisherian tradition is arguably the most popular methodology in statistical practice. But there are important distinctions between Fisher's original view, discussed above, and the practice of significance testing in the sciences, which is a hybrid between the Fisher and the Neyman-Pearson school of hypothesis testing, and where the concept of the **p-value** plays a central role.

To explain these differences, we distinguish between a **one-sided** and a

**two-sided testing problem**. The one-sided problem concerns the question of whether an unknown parameter is greater or smaller than a particular value ($\theta \leq \theta_0$ vs. $\theta > \theta_0$), whereas the two-sided testing problem (or point null hypothesis test) concerns the question of whether or not parameter $\theta$ is exactly equal to $\theta_0$: $H_0 : \theta = \theta_0$. vs. $H_1 : \theta \neq \theta_0$. The two-sided test can be used for asking different questions: first, whether there is "some effect" in the data (e.g., if the null denotes the absence of a causal relationship), second, whether $H_0$ is a suitable proxy for $H_0 \vee H_1$, that is, whether the null is a predictively accurate idealization of a more general statistical model.

The central concept of modern significance tests—the p-value—is now illustrated in a two-sided testing problem. Again, we want to infer the presence of a significant effect in the parameter $\theta$ if the discrepancy between data $x := (x_1, \ldots, x_N)$, corresponding to $N$ realizations of an experiment, and null hypothesis $H_0 : \theta = \theta_0$ is large enough. Assume now that the variance $\sigma^2$ of the population is known. Then, one measures the discrepancy in the data $x$ with respect to the postulated mean value $\theta_0$ by means the standardized statistic

$$z(x) := \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N \cdot \sigma^2}} \tag{2}$$

We may re-interpret equation (2) as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}} \tag{3}$$

Determining whether a result is significant or not depends on the p-value or **observed significance level**, that is, the "tail area" of the null under the observed data. This value depends on $z$ and can be computed as

$$p := p(|z(X)| \geq |z(x)|), \tag{4}$$

that is, as the probability of observing a more extreme discrepancy under the null than the one which is actually observed. Figure 1 displays an observed significance level $p = 0.072$ as the integral under the probability distribution function.

For the frequentist practitioner, p-values are practical, replicable and objective measures of evidence against the null: they can be computed automatically once the statistical model is specified, and only depend on the sampling distribution of the data under $H_0$. Fisher interpreted them as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher 1956, 43).

The virtues and vices of significance testing and p-values have been discussed at length in the literature, and it would go beyond the scope

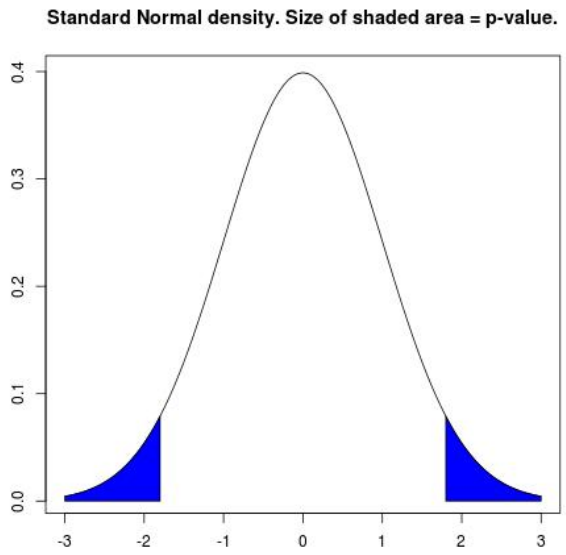


Figure 1: The probability density function of the null $H_0 : X \sim N(0,1)$, which is tested against the alternative $H_1 : X \sim N(\theta,1)$, $\theta \neq 0$. The shaded area illustrates the calculation of the p-value for observed data $x = 1.8$ ($p = 0.072$).

of this article to deliver a comprehensive discussion (see e.g., Cohen 1994; Harlow et al. 1997). However, some important criticisms and the relation to Bayesian inference will be discussed below.

### 3.1 The Base Rate Fallacy

The biggest problem with p-values is arguably practical: many researchers are unable to interpret them correctly. Quite often, a low p-value (e.g., $p < 0.001$) is taken as the statement that the null hypothesis has a posterior probability smaller than that number (e.g., Oakes 1986; Fidler 2005). But of course, the p-value is essentially the probability of some piece of evidence given the null hypothesis, $p(E|H_0)$, which is quite different from the conditional probability of the null hypothesis given the evidence, $p(H_0|E)$.

Let us illustrate this fallacy, and the reasons why it may be attractive, in a simple example. Consider a blood donor who is routinely tested for an HIV infection. Let the null hypothesis state that the donor has not contracted HIV. The test returns the correct result in 99% of all cases, regardless of whether an HIV infection is present or not. Now, the test returns a positive result. Under the null, this hypothesis certainly constitutes an exceptionally rare chance whereas under the alternative, it is very likely. Should the donor now be convinced that he has contracted HIV, given a general HIV prevalence of 0.01% in the population?

A Bayesian calculation yields, perhaps surprisingly, that he should still be quite certain of *not* having contracted HIV:

$$\begin{aligned} & p(\text{HIV contraction}|\text{positive test}) \\ = & \left(1 + \frac{p(\text{positive test}|\text{no contraction})}{p(\text{positive test}|\text{HIV contraction})} \frac{p(\text{no contraction})}{p(\text{HIV contraction})}\right)^{-1} \\ = & \left(1 + \frac{0.01}{0.99}\frac{0.9999}{0.0001}\right)^{-1} \approx 0.01 \end{aligned}$$

In other words, the evidence for a contraction is more than cancelled by the very low base rate of HIV infections in the relevant population. Therefore, straightfowardly rejecting the null hypothesis on the basis of a "significant" finding is no valid probabilistic inference, even if the findings are likely under the alternative. Since the fallacy is caused by neglecting the HIV base rate in the populations, it is called the **Base Rate Fallacy** (cf. Goodman 1999a).

Despite persistent efforts to erase the Base Rate Fallacy, it continues to haunt statistical practitioners. Some have argued that this is an effect of the unintuitive features of the entire frequentist framework. For example, the German psychologist Gerd Gigerenzer (1993) argues that scientists are

primarily interested in the tenability or credibility of a hypothesis, not in the probability of the data under the null. The question is then: how should we relate p-values to Bayesian measures of evidence? After all, a Bayesian and a frequentist analysis should agree when prior probability distributions can be objectively grounded. The comparison of Bayesian and frequentist measures of evidence is therefore not only a mathematical, but also a philosophically significant endeavor.

### 3.2 p-values and Bayesian measures of evidence

Bayesians base their beliefs and decisions on the posterior probability distribution over the hypotheses of interest, but what do they use as a measure of evidence, as a way of expressing how much the data support the null over the alternative (or vice versa)? In Bayesian statistics, a particular measure is used almost universally: the **Bayes factor**, that is, the ratio of prior and posterior odds between hypothesis $H_0 : \theta \in \Theta_0$ and alternative $H_1 : \theta \in \Theta_1$ conditional on data $x$ (Kass and Raftery 1995).

$$B_{01}(x) \quad := \quad \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{\int_{\theta \in \Theta_0} p(x|\theta)p(\theta)d\theta}{\int_{\theta \in \Theta_1} p(x|\theta)p(\theta)d\theta}.^{2} \tag{5}$$

Thus, for two composite hypotheses $H_0$ and $H_1$, the Bayes factor can be written as the ratio of the integrated likelihoods, weighted with the prior plausibility of the individual hypotheses. This measure is appealing for several reasons. Crucially, one can derive the posterior probability of a hypothesis $H$ when one knows its prior $p(H)$ and the Bayes factor of $H$ vs. $\neg H$. In the case of simple point hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, the Bayes factor reduces to the **likelihood ratio** $L(x, H_0, H_1) = p(x|H_0)/p(x|H_1)$ which possesses some interesting optimality properties as a measure of evidence (Royall 1997; Lele 2004).

Notably, Bayes factors and p-values can disagree substantially. To illustrate the paradox, we give an example from parapsychological research (Jahn, Dunne and Nelson 1987). The case at hand involved the test of a subject's claim to affect a series of randomly generated zeros and ones ($\theta_0 = 0.5$) by his extrasensory capacities (ESP). The subject claimed that by sheer mental power, he would make the sample mean differ significantly from 0.5.

A very large dataset (N = 104.490.000) was collected to test this hypothesis. The sequence of zeros and ones, $X_1, \ldots, X_N$, was described by a Binomial model $B(\theta, N)$. The null hypothesis asserted that the results were generated by a machine operating with a chance of $H_0 : \theta = \theta_0 = 1/2$, whereas the alternative was the unspecified hypothesis $H_1 : \theta \neq 1/2$.

Jahn, Dunne and Nelson (1987) report that in 104.490.000 trials, 52.263.471 ones and 52.226.529 zeros were observed. Frequentists would now calculate the $z$-statistic which is

$$z(x) := \sqrt{\frac{N}{\theta_0(1-\theta_0)}} \left( \frac{1}{N} \sum_{i=1}^{N} x_i - \theta_0 \right) \approx 3.61$$

and reject the null hypothesis on the grounds of the very low *p-value* it induces:

$$p := P_{H_0}(|z(X)| \geq |z(x)|) \ll 0.01$$

Thus, the data would be interpreted as strong evidence for the presence of extrasensory capacities.

Compare this now to the result of a Bayesian analysis. Jefferys (1990) assigns a conventional positive probability $p(H_0) = \varepsilon > 0$ to the null hypothesis, a uniform prior over the alternative, and calculates the evidence that $x$ provides for $H_0$ vis-à-vis $H_1$. This is quantified by the Bayes factor $B_{01}$:

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} \approx 12$$

Hence, the data clearly favor the null over the alternative and do *not* provide evidence for the presence of ESP.

That Bayesian and frequentist inferences may pull into completely different directions as sample sizes increases is a phenomenon known as **Lindley's paradox** (Lindley 1957). But how can we explain it? An important reason is that statistically significant results are not automatically good indicators of the **size of an effect**. The random generator which generates the sequences of zeros and ones will in practice not be perfect and have a tiny bias. A test of significance will detect this in a very large sample and will, with high confidence, reject the null hypothesis $H_0$ that the population mean is *precisely* equal to .5. But such a conclusion blurs the difference between statistical and scientific significance: the effect may be negligible. With respect to the conglomerate of alternative hypotheses (some of them including non-trivial effect sizes), $H_0$ may still be supported. This is the intuition which feeds the Bayes factor analysis and which explains the divergence between both results (Sprenger 2013a; Robert 2014b). To wrap up, a low p-value does not indicate, from a Bayesian point of view, that the hypothesis has become less tenable.

This phenomenon is more than a purely theoretical problem: it leads to frequent misperceptions of statistical significance as a substantial effect, jeopardizing the validity of the drawn conclusions (McCloskey and Ziliak 1996; Ziliak and McCloskey 2008). By scrutinizing statistical practice in the leading economics journal *American Economic Review*, as well as by

surveying the opinion of economists on the meaning of statistical significance, McCloskey and Ziliak derive the conclusion that most economists are unaware of the proper meaning of statistical concepts. In practice, "asterisking" prevails: e.g., in correlation tables, the most significant results are marked with an asterisk, and these results are the ones that are supposed to be real, big, and of practical importance. But an effect need not be statistically significant to be big and remarkable, and a statistically significant effect can be quite small and uninteresting (like in the ESP example).

There is also theoretical research on relating p-values to posterior probabilities. It turns out that in the one-sided testing problem, p-values can often be related to posterior probability (Casella and Berger 1987) whereas in the two-sided testing problem, the two usually diverge. More precisely, Berger and Sellke (1987) show that the p-value is proportional to a *lower bound* on the posterior probability of the null, thus systematically overstating the evidence against the null. This suggests a principal incompatibility between frequentist and Bayesian measures of evidence in the two-sided testing problem. See the chapter on Bayesian statistics by Christian Robert (2014a) for more detail on this point.

### 3.3 Are most published research findings false?

A methodological problem with p-values, stemming from their roots in Fisherian significance testing, is that insignificant results (=p-values greater than .05) barely have a chance of getting published. This is worrisome for at least two reasons: first, even a statistically insignificant result may conceal a big and scientifically relevant effect; second, it prevents an appraisal of the evidence *in favor of the null hypothesis*. As a consequence, valuable resources are wasted because different research teams replicate insignificant results over and over again, not knowing of the efforts of the other teams. In addition, the frequentist provides no logic of inference for when an insignificant result supports the null, rather than just failing to reject it.

This asymmetry in frequentist inference is at the bottom of Ioannidis' (2005) famous thesis that "most published research findings are false". Ioannidis reasons that there are many false hypotheses that may be erroneously supported and yield a publishable research finding. If we test for significant causal relationships in a large set of variables, then the probability of a false positive report is, for type I and type II error rates $\alpha$ and $\beta$, normally larger than the probability that a true hypothesis is found. In particular, if $R$ denotes the ratio of true to false relationships that are tested in a field of scientific inquiry, $T$ a particular causal relationship and

$E$ significant evidence in favor of $T$, then

$$p(T|E) = \frac{p(E|T) \cdot p(T)}{p(E|T) \cdot p(T) + p(E|\neg T) \cdot p(\neg T)} = \frac{(1-\beta) \cdot R}{(1-\beta) \cdot R + \alpha} \qquad (6)$$

This quantity is smaller than $1/2$ if and only if $R < \alpha/(1-\beta)$ which will typically be satisfied, given that $\alpha = .05$ is the standard threshold for publishable findings, and that most causal relationships that scientists investigate are not substantial. Thus, most published research findings are indeed artifacts of the data and plainly false. This effect is augmented by bad research practices that lead to bias in selecting, processing and analyzing data sets (cf. Francis 2014). Notably, this phenomenon is not a feature of scientific inquiry in general, but specifically due to the frequentist logic of statistical inference: the achievement of a significant result is just not a very good indicator for the objective credibility of a hypothesis. Indeed, researchers often fail to replicate findings by another scientific team, and periods of excitement and subsequent disappointment are not uncommon in frontier science. To counter this bias, the use of Bayesian measures of evidence may be a suitable antidote since they naturally trade off theoretical expectations with the observed findings (e.g., Goodman 1999b).

### 3.4 Confidence intervals

This section concludes with a brief note on confidence intervals: functions that map an observed value $x_0$ to an interval $C_{x_0}$. Often, they are recommended as an improvement over hypothesis tests and as a solution of the foundational problems of the frequentist framework (e.g., Cumming and Finch 2005).

A confidence interval at level $\alpha \in [0, 1]$ does not mean that upon observing $x_0$, parameter $\theta$ lies in the interval $C_{x_0}$ with probability $\alpha$. After all, frequentists do not assign posterior probabilities to specific values of the parameters of interest. Rather, the level of the confidence interval says something about the procedure used to construct it: for each $\theta$, we construct an interval $C_\theta$ such that the data $x$ will, in the long run, be *consistent* with $\theta$ in $100 \cdot \alpha\%$ of all cases. Projecting the set $\{(x|\theta)|x \in C_\theta\}$ to the actually observed value $x_0$ delivers the confidence interval $C_{x_0}$ for $\theta$.

A crucial advantage of confidence intervals over significance tests is that considerations relating to effect size are taken into account. In the above ESP example, where a very low p-value contrasted with a high posterior probability of the null hypothesis, the 95% or 99% confidence interval for $\theta$ would have been a very narrow interval in the neighborhood of $\theta_0$. That is, under the conditions of large sample size with low effect

size, a confidence interval would avoid the false impression that the null was substantially mistaken and should be rejected.

However, confidence intervals cannot be commended as an ideal solution. First, the idea that scientists conduct real *tests* in order to check the adequacy of a statistical model has completely vanished. But this is a vital and omnipresent aspect of scientific practice. Second, confidence intervals rather fulfill the function of a consistency check than of inspiring trust in a specific estimate. They list the set of parameter values for which the data fall would not have been rejected at level $1 - \alpha$. This is in essence a pre-experimental perspective. But this do not warrant, post-experimentally, that the parameter of interest lies "probably" in the confidence interval. It is precisely the problem of not having a foundationally sound post-experimental measure of evidence that troubles frequentist statistics, and that might, in the long run, tilt the balance in favor of the Bayesian approach.

## 4 The Likelihood and Stopping Rule Principles

The previous arguments against frequentist inference all have, either implicitly or explicitly, presupposed a Bayesian perspective. However, is there a way to ground Bayesian inference without assuming what is supposed to be shown?

A famous attempt to this end was made by Birnbaum (1962). His argument was built on the **Conditionality Principle**: it states that evidence gained in a probabilistic mixture of experiments is equal to the evidence in the actually performed experiment. Less abstractly, assume that we have to choose which of two different clinical trials, $\mathcal{E}_1$ and $\mathcal{E}_2$, that we plan to conduct. To decide the issue, we throw a fair coin. The coin turns up heads and we conduct $\mathcal{E}_2$, but not $\mathcal{E}_1$. Birnbaum now demands that in our evaluation of the results, only the evidence obtained from $\mathcal{E}_2$ should count. That is, the fact that another experiment would have been performed if the coin had turned up tails should be immaterial to the conclusions we draw.

The plausible idea behind this principle is that the coin toss does not matter at all to the inference problem we are trying to solve. Hence, we can condition on its outcome. If we find the Conditionality Principle plausible, then we can combine it with the Sufficiency Principle—an innocuous principle of statistical inference that both Bayesians and frequentist accept—to derive the

> **Likelihood Principle (LP):** Consider a statistical model $\mathcal{M}$ with a set of probability measures $p(\cdot|\theta)$ parametrized by $\theta \in \Theta$. Assume we conduct an experiment $\mathcal{E}$ in $\mathcal{M}$. Then,

> all evidence about $\theta$ generated by $\mathcal{E}$ is contained in the *likelihood function* $p(x|\theta)$, where the observed data $x$ are treated as a constant. (Birnbaum 1962; Berger and Wolpert 1984).

To clarify, the **likelihood function** takes as argument the parameters of a statistical model, yielding the probability of the actually observed data under those parameter values. In particular, the LP entails that the probability of outcomes which have not been observed does not matter for the statistical interpretation of an experiment. See Berger and Wolpert (1984); Mayo (2010); Gandenberger (2014) for discussions of the principle and the various ways of proving it. Notably, a subjective Bayesian automatically endorses the LP: all that is needed to update a prior to a posterior is the likelihood of $H$ and $\neg H$ given the observed data. In a statistical inference problem, this corresponds to the probability of the data $x$ under various values of the unknown parameter $\theta$. Frequentists, however, also consider information beyond the likelihood function to be evidentially relevant (e.g., the chance of observing an even less likely result under the null) and therefore disagree with the Conditionality Principle and the LP.

The LP is more than a purely theoretical principle: it is vital for the interpretation of sequential trials in medicine. There, **stopping rules** describe under which conditions an experiment about the efficacy of a medical drug should be terminated. For example, we may terminate the trial when a certain sample size is reached, or whenever the results clearly favor one of the two tested hypotheses. The dissent between defenders and opponents of the LP, and between Bayesians and frequentists, concerns the question of whether an inference about the efficacy of the drug should be sensitive to the specific stopping rule used.

For adherents of LP, stopping rules have no evidential role. Berger and Berry (1988, 34) call this the **Stopping Rule Principle**. Since the likelihood functions of the parameter values under different stopping rules are proportional to each other (proof omitted), considerations pertaining to the design of an experiment are evidentially irrelevant. The justification is that

> The design of a sequential experiment is [..] what the experimenter actually *intended* to do. (Savage 1962, 76. Cf. Edwards, Lindman and Savage (1963, 239).)

In other words, since such intentions are "locked up in [the experimenter's] head" (ibid.), not verifiable for others, and apparently not causally linked to the data-generating process, they should not matter for sound statistical inference.

However, from a frequentist point of view, certain stopping rules, such as sampling on until the result favors a particular hypothesis, lead us to biased conclusions (cf. Mayo 1996, 343–345). In other words, neglect of stopping rules in the evaluation of an experiment can make us *reason to a foregone conclusion*. Consider a stopping rule that rejects a point null $H_0 : \theta = \theta_0$ in favor of $H_0 : \theta \neq \theta_0$ whenever the data are significant at the 5% level. With probability one, this event will happen at *some* point, independent of the true value of $\theta$ (Savage 1962; Mayo and Kruse 2001). In addition, it has been argued that certain stopping practices in medicine (e.g., stopping when a drug proves to be superior to a placebo) leads to implausibly high effect size estimates.

On the other hand, it has been shown that reasoning to a foregone conclusion is only possible if frequentist rather than Bayesian measures of evidence are adopted (Kadane et et al. 1996). That is, the frequentist charge that optional stopping introduces bias is countered by the argument that *this is only so if a frequentist understanding of statistical evidence is adopted*. If the Bayesian take on optional stopping is combined with Bayesian measures of evidence, there is no problem (Sprenger 2009). Like the debate about frequentist vs. Bayesian measures of evidence, that is p-values vs. Bayes factors, the methodological question about the evidential relevance of stopping rules seems to be caught in a stalemate. Each side of the debate presupposes (or has to presuppose) their own inferential framework in order to criticize their rivals. The next section therefore deals with intermediate positions between subjective Bayesianism and classical frequentism that aim at an objective concept of statistical evidence without wearing the lens of a specific framework.

## 5 Discussion: The Search for Objectivity

Given all the benefits of Bayesianism that we have seen so far, why is there so much resistance to a widespread introduction of Bayesian methods? Why do frequentist methods, with all their peculiarities and problems, still dominate scientific practice? In my opinion, three points can be identified.

First, there are principled reservations against the Bayesian approach because it seems to threaten the objectivity, impartiality and epistemic authority of science. Although the ideal of objective statistical inference as free from personal perspective has been heavily criticized (e.g., Daston and Galison 2007; Douglas 2009) and may have lost its appeal for many philosophers, it is still influential for many scientists and regulatory agencies who are afraid of external interests influencing the inference process. For a long time, bodies such as the FDA were afraid that Bayesian analysis

would be misused for discarding hard scientific evidence on the basis of prejudiced a priori attitudes. Only quite recently, the FDA has opened up to a Bayesian analysis of clinical trials.

Second, scientific institutions such as editorial offices, regulatory bodies and professional associations are inert: they tend to stick to practices which have been "well probed" and to which they are familiar. Take experimental psychology as an example: even implementing the most basic changes, such as accompanying p-values by effect size estimates and/or power calculations, was a cumbersome process that took a lot of time. Changing the relevant textbook literature and the education of young scientists may take even more time. On a positive note, a more pluralist climate has developed over the last years, and there is now an increasing interest in Bayesian and other non-orthodox statistical methods.

Third, even some well-known Bayesians modelers like Gelman and Shalizi (2013) confess that while they apply Bayesian statistics as a technical tool, they would not qualify themselves as subjectivists. Rather, their methodological approach is closer to the hypothetico-deductive approach of testing models by means of their predictions. This is again similar to the frequentist rationale of hypothesis testing. So it may appear that while Bayesians may have the winning hand from a purely foundational point of view, it is by no means obvious that their methods provide the best answer in scientific practice. This points us to the task of telling a story of how Bayesian inference relates to statistical model checking in a hypothetico-deductive spirit, and more generally, to investigating the relationship between qualitative and quantitative, between subjective and objective accounts of theory confirmation (Sprenger 2013b).

Can the Bayesian scheme of inference can be modified in order to make it more objective? To this difficult question, three possible answers, that correspond to three different research programs, shall be sketched.

**Objective Priors** As explained in Christian Robert's (2014a) chapter on Bayesian statistics, there are various ways to choose a prior probability distribution. The method of objective priors tries to get the sting out of the subjectivity objection to Bayesianism by adopting priors that implement a sort of Principle of Indifference between the hypotheses under consideration (e.g., each parameter value gets equal probability). The problem with this approach is that the underlying Principle of Indifference is philosophically shaky (e.g., Hájek 2011). Other forms of objective priors, such as those motivated by transformation invariance (Jeffreys 1939; Bernardo 2012), may be more promising—see Sprenger (2012) for a discussion of the underlying philosophical issues.

**The Principle Maximum Entropy** This approach differs from Bayesian inference with objective priors since the entire principle of conditionalizing is rejected. Rather, an agent's rational degrees of belief should satisfy three constraints (Jaynes 1968; Williamson 2010): they should conform to the axioms of probability, satisfy empirical constraints on our rational degrees of belief and at last, they should be **equivocal**, that is, as middling as possible. This latter constraint amounts to maximizing the entropy of the probability distribution in question. If $\omega$ denotes the 'atoms' of the relevant $\sigma$-algebra, the entropy is given by the term

$$H = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega). \tag{7}$$

While the Principle of Maximum Entropy is of great help in many practical problems in engineering, computer science, and related disciplines, it is hard to find a waterproof epistemic or decision-theoretic justification for why degrees of belief should in general be as middling as possible.

**Conditioning on Evidence Strength** Of the three approaches discussed, this is the least known one. The idea is to give a valid Bayesian interpretation to frequentist error probabilities by appropriate conditioning on the strength of the observed evidence (Berger 2003). The principle of conditioning could therefore be a possible bridge across the gap that divides Bayesians and frequentists. Moreover, it is directly applicable to salient problems in the analysis of sequential trials in medicine (Berger, Brown and Wolpert 1994; Nardini and Sprenger 2013). Such attempts to find a compromise between Bayesian and frequentist inference are, for the most part, still terra incognita from a philosophical point of view. But in my perspective, there is a lot to gain from carefully studying how these approaches try to find a middle ground between two competing schools of inference that are often described as incompatible.

## Conclusion

This paper has introduced the principles of Bayesian and frequentist inference and compared them in a number of respects, most prominently the measures of statistical evidence which they advocate. We have seen that frequentist inference suffers from a variety of conceptual, epistemological and practical problems which seem to favor a subjective Bayesian approach. However, when it comes to objectivity in statistical inference

and accounting for crucial elements of scientific practice (e.g., hypothesis testing), it is not so easy for subjective Bayesianism to make ends meet. The varieties of objective Bayesian inference that are pursued in the statistical literature are a fruitful and exciting research program that could find a philosophically sound and practically viable middle ground between the two grand schools of statistical inference.

## References


Aldrich, J. (2013): "The Origins of Modern Statistics", in: A. Hájek and C. Hitchcock (eds.), *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Berger, J.O. (2003): "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?", *Statistical Science* 18, 1–32.

J.O. Berger, D. Berry (1988): "The Relevance of Stopping Rules in Statistical Inference (with discussion)", in: S. Gupta and J. O. Berger (eds.), *Statistical Decision Theory and Related Topics IV*, 29–72. Springer, New York.

Berger, J.O., L.D. Brown, and R.L. Wolpert (1994): "A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing", *Annals of Statistics* 22, 1787–1807.

Berger, J.O., and T. Sellke (1987): "Testing a point null hypothesis: The irreconciliability of P-values and evidence", *Journal of the American Statistical Association* 82, 112–139.

Berger, J.O., and R.L. Wolpert (1984): *The Likelihood Principle*. Hayward/CA: Institute of Mathematical Statistics.

Bernardo, J.M. (2012): "Integrated objective Bayesian estimation and hypothesis testing", in J.M. Bernardo et al. (eds.): *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, 1–68. Oxford: Oxford University Press.

Bernardo, J.M., and A.F.M. Smith (1994): *Bayesian Theory*. Chichester: Wiley.

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

Carnap, R. (1950): *Logical Foundations of Probability*. The University of Chicago Press, Chicago.

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, 106–111.

Cohen, J. (1994): "The Earth is Round ($p < .05$)", *American Psychologist* 49, 997–1001.

Cumming, G., and S. Finch (2005): "Inference by eye: Confidence intervals, and how to read pictures of data", *American Psychologist* 60, 170–180.

Daston, L., and Galison, P. (2007): *Objectivity*. New York: Zone Books.

Douglas, H. (2004): "The irreducible complexity of objectivity", *Synthese* 138, 453–473.

Douglas, H. (2009): *Science, Policy and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Edwards, A.W.F. (1972): *Likelihood*. Cambridge: Cambridge University Press.

Edwards, W., H. Lindman and L.J. Savage (1963): "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70, 450–499.

Fidler, F. (2005): *From Statistical Significance to Effect Estimation*. Ph.D. Thesis: University of Melbourne.

Fisher, R.A. (1925): *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1935): *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. To appear in *Psychonomic Bulletin & Review*.

Gandenberger, G. (2014): "A New Proof of the Likelihood Principle", forthcoming in *British Journal for the Philosophy of Science*.

Gelman, A., and C. Shalizi (2013): "Philosophy and the practice of Bayesian statistics (with discussion)", *British Journal of Mathematical and Statistical Psychology* 66, 8–18.

Gillies, D. (1971): "A Falsifying Rule for Probability Statements", *British Journal for the Philosophy of Science* 22, 231–261.

Goodman, S.N. (1999a): "Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy", *Annals of Internal Medicine* 130, 995–1004.

Goodman, S.N. (1999b): "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor", *Annals of Internal Medicine* 130, 1005–1013.

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

Hájek, A. (2011): Interpretations of Probability. *Stanford Encyclopedia of Philosophy*, ed. E. Zalta.

Harlow, L.L., S.A. Mulaik, and J.H. Steiger (eds.) (1997): *What if there were no significance tests?*. Mahwah/NJ: Erlbaum.

Hoover, K.D., and M.V. Siegler (2008): "The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey", *Journal of Economic Methodology* 15, 57–68.

Howson, C. and P. Urbach (2006): *Scientific Reasoning: The Bayesian Approach*. Third Edition. La Salle: Open Court.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine* 2(8): e124. doi:10.1371/journal.pmed.0020124 (electronic publication).

Jaynes, E.T. (1968): "Prior Probabilities", *IEEE Transactions on Systems Science and Cybernetics* SSC4, 227–241.

Jahn, R.G., B.J. Dunne and R.D. Nelson (1987): "Engineering anomalies research", *Journal of Scientific Exploration* 1, 21–50.

Jefferys, William H (1990): "Bayesian Analysis of Random Event Generator Data", *Journal of Scientific Exploration* 4, 153–169.

Jeffreys, H. (1939): *Theory of Probability*. Oxford: Clarendon Press.

Kadane, J.B., M.J. Schervish, and T. Seidenfeld (1996): "When Several Bayesians Agree That There Will Be No Reasoning to a Foregone Conclusion", *Philosophy of Science* 63, S281–S289.

Kass, R. and A. Raftery (1995): "Bayes Factors", *Journal of the American Statistical Association* 90, 773–790.

Krüger, L., G. Gigerenzer, and M. Morgan (eds.) (1987): *The Probabilistic Revolution, Vol. 2: Ideas in the Sciences*. Cambridge/MA: The MIT Press.

Lele, S. (2004): "Evidence Functions and the Optimality of the Law of Likelihood (with discussion)", in: Mark Taper and Subhash Lele (eds.), *The Nature of Scientific Evidence*, 191–216. The University of Chicago Press, Chicago & London.

Lindley, D.V. (1957): "A statistical paradox", *Biometrika* 44, 187–192.

Mayo, D.G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago & London: The University of Chicago Press.

Mayo, D.G. (2010): "An error in the argument from conditionality and sufficiency to the likelihood principle", in: D. Mayo, A. Spanos (eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*, 305–314. Cambridge: Cambridge University Press.

Mayo, D.G., and M. Kruse (2001): "Principles of inference and their consequences", in: D. Cornfield, J. Williamson (eds.), *Foundations of Bayesianism*, 381–403. Kluwer, Dordrecht.

Mayo, D.G., and A. Spanos (2006): "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *The British Journal for the Philosophy of Science* 57, 323–357.

McCloskey, D.N., and S.T. Ziliak (1996): "The Standard Error of Regressions", *Journal of Economic Literature* 34, 97–114.

Nardini, C., and J. Sprenger (2013): "Bias and Conditioning in Sequential Medical Trials", *Philosophy of Science* 80, 1053–1064.

Neyman, J., and E. Pearson (1933): "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society A* 231, 289–337.

Neyman, J., and E. Pearson (1967): Joint Statistical Papers. Cambridge: Cambridge University Press.

Oakes, M. (1986): *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

O'Hagan, T. (2012): Posting on the statistical methods used in the discovery of the Higgs Boson, made via the email list of the International Society for Bayesian Analysis (ISBA). Retrieved from `www.isba.org` on January 6, 2013.

Popper, K.R. (1934/59): *Logik der Forschung*. Berlin: Akademie Verlag. English translation as *The Logic of Scientific Discovery*. New York: Basic Books, 1959.

Robert, C. (2014a): "Des spécificités de l'approche bayesienne et de ses justifications en statistique inféréntielle", forthcoming in *this volume*.

Robert, C. (2014b): "On the Jeffreys-Lindley-paradox", forthcoming in *Philosophy of Science*.

Royall, R. (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Savage, L.J. (1962): *The foundations of statistical inference*. London: Methuen.

Spanos, A. (2010): "Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?", *Philosophy of Science* 77, 565–583.

Spielman, S. (1974): "The Logic of Significance Testing", *Philosophy of Science* 41, 211–225.

Spielman, S. (1978): "Statistical Dogma and the Logic of Significance Testing", *Philosophy of Science* 45, 120–135.

Sprenger, J. (2009): "Evidence and Experimental Design in Sequential Trials", *Philosophy of Science* 76, 637–649.

Sprenger, J. (2012): "The Renegade Subjectivist: Jose Bernardo's Reference Bayesianism", *Rationality, Markets and Morality* 3, 1–13.

Sprenger, J. (2013a): "Testing a Precise Null Hypothesis: The Case of Lindley's Paradox", *Philosophy of Science* 80, 733–744.

Sprenger, J. (2013b): "A Synthesis of Hempelian and Hypothetico-Deductive Confirmation", *Erkenntnis* 78, 827–838.

Williamson, J. (2010): *In defense of objective Bayesianism*. Oxford: Oxford University Press.

Ziliak, S.T., and D.N. McCloskey (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.