Bayesianism vs. Frequentism in Statistical Inference

Jan Sprenger*

September 30, 2013

Contents

*Contact information: Tilburg Center for Logic and Philosophy of Science, 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.

To better describe the scope and goals of these approaches, I follow Royall (1997, 6) in his distinction of three main questions in statistical analysis :

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

These questions are closely related, but distinct. Bayesians focus on the first question—rational belief—because for them, scientific hypotheses are an object of personal, subjective uncertainty. Therefore, Bayesian inference is concerned with the question of how data should change our degree of belief in a 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 are united in rejecting the use of subjective uncertainty in the context of scientific inquiry. Still, they considerably disagree on the foundations of statistical inference. Behaviorists such as Jerzy Neyman and Egon Pearson build their statistical framework on reliable decision procedures, thus emphasizing the second question, while others, such as Ronald A. Fisher or Deborah Mayo, stress the relevance of (post-experimental) evidential assessments.

The purpose of this article is to make the reader understand the principles of the two major schools of statistical inference—Bayesianism and frequentism—and to recognize the scope, limitations and weak spots of either approach. Notably, the divergences between both frameworks 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 does not exist, the experimental results deviate 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 (and the press department) 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 [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 many particle physicists taking part in the discussion. In the first place, the debate concerned a specific standard of evidence, but since the notion of strong evidence 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 include issues as applied as the admission process for medical drugs.

Statistical methodology is thus a significant topic for philosophy, science and public policy. In this contribution, we focus on how statistical evidence

should be interpreted. 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 and 3 contrast behavioral and evidential interpretations of frequentist tests. Section 4 deals with the notorious p-values. Section 5 discusses confidence intervals as an alternative to significance tests and p-values whereas 6 deals with Mayo's error-statistical approach. Section 7 briefly exposes the optional stopping problem, and Section 8 concludes with a general discussion.

1 Bayesian Inference

Bayesian reasoners 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. These degrees of belief can be 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: $p_{\text{new}}(H) = p(H|E)$.[1]

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

$$p_{\text{new}}(H) := p(H|E) = p(H)\frac{p(E|H)}{p(E)} = \left(1 + \frac{p(\neg H)}{p(H)} \cdot \frac{p(E|\neg H)}{p(E|H)}\right)^{-1} \quad (1)$$

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" usually refers to the conjunction of the following principles:

- The representation of subjective degrees of belief in terms of probabilities.

[1] See the handbook entry on the subjective interpretation of probability (Zynda 2013) for a defense of Conditionalization, and for arguments that degrees of belief should satisfy the probability calculus.

- 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. Carnap's (1950) system of logical probability and Jeffreys' (1939) objective priors violate the first principle. Conditionalization is rejected by Bayesians who accept the Principle of Maximum Entropy (Williamson 2010). In this paper, however, we focus on the standard subjectivist position in Bayesian statistics that is built on the conjunction of these three principles.[2]

A consequence of Bayesianism that is of particular importance in statistical inference is 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 θ 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).[3]

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.

From the perspective of Bayes' Theorem, 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 θ. Therefore, it is absolutely logical that a subjective Bayesian endorses the LP.

As Birnbaum (1962) showed in a celebrated paper, the Likelihood Principle can be derived from two different and more basic principles: Sufficiency and Conditionality. We begin with the first one. A statistic (i.e., a function

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

[3] We follow the convention of using capital letters for random variables and regular letters for their realizations.

of the data X) $T(X)$ is *sufficient* if the distribution of the data X does not depend on the unknown parameter θ, conditional on T. In other words, sufficient statistics are compressions of the data set that do not lose any relevant information about θ. An example is an experiment about the bias of a coin. Assuming that the tosses are independent and identically distributed, the overall number of heads and tails is a sufficient statistics for an inference about the bias of the coin. Thus, we can neglect the precise order in which the results occurred. Formally, the **Sufficiency Principle** states that any two observations x_1 and x_2 are evidentially equivalent with regard to the parameter of interest θ as long as $T(x_1) = T(x_2)$ for a sufficient statistic T. Therefore, the principle is usually accepted by Bayesians and frequentists alike.

The **Conditionality Principle** is more controversial: it states that evidence gained in a probabilistic mixture of experiments is equal to the evidence in the actually performed experiment. In other words, if we throw a die to decide whether experiment $\mathcal{E}_1$ is conducted (in case the die comes up with an odd number) or experiment $\mathcal{E}_2$ (even number) and we throw a six, then the evidence from the overall experiment $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ is equal to the evidence from $\mathcal{E}_2$. Frequentists usually reject Conditionality since their measures of evidence take the entire sample space into account.[4]

According to many, it is the task of science to state the evidence for hypotheses of interest, instead of reporting degrees of belief in their truth. To address this challenge, the Bayesian needs a **measure of evidence**, that is, a numerical representation of the impact of the data on the hypotheses of interest. 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).

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

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

[4]See Section 7. A thorough discussion of these principles goes beyond the scope of this article although some issues about Conditionality also return in the section on optional stopping. Recently, Mayo (2010) has challenged Birnbaum's proof of the LP from the Sufficiency and Conditionality principles.

plausibility of the individual hypotheses.

The Bayes factor 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 (Lele 2004).

Some statisticians and methodologists use $L(x, H_0, H_1)$ as a contrastive measure of evidence without using the subjective probability interpretation. The reason is that they have doubts about whether subjective degrees of belief should be used in quantifying statistical evidence. These **likelihoodists** answer the third grand question—measuring evidence—by means of the thesis that L, or appropriate amendments thereof, provide the best measure of evidence (e.g., Royall 1997; Lele 2004). The likelihoodist framework, sometimes also called "Bayesianism without priors", has been anticipated by Hacking (1965) and elaborated by Edwards (1972) and most recently by the methodologist and biostatistician Richard Royall (1997).

Since likelihoodists and Bayesians agree on a lot of foundational issues—e.g., both camps accept the LP and raise similar objections against frequentism—I do not give a separate treatment of this approach. Certainly, it is conceptually and foundationally appealing, especially because it seems to do justice to the idea that statistical evidence is objective. However, likelihood-based inference is hard to implement in practice if the inference problems involve composite hypotheses and nuisance parameters. In those cases, computing $L(x, H_0, H_1)$ seems to require calculation of the marginal likelihoods, and thus, prior weights over the elements of the hypothesis space: $p(x|H_0) = \int_{\theta \in \Theta_0} p(x|\theta)p(\theta)$. So the likelihoodist either has to compromise the objectivity of her approach, effectively becoming a Bayesian, or to take refuge in other measures of evidence, such as conditional likelihood ratios. Royall (1997, ch. 7) discusses several *ad hoc* techniques that take care of important applications, but the fundamental philosophical problem remains.

2 Frequentism: Neyman and Pearson's Behavioral Approach

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 statistics. Yet, statistics was not as strongly linked to inductive inference as it is today. For instance, an eminent statistician and biologist like Francis Galton conceived of statistics as a *descriptive* tool for meaningfully compressing data, and summarizing general trends (cf. Aldrich 2013). Moreover, in those days the *nomothetic ideal*—to strive for certainty, for invariable, deterministic laws—had a great impact on scientific practice. Probability was used to quantify the lack of precision in measurement and conclusions, but not as a meaningful part of scientific theorizing (see the contributions in Krüger et al. 1987).

This attitude changed at the beginning of the 20th century with the groundbreaking discoveries of statisticians such as Karl Pearson and William Gosset ("Student"). Pearson discovered the χ^2-test (1900) for testing the goodness of fit between a hypothesized distribution and the observed data, Gosset discovered the t-test (1908) for making inferences about the mean of a Normally distributed population. These techniques, which are still widely used today, were invented in response to applied research questions and mark the transition from **descriptive** to **inferential statistics**. Statistics became a discipline concerned with making inferences about a parameter of interest, predictions and decisions, rather than just summarizing data.

Given the aforementioned nomothetic, objectivist ideals, many scientists had issues with the Bayesian approach to probabilistic inference. After all, subjective degrees of belief are hard to measure, and apparently lack the impartiality and objectivity of scientific findings. 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.

The need to develop a coherent non-Bayesian theory of probabilistic inference was therefore badly felt. One famous answer was given by the

British statisticians Jerzy Neyman and Egon Pearson who connected statistical analysis to rational decision-making. In their groundbreaking 1933 paper, they developed a genuinely frequentist theory of hypothesis testing: statistical tests should be constructed as to minimize the relative frequency of wrong decisions in a hypothetical series of repetitions of the test. In particular, Neyman and Pearson linked the interpretation of statistical experiments tightly to their design.

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**.

For making a sound decision, Neyman and Pearson suggest the following route: first, the scientist chooses a default or **null hypothesis** H_0 for which a type I error rate is fixed. In medicine, the null usually states that the new treatment brings no improvement over the old one. After all, admitting an inefficient or even harmful drug is worse than foregoing a more effective treatment—at least from a regulatory point of view. By contrast, the **alternative** H_1 states that the drug is a genuine improvement. 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. In good frequentist spirit, Neyman and Pearson devise a decision procedure such that (i) in not more than 5%/1%/0.1% of all cases where the null hypothesis is true, it will be rejected; (ii) the **power** of the test—its ability to discern the alternative when it is true—is maximal for the chosen level of the test. In other words, given a fixed type I error rate (e.g., 1%), we design the test such that the type II error rate is minimized. Then, we are 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)

But how do we find the optimal test? For the case of two point hypotheses ($\theta = \theta_0$ vs. $\theta = \theta_1$) being tested against each other, Neyman and Pearson have proved an elegant result:

> **Fundamental Lemma of Neyman and Pearson (1933):** When testing two point hypotheses against each other, the most powerful test at any level α is the *likelihood ratio test*. This is a test T for which there is a $C(\alpha) \in \mathbb{R}$ such that for data x:

$$T(x) = \begin{cases} \text{accept } H_0 & \text{if } L = \frac{p(x|\theta=\theta_0)}{p(x|\theta=\theta_1)} \geq C(\alpha) \\ \text{reject } H_0 & \text{if } L = \frac{p(x|\theta=\theta_0)}{p(x|\theta=\theta_1)} < C(\alpha) \end{cases} \tag{3}$$

Hence, the uniformly most optimal test in Neyman and Pearson's sense depends on the weight of evidence as measured by the likelihood ratio L. If L strongly favors H_1 over H_0, we will reject the null, otherwise we will accept it. This result is at the bottom of a powerful mathematical theory of hypothesis testing and has greatly influenced statistical practice. However, Neyman and Pearson's approach has been attacked from a methodological point of view: according to Fisher, such tests are clever decision tools, but miss the point of scientific research questions. The next section explains this criticism.

3 Frequentism: Significance Tests and Fisher's Disjunction

The second grand tradition in frequentist statistics emerged with Ronald A. Fisher, eminent geneticist and statistician, who violently opposed Neyman and Pearson's behavioral, decision-theoretic approach. In determining an acceptable type I error rate, Neyman and Pearson implicitly determine 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, we cannot quantify the utility that correctly accepting or rejecting a hypothesis will eventually have for the advancement of science. The far-reaching consequences of such a decision lie beyond our horizon. Second, statistical hypothesis tests should state the *evidence* for or against the tested hypothesis: a scientist is interested in whether she has reason to believe that a hypothesis is *true* or *false*. Her judgment should not be obscured by the practical consequences of working with this rather than that hypothesis. Therefore, Neyman-Pearson tests may be helpful in industrial quality control and other applied contexts, but not in finding out the truth about a scientific hypothesis.[6]

For Fisher (1956) himself, 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 there being no effect of interest, no causal relationship between two variables, etc. In other words, the null denotes the absence of a phenomenon to be demonstrated. Then, the null is tested for being compatible with the data—notably, without considering explicit alternatives. This is called a **significance test**. Thus, 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 with sufficiently high power are essentially symmetric in the interpretation of the outcome.

The basic rationale of significance testing, called "Fisher's Disjunction" by Hacking (1965), is as follows: a very unlikely result undermines the (objective) tenability of the null hypothesis.

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

The occurrence of such an exceptionally rare chance has both epistemological and practical consequences: first, the null hypothesis is rendered "objectively

[6]The classification of Neyman-Pearson tests as purely behavioral is not without contention. Following the representation theorems in Savage (1962), one might link Neyman-Pearson tests to a general theory of belief attitudes and rational decision-making. Romeijn (2010, Section 9) also investigates an embedding of Neyman-Pearson tests into Bayesian statistics.

incredible" (Spielman 1974, 214), second, the null should be treated as if it were false. Naturally, this judgment is not written in stone, but may be overturned by future evidence. Below, there is a graphical representation of Fisher's main idea.

p(`Data`|`Null Hypothesis`) is low.
`Data` is observed.

`Null Hypothesis` is discredited.

Notably, Fisher's ideas are close to Popper's falsificationism, albeit with more inductivist inclinations. They both agree the only purpose of an experiment is to "give the facts a chance of disproving the null hypothesis" (Fisher 1925, 16). They also agree 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.

The above scheme of inference (cf. also Gillies 1971) has been criticized frequently. Hacking (1965, 81–82) has pointed out that the explication of the term "exceptionally rare chance" inevitably leads into trouble. A prima facie reading of Fisher's above quote seems to suggest that the chance of the observed event must be exceptionally low compared to other events that could have been observed. But in that case, some statistical hypotheses could never be tested. For instance, a uniform distribution over a finite set of events assigns equal likelihood to all observations, so there is no exceptionally rare chance. How should we test—and possibly reject—such a hypothesis?

To expand on this point, imagine that we are now testing the hypothesis that a particular coin is fair. Compare now two series of independent and identically distributed tosses: '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*: namely, we are interested in the propensity θ of the coin to come up tails. Since T, the number of tails, is a sufficient statistic with respect to θ, we can restrict our attention to the value of T. Then, $\{T = 0\}$ 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 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. Otherwise we would have no reason to appraise that effect. Indeed, modern frequentist approaches, such as Mayo's (1996) error-statistical account, 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.

However, does this modification suffice to save the logic of significance testing? 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 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 inference, even if the findings are likely under the alternative. Since the fallacy is caused by neglecting the base rates in the populations, it is called the **Base Rate Fallacy** (cf. Goodman 1999).

Thus, if a significance test is supposed to deliver a valid result, the null must not be too credible beforehand (cf. Spielman 1974, 225). But if we make all these restrictions to Fisher's proposal, it is questionable why we

should not switch to a straight Bayesian approach. After all, both approaches involve judgments of prior credibility, and the Bayesian framework is much more systematic and explicit in making and revising such judgments, and in integrating various sources of information.[7]

The above criticisms show that significance testing is logically invalid. To rescue it, we have to make additional premises, some of which adopt a Bayesian viewpoint. But if all this is right, why is significance testing such a widespread tool in scientific research? This question will be addressed in the next section.

4 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 θ is exactly equal to θ_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 (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.

This use of hypothesis tests differs from Fisher's since he considered inferences within a parametric model primarily as a problem of **parameter estimation**, not of hypothesis testing (cf. Spielman 1978, 122). His method of significance testing was devised for testing hypotheses without considering

[7]Spanos (2010, 576–580) objects that properly conceptualized frequentist tests do not fall prey to the Base Rate Fallacy: frequentist hypotheses describe an unknown data-generating mechanism. Whereas the hypothesis of interest in the above example (whether the donor has contracted HIV) is just an event in a more general statistical model that describes contraction status and test results of the entire population.

alternative hypotheses. But due to the problems mentioned in the previous section and the influence of Neyman and Pearson, modern significance tests require the specification of an alternative hypothesis. However, their interpretation is not behavioral, as Neyman and Pearson would require, but evidential, as Fisher would have requested.

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 θ 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 σ^2 of the population is known. Then, one measures the discrepancy in the data x with respect to the postulated mean value θ_0 by means the standardized statistic

$$z(x) := \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N \cdot \sigma^2}} \tag{4}$$

We may re-interpret equation (4) as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}} \tag{5}$$

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

$$p := p(|z(X)| \geq |z(x)|), \tag{6}$$

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).[8]

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 of this

[8]See Romeijn (2010) for further exposition of an epistemic reading of frequentist statistics, including Fisher's *fiducial argument*.

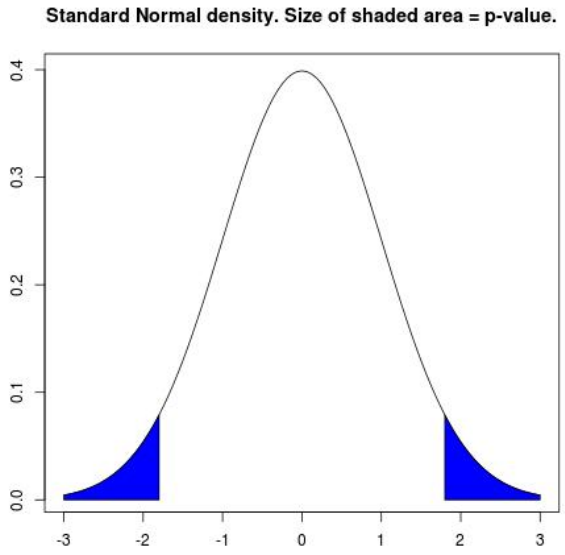

Figure 1: The probability density function of the null $H_0 : X \sim N(0, 1)$, which is tested against the alternative $H_1 : X \sim N(\theta, 1)$, $\theta \neq 0$. The shaded area illustrates the calculation of the p-value for observed data $x = 1.8$ ($p = 0.072$).

article to deliver a comprehensive discussion (see e.g., Cohen 1994; Harlow et al. 1997). The most important criticisms are discussed below and in Section 7 where the **sample space dependence** of frequentist inference will be thematized (cf. Hartmann and Sprenger 2011).

4.1 p-values and posterior probabilities

The arguably biggest problem with p-values is 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). Of course, this is just an instance of the Base Rate Fallacy: subjects conflate the conditional probability of the evidence given the hypothesis, $p(E|H)$, with the conditional probability of the hypothesis given the evidence, $p(H|E)$. In other words, they conflate statistical evidence with rational degree of belief.

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 posterior probabilities? After all, a Bayesian and a frequentist analysis should agree when prior probability distributions can be objectively grounded.

It turns out that in the one-sided testing problem, p-values can often be related to posterior probability (Casella and Berger 1987, more on this in Section 6) whereas in the two-sided or point null testing problem, the two measures of evidence diverge. When the prior is uninformative, a low p-value may still entail a high posterior probability of the null. More precisely, Berger and Sellke (1987) show that the p-value is often 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 reasoning in the two-sided testing problem. We expand on this point in a later subsection when discussing Lindley's paradox.

4.2 p-values vs. effect size

Another forceful criticism of p-values and significance tests concerns their relation to effect size. The economists Deirdre McCloskey and Stephen Ziliak have launched strong attacks against significance tests in a series of papers and books (McCloskey and Ziliak 1996; Ziliak and McCloskey 2008). Let us give their favorite example. Assume that we have to choose between two diet cures, based on pill A and pill B. Pill A makes us lose 10 pounds on average, with an average variation of 5 pounds.[9] Pill B makes us lose 3 pounds on average, with an average variation of 1 pound. Which one leads to more significant loss? Naturally, we opt for pill A because the effect of the cure is so much larger.

However, if we translate the example back into significance testing, the order is reversed. Assume the standard deviations are known for either pill. Compared to the null hypothesis of no effect at all, observing a three pounds weight loss after taking pill B is a more significant result evidence for the efficacy of that cure than observing a ten pounds weight loss after taking pill A:

$$z_A(10) = \frac{10-0}{5} = 2 \qquad\qquad z_B(3) = \frac{3-0}{1} = 3$$

Thus, there is a notable discrepancy between our intuitive judgment and the one given by the p-values. This occurs because statistical significance is supposed to be "a measure of the strength of the signal relative to background noise" (Hoover and Siegler 2008, 58). On this score, pill B indeed performs better than pill A, because of the favorable signal/noise ratio. But pace Ziliak and McCloskey, economists, businesspersons and policy-makers are interested in the effect size, not the signal/noise ratio: they do not want to ascertain the presence of *some* effect, but to demonstrate a *substantial* effect, as measured by **effect size**.

This fundamental difference is, however, frequently neglected. By scrutinizing the statistical practice in the top 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

[9] The concept of "average variation" is intuitively explicated as the statistical concept of standard deviation, which is, for a random variable X, defined as $\sqrt{E[(X - E(X))^2]}$.

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 (like pill A), and a statistically significant effect can be quite small and uninteresting (like pill B).

4.3 p-values and Lindley's Paradox

The tension between effect size and statistical significance is also manifest in one of the most famous paradoxes of statistical inference, Lindley's Paradox. Classically, it is stated as follows:

> **Lindley's Paradox:** Take a Normal model $N(\theta, \sigma^2)$ with known variance σ^2 and a two-sided testing problem $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Assume $p(H_0) > 0$, and any regular proper prior distribution on $\{\theta \neq \theta_0\}$. Then, for any testing level $\alpha \in [0, 1]$, we can find a sample size $N(\alpha)$ and independent, identically distributed data $x = (x_1, \ldots, x_N)$ such that
>
> 1. The sample mean $\bar{x}$ is significantly different from θ_0 at level α;
> 2. $p(H_0|x)$, that is, the posterior probability that $\theta = \theta_0$, is at least as big as $1 - \alpha$. (cf. Lindley 1957, 187)

In other words, a Bayesian and a frequentist analysis of a two-sided test may reach completely opposite conclusions. The reason is that the combination of statistical significance and large sample size (=high power) is highly misleading. In fact, as sample size increases, an ever smaller discrepancy from the null suffices to achieve a statistically significant result against the point null. The reader will thus be lured into believing that a "significant" result has substantial scientific implications although the effect size is very small. The high power of a significance test with many observations provides no protection against inferring to insignificant effects, quite to the contrary. Therefore, Lindley's Paradox lends forceful support to Ziliak and McCloskey's claim that statistical significance is a particularly unreliable guide to scientific inference.

This is not to say that all is well for the Bayesian: Assigning a strictly positive degree of belief $p(H_0) > 0$ to the point null hypothesis $\theta = \theta_0$ is

a misleading and inaccurate representation of our subjective uncertainty. After all, $\theta = \theta_0$ is not much more credible than any value $\theta_0 \pm \varepsilon$ in its neighborhood. Therefore, assigning a strictly positive prior to H_0, instead of a continuous prior, seems unmotivated (cf. Bernardo 2012).

But if we set $p(H_0) = 0$, then for most priors (e.g., an improper uniform prior) the posterior probability distribution will not peak at the null value, but somewhere else. Thus, the apparently innocuous assumption $p(H_0) > 0$ has a marked impact on the result of the Bayesian analysis. Attempts to consider it as a mathematical approximation of testing the hypothesis $H_0 : |\theta - \theta_0| < \varepsilon$ break down as sample size increases (cf. Berger and Delampady 1987).

The choice of prior probabilities, for H_0 as well as over the elements of H_1, is therefore a very sensitive issue in Lindley's paradox. Quite recently, the Spanish statistician José M. Bernardo (1999, 2012) has suggested to replace the classical Bayesian focus on posterior probability as a decision criterion by the Bayesian Reference Criterion (BRC), which focuses on the predictive value of the null in future experiments. This move avoids assigning strictly positive mass to a set of measure zero $\{\theta = \theta_0\}$ and reconciles Bayesian and frequentist intuitions to some extent. Sprenger (2013a) provides a more detailed discussion of this approach.

4.4 p-values and the assessment of research findings

A methodological problem with p-values, stemming from their roots in Fisherian significance testing, is that insignificant results (=p-values greater than .05) have barely 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, as indicated by Ziliak and McCloskey; 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 Ioannides' (2005) famous thesis that "most published research findings are false". Ioannides 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 α and β, 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 and a "significant" causal relationship is found, then

$$p(\text{the supposed causal relationship is true}) = \frac{(1-\beta)\cdot R}{(1-\beta)\cdot R + \alpha} \qquad (7)$$

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—an effect that is augmented by the experimenter's bias in selecting and processing his or her data set.

This finding is not only a feature of scientific inquiry in general, but specifically due to the frequentist logic of inference: the one-time 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. Th e problems with frequentist inference affect the success of entire research programs.

5 Confidence Intervals as a Solution?

The above criticisms dealt severe blows to classical significance tests and the use of p-values. In the last decades, frequentists have therefore adapted their tools. Nowadays, they often replace significance tests by confidence intervals, allegedly a more reliable method of inference (e.g., Cumming and Finch 2005; Fidler 2005). Confidence intervals are interval estimators that work as follows: Let $C(\cdot,\cdot)$ be a subset of $\Theta \times \mathcal{X}$ for parameter space Θ and sample space $\mathcal{X}$. Then consider the set $C(\theta_0,\cdot)$ that comprises those (hypothetical) data points for which the hypothesis $\theta = \theta_0$ would *not* be rejected at the level α. In other words, $C(\theta_0,\cdot)$ contains the data points that are *consistent* with θ_0.

If we construct these sets for all possible values of θ, then we obtain a two-dimensional set C with $(\theta, x) \in \Theta \times \mathcal{X}$. Assume further that we

observe data x_0. Now define the projection of C on the data x_0 by means of $C_{x_0} := \{\theta | x_0 \in C(\theta, x_0)\}$. This set $C_{x_0} \subset \Theta$ is called the **confidence interval** for parameter θ at level α, on the basis of data x_0.

Confidence intervals should not be understood in the literal sense that upon observing x_0, parameter θ lies in the interval C_{x_0} with probability $1 - \alpha$. After all, the frequentist does not assign any posterior probability to the parameters of interest. Rather, the level of the confidence interval says something about the procedure used to construct it: in the long run, the observed data x will be consistent with the constructed intervals for θ in $100 \cdot (1 - \alpha)\%$ of all cases, independent of the actual value of θ.

The advantage of confidence intervals over significance tests can be illustrated easily in the case of Lindley's Paradox. If we constructed a 95% confidence interval for θ, it would be a very narrow interval in the neighborhood of θ_0. 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.

However, confidence intervals do not involve a decision-theoretic component; they are interval estimators. If we take seriously that scientists want to conduct real *tests*, instead of estimating parameters, then confidence intervals cannot alleviate the worries with frequentist inference. Rather than solving the problem, they shift it, although they are certainly an improvement over naïve significance testing.

That said, 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 into the acceptance region at a certain level. This is in essence a pre-experimental perspective. But this do not warrant, post-experimentally, that the parameter of interest is "probably" in the confidence interval. Therefore, some frequentists are not happy with confidence intervals either. In recent years, the philosopher of statistics Deborah Mayo (1996) has tried to establish *degrees of severity* as superior frequentist measures of evidence. The next section is devoted to discussing her approach.

6 Mayo's Error-Statistical Account

In her 1996 book "Error and the Growth of Experimental Knowledge", Deborah Mayo works out a novel account of frequentist inference. Mayo's key concept, degrees of severity, combines Neyman and Pearson's innovation regarding the use of definite alternatives and the concept of power with Fisher's emphasis on post-experimental appraisals of statistical evidence.

Mayo's model of inference stands in a broadly Popperian tradition: for her, it is essential to scientific method that a hypothesis that we appraise has been well probed (=severely tested). Why should passing a test count in favor of a hypothesis? When are we justified to rely on such a hypothesis?

Popper (1934/59, 282) gave a skeptical reply to this challenge: he claimed that corroboration—the survival of past tests—is just a report of past performance and does not warrant any inference to future expectations. Mayo wants to be more constructive and to entitle an *inference* to a hypothesis:

> evidence E should be taken as good grounds for H to the extent that H has passed a severe test with E (Mayo 1996, 177)

Regarding the notion of what it means (for a statistical hypothesis) to pass a severe test, she adds:

> a passing result is a severe test of hypothesis H just to the extent that it is very improbable for such a result to occur, were H false (loc. cit., 178)

Notably, a null hypothesis which passes a significance test would, on Mayo's account, not necessarily count as being severely tested. For example, in tests with low sample size, the power of the test would typically be small, and even a false null hypothesis would probably pass the test. This is one of the reasons why she insists that hypotheses are *always tested against definite alternatives*. By means of quantifying how well a statistical hypothesis has been probed, we are entitled to inferences about the data-generating process. This exemplifies the basic frequentist idea that statistical inferences are valid if they are generated by reliable procedures.

For Mayo, a hypothesis H is *severely tested with data* x if (S-1) the data agree with the hypothesis, and (S-2) with very high probability, the data would have agreed less well with H if H were false (Mayo and Spanos 2006,

329).[10]

We illustrate her approach with an example of a Normally distributed population $N(\theta, \sigma^2)$ with known variance σ^2. Assume that we want to quantify the **severity** with which the hypothesis $H_0 : \theta \leq \theta_0$ passes a test T with observed data x (against the alternative $H_1 : \theta > \theta_0$). First, we measure the discrepancy of the data from the hypotheses by means of the well-known statistics

$$z_{\theta_0}(X) = \sqrt{N}(\bar{X} - \theta_0)/\sigma$$

z measures the distance of the data from H_0 *in the direction of* H_1 (cf. Mayo and Spanos 2006, 331): a large value of $\bar{X} - \theta_0$ yields large values of z and thus, evidence against the null. Then, the severity with which H_0 passes a test with data x is defined as the probability that $z_{\theta_0}(X)$ would have taken a higher value if the alternative $H_1 : \theta > \theta_0$ had been true. Mathematically:[11]

$$\text{SEV}(\theta \leq \theta_0)(x, H_1) = p(z_{\theta_0}(X) > z_{\theta_0}(x); \theta > \theta_0). \tag{8}$$

As the alternative $H_1 : \theta > \theta_0$ comprises a large set of hypotheses that impose different sampling distributions on the z-statistic, there is an ambiguity in (8). Which element of H_1 should be used for calculating the probability of the right hand side? To resolve this problem, Mayo observes that a lower bound on the test's severity is provided by calculating severity with respect to the hypothesis $\theta = \theta_0$. Thus, equation (8) becomes

$$\text{SEV}(\theta \leq \theta_0)(x, H_1) = p(z_{\theta_0}(X) > z_{\theta_0}(x); \theta = \theta_0). \tag{9}$$

This is, however, only half of the story. Mayo would also like to calculate at which severity the claim $\theta \leq \theta_0$ passes a test as a function of θ_0 when x is kept constant. Therefore, she calculates the **severity function** for θ_0 indicating which discrepancies from the null are warranted by the actual data, and which are not. Figure 2 gives an illustration.

[10] The precise meaning of (S-1) remains a bit unclear; Mayo and Spanos (2006, 336) say in passing that statistically insignificant results "agree" with the null. This definition may be contested, however: depending on the choice of the alternative, insignificant results may strongly discredit the null.

[11] I modify the notation in Mayo and Spanos (2006) to some extent. However, I follow them in using the semicolon for separating event and hypothesis in the calculation of the degree of severity because for them, the difference to the vertical dash (and to conditional probability) carries philosophical weight.

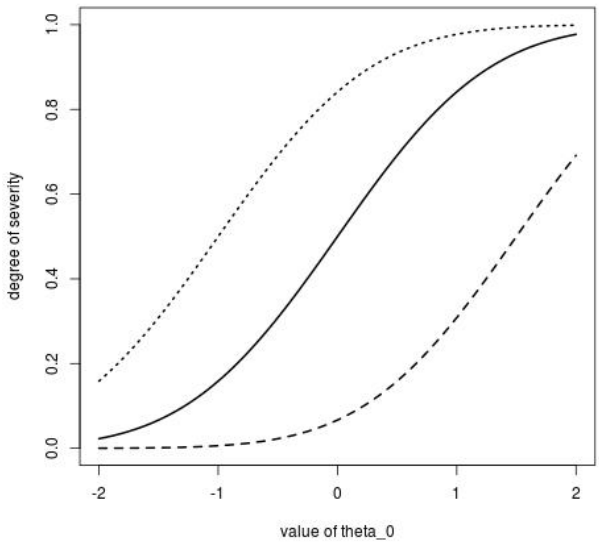

Figure 2: Inference about the mean θ of a normal population with variance $\sigma^2 = 1$. The three curves show the degrees of severity at which the hypothesis $\theta \leq \theta_0$ is accepted for three different data points. Dotted line: severity function for data $x = -1$, full line: $x = 0$, dashed line: $x = 1.5$.

The main merit of Mayo's approach consists in systematizing the various intuitions, heuristics and rationales in frequentist statistics. Practice is often a hodgepodge of methods, inspired by ideas from both the Neyman-Pearson and the Fisherian school. In particular, practitioners often combine decision procedures—calculating the power of a test, accepting/rejecting a null, etc.—with post-data evidential assessments, such as "hypothesis H_0 was rejected in the experiment (p=.026, power=.87)". Strictly speaking, this mix of Fisherian and Neyman-Pearson terminology is incoherent. With the error-statistical philosophy of inference and the concept of degree of severity, there is now a philosophical rationale underlying this practice: Mayo supplements Neyman and Pearson's pre-experimental methodology for designing powerful tests with a post-experimental measure of evidence. Therefore Mayo's approach is reasonably close to a lot of scientific practice carried out in the framework of frequentist statistics.

That said, there are also a number of problems for Mayo's approach, mainly due to foundational problems that are deeply entrenched in the entire frequentist framework.

First, error statistics reduces the testing of composite hypotheses against each other (e.g., $\theta \leq \theta_0$ vs. $\theta > \theta_0$) to testing a hypothesis against that particular hypothesis which provides the most severe test (in this case, $\theta = \theta_0$). Thus, it may be asked whether degrees of severity really improve on a traditional frequentist or a likelihoodist analysis.

Second, there is a close mathematical relationship between degrees of severity and (one-sided) p-values. Both are derived from the cumulative distribution function of the z-statistic, as equation (9) indicates. Therefore, degrees of severity share the problems of p-values and are poor indicators of rational credibility, at least from a Bayesian point of view.

Mayo might counter that a result by Casella and Berger (1987) shows the convergence of Bayesian and frequentist measures of evidence in the one-sided testing problem, effectively alleviating the Bayesian's worries. But I am skeptical that this response works. The reconciliationist results by Casella and Berger make substantial demands on the statistical model, e.g., probability density functions must be symmetric with monotone likelihood ratios (e.g., Theorem 3.1). Even then, the p-value can still substantially deviate from the posterior probability. Only for very large datasets, we will finally have agreement between Bayesian and frequentist measures.

Third, Mayo only provides an evidential interpretation of *directional* tests, not a rebuttal of the objections raised against two-sided frequentist tests (e.g., in Lindley's paradox). In particular, the question of whether a specific model can be treated as a proxy for a more general model is not addressed in the error-statistical framework: it only specifies warranted differences from a point hypothesis in a particular direction. However, a statistical framework that aims at resolving the conceptual problems in frequentist inference should address these concerns, too.

Fourth, as I will argue in the following section, the error-statistical theory fails, like any frequentist approach, to give a satisfactory treatment of the optional stopping problem.

7 Sequential Analysis and Optional Stopping

Sequential analysis is a form of experimental design where the sample size is not fixed in advance. This is of great importance in clinical trials, e.g. when we test the efficacy of a medical drug and compare it to the results in a control group. In those trials, continuation of the trial (and possibly the decision to allocate a patient to either group) depends on the data collected so far. For instance, data monitoring committees will decide to stop the trial as soon as there are substantial signs that the tested drug has harmful side effects.

A **stopping rule** describes under which conditions a sequential trial is terminated, as a function of the observed results. For example, we may terminate a trial when a certain sample size is reached, or whenever the results clearly favor one of the two tested hypotheses.[12] The dissent between Bayesians and frequentists concerns the question of whether our inference about the efficacy of the drug should be sensitive to the specific stopping rule used.

From a frequentist point of view, the significance of a result may depend on whether or not it has been generated by a fixed sample-size experiment. Therefore, regulatory bodies such as the Food and Drug Administration (FDA) require experimenters to publish all trial properties in advance, *including the stopping rule they are going to use.*

[12]Formally, stopping rules are functions $\tau : (\mathcal{X}^\infty, \mathcal{A}^\infty) \to \mathbb{N}$ from the measurable space $(\mathcal{X}^\infty, \mathcal{A}^\infty)$ (=the infinite product of the sample space) to the natural numbers such that for each $n \in \mathbb{N}$, the set $\{x \in \mathcal{X}^\infty | \tau(x) = n\}$ is measurable.

For a Bayesian, the LP implies that only information contained in the likelihood function affects a post-experimental inference. Since the likelihood functions of the parameter values under different stopping rules are proportional to each other (proof omitted), stopping rules can have no evidential role. Berger and Berry (1988, 34) call this the **Stopping Rule Principle**. To motivate this principle, Bayesians argue 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. This is the **sample space dependence** of frequentist inference mentioned in Section 4.

This position has substantial practical advantages: if trials are terminated for unforeseen reasons, e.g. because funds are exhausted or because unexpected side effects occur, the observed data can be interpreted properly in a Bayesian framework, but not in a frequentist framework. As externally forced discontinuations of sequential trials frequently happen in practice, claims to the evidential relevance of stopping rules would severely compromise the proper interpretation of sequential trials.

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 makes 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 θ (Savage 1962; Mayo and Kruse 2001).[13] In this case, the type I error is apparently 0.05 while it actually approaches unity since rejection of the null is bound to happen at some point. Not only this: a malicious scientist who wants to publish a result where a certain null hypothesis is rejected, can design an experiment where this will almost certainly happen, with an arbitrarily high level of statistical significance (provided she does not run out

[13] As we saw in the case of Lindley's Paradox, an ever smaller divergence from the null is sufficient to trigger statistical significance as sample size increases.

of money before). Should we trust the scientist's conclusion? Apparently no, but the Bayesian cannot tell why. Frequentists such as Mayo read this as the fatal blow for positions that deny the post-experimental relevance of stopping rules.

The Bayesian response is threefold. First, the posterior probability of a hypothesis cannot be arbitrarily manipulated (Kadane et et al. 1996). If we stop an experiment if and only if the posterior of a hypothesis exceeds a certain threshold, there will be a substantial chance that the experiment never terminates. It is therefore not possible to reason to a foregone conclusion with certainty by choosing a suitable stopping rule. Similar results that bound the probability of observing misleading evidence have been proved by Savage (1962) and Royall (2000).

Second, the frequentist argument is valid only if frequentist evidence standards are assumed. But from a Bayesian point of view, even biased experiments can produce impressive evidence—provided the design of the experiment did not interfere with the data-generating mechanism. If scientists had to throw away arduously collected data just because the experimental design was not properly controlled, scientific knowledge would not be at the point where it is now.

Third, preferring a (post-experimental) decision rule that is sensitive to the used stopping rule leads to incoherence, in the sense that a Dutch Book can be construed against such preferences. This result by Sprenger (2009) is derived from a more general, quite technical theorem by Kadane, Schervish and Seidenfeld (2003).

These arguments demonstrate the coherence of the Bayesian approach to stopping rules, and show that they should not matter post-experimentally if statistics is supposed to be consistent with standard theories of rational preferences and decisions. That said, there is a valid core in the frequentist argument: sequential trials are often costly and require careful pre-experimental design for efficient experimentation. Also, the termination of a sequential trial often involves complex ethical issues. Here, the choice of a stopping rule can make a great difference to frequentists *and* Bayesians.

8 Discussion: Some Thoughts on Objectivity

We have introduced the Bayesian and the frequentist paradigm as well as their philosophical foundations, and focused on three grand questions: what should be believe, what should we do, and how should we measure statistical evidence? In particular the last question sparks fierce debates between Bayesians and frequentists, as well as between different strands of frequentism.

The author has not concealed his inclinations toward a broadly Bayesian view on inductive inference. This position is supported by the numerous inadequacies of significance tests and p-values, among which the mathematical incompatibility with posterior probabilities, the neglect of effect size, and Lindley's Paradox. Moreover, the frequentist stance on stopping rules appears to lead to unacceptable consequences.

In light of these arguments, it may be surprising that frequentist statistics is still the dominating school of inductive inference in science. However, two points have to be considered. First, there are still principled reservations against the subjectivist approach because it apparently threatens 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., 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, and only 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).

Finally, I would like to mention some compromises between Bayesian and frequentist inference that Bayesians have invented for meeting objectivity demands. First, there is the **conditional frequentist approach** of Berger (2003) and his collaborators (e.g., Berger, Brown and Wolpert 1994). The idea of this approach is to supplement frequentist inference by conditioning on the observed strength of the evidence (e.g., the value of the Bayes factor). The resulting hypothesis tests have a valid interpretation from a Bayesian and a frequentist perspective and are therefore acceptable for either camp. Nardini and Sprenger (2013) describe how this approach can ameliorate the practice on sequential trials in medicine. Second, there are José Bernardo's (2012) **reference priors** which are motivated by the idea of maximizing the information in the data vis-à-vis the prior and posterior distribution (see Sprenger 2012, for a philosophical discussion).

Attempts to find a compromise between Bayesian and frequentist inference are, for the most part, still terra incognita from a philosophical point of view. In my perspective, there is a lot to gain from carefully studying how these approaches try to find a middle ground between subjective Bayesianism and frequentism.

References

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

ford: 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 M. Delampady (1987): "Testing Precise Hypotheses", *Statistical Science* 2, 317–352.

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. (1999): "Nested Hypothesis Testing: The Bayesian Reference Criterion", in J. Bernardo et al. (eds.): *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting*, 101–130. Oxford: Oxford University Press.

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.

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.

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. (1999): "Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy", *Annals of Internal Medicine* 130, 1005–1013.

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

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

Hartmann, S., und J. Sprenger (2011): "Mathematics and Statistics in the Social Sciences", in: I.C. Jarvie and J. Zamora Bonilla (eds.), *SAGE Handbook of the Philosophy of Social Sciences*, 594–612. London: SAGE.

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.

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

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", forthcoming in *Philosophy of Science*.

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.

Romeijn, J.W. (2010): "Inductive Logic and Statistics", in: D. Gabbay, S. Hartmann and J. Woods (eds.), *Handbook of the History of Logic, Volume 10 (Inductive Logic)*, 625–650. Amsterdam: Elsevier.

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

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

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", forthcoming in *Philosophy of Science*.

Sprenger, J. (2013b): "A Synthesis of Hempelian and Hypothetico-Deductive Confirmation", forthcoming in *Erkenntnis*.

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.

Zynda, L. (2013): "Subjectivism", in: A. Hájek and C. Hitchcock (eds.), *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.