Evidence and Experimental Design in Sequential Trials

Jan Sprenger[†][‡]

To what extent does the design of statistical experiments, in particular sequential trials, affect their interpretation? Should postexperimental decisions depend on the observed data alone, or should they account for the used stopping rule? Bayesians and frequentists are apparently deadlocked in their controversy over these questions. To resolve the deadlock, I suggest a three-part strategy that combines conceptual, methodological, and decision-theoretic arguments. This approach maintains the pre-experimental relevance of experimental design and stopping rules but vindicates their evidential, post-experimental irrelevance.

1. Exposition. Which relevance does the *design* of a statistical experiment in science have, once the experiment has been performed and the data have been observed? Do data speak for themselves, or do they have to be assessed in conjunction with the design that was used to generate them? Few questions in the philosophy of statistics are the subject of greater controversy.

The paradigmatic example is the inferential role of *stopping rules* in *sequential trials*. Those trials, which can be compared to the repeated toss of a coin, accumulate evidence from several independent and identically distributed trials. Sequential trials are standardly applied in medicine when the efficacy of a drug is tested by giving it to several patients after each other. The stopping rule describes under which circumstances the trial is terminated and is thus a centerpiece of the experimental design. Possible stopping rules could be "give the drug to 100 patients," "give the drug

[†]To contact the author, please write to: Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: j.sprenger@uvt.nl.

[‡]I would like to thank José Bernardo, Bruce Glymour, Valeriano Iranzo, Kevin Korb, Deborah Mayo, Jonah Schupbach, Gerhard Schurz, Aris Spanos, Kent Staley, Roger Stanev, Carl Wagner, the referees of *Philosophy of Science*, and especially Teddy Seidenfeld for their helpful and stimulating feedback.

Philosophy of Science, 76 (December 2009) pp. 637–649. 0031-8248/2009/7605-0031$10.00

until the number of failures exceeds the number of recoveries," or "give the drug until funds are exhausted." In other words, they indicate the number of repetitions of the trial as a function of some feature of the observed data; that is, technically speaking, a stopping rule τ is a function from a measurable space $(\mathcal{X}^{\infty}, \mathcal{A}^{\infty})$—the infinity product of the sample space—into the natural numbers such that, for each $n \in \mathbb{N}$, the set $\{x|\tau(x) = n\}$ is measurable.[1]

In the above example, the question about the relevance of stopping rules can be recast as the question whether our inference about the efficacy of the drug should be sensitive to the proposed ways to conduct and to terminate the experiment. If stopping rules were really indispensable and if we performed fewer trials than the stopping rule prescribed, for example, because funds are exhausted or because unexpected side effects occur, a proper statistical interpretation of the observed data would be difficult, if not impossible. The (ir)relevance of stopping rules thus has severe implications for scientific practice and the proper interpretation of sequential trials. Therefore, both scientists and philosophers of science should pay great attention to the question whether stopping rules are a crucial and indispensable part of statistical evidence or not.

The statistical community is deeply divided over that question. From a frequentist (Neyman-Pearsonian, error-statistical) point of view, a biased stopping rule, such as sampling on until the result favors our pet hypothesis, will lead us to equally biased conclusions (Mayo 1996, 343–345). Bayesians, however, claim that "the design of a sequential experiment is . . . what the experimenter actually *intended* to do" (Savage 1962, 76). Since such intentions are "locked up in [the experimenter's] head" (76), stopping rules cannot matter for sound inference (see also Edwards, Lindman, and Savage 1963, 239). The following principle captures the Bayesian position in a nutshell:[2]

> **Stopping Rule Principle (SRP)**. In a sequential experiment with observed data $x = (x_1, \ldots, x_n)$, all experimental information *about* ϑ is contained in the function $P_n(x|\vartheta)$; the stopping rule τ that was used *provides no additional information about* ϑ. (See Berger and Berry 1988, 34.)

However, the debate is characterized by a mutual deadlock, because each

[1] I confine myself to *noninformative stopping rules*—stopping rules that are independent of the prior distribution of the parameter. This means that for a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ representing the trial results, the event $\{\tau = n\}$ is measurable with respect to $\sigma(X_1, \ldots, X_n)$. See Schervish 1995, 565.

[2] See also Royall 1997, 68–71. Note that the first part of the SRP contains the Likelihood Principle (Birnbaum 1962; Berger and Wolpert 1984).

side presupposes its own inferential framework and measures by its own standards. For instance, Howson and Urbach (2006, 251) bluntly claim that unless the Bayesian position is led into absurdity in Bayesian terms, "there is no case whatever for the Bayesian to answer." Frequentists respond in a similar way to the Bayesian charge, pointing out that error probabilities are the hallmark of a sound inference and that they do depend on stopping rules (Mayo 1996, 348). But do we really have to wear Bayesian or frequentist glasses in order to enter the debate? Isn't there a strategy to overcome the stalemate between Bayesians and frequentists?

I believe that we can break the deadlock, and here I outline my strategy. First, we make explicit the distinction between pre-experimental and post-experimental evidential relevance. This will help us to disentangle and to classify the existing arguments. Second, we elicit which conception of statistical evidence best responds to the practical needs of empirical scientists. This has immediate consequences for the relevance of stopping rules. Third, we assert the pre-experimental relevance and postexperimental irrelevance of stopping rules and vindicate this standpoint from a decision-theoretic perspective. Thus, instead of solely relying on foundational intuitions, we combine arguments from mathematical statistics and decision theory with a methodological perspective on the needs of experimental practice.

2. Measures of Evidence: A Practitioner's Perspective. The two senses in which stopping rules can be relevant correspond to two stages of a sequential trial: first, the *pre-experimental* stage, where the trial is planned and the stopping rule is determined, and second, the *postexperimental* stage, where observed data are interpreted and transformed into an evidential assessment. For the latter project, we need evidence measures that summarize raw data to make us see which of two competing hypotheses is favored over its rival. Such quantifications help us to endorse or to reject scientific hypotheses or to make policy-relevant decisions. For instance, frequentist statistics is concerned with statistical testing and the comparison of two mutually exclusive hypotheses, the *null hypothesis* H_0 and the *alternative* H_1. After looking at the data, one of them is accepted and the other one is rejected. Such decision rules are characterized and ranked according to their *error probabilities*, that is, the probability of erroneously rejecting the null hypothesis (type I error) and the probability of erroneously rejecting the alternative hypothesis (type II error).[3] However, such error probabilities characterize (pre-experimen-

[3] Other frequentist procedures, such as constructing confidence intervals, are equally

tally) a particular testing procedure and do not directly tell us (postexperimentally) the strength of the observed evidence. For this reason, frequentists supplement their error analysis by a measure of evidence, such as p-values, significance levels, or, most recently, degrees of severity (Mayo and Spanos 2006, 337–346). Neglecting subtle differences between those measures, they are united in measuring the evidence against H_0 by summing up the H_0 likelihoods of those observations that have a greater discrepancy from the null hypothesis than the observed value x:

$$p := P_{H_0}(T(X) \geq T(x)). \tag{1}$$

Here T is a suitable (minimally sufficient) transformation of the data that indicates the discrepancy between the data and H_0. In other words, p-values (taken *pars pro toto*)[4] give the probability that, if the null hypothesis were true and the experiment were repeated, the results would speak at least as much against H_0 as the actual data do. Thus, p-values summarize the evidential import of the data and measure the tenability of the null hypothesis in the light of incoming evidence such that we can base further decisions on them. Indeed, p-values are widespread in the empirical sciences and are often a compulsory benchmark for experimental reports, for example, in medicine or experimental psychology. In particular, only results with a p-value lower than .05 are generally believed to be (statistically) significant and publishable (Goodman 1999).

All those measures of evidence are sensitive to the used stopping rule. This comes as no surprise since each stopping rule shapes up a different sample space; for example, in a fixed sample size scheme and a variable sample size scheme, different observations are possible. In other words, p-values depend not only on the likelihoods of the actually observed results but also on the likelihood of results that *could have been observed under the actual experimental design* as equation (1) makes clear. Hence, for a frequentist statistician who works with p-values, significance levels, degrees of severity, or the like, the strength of the observed evidence depends on the used stopping rule.

To see whether this is a desirable or embarrassing property, we should clarify our expectations of a measure of statistical evidence. Evidence about a parameter is required for inferences about that parameter, for example, for sensible estimates and decisions to work with this rather than that value (e.g., $\vartheta = \vartheta_0$ instead of $\vartheta = \vartheta_1$). An evidence measure transforms the data to provide the basis for a scientific inference. In order

justified by the error probabilities that characterize that procedure. Thus, in the remainder of the paper, I focus on the hypothesis testing framework.

[4] In Mayo and Spanos's (2006, 342) framework, the severity with which the alternative passes a test against the null is equal to $1 - p$.

to be suitable for public communication in the scientific community and for use in research reports, a measure of evidence should be *free of subjective bias and immune to manipulation*, as well as independent of prior opinions. While we can disagree on the a priori plausibility of a hypothesis, we should agree on the strength of the observed evidence; that is the very point of evidence-based approaches in science and policy making. Therefore, we need a method to quantify the information contained in the data that is independent of idiosyncratic convictions and immune to deliberate manipulations.

To clarify the point, consider an example. A malicious experimenter conducts a sequential trial with a certain stopping rule, but the evidence against the null that she finds is not as strong as desired. In particular, the p-value is not significant enough to warrant rejection of the null and publication of the results ($p \approx .051$). What can she do? The first option consists in outright fraud: she could fake some data (e.g., replace some observed failures by successes) and make the results significant in that way. While tempting, such a deception of the scientific community is risky and would be heavily punished if discovered. The career of our experimenter would be over once and for all. Therefore, a second option looks more attractive: not to report the true stopping rule τ_1 (fixed sample size), but a modified stopping rule τ_2 under which the data D yield a p-value smaller than .05.[5] The results are now "statistically significant" and get published. But clearly, as readers of a scientific journal, we want to be protected against such tricks. The crucial point is that the malicious experimenter did not manipulate the *data*: she was just insincere about her *intentions* when to terminate the experiment. Using fake data involves considerable risk: if continued replications fail to reproduce the results, our experimenter will lose all her reputation. By contrast, she can never be charged for insincerely reporting her intentions. The crucial point here is not the intuition that "intentions cannot matter for strength of evidence," but rather that the scientific community is unable to *control* whether these intentions have been correctly reported. This inability to detect subjective distortion and manipulation of statistical evidence is a grave problem for frequentist methodology.

What kind of answers could the frequentist give? To propose a standardized stopping rule τ, such as fixed sample size, does not help: experimenters could still use another stopping rule τ' and report the results as if they had been generated by τ. What about the (actually made) proposal

[5] For instance, she could have tested the null hypothesis $\vartheta = 0.5$ in 46 Bernoulli (success/failure) trials with a fixed sample size and have obtained 29 successes with $p = .052$. However, under a *negative* binomial stopping rule (sample until you get 17 failures), the p-value would have been $p = .036 < .05$.

to fix and to publicly declare a stopping rule in advance? This sounds good, but a stopping rule that covers all eventualities in advance is hard, if not impossible, to find. What if research funds expire because the trial proves to be more expensive than thought? What if unexpected technical problems blur the measurements or force the termination of the experiment? These problems are certainly no remote thought experiments: they frequently occur in scientific practice. Considering all those external influences in advance and assigning probabilities to them (!), as required for explicit stopping rules, is certainly not feasible. Should we then consider data from early stopped experiments as entirely worthless because this course of events was not accounted for in planning the experiment and formulating the stopping rule?

At this point, we cannot just be a little bit frequentist: if we believe in the evidential, postexperimental relevance of stopping rules, then we have to be silent on the meaning of data where the stopping rule is unavailable. But if we throw the data into the trash bin, we give away a great deal of what reality tells us, impeding scientific progress as well as responsible, evidence-based policy making. In fact, no journal article that reports *p*-values (and is implicitly committed to the relevance of stopping rules) ever bothers about fine-tuning the stopping rule to the external circumstances under which the experiment was conducted. Thus, empirical scientists do not take the relevance of stopping rules as seriously as their widespread adherence to the frequentist framework of statistical inference suggests. In fact, they have no other choice when they want to maintain ordinary experimental practice. Specifying the stopping rule in advance sounds good, but specifying the correct, comprehensive stopping rule (which we need to interpret the results properly) is practically impossible. Thus, the frequentist understanding of evidence, whether explicated as *p*-values, significance levels, or degrees of severity, is unable to cope with the practical problems that arise when the relevance of stopping rules is taken seriously.

The nonfrequentist alternatives, such as likelihood ratios, or their generalization, Bayes factors, fare much better. These measures of evidence merely build on publicly accessible factors, such as the likelihood of observed data under competing hypotheses and possibly explicit prior distributions:

$$B(H_1, H_0, x) := \frac{P(H_1|x)}{P(H_1)}\,\frac{P(H_0)}{P(H_0|x)} = \frac{\int_{\vartheta\in H_1} P(\vartheta|H_1)P(x|\vartheta, H_1)d\vartheta}{\int_{\vartheta\in H_0} P(\vartheta|H_0)P(x|\vartheta, H_0)d\vartheta}. \quad (2)$$

For the case of two competing point hypotheses H_0 and H_1, the Bayes factor collapses into the likelihood ratio of the two hypotheses:

$$L(H_1, H_0, x) = \frac{P(x|\vartheta = \vartheta_1)}{P(x|\vartheta = \vartheta_0)}. \tag{3}$$

It is easy to check that both (2) and (3) conform to the SRP and remain unaffected by stopping rules.[6] Furthermore, Lele (2004) has shown that in comparing point hypotheses, the likelihood ratio is the only measure of evidence that satisfies a number of reasonable invariance conditions.[7]

The preceding arguments have dealt with the evidential, postexperimental irrelevance of experimental design. Now we have to integrate this position into a decision-theoretic framework and defend it against attempts to render it incoherent. Furthermore, we have to explore why stopping rules often appear to be relevant and whether they are preexperimentally relevant, that is, relevant for responsible and efficient planning of an experiment.

3. Coherent Testing: A Decision-Theoretic Argument. How can frequentist statisticians respond to the charge? Usually, they aim at a *reductio ad absurdum* of the evidential irrelevance of stopping rules (e.g., Mayo 1996; Mayo and Kruse 2001); that is, they try to beat the Bayesians and their allies in their own game. One example is the following: Medical scientists conduct a phase II trial, that is, a trial with 100–300 participants that test the efficacy of a newly invented drug. If the drug proves to be effective in the phase II trial, a large-scale randomized controlled (phase III) trial will take place. Because of the costs of the experiments, the desire for subsequent funding, pressure from the pharmacy industry, and so forth, the scientists would be happier with a significant result (i.e., rejecting the null hypothesis that the new drug is not effective) than an insignificant one. Thus, upon learning that the collected results do not achieve the required significance level, our scientists decide to sample on and to include new patients. Finally, they obtain a result that, *if reported as a fixed sample size experiment*, would move the test into phase III. Shouldn't we be suspicious about such a move? Isn't such a conclusion less trustworthy than a conclusion drawn from the same data, but achieved in an "honest" way, without any interim decisions? Mayo writes:

> the try-and-try-again method allows experimenters to attain as small

[6] Note that this does not hold for *improper priors* where the integral over the probability density is not equal to one; see Mayo and Kruse 2001.

[7] A prima facie counterargument against Bayes factors consists in the "subjectivity" of the prior probabilities in $B(H_1, H_0, \cdot)$. But priors can be reported separately and disentangled from the "impact of the evidence."

> a level of significance as they choose (and thereby reject the null hypothesis at that level), even though the null hypothesis is true. (1996, 343)

Because of the dodgy way in which the conclusion was achieved, frequentist statisticians are ostensibly justified to assert that such data do not provide genuine evidence against the null hypothesis, whereas Bayesians are allegedly unable to detect that the experiment was biased toward a particular conclusion (see the discussion in Savage 1962). Evidently, the above example can be easily transferred to other testing problems in science.

Counterexamples of the above type raise two kinds of worries. The first is a *pre-experimental* one, namely, that certain stopping rules inevitably drive our inference into a particular direction. Hence, Bayesians apparently neglect bias and manipulation as a source of impressively high posterior probabilities. This worry is addressed by the results of Kadane, Schervish, and Seidenfeld (1996), who prove that the posterior probability of a hypothesis cannot be arbitrarily manipulated. If we stop an experiment if and only if the posterior of a hypothesis rises above 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 and to appraise a wrong hypothesis, or to discredit a true hypothesis, come what may.

Of course, this does not mean that Bayesians should deny the importance of experimental design. By contrast, when each single sample comes at a certain cost (such as in medical trials where surveillance is expensive), Bayesians and frequentists alike have to design the experiment in a way that the expected sample size is minimized. Indeed, a huge pile of literature deals with designing sequential experiments, from both Bayesian and frequentist perspectives (e.g., Wald 1947; Armitage 1975; Berry 1987). So both sides are well advised to affirm the *pre-experimental* relevance of stopping rules and also of error probabilities. The crucial question is the *postexperimental* issue: once we have observed the data, do we gain anything from learning the stopping rule according to which they have been produced?

To decide the question, note that in science, hypothesis testing is used to substantiate *decisions* of all kinds, such as establishing a working hypothesis for further research, moving a trial into the next stage, or approving a new medical drug. Thus, we should adopt a decision-theoretic perspective where gains and losses for right and wrong decisions, and the risk of various testing strategies, are taken into account. As I do not want to beg the question, I focus on a frequentist understanding of risk with respect to hypothesis tests and decision rules. In the remainder of the section I demonstrate that not the Bayesians but the frequentists are beaten in their own game. Let me elaborate.

In testing hypotheses and making decisions, frequentists rely on prespecified error probabilities. In particular, they specify the level of the type I error—the probability of erroneously rejecting the null hypothesis (e.g., $\alpha = .05$)—and aim at the most powerful test (i.e., the test with the lowest type II error) at this level. This gives a decision rule for accepting or rejecting the null hypothesis. In particular, upon learning that stopping rule τ_1 was used, frequentist inference interprets the data as produced by the statistical model induced by τ_1. In other words, the frequentist hypothesis test—usually the most powerful test at level α—and the associated decision rule are based on calculations in that model, and vice versa if they learn that τ_2 was used, and so forth. Such a stopping rule–sensitive procedure is, in the frequentist understanding, *preferred* to a procedure that interprets the data as generated by an arbitrary stopping rule. I take this to be the canonical way to phrase the postexperimental relevance of stopping rules in frequentist terms (see Schervish, Seidenfeld, and Kadane 2002 on fixed-level testing). In the calculations below, this standpoint is expressed in the decision rule δ_S.

Now, assume that the following conditions are met:

1. Let $\vartheta \in \mathbb{R}^m$ be the parameter of interest, with $H_0 : \vartheta \in \Theta_0 \subset \mathbb{R}^m$ and $H_1 : \vartheta \in \Theta_1 \subset \mathbb{R}^m$. Let $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\vartheta, \vartheta \in \Theta_0 \cup \Theta_1)$ be the corresponding statistical model, with observed data $x \in \mathcal{X}^n$.
2. Let S_x be the set of noninformative stopping rules τ such that, $\forall$ $y \in \mathcal{X}^\infty$, if $y_i = x_i \; \forall \; i \leq n$, then $\tau(y) = n$. In other words, S_x is the set of (noninformative) stopping rules that could have been used to generate the data x.
3. Let μ be a probability measure on $(S_x, \mathcal{B})$, and let $\delta_S : S_x \rightarrow \{0, 1\}$, for each $\tau \in S_x$, be the following 0-1 decision rule: H_0 is rejected if and only if H_1 passes, conditional on observed data x, an α-level significance test against H_0 in the model $(\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}^\tau_{\Theta_0}, \mathbb{P}^\tau_{\Theta_1})$.
4. Let δ_τ be a 0-1 decision rule that interprets data x invariably as a result of an experiment with stopping rule τ and rejects H_0 if and only if H_1 passes, conditional on observed data x, an α-level significance test against H_0 in the model $(\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}^\tau_{\Theta_0}, \mathbb{P}^\tau_{\Theta_1})$. Since τ is treated as a constant, either $\delta_\tau = 0$ or $\delta_\tau = 1$.
5. Let $L = (l_{ij})_{i,j \in \{0,1\}}$ be the loss matrix—l_{ij} being the loss suffered by opting for H_i when H_j is true—with $l_{00} < l_{10}$ and $l_{01} > l_{11}$.

Proposition. Assume conditions 1–5. Let $R(\vartheta, \cdot)$ be the frequentist risk of a decision rule, understood as the expected loss if ϑ happens to be the true parameter. Then:

- For each $\tau \in S_x$, either $R(\vartheta, \delta_\tau) < R(\vartheta, \delta_S)$ $\forall \vartheta \in \Theta_0$ or $R(\vartheta, \delta_\tau) < R(\vartheta, \delta_S)$ $\forall \vartheta \in \Theta_1$.
- For each $\vartheta \in \Theta_0 \cup \Theta_1$, either $R(\vartheta, 0) < R(\vartheta, \delta_S)$ or $R(\vartheta, 1) < R(\vartheta, \delta_S)$.

Proof. For each $\tau \in S_x$, either $\delta_\tau = 0$ or $\delta_\tau = 1$. Assume first $\delta_\tau = 0$. Then let $\vartheta \in H_0$. Then $R(\vartheta, \delta_\tau) = l_{00}$ and

$$R(\vartheta, \delta_S) = \mu_0 l_{00} + (1 - \mu_0) l_{10},$$

where $\mu_0 := \mu\{\tau \in S_x | H_0$ is accepted on the basis of x in $\mathbb{P}^{\tau}_{\Theta_0}\}$;

$$\begin{aligned} R(\vartheta, \delta_S) - R(\vartheta, \delta_\tau) &= \mu_0 l_{00} + (1 - \mu_0) l_{10} - l_{00} \\ &= (1 - \mu_0)(l_{10} - l_{00}) \\ &> 0. \end{aligned}$$

Similarly, for $\delta_\tau = 1$, where we choose $\vartheta \in H_1$:

$$\begin{aligned} R(\vartheta, \delta_S) - R(\vartheta, \delta_\tau) &= \mu_0 l_{01} + (1 - \mu_0) l_{11} - l_{11} \\ &= \mu_0 (l_{01} - l_{11}) \\ &> 0. \end{aligned}$$

The second part of the proposition follows immediately. □

Corollary. Preferring δ_S over $\delta_\tau = 0$ and $\delta_\tau = 1$ leads to *incoherence*, for any value of ϑ, in the sense that a Dutch book (namely, a sure loss) can be construed against these preferences.

Proof. Follows straightforwardly from the second part of the proposition. Compare to the argument given in Section 5.1. of Schervish, Kadane, and Seidenfeld 2003.

Remark 1. The proposition sounds complicated, but it merely captures the intuitive conjecture that which decision rule minimizes the frequentist risk depends on the true value of ϑ. Also note that the result is *independent of* μ; that is, when we get to know the used stopping rule postdata, it does not matter whether this particular stopping rule was likely to be chosen at the outset.

Remark 2. The frequentist's dilemma bears a close relationship to the problem of testing a hypothesis at a fixed level when a random choice between different experiments is made (Cox 1958) or when

> the value of a nuisance parameter is unknown (Schervish et al. 2003). In both cases, sticking to fixed-level testing leads to incoherence.[8]

A practical application of this result is an experiment in which we are told the data but not the stopping rule. Assume that postmortem elicitation would take some time and effort. The above results tell us that if we decide to treat the data as generated by, for example, a fixed sample size experiment, we will do better than waiting for the true stopping rule to be reported for some values of ϑ, while doing worse for others. Thus, in a frequentist framework, there can be no general argument for taking into account the stopping rule, as opposed to neglecting it. More precisely, if a frequentist prefers stopping rule–sensitive fixed-level testing to a fixed-level test of H_0 with respect to arbitrary stopping rules, her set of preferences is incoherent. In the medical trial example given at the outset, this implies that caring for the actually used stopping rule (instead of treating the trial as, say, a fixed-sample experiment) *makes certain presuppositions on our beliefs about the drug efficacy* ϑ: for certain values of ϑ, the expected loss will decrease, while for others, it will increase. Thus, prior expectations on ϑ have to be formulated to decide between both options. But these kinds of expectations on ϑ (such as prior distributions) are what frequentist statisticians or philosophers of statistics, by the very nature of their approach, want to avoid.[9]

Bayesians, on the other hand, avoid these troubles by assessing evidence in terms of Bayes factors and posterior probabilities that are not at all affected by stopping rules. Hence, the *practical* argument against the post-experimental relevance of stopping rules from Section 2 obtains a *theoretical*, decision-theoretic vindication.

4. Evaluation: A Philosopher's Conclusion. The debate about the relevance of experimental design and stopping rules is blurred by the lack of clarity about which kind of relevance is meant. Equivocation and confusion result. Moreover, the debate is characterized by a mutual deadlock. To resolve it, I have suggested to distinguish pre- and postexperimental relevance and to choose a position that corresponds to the practical needs

[8] Teddy Seidenfeld reminded me that δ_S is an *inadmissible* (dominated) decision rule, in the sense that a test that is *randomized* over the elements of S_x could achieve a lower type II error than δ_S, while maintaining the same type I error level α (see Cox 1958). However, since μ is in general unknown, this remains a result of purely theoretical interest.

[9] Frequentists, while conceding that their decision rule is strictly speaking incoherent, might maintain that it is at least *risk averse* in the following sense: the expected loss of δ_S will always figure between the expected losses of $\delta_\tau = 0$ and $\delta_\tau = 1$. This argument will be pursued in further work.

of empirical science. Such a position has to reject the postexperimental, evidential relevance of stopping rules: First, such a standpoint would yield measures of evidence that are easily manipulable, without any means of control on behalf of scientific institutions. Thus, such measures of evidence cannot play their proper role in scientific communication. Second, such a standpoint would also lead to decision-theoretic incoherence. In particular, a frequentist who claims the postexperimental relevance of stopping rules cannot avoid referring to prior expectations on the unknown parameter, undermining the very foundations of frequentist inference.

The valid core of the frequentist argument is the pre-experimental relevance of stopping rules as a means of providing for efficient, cost-minimizing sampling. The lack of disentanglement between both concepts of relevance has obfuscated the debate and led to the belief that stopping rules should matter postexperimentally, too. This belief is, however, fallacious. Hence, experimental design—and, in particular, the design of stopping rules—remains indispensable for scientific inference, but in a more narrow sense than frequentist statisticians and philosophers of science believe.

REFERENCES

Armitage, Peter (1975), *Sequential Medical Trials*. Oxford: Blackwell.

Berger, James O., and Donald A. 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*. New York: Springer, 29–72.

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

Berry, Donald A. (1987), "Statistical Inference, Designing Clinical Trials, and Pharmaceutical Company Decisions", *Statistician* 36: 181–189.

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

Cox, David R. (1958), "Some Problems Connected with Statistical Inference", *Annals of Mathematical Statistics* 29: 357–372.

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

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

Howson, Colin, and Peter Urbach (2006), *Scientific Reasoning: The Bayesian Approach*. 3rd ed. La Salle, IL: Open Court.

Kadane, Joseph B., Mark J. Schervish, and Teddy Seidenfeld (1996), "When Several Bayesians Agree That There Will Be No Reasoning to a Foregone Conclusion", in Lindley Darden (ed.), *PSA 1996: Proceedings of the 1996 Biennial Meeting of the Philosophy of Science Association*, vol. 1. East Lansing, MI: Philosophy of Science Association, S281–S289.

Lele, Subhash (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*. Chicago: University of Chicago Press, 191–216.

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

Mayo, Deborah G., and Michael Kruse (2001), "Principles of Inference and Their Consequences", in D. Cornfield and J. Williamson (eds.), *Foundations of Bayesianism*. Dordrecht: Kluwer, 381–403.

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

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

Savage, Leonard J. (1962), *The Foundations of Statistical Inference: A Discussion*. London: Methuen.

Schervish, Mark (1995), *Theory of Statistics*. New York: Springer.

Schervish, Mark J., Joseph B. Kadane, and Teddy Seidenfeld (2003), "Measures of Incoherence: How Not to Gamble If You Must", in J. Bernardo et al. (eds.), *Bayesian Statistics 7: Proceedings of the 7th Valencia Conference on Bayesian Statistics*. Oxford: Oxford University Press, 385–402.

Schervish, Mark J., Teddy Seidenfeld, and Joseph B. Kadane (2002), "A Rate of Incoherence Applied to Fixed-Level Testing", *Philosophy of Science* 69 (Proceedings): S248–S264.

Wald, Abraham (1947), *Sequential Analysis*. New York: Wiley.