How much evidence should one collect?

Heesen, Remco

Published in:
Philosophical Studies

DOI:
10.1007/s11098-014-0411-z

Publication date:
2015

How Much Evidence Should One Collect?*

Remco Heesen†

October 19, 2014

Abstract

A number of philosophers of science and statisticians have attempted to justify conclusions drawn from a finite sequence of evidence by appealing to results about what happens if the length of that sequence tends to infinity. If their justifications are to be successful, they need to rely on the finite sequence being either indefinitely increasing or of a large size. These assumptions are often not met in practice. This paper analyzes a simple model of collecting evidence and finds that the practice of collecting only very small sets of evidence before taking a question to be settled is rationally justified. This shows that the appeal to long run results can be used neither to explain the success of actual scientific practice nor to give a rational reconstruction of that practice.

Keywords: philosophy of science; evidence; rational choice; formal epistemology; Bayesian epistemology; sequential decision problems

*This paper has been accepted by *Philosophical Studies*. The final publication is available at Springer via `http://dx.doi.org/10.1007/s11098-014-0411-z`. Thanks to Kevin Zollman, Kevin Kelly, Liam Bright, Adam Brodie, and an anonymous referee for valuable comments and discussion.

†Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: `rheesen@cmu.edu`.

1 Introduction

An important question in the philosophy of science concerns the relation between hypothesis and evidence. Positive evidence is variously said to support, confirm, or prove a hypothesis, whereas negative evidence may detract from, disconfirm, or refute it. But what exactly these relations consist in remains an open question (Hempel 1945a,b, Popper 1959, Howson & Urbach 1989).

A number of approaches set up the problem as follows. The scientist is faced with a potentially infinite sequence of evidence. At any given time she has observed finitely many pieces of evidence, which is not sufficient to pronounce on the hypothesis with certainty. The scientist may be forced to act; what should she do?

On this setup, it is tempting to define successful methods as those that get it right in the limit. Among philosophers of science, Reichenbach and Peirce are perhaps the most prominent proponents of this line of thinking. Reichenbach (1938) attempts to address the problem of induction by comparing predictive methods based on their long run behavior. Peirce (1931 [1878]) even goes so far as to define success in terms of the long run results of certain methods of inquiry.

I discuss Reichenbach and Peirce in more detail in section 2. But they are not the only ones to make recommendations and comparisons based on limits of infinite sequences. Statisticians both on the frequentist and the Bayesian side use laws of large numbers and Central Limit Theorems to justify conclusions from finite sets of evidence (Casella & Berger 2001, Earman 1992). And formal learning theorists compare learning methods based on their performance on an infinite sequence (Kelly 1996).

What all of the above have in common is that limiting results are used to make or justify claims about the present. Friedman (1979, section I) argues that such long run justifications are useless to science, as they never provide any guarantees on the truth or approximate truth of science's current results. "[Long run justifications] do not and cannot show that scientific method tends

to produce true theories in actual practice." (Friedman 1979, p. 368)

How can a defender of long run justifications respond to this? How can limiting results and claims about the present be linked?

The link cannot be that there will come a time at which the entire infinite sequence of evidence has been observed. This is of course impossible (and none of the authors mentioned suggest it): various practical constraints put a finite upper bound on the amount of evidence a scientist (or all of science) could obtain.

But then what makes results concerning infinite evidence relevant to scientific method or practice? Two suggestions have been given: either (i) the limiting result is relevant because scientists collect evidence indefinitely, thus getting ever closer to the limit, or (ii) it is relevant because the amount of evidence collected (while finite) is large, so that the limiting result holds approximately. As I show in section 2, Reichenbach and Peirce rely on these suggestions.

The problem is that scientists do not act in accordance with either (i) or (ii). Often a single experiment (or a small number of replications) is taken to decide a question. I will illustrate this with a few cases in section 3. But the main point of this paper is to show that the practice of gathering small amounts of evidence is (at least sometimes) rational.

I show first that scientists should not collect evidence indefinitely in a fairly general model (section 4). Next I obtain some more specific results in a model where evidence takes the form of Bernoulli trials, showing that in most circumstances it is rational to do either zero, one, or a small handful of experiments (section 5).

Sections 4 and 5 together show that there are cases where suggestions (i) and (ii) both fail to provide a justification for appealing to long run results. As long as no other suggestions are forthcoming, any explanation or rational reconstruction of scientific practice that relies on long run results fails. This forces us to rethink the arguments of many philosophers of science and statisticians.

2 Philosophers Appealing to the Long Run

Reichenbach (1938, § 39) is a classic example of a philosopher using long run results to justify short run behavior. He characterizes the problem of induction as the attempt to extrapolate from a finite sample the limiting relative frequency of a certain type of event in an infinite sequence of events. His proposal is to use the current relative frequency (based on the finite sample) as an estimate of the limit.

His justification for this proposal depends on a convergence result: if the size of the sample is made large enough the relative frequency in the sample must approach the limit. Reichenbach admits that one can never be sure if the current size of the sample is large enough to be close to the limit. Moreover, I will show in section 5 that it can be irrational to collect a sample of large size.

He replies: "We are not bound to stay at [the current sample size]; we may continue our procedure and shall always consider the last [relative frequency] obtained as our best value" (Reichenbach 1938, p. 351). So for Reichenbach it is fine if the sample is not very large, as long as one keeps increasing its size indefinitely. I address this argument in section 4.

Reichenbach's approach to the problem of induction is not obscure. Among contemporary philosophers, Kelly (1991, 1996) and Schurz (2008) have defended similar views, both explicitly drawing on Reichenbach.

Similarly, Bayesian philosophers of science have used limiting results to address the criticism that their views depend on arbitrary prior probabilities. They point out that two scientists with different priors will eventually find themselves with closely agreeing posteriors, assuming enough evidence is collected (see Edwards et al. 1963 for a classic defense of this position and Earman 1992, chapter 6 for critical discussion). This kind of argument is also vulnerable to the issues I raise in sections 4 and 5.

Peirce represents a somewhat different example of a philosopher appealing to the long run. He defines to be true (in the present) that which would be believed at the limit of inquiry. But he does not want to be a relativist, so

he rejects the notion that truth depends on the beliefs of actual people.

> Our perversity and that of others may indefinitely postpone the settlement of opinion (...). Yet even that would not change the nature of the belief, which alone could be the result of investigation carried sufficiently far. (Peirce 1931 [1878], p. 5.408)

So the right (not "perverse") way of acquiring true belief is to carry investigation sufficiently far. What is sufficiently far? Either some given (large) finite amount suffices, or one simply needs to keep going indefinitely. The results from my model in sections 4 and 5 show that rational scientists may fail to do either.

Before discussing what is rational, I will discuss some cases from the history of science. In these cases, scientists took a single experiment or study to be decisive for some hypothesis, contrary to the requirements of the philosophers I just discussed.

3 Scientists Working in the Short Run

On December 29, 1849, Hermann von Helmholtz performed his first experiment to measure the speed of the nervous impulse (Olesko & Holmes 1993, p. 88). He hooked a muscle from a frog's leg and some of the nerve attached to it onto his newly invented experimental apparatus (later to become known as the myograph), stimulated the nerve, and measured the time until the muscle contracted. By varying the location of the stimulation (nearer or farther from the muscle) and observing the difference in reaction time he obtained an estimate of the speed at which the signal propagates along the nerve.

Helmholtz repeated the experiment on January 4 and January 6, 1850, using different frogs and varying the weight that the contracting muscle was lifting. Over the course of the three days, the values he found for the speed of the nervous impulse varied between 24.6 and 38.4 m/s.

After this, rather than doing further experiments to get a more precise estimate of the speed or of the conditions that influence it, he wrote up his results and announced them to his peers. The first presentation occurred on January 21, little more than three weeks after the first experiment (Olesko & Holmes 1993, fn. 86). The report was published in a number of venues over the next few months. Edwin Boring, a historian of psychology, describes its impact as follows.

> Every one thought (...) of his hand as of a piece with himself. To move his finger voluntarily was an act of mind itself, not a later event caused by a previous act of mind. To separate the movement in time from the event of will that caused it was in a sense to separate the body from the mind (...). Helmholtz's discovery was a step in the analysis of bodily motion that changed it from an instantaneous occurrence to a temporal series of events. (Boring 1950, p. 42)

So Helmholtz's experiment turned the accepted idea that sensation and bodily motion are essentially instantaneous events into an untenable view. While further experiments would be done later by Helmholtz and others, this first set of results was sufficient to prove that the speed of the nervous impulse was measurable, with Johannes Müller calling it "a great stride" and Alexander von Humboldt "a noteworthy discovery" that had "stimulated astonishment" (Olesko & Holmes 1993, pp. 90–91). It is an especially dramatic example of a single experiment settling a hypothesis, since the result is so different from what most scientists at the time would have expected.

It might be asked whether this really represents a single experiment. How much evidence did Helmholtz actually collect before publishing his discovery?

During three days of experimental work, he used six different frog muscles (one frog each day) and performed a total of 89 measurements of the reaction time. This is not the same as a measurement of the speed of the nervous impulse, as that is obtained by comparing the difference in reaction time

when different parts of the nerve are stimulated. Helmholtz reports the 89 measurements in five groups, obtaining only five estimates of the speed (von Helmholtz 1850, pp. 339-344).

Whether this should count as one, five, six, or 89 experiments is perhaps a subjective matter. But in any case, a small set of evidence was used to settle the hypothesis that the speed of the nervous impulse is measurable, and both the experiment and the conclusion stand as paradigm cases of good science to this day.

Other examples abound. Perrin's experiment in 1908 was the first to confirm Einstein's theoretical work on Brownian motion. For physicists, this settled the question whether atoms and molecules existed. In 1668, Redi disproved the commonly accepted hypothesis that maggots arise spontaneously from rotting meat with a simple experiment. He put some rotting meat in two jars, covering one but not the other, and observed a few days later that maggots appeared only on the latter.

The model to be introduced and analyzed in the next sections shows that it is sometimes rational for scientists to take a single study or experiment as decisive evidence for or against a hypothesis. The cases above are among many that suggest that real scientists are sometimes willing to do this. But they are merely illustrative: the argument in the rest of this paper does not rely on the correctness of my analysis of these cases.

4 Should Scientists Collect Evidence Indefinitely?

Consider a scientist who wants to know whether some hypothesis h is true or false. For example, Helmholtz wanted to test the hypothesis that bodily motion is instantaneous. Evidence about h takes the form of experimental outcomes (represented in the model by numerical values) which have probabilities associated with them that depend on the truth-value of h. That is, a piece of evidence is a realization of a random variable that follows some

given distribution X. Realizations of X are assumed to be independent given either h or $\neg h$, so any collection of evidence is independent and identically distributed (i.i.d.; the role of this assumption is discussed at the end of this section).

At a cost $c > 0$, the scientist gains one piece of evidence (i.e., one realization of X). Think of a piece of evidence as an experiment done by the scientist, although it may also reflect what she learns by reading or talking to other scientists.

The cost c may represent the costs of buying equipment or paying researchers' wages. Especially relevant in the present context, however, is the opportunity cost: whatever time the scientist spends collecting evidence related to h is time she fails to spend on other questions. The cost is thus (at least partially) an epistemic cost: it reflects the knowledge about other questions lost to the scientist because she chose to gain knowledge about h.

The scientist collects evidence sequentially. That is, the decision whether or not to collect a next piece of evidence may depend on what is learned from previous pieces of evidence.

Whenever the scientist decides to stop collecting evidence, she has to choose one of two terminal decisions: d_1 represents the decision to believe that h is true (and to act on that belief when appropriate), and d_2 represents the decision to believe that h is false. For example, after three days, Helmholtz decided that no more experiments were needed: bodily motion was not instantaneous.

The scientist is faced with a trade-off. Collecting more evidence reduces the chance of drawing the wrong conclusion about the truth-value of h (as represented by the terminal decision), but increases the accumulated costs. Collecting less evidence reduces costs, but increases the chance of drawing the wrong conclusion about h.

What should a rational scientist do in this situation? I assume the scientist acts as if she were a Bayesian statistician (I will argue later on that my results will be the same if she acts like a frequentist statistician instead).

This means the following.

The scientist has a subjective probability $\xi \in [0, 1]$ that reflects how likely she thinks it is that h is true. In response to evidence she updates this probability using Bayes' rule. She has a loss function that puts a numerical value on each decision. And she makes decisions that minimize risk, where the risk is the expected value (relative to her subjective beliefs) of the loss plus costs.

In this model, the loss ℓ associated with the terminal decision is zero if the decision is "correct" (d_1 if h and d_2 if $\neg h$), and $\beta > 0$ if the decision is "incorrect" (d_2 if h and d_1 if $\neg h$, see table 1). The total loss is ℓ plus the number of pieces of evidence collected times c.

$\ell(\cdot, \cdot)$	h	$\neg h$
d_1	0	β
d_2	β	0

Table 1: The loss function ℓ.

The risk function ρ gives the expected value of the total loss: $\rho(\xi, \delta)$ denotes the expected total loss relative to a scientist's subjective probability ξ if she chooses the sequential decision procedure δ.

A sequential decision procedure δ specifies at each decision point whether the scientist collects an additional piece of evidence (and which terminal decision to choose if she does not) as a function of the evidence obtained so far. Let $d(\delta)$ denote the terminal decision and $N(\delta)$ the number of observations taken under δ.

The problem that the scientist needs to solve is that of finding an optimal sequential decision procedure or optimal stopping rule (where optimal means minimizing the risk function ρ). DeGroot (2004, sections 12.14–12.16) provides an analysis of this situation.

Let f_1 denote the likelihood function associated with a single realization of X if h is true, and f_2 the likelihood function if h is false. So f_1 and f_2

are probability density functions if X is continuous and probability mass functions if X is discrete. Let

$$Z_i := \log \frac{f_2(X_i)}{f_1(X_i)}.$$

Consider the sequential decision procedure $\delta(a, b)$ that, for any n, takes an $n + 1$-st observation if

$$a < \sum_{i=1}^{n} Z_i < b,$$

and stops otherwise. A procedure of this form is known as a "sequential probability-ratio test". The optimal procedure for the sequential decision problem described above takes this form, unless it is optimal to take no observations at all.

Proposition 1 (Wald & Wolfowitz (1948))**.** *The optimal sequential decision procedure among those that take at least one observation is* $\delta(a, b)$ *for some* $a < 0$ *and* $b > 0$.

This shows that a sequential probability-ratio test is optimal for a Bayesian scientist. What if the scientist is a frequentist instead? In that case she disavows priors and is instead interested in controlling the error probabilities $\Pr(d(\delta) = d_1 \mid \neg h)$ and $\Pr(d(\delta) = d_2 \mid h)$ directly. Since observations are costly, she wants to do so with a minimal number of observations.

Theorem 2 (Wald & Wolfowitz (1948))**.** *Let* $a < 0$ *and* $b > 0$. *Let* δ_0 *be a sequential decision procedure that takes at least one observation. If*

$$\Pr(d(\delta_0) = d_2 \mid h) \leq \Pr(d(\delta(a, b)) = d_2 \mid h),$$
$$\Pr(d(\delta_0) = d_1 \mid \neg h) \leq \Pr(d(\delta(a, b)) = d_1 \mid \neg h),$$

then

$$\mathbb{E}[N(\delta(a, b)) \mid h] \leq \mathbb{E}[N(\delta_0) \mid h],$$
$$\mathbb{E}[N(\delta(a, b)) \mid \neg h] \leq \mathbb{E}[N(\delta_0) \mid \neg h].$$

If the evidence follows a continuous probability distribution, the error levels of a sequential probability-ratio test $\delta(a, b)$ vary continuously as a function of a and b. Thus, no matter what the desired error probabilities are, some sequential probability ratio-test achieves them exactly. By theorem 2 it does so with a minimal number of observations. Thus the optimal sequential decision procedure for a frequentist is $\delta(a, b)$ for some $a < 0$ and $b > 0$.

If the evidence follows a discrete probability distribution, there might not be a sequential probability-ratio test that achieves the desired error probabilities exactly. However, Wald (1947, section 3.3) provides a way of determining a and b such that $\delta(a, b)$ approximates the desired error probabilities. Any test that is deemed superior to $\delta(a, b)$ must achieve either comparable error probabilities with fewer observations, lower error probabilities with a comparable number of observations, or lower error probabilities with fewer observations. In any case, the expected number of observations for $\delta(a, b)$ can act as an upper bound (at least approximately) for the number of observations taken by the test actually chosen. This means that the test actually chosen by the frequentist inherits the features of $\delta(a, b)$ that I investigate below and in section 5.

So regardless of whether the evidence is continuous or discrete, the Bayesian nature of my analysis does not involve a substantive assumption.

Note that the sequence $Z_1, Z_2, \ldots$ is i.i.d. because the sequence $X_1, X_2, \ldots$ is i.i.d. Unless one is in the trivial case where the likelihood functions f_1 and f_2 are identical almost everywhere (which would mean that $Z_i = 0$ with probability 1) this is sufficient to guarantee that $\delta(a, b)$ will terminate with probability 1.

Theorem 3 (Stein (1946))**.** *Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. random variables with $\Pr(Z_i = 0) < 1$. It follows that $\Pr(N(\delta(a, b)) < \infty) = 1$ for all $a < 0$ and $b > 0$.*

Proposition 1 and theorem 3 together establish that the optimal sequential decision procedure either takes no observations at all or it takes a finite number of observations with probability 1. In either case it is established that

the scientist should not collect evidence indefinitely: for a given hypothesis, there always comes a time at which it is better to stop collecting evidence. This runs contrary to Reichenbach and Peirce's claim that good scientists will continue to gather more evidence for their hypotheses indefinitely.

To establish this result I have relied on the assumption that the evidence gathered by scientists on a particular question is i.i.d. This assumption will not always be appropriate. Experimental results may not be independent when scientists build on each other's work or reuse experimental setups. They may not be identically distributed if different kinds of experiments are used to test the same hypothesis.

However, it is sufficient for my argument that some scientific investigations are well-described by an i.i.d. model. This shows that scientists are sometimes rational not to collect evidence indefinitely. As a result, the justification of their methodology, at least in these cases, cannot be as Reichenbach and Peirce suggest.

Moreover, it is not clear that the results established here (optimality of the sequential probability-ratio test and the fact that it terminates with probability 1) crucially depend on the assumption that evidence is i.i.d. Conditions under which these results generalize have been investigated by Lai (1981) and Liu & Blostein (1992).

5 Should Scientists Collect a Large Set of Evidence?

Given that the scientist will not collect evidence indefinitely, the next question is whether she will at least collect a large set of evidence. In that case results that apply to infinite sets of evidence may perhaps have approximate validity, so that appeals to such results by philosophers like Reichenbach may be justified that way.

In order to get the more specific results that are needed to answer this question, some assumption on the probability distribution of the evidence

needs to be made. Here I will assume that evidence about h is Bernoulli-distributed:

$$X \mid h \sim Ber(1-\varepsilon),$$
$$X \mid \neg h \sim Ber(\varepsilon),$$

for some given $\varepsilon \in (0, 1/2)$. So if h is true it is more likely that $X = 1$ than that $X = 0$, while if h is false this is reversed. As a result, $X = 1$ is positive evidence for h, and $X = 0$ is negative evidence, while neither type of evidence settles the truth-value of h conclusively.

The assumption that the evidence is Bernoulli-distributed is restrictive. However, the following three considerations should be kept in mind when evaluating this restrictiveness.

First, the assumption allows for a range of possibilities for the informational value of a piece of evidence. As ε approaches $1/2$ a piece of evidence provides almost no information for or against h. If ε is close to zero a single piece of evidence can settle the truth-value of the hypothesis with near-certainty. Intermediate values of ε can model any situation in between those two extremes.

Second, it seems unlikely that the particular form of the distribution drives the results I will obtain in this section. To illustrate this point, I will briefly consider evidence that follows a normal distribution at the end of this section.

Finally, recall that I am trying to show that it is sometimes rational for a scientist to gather a small amount of evidence. For my argument to work, it suffices that there exist some scientific hypotheses for which the evidence takes this form. I do not need to argue that all evidence is Bernoulli-distributed.

For this distribution the likelihood functions f_1 and f_2 are given by

$$f_1(x) = \Pr(X_i = x \mid h) = \varepsilon^{1-x}(1-\varepsilon)^x, \qquad x = 0, 1,$$
$$f_2(x) = \Pr(X_i = x \mid \neg h) = \varepsilon^x(1-\varepsilon)^{1-x}, \qquad x = 0, 1.$$

Plugging this into the definition of Z_i yields

$$Z_i = \log \frac{f_2(X_i)}{f_1(X_i)} = (1 - 2X_i) \log \frac{1-\varepsilon}{\varepsilon}.$$

Note that Z_i can take only two values: $\log \frac{1-\varepsilon}{\varepsilon}$ if $X_i = 0$ and $-\log \frac{1-\varepsilon}{\varepsilon}$ if $X_i = 1$. Thus, for any n, $\sum_{i=1}^{n} Z_i$ only takes values that are integer multiples of $\log \frac{1-\varepsilon}{\varepsilon}$.

By proposition 1 the optimal procedure is $\delta(a, b)$ for some $a < 0$ and $b > 0$ (unless it is optimal to take no observations at all). But now without loss of generality only integer multiples of $\log \frac{1-\varepsilon}{\varepsilon}$ need to be considered as possible values for a and b.

Proposition 4 (DeGroot (2004))**.** *Suppose the random variables Z_i can only take the values z and $-z$ for some z and $a < 0$ and $b > 0$ are integer multiples of z. Then the risk of the sequential decision procedure $\delta(a, b)$ is*

$$\begin{aligned}
\rho(\xi, \delta(a,b)) &= \beta\xi \Pr(d(\delta(a,b)) = d_2 \mid h) + \beta(1-\xi) \Pr(d(\delta(a,b)) = d_1 \mid \neg h) \\
&\quad + c\xi \mathbb{E}[N(\delta(a,b)) \mid h] + c(1-\xi)\mathbb{E}[N(\delta(a,b)) \mid \neg h] \\
&= \beta\xi \frac{1-e^a}{e^b - e^a} + \beta(1-\xi)\frac{e^a(e^b-1)}{e^b - e^a} + c\xi \frac{a(e^b-1) + b(1-e^a)}{(e^b - e^a)\mathbb{E}[Z_i \mid h]} \\
&\quad + c(1-\xi)\frac{ae^a(e^b-1) + be^b(1-e^a)}{(e^b - e^a)\mathbb{E}[Z_i \mid \neg h]}. \qquad (1)
\end{aligned}$$

From propositions 1 and 4 it follows that the optimal sequential decision procedure among those that take at least one observation is $\delta(a^*, b^*)$, where $a^* < 0$ and $b^* > 0$ are those integer multiples of $\log \frac{1-\varepsilon}{\varepsilon}$ that minimize (1). So I can restrict attention to procedures that take the form

$$\delta_{m,k} := \delta\left(-m \log \frac{1-\varepsilon}{\varepsilon}, k \log \frac{1-\varepsilon}{\varepsilon}\right),$$

where m and k are positive integers. To keep the notation uniform, I also take $\delta_{m,k}$ to be well-defined when m or k is non-positive. In this case no observations are taken and the scientist chooses a decision immediately.

Here is how to interpret a procedure of the form $\delta_{m,k}$ for positive m and k. The scientist should keep track of the following quantity: the number of X_i so far observed that took the value zero minus the number of X_i so far observed that took the value one. The procedure tells her to continue to take observations as long as that quantity is strictly between $-m$ and k. If the quantity hits $-m$ she stops taking observations and chooses decision d_1, and if it hits k she stops and chooses decision d_2.

Let g_k be defined by

$$g_k(\varepsilon) := \frac{(1-\varepsilon)^{2k+1} - \varepsilon^{2k+1}}{(1-2\varepsilon)^2 \varepsilon^k (1-\varepsilon)^k} + \frac{2k+1}{1-2\varepsilon},$$

for all non-negative integers k and $\varepsilon \in (0, 1/2)$. Since $g_k(\varepsilon)$ is increasing in k for any fixed ε, there is a unique k^* such that

$$g_{k^*-1}(\varepsilon) < \frac{\beta}{c} \leq g_{k^*}(\varepsilon),$$

unless $\beta/c \leq g_0(\varepsilon)$; in that case define $k^* = 0$ (see also table 2). The following results, proved in the appendix, specify the optimal sequential decision procedure in terms of k^*.

k^*	β/c
0	$(0, g_0(\varepsilon)]$
1	$(g_0(\varepsilon), g_1(\varepsilon)]$
$\vdots$	$\vdots$
k	$(g_{k-1}(\varepsilon), g_k(\varepsilon)]$
$k+1$	$(g_k(\varepsilon), g_{k+1}(\varepsilon)]$
$\vdots$	$\vdots$

Table 2: k^* is determined by finding an interval of the form $(g_{k-1}(\varepsilon), g_k(\varepsilon)]$ such that β/c is in that interval.

Proposition 5. *If $\xi = 1/2$, the optimal sequential decision procedure is δ_{k^*,k^*}.*

This proposition applies to a scientist who starts out thinking h is equally likely to be true or false. What if the scientist has a different prior?

Proposition 6. *Let $d \in \mathbb{Z}$. If*

$$\xi = \xi_d := \frac{\varepsilon^d}{\varepsilon^d + (1-\varepsilon)^d},$$

the optimal sequential decision procedure is δ_{k^*+d,k^*-d}.

Corollary 7. *For any $\xi \in (0,1)$ not covered by proposition 6 there must be a $d \in \mathbb{Z}$ such that $\xi_d < \xi < \xi_{d-1}$. The optimal sequential decision procedure for such a ξ is either δ_{k^*+d,k^*-d}, δ_{k^*+d-1,k^*-d+1}, δ_{k^*+d-1,k^*-d}, or δ_{k^*+d,k^*-d+1}.*

One can derive general inequalities to determine which of these four procedures is optimal for given values of ξ, β, c, and ε, but this is not important for my purposes here.

What proposition 6 and its corollary show is that independent of ξ a larger value of k^* indicates that more observations will be needed to come to a decision on the truth-value of h. The value of ξ biases the process towards one conclusion or the other but it does not change this general level k^*. I will focus on the value of k^* in the remainder of this section. This value depends on ε and the ratio β/c (see figure 1).

How does k^* respond to changes in the parameter values? All else being equal, if β increases or c decreases the scientist takes more observations before making a decision. These results are reasonable: a higher loss associated with a wrong decision gives the scientist an incentive to play it safe by taking many observations, while increased costs of observations encourage coming to a decision quickly, at the expense of a higher risk of a wrong decision.

Now consider the reliability of the evidence. If ε is close to 0 or 1/2, it is optimal to take at most one observation. In this case the evidence is either

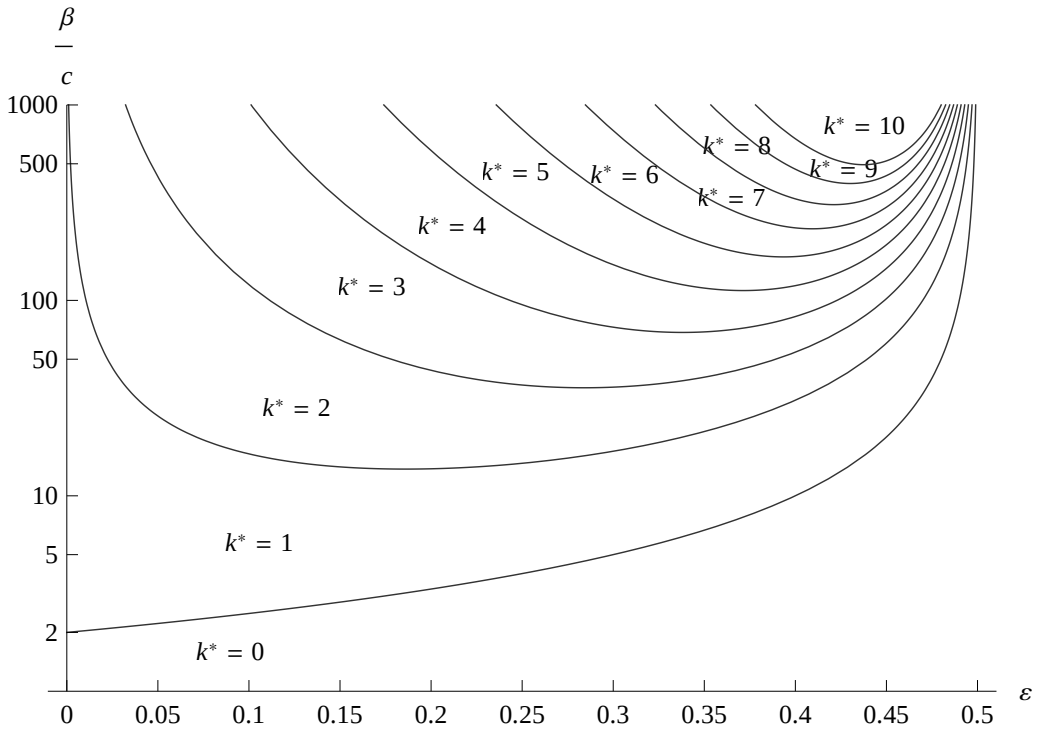

Figure 1: k^* as a function of β, c, and ε (for $\beta/c \leq 1000$ and $0 < \varepsilon < 1$). The lines are indifference curves as defined by the family of functions g_k. Note that the β/c-axis is logarithmic.

so good that it settles the hypothesis immediately, or so bad that it is not worth collecting.

For moderate values of ε, there is more interesting behavior. At values of β/c greater than 13.7, more complicated decision procedures than "decide immediately" or "take one observation and then decide" start appearing. It turns out that quite large values of β/c are needed before procedures that wait for a larger difference than a few between the number of observations favoring h's truth and h's falsity come into the picture. For instance, if $\beta/c \leq 100$, it is never optimal to wait for a larger difference than 4, whatever the value of ε (i.e., no matter how informative a single piece of evidence is

about the truth-value of h).

Are these results peculiar to the Bernoulli distribution? A brief investigation with the normal distribution suggests that they are not. Suppose that the hypothesis h and its negation both imply that the evidence is normally distributed, but they disagree about its mean:

$$X \mid h \sim N(\mu_1, \sigma^2),$$
$$X \mid \neg h \sim N(\mu_2, \sigma^2),$$

with $\mu_1 \neq \mu_2$ and $\sigma^2 > 0$ known.

Proposition 1 applies so the optimal sequential decision procedure is either $\delta(a, b)$ for some $a < 0$ and $b > 0$ or takes no observations at all. DeGroot (2004, section 12.16) provides a formula for approximating the expected number of observations $\mathbb{E}[N(\delta(a^*, b^*))]$ taken by the optimal procedure.

The value of $\mathbb{E}[N(\delta(a^*, b^*))]$ depends on the prior ξ, the difference between μ_1 and μ_2, the variance σ^2, and the value of β/c. However, if $\beta/c \leq 100$, then $\mathbb{E}[N(\delta(a^*, b^*))] < 14$ for all possible values of the other parameters. That is, even in the "worst case", the optimal procedure takes on average no more than fourteen observations.

The amounts of evidence collected if evidence follows a Bernoulli or normal distribution are so small that it seems implausible to claim that limiting results (e.g., the law of large numbers implicitly appealed to by Reichenbach 1938, § 39) may apply to them. Thus it is shown that there are cases (namely, those where either of the models of this section applies) where it is rational for scientists to stop collecting evidence after a small number of observations. This completes my argument against the use of long run behavior as a generic method of evaluation of scientific methods.

6 Conclusion

I started this paper by observing that various philosophers of science have been interested in identifying good methods based on their behavior as the amount of evidence collected goes to infinity (e.g., Reichenbach 1938, who recommends the straight rule for this reason). This makes sense only if scientists collect evidence either (i) indefinitely (so they will eventually come arbitrarily close to the limit) or (ii) in large finite amounts (so that the limit applies approximately).

Using some cases from the history of science, I showed that scientists do not always collect evidence indefinitely or in large finite amounts. I then analyzed a model of a scientist trying to learn the truth-value of a hypothesis by collecting evidence sequentially. The two conclusions were as follows.

First, the scientist should not collect evidence indefinitely, contra (i). There is some finite number such that it is better for the scientist to stop collecting evidence (perhaps so she can focus on some other question) after seeing that number of pieces of evidence.

Second, if it is assumed that the evidence follows a Bernoulli distribution, more specific conclusions about the amount of evidence that should be collected can be drawn. If the loss associated with a wrong decision is no higher than the cost of thirteen observations ($\beta \leq 13c$) then it is optimal to take no more than one observation. Even if the loss is as high as the cost of a hundred observations it is not optimal to wait for a difference larger than four between the number of observations favoring one conclusion and the number of observations favoring the other. So in these situations it is rational to collect only a small amount of evidence, contra (ii).

This suggests that only the most important hypotheses (where the consequences of having the wrong belief about it are many times worse than the costs of collecting additional evidence) merit extensive investigation. For less important hypotheses collecting a single piece of evidence (or simply guessing the truth-value based on no evidence at all) is the best strategy.

The analysis in this paper casts serious doubt on the kind of view in

philosophy of science that wants to justify scientific practices in terms of the features those practices would have when applied to an indefinitely increasing sequence of evidence. As the old saying goes: "In the long run we are all dead". I would add to this that scientists stop paying attention in the short run already, and are rational to do so. Our philosophy of science should reflect this fact.

A Proofs

From proposition 1 it follows that the optimal procedure that takes at least one observation takes the form $\delta(a, b)$, where a is a negative and b a positive integer multiple of $\log \frac{1-\varepsilon}{\varepsilon}$. If $\xi = 1/2$, the symmetry of the problem (the loss for a wrong decision β and the cost per observation c are the same whether h is true or false) implies that in the optimal solution $a = -b$. So I can restrict attention to procedures of the form

$$\delta_{k,k} := \delta\left(-k \log \frac{1-\varepsilon}{\varepsilon}, k \log \frac{1-\varepsilon}{\varepsilon}\right)$$

for some positive integer k. Note also that

$$\mathbb{E}[Z_i \mid \neg h] = (1-2\varepsilon) \log \frac{1-\varepsilon}{\varepsilon} = -\mathbb{E}[Z_i \mid h].$$

Applying equation (1) to $\delta_{k,k}$ yields

$$\rho\left(\frac{1}{2}, \delta_{k,k}\right) = \beta \frac{\varepsilon^k}{(1-\varepsilon)^k + \varepsilon^k} + c \frac{k}{1-2\varepsilon} \frac{(1-\varepsilon)^k - \varepsilon^k}{(1-\varepsilon)^k + \varepsilon^k}.$$

Note that $\rho(1/2, \delta_{0,0}) = \beta/2$ correctly gives the risk of the procedure that takes no observations. So the optimal procedure (without the caveat "among those that take at least one observation") is of the form $\delta_{k,k}$ for some non-negative integer k.

Next, fix a value of k and ask whether $\delta_{k+1,k+1}$ is better than $\delta_{k,k}$. Some algebra shows that $\rho(1/2, \delta_{k+1,k+1}) < \rho(1/2, \delta_{k,k})$ if and only if

$$\frac{\beta}{c} > g_k(\varepsilon) = \frac{(1-\varepsilon)^{2k+1} - \varepsilon^{2k+1}}{(1-2\varepsilon)^2 \varepsilon^k (1-\varepsilon)^k} + \frac{2k+1}{1-2\varepsilon}.$$

Note that $g_k(\varepsilon)$ is increasing in k, so either there is a unique positive integer k^* such that

$$g_{k^*-1}(\varepsilon) < \frac{\beta}{c} \leq g_{k^*}(\varepsilon),$$

or $\beta/c \leq g_0(\varepsilon)$; in that case set $k^* = 0$. In either case δ_{k^*,k^*} is the optimal sequential decision procedure. This proves proposition 5.

Now consider a prior of the form ξ_d for some $d \in \mathbb{Z}$ (where ξ_d is as defined in proposition 6). This might be called a conjugate prior for this decision problem: the posterior after conditioning on evidence X_1 is ξ_{d-1} if the evidence is $X_1 = 1$ and ξ_{d+1} if $X_1 = 0$.

Note that $\xi_0 = 1/2$ so the optimal sequential decision procedure for ξ_0 is δ_{k^*,k^*} by proposition 5. In light of the above this statement is equivalent to the following: it is optimal to continue taking observations as long as the posterior remains between ξ_{k^*-1} and ξ_{1-k^*}, and it is optimal to stop if the posterior is ξ_{k^*} or smaller, or ξ_{-k^*} or larger.

But the latter statement does not depend on the prior one started with. So for any prior ξ_d it is optimal to take observations if and only if the posterior remains strictly between ξ_{k^*} and ξ_{-k^*}. This is exactly the sequential decision procedure δ_{k^*+d,k^*-d} (which takes no observations if either $k^* + d \leq 0$ or $k^* - d \leq 0$). This proves proposition 6.

If $\xi_d < \xi < \xi_{d-1}$ then observing $X_i = 0$ k^*-d+1 times forces the posterior to be less than ξ_{k^*}, at which point it is optimal to stop taking observations. Observing $X_i = 0$ less than $k^* - d$ times forces the posterior to be larger than ξ_{k^*-1}, so continuing to take observations is optimal.

Similarly, observing $X_i = 1$ $k^* + d$ times forces the posterior to be greater than ξ_{-k^*}, and observing $X_i = 1$ less than k^*+d-1 times forces the posterior to be less than ξ_{-k^*+1}. Hence one of δ_{k^*+d,k^*-d}, δ_{k^*+d-1,k^*-d+1}, δ_{k^*+d-1,k^*-d}, or δ_{k^*+d,k^*-d+1} is the optimal sequential decision procedure. This proves the corollary.

References

Boring, E. G. (1950). *A History of Experimental Psychology*. New Jersey: Prentice-Hall, second edn.

Casella, G., & Berger, R. L. (2001). *Statistical Inference*. Belmont: Duxbury, second edn.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. New Jersey: John Wiley & Sons.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: MIT Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242.

Friedman, M. (1979). Truth and confirmation. *The Journal of Philosophy*, 76(7), 361–382.

Hempel, C. G. (1945a). Studies in the logic of confirmation (I.). *Mind*, 54(213), 1–26.

Hempel, C. G. (1945b). Studies in the logic of confirmation (II.). *Mind*, 54(214), 97–121.

Howson, C., & Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach*. La Salle: Open Court.

Kelly, K. T. (1991). Reichenbach, induction, and discovery. *Erkenntnis*, 35(1/3), 123–149.

Kelly, K. T. (1996). *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.

Lai, T. L. (1981). Asymptotic optimality of invariant sequential probability ratio tests. *The Annals of Statistics*, 9(2), 318–333.

Liu, Y., & Blostein, S. D. (1992). Optimality of the sequential probability ratio test for nonstationary observations. *IEEE Transactions on Information Theory*, 38(1), 177–182.

Olesko, K. M., & Holmes, F. L. (1993). Experiment, quantification, and discovery: Helmholtz's early physiological researches, 1843–50. In D. Cahan (Ed.), *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*, chap. 2 (pp. 50–108). Berkeley: University of California Press.

Peirce, C. S. (1931 [1878]). How to make our ideas clear. In C. Hartstone & P. Weiss (Eds.), *Collected Papers of Charles Sanders Peirce*, vol. V (pp. 5.388–5.410). Cambridge: Harvard University Press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: The University of Chicago Press.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, 75(3), 278–305.

Stein, C. (1946). A note on cumulative sums. *The Annals of Mathematical Statistics*, 17(4), 498–499.

von Helmholtz, H. L. F. (1850). Messungen über den zeitlichen Verlauf der Zuckung animalischer Muskeln und die Fortpflanzungsgeschwindigkeit der Reizung in den Nerven. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, pp. 276–364.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3), 326–339.