Statistical Inference: Small Probabilities and Errors

Daniel C. Sloughter
Furman University

1 The Howland trial

In the 1867 trial contesting the will of Sylvia Ann Howland in New Bedford, Massachusetts, the executor alleged that one of Howland's signatures on a a will written in 1862 was a tracing of her signature from an earlier page of the will. Benjamin Peirce and his son Charles studied 42 of Sylvia Howland's signatures, from which they estimated the probability that corresponding "downstrokes" would agree in both length and position to be $\frac{5,325}{25,830}$. Since her signature included 30 such "downstrokes," and the disputed signature agreed in all 30 cases with the undisputed signature, the elder Peirce concluded that the probability that the two signatures would agree in such a way if they had been written independently by Sylvia Howland herself was

$$q = \left(\frac{5,325}{25,830}\right)^{30} \approx 2.67 \times 10^{-21}.$$

In his comments, Peirce noted:

> This number far transcends human experience. So vast an improbability is practically an impossibility. Such evanescent shadows of probability cannot belong to actual life. They are unimaginably less than those least things which the law cares for. (Meier and Zabell 1980, 499)

This statement is very similar to what Émile Borel calls the *single law of chance*: "Phenomena with very small probabilities do not occur" (Borel 1962, 1). That is, if the signature in question is indeed that of Sylvia Howland, then we must accept that an event of exceedingly small probability has occurred. Since such events, at least in this context, are "practically an impossibility," we should seek elsewhere for an explanation of the similarity between the two signatures.

Leaving aside questions about the validity of the probabilistic model employed by the Peirces in this case, Meier and Zabell remark that the probability q which they compute may be misinterpreted as the posterior probability of the hypothesis that Sylvia Howland wrote both signatures.* This is in keeping with their subjectivist view of probability. From this perspective, a probability represents the strength of one's opinion about the truth of a given proposition. Given a hypothesis H, in this case, the statement that Sylvia Howland wrote both signatures, a prior probability for H and all alternatives to H, and data D, one computes $P(H|D)$, the posterior probability of H given the evidence D. Since this computation is based on Bayes' theorem, those who use a subjectivist interpretation of probability in statistical inference are referred to as Bayesians.

* Although they do note that, in this case, with any reasonable prior probabilities, the posterior probability will be of the same order of magnitude as the computed probability.

Those who hold a frequentist view of probability, that is, the view that a statement of probability is always a reference to the frequency of occurrence of some event, call q the p-value for the test of the hypothesis H. Bayesians have criticized the use of p-values in statistical inference on many fronts, the one mentioned by Meier and Zabell being one of the weakest (we would have to abandon much of mathematics and statistics if we could use only those parts which are seldom misinterpreted). Less commonly discussed are the disagreements among frequentists themselves on how to interpret p-values.

2 Frequentist probabilities: Neyman

A frequentist interpretation of probability requires that the probability of an event A refer in some way to the proportion of times the event A occurs in a sequence of repetitions of a specified experiment. However, frequentists differ on what they accept as repetitions. For Jerzy Neyman and E. S. Pearson, the founders of the principal school of modern frequentist theories of hypothesis testing, the repetitions need not be actual, but they must at least be potentially actual.

Neyman identifies the probability of an event A with the ratio of the measure of the set A to that of the set which he calls the fundamental probability set (Neyman 1952, 3), where the measure is just cardinality in the case of a finite fundamental probability set. This at first appears to be a circular definition, in the same way as Laplace's definition of probability in terms of events assumed to be "également possible," but Neyman immediately counters that claim. In two examples involving a standard die, he distinguishes between "the probability of a side of the die having six points on it," a probability which he claims is always $\frac{1}{6}$, and "the probability of getting six points on the die when the die is thrown" (Neyman 1952, 5). He considers the latter statement ambiguous. It might refer to a completed sequence of throws, in which case the probability is simply the ratio of the number of observed sixes to the total number of throws; to a sequence of throws to take place in the future, in which case the probability is unknown until the throws are actually carried out; or to a hypothetical sequence of throws. He considers the latter to be the most fruitful because it leads to the discussion of how to deduce various probabilities from hypothesized values of other probabilities. As an example of this latter case, he considers the hypothetical experiment of tossing a die n times. To say the probability of getting a six on one throw is $\frac{1}{6}$ then means, to Neyman, "that among the the n throws in F_1 [the fundamental probability set] there are exactly $\frac{n}{6}$ with six on the top face of the die" (Neyman 1952, 6). Hence, for Neyman, the probability of an event refers to the frequency with which the event has either occurred in some sequence of repetitions, or to the frequency with which the event will occur in a sequence of repetitions.

3 Frequentist probabilities: Pearson

If probability is nothing more than a name we give to the ratio of the number of occurrences of an event to the total number of repetitions of some experiment, then how do we use it to make statistical inferences when the basic data does not come from a repeatable experiment? For example, what does probability have to say about the Howland will,

given that the data consists of a single signature, whose creation cannot be repeated? Pearson, in a discussion of 2×2 contingency tables, attempts an explanation:

> It seems clear that in certain problems probability theory is of value because of its close relation to frequency of occurrence ... In other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules to guide our decision, following the analysis of an isolated set of numerical data. Why do we do this? What are the springs of decision? Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgement? Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgement which we control at a low figure? (Pearson 1947, 142)

Although he protests that "I should not care to dogmatize," it is clear from these statements and what follows that Pearson considers a probability statement in a singular situation to have two possible meanings: (1) a way of providing a paradigm to guide our thinking, or (2) one event in a sequence of events in which the individual, or, perhaps, the community, will make statistical decisions. The latter view is very reminiscent of the position of C. S. Peirce in his early writings.

4 Frequentist probabilities: Peirce

In his 1878 paper "The Doctrine of Chances," Peirce states:

> According to what has been said, the idea of probability essentially belongs to a kind of inference which is repeated indefinitely. An individual inference must be either true or false, and can show no effect of probability; and, therefore, in reference to a single case considered in itself, probability can have no meaning. Yet if a man had to choose between drawing a card from a pack containing twenty-five red cards and a black one, or from a pack containing twenty-five black cards and a red one, and if the drawing of a red card were destined to transport him to eternal felicity, and that of a black one to consign him to everlasting woe, it would be folly to deny that he ought to prefer the pack containing the larger proportion of red cards, although, from the nature of the risk, it could not be repeated. (Peirce 1992b, 147)

Peirce concludes that "[i]t is not easy to reconcile this with our analysis of the conception of chance" (Peirce 1992b, 147), and in fact he is able to reconcile the two only by enlarging the interests of the individual to include the interests of the community:

> But what, without death, would happen to every man, with death must happen to some man. At the same time, death makes the number of our risks, of our inferences, finite, and so makes their mean result uncertain. The very idea of probability and of reasoning rests on the assumption that this number is indefinitely great. ... It seems to me that we are driven to this, that logicality inexorably requires that our interests shall *not* be limited. They must not stop at our own fate, but must embrace the whole community. ... Logic is rooted in the social principle. (Peirce 1992b, 149)

This narrow view of frequencies formulated by Neyman, Pearson, and Peirce in his early years, involving only sequences which are actually or potentially instantiated, fits well with the prototype of Neyman-Pearson statistical inference, quality control sampling. If θ is some measurable characteristic of an object manufactured in an assembly line setting, one may set up a fixed testing procedure which will result in incorrect inferences (the type I and type II errors, the inference that the process is not within the control bounds when it in fact is, and the inference that the process is within the control bounds when it in fact is not, respectively) with known probabilities. In such a situation the license for, and the consequences of, the inferences are clear: a known percentage of all the inferences made will be in error. But the situation is not so clear when the same ideas are applied to scientific inferences, inferences about the state of nature. In this realm repetitions, even if possible, are not intended. The goal is not to minimize risk through adjustments of testing parameters to control the frequency of type I and type II errors, but to make a statement of fact about the world. Neyman, Pearson, and Peirce are aware of this weakness in their conception of frequencies, and so Neyman prefers to speak of "inductive behavior" instead of "inductive inference" (see Neyman and Pearson 1933, 291, and Neyman 1950, 1) and Pearson speaks of statistical testing procedures as a guide to our decisions, or, with Peirce, treats an individual inference as but one among all the inferences made by a larger community.

In his later writings, Peirce recognized the deficiencies in his original view of frequencies. In a letter to Paul Carus in 1910, Peirce considers what is meant by the statement that the probability of obtaining a three or a six on the roll of a die is $\frac{1}{3}$:

> I mean, of course, to state that the die has a certain habit or disposition of behaviour in its present state of wear. It is a *would be* and does not consist in actualities of single events in any multitude finite or infinite. Nevertheless a habit does consist in what *would* happen under certain circumstances if it should remain unchanged throughout an endless series of actual occurrences. I must therefore define that habit of the die in question which we express by saying that there is a probability of $\frac{1}{3}$ (or odds of 1 to 2) that if it be thrown it will turn up a number divisible by 3 by saying how it *would* behave if, while remaining with its shape, etc. just as they are now, it *were to be* thrown an endless succession of times. (Peirce 1958, 8.225)

Earlier in this same letter, Peirce declares that the "principal positive error" in his early essays "The Fixation of Belief" and "How to Make Our Ideas Clear" is their nominalism. In particular, he claims that:

> I must show that the *will be's*, the actually *is's*, and the *have beens* are not the sum of the reals. They only cover actuality. There are besides *would be's* and *can be's* that are real. (Peirce 1958, 8.216)

Thus Peirce now sees a statement of probability as a statement of the inclination of a mechanism to behave in a certain way under certain circumstances. This inclination, a propensity or habit, is a real property, and would reveal itself in an endless sequence of identical repetitions, although such repetitions are physically impossible. Indeed, if pushed to the limit, no experiment, not even the simple one of throwing a die, is repeatable even

a finite number of times, for, if nothing else, the die thrown on the second toss is not identical to the die thrown on the first toss.

5 Frequentist probabilities: Fisher

R. A. Fisher, the geneticist and founder of the modern frequentist theory of statistical inference, gives an account of probability which has much in common with that of Peirce's later view. Although Fisher talks of probabilities in terms of frequencies, it is clear that the frequencies do not refer to actual sequences or even potentially actual sequences. For example, Fisher prefaces his 1925 paper "Theory of Statistical Inference" with the following comments:

> If, in a Mendelian experiment, we say that the probability is one half that a mouse born of a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not the hypothesis that we wish to consider; moreover, the probability might not always be a rational number. Being infinite the population is clearly hypothetical, for not only must the actual number produced by any parents be finite, but we might wish to consider the possibility that the probability should depend on the age of the parents, or their nutritional conditions. We can, however, imagine an unlimited number of mice produced upon the conditions of our experiment, that is, by similar parents, of the same age, in the same environment. The proportion of white mice in this imaginary population appears to be the actual meaning to be assigned to our statements of probability. (Fisher 1925, 700)

Fisher is thinking along the same lines as Peirce, his "unlimited number of mice produced upon the conditions of our experiment" corresponding to the results from Peirce's die if, "while remaining with its shape, etc. just as they are now, it *were to be* thrown an endless succession of times."

It is not surprising then that Fisher rejects the Neyman-Pearson interpretation of statistical inferences. In discussing a hypothesis concerning the random distribution of stars, Fisher says,

> The force with which such a conclusion is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true. (Fisher 1973, 42)

Using logic similar to that used by Benjamin Peirce in the Howland will case, Fisher argues that, when the supposition of a certain hypothesis implies that an event of small probability has occurred, the rational reaction is to seek an explanation elsewhere. The license for such an inference lies not in the rate of errors which we will commit using such a reasoning procedure, but in the nature of our understanding of what constitutes a small probability.

6 Deborah Mayo and error statistics

Deborah Mayo's error statistics is, in part, a vigorous defense of the frequentist point of view in statistical inference. Her work has not only revealed serious shortcomings in the Bayesian approach, but has provided insightful replies to Bayesian criticisms of frequentist statistics. Her arguments are grounded in a reinterpretation of the Neyman-Pearson school of frequentist statistics, a reinterpretation more heavily influenced by Pearson than by Neyman.

Mayo jettisons the rigid decision-theoretic machinery of the fully developed Neyman-Pearson school (such as fixed critical regions and randomized tests) in favor of a more flexible inference scheme centered on her notion of severity. Considering a hypothesis H and given data $\mathbf{x}$, Mayo says,

> Hypothesis H passes a **severe test** with $\mathbf{x}$ if, (i) $\mathbf{x}$ agrees with H (for a suitable notion of agreement), and (ii) with very high probability, test T would have produced a result that fits H less well than $\mathbf{x}$ does, if H were false or incorrect, or equivalently, (ii') with very low probability, test T would have produced a result that fits H as well as (or better than) $\mathbf{x}$ does, if H were false or incorrect. (Mayo and Spanos 2000, 7)

In the simplest situation, that of testing the value of a single parameter, we may summarize the idea of severity as follows: Let $\Theta \subset \mathbb{R}$ (the set of real numbers) and, for fixed $\theta \in \Theta$, suppose $X_1, X_2, \ldots, X_n$ is a random sample with joint probability measure P_θ. Moreover, consider a function

$$T : \mathbb{R}^n \times \Theta \to \mathbb{R}$$

which, for fixed $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $\theta \in \Theta$, measures the "goodness-of-fit" of $x_1, x_2, \ldots, x_n$ to θ in the sense that we expect $T(x_1, x_2, \ldots, x_n; \theta_0)$ to be near 0 when $\theta = \theta_0$, $T(x_1, x_2, \ldots, x_n; \theta_0) < 0$ when $\theta < \theta_0$, and $T(x_1, x_2, \ldots, x_n; \theta_0) > 0$ when $\theta > \theta_0$. For example, if $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$ for $i = 1, 2, \ldots, n$ and some fixed $0 \leq \theta \leq 1$, then we might define

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

the maximum likelihood estimator for θ, and let

$$T(x_1, x_2, \ldots, x_n; \theta) = \hat{\theta}(x_1, x_2, \ldots, x_n) - \theta.$$

For given observations $x_1, x_2, \ldots, x_n$ of $X_1, X_2, \ldots, X_n$, we say the hypothesis $H : \theta < \theta_0$ passes a severe test if

$$P_{\theta_0}(T(X_1, X_2, \ldots, X_n; \theta_0) \leq T(x_1, x_2, \ldots, x_n; \theta_0))$$

is "small" (that is, we could reject the hypothesis $H' : \theta \geq \theta_0$), and we say a hypothesis $H : \theta > \theta_0$ passes a severe test if

$$P_{\theta_0}(T(X_1, X_2, \ldots, X_n; \theta_0) \geq T(x_1, x_2, \ldots, x_n; \theta_0))$$

is "small" (that is, we could reject the hypothesis $H' : \theta \leq \theta_0$). In both cases, one minus the given probability is the severity of the test.

The line of argument is, for example: If the inference that $\theta < \theta_0$ is in error, then $\theta \geq \theta_0$; but if we find that the probability of observing our data, under the assumption that $\theta \geq \theta_0$ is true, is small, then we should rule out $\theta \geq \theta_0$ as not conforming to observation. But what is it about the small probability that licenses this inference? Mayo is never explicit about her notion of frequency, but, in the context of justifying the use of the apparently behavioristic schema of the Neyman-Pearson approach in scientific inference, she is thinking along the lines of Pearson:

> Why are NP methods so productively used in science despite their "rule of behavior" formulation? How, paraphrasing Neyman, do they manage to correspond precisely to the needs of applied research? There seem to be two main reasons: First, many scientific tasks fit the "assembly line" behavioral-decision model. At many junctures in the links between experimental and data models there is a need for standardized approaches to data analysis that allow us to get going with few assumptions, enable results to be communicated uniformly, and help ensure that we will not too often err in declaring "normal puzzles" solved or not. Second, the behavioral decision approach provides canonical models for nonbehaviorial and non-decision-theoretic uses. The behavioral concepts simply serve to characterize the key features of the NP tools, and *these features* are what enable them to perform the nonbehaviorial tasks to which tests are generally put in science. (Mayo 1996, 374-5)

The problem with this approach is that it ultimately reduces scientific inference to reasoning by analogy. For example, suppose we have 30 independent Bernoulli trials, each with probability of success θ. If we observe 15 successes and wish to test the hypothesis $H_0 : \theta \leq \frac{1}{5}$, we obtain a p-value of at most

$$\begin{aligned} q &= P\left(\hat{\theta}(X_1, X_2, \ldots, X_{30}) \geq \frac{1}{2} \;\middle|\; \theta = \frac{1}{5}\right) \\ &= P\left(X_1 + X_2 + \cdots + X_{30} \geq 15 \;\middle|\; \theta = \frac{1}{5}\right) \\ &= \sum_{i=15}^{30} \left(\frac{1}{5}\right)^i \left(\frac{4}{5}\right)^{30-i} \approx 0.00023. \end{aligned}$$

Mayo would say that the hypothesis $H : \theta > \frac{1}{5}$ has passed a severe test since q is so small, on the order of 2 chances out of 10,000. If our data were coming from an assembly line (say, the number of defective widgets in a batch of 30), then we could reason that if H_0 were indeed true, then we would be committing the error of accepting H when it is false no more than 2.3 times out of every 10,000 such inferences. However, suppose that we are considering the Howland will and that the Peirces had found, not 30 agreements in "downstrokes," but only 15. Now we are in a singular case, and if we reject H_0, we cannot say that we are making an inference which fails to be true in no more 2.3 out of every 10,000 such inferences. In Mayo's approach, we can say only that our inference is

analogous to the inference we could make in the assembly line case. With Pearson (or with Peirce in his early years), we could say that, as part of the community of rational beings, we are making an inference of a style which will fail to be true no more than 2.3 out of every 10,000 such inferences, but Mayo does not want to pursue that line of thought.

7 Small probabilities

The force of the argument in the previous example comes from Fisher's disjunction and Borel's single law of chance: Either H_0 is true and we have observed an event of very small probability, or H_0 is false. Since "[p]henomena with very small probabilities do not occur," it is rational to conclude that H_0 is false. Of course, events of very small probability do occur, and, consequently, this attempt to justify statistical inference has often been dismissed. Fisher himself admits that "the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*" (Fisher 1971, 13-4). Borel is well aware that events of even fantastically small probabilities do occur and discusses in detail the contexts in which events of differing degrees of small probability might be deemed negligible.

One common form of criticism assumes that there is a fixed value (often taken to be 0.05) below which all probabilities are considered to be small for the purposes of statistical inference. But Fisher is very critical of those who would hold to a fixed level of significance for rejecting hypotheses: "... for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1973, 45). Moreover, for Fisher, as opposed to the Neyman-Pearson decision-theoretic school, the decision on what constitutes a small probability in a given situation is not a matter of the perceived risks involved, for scientific research is

> an attempt to improve *public* knowledge undertaken as an act of faith to the effect that, as more becomes known, or more surely known, the intelligent pursuit of a great variety of aims, by a great variety of men, and groups of men, will be facilitated. We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any sort of currency. (Fisher 1955, 77)

Another criticism of Fisher's disjunction is the claim that it does not account for the relative likelihood of rival hypotheses. This appears to be Mayo's primary objection to Fisher's approach to statistical inference. In a discussion of Pearson's response to Fisher's criticisms of the Neyman-Pearson approach, Mayo says that Pearson's "original heresy" (from the Fisherian model of statistical inference) was

> the break Pearson made (from Fisher) in insisting that tests explicitly take into account alternative hypotheses, in contrast with Fisherian significance tests, which did not. With just the single hypothesis (the null hypothesis) of Fisherian tests, the result is either reject or fail to reject according to the significance level of the result. However, just the one hypothesis and its attended significance level left too much latitude in specifying the test, rendering the result too arbitrary. With the inclusion of a set of admissible alternatives to H, it was possible to consider type II as well as

type I errors, and thereby to constrain the appropriate tests. (Mayo 1996, 381)

Richard Royall uses the following example to illustrate the point (Royall 1997, 67): Suppose I am presented with an urn containing 100 balls. Letting ω represent the number of white balls in the urn, I propose to test the hypothesis $H : \omega = 2$ by drawing one ball out of the urn. Since the probability of drawing a white ball is only 0.02 if H is in fact true, the fact of drawing a white ball provides evidence, under Royall's reading of Fisher, that H is false. Royall counters that the 0.02 tells us nothing of the strength of our argument in the absence of any information about possible alternative hypotheses. Mayo would agree; in the absence of any knowledge of alternative hypotheses, we cannot say that the inference "H is false" has passed a severe test (since it is not possible to determine if the data is consistent with the conclusion that H is false, and so we cannot verify part (i) of the definition of a severe test). For example, Royall says, suppose there were only two urns from which to choose, one containing two white balls and the other containing no white balls. In that case, drawing a white ball is in fact conclusive evidence that I am drawing from the urn with two white balls. Mayo would concur, for now H has in fact passed a severe test.

Royall argues that the mistake in Fisher's logic stems from considering the probability of the data given H in isolation from the probability of possible alternative hypotheses. That is, although drawing a white ball is rare under H, it is even rarer (in fact, impossible) under the only viable alternative. Mayo, along with Neyman and Pearson, would say that we must first identify alternative hypotheses which are compatible with the data before we make inferences about H. But there is no mistake in Fisher's logic. We often know little about the possible states of a physical system. Given that this urn has yielded a white ball from a single draw, and in the absence of further knowledge, it is reasonable to infer that more than 2% of the balls in the urn are white, although, with Fisher, we must understand that "conclusions drawn by a scientific worker from a test of significance are *provisional*, and involve an intelligent attempt to *understand* the experimental situation" (Fisher 1955, 74). However, if we indeed have further knowledge, that is, that we are drawing from one of only two urns and that the second has no white balls, then Fisher's disjunction still holds, only now logic forces us to the first disjunct, that is, H is true and a rare event has occurred. The additional knowledge enables deductive logic to intercede and compel us to conclude that a rare event has indeed occurred. Royall's mistake is to assume that a p-value is to be interpreted as some absolute measure of the evidentiary weight of the data. But in fact, the experimenter must evaluate the p-value of her test, in light of Fisher's disjunction and the context of the experiment, anew each time she investigates a hypothesis.

It also follows that if small probabilities do not occur and the hypothesis H is in fact false, we should be able to design an experiment which will reject H almost without fail. As Fisher says, "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (Fisher 1971, 14). For the example above, such an experiment would involve taking a much larger sample than a single ball. Indeed, the experiment of the example is almost worthless for testing H, for, most of the time, the sample will consist

of a single non-white ball, providing no realistic support for or against H. As much as some will insist that we draw definite conclusions from whatever bit of data might come our way, it is clear that this is not always possible. In this example, the only way that any useful conclusion might be drawn from observing a single ball is if that ball is white, in which case there is no need of statistics of any school.

The above is all part of the oft repeated criticism of Fisher that his methodology does not consider the power of a statistical test (that is, the probability of rejecting a specified hypothesis when it is false). But the design of experiments is a central part of his statistical theory, and the design of experiments is, in large, about how to set up tests which will realistically discriminate between various alternatives to the hypothesis being tested. Much of the confusion on this issue stems from Fisher's care in always phrasing his conclusions in terms of the evidence against a proposed hypothesis, and never in terms of accepting one hypothesis over another:

> [I]t is a fallacy ... to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened. (Fisher 1955, 73)

In a sense, Mayo's notion of severity and, similarly, the Neyman-Pearson notion of power, interchange the roles of the inference "H is true" and the inference "H is false." That is, Mayo will say that H has passed a severe test if, in part, we can reject the conclusion that H is false. I see nothing in this that Fisher would find objectionable, nor do I see anything that requires the decision-theory analogies of the Neyman-Pearson philosophy of statistics.

8 Conclusion

In a conversation with Morris DeGroot, David Blackwell relates how a conversation with L. J. Savage led to his conversion to subjective probability and Bayesian statistics. An economist at Rand had approached Blackwell and asked him how one would go about estimating the probability that there would be a major war within the next five years (an estimate the economist needed to help with the allocation of the Air Force research budget). Blackwell glibly responded,

> that question doesn't make sense. Probability applies to a long sequence of repeatable events, and this is clearly a unique situation. The probability is either 0 or 1, but we won't know for five years ... (DeGroot 1986, 44)

The economist responded that several other statisticians had given him the same reply. Blackwell was uncomfortable with his response, "a frivolous, sort of flip, answer" to a serious question and, after a subsequent conversation with Savage, started thinking in terms of the personal probability of deFinetti.

Under a strict, nominalist interpretation of frequencies in a sequence of repeatable experiments, we can never really speak of any probabilities other than 0 or 1. For, just as Heraclitus found he could not step into the same river twice, we can never roll the same die twice. If we do speak of relative frequencies in the sense of Neyman, then the ground for our inferences outside of the decision-theory framework seems weak, being little more

than a general guide to what would be the case if we were in the realm of "assembly line" statistics. No wonder many probabilists and statisticians have, like Blackwell, turned to Bayesian methodology to find a foundation for their work.

Mayo has made a major contribution to the understanding of frequentist statistics with her deft descriptions and explanations of how positive knowledge may be built up from what at first appears to be an essentially negative idea, that of the rejection of a hypothesis. Fisher called attention (in his usual forceful way) to the problem of reaching positive conclusions from negative results when he called into question the conclusions of those investigating causal links between smoking and lung cancer. In his letters to the *British Medical Journal* (Fisher 1957a and 1957b), Fisher emphasizes that the rejection of the hypothesis of no association between rates of smoking and rates of lung cancer does not imply that smoking is the cause of lung cancer; rather, it shows what other hypotheses (for example, that perhaps there is a genetic component to both) must now be investigated. Mayo has provided a conceptual framework, along with examples from the work of scientists, on exactly how, contrary to the claims of the Bayesian school, one might use frequentist statistics to carry out such project. Yet to get such a project off the ground, one must understand how it is that small probabilities provide a license for statistical inference. I suggest that the Neyman-Pearson framework is not strong enough to carry the burden, and that one must develop a more robust concept of frequencies, such as that described in Peirce's later work, to support the work of frequentist statistics.

References

Borel, É. 1962. *Probabilities and Life.* Translated by M. Baudin. New York: Dover.

Fisher, R. A. 1925. Theory of statistical inference. *Proceedings of the Cambridge Philosophical Society* 22:700-725.

Fisher, R. A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)* 17:69-78.

Fisher, R. A. 1957a. Letter to the editor of the *British Medical Journal* 2:43.

Fisher, R. A. 1957b. Letter to the editor of the *British Medical Journal* 2:297-8.

Fisher, R. A. 1971. *The design of experiments.* 8th ed. New York: Hafner.

Fisher, R. A. 1973. *Statistical methods and scientific inference.* 3rd ed. New York: Hafner.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge.* Chicago and London: University of Chicago Press.

Mayo, D. and A. Spanos. 2000. A post-data interpretation of Neyman-Pearson methods based on a conception of severe testing. *Measurements in Physics and Economics Paper Series*, Centre for Philosophy of Natural & Social Science, London School of Economics.

Meier, P. and Zabell, S. 1980. Benjamin Peirce and the Howland will. *Journal of the American Statistical Association.* 75:497-506.

Neyman, J. 1950. *First course in probability and statistics.* New York: Henry Holt.

Neyman, J. 1952. *Lectures and conferences on mathematical statistics and probability.* 2d ed. Washington, D.C.: U.S. Department of Agriculture.

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

Pearson, E. S. 1947. The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika* 34:139-67.

Peirce, C. S. 1958. *Collected Papers.* Vols. 7-8. Edited by A. Burks. Cambridge: Harvard University Press.

Peirce, C. S. 1992a. *The Essential Peirce.* Vol. 1. Edited by N. Houser and C. Kloesel. Bloomington and Indianapolis: Indiana University Press.

Peirce, C. S. 1992b. The doctrine of chances. In *The Essential Peirce*, 142-54 (Bloomington and Indianapolis: Indiana University Press). First published in *Popular Science Monthly* 12 (1878):604-15.

Royall, R. 1997. *Statistical evidence: a likelihood paradigm.* London: Chapman and Hall.