CHAPTER TWELVE

Error Statistics and Peircean Error Correction

> Induction (at least, in its typical forms) contributes nothing to our knowledge except to tell us approximately how often, in the course of such experience as our experiments go towards constituting, a given sort of event occurs. It thus simply evaluates an objective probability. Its validity does not depend upon the uniformity of nature, or anything of that kind. The uniformity of nature may tend to give the probability evaluated an extremely great or small value; but even if nature were not uniform, induction would be sure to find it out, *so long as inductive reasoning could be performed at all.* . . .
>
> But all the above is at variance with the doctrines of almost all logicians. . . . They commonly teach that the inductive conclusion approximates to the truth because of the uniformity of nature.
>
> —C. S. Peirce, *Collected Papers*, vol. 2, par. 775

I OPENED CHAPTER 1 with a quote from Popper: "The essays and lectures of which this book is composed are variations upon one very simple theme—the thesis that we can *learn from our mistakes.*" The theme of learning from error has a central place in the experimental program based on error statistics that I have been sketching. Nevertheless, from all we have said, is it apparent that Popper's account falls far short of showing how reliable knowledge is obtained from experiment or how that knowledge grows. The present account does not find its home in a Popperian framework. It is quite at home, however, within the experimental framework of another philosopher who also developed an account wherein scientific inference is based on learning from error and error correction, namely, C. S. Peirce. Nevertheless, the Peircean error correction thesis has been soundly criticized and found wanting. Indeed, as Nicholas Rescher (1978) remarks, "No part of Peirce's philosophy of science has been more severely criticized, even by his most sympathetic commentators, than this attempted validation of inductive methodology on the basis of its purported self-correctiveness" (p. 2).

Despite the hard times on which Peirce's validation of inductive methodology has fallen, I propose to revive the Peircean self-correcting doctrine. I want to do so not only to defend a great philosopher from whom I have gained many insights, but also, more selfishly, because by developing my view of Peirce's error-correcting justification of induction I will at the same time be developing the justification I need for error statistical methods in science. The justification for these methods lies in their ability to control error probabilities, hence sustain learning from error, hence provide for the growth of experimental knowledge. It seems to me that this is also the essence of Peirce's self-correcting rationale of inductive methods—when that thesis is properly understood. While on the one hand Neyman-Pearson and other contemporary methods increase the mathematical rigor and generality of Peirce's assertions about self-correcting methods, on the other, Peirce provides something the formal statistical tools lack (and Pearson only hinted at): an account of inductive inference and a philosophy of experiment ready-made for just such tools.

12.1 PEIRCEAN INDUCTION AND NEYMAN-PEARSON STATISTICS

Peirce's philosophy of experimental testing shares a number of key features with the Neyman and Pearson theory. For both, statistical methods provide not means for assigning degrees of probability, evidential support, or confirmation to hypotheses, but procedures for testing (and estimation) whose rationale is their predesignated high frequencies of leading to correct results in some hypothetical long run. The key similarities between Peirce and the methods later developed by Neyman and Pearson were first unearthed by Isaac Levi (1980b).[1]

> Peirce's inductions are inferences according to rules specified in advance of drawing the inferences where the properties of the rules which make the inferences good ones concern the probability of success in using the rules. These are features of the rules which followers of the Neyman-Pearson approach to confidence interval estimation would insist upon. (P. 138)

In describing his theory of inference, Peirce could be describing that of the error statistician:

1. Levi also relates Peirce's work to that of R. B. Braithwaite, who developed a theory of chance based on rules of testing akin to Neyman and Pearson tests. I regret being unable to discuss Braithwaite's work here, but good discussions exist in Hacking 1965 and Mellor 1980.

> The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently *that conclusion* would be found true. It does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different—in meaning, numerical value, and form—from that of those who would apply to ampliative inference the doctrine of inverse chances. (Peirce 2.748)[2]

One finds specific examples in Peirce that anticipate Neyman-Pearson hypothesis tests and confidence interval methods. A study of the statistical mathematics found in Peirce is of interest in its own right, but that is not my purpose here. My purpose is to explore how in Peirce's philosophy of experiment the formal NP tools become tools for scientific induction. (Neyman, remember, had denied them that function, and Pearson never fully worked out his "evidential" interpretation of NP tools.)

The place to begin is with the contrast that Peirce is at pains to draw between his view of induction and the more popular inductive accounts of his day. The most popular accounts of the time, Peirce tells us, are those of the "conceptualists"—the Bayesian theorists of Peirce's day—and the followers of Mill—essentially those who viewed induction as the straight rule, coupled with a premise as to the uniformity of nature. With admirable clarity Peirce compares these opposing views and forcefully argues against the popular ones. The main contrasts show up in the form of conclusion or inference (severe tests); the type of relevant information (preliminary planning, predesignation, random sampling); and the nature of its justification (self-correcting, growth of experimental knowledge). In each, Peirce takes the position of our error statistician.

What We Really Want to Know . . . Error Probabilities

The key disagreement was and is over the function of probability in statistical inference in science: whether probability provides a measure of evidential strength in a hypothesis, or whether it should be used only to characterize error probabilities of test procedures. Although the terminology has changed, it is clear that Peirce adopts the

2. All Peirce references are to C. S. Peirce, *Collected Papers*. References are cited by volume and paragraph number. For example, Peirce 2.777 refers to volume 2, paragraph 777.

second use of probability as the appropriate one for experimental inference. The supposition that the probability of the conclusion is needed, Peirce recognizes, stems from a faulty analogy with deductive inference. Since deductive inference tells us that if such and such premises are true, then a given conclusion is true, it might be thought that inductive inference should tell us that if such and such premises are true, then a given conclusion is probable. Peirce denies this:

> In the case of analytic inference we know the probability of our conclusion (if the premisses are true), but in the case of synthetic inferences we only know the degree of trustworthiness of our proceeding. (2.693)

Those who modeled induction on the analogy with deduction were to Peirce what Bayesians or other E-R theorists are to NP or error statisticians of today, and the key themes of Peirce's work on induction mirror the key issues that divide E-R theorists from error statisticians: the importance of error probabilities, the rejection of prior probabilities, and the centrality of the mode of data and hypothesis generation to the analysis of test results.[3]

In Peirce's testing model, like that of Neyman and Pearson, the experimental conclusion concerns a hypothesis that either is or is not true about this one universe, and so the only probability that a frequentist could assign it is a trivial one, 1 or 0.[4] Assigning a probability to a particular conclusion, for Peirce (recall chapter 3), makes sense only "if universes were as plenty as blackberries" (2.684). If people had only been careful to keep to the relative frequency notion of probability, Peirce scolds, the mistake in analogizing induction to deduction would have been apparent. To view statistical inference as a matter of assigning a probability to a conclusion (an a posteriori probability), is, for a frequentist like Peirce, tantamount to seeing the problem as follows:

> Given a synthetic conclusion; required to know out of all possible states of things how many will accord, to any assigned extent with this conclusion. (2.685)

3. An interesting article of Hacking's is relevant in this connection. Hacking (1980), discussing Peirce and Braithwaite, admits to having promoted the rejection of Neyman-Pearson statistics on the grounds that, failing to provide an E-R account, it could not be seen as an account of inductive inference. With this article Hacking announces that he has changed his mind on this point, and allows that an error-statistical account does provide us with an account of inductive inference.

4. Peirce also has an account of probabilistic inference where that is appropriate. But this is not induction.

Here "all possible states of things" refers to all possible universes or possible ways in which this universe could be. This Peirce sees as "an absurd attempt to reduce synthetic to analytic reason, and that no definite solution is possible" (ibid.). Moreover, it "implies that we are interested in all possible worlds, and not merely the one in which we find ourselves placed" (2.686).

What we really want to know, according to Peirce, is this:

> Given a certain state of things, required to know what proportion of all synthetic inferences relating to it will be true within a given degree of approximation. (2.686)

In more modern terminology, what we want to know are the error probabilities associated with particular methods of reaching conclusions about this world. Peirce continues:

> Now, there is no difficulty about this problem (except for its mathematical complication); it has been much studied, and the answer is perfectly well known. And is not this, after all, what we want to know much rather than the other? (Peirce 2.686)

Peirce goes on to illustrate how, even in his day, "the answer is perfectly well known." His numerical illustration is important for us, and I will return to it in a later section.

Denying That Belief Has Anything to Do with It

A further reason that error statistical methods are congenial to Peirce's picture is that their error probability characteristics do not depend on subjective probabilities. Peirce held that

> subjective probabilities, or likelihoods, . . . express nothing but the conformity of a new suggestion to our prepossessions; and these are the source of most of the errors into which man falls, and of all the worst of them. (2.777)

An important part of Peirce's rejection of subjective probabilities is his insistence upon a distinction between the proper procedure for a scientific investigation and that for an individual seeking a practical basis for action. Peircean pragmatism (or pragmaticism) is not at all to be identified with practicalism!

While allowing that subjective beliefs and personal opinions may have to be appealed to in the area of practical conduct, where expediency is the rule, and where personal beliefs matter, Peirce thinks that "the word belief is out of place in the vocabulary of science" (7.185), except when considering actions based on science. In a scientific inves-

tigation Peirce declares, "I would endeavor to get to the bottom of the question, without reference to my preconceived notions" (7.177). The aim of science is to predict the future "or the means of conditionally predicting what would be perceived were anybody to be in a situation to perceive it" (7.186). Its aim is to predict what would be expected to occur with various relative frequencies were specified experiments carried out—in short, to obtain what I call experimental knowledge. Indeed, for Peirce "the essential character of induction is that it infers a *would-be* from actual singulars" (8.236).

Interestingly, Peirce's central arguments against the use of subjective probabilities have a naturalistic flavor: inferences based on subjective probabilities, he finds, make a poor showing when they themselves are put to the test of experiment. As an example Peirce considers their track record in archaeology. Finding the conclusions sanctioned by the practitioners of the subjective method "to be more or less fundamentally wrong in nearly every case," he declares the method "condemned by those tests" (7.182).[5]

This much has so far been brought to light about Peirce's theory of induction:

> In the case of analytic inference we know the probability of our conclusion (if the premisses are true), but in the case of synthetic inferences we only know the degree of trustworthiness of our proceeding. As all knowledge comes from synthetic inference, we must equally infer that all human certainty consists merely in our knowing that the processes by which our knowledge has been derived are such as must generally have led to true conclusions. (2.693)

12.2 PEIRCEAN INDUCTION AS SEVERE TESTING

The scientific procedure in whose trustworthiness we are interested is, for Peirce, induction, but induction is to be understood as testing. The trustworthiness of inductive procedures, I maintain, is a matter of the test's severity, as measured formally (quantitative induction) or informally (qualitative induction). What is my evidence for this reading of Peirce?

First, there is the evidence that Peirce regards induction as severe testing. Induction, Peirce tells us, begins with a question or theory:

> The next business in order is to commence deducing from it whatever experiential predictions are extremest and most unlikely . . . in order

5. Peirce elsewhere gives astute criticisms of the use of the principle of indifference in assigning equal subjective probabilities, which I will not discuss.

> to subject them to the *test of experiment.* (Peirce 7.182; emphasis added)
>
> The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true. (7.231)
>
> When the hypothesis has sustained a testing as severe as the present state of our knowledge . . . renders imperative, it will be admitted provisionally . . . subject of course to reconsideration. (Ibid.)[6]

Further passages to the same effect could easily be multiplied.

While these and other passages are redolent of Popper, Peirce differs from Popper in crucial ways—the same ways in which my own account differs. Peirce, unlike Popper, is primarily interested in the positive pieces of information provided by tests, that is, with the hypotheses, modified or not, that manage to pass severe tests. Indeed, Peirce often suggests that he equates the proper inductive part of a test of experiment with the inference that is reached when a hypothesis passes several stringent tests:

> When, however, we find that prediction after prediction, notwithstanding a preference for putting the most unlikely ones to the test, is verified by experiment, whether without modification or with a merely quantitative modification, we begin to accord to the hypothesis a standing among scientific results. This sort of inference it is, from experiments testing predictions based on a hypothesis, that is alone properly entitled to be called *induction.* (7.206)

A Peircean inductive inference, then, accords well with the thesis I have advocated: an inductive inference—that which is warranted to infer—is what passes a severe test. Whereas one could say nothing about the reliability of a Popperian corroboration procedure—the very reason I denied Popper supplies a genuine account of learning from error—the centerpiece of Peirce's experimental philosophy is his argument for the trustworthiness of proper inductive test procedures.

It is impossible to understand Peirce's argument, however, without understanding Peirce's doctrine of induction as self-correcting or as error-correcting. This requires us to open a door that many Peirce scholars already regard as closed, or at least to open it just far enough

6. Here Peirce is talking about historical hypotheses. See also (Peirce 1958, vol. 7, p. 89).

to give a different reading of the error-correcting doctrine. This new reading of the error-correcting doctrine, I believe, shows how the criticisms of the usual reading are avoided.

12.3 REVISITING PEIRCE'S ERROR-CORRECTING DOCTRINE

According to Peirce:

> The validity of induction is entirely different [from deduction]. . . . In the majority of cases, the method would lead to *some* conclusion that was true, and that in the individual case in hand, if there is any error in the conclusion, that error will get corrected by simply persisting in the employment of the same method. (2.781)

Throughout Peirce's work, a multitude of such passages can be found, each offering different clues to and different facets of his self-correcting doctrine. (Several are noted in Laudan 1981a.) What must be kept in mind, and often is not, is that induction for Peirce is testing, and testing of a certain sort (severe or reliable); it is testing (done severely) that he is claiming is self-corrective, and not other methods that philosophers often regard as inductive:

> Induction is the experimental testing of a theory. The justification of it is that, although the conclusion at any stage of the investigation may be more or less erroneous, yet the further application of the same method must correct the error. The only thing that induction accomplishes is to determine the value of a quantity. It sets out with a theory and it measures the degree of concordance of that theory with fact. (5.145)

Can Peirce sustain his self-correcting thesis as a way of giving a rationale for scientific induction? Critics and followers alike say no. The literature on this issue is too large to consider here, but fortunately, Rescher's excellent discussion (1978) lets me zero in on the key criticism, as waged by Larry Laudan and others. I will follow Laudan's abbreviation of the self-correcting thesis: (SCT). Let me begin with a brief summary of the main criticism and how I propose to deal with it.

The main criticism of the SCT is this: whereas Peirce claims to have substantiated the SCT for induction generally, he has at most done so regarding a certain species of induction, namely, quantitative or statistical induction. This criticism rests on two assumptions: the first concerns the nature of inductive testing for Peirce, of both the "quantitative" and "qualitative" varieties; the second concerns the question of what substantiating the SCT requires. As to the first, Peirce's critics

typically construe quantitative induction as classic enumerative induction or "the straight rule" (i.e., inference about a population proportion from a sample proportion). By qualitative induction, critics understand Peirce to mean hypothetico-deductive inference (Laudan 1981b, 238). But from all we have already seen, it is clear that neither of these modes of inference suffices for a test procedure that is trustworthy or, in my terms, reliable or severe. So the first thing we need to do is to revise the standard interpretation of Peirce's two types of induction.

I will be arguing that what distinguishes Peircean quantitative from qualitative induction is not that the former is the straight rule while the latter is a hypothetico-deductive inference. *Both* types of inference, in so far as they qualify as Peircean inductions, are inferences based on tests with various degrees of severity. What distinguishes them is the extent to which their severity or reliability can be quantitatively or only qualitatively determined. If the severity is quantitatively specified, as in the case of the statistical significance test, then the inference is a quantitative induction. If severity is only qualitatively assessed, as for example in one of the informal arguments from coincidence we have considered, then it counts as a qualitative induction. The difference is a matter of degree.

Turning to the second issue, critics are fairly clear on what they suppose is required for an inductive method to be self-correcting: (*a*) it must be capable of (eventually) rejecting false hypotheses, and (*b*) it must provide a method of replacing rejected hypotheses with a better (truer) one (Laudan 1981b, 229).[7] Their criticism of Peirce's SCT, in the light of *their* understanding of Peircean quantitative and qualitative induction, is this: although quantitative induction pretty well satisfies both (*a*) and (*b*), qualitative induction only satisfies (*a*). Laudan puts it plainly:

> Such qualitative inductions clearly satisfy the first condition for an SCM [self-correcting method], insofar as persistent application of the method of hypothesis will eventually reveal that a false hypothesis is, in fact, false. But the method . . . provides no machinery whatever for satisfying the second necessary condition. . . . Given that an hypothesis has been refuted, qualitative induction specifies no technique for generating an alternative which is (or is likely to be) closer to the truth than the refuted hypothesis. (Laudan 1981b, 238–39)

Ilkka Niiniluoto (1984), in like fashion, assimilates Peircean self-correcting to a view of scientific progress as replacing earlier theories

7. Laudan regards statement *b* as the strong thesis of self-correcting. A weaker thesis would replace *b* with *b'*: science has techniques for unambiguously determin-

with those closer to the truth, leading him also to criticize Peirce for not having told us how induction affords such progress.[8] The technique for discovering a better alternative, moreover, is supposed to be mechanical or routine, and, not surprisingly, critics find that Peirce has not provided such a routine. Rescher (1978) objects to this requirement and defends the Peircean SCT as claiming only that it is the conglomeration of scientific methods that serves to find better alternatives. Rescher is right to object, but I think we can show that Peirce is saying something more specific about the error correcting role of inductive methodology in science. Inductive methods, properly construed, are very good at uncovering mistakes and this is what allows them to carry out effective tests to begin with. Their effectiveness consists in this: when they regard a hypothesis as having passed a test sufficiently well, that constitutes good grounds for that hypothesis.

These points lead to a reworking of the critics' two assumptions about the SCT. From the severity requirement we actually get a strengthened form of condition *a:* the inductive test procedure must have a *high,* not merely some, probability of rejecting false hypotheses. But we must not overlook, as critics seem to, the emphasis Peirce places on what is learned when such severe tests do not reject but instead pass their hypotheses. For Peirce, as I read him, the SCT is called upon to justify the *acceptance* of a hypothesis that has passed a severe test (e.g., 2.775). Inductive inference is the inference that is warranted when predictions hold up to severe testing. So the proper requirement for the SCT is not condition *b,* as the critics state it, but rather a condition that takes more literally what error-correction means.

A reworked condition *b* would have two parts: First, the method should be sufficiently good at detecting errors such that when no error is detected, when, try as we might, the effect will not go away, experimental knowledge (as we have defined it) is gained. Second, the method should be able to detect its own errors in the sense of checking its own assumptions or its "own premises" as Peirce puts it (i.e., assumptions of experimental tests and data), and it should be able to correct violations or "subtract them out" in the analysis. To show that scientific induction is self-correcting comes down to showing that severe testing methods exist and that they enable the growth of experi-

---

ing whether an alternative T' is closer to the truth than a refuted T. I reject both of these.

8. For some commentators, for example, Lenz (1964), what Peirce says about qualitative induction is so unclear that they restrict themselves to quantitative.

mental knowledge. The progress is not of the theory-dominated but of the experimentalist variety. My task now is to justify these claims.

The Path from Qualitative to Quantitative Induction

First I will argue my thesis about Peirce's notions of quantitative and qualitative induction. A major problem in understanding the self-correcting doctrine is that induction, for Peirce, takes several different forms corresponding to different types of test procedures. These different test procedures, in turn, are associated with different types of assessments of trustworthiness (i.e., of error probabilities) as well as different types of error-correcting tasks. What is more, throughout Peirce's work one finds a variety of attempts to delineate types of induction, and one may wonder which delineation to work with. In fact, Peirce does not think there is anything hard and fast about his classification attempts. Although critics are right to notice some shifts in Peirce's view on induction, his different schemes for classifying types of induction are almost entirely due to his directing himself to different kinds of experimental tests in different essays. Most important, if one looks at the big picture, a fairly clear-cut image emerges. Induction is testing, some qualitative, some quantitative or statistical—all agree on this. Where my reading of Peircean induction is new is that I view Peirce's delineation into quantitative and qualitative induction as a matter of classifying tests according to whether their trustworthiness (or severity) is quantitatively or only qualitatively ascertained. (This is the same construal, recall chapter 2, that I suggested for Kuhn's use of quantitative inference in normal science.)

In this reading of Peirce, the difference between qualitative and quantitative induction is really a matter of degree, and the degree is a function of how well developed its associated measures of trustworthiness are—in particular severity. This reading not only neatly organizes the long stories Peirce tells in classifying and subclassifying types of induction, it explains the way in which Peirce further subdivides types of inductions by their "strength" within a given classification.

First-order, rudimentary or crude induction. Take Peirce's delineation of types of induction in discussing scientific method. Here Peirce divides nonstatistical or qualitative induction into first and second orders. The first order is the lowest, most *rudimentary induction,* the so-called "pooh-pooh" argument. It is essentially an argument from ignorance: Lacking evidence for the falsity of H, provisionally adopt H—where H is some general claim or regularity. While Peirce holds this type of "crude induction" to be uneliminable in ordinary life, it has little place in scientific inquiry. (It corrects itself—but with a bang!) It is only in

this very weakest sort of induction, crude induction, that one is limited to saying that a hypothesis would eventually be falsified if false. Crude induction, Peirce says, is "as weak an inference as any that I would not positively condemn" (8.237), and does not even make it into science. Once positive information is available, this most rudimentary induction is to go by the board. Hence, following Peirce, rudimentary induction is not to be included as scientific induction. It is, however, worthwhile to recognize why not: without some reason to think that evidence of *H*'s falsity would probably have been detected, failure to detect it is poor evidence for *H*. It is a highly unreliable error probe.

Second order (qualitative) induction. It is only with what Peirce calls the "Second Order" of induction that we arrive at a genuine test, and thereby scientific induction. Within second-order inductions, a stronger and a weaker type exist, and they correspond neatly to viewing the strength of a testing procedure as reflecting severity.

> The weaker of these is where the predictions that are fulfilled are merely of the continuance in future experience of the same phenomena which originally suggested and recommended the hypothesis. (7.116)

> The other variety of the argument from the fulfillment of predictions is where [they] . . . lead to new predictions being based upon the hypothesis of an entirely different kind from those originally contemplated and these new predictions are equally found to be verified. (7.117)

The weaker type, to put it in our terminology, occurs where violating use-novelty destroys the severity requirement. The stronger type is stronger because it generally yields a higher severity test. Peirce's divisions by strength within second-order inductions are also a function of severity, but the assessment of severity is qualitative, for example, very strong, weak, very weak.

> The strength of any argument of the Second Order depends upon how much the confirmation of the prediction runs counter to what our expectation would have been without the hypothesis. It is entirely a question of how much; and yet there is no measurable quantity. *For when such measure is possible the argument . . . becomes an induction of the Third Order* [statistical induction]. (7.115; emphasis added)

It is upon these and numerous like passages that I base my reading of Peirce. Furthermore, a qualitative induction, Peirce is quite clear, *becomes* a quantitative induction when the severity is quantitatively de-

termined, when, as we might say, an objective error probability can be given.

Third order, statistical (quantitative) induction. This takes us to the third-order, statistical or quantitative induction. We enter the third order of induction when, to paraphrase Peirce, it is possible to quantify "how much" the prediction runs counter to what our expectation would have been without the hypothesis. Quantifying how much, as I hope is already clear from earlier discussions, permits quantifying trustworthiness by quantifying error probabilities.

To remind us, consider how a significance level measures how much a prediction runs counter to what is expected "without the hypothesis," where this refers to a simple null hypothesis H_0. As always, we see the following inversion: the lower the significance level, the more the prediction runs counter to the null hypothesis. Hence, the lower the significance level required before rejecting H_0 and accepting the nonnull hypothesis—call it H—the more improbable such an acceptance of H is, when in fact H_0 is true. And the more probable such an erroneous acceptance of H is, the higher the severity is of a result taken to pass H. This just rehearses what we already know. Other associated measures of "how much" are given by standard errors and probable errors, error probabilities all.

Notice that it is in order for the inductive *acceptance* of a hypothesis H to have strength that we meet the requirement that there be a high probability of rejecting hypothesis H, were H false. That is, Peircean induction refers to the positive inference—to what can be said to have passed a severe test:

> When we adopt a certain hypothesis, it is not alone because it will explain the observed facts, but also because the contrary hypothesis would probably lead to results contrary to those observed. So, when we make an induction, it is drawn not only because it explains the distribution of characters in the sample, but also because a different rule *would probably have led to the sample being other than it is.* (Peirce 2.628; emphasis added)[9]

This concern with the probability that the sample *would have been other than it is* in reasoning from the actual sample obtained puts Peirce squarely in the error statistics camp. And because one need not be able to point to some precise probability, the same self-correcting rationale is open to quantitative and qualitative tests.

As further evidence that Peirce understood the strength of an in-

9. By a "rule" here Peirce means a hypothesis such as most *As* are *Bs*.

duction in this way, Peirce often links the strength of induction—even in qualitative cases—to achieving what we would term a low standard deviation or low standard error (and, correspondingly, to a high severity):

> The results of non-quantitative researches also have an inexactitude or indeterminacy which is analogous to the probable error of quantitative determinations. To this inexactitude, although it be not numerically expressed, the term "probable error" may be conveniently extended. (7.139)

(A probable error is approximately .7 of a standard deviation.) It is convenient to extend the notion of a probable error for the same reason we found it convenient to use the term "severity" both when there was a numerical error probability that could be assigned to a test and when we could only argue that there was clearly a very high or a very low chance of error. They serve analogous roles in argument and, accordingly, qualitative and quantitative inductions are improved upon in analogous ways. The factors Peirce takes to increase or diminish the strength of procedure further illuminate the correspondence between the "strength of a proceeding" and our severity concept. Peirce explains that arguments are strengthened when certain invariabilities exist—in effect, factors that by diminishing a standard deviation would increase the chance of rejecting a false hypothesis. (See, for example, 7.125.)

Scientific induction, for Peirce, is inferring or accepting hypotheses that pass severe or trustworthy tests. The move from qualitative to quantitative induction is achieved by the acquisition of quantitative assessments of severity.

The SCT and Quantitative Induction

In inductive inference, unlike deductive or analytic inference, Peirce declared, what we really want to know is the trustworthiness of the proceeding or, in more modern terms, the error probabilities. Moreover, in the case of quantitative induction, Peirce said, the answer to the question we really want to know "is perfectly well known" (2.686). Let us now pick up Peirce where we left him in section 12.1.

In Peirce's example, the inductive inference estimates a Binomial parameter p on the basis of the number n of white balls observed in a sample of s balls. Referring to the difference between the observed proportion $\frac{n}{s}$ and the true proportion p as "the error", Peirce (2.686) explains that

> it is found that, if the true proportion of white balls is p, and s balls are drawn, then the error of the proportion obtained by the induction will be—
>
> | | |
> |---|---|
> | half the time within | $0.477e$ |
> | 9 times out of 10 within | $1.163e$ |
> | 99 times out of 100 within | $1.821e$ |
> | 999 times out of 1,000 within | $2.328e$ |
> | 9,999 times out of 10,000 within | $2.751e$ |
> | 9,999,999,999 times out of 10,000,000,000 within | $4.77e$, |

where I have substituted e for the square root of $[\frac{2\,p(1-p)}{s}]$.

Whereas simple enumerative induction, which is how critics construe quantitative induction, would merely estimate p to be the sample proportion, Peirce insists on a second step: attaching an error to this estimate. It may be in terms of the "probable error" concept of Peirce's day, or the more modern standard error, or, as in contemporary polls, a margin of error.

The SCT and Confidence Interval Estimation Procedures. The data from the above chart may be used to form confidence interval estimates (discussed in chapters 8 and 10). The inductive conclusion in the case of the interval estimation asserts that the observed proportion $\frac{n}{s}$ is within a certain distance from the true value of p, and attaches to that estimate a statement of the overall reliability of that method (as given by the confidence level). An example of such an estimate would assert that the observed proportion is within $1.821e$ of the true value p. Although the method does not assign a probability to this particular estimate being true, that probability being seen as either 0 or 1, the method can say that the inference comes from a procedure with .99 probability of covering the true value of p. The inferred estimate, that parameter p is within the interval formed, passes a severe test. So Neyman and Pearson confidence interval estimation satisfies Peirce's model of induction. In contrast to induction by simple enumeration (or the straight rule) as Peirce never tires of reminding us, the induction he espouses depends entirely on "the manner in which the instances have been collected" (2.765). But critics seem to overlook this contrast.

Isaac Levi puts his finger on how self-correcting works in the case of statistical estimation:

> Peirce is not claiming that induction is self correcting in the sense that following an inductive rule will, in the messianic long run, reveal the

> true value of p. His thesis can be put this way: Either the conclusion reached *via* an inductive rule is correct or, if wrong, the revised estimate emerging from a new attempt at estimation based on a different sample will with probability at least equal to k be correct. (Levi 1980b, 138)

Suppose the induction is to a confidence interval with level k. The idea is that if one continues to sample (with replacement) and form a confidence interval with confidence level k, "he would be right with a relative frequency which would converge on k in the long run" (p. 136). The same type of argument is available for other cases of statistical estimation.

In the case of hypothesis testing, a claim parallel to Levi's on estimation can be made: if a particular conclusion is wrong, subsequent severe (or highly powerful) tests will with high probability detect this. For example, in a good test hypothesis H_0 is rejected by results improbably far from what is expected were H_0 true. Then, if we are wrong to reject H_0 (and H_0 is actually true), we would find we were rarely able to get so statistically significant a result to recur, and in this way we would discover our original error. If, on the other hand, we find that it is easy to keep getting results statistically significantly far from H_0, then we have grounds for saying that a real departure from H_0 exists. To say we have experimental knowledge of a real or systematic departure from H_0 is to say that H_0 would be rejected about as often as expected if such a departure exists. (The expectation comes from the laws of large numbers, discussed in chapter 5.)

Peirce discussed the Gaussian or Normal case as well as the Binomial. Modern statistical theory greatly extends the cases for which "the answer is well-known," but the rationale for the inferences it licenses is essentially the one that Peirce had already articulated:

> While the induction is probable in this sense, that though it may happen to give a false conclusion, yet in most cases in which the same precept of inference was followed, a different and approximately true inference (with the right value of p) would be drawn. (2.703)

More needs to be said about how formal statistical arguments supply tools for substantive error-correcting and learning. Here is where Peirce's stress on the *intended use* of these methods comes in.

Quantitative Induction and Canonical Models of Error. In developing the error statistical approach to testing, I have urged that the role of quantitative models such as the Binomial goes far beyond the case in which the primary aim is to infer the proportion of *B*s in a population of *A*s—

even though Binomial inference is formally couched in those terms. This formal statistical case serves largely as a canonical model for imaginatively asking questions about errors, about experimental assumptions, about the reality of a given effect, about quantities in laws and theories, about causes. I find evidence of this idea in Peirce, if not explicitly, then by considering how he applies statistical models in his examples.

Immediately after listing the different error ranges for the Binomial case above, Peirce remarks that "the use of this may be illustrated by an example" (2.687) that sounds very much like running an NP statistical significance test. As in many other cases, Peirce applies it to testing if a difference is real or systematic as opposed to due to chance. Peirce reports that an observed proportion of white males under one year (according to the census of 1870) is .5, while that of nonwhite children is .498, the difference being about .01. Peirce asks, "Can this be attributed to chance," or is it systematic? The largeness of the observed difference excludes it even from the largest interval formed; it falls beyond 4.77*e*, "and such a result would happen, according to our table, only once out of 10,000,000,000 censuses, in the long run" (2.687). In short, the observed difference is indicative of a real rather than a chance difference. Were it due to chance it would, with high probability, have been included in the interval. The procedure was a reliable probe of the error of ruling out chance; so we can argue that this error is absent.

In the above illustration, the hypothesis concerned the ordinary kind of Binomial population. But Peirce extends this analysis to assess hypotheses that are not themselves statistical, but where *introducing* statistical considerations enables a question of interest to be modeled as inquiring about a Binomial parameter p. In particular, a question that can often be framed by means of parameter p is to let p be the probability with which a given agreement or fit between the experiment and a given hypothesis H would occur. Such a question may be probed statistically, even where hypothesis H itself is not statistical. Let us see how the SCT enters:

> It is true that the observed conformity of the facts to the requirements of the hypothesis may have been fortuitous. But if so, we have only to persist in this same method of research and we shall gradually be brought around to the truth. (Peirce 7.115)

But the correction is not a matter of getting estimates closer to p. It is a matter of finding out whether the agreement is fortuitous; whether it is generated *about as often as would be expected* were the agreement of the chance variety. The measure of severity reflects how fast the correction is likely to be.

The SCT and the Importance of Hypothesis and Data Generation

This error-correcting capacity, Peirce stresses, depends upon the *predesignation* of the Binomial property p (or, at least on an argument that its violation does not vitiate the induction).[10] I limit myself to one of Peirce's many instructive examples: that of Dr. Lyon Playfair. It illustrates both a mistake resulting from violating predesignation, as well as how, arguing from cases where error probabilities are sustained, the original mistake is corrected. Error correction is not a hope for tomorrow, it *is* the inductive conclusion of tests we run today.

The Example of Dr. Playfair. Peirce describes how "so accomplished a reasoner" as Dr. Playfair violates predesignation in testing a hypothesis about a regularity between the specific gravity of a metal and its atomic weight (2.738). Looking at the specific gravities of 3 forms of carbon, Peirce tells us, Playfair seeks and discovers a formula connecting them: each is a root of the atomic weight of carbon, which is 12. Peirce describes the test Playfair carries out to judge whether this regularity can be expected to hold generally for metals, showing that several alleged instances of the formula really involve modifications not specified in advance. If one limits the instances to ones for which the formula is predesignated, only half satisfy Playfair's formula. Peirce reasons:

> Having thus determined [the] ratio, we proceed to inquire whether an agreement half the time with the formula constitutes any special connection between the specific gravity and the atomic weight of a metalloid. (2.738)

Of particular interest here is the creative use of a canonical test of a proportion. The proportion refers to the *proportion or probability of agreements* with the formula. There is hardly a limit to the kinds of cases where a question about this proportion could be posed.

Peirce then subjects the hypothesis that there *is* a special connection (between the specific gravity and the atomic weight of a metal) to a test of experiment. The falsity of this hypothesis is that the observed agreement is "due to chance" (2.738)—a variant of the standard null hypothesis. Peirce asks, How often would such an agreement be found even if it were due to chance? To answer this question, Peirce *introduces* statistical considerations into an otherwise nonstatistical case.

Peirce introduces a hypothetical chance distribution by matching the specific gravity of a set of elements not with its own atomic weight but with the atomic weight of some other element with which it is

10. Peirce qualifies this. It is sufficient that the Binomial property to be estimated or tested be prespecified; the value of the proportion need not be.

arbitrarily paired. For example, the specific gravity of carbon is compared with the atomic weight of iodine. Note that Peirce is not running more trials of Playfair's experiment, but considering "on paper" how often agreements with Playfair's formula would occur in a case designed so that such agreements could only be due to chance, and using this information about *what would occur* to argue about the cause of the agreements actually found. This strategy is analogous to the other introductions of statistics we have seen, whether they are by random pairings of treatments and subjects, by manipulations done on paper (e.g., Perrin), or by simulation (neutral currents). The logic applied to the results is the same as well.

Peirce finds about the same number of cases satisfying Playfair's formula in this chance pairing of elements as Playfair found in comparing the specific gravities and atomic weights of a given element. Peirce concludes,

> It thus appears that there is no more frequent agreement with Playfair's proposed law than what is due to chance. (2.738)

So Playfair was mistaken in thinking that the evidence showed a special or systematic connection. This example, which merits more attention than I can give it here, is used by Peirce to make a point about predesignation. His point is that the popular inductive accounts are insensitive to the effects of violating predesignation, and as a result they allow one to persist in Playfair's error.

While it would be going too far to see in Peirce the anticipation of Armitage (chapter 10), it is no stretch to see that error probability considerations play identical roles for Peirce and for the error statistician: before the trial their role is to ensure the severity of the test, after the trial it is to assess what induction is warranted. Peirce recognized that violations of predesignation need not preclude severity (see chapter 9). By introducing the hypothetical chance element, Peirce is ascertaining whether Playfair's inference is warranted *despite* the violation of predesignation. He shows that it is not. Whether it is Peirce's handwritten pairings on paper or twentieth-century Monte Carlo simulations in high energy physics, the basic strategy is the same. We find a way of modeling *what it would be like* (in this case, in terms of proportions of agreements) if the agreement is accidental or "due to chance." In Playfair's case, the actual situation is much like what we would expect were the observed agreements accidental.

The SCT and the Relevance of Preliminary Planning

The concern with "the trustworthiness of the proceeding" for Peirce, like the concern with error probabilities for Pearson and error

statisticians generally, is directly tied to their view that statistical method should closely link experimental design and data collection with subsequent inferences. Pearson, remember, railed against the tendency to see statistical inference as beginning once "data are thrown at the statistician and he is asked to draw a conclusion" (Pearson 1966e, 278). Peirce had the same problem with the popular inductive accounts of his day. Peirce regarded as a conclusive refutation of Mill that "an induction, unlike a demonstration, does not rest solely upon the facts observed, but upon the manner in which those facts have been collected" (2.766). Peirce even introduces a term, "quasi-experimentation," to include the entire process of generating and analyzing the data *and* using them to test a hypothesis. And "this whole proceeding," Peirce declares, "I term Induction" (7.115, editor's note). Accordingly, for Peirce, the "true and worthy" task of logic is to "tell you how to proceed to form a plan of experimentation" (7.59).

It is this emphasis on the manner in which the data and hypotheses to test are generated, Peirce stresses, that really distinguishes his view of scientific induction from the two far more popular views of his day (Mill and the conceptualists). That is why the rationale for Peircean induction cannot be divorced from experimental rules for controlling error probabilities.

> This account of the rationale of induction is distinguished from others in that it has as its consequences two rules of inductive inference which are very frequently violated. . . . The first . . . is that the sample must be a random one. . . . The other rule is that the character [about which claims are to be tested] must not be determined by the character of the particular sample taken. (Peirce 1.95)

Hence induction, Peirce says, "must by the rule of predesignation, be a deliberate experiment" (5.579). One wishes that Peirce's critics had made more of the importance he attaches to these rules of data and hypothesis generation. Recognizing their importance is the key to understanding Peirce's SCT: they show that this self-correcting rationale has to do with the control of error probabilities.

Peirce's arguments for these rules are strikingly similar to those arising from the contemporary controversy between Neyman-Pearson "sampling" and nonsampling philosophies, that is, between error probability principles and the likelihood principle. As we saw in previous chapters (e.g., chapter 10), for those who accept the likelihood principle (e.g., Bayesians), once the data are obtained, it is irrelevant for assessing their evidential import how they were selected, or whether the hypothesis was predesignated (the so-called irrelevance of the sampling rule). For these do not alter likelihoods. But they do alter error

probabilities. Just as NP theorists insist on the relevance of predesignation—along the lines detailed in chapter 9—Peirce is highly critical of predesignation being "singularly overlooked by those who have treated of the logic of [induction]" (2.738).

> It is of the essence of induction that the consequence of the theory should be drawn first in regard to the unknown . . . result of experiment. . . . For if we look over the phenomena to find agreements with the theory, it is a mere question of ingenuity and industry how many we shall find. (2.775)

Just as it is only by planning ahead of time that a test can be regarded as a reliable error probe, for Peirce "reasoning tends to correct itself, and the more so, the more wisely its plan is laid" (5.575).

Learning from Qualitative Induction

Now critics claim that for qualitative inductive testing to be self-corrective, it would have to provide a method of replacing substantive hypotheses with better ones, for example, condition *b*. But Peirce calls inferences from data to substantive hypotheses abduction or presumption, not induction. As abductive inference is free to violate predesignation, Peirce holds, it enjoys no such general error-correcting guarantee. Since induction is said only to have the power to correct any errors into which *it* may lead, it is no part of the SCT to show that abduction is trustworthy. However, the self-correcting rationale is all-important when it comes to putting an abductively arrived at hypothesis to the test of experiment.

Inductive testing of the qualitative variety has to do not with replacing falsified hypotheses with brand new ones, but with learning from rejected hypotheses. A central aim is to learn what modifications are called for by the experiments. The rationale for subjecting such an abduction to a severe test of experiment is to learn about these modifications. Among types of qualitative induction, Peirce places a case that "tests a hypothesis by sampling the possible predictions that may be based upon it. . . . We cannot say that a collection of predictions drawn from a hypothesis constitutes a strictly random sample of all that can be drawn. Sometimes we can say that it appears to constitute a very fair, or even a severe sample of the possible prediction" (7.216). Here the correction of hypotheses is expected to come about through gradual modification. Peirce illustrates with the case of the kinetic theory of gases.

> It began with a number of spheres almost infinitesimally small occasionally colliding. It was afterward so far modified that the forces be-

> tween the spheres, instead of merely separating them, were mainly attractive, that the molecules were not spheres, but systems. (7.216)

These modifications "were partly merely quantitative, and partly such as to make the formal hypothesis represent better what was really supposed to be the case, but which had been simplified for mathematical simplicity" (7.216). Peirce grants that there is "no new hypothetical element in these modifications," but it is precisely with these kinds of modifications that induction is concerned. One poses a question, say, "Suppose I tried to model molecules as having uniform radius?" and then learns from the given experimental data how similar or divergent that model would be from the experimental phenomena.

The quantity of interest is not how much the evidence confirms the hypothesis tested—in any of the senses of confirmation—but *how discordant* evidence shows a given model to be in a specified respect. The problem of assessing the approximate accordance of a model is quite different from that of assigning it some E-R measure. There are a handful of methods for putting forward deliberately oversimplified or canonical hypotheses, because, with the appropriate methodology of testing, they serve for learning about these modifications (from rejected hypotheses). Experimental learning requires not some update of the probability assignment that I start out with, but tools to build, correct, and fill out a model. What justifies Peirce's SCT is that induction—understood as severe testing—supplies such tools.

Economy and the Piecemeal Breakdown of Inquiries

Having identified the aim of inductive testing, it is easy to understand Peirce's advice as to the type of hypotheses that are useful to test. Peirce considers "what principles should guide us in abduction, or the process of choosing a hypothesis" (7.219). He lists three: First, the hypothesis selected for tests "must be capable of being subjected to experimental testing"; second, the hypothesis must explain surprising facts. "In the third place," Peirce continues, "is the consideration of economy" (7.220).

The first two are familiar, but the third is rather unique to Peirce. While a concern for economy sounds as if pragmatic or practical considerations are being appealed to, Peirce's concern is in fact wholly epistemological. Considering "economy" in choosing a hypothesis to test means we should consider strategically what questions can be put to a reasonably severe test with the data that are likely to be actually obtainable. This aim, I have argued, leads to "getting small" and to a piecemeal approach to inquiry. It is likewise for Peirce. Under economy

Peirce cites the kind of strategy that makes for a shrewd playing of 20 questions:

> Twenty skillful hypotheses will ascertain what two hundred thousand stupid ones might fail to do. The secret of the business lies in the caution *which breaks a hypothesis up* into its smallest logical components, and only risks one of them at a time. (7.220; emphasis added)

These are questions amenable to the yes/no types of answers typical of standard statistical tests.

Considerations of economy, Peirce says, also direct one to try the same kind of model to account for the same kinds of phenomena, but in different areas. Peirce considers how the model used in the kinetic theory (chapter 7) "accounts for those phenomena . . . by representing that they are results of chance; or . . . of the law of high numbers" (7.221).

Giving Good Leave. A third important consideration under economy is "that it may give a good 'leave,' as the billiard-players say. If it does not suit the facts, still the comparison with the facts may be instructive with reference to the next hypothesis" (7.221). Even if we primarily want to know whether a quadratic equation holds between quantities, we would do well to test a linear model first "because the residuals will be more readily interpretable." The residuals, or errors—the differences between the observed and predicted value—may teach more about the next hypothesis to try. Hence, "even although we imagine that by complicating the hypothesis it could be brought nearer the truth" (ibid.), testing a simpler one may be justified because it will teach us more.

An adequate philosophy of experiment, I agree with Peirce, should include methodological rules directed at asking fruitful questions and arriving at local hypotheses to test, as well as rules for data generation and modeling. The former type has generally been left out of discussions of philosophy of statistics, and yet an important asset of standard statistical methods is that they can offer canonical models and rules for both of these types of rules. The nature and aims of the rules are very much in the spirit of Peirce's considerations. They are not mechanical or algorithmic, but neither are they mere guesswork. The logic of science, for Peirce, is not formal but a systematic methodology for experiment. In a favorite passage, Peirce describes the aim of a theory of experiment thus:

> It changes a fortuitous event which may take weeks or may take many decennia into an operation governed by intelligence, which will be finished within a month. (7.78)

The idea that a central aim of statistical method is to speed things up in this way, while overlooked in philosophical discussions, is at the heart of the rationale of error statistical methods. The concern is not with the kind of speeding up of production that Fisher so disliked (chapter 11), but rather, we might say, with making good on the "long-run" claims in the short long run, if not "within a month," then within a year or the usual amount of the time for a given scientific research project.

That we have a workable theory of experiment, that we make progress with this theory is what the SCT is all about. However, we are not quite finished with justifying this thesis; we have to go back down to the models of data, experimental design, and data generation.

12.4 INDUCTION CORRECTS ITS PREMISES

Justifying experimental inferences depends on being able to justify the assumptions of the experimental and data models required. Self-correcting, or error-correcting, enters here too, and precisely in the way that Peirce recognized. This leads me to consider something apparently missed by critics of the SCT, namely, Peirce's insistence that induction "not only corrects its conclusions, *it even corrects its premises*" (3.575; emphasis added).

Induction corrects its premises by checking, correcting, or validating its own assumptions. One way that induction corrects its premises is by correcting and improving upon the accuracy of its data. The idea is a fundamental part of what allows induction—understood as severe testing—to be genuinely ampliative. It is why, in an important sense, statistical considerations allow one to come out with more than is put in. At times, even "garbage in" need not mean "garbage out."

Peirce comes to his philosophical stances from his experiences with astronomical observations:

> Every astronomer, however, is familiar with the fact that the catalogue place of a fundamental star, which is the result of elaborate reasoning, is far more accurate than any of the observations from which it was deduced. (5.575)

Daily use of the method of least squares taught Peirce how knowledge of errors of observation can be used to infer an accurate observation from highly shaky data.[11] Peirce proceeds to apply the same strategy

11. The method of least squares is a method of finding the best estimate of a parameter value. Given a set of observations made independently, the differences of the observed values from the best estimate are the residuals or errors. The theory of least squares directs one to find the value for which the sum of the squares of

from astronomy to an informal, qualitative example to illustrate how "a properly conducted Inductive research corrects its own premisses":

> That Induction tends to correct itself, is obvious enough. When a man undertakes to construct a table of mortality upon the basis of the Census, he is engaged in an inductive inquiry. And lo, the very first thing that he will discover from the figures . . . is that those figures are very seriously vitiated by their falsity. (5.576)

The premises here are reports on age, and it is discovered that there are systematic errors in these reports. How? By noticing, Peirce explains, that the number of men reporting their age as 21 far exceeds those who are 20, while in all other cases ages are much more likely to be expressed in round numbers. How is it that induction helps to uncover that there is this subject bias, that those under 21 tend to put down that they are 21? It does so by means of formal models of age distributions along with informal background knowledge of the root causes of such bias. "The young find it to their advantage to be thought older than they are, and the old to be thought younger than they are" (5.576). Moreover, statistical considerations often allow one to correct for bias, that is, by estimating the number of "21" reports that are likely to be attributable to 20-year-olds. As with the star catalogue in astronomy, the data thus corrected are *more accurate* than the original data. That is Peirce's main point. The thrust of the thesis that induction corrects its own premises is easy to put in terms of our error statistical framework: by means of an informal tool kit of key errors and their causes, coupled with systematic tools to model them, experimental inquiry checks and corrects its own assumptions for the purpose of carrying out some other (primary) inquiry.

These cases of correcting premises underscore what I have maintained for Peircean self-correction generally. It is not a matter of saying that with enough data we will get better and better estimates of the star positions or the distribution of ages in a population. It is a matter of being able to employ methods right now to detect and correct mistakes in a given inquiry. The methods stem from canonical models of error, here for errors in observations of different types (e.g., from instruments, subjects, etc.). To get such methods off the ground, we need not build a careful tower where evidence rests on a pile of inferences, each as shaky as the ones before (like piles driven into a swamp). Properly collected and cleverly used, inaccurate observations give way to far more accurate data.

---

residuals is minimum. That is the best estimate of the value. This canonical method was used in the eclipse experiments discussed in chapter 8.

Induction Fares Better than Deduction at Correcting Its Errors

Consider how this reading of Peirce makes sense of his holding inductive science as better at self-correcting than deductive science.

> Deductive inquiry . . . has its errors; and it corrects them, too. But it is by no means so sure, or at least so swift to do this as is Inductive science. (5.577)

An example he gives is that the error in Euclid's elements was undiscovered until non-Euclidean geometry was developed. Other everyday examples arise in checking and rechecking calculations to uncover arithmetical errors. "It is evident that when we run a column of figures down as well as up, as a check," or look out for possible flaws in a demonstration, "we are acting precisely as when in an induction we enlarge our sample for the sake of the self-correcting effect of induction" (5.580). In both cases we are appealing to various methods we have devised because we find they increase our ability to correct our mistakes.

What is distinctive about the methodology of inductive testing is that it deliberately directs itself to devising tools for reliable error probes. This is not so for mathematics. Granted, "once an error is suspected, the whole world is speedily in accord about it" (5.577) in the case of deductive reasoning. But for the most part mathematics itself does not supply tools for uncovering flaws. (Consider, in this connection, the recent dispute about the correctness of an alleged proof of Fermat's last theorem.)

> So it appears that this marvelous self-correcting property of Reason . . . belongs to every sort of science, although it appears as essential, intrinsic, and inevitable only in the highest type of reasoning, which is induction. (5.579)

In one's inductive or experimental tool kit, one finds explicit models and methods whose single purpose is the business of detecting patterns of irregularity, checking assumptions, assessing departures from canonical models, and so on. Where experimental tests are unable to do this—where methods are unable to mount severe tests—then they fail to count as scientific induction.

Random Sampling and the Uniformity of Nature

In addition to the rule of predesignation, Peirce's SCT requires that the selection of the experimental sample be random or approximately

so.[12] In fact Peirce is generally credited with defining a random sample. Yet the assumption of random sampling is often thought to be an obstacle to justifying statistical inference. In an interesting footnote, Hans Reichenbach (1971) makes this remark about Peirce:

> The self-corrective nature of induction was emphasized by C. S. Peirce. . . . I have not been able . . . to find a passage in Peirce's work where he clearly states a reason for his contention. The fact that he constantly connects the problem of induction with that of a fair sample . . . seems to indicate that he bases the self-corrective nature of induction on Bernoulli's theorem. . . . Such an argument is invalid, of course, since the justification of induction must be given before the use of probability considerations. (P. 446, n. 1)

There are many intriguing similarities between Peirce and Reichenbach that merit attention, but here I want to dwell on a key point of contrast that this passage points up. For it is this classical view of what is required to justify induction that Peirce is anxious to deny.

Peirce views the problem of justifying induction as explaining why inductive testing is so successful when it is. He contrasts his explanation with those favored by followers of Mill and "almost all logicians" of his day, who "commonly teach that the inductive conclusion approximates to the truth because of the uniformity of nature" (2.775). Inductive inference, as Peirce conceives it (i.e., severe testing) does not use the uniformity of nature as a premise. Rather, the justification is sought in the manner of obtaining data and specifying hypotheses to test. It is a matter of showing that methods exist with good error probabilities. For this it suffices that randomness be met only approximately, that inductive methods check their own assumptions, and that inductive methods can often detect and correct departures from randomness. Says Peirce:

> A sample is a *random* one, provided it is drawn by such machinery . . . that in the long run any one individual of the whole lot would get taken as often as any other. Therefore, judging of the statistical composition of a whole lot from a sample is judging by a method which will be right on the average in the long run, and, by the reasoning of the doctrine of chances, will be nearly right oftener than it will be far from right.
>
> It has been objected that the sampling cannot be random in this sense. But this is an idea which flies far away from the plain facts. Thirty

12. All that is really required is that a statistical relationship between the sampling and the population of interest be known approximately.

> throws of a die constitute an approximately random sample of all the throws of that die; and that the randomness should be approximate is all that is required. (Peirce 1.94)

This again shows that Peirce was in the know about mathematical results (the central limit theorem). (Thirty is the magic number for which the distribution of the sample mean is nearly normal, regardless of the underlying distributions.)

Peirce backs up his defense with robustness arguments. For example, in an (attempted) Binomial induction Peirce asks, "What will be the effect upon inductive inference of an imperfection in the strictly random character of the sampling?" (2.728). What if, for example, a certain proportion of the population had twice the probability of being selected? Peirce shows that "an imperfection of that kind in the random character of the sampling will only weaken the inductive conclusion, and render the concluded ratio less determinate, but will not necessarily destroy the force of the argument completely" (2.728). This is particularly so if the sample mean is near 0 or 1. Yet a further safeguard is at hand, Peirce reminds us:

> Nor must we lose sight of the constant tendency of the inductive process to correct itself. This is of its essence. This is the marvel of it. . . . Even though doubts may be entertained whether one selection of instances is a random one, yet a different selection, made by a different method, will be likely to vary from the normal in a different way, and if the ratios derived from such different selections are nearly equal, they may be presumed to be near the truth. (2.729)

Here the "marvel" is its ability to correct the attempt at random sampling. Numerous, even more marvelous methods exist today to check randomness and other assumptions. Still, Peirce cautions, we should not depend so much on the self-correcting virtue that we relax our efforts to get a random and independent sample. But if our effort is not successful, and our method not robust, we will probably discover it. "This consideration makes it extremely advantageous in all ampliative reasoning to fortify one method of investigation by another" (ibid.).

"The Supernal Powers Withhold Their Hands and Let Me Alone"

Peirce turns the tables on those skeptical about satisfying random sampling—or, more generally, about satisfying the assumptions of a statistical model. He declares himself "willing to concede, in order to concede as much as possible, that when a man draws instances at random, all that he knows is that he *tries* to follow a certain precept" (2.749). There might be a "mysterious and malign connection between

the mind and the universe" that deliberately thwarts such efforts. Peirce considers betting on the game of *rouge et noire*. "Could some devil look at each card before it was turned, and then influence me mentally" to bet or not, the ratio of successful bets might differ greatly from .5 (ibid.). But this would equally vitiate *deductive* inferences about the expected ratio of successful bets. We would find systematic departures from the Binomial model with $p = .5$, even where the card game did have an equal chance of a red or black card.

Peirce's argument can be seen as the counterpart to Neyman's justification for the use of mathematical models of random experiments (from chapter 5). Neyman (1952), recall, had explained how probabilistic models adequately represent certain real experimental procedures "whenever we succeed in arranging the technique of a random experiment, such that the relative frequencies of its different results in long series approach" sufficiently the mathematical probabilities in the sense of the law of large numbers (Neyman 1952, 19). We can check whether we have succeeded in satisfying the statistical model sufficiently. But the experimental procedure whose assumptions are found to be satisfied where p is known should work as well when p is unknown. To suppose otherwise, Peirce is saying, would be akin to supposing a mysterious power can read my mind and deliberately thwart my efforts to satisfy assumptions just when p is unknown.

Peirce therefore grants that the validity of induction is based on assuming "that the supernal powers withhold their hands and let me alone, and that no mysterious uniformity . . . interferes with the action of chance" (2.749). But this is very different from the uniformity of nature assumption.

> The negative fact supposed by me is merely the denial of any major premiss from which the falsity of the inductive or hypothetic conclusion could . . . be deduced. Nor is it necessary to deny altogether the existence of mysterious influences adverse to the validity of the inductive . . . processes. So long as their influence were not too overwhelming, the wonderful self-correcting nature of the ampliative inference would enable us, even so, to detect and make allowance for them. (2.749)

This is the reason for having standard mechanisms, for example, a coin-tossing mechanism such as our canonical Binomial experiment with $p = .5$. Finding systematic departures from the deductively derived statistical distribution would be one way of detecting that we had failed in a particular case to get the experiment to accord with the standard Binomial. We could then subtract out its influence.

Not only do we not need the uniformity of nature assumption, but also, Peirce declares, "That there is a general tendency toward uniformity in nature is not merely an unfounded, it is an absolutely absurd, idea in any other sense than that man is adapted to his surroundings" (2.750). But the validity of inductive inference does not depend on this.

The ability to make successful inductions, our success in obtaining experimental knowledge, is explained by the properties of our methods. The properties of the methods are error probabilities. Because we can frame questions of interest in term of hypotheses amenable to severe testing, we are able to learn from error and in so doing obtain experimental knowledge. That is what Peirce's SCT requires and what Peirce means by saying that "the true guarantee of the validity of induction" is that it is a method of reaching a conclusion that is able to detect errors:

> This it will do . . . because it is manifestly adequate . . . to discovering any regularity there may be among experiences, while utter irregularity is not surpassed in regularity by any other relation of parts to whole, and is thus readily discovered by induction to exist where it does exist, and the amount of departure therefrom to be mathematically determinable from observation. . . . The doctrine of chances . . . is nothing but the science of the laws of irregularities. . . . There is no possibility of a series of experiences so wanting in uniformity as to be beyond the reach of induction, provided there be sufficiently numerous instances of them, and provided the march of scientific intelligence be unchecked. (2.769)[13]

In the final chapter, I shall have more to say about how the error statistical program explains the success of scientific induction.

13. In aligning myself with Peirce, it should not be thought that I agree with a position popularly attributed to him, namely, that truth is the final opinion to which inquiry would eventually lead. One gloss is unproblematic, however: the true but *fixed* value of a population mean is the average of all the possible sample means. Hence, the average of sample means would eventually equal it.