Peirce’s philosophy of inductive inference in science is based on the idea that what permits us to make progress in science, what allows our knowledge to grow, is the fact that science uses methods that are self-correcting or error correcting:

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. (5.145)

Inductive methods — understood as methods of experimental testing — are justified to the extent that they are error-correcting methods. We may call this Peirce’s error-correcting or self-correcting thesis (SCT):

Self-Correcting Thesis SCT: methods for inductive inference in science are error correcting; the justification for inductive methods of experimental testing in science is that they are self-correcting.

Peirce’s SCT has been a source of fascination and frustration. By and large, critics and followers alike have denied that Peirce can sustain his SCT as a way to justify scientific induction: “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” (Rescher 1978, p. 20).

In this paper I shall revisit the Peircean SCT: properly interpreted, I will argue, Peirce’s SCT not only serves its intended purpose, it also provides the basis for justifying (frequentist) statistical methods in science. While on the one hand, contemporary statistical methods increase the mathematical rigor and generality of Peirce’s SCT, on the other, Peirce provides something...
current statistical methodology lacks: an account of inductive inference and a philosophy of experiment that links the justification for statistical tests to a more general rationale for scientific induction. Combining the mathematical contributions of modern statistics with the inductive philosophy of Peirce, sets the stage for developing an adequate justification for contemporary inductive-statistical methodology.

2. Probabilities are assigned to procedures not hypotheses

Peirce’s philosophy of experimental testing shares a number of key features with the contemporary (Neyman and Pearson) Statistical Theory: 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. A Neyman and Pearson (N-P) statistical test, for example, instructs us “To decide whether a hypothesis, \( H \), of a given type be rejected or not, calculate a specified character, \( x_0 \), of the observed facts; if \( x > x_0 \) reject \( H \); if \( x < x_0 \) accept \( H \).” Although the outputs of N-P tests do not assign hypotheses degrees of probability, “it may often be proved that if we behave according to such a rule ... we shall reject \( H \) when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject \( H \) sufficiently often when it is false” (Neyman and Pearson, 1933, p.142).1

The relative frequencies of erroneous rejections and erroneous acceptances in an actual or hypothetical long run sequence of applications of tests are error probabilities; we may call the statistical tools based on error probabilities, error statistical tools. In describing his theory of inference, Peirce could be describing that of the error-statistician:

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. (2.748)
The doctrine of "inverse chances" alludes to assigning (posterior) probabilities in hypotheses by applying the definition of conditional probability (Bayes's theorem) — a computation requires starting out with a (prior or "antecedent") probability assignment to an exhaustive set of hypotheses:

If these antecedent probabilities were solid statistical facts, like those upon which the insurance business rests, the ordinary precepts and practice [of inverse probability] would be sound. But they are not and cannot be statistical facts. What is the antecedent probability that matter should be composed of atoms? Can we take statistics of a multitude of different universes? (2.777)

For Peircean induction, as in the N-P testing model, the conclusion or inference concerns a hypothesis that either is or is not true in this one universe; thus, assigning a frequentist probability to a particular conclusion, other than the trivial ones of 1 or 0, for Peirce, makes sense only "if universes were as plentiful as blackberries" (2.684). Thus the Bayesian inverse probability calculation seems forced to rely on subjective probabilities for computing inverse inferences, but "subjective probabilities" Peirce charges "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 worse of them" (2.777).

Hearing Pierce contrast his view of induction with the more popular Bayesian account of his day (the Conceptualists), one could be listening to an error statistician arguing against the contemporary Bayesian (subjective or other) — with one important difference. Today's error statistician seems to grant too readily that the only justification for N-P test rules is their ability to ensure we will rarely take erroneous actions with respect to hypotheses in the long run of applications. This so called inductive behavior rationale seems to supply no adequate answer to the question of what is learned in any particular application about the process underlying the data. Peirce, by contrast, was very clear that what is really wanted in inductive inference in science is the ability to control error probabilities of test procedures, i.e., "the trustworthiness of the proceeding". Moreover it is only by a faulty analogy with deductive inference, Peirce explains, that many suppose that inductive (synthetic) inference should supply a probability to the conclusion: "... in the case of analytic inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic inferences we only know the degree of trustworthiness of our proceeding ("The Probability of Induction" 2.693).
Knowing the "trustworthiness of our inductive proceeding", I will argue, enables determining the test's probative capacity, how reliably it detects errors, and the severity of the test a hypothesis withstands. Deliberately making use of known flaws and fallacies in reasoning with limited and uncertain data, tests may be constructed that are highly trustworthy probes in detecting and discriminating errors in particular cases. This, in turn, enables inferring which inferences about the process giving rise to the data are and are not warranted: an inductive inference to hypothesis H is warranted to the extent that with high probability the test would have detected a specific flaw or departure from what H asserts, and yet it did not.

3. So why is justifying Peirce's SCT thought to be so problematic?

Thanks to the excellent discussion of Rescher (1978) we can zero right in on the heart of the key criticism. It is this: whereas Peirce claims to have substantiated the SCT for induction generally, he has at most done so for "quantitative" or statistical induction, but he has not done so for "qualitative" induction. The two chief assumptions on the part of critics concern:

(1) the nature of inductive testing (of both types) for Peirce
(2) what is required for a method to be self-correcting

(1) As to the first, for Peirce's critics, quantitative induction is construed as classic enumerative induction or "the straight rule". Here, one infers from observing n A's that are B's to the proportion of A's that are B's in the population from which the observations derive. By qualitative induction, critics understand Peirce to mean hypothetico-deductive (H-D) inference. In its most rudimentary form, a hypothesis H is tested by deducing from it a prediction x, if x does not occur, H is falsified; if x does occur H is accepted or supported or confirmed in some sense. However, both of these modes of inference fall short of Peircean induction which requires a test procedure that is trustworthy, or, in more modern terms, reliable or severe. Thus to adequately appraise the SCT, the first thing that is needed is a revision of this construal of Peirce's two types of induction.

(2) As for the second issue, critics are fairly clear as to what they suppose is required for an inductive method M to be self-correcting:

(a) M must eventually arrive at the truth, it asymptotically approaches truth in the long run;
(b) M must provide a (mechanical?) method of replacing rejected hypotheses with better (truer) ones. (Laudan, p. 229)²
We can flesh out the charge that Pierce fails to deliver what the SCT promises. The charge, more particularly, is that while quantitative or statistical induction is taken to pretty well satisfy both (a) and (b), qualitative induction only satisfies (a).

Such qualitative inductions clearly satisfy the first condition for [a self-correcting method], insofar as persistent application of the method of hypothesis testing 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 discovering an alternative which is (or is likely to be) closer to the truth than the refuted hypothesis. 

(Laudan 1981, pp. 238-239)

Ilkka Niiniluoto (1984) in like fashion, assimilates Peircean self-correcting to a view of scientific progress as replacing earlier theories with ones closer to the truth, leading him also to criticize Peirce for not having told us how induction affords such progress. The technique for discovering a better alternative, moreover, is supposed to be mechanical or routine, and it is clear that induction, even correctly conceived, fails to provide such a technique. But need a method provide such a technique to be self-correcting? Rescher, for one, thinks not. It suffices, he argues, that the whole conglomeration of scientific methods serves to find better alternatives.

Science is autonomous. Corrections to science must come from science ... The mistaken results of science can be improved or corrected only by further results of science. There can be no recourse at this point to tea leaf reading, numerology, the Delphic oracle or the like. (p. 160)

While I agree, I think Peirce is saying something more specific about the kinds of inductive methods and strategies in science: that the inductive methods of the type he is endorsing are very good at uncovering mistakes and learning from errors. This, of course, takes us to the first issue of revising the construal of Peirce’s two types of induction.

Neither ‘the straight rule’ of induction, nor H-D inference — as these have been understood — are what Peirce is recommending for quantitative or qualitative induction! Induction for Peirce is a matter of “trustworthy” or reliable experimental testing. Evaluating the “trustworthiness of inductive
procedures" requires determining how reliably they detect error, and the severity of the tests a hypothesis withstands, as measured formally (quantitative induction) or informally (qualitative induction).

4. Peircean induction as severe testing

What is my evidence for this reading of him? For starters, induction, for Peirce, is a matter of subjecting hypotheses to "the test of experiment" (7.182).

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 ... which would be very unlikely or surprising in case the hypothesis were not true. (7.231)

When, however, we find that prediction after prediction, notwithstanding a preference for putting the most unlikely ones to the test, is verified by experiment, ... 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)

While these and other passages are redolent of Popper, Peirce differs from Popper in crucial ways. Peirce, unlike Popper, is primarily interested not in falsifying claims but in the positive pieces of information provided by tests, with "the corrections called for by the experiment" and with the hypotheses, modified or not, that manage to pass severe tests. For Popper, even if a hypothesis is highly corroborated (by his lights), he regards this as at most a report of the hypothesis' past performance and denies it affords positive evidence for its correctness or reliability. Further, Popper denies that he could vouch for the reliability of the method he recommends as "most rational" — conjecture and refutation. Indeed, Popper's requirements for a highly corroborated hypothesis are not sufficient for ensuring severity in Peirce's sense (Mayo 1996, 2003, 2004). Where Popper recoils from even speaking of warranted inductions, Peirce conceives of a proper inductive inference as what had passed a severe test — one which would, with high probability, have detected an error if present.

In Peirce's inductive philosophy, we have evidence for inductively inferring a claim or hypothesis H when not only does H "accord with" the
data $x$, but also, so good an accordance would very probably not have resulted, were $H$ not true. In other words, we may inductively infer $H$ when it has withstood a test of experiment that it would not have withstood, or withstood so well, were $H$ not true (or were a specific flaw present). This can be encapsulated in the following severity requirement for an experimental test procedure, $ET$, and data set $x$:

**Hypothesis $H$ passes a severe test with $x$ iff** (firstly) $x$ accords with $H$ and (secondly) the experimental test procedure $ET$ would, with very high probability, have signaled the presence of an error were there a discordancy between what $H$ asserts and what is correct (i.e., were $H$ false).

The test would "have signaled an error" by having produced results less accordant with $H$ than what the test yielded. Thus, we may inductively infer $H$ when (and only when) $H$ has withstood a test with high error detecting capacity, the higher this probative capacity, the more severely $H$ has passed. What is assessed (quantitatively or qualitatively) is not the amount of support for $H$ but the probative capacity of the test of experiment $ET$ (with regard to those errors that an inference to $H$ is declaring to be absent).

These observations about Peirce's inductive philosophy lead us to a reworking of the two assumptions of critics of the SCT regarding (1) the nature of inductive testing (of both types) and (2) what is required for a method to be self-correcting. (1) Firstly, what distinguishes "quantitative" from "qualitative" induction is not that the former is the straight rule while the latter a H-D inference. Both types of inference — in so far as they qualify as Peircean inductions — are arguments based on tests with various degrees of severity, what distinguishes them is the extent to which their severity or reliability or error probing capacity can be quantitatively or only qualitatively determined.

(2) Secondly, from the severity requirement we get a strengthened form of condition (a), the inductive test procedure must not merely asymptotically approach truth, it must have a high 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). 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) involves two main capacities for self-correcting: (i) learning from test results: First, the methods 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, we learn about the process or phenomenon from which the data arose. The overall methodology should supply systematic, though not necessarily mechanical, methods for learning from hypotheses that fail as well as those that pass probative tests. (ii) Correcting its own assumptions: Second the method should be able to detect its own errors in the sense of checking its own assumptions (its “own premises”); it should be able to correct violations or subtract them out in the analysis of data. Moreover, having uncovered inadequacies in either the hypotheses tested, or in the test procedures themselves (e.g., determining the test fails to severely probe a full theory of interest), the self-critical scrutiny of (i) and (ii) should form the basis by which the methods teach about fruitful hypotheses or models to try next (taking into account considerations of efficiency and economy). To sum up my reworking of (2) to show that scientific induction is self-corrective is to show that severe testing methods exist and that they enable reliable means for learning from error. My task is to justify these claims, or at least sketch how a full justification would go.

5. The path from qualitative to quantitative induction

In my understanding of Peircean induction, the difference between qualitative and quantitative induction is really a matter of degree, according to whether their trustworthiness or severity is quantitatively or only qualitatively ascertainable. This reading not only neatly organizes Peirce’s typologies of the various types of induction, it underwrites the manner in which, within a given classification, Peirce further subdivides inductions by their “strength”.

(I) First-Order, Rudimentary or Crude Induction

Consider Peirce’s First Order of induction: the lowest, most rudimentary form that he dubs, the “pooh-pooh argument”. It is essentially an argument from ignorance: Lacking evidence for the falsity of some hypothesis or claim H, provisionally adopt H. In this very weakest sort of induction, crude induction, the most that can be said is that a hypothesis would eventually be falsified if false. (It may correct itself — but with a bang!) It “is as weak an inference as any that I would not positively condemn” (8.237). While uneliminable in ordinary life, Peirce denies that rudimentary induction is to be included as scientific induction. Without some reason to think evidence of H’s falsity would probably have been detected, were H false, finding no evidence against H is poor inductive evidence for H. H has passed only a highly unreliable error probe.

(II) 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, corresponding neatly to viewing
strength as the severity of a testing procedure.

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 ... is where [results] 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 occurs where the predictions, though fulfilled, lack novelty; whereas, the stronger type reflects a more stringent hurdle having been satisfied: the hypothesis has had “novel” predictive success, and thereby higher severity. (For a discussion of the relationship between types of novelty and severity see Mayo 1991, 1996). Note that within a second order induction the assessment of strength is qualitative, e.g., 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)

It is upon these and like passages that I base my reading of Peirce. A qualitative induction, i.e., a test whose severity is qualitatively determined, becomes a quantitative induction when the severity is quantitatively determined; when an objective error probability can be given.

(III) Third Order, Statistical (Quantitative) Induction

We enter the Third Order of statistical or quantitative induction when it is possible to quantify “how much” the prediction runs counter to what our expectation would have been without the hypothesis. In his discussions of such quantifications, Peirce anticipates to a striking degree later developments of statistical testing and confidence interval estimation (Hacking 1980, Mayo 1993, 1996). Since this is not the place to describe his statistical contributions, I move to more modern methods to make the qualitative-quantitative contrast.
6. Quantitative and qualitative induction: significance test reasoning

Quantitative Severity

A statistical significance test illustrates an inductive inference justified by a quantitative severity assessment. The significance test procedure has the following components: (1) a null hypothesis $H_0$, which is an assertion about the distribution of the sample $X = (X_1, ..., X_n)$, a set of random variables, and (2) a function of the sample, $d(x)$, the test statistic, which reflects the difference between the data $x = (x_1, ..., x_n)$, and null hypothesis $H_0$. The observed value of $d(x)$ is written $d(x)$. The larger the value of $d(x)$ the further the outcome is from what is expected under $H_0$, with respect to the particular question being asked. We can imagine that null hypothesis $H_0$ is

$$H_0: \text{there are no increased cancer risks associated with hormone replacement therapy (HRT) in women who have taken them for 10 years.}$$

Let $d(x)$ measure the increased risk of cancer in $n$ women, half of which were randomly assigned to HRT. $H_0$ asserts, in effect, that it is an error to take as genuine any positive value of $d(x)$ — any observed difference is claimed to be "due to chance". The test computes (3) the p-value, which is the probability of a difference larger than $d(x)$, under the assumption that $H_0$ is true:

$$p\text{-value} = \text{Prob}(d(X) > d(x); H_0).$$

If this probability is very small, the data are taken as evidence that

$$H^*: \text{cancer risks are higher in women treated with HRT}$$

The reasoning is a statistical version of *modus tollens*:

If the hypothesis $H_0$ is correct then, with high probability, $1-p$, the data would *not* be statistically significant at level $p$.

$x$ is statistically significant at level $p$.

Therefore, $x$ is evidence of a discrepancy from $H_0$, in the direction of an alternative hypothesis $H$.

(*i.e., $H^*$ severely passes, where the severity is 1 minus the p-value)*.

For example, the results of recent, large, randomized treatment-control
studies showing statistically significant increased risks (at the 0.001 level) give strong evidence that HRT, taken for over 5 years, increases the chance of breast cancer, the severity being 0.999. If a particular conclusion is wrong, subsequent severe (or highly powerful) tests will with high probability detect it. In particular, 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.

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. (7.115)

The correction is not a matter of getting higher and higher probabilities, 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.

**Semi-formal or Qualitative Severity:**

The quantitative statistical significance assessment, however, assumes that the underlying statistical requirements for the calculation are met. In criticizing such assumptions, one may again use severity reasoning. An example might be the critique of earlier, observational, studies that led scientists to suppose, not only that risks were negligible or absent, but that

\[
H: \text{HRT is advantageous for post-menopausal women.}
\]

An informal (or semi-formal) critique might note that earlier observational studies on HRT had little capacity to distinguish benefits due to HRT from confounding factors separately correlated with the beneficial outcomes (e.g., women treated with HRT are healthier, have better access to medical care, and are better educated than women not taking HRT). The agreement (between \( x \) and \( H \)) from the observational studies failed to provide evidence in support of \( H \) because \( H \) did not thereby pass a probative or severe test.

In ordinary day-to-day reasoning, we infer that errors are absent if we would almost surely have detected any errors, and we do so without any formal assessment of severity. Indeed, the strongest severity arguments are those where no formal assessment of probativeness is required.

There are two other points of confusion in critical discussions of the SCT, that we may note here:
I. The SCT and the Requirements of Randomization and Predesignation

The concern with “the trustworthiness of the proceeding” for Peirce like the concern with error probabilities (e.g., significance levels) for error-statisticians generally, is directly tied to their view that inductive method should closely link inferences to the methods of data collection as well as to how the hypothesis came to be formulated or chosen for testing.

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 (1.95) namely, that the sample be (approximately) random and that the property being tested not be determined by the particular sample — i.e., predesignation.

The picture of Peircean induction that one finds in critics of the SCT disregards these crucial requirements for induction: Neither enumerative induction nor H-D testing, as ordinarily conceived, requires such rules. Statistical significance testing, however, clearly does.

Suppose, for example that researchers wishing to demonstrate the benefits of HRT search the data for factors on which treated women fare much better than untreated, and finding one such factor they proceed to test the null hypothesis:

\[ H_0: \text{there is no improvement in factor F (e.g. memory)} \]

among women treated with HRT.

Having selected this factor for testing solely because it is a factor on which treated women show impressive improvement, it is not surprising that this null hypothesis is rejected and the results taken to show a genuine improvement in the population. However, when the null hypothesis is tested on the same data that led it to be chosen for testing, it is well known, a spurious impression of a genuine effect easily results. Suppose, for example, that 20 factors are examined for impressive-looking improvements among HRT-treated women, and the one difference that appears large enough to test turns out to be significant at the 0.05 level. The actual significance level — the actual probability of reporting a statistically significant effect when in fact the null hypothesis is true — is not 5% but approximately 64% (Mayo 1996, Mayo and Kruse 2001, Mayo and Cox 2005). To infer the denial of \( H_0 \), and infer there is evidence that HRT improves memory, is to make an inference with low severity (approximately 0.36).
II. Understanding the "long-run error correcting" metaphor

Discussions of Peircean 'self-correction' often confuse two interpretations of the 'long-run' error correcting metaphor, even in the case of quantitative induction: (a) Asymptotic self-correction (as n approaches ∞): In this construal, it is imagined that one has a sample, say of size n=10, and it is supposed that the SCT assures us that as the sample size increases toward infinity, one gets better and better estimates of some feature of the population, say the mean. Although this may be true, provided assumptions of a statistical model (e.g., the Binomial) are met, it is not the sense intended in significance-test reasoning nor, I maintain, in Peirce's SCT. Peirce's idea, instead, gives needed insight for understanding the relevance of 'long-run' error probabilities of significance tests to assess the reliability of an inductive inference from a specific set of data. (b) Error probabilities of a test: In this construal, one has a sample of size n, say 10, and imagines hypothetical replications of the experiment — each with samples of 10. Each sample of 10 gives a single value of the test statistic d(X), but one can consider the distribution of values that would occur in hypothetical repetitions (of the given type of sampling). The probability distribution of d(X) is called the sampling distribution, and the correct calculation of the significance level is an example of how tests appeal to this distribution: Thanks to the relationship between the observed d(x) and the sampling distribution of d(X), the former can be used to reliably probe the correctness of statistical hypotheses (about the procedure) that generated the particular 10-fold sample. That is what the SCT is asserting.

It may help to consider a very informal example. Suppose that weight gain is measured by 10 well-calibrated and stable methods, possibly using several measuring instruments and the results show negligible change over a test period of interest. This may be regarded as grounds for inferring that the individual's weight gain is negligible within limits set by the sensitivity of the scales. Why? While it is true that by averaging more and more weight measurements, i.e., an eleventh, twelfth, etc., one would get asymptotically close to the true weight, that is not the rationale for the particular inference. The rationale is rather that the error probabilistic properties of the weighing procedure (the probability of ten-fold weighings erroneously failing to show weight change) inform one of the correct weight in the case at hand, e.g., that a 0 observed weight increase passes the "no-weight gain" hypothesis with high severity.

7. Induction corrects its premises

Justifying the severity, and accordingly, the error-correcting capacity, of tests depends upon being able to justify sufficiently test assumptions, whether in the quantitative or qualitative realms. In the former, a typical assumption would be that the data set constitutes a random sample from the appropriate population; in the latter, assumptions would include such things as "my
instrument (e.g., scale) is working”. The problem of justifying methods is often taken to stymie attempts to justify inductive methods. Self-correcting, or error-correcting, enters here too, and precisely in the way that Peirce recognized. This leads me to consider something apparently overlooked by his critics; namely, Peirce’s insistence that induction “not only corrects its conclusions, it even corrects its premises” (3.575).

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. This idea is at the heart of what allows induction — understood as severe testing — to be genuinely ampliative: to come out with more than is put in. 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)

His day-to-day use of the method of least squares made it apparent to him how knowledge of errors of observation can be used to infer an accurate observation from highly shaky data.

It is commonly assumed that empirical claims are only as reliable as the data involved in their inference, thus it is assumed, with Popper, that “should we try to establish anything with our tests, we should be involved in an infinite regress” (Popper 1962, p. 388). Peirce explicitly rejects this kind of “tower image” and argues that we can often arrive at rather accurate claims from far less accurate ones. For instance, with a little data massaging, e.g., averaging, we can obtain a value of a quantity of interest that is far more accurate than individual measurements.

**Qualitative Error Correction**

Peirce applies the same strategy from astronomy to a qualitative example:

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)

How is it discovered that there are systematic errors in the age reports? By
noticing 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). Induction, as Pierce understands it, helps to uncover 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 correcting for bias, i.e., by estimating the number of “21” reports that are likely to be attributable to 20 year olds. As with the star catalogue, the data thus corrected is more accurate than the original data report.

By means of an informal tool kit of key errors and their causes, coupled with formal or systematic tools to model them, experimental inquiry checks and corrects its own assumptions for the purpose of carrying out some other inquiry. As I have been urging for Peircean self-correction generally, satisfying the SCT is not a matter of saying 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 in a given inquiry to detect and correct mistakes in that inquiry, or that data set. To get such methods off the ground there is no need to build a careful tower where inferences are piled up, each depending on what went on before: Properly exploited, inaccurate observations can give way to far more accurate data. By building up a “repertoire” of errors and means to check, avoid, or correct them, scientific induction is self-correcting.

**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. Or again, “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, and thus increase the error probing power of our reasoning. 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 deductive reasoning. But, for the most part mathematics does not itself supply tools for uncovering flaws.

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. If an experimental test is unable to do this — if it is unable to mount severe tests — then it fails to count as scientific induction.

8. Random sampling and the uniformity of nature

We are now at the point to address the final move in warranting Peirce’s SCT. The severity or trustworthiness assessment, on which the error correcting capacity depends, requires an appropriate link (qualitative or quantitative) between the data and the data generating phenomenon, e.g., a reliable calibration of a scale in a qualitative case, or a probabilistic connection between the data and the population in a quantitative case. Establishing such a link, however, is regarded as assuming observed regularities will persist, or making some “uniformity of nature” assumption — the bugbear of attempts to justify induction.

But Peirce contrasts his position 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. Justifying induction is a matter of showing that there exist methods with good error probabilities. For this it suffices that randomness be met only approximately, that inductive methods check their own assumptions, and that they can often detect and correct departures from randomness.

... 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 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. (1.94)
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? He 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. In other words, violating experimental assumptions may be shown to weaken the trustworthiness or severity of the proceeding, but this may only mean we learn a little less.

Yet a further safeguard is at hand:

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 an inductive method’s ability to correct the attempt at random sampling. 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 neither is our method 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, 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 tried 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. He 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 0.5. But, as Peirce is quick to point out, this would equally vitiate
deductive inferences about the expected ratio of successful bets.

Consider our informal example of weighing with calibrated scales. If I check the properties of the scales against known, standard weights, then I can check if my scales are working in a particular case. Were the scales infected by systematic error, I would discover this by finding systematic mismatches with the known weights; I could then subtract it out in measurements. That scales have given properties where I know the object’s weight indicates they have the same properties when the weights are unknown, lest I be forced to assume that my knowledge or ignorance somehow influences the properties of the scale. More generally, Peirce’s insightful argument goes, the experimental procedure thus confirmed where the measured property is known must work as well when it is unknown unless a mysterious and malign demon deliberately thwarts my efforts.

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” (ibid.). But this is very different from the uniformity of nature assumption.

...the negative fact supposed by me [no mysterious force interferes with the action of chance] is merely the denial of any major premise from which the falsity of the inductive conclusion could be deduced. Actually so long as the influence of this mysterious source not be overwhelming, the wonderful self-correcting nature of the ampliative inference would enable us, even so, to detect and make allowance for them. (2.749)

Not only do we not need the uniformity of nature assumption, 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). In other words, it is not nature that is uniform, it is we who are able to find patterns enough to serve our needs and interests. But the validity of inductive inference does not depend on this.

9. Conclusion

For Peirce, “the true guarantee of the validity of induction” is that it is a method of reaching conclusions which corrects itself; inductive methods — understood as methods of severe testing — are justified to the extent that they are error-correcting methods (SCT). I have argued that the well-known skepticism as regards Peirce’s SCT is based on erroneous views concerning the nature of inductive testing as well as what is required for a method to be self-correcting. By revisiting these two theses, justifying the SCT boils down to
showing that severe testing methods exist and that they enable reliable means for learning from error.

An inductive inference to hypothesis H is warranted to the extent that H passes a severe test, that is, one which, with high probability, would have detected a specific flaw or departure from what H asserts, and yet it did not. Deliberately making use of known flaws and fallacies in reasoning with limited and uncertain data, tests may be constructed that are highly trustworthy probes in detecting and discriminating errors in particular cases. Modern statistical methods (e.g., statistical significance tests) based on controlling a test's error probabilities provide tools which, when properly interpreted, afford severe tests. While on the one hand, contemporary statistical methods increase the mathematical rigor and generality of Peirce's SCT, on the other, Peirce provides something current statistical methodology lacks: an account of inductive inference and a philosophy of experiment that links the justification for statistical tests to a more general rationale for scientific induction. Combining the mathematical contributions of modern statistics with the inductive philosophy of Peirce sets the stage for developing an adequate solution to the age-old problem of induction. To carry out this project fully is a topic for future work.

Virginia Polytechnic Institute and State University
mayod@vt.edu

REFERENCES


NOTES


2. This statement of (b) is regarded by Laudan as the strong thesis of self-correcting. A weaker thesis would replace (b) with (b'): science has techniques for determining unambiguously whether an alternative T' is closer to the truth than a refuted T.
3. If the p-value were not very small, then the difference would be considered statistically insignificant (generally small values are 0.1 or less). We would then regard $H_0$ as consistent with data $x$, but we may wish to go further and determine the size of an increased risk $r$ that has thereby been ruled out with severity. We do so by finding a risk increase, such that, $\text{Prob}(d(x) > d(x); \text{risk increase } r)$ is high, say. Then the assertion: the risk increase $< r$ passes with high severity, we would argue.

If there were a discrepancy from hypothesis $H_0$ of $r$ (or more), then, with high probability, $1-p$, the data would be statistically significant at level $p$.

$x$ is not statistically significant at level $p$.

Therefore, $x$ is evidence than any discrepancy from $H_0$ is less than $r$.

For a general treatment of effect size, see Mayo and Spanos (forthcoming).