Brit. J. Phil. Sci. **53** (2002), 69–76

Resolving Neyman's Paradox

Max Albert

ABSTRACT

According to Fisher, a hypothesis specifying a density function for X is falsified (at the level of significance α) if the realization of X is in the size-α region of lowest densities. However, non-linear transformations of X can map low-density into high-density regions. Apparently, then, falsifications can always be turned into corroborations (and vice versa) by looking at suitable transformations of X (Neyman's Paradox). The present paper shows that, contrary to the view taken in the literature, this provides no argument against a theory of statistical falsification.

1 The problems of statistical falsification

From a theoretical perspective, statistical inference is a serious problem for falsificationism. Actual tests of scientific hypotheses in most cases involve statistical arguments. Falsificationism is therefore plausible only if it can deal with statistical hypotheses. However, there is no falsificationist theory of statistical inference that is accepted even by falsificationists.[1]

Except for the trivial case where a zero-probability event is observed, a falsification of a statistical hypothesis presupposes the choice of a *rejection region*, that is, an event that is possible under the hypothesis in question but the observation of which is nevertheless taken as a falsification.[2] Following Popper ([1984]) and Gillies ([1971], [1973]), the rejection region is to be

[1] Forster ([1995], p. 402) explains the attraction of Bayesianism in philosophy of science by the fact that the logic-based accounts of scientific inference (which include falsificationism) have nothing to say about statistical problems. Nevertheless, statistical practice is often quite falsificationist in spirit; cf. Gillies ([forthcoming]) on Neyman's use of the χ^2 test. This causes problems to non-falsificationists, who have to argue that standard and successful practices are in fact illegitimate.

[2] 'Rejection' and 'falsification' are used as synonyms throughout, although this is not always in accordance with the use of the term 'rejection' in the statistical literature.

chosen according to a methodological rule, a falsifying rule for statistical inference (FRSI).

According to Neyman and Pearson's ([1933]) theory of statistical inference (NPT), such a rule should take two kinds of errors into account, usually called error of the first and error of the second kind. A first-kind error is an erroneous falsification of a true hypothesis, while a second-kind error is the failure to falsify a false hypothesis (erroneous corroboration). The risk of a first-kind error can be controlled by choosing a rejection region that has a sufficiently small probability α, also called level of significance, under the hypothesis. The risk of a second-kind error has to be controlled by considering at least one explicit alternative to the hypothesis under test. The optimal test in the case of only two alternatives minimizes the probability β of not falsifying the original hypothesis if the alternative is true. Equivalently, it maximizes the power $1 - \beta$.

The NPT is acceptable to falsificationists if and only if the complete set of alternative hypotheses considered is part of the background knowledge, i.e. can be taken for granted in the context of the inquiry. If, for example, it is known that a certain procedure guarantees random sampling from a finite population (in the sense of every element of the population having the same probability to be included in the sample), then there is a fixed set of hypotheses concerning the population average of any variable, and these hypotheses give rise to a fixed set of statistical hypotheses. The random-sampling hypothesis is part of the background knowledge, and the set of hypotheses in which we can use an NP test is determined by the logical consequences of this basic hypothesis together with certain non-statistical hypotheses about the population. In such cases, a falsificationist can agree with Neyman ([1965], p. 448) that there is no difference between estimation (choosing the in some sense best hypothesis from a given set) and testing.[3]

When, however, the question arises (as it must at some stage) as to whether the basic hypothesis in the background is true, the situation is different. One may be able to find some test based on a still more general basic hypothesis, but if the hypothesis under scrutiny becomes more and more general, the NPT quickly runs out of tests. In practice, therefore, some assumptions are tested with the help of tests that are quite general but questionable from the standpoint of the NPT, like the χ^2 test.

Thus, a falsificationist interpretation of the NPT must be based on some other, more fundamental approach answering the question of how to assess,

[3] For these cases, the interpretation and extension of the NPT by Mayo ([1996]) fits into a falsificationist framework. See also Mayo and Spanos ([2000]) for further progress in rendering the idea of the severity of a test more precise. Note that the falsificationist idea of severity of tests relies on background knowledge that can be (or already has been) tested independently; cf. Musgrave's ([1974]) discussion of Hempel's raven paradox.

as econometricians are wont to say, the assumptions of the statistical model. At this level, at least, tests should be independent from alternative hypotheses (Gillies [1971], [1973], [forthcoming]; Albert [1992]).

The obvious candidate for a theory of statistical falsification, Gillies' ([1971], [1973]) theory, is open to three important objections that have already been raised against Fisher's theory of significance testing (cf. Fisher [1990], Spielman [1974]). Consider a statistical hypothesis stating a distribution function for a one-dimensional random variable (rv). Let us assume that we have an FRSI specifying, without recourse to alternative hypotheses, a rejection region for a single observation of the rv. Then the following three problems must be solved.[4]

1. Selection of a test statistic. It is not obvious how to extend the FRSI to the case of $n > 1$ observations. The solution is to select a one-dimensional test statistic; however, there are many candidates.

2. Optional stopping. The rule determining the number of observations, the stopping rule, has a potential influence on the level of significance if the decision to stop is not independent from the observations (optional stopping). For any experiment, there are always several interpretations yielding different significance levels: the conventional interpretation that assumes stopping to be independent from observations, and many alternative interpretations specifying different ways of optional stopping. If one concludes that optional stopping makes a difference, the experimenter's *intentions* matter for the evaluation of otherwise indistinguishable observations.

3. Neyman's Paradox. Assume that the rv is continuously distributed with a density. Fisher's and Gillies' theories select the lowest-density regions as rejection regions. Non-linear transformations of the rv can map low-density into high-density regions and vice versa, leading to different decisions on the basis of the same FRSI and the same observations.

The NPT solves the first and the third problem by appealing to alternative hypotheses. As already argued, this is not always an option for falsificationists. The second problem also arises for the NPT.

However, these problems pose no insurmountable obstacles for a theory of statistical falsification. Just adopting some test by convention (e.g. the χ^2

[4] The problem of selecting a test statistic originally motivated Neyman and Pearson ([1930], [1933]) to modify Fisher's theory; see also Hacking ([1965], pp. 75–81). On optional stopping, see Hacking ([1965], pp. 107–9) and Berger and Berry ([1988]). Neyman's Paradox, originally formulated in connection with the *t* test, goes back to a 1929 paper of Neyman; cf. Neyman and Pearson ([1930], p. 101n) and Neyman ([1952], pp. 45–51). The general formulation is due to Redhead ([1974]).

goodness-of-fit test for discrete hypotheses) goes far in solving the first problem.[5]

Mayo ([1996]) argues that one actually should worry about optional stopping, thus attacking the position of adherents of the likelihood principle (Bayesians and others) that the insensitivity of likelihood-based inferences to optional stopping is a point in their favor.

The present note argues that Neyman's Paradox does not arise if one accepts a reasonable condition concerning the scope of an FRSI.

2 Redhead's version of Neyman's Paradox

The very general nature of Neyman's Paradox has been argued most forcefully by Redhead ([1974]) in his critique of Gillies' rule. Consider the hypothesis $X \sim N$, meaning that the rv X is distributed according to the cumulative distribution function N of the standard normal distribution, and restrict the analysis to the case of a single observation. The so-called Gauss test at a level of significance 0.05 then requires rejection if $|x| \geqslant 1.96$. This test is accepted by Fisher and Gillies, at least in principle. Consider the rv $Y = t(X)$, where

$$t(X) \stackrel{\text{def}}{=} \begin{cases} N^{-1}\left[\frac{3}{2} - N(X)\right], & X > 0 \\ N^{-1}\left[\frac{1}{2} - N(X)\right], & X < 0 \end{cases}$$

and where therefore $Y \sim N$.[6] The same Gauss test can be applied to Y as well. If X is in the rejection region, Y is not, and vice versa. Thus one can always reformulate the hypothesis in such a way that it is not rejected by mathematically the same test.

Redhead's argument is completely general. Transformations that map the tails of the original distribution to the centre of the new distribution can always be found. Moreover, Redhead's transformation guarantees for any unimodal distribution with mode 0 that X and $t(X)$ are distributed identically.

3 A Counterargument

The present paper argues that Neyman's Paradox, taken as a general criticism of a theory of statistical falsification, misses its target. An FRSI has only one purpose, namely to render statistical hypotheses falsifiable. Thus, I suggest that a theory of statistical falsification is based on the following adequacy condition.

[5] Albert ([1992]) proposes an FRSI for discrete rvs similar in spirit to Gillies' rule but covering the test-selection problem. This rule selects the multinomial goodness-of-fit test, which can be approximated by the χ^2 test under certain conditions. On continuous distributions, see below.

[6] $t(0)$ is not defined. This is not important, however, since distributions differing only w.r.t. zero-probability events are taken to be equivalent.

Scope of an FRSI. An FRSI only applies to hypotheses that can immediately be confronted with data (low-level hypotheses).

Real-world measurement processes have finite precision. For this reason, continuous distributions, as far as they are part of the low-level hypotheses under test, are approximations. This means that Neyman's Paradox does not arise since it depends on the assumption that real variables can be observed with infinite precision. The adequacy condition destroys the basis for the paradox.

Let us make this clear by reference to an example. Since we consider just Neyman's Paradox and not the other objections against statistical falsifications, assume that there is an FRSI selecting a test for discrete rvs.

Assume that the theory from which the low-level hypothesis under test is derived postulates a continuous distribution. For example, one might argue that the height of humans at age 25 (X) is a continuous rv. We want to test the hypothesis that this rv obeys a continuous distribution with density f. Since we focus on Neyman's Paradox, assume that we test the hypothesis on the basis of a single observation; the argument is completely analogous if we consider, for example, the mean value of a sample. The actual process of measurement yields intervals on a specific scale, and these intervals have a certain probability according to the hypothesis. We are back to a test of a discrete hypothesis where, by assumption, the FRSI unambiguously picks a rejection region.

4 Different measurement processes

In a presentation of this argument, the following point was raised. It is physically possible to measure $Y = a/X$ instead of X (where a is some constant). Let us make this clear by an example.

Assume that for some obscure reason a person's height is measured in the following way. The person stands upright against a wall. A sliding bar that projects 20 cm from the wall at a right angle is moved down until it touches the top of the person's head. The tip of the bar (point A), the point exactly below it on the ground (point B), and a third point C on the ground 320 cm from point B form a right-angled triangle (see Figure 1). The length of $\overline{AB}$ is the height of the person, the length of $\overline{BC}$ is known to be 300 cm. From A to C, a string is stretched. We mark the point D where the height of the string is exactly 100 cm; the distance of D from C is the rv Y. We have $Y = a/X$ where $a = 30{,}000$.

With Y, we have a transformed variable with density $f(a/Y)$. Actual measurements again yield intervals on a certain scale, resulting in a different discrete hypothesis that can be tested by using the FRSI. It is possible that the

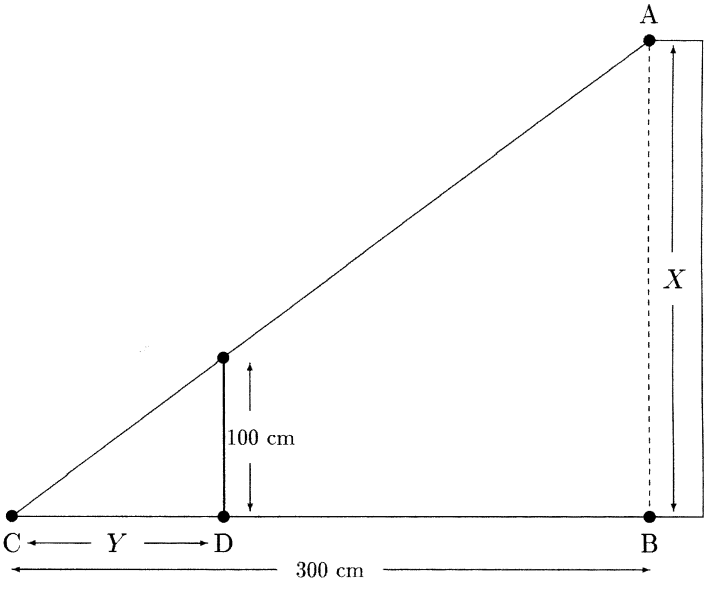

Figure 1. Measuring $Y = a/X$, where X is the height of a person.

same observation gives rise to a falsification if X is measured, while a corroboration occurs if Y is measured. At first glance, then, it seems that Neyman's Paradox again raises its ugly head.

However, a closer examination shows that this is not true. While there is one observation, there are two different low-level hypotheses since two different measurement processes are used. These hypotheses differ not only w.r.t. the rv and the density function, but also w.r.t. the if-clause describing the experimental set-up. It is even possible that one of the hypotheses is true and the other is false since they are necessarily derived from different auxiliary assumptions.

For instance, the more complicated process of observing Y may introduce errors of measurement that are absent when X is observed. The hypothesis that the density of Y is $f(a/Y)$ is derived under the implicit assumption that no such measurement errors are introduced by observing Y instead of X, an assumption which might be false.

But let us grant the truth of all auxiliary assumptions. The situation is nevertheless different from Neyman's Paradox. In the case of Neyman's Paradox, we assume that measurements of X and Y result in real values. We then have to *decide* how to *interpret* these results. We do not gain anything by combining both measurements: we can compute one from the other.

With finite precision, the situation is quite different. If we can subject the same person to both measurement processes, we get a more precise result in

all those cases where the observed X interval and the observed Y interval are not mapped onto each other by taking inverse values.

Assume that, in our example, the realization of X is found to be in the interval [110, 111]. Measuring Y with the same precision, we find an interval [270, 271]. Combining the information, we learn that the true height is in the interval [110.70, 111]. Hence, we can test a discrete hypothesis that gives the probabilities for the smaller intervals for the height of a person ascertainable by combining both measurements.

Thus, the cases where the results of the tests differ are just the cases where we get more precise results from using both methods of measurement. The combined measurement obviously allows for a more severe test than any of the single measurements.

Of course, if we have to decide between the two experiments because only one of them is feasible, it may depend on our decision whether we get a falsification or a corroboration. But this is only reasonable. The combined experiment delivers maximum precision. If there are several second-best methods of measurement, it is plausible that there are cases where these measurements taken in isolation come to different results. This is not different from testing a deterministic hypothesis. If you want to test the hypothesis that swans are white, and if you can only afford a trip either to Germany or to New Zealand, it depends on your decision whether you falsify or corroborate the hypothesis.

There is only one strange feature of the statistical setting: we are better off (in that we can avoid Neyman's Paradox) because we know less (namely, intervals instead of exact values). But since no problem results from this, we should be grateful. The working properties of the tests are not affected. There is no ambiguity. Thus, Neyman's Paradox does not exist after all if one accepts the very reasonable adequacy condition for the scope of an FRSI. It comes up only in thought experiments where one confronts hypotheses with actually unobservable events.

Acknowledgements

For useful hints and discussions, I am grateful to Donald Gillies and the participants, especially Joseph Berkovitz, Jeremy Butterfield and Michael Redhead, of a Cambridge University Σ Club session.

Department of Economics
University of Koblenz–Landau
August-Croissant-Str. 5
D-76829 Landau, Germany
albert@uni-landau.de

References

Albert, M. [1992]: 'Die Falsifikation statistischer Hypothesen', *Journal for General Philosophy of Science*, **23**, pp. 1–32.

Berger, J. O. and Berry, D. A. [1988]: 'The Relevance of Stopping Rules in Statistical Inference (with discussion)', in S. S. Gupta and J. O. Berger (*eds*), 1988, *Statistical Decision Theory and Related Topics IV*, Vol. I, New York: Springer, pp. 29–72.

Fisher, R. A. [1990]: *Statistical Methods, Experimental Design and Scientific Inference*, Oxford: Oxford University Press.

Forster, M. R. [1995]: 'Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation', *British Journal for the Philosophy of Science*, **46**, pp. 399–424.

Gillies, D. A. [1971]: 'A Falsifying Rule for Probability Statements', *British Journal for the Philosophy of Science*, **22**, pp. 231–61.

Gillies, D. A. [1973]: *An Objective Theory of Probability*, London: Methuen.

Gillies, D. A. [forthcoming]: 'Bayesianism and the Fixity of the Theoretical Framework', in D. Corfield and J. Williamson (*eds*), *Foundations of Bayesianism*, Kluwer.

Hacking, I. [1965]: *Logic of Statistical Inference*, Cambridge: Cambridge University Press.

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

Mayo, D. G., and Spanos, A. [2000]: 'A Post-Data Interpretation of Neyman–Pearson Methods Based on a Conception of Severe Testing', unpublished, Blacksburgh, VA: Virginia Polytechnic Institute and State University.

Musgrave, A. [1974]: 'Logical versus Historical Theories of Confirmation', *British Journal for the Philosophy of Science*, **25**, pp. 1–23.

Neyman, J. [1952]: *Lectures and Conferences on Mathematical Statistics*, Washington: Graduate School of the U.S. Department of Agriculture.

Neyman, J. [1965]: 'Behavioristic Points of View on Mathematical Statistics', in (no *ed.*), 1965, *On Political Economy and Econometrics—Essays in Honor of Oskar Lange*, Oxford: Pergamon Press, pp. 445–62.

Neyman, J., and Pearson, E. S. [1930]: 'On the Problem of Two Samples', in J. Neyman and E. S. Pearson, 1967, *Joint Statistical Papers*, Cambridge: Cambridge University Press, pp. 99–115.

Neyman, J., and Pearson, E. S. [1933]: 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', in J. Neyman and E. S. Pearson, 1967, *Joint Statistical Papers*, Cambridge: Cambridge University Press, pp. 140–85.

Popper, K. R. [1984]: *Die Logik der Forschung*, 8th impr. and enh. edn, Tübingen: Mohr (Siebeck).

Redhead, M. L. G. [1974]: 'On Neyman's Paradox and the Theory of Statistical Tests', *British Journal for the Philosophy of Science*, **25**, pp. 265–71.

Spielman, S. [1974]: 'The Logic of Tests of Significance', *Philosophy of Science*, **41**, pp. 211–26.