On the Likelihood Principle and a Supposed Antinomy

Barry Loewer, Robert Laddaga, and Roger Rosenkrantz

It is natural to regard a given datum, x, as favoring that one of two rival hypotheses, H or K, which affords x higher probability, so that H is favored iff the likelihood ratio: $P(x|H):P(x|K)$ exceeds one. More generally, the relative support x accords the different members of a partition of hypotheses, $H_1,\ldots,H_n$, is assessed by comparing their likelihoods, $P(x|H_i)$. Considered as a function of the H_i, $L(H_i) = P(x|H_i)$ is called the likelihood function. If we view the likelihood function associated with an observed experimental outcome, x, as conveying the entire evidential import of the experiment, then it follows that the two outcomes, x and y, whether of the same or different experiments, are evidentially equivalent just in case they give rise to the same relative likelihoods, $P(x|H_i):P(x|H_j) = P(y|H_i):P(y|H_j)$ for all i,j, or , in other words, iff $P(x|H_i) \propto P(y|H_i)$ for all i. Writing Ev(E,x) for the evidential import of outcome x of experiment E, we may formulate this as the LIKELIHOOD PRINCIPLE: Ev(E′,x′) = Ev(E″,x″) iff x′ and x″ give rise to proportional likelihood functions.

The Likelihood Principle is an immediate consequence of Bayes′ Rule for updating the probabilities of hypotheses. For that rule tells us that the posterior probability of H_i is proportional to the prior probability, $P(H_i)$, and the likelihood, $P(x|H_i)$, hence, if two experimenters assign the same prior probabilities to the H_i, and the outcomes they observe give rise to proportional likelihoods, they will obtain proportional posterior probabilities, $P(H_i|x) = kP(H_i|y)$, for all i. Since the posterior probabilities must continue to sum to one, k = 1, and the posterior probabilties of the two experimenters are the same.

Orthodox (non-Bayesian) statistical procedures are easily seen to violate the Likelihood Principle. Consider the following example. Two orthodox statisticians, Moe and Joe, are interested in the lot proportion defective in a batch of manufactured articles. Moe elects to sample nine items and finds two defectives. Joe decides to sample until he finds two defectives, but, by coincidence, finds his second defective

---

PSA 1978, Volume 1, pp. 279-286

on the ninth trial (where sampling is with replacement). Their likelihood functions are, respectively, binomial and negative binomial, viz., $36p^2(1-p)^7$ for Moe, and $8p^2(1-p)^7$ for Joe. And since these are proportional, the Likelihood Principle qualifies their results as equivalent. (Had they started out with the same prior distribution of p, they would be led to the same posterior distribution.) But now suppose that they are interested in testing the hypothesis H that $p = 0.5$ against the alternative K that $p = 0.1$, using a test whose probability of erroneously rejecting H (the so-called 'type I error probability') is $< .05$. As you can easily check, Moe's test will reject H when x, the observed number of defectives, is less than two, while Joe's test will reject H when the second defective occurs on or after the ninth trial. Hence, Moe's test accepts H while Joe's test rejects H. Their different acceptance decisions are traceable to the fact that there are more ways, viz., 36, of obtaining two defectives in nine Bernoulli trials than there are ways, viz., eight, of obtaining the second defective on the ninth trial. But that consideration seems irrelevant, and all the more so if Moe and Joe happen to observe the same sequence of trial outcomes, say one in which the two defectives occur on the first and ninth trials. Only their intentions when to stop sampling would then be different, and, as Edwards, Lindman and Savage insist, "the intentions of the experimenter are irrelevant to the interpretation of the data once collected, though of course they are crucial to the design of experiments."([7], p. 565).

Another sort of violation might occur in our example. Suppose that Moe wants a test whose type I error probability, α is exactly .05 (a so-called exact 5% test). If he rejects H when $x < 2$, $\alpha = .0195$, while $\alpha = .0898$ if he rejects H when $x < 3$. If, however, he rejects H when $x < 2$ and with probability .4339 when $x = 2$, then $\alpha = P(x<2|H) + .4339P(x=2|H) = .05$. This trick is called randomizing at the boundary, and it clearly violates the Likelihood Principle, for, in this example, it makes the import of $x = 2$ depend on the outcome of an extraneous random experiment whose two outcomes have probabilities .4339 and $1 - .4339 = .5661$. Even many orthodox statisticians have balked at allowing the import of the observed outcome to depend on the outcome of an unrelated random device, like a coin flip. But there are stronger criticisms of this than the bald claim that it is counterintuitive.

Moe's willingness to randomize connotes his preference for the resulting error probabilities over those attainable without randomizing. He considers that $\alpha = .05$ provides just the right amount of protection against erroneous rejections of H, while $\alpha = .0898$ provides too little. Again, the test which rejects H when $x < 2$ provides even better protection against erroneous rejections (viz., .0195), but it purchases this greater security only by running a higher risk of erroneous acceptances of H (the so-called 'type II error'), and Moe considers this higher type II error probability too high a price to pay for the reduction of type I error probability to an unnecessarily low level.

In effect, statistical tests or rejection rules are being compared like commodity bundles, where the commodities in each bundle are the

type I and type II error probabilities. They can be considered as points in a Cartesian plane, where the coordinates of a point are its two error probabilities, usually denoted α and β. The ordinary rationality assumptions should govern preferences among these commodity bundles. But suppose that Moe prefers an exact test, Q, to tests P′ and P″, for reasons of the sort sketched above. Then the usual mixture assumption states that P′ should be preferred (respectively dispreferred) to any mixture, kP′+(1-k)P″, of P′ and P″ if P′ is preferred (respectively dispreferred) to P″. (A 'mixture', kP′+(1-k)P″, refers to a rejection rule which applies P′ or P″ with probabilities k and 1-k, $0<k<1$.) Either way, by transitivity of preference (a normative assumption!), Q should be preferred to any mixture of P′ and P″. But it may happen that Q lies above the line joining P′ to P″ (figure 1). In that case, the line from the origin to Q will intersect P′P″ in a point P whose coordinates are both smaller than the corresponding coordinate of Q. P then has smaller type I error probability than Q and smaller type II error probability than Q, and so, the Principle of Admissibility requires that P be preferred to Q, contradicting what we found above.

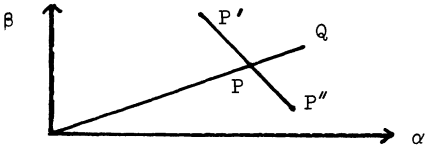

FIGURE 1.

Allan Birnbaum [2] proposes to meet this objection by rejecting the decision theoretic interpretation of a test in favor of an evidential interpretation. That is, he views the proper output of a test as a statement of the form, (Reject H for K, α , β), where the indicated error probabilities provide an informal index of the strength of the evidence against H and in favor of K. Birnbaum then argues that mixing assumptions fail when tests are evidentially interpreted. In testing whether a gene lies on chromosome 1 or chromosome 2, for example, I am interested in decisive results without regard to which hypothesis is favored. In that case, I might strongly prefer a test with error probabilities (.05,.05) to a test with error probabilities (.1,0), and, given the symmetry of the situation, the latter might be indifferent to a test with error probabilities (0,.1). "In such situations", Birnbaum writes, "I particularly value the guarantee, which is provided by use of (.05,.05), that strong evidence will be obtained." ([2], pp. 37 ff.). This pattern of preferences is clearly incompatible with the mixing assumption. For if the tests (.1,0) and (0,.1) are indifferent, either should be indifferent to a 50-50 mixture of the two, and hence, the latter should be dispreferred to (.05,.05). Yet, the error probabilities associated with a 50-50 mixture of (.1,0) and (0,.1) are (.05,.05), and consequently, equating a test with its error probabilities would lead to

the absurdity that (.05,.05) is dispreferred to itself. On the evidential interpretation, then, both the mixing assumption and the identification of a test with its error probabilities fail. Two tests with the same error probabilities, even when they lead to the same evidence statements of the form, (Reject H for K, α , β), may be indicative of quite disparate degrees of evidential support for an accepted hypothesis. This point is further accentuated by examples of the following sort.

We are given two normal populations with the same (unknown) mean θ , but quite different standard deviations, $\sigma_1 >> \sigma_2$. A fair coin is flipped, and we sample one population or the other according as it lands heads or tails. We wish to test the null hypothesis $\theta = 0$ against the simple alternative $\theta = \theta' \doteq \sigma_1$, at, say, the 5% level of significance. It is assumed that we know which of the two populations it is we are sampling. Now a conditional test bases calculations of the error probabilities on the particular population known to have been sampled, and thus rejects the null hypothesis if $x > 1.64\ \sigma_1$ when the first population is sampled, and rejects it if $x > 1.64\ \sigma_2$ when the second population is sampled. However, the test which rejects when $x > 1.28\ \sigma_1$, if the first population is sampled, $x > 5\ \sigma_2$, if the second population is sampled, is readily seen to have the same type I error probability and a smaller type II error probability.

The example is due to D.R. Cox, who writes: "...if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order... . If, however, our object is to say `what we can learn from the data we have', the unconditional test is surely no good. Suppose we have an observation from (θ , σ_1). The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution with variance σ_1^2." ([5], p.360-361).

Cox's intuition about his example can be formulated as the ANCILLARITY PRINCIPLE: $Ev(E,(E',x')) = Ev(E',x')$ and $Ev(E,(E'',x'')) = Ev(E'',x'')$, when E is a mixture of E' and E''. Here we write (E,(E',x')) when the mixed experiment leads to performance of E' and the latter issues in outcome x'. The Ancillarity Principle was introduced by R.A. Fisher, an outspoken and often acerbic critic of the decision theoretic interpretation of a test, and drives a wedge between that interpretation and the evidential interpretation which Fisher [8] favored. Birnbaum has shown [3] that, together with another almost universally accepted principle (sufficiency), the Ancillarity Principle implies the Likelihood Principle. Given that the Ancillarity Principle is virtually what distinguishes the evidential from the decision theoretic interpretation of an orthodox test, Birnbaum's derivation can be taken (and certainly was taken) to have shown that orthodox statisticians who espouse the evidential interpretation are, in essence, committed to the Likelihood Principle. And espousal of that principle

is nearly an espousal of the Bayesian approach. But not quite. One can regard the likelihood function as conveying the entire import of the evidence and still eschew the use of prior probabilities. That position on the board was occupied by R.A. Fisher [8], later by George Barnard [1], and more recently by Ian Hacking [9], A.W.F. Edwards [6] and others.

Given his derivation of the Likelihood Principle and his espousal of the evidential interpretation, Allan Birnbaum became identified as one of the leading spokesmen of this `likelihood approach'. But, suprisingly, Birnbaum did not favor this approach and soon said so in print [4]. Evidently, he took the Likelihood Principle to be incompatible with a requirement he regarded as even more fundamental, namely, the requirement that an adequate concept of statistical evidence interpret given data as being strongly opposed to a true hypothesis (and favorable to a false hypothesis) with low probability. He regarded this clash of fundamental principles as a genuine antinomy, and indicative of a real crisis in the foundations of statistics. The main burden of this paper is to argue that Birnbaum's alleged counterexample to the Likelihood Principle is unconvincing. We begin with an example that is rather transparently not a genuine counterexample. It will then be seen that our example is like Birnbaum's in all essentials.

A deck of cards is either normal (H) or anomalous (K), where all 52 cards of an anomalous deck are identical. A card is drawn at random and proves to be an ace of spades. This outcome has probabilities 1/52 on H and 1 on the hypothesis that all 52 cards in the deck are aces of spades, indicating strong evidence in favor of the latter hypothesis over all others, and (paralleling Birnbaum's analysis) this carries over to evidence that the deck is anaomalous rather than normal. But strong evidence of anomaly would be obtained, on a likelihood criterion, whatever card we drew from the deck, even if the deck were in fact normal. This leads Birnbaum to conclude that "the likelihood concept cannot be construed so as to allow useful appraisal and thereby possible control, of probabilities of erroneous interpretations." ([4], p. 128).

Birnbaum's very similar example involves testing the hypothesis H' that a random variable has mean μ and zero variance against the hypothesis K' that it has mean μ and a much larger variance, where the mean is unknown. Our initial reaction to his example was to doubt whether anything could be learned about the variance on the basis of a single measurement when the mean is unknown, and to distinguish misleading (or uninformative) data from misleading interpretations of non-misleading data. However, a better answer can be given by noting that the analysis of the card example is defective from a Bayesian point of view. For K is a composite hypothesis, one with 52 simple constituents. To adequately assess its evidential support, we must average the likelihoods of its 52 special cases (viz., all cards are aces of spades, all cards are aces of clubs, etc.), and the result, assuming all 52 cases to be equiprobable is, of course, 1/52. For `all cards are aces of spades' has likelihood one, while the other 51 simple constituents of K have likelihood zero. Hence, both H (a simple

hypothesis) and K have the same average likelihood, and we see that, far from providing strong evidence against H in favor of K, the drawing of a single card is completely inconclusive. This both accords with and explains our intuition that no light can be shed upon the question whether the deck is normal or anomalous by drawing just a single card. If two cards were drawn (without replacement) and both proved to be the same (say, aces of hearts), the average likelihood of K would still be 1/52, but the likelihood of H would decrease to $(1/52)^2$, providing fairly strong evidence in favor of K.

The use of average likelihoods has an obvious Bayesian rationale. If H is a composite hypothesis with simple constituents, h_i, $i = 1,\ldots,n$, then the prior (respectively posterior) probability of H is the sum of the prior (respectively posterior) probabilities of the h_i. To obtain the posterior probability of H, therefore, we sum those of the h_i, and the latter are proportional to $P(x|h_i)P(h_i)$. But if all h_i have equal prior probabilities, $P(H|x)$ is proportional to $P(h_i)\ \Sigma\ P(x|h_i) = P(H)\ \Sigma\ P(x|h_i)/n$ so that $P(H|x)$ is proportional to the product of $P(H)$ by the average likelihood, $\Sigma\ P(x|h_i)/n$. The same justification applies to a composite hypothesis with continuum many simple constituents (i.e., to probability models or distributions with one or more continuous parameters), the summations giving way to integrals. This case is needed for Birnbaum's example ([4], pp. 127-128).

The mean of a random variable, X, is known to lie between 10^{-10} and 10^{10}. The standard deviation, σ, of X, is either zero or quite large. Specifically, let H assert that the distribution of X has mean μ and zero variance, and let K assert that X has a wide triangular distribution centered at μ, where μ is unknown. (The triangle must, of course, have unit area for this to be a probability distribution.) Given a single measurement x of X, the parameter values $\mu = x$, $\sigma = 0$ have likelihood one, while the likelihoods of the simple constituents of K vary from 0 to the height of the triangle, which is quite small (e.g., it is .01 when the base of the triangle has length 200). Consequently, Birnbaum contends, x provides strong evidence in favor of $\mu = x$ and $\sigma = 0$, and he says: "This includes, in particular, that the value $\sigma = 0$ seems to be supported against the alternative $\sigma = 100$. Evidently the numerical values of the likelihood ratios referred to, all exceeding 100, represent strong evidence in some sense." ([4], p. 128).

The example is parallel to our card example and lends itself to a similar analysis in terms of average likelihood. Here, both H and K are composite hypotheses (with parameter μ). Let us compute and compare their average likelihoods.

We obtain the special cases of K by sliding the unit triangle in question along the x-axis between 10^{-10} and 10^{10}, μ being the midpoint of the base. The likelihood of each member of this family of triangles is 0 if x lies outside the triangle and is otherwise equal to the height of the triangle at the abscissa x. Thus, the likelihood function over the constituents of K is zero everywhere in the interval, except along

the base of a triangle centered at x, where it is given by the two sides of the triangle. The average height of this curve (i.e., the average likelihood of K) is equal to the area under the curve divided by the length of the interval, or $1/(2 \times 10^{10})$ since the triangle has unit area.

To compute the average likelihood of H, consider first a less extreme (more realistic) hypothesis, H′, which assigns X a very narrow triangular distribution centered at its unknown mean, μ . However wide or narrow, this triangle must have unit area to qualify as a distribution, and so the above analysis shows its average likelihood to be the same as that of K. H itself is best thought of as a limiting triangular distribution, letting the base shrink to zero and the height climb to one. Its average likelihood must then be the limit of the average likelihoods of these ever narrower triangular distributions, hence, it, too, must be the same as that of K. (This limit argument is the only plausible way of assigning an average likelihood to H.) Again we conclude that, far from providing strong evidence for H against K, a single measurement is completely equivocal, and, in particular, it tells us nothing about the variance. (As in the card example, two or more observations may, of course, tell us a good deal.)

We conclude with one or two brief remarks about the evidential interpretation. As Lindley [10] and Pratt [11] point out in their discussions of Birnbaum′s recent paper, evidence statements of the form Birnbaum favored can be quite unrevealing regarding the strength of the evidence. For example, if the observation lies midway between two normal means, μ_1 and μ_2, where both populations have unit variances, σ_1 and σ_2, the usual test, which rejects $\mu = \mu_1$ when $X > \mu_1 + 1.64$, would lead to the evidence statement, (Reject H for K, .05,.05), and yet, the observation is quite equivocal ([11], pp. 63-4). Similarly, let H assert that X is uniformly distributed in the interval [0,1], and let K assert that X is uniform in [0.9,1.9]. Let X = 0.97 be observed, again, a very equivocal outcome, since it falls in the overlap of the two intervals. But, again, it leads to the (presumably) strong evidence statement, (Reject H for K,.05,.05), using the test which rejects H iff X > 0.95. Moreover, the very same evidence statement is forthcoming if X = 1.01 is observed, but this outcome is entirely decisive. In short, use of such evidence statements or of error probabilities per se, fails to distinguish evidence of very different strengths. This criticism, together with others we could lodge, suggests that the evidential interpretation requires a good deal of unpacking before it can serve as a very firm foundation for statistical practice. Lacking any strong intuitions about how to go about this task, we are more inclined to believe that the Likelihood Principle captures what is sound in the evidential interpretation, and that Birnbaum′s supposed antinomy disappears upon closer inspection.

References

[1] Barnard, G. A., Jenkins, G. M., and Winsten, C. B. "Likelihood inference and time series." Journal of the Royal Statistical Society 125(1962):321-327.

[2] Birnbaum, A. "The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory." Synthese 36(1977): 19-50.

[3] ----------. "On the foundations of statistical inference." Journal of the American Statistical Association 57(1962): 269-306 (with discussion).

[4] ----------. "Concepts of statistical evidence." In Philosophy, Science and Method. Edited by S. Morgenbesser, P. Suppes, and M. White. New York: St. Martin's Press, 1969. Pages 112-143.

[5] Cox, D.R. "Some problems connected with statistical inference." Annals of Mathematcal Statistics 29(1958): 357-372.

[6] Edwards, A. W. F. Likelihood. Cambridge: Cambridge University Press, 1972.

[7] Edwards, W., Lindman, H., and Savage, L.J. "Bayesian statistical inference for psychological research." Psychological Review 70(1963): 193-242.

[8] Fisher, R. A. Statistical Methods and Scientific Inference. Edinburgh: Oliver and Boyd, 1956.

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

[10] Lindley, D.V. "The distinction between inference and decision." Synthese 36(1977): 51-58.

[11] Pratt, J.W. "`Decisions' as statistical evidence and Birnbaum's `confidence concept'." Synthese 36(1977): 59-70.