JOHN W. PRATT

'DECISIONS' AS STATISTICAL EVIDENCE AND BIRNBAUM'S 'CONFIDENCE CONCEPT'

1. INTRODUCTION

The statistical profession has suffered grievously in the loss of one whose pursuit of the true purposes of statistical inference was arduous, penetrating, persistent, illuminating, and unflinchingly honest. The comments to follow were prompted by Birnbaum (1977), and references are to this paper unless otherwise specified, but I will try not to be unfair to Birnbaum's thought as it evolved. My discussion has two principal parts. The first concerns the consequences of accepting the view urged by Birnbaum that the interpretation of 'decisions' which is relevant to scientific reporting and typical standard practice is evidential rather than behavioral. The second concerns the consequences of including power as part of the evidential interpretation of a statistical test, as in Birnbaum's examples of the 'confidence concept' of statistical evidence. These are followed by a short *reductio ad absurdum* and a brief summary of my conclusions.

2 STATISTICAL EVIDENCE

One version of history, by a non-historian, is this. In the 1920's, the longstanding inadequacy and obscurity of the foundations of the theory of statistical inference was more than ever found troubling and was beginning to be a serious barrier to progress otherwise ready to be made. The Neyman-Pearson theory introduced a formalization or paradigm which greatly clarified the situation and appeared to improve it. It provided a basis for the tremendous advances in statistical methodology which ensued. It also led to decision theory of the Wald type. This, in turn, merged with other streams of thought and emerged as Bayesian decision theory.

As a theory of individual decision-making, Bayesian decision theory has no real competition. It is thus an appropriate point of reference or

Synthese **36** (1977) 59–69. *All Rights Reserved.*

comparison for other, more practical or popular methods and for methods dealing with other but related or similar problems.

The theory of inference which has grown up in the Neyman-Pearson tradition, however, despite extensive development at all levels and widespread dissemination and acceptance, has remained incompletely defined and turns out only to have pushed the fundamental difficulties further into the background without really solving them.

Most of this is not in contradiction to Birnbaum, and I agree with much that he has said. In particular, I welcome his emphasis that there is a problem of inference, of evidential interpretation, distinct from decision making. The high road to salvation by no means follows, however, as he knew.

Since inferences should be able to be used in decision making, the more readily a theory of inference relates to the Bayesian theory of individual decisions the better, other things being equal. Where practical, this makes Bayesian inference the obvious choice – Bayesian inference with due respect for the variety of prior distributions and models the ultimate decision-makers might contemplate and for the dangers of improper priors. Some standard statistical methods are compatible with Bayesian decision theory in the sense that any report from which sufficient statistics can be recovered is compatible with it, but they do not relate at all readily except by way of mathematical (not conceptual) identities and approximations (Pratt, 1965). Birnbaum develops them here in a direction if anything still further from a ready relationship, important as it may be to recognize that α is not everything.

That is my view, through Bayes-colored glasses. With history in mind, however interpreted, suppose we accept Birnbaum's reasoning that the Lindley-Savage argument is not directly applicable to standard practice – what then? Lindley and Savage worked within the formalization chosen by Neyman and Pearson and showed that it led to Bayesianism, contrary to Neyman and Pearson's intention. Essentially, Birnbaum has questioned the formalization. If this makes the Lindley-Savage argument irrelevant, at the same time it undermines the Neyman-Pearson position. It does not undermine the Bayesian position, which has many supports.

As a basis for much of standard practice and thinking, the Neyman-Pearson formalization is the best and perhaps the only widely accepted one providing a framework and criteria definite enough to bring some order out of chaos. Rejecting it returns us to the chaos of the 20's – with a deeper understanding, to be sure, but more of difficulties than of resolutions. Perhaps it can be carefully circumscribed somehow to retain some usefulness without leading to Bayesianism. When methods are prevented from falling by such feats of balance, however, I am filled more with admiration than with confidence, especially if the methods are for general use. (A similar feat is accomplished by those who profess happiness with empty confidence intervals because confidence intervals are not to have confidence in but to specify parameter values which would be acceptable if tested.) Furthermore, there is no reason to suppose that a balancing act will remove the difficulties.

At the symptomatic level, these difficulties manifest themselves in questions of choice of test statistic and of the proper conditioning to use in citing significance or confidence levels, and in a variety of anomalies and paradoxes which I won't rehearse here (see, e.g., Pratt, 1961). They will surely not disappear, though they may be obscured, if the only clear, generally accepted framework in which to discuss them is rejected or made unavailable by circumscription. Birnbaum himself repeatedly calls the procedures 'ad hoc' (§ 7, pp. 34–35; note also top of p. 36) and likens his usage of the first person to Bayesian usage (p. 24).

This implies a recognition that the problem is more fundamental than interpreting tests as evidence and explaining away their anomalies. Why is an 'objective' probability of 'rejection' so vital if we are not really rejecting anything (not really enough to mix probabilities)? Why not report probabilities of observed values rather than tails? They too are objective. They even provide a 'plausible' concept of inference in Birnbaum's sense for simple hypotheses (since the probability under H_1 of a likelihood ratio in favor of H_2 of k or more is at most $1/k$), but not for composite hypotheses (Birnbaum, 1969). But Birnbaum's definition of plausibility already assumes far too much. The problem is how evidence should be expressed. Why should it be in anything like the traditional formats? If even the proper posing of the question is wide open, why should the standard answer be right?

At this more fundamental level, Birnbaum gave an argument in 1962 which was carefully couched in terms of evidence, avoiding assumptions implicit in terms such as 'decision' or 'rejecting hypotheses.' This argument shows very convincingly that inferences should depend only on the likelihood function obtained, and not on what else might have happened. (In the discussion, I described and illustrated a direct argument which also applies in an evidential framework.) This is seriously at variance with the thinking behind standard practice. Thus there is a substantial theoretical deficiency (of some practical importance, too) in standard methods, in any interpretation, including Birnbaum's. Birnbaum (1970) rejected these arguments, but unfortunately he never explained why, as far as I know, and they are as convincing as ever to me.

3. BIRNBAUM'S INTERPRETATION OF TESTS

Birnbaum used the term 'confidence concept' extensively. I particularly regret that he did not live to clarify it further, even in the absence of "any precise systematic theory of statistical inference" (p. 23). I am bothered and hampered by uncertainty about what beyond standard ideas he intended to convey or support. The 'confidence concept' refers *inter alia* to evidential, but not behavioral, interpretations of confidence limits (1970; 1977 pp. 35–36) but is illustrated here primarily by a sophisticated form of presentation of the outcome of a test. I am unsure of his attitude even toward this illustration. Nevertheless, I will raise two questions and make some comments about this manner of reporting and interpreting tests, while recognizing that I may not be speaking directly to Birnbaum's ideas or views.

Even for simple hypotheses, the question arises whether the tail probabilities cited (as on pp. 24–25) are to correspond to the particular data observed, or are to be fixed in advance with only 'accept' or 'reject' determined by the data. The latter seems to me clearly a very inadequate expression of the evidence. (The Principle of Adequacy: a concept of statistical evidence is (very) inadequate if it does not distinguish evidence of (very) different strengths.) This accords with the view that a P-value (critical level) is preferable to a report of 'significant' or 'not significant' in usual current practice where only tail probabilities under the null hypothesis are seriously considered in the final analysis. Dichotomous reports seem less satisfactory the less behaviorally and more evidentially the 'decisions' are interpreted. Incidentally, if tail probabilities corresponding to the data are to be cited, then mixtures of experiments are even more complicated than Birnbaum had occasion to mention, since the (α, β) pairs not only don't represent decision functions, they don't even represent experiments, being merely one possible outcome of an experiment.

For composite hypotheses, another question arises. It is usual to think of the Type II error β (or the power) as a function defined over the alternative hypothesis, but the size α as a single number, the maximum probability of rejection under the null hypothesis. The question is whether α, too, should not be reported as a function, as the principle of adequacy clearly suggests. The question is applicable whether α is preselected or corresponds to the data. If the null hypothesis is rejected at the level 0.1, the evidence against it is much weaker when the probability of rejection is 0.1 for all null distributions than when it is far below 0.1 for most null distributions. Note that, while it seems important for the completeness of the evidence to report the whole function α or β if the null or alternative hypothesis is composite, it also makes it hard to derive much evidential *interpretation* from the report. For example, in typical applications, the values of the function β would cover the entire range from 0 to 1 minus the significance level or P-value, at least.

When we go deeper into the evidential interpretation of tests, some surprises appear. They are especially clear in Birnbaum's format when observed tail probabilities are cited. I will therefore discuss this case primarily, later indicating briefly how similar points apply to more ambiguous expressions of evidence employing preselected α and β or α without β. I will assume throughout that H_1 and H_2 are simple hypotheses and that the test statistic and the likelihood ratio obtained from it are continuously distributed under each. The complications of composite hypotheses and discrete probabilities would be an irrelevant distraction here.

Consider first an outcome which is barely significant at $\alpha = 0.05$ with $\beta = 0.05$, expressible in Birnbaum's format as (reject H_1 for H_2, 0.05, 0.05). Note that these are observed tail probabilities – if the outcome might be more extreme, then the full evidence has not yet been given. Such an outcome is equally discordant with H_1 and with H_2 (insofar as tail probabilities measure discordance, which has to be the assumption here). In a symmetric situation, this would be an outcome exactly half way between H_1 and H_2, giving the same evidence on H_2 versus H_1 as on H_1 versus H_2. An example would be an observation of 1.645 on a random variable which is $N(0, 1)$ under H_1 and $N(3.290, 1)$ under H_2, where $N(M, 1)$ means normal with mean M and variance 1. In such symmetric situations (it is symmetry between H_1 and H_2 which matters, not symmetry of the individual distributions), such an outcome clearly stands in the same relationship to (H_2, H_1) as to (H_1, H_2) by any criterion, not merely by tail probabilities.

The same would be true, of course, of the outcome (reject H_1 for H_2, ε, ε) for $\varepsilon = 0.01$ or 0.001 or any other value. Thus a small α and β do not necessarily imply strong evidence, and an outcome with $\alpha = \beta$ (equal observed tail probabilities) is highly equivocal. Thus one thing we must remember is that small α and β, though they indicate a very powerful experiment, nevertheless may correspond to a very equivocal outcome. If we test $N(0, 1)$ versus $N(5, 1)$ and observe an outcome of 2.5, then the experiment was very powerful but turned out by very bad luck to have been useless (for distinguishing these hypotheses). However good the prospect may have been, the outcome was a bust.

Another thing to remember is that (reject H_1 for H_2, α, β) is ordinarily stronger evidence for H_2 as against H_1 the *larger* β is (for $\beta < 0.5$). For example, if x is $N(0, 1)$ under H_1 and $N(\theta, 1)$ under H_2 and $x = 1.645$ is observed ($\alpha = 0.05$), then H_2 is more plausible if $\theta = 2$ ($\beta = 0.36$) than if $\theta = 4$ ($\beta = 0.009$). In general, since β is the tail probability under H_2, the smaller β is (below 0.5), the less compatible the observations are with H_2 and hence the weaker the evidence is for H_2 as against H_1, other things (α) being equal. Again, smaller β means more power for an experiment, but as an observed tail probability it means weaker evidence against H_1. From an *ex ante* viewpoint, if the experiment is powerful and H_2 is true, then ordinarily the outcome will correspond to a very small α and a moderate β (β being uniformly distributed between 0 and 1), and observing instead a very small β is relatively unfavorable to H_2.

A word on the case $\alpha = 0$. The continuity assumption above precludes it, but it arises in Birnbaum's examples. Clearly (reject H_1 for H_2, 0, β) is conclusive evidence against H_1 regardless of β. Thus, in this case, β has no effect on the strength of the evidence (though again, the β of an experiment with $\alpha = 0$ measures the chance under H_2 of obtaining conclusive evidence against H_1).

When α and β are observed tail probabilities, the evidence (reject H_1 for H_2, α, β) is, in fact, equivalent to (reject H_2 for H_1, α, β). Which should we say, if α and β are both small? Since Birnbaum's (1977) paper does not treat the two statements as equivalent, the α and β cited there must be preselected. We turn briefly to this case.

If α and β are preselected and a statistically significant outcome (reject H_1 for H_2, α, β) is obtained, then it is still true that small α and β do not necessarily imply strong evidence, and that the evidence against H_1 is stronger the *larger* β is. However, if $\alpha = \beta$, for example, we cannot say that the evidence favors H_1 and H_2 equally, but only that it favors H_1 at least equally and possibly very strongly. If we select α much less than β, then we can be sure that a statistically significant outcome (reject H_1 for H_2, α, β) is strong evidence for H_2 as against H_1, but there is no outcome which will give us evidence we can be sure is strong for H_1 as against H_2. (All this assumes that the tail probabilities measure evidence, on which more shortly.)

This illustrates the disadvantage of dichotomous reporting with respect to preselected α and β rather than reporting observed tail probabilities. Suppose, for example, that H_1 is $N(0, 1)$ and H_2 is $N(3, 1)$. If we choose $\alpha = \beta = 0.067$ (critical value 1.5), then even a report of statistical significance may, for all we know, have resulted from an outcome such as 1.6 favoring H_2 only slightly, so the report is sure to leave room for doubt whatever occurs. If we choose $\alpha = 0.01$, $\beta = 0.25$ (critical value 2.33), then a statistically significant outcome must strongly favor H_2, but an outcome such as -1, which strongly favors H_1, must be reported merely not significant, as must an almost significant outcome such as 2.2 ($P = 0.014$).

More usual practice is not to report β at all, but only the P-value or whether a preselected significance level was reached. Such a report is ambiguous whatever the outcome. If the outcome is significant at an extreme level, it is still possible that the experiment was extremely powerful and hence that $\beta < \alpha$, in which case the outcome favors the null hypothesis. If the ϙutcome is not significant even at a moderate level, it may favor the null hypothesis strongly or slightly or even favor the alternative hypothesis. One could argue that the outcome is no more favorable to the alternative than $(\alpha, 0.5)$, i.e., than the evidence (reject H_1 for $H_2, \alpha, 0.5$), but this doesn't take us far.

The relationship of the observed tail probabilities α and β to the likelihood ratio also casts light on the situation. In problems with two simple hypotheses, many people find the likelihood ratio meaningful either in itself or as one factor in the posterior odds, the other factor being the prior odds.

Consider, for instance, Birnbaum's first example, (reject H_1 for H_2, 0.06, 0.08), which he interprets as strong evidence for H_2 as against H_1. This would arise if the test statistic x is $N(0, 1)$ under H_1 and $N(2.960, 1)$ under H_2 and $x = 1.555$ is observed. Intuitively this seems close to the half-way point $x = 1.480$, raising a question how strong the evidence is. The likelihood ratio in favor of H_2 is only 1.25, so it is not strong at all. (If Birnbaum's 0.06 and 0.08 are preselected, we know only that the evidence favors H_2 to at least this extent, i.e., perhaps strongly, perhaps only slightly. It would seem highly objectionable to treat this as strong evidence for H_2 as against H_1 if in fact the observation exceeded 1.555 only a little.)

Conversely, consider a likelihood ratio of 2, say. This is hardly impressive, but for two simple normal hypotheses with common variance, it corresponds to the following observed (α, β) pairs: (0.12, 0.50), (0.10, 0.31), (0.05, 0.13), (0.01, 0.022), and for that matter, (0.10, 0.69), (0.05, 0.87), and (0.01, 0.978). Thus data which are just significant at $\alpha = 0.01$ with $\beta = 0.022$ are rather weak evidence for H_2 as against H_1, and no stronger (or weaker) than data which are just significant at $\alpha = 0.10$ with $\beta = 0.31$. Also, any data with $\alpha > 0.12$ are weaker evidence than this. However, no matter how small α may be, the evidence for H_2 as against H_1 may be this weak, or weaker, if β is small enough.

A likelihood ratio of 19 might appear to correspond to $\alpha = 0.05$. However, for two simple normal hypotheses with common variance it can only occur for $\alpha \leqslant 0.0076$, and it corresponds to $(\alpha, \beta) = (0.0076, 0.5)$, (0.005, 0.20 or 0.80), (0.001, 0.028 or 0.972), etc., while the likelihood ratio corresponding to $\alpha = 0.05$ cannot exceed 3.87.

Whatever the hypotheses (still assumed simple), it is familiar and easily proved that the (α, β)-curve of any admissible test is convex and hence that, on this curve,

$$(1) \qquad \alpha \leqslant \frac{\alpha}{1-\beta} \leqslant \frac{d\alpha}{d\beta} = \text{likelihood ratio} \leqslant \frac{1-\alpha}{\beta} \leqslant \frac{1}{\beta}.$$

It follows that, if the likelihood ratio is small, then α must be small, and if the likelihood ratio is large, then β must be small. The converse does not hold, however. In fact, the inner inequalities above are tight: given any α and β, the likelihood ratio can have any value between $\alpha/(1-\beta)$ and $(1-\alpha)/\beta$ for suitable hypotheses. Therefore a small α or β or both does not imply an extreme likelihood ratio. For instance, $\alpha = 0.06$ and $\beta = 0.08$ (Birnbaum's first example) could be associated with a likelihood ratio anywhere from 0.0652 to 11.75. If the likelihood ratio is accepted as measuring the extent to which the data favor one hypothesis or the other, then (α, β) is far from an expression of the evidence (although, of course, for any particular hypotheses, one can determine the likelihood ratio from the observed α, β, or from the observed α alone, for that matter). The foregoing refers to observed values of α and β. If they are preselected, then they convey even less of the evidence, in that the likelihood ratio has no upper bound if the observations are 'significant,' no lower bound if they are 'not significant.'

For composite hypotheses qualitatively similar ideas apply but quantitative results might be quite different. In particular, it is still generally true that "the more powerful the test, the more a just significant result favors the null hypothesis" (Pratt, 1961, p. 166).

4. REDUCTIO AD ABSURDUM

The following argument reduces to absurdity the idea that the evidence can be expressed by means of α and β alone. Since many will doubtless find good or bad reasons not to accept the argument, I give it separately and wish to emphasize that nothing else in this discussion in any way relies on it, though of course it is further confirmation of my misgivings.

Suppose a random variable x is observed whose mass function under H_1 and H_2 is given by

x	0	1	2
$f_1(x)$	0	p_1	$1-p_1$
$f_2(x)$	$1-p_2$	p_2	0.

Suppose $x = 1$ has been observed. Suppose that a random variable u has also been observed which is independent of x and uniformly distributed on $(0, 1)$. Clearly u supplies no evidence. However, the uniformly most powerful test at level α can be based on the test statistic $x+u$, and has

$$(2) \qquad (\alpha, \beta) = (tp_1, (1-t)p_2)$$

if the critical value is $1+t$, $0<t<1$. Therefore all (α, β) of this form supply the same evidence in this experiment.

If the same (α, β) supplies the same evidence whatever experiment gave rise to it, then it follows from the above that all (α, β) with $\alpha+\beta<1$ supply the same evidence. Specifically, given (α, β) and (α', β') it is easy to see geometrically that there exist $p_1, p_2, p_1', p_2', t, t', s, s'$, such that (2) holds, the same holds with primes, and

$$(3) \qquad (sp_1, (1-s)p_2) = (s'p_1', (1-s')p_2');$$

since (α, β) is equivalent evidence to the left side by the statement following (2) and (α', β') to the right side similarly, they are equivalent to each other.

5. CONCLUSION

To whatever extent the use of a behavioral, not an evidential, interpretation of decisions in the Lindley-Savage argument for Bayesian theory undermines its cogency as a criticism of typical standard practice, it also undermines the Neyman-Pearson theory as a support for typical standard practice. This leaves standard practice with far less theoretical support than Bayesian methods. It does nothing to resolve the anomalies and paradoxes of standard methods. (Similar statements apply to the common protestation that the models are not real anyway.) The appropriate interpretation of tests as evidence, if possible at all, is difficult and counterintuitive. Any attempt to support tests as more than rules of thumb is doomed to failure.

Graduate School of Business Administration, Harvard University

REFERENCES

Birnbaum, Allan: 1962, 'On The Foundations of Statistical Inference', *Journal of the American Statistical Association* **57**, 269–326 (with discussion).

Birnbaum, Allan: 1969, 'Concepts of Statistical Evidence', in *Philosophy Science, and Method: Essays in Honor of Ernest Nagel*, edited by Sidney Morgenbesser, Patrick Suppes, and Morton White. New York: St. Martin's Press.

Birnbaum, Allan: 1970, 'Statistical Methods in Scientific Inference' (Letter), *Nature* **225**, 1033 (with author's corrections not in publication).

Birnbaum, Allan: 1977, 'The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage argument for Bayesian Theory', *Synthese*, this issue, pp. 19–49.

Pratt, John W.: 1961, Review of *Testing Statistical Hypotheses* by E. L. Lehmann, *Journal of the American Statistical Association* **56**, 163–167.

Pratt, John W.: 1965, 'Bayesian Interpretation of Standard Inference Statements', *Journal of the Royal Statistical Society* Series B **27**, 169–203 (with discussion).