Criticism Against Einhorn and Hogarth's Theory of Diagnostic Inference

顏 乃 欣

（作者為本校心理系專任副教授）

摘 要

所謂診斷性推論，是指人們會依所發生的事件，來推測造成這些現象之原因的推測方式。艾宏及侯佳士（1982）認爲診斷性推論是一個主動的，強調因果關係的，由果推因的歷程，所關心的是對特殊事件而非一般性事件的推論。他們批評一般統計模式無法處理有這些特質的診斷性推論，所以提出一個新的描述性模式說明人們處理這類問題的歷程。事實上，他們對於一般統計模式的批評是不合理的，他們錯將所處理的貝氏問題以點估計問題視之，以至於提出很多錯誤的論點。同時他們提出的新模式雖然得到研究結果的支持，但在研究方法上卻有缺失之處。這篇文章主要就艾宏——侯佳士文中有關統計及研究方法方面提出批判。

ABSTRACT

The particular issue with which this paper is concerned is diagnostic inference. That is, given the occurrence of a set of outcomes/results/symptoms, one has to infer to what extent is a particular action or event responsible for the observed effects. Einhorn and Hogarth (1982) argued that the essential aspects of such inferences are that they are causal rather than correlational, backward rather than forward (one goes from effects to prior causes), concerned with specific rather than the general cases, and constructive rather than nonconstructive (one can synthesize, enlarge, or otherwise develop new hypotheses). They further argued that the most common statistical model (e.g., Peterson & Beach, 1967) involving inferences does not consider these four aspects, and they developed a new model to describe how people assess the likelihood that one of two hypotheses is true on the basis of varying amount of evidence for each. I shall show, however, that their claims against the usual statistical model are unfounded and that they, in fact, misconceive the type of statistical problem with which they are faced. They think they are dealing with point estimation problems, when, in fact, the diagnostic problems with which they are dealing are Bayesian problems. Furthermore, even though they concluded that their model fitted the data reasonably well, some methodological considerations provide questions about their conclusion. The main purpose of this paper is to critique Einhorn and Hogarth's arguments and model in statistical and methodological terms.

Criticism against Einhorn and Hogarth's theory of diagnostic inference

Einhorn and Hogarth (1982) proposed a theory of diagnostic inference which involves the assessment and generation of casual hypotheses to account for observed evidence. They define diagnostic inference as a constructive process that is causal, backward in its direction, and concerned with specific cases. A model was developed by them for describing how people assess varying amounts of evidence for each. They suggested that subjects might process information with an anchoring and adjustment strategy. Specifically, they anchor on the evidence observed and adjust on the basis of the imagined evidence that might have been. Several factors influence adjustment, e.g., the total amount of evidence at hand, attentional shifts due to rephrasing likelihood questions, the number and specificity of alternative hypotheses, and perception of missing evidence. They performed experiments to test their theory and model and concluded that the model fitted the data reasonably well. However, some statistical and methodological considerations provide questions about their model and conclusions.

First, Einhorn and Hogarth argued that the statistical model is a-causal, forward in direction of inference, and concerned with the general case, thus it is inappropriate as a model for diagnostic inference. As they state,

> "the statistical model does not formally consider causal ideas in its language. . . . statisticians do not encourage causal thinking, as for example in warning that correlation does not imply causation (although what *does* imply causation is never made clear). (p. 1)
>
> Statistical methods are greatly concerned with forecasting events or consequences and can thus be characterized as involving 'forward' inference. . . . We call inferences that are both backward and causal, 'diagnostic'. (p. 1)
>
> Statistical concepts such as average, variability, relative frequency, population, and so on, clearly indicate that the domain to which inferences are being made are aggregates of some sort. Therefore, one is concerned with the general case or with classes of cases. . . . as when considering one-of-a-kind events (such as the likelihood of Russian invasion of Poland), controversy exists regarding the meaning and meaningfulness of probability statements. Now consider the domain of inference of the lawyer, detective, or historian. Here one is concerned solely with the specific case – did Mr. X commit the crime? Is Mr. Y responsible for the accident? What were the causes of World War

> I? . . . the relevance of such evidence is often questioned on the grounds that the specific case is not a member of this or that class. (p.2)"

However, statisticians are often concerned with causal thinking. Such thinking is involved in the concept of experiment, for example, in particular, experimental control enables scientists to test for causal relations. Consider the direction of inference, statistical methods are concerned not only with forward inferences, but also backward inferences. In fact, statistics is directionless and timeless. The purpose of statistical method is to evaluate evidence regardless of whether the question is concerned with forward or backward inference. Furthermore, the claim with regard to the domain of statistical inferences is not necessarily true. In fact, there are examples that statistical analysis can be applied to inference which is causal, backward, and concerned with specific cases and the probability statements are meaningful for them. For example, Mosteller and Wallace (1972) tried to settle the disputed authorship of several of the Federalist papers. These papers were written by Alexander Hamilton, James Madison, and John Jay to try to convince New York State to ratify the U.S. constitution. For twelve of the papers there has been uncertainty whether the author was Hamilton or Madison. Mosteller and Wallace differentiated the two authors on stylistic features of the papers. It was found that a Bayesian analysis yielded overwhelming support for Madison's authorship of the disputed papers. Thus the study of Mosteller and Wallace provided a good example that statistical models are appropriate for dealing with inference which is causal, backward, and concerned with specific cases.

Einhorn and Hogarth proposed a model for the evaluation of the net strength of evidence. Suppose subjects are asked to assess the likelihood that one of two hypotheses is true. Consider that there are n equally strong pieces of evidence that consist of f favorable and c unfavorable arguments, where $n = f + c$. Moreover, let $p = f/n$, the proportion of favorable evidence. They proposed that one would evaluate the net strength of evidence on the basis of an anchoring and adjustment process. Specificially, it is assumed that in evaluating conflicting evidence, one first anchors on p, and adjusts to p by imagining a piece of favorable or unfavorable evidence. Einhorn and Hogarth suggested that several factors would affect the adjustment. One factor is the amount of evidence on hand, n. They proposed that people would anchor on p and adjust for n by imagining a worse case in which one f is shifted to one c. The model can be written as

$$Sn(f:c) = a_1(f/n) + a_2((f-1)/n) \qquad (1)$$

where $Sn(f{:}c)$ is the net strength of f and c pieces of evidence, a_1 is the weight for the anchor, and a_2 is the weight for the adjustment. Assuming $a_1 + a_2 = 1$, equation (1) can be rewritten as

$$\begin{aligned} Sn(f{:}c) &= a_1(f/n) + a_2((f-1)/n) \\ &= (a_1f + a_2f - a_2)/n \\ &= (f(a_1+a_2) - a_2)/n \\ &= p - a_2/n \qquad (2) \end{aligned}$$

Thus the model suggests a tradeoff between p and n such that one would accept less p for greater n. That is, the model predicts that as the total amount of evidence increases, the net strength approaches p as an asymptote with a rate determined by a_2.

However, when $p = 0$, the equation results in net strength being negative, which makes no sense. Thus they modified the equation when p is at, or close to zero. When $0 \leqslant p \leqslant P_c$, where P_c represents some small value of p, they proposed that people would adjust for n by imagining a favorable case instead of imagining an unfavorable case. Then the equation becomes

$$\begin{aligned} S\acute{n}(f{:}c) &= a_1(f/n) + a_2((f+1)/n) \\ &= p + a_2/n \qquad (3) \end{aligned}$$

Thus, when $p = 0$, $Sn(f{:}c) = a_2/n$. As $n \to \infty$, net strength approaches an asymptote of zero. And the general equation, which they called the evidence function, can be written as

$$Sn(f{:}c) = p + \beta(a_2/n) \qquad (4)$$

where $\beta = \begin{cases} 1 \text{ if } p \leqslant P_c \\ -1 \text{ if } p > P_c \end{cases}$

Einhorn and Hogarth considered several problems which varied in the values of n, f, and c to illustrate their theory and model. However, whether their arguments are supported in these problems is questionable. One example they gave is as follows

> "imagine that there has been a hit-and-run accident where f witnesses say the offending car was blue while c witnesses claim it was green. We are

> interested in the evaluation of the likelihood that a blue car caused the accident as a function of f, p, and n.·. . . for example, consider that there were only two witnesses who both said that the hit-and-run car was blue vs. a situation in which 9 witnesses said blue and one said green. Many people find the latter evidence stronger than the former in supporting the proposition that a blue car caused the accident. Why? We argued that when the total amount of evidence is meager, it is quite easy to imagine a different result by simply changing one piece of evidence. Thus, an outcome of (2:0) could easily be (1:1); or, (2:1) become (1:2) if only one witness changes his/her mind. (p. 9)"

Einhorn and Hogarth assert that subjectively evaluating the (9:1) case as stronger evidence than the (2:0) case for hypothesis 'blüe' is statistically improper. They do so because they see the problem, statistically, as one of point estimation. If one grants, for the purpose of argument, that the statistical problem is one of point estimation, their argument is shakey when one considers confidence intervals about the point estimates. Given the evidence that 9 witnesses said blue and one said green ($n = 10, p = .9$) vs. a situation in which only two witnesses both said blue ($n = 2, p = 1.0$), what is the 95% confidence interval for the estimate of the proportion of witnesses who will testify that a blue car caused the accident? For large samples, one can use observable proportion to estimate $\hat{\sigma}_P$. However, for small samples, one can get confidence intervals by reading tables which give confidence intervals for the probability of success in small samples (e.g., Hollander & Wolfe, 1973). Thus, for $n = 10$ and $p = .9$, the 95% confidence interval is the interval between .9975 and .5550. For $n = 2$, $p = 1.0$, the 95% confidence interval is the interval between 1.0000 and .1581. Thus, the 95% confidence interval for evidence (9:1) includes only values greater than .5 while that of evidence (2:0) includes values less that .5, which can be interpreted as indicating that the evidence (9:1) is stronger than the evidence (2:0) in supporting the proposition that a blue car caused the accident.

In fact, however, Einhorn and Hogarth mispercieve the problem. The problem presented is a Bayesian problem with incomplete information. According to the problem, there are two mutually exclusive and exhaustive hypotheses: a blue car caused the accident (H_B) or a green car caused the accident (H_G). One is asked to evaluate the likelihood of H_B given the observed evidence, that is, $P(H_B \mid D)$. However, the priors, $P(H_B)$ and $P(H_G)$, and the likelihood of each piece of evidence given that each hypothesis is true, $P(b \mid H_B)$, $P(g \mid H_B)$, $P(b \mid H_G)$, and $P(g \mid H_G)$, are not given in the problem. A way to assign probability under uncertainty is to use the maximum entropy rule (Jaynes, 1968; Rosenkrantz, 1983). Following this rule, priors are assigned in such a way as to be consistent with the constraints of the relevant variable, but beyond that they must maximize entropy. The maximum entropy priors can be derived by maximizing the entropy, or uncertainty, of the prior probability distribution $P(Hj) = -\sum_{j=1}^{a} P(Hj) \ln P(Hj)$ subject to the constraint that the sum of the probabilities must be unity $(\sum_{j=1}^{a} P(Hj) = 1)$. Thus, one should maximize the following system of equations with respect to the P(Hj):

$$F[P(Hj)] = -\sum_{j=1}^{a} P(Hj) \ln P(Hj) + \lambda \sum_{j=1}^{a} P(Hj)$$

where λ is a LaGrange mutiplier. It is nontrivial but easy to show that:

$$P(Hj) = e^{\lambda - 1}$$

It follows directly that, the maximum entropy priors for the case in which the only constraint is $\sum_{j=1}^{a} P(Hj) = 1$ are 1/a, where a is the number of mutually exclusive and exhaustive hypotheses in question.

In Einhorn and Hogarth's problem, a = 2. Therefore, the derived maximum entropy priors are $P(H_B) = 1/2$ and $P(H_G) = 1/2$. However, it is not appropriate to use the maximum entropy rule to get $P(b|H_B)$, $P(g|H_B)$, $P(b|H_G)$, and $P(g|H_G)$ for the cab problem discussed above. To use the maximum entropy rule would be equivalent to assuming that eyewitness testimony is not informative. But we usually believe that eyewitness testimony does give us some information. The impact of each piece of evidence given that each hypothesis is true is subjective and may depend on some variables such as the reliability of the witness (e.g., Schum, 1980), the dissimilarity between the hypotheses (e.g., Einhorn & Hogarth, 1985), etc.

If the problem is reduced to the symmetric binomial case, that is, $P(b|H_B) = P(g|H_G)$, $P(g|H_B) = P(b|H_G)$, and $P(b|H_B) + P(g|H_B) = P(b|H_G) + P(g|H_G) = 1$, diagnosticity is a function of the difference between the number of favorable and unfavorable witnesses. The difference of evidence (9:1) is 8 and the difference of evidence (2:0) is 2. Again, the evidence (9:1) is stronger than the evidence (2:0) in supporting the proposition that a blue car caused the accident.

Einhorn and Hogarth suggested that one of the implications of their model is that the addition of positive evidence has less effect on the net strength of evidence than the reduction of an equal amount of negative evidence when a majority or neutral position is evaluated (i.e., $2f \geqslant n$). As they state,

> " . . . compare the addition of one positive argument to make evidence of (3:2) into (4:2), vs. the reduction of one negative argument to yield (3:1). According to probability theory and our model, (3:1) is stronger evidence than (4:2). (p. 13)"

However, the statements are not necessarily correct. Although their model predicts that (3:1) is stronger evidence than (4:2), probability theory may not. In the symmetric binomial case, the Bayesian model suggests that (3:1) and (4:2) are equally informative because the difference between the number of favorable and unfavorable witnesses is two for both cases. Again, Einhorn and Hogarth mispercieved the problem as point estimation and therefore suggested that (3:1) has larger p than (4:2).

Einhorn and Hogarth further suggested that their model implies that the deletion of negative arguments results in a loss of n such that large downward adjustments to p may occur. That is, deletions that substantially reduce n can lower net strength and thus work against increases in p. However, they argued that probability theory and their model diverage on this point. They stated:

> "consider initial evidence of (1:1) and compare (2:1) to (1:0). If a_2 = .4, Sn = .58 and Sn = .60, which are much closer than would be the case if probability were used as a measure of evidentiary strength. (p. 13)"

In fact, probability theory suggests the same thing, that is, (2:1) and (1:0) may be equally informative according to the Bayesian model.

In the process of developing their model, Einhorn and Hogarth proposed a minimum P_c (a small value of p) in order to get rid of the undesirable result that net strength is negative when p = 0. The development of P_c resulted from an ad hoc assumption. Although ad hoc assumptions are not always bad (e.g., Kitcher, 1982; Popper, 1959), the introduction of P_c seems scientifically improper. For example, Popper suggested that ad hoc hypotheses are acceptable when their introduction does not decrease but rather increases the degree of falsifiability or testibility of the theory. That is, the modified theory should rule out more logically possible events and thus restrict the range of permitted events. However, introducing Pc into Einhorn and Hogarth's theory does not increase the falsifiability of the theory. The theory with Pc does not narrow the range of confirming events in the empirical world.

Einhorn and Hogarth further proposed a second factor which influences adjustments to p, namely, whether one is evaluating a particular hypothesis or its complement. Their model specifies that attentional shifts due to rephrasing likelihood questions will lead to the subadditivity of complementary probabilities. They postulated it a "focus effect" which can be expressed as

$$\begin{aligned} &Sn\,(c{:}f) < 1 - Sn(f{:}c) \quad \text{or} \\ &Sn'(c{:}f) < 1 - Sn(f{:}c). \end{aligned}$$

When p and (1−p) are greater than Pc,

$$\begin{aligned} Sn(f{:}c) + Sn(c{:}f) &= (p-a_2/n) + [(1-p)-a_2/n] \\ &= 1 - 2a_2/n \quad , \end{aligned}$$

thus the focus effects occurs if $a_2 > 0$. However, when either p or (1−p) is less than or equal to Pc,

$$Sn(f{:}c) + Sn'(c{:}f) = (p-a_2/n) + [(1-p)+a_2/n] = 1 \quad ,$$

thus no focus effect should occur.

Einhorn and Hogarth ran experiments to test their model and examined the effects of the amount of evidence on strength of evidence and tested for focus effects. Thirty-two subjects were presented with a set of scenarios that involved a hit-and-run accident seen by varying numbers of witnesses who were asked to judge how likely the accident was caused by a particular colored car. Each stimulus contained the same basic story but varied in the total number of witnesses, n, the number of witnesses saying it was a green car, f, or a blue car, c, and whether one was to judge the likelihood that the majority or minority position was true. For aggregate analyses, the predicted mean net strength, $\hat{\bar{S}}n(f{:}c)$, can be written as

$$\hat{\bar{S}}n(f{:}c) = p - \hat{a}_2\,(1/n)$$

where $\hat{a}_2$ is the estimated weight for the hypothesized adjustment process.

The parameters in the model, a_2 and Pc, have to be estimated from the data. Pc can be located by finding where $\overline{Sn} > p$ at small n, since the sign of a_2 is positive when $p \leqslant Pc$ in their model. However, they noticed a problem in estimating a_2. As they state,

> "A statistical problem in estimating a_2 from (9) [$Sn(f:c) = p - a_2(1/n)$] is that p and 1/n must be highly correlated since $p = f(1/n)$. This makes the determination and testing of a_2 problematic. (p. 21)"

They considered that they had a multicolinearity problem in estimating a_2. However, the argument that p and 1/n must be correlated is wrong. The amount of evidence should influence the adjustment to p but not the p value itself.

Because they perceived that p and 1/n are highly correlated. They used a two-step procedure to handle this "multicolinearity problem". First, $\overline{Sn}$ was regressed onto p to test for the importance of p as an anchor. Then by regressing the difference, $p - \overline{Sn}$, onto 1/n, they estimated a_2 and tested whether the hypothesized adjustment process predicts the differences between mean net strength and p. In the first step, they got a high correlation between $\overline{Sn}$ and p ($r = .98$) and interpreted the result as indicating people anchor on p in the assumed evaluation process. However, the analysis does not have to give support for anchoring on p. In the Bayesian symmetric binomial case, the strength of evidence and p are perfectly correlated. It is possible that people anchor on prior probability and the result that $\overline{Sn}$ correlates high with p still can be found.

On the whole, Einhorn and Hogarth showed that their model fits the data. However, the fit is not surprising since almost any theory intelligently (not perversely) conceived will fit the data reasonably well (e.g., Dawes & Corrigan, 1974). A better strategy is to compare models in order to discover which model fits the data better under that particular condition (e.g., Commbs, Dawes, & Tversky, 1970; Platt, 1964).

A third factor which influences adjustments to p is the number and specificity of alternative hypotheses. Einhorn and Hogarth suggested that a diffusion effect occurs when the total amount splits or diffuses into multiple categories or hypotheses. They argued that the diffusion effect violates the evaluation of evidence in probability theory. As they state,

> " . . . recall our hit-and-run scenario and imagine that four witnesses reported a green car and four reported the color as blue, i.e., (4G:4B). Now consider

> a second situation in which four witnesses reported green, two reported blue, and two reported red; i.e., (4G:2B:2R). In this second case, is it more, less, or equally likely that a green car was responsible for the accident? We hypothesize that for many people, the strength of evidence for a green car will not be the same. . . . Note that a diffusion effect violates the evaluation of evidence in standard probability theory. That is, the probability of a hypothesis should be unaffected by the number or composition of alternative hypotheses. Thus, if the probability of some hypothesis H is p, the fact that H is made up of one or more alternatives is irrelevant to the probability of H (and therefore H). (p. 26)"

Again, the argument against probability theory is wrong. According to the Bayesian model, the probability of a hypothesis is affected by the number of alternative hypotheses. (4G:4B) and (4G:2B:2R) may provide different results in evaluating the likelihood that a green car caused the accident. With evidence (4G:4B), it is assumed that one has to evaluate the strength of evidence in supporting hypothesis H_G between two hypotheses, H_G and H_B. On the other hand, with evidence (4G:2B:2R), it is assumed that one has to evaluate the strength of evidence in supporting hypothesis H_G among three hypotheses, H_G, H_B, and H_R. The maximum entropy priors in the former case are $P(H_G) = 1/2$ and $P(H_B) = 1/2$, those in the latter case are $P(H_G) = 1/3$, $P(H_B) = 1/3$ and $P(H_R) = 1/3$. Thus the two situations can lead to different results in evaluating the strength of evidence in supporting hypothesis H_G. For example, when two hypotheses, H_G and H_B, with evidence (4G:4B), are considered, and $P(g|H_G) = P(b|H_B)$, the posterior probability of H_G is .50 since the difference between the number of favorable and unfavorable witnesses is zero. When three hypotheses, H_G, H_B, and H_R, with evidence (4G:2B:2R) are considered, and the conditional probabilities assumed to be $P(g|H_G) = P(b|H_B) = P(r|H_R) = .80$, $P(g|H_B) = P(g|H_R) = P(b|H_G) = P(b|H_R) = P(r|H_G) = P(r|H_B) = .10$, the posterior probability of H_G is .97. Thus the two situations can lead to different results in supporting H_G.

In summary, Einhorn and Hogarth showed statistical and methodological shortcomings in their paper. Even though their theory and model can deal with diagnostic inference which is causal, backward in its direction, and concerned with specific cases, it is not correct to state that statistical models cannot deal with similar situations. In fact, they mispercieved the problem with which they were dealing and did not use the appropriate statistical procedure in interpreting the problem. The problems on which they were working are essentially Bayesian problems with incomplete information. Although their model fits the data well, it does not mean that their model is good or correct. To improve in methodology, it would be better to compare several, at least two, models at the same time and to see which one is better.

References

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Einhorn, H. J., & Hogarth, R. M. (1982). *A theory of diagnostic inference: I. Imagination and the psychophysics of evidence.* Unpublished manuscript, Center for Decision Research, Graduate School of Business, University of Chicago.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92, 433-461.*

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods.* New York, NY: John Wiley & Sons.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems Sci. Cybernetics, SSC-4*, 227-241.

Kitcher, P. (1982). *Abusing science: The case against creationism*. Cambridge, MA: MIT.

Mosteller, F., & Wallace, D. L. (1972). Deciding authorship. In J. M. Tanur, F. Mosteller, W. Kruskal, R. F. Link, R. S. Pietters, & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 164-175). San Francisco, CA: Holden-Day.

Peterson, C. R., & Beach, L. R. (1967). Conservatism in simple probability inference tasks. *Journal of Experimental Psychology, 76*, 236-243.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Science.

Rosenkrantz, R. D. (Ed.) (1983). *E.T. Jaynes: Paper on probability statistics, and statistical physics*. Dordrecht, Holland: D. Reidel.

Schum, D. A. (1980). Current developments in research on cascaded inference processes. In T. S. Wallsten (Ed.), *Cognitive processes and decision behavior* (pp. 179-210). Hillsdale, NJ: Erlbaum.