OPERATIONS RESEARCH
Vol. 26, No. 3, May-June 1978

0030-364X/78/2603-0406 $01.25

Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach

JOHN R. HAUSER

Northwestern University, Evanston Illinois

(Received February 1977; accepted September 1977)

Disaggregate demand models predict the choice behavior of individual consumers. But while such models predict choice probabilities $(0<p<1)$, they must be tested against (0, 1) choice behavior. This paper uses information theory to derive three complementary tests that help analysts select a "best" disaggregate model. "Usefulness" measures the percentage of uncertainty (entropy) explained by the information the model provides. It provides theoretic rigor and intuitive appeal to the commonly used likelihood ratio index and leads to important practical extensions. "Accuracy" is a new two-tailed normal test that determines whether the (0, 1) observations are reasonable under the hypothesis that the model is valid. "Significance" is the standard chi-squared test to determine whether a null model can be rejected. This paper also extends the information test to examine the relationships among successively more powerful null hypotheses. For example, in a logit model one can quantify (1) the contribution due to knowing aggregate market shares, (2) the incremental contribution due to knowing choice set restrictions, and (3) *the final incremental contribution due to the explanatory variables.* Further extensions provide "explanable uncertainty" measures applicable if choice frequencies are observed. Market research and transportation analysis empirical examples are given.

THE DESIGN of successful products and services requires valid predictions of how consumers will respond to changes in product or service strategy. Recently in marketing research and in transportation planning, demand models have been developed that base their predictions on causal hypotheses that model the behavior of individual consumers (logit analysis, McFadden [19]; probit analysis, Finney [3]; discriminant analysis, Fisher [4], etc.). Because of their behavioral content and because of the rich, individual specific data on which these models are based, analysts expect these "disaggregate behavioral demand models" to provide accurate predictions of consumer behavior and to provide useful diagnostics that help understand the consumers' choice process. But how accurate are these models? This question, which must be answered to the satisfaction of both the analytic modeler and the marketing or transportation manager, is the subject of this paper.

Disaggregate models predict group response, e.g., the number of bus riders from zone to zone, by aggregating together predictions of how individual consumers behave (Koppelman [16]). But because of potential errors in modeling, in measurement, in estimation, and because of random influences on consumer behavior, these models cannot predict with certainty. Instead, for each individual i, they predict choice probabilities. For example, in modeling choice among modes of transportation a model might predict the probability that a particular consumer will choose public transit, drive, walk, or will not travel. The fundamental problem in testing is that while the models predict probabilities, they must be tested on observed events. In a given instance individual i either rides, drives, walks, or stays put. Suppose a model predicts that i will ride the bus with probability 0.7 and i does ride the bus. To assess the validity of such a model, a test must quantify how much "rightness" or "wrongness" there was in the prediction. Furthermore, suppose a model makes individual predictions, but for 1,000 individuals. Analysts need a test to measure that model's predictive ability and to select a "best" model.

1. EXISTING TESTS

The problem of testing predicted probabilities as observed events is not new, and there are a number of tests now in use. Aggregate tests compare average probabilities with aggregate statistics such as market shares. Disaggregate tests compare individual probabilities with individual events. This section reviews both types of tests and then discusses their relative merits.

Aggregate tests have strong intuitive appeal and are useful aids to communication between analysts and managers. Managers can internalize the meaning of these tests, compare the model to their prior beliefs, and assess the accuracy of a model in a way that can be readily communicated to others. For example, *first preference recovery*, r_1, which computes the percentage of individuals that select their first-preference alternative, is easy to understand and can be readily compared to chance recovery, $r_c = 1/(\text{number of alternatives})$, or market share recovery, $r_{ms} = \sum_j (ms_j^2)$ where ms_j = market share of product j. In most probabilistic models, maximum probabilities are substituted for first preference because choice probabilities are monotonic in preference.

Other useful aggregate tests compare predicted market shares, $\widehat{ms}_j$, with observed market shares, ms_j. $[\widehat{ms}_j = (1/n) \sum_i p_{ij}$, where p_{ij} is the predicted probability that i chooses j and n is the total number of individuals.] For example, *root mean square percent error* in predicted market shares, e_p, has been used by Koppelman [16] to compare aggregation methods for mode choice predictions in Washington, D.C. He reports errors in the range of 25–35%. Hauser and Urban [10] report that percent error was a better discriminator than first-preference recovery between von Neumann-Morgenstern utility assessment and logit analysis ($e_p=18\%$ vs. 36% while $r_1=50\%$ vs. 46%). Similar tests such as weighted percent error, mean absolute error, least-squares error, and weighted least-squares error have all been used with varying success (see [16]).

Disaggregate tests address the basic testing problem by comparing pre dictions and events on an individual level. These tests can discriminate between models that predict aggregate market shares well but miss the individual choice process and those that capture individual differences. For example, any logit choice model with $J-1$ choice specific constants ($J=$the number of alternatives) will predict aggregate market shares exactly on the "calibration" data, but different models within this class may be "better" than others. Disaggregate tests quantify the concept of "better."

A common test is the *log-likelihood chi-squared significance test* (Mood and Graybill [20]). In this test the probabilistic model is compared to a null model. If the null model can be formulated as a restriction (subset) on the parameters of the tested model, then $L=2$ log [likelihood ratio of tested model to null model] is χ^2 distributed with degrees of freedom equal to the difference in degrees of freedom between the tested model and the null model. In logit applications the most common null model is the equally likely model (all choice parameters set equal to zero), but some researchers use the market-share proportional model (choice specific constants only) when a full set ($J-1$) of choice specific constants are used in the estimated model.

The chi-squared test can reject a null model, but it cannot give an indication of how well a model predicts nor can it compare two models unless one model is a restriction of the other. The most common disaggregate test used to measure a model's predictive ability is the *likelihood ratio index* [19]. This test, $\rho^2=1-L(X)/L_0$ where $L(X)$ is the log-likelihood of the tested model (explanatory variables X) and L_0 is the log-likelihood of the null model, acts like a pseudo-R^2 since $\rho^2=0$ when $L(X)=L_0$ and $\rho^2=1$ when the model predicts perfectly, otherwise $0<\rho^2<1$. In related tests Kendall [14] suggests a correlation coefficient similar to that for regression and Cragg [1] suggests a correlation-like coefficient. Stopher [25] uses the correlation ratio (Weatherburn [27], Neter and Maynes [21], Johnson and Leone [13]) to augment the correlation coefficient, but his use requires that individuals be grouped. Results are extremely sensitive to the grouping.

Although the aggregate tests are intuitive and aid communication between managers and analysts, they can be misleading. For example, a first-preference recovery of 55% is usually good, but not in a market of two products. A recovery of 90% is usually good in a two-product market but not if one product has a market share of 95% ($r_{ms}=90.5\%$). Similarly, e_p is identically zero for the market-share proportional null model, but $e_p>0$ for most models, which may be more realistic representations of the true choice process. (For example, most logit models without choice specific constants will not predict market shares exactly. But choice specific constants are often undesirable because they make it difficult to project a model from the "calibration" situation to a new situation. In particular, if new products are introduced, there is no way to know the choice specific constant for the new product.) These restrictions on aggregate tests caution the analyst to use aggregate tests with great care. Furthermore, because aggregate tests do not address the fundamental problem of testing individual probabilities against observed events, they may not be able to discriminate between models to select the "best" model of individual choice behavior.

The disaggregate tests do address the fundamental testing problem. The chi-squared test can statistically reject properly formulated null hypotheses, and the likelihood ratio index can give an R^2-like measure of the predictive ability of a probabilistic model. In many cases these tests nicely complement the aggregate tests. Disaggregate tests are not used alone because they are theoretically sensitive to the problem that $\lim_{p_{ij}\to 0}[\log p_{ij}]=-\infty$. Aggregate tests are not as sensitive to zero probabilities.

This battery of aggregate and disaggregate tests can address many problems in testing probabilistic models, but there are importance problems that this battery does not address. For example: (1) The likelihood ratio index behaves nicely at the limits ($\rho^2=0$, $\rho^2=1$), but it does not have an intuitive interpretation between the limits. Managers need an intuitive interpretaton that is naturally related to a measure of probabilistic uncertainty. (2) ρ^2 can be computed relative to any null hypothesis, L_0, but no deductive theory indicates whether that simple computation is the appropriate generalization for complex null hypotheses. (3) The choice of a null hypothesis is based on judgment. A good test should indicate which null hypothesis is best and indicate the relationship among null hypotheses. (4) The null hypothesis sets the lower bound for ρ^2, but $\rho^2=1$ may not be the appropriate upper bound. If individuals make repeated choices and do not always select the same alternative, then $\rho^2=1$ is not possible, even in theory. (Perfect prediction would require different probabilities for different occasions. Such predictions are not possible without situational variables.) A theory-based test should indicate how to incorporate upper-bound information. Finally, (5) the chi-squared test can reject a null hypothesis but does not test the accuracy of predictions. A test of "accuracy," which can accept or reject the *tested* model, is necessary to complement the chi-squared test of "significance" and the ρ^2 test (or its generalization) of "usefulness."

These problems and others can be effectively resolved by considering probabilistic models as an information system where the predicted probabilities (or null hypotheses) represent the best information derived from the set of explanatory variables, X.

2. INFORMATION THEORY: AN INTERPRETABLE TEST OF MODEL USEFULNESS

Suppose there is a set of choice alternatives, $A=\{a_1,a_2,\cdots a_j\}$, and suppose there is a set of explanatory variables, X, which take on specific values, X_i, each individual, i. Suppose that through some mathematical analysis a conditional probability model, $p(a_j \mid X_i)$, has been developed to estimate choice probabilities, $p_{ij}=p(a_j \mid X_i)$, from the explanatory variables. Suppose that to test the model each individual's choice behavior, as represented by δ_{ij}, has been observed. ($\delta_{ij}=1$ if i chooses j, $\delta_{ij}=0$ otherwise.) This section will derive the information test for such a probabilistic model of consumer behavior. Later sections will extend the test to cases where the choice set varies and the number of choice occasions is greater than one.

The probability model can be viewed as an information system. In other words, the "observable occurrence," e.g., the attributes of the choice alternatives, provides information about "unobservable events," i.e., about the choice outcome. Thus a test uses the information measure, $I(a_j,X_i)$, (Gallagher [5]) to quantify the information provided by X_i. Formally we have $I(a_j,X_i)=\log\,[p(a_j \mid X_i)/p(a_j)]$, where $p(a_j)$ is the prior likelihood of the outcome, i.e., the event that a_j is chosen.

First observe that the information criteria provide managerially interpretable benchmarks. The first benchmark is the expected information provided by the model, $EI(A;X)$, where

$$EI(A;X)=\textstyle\sum_{X_i\epsilon X}\sum_j p(a_j,X_i)\log\,[p(a_j \mid X_i)/p(a_j)] \qquad (1)$$

with $p(a_j, X_i)$ the joint probability of an "observation" of X_i and an "event," a_j chosen.

Another benchmark is the total uncertainty in the system, which is measured by the prior entropy, $H(A)$, where

$$H(A) = -\textstyle\sum_j p(a_j)\log p(a_j). \qquad (2)$$

The prior entropy measures the uncertainty before "observing" X_i. After observing X_i the uncertainty is reduced to the posterior entropy, $H(A \mid X)$, where

$$H(A \mid X) = -\textstyle\sum_{X_i\epsilon X}\sum_j p(a_j,X_i)\log p(a_j \mid X_i). \qquad (3)$$

Note that for a sample of n individuals the test can use $p(a_j, \mathbf{X}_i) = p(a_i \mid \mathbf{X}_i)p(\mathbf{X}_i)$ by setting $p(\mathbf{X}_i) =$ (number of times $\mathbf{X}_i$ occurs)$/n$ and by setting $p(a_j)$ either equal to the observed market-share fraction, ms_j, or equal to 1/(number of alternatives) (equally likely model), or to any other prior belief of $p(a_j)$. For comparing $p(a_j \mid \mathbf{X}_i)$ against the market share model

$$EI(A;X) = \textstyle\sum_i \sum_j (1/n)p(a_j \mid \mathbf{X}_i) \log [p(a_j \mid \mathbf{X}_i)/p(a_j)] \quad (4)$$

and

$$H(A) = -\textstyle\sum_j ms_j \log (ms_j). \quad (5)$$

Note that since $0 \leqq ms_j \leqq 1$, $H(A)$ is positive.

The accuracy of the model can be calculated by comparing the empirical information, $I(A;X)$, with the expected information. (Note: We will discuss this point further.) To compute the empirical information we use

$$I(A;X) = (1/n) \textstyle\sum_i \sum_j \delta_{ij} \log [p(a_j \mid \mathbf{X}_i)/ms_j]. \quad (6)$$

Equations 4, 5, and 6 show that information theory can be formulated to test probabilistic models. But before this test can be used for probabilistic choice models, (4)–(6) must be given more intuitive meanings. Consider the following theorems:

THEOREM 1. *The entropy of a system is numerically equal to the information that would be observed, given perfect knowledge, i.e.,* $H(A) = I(A;$ *perfect knowledge*$)$.

Proof. Under the assumption of perfect knowledge, $p(a_j \mid \mathbf{X}_i) = \delta_{ij}$ for all j. Thus, $I(A;$ perfect knowledge$) = (1/n) \sum_i \sum_j \delta_{ij} \log [\delta_{ij}/p(a_j)]$. Switching summations and recognizing $\delta_{ij}=0$ if i does not choose j give:

$$I(A;\text{ perfect knowledge}) = (1/n) \textstyle\sum_j \sum_{i\epsilon C(j)} \log [1/p(a_j)] + \sum_{i\notin C(j)} 0 \log [0/p(a_j)]$$

where $C(j)$ is the set of individuals who choose j. Since under the null hypothesis the number of individuals who choose i is $np(a_j)$ and since $\lim_{x\to 0} [x \log x] = 0$, this gives

$$I(A;\text{ perfect knowledge}) \textstyle\sum_j [n \cdot p(a_j)/n] \log [1/p(a_j)] = -\sum_j p(a_j) \log p(a_j) = H(A).$$

THEOREM 2. *If the probability model is aggregately consistent with the null hypothesis, i.e.* $\sum_{\mathbf{X}_i \epsilon X} p(a_j, \mathbf{X}_i) = p(a_j)$, *then the expected information is equal to the reduction in uncertainty, i.e.,* $EI(A;X) = H(A) - H(A \mid X)$.

Proof. Expanding $EI(A;X)$ as defined in (1) gives

$$EI(A;X) = \sum_{X_i \epsilon X} \sum_j p(a_j, X_i) \log p(a_j \mid X_i) - \sum_{X_i \epsilon X} \sum_j p(a_j, X_i) \log p(a_j).$$

The first term on the right-hand side is $-H(A \mid X)$, as given by (3). Using $\sum_{X_i \epsilon X} p(a_j, X_i) = p(a_j)$, the second term can be shown to be $-H(A)$, as given by (2). Thus $EI(A;X) = H(A) - H(A \mid X)$.

THEOREM 3. *Suppose that the true choice probabilities are given by* $p_{ij} = q_{ij}$; *then* $EI(A;X)$ *attains its maximum value for* $p_{ij} = q_{ij}$.

Proof. Let $Q = \max_{p_{ij}, \lambda_i} \{\sum_i \sum_j (1/n)\ q_{ij} \log\ [p_{ij}/p(a_j)] + \sum_i \lambda_i (1 - \sum_j p_{ij})\}$, where the Lagrange multipliers, λ_i, have been used to incorporate the constraint that $\sum_j p_{ij} = 1$ for all i. Since q_{ij} are the true probabilities, $p(a_j, X_i) = q_{ij}$ (number of times X_i occurs)$/n$. The conditions for optimality are then $\lambda_i = (1/n)$ and $p_{ij} = q_{ij}$ and second-order conditions indicate a maximum.

Theorems 1, 2, and 3 together give intuitive meaning to the information measure. The entropy, $H(A)$, is a naturally occurring measure of uncertainty in thermodynamics (Reif [23]), in statistics (Jaynes [12]), and in marketing (Herniter [11]). It measures the total uncertainty of the system, and by Theorem 1 it represents the maximum uncertainty that can be explained with perfect information. Furthermore, if the model is less than perfect, then the expected information represents the reduction in uncertainty due to the model. Thus, $EU^2 = EI(A;X)/H(A)$ can be used to measure the percentage of uncertainty explained by the model. $1 - EU^2 = H(A \mid X)/H(A)$ gives the residual uncertainty. (Note that $H(A)$, and hence EU^2, depend on the null hypothesis. Since $p(a_j) = 1/J$ maximizes $H(A)$, the equally likely null model represents maximum uncertainty or, conversely, minimum knowledge.)

Finally, if knowledge is limited by the explanatory variables, X, and if there are some true probabilities, q_{ij}, known only to a clairvoyant, then the best value for the expected information is attained by setting $p_{ij} = q_{ij}$. Thus the expected information is indeed an "honest reward function" (Raiffa [22]) in the sense that the "reward" structure would force a clairvoyant to divulge the true probabilities. Note that some commonly used measures such as least squares, R^2, can be shown to be dishonest for testing probabilities against events. (A clairvoyant would maximize R^2 by setting $p_{im} = 1$ for alternative m such that $q_{im} = \max_j q_{ij}$ and $p_{ij} = 0$ for $j \neq m$.)

A problem with EU^2 is that it is computed independently of the observed data, δ_{ij}. In fact, it is the expected value of a test statistic, $U^2 =$

$I(A;X)/H(A)$. Thus, in practice, an analyst can either (1) use the empirical uncertainty explained, U^2, to measure the predictive usefulness of a model, or (2) use the expected uncertainty explained, EU^2, for usefulness and then test the "closeness" of U^2 to EU^2. The "closeness" test can be interpreted as a test of accuracy and will be explained in the next section.

All that remains is to show that U^2 is the appropriate generalization of of the likelihood ratio index, ρ^2. This is shown by

THEOREM 4. *If the null hypothesis is independent of i, i.e.,* $p_{ij}^0 = p(a_j)$, *then the likelihood ratio index,* ρ^2, *is numerically equal to the empirical percent uncertainty explained,* U^2.

Proof. $\rho^2 = 1-r$, where r is the logarithm of the likelihood function for the probabilistic model, call it $L(X)$, divided by the logarithm of the likelihood function for the null hypothesis, call it L_0. Thus $\rho^2 = 1-L(X)/L_0 = (L_0-L(X))/L_0$. Now $L_0 = \log \prod_{i=1}^{n} \prod_{j=1}^{J} p(a_j)^{\delta_{ij}} = \sum_i \sum_j \delta_{ij} \log p(a_j)$. Similarly, $L(X) = \sum_i \sum_j \delta_{ij} \log p(a_j \mid \mathbf{X}_i)$; thus $L_0-L(X) = \sum_i \sum_j \delta_{ij} [\log p(a_j) - \log p(a_j \mid \mathbf{X}_i)] = -n\, I(A;X)$.

Now since $p(a_j)$ is independent of i: $L_0 = \sum_j \sum_{i \varepsilon C(j)} \delta_{ij} \log p(a_j) = \sum_j n\, p(a_j) \log p(a_j) = -n\, H(A)$. Thus $\rho^2 = -n\, I(A;X)/[-n\, H(A)] = I(A;X)/H(A) = U^2$.

In summary, the information test, EU^2 or U^2, provides a natural measure of uncertainty and a natural intuitive managerial interpretation of uncertainty explained. Furthermore, it is an "honest" reward function and in the case of simple null hypotheses it reduces to the likelihood ratio index. Thus, EU^2 or U^2 provides the first stage of a three-stage disaggregate test. The next two sections will develop accuracy and significance tests to complement this test of usefulness. Section 5 will then shown how this test extends naturally to successively more powerful null hypotheses and Section 6 will show how to shift the upper bound when frequencies rather than single events are observed.

3. NORMAL DISTRIBUTION: A COMPANION TEST FOR ACCURACY

It is tempting to use EU^2 as a measure of uncertainty, but EU^2 can be easily maximized for a completely inaccurate model (i.e., set $p_{i1}=1$ and $p_{ij}=0$ for $j \neq 1$). Thus a test must be devised to compare an observed statistic, U^2, to its expected value, EU^2. Fortunately, under reasonable assumptions $I(A;X)$ is normally distributed.

THEOREM 5. *Suppose that the model is accurate, i.e., the observed events,* δ_{ij}*'s, are Bernoulli random variables with probabilities given by* $p(a_j \mid \mathbf{X}_i)$, *and individuals are independent. Then for large samples* $I(A;X)$ *is a normal* *random variable with, mean $EI(A;X)$ and variance*

$$V(A;X) = (1/n)\sum_i\{\sum_j p(a_j \mid X_i)[\log (p(a_j \mid X_i)/p(a_j))]^2 \\ -[\sum_j p(a_j \mid X_i)\log (p(a_j \mid X_i)/p(a_j))]^2\} \tag{7}$$

Proof. First recognize that $I(A;X) = \sum_i (1/n) \sum_j \delta_{ij} \log [p(a_j \mid X_i)/p(a_j)]$ is the sum of n independent random variables, e.g., the first random variable takes on a value $(1/n)$ log $[p(a_j \mid X_i)/p(a_j)]$ with probability $p(a_j \mid X_1)$. Under reasonable conditions this sum of independent random variables is asymptotically normal. The reasonable conditions require that no term dominates the sum and that the individual terms are not uniformly skewed (Drake [2]). Although algebraically complex, these conditions reduce to the condition that the $p(a_j \mid X_i)$'s are not arbitrarily close to 1 or 0. This condition is met in any reasonable empirical probability of choice model, such as the logit model. The mean and variance are then directly computed.

Thus a two-tailed test can be applied to determine whether $I(A;X)$ is a reasonable observation from the model. If $I(A;X)$ is statistically far from $EI(A;X)$, reject the probabilistic model as unable to explain the empirical observations.

4. STATISTICAL SIGNIFICANCE: ITS RELATIONSHIP TO USEFULNESS AND ACCURACY

Based on Section 2, the information measure provides a useful interpretation and extension of the commonly applied likelihood ratio index, and based on Section 3, this measure provides a new test of accuracy that allows the analyst to accept or reject the hypothesis that the observations could have been generated by the model.

By recognizing that $L = 2nI(A;X)$, we can add a third stage to the disaggregate information test. This third stage, *significance,* is simply the standard chi-square significance test reviewed in Section 1. In this test the analyst sees whether the model, the $p(a_j \mid X_i)$'s, and the observations, δ_{ij}'s, are reasonable under the hypothesis that the null model is true. Too large a χ^2 statistic rejects the null model. Note that $I(A;X)$ is normally distributed in the accuracy test because only the δ_{ij}'s are random variables under the hypothesis that the probabilistic model is correct, while $2nI(A;X)$ is chi-squared distributed in the significance test because both the δ_{ij}'s and the $p(a_j \mid X_i)$'s are random variables under the hypothesis that the null model is correct.

This three-part test of "usefulness," "accuracy," and "significance" is illustrated in Figure 1. The model is a standard logit model without choice specific constants. The choice set consists of seven shopping centers in the suburbs north of Chicago and the explanatory variables are factor scores for each individual along six dimensions: variety, quality, atmosphere, value, layout, and parking. The dependent variable was first preference and the sample size is 99.

The model is overwhelmingly significant with respect to the equally likely null model, N_0, but it only explains 44% of the uncertainty. The model is clearly accurate with respect to N_0 (99% level), but less accurate with respect to the market-share null model, N_1 (80% level). Note that the accuracy test is a relative test because the null model appears in it. $[EI(A;X)$ depends on $p(a_j)$.] In this application the model was not statistically rejected, but the accuracy test relative to N_1 was one stimulus that led to further investigation. The final model, presented in Hauser

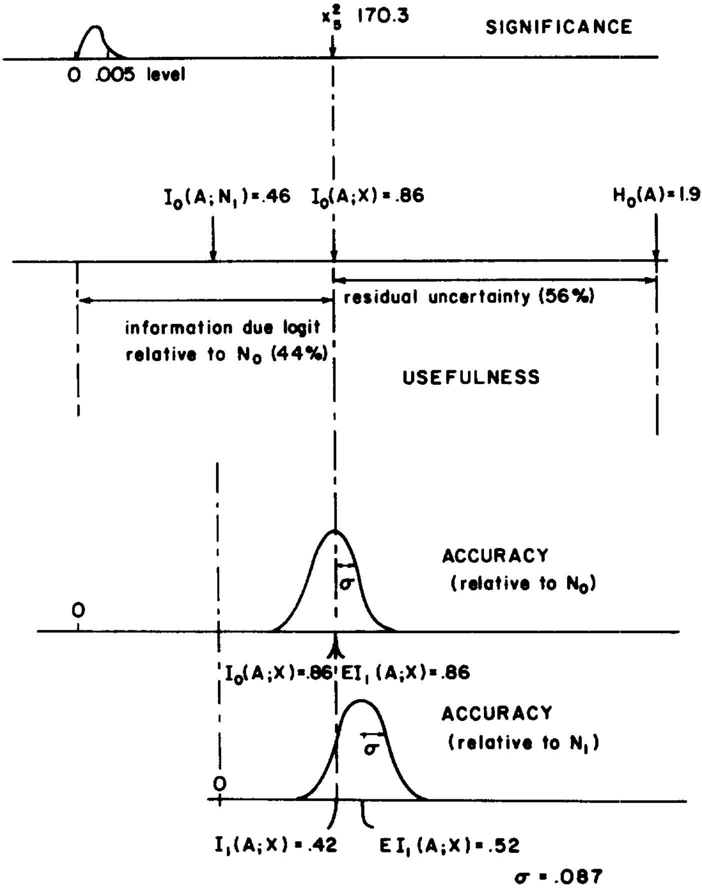

Figure 1. Information test applied to shopping center preference prediction.

and Koppelman [8], required statistical corrections for choice-based sampling (Manski and Lerman [18]).

5. SUCCESSIVELY MORE POWERFUL NULL HYPOTHESES

The example in Figure 1 illustrates how important null hypotheses are in the choice of a test. Fortunately, the information test provides a useful generalization that helps overcome the problem of selecting a null hypothesis. To begin this discussion, consider the following formal notation.

Call the equally likely null hypothesis N_0 ($p(a_j) = 1/j$), and call the market-share proportional null hypothesis N_1 ($p(a_j) = ms_j$). Using the theory introduced in (1)–(6), one can compute the observed information and the entropy relative to either null model. Let $I_1(A; X)$ be the observed information relative to N_1, and let $H_1(A)$ be the entropy relative to N_1 (see (5) and (6)). Similarly, let $I_0(A; X)$ and $H_0(A)$ be computed relative to N_0. (Substitute $p(a_j) = 1/J$ in (5) and (6).) Finally, let $I_0(A; N_1)$ be the observed information of N_1 relative to N_0. (Substitute $p(a_j \mid \mathbf{X}_i) = ms_j$ and $p(a_j) = 1/J$ in (6).)

The first important results are that $I_0(A; N_1)$ can be more simply represented and that $I_0(A; X)$ can be computed from component parts.

THEOREM 6. *The incremental information of N_1 relative to N_0 is equal to the reduction in entropy in going from N_0 to N_1, i.e., $I_0(A; N_1) = H_0(A) - H_1(A)$.*

Proof. $I_0(A; N_1) = (1/n) \sum_i \sum_j \delta_{ij} \log [ms_j/(1/J)]$. Thus

$$I_0(A; N_1) = (1/n) \sum_i \sum_j \delta_{ij} \log ms_j - (1/n) \sum_i \sum_j \delta_{ij} \log (1/J).$$

Switching the order of the summation and noting that, for fixed j $\sum_i \delta_{ij}$ is equal to choosing ms_j and that $\sum_i \sum_j \delta_{ij} = n$ yield: $I_0(A; N_1) = -(-\sum_j ms_j \log ms_j) + (-\sum_j (1/J) \log (1/J)) = -H_1(A) + H_0(A)$, which completes the proof.

THEOREM 7. *The information relative to N_0, $I_0(A; X)$, is equal to the information relative to N_1, $I_1(A; X)$, plus the information of N_1 relative to N_0, $I_0(A; N_1)$, i.e., $I_0(A; X) = I_0(A; N_1) + I_1(A; X)$.*

Proof. Similar to that of Theorem 6.

Together, Theorems 6 and 7, which can be proven for any set of null hypotheses, provide very useful results. Taking N_0, the equally likely hypothesis, as the state of no knowledge, one can view information as coming successively first by the hypothesis, N_1, which tells only market shares, and then incrementally from the model $p(a_j|\mathbf{X}_i)$ for all i. Furthermore, the "no knowledge entropy," $H_0(A)$, can be viewed as being successively reduced, first to $H_1(A)$ by the market-share information, N_1, and then to the estimated residual entropy, $\tilde{H}(A|X)$. (Note that $\tilde{H}(A|X)$ is independent of both N_0 and N_1.)

A practical advantage of Theorems 6 and 7 is that while $I_0(A;X)$ may be difficult to compute, $H_0(A)$ and $I_0(A;N_1)$ are given by simple formulas. Thus $I_*(A;X)$ can be computed relative to any null hypothesis by simple addition and subtraction once $I_0(A;X)$ is known. For example, $H_0(A) = -\sum_j (1/J) \log (1/J) = \log J$ and $I_0(A;N_1) = \log J + \sum_j ms_j \log ms_j$.

A point of further interest is that Theorems 6 and 7 apply to the expected information measure only if $(1/n)\sum_i p(a_j, \mathbf{X}_i) = ms_j$, i.e., only if the model is constrained to correctly predict the market shares. Thus $I_0(A;X) - I_1(A;X)$ is only equal to $EI_0(A;X) - EI_1(A;X)$ when the predicted market shares are constrained. This is why the test of accuracy is actually a test relative to the null hypothesis.

Once the information measure is extended to test the comparison between simple null hypotheses like N_0 and N_1, the generalization is straightforward to other null hypotheses or to successively stronger models. For example, when choice specific constants were added to the model in Figure 1, they explained an additional 3.1% of the uncertainty.

An important problem in practice is when the choice set varies across individuals. This problem can be addressed with the information test by selecting a null model, N_2, which assumes that the choice set and nothing else is known. This test is illustrated in a study by Silk and Urban [24] on deodorants. There were 18 brands on the market, but the average size of the choice set was only three brands.

Define the null model, N_2, as follows: Let J_i equal the number of alternatives in individual i's choice set; then the null probabilities are given by $p_{ij}^0 = 1/J_i$ if alternative a_j is in individual i's choice set and $p_{ij}^0 = 0$ otherwise.

In the study the explanatory variable was a ratio-scaled preference measure calculated from constant sum-paired comparisons (Torgenson [26]). The dependent variable was the last brand purchased. The model was a one-parameter logit model linking preference to probability of choice. First-preference recovery was 83%.

Relative to the equally likely hypothesis (N_0, $p_j = 1/18$), the logit model explained 80% of the total uncertainty. But N_2 explains 62% of that uncertainty and the logit model adds only 18% to that. Thus N_2 represents a significant amount of information and is an extremely strong assumption. (In a category like deodorants, where the choice set is determined more by each consumer's interest than by product availability, knowledge of everyone's choice set contains considerable preference information.)

Finally, as is shown in Figure 2, the information test can compute information as coming first from N_1 relative to N_0, then incrementally from N_2 relative to N_1, and finally from the logit model (X) relative to N_2. This can be done even though the implicit parameters for N_1 are not a subset of N_2 or of those for the logit model.

6. FREQUENCY OF CHOICE

A final problem that the information test can address is the problem encountered when market-research data is collected from a consumer panel. In this case observed choice is not a one-time occasion, but rather the consumer makes repeated purchases over time. Frequencies rather than (0, 1) events are observed.

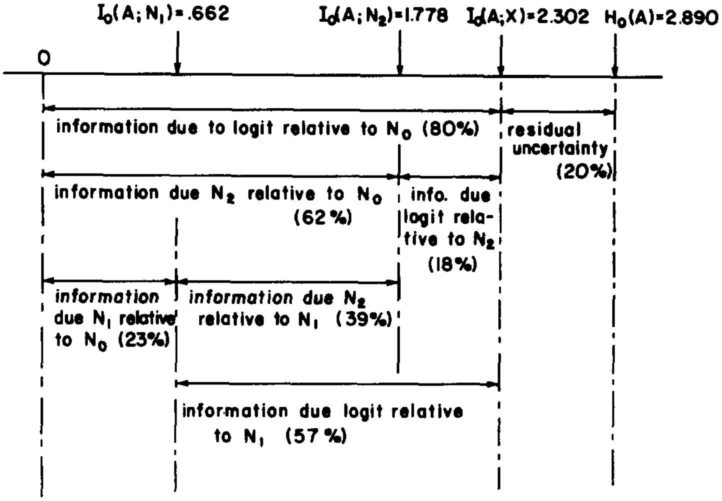
I₀(A; N₁)=.662
I₀(A; N₂)=1.778
I₀(A;X)=2.302
H₀(A)=2.890
0
information due to logit relative to N₀ (80%)
residual uncertainty (20%)
information due N₂ relative to N₀ (62%)
info. due logit relative to N₂ (18%)
information due N₁ relative to N₀ (23%)
information due N₂ relative to N₁ (39%)
information due logit relative to N₁ (57%)

Figure 2. Information test when the choice set varies (deodorant example).

Perfect information would result from correctly predicting every choice occasion for every individual; i.e., $p_{ijk}=\delta_{ijk}$ where k indexes the choice occasion. Unfortunately, without situational variables probabilities, p_{ij}, that are predicted by probabilistic choice models are independent of choice occasion. Thus $H_0(A)=I_0(A;$ perfect information$)$ is not possible even in theory.

This problem can be addressed by defining a new perfect model, P_2, such that $p_{ijk}=f_{ij}$, where f_{ij} is the observed frequency. The new entropy, $G_0(A)=I_0(A;P_2)$, then becomes the base uncertainty, and a new measure, $V_0^2=I_0(A;X)/G_0(A)$, gives the percentage of "explainable" uncertainty that is actually explained by the model. Alternatively, a figure such as Figure 2 can be produced and $G_0(A)$ can be compared to $H_0(A)$ to determine the percentage of unexplainable uncertainty.

Thus, in addition to indicating the relationships between the lower bounds (null hypothesis), the information test is readily extendable to indicate the relationships among the upper bounds (explainable uncertainty).

7. SUMMARY

This paper addresses the fundamental problem of testing probabilistic predictions against 0, 1 observed events by deductively deriving an information-theoretic test. Under standard null hypotheses this test reduces to the likelihood ratio index, ρ^2, now in common use. One advantage of the information-theoretic approach is that it gives both theoretic rigor and an intuitive appeal to this hiterto heuristic measure. But the information test goes beyond that. It indicates how to extend ρ^2 to complex null hypotheses and how to change the upper bound on explainable uncertainty. Together these extensions make clear many interesting and complex effects. For example, the contribution of choice set restrictions is quantified in Figure 2.

The information test measures usefulness, but it also statistically measures accuracy. A two-tailed normal test indicates whether the information statistic is reasonable under the hypothesis that the probabilistic model is correct. This test, which is relative to the chosen null hypothesis, provides the model builder with an important diagnostic tool to assess the validity of a probabilistic model.

Finally, under the appropriate null hypothesis, $2nI(A;X)$ is the standard χ^2 statistic used to measure statistical significance.

Thus the information test gives a three-stage disaggregate test of usefulness, accuracy, and significance. It provides useful generalizations for existing disaggregate tests, makes possible new comparisons among models and hypotheses, and indicates the intuitive and statistical relationships among model tests. These advantages are sufficient to add this test to those tests that modelers use to select probabilistic models. To date, the test has been used to test a new ranked probability model (Hauser [6]), to test independence of irrelevant alternatives (Silk and Urban [24]), to compare various means to model consumer perceptions (Hauser and Koppelman [7]), to test the relative effects of attitudinal and engineering variables in logit models (Lavery [17]), to test a bargain-value model of brand choice (Keon [15]), to test improved new product models (Hauser and Urban [9]), and to test location models for financial services.

ACKNOWLEDGMENTS

Special thanks to Professors Glen L. Urban and John D. C. Little for stimulating discussion on the initial formulation of the information test, and to Professors Frank Koppelman and Andy Daughety for probing questions on many aspects of the test. I also wish to thank Glen Urban and Al Silk for providing me with access to and publication of tests performed on the data collected as part of their excellent study of frequently purchased consumer goods. Finally, I wish to thank referees for insightful comments on content and style.

Parts of this study were funded by a university research Grant DOT-OS-40001 from the U.S. Department of Transportation.

REFERENCES

1. J. G. Cragg, "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," Discussion Paper 8, Vancouver, University of British Columbia, 1968.
2. A. Drake, *Fundamentals of Applied Probability Theory*, McGraw-Hill, New York, 1967.
3. D. J. Finney, *Probit Analysis*, Cambridge University Press, New York, 1964.
4. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugenics* **7**, 179–188 (1936).
5. R. Gallagher, *Information Theory and Reliable Communication*, John Wiley & Sons, New York, 1968.
6. J. R. Hauser, "Consumer Preference Axioms: Behavioral Postulates for Describing and Predicting Stochastic Choice," Working Paper, Department of Marketing, Northwestern University, Evanston, Ill., 1976.
7. J. R. Hauser and F. S. Koppelman, "Effective Marketing Research: An Empirical Comparison of Techniques to Model Consumers' Perceptions and Preferences," Technical Report, Transportation Center, Northwestern University, Evanston, Ill., 1976.
8. J. R. Hauser and F. S. Koppelman, "An Empirical Model of Consumer Shopping Behavior," Transportation Center Working Paper, Northwestern University, Evanston, Ill., 1977.
9. J. R. Hauser and G. L. Urban, "A Normative Methodology for Modeling Consumer Response to Innovation" *Opns. Res.* **25**, 579–619 (1977).
10. J. R. Hauser and G. L. Urban, "Direct Assessment of Consumer Utility Functions: von Neumann-Morgenstern Utility Theory Applied to Marketing," Working Paper, M.I.T. Sloan School of Management, Cambridge, Mass., 1977.
11. J. Herniter, "An Entropy Model of Brand Purchase Behavior," *J. Market. Res.* **10**, 361–375 (1973).
12. E. T. Jaynes, "Information Theory and Statistical Mechanics," *Phys. Rev.* **106**, 620 (1957).
13. N. L. Johnson and F. C. Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences*, Vol. II, John Wiley & Sons, New York, 1964.
14. M. G. Kendall, *A Course in Multivariate Analysis*, Charles Griffin, London, 1965.
15. J. Keon, "Bargain-Value Model of Brand Choice," Dissertation Proposal, Wharton School, University of Pennsylvania, Philadelphia, 1977.

16. F. S. Koppelman, "Travel Prediction with Models of Individual Choice Behavior," Center for Transportation Studies, MIT CTS Report No. 7S-7, Cambridge, Mass., 1975.

17. L. Lavery, "Logit Mode Choice Model Calibration Results," Technical Memorandum, Barton-Aschman Assoc., Inc., Minneapolis, Minn., 1976.

18. C. F. Manski, and S. R. Lerman, "The Estimation of Choice Probabilities from Choice Based Samples," Working Paper, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, Pa., 1976.

19. D. McFadden, "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, P. Zarenblea (Ed.), pp. 105–142, Academic Press, New York, 1970.

20. A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics.* McGraw-Hill, New York, 1963.

21. J. Neter and E. S. Maynes, "On the Appropriateness of the Correlation Coefficient with a 0,1 Dependent Variable," *J. Am. Stat. Assoc.* **65,** 501–509 (1970).

22. H. Raiffa, "Assessments of Probabilities," unpublished manuscript, January 1969.

23. F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, New York, 1965.

24. A. J. Silk and G. L. Urban, "Pretest Market Evaluation of New Packaged Goods: A Model and Measurement Methodology," Working Paper, Sloan School of Management, MIT, Cambridge, Mass., November 1975.

25. P. R. Stopher, "Goodness-of-fit Measures for Probabilistic Travel Demand Models," *Transportation* **4,** 67–83 (1975).

26. W. S. Torgenson, *Methods of Scaling*, John Wiley & Sons, New York, 1958.

27. C. E. Weatherburn, *A First Course in Mathematical Statistics*, Cambridge University Press, New York, 1962.