The API and the Design of Experiments
Author(s): James M. Patell
Source: Journal of Accounting Research, Vol. 17, No. 2 (Autumn, 1979), pp. 528-549
Published by: Wiley on behalf of Accounting Research Center, Booth School of Business, University of Chicago
Stable URL: https://www.jstor.org/stable/2490517
Accessed: 14-08-2018 20:37 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Wiley is collaborating with JSTOR to digitize, preserve and extend access to Journal of Accounting Research
The API and the Design of Experiments

JAMES M. PATELL*

1. Introduction

Residual analysis techniques, generically labeled Abnormal Performance Index (API) tests, have served as the primary experimental procedure in a wide variety of empirical financial accounting research studies. Through various types of API calculations, investigators have observed the security price behavior which preceded and accompanied such events as stock splits, secondary distributions, annual earnings announcements, accounting changes, and earnings forecasts.1 Other API studies have drawn inferences about the content of quarterly earnings reports, tax accounting procedures, depreciation methods, and product-line reporting.2 The range of interpretations has included the demonstration of statistically unusual behavior, the imputation of information content or relevance to investors, and the expression of investor preference for a particular accounting technique.

The discussion presented here attempts analytically to characterize the experimental design of API studies and to focus attention on the role of models of investor expectations. I address the questions of what constitutes an adequate model, what makes one model better than

* Assistant Professor, Stanford University. This research was sponsored by the Stanford Program in Professional Accounting, major contributors to which are: Arthur Andersen & Co.; Coopers & Lybrand; Deloitte Haskins & Sells Foundation, Inc.; Ernst & Ernst; Peat, Marwick, Mitchell & Co.; Price Waterhouse & Co.; and Arthur Young & Co. I gratefully acknowledge the comments of William Beaver, Joel Demski, Nicholas Gonedes, George Foster, Robert Kaplan, Richard Lambert, David Ng, Eric Noreen, James Ohlson, and Mark Wolfson. [Accepted for publication April 1979.]

1 See Fama et al. [1969], Scholes [1972], Ball and Brown [1968], Ball [1972] or Kaplan and Roll [1972], and Patell [1976].

2 See Brown and Kennelly [1972], Beaver and Dukes [1972], Beaver and Dukes [1973], and Collins [1975].
another, and how one would empirically identify a superior model. Explicit consideration of modeling as experimental design suggests the definition of two separate “association” concepts, information content and model evaluation, which have fundamentally different measurement properties. The analysis proceeds through the construction of sets of sufficient conditions under which the API represents a valid association measure; these conditions highlight the implicit assumptions that API tests make concerning accounting signals, expectation models, and the probability distributions of security prices.

The analysis draws heavily from the work of Blackwell [1951; 1953] on the design of experiments and the subsequent study by Marschak and Miyasawa [1968] of the economic comparability of information systems. It treats the analysis of security returns as an experiment and applies concepts developed in the Blackwell papers to determine what makes an experiment informative and whether one experimental design is potentially more informative than another. The term “model” is used in the same sense that Blackwell employs “experimental design” and Marschak and Miyasawa (hereafter MM) use “information system.” The terminology is intended to be consistent with the Marshall [1975] study, in which API tests are criticized through the construction of a hypothetical example in which the API performs poorly as an association metric. In an ensuing comment, Barefield, Foster, and Vickrey [1976] propose a sufficiency condition which remedies the deficiencies present in the Marshall example. It is interesting to note that Marshall both opens his original section entitled “Association and the API” and closes his reply (Marshall [1976]) to the Barefield et al. comment with the statement that the term “association” has not been defined in the API context. The construction of such a definition is a central task of this study.

Section 2 introduces the concept of an unobservable variable (investor beliefs) and the necessary reliance on observable models or surrogates. In Section 3, various homogeneity issues are introduced, and in Section 4 I construct a set of sufficient conditions under which the API can provide a measure of information content. Section 5 presents a more detailed examination of the model evaluation association concept and supplies alternative sets of sufficient conditions for API measurement of relative model informativeness. Repeated reference is made to an API study of earnings announcements in order to exemplify various concepts and terms. The underlying analysis should be applicable to a wide variety of empirical accounting research topics and, at least in part, to other statistical techniques similar to Abnormal Performance Index residual tests.

2. Association: Information Content versus Model Evaluation

In an information-economics framework, investment information about a firm is defined as data whose receipt may alter a decision maker’s
probability distribution over future states of the world relevant to the utility of possessing a share of the firm's equity or debt. The use of security price analysis to detect information has led to an operational definition of information as any stochastic signal which may assume a realization such that the probability distributions of security returns conditioned upon the signal differ from the unconditional return distributions. However, this operational definition may obscure the basic conceptual point that prices are manifestations of investors' beliefs about the probability distributions of future states. Gonedes and Dopuch [1974] extensively detail conditions under which this linking of return distributions and investor beliefs is appropriate for information detection. The emphasis here is on the operational technique: when prices (an observable variable) change in a fashion unexplained by other observable events, we infer that investors' beliefs about future states (an inherently unobservable variable) have changed due to the receipt of information.

Figure 1 displays the relationships under consideration. A realized state \( s \in S \) serves as an input to the firm's financial accounting system \( \eta \). The system \( \eta \) produces a signal \( \theta \in \Theta \) (e.g., an annual earnings report) which is provided to investors. I depict investors as maintaining a partition \( X = \{ x_1, \ldots, x_n \} \) of the space \( \Theta \) and as assessing a probability distribution \( P(\cdot) \) over the events \( x_1, \ldots, x_n \).

Since investors' partitions and probability distributions are unobservable, I portray the researcher as constructing an observable model \( Y \), which provides an algorithm for assigning the signal \( \theta \) to an element of the partition \( Y = \{ y_1, \ldots, y_m \} \). Finally, \( Z \) represents a generalized space of security prices or returns; \( Z \) encompasses transformations such as the various methods of computing return residuals (e.g., the Sharpe-Lintner market model) and the various additive or multiplicative aggregations of those residuals over a test period. Elements of \( Z \) will be denoted as the random variable \( \hat{z}(s) \).
It is important to clarify the distinction between $X$ and $Y$. $X$ must be treated as unobservable, not because the concept of earnings may be difficult to define, but because investors' partitions of the space $\Theta$ into events and their probability distributions over these events cannot be directly observed. $Y$ is an observable surrogate for $X$ in the sense that it supplies a method for assigning observed signals to events and for modeling investors' probability beliefs (typically their expected value of the signal). Most $Y$ models use available financial data (e.g., the firm's previous earnings and the current value of a market index of earnings) to label $Y$ realizations as indicators of $X$ events. For example, particular realizations might be:

**$X$ event:** Announced annual earnings per share of $2.50$ are between $0.20$ and $0.30$ higher than the aggregate market expected value, as assessed one month prior to the announcement.

**$Y$ event:** Announced annual earnings per share of $2.50$ are $0.12$ higher than the five-year arithmetic average of prior annual earnings per share, adjusted for stock splits and stock dividends.

One could avoid introducing $X$ events and merely view this type of empirical research as a descriptive investigation of the relation between security prices and various mathematical functions of accounting data. It is when we seek to interpret these phenomena in terms of the revision of probability beliefs that the introduction of variables to denote these beliefs becomes necessary. The analysis which follows presumes that the investigator is examining hypotheses concerning the revision of probabil-

---

8 Beaver and Demski [1979] examine the concept of earnings in a measure-theoretic context and argue for the general impossibility of earnings providing a measure of value or preference.

9 Investor beliefs concerning probability distributions are frequently referred to as investor expectations. That terminology is avoided hereafter to prevent confusion with the mathematical expectation or expected value of a random variable.
ity assessments and emphasizes the interpretation of the $Y$ event algorithm as a model of investor belief. As figure 1 indicates, API measures are computed from $Y$ and $Z$ realizations, but two different sorts of inferences can be drawn: one may examine the association between security prices and investor beliefs concerning signals ($X-Z$ association), which I refer to as information content, or one may examine the association between investor beliefs and various models of those beliefs ($X-Y$ association), which I refer to as model evaluation. In general, the investigator must assume that one of the two forms of association exists in order to test the other.

INFORMATION CONTENT

In a test for information content, the researcher assumes that investors assess a probability distribution over possible $X$ events, and he systematically attempts to determine whether there exist $X$ events whose occurrence appears to alter the probability distribution of security returns. Since $X$ events are not directly observable, the $Y$ algorithm is used to model investors' partition of the signal space and to assign observed signals to particular events. The firm's return series is then examined to determine whether certain events coincide with changes in the return distribution.

The information content concept of association is thus defined as statistical dependence between $\tilde{\xi}(s)$ and $X$ events; the investigator tests for association or statistical dependence between the return probability distribution and events in investors' partition of the accounting signal space under study.\footnote{Statistical independence is typically defined either between discrete events or between random variables. As an intermediate step, we can define the random variable $\tilde{\xi}$ and the partition $X$ to be independent if, and only if, $\phi(\tilde{\xi} | x) = \phi(\tilde{\xi})$ for every $x \in X$. Alternatively we may define $\tilde{\xi}$ to be a discrete indicator random variable denoting which $x \in X$ has occurred. Then we can consider independence between the two random variables $\tilde{\xi}$ and $\tilde{X}$. A similar indicator random variable, $\tilde{Y}$, may be defined for the $Y$ partition.} Such a test requires only a classificatory measurement scale; an API metric will be a suitable instrument if it takes on one of two values, a zero indicating independence (no association) and a nonzero indicating dependence. Section 4 presents conditions under which the API provides such a measurement.

MODEL EVALUATION

A model evaluation API analysis, of which the Beaver and Dukes [1972] interperiod tax allocation study is a good example, changes the object of the experiment. The researcher assumes that the accounting signal possesses information content and uses this assumption to compare the relative ability of competing models to represent unobservable investor beliefs. A specified security price event (e.g., a positive abnormal return) is taken as evidence that the signal realization fell within a particular $X$ event (positive unexpected earnings). Through its API score,
each competing Y model is examined to determine whether it correctly classified the signal. Model Ya may prove superior to model Yb because Ya transforms the signal (e.g., deferral vs. flow-through earnings) or because Yb is mathematically misspecified relative to Ya (e.g., stationary vs. martingale earnings process).

The model evaluation concept of association is therefore a measure of the congruence between the Y model and investor beliefs concerning $\theta$, with security prices playing the role of an instrumental variable. Such an analysis could be conducted to test the validity of a single Y model, that is, to determine whether X events and Y events are statistically dependent. In order to compare several competing Y models, however, the API must provide a (partial) ordering; if model Ya achieves a higher API score than model Yb, we seek to infer that Ya is a closer approximation to investor beliefs. The Blackwell and MM analyses provide linkages between the statistical properties of a Y model in relation to X and a decision maker’s or researcher’s preferences for models, which lead to an unambiguous definition of model superiority. Section 5 contains a complete presentation of the ranking scheme, together with sufficient conditions for an API representation.

3. Homogeneity and Terminology

Many of the sufficient conditions developed in succeeding sections deal in one way or another with assumptions of homogeneity, and it is convenient to group the assumptions into three classes.

A. Homogeneity across investors with regard to each firm: Investors’ beliefs about X events are treated as though there were a single probability distribution for each firm. This simplifying assumption will be common to all sets of sufficient conditions and the analysis will not incorporate the growing number of theoretical models which explicitly consider heterogeneity of investors’ beliefs.

B. Homogeneity of investor belief across firms: Possible diversity across firms in the partition X, in investors’ probability distribution over the events in X, and in investors’ conditional probability distribution over Z given the X realization, must be considered. Alternative sets of sufficient conditions making different assumptions in this regard are developed in Section 5.

C. Homogeneity of model effectiveness across firms: Model effectiveness is described in terms of joint and conditional distributions involving X and Y. While I generally wish to maintain the notion that “a model works with equal effectiveness on all firms in the sample,” this assumption can be stated with varying levels of strictness.

The analysis considers a binary experiment; it assumes that the signal space $\Theta$ has been partitioned by investors into two events, x and $\bar{x}$ (e.g., positive and negative unexpected earnings) and that all models assign signals to the two events y and $\bar{y}$. This simplification is characteristic of many API studies and leads to simplified exposition and derivation.
However, some of the definitions, test criteria, and sufficient conditions would be significantly more complex if the number of X or Y events increased.\textsuperscript{11}

In correspondence with MM terminology, define the following probability vectors and matrices (firms are indexed by $k = 1, \ldots, N$):\textsuperscript{12}

\begin{equation}
\begin{bmatrix}
P(x_k) \\
P(\bar{x}_k)
\end{bmatrix} = r_k
\end{equation}

\begin{equation}
\begin{bmatrix}
P(y_k) \\
P(\bar{y}_k)
\end{bmatrix} = q_k
\end{equation}

\begin{equation}
\begin{bmatrix}
P(x_k | y_k) & P(x_k | \bar{y}_k) \\
P(\bar{x}_k | y_k) & P(\bar{x}_k | \bar{y}_k)
\end{bmatrix} = \Pi_k, \text{ the posterior probability matrix.}
\end{equation}

\begin{equation}
\begin{bmatrix}
P(y_k | x_k) & P(y_k | \bar{x}_k) \\
P(\bar{y}_k | x_k) & P(\bar{y}_k | \bar{x}_k)
\end{bmatrix} = \Lambda_k, \text{ the likelihood matrix.}
\end{equation}

Note necessary consistency conditions:

\begin{equation}
\Pi_k q_k = r_k,
\end{equation}

\begin{equation}
\Lambda_k^t r_k = q_k, \text{ where } \Lambda_k^t \text{ is the transpose of } \Lambda_k.
\end{equation}

Relations 5 and 6 can be combined as:

\begin{equation}
[\Pi_k \Lambda_k^t - I] r_k = 0,
\end{equation}

where I is an identity matrix and 0 is a zero vector.\textsuperscript{7}

With regard to the homogeneity of model effectiveness, the investigator may, for example, specify that the $\Pi_k$ matrix is homogeneous across firms, or he may strengthen the assumption to include homogeneity of $\Lambda_k$ as well.

API measures are constructed by partitioning a sample of N firms on the basis of their Y event realization and aggregating the Z realizations (return residuals) accordingly:

\begin{equation}
K^+ = \{k \mid \theta_k \in y_k\}
\end{equation}

\begin{equation}
K^- = \{k \mid \theta_k \in \bar{y}_k\}.
\end{equation}

A composite API score is computed as:\textsuperscript{13}

\begin{equation}
\text{API} = \frac{1}{N} \left[ \sum_{k \in K^+} z_k - \sum_{k \in K^-} z_k \right].
\end{equation}
4. Tests of Information Content

Tests of information content examine the association between the random variable $\tilde{z}(s)$ and events $x$ in the partition of the signal space $\Theta$. The null hypothesis of the test is statistical independence.

$$H_n: \phi(\tilde{z} \mid x) = \phi(\tilde{z}) \quad \text{for all} \quad x \in X.$$  \hfill (10)

Most API analyses have concentrated on the first moment of the density function and have drawn the inference that:

$$E(API) \neq 0 \rightarrow \phi(\tilde{z} \mid x) \neq \phi(\tilde{z}),$$  \hfill (11)

where $E(\cdot)$ is the expectation operator. The following four assumptions are sufficient conditions for the API to serve as a classificatory measure of association, as defined by the inference in inequality (11).

A1. HOMOGENEOUS X BELIEFS

All investors agree on the partition $X_k$ of the signal space and on the probability assigned to events $x_k$ and $\bar{x}_k$, for each $k = 1, \ldots, N$.

A2. MODEL PROPERTIES

Model $Y_k$ assigns signals to the partition $\{y_k, \bar{y}_k\}$ such that:

A2.1. $P(x_k \mid y_k) > P(x_k \mid \bar{y}_k)$, \hspace{1cm} $k = 1, \ldots, N$ (II labeling),

A2.2. $\phi(\tilde{z}_k \mid x, y) = \phi(\tilde{z}_k \mid x)$, \hspace{1cm} all $x \in X_k, y \in Y_k$,

$$k = 1, \ldots, N$$ \hspace{1cm} (surrogation).

A3. ZERO EXPECTED ABNORMAL RETURN

$$E(\tilde{z}_k) = E(\tilde{z}_k \mid x_k) P(x_k) + E(\tilde{z}_k \mid \bar{x}_k) P(\bar{x}_k) = 0, \hspace{1cm} k = 1, \ldots, N.$$  

A4. CONDITIONAL EXPECTATION SIGN HOMOGENEITY

$$E(\tilde{z}_k \mid x_k)$$ \text{is of the same sign for all firms $k = 1, \ldots, N$.}

Assumption A1 allows one to consider investors as unanimously classifying signals into events (e.g., positive unexpected earnings) and as maintaining a common belief about the probability of these events. A2.1 asserts both that the model does discriminate between $X$ events (i.e., $X_k$ and $\bar{Y}_k$ are not independent) and that it is properly labeled for each firm. In a binary system, A2.1 also implies:

$$A2.1'. P(\bar{x}_k \mid \bar{y}_k) > P(\bar{x}_k) > P(\bar{x}_k \mid y_k), \hspace{1cm} k = 1, \ldots, N.$$
Thus $y_k$ signifies the event $x_k$ and $\tilde{y}_k$ signifies the event $\tilde{x}_k$. We do not, however, assume that $Y_k$ is a perfect model ($P(x_k \mid y_k) = 1$), but only that it is better than no model at all.\footnote{Marschak and Miyasawa [1968] refer to the condition where the columns of the $\Pi$ matrix each contain 1 one and $m - 1$ zeros as perfect information. This is a special case of noiseless information, in which each row of $\Lambda$ consists of 1 one and $n - 1$ zeros. See Marschak and Miyasawa [1968, p. 143, n. 7]. In a binary system ($m = n = 2$), the conditions are equivalent.} A2.1 also conveys a form of homogeneous model effectiveness assumption, for although it does not require that $r_k, q_k, \Pi_k, \text{ or } \Lambda_k$ be equal for all firms, it assumes that the relation A2.1 between the elements of $r_k$ and $\Pi_k$ holds for all firms.

A2.2 asserts that $Y$ is a surrogate for $X$ alone, implying that the $Y$ event conveys no more information relevant to valuing the firm than would a perfect revelation of the $X$ event itself.\footnote{In the Marschak and Miyasawa [1968] terminology, $Y$ is a garbling of $X$.} Thus, in testing for the information content of an $X$ event such as positive unexpected earnings, the researcher assumes that events in model partition $Y$ are not statistically associated with a separate potentially $Z$-relevant event, such as dividend policy.

Assumption A3 is stronger than necessary for tests of information content on two counts. In testing for $\phi(\hat{z} \mid X) \neq \phi(\hat{z})$, one need only be able to specify or estimate the marginal distribution $\phi(\hat{z})$, rather than set its mean at zero. Further, a change in the mean of a distribution is a sufficient but not a necessary condition for a change in the distribution itself. However, A3 proceeds naturally from the assumption of zero marginal expectation of the residuals of the market model.

A4 is not precisely equivalent either to the Marshall [1975] $w^*$ weighting scheme or to the Barefield et al. [1976] sign-preserving condition. The Marshall $w^*$ scheme demands a common sign of the expectation of $\hat{z}$ for all signals classified into the same $Y$ category. A4 demands a common sign of the expectation of $\hat{z}$ within a particular $X$ (investor perceived) event, but A2.1 leaves open the possibility that the $Y$ model will incorrectly indicate which $X$ event occurred (i.e., $P(x \mid y) \leq 1$).\footnote{Barefield et al. [1976, p. 174, n. 4] acknowledge this alternative representation of their sign-preserving condition.} Note that if both $E(\hat{z} \mid x)$ and $P(x)$ are nonzero, A3 requires that the conditional expectation of $\hat{z}$ be of different signs for the two $X$ events. A4 must be substantively altered when the $X$ partition contains more than two distinct events, perhaps to a statement of the relative magnitudes of the various $E(\hat{z} \mid \tilde{X})$.

The consequences of assumptions A1 through A4 can be demonstrated by taking the expectation of the composite API defined in equation (9).\footnote{Details are presented in Section 1 of Appendix A.}

\begin{equation}
E(API) = \frac{1}{N} \sum_{k=1}^{N} [E(\hat{z}_k \mid y_k)P(y_k) - E(\hat{z}_k \mid \tilde{y}_k)P(\tilde{y}_k)]
\end{equation}

\begin{equation}
= \frac{2}{N} \sum_{k=1}^{N} E(\hat{z}_k \mid x_k)P(x_k)[P(y_k \mid x_k) + P(\tilde{y}_k \mid \tilde{x}_k) - 1].
\end{equation}
Assumption A2.1 ensures that the bracketed probability term in equation (12) is strictly positive for all firms. Therefore, evidence that $E(\text{API}) \neq 0$ implies that there exists at least one firm for which $E(\tilde{z}_k | x_k) \neq 0$, and significance testing over a large sample would extend the inference of association between $X$ and $Z$ to the sample as a whole.

The derivation of equation (12) illustrates the sensitivity of API information content tests to the quality of the model of investor beliefs. If the $Y$ model is weak, that is, if $P(x | y) - P(x)$ is small, the bracketed probability term approaches zero, and even situations in which $E(\tilde{z} | x)$ is significantly different from zero may yield insignificant API scores, preventing the proper rejection of the null hypothesis.

On the other hand, in A2.2 we assume that $Y$ events are not related to signals other than the $\theta \in \Theta$ partitioned by $X$, thereby allowing the substitution:

$$E(\tilde{z} | x, y) = E(\tilde{z} | x). \quad (13)$$

If A2.2 is violated such that $Y$ serves as a surrogate for two different types of signals, there exist examples where $\phi(\tilde{z} | x) = \phi(\tilde{z})$ but $E(\tilde{z} | x, y) \neq E(\tilde{z})$, and the corresponding nonzero expectation of the API results in a misleading inference about the marginal information content of $X$ events.\textsuperscript{18} Thus, the “quality” of the $Y$ model plays a major role in the interpretation of any API analysis, and a “poor” model can lead to either type I or type II errors.

The fact that A1 through A4 are sufficient rather than necessary conditions for the implication given in (11) is apparent. For example, A4 assumes a common sign for $E(\tilde{z}_k | x_k)$, although any “cancelling out” which might occur if the sign of $E(\tilde{z}_k | x_k)$ varied across firms would not invalidate the implication in (11), but merely reduce the power of the API as a test of the null hypothesis. Similarly, demonstration that $E(\tilde{z}_k | x_k) \neq E(\tilde{z}_k)$ is a sufficient but not necessary condition for $\phi(\tilde{z}_k | x_k) \neq \phi(\tilde{z}_k)$. That is, $X$ events may substantially alter the $\phi(\tilde{z})$ distribution in a manner which leaves its expectation unchanged. Nevertheless, the assumptions of homogeneous $X$ beliefs, a proper model, zero marginal expected abnormal return, and sign homogeneity of the conditional expected abnormal return seem to be consistent with the spirit of most API studies and permit the use of the API as a measure of the information content association concept.

5. Model Evaluation

In model evaluation API studies, the investigator seeks to infer whether one model of investor beliefs is superior to another. The goal of this section is the development of sets of sufficient conditions under which API “scores” can rank models according to a researcher’s preferences. In

\textsuperscript{18} See especially Gonedes [1978] for a discussion of this phenomenon and tests of the marginal and joint information content of earnings, dividends, and extraordinary items.
addition, I show that different sets of sufficient conditions imply different experimental designs or calculations.

The Blackwell [1951; 1953] and Marschak and Miyasawa [1968] analyses relate the statistical properties of models or experimental designs to a partial ordering of models by researcher preference. The researcher is depicted as a decision maker who must choose an action and who will receive a payoff which depends both on the action chosen and on the realization of the X event. The X event is not directly observable at the decision point, but the researcher may choose among various models which are stochastically related to X, and he engages in a model evaluation API analysis of security returns in order to identify a superior model. I assume that the researcher will rank model $Y_a$ superior to model $Y_b$ if the use of $Y_a$ (as an aid in action choice) leads to an expected utility which is greater than or equal to the expected utility attainable with $Y_b$.

This procedure is entirely analogous to the selection of an estimation technique which minimizes mean-squared error. In that case, the investigator prefers an estimator which maximizes his expected utility in situations where his action is the estimation of a parameter and his payoff will be reduced by an amount proportional to the squared error of estimation. Blackwell expands this concept to all payoff functions for which the partition $[x, x^{-}]$ is payoff-relevant.

As in Sections 3 and 4, attention is restricted to binary systems in which all models are labeled as in Assumption A2.1. Following the MM condition (II), define the model evaluation concept of association:

Model $Y_a$ is more closely associated with X than is model $Y_b$ (written $Y_a > Y_b$) if the researcher's expected payoff using model $Y_a$ is greater than his expected payoff using $Y_b$ for all payoff functions for which the partition $[x, x^{-}]$ is payoff relevant, and for all $q_a, q_b, r$ vectors satisfying the consistency relation (7).

MM prove the following theorem:

Condition K: If binary models $Y_a$ and $Y_b$ are both labeled as in A2.1, then:

$$Y_a > Y_b \text{ if } \left\{ \begin{array}{l} P(x | y_a) \geq P(x | y_b) \\ P(\tilde{x} | \tilde{y}_a) \geq P(\tilde{x} | \tilde{y}_b). \end{array} \right. \quad (14a)$$

19 MM define the payoff function as being measured in units of utility, and except for a change in the notational representation of the arguments (actions and state outcomes), the terms payoff function and utility function are interchangeable. See Marschak and Miyasawa [1968, p. 139, eq. 2.2].

20 The MM analysis considers all payoff functions for which the partition $[x, x^{-}]$ is payoff-adequate, where a payoff-adequate partition is a subpartition of a payoff-relevant partition. However, a dichotomous partition is a subpartition only of the sample space as a whole.

21 Marschak and Miyasawa [1968, theorem 10.2]. Theorem 10.2 follows from theorems 10.1 and 9.8 and recognition that in a binary system, the columns of $\Pi$ will be linearly independent except in the case of a totally noninformative model, i.e., $P(x | y) = P(x | \tilde{y})$. 

This content downloaded from 165.230.34.196 on Tue, 14 Aug 2018 20:37:35 UTC
All use subject to https://about.jstor.org/terms
Condition $K$ allows one to rank models either by directly comparing elements of their $\Pi$ matrices or by analyzing the models' relationships with security price behavior in a manner which implies the comparisons in (14a) and (14b).\textsuperscript{22} An immediate first step, common to all the sufficient condition sets which follow, is the assumption that each firm represents a drawing from the same $\Pi$ matrix. Assumption $A2$ is therefore strengthened to include $\Pi$ homogeneity.

$A2^*$. MODEL PROPERTIES

Models $Y_a$ and $Y_b$ assign signals to the partition $[y_k, \bar{y}_k]$ such that:

$$
A2.0^* \begin{cases} 
\Pi_{ka} = \Pi_a; \\
\Pi_{kb} = \Pi_b,
\end{cases} \quad k = 1, \ldots, N. \quad (\Pi \text{ homogeneity})
$$

$$
A2.1^* \begin{cases} 
P(x | y_a) > P(x_k) > P(x | \bar{y}_a); \\
P(x | y_b) > P(x_k) > P(x | \bar{y}_b),
\end{cases} \quad k = 1, \ldots, N. \quad (\Pi \text{ labeling})
$$

$$
A2.2^* \begin{cases} 
\phi(\tilde{z}_k | \bar{X}_k, \bar{Y}_{ka}) = \phi(\tilde{z}_k | \bar{X}_k); \\
\phi(\tilde{z}_k | \bar{X}_k, \bar{Y}_{kb}) = \phi(\tilde{z}_k | \bar{X}_k),
\end{cases} \quad k = 1, \ldots, N. \quad \text{(surrogation)}\textsuperscript{23}
$$

$A2.0^*$ treats the $X$ event realizations for firms in the same $Y$ event category as identically distributed.\textsuperscript{24} Note that this does not set the $r_k$ vector (the prior probabilities of $x_k$ and $\bar{x}_k$) equal for all firms, but $A2.1^*$ does restrict its domain.

As indicated in figure 1, model evaluation $API$ studies use security price behavior in order to draw inferences about the association between observable $Y$ events and unobservable $X$ events. Therefore we must assume that, in some way, observable security prices are related to $X$ events.

$A5$. X-RELEVANT ABNORMAL RETURNS

$$
E(\tilde{z}_k | x_k) > 0. \quad k = 1, \ldots, N.
$$

$A5$ states that $X$ events do possess information content; it assumes that the null hypothesis in (10) is false.\textsuperscript{25} While $A5$ is presented as a sufficient condition, and will later be stated in a different form, it seems that some type of $X$-relevance of security returns is necessary for model evaluation tests. One either assumes that a $Y$ model is valid and uses it to test for security price association with $X$ events (information content), or one assumes that security returns are associated with $X$ events and tests the

\textsuperscript{22} Condition $K$ allows a partial ordering, in the sense that two models may exist for which one but not both of (14a) and (14b) are satisfied. To test whether a single model $Y_a$ is “valid,” consider $Y_b$ to be the null or noninformative model with $P(x | y) = P(x) = P(x | \bar{y})$.

\textsuperscript{23} The notation of $A2.2^*$ uses the indicator random variables $X$ and $Y$ defined in n. 10.

\textsuperscript{24} The investigator may wish to extend this assumption to include both identically and independently distributed conditional $X$ events or to make specific allowance for assumed interfirm dependencies, in performing significance tests.

\textsuperscript{25} $A5$ also implies $A4$.  

This content downloaded from 165.230.34.196 on Tue, 14 Aug 2018 20:37:35 UTC
All use subject to https://about.jstor.org/terms
ability of a model to "match up" to the $X$ events which observed price changes have instrumentally categorized.

I now construct three alternative sets of sufficient conditions under which the API provides an ordinal measure of model superiority, or $X-Y$ association. The sets differ with respect to interfirm homogeneity, either of investor belief or model effectiveness; each set implies different API computations in order to produce the condition $K$ comparisons.

Set I

To assumptions A1, A2*, A3, and A5, append assumption A6*.

A6*. STRONG MODEL HOMOGENEITY

Models $Y_a$ and $Y_b$ assign signals to the partition $[y_k, \bar{y}_k]$ such that:

$$
A6.1^* \begin{cases} 
\Lambda_{ka} = \Lambda_a; \\
\Lambda_{kb} = \Lambda_b,
\end{cases} 
\quad k = 1, \ldots, N.
$$

$$
A6.2^* \begin{cases} 
P(y_a | x) = P(\bar{y}_a | \bar{x}); \\
P(y_b | x) = P(\bar{y}_b | \bar{x}).
\end{cases}
$$

When A6.1* is combined with A2*, the investigator has assumed that, for each model, both the posterior matrix and the likelihood matrix do not vary across firms. While these conditions do not appear overly restrictive in themselves, their joint effect is to fix the $r_k$ vector (the prior probability of $X$ events) at a common value for all firms, through the consistency relation (7). A6.2* states further that a model's likelihood of generating the "correct" signal is the same for both events, $x$ and $\bar{x}$.

Assumption-set I allows the use of the composite API defined by equations (8) and (9). Composite scores $API_a$ and $API_b$ are constructed for each model, and the expected value of their difference can be derived from equation (12):

$$
E(API_a) - E(API_b) = \frac{2}{N} \sum_{k=1}^{N} E(\hat{z}_k | x_k) P(x_k) \left[ P(y_a | x) + P(\bar{y}_a | \bar{x}) - P(y_b | x) - P(\bar{y}_b | \bar{x}) \right]
$$

$$
= 4 \sum_{k=1}^{N} E(\hat{z}_k | x_k) P(x_k) [P(y_a | x) - P(y_b | x)].
$$

Assumption A5 puts a common positive sign on the conditional expectation of $\hat{z}_k$, and A6.1* equates each model's $\Lambda$ probabilities across firms.

---

26 A6.1 and A2* fix $r_k = r, k = 1, \ldots, N$ except when both $Y_a$ and $Y_b$ are perfect models (see n. 14), in which case $\Pi \Lambda' = I$, and $r_a$ remains a free parameter.
A6.2* then allows the following two implications:

\[ E(API_a) > E(API_b) \rightarrow \begin{cases} 
P(y_a | x) > P(y_b | x) \\
\text{and} \\
P(\tilde{y}_a | \tilde{x}) > P(\tilde{y}_b | \tilde{x}).
\end{cases} \tag{16a,b} \]

It is easy to demonstrate that (16a) and (16b) imply condition \( K \), and therefore:

\[ E(API_a) > E(API_b) \rightarrow Y_a > Y_b. \tag{17} \]

Thus, the additional assumptions of \( X \)-relevant abnormal returns and strong model homogeneity are sufficient for the composite \( API \) to provide a ranking of models by a very general form of researcher preference. A6.2* is motivated by the existence of examples for which, in the absence of A6.2*, the expected composite \( API \) difference in equation (15) is positive, but only one of the two inequalities of condition \( K \) is satisfied.\(^{28}\)

In set II, the strong model homogeneity condition is partially relaxed, but in return a stronger form of interfirm return distribution homogeneity must be assumed. These changes motivate a different \( API \) computation.

**Set II**

To assumptions A1, A2*, and A3, append assumptions A5* and A6.

**A5*. HOMOGENEOUS ABNORMAL RETURN DISTRIBUTIONS**

\[ \begin{align*}
A5.1^* & \quad \phi(\tilde{z}_k | x_k) = \phi(\tilde{z} | x); \\
& \quad \phi(\tilde{z}_k | \tilde{x}_k) = \phi(\tilde{z} | \tilde{x}), \quad k = 1, \ldots, N.
\end{align*} \]

\[ A5.2^* \quad E(\tilde{z} | x) > 0. \]

**A6. MODEL HOMOGENEITY**

Models \( Y_a \) and \( Y_b \) are constructed such that:

\[ \begin{align*}
A6.1 & \quad \Lambda_{ka} = \Lambda_a; \\
& \quad \Lambda_{kb} = \Lambda_b, \quad k = 1, \ldots, N.
\end{align*} \]

A6 relaxes the assumption of equal model effectiveness in both the \( x \) and \( \tilde{x} \) events, and therefore separate \( API \) scores are computed for each \( Y \) event for each model.

\[ \begin{align*}
K^+_a &= \{ k | \theta_k \in y_a \}; \quad K^+_a \text{ contains } N^+_a \text{ elements.} \tag{18a} \\
K^+_b &= \{ k | \theta_k \in y_b \}; \quad K^+_b \text{ contains } N^+_b \text{ elements.} \tag{18b}
\end{align*} \]

\(^{27}\) Details are presented in section 2 of Appendix A.

\(^{28}\) Such an example is presented in section 3 of Appendix A.
\( K_a^- = \{ k | \theta_k \in \bar{y}_a \} \); \( K_a^- \) contains \( N_a^- \) elements. (18c)

\( K_b^- = \{ k | \theta_k \in \bar{y}_b \} \); \( K_b^- \) contains \( N_b^- \) elements. (18d)

Denote the separate API scores for the \( y \) and \( \bar{y} \) events as \( API^+ \) and \( API^- \). The expected value of the difference between the \( API^+ \) scores achieved by the two models is given by:

\[
E(API_{a+}) - E(API_{b+}) = E\left[ \frac{1}{N_a^+} \sum_{k \in K_a^+} z_k \right] - E\left[ \frac{1}{N_b^+} \sum_{k \in K_b^+} z_k \right].
\] (19)

Despite a sample of fixed total size \( N \), \( N_a^+ \) and \( N_b^+ \) are random variables. Further, examples can be constructed where variations in \( E(z_k | x_k) \) across firms can lead to an expected difference in API scores which is positive, negative, or zero, regardless of relative model effectiveness. Assumption A5.1* removes these problems by allowing each term in equation (19) to be represented as the expected value of the average of a random number of drawings from a common distribution. Given event \( y_a, \bar{z} \) will be drawn from \( \phi(\bar{z} | x) \) with probability \( P(x | y_a) \), or from \( \phi(\bar{z} | x) \) with probability \( P(\bar{x} | y_a) \). The random nature of \( N_a^+ \) and \( N_b^+ \) can be ignored in taking expectations, and the expected value of the average of any number of drawings in \( K_a^+ \), for example, is given by:

\[
E(API_{a+}) = \int_{\bar{z}} \bar{z} [\phi(\bar{z} | x)P(x | y_a) + \phi(\bar{z} | x)P(\bar{x} | y_a)]
\]

\[
= E(\bar{z} | x)P(x | y_a) + E(\bar{z} | x)P(\bar{x} | y_a)
\]

\[
= E(\bar{z} | x) \left[ \frac{P(x | y_a) - P(x)}{P(\bar{x})} \right].
\] (20)

The expected value of the difference in API scores then reduces to:

\[
E(API_{a+}) - E(API_{b+}) = E(\bar{z} | x) \left[ \frac{P(x | y_a) - P(x | y_b)}{P(\bar{x})} \right].
\] (21a)

By a completely symmetric derivation, one obtains:

\[
E(API_{a^-}) - E(API_{b^-}) = -E(\bar{z} | \bar{x}) \left[ \frac{P(\bar{x} | y_a) - P(\bar{x} | y_b)}{P(x)} \right].
\] (21b)

Thus, if we compare separately the expected values of the \( API^+ \) and \( API^- \) scores of two competing models and find superiority in both regions, assumption-set II allows the inference of researcher preference:

\[
E(API_{a+}) > E(API_{b+}) \quad \text{and} \quad E(API_{a^-}) > E(API_{b^-}) \quad \Rightarrow \quad Y_a > Y_b.
\] (22)

In passing from set I to set II, the investigator exchanges homogeneity of model effectiveness across \( X \) events for homogeneity of the conditional
distribution of abnormal returns across firms. Even with procedures which standardize the estimated residuals, A5.1* may be overly restrictive in some experimental settings.\textsuperscript{29} In set III, the assumption of homogeneity of the conditional abnormal return distribution is relaxed by concentrating only on the sign of the $\tilde{z}$ realizations. Furthermore, the strengthened model homogeneity assumptions A6 or A6* are completely removed. The subsequent derivation of an API “score” reveals a particular sampling problem of model evaluation studies.

**Set III**

To assumptions A1 and A2*, append assumption A5**.

A5**. WEAK ABNORMAL RETURN DISTRIBUTION HOMOGENEITY

\begin{equation}
\begin{align*}
A5.1** & \quad P(\tilde{z}_k > 0 \mid x_k) = P(\tilde{z} > 0 \mid x); \\
& \quad P(\tilde{z}_k < 0 \mid \tilde{x}_k) = P(\tilde{z} < 0 \mid \tilde{x}), \quad k = 1, \ldots, N. \\
A5.2** & \quad P(\tilde{z} > 0 \mid x) > P(\tilde{z} > 0 \mid \tilde{x}).
\end{align*}
\end{equation}

A5.1** is a relaxation of A5.1* in that it requires homogeneity only of the conditional probability of a positive abnormal return, rather than homogeneity of the entire conditional abnormal return distribution. A5.2** is similar to A5.2*; it assumes that a positive abnormal return is more likely given an $x$ event (positive unexpected earnings) than given an $\tilde{x}$ event. A5.2** assumes, in essence, that the probability of a “misleading” abnormal return is small, and A5.1** equates the probability of this type of binary measurement error across firms.

Consider a comparison of the relative frequencies with which two competing models correctly classify the signs of abnormal returns, as suggested by the matrices in figure 2.

The expected relative frequency of correct classification of positive abnormal returns by model $Y_a$ is given by:

\begin{equation}
P(\tilde{z} > 0 \mid y_a) = P(\tilde{z} > 0 \mid x)P(x \mid y_a) + P(\tilde{z} > 0 \mid \tilde{x})P(\tilde{x} \mid y_a) \\
= P(x \mid y_a)[P(\tilde{z} > 0 \mid x) - P(\tilde{z} > 0 \mid \tilde{x})] + P(\tilde{z} > 0 \mid \tilde{x}).
\end{equation}

The expected differences in relative frequency of correct classification by competing models in each region are given by:

\begin{align*}
P(\tilde{z} > 0 \mid y_a) & - P(\tilde{z} > 0 \mid y_b) \\
& = [P(x \mid y_a) - P(x \mid y_b)][P(\tilde{z} > 0 \mid x) - P(\tilde{z} > 0 \mid \tilde{x})]. \quad (24a) \\
P(\tilde{z} < 0 \mid \tilde{y}_a) & - P(\tilde{z} < 0 \mid \tilde{y}_b) \\
& = [P(\tilde{x} \mid \tilde{y}_a) - P(\tilde{x} \mid \tilde{y}_b)][P(\tilde{z} < 0 \mid \tilde{x}) - P(z < 0 \mid x)]. \quad (24b)
\end{align*}

\textsuperscript{29} For examples of such standardization procedures, see Gonedes, Dopuch, and Penman [1976] or Patell [1976].
A5** ensures that the rightmost bracketed terms are positive and equal for all firms. Therefore, comparison of the relative frequencies with which two models have correctly classified the direction of abnormal price change (i.e., separate comparison of the corresponding columns of the classification matrices in fig. 2) may reveal whether one model is uniformly preferable.

\[
P(\tilde{z} > 0 \mid y_a) > P(\tilde{z} > 0 \mid y_b) \quad \text{and} \quad P(\tilde{z} < 0 \mid \tilde{y}_a) > P(\tilde{z} < 0 \mid \tilde{y}_b) \implies Y_a > Y_b. \tag{25a}
\]

Examination of the binary classification matrices in figure 2, coupled with an insight noted by Beaver and Dukes [1972], highlights a potential weakness of model evaluation API studies. Whether we compare the magnitudes of API scores, as in equations (15) or (21), or the relative frequencies of correct classification, as in equation (24), the observed differences between models derive solely from those firms for which the models disagree about which \( x \) event occurred; the only relevant observations lie in two groups which may be very small subsets of the full sample. Define the model disagreement events:

\[
D^1 = \{ \theta \mid \theta \in y_a \cap \tilde{y}_b \}. \tag{26a}
\]
\[
D^2 = \{ \theta \mid \theta \in \tilde{y}_a \cap y_b \}. \tag{26b}
\]
\[
D = D^1 \cup D^2. \tag{26c}
\]
\[
K_{D1} = K_a^+ \cap K_b^- \tag{27a}
\]
\[
K_{D2} = K_a^- \cap K_b^+ \tag{27b}
\]

If A1, A2*, and A5** hold, one can draw precisely the same inferences made in relations (24) and (25) by restricting attention only to the disagreement subsets; the following test is equivalent to implication (25):

\[
P(y_a \mid D, \tilde{z} > 0) > P(\tilde{y}_a \mid D, \tilde{z} > 0) \quad \text{and} \quad P(y_a \mid D, \tilde{z} < 0) > P(\tilde{y}_a \mid D, \tilde{z} < 0) \implies Y_a > Y_b. \tag{28a}
\]

To conduct such a test, we need only examine the rows of the model disagreement classification matrix illustrated in figure 3.\(^{31}\) If the frequency in cell 1 significantly exceeds the frequency in cell 2 and the frequency in cell 4 exceeds that of cell 3, condition \( K \) is satisfied; we infer that model \( Y_a \) is preferable to model \( Y_b \).

The form of the implication in (28) makes explicit the notion that one

\(^{30}\) A proof of equivalence is presented in section 4 of Appendix A.

\(^{31}\) Note that 28 reverses the order of the conditioning in 25, shifting the comparison from columns in figure 2 to rows in figure 3.
compares the relative $X$-$Y$ association of two competing models by examining the cases in which they disagree. Unfortunately, these observations may form a small fraction of the total sample, and they may consist of exactly those firms for which homogeneity assumptions, either of model effectiveness or of conditional return probabilities, may be tenuous.\textsuperscript{32} Further, condition $K$ requires model superiority in both disagreement events $D^1$ and $D^2$ in order to infer researcher preference.

5. Summary

I began with a representation of API experimental design that involves three variables: investor beliefs about accounting signals, observable models of investor beliefs, and security prices. Association between beliefs and prices (information content) is distinguished from association between beliefs and models (model evaluation) both by basic definition and by measurement properties. In general, an API study must adopt one form of association as a maintained assumption in order to test the other.

Information content is defined as statistical dependence between the probability distribution of security returns and investor beliefs about accounting signals. The API can provide a classificatory (nominal) mea-

\textsuperscript{32} In the Beaver and Dukes [1972] study of interperiod tax allocation, the model disagreement sets were approximately 10 to 15 percent of the original sample. See Beaver and Dukes [1972, pp. 329-31].
sure of this type of association under a set of sufficient conditions concerning homogeneity of investors’ beliefs concerning signals, the adequacy of the model of investors’ beliefs, and the probability distribution of abnormal returns. The analysis focused attention on the role of the model and indicated that a weak model reduces the power of the test, while a model correlated with extraneous signals can lead to false rejection of the null hypothesis of no information content.

Many API studies have involved several different models of investors’ beliefs, and the Blackwell and Marschak and Miyasawa analyses provide a criterion for model evaluation in terms of the researcher’s preferences. Three alternative sets of sufficient conditions were developed under which an API test can provide an ordinal measure of model ability; the sets differ with respect to the homogeneity of model effectiveness across firms and the characterization of the assumed information content of the accounting signal.

One can envision the combination and extension of these two association concepts to consider the question of whether one type of accounting information system is more informative than another. A criterion similar to the Blackwell-Marschak-Miyasawa condition $K$ could be invoked, either by considering $\theta$ to be a very general space of accounting data on which different systems impose different partitions, or by considering separate $\theta$ spaces with separate (although stochastically linked) event partitions. In either case, the problem of constructing expectation models of equivalent effectiveness for each system must be solved in order to ensure that the system deemed inferior is not merely the victim of a poorly specified model of investors’ beliefs.

The two definitions of association discussed here are not the only ways in which the concept could be considered, and different definitions would lead to different sufficient conditions and different measurement techniques. In particular, the model evaluation criterion is very strong, and it is possible that many sets of competing models could not be completely ordered according to the criterion through security price analysis. However, the conditions developed here appear to be consistent with the methodology and inferences of prior API studies, and they make explicit those assumptions which have served as an implicit framework for much of the extant empirical research in financial accounting.

APPENDIX A

1. $E(API) = \frac{1}{N} \sum_{k=1}^{N} \left[ E(\tilde{z} \mid y) P(y) - E(\tilde{z} \mid \bar{y}) P(\bar{y}) \right]$

   $= \frac{1}{N} \sum_{k=1}^{N} \left[ E(\tilde{z} \mid x, y) P(x, y) + E(\tilde{z} \mid \bar{x}, y) P(\bar{x}, y) \right]$

---

33 See Gonedes [1978] for an experimental design which considers the marginal and joint information content of three types of accounting signals: earnings, dividends, and extraordinary items.
\[ - E(\tilde{z} \mid x, \tilde{y}) P(x, \tilde{y}) - E(\tilde{z} \mid \bar{x}, \tilde{y}) P(\bar{x}, \tilde{y}) \]

\[ = \frac{1}{N} \sum_{k=1}^{N} \left\{ E(\tilde{z} \mid x) \{ P(x, y) - P(x, \tilde{y}) \} - E(\tilde{z} \mid x) \{ P(\bar{x}, \tilde{y}) - P(\bar{x}, y) \} \right\} \]

\[ = \frac{1}{N} \sum_{k=1}^{N} E(\tilde{z} \mid x) \left\{ \left( P(x, y) - P(x, \tilde{y}) \right) + \frac{P(x)}{P(\bar{x})} \left( P(\bar{x}, \tilde{y}) - P(\bar{x}, y) \right) \right\} \]

\[ = \frac{1}{N} \sum_{k=1}^{N} E(\tilde{z} \mid x) \left\{ \left( P(x, y) - [P(x) - P(x, y)] \right) \right\} \]

\[ + \frac{P(x)}{P(\bar{x})} \left\{ P(\bar{x}, \tilde{y}) - [P(\bar{x}) - P(\bar{x}, \tilde{y})] \right\} \]

\[ = \frac{1}{N} \sum_{k=1}^{N} E(\tilde{z} \mid x) P(x) \left\{ \left[ \frac{2P(x, y)}{P(x)} - 1 \right] + \left[ \frac{2P(\bar{x}, \tilde{y})}{P(\bar{x})} - 1 \right] \right\} \]

\[ = \frac{2}{N} \sum_{k=1}^{N} E(\tilde{z} \mid x) P(x) \left[ P(y \mid x) + P(\tilde{y} \mid \bar{x}) - 1 \right]. \]

To show that the bracketed probability term must be positive, note:

\[ A2.1 \rightarrow \frac{P(x, y)}{P(y)} > P(x) \rightarrow P(y \mid x) > P(y). \]

\[ A2.1' \rightarrow \frac{P(\bar{x}, \tilde{y})}{P(\tilde{y})} > P(\bar{x}) \rightarrow P(\tilde{y} \mid \bar{x}) > P(\tilde{y}). \]

Addition of the two rightmost inequalities yields:

\[ P(y \mid x) + P(\tilde{y} \mid \bar{x}) > P(y) + P(\tilde{y}) = 1. \]

2. Given \( P(y_a \mid x) > P(y_b \mid x) \) and \( P(\tilde{y}_a \mid \bar{x}) > P(\tilde{y}_b \mid \bar{x}) \),

\[ \frac{P(y_a \mid x)}{P(y_b \mid x)} > 1 > \frac{P(y_a \mid \bar{x})}{P(y_b \mid \bar{x})} \]

\[ P(y_a \mid x) P(y_b \mid \bar{x}) > P(y_a \mid \bar{x}) P(y_b \mid x) \]

\[ P(y_a \mid x) P(y_b \mid \bar{x}) P(\bar{x}) > P(y_a \mid \bar{x}) P(y_b \mid x) P(\bar{x}) \]

\[ + P(y_a \mid \bar{x}) P(y_b \mid x) P(\bar{x}) \frac{P(\bar{x})}{P(x)} > P(y_a \mid x) P(y_b \mid x) P(x) \]

\[ + P(y_a \mid \bar{x}) P(y_b \mid x) P(\bar{x}) \frac{P(\bar{x})}{P(\bar{x})} > P(y_a \mid x) P(y_b \mid x) P(x) \]

\[ > P(y_b \mid x) \left[ P(y_a \mid x) P(x) + P(y_a \mid \bar{x}) P(\bar{x}) \right] \]

\[ \frac{P(y_a \mid x)}{P(y_a)} > \frac{P(y_b \mid x)}{P(y_b)} \]
\[ P(x | y_a) > P(x | y_b). \]

An entirely symmetric development proves that \( P(\tilde{x} | \tilde{y}_a) > P(\tilde{x} | \tilde{y}_b). \)

3. Consider the example:

\[
\begin{align*}
P(x, y_a, y_b) &= 5/20 & P(\tilde{x}, y_a, y_b) &= 1/20 \\
P(x, y_a, \tilde{y}_b) &= 2/20 & P(\tilde{x}, y_a, \tilde{y}_b) &= 3/20 \\
P(x, \tilde{y}_a, y_b) &= 1/20 & P(\tilde{x}, \tilde{y}_a, y_b) &= 2/20 \\
P(x, \tilde{y}_a, \tilde{y}_b) &= 1/20 & P(\tilde{x}, \tilde{y}_a, \tilde{y}_b) &= 5/20 \\
P(x) &= 9/20 & P(\tilde{x}) &= 11/20 \\
P(x | y_a) &= 7/11 > P(x) \\
P(\tilde{x} | \tilde{y}_a) &= 7/9 > P(\tilde{x}) \rightarrow y_a \text{ properly labeled.} \\
P(x | y_b) &= 6/9 > P(x) \\
P(\tilde{x} | \tilde{y}_b) &= 8/11 > P(\tilde{x}) \rightarrow y_b \text{ properly labeled.} \\
P(x | y_a) < P(x | y_b) \rightarrow y_b \text{ superior for } x. \\
P(\tilde{x} | \tilde{y}_a) > P(\tilde{x} | \tilde{y}_b) \rightarrow \tilde{y}_a \text{ superior for } \tilde{x}. \\
P(y_a | x) + P(\tilde{y}_a | x) - P(y_b | x) - P(\tilde{y}_b | \tilde{x}) = +2/99.
\end{align*}
\]

4. Using simplified notation, let:

\[
\begin{align*}
P(\tilde{z} > 0, y_a, y_b) &= a \\
P(\tilde{z} > 0, y_a, \tilde{y}_b) &= b \\
P(\tilde{z} > 0, \tilde{y}_a, y_b) &= c \\
P(\tilde{z} > 0, \tilde{y}_a, \tilde{y}_b) &= d \\
P(\tilde{z} < 0, y_a, y_b) &= e \\
P(\tilde{z} < 0, y_a, \tilde{y}_b) &= f \\
P(\tilde{z} < 0, \tilde{y}_a, y_b) &= g \\
P(\tilde{z} < 0, \tilde{y}_a, \tilde{y}_b) &= h.
\end{align*}
\]

In this notation:

\[
\begin{align*}
(28a) \quad \frac{b}{b + c} &> \frac{c}{b + c} \quad \text{or} \quad b > c, \\
(28b) \quad \frac{g}{f + g} &> \frac{f}{f + g} \quad \text{or} \quad g > f,
\end{align*}
\]

and (25a) and (25b) follow trivially:

\[
\begin{align*}
(25a) \quad \frac{a + b}{a + b + e + f} &> \frac{a + c}{a + c + e + g} \quad \text{or} \quad \frac{e + f}{a + b} < \frac{e + g}{a + c} \\
(25b) \quad \frac{g + h}{c + d + g + h} &> \frac{f + h}{b + d + f + h} \quad \text{or} \quad \frac{c + d}{g + h} < \frac{b + d}{f + h}
\end{align*}
\]

REFERENCES

API AND DESIGN OF EXPERIMENTS 549


