JOURNAL OF MATHEMATICAL PSYCHOLOGY 12, 387–415 (1975)

The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph

DONALD BAMBER

Psychology Service, Veterans Administration Hospital, St. Cloud, Minnesota 56301

Receiver operating characteristic graphs are shown to be a variant form of ordinal dominance graphs. The area above the latter graph and the area below the former graph are useful measures of both the size or importance of a difference between two populations and/or the accuracy of discrimination performance. The usual estimator for this area is closely related to the Mann–Whitney U statistic. Statistical literature on this area estimator is reviewed. For large sample sizes, the area estimator is approximately normally distributed. Formulas for the variance and the maximum variance of the area estimator are given. Several different methods of constructing confidence intervals for the area measure are presented and the strengths and weaknesses of each of these methods are discussed. Finally, the Appendix presents the derivation of a new mathematical result, the maximum variance of the area estimator over convex ordinal dominance graphs.

ORDINAL DOMINANCE GRAPHS

Suppose two random variables X and Y are given. Let c be an arbirary constant. Consider a graph in which a point is plotted having as its horizontal coordinate $P(X \leqslant c)$ and as its vertical coordinate $P(Y \leqslant c)$. Let this point be denoted by $T(c)$. Suppose that, for all values of c from $-\infty$ to $+\infty$, a point $T(c)$ is plotted on this graph. Following Darlington (1973), this graph will be called the ordinal dominance (*OD*) graph for X and Y, or the (X, Y) *OD* graph. More specifically, this graph is a particular type of *OD* graph, namely a population *OD* graph. A second type of *OD* graph, the sample *OD* graph, will be discussed later in this paper.

One interesting property of *OD* graphs is that they are invariant under order-preserving transformations. Let m be a strictly increasing function which is defined over all possible values of the random variables X and Y. Then, the *OD* graph for X and Y is identical to the *OD* graph for the random variables $m(X)$ and $m(Y)$.

OD Graphs for Continuous and Finitely Discrete X and Y

This paper is concerned primarily with the *OD* graphs of random variables having either continuous or finitely discrete distributions. By a finitely discrete random variable is meant a random variable which can equal only a finite number of values with nonzero probability. There are two reasons for this restricted concern. First, it is convenient. Second, most distributions encountered in psychological applications are either continuous or finitely discrete. Occasionally, however, other types of distributions will be considered here.

Continuous X and Y. Suppose X and Y are continuous. Consider the point $T(c)$. The two coordinates $P(X \leqslant c)$ and $P(Y \leqslant c)$ of $T(c)$ are always between zero and one. Thus, $T(c)$ is always located within the unit square whose corners are the points $(0, 0)$, $(1, 0)$, $(0,1)$ and $(1, 1)$. When c equals $-\infty$, $T(c)$ is at $(0, 0)$. Then, as c is increased, $T(c)$ traces out a continuous curve which finally ends at $(1, 1)$ when c equals $+\infty$. Thus, when X and Y are continuous, their *OD* graph is a continuous curve running from $(0, 0)$ to $(1, 1)$. This is illustrated on the left side of Fig. 1.

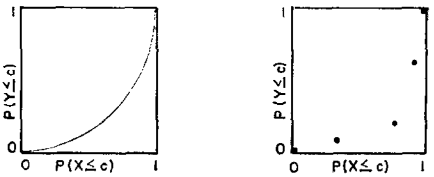

FIG. 1. Two examples of *OD* graphs. Left: Continuous X and Y. Right: Finitely discrete X and Y.

Finitely Discrete X and Y. Suppose X and Y are finitely discrete. Let $c_1 ,..., c_k$ denote the ordered set of constants (i.e., $-\infty < c_1 < \cdots < c_k < +\infty$) which comprise the sole values which either X or Y can assume with nonzero probability. Thus, for $i = 1,..., k$, either $P(X = c_i) > 0$ or $P(Y = c_i) > 0$. In addition, let c_0 and c_{k+1} denote $-\infty$ and $+\infty$, respectively. As before, $T(c)$ is the point having coordinates $P(X \leqslant c)$ and $P(Y \leqslant c)$. Note that $T(c_0) \neq T(c_1) \neq \cdots \neq T(c_k)$ whereas $T(c_k) = T(c_{k+1})$. Also, note that, if $c_i \leqslant c < c_{i+1}$, then $T(c) = T(c_i)$ $(i = 0,..., k)$. Thus, although there are an infinite number of possible values of c, nevertheless the *OD* graph for X and Y consists of only $k + 1$ distinct points: $T(c_0)$, $T(c_1)$,..., $T(c_k)$. Moreover, two of these points, namely $T(c_0)$ and $T(c_k)$, equal $(0, 0)$ and $(1, 1)$, respectively. An example of an *OD* graph for finitely discrete X and Y is given on the right side of Fig. 1. Any two distinct points of the form $T(c_i)$ and $T(c_{i-1})$ $(i = 1,..., k)$ will be referred to as adjacent points of the *OD* graph.

Generation of a Finitely Discrete OD Graph From a Continuous OD Graph. Let X and Y be two continuous random variables. Thus, their *OD* graph is a continuous curve running from $(0, 0)$ to $(1, 1)$. Let s be an arbitrary increasing step function which is defined over the real line and which consists of only a finite number of steps. Specifically, let the domain of the step function s be divided into k intervals by the ordered set of constants $\gamma_1 ,..., \gamma_{k-1}$. Let $c_1 ,..., c_k$ be an ordered set of constants comprising the range of the step function s. In addition, let γ_0 and γ_k denote $-\infty$ and $+\infty$, respectively, and let c_0 be an arbitrary value less than c_1. Then, if $\gamma_{i-1} < \gamma \leqslant \gamma_i$, $s(\gamma) = c_i$ $(i = 1,..., k)$.

Obviously, $s(X)$ and $s(Y)$ are finitely discrete random variables. The *OD* graph for $s(X)$ and $s(Y)$ consists of the $k + 1$ points, respectively, having horizontal coordinates $P[s(X) \leqslant c_i]$ and vertical coordinates $P[s(Y) \leqslant c_i]$, $i = 0,..., k$. (Note however that these points are not necessarily all distinct.) Now it follows from the definition of s that, for $i = 0,..., k$,

$$P[s(X) \leqslant c_i] = P[X \leqslant \gamma_i]$$

and

$$P[s(Y) \leqslant c_i] = P[Y \leqslant \gamma_i].$$

This shows that every point of the $[s(X), s(Y)]$ *OD* graph is also a point on the (X, Y) *OD* graph. Thus, the latter *OD* graph, which is a continuous curve, passes through every one of the finite number of points comprising the former *OD* graph. This result may be termed the superposition principle.

Generation of a Continuous OD Graph From a Finitely Discrete OD Graph. Let X and Y be finitely discrete random variables. Let c_1 ,..., c_k be the ordered set of values which either X or Y can assume with nonzero probability. In addition, let c_0 be an arbitrarily chosen value less than c_1. Define the function f by setting $f(c_i) = c_{i-1}$ for $i = 1,..., k$. Let I and J be two independent random variables uniformly distributed over the interval zero to one. Then let

$$X^* = IX + (1 - I)f(X) \tag{1}$$

$$Y^* = JY + (1 - J)f(Y). \tag{2}$$

Thus, X^* and Y^* are continuous random variables. The probability that had been concentrated at the point c_i by the random variable X has now been uniformly distributed over the interval c_{i-1} to c_i by the random variable X^*. The same may be said of Y and Y^*. Therefore, the (X^*, Y^*) *OD* graph may be generated from the (X, Y) *OD* graph simply by interpolating straight-line segments between the adjacent points of the latter *OD* graph. This is illustrated in Fig. 2.

Equations (1) and (2) imply that if $X < Y$ then $X^* < Y^*$ and if $Y < X$ then $Y^* < X^*$. However, if $X = Y$ then either $X^* < Y^*$ or $Y^* < X^*$, each with probability $\frac{1}{2}$.

Shapes of OD Graphs and Their Significance

One-Sided OD Graphs. Following Birnbaum and Klose (1957), a random variable X will be said to be stochastically smaller than or equal to another random variable Y if, for every constant c,

$$P(X \leqslant c) \geqslant P(Y \leqslant c).$$

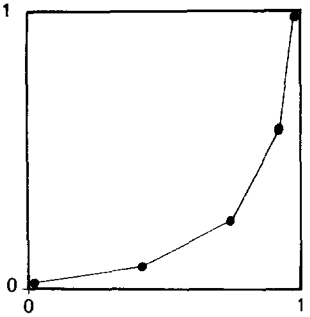

FIG. 2. The (X, Y) *OD* graph consists of the five dots. The (X^*, Y^*) *OD* graph consists of the line segments between adjacent dots.

Two random variables X and Y will be said to be stochastically comparable if either X is stochastically smaller than or equal to Y or vice versa. Let the straight-line segment connecting the points (0, 0) and (1, 1) be termed the positive diagonal. Thus, two random variables X and Y are stochastically comparable if and only if their *OD* graph lies entirely on or to one side of (i.e., does not cross) the positive diagonal. Such an *OD* graph will be said to be one-sided. Alternative terminology has been employed by Darlington (1973). If the (X, Y) *OD* graph lies entirely on or below the positive diagonal, then Y is said to completely dominate X. If the *OD* graph crosses the positive diagonal, then a situation of mixed dominance is said to exist. Moreover, he has proposed statistical procedures for determining whether the situation is one of complete or mixed dominance. One particularly interesting feature of complete dominance has been noted by Darlington. Namely, if Y completely dominates X and if m is any monotonic increasing function, then the expectation $E[m(X)]$ is guaranteed to be less than or equal to $E[m(Y)]$. This result is of particular interest in the case where X and Y represent the "true" scores of two populations on a trait and where the measured scores are monotonically related to the "true" scores.

Convex OD Graphs. Suppose that a random observation is to be sampled from either the X distribution or the Y distribution. Let W be random variable representing the value of the observation. Let S_X denote the event that the observation was sampled from the X distribution. Assume that the prior probability $P(S_X)$ is $\frac{1}{2}$. Suppose that the value of the observation is c. What is the posterior probability $P(S_X \mid W = c)$? If X and Y are finitely discrete, then $P(S_X \mid W = c)$ equals $1/[1 + P(Y = c)/P(X = c)]$. On the other hand, if X and Y are continuous, $P(S_X \mid W = c)$ equals $1/[1 + f_Y(c)/f_X(c)]$, where f_X and f_Y are the probability density functions of X and Y.

Let the likelihood ratio $L(c)$ be defined as being $P(Y = c)/P(X = c)$ for finitely discrete X and Y, and as being $f_Y(c)/f_X(c)$ for continuous X and Y. Thus, no matter whether X and Y are finitely discrete or continuous, the posterior probability $P(S_X \mid W = c)$ equals $1/[1 + L(c)]$.

Suppose X and Y are finitely discrete. Let $c_1, \ldots, c_k$ denote the ordered set of values which either X or Y can assume with nonzero probability. In addition, let c_0 be an arbitrary value less than c_1. Then, for $i = 1, \ldots, k$, $L(c_i)$ equals the interpolated slope between the OD graph point having horizontal and vertical coordinates $P(X \leqslant c_{i-1})$ and $P(Y \leqslant c_{i-1})$ and the adjacent point having coordinates $P(X \leqslant c_i)$ and $P(Y \leqslant c_i)$. Thus, if X and Y are finitely discrete, then $L(c_i)$ equals the slope of the (X^*, Y^*) OD graph between the points $T(c_{i-1})$ and $T(c_i)$. On the other hand, if X and Y are continuous, then $L(c)$ equals the instantaneous slope of the OD graph as it passes through the point $T(c)$.

A continuous OD graph will be said to be convex if its shape is either entirely convex upward or entirely convex downward. In other words, the OD graph must have consistent curvature so that it is not partly convex upward and partly concave upward. Similarly, if X and Y are finitely discrete, the (X, Y) OD graph will be said to be convex if and only if the (X^*, Y^*) OD graph is convex. Obviously, if an OD graph is convex, then it is also one-sided.

A desirable property for two random variables X and Y is that their posterior probability $P(S_X \mid W = c)$ be a monotonic function of c. If X and Y have this property, they will be said to have a monotonic posterior. For example, suppose that X is stochastically less than or equal to Y and that the posterior probability $P(S_X \mid W = c)$ is a monotonic decreasing function of c. Then, the lower the value of the observation c, the more certain it is that this observation was sampled from the X distribution rather than the Y distribution.

Now, since $P(S_X \mid W = c)$ equals $1/[1 + L(c)]$, the posterior probability will be a monotonic function of c if and only if the likelihood ratio is a monotonic function. Furthermore, since the likelihood ratio equals either the interpolated or the instantaneous slope of the OD graph, the likelihood ratio will be monotonic if and only if the OD graph is convex. Thus, X and Y will have a monotonic posterior if and only if the (X, Y) OD graph is convex.

Rotated OD Graphs. Consider a graph consisting of the points having horizontal coordinates $P(X \geqslant c)$ and vertical coordinates $P(Y \geqslant c)$ for all values of the constant c. If X and Y are both continuous or if they are both finitely discrete, then it is readily shown that this graph is geometrically congruent to the (X, Y) OD graph. Specifically, it is identical to the (X, Y) OD graph rotated 180° about the point $(\frac{1}{2}, \frac{1}{2})$. For this reason, such a graph will be termed a rotated OD graph.

Area Above the OD Graph

Suppose that an observation is randomly sampled from the X distribution and another random observation is independently sampled from the Y distribution. Let $P(X < Y)$ denote the probability of the event that the X observation is less than the Y observation. The quantities $P(X \leqslant Y)$ and $P(X = Y)$ may be defined similarly.

Continuous X and Y. Suppose X and Y are continuous. Consider the area above the (X, Y) *OD* curve. Specifically, this is the area of the region bounded on the lower right by the *OD* curve, bounded on the left by the line running (0, 0) to (0, 1), and bounded on the top by the line running from (0, 1) to (1, 1). This area is illustrated in Fig. 3.

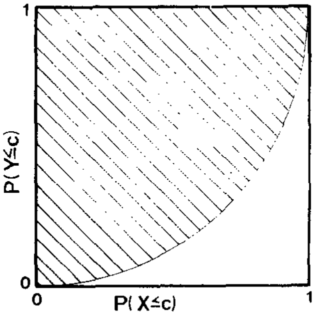

FIG. 3. The area above the continuous (X, Y) *OD* graph is striped.

Let $A(X, Y)$ denote the area above the (X, Y) *OD* graph. Let f_Y denote the probability density function of Y. Then,

$$\begin{aligned} A(X, Y) &= \int_0^1 P(X \leqslant c)\, dP(Y \leqslant c) \\ &= \int_{-\infty}^{+\infty} P(X \leqslant c) f_Y(c)\, dc \\ &= P(X \leqslant Y). \end{aligned}$$

Since X and Y are continuous, $P(X = Y)$ equals zero. Thus, $P(X < Y)$, $P(X \leqslant Y)$ and $A(X, Y)$ are all equal.

Finitely Discrete X and Y. When X and Y are finitely discrete, the (X, Y) *OD* graph consists of only a finite number of points. Consequently, the area above the *OD* graph cannot be defined in the same way as for continuous X and Y. Instead, this area is defined as being identical to the area above the (X^*, Y^*) *OD* graph. Note that this definition of the area above the (X, Y) *OD* graph is equivalent to the area as defined by the trapezoidal rule. Thus, $A(X, Y)$ equals $A(X^*, Y^*)$.

Let $c_1, \ldots, c_k$ denote the ordered set of values which either X or Y can assume with nonzero probability. In addition, let c_0 be an arbitrary value less than c_1. The area above the *OD* graph may be computed by dividing it into trapezoids and computing the area of each trapezoid. Consider the area A_i of the trapezoid indicated in Fig. 4.

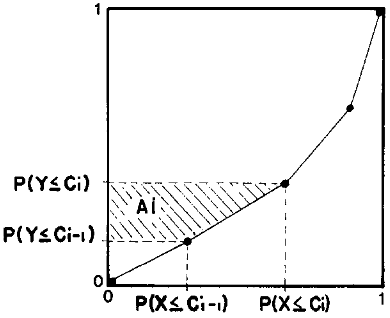

FIG. 4. Calculation of $A(X, Y)$ by the trapezoidal rule.

The height of this trapezoid is $P(Y \leqslant c_i) - P(Y \leqslant c_{i-1})$ or $P(Y = c_i)$. The lengths of the top and bottom edges of the trapezoid are $P(X \leqslant c_i)$ and $P(X \leqslant c_{i-1})$, respectively. Thus,

$$\begin{aligned} A_i &= P(Y = c_i)[\tfrac{1}{2}P(X \leqslant c_i) + \tfrac{1}{2}P(X \leqslant c_{i-1})] \\ &= P(Y = c_i)[P(X \leqslant c_{i-1}) + \tfrac{1}{2}P(X = c_i)]. \end{aligned}$$

Then, since $A(X, Y)$ equals the sum of the A_i's,

$$\begin{aligned} A(X, Y) &= \sum_{i=1}^{k} P(Y = c_i)\, P(X \leqslant c_{i-1}) + \tfrac{1}{2} \sum_{i=1}^{k} P(Y = c_i)\, P(X = c_i) \\ &= P(X < Y) + \tfrac{1}{2}P(X = Y). \end{aligned} \tag{3}$$

Recall that, when X and Y are continuous, $P(X = Y)$ equals zero. Thus Eq. (3) is valid both for finitely discrete and for continuous X and Y.

$A(X, Y)$ as a Measure of the Size of the Difference Between Two Populations. Note that $A(X, Y)$ measures the extent to which the X distribution lies below the Y distribution. Thus, $A(X, Y)$ can take on any value from a minimum of zero to a maximum of one. The maximum value is attained if and only if the X distribution lies entirely below the Y distribution with no overlap between the two distributions. Similarly, $A(X, Y)$ equals zero if and only if the X distribution lies entirely above the Y distribution with no overlap. On the other hand, if X and Y are identically distributed, then $A(X, Y)$ equals $\frac{1}{2}$. Note that $A(X, Y)$ and $A(Y, X)$ are complementary, in that they always sum to one.

These properties of $A(X, Y)$ make it a useful measure of the size or importance of a difference between two populations (Darlington, 1973). For example, are urban children more or less aggressive than rural children? If there is a difference, is it large or small? Let X and Y, respectively, represent the scores of urban and rural children on an agression scale. Equation (3) shows that $A(X, Y)$ may be interpreted as the probability that a randomly selected urban child will be less aggressive than a randomly selected rural child plus one half the probability that they will be equally aggressive. The closer $A(X, Y)$ is to zero or one, the larger and more important is the difference between the two populations. On the other hand, the closer $A(X, Y)$ is to $\frac{1}{2}$ the smaller and less important is the difference.

$A(X, Y)$ as a Measure of Discrimination Accuracy. In addition, $A(X, Y)$ may be used to measure how accurately a given test differentiates two populations. Consider, for example, a test designed to diagnose brain damage. Suppose that it was intended by the designers of the test that brain-damaged individuals should score low and normal individuals should score high on the test. Let X and Y denote the scores of brain-damaged and normal individuals, respectively. Then, $A(X, Y)$ measures the extent to which the test designers succeeded in their goal. Suppose that $A(X, Y)$ equals one. This would mean that the test discriminates the two populations perfectly. It would mean that there is some critical score below which all brain-damaged individuals score and above which all normal individuals score. Once this critical score has been determined, the test could be used to diagnose brain damage with 100% accuracy. If $A(X, Y)$ were close to one, then brain damage could be diagnosed with almost 100% accuracy. On the other hand, if the value of $A(X, Y)$ were only slightly greater than $\frac{1}{2}$, then diagnoses based on this test would be only slightly more accurate than chance. Thus, the test would have little value in diagnosing individual cases.

Receiver Operating Characteristic Graphs

Consider an observer in a signal detection experiment. On each trial of the experiment, he may be presented with either of two events: a signal event or a noise event. His task is to discriminate the two. On each trial, the observer obtains a sensory impression of the event presented to him. The strength of this impression is the basis on which the observer infers whether signal or noise was presented. Signal events tend to produce stronger impressions than noise events. However, the impression strength induced by both signal and noise events varies greatly from trial to trial. Thus, impression strength is only a fallible guide as to whether signal or noise was presented.

The Yes–No ROC Curve. Consider the yes–no signal detection task. In this task, the observer is told to respond "yes" if he thinks that a signal was presented on that trial and to respond "no" otherwise. It is assumed that the observer performs this task as follows. First, he adopts an impression strength criterion. Then, on each trial, if the impression strength reaches or exceeds his criterion, he responds "yes." Otherwise, he responds "no." Let $P(\text{yes} \mid \text{signal})$ and $P(\text{yes} \mid \text{noise})$ denote the probability of a "yes" response on signal trials and noise trials, respectively. These two probabilities suffice to summarize the results of a yes–no experiment.

Now, there are various manipulations which the experimenter may employ to induce the observer to change his impression strength criterion (Green & Swets, 1966, pp. 87–88). If the observer can be induced to lower his criterion, then $P(\text{yes} \mid \text{signal})$ and $P(\text{yes} \mid \text{noise})$ will both become larger. Conversely, they will both become smaller if the observer raises his criterion.

Consider the following *theoretical* graph. Let the horizontal axis denote $P(\text{yes} \mid \text{noise})$ and the vertical axis denote $P(\text{yes} \mid \text{signal})$. Let a separate point be plotted for every criterion that the observer could adopt. For extremely low criteria, the plotted point is located at (1, 1). Then, as the observer's criterion is raised, both coordinates decrease and a continuous curve is traced out. Finally, for very high criteria, the curve ends at (0, 0). Let this graph be called the yes–no receiver operating characteristic (*ROC*) curve. This is illustrated in Fig. 5.

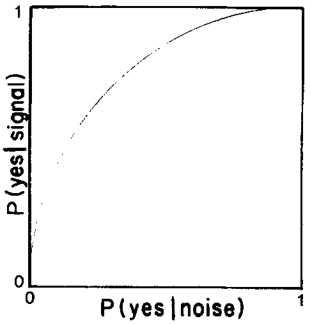

FIG. 5. An example of a yes–no *ROC* curve.

The (I_n, I_s) *ROC curve.* Let I_s and I_n be random variables denoting the strengths of sensory impressions aroused by signal events and noise events, respectively. It is assumed that I_s and I_n are continuous random variables. Let c be the observer's criterion. Then,

$$P(\text{yes} \mid \text{signal}) = P(I_s \geqslant c)$$

and

$$P(\text{yes} \mid \text{noise}) = P(I_n \geqslant c).$$

Thus, the yes–no *ROC* curve is identical to a graph in which $P(I_s \geqslant c)$ is plotted on the vertical axis vs $P(I_n \geqslant c)$ on the horizontal axis, for all values of the constant c. The latter graph will be termed the (I_n, I_s) *ROC* curve. Note that it is a rotated *OD* graph. Thus, the yes–no *ROC* curve and the (I_n, I_s) *ROC* curve are both variant forms of an *OD* graph.

Consider the area *below* the yes–no *ROC* curve. This is identical to the area *below* the (I_n, I_s) *ROC* curve, which is identical to the area *above* the (I_n, I_s) *OD* graph. Thus, the area below the yes–no *ROC* curve equals $A(I_n, I_s)$. Recall that, since I_n

and I_s are continuous, $A(I_n, I_s)$ equals $P(I_n < I_s)$. Thus, $A(I_n, I_s)$ measures the extent to which the I_n distribution lies below the I_s distribution. If $A(I_n, I_s)$ equals one, then the I_n distribution lies entirely below the I_s distribution and, consequently, it is possible for the observer to be errorless in distinguishing signal events from noise events. However, if I_n and I_s are identically distributed, then $A(I_n, I_s)$ equals $\frac{1}{2}$. In this case, impression strength provides no information and consequently the observer is forced to guess whether signal or noise was presented. Thus, $A(I_n, I_s)$ might be considered to be a measure of the fidelity of information transmitted by the observer's sensory system; or, equivalently, it may be considered a measure of the fidelity of information available to the observer for inferring whether signal or noise was presented.

The Rating Method

Now, the (I_n, I_s) *ROC* curve is a strictly theoretical construct. There is no way to observe this curve in its entirety. However, there are two experimental methods that make it possible to observe the location of a limited number of points on this curve. These are the yes–no method and the rating method (Green & Swets, 1966). The yes–no method consists of performing a number of separate yes–no experiments and inducing the observer to utilize a different criterion for each experiment. Each experiment generates a separate point on the (I_n, I_s) *ROC* curve. The rating method is described below. From this point onward, this paper is concerned exclusively with the rating method.

The rating method works as follows. On each experimental trial, the observer is presented with either a signal event or a noise event. The observer is required to rate his degree of confidence that a signal was presented on that trial. For this purpose, he is given a confidence scale consisting of k confidence levels. These confidence levels may be coded by the integers one through k. The integer one is assigned to the confidence level representing lowest confidence that the signal was presented (i.e., highest confidence that the signal was *not* presented). The integer k is assigned to the confidence level representing highest confidence that the signal was presented.

It is assumed that the observer performs his rating task as follows. First, he selects $k - 1$ boundary points on the scale of possible impression strengths. These boundary points divide the impression strength scale into k mutually exclusive and exhaustive intervals. Then, on each trial, if the impression strength aroused by the presented event falls within the mth interval, the observer responds with confidence level m. Thus, the observer's confidence rating is an increasing step function of the event's impression strength. Moreover, this step function has only a finite number of steps.

The (R_n, R_s) *ROC Graph.* Let R_s and R_n be random variables denoting the observer's confidence ratings on signal trials and noise trials, respectively. Thus, R_s and R_n are finitely discrete random variables. Moreover, R_s and R_n are a monotonic step function of I_s and I_n, respectively. Consider a graph in which $P(R_s \geqslant c)$ is plotted on the ordinate vs $P(R_n \geqslant c)$ on the abscissa, for all values of the constant c. This graph is a rotated *OD* graph and is termed the (R_n, R_s) *ROC* graph. It consists of a finite number of points. Because R_s and R_n are a monotonic step function of I_s and I_n, it follows from the superposition principle that every point of the (R_n, R_s) *ROC* graph is also a point on the continuous (I_n, I_s) *ROC* curve. Since there are a total of k confidence levels, the (R_n, R_s) *ROC* graph consists of $k + 1$ points. Two of these points are (0, 0) and (1, 1). The remaining $k - 1$ points may be located anywhere along the (I_n, I_s) *ROC* curve. Their precise locations on this curve are determined by the observer's choice of $k - 1$ boundary points on the impression strength scale.

Thus, the $k + 1$ (R_n, R_s) *ROC* graph points are like beads strung on a wire track. One bead is fixed at each end of the track. The $k - 1$ beads in the middle cannot move off the track. However, they are free to slide along the track subject only to the constraint that they always be strung in the same order. The observer, by his choice of boundary points on the impression strength scale, indirectly controls the position of the $k - 1$ movable beads.

$A(R_n, R_s)$ Considered as an Approximation to $A(I_n, I_s)$. Now, $A(R_n, R_s)$ is the area below the (R_n, R_s) *ROC* graph as determined by the trapezoidal rule. Since the (I_n, I_s) *ROC* curve passes through every point of the (R_n, R_s) *ROC* graph, it follows that $A(I_n, I_s)$ and $A(R_n, R_s)$ should be approximately equal. Thus, $A(R_n, R_s)$ may be used as an approximation to $A(I_n, I_s)$. This approximation has practical value since $A(I_n, I_s)$ is not observable, whereas $A(R_n, R_s)$ is observable. How accurate this approximation will be depends in large part upon the spacing between the points of the (R_n, R_s) *ROC* graph. This is illustrated in Fig. 6. Both (R_n, R_s) *ROC* graphs in this figure consist of only five points. However, $A(R_n, R_s)$ is only slightly smaller than $A(I_n, I_s)$ on the left of the figure, whereas $A(R_n, R_s)$ is considerably smaller than $A(I_n, I_s)$ on the right.

If the (I_n, I_s) *ROC* graph is convex upward then,

$$\tfrac{1}{2} \leqslant A(R_n, R_s) \leqslant A(I_n, I_s).$$

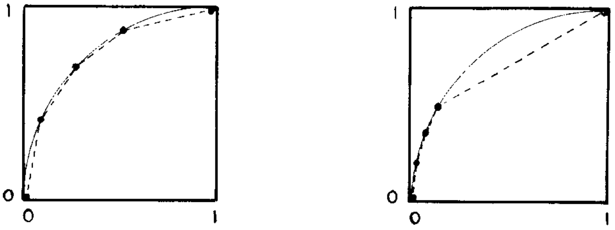

FIG. 6. Two (R_n, R_s) *ROC* graphs superimposed on the same (I_n, I_s) *ROC* curve. The area below the solid curve is $A(I_n, I_s)$. The area below the dashed line is $A(R_n, R_s)$.

Thus, $A(R_n, R_s)$ will usually be somewhat less than $A(I_n, I_s)$. However, if the number k of confidence intervals is made large and if there is approximately equal spacing between the $k + 1$ points of the (R_n, R_s) *ROC* graph, then $A(R_n, R_s)$ can be brought arbitrarily close to $A(I_n, I_s)$. Now, the number of confidence levels is determined by the instructions given the observer, while the spacing between points of the (R_n, R_s) *ROC* graph is under the observer's control. For this reason, $A(I_n, I_s)$ may be considered to be a measure of the observer's discrimination *capacity*, whereas $A(R_n, R_s)$ may be considered to be a measure of the observer's discrimination *performance*. Unless the observer has made a poor choice of the step function relating his confidence ratings to impression strength, the performance measure should be only slightly smaller than the capacity measure.

A well-known application of this approximation is due to Green. Consider a two-alternative forced-choice signal detection task. On each trial, the observer is presented with a signal event followed by a noise event or vice versa. His task is to indicate which of the two events was the signal. Presumably the observer selects the event which aroused the stronger sensory impression and identifies that event as the signal. The probability of the observer being correct $P(\text{correct})$ should equal $P(I_s > I_n)$. But, $P(I_s > I_n)$ equals $A(I_n, I_s)$. So, $P(\text{correct})$ should equal $A(I_n, I_s)$ (Green, 1964; Green & Swets, 1966, pp. 45–47). Now, $A(I_n, I_s)$ is not observable. However, it may be approximated by $A(R_n, R_s)$ as measured in a confidence rating task. Thus, $P(\text{correct})$ from a two-alternative forced-choice test and $A(R_n, R_s)$ from a confidence-rating task should be approximately equal. Green and Moses (1966) tested this prediction in a recognition memory experiment and found it to be valid.

$A(R_n, R_s)$ Considered on Its Own Merits. Up to this point, $A(R_n, R_s)$ has been discussed only within the context of signal detection theory. It would be a mistake, however, to think that $A(R_n, R_s)$ has a meaningful interpretation only within the context of this theory. Moreover, it would be doubly wrong to think that $A(R_n, R_s)$ is meaningful only insofar as it approximates $A(I_n, I_s)$. In fact, $A(R_n, R_s)$ may be sensibly interpreted, without any reference to signal detection theory, as a measure of the observer's discrimination performance. Obviously, the less overlap there is between the R_s and R_n distributions, the better the observer is discriminating signal events from noise events. Recall that

$$A(R_n, R_s) = P(R_n < R_s) + \tfrac{1}{2}P(R_n = R_s).$$

Thus, $A(R_n, R_s)$ measures the extent to which the R_n distribution lies below the R_s distribution. Consequently, it is a measure of the extent to which the observer's confidence ratings separate signal events from noise events.

This interpretation of $A(R_n, R_s)$ is exactly analogous to the interpretation of $A(X, Y)$ as a measure of how accurately a diagnostic test discriminates brain-damaged from normal individuals. In both cases, elements from two populations are examined by a device. The area above the *OD* graph measures the accuracy with which the device, by means of an ordinally-scaled output, discriminates elements from the two populations. In one case, the two populations are signal and noise events, the device is a human observer, and the ordinally-scaled output are his confidence ratings. In the other case, the two populations are brain-damaged and normal individuals, the device is a diagnostic test, and the ordinally-scaled output are the test scores. Thus, even if signal detection theory had never been invented, $A(R_n, R_s)$ would be a meaningful measure of the observer's performance at discriminating signal events from noise events.

The $D'(X, Y)$ Measure

Simpson and Fitter (1973), on the basis of suggestions made by Schulman and Mitchell (1966) and Taylor (1967, p. 394), proposed the measure of discrimination d_a. Suppose that I_n and I_s are both normally distributed. Let μ_n and μ_s denote the respective means of I_n and I_s while σ_n and σ_s denote the respective standard deviations. Then, d_a is defined by

$$d_a = (\mu_s - \mu_n)/[(\sigma_n{}^2 + \sigma_s{}^2)/2]^{1/2}.$$

Note, that if σ_s equals σ_n, then d_a reduces to the well-known measure d' (Green & Swets, 1966, Chap. 3).

Simpson and Fitter showed that d_a is a monotonic function of $A(I_n, I_s)$. Let the function z denote the inverse of the cumulative distribution function of a normal random variable with mean zero and unit standard deviation. Specifically, Simpson and Fitter showed that, if I_n and I_s are normally distributed, then

$$d_a = 2^{1/2}z[A(I_n, I_s)].$$

These results suggest the following generalization of Simpson and Fitter's proposals. Suppose, for example, that X and Y represent scores of brain-damaged and normal individuals, respectively, on a diagnostic test. Let

$$D'(X, Y) = 2^{1/2}z[A(X, Y)].$$

Then $D'(X, Y)$ is a measure of the accuracy with which the test discriminates the two populations. Moreover, $D'(X, Y)$ is commensurate with d' in the following sense. Consider a human observer in a signal detection task whose d' equals $D'(X, Y)$ for the diagnostic test. Then, the diagnostic test may be said to discriminate brain-damaged from normal individuals with the same accuracy that the human observer discriminates signal events from noise events.

Point Estimation of $A(X, Y)$

The remainder of this paper is concerned with the following questions. Suppose that random samples of observations of each of the random variables X and Y have been obtained. How should $A(X, Y)$ be estimated? What are the properties of the statistic used to estimate $A(X, Y)$? How should confidence intervals for $A(X, Y)$ be constructed?

Sample OD Graphs

Recall the definition of a *population OD* graph. For any constant c, $T(c)$ is the point having the horizontal coordinate $P(X \leqslant c)$ and the vertical coordinate $P(Y \leqslant c)$. The population OD graph for X and Y consists of the points $T(c)$ for all c from $-\infty$ to $+\infty$.

A *sample OD* graph may now be defined. Suppose that a random sample of N_X observations of the random variable X and a random sample of N_Y observations of Y have been obtained. Let $p(X \leqslant c)$ denote the *proportion* of the N_X observations of X which are less than or equal to the constant c. Let $p(Y \leqslant c)$ be defined similarly. For any c, let $t(c)$ be the point having the horizontal coordinate $p(X \leqslant c)$ and the vertical coordinate $p(Y \leqslant c)$. Now, the sample OD graph for X and Y consists of the points $t(c)$ for all c from $-\infty$ to $+\infty$. For any c, each coordinate of the point $t(c)$ is an unbaised estimator of the corresponding coordinate of $T(c)$. In this sense, the entire sample OD graph for X and Y may be considered to be an unbiased estimator of the population OD graph for X and Y.

Note that, because it is constructed on the basis of finite samples of X and Y observations, a sample OD graph always consists of a finite number of points. If there are no ties in the combined sample of X and Y observations, then the sample OD graph will consist of $N_X + N_Y + 1$ points. If there are ties in the combined sample, there will be fewer points in the sample OD graph. Two of these points are always $(0, 0)$ and $(1, 1)$. This property is a basic difference between sample and population OD graphs. A population OD graph for continuous X and Y is a continuous curve, whereas a sample OD graph for continuous X and Y can never be anything but a finite set of points.

Area Above the Sample OD Graph

A total of N_X observations of X and N_Y observations of Y have been randomly sampled. Thus, there are a total of $N_X N_Y$ ways of pairing an X observation with a Y observation. Let $p(X < Y)$, $p(X = Y)$ and $p(X \neq Y)$ denote the proportion of these $N_X N_Y$ pairs for which $X < Y$, $X = Y$, and $X \neq Y$, respectively. It is evident that $p(X < Y)$, $p(X = Y)$, and $p(X \neq Y)$ are unbiased estimators of $P(X < Y)$, $P(X = Y)$, and $P(X \neq Y)$, respectively.

Let $a(X, Y)$ denote the area above the sample *OD* graph for X and Y as computed by the trapezoidal rule. It is easily shown that

$$a(X, Y) = p(X < Y) + \tfrac{1}{2}p(X = Y).$$

(The proof of this is quite similar to the proof of Eq. (3).) Taking the expected values of both sides of this equation reveals that $a(X, Y)$ is an unbiased estimator of $A(X, Y)$. Thus, the area above a sample *OD* graph is an unbiased estimate of the area above the population *OD* graph.

The Mann–Whitney U and $a(X, Y)$. Mann and Whitney's (1947) U statistic is defined as being the total number of (X, Y) pairs in which $X < Y$. From this definition, it can be seen that $a(X, Y)$ and the Mann–Whitney U statistic are closely related. Thus, if X and Y are continuous, $a(X, Y)$ equals $U/N_X N_Y$. This is a fortunate relationship as there is a large body of literature dealing with the Mann–Whitney U and related statistics. Some of that literature is reviewed in succeeding sections of this paper.

From this point on, when there is no danger of ambiguity, $a(X, Y)$ and $A(X, Y)$ will be abbreviated by a and A, respectively.

Variance of $a(X, Y)$

This section of the paper discusses the variance of a. A variance formula, variance estimator, and upper bounds on the variance of a are all given. These results will be useful in calculating confidence intervals for A.

It was mentioned above that a is an unbiased estimate of A. Moreover, Lehmann (1951, Lemma 3.2) has stated without giving the proof that, if X and Y are continuous, then a has the smallest variance of all unbiased estimators of A. Whether this result also holds for finitely discrete X and Y is not known to the present author.

Variance Formula

A number of writers have derived formulas for the variance of the Mann–Whitney U and related statistics. Unfortunately for present purposes, most of these formulas were derived under the assumption of continuous X and Y. However, Noether (1967) has derived the variance of a statistic which is a linear function of a without assuming continuity of X and Y. Let $\sigma_a{}^2$ denote the variance of a. Then, from Noether's Eq. (6.5), it is seen that

$$\sigma_a{}^2 = (1/4N_X N_Y)[P(X \neq Y) + (N_X - 1)\, B_{XXY} + (N_Y - 1)\, B_{YYX} \\ - 4(N_X + N_Y - 1)(A - \tfrac{1}{2})^2], \qquad (4)$$

where B_{XXY} and B_{YYX} are defined as follows. Let Y_1 and Y_2 be two independent observations randomly sampled from the Y distribution and let X be a random observation independently sampled from the X distribution. Then, B_{YYX} is defined by

$$\begin{aligned} B_{YYX} &= P(Y_1, Y_2 < X) + P(X < Y_1, Y_2) \\ &\quad - P(Y_1 < X < Y_2) - P(Y_2 < X < Y_1) \qquad (5) \end{aligned}$$

and B_{XXY} is defined analogously. Note, that, as N_X and N_Y approach infinity, σ_a^2 goes to zero. This shows that a is a consistent estimator of A.

Estimating σ_a^2

In order to estimate σ_a^2, it is first necessary to estimate B_{YYX} and B_{XXY}. This may be done as follows. Consider triplets of the form (Y_1, Y_2, X) where Y_1 and Y_2 are independent. The total number of such triplets that can be formed from the X and Y samples is $N_Y(N_Y - 1)N_X$. Let $p(Y_1, Y_2 < X)$ denote the proportion of these triplets for which both Y_1 and Y_2 are less than X. Let $p(X < Y_1, Y_2)$ and $p(Y_1 < X < Y_2)$ be defined similarly. Then, set

$$b_{YYX} = p(Y_1, Y_2 < X) + p(X < Y_1, Y_2) - 2p(Y_1 < X < Y_2).$$

Obviously, b_{YYX} is an unbiased estimate of B_{YYX}. It may be computed as follows. First, rank order the combined samples of X and Y observations. Then, for each X observation, count the number of Y observations that are less than that X and the number that are greater than that X. Let these two quantities be denoted by u_X and v_X, respectively. Then,

$$b_{YYX} = \sum [u_X(u_X - 1) + v_X(v_X - 1) - 2u_X v_X]/[N_Y(N_Y - 1)\, N_X],$$

where the summation is carried out over all X observations in the sample. In an analogous manner, b_{XXY} an unbiased estimate of B_{XXY} may be defined.

Now, $p(X \neq Y)$ and a unbiasedly estimate $P(X \neq Y)$ and A. This provides unbiased estimators for $P(X \neq Y)$, B_{XXY}, B_{YYX}, and A which may be substituted into Eq. (4). Note, however, that the expected value of $(a - \frac{1}{2})^2$ is $(A - \frac{1}{2})^2 + \sigma_a^2$. This introduces a bias which may be corrected by multiplying Eq. (4) by $N_X N_Y/(N_X - 1)(N_Y - 1)$. So, let

$$\begin{aligned} s_a^2 = {} & [1/4(N_X - 1)(N_Y - 1)][p(X \neq Y) + (N_X - 1)\, b_{XXY} + (N_Y - 1)\, b_{YYX} \\ & - 4(N_X + N_Y - 1)(a - \tfrac{1}{2})^2]. \qquad (6) \end{aligned}$$

Then, s_a^2 is an unbiased estimate of σ_a^2.

Maximum Value of $\sigma_a{}^2$

Let $\sigma^2_{\max}$ denote the maximum possible value of $\sigma_a{}^2$ for fixed A and fixed sample sizes N_X and N_Y. That is, all combinations of random variables X and Y yielding a fixed value of $A(X, Y)$ are examined. For each such combination of X and Y, $\sigma_a{}^2$ is calculated for fixed sample sizes N_X and N_Y. The largest $\sigma_a{}^2$ obtained is denoted $\sigma^2_{\max}$.

Three different expressions for $\sigma^2_{\max}$ are given here. The first expression gives the maximum value of $\sigma_a{}^2$ for the general case of any continuous X and Y. The second expression gives the maximum value of $\sigma_a{}^2$ for the case where X and Y are continuous and stochastically comparable. The third expression gives the maximum value of $\sigma_a{}^2$ for the case where X and Y are continuous and have a monotonic posterior. Obviously, $\sigma^2_{\max}$ for the general case is larger than $\sigma^2_{\max}$ for stochastic comparability which, in turn, is larger than $\sigma^2_{\max}$ for a monotonic posterior.

In deriving all three of these upper bounds for the variance of $a(X, Y)$, X and Y were assumed to be continuous. Nevertheless, these three upper bounds are also valid for finitely discrete X and Y. Suppose that X and Y are finitely discrete. Let X^* and Y^* be defined by Eqs. (1) and (2). Then X^* and Y^* are continuous. Let $\sigma_a{}^2$ and $\sigma_*{}^2$ denote the variances of $a(X, Y)$ and $a(X^*, Y^*)$, respectively. An examination of each term in Eq. (4) shows that

$$\sigma_a{}^2 \leqslant \sigma_*{}^2 \leqslant \sigma^2_{\max} .$$

General Case. Pollack and Hsieh (1969) investigated the variance of $a(X, Y)$ using Monte Carlo techniques. They studied $\sigma_a{}^2$ under conditions where N_X and N_Y were equal and where X and Y had either both normal, both rectangular, or both negative exponential distributions. They found that $\sigma_a{}^2$ was usually close to but less than $A(1 - A)/N_X$.

Van Dantzig (1951, footnote 4) and Birnbaum and Klose (1957) have published mathematical proofs that, if X and Y are continuous, then

$$\sigma^2_{\max} = A(1 - A)/N_L , \tag{7}$$

where N_L denotes the lesser of N_X and N_Y. Thus, the theorems of Van Dantzig and Birnbaum and Klose confirm and extend the results of Pollack and Hsieh's Monte Carlo investigations.

Stochastic Comparability. Let N_G denote the greater of N_X and N_Y and let A_G denote the greater of A and $1 - A$. Birnbaum and Klose (1957) have proven that, if X and Y are continuous and stochastically comparable, then

$$\sigma^2_{\max} = (1/N_G N_L)\{(2N_L - 1)A_G(1 - A_G) - (N_G - N_L)(1 - A_G)^2 + \tfrac{1}{3}(N_G - 2N_L + 1)[1 - (2A_G - 1)^{3/2}]\}. \tag{8}$$

Monotonic Posterior. It is shown in the Appendix that, if X and Y are continuous and have a monotonic posterior, then

$$\sigma^2_{\max} = (1/3N_GN_L)[(2N_G + 1)\,A_G(1 - A_G) - (N_G - N_L)(1 - A_G)^2]. \tag{9}$$

OD Curves That Maximize $\sigma_a{}^2$. What is the shape of the OD curves that maximize $\sigma_a{}^2$ for fixed A and for fixed sample sizes? Suppose that $A \geqslant \frac{1}{2}$ and $N_X \leqslant N_Y$. The particular OD curves that maximize $\sigma_a{}^2$, for the general case, for stochastic comparability, and for a monotonic posterior, are shown in Fig. 7. Using Eq. (4) to calculate $\sigma_a{}^2$ for these three OD curves yields the expressions for $\sigma^2_{\max}$ given in Eqs. (7), (8), and (9).

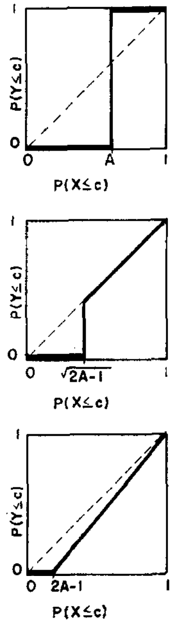

FIG. 7. OD curves that maximize $\sigma_a{}^2$ for fixed $N_X \leqslant N_Y$ and fixed $A \geqslant 1/2$. Upper: General case. Middle: Stochastic comparability (one-sided OD curve). Lower: Monotonic posterior (convex OD curve).

Estimating $\sigma^2_{\max}$. How should $\sigma^2_{\max}$ be estimated? Let $s^2_{\max}$ denote an estimate of $\sigma^2_{\max}$. Let a_G denote the greater of a and $1 - a$. The three $\sigma^2_{\max}$'s may be estimated as follows. First, replace A with a everywhere in Eq. (7). Replace A_G with a_G everywhere in Eqs. (8) and (9). Second, change the denominator in Eq. (7) from N_L to $N_L - 1$. Change the denominator in Eq. (9) from $3N_GN_L$ to $3N_GN_L - N_G - N_L + 1$. The purpose of these denominator changes is to convert the estimator $s^2_{\max}$ from a bias toward underestimation to a bias toward overestimation of $\sigma^2_{\max}$. (Naturally, when estimating an upper bound, the conservative strategy is to overestimate rather than underestimate.) It is unclear whether and how a corresponding adjustment should be made in Eq. (8).

ASYMPTOTIC NORMALITY OF $a(X, Y)$

A particularly useful property of a is that it is asymptotically normal. Mann and Whitney (1947) showed that, if X and Y are continuous and identically distributed, then, as N_X and N_Y approach infinity, the distribution of U approaches the normal distribution. Lehmann (1951, Theorem 3.2) considerably extended Mann and Whitney's results. Lehmann examined a large class of statistics. This class includes both the Mann–Whitney U statistic and the a statistic. Lehmann showed that, irrespective of how X and Y are distributed, the distribution of all statistics in this class is asymptotically normal. In particular, asymptotic normality holds even if X and Y are discrete and are not identically distributed. Specifically, it follows from Lehmann's theorem that, if N_X and N_Y are held in constant ratio, then as N_X and N_Y approach infinity, the distribution of $(N_X)^{1/2}(a - A)$ approaches a normal distribution with mean zero and fixed variance.

How large should N_X and N_Y be in order to ensure that the a distribution is adequately approximated by the normal distribution? Mann and Whitney showed that, if X and Y are continuous and identically distributed, then the approximation is adequate for most purposes whenever N_X and N_Y are both eight or larger. Note that, if X and Y are identically distributed, then A equals $\frac{1}{2}$. Birnbaum (1956) opined that, if A is close to either zero or one, then considerably larger values of both N_X and N_Y may be required. Thus, in most cases at present, there is no good way to determine how large N_X and N_Y should be to justify the use of the normal approximation.

There are, however, two principles that can be of some use in this regard. Now, the a distribution has a mean of A and a standard deviation of σ_a. As one travles more and more standard deviations away from A into either tail of the distribution, eventually a point is reached where the normal approximation breaks down and breaks down badly. The reason for this is simple. The two tails of the normal distribution stretch off to infinity, whereas the tails of the a distribution end at the values zero and one. Therefore, somewhere along the way as zero or one are approached, the normal approximation will break down. Thus, the first principle is that, the more s_a's (or $s_{\max}$'s) that a is from both zero and one, the better the normal approximation will tend to be. The second principle is that what may be a sufficiently adequate normal approximation for one purpose may not be for another. For example, a given value of N_X and N_Y may be sufficiently large to justify using the normal approximation to calculate a 50% confidence interval, but not a 99% confidence interval. To calculate the 99% confidence interval, it is necessary that the normal approximation be accurate further out into the tails of the distribution than is necessary for the 50% confidence interval. Thus, while these two principles do not answer the question of how large N_X and N_Y should be, they do provide some guidance on the issue.

Interval Estimation of $A(X, Y)$

Overview

Several different methods for constructing confidence intervals or confidence bounds for A are given below. Each of these methods has its good points and bad points. The methods based upon the normal approximation require larger sample sizes but yield narrower confidence intervals than the methods which are not based upon the normal approximation. Thus, there is a tradeoff involved in the choice of method: The methods which are less dependent upon large sample sizes and upon possibly questionable assumptions yield the wider confidence intervals. Unfortunately, it is often difficult to determine whether the sample sizes are large enough to justify the use of the normal approximation. This can make the choice of the best method quite difficult.

Application to Psychophysics. The methods described below may be applied in rating studies of signal detection to construct a confidence interval for $A(R_n, R_s)$. It should be noted, however, that a confidence interval for $A(R_n, R_s)$ is just that; it is *not* a confidence interval for $A(I_n, I_s)$. An $A(R_n, R_s)$ confidence interval is constructed to answer the question: How accurately has $A(R_n, R_s)$ been estimated? It does not provide an answer to the entirely separate question: How accurately does $A(R_n, R_s)$ approximate $A(I_n, I_s)$? Thus, an $A(R_n, R_s)$ confidence interval does *not* tell how accurately $A(I_n, I_s)$ is known.

Estimating $D'(X, Y)$. The methods given in this paper may also be used to provide point and interval estimators for $D'(X, Y)$. Suppose A has been estimated by a. Then, $(2)^{1/2} z(a)$ is a point estimator for $D'(X, Y)$. Moreover, if the confidence interval for A at confidence level γ runs from a_{low} to a_{high}, then the interval from $(2)^{1/2} z(a_{\text{low}})$ to $(2)^{1/2} z(a_{\text{high}})$ is a confidence interval for $D'(X, Y)$ at confidence level γ.

Confidence Intervals Based Upon Asymptotic Normality and s_a

Sen (1967) has proposed the following method of obtaining confidence intervals for A. In developing this method, Sen assumed that X and Y were continuous. However, as presented here, Sen's method has been slightly modified so that it is valid for both continuous and noncontinuous X and Y. Govindarajulu (1968) has also proposed confidence intervals which differ from Sen's only in that a biased rather than unbiased estimate of σ_a^2 is employed.

Let λ denote $N_X/(N_X + N_Y)$ and let N_0 denote λN_Y. It is evident from Eq. (4) that, if N_X and N_Y go to infinity in such a manner that λ remains constant, then $N_0\sigma_a{}^2$ approaches a limiting value which is a function of λ. Let this limiting value be denoted by $w_a{}^2(\lambda)$. Now, consider $s_a{}^2$ the unbiased estimator of $\sigma_a{}^2$. As N_X and N_Y go to infinity in constant ratio, $N_0s_a{}^2$ is a consistent estimator of $w_a{}^2(\lambda)$. Consider the ratio $(N_0)^{1/2}(a - A)/(N_0)^{1/2} s_a$. It is known from Lehmann (1951, Theorem 3.2) that, as N_X and N_Y go to infinity in constant ratio, the numerator is asymtptotically normal with mean zero and standard deviation $w_a(\lambda)$. The denominator is a consistent estimate of $w_a(\lambda)$. Then, it follows from a basic theorem (Cramér, 1946, Theorem 20.6) on the convergence of random variables that $(a - A)/s_a$ is asymptotically normal with mean zero and standard deviation one.

As previously, let the function z denote the inverse of the cumulative distribution function of a normal random variable with mean zero and standard deviation one. Then, for sufficiently large N_X and N_Y,

$$P[|(a - A)/s_a| \leqslant z[(1 + \gamma)/2]] \approx \gamma.$$

So, for sufficiently large N_X and N_Y, $a \pm z(\frac{1}{2} + \gamma/2)\, s_a$ may be taken as the two ends of a *confidence interval* for A at confidence level γ. Similarly, $a + z(\gamma)\, s_a$ may be taken as an upper *confidence bound* for A at confidence level γ. The problem with Sen's confidence intervals, of course, is that it is difficult to determine whether the sample sizes N_X and N_Y are large enough to justify use of the method.

Confidence Intervals Based Upon Asymptotic Normality and $s_{\max}$

Birnbaum (1956; Birnbaum & Klose, 1957) has suggested that the maximum variance of a, rather than the variance itself, be used to calculate confidence intervals for a. Now, three different expressions for $\sigma^2_{\max}$ (Eqs. (7), (8), and (9)) were given above. Which of these $\sigma^2_{\max}$'s should be used to calculate a confidence interval depends upon whether one is willing to make the strong assumption of a monotonic posterior, the weak assumption of stochastic comparability, or no assumption at all. The following remarks are valid whichever expression for $\sigma^2_{\max}$ is employed.

As N_X and N_Y go to infinity in constant ratio, $N_0\sigma^2_{\max}$ approaches a limiting value which is a function of λ. Let this limiting value be denoted $w^2_{\max}(\lambda)$. Furthermore, $N_0s^2_{\max}$ is a consistent estimator of $w^2_{\max}(\lambda)$. Then, it follows from the previously cited theorems of Lehmann and Cramér that $(a - A)/s_{\max}$ is asymptotically normal with mean zero and standard deviation $w_a(\lambda)/w_{\max}(\lambda)$.

So, for sufficiently large N_X and N_Y,

$$P\left[\left|\frac{a - A}{s_{\max}}\right| \leqslant z\left(\frac{1 + \gamma}{2}\right)\frac{w_a(\lambda)}{w_{\max}(\lambda)}\right] \approx \gamma. \tag{10}$$

Then, since $w_a(\lambda) \leqslant w_{\max}(\lambda)$,

$$P[|(a - A)/s_{\max}| \leqslant z((1 + \gamma)/2)] \geqslant \gamma. \tag{11}$$

So, for sufficiently large N_X and N_Y, $a \pm z(\frac{1}{2} + \gamma/2)\, s_{\max}$ may be *conservatively* taken as the two ends of a confidence interval for A at confidence level γ. Similarly, $a + z(\gamma)s_{\max}$ may be conservatively taken as an upper confidence bound for A at confidence level γ.

Thus, confidence intervals based upon $s_{\max}$ are wider than Sen's confidence intervals which are based upon s_a. However, the need for large sample sizes is presumably less acute when confidence intervals are calculated using $s_{\max}$ rather than s_a. (Note that, even if the sample sizes are not large enough for Eq. (10) to be valid, nevertheless the inequality of Eq. (11) may still hold.) A further advantage for confidence intervals calculated from $s_{\max}$ is that $s_{\max}$ is easier to compute than s_a.

Birnbaum–McCarty Confidence Bounds

It is evident that the Achilles heel of the previous methods of constructing confidence bounds for A is the difficulty in determining whether N_X and N_Y are large enough to justify the assumption of approximate normality. In order to avoid this difficulty, Birnbaum and McCarty (1958) developed a method for constructing confidence bounds which does not depend upon asymptotic normality. These confidence bounds, however, are highly conservative.

Background. Suppose N_X observations have been randomly sampled from the X distribution. Define the random variables:

$$D^+(N_X) = \sup_{-\infty<c<+\infty} [P(X \leqslant c) - p(X \leqslant c)]$$
$$D^-(N_X) = \sup_{-\infty<c<+\infty} [p(X \leqslant c) - P(X \leqslant c)].$$

If X is a continuous random variable, then the distributions of $D^+(N_X)$ and $D^-(N_X)$ are identical and are independent of the distribution of X. Let $M(\epsilon)$ denote $1 - \exp(-2\epsilon^2)$. Then, if X is continuous,

$$\lim_{N_X \to \infty} P[D^+(N_X) \leqslant \epsilon/N_X^{1/2}] = M(\epsilon) \tag{12}$$

(Wilks, 1962, pp. 336–339).

Upper Confidence Bounds for Continuous X and Y. Suppose that N_X and N_Y observations have been randomly sampled from the X and Y distributions. Birnbaum and McCarty (1958) showed that, if X and Y are continuous, then

$$P(X < Y) - p(X < Y) \leqslant D^+(N_X) + D^-(N_Y). \tag{13}$$

Consequently,

$$P[P(X < Y) \leqslant p(X < Y) + \epsilon] \geqslant P[D^+(N_X) + D^-(N_Y) \leqslant \epsilon]. \tag{14}$$

Then, it follows from Eq. (12) that, for large N_X and N_Y,

$$P[D^+(N_X) + D^-(N_Y) \leqslant \epsilon] \approx \int_0^\epsilon M[N_X^{1/2}(\epsilon - \eta)]\, M'(N_Y^{1/2}\eta)\, d\eta, \tag{15}$$

where M' denotes the derivative of M. Thus, for sufficiently large N_X and N_Y, the probability that $P(X < Y)$ is smaller than $p(X < Y) + \epsilon$ is at most equal to the right side of Eq. (15). Confidence bounds constructed in this way are highly conservative because the inequality in Eq. (14) is quite crude. Note that both upper and lower confidence bounds may be obtained. Thus, for continuous X and Y, finding an upper confidence bound for $P(Y < X)$ is equivalent to finding a lower confidence bound for $P(X < Y)$.

Now, the Birnbaum–McCarty confidence bounds are based upon the approximation in Eq. (15). How large must N_X and N_Y be for this approximation to be accurate? Birnbaum and McCarty stated that the approximation is quite accurate when N_X and N_Y are both 50 or larger. Moreover, they claimed (but could not prove) that, for smaller values of N_X and N_Y, the left side of Eq. (15) is larger than the right. This has the effect of making the confidence bounds still more conservative. Thus, the Birnbaum–McCarty confidence bounds are valid for both small and large sample sizes.

Extension to Noncontinuous X and Y. Owen, Craswell, and Hanson (1964) constructed tables for the Birnbaum–McCarty confidence bound. As previously, let λ equal $N_X/(N_X + N_Y)$. In their table, one enters the value of λ and the confidence level γ and reads out a value $\delta(\lambda, \gamma)$ such that

$$P[P(X < Y) \leqslant p(X < Y) + \delta(\lambda, \gamma)/(N_X + N_Y)^{1/2}] \geqslant \gamma. \tag{16}$$

Owen *et al.* extended Birnbaum and McCarty's results in two ways. First, they showed that Eq. (16) is valid for any X and Y, not just continous X and Y. Second, they showed that Eq. (13) and, therefore, Eq. (16) remain valid when $P(X < Y)$ and $p(X < Y)$ are replaced with $P(X \leqslant Y)$ and $p(X \leqslant Y)$, respectively. From this, it can be seen that Eqs. (13) and (16) are still valid when $P(X < Y)$ and $p(X < Y)$ are replaced with A and a, respectively. Thus, given a sample value of a, an upper confidence bound for A may be constructed using Birnbaum and McCarty's method.

Confidence Intervals. There is a natural way of extending Birnbaum and McCarty's methods in order to obtain confidence intervals, at least for continuous X and Y (Birnbaum, 1956). However, Govindarajulu (1968) has opined that the confidence intervals obtained by this method are quite crude because they are so conservative.

Moreover, to the present author's knowledge, tables for the computation of these intervals have not been published.

An alternative method of computing confidence intervals is to combine a Birnbaum–McCarty upper confidence bound with a lower confidence bound. Thus,

$$P(A \leqslant a + \epsilon) \geqslant \gamma$$

and

$$P(A \geqslant a - \epsilon) \geqslant \gamma$$

imply that

$$P(a - \epsilon \leqslant A \leqslant a + \epsilon) \geqslant 2\gamma - 1.$$

Ury's Confidence Intervals

A method of constructing confidence intervals for A which does not depend upon asymptotic normality has been proposed by Ury (1972). Ury's method involves the use of Chebyshev's inequality in place of asymptotic normality. This inequality states that the probability that a random variable will deviate from its mean by more than k standard deviations is at most $1/k^2$. Now, in the case of a *normal* random variable, the probability that it will deviate from its mean by more than k standard deviations is considerably less than $1/k^2$. Thus, use of Chebyshev's inequality in place of asymptotic normality yields a highly conservative confidence interval.

Now, the expression for $\sigma^2_{\max}$ for the general case (Eq. (7)) implies that $\sigma^2_{\max}$ is at most $1/(4N_L)$. Together with the Chebyshev inequality, this implies that

$$P[| a - A | \leqslant 1/[4N_L(1 - \gamma)]^{1/2}] \geqslant \gamma.$$

Thus, $a \pm 1/[4N_L(1 - \gamma)]^{1/2}$ may conservatively be taken as the two ends of the γ-level confidence interval for A.

Under certain circumstances, the upper end of this γ-level confidence interval will be less (i.e., better) than the γ-level Birnbaum–McCarty upper confidence bound. In particular, for equal N_X and N_Y, the upper end of Ury's interval will be less than the Birnbaum–McCarty upper bound when $\gamma \leqslant 0.925$. The reverse is true when $\gamma \geqslant 0.95$.

For most distributions, Chebyshev's inequality is quite crude. Usually, this inequality is considered to have only theoretical significance rather than any practical value. Thus, the fact that a confidence interval based upon Chebyshev's inequality has been proposed as a serious competitor to the Birnbaum–McCarty confidence bound serves to underline the extreme conservatism of the latter bound.

Width of the Various Confidence Intervals

Consider the confidence intervals based upon asymptotic normality. As a approaches either zero or one, $s_a{}^2$ and all three forms of $s^2_{\max}$ approach zero. Thus, these confidence intervals are narrowest when a is near zero or one and are widest when a is near $\frac{1}{2}$. If N_X and N_Y are equal, $s_{\max}^2$ for the case of a monotonic posterior is roughly $\frac{2}{3}$ of $s_{\max}^2$ for the general case. Thus, for equal N_X and N_Y, confidence intervals calculated from $s_{\max}$ for a monotonic posterior are roughly $(\frac{2}{3})^{1/2}$ or 0.82 the width of confidence intervals calculated from $s_{\max}$ for the general case. Depending on the shape of the *OD* graph, confidence intervals calculated from s_a are likely to be somewhat narrower still.

Suppose a_{high} is an upper confidence bound for A at confidence level γ. Let the difference $a_{\text{high}} - a$ be termed the γ-level confidence margin. Thus, the smaller the confidence margin, the better. How do the confidence margins for upper bounds based upon asymptotic normality compare with the confidence margins for Birnbaum–McCarty upper bounds? For the $s_{\max}$ procedure, the γ-level confidence margin is $z(\gamma)\, s_{\max}$. For the general case, $s_{\max}$ reaches its greatest value, which is $1/[4(N_L - 1)]^{1/2}$, when a equals $\frac{1}{2}$. So, for the $s_{\max}$ procedure, the widest possible confidence margin is $z(\gamma)/[4(N_L - 1)]^{1/2}$. For the Birnbaum–McCarty upper bound, the confidence margin is $\delta(\lambda, \gamma)/[N_X + N_Y]^{1/2}$. For equal N_X and N_Y, the Birnbaum–McCarty confidence margin is roughly twice the largest possible confidence margin for the $s_{\max}$ procedure (Govindarajulu, 1968, Table 2.1). Moreover, as a approaches either zero or one, the $s_{\max}$ confidence margin goes to zero, whereas the Birnbaum–McCarty confidence margin remains constant. Thus, for values of a near zero or one, the Birnbaum–McCarty confidence margin is many times wider than the $s_{\max}$ confidence margin.

Thus, whenever N_X and N_Y are large enough to justify it, it is always preferable to compute confidence intervals or bounds by means of the s_a or $s_{\max}$ procedures rather than the Birnbaum–McCarty or Ury procedures. Unfortunately, this means that in some cases there will be no good way to compute a confidence interval or bound. Consider the following example. Suppose N_X and N_Y both equal 25 and a equals 0.990. It is desired to compute a lower confidence bound for A with 0.95 confidence. Now, $s_{\max}$ for the general case equals 0.020. Thus, the lower confidence bound obtained with the $s_{\max}$ procedure is 0.958. However, since a is only one half of a standard deviation distant from one, the assumption of approximate normality is not justified. On the other hand, using the Birnbaum–McCarty procedure, the lower confidence bound is found to be 0.576, a result which is so conservative that it is almost useless. Thus, in this example, the $s_{\max}$ procedure is efficient but not valid, whereas the Birnbaum–McCarty procedure is valid but not efficient. At present, there is no procedure that yields a confidence bound which is both valid and efficient in a case where the sample sizes are small and a is close to zero or one.

Appendix

Derivation of $\sigma^2_{\max}$ for a Monotonic Posterior

The derivation of the value of $\sigma^2_{\max}$ (Eq. (9)) for continuous X and Y having a monotonic posterior is sketched here. Recall that X and Y have a monotonic posterior if and only if their *OD* graph is convex. Thus, it is desired to find the convex *OD* curve that maximizes $\sigma_a{}^2$ while A (the area above the curve) and the sample sizes N_X and N_Y are held constant. Examination of Eq. (4) shows that $\sigma_a{}^2$ will be maximized if and only if $(N_X - 1)\, B_{XXY} + (N_Y - 1)\, B_{YYX}$ is maximized.

Let x and y denote $P(X \leqslant c)$ and $P(Y \leqslant c)$, respectively. Let the *OD* curve for X and Y be described by the equation $y = g(x)$. Let R denote the region above the *OD* curve and let dA denote the differential of area. Thus

$$A(X, Y) = \int_R dA.$$

It follows from the definition of B_{YYX} (Eq. (5)) that

$$\begin{aligned} B_{YYX} &= \int_0^1 g(x)^2\, dx + \int_0^1 [1 - g(x)]^2\, dx - 2 \int_0^1 g(x)[1 - g(x)]\, dx \\ &= \int_0^1 [4g(x)^2 - 4g(x) + 1]\, dx \\ &= 1 - 8 \int_0^1 \int_{g(x)}^1 (y - 1/2)\, dy\, dx \\ &= 1 - 8 \int_R (y - 1/2)\, dA. \end{aligned}$$

Similarly, it can be shown that

$$B_{XXY} = 1 + 8 \int_R (x - 1/2)\, dA.$$

Now, define

$$J^- = \int_R (x - y)\, dA$$

and

$$J^+ = \int_R (x + y - 1)\, dA.$$

Then,

$$(N_X - 1)\, B_{XXY} + (N_Y - 1)\, B_{YYX} = (N_X + N_Y - 2) + 4[(N_X + N_Y - 2)\, J^- + (N_X - N_Y)\, J^+]. \qquad (17)$$

Assume that $A(X, Y)$ is at least one-half. This entails no loss of generality since either $A(X, Y)$ or $A(Y, X)$ will be greater than or equal to one-half. Let the value of $A(X, Y)$ be denoted by A_0. Consider an arbitrary convex *OD* curve having area A_0 above it. This curve is represented by the smooth convex curve in Fig. 8. Let A and B be two arbitrary points on the arbitrary convex *OD* curve which have the property that they are equidistant from the positive diagonal *CD*. Let this common distance be denoted by h. Let the distance between A and B be denoted by w. Consider a line drawn through the points A and C and another line drawn through the points B and D. Let E be the intersection of these two lines. All of the above points and lines are represented in Fig. 8.

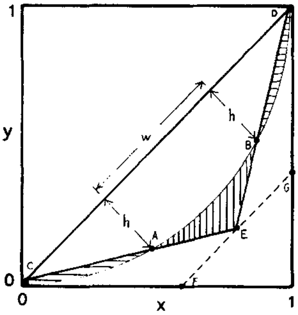

FIG. 8. The *OD* graph *CED* and an arbitrary convex *OD* curve.

Consider the *OD* graph which consists of the straight line segments *CE* and *ED*. Now, the area above the *OD* graph *CED* is a continuous function of w. Let this area be denoted by $CED(w)$. Now, w can assume any value from zero to a maximum of $2^{1/2}$ (the length of the diagonal *CD*). Recall that the area above the arbitrary convex *OD* curve in Fig. 8 is A_0. Since the arbitrary *OD* curve is convex, it can be seen that

$$CED(0) \leqslant A_0 \leqslant \lim_{w \to 2^{1/2}} CED(w).$$

Consequently, there must exist some particular value of w such that

$$CED(w) = A_0. \tag{18}$$

Let the points A and B be chosen such that Eq. (18) is satisfied. Thus, the areas above the arbitrary convex *OD* curve and above the *OD* graph *CED* must be equal. Therefore, in Fig. 8, the sum of the areas of the two horizontally striped regions must equal the area of the vertically striped region. Recall that the arbitrary *OD* curve is convex and that the points A and B are equidistant from the diagonal *CD*. Consequently, everywhere in the vertically striped region, $x - y$ is at least as great as it is everywhere in the horizontally striped regions. Therefore, J^- for the OD graph CED must be greater than or equal to J^- for the arbitrary convex OD curve.

Let the line FG be drawn through the point E and parallel to the diagonal CD. This is illustrated in Fig. 8. Suppose that the point E is moved back and forth along the line between F and G. No matter where E moves on this line, the area above the OD graph CED and the value of J^- for this OD graph will remain constant. Thus, no matter where E is located between F and G, the OD graph CED maximizes J^- over all convex OD graphs having area A_0 above them.

Let E be moved all the way over to the point F. Consider the OD graph CFD. This is illustrated in Fig. 9. The smooth convex curve in Fig. 9 is the same (arbitrary) convex OD curve as in Fig. 8. The areas above the arbitrary convex OD curve and above the OD graph CFD both equal A_0. Therefore, the area of the vertically striped region in Fig. 9 must equal the area of the horizontally striped region. Note that, everywhere in the vertically striped region, $x + y - 1$ must be at least as small as everywhere in the horitontally striped region. Therefore, the OD graph CFD minimizes J^+ over all convex OD graphs having area A_0 above them. Similarly, it can be shown that the OD graph CGD maximizes J^+ over all such graphs.

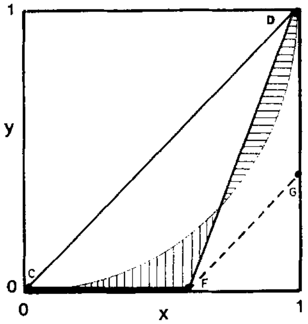

Fig. 9. The OD graph CFD and an arbitrary convex OD curve.

Thus, CFD simultaneously maximizes J^- and minimizes J^+, while CGD simultaneously maximizes both J^- and J^+. It follows from Eq. (17) that CFD maximizes $\sigma_a{}^2$ when $N_X \leqslant N_Y$ and CGD maximizes $\sigma_a{}^2$ when $N_Y \leqslant N_X$. Calculating $\sigma_a{}^2$ for the OD graphs CFD and CGD from Eq. (4) yields the expression for $\sigma^2_{\max}$ for a monotonic posterior given in Eq. (9).

Acknowledgment

The author thanks Raymond O. Collier, Jr., for his helpful comments on earlier drafts of this paper and Teresa Kucala, David Plantenberg, Marcia Deuth, and Helen Buerkle for preparing the manuscript and figures.

REFERENCES

BIRNBAUM, Z. W. On a use of the Mann–Whitney statistic. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 1956, Pp. 13–17.

BIRNBAUM, Z. W., & KLOSE, O. M. Bounds for the variance of the Mann–Whitney statistic. *Annals of Mathematical Statistics*, 1957, **38**, 933–945.

BIRNBAUM, Z. W., & MCCARTY, R. C. A distribution-free upper confidence bound for $\Pr\{Y < X\}$, based on independent samples of X and Y. *Annals of Mathematical Statistics*, 1958, **29**, 558–562.

CRAMÉR, H. *Mathematical methods of statistics*. Princeton: Princeton University Press, 1946.

DARLINGTON, R. B. Comparing two groups by simple graphs. *Psychological Bulletin*, 1973, **79**, 110–116.

GOVINDARAJULU, Z. Distribution-free confidence bounds for $P(X < Y)$. *Annals of the Institute of Statistical Mathematics*, 1968, **20**, 229–238.

GREEN, D. M. General prediction relating yes–no and forced-choice results. *Journal of the Acoustical Society of America* (Abstract), 1964, **36**, 1042.

GREEN, D. M., & MOSES, F. L. On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, 1966, **66**, 228–234.

GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

LEHMANN, E. L. Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics*, 1951, **22**, 165–179.

MANN, H. B., & WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947, **18**, 50–60.

NOETHER, G. E. *Elements of nonparametric statistics*. New York: Wiley, 1967.

OWEN, D. B., CRASWELL, K. J., & HANSON, D. L. Nonparametric upper confidence bounds for $\Pr\{Y < X\}$ and confidence limits for $\Pr\{Y < X\}$ when X and Y are normal. *Journal of the American Statistical Association*, 1965, **59**, 906–924.

POLLACK, I., & HSIEH, R. Sampling variability of the area under the ROC-curve and of d_e'. *Psychological Bulletin*, 1969, **71**, 161–173.

SCHULMAN, A. I., & MITCHELL, R. R. Operating characteristics from yes–no and forced-choice procedures. *Journal of the Acoustical Society of America*, 1966, **40**, 473–477.

SEN, P. K. A note on asymptotically distribution-free confidence bounds for $P\{X < Y\}$, based on two independent samples. *Sankhyā: The Indian Journal of Statistics, Series A*, 1967, **29**, 95–102.

SIMPSON, A. J., & FITTER, M. J. What is the best index of detectability? *Psychological Bulletin*, 1973, **80**, 481–488.

TAYLOR, M. M. Detectability theory and the interpretation of vigilance data. *Acta Psychologica*, 1967, **27**, 390–399. (Reprinted in A. F. Sanders (Ed.), *Attention and performance I*. Amsterdam: North-Holland Publishing Co., 1967. Pp. 390–399.)

URY, H. K. On distribution-free confidence bounds for $\Pr\{Y < X\}$. *Technometrics*, 1972, **14**, 577–581.

VAN DANTZIG, D. On the consistency and power of Wilcoxon's two sample test. *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings, Series A*, 1951, **54** (Also: *Indagationes Mathematicae*, **13**), 1–8.

WILKS, S. S. *Mathematical statistics*, New York: Wiley, 1962.

RECEIVED: October 2, 1973