STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:543–557
Published online 26 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2323

Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: General issues and tail-area-based methods

Robert G. Newcombe[*,†]

Department of Epidemiology, Statistics and Public Health, Wales College of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, U.K.

SUMMARY

For two random variables X and Y, $\theta = \Pr[Y > X] + \frac{1}{2}\Pr[Y = X]$ is advocated as a general measure of effect size to characterize the degree of separation of their distributions. It is estimated by U/mn, a generalization of the Mann–Whitney U statistic, derived by dividing U by the product of the two sample sizes. It is equivalent to the area under the receiver operating characteristic curve. It is readily visualized in terms of two Gaussian distributions with appropriately separated peaks. The effect of discretization of a continuous variable is explored. Tail-area-based confidence interval methods are developed which can be applied to very small samples or extreme outcomes. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: confidence intervals; effect size; generalized Mann–Whitney statistic

1. INTRODUCTION

It is now the established policy of many leading health research journals to prefer point and interval estimates of effect size to p-values as an expression of the uncertainty resulting from limited sample sizes. The p-value is a probabilistic abstraction, which is commonly misinterpreted, particularly when dichotomized at 0.05 or some other conventional α level: ‘significant’ is interpreted as ‘real’ and ‘not significant’ as ‘null’. Lack of awareness of issues such as power remains prevalent in the research community. The common interpretation of ‘not significant’ as ‘the study was too small’ has led to publication bias seriously compromising both the interpretability of the whole corpus of published research and the feasibility of unbiased meta-analyses. Furthermore, commonly the interpretation of p-values requires to heed the issue of multiple comparisons, for which there is no entirely satisfactory answer. Conversely, when the emphasis is shifted to point and interval estimation, correctly interpreted, i.e. not

[*] Correspondence to: Robert G. Newcombe, Department of Epidemiology, Statistics and Public Health, Wales College of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, U.K.
[†] E-mail: newcombe@cf.ac.uk

Received 23 August 2004
Accepted 23 May 2005

merely 'hypothesis testing by the back door' by examining whether the interval includes the null hypothesis value, the resulting figures are expressed on a measurement scale that is directly interpretable by researchers: for quantitative variables, in units such as mm Hg for blood pressure; for binary outcomes, in terms of naturally interpretable proportions. This is more informative than the p-value.

In a study comparing two independent groups, in the case of a binary outcome several measures of effect size are available, notably the difference of proportions and its reciprocal, the number needed to treat, which are *absolute* measures, and relative risk and odds ratio which are *relative* measures. For a continuous outcome measure, the natural absolute measure is the difference in means, for which a confidence interval is readily obtained corresponding to either the Student or Welch test. In the event of seriously non-Gaussian distributional form, or for an ordinal outcome, a median difference may be estimated, with a confidence interval [1]. Once again, use of an absolute measure leads to point and interval estimates on the original scale of measurement. Conversely, when the binary variable is the outcome and the continuous one is the explanatory variable, the logistic regression coefficient can characterize the strength of relationship. The units are then based on but inverse to the original scale. For example, in a logistic regression of presence of disease by age in a cross-sectional study, the regression coefficient might be 0.070 per year, interpreted as an increase in odds of presence of disease by a factor $e^{0.070} = 1.073$ per year.

However, such measures are only helpful as a communication of the results if the scale of measurement is familiar to the relevant research community. Often this is not the case. Many comparisons involve outcomes such as visual analogue scales for self-rated pain or other symptom levels, or psychometric rating scales, albeit validated according to accepted criteria; in both instances, readers not involved in the original research may be at a loss to interpret a 1- or a 10-unit difference for clinical importance. It is then more informative to quote a relative measure of effect size. For the homoscedastic Gaussian case, the natural relative measure is the standardized difference δ obtained by dividing the difference of means by the (pooled) standard deviation.

An analogous relative effect size measure which does not embody parametric assumptions, applicable to both continuous and ordinal data, is $\theta = \Pr[Y > X] + \frac{1}{2}\Pr[Y = X]$, where X and Y denote independent random variables on the same support. It is estimated by $\hat{\theta} = U/mn$, Mann–Whitney statistic U divided by the product of the two sample sizes m and n. U is defined in the usual manner as $\sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}$ where $U_{ij} = 1, \frac{1}{2}$ or 0 according as Y_j is greater than, equal to or less than X_i. Then U/mn serves as an obvious empirical estimate of θ. It can be regarded as expressing the degree of overlap (or, conversely, separation) between the values constituting the two samples, and is applicable to both continuous and ordinal cases.

Hitherto, U/mn has seen little use as a measure of effect size. This is largely because statisticians and software have done little to promote its use, so that the wider research community remains unaware of its usefulness as a widely applicable measure, more informative than a p-value. Furthermore, it is not obvious to users how to obtain a confidence interval from widely used software such as SPSS, and the method implemented there does not cope well with small sample sizes or extreme outcomes. The purpose of this and an accompanying article [2] is to show the feasibility of calculating appropriate confidence intervals for this measure by convenient methods, enabling this measure to take its place alongside other well-established though arguably not immediately intuitive measures such as the odds ratio.

Both U/mn and the corresponding theoretical value θ range from 0 to 1, with values of 0 and 1 indicating no overlap. On the null hypothesis that X and Y are identically distributed, $\theta = 0.5$, but the converse does not hold. θ can be regarded as a measure of separation, or equally, a measure of discriminatory ability [3]. It is equivalent to AUROC, the area under the receiver operating characteristic curve, and the mean ridit [4]. It has been termed the 'probability of concordance', 'common language effect size' [5] and 'measure of stochastic superiority' [6]. It is discussed in the form presented here by Fay and Gennings [7], and is linearly related to Somers' D which is $2\theta - 1$ [8–10]. Furthermore, in a study such as Ukoli *et al.* [11] evaluating height and weight of children against international norms, the mean centile score is equivalent to the U/mn value that would be obtained by comparing the series of interest against the normative series. A related index, Agresti's α defined as $\Pr[Y > X]/\Pr[X > Y]$ [12–14] will not be considered further here as it is potentially unbounded.

Hanley and McNeil [15] presented a Wald or delta method for calculating a confidence interval for the AUROC without parametric assumptions. A procedure to plot the ROC curve and obtain this confidence interval is available in SPSS from the Graphs menu. Hanley and McNeil also developed a modification based on assumed exponential distributions. Both methods have the deficiencies of producing zero width intervals in extreme cases and limits outwith [0,1] in near-extreme ones. SPSS does not issue any warning when this occurs. The Wald variance imputed to U/mn can also be used in hypothesis testing, either for a one-sample test of H_0: $\theta = \theta_0$ for some specified value θ_0 or a two-sample test of H_0: $\theta_1 = \theta_2$. A confidence interval for θ_1–θ_2 may be calculated analogously.

An alternative asymptotic approach was developed by Halperin *et al.* [16] and Mee [17]. This method is closely analogous to the Wilson [18] score interval for the single proportion: confidence limits are obtained by inversion, i.e. solving a quadratic of the form $|\theta - \hat{\theta}| = z\sqrt{\{\theta(1-\theta)/\hat{N}_J\}}$. This avoids the deficiency of the Wald method in extreme and near-extreme cases, but only because the formula for the pseudo-sample size $\hat{N}_J$ incorporates an *ad hoc* shift modification to cope with these cases.

Work on boundary-respecting methods has been limited. Obuchowski and Lieber [19] presented lower confidence limits for $\hat{\theta} = 1$, only. A $\tanh^{-1}$ transformation for D recommended by Simonoff *et al.* [20], equivalent to a logit transformation for θ, was implemented by Edwardes [8] for D estimated from clustered data. In extreme cases, the variance estimate is calculated by a shift modification similar to the Mee method above.

The measure $\hat{\theta} = U/mn$ studied here is mathematically identical to the AUROC, nevertheless, calculating it and plotting the ROC curve should be regarded as meeting rather different objectives. When the purpose is to characterize the trade-off between sensitivity and specificity of a single test as the cutpoint is altered, the ROC curve is the obvious summary of the data. But it should be borne in mind that in this situation the AUROC has the deficiency that it disregards the relative importance of the two types of errors. (Moreover, use of the curve to identify an optimal cutpoint, as, e.g. in Reference [21] is logically flawed due to overfitting to the vagaries of fine structure of the training data set. Arguably, as in Reference [10], the usefulness of the ROC area is mainly that it can be compared with other ROC areas for different diagnostic tests for the same disease in the same population, because the difference between two ROC areas for different tests for the same disease in the same population is governed by differences in disease status in pairs of patients in that population whose results for two different diagnostic tests are non-concordant. Comparisons of ROC areas for the same disease between populations are more likely to be misleading, as they may reflect differences in the population distributions of the diagnostic indicators, rather than differences in the predictive performance of the diagnostic indicators.)

The present paper explores some issues relating to the U/mn statistic, and shows the feasibility of developing confidence interval methods based on tail areas, albeit of limited practical usefulness. The accompanying paper develops more widely applicable asymptotic methods and evaluates their performance.

2. AN EXAMPLE

Fearnley and Williams [22] studied a series of 19 men charged with offence against the person and referred to the forensic psychiatry service. Subjects were rated for impulsivity using the Monroe dyscontrol scale (RDS) [23, 24] that can range from 0 to 54. Scores for the 19 subjects were 4*, 5, 8, 12, 14, 16, 18, 19, 23, 24, 28, 29, 35, 36*, 38*, 39*, 39, 44* and 45*. Scores for 6 subjects with history of head injury leading to loss of consciousness for over 48 hours are asterisked. The distributional form does not appear close to Gaussian, especially for the head injury cases.

Using Minitab, the estimated median difference between subjects with and without head injury is 17, with a default confidence interval reported as 1–27. This is a 95.2 per cent interval; construction of a 95 per cent interval is precluded by the highly discrete behaviour with such small sample sizes. But these figures are only readily interpreted by those familiar with the RDS. Alternatively, the results may be summarized by calculating $\hat{\theta}=0.801$. For this, a $100(1-\alpha)$ per cent interval can be calculated for any α; a 95 per cent confidence interval is 0.515–0.933 using method 5 of the accompanying paper [2]. These figures clearly indicate a very substantial observed effect size, with a wide confidence interval reflecting small sample uncertainty. The upper limit of 0.933 indicates that there may possibly be little overlap between the two distributions in the population from which these observations have been drawn. The lower limit is just above 0.5, corresponding to an obviously trivial difference, yet implying rejection of the null hypothesis at a conventional α level very much in the same way as the corresponding 2-tailed exact conditional p-value of 0.036. Familiarity with the RDS is not a prerequisite to interpreting these relative measures.

3. RELATIONSHIP BETWEEN U/mn AND STANDARDIZED DIFFERENCE

For the general Gaussian case with $X\sim\mathrm{N}(\mu_1,\sigma_1^2)$ and $Y\sim\mathrm{N}(\mu_2,\sigma_2^2), \theta=\Phi((\mu_2-\mu_1)/\sqrt{\sigma_1^2+\sigma_2^2})$ [25, 26] where Φ denotes the cdf of the standard Gaussian distribution. In particular, for the homoscedastic Gaussian case, expressed without loss of generality as $X\sim\mathrm{N}(0,1)$ and $Y\sim\mathrm{N}(\delta,1), \theta$ reduces to $\Phi(\delta/\sqrt{2})$ [3]. Table I gives values of δ corresponding to selected values of θ and vice versa. We choose to estimate θ rather than δ, as it is less dependent on distributional assumptions, thus more satisfactory than the standardized difference in extreme cases, as $\hat{\theta}=1$ suggests δ is large without implying any specific value. The degree of separation implied by a particular value of θ is readily visualized by plotting two Gaussian

Table I. Correspondence between proposed measure θ and standardized difference δ in the homoscedastic Gaussian case.

θ	δ	δ	θ
0.50	0.000	0.00	0.500
0.55	0.178	0.25	0.570
0.60	0.358	0.50	0.638
0.65	0.545	0.75	0.702
0.70	0.742	1.00	0.760
0.75	0.954	1.25	0.812
0.80	1.190	1.50	0.856
0.85	1.466	1.75	0.892
0.90	1.812	2.00	0.921
0.95	2.326	2.50	0.961
0.99	3.290	3.00	0.983
0.999	4.370	3.50	0.993
		4.00	0.998

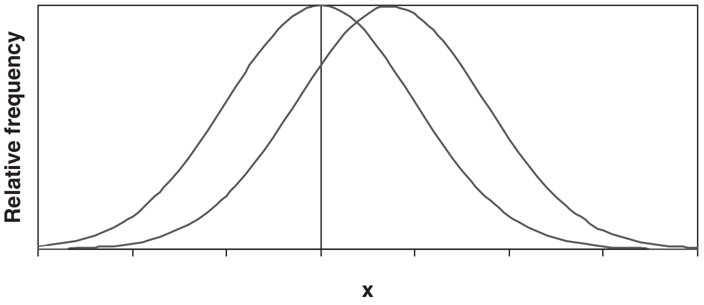

Figure 1. Relative frequency curves for Gaussian distributions N(0,1) and N(δ, 1) with $\delta = 0.7416$ corresponding to $\theta = 0.7$.

curves with equal standard deviations and peaks separated by the corresponding standardized difference (Figure 1).

At the planning stage of a study, best practice is to elicit separately from the investigator the difference between group means that is judged both plausible in the light of current understanding and importantly large in the context, and the relevant standard deviation, calculate the corresponding δ and use, for example, the Altman nomogram [27] to obtain the sample size required to achieve the desired power. Sometimes the investigator is reluctant to specify these two parameters separately, but only δ as a relative effect measure. In this situation some investigators may be more ready to specify the target degree of separation in terms of θ. This θ could then be converted to the corresponding δ assuming a homoscedastic Gaussian model, and the sample size assessed accordingly. Alternatively, it is used directly as a basis for planning sample size based on the Mann–Whitney test, for example, in nQuery Advisor 4.0. Though θ is specified as $\Pr[X<Y]$ for both the cases ‘continuous outcome’ and ‘ordered categories’, and, moreover, the formula used does not incorporate tie correction, and so may be incorrect for the latter case.

4. CONDORCET AND RELATED PARADOXICAL EFFECTS

Consider three independent random variables X, Y and Z. In the simplest, homoscedastic Gaussian case with $X{\sim}\mathrm{N}(\mu_1, 1), Y{\sim}\mathrm{N}(\mu_2, 1)$ and $Z{\sim}\mathrm{N}(\mu_3, 1)$, the standardized differences δ_{XY}, δ_{XZ} and δ_{YZ} expressing their separation are simply related by $\delta_{XZ} = \delta_{XY} + \delta_{YZ}$. From the relationship $\theta = \Phi(\delta/\sqrt{2})$, the corresponding θ_{XY}, θ_{XZ} and θ_{YZ} are related by $\Phi^{-1}(\theta_{XZ}) = \Phi^{-1}(\theta_{XY}) + \Phi^{-1}(\theta_{YZ})$. We expect a similar relationship to hold approximately for the $\hat{\theta}$'s, in particular, that when $\hat{\theta}_{XY}$ and $\hat{\theta}_{YZ}$ are both greater than $\frac{1}{2}$, $\hat{\theta}_{XZ} \geqslant \max(\hat{\theta}_{XY}, \hat{\theta}_{YZ})$, and also to some degree when the assumptions of Gaussian distributional form and homoscedasticity are relaxed. But very different patterns can sometimes occur.

Since U/mn is simply an alternative way of expressing the information from the Mann–Whitney U statistic, the classical Condorcet (or Escher staircase) non-transitive dominance paradox [28] applies to U/mn just as to the test. It is exemplified by the data sets $\mathbf{X} = \{3, 5, 7\}$, $\mathbf{Y} = \{1, 6, 8\}$ and $\mathbf{Z} = \{2, 4, 9\}$. Here, a Kruskal–Wallis test comparing the three groups on an equal footing would assign identical rank sums to all. But ranks are altered by omitting uninvolved groups, so here the Mann–Whitney criterion ranks pairs of groups $\mathbf{X} < \mathbf{Y} < \mathbf{Z} < \mathbf{X}$. From the fact that $\hat{\theta}_{XY} = \hat{\theta}_{YZ} = 5/9$, one would expect $\hat{\theta}_{XZ}$ to be greater than 5/9, around 0.610 by summating the corresponding δ's; in fact, it is only 4/9.

Conversely, for any n, consider $\mathbf{X} = \{1, 2, \ldots, n\}$, $\mathbf{Z} = \{n+1, n+2, \ldots, 2n\}$, and $\mathbf{Y} = \mathbf{X} \cup \mathbf{Z}$. Then $\hat{\theta}_{XY} = \hat{\theta}_{YZ} = \frac{3}{4}$ whereas $\hat{\theta}_{XZ} = 1$. A real example approaching this situation involving three independent groups occurs in a parasitology study [29] comparing *Trinoton luridum* (X), *Anatoecus dentatus* (Y) and *Carnus hemapterus* (Z) for which $\hat{\theta}_{XY} = 0.759$, $\hat{\theta}_{YZ} = 0.725$ and $\hat{\theta}_{XZ} = 0.968$. Similar behaviour could result from effective triage of a presenting series of subjects Y into subgroups X and Z, for example, when a telephone ambulance dispatch prioritization algorithm for injuries is validated by reference to an injury severity scale. Or equivalently, for a non-age-dependent biochemical gene product marker for the autosomal dominant disorder, Huntington's disease, for which penetrance increases from 0 at birth in a sigmoid fashion to approach 100 per cent in elderly heterozygotes [30]. Here, group Z comprises known affected subjects, group Y young clinically unaffected offspring of established heterozygotes, whose genotype is unknown but whose posterior risk of heterozygosity is still at its 50 per cent prior level, and group X their elderly counterparts whose posterior risk has declined to near zero.

The hypothetical examples above may be extremes for the behaviour of $\hat{\theta}_{XZ}$ given $\hat{\theta}_{XY}$ and $\hat{\theta}_{YZ}$. The possibility of behaviour so far from additivity on the δ scale should be regarded as a curiosity rather than a serious limitation in practice.

5. EFFECT OF DISCRETIZATION ON *U/mn*

Hanley and McNeil [15] pointed out that $\hat{\theta}$ values estimated directly as U/mn are not interchangeable with those obtained by fitting continuous distributions to the data. They illustrated this with numerical results and with ROC 'curves' for the same data constructed as line segments and smooth convex curves through points. It is well recognized [31, 32] that discretizing continuous data reduces the power available in hypothesis testing with a shift alternative; this occurs predominantly because discretization tends to reduce θ. It is important to distinguish three cases:

CF continuous distributions, leading (theoretically) to data free of ties
CR continuous distributions, leading (in practice) to rounded or discretized data
DT discrete distributions, leading naturally to tied data.

CF and DT should be regarded as genuinely distinct cases, and the $\hat{\theta}$ for each should be regarded as a meaningful estimate, not requiring adjustment. Hanley and McNeil's exemplar data set comprised $m = 58$ 'normal' and $n = 51$ 'abnormal' ratings of computed tomography images expressed on a 5-point Likert-scale. $\hat{\theta}$ calculated directly from the data was 0.893, contrasting with an estimate $\tilde{\theta} = 0.911$ estimated by fitting a smooth Gaussian-based ROC curve by maximum likelihood. Hanley and McNeil do not indicate clearly whether the 0.893 or the 'adjusted' figure of 0.911 should be quoted. It appears that here it would be less appropriate to quote 0.911, for two reasons. The data are poorly fitted by a homoscedastic Gaussian model. Furthermore, there is the more basic reason that the investigators chose to use a 5-point scale, not considering it meaningful to elicit ratings on a more finely subdivided or continuous scale. Under these circumstances the value 0.893 obtained by regarding the data as arising from case DT is the appropriate point estimate. Subject to distributional reservations, 0.911 represents the degree of separation that would have been obtained had it been meaningful to express on a continuous scale, but this is regarded as counterfactual. It is in the hybrid case CR that it is important to examine the effect of bias that may have been introduced by discretization. In case DT, the number of categories used is an inherent part of the structure of the data, and ideally confidence interval methods for this case should be developed which incorporate this information. Nevertheless, the evaluation in the subsequent paper [2] shows that asymptotic confidence interval methods developed primarily for case CF, with the assumption that the underlying distributions of the two random variables X and Y are absolutely continuous so that ties occur with probability zero, also perform well on data generated from a discrete distribution with 5 categories.

To examine the degree of bias introduced when data from inherently continuous distributions is discretized or rounded, we develop analyses analogous to those in Reference [32] by considering three readily tractable models:

Double Gaussian: $X \sim \mathrm{N}(-\delta/2, 1)$, $Y \sim \mathrm{N}(\delta/2, 1)$,
Double beta: $X \sim \mathrm{B}(1, \alpha)$, $Y \sim \mathrm{B}(\alpha, 1)$,
Beta-uniform: $X \sim \mathrm{B}(1, 1)$, $Y \sim \mathrm{B}(\alpha, 1)$.

For the double Gaussian model, theoretically the support is X, $Y \in (-\infty, \infty)$, but we use an effective range between $\pm(\delta/2 + 3)$, of width $r = \delta + 6$. The support for double beta and beta-uniform models is X, $Y \in$ [0,1], with effective range width $r = 1$.

We examine the effect of discretizing the sample space by cutpoints at equal intervals of width r/k. For distributions on [0,1], and $k = 3$ and 4, we use 4 sets of cutpoints, labelled as offsets 0 to 3 defined modulo 4:

$k = 3$:	Offset 0: 1/6, 1/2, 5/6	$k = 4$:	Offset 0: 1/4, 1/2, 3/4
	Offset 1: 3/12, 7/12, 11/12		Offset 1: 1/16, 5/16, 9/16, 13/16
	Offset 2: 1/3, 2/3		Offset 2: 1/8, 3/8, 5/8, 7/8
	Offset 3: 1/12, 5/12, 9/12		Offset 3: 3/16, 7/16, 11/16, 15/16.

Similarly, for other odd and even values of k. Thus, offset 0 places a boundary at $\frac{1}{2}$, and offset 2 produces an interval centred at $\frac{1}{2}$. The number of bins produced is generally $k+1$, but only k for offset 0 with even k and offset 2 with odd k. For the Gaussian model, we split the interval $(-\delta/2-3,\ \delta/2+3)$ in these proportions, with the first and last categories open ended. Thus, for each combination of the three models for the CF case with $\theta=0.55$, 0.75, 0.90 and 0.95, we generate the relevant pair of theoretical distributions. Then for each combination of $k=3$, 4, 6 and 12 and offsets 0–3, we discretize them at the relevant cutpoints and calculate the resulting θ value that would occur in case CR, θ^* exactly from the discretized distributions.

Table II shows the resulting discretized θ^* values. In every case θ^* underestimates θ. The bias is small if k is large and for midrange θ. It is serious if k is small and θ is close to 1. In this situation the degree of bias depends on distributional form, being least for the double beta model and greatest for the Gaussian, with beta-uniform generally intermediate—in effect, $k=3$ is a more drastic discretization for the Gaussian model with its long, thin tails. Furthermore, the offset can substantially affect the degree of bias—not surprisingly, this occurs particularly when k instead of $k+1$ bins ensue. In practice, of course, a data set will give limited information on what offset applies—hence, in interpreting an observed U/mn value in case CR, it is important to bear in mind the resulting uncertainty in the degree of underestimation.

6. TAIL AREA MODELLING APPROACH

In order to develop confidence interval methods, assume that $X_1, X_2, \ldots, X_m$ are iid, drawn from some density $f(x)$, and $Y_1, Y_2, \ldots, Y_n$ are iid, with density $g(y)$. The corresponding cdfs are F and G. The support for f and g may be the doubly infinite real line or a finite subset such as [0,1]—these two cases are equivalent as it is desirable for a method to be invariant under monotonic transformation of the support space, such as logit. The supports of f and g need not be identical, but must overlap, else $\theta\equiv 1$ (or 0). Then θ is defined as

$$\Pr[Y>X]=\int f(x)(1-G(x))\,\mathrm{d}x=\int g(y)F(y)\,\mathrm{d}y$$

Though the methods developed here are intended primarily for case CF, it is preferable to redefine θ as $\Pr[Y>X]+\frac{1}{2}\Pr[Y=X]$ in order to accommodate tied data.

In practice we do not know the distributional form of X and Y. Nevertheless, we investigate the feasibility of constructing tail-based CI's for θ for various simple distributional assumptions. The primary objective of U is to detect a shift in location, so an obvious choice of model is $X\sim\mathrm{N}(0,1)$, $Y\sim\mathrm{N}(\delta,1)$ for some δ. As above, $\theta=\Phi(\delta/\sqrt{2})$, so that values of δ above and below 0 correspond to θ above or below $\frac{1}{2}$. But to develop tail-based confidence intervals, we also need to obtain probabilities of each possible sequence of X's and Y's. Unfortunately, in general, order statistics from a Gaussian model are not closed form [33], and the issue is a similar one here.

Accordingly, we seek more tractable alternative models. Ones experimented with include: *Uniform*:

$$f(x)=1\ \ (0\leqslant x\leqslant 1)\ \text{else}\ 0$$

$$g(y)=1\ \ (0\leqslant y-a\leqslant 1)\ \text{else}\ 0\quad \text{for some } a\in[-1,1]$$

Table II. Effect on θ of discretizing continuous distributions based on homoscedastic double Gaussian, double beta and beta-uniform models.

			Degree of discretization			
		Offset	$k=3$	$k=4$	$k=6$	$k=12$
Double Gaussian model						
θ	δ					
0.55	0.178	0	0.5382	0.5418	0.5459	0.5489
		1	0.5368	0.5416	0.5459	0.5489
		2	0.5354	0.5415	0.5459	0.5489
		3	0.5368	0.5416	0.5459	0.5489
0.75	0.954	0	0.6919	0.7070	0.7279	0.7441
		1	0.6797	0.7050	0.7279	0.7441
		2	0.6673	0.7030	0.7279	0.7441
		3	0.6797	0.7050	0.7279	0.7441
0.90	1.812	0	0.8256	0.8428	0.8701	0.8922
		1	0.7992	0.8373	0.8700	0.8922
		2	0.7727	0.8318	0.8699	0.8922
		3	0.7992	0.8373	0.8700	0.8922
0.95	2.326	0	0.8841	0.8990	0.9235	0.9432
		1	0.8529	0.8919	0.9233	0.9432
		2	0.8213	0.8848	0.9231	0.9432
		3	0.8529	0.8919	0.9233	0.9432
Double beta model						
θ	α					
0.55	1.104	0	0.5456	0.5456	0.5479	0.5494
		1	0.5450	0.5469	0.5484	0.5495
		2	0.5426	0.5472	0.5486	0.5496
		3	0.5450	0.5469	0.5484	0.5495
0.75	1.647	0	0.7279	0.7317	0.7417	0.7479
		1	0.7255	0.7352	0.7429	0.7481
		2	0.7181	0.7363	0.7433	0.7481
		3	0.7255	0.7352	0.7429	0.7481
0.90	2.431	0	0.8705	0.8799	0.8912	0.8978
		1	0.8688	0.8818	0.8917	0.8979
		2	0.8636	0.8824	0.8918	0.8979
		3	0.8688	0.8818	0.8917	0.8979
0.95	3.000	0	0.9236	0.9336	0.9429	0.9483
		1	0.9228	0.9345	0.9431	0.9483
		2	0.9198	0.9348	0.9431	0.9483
		3	0.9228	0.9345	0.9431	0.9483
Beta uniform model						
θ	α					
0.55	1.222	0	0.5458	0.5460	0.5481	0.5495
		1	0.5449	0.5472	0.5486	0.5496
		2	0.5432	0.5473	0.5487	0.5496
		3	0.5455	0.5469	0.5485	0.5496

Table II. *Continued.*

			Degree of discretization			
		Offset	$k=3$	$k=4$	$k=6$	$k=12$
0.75	3.000	0	0.7292	0.7344	0.7431	0.7483
		1	0.7271	0.7368	0.7438	0.7484
		2	0.7222	0.7373	0.7439	0.7484
		3	0.7277	0.7364	0.7436	0.7483
0.90	9.000	0	0.8676	0.8557	0.8797	0.8948
		1	0.8605	0.8711	0.8855	0.8957
		2	0.8246	0.8775	0.8873	0.8959
		3	0.8530	0.8726	0.8852	0.8956
0.95	19.000	0	0.9088	0.8739	0.9114	0.9394
		1	0.9184	0.9020	0.9257	0.9425
		2	0.8332	0.9226	0.9337	0.9436
		3	0.8738	0.9227	0.9308	0.9425

Triangular:

$$f(x) = 1 - |x| \;\; (-1 \leqslant x \leqslant 1) \text{ else } 0$$

$$g(y) = 1 - |y-a| \;\; (-1 \leqslant y-a \leqslant 1) \text{ else } 0 \quad \text{for some } a \in [-1, 1]$$

Bi-exponential:

$$f(x) = \tfrac{1}{2}\,\mathrm{e}^{-|x|} \;\; (-\infty < x < +\infty)$$

$$g(y) = \tfrac{1}{2}\,\mathrm{e}^{-|y-a|} \;\; (-\infty < y < +\infty) \quad \text{for some } a \in (-\infty, +\infty)$$

All three lead to closed form outcome probabilities, tractable for very small m and n, nevertheless rapidly become unwieldy as m and n increase.

A more promising approach is the double beta model, $X \sim \mathrm{B}(1, \alpha)$, $Y \sim \mathrm{B}(\alpha, 1)$, for $\alpha \geqslant 1$, representing θ between $\frac{1}{2}$ and 1. (For θ between 0 and $\frac{1}{2}$, we interchange these.) Unlike the models above, these are asymmetrical, to equal and opposite degrees, but if we consider the equivalent distributions for logit X and logit Y, say, the degree of evident asymmetry is greatly reduced. With this model, probabilities of the extreme outcomes $U=0$ and $U=mn$ are readily derived for any m and n, as finite alternating sums of beta functions. Furthermore, probabilities of all outcomes can be obtained for the cases $m=1$ or $n=1$, though these are of course of little practical use.

The final model chosen is the beta-uniform model, as described below. This is chosen purely on grounds of expediency, as all outcome probabilities are readily calculated. The obvious precedent for such an unashamedly pragmatic approach is the choice of the same distribution, beta, as conjugate prior for binomial parameter estimation in the Bayesian paradigm, and for the same reasons.

The beta-uniform model is not symmetrical. Moreover, except in the H_0 case the two distributions are qualitatively different in shape, one uniform, the other skew. Two different versions, models 1 and 2 can be fitted, which generally lead to different results. The ambiguity is resolved by calculating the probability for each outcome on both models, and averaging these, to give the final model 3. The resulting method is equivariant in the sense of Blyth and Still [34], i.e. if (θ_1, θ_2) is the calculated interval corresponding to $U=u$, then $(1-\theta_2, 1-\theta_1)$ is the interval corresponding to $U=mn-u$.

7. THE BETA-UNIFORM MODEL

All models have support [0,1] and are indexed by a parameter λ that ranges on $(-\infty, +\infty)$. Values of λ above and below 0 correspond to θ above and below $\frac{1}{2}$. If $\lambda \geqslant 0$, let $\delta = 1+\lambda \geqslant 1$. If $\lambda \leqslant 0$, let $\gamma = 1-\lambda \geqslant 1$.

$$\begin{array}{llll}
\textit{Model } 1: & \forall\lambda, & X \sim \mathrm{B}(1,1) & f(x)=1 \\
\text{Also:} & \text{for } \lambda \geqslant 0, & Y \sim \mathrm{B}(\delta,1) & g(y)=\delta y^{\delta-1} \\
 & \text{for } \lambda \leqslant 0, & Y \sim \mathrm{B}(1,\gamma) & g(y)=\gamma(1-y)^{\gamma-1} \\
\textit{Model } 2: & \forall\lambda, & Y \sim \mathrm{B}(1,1) & g(y)=1 \\
\text{Also:} & \text{for } \lambda \geqslant 0, & X \sim \mathrm{B}(1,\delta) & f(x)=\delta(1-x)^{\delta-1} \\
 & \text{for } \lambda \leqslant 0, & X \sim \mathrm{B}(\gamma,1) & f(x)=\gamma x^{\gamma-1}
\end{array}$$

On both models,

$$\theta = \delta/(\delta+1) = (\lambda+1)/(\lambda+2) \quad \text{if } \lambda \geqslant 0$$

$$\theta = 1/(\gamma+1) = 1/(2-\lambda) \quad \text{if } \lambda \leqslant 0$$

Then the probability of any sequence of X's and Y's is easily derived.

$$\left.\begin{array}{r}\text{Define } q_i = 1 \text{ if the } i\text{th element of the sequence is a } Y \\ q_i = 0 \text{ if the } i\text{th element of the sequence is an } X\end{array}\right\} \text{for } i=1,2,\ldots,t=m+n.$$

Then, for example, with $m=2$ and $n=3$, the vector $\mathbf{q}=\{0,1,1,0,1\}$ corresponds to the outcome *XYYXY*, or in order statistics notation, $x_{(1)} < y_{(1)} < y_{(2)} < x_{(2)} < y_{(3)}$.

Define $p_i = 1-q_i, i=1,2,\ldots,t$.
Let $r_i = \sum_{j=1}^{i} q_j$ — a partial sum, or the number of Y's among the first i observations.
Similarly, let $s_i = \sum_{j=1}^{i} p_j$, $v_i = \sum_{j=i}^{t} p_j$, $w_i = \sum_{j=i}^{t} q_j$.
Then the expressions for $\Pr[\mathbf{q}]$ are remarkably simple and highly tractable:

$$\begin{array}{ll}
\text{Model } 1, \lambda \geqslant 0: m!n!\delta^n/\prod_{i=1}^{t}(i+r_i\lambda) & \equiv \prod_{i=1}^{t}(p_i s_i + q_i r_i(1+\lambda))/(i+r_i\lambda) \\
\text{Model } 1, \lambda \leqslant 0: m!n!\gamma^n/\prod_{i=1}^{t}(i-w_{t+1-i}\lambda) & \equiv \prod_{i=1}^{t}(p_i s_i + q_i r_i(1-\lambda))/(i-w_{t+1-i}\lambda) \\
\text{Model } 2, \lambda \geqslant 0: m!n!\delta^m/\prod_{i=1}^{t}(i+v_{t+1-i}\lambda) & \equiv \prod_{i=1}^{t}(p_i s_i(1+\lambda) + q_i r_i)/(i+v_{t+1-i}\lambda) \\
\text{Model } 2, \lambda \leqslant 0: m!n!\gamma^m/\prod_{i=1}^{t}(i-s_i\lambda) & \equiv \prod_{i=1}^{t}(p_i s_i(1-\lambda) + q_i r_i)/(i-s_i\lambda).
\end{array}$$

It is easily verified that the Pr[$\mathbf{q}$] for all possible $\mathbf{q}$ summate to 1, and that substituting $m = n = 1$ leads back to the previous formulae for θ. The second expressions are manageable for large m and n.

We then define $\mathrm{Pr}_{\mathrm{model3}}[\mathbf{q}] = \frac{1}{2}(\mathrm{Pr}_{\mathrm{model1}}[\mathbf{q}] + \mathrm{Pr}_{\mathrm{model2}}[\mathbf{q}]) \quad \forall \mathbf{q}$.

Hence, confidence limits for θ are obtained iteratively by seeking values of λ that make the appropriate aggregated tail areas equal to $\alpha/2$, i.e. 0.025, etc. The procedure is analogous to that used to obtain 'exact' [35] intervals for the binomial proportion, which align the minimum coverage with $1 - \alpha$. Mid-p intervals [36] which aim to align the mean coverage with $1 - \alpha$ may also be obtained by giving the observed outcome a weight of $\frac{1}{2}$ when accumulating tail probabilities. (As an alternative to what is essentially the construction of two 1-sided $1 - \alpha/2$ limits, shortened, inherently 2-sided intervals analogous to the Sterne [37] and Blaker [38] intervals for the binomial parameter could also be developed.)

However, an additional complexity arises in defining tail areas. Generally, several different $\mathbf{q}$ outcomes correspond to the same value of U. We could accumulate 'exact' tail areas by two approaches: summating probabilities of U values that are as or more extreme than that observed; or by summating probabilities of $\mathbf{q}$ vectors that either have a more extreme U, or else have the same U, but a similar or lower probability to that observed. The latter approach does not lead to a workable method, because the behaviour of tail probabilities accumulated in this way as the centre of symmetry $\lambda = 0$ is approached or passed is discontinuous, and moreover non-monotonic. Thus, tail areas are defined entirely from the distribution of U.

In general, this, and any tail area modelling approach, involves calculating probabilities for a large number of outcomes, e.g. 184 756 if $m = n = 10$. Only for extreme outcomes can we bypass enumerating these. So the tail area modelling approach is really only suitable for small samples or extreme outcomes. In general an asymptotic method analogous to the score interval [18] for the single proportion would be preferable. Such methods are developed and evaluated in the accompanying paper.

Specimen results for the beta-uniform model are shown in Table III. These are lower limits cutting off a 0.025 tail area, as normally used in forming 95 per cent confidence intervals. For the first two blocks of the table, upper limits may be found from entries for complementary values of U, thus the 95 per cent 'exact' interval for $m = 2$, $n = 3$, $U = 1$, $\hat{\theta} = 0.1667$ runs from 0.006202 to 1 – 0.216181, i.e. 0.783819. Throughout the final block, the upper limit is 1. The three values in bold are exact and readily verified from first principles.

8. DISCUSSION

I have sought to present $\hat{\theta} = U/mn$ as a very general and widely applicable, indeed much needed measure of separation of two frequency distributions. Any novel measure will only achieve wide currency and fulfil its potential if it becomes accepted by both statisticians and the wider research community as appropriate and useful.

Scepticism has been expressed on several grounds. Firstly, that researchers will be unable to visualize or identify the implication of any specified value of U/mn. Also, for another relative measure of effect size, Cohen's κ [39] in an otherwise excellent paper Landis and Koch [40] tentatively identified six zones ranging from poor ($\kappa < 0$) through slight, fair, moderate and substantial to almost perfect ($\kappa > 0.8$). Unfortunately, these labels have been seized upon by the user community and applied widely without heed to the fact that κ values

Table III. Specimen results for the beta-uniform model. Lower limits cutting off a 0.025 tail area are shown, corresponding to both 'exact' and 'mid-*P*' accumulations of tail areas.

m	*n*	*U*	*U*/*mn*	Exact	Mid-*P*
2	3	0	0.0	0.000000	0.000000
2	3	1	0.1667	0.006202	0.008895
2	3	2	0.3333	0.015320	0.024220
2	3	3	0.5	0.051290	0.074019
2	3	4	0.6667	0.121315	0.151274
2	3	5	0.8333	0.216181	0.248602
2	3	6	1.0	0.303475	0.391775
3	3	0	0.0	0.000000	0.000000
3	3	1	0.1111	0.004606	0.006310
3	3	2	0.2222	0.009866	0.013861
3	3	3	0.3333	0.022753	0.037235
3	3	4	0.4444	0.072245	0.086748
3	3	5	0.5556	0.112141	0.129628
3	3	6	0.6667	0.157500	0.189725
3	3	7	0.7778	0.245262	0.275498
3	3	8	0.8889	0.328589	0.361852
3	3	9	1.0	0.412762	**0.500000**
6	6	36	1.0	0.722647	0.775969
10	10	100	1.0	0.854912	0.884312
19	19	361	1.0	0.937114	**0.950000**
39	39	1521	1.0	**0.975000**	0.980076
60	60	3600	1.0	0.985433	0.988368
100	100	10 000	1.0	0.992230	0.993782
100	10	1000	1.0	0.941073	0.954760
10	100	1000	1.0	0.941073	0.954760
1000	1000	1 000 000	1.0	0.999584	0.999683

(a) tend to be highest when κ is used in weighted form for an ordinal scale, intermediate in the binary case, and lowest in the unordered categorical case, and (b) tend to be highest for within-observer variation, intermediate for between-observers and lowest for between-methods variation. Analogously, concern has been expressed that it may in time become common practice for the range of possible values of *U*/*mn* to be divided into zones delimited by round but essentially arbitrary values, with adjectival labels such as the above, following some publication advocating this practice. Similarly, it might become customary for the same essentially arbitrary value of θ to appear in most power calculations, with little consideration of the appropriateness in the particular context.

The response to the above points is, firstly, that a measure of this form is definitely needed, especially when the original scale of measurement is not readily interpreted. Alternative measures to achieve the same objective could be developed, for example, based on Normal scores, but would be more complex and thus less likely to gain wide acceptance. Furthermore, the danger that a proposed statistical method will be abused by being applied naively is an argument not for avoiding developing such a tool but rather for promoting understanding of its appropriate use, alongside a greater understanding of statistical issues in general throughout the user community.

Another possible criticism is that U/mn might dampen the influence of extreme values. In fact, the converse is true. Suppose, in Normal order statistics notation, $y_{(n)} > x_{(m)}$. Then as $y_{(n)} \to \infty$, holding $y_{(1)}, y_{(2)}, \ldots, y_{(n-1)}$ and $\mathbf{x}$ constant, $\hat{\theta}$ remains unaltered whilst $\hat{\delta}$ asymptotes down to a finite value, $\sqrt{((m+n-2)/(n(n-1)))}$, and the pooled variance t (albeit inappropriate) asymptotes down to $\sqrt{(m(1-2/(m+n))/(n-1))}$, a low value, only significant at the conventional $\alpha = 0.05$ level when m is around $2n$ or higher.

Of course, there are ways in which two distributions or two samples can differ, that are far from a shift alternative. For instance, X and Y might have very different degrees of dispersion, so that the support for one variable is entirely within the support for the other. The measure developed here is less useful in such a situation.

An Excel spreadsheet is available at the author's website http://www.cardiff.ac.uk/medicine/epidemiology_statistics/research/statistics/newcombe.htm which facilitates visualization of any $\theta \in (0, 1)$ in terms of pairs of Gaussian distributions with peaks separated by the corresponding δ. It displays the value of δ corresponding to θ and Gaussian curves for N(0,1) and N(δ, 1) as in Figure 1 which corresponds to $\theta = 0.7$. No units are shown as they are redundant, any pair $N(\mu, \sigma^2)$ and $N(\mu + \delta\sigma, \sigma^2)$ would be equivalent.

ACKNOWLEDGEMENTS

I wish to thank Ying Lu, Jenő Reiczigel and Ruth Pickering for helpful discussion, and two referees for painstaking comments.

REFERENCES

1. Conover WJ. *Practical Non-parametric Statistics* (2nd edn). Wiley: New York, 1980.
2. Newcombe RG. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. 2. Asymptotic methods and evaluation. *Statistics in Medicine* 2005, this issue.
3. Newcombe RG. A critical review of risk prediction, with special reference to the perinatal period. *Ph.D. Thesis*, University of Wales, 1979.
4. Bross IDJ. How to use ridit analysis. *Biometrics* 1958; **14**:18–38.
5. McGraw KO, Wong SP. A common language effect size statistic. *Psychological Bulletin* 1992; **111**:361–365.
6. Vargha A, Delaney HD. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 2000; **25**:101–132.
7. Fay MP, Gennings C. Non-parametric two-sample tests for repeated ordinal responses. *Statistics in Medicine* 1996; **15**:429–442.
8. Edwardes MD deB. A confidence interval for $\Pr(X < Y) - \Pr(X > Y)$ estimated from simple cluster samples. *Biometrics* 1995; **51**:571–578.
9. Edwardes MD deB. Distribution-free tests for cluster samples of ordinal responses. *Journal of Statistical Planning and Inference* 2002; **105**:393–404.
10. Newson R. Parameters behind 'nonparametric' statistics: Kendall's tau, Somers'D and median differences. *Stata Journal* 2002; **2**:45–64.
11. Ukoli FA, Adams-Campbell LL, Ononu J, Nwankwo MU, Chanetsa F. Nutritional status of urban Nigerian school children relative to the NCHS reference population. *East African Medical Journal* 1993; **70**:409–413.
12. Agresti A. Generalized odds ratios for ordinal data. *Biometrics* 1980; **36**:59–67.
13. Wellek S, Hampel B. A distribution-free two-sample equivalence test allowing for tied observations. *Biometrical Journal* 1999; **41**:171–186.
14. Fujii Y. Inference based on $P(X < Y)/P(X > Y)$ in two sample problems. *Bulletin of Informatics and Cybernetics* 2004; **36**:137–145.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
16. Halperin M, Gilbert PR, Lachin JM. Distribution-free confidence intervals for $\Pr(X_1 < X_2)$. *Biometrics* 1987; **43**:71–80.
17. Mee RW. Confidence intervals for probabilities and tolerance regions based on a generalisation of the Mann–Whitney statistic. *Journal of the American Statistical Association* 1990; **85**:793–800.

18. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
19. Obuchowski NA, Lieber ML. Confidence bounds when the estimated ROC area is 1.0. *Academic Radiology* 2002; **9**:526–530.
20. Simonoff JS, Hochberg Y, Reiser B. Response to Brownie. *Biometrics* 1988; **44**:621.
21. Kyle PM, Campbell S, Buckley D, Kissane J, de Swiet M, Albano J, Millar JG, Redman CWG. A comparison of the inactive urinary kallikrein:creatinine ratio and the angiotensin sensitivity test for the prediction of pre-eclampsia. *British Journal of Obstetrics and Gynaecology* 1996; **103**:981–987.
22. Fearnley D, Williams T. A prospective study using the Monroe dyscontrol scale as a measure of impulsivity in referrals to a forensic psychiatry service. *Medicine*, *Science and the Law* 2001; **41**:58–62.
23. Monroe RR. *Episodic Behavioral Disorder*: *A Psychodynamic and Neurological Analysis*. Harvard University Press: Cambridge, 1970.
24. Monroe RR. *Brain Dysfunction in Aggressive Criminals*. Lexington Books: Lexington, 1978.
25. Simonoff JS, Hochberg Y, Reiser B. Alternative estimation procedures for $\Pr(X<Y)$ in categorized data. *Biometrics* 1986; **42**:895–907.
26. Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 1987; **82**:645–647.
27. Altman DG. Statistics and ethics in medical research. III. How large a sample? *British Medical Journal* 1980; **281**:1336–1338.
28. http://en.wikipedia.org/wiki/Condorcet_paradox. Accessed 19 May 2005.
29. Reiczigel J, Rozsa L, Zakarias I. Bootstrap Wilcoxon–Mann–Whitney test for use in relation with non-shift alternatives. *Controlled Clinical Trials* 2003; **24**(Suppl. 3):291.
30. Newcombe RG. A life table for onset of Huntington's Chorea. *Annals of Human Genetics* 1981; **45**:375–385.
31. Pearson K. On the measurement of the influence of 'broad categories' on correlation. *Biometrika* 1913; **9**: 116–139.
32. Agresti A. The effect of category choice on some ordinal measures of association. *Journal of the American Statistical Association* 1976; **71**:49–55.
33. Kendall MG, Stuart A. *The Advanced Theory of Statistics. Volume 1. Distribution Theory* (3rd edn). Griffin: London, 1969; 325.
34. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**:108–116.
35. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404–413.
36. Lancaster HO. The combination of probabilities arising from data in discrete distributions. *Biometrika* 1949; **36**:370–382.
37. Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1956; **43**:423–435.
38. Blaker H. Confidence curves and exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**:783–798.
39. Cohen J. A coefficient of agreement for binomial scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
40. Landis JR, Koch GG. The measurement of agreement for categorical data. *Biometrics* 1977; **33**:159–174.