International Journal of Computational and Theoretical Statistics
ISSN (2384-4795)
Int. J. Comp. Theo. Stat. 1, No. 1 (Nov-2014)

A Solution of the Two-Sample Problem

AD Tsakok

AD Tsakok Mathematical Centre, London, UK

Received March 23, 2014, Revised May 24, 2014, Accepted May 30, 2014 Published 1 Nov. 2014

Abstract: The Tsakok test of fit is used to obtain an exact interval estimate of distributions and achieves exact two-sample tests, UMPU in a class of two-sided alternatives and not conditional on marginal totals. Comparisons and illustrations are made, an investigation into independent events is undertaken and an algorithm for computing the significance level of the Tsakok test of fit is given.

Keywords: UMPU, two-sample problem

1. INTRODUCTION

The chi-squared test by Pearson [1] has been extensively used to estimate distributions and in contingency tables, but as Kendall and Stuart [2, p. 453] pointed out, it cannot be expected to be unbiased in general; other than being inexact and uses distributions conditional on marginal totals. These are overcome by the Tsakok [3] test of fit, since it is exact (with test sizes easily calculated directly from the multinomial distribution) and Uniformly Most Powerful Unbiased (UMPU) in a class of two-sided alternatives.

The Tsakok test of fit enables an exact UMPU two-sample test, using distributions which are not conditional on marginal totals, and so compares favorably to other alternatives. It is a frequent practical problem, as the example illustrates, and solves the two-sample problem; considered by Euler (Kendall and Stuart [2], p 512).

2. ESTIMATING DISTRIBUTIONS

The Tsakok [3] test of fit has been shown to exist as specified: Let independent observations be made on a random vector X in $\mathbf{R}^N$, with unknown, possibly continuous, distribution F. The range along which X varies is partitioned into k mutually exclusive pre-specified classes i = 1, . . .k. For each class i, let n_i be the number of observations falling into i, and p_i be the probability of an observation occurring in class i. So $\sum_{i=1}^{k} p_i = 1$. Let $\sum_{i=1}^{k} n_i = n$. If $F = F_0$ for some distribution F_0, then $p_i = p_{0i}$ $\forall$ i. $n_i = x_i$ from the data. So x_i is a particular value of n_i.

To test H_0: $p_i = p_{0i}$ (i = 1, ...k) against H_1 : $p_i \neq p_{0i}$ for at least one i; the test f is, in the non-randomized case:

$f = 1$ when $n_i \leq c_{i1}$ or $n_i \geq c_{i2}$ given n_j $\forall$ j ≠ i, k for at least one i = 1, ...k — 1;

$f = 0$ otherwise;

where the integer interval acceptance region $[(c_{i1},c_{i2})|n_j$ $\forall$ j ≠ i, k] for which $0 \leq c_{i1} \leq n$, $0 \leq c_{i2} \leq n$ is subject to

$$E(f|H_0) = \alpha \quad 0 < \alpha < 1 \tag{1}$$

$$E(n f|H_0) = \alpha E(n_i|H_0) \tag{2}$$

for i = 1, ...k — 1, at the test size α. This test is UMPU against alternatives under H_1 for which p_i differs from p_{0i} for only one i (i = 1, ...k — 1).

Lemma 2.1. Properties of unbiasedness condition (2) given n and n_j $\forall$ j ≠ i,k.

(i) The Tsakok test f satisfies the unbiasedness condition when $c_{i1} = np_{0i} — d_{i1}$, $c_{i2} = np_{0i} + d_{i2}$, for some integers d_{i1}, $d_{i2} > 0$ for which $n \geq c_{i1} \geq 0$ and $0 \leq c_{i2} \leq n$.

(ii) The unbiasedness condition for f is satisfied iff there exists real functions $I_1(d_{i1})$, $I_2(d_{i2})$ such that $E((n_i — np_{0i}) f|H_0) = I_2(d_{i2}) — I_1(d_{i1}) = 0$. $I_1(d_{i1}) \geq 0$, $I_2(d_{i2}) \geq 0$ and vary monotonically with d_{i1}, d_{i2} respectively, with d_{i1}, d_{i2} sufficiently large to prevent them from satisfying the conditions:

$d_i = d_{i1}$ or d_{i2} for which $n_i = np_i + d_i \in \mathbf{Z}$, the set of integers (i = 1, ...k), $\sum_{i=1}^{k} d_i = 0$, $|d_r — d_s| < 1$, and $p_s(np_r + d_r) \neq p_r(np_s + d_s + 1)$ for any distinct pair of classes r and s.

E-mail:adtsakok@hotmail.com

(iii) There exists α such that integers $d_{ij} = m_{ij}$ and $m_{i1} < np_{0i}$, $m_{ij}/n < 1$ (j = 1, 2) for which $c_{i1} = np_{0i} - m_{i1}$, $c_{i2} = np_{0i} + m_{i2}$ for any choice of p_{0i} (i = 1, ...k — 1) with $\sum_{i=1}^{k} p_{0i} = 1$.

Proof. (i) and (ii). Putting $c_{i1} = np_{0i} - d_{i1}$, $c_{i2} = np_{0i} + d_{i2}$ for some integers $d_{i1}, d_{i2} > 0$, given $n_j \ \forall \ j \neq i, k$; the d_{i1}, d_{i2} are chosen to satisfy the unbiasedness condition as follows:

$E(n_i\ f\ |H_0) = E((n_i - np_{0i})\ f\ |H_0) + np_{0i}E(f\ |H_0) = np_{0i}E(f\ |H_0) = \alpha E(n_i\ |H_0)$

iff $E((n_i - np_{0i})\ f\ |H_0) = 0$, given that $np_{0i}E(f\ |\ H_0)$ is bounded, with n fixed. With randomization constants γ_1, γ_2, this can be expressed as:

$E((n_i - np_{0i})\ f\ |H_0) = I_2(d_{i2}) - I_1(d_{i1}) = 0$

where, if $P_i = P(n_i \leq x\ |H_0, n_j \ \forall \ j \neq i, k)$ for some real $x \geq 0$, and $0 \leq \gamma_1 \leq 1$, $0 \leq \gamma_2 \leq 1$;

$$I_2 = \int n_i - np_{0i} dP_i + \gamma_1 d_{i2} P(n_i - np_{0i} = d_{i2}\ |\ H_0, n_j \ \forall\ j \neq i, k) \quad (3)$$

$(n_i : np_{0i} + d_{i2} < n_i \leq n - \sum_{(j:\ \text{for all}\ j \neq i,k)} n_j)$

$$I_1 = \int np_{0i} - n_i dP_i + \gamma_2 d_{i1} P(np_{0i} - n_i = d_{i1}\ |H_0, n_j \ \forall\ j \neq i, k) \quad (4)$$

$(n_i : np_{0i} - d_{i1} > n_i \geq 0)$

By inspection, I_1 and I_2 are both positive and vary monotonically with d_{i1}, d_{i2} respectively, where $d_{i1}, d_{i2} > 0$, when d_{i1}, d_{i2} are sufficiently large so that $(np_{0i} - d_{i1}, np_{0i} + d_{i2})$ contains the maximal value of Prob.$(n_1, ...n_k\ |H_0)$, given i. According to Tsakok [3], these conditions are: integer values of $d_i = d_{i1}$ or d_{i2} for which $n_i = np_i + d_i \in \mathbf{Z}$ (i = 1, ...k), where $\sum_{i=1}^{k} d_i = 0$, $|d_r - d_s| < 1$, and $p_s(np_r + d_r) \neq p_r(np_s + d_s + 1)$ for any distinct pair of classes r and s.

(iii) Since $1 > \alpha > 0$, $c_{i1} = np_{0i} - d_{i1}$ and $c_{i2} = np_{0i} + d_{i2}$ given $n_j \ \forall\ j \neq i, k$, this means $0 \leq c_{i1} \leq np_{0i}$ and $np_{0i} \leq c_{i2} \leq n$, as $c_{i1} \geq 0$ and $c_{i2} \leq n$; being possible values of n_i to satisfy (1). Setting $d_{i1} = m_{i1}$ and $d_{i2} = m_{i2}$ when $c_{i1} = np_{0i} - d_{i1}$ and $c_{i2} = np_{0i} + d_{i2}$, for some integers m_{i1} and m_{i2}, α is then chosen to be sufficiently large so that $m_{i1} < np_{0i}$, $m_{i1}/n < 1$ and $m_{i2}/n < 1$ (i = 1, ...k — 1), for any choice of p_{0i} (i = 1, ...k)

↑

Definition 2.2. A value of p_i is said to be acceptable by a test *f* at the test size α if it can be accepted to be a value of p_i if tested using the test *f* at the test size α.

Definition 2.3. A test size $\alpha \in (0,1)$ which satisfies the conditions of Lemma 2.1 is said to be sufficiently large.

Theorem 2.4. For sufficiently large test size α given n:

(i) There exists an interval estimate (p_{li}, p_{ui}) of p_i (i = 1, ...k — 1) formed from acceptable values of p_i using the test *f*.

(ii) There exists a sequence of values p_{ui1r} and p_{ui2r} such that $p_{ui1r} \geq p_{ui} \geq p_{ui2r}$ with $p_{ui1r} \rightarrow p_{ui}$ and $p_{ui2r} \rightarrow p_{ui}$ as $r \rightarrow \infty$ (i = 1,...k — 1). Similarly, there exists a sequence of values p_{li1r} and p_{li2r} such that $p_{li1r} \geq p_{li} \geq p_{li2r}$ with $p_{li1r} \rightarrow p_{li}$ and $p_{li2r} \rightarrow p_{li}$ as $r \rightarrow \infty$ (i = 1, ...k — 1).

Proof. (i) In the test *f*, the unbiasedness condition is met by choosing d_{i1}/d_{i2} as follows: If d_{i1} is such that $I_1(d_{i1})$ is too large when compared to $I_2(d_{i2})$, $E((n_i - np_{0i})\ f\ |H_0) < 0$, while if d_{i2} is such that $I_2(d_{i2})$ is too large in relation to $I_1(d_{i1})$, $E((n_i - np_{0i})\ f\ |H_0) > 0$. Thus as $E((n_i - np_{0i})\ f\ |H_0)$ varies with d_{i1}, d_{i2} ($d_{i1}, d_{i2} > 0$); there is a value of d_{i1}/d_{i2} for which $E(n_i - np_{0i})\ f\ |\ H_0) = 0$, possibly using randomization constants γ_1, γ_2.The actual values of d_{i1}, d_{i2} can then be chosen to satisfy the required test size. α must moreover be sufficiently large with $P(n_i > np_{0i} + d_{i1}\ |H_0, n_j \ \forall\ j \neq i,k) > 0$ and $P(n_i < np_{0i} - d_{i1}\ |H_0, n_j \ \forall\ j \neq i, k) > 0$, since otherwise it would not be a two-sided test.

d_{i1} and d_{i2} (i = 1, ...k — 1) thus chosen meets the unbiasedness requirement for any $p_{0i} \in (0, 1)$, and are bounded above by some integers m_{i1}, m_{i2} respectively such that $m_{ij}/n < 1$ (j = 1, 2), for sufficiently large α.

Suppose changing p_{0i} to $p_{0i} + \Delta p_{0i}$, for some real $\Delta p_{0i} > 0$, causes $E((n_i - np_{0i})\ f\ |H_0) \neq 0$. Unbiasedness can then be restored by varying d_{i1}/d_{i2}, at the given constant test size. If $p_{0i} = x_i/n$ initially, successive increments of Δp_{0i} may then require corresponding adjustments in d_{i1}/d_{i2} at the required test size. This process can continue until x_i/n does not belong to the interval $(x_i/n + \delta - m_{i1}/n, x_i/n + \delta + m_{i2}/n)$; where $p_{0i} = x_i/n + \delta$, for some real δ obtained from increases of Δp_{0i} such that $x_i/n + \delta > m_{i1}/n$. Similar considerations apply if variations $p_{0i} - \Delta p_{0i}$ are made, where $\Delta p_{0i} > 0$, starting from when $p_{0i} = x_i/n$. This shows that, for a given sample of observations, values of p_{0i} which may be accepted for p_i vary within an interval for each i, using *f* given sufficiently large α and m_{i1}, m_{i2} (i = 1, ...k — 1).

(ii) Let (p_{li}, p_{ui}) be an interval along which acceptable values of p_i are allowed to vary at a sufficiently large α, with *f*. An iterative procedure is now used to determine (p_{li}, p_{ui}). Using an acceptance region of p_i (using *f*) as an initial estimate of (p_{li}, p_{ui}), an upper and lower bound is obtained for p_{li} and p_{ui} by trial and improvement.

Let the upper and lower bound of p_{ui} be p_{uiUr} and p_{uiLr} respectively at the r^{th} iteration, with *f*. Denote the corresponding acceptance regions for p_{uiUr} and p_{uiLr} by $(c_{i1}(U, r), c_{i2}(U, r))$ and $(c_{i1}(L, r), c_{i2}(L, r))$ respectively.

With (c_{i1},c_{i2}) as above with p_{0i} varying with the r^{th} iteration, an upper bound of p_{ui} is a hypothetical value of p_{ui} which is rejected, with $n_i < c_{i1}(U, r)$ for rejection of p_{uiUr}. A lower bound of p_{ui} is a hypothetical value of p_{ui} which is accepted, with $(c_{i1}(L, r) < n_i < c_{i2}(L, r))$ for acceptance of p_{uiLr}. The inequality $p_{uiUr} \geq p_{ui} \geq p_{uiLr}$ is now constructed as follows:

An initial test is made using f to obtain an acceptance region (c_{i1}, c_{i2}) of p_i (i = 1,...k — 1).

Then test $H_0 : p_i = p_{0i}$ (i = 1, ...k) against $H_1 : p_i \neq p_{0i}$ for at least one i; where p_{0i} is chosen such that:

When r = 1, $p_{0i} \approx c_{i2}(L, 1)/n = c_{i2}/n$ (i = 1, ...k — 1), where the approximation is made subject to $\sum p_{0i} = 1$. If p_{0i} (i = 1, ...k) is accepted, then $p_{uiL1} = p_{0i}$ (i = 1, ...k). A new set of hypothetical values of p_{0i} (i = 1,...k —1) is chosen such that $p_{0i} > p_{uiL1}$ (i = 1, ... k— 1) which are rejected when tested. For these rejected p_{0i} (i = 1,...k)), set $p_{uiU1} = p_{0i}$ (i = 1,...k).

Otherwise (r = 1), if p_{0i} (i = 1,...k) is rejected, then $p_{uiU1} = p_{0i}$ (i = 1,...k). A new set of hypothetical values of p_{0i} (i = 1,...k — 1) is chosen such that $p_{0i} < p_{uiU1}$ (i = 1,...k — 1) which are accepted when tested. For these accepted p_{0i} (i = 1, ...k), set $p_{uiL1} = p_{0i}$ (i = 1, ...k).

With $r \geq 1$, p_{0i} (i = 1, ...k — 1) is chosen at the $r + 1^{th}$ iteration such that $p_{uiUr} > p_{0i} > p_{uiLr}$ if $p_{uiUr} > p_{uiLr}$; and otherwise $p_{0i} = p_{ui} = p_{uiUr}$ if $p_{uiUr} = p_{uiLr}$, with the iteration stopped.

If p_{0i} (i = 1,...k — 1) is accepted, then $p_{uiL(r+1)} = p_{0i}$ (i = 1, ...k — 1) and $p_{uiU(r+1)} = p_{uiUr}$ (i = 1, ...k — 1).

Otherwise, $p_{uiL(r+1)} = p_{uiLr}$ (i = 1,...k — 1) and $p_{uiU(r+1)} = p_{0i}$ (i =1,...k — 1) .Therefore $p_{uiU(r+1)} — p_{uiL(r+1)} > 0$ (i = 1, ...k —1) and is decreased at the $r + 1^{th}$ iteration. As $r \to \infty$, $p_{uiUr} — p_{uiLr} \to 0$ (i = 1, ...k — 1). But $p_{uiUr} — p_{uiLr} = (p_{uiUr} — p_{ui}) + (p_{ui} — p_{uiLr})$. So $p_{uiUr} \to p_{ui}$ and $p_{uiLr} \to p_{ui}$ as $r \to \infty$ (i = 1, ...k — 1) since $p_{uiUr} \geq p_{ui} \geq p_{uiLr}$. A similar result applies to p_{li} (i = 1, ...k — 1).

So p_{ui} and p_{li} (i = 1, ...k — 1) can be estimated to any accuracy, at a given sufficiently large test size α. This is possible because the above intervals (c_{i1}, c_{i2}) can be computed from the data and the null distribution, as shown by Lemma 2.1 and Algorithm 5.6 of the Appendix.

Example: Consider some of the data of Topp et al. [4]. This is summarized in Table 1. For the above case, i = 1 and p_1 is the probability of an occurrence of diplegia in the sample. Thus (p_{l1}, p_{u1}) is the interval estimate of p_1 with a probability of .976 for the occurrence of diplegia for Preterm 1983-86; so that the Tsakok test of fit was used at a significance level of 1-.976 (i.e., .024). The small variations in the confidence levels arise out of the non-randomized tests applied to discrete data. Normality is not assumed.

TABLE I. PRETERM AND TERM DIPLEGIA OCCURRENCE

	Preterm 1983-80	Term 1987-90
Sample Size	170	171
No (%) Diplegia	113(66)	69(40)
(p_{l1}, p_{u1})	(.58,.70)	(.36,.49)
Confidence level	.976	.979

3. COMPARING THE DISTRIBUTIONS

As indicated by Tsakok [5], the interval estimates of distributions immediately enable comparisons between them to be made. This is also discussed in Tsakok [6]. The technique is similar to that used by Tsakok [7] to solve the Behrens-Fisher problem.

Thus if F_1 and F_2 are two distributions from independent populations 1 and 2, the problem is to decide with a test f_{12} between the hypotheses:

H: $F_1 = F_2 = F$, say, and K: $F_1 \neq F_2$;

from independent random samples of size n_j for each population j (j = 1, 2). With the above method, let C_j be the set of distributions that are acceptable hypothetical distributions for population j using a Tsakok test of fit f_j at significance level α_j. Thus $P(F \in C_j|F_j = F)=1— \alpha_j = 1— E(f_j|F_j = F)$. Since the samples are independent, $P((F \in C_1) \bigcap (F \in C_2)|F_1 = F_2 = F) = (1 — \alpha_1)(1 — \alpha_2)$. With $f_{12}= 1— (1 — f_1)(1 — f_2)$, the test size α of H is $E(f_{12}|H)=E(1 — (1 — f_1)(1 — f_2)) = \alpha = 1 — (1 — \alpha_1)(1 — \alpha_2)$, being the probability that at least one of the events $(F \in C_j|F_j = F)$ does not occur.

Let $Q = C_1 \bigcap C_2$. If Q is empty, at least one of the events $(F \in C_j|F = F_j)$ will not occur, leading to:

Reject H if Q is empty.

When Q is not empty, H may be either accepted or rejected, depending on whether or not the F chosen by the test of fit belongs to Q, showing that the concept of acceptance regions is inadequate. To overcome this inadequacy, the following complementary concept is proposed:

Definition 3.1. With hypothetical distribution F_0 fully specified, $P(Q) = P(F_0 \in Q|F = F_0)$.

Theorem 3.2. Using the test $f_{12} = 1 — (1 — f_1)(1 — f_2)$ at the test size α, $0 \leq P(Q) \leq 1 — \alpha$.

Proof. For f_j, C_j is of the form $(p_i, (i = 1, ...k — 1) : p_i \in (p_{lij}, p_{uij}), i = 1, ...k — 1)$, where the classes i are mutually exclusive, and (p_{lij},p_{uij}) is a real interval for which $p_{lij} < pi < p_{uij}$ with probability $1 — \alpha_j$. Let $(a_i, b_i) = \bigcap_j (p_{lij}, p_{uij})$, i = 1, ...k — 1. Then $Q = (p_i, (i = 1, ...k — 1): p_i \in (a_i, b_i), i=1, ...k — 1)$. For a member F_0 of C_j with (c_{ij1}, c_{ij2}) as the acceptance region for population j (i =

1,,...k — 1) using f_j at the test size α_j, let $A_j = ((c_{ij1}/n_j, c_{ij2}/n_j)$, i = 1, ...k — 1) for population j using f_j for the hypothesis that $F_j = F_0$. So event ($F_0 \in A_j|F_j = F_0$) occurs whenever event ($f_j = 0|F_j = F_0$) occurs at the test size α_j, with F_0 partitioned into mutually exclusive class probabilities p_{0i} (i = 1, ...k) as above, since then $p_{0i} \in (c_{ij1}/n_j, c_{ij2}/n_j)$ (i = 1, ...k — 1), with the hypothesis ($F_j = F_0$) accepted. So $P(F_0 \in Q|F = F_0) = P(F_0 \in \bigcap_j A_j|F_1 = F_2 = F_0) = P(F_0 \in \bigcap_j A_j |F_1 = F_0)P(F_0 \in \bigcap_j A_j |F_2 = F_0)$. Hence $0 \leq P(Q) \leq 1 - \alpha$. In particular, $P(Q) = 0$ if $Q = 0$.

The above expression for P(Q) uses the fact that the events $E_s = (F_0 \in \bigcap_j A_j |F_s = F_0)$ are independent, for s = 1, 2. This is true because E_s is contained in G_s only, where $G_s = (F_0 \in A_s|F_s = F_0)$, and G_1, G_2 are independent by hypothesis. For if E_1 and E_2 were not independent, then when $E_s = G_s$ (s = 1, 2), this contradicts the hypothesis of independence of G_1, G_2.

The Appendix proves a more general result.

Values of P(Q) close to its maximum or indicating considerable (at least 95%) overlap with at least one of the confidence intervals support the view that H should be accepted but not otherwise, as it suggests that a Type II error would then be made. Thus corrective action can be taken if necessary. P(Q) is not needed if Q = 0 (i.e. Q is empty). The asymptotic properties of P(Q) have been considered by Tsakok [6].

Since the method is to measure the amount of overlap between confidence intervals for the comparisons, there is no p value, but P(Q) instead. The method described is now applied to the above data. Hence

F_1 = distribution of the incidence of diplegia for Preterm 1983-86

F_2 = distribution of the incidence of diplegia for Term 1987-90

Q = 0, from Table 1.

Thus H is rejected at .04 (1 significant figure) significance level.

The two-sample test using the Tsakok technique has shown that, due to its exact UMPU properties, it is able to detect differences where Topp et al. [4] have failed using chi-squared tests at the .05 significance level; as the latter does not share these characteristics.

This is just one of several significant differences which Topp et al. [4] failed to discover because they relied on the asymptotic approximations of chi-squared tests and marginal totals.

The choice of class probabilities is not unique, but since for given k, the number of classes, each set of possible mutually exclusive pre-specified class probabilities is different from the other, it is not possible to compare the outcome of the above tests with each other; resulting in no possible contradictions between these above tests. In fact, the different sets of class probabilities complement each other in the data analysis.

4. COMPARISONS WITH OTHER METHODS

The interval estimates of distributions, up to their class probabilities, are exact and inherit the optimal properties of the Tsakok test of fit with which they were obtained. By comparison, the Kolmogorov single-sample statistic has been shown to be biased (Massey [8]).

The results of Massey [8] also show that the Kolmogorov-Smirnov two-sample test is biased, since it cannot be unbiased for large sample sizes of one population.

Theorem 4.1. If the Tsakok test of fit is used for each f_j (j = 1,2), the test f_{12} for testing H: $F_1 = F_2$ against K : $F_1 \neq F_2$ is UMPU against some two-sided alternatives for which $p_{ji} \neq p_{0i}$ for one j > 0 and one i only under K; and unbiased against all alternatives under K.

Proof. If $F_1 \neq F_2$ when comparing samples, then $F_j \neq F_0$ for at least one j for any F_0 (j = 1,2). Thus if p_{ji} is the probability of an observation falling in class i given F_j (j = 0, 1, 2), it is possible that $p_{ji} \neq p_{0i}$ for one $j \neq 0$ and one i only under K. Hence the optimal properties of the Tsakok test of fit f_j result in a test of H which is UMPU against some two-sided alternatives for which $p_{ji} \neq p_{0i}$ for one j > 0 and one i only under K; and unbiased against all alternatives under K.

This compares with the Wilcoxon test [9, 10] or other similar tests based on ordered samples, such as those by Gehan [11], Mantel [12] and Cox [13], which have been shown to be biased against the two-sided alternative (Lehmann [14]). The pre-test by Martinez and Narano [15] does not address this problem of biasedness.

In the chi-squared statistic involving all the k classes, one of the classes k is determined by the other k - 1 classes, and so is redundant in that sense. Apart from being approximate and biased (Kendall and Stuart [2, p. 453]), chi-squared tests on contingency tables are conditional upon the marginal totals, unlike the method used here. So the method proposed here gives results that are more generally applicable.

To achieve tests not based on conditional distributions, Suissa and Shuster [16] chose $p_{sup} = \sup_{\pi \in [0,1]} p(\pi)$, where π is the common proportion of 2 independent binomial distributions, whose proportions are being compared. This suffers from the difficulty that the π thus chosen may not be that of the distributions under study.

A similar criticism applies to the approach used by Storer and Kim [17], since the maximum likelihood estimate of π may not be that of the binomial distributions concerned when comparing their proportions, and is therefore irrelevant.

Berger and Boos [18], in comparing binomial proportions with distributions which are not conditional, maximised the p value by a choice of π over a confidence set. It thus suffers from the same criticism that the π chosen may not be that of the binomial populations of interest, and so is irrelevant.

Berger and Boos [18] also refer to their computational difficulties. The above numerical example illustrates no computational problems with the method used here.

When comparing samples, Bithell and Stone [19] proposed a Maximum Likelihood Ratio (MLR) statistic for comparing populations. The problem with their method is that the null distribution of their MLR statistic is unknown for small samples.

The Linear Risk Scores (LRS) statistic of Bithell et al. [20] for comparing populations is hard to interpret, apart from having an unknown distribution; as they noted. Consequently Bithell et al. [20] only made qualitative comparisons using LRS and MLR statistics.

By comparison, the method proposed here is exact, and gives clear results; using incidence rates of the populations concerned.

As the principle used for comparing distributions is similar to that used by Tsakok [7] for solving the Behrens-Fisher problem, other publications on it are relevant. In particular, Matuszewski and Sotres [21] proposed a test for a variation of the Behrens-Fisher problem: They tested $H : \mu_1 = \mu_2$ of two independent normal populations with means μ_1 and μ_2 against $K : \mu_1 < \mu_2$. Let (a_j, b_j) be the 80% confidence interval for μ_j (j= 1,2).

The test then rejects H at the nominal .05 significance level if and only if $b_1 < a_2$, assuming unknown variances.

This test differs from Tsakok [7] in two respects: Unlike Tsakok [7], it does not test H against the two sided alternative that the means are unequal, which is the standard formulation of the Behrens-Fisher problem found in most texts, such as Kendall and Stuart [2].

Secondly, it rejects H if and only if the confidence intervals do not overlap, with $b_1 < a_2$. This differs from the Tsakok [7] test for the Behrens- Fisher problem, which rejects H if (but not only if) the confidence intervals for the means do not overlap.

Therefore the characteristics of the two tests differ. Matuszewski and Sotres [21] do not claim that their test is exact, and only applies to the .05 nominal significance level. By comparison, the Tsakok [7] test is exact and optimal, at any significance level in (0, 1) in the frequentist sense.

The two-sample problem amounts to a problem of deciding whether two groups of data, as specified above, should be classified into the same or different groups. For this problem, Kline [22] made restricted comparisons with numerical examples between the nonparametric Bayesian Data Reduction Algorithm (BDRA), the Linear Discriminant Analysis and the Quadratic Analysis. He admits that BDRA lacks explanatory power and its theoretical operating characteristics are unclear. He does not claim that the BDRA has theoretical optimal properties, unlike the proposal presented here. Kayano and Dozono [23] use cluster analysis to the classification problem, under the normality assumption. They also make limited numerical investigations into the effectiveness of their approach, with no optimality claims. The normality assumption is here shown to be unnecessary for classification, and the theoretical effectiveness of the present proposal is established with its optimal properties.

These clear advantages of the method proposed here over its competitors make further comparisons unnecessary.

5. APPENDIX

Definition 5.1. A' will be said to be contained in event A only in the well-known sense that A' consists of sample points of A only, excluding any sample point that may simultaneously belong to another event B. Thus for the event $(A \bigcap B)$, A' is not $(A \bigcap B)$ if events A and B are unequal.

Theorem 5.2. Let A and B be independent events, and A' be contained in A only. Then A' cannot be dependent on B, if $P(B) > 0$ and $A - A'$ is independent of B.

Proof. If A and B are independent events, then $P(A \bigcap B) = P(A)P(B)$. Suppose A contains A' only, but A' is dependent on B. If $A - A'$ is empty, then $A = A'$, so that a contradiction is established. If it is not empty, $A - A'$ must be independent of B by hypothesis. So $P((A - A') \bigcap B) = P(A - A')P(B)$. Thus $P(A \bigcap B) = P((A' \bigcup (A - A')) \bigcap B) = P((A' \bigcap B) \bigcup ((A - A') \bigcap B)) = P(A' \bigcap B) + P((A - A') \bigcap B) = P(A'|B)P(B) + P(A - A')P(B)$. Since $P(A \bigcap B) = P(A)P(B)$, then $P(A \bigcap B) = P(A' \bigcup (A - A'))P(B) = (P(A') + P(A - A'))P(B) = P(A')P(B) + P(A - A')P(B)$. So $P(A') = P(A'|B)$, and A' is independent of B (Feller [24], p125); contrary to the supposition that A' is dependent on B. ↑

Using Feller's [24, p.128] definition of mutual independence, a similar argument readily establishes:

Theorem 5.3. Let $A_1, ...A_n$ be mutually independent events, for which $P(A_i) > 0$ (i = 1, ...n), and event A'_i be contained in A_i only. Then event A'_i cannot be dependent on $\bigcap_{(j:j \neq i, j=1,...n)} A_j$; when $A_i - A'_i$ is mutually independent with $(A_j)_{(j:j \neq i)}$.

If the Tsakok test of fit *f* is used, the following considerations should assist in its use.

Definition 5.4. A scan of class i is the process of alternately determining $n_i = np_{0i} - d_{i1}$ and $n_i = np_{0i} + d_{i2}$ given $n_j \; \forall \; j \neq i, k$, as integers d_{i1} and d_{i2} are each

increased from 0 to at most z in that order, for some integer z. z is chosen to satisfy the significance level of the non-randomised test *f* and the unbiasedness condition for class i as closely as possible. z must minimise $|E(n_i - np_{0i})\ f\ |H_0)|$.

Definition 5.5. A multi scan of class i > 1 is a scan of class i where after each ordered increase of d_{i1} or $d_{i2,}$ a scan of classes 1 to i —1 are made in that order, with each ordered change in d_{sj} during the scan of class s > 1 similarly resulting in a scan of classes 1 to s — 1 in that order for any s < i, j = 1, 2.

Algorithm 5.6. With classes i ordered from 1 to k —1, a scan of class i = 1 is made, followed by multi scans of classes 2 to k —1 in that order. Then for each value of n_2 formed during the multi scan of class i = 2, there will be a pair (n_1, n_2) with each of the values of n_1 formed during the scan of class i = 1. Similarly, each value of n_3 formed during the multi scan of class i = 3 will form a triple (n_1, n_2, n_3) with each of the pairs (n_1, n_2) formed during the multi scan of class i = 2. Thus by induction hypotheses the multi scan of class i = k —1 will create a set S of k — 1-tuples (n_1, n_2, $_{...}n_{k-1}$) with all the values of n_1, n_2, ...n_{k-1} formed during the multi scan. The interval of values for each n_i (i = 1, ...k - 1) thus formed are (c_{i1}, c_{i2}) since these satisfy the unbiasedness condition by construction, within the limits of non-randomised tests. Moreover, since *f* = 0 exactly when $c_{i1} < n_i < c_{i2}$ (i = 1,. ..k - 1) for the non-randomised case, $\alpha \geq 1 - \sum_S P(n_1,n_2...n_k)$, where the summation is over all the elements (n_1,n_2...n_k) of set S, and $P(n_1,n_2...n_k)$ is the multinomial distribution. Integer z is maximised to satisfy the above inequality for α for the non-randomised test *f* ; and is decreased otherwise if this inequality is not satisfied.

REFERENCES

[1] Pearson, K.,"On a criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen out of random sampling", Phil. Mag., vol. 50(5), p. 157, 1900.

[2] Kendall, M.G. and Stuart, A., The Advanced Theory of Statistics, vol. 2. London: Charles Griffin and Co. Ltd., 1973.

[3] Tsakok, A.D., "A test of fit satisfying some optimality criteria non- asymptotically", Metron, vol. 36, p 105, 1978.

[4] Topp, M., Uldall, P., Greisen, G., "Cerebral palsy births in Eastern Denmark 1987-90: implications for neonatal care", Paediatric and Perinatal Epidemiology, vol. 15, p. 271, 2001.

[5] Tsakok, A.D., "Comment on visual acuity" Ophthalmic Epidemiology, vol. 9, p. 347, 2002.

[6] Tsakok, A.D., Statistics and the Unified Field. London: AD Tsakok Mathematical Centre, 1993.

[7] Tsakok, A.D., "A solution to the generalized Behrens-Fisher problem", Metron, vol. 36, p. 79, 1978.

[8] Massey, F.J. Jr., "A note on the power of a non-parametric test", Ann. Math. Statist., vol. 21, p. 440, 1950; and vol. 23, p. 637, 1952.

[9] Wilcoxon, F., "Individual comparisons by ranking methods", Biometrics Bull., vol. 80, 1945.

[10] Wilcoxon, F., "Probability tables for individual comparisons by ranking methods", Biometrics, vol. 3, p. 119, 1947.

[11] Gehan, E.A., "A generalized Wilcoxon test for comparing arbitrarily single-censored samples", Biometrika, vol. 52, pp. 203-223, 1965.

[12] Mantel, N., "Evaluation of survival data and two new rank order statistics arising in its consideration", Cancer Chemotherapy Rep., vol. 50, pp. 167-170, 1967.

[13] Cox, D.R., "Regression models and life tables (with discussion)", Jour. Roy. Statist. Soc. B, vol. 34, pp. 187-220, 1972.

[14] Lehmann, E.L., Testing Statistical Hypotheses. New York: John Wiley and Sons Inc., 1959, p. 187.

[15] Martinez, R.L.M.C. and Narano, J., "A pretest for choosing between logrank and Wilcoxon tests in the two-sample problem", Metron, vol. 68, pp. 111- 125, 2010.

[16] Suissa, S. and Shuster, J., "Exact unconditional sample sizes for the 2x2 binomial trial", Jour. Roy. Statist. Soc. A, vol. 148, p. 317, 1985.

[17] Storer, B.E. and Kim, C., "Exact proportions of some exact test statistics for comparing two binomial proportions", Jour. Amer. Statist. Ass., vol. 85, p. 146, 1990.

[18] Berger, R.L. and Boos, D.D. "P values maximized over a confidence set for the nuisance parameter", Jour. Amer. Statist. Ass., vol. 89, p. 1012, 1994.

[19] Bithell J.F. and Stone R.A. "On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations", J. Epidemiol. Community Health, vol. 43, pp. 79-85, 1989.

[20] Bithell J.F., Dutton S.J., Draper G.J., Neary N.M., "Distribution of childhood leukemia and non-Hodgkin's lymphomas near nuclear installations in England and Wales", BMJ, vol. 309, p. 501, 1994.

[21] Matuszewski, A. and Sotres, D., "A simple test for the Behrens-Fisher problem", Computational Statistics and Data Analysis, vol. 3, p. 241, 1986.

[22] Kline, D.M., "Two-group classification using the Bayesian data reduction algorithm", Complexity, vol. 15, pp. 43-49, 2010.

[23] Kayano, M. and Dozono, K., "Functional cluster analysis via orthonormal Gaussian basis expansion and its applications", Jour. of Classification, vol. 27, pp. 211-230, 2010.

[24] Feller, W. An Introduction to Probability Theory and Its Applications, Vol.1. New York: John Wiley and Sons Inc., 1968.