Bringing Game Theory to Hypothesis Testing: Establishing Finite Sample Bounds on Inference[1]

Karl H. Schlag[2]

June 25, 2008

[1]The author would like to thank Joachim Röhmel for feedback on noninferiority tests.

[2]Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, E-08005 Barcelona, karl.schlag@upf.edu

Abstract

Small sample properties are of fundamental interest when only limited data is available. Exact inference is limited by constraints imposed by specific nonrandomized tests and of course also by lack of more data. These effects can be separated as we propose to evaluate a test by comparing its type II error to the minimal type II error among all tests for the given sample.

Game theory is used to establish this minimal type II error, the associated randomized test is characterized as part of a Nash equilibrium of a fictitious game against nature.

We use this method to investigate sequential tests for the difference between two means when outcomes are constrained to belong to a given bounded set. Tests of inequality and of noninferiority are included. We find that inference in terms of type II error based on a balanced sample cannot be improved by sequential sampling or even by observing counter factual evidence providing there is a reasonable gap between the hypotheses.

Keywords: exact; distribution-free; nonparametric; independent samples; matched pairs; Z test; unavoidable type II error; noninferiority.

JEL classification: C12, C14.

1 Introduction

Data available for inference is often very limited, since small samples are common in many disciplines. Inference can have important consequences so we consider exact hypothesis testing which involves rigorously analyzing tests, proving properties for instance in terms of level analytically and not approximately. Moreover we consider *distribution-free inference* (cf. Kendall and Sundrum, 1953) as we wish to derive implications that do not rely on unverifiable assumptions imposed by the investigator. What can the data tell us directly? As a related issue, how can we compare tests and evaluate the conclusions in a given finite sample?

We show here how game theory can help answer these questions. The key is to develop statistical hypothesis testing as a strategy in a game against nature where nature chooses the data generating process. It is a zero-sum game between the statistician and nature in which wrong recommendations are recorded as losses for the statistician and as gains for nature. A test that is part of a Nash equilibrium of this game generates the most powerful inference in terms of minimizing the type II error. By finding such an equilibrium one can establish tight bounds to inference in terms of type II error. Bounds on inference are of interest in their own right as they answer the question "What is the most a finite sample of data can tell us?". These bounds provide a natural benchmark for evaluating and comparing tests, thus answering our second question "How can we evaluate the performance of a test in light of many different alternatives?"

As pointed out by Savage (1954), game theory can be used to solve problems in statistics. The underlying idea is to solve worst case problems by invoking the minimax theorem for zero-sum games developed by von Neumann (1928). However game theory methods have not yet been used in hypothesis testing. Why not? For hypothesis testing where search is for a level α test that minimizes type II error this would mean to perform a worst case analysis among the level α tests. The problem with this is that the characterization of the set of level α tests itself is typically a very difficult problem. Here we proceed to formulate hypothesis testing as a zero-sum game

against nature without adding constraints on the level of test chosen in this game. The desired level condition is fulfilled in equilibrium provided penalties are assigned appropriately. In particular one need not be able to characterize all level α tests in order to apply this method. This allows for deriving most powerful recommendations for small sample problems.

Solving this game against nature produces a test that minimizes type II error. Such a test is typically randomized and hence is not very useful in practice. However the induced lower bounds can be used to evaluate the inference of other tests. New terminology needs to be introduced to reflect the new standards. The minimal type II error achievable for a given pair of hypotheses will be called the *unavoidable type II error*. The *added type II error* then measures the amount that the type II error of a given test is above the unavoidable type II error.

We apply the new methodology by considering tests for comparing two means or two distributions just given the interval or ratio data without making added distributional assumptions. The power of the game theory approach is that we are able to analyze the unavoidable type II error among the most general tests, namely among all tests that are based on sequential sampling.

An important condition for the environments we consider in this application is that outcomes belong to a known bounded set. Given this condition we need not make distributional assumptions and can focus on pure inference. Most environments satisfy this condition, a property that emerges whenever measuring outcomes on a bounded scale. Following arguments of by Bahadur and Savage (1956) we know that nontrivial inference is not possible if there is no restriction on the underlying distributions. Due to the possibility of fat tails, any test that has level α will have a type II error bounded below by $1-\alpha$. Alternative constraints on the data generating process added to ensure non trivial inference such as bounding moments are typically not verifiable. We wish to consider inference that is based on the properties of the data without adding additional assumptions. "Let the data speak!".

Our application retains quite general features as we include tests of inequality and of noninferiority, samples of matched pairs as well as sequential sampling. Our

understanding of hypothesis testing as a game against nature allows us to investigate inference among all tests based on sequential sampling. Analysis is greatly simplified as one need not understand the performance of all tests in all environments but only of all tests when facing the equilibrium strategy of nature. Beyond the novel game theory methodology another tool is introduced. A randomization trick is employed to extend results for environments with binary outcomes to those with known bounded outcome spaces.

We illustrate the new insights available due to the findings of this paper.

(a) Tests can be evaluated. Consider testing $H_0 : EY_1 \leq EY_2$ against $H_1 : EY_1 \geq EY_2 + d$ when Y_1 and Y_2 are Bernoulli distributed. Given a balanced sample of $n = 25$ independent observations of each variable the maximal added type II error of the test of Boschloo (1970) across all $d > 0$ is 0.046 when level is equal to 5%. So it is not possible to outperform this test in terms of type II error by more than 0.046 by any test based on sequential sampling provided at most 50 observations may be gathered.

(b) Tests can be compared. The corresponding value for the Z test (Suissa and Shuster, 1984, 1985) is 0.066.

(c) Tight minimal sample sizes can be derived. The unavoidable type II error of any sequential noninferiority test for testing $EY_1 \leq EY_2 - 0.3$ against $EY_1 \geq EY_2$ at level 2.5% is strictly above 20% if there are at most 84 independent observations while it is below 20% if the sample size is at least 86. This means that any 95% (equi-tailed) confidence interval of the difference between the two means based on at most 84 observations will wrongly cover the value 0.3 more than 20% of the time when the two means are equal. These statements on sample size hold if one assumes a balanced sample or if instead one allows for any sequential test. They hold if Y_1 and Y_2 are both Bernoulli distributed as well as when the only condition is that $Y_1, Y_2 \in [0, 1]$. The approximate formulae of Rodary et al. (1989) for the binary valued case underestimate this minimal sample size.

(d) Restrictions on inference can be evaluated. Consider testing the inequality of two distributions that contain outcomes in $[0, 1]$ when $n = 25$ and $\alpha = 0.05$. The loss when restricting selection to unbiased tests in terms of increased maximal unavoidable

type II error is found to be 0.0066.

(e) The value of having more information available is measurable. Instead of observing only one outcome assume that the outcomes of both variables are observed each time, creating a sample of matched pairs. All statements above remain true. More generally, consider inference when the difference in means is the parameter of interest and assume that there is a reasonable gap between the two hypotheses. Then this additional information that can also be interpreted as counter factual evidence does not to have any added value.

We add a few comments on the related literature. The only existing finite sample lower bounds on type II error for the setting of this paper are those implicitly given by the uniformly most powerful unbiased test for comparing two binomial proportions due to Tocher (1950). Connections between matched pairs and sequential sampling via properties of the least favorable distribution were first established by Schlag (2006) in the context of statistical decision making. The random transformation used to extend results for binary valued distributions to a nonparametric setting was independently developed in four special cases. It has been used by Cucconi (1968) for constructing a nonparametric randomized probability ratio test, by Gupta and Hande (1990) and by Schlag (2006) in the context of statistical decision making, by Schlag (2003) in the context of repeated decision making and by Schlag (2007a, 2007b) to design exact nonrandomized tests.

We proceed as follows. In Section 2 we characterize hypothesis testing as an equilibrium of a zero-sum game for a very general environment. In Section 3 we apply this methodolgy to tests for comparing two binomial proportions. After formulating the hypotheses in Section 3.1 we consider matched pairs in Section 3.2 and sequential sampling in Section 3.3. In Section 4 we extend the above results to distributions that have a support that is contained in a known bounded set. In Section 5 we conclude.

2 Hypothesis Testing as a Game against Nature

We start by showing how one can use game theory to derive tests that minimize type II error. We present the result in an abstract context. In the later sections we then show how to apply this to various testing and sampling scenarios involving bivariate distributions.

Let $\mathcal{Y}$ be the set of possible data generating processes. Assume that $\mathcal{Y}$ is a topological space that is compact. A typical element of $\mathcal{Y}$ will be denoted by Y. Consider two sets H_0 and H_1, identified with the null hypothesis and the alternative hypothesis respectively where $H_0, H_1 \subset \mathcal{Y}$ and H_0 and H_1 are nonempty, closed and disjoint. A statistician uses information coming from the underlying data generating process to make a recommendation whether or not to reject the null hypothesis in favor of the alternative hypothesis. To keep notation simple we do not explicitly specify the information available to the statistician. In particularly we allow for the statistician to influence the information available as in sequential testing. Instead we consider a reduced form approach and describe a test ϕ as a mapping from the set of data generating processes to a randomized choice of whether or not to reject the null hypothesis. Formally, $\phi : \mathcal{Y} \to [0, 1]$ where the value of ϕ specifies the probability of rejecting the null hypothesis. Let $\mathcal{F}$ be a set of tests available to the statistician. We assume that $\mathcal{F}$ contains the two simplest tests, namely always reject $\phi \equiv 1$ and never reject $\phi \equiv 0$.

Let $E_Y(\phi)$ denote the probability of rejection or *power* of ϕ when Y is the true data generating process and let $E_Y(1-\phi) = 1 - E_Y(\phi)$ for $Y \in \mathcal{Y}$. $\sup_{Y \in H_0} E_Y(\phi)$ then represents the *type I error* of ϕ. The test ϕ has *level* α for $\alpha \in (0, 1)$ if its type I error lies below α. The *type II error* of ϕ is given by $\sup_{Y \in H_1} E_Y(1-\phi)$. We call $\inf_{\phi \in \mathcal{F}} \sup_{Y \in H_1} E_Y(1-\phi)$ the *unavoidable type II error*. A so-called *least favorable distribution* corresponds to a data generating process under which the unavoidable type II error is attained. We call the difference between the type II error of a test ϕ and the unavoidable type II error the *added type II error of* ϕ. The type II error of ϕ is therefore the sum of the unavoidable and the added type II error.

Note that the unavoidable type II error refers to the set of all randomized tests as this error is meant to measure limits to inference. Practitioners naturally only choose among nonrandomized tests. We prefer to evaluate such additional constraints on inference by the added type II error of the respective tests.

Let $\Gamma(\alpha,\beta)$ be the following zero-sum game defined for given constants $\alpha,\beta \in (0,1)$, a game we visualize as being played between the statistician and nature. Simultaneously, the statistician chooses a test $\phi \in \mathcal{F}$ and nature chooses a data generating process $Y \in H_0 \cup H_1$. The outcome resulting from this simultaneous choice is a nonnegative penalty for the statistician. Penalites are defined as follows. The penalty of wrongly not rejecting the null hypothesis is α, the penalty of not rejecting the null hypothesis when the alternative hypothesis is true is $1-\beta$, and there is no penalty when making the correct recommendation.[1] Let $\pi(\phi, Y; \alpha, \beta)$ denote the expected penalty attained by ϕ when facing Y, so

$$\pi(\phi, Y; \alpha, \beta) = \begin{cases} \beta E_Y(\phi) & \text{if} \quad Y \in H_0 \\ \alpha E_Y(1-\phi) & \text{if} \quad Y \in H_1 \end{cases}.$$

In this game it is assumed that the statistician aims to minimize the expected penalty while nature aims to maximize the expected penalty (of the statistician). This makes Γ a zero-sum game. Both players are also allowed to randomize.[2] Thus the statistician chooses a possibly randomized test ϕ belonging to $\Delta\mathcal{F}$ and nature may choose an element of $\Delta(H_0 \cup H_1)$ which will be typically denoted by η.[3] When $\eta \notin \Delta H_0 \cup \Delta H_1$ then we let $\eta_0 \in \Delta H_0$, $\eta_1 \in \Delta H_1$ and $\lambda \in (0,1)$ be defined such that $\eta = (1-\lambda)\eta_0 + \lambda\eta_1$. Above definitions of E and π extend from Y to $\eta \in \Delta(H_0 \cup H_1)$ by taking expectations, for instance,

$$\pi(\phi, \eta; \alpha, \beta) = (1-\lambda)\beta E_{\eta_0}(\phi) + \lambda\alpha E_{\eta_1}(1-\phi).$$

[1] Note that we could have allowed nature to also choose $Y \in \mathcal{Y} \setminus (H_0 \cup H_1)$, in which case the penalty for such choices would be equal to 0. However this would have unnecessarily complicated notation.

[2] More generally, for the result below to hold we only need that the strategy set of each player is convex.

[3] ΔA denotes the set of all distributions with support contained in the set A.

We now present our characterization of tests that attain the unavoidable type II error in terms of being part of a Nash equilibrium of the game $\Gamma(\alpha, \beta)$ for appropriately defined α and β.

Proposition 1 *Assume that π is continuous in ϕ and η. The following statements are equivalent:*

(i) ϕ^* attains the unavoidable type II error among the tests in $\Delta\mathcal{F}$ that have level α.

(ii) There exists $\eta^* \in \Delta(H_0 \cup H_1)$ with $\lambda^* \in (0,1)$ such that (ϕ^*, η^*) is a Nash equilibrium of $\Gamma(\alpha, \beta)$ when $\beta = E_{\eta_1^*}(1 - \phi^*)$.

(iii) There exists $\eta^* \in \Delta(H_0 \cup H_1)$ with $\lambda^* \in (0,1)$ such that $E_{\eta_0^*}(\phi^*) = \alpha$, $E_Y(1 - \phi^*) \leq E_{\eta_1^*}(1 - \phi^*)$ for all $Y \in H_1$ and $\pi(\phi^*, \eta^*; \alpha, \beta) \leq \pi(\phi, \eta^*; \alpha, \beta)$ for all $\phi \in \Delta\mathcal{F}$ when $\beta = E_{\eta_1^*}(1 - \phi^*)$.

Continuity of π is only needed to show that (i) implies either (ii) or (iii). Neither in (ii) nor in (iii) do we assume that ϕ^* has level α. (ii) and (iii) are useful to evaluate whether a candidate test ϕ^* attains the unavoidable type II error, set $\beta = \max_{Y \in H_1} E_Y(1 - \phi^*)$. However the above is not useful for deriving the unavoidable type II error when one does not have such a candidate test. This is because the game $\Gamma(\alpha, \beta)$ depends via the parameter β on the equilibrium strategies. In the following formulation this is no longer the case. Here the exogenous parameters α' and β' that enter $\Gamma(\alpha', \beta')$ determine the ratio of the two errors of the test ϕ^*.

Corollary 1 *Assume that (ϕ^*, η^*) is a Nash equilibrium of the game $\Gamma(\alpha', \beta')$ for some given $\alpha', \beta' \in (0,1)$. Then ϕ^* has size $\alpha = E_{\eta_0^*}(\phi^*)$ and the unavoidable type II error among the tests in $\Delta\mathcal{F}$ that have level α is attained by ϕ^* and is equal to $E_{\eta_1^*}(1 - \phi^*) = \beta'\alpha/\alpha'$.*

One implication from understanding hypothesis testing in terms of a Nash equilibrium of a zero-sum game is that it is very easy to establish necessary conditions for which alternative tests may attain the unavoidable type II error. Here we iterate on a well known result for zero-sum games, namely that the set of Nash equilibria has a product structure.

Corollary 2 *Assume that* (ϕ^*, η^*) *is a Nash equilibrium of* $\Gamma(\alpha, \beta)$ *and that* $\beta = E_{\eta_1^*}(1 - \phi^*)$. *If* ϕ' *attains the unavoidable type II error then* (ϕ', η^*) *is also a Nash equilibrium of* $\Gamma(\alpha, \beta)$ *and* $E_{\eta_1^*}(1 - \phi^*) = E_{\eta_1^*}(1 - \phi')$.

Finally we show how to derive a lower bound on the type II error for the case where a Nash equilibrium of Γ is not known.

Proposition 2 *If there exists* ϕ^* *and* η^* *such that* $\lambda^* > 0$ *and* $\pi(\phi^*, \eta^*; \alpha, \beta) \leq \pi(\phi, \eta^*; \alpha, \beta)$ *for all* $\phi \in \Delta\mathcal{F}$ *when* $\alpha = E_{\eta_0^*}(\phi^*)$ *and* $\beta = E_{\eta_1^*}(1 - \phi^*)$ *then* $E_{\eta_1^*}(1 - \phi^*)$ *is a lower bound on the unavoidable type II error among all tests that have level* α.

Proof. We first prove Proposition 2. Let ϕ^* and $\eta^* = (\eta_0^*, \eta_1^*, \lambda^*)$ satisfy the conditions of the "if statement" of Proposition 2. Let ϕ be a test that has level α. Then $E_{\eta_0^*}(\phi) \leq \alpha$, $\pi(\phi, \eta^*; \alpha, \beta) \geq \pi(\phi^*, \eta^*; \alpha, \beta) = \alpha\beta$ and $\lambda^* > 0$ imply $E_{\eta_1^*}(1 - \phi) \geq E_{\eta_1^*}(1 - \phi^*)$. Hence $E_{\eta_1^*}(1 - \phi^*)$ is a lower bound on the unavoidable type II error.

We now prove Proposition 1. Note that it is easy to show equivalence of (ii) and (iii). For instance $E_{\eta_0^*}(\phi^*) = \alpha$ follows directly from the indifference of nature, that η_0^* and η_1^* both yield the same expected penalty. Otherwise nature would not randomize between them.

We now wish to prove that (ii) implies (i). Since nature is indifferent between η_0^* and η_1^* it follows that $\pi(\phi^*, \eta^*; \alpha, \beta) = \alpha\beta$. Together with the fact that $\pi(\phi^*, \eta^*; \alpha, \beta) \geq \pi(\phi^*, Y; \alpha, \beta) = \beta E_Y(\phi)$ for $y \in H_0$ it follows that ϕ^* has size α. Similarly it follows from $\pi(\phi^*, \eta^*; \alpha, \beta) \geq \alpha E_Y(1 - \phi)$ for $Y \in H_1$ that the type II error of ϕ^* is equal to β which together with Proposition 2 proves (i).

Finally, we prove that (i) implies (ii). Let β be the unavoidable type II error. Since both $\Delta\mathcal{F}$ and $\Delta\mathcal{Y}$ are compact and convex Hausdorff spaces and π is continuous there exists a Nash equilibrium of $\Gamma(\alpha, \beta)$ (Glicksberg, 1952). Let (ϕ', η^*) be such an equilibrium. Since ϕ^* has size α and attains the unavoidable type II error it follows that $\pi(\phi^*, \eta^*; \alpha, \beta) \leq \alpha\beta$. Since (ϕ', η^*) is a Nash equilibrium, $\pi(\phi', \eta^*; \alpha, \beta) \leq \pi(\phi^*, \eta^*; \alpha, \beta)$ and hence $\pi(\phi', \eta^*; \alpha, \beta) \leq \alpha\beta$. Given $\pi(\phi', \eta^*; \alpha, \beta) \geq \pi(\phi', \eta; \alpha, \beta)$

holds for all η it follows that ϕ' has level α and that ϕ' attains the unavoidable type II error. This proves (ii).

The statements in Corollaries 1 and 2 follow immediately. ■

3 Tests for Comparing the Means of Two Bernoulli distributions

We now use our above insights to investigate inference when testing the inequality of the means of two Bernoulli distributed random variables.

3.1 The Setting

Let $Y = (Y_1, Y_2)$ be a binary valued bivariate random variable so $Y \in \{0,1\}^2$. Thus Y_1 and Y_2 are two Bernoulli random variables with means (or success probabilities) we denote by p_1 and p_2 respectively. In a later section we extend our results to the case where the set of possible outcomes is only constrained to be contained in a known bounded set.

We wish to test the one-sided null hypothesis $H_0 : p_1 + d_0 \geq p_2$ against the composite alternative hypothesis $H_1 : p_1 + d \leq p_2$ for some given d_0 and d with $d_0 < d$. We refer to $d - d_0$ as the *gap* between the null and the alternative hypothesis. In terms of inference the difference between the two means is assumed to be the only parameter of interest.

Tests of *inequality* emerge when setting $d_0 = 0$. Tests of *superiority* (of Y_2 over Y_1) result when $d_0 > 0$, tests of *non-inferiority* or equivalence refer to the case where $d_0 < 0$ with focus typically on $d = 0$ (e.g. see Röhmel and Mansmann, 1999, Röhmel, 2005).[4] Tests for each $d_0 \in (-1, 1)$ are typically used when constructing confidence intervals for the difference between the two underlying means.

[4] The underlying story is that there is a new treatment whose outcome is given by Y_1 that should be compared to a reference treatment corresponding to Y_2.

3.2 Matched Pairs

Consider inference based on *matched pairs* where the statistician observes N independent realizations $y^j \in \{0,1\}^2$ of Y for $j = 1, ..., N$. Data generating processes that belong to $\Delta\{(1,0),(0,1)\}$ will play a special role and will be denoted by $Y(p_2)$.

A (randomized) test ϕ is formally given by

$$\phi : \left(\{0,1\}^2\right)^N \to [0,1]$$

where $\phi\left(y^1, ..., y^N\right)$ is the probability of recommending a rejection based on the sample $\left(y^1, ..., y^N\right)$.

Consider first tests of inequality so $d_0 = 0$. The natural candidate is the randomized version of McNemar's test (McNemar, 1947, see Lehmann and Romano, 2005, p. 138). This test, denoted in the following by ϕ^u, evaluates whether there are significantly more observations of $(0,1)$ than of $(1,0)$ in the data set. ϕ^u is uniformly most powerful among the unbiased tests (UMPU).[5] We show that ϕ^u attains the unavoidable type II error and hence that the property of being unbiased here does not constrain inference.

To also understand the case of $d_0 \neq 0$ we construct a new test that attains the unavoidable type II error.[6] This test emerges when applying the following two steps. First randomly transform the data set into one that contains only outcomes $(1,0)$ and $(0,1)$ in a way that leaves $EY_2 - EY_1$ unchanged. Then reject the null hypothesis if there are sufficiently more observations of $(0,1)$ than of $(1,0)$ in the transformed data set. The transformation independently replaces observations $(0,0)$ and $(1,1)$ equally likely with $(1,0)$ and with $(0,1)$. We now describe the recommendation of this test denoted by ϕ^+ for a sample that contains only observations $(1,0)$ and $(0,1)$. There is some $t \in \mathbb{Z}$ and $\tau \in [0,1)$ such that in a sample that contains z_1 observations of

[5] Recall that a test ϕ is *unbiased* if $E_{Y'}(\phi) \geq E_Y(\phi)$ when $Y \in H_0$ and $Y' \in H_1$. A test ϕ' is *uniformly more powerful* than a test ϕ if $E_{Y'}\left(\phi'\right) \geq E_{Y'}(\phi)$ for all $Y' \in H_1$.

[6] We remind the reader that the objective here is not to design practical tests but to uncover benchmarks useful to evaluate such practical tests.

$(1,0)$ and z_2 of $(0,1)$ the test ϕ^+ satisfies

$$\phi^+ = \begin{cases} 1 \text{ if } z_2 > t \\ \tau \text{ if } z_2 = t \quad \text{if } z_1 + z_2 = N \\ 0 \text{ if } z_2 < t \end{cases} \tag{1}$$

and

$$E_{Y\left(\frac{1}{2}(1+d_0)\right)}\left(\phi^+\right) = \alpha. \tag{2}$$

Note that the parameters t and τ as defined above are unique. Note also that if $d_0 = 0$ and $Y \in \Delta\{(1,0),(0,1)\}$ then $\phi^+ = \phi^u$. It follows from the proof below that ϕ^+ is unbiased and has size α. We apply Proposition 1 and Corollary 2.

Proposition 3 *(i) The unavoidable type II error is given by*

$$E_{Y\left(\frac{1}{2}(1+d)\right)}\left(1-\phi^+\right). \tag{3}$$

(ii) If $d_0 = 0$ then (3) is attained by the UMPU test ϕ^u.

(iii) (1) and (2) are necessary conditions for a test to attain the unavoidable type II error.

In particular we have shown that there is a least favorable distribution contained in $\Delta\{(1,0),(0,1)\}$. This will play an important role in later sections.

Proof. Let $\eta_0^* = Y\left(\frac{1}{2}(1+d_0)\right)$ and $\eta_1^* = Y\left(\frac{1}{2}(1+d_0)\right)$. We will first show that one can choose $\lambda^* \in (0,1)$ such that ϕ^+ is a best response against η^* in $\Gamma(\alpha,\beta)$ when $\beta = E\left(1-\phi^+|\eta_1^*\right)$.

Since $\eta_0^*, \eta_1^* \in \Delta\{(1,0),(0,1)\}$ we obtain that z_2 is a sufficient statistic for the information contained in the sample when facing η^* where $z_1 = N - z_2$. The expected penalty from rejecting the null hypothesis conditional on z_2 is equal to $\beta \Pr(H_0 \text{ true}|z_2)$. The expected penalty from not rejecting the null hypothesis conditional on y is equal to $\alpha \Pr(H_1 \text{ true}|z_2)$. We derive the ratio of these two expected penalties:

$$\begin{aligned} \frac{\beta \Pr(H_0 \text{ true}|z_2)}{\alpha \Pr(H_1 \text{ true}|z_2)} &= \frac{\beta\binom{2n}{z_2}\left((1+d_0)/2\right)^{2n-z_2}\left((1-d_0)/2\right)^{z_2}(1-\lambda)}{\alpha\binom{2n}{z_2}\left((1+d_1)/2\right)^{2n-z_2}\left((1-d_1)/2\right)^{z_2}\lambda} \\ &= \left(\frac{(1-d_0)(1+d_1)}{(1+d_0)(1-d_1)}\right)^{z_2}\left(\frac{1+d_0}{1+d_1}\right)^{2n}\frac{\beta(1-\lambda)}{\alpha\lambda}. \end{aligned} \tag{4}$$

Now set λ^* equal to the solution λ of

$$\left(\frac{(1-d_0)(1+d_1)}{(1+d_0)(1-d_1)}\right)^t \left(\frac{1+d_0}{1+d_1}\right)^{2n} \frac{\beta(1-\lambda)}{\alpha\lambda} = 1 \tag{5}$$

which exists and is necessarily contained in $(0,1)$.

Since the right hand side in (4) is increasing in z_2 it follows from the definition of λ^* that ϕ^+ is a best response to the strategy η^* of nature conditional on z_2 in the sense that ϕ^+ minimizes the expected penalty of the statistician among all possible tests ϕ. In particular, when $z_2 = t$ then the definition of λ^* ensures that the statistician is indifferent between rejecting and not rejecting the null hypothesis.

We will now establish the remaining statements in Proposition 1(iii). Given the way ϕ^+ is defined when either $(0,0)$ or $(1,1)$ is contained in the sample it is as if the statistician facing Y is really facing $Y' \in \Delta\{(1,0),(0,1)\}$ with $EY_2' - EY_1' = EY_2 - EY_1$. This is because the random transformation is performed independently for each matched pair in the sample and because this transformation preserves the expected difference between the two variables. Hence, for such Y and Y' we find that $E_Y\left(\phi^+\right) = E_{Y'}\left(\phi^+\right)$. Using the properties of ϕ^+ we thus obtain

$$\max_{Y \in H_1 \cap \Delta\{(1,0),(0,1)\}} E\left(1-\phi^+\right) = E_{Y\left(\frac{1}{2}(1+d_1)\right)}\left(1-\phi^+\right).$$

This establishes Proposition 1(iii) which completes the proof of (i).

Concerning part (iii), if ϕ' attains the unavoidable type II error then following Corollary 2 ϕ' hat to be a best response to η^* which means that it has to satisfy (1) and (2).

We now prove part (ii) so assume $d_0 = 0$. Note that ϕ^+ is unbiased. Hence the UMPU test ϕ^u is uniformly more powerful than ϕ' and hence ϕ^u also attains the unavoidable type II error. ■

The above proof reveals that unbiasedness does not here constrain inference:

Corollary 3 *All statements in Proposition 3 remain true if one restricts attention to unbiased tests.*

Following Pratt (1961) the unavoidable type II error can be used to derive a tight

lower bound on the maximal expected width of any family of confidence intervals for the difference between the two underlying means.

Given space constraints we numerically illustrate our findings only in the following more intricate setting.

3.3 Independent Observations and Sequential Sampling

We now consider inference based on N independent observations where for simplicity we focus on the case where N is even. Let $n = N/2$. Each observation consists of an outcome realized by one of the two random variables. The sample can thus be described as $\left(\left(i_k, y^k_{i_k}\right), k = 1, .., 2n\right) \in \left(\{1,2\} \times \{0,1\}\right)^{2n}$ where $y^k_{i_k} \in \{0,1\}$ has been drawn from Y_{i_k}, $k = 1, ..., 2n$. The sample is *balanced* if $|\{k : i_k = 1\}| = n$.[7] Let y_i be the number of times that Y_i realized 1 in this sample, so $y_i = \left|\left\{k : i_k = i, y^k_i = 1\right\}\right|, i = 1, 2$. We allow the statistician to choose sequentially which random variable to observe an outcome from, hence to determine i_k conditional on $\left(\left(i_j, y^j_{i_j}\right), j = 1, .., k-1\right)$. This we call *sequential sampling*. Formally a test ϕ now describes how to gather the sample, so

$$\phi : \cup_{k=0}^{2n-1} \left(\{1,2\} \times \{0,1\}\right)^k \to \Delta\{1,2\}$$

where ϕ describes the index of the random variable from which the next outcome should be realized. As in the setting with matched pairs, the test ϕ also specifies the probability of making a rejection once the entire sample has been gathered, hence additionally we have that

$$\phi : \left(\{1,2\} \times \{0,1\}\right)^{2n} \to [0,1].$$

Under *simultaneous* sampling the statistician determines ex-ante how many times to observe each variable. This can be formally embedded in sequential sampling by asserting that $|\{k : i_k = 2\}|$ is a constant and hence does not depend on the observed outcomes. An important representative is *balanced sampling* where $|\{k : i_k = 2\}| = n$.

Clearly sequentially sampling $2n$ independent observations generates less information than sampling $2n$ matched pairs. Consequently the unavoidable type II error

[7] $|A|$ denotes the cardinality of the finite set A.

under matched pairs (see (3)) is a lower bound on the type II error under sequential sampling. In the following we show that the two unavoidable type II errors can coincide.

Let ϕ^{**} be any test that has the following three properties. (i) ϕ^{**} generates a balanced sample. (ii) There exists $b \in \mathbb{Z}$ such that

$$\phi^{**}\left(\left(i_k, y_{i_k}^k\right)_{k=1}^{2n}\right) = \begin{cases} 1 \text{ if } y_2 - y_1 > b \\ 0 \text{ if } y_2 - y_1 < b \end{cases}. \tag{6}$$

(iii) The power of ϕ^{**} is equal to α when $p_1 = \frac{1}{2}(1 - d_0)$ and $p_2 = \frac{1}{2}(1 + d_0)$, formally

$$E_{Y\left(\frac{1}{2}(1+d_0)\right)}\left(\phi^{**}\right) = \alpha. \tag{7}$$

Notice that ϕ^{**} is constructed similarly to ϕ^{+}. In fact, it follows that $b = t - n$. Moreover, if $\phi^{**} = \tau$ when $y_1 - y_2 = b$ then ϕ^{**} has the same behavior as ϕ^{+} whenever $Y \in \Delta\{(1,0),(0,1)\}$. This particular representative will be called $\phi_{b,-1}^{**}$.

We combine Propositions 1 and 3 to derive necessary and sufficient conditions for when inference (in terms of type II error) based on a sequential test is as good as when based on matched pairs.

Proposition 4 *(i) (3) is a lower bound on the type II error of any sequential test that has level α.*

(ii) If ϕ^{**} has size α and attains its type II error when $Y = Y\left(\frac{1}{2}(1+d)\right)$ then ϕ^{**} attains the unavoidable type II error which is equal to (3).

(iii) (6) and (7) are necessary conditions for a sequential test ϕ with level α to have a type II error equal to (3).

If $d_0 = 0$ then we find a related though more specific result to Proposition 4(ii) in Lehmann and Romano (2005, Problem 3.59): $\phi_{b,-1}^{**}$ is uniformly most powerful among all tests that gather a balanced sample when testing $H_0 : p_1 = p_2 = 1/2$ against $H_1 : p_2 = 1 - p_1 > 1/2$.

Proof. Part (i) follows immediately as matched pairs generates more information than sequential testing. For parts (ii) and (iii) consider the game as defined in the

proof of Proposition 3. Concerning part (ii), note that property (6) and (7) ensure that ϕ^{**} is a best response to the strategy of nature. The remaining assumptions ensure that Proposition 1(iii) can be applied. The proof of part (iii) is analogous to that of Proposition 3(ii). ■

Numerical calculations for many values of n and α reveal that $\phi^{**}_{b,-1}$ has size α and that if the gap $d - d_0$ is sufficiently large then $\phi^{**}_{b,-1}$ attains its type II error when $Y = Y\left(\frac{1}{2}(1+d)\right)$. Thus we have found that inference based on a balanced sample is as powerful as when based on matched pairs provided there is sufficient gap between the hypotheses. For instance, when $n = 20$, $\alpha = 0.05$ and $d_0 = 0$ then $d \geq 0.27764$ is sufficient for this to be true. In the following we present alternative tests within the class ϕ^{**} that have the potential to generate the same power of inference as under matched pairs for smaller values of d than under $\phi^{**}_{b,-1}$. The idea is to vary the recommendation on the border of the critical region where $y_2 - y_1 = b$.

3.3.1 The L Test

Let $\phi^{**}_{b,v}$ be a test defined as follows. $\phi^{**}_{b,v}$ gathers a balanced sample and satisfies (6) and (7). When $y_2 - y_1 = b$ this test rejects the null hypothesis on the $2(v+1)$ data points that are closest to the border and rejects with a constant probability in the interior. Specifically $\phi^{**}_{b,v}(y_1, y_1 + b) = 1$ if $y_1 \leq v$ or $y_1 \geq n - v - b$ and $\phi^{**}_{b,v}(y_1, y_1 + b) = \eta$ if $v < y_1 < n - v - b$ where $v \in \{-1, 0, 1, .., \lfloor (n-b)/2 \rfloor\}$. Here v and η have to be chosen such that (7) holds. Then $\phi^{**}_{b,v}$ belongs to the class of tests ϕ^{**} by construction.

It turns out in all numerical examples that $\phi^{**}_{b,v}$ attains the unavoidable type II error if and only if $d \geq d^*$ for appropriately chosen threshold d^*. The value of v that minimizes the threshold d^* will be denoted by v^*, ϕ^{**}_{b,v^*} will also be called the *L test*.

We illustrate for $n = 20$, $\alpha = 0.05$ and $d_0 = 0$. We find that $b = 6$ and $v \leq 5$ are necessary to satisfy (6) and (7). However we find that $\phi^{*}_{b,v}$ only has size 0.05 if $v \leq 3$. The next step is then to search for each value of v for the values of d under which the type II error is attained when $Y = Y\left(\frac{1}{2}(1+d)\right)$. We then select $v =: v^*$ which has this property for the smallest values of d. Here it turns out that $\phi^{**}_{5,3}$ is uniformly

more powerful than $\phi^{**}_{5,v}$ for $-1 \leq v \leq 2$. Hence $v^* = 3$ and we find that $d^* = 0.18969$. Remember that $d^* = 0.27764$ under $\phi^{**}_{5,-1}$.

In Figure 1 we plot the unavoidable type II error under matched pairs together with the type II error of the L test $\phi^{**}_{5,3}$. It turns out for $d < d^*$ that the type II error under $\phi^{**}_{5,3}$ is attained when $p_1 = 0$ and $p_2 = d$. Note that in the region where $d > d^*$, and hence where the two graphs coincide, the graphs show the unavoidable type II error under balanced sampling.

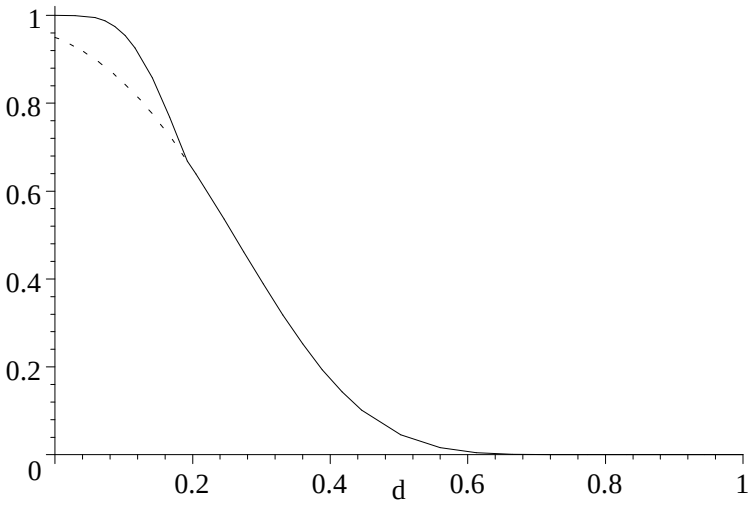

Figure 1: Type II error of $\phi^{**}_{5,3}$ (solid) and unavoidable type II error under matched pairs (dotted) as function of d when $n = 20$, $\alpha = 0.05$ and $d_0 = 0$.

In Table 1a we provide the values of b and v^* of the L test together with the threshold d^* and the type II error attained at this threhold for various values of n when $\alpha = 0.05$. This means that the L test attains the unavoidable type II error whenever its type II error is below the indicated value at the threshold $d = d^*$.

Table 1a: Some Parameters of the L Test when $\alpha = 0.05$ and $d_0 = 0$

n	5	10	15	20	25
b	3	4	4	5	6
v^*	0	2	1	3	6
d^*	0.4894	0.3343	0.2735	0.1897	0.2116
Type II error for $d = d^*$	0.52	0.56	0.55	0.67	0.56
n	30	40	50	60	70
b	6	7	8	9	10
v^*	5	9	13	26	24
d^*	0.2053	0.1607	0.1404	0.1374	0.1361
Type II error for $d = d^*$	0.52	0.58	0.6	0.56	0.51

In Table 1b we illustrate how these parameters change with d_0 when $n = 30$ and $\alpha = 0.025$. Note that the L test attains the unavoidable type II error in this table whenever this is above 0.475 unless d_0 is very large.

Table 1b: Some Parameters of the L Test when $n = 30$ and $\alpha = 0.025$

d_0	-0.98	-0.95	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4
d^*	-0.933	-0.87	-0.789	-0.645	-0.573	-0.394	-0.317	-0.224
b, v^*	$-28, -1$	$-26, -1$	$-23, -1$	$-19, -1$	$-15, 4$	$-12, 1$	$-8, 7$	$-5, 7$
type $\text{II}_{d=d^*}$	0.667	0.615	0.587	0.556	0.735	0.527	0.646	0.693
d_0	-0.3	-0.25	-0.234	-0.2	-0.1	0	0.1	0.2
d^*	-0.069	-0.0284	0	0.0579	0.156	0.256	0.312	0.397
b, v^*	$-2, 7$	$0, 10$	$0, 8$	$2, 13$	$5, 11$	$8, 10$	$10, 4$	$13, 3$
type $\text{II}_{d=d^*}$	0.548	0.586	0.546	0.475	0.488	0.484	0.614	0.65
d_0	0.3	0.4	0.5	0.6	0.65	0.7	0.8	0.841
d^*	0.447	0.573	0.693	0.735	0.671	0.889	0.96	1
b, v^*	$16, 3$	$19, 3$	$21, 0$	$24, 0$	$25, 0$	$26, -1$	$28, -1$	$29, -1$
type $\text{II}_{d=d^*}$	0.779	0.683	0.562	0.737	0.961	0.359	0.275	0

where type $\text{II}_{d=d^*}$ denotes the type II error when $d = d^*$

3.3.2 Bounds to Inference among Noninferiority Tests

In the context of noninferiority tests where $d_0 < 0$ there has been special interest in deriving minimal sample sizes necessary for particular inference, mostly for the case where $d = 0$. For a given pair of hypotheses the *unavoidable sample size* refers to the smallest sample under which the unavoidable type II error is below a given threshold. Here we focus on balanced samples and report in Table 2 the value of n corresponding to the unavoidable sample size necessary to achieve a type II error of 0.2 when $\alpha = 0.025$. These values are derived by first calculating lower bounds on the sample sizes using Proposition 4(i) and then verifying for those sample sizes and for the given value of d_0 that the conditions in Proposition 4(ii) hold when considering $\phi^{**} = \phi^{**}_{b,v^*}$. We include in Table 2 the values, denoted by n_{asym}, that result from the asymptotic formula of Rodary et al. (1989) (see also Farrington and Manning, 1990).

Table 2: Testing Non-Inferiority with a Balanced Sample:
Achieving Type II Error below 0.2 when $\alpha = 0.025$ and $d = 0$

d_0	-0.3	-0.25	-0.2	-0.15
Unavoidable sample size n	43	62	97	173
n_{asym}	41	61	96	172

3.3.3 Unbiased Tests for Inequality

Here we briefly investigate inference under sequential sampling within the class of unbiased tests when $d_0 = 0$. The randomized version of Fisher's (1935) exact test due to Tocher (1950), denoted here by ϕ^T, is UMPU under simultaneous sampling. Hence it attains the unavoidable type II error among the unbiased tests under simultaneous sampling. If its type II error (given a balanced sample) is attained for $Y\left(\frac{1}{2}(1+d)\right)$ we obtain more, namely that ϕ^T attains the unavoidable type II error among unbiased tests for sequential sampling. However, ϕ^T never attains the lower bound (3) on the type II error as it does not satisfy (6). Extending Proposition 2 to unbiased tests we then obtain the following.

Corollary 4 *Consider $d_0 = 0$. Then $E_{Y\left(\frac{1}{2}(1+d)\right)}\left(\phi^T\right)$ is a lower bound on the type II error of any unbiased test which is strictly above (3). In particluar, if the type II error of ϕ^T under a balanced sample is attained at $Y\left(\frac{1}{2}(1+d)\right)$ then the unavoidable type II error among unbiased tests is attained by ϕ^T and equal to $E_{Y\left(\frac{1}{2}(1+d)\right)}\left(\phi^T\right)$.*

It follows that unbiasedness constrains inference under independent observations when $d_0 = 0$. All numerical examples have revealed that ϕ^T based on a balanced sample attains the unavoidable type II among the unbiased tests for all $d > 0$. Moreover, we have found that the type II error of ϕ^T is very close to the bound (3) if n is not too small. For instance, we find for $n = 20$ and $\alpha = 0.05$ that the maximal distance between the type II error of ϕ^T and (3) is 0.0072. For Table 3 we have also calculated the maximal distance for other values of n. Importantly, we have verified each time using the L test that this maximal distance is attained for a value of d where (3) is

equal to the unavoidable type II error. Hence Table 3 lists the maximal added type type II error of ϕ^T.

3.3.4 Evaluating Specific Tests for Inequality in Balanced Samples

Next we calculate the added type II error of two special nonrandomized tests for the case where $d_0 = 0$. We consider the B test (Boschloo, 1970) which is uniformly more powerful than Fisher's (1935) exact test and the Z test (Suissa and Shuster, 1984)[8]. We only evaluate these tests for values of d where the lower bound on the type II error is below 0.8 for following reason. These tests are nonrandomized and hence their type II error is equal to 1 when $p_1 = p_2 = 0$. On the other hand, the maximal unavoidable type II error is equal to $1 - \alpha$. Thus, α is the lower bound on the maximal type II error of any nonrandomized test. To dampen this disadvantage of being nonrandomized we only evaluate the tests over those values of d where the lower bound (3) is below 0.8. For instance, for $n = 20$ and $\alpha = 0.05$ this means that we only consider the added type II error for $d \geq 0.12817$. With this restriction on the possible alternative hypotheses we find for the chosen values of n that the upper bound on the added type II error is largest in the region of d where it is tight. Thus we are able to present maximal added type II errors in Table 3. Note that an analysis with only $\phi_{b,-1}$ would not have generated this result as for instance the maxima are attained under $n = 20$ in the region where $\phi_{b,-1}$ does not attain the unavoidable type II error (so where $d < 0.27764$).

Table 3: Maximal Added Type II Error when $\alpha = 0.05$

n	5	10	15	20	25	30	40	50
UMPU	0.087	0.029	0.026	0.0071	0.0066	0.0093	0.0067	0.0032
Z test*	0.136	0.031	0.048	0.034	0.066	0.013	0.013	0.0205
B test*	0.136	0.08	0.034	0.033	0.046	0.012	0.012	0.0204

(* when the unavoidable type II error is below 0.8)

[8] Suissa and Shuster (1985) have verified that the Z test is uniformly more powerful than Fisher's exact test when $\alpha \in \{1\%, 2.5\%, 5\%\}$ and $10 \leq n \leq 150$.

We hasten to point out that while the added type II error is a useful means for comparing tests it should not be the sole measure for selecting a test. For instance, in this table the Z test always attains a slightly higher added type II error than the B test. However, if $n \in \{20, 25, 30, 40\}$ then while the power of the Z test is never lower than that of the B test by more than 0.043 it can be up to 0.22 higher. For these parameters the Z test seems preferable. On the other hand, when $n = 15$ then it turns out that the B test is uniformly more powerful than the Z test.

An alternative means to evaluate a test ϕ is to compare its *minimal sample size* to the unavoidable sample size. The minimal sample size of a test ϕ with level α refers to the smallest value of n for which its type II error is below a given threshold β. In Table 5 we present the minimal sample size of the Z test and the unavoidable sample size for various values of d when $d_0 = 0$, $\alpha = 0.05$ and $\beta = 0.2$. Note that we find in each case as in numerical examples of Section 3.3.2 that the lower bound derived using (3) is tight.

Table 5: Minimal Sample Sizes for Type II Error Below 0.2 when $\alpha = 0.05$ and $d_0 = 0$

d	0.5	0.4	0.3	0.25	0.2
Unavoidable sample size	12	19	34	49	77
Z test	13	20	37	51	79

Following Table 5, given $d = 0.3$, there is no sequential test with level 0.05 that yields a type II error below 0.2 when $n \leq 33$. If instead $n \geq 34$ then the unavoidable type II error is below 0.2. The Z test based on a balanced sample requires $n = 37$ to attain a type II error below 0.2.

4 Testing given Multiple Outcomes

Here we consider the more general setting where outcomes belong to some known bounded set Z which contains more than two different outcomes. It will be enough

to consider the case where $\{0,1\} \subsetneqq Z \subseteq [0,1]$.[9] In addition to comparing means we will now also consider testing the equality of the two distributions. If Z is not finite, such as when $Z = [0,1]$, then our hypotheses will be *nonparametric* (cf. Kendall and Sundrum, 1953). We show that our previous results extend.

Let P denote the distribution of the joint random variable $Y = (Y_1, Y_2)$, let F_{Y_i} be the cdfs of the marginal distribution with respect to Y_i and let EY_i be the expected value, $i = 1, 2$. Let $g : [0,1] \to \Delta\{0,1\}$ be the so-called *binomial transformation* where $g(z) = 1$ with probability z and $g(z) = 0$ with probability $1 - z$, $z \in [0,1]$. Note that $\int g(z)\, dP_i(z) = EY_i$ and that g is the identity on $\{0,1\}$. For a given test ϕ defined for binary valued data let $\phi \circ g$ be the test for data contained in $[0,1]$ defined by first transforming each observation independently into $\{0,1\}$ using g and then applying ϕ to the transformed sample.

The first step is to show that the power of inference in terms of type II error remains unchanged when intermediate outcomes are possible. In particular we find that there is always a least favorable distribution that puts only weight on the extreme outcomes. The result holds whenever both hypotheses only depend on means or when the alternative hypothesis has this property while the null hypothesis postulates the identity of the two distributions. Here we utilize that Bernoulli distributions are identical if and only if their means are equal.

Proposition 5 *Consider either tests of* $H_0 : F_{Y_1} \equiv F_{Y_2}$ *against* $H_1 : (EY_1, EY_2) \in W$ *for some* $W \subset [0,1]^2 \setminus \{(w,w), w \in [0,1]\}$ *or tests of* $H_0 : (EY_1, EY_2) \in W_0$ *against* $H_1 : (EY_1, EY_2) \in W_1$ *for some* $W_0, W_1 \subset [0,1]^2$ *with* $W_0 \cap W_1 = \emptyset$. *If* ϕ *attains the unavoidable type II error for binary valued data then* $\phi \circ g$ *attains the unavoidable type II error when outcomes belong to* Z.

Proof. Given $P \in \Delta[0,1]^2$ let $P^0 \in \Delta\{0,1\}^2$ satisfy $P_i^0(1) = EY_i$, $i = 1, 2$. Then $E_P(\phi \circ g) = E_{P^0}(\phi)$. So for given $w \in [0,1]^2$ it follows that

$$\max_{P \in \Delta[0,1]^2 : (EY_1, EY_2) = w} E_P(\phi) \geq \max_{P \in \Delta\{0,1\}^2 : (EY_1, EY_2) = w} E_P(\phi) = \max_{P \in \Delta[0,1]^2 : (EY_1, EY_2) = w} E_P(\phi \circ g).$$

[9] For general Z first transform all outcomes linearly, mapping the extreme points into 0 and 1 respectively.

Similarly,

$$\max_{P\in\Delta[0,1]^2:F_{Y_1}\equiv F_{Y_2}} E_P(\phi) \geq \max_{P\in\Delta\{0,1\}^2:F_{Y_1}\equiv F_{Y_2}} E_P(\phi)$$
$$= \max_{P\in\Delta\{0,1\}^2:EY_1=EY_2} E_P(\phi) = \max_{P\in\Delta[0,1]^2:EY_1=EY_2} E_P(\phi\circ g).$$

Given these observations the claim is immediate. ■

Now we combine Proposition 5 with some of our previous results for binary valued distributions. In particular we gain insights to inference among permutation tests as these are particular unbiased tests for testing the identity of two distributions.

Corollary 5 *Consider either tests of* $H_0 : F_{Y_1} \equiv F_{Y_2}$ *against* $H_1 : EY_1 + d \leq EY_2$ *for some* $d > 0$ *or tests of* $H_0 : EY_1 + d_0 \geq EY_2$ *against* $H_1 : EY_1 + d \leq EY_2$ *for some* $d_0 < d$. *Consider* $2n$ *independent realizations generated from a balanced sample, from sequential testing or from sampling matched pairs. (3) is a lower bound on the unavoidable type II error which is tight whenever it is tight for the setting with binary valued outcomes, in particular when sampling matched pairs. (3) can be attained with an unbiased test when sampling matched pairs while this is not true under sequential sampling.*

5 Conclusion

The knowledge of a compact set that contains all outcomes plays a central role in our analysis. Given that we make no distributional assumptions the statistician may face distributions that only put weight on the extreme outcomes in the support. In fact it turns out that a least favorable distribution is contained among these particular distributions. In other words, distribution-free inference is not limited per se by the number of possible outcomes but by its range.

We find particular distributions to be least favorable and use their property to make statements about inference among all sequential tests. The particular property is that the two random variables almost surely yield different outcomes. The intuition is that these distributions generate the most variance and hence make learning most difficult when interested in the difference between the two means.

The strategic component of trying to outguess the opponent in the underlying zero-sum game naturally leads to mixed strategies being played in equilibrium. The consequence is that the unavoidable type II error is typically realized by a randomized test. Randomized tests have only received little attention in statistics but here we find that understanding their properties is insightful to deriving bounds to inference. Randomized tests for data with binary valued outcomes along with insights from this paper are also used by Schlag (2007b) in the construction of nonrandomized nonparametric tests and thus attaining the first exact solution to a nonparametric Behrens Fisher problem.

Game theoretic methodology and thinking is generating new insights and results in distribution-free hypothesis testing. The existence of equilibria in which opponent's strategy set is implicitly limited by own play (e.g. underlying Corollary 5) does not come at a surprise to game theorists. The extreme example of this phenomenon arises in the babbling equilibrium of games with cheap talk where ignoring messages makes opponent's messages that have no meaning optimal and vice versa (Crawford and Sobel, 1982). For the first time in hypothesis testing we can compare tests based on sequential sampling. Instead of needing to compute type II errors for each sampling sequence one only needs to consider the best responses to nature's strategy. Inference focuses on specific pairs of composite hypotheses and often generalizes to a large class of hypotheses. Uniqueness results are easily established. For instance, given the results in Section 2 it is an easy excersize for any game theorist to show that the binomial test is the unique uniformly most powerful test. This follows when investigating best response behavior which turns out to resemble that of "Matching Pennies".

References

[1] Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122.

[2] Boschloo, R. D. (1970). Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities. *Statist. Neerlandica* **24** 1–35.

[3] Crawford, V. P and Sobel, J. (1982). Strategic information transmission. *Econometrica* **50** 1431–1451.

[4] Cucconi, O. (1968). Contributi all'analisi sequenziale nel controllo di accettazione per variabili. *Atti dell' Ass. Italiana per il Controllo della Qualità* **6** 171–186.

[5] Farrington, C.P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **9** 1447–1454.

[6] Fisher, R. A. (1935). The logic of inductive inference. *J. Roy. Stat. Soc.* **98** 39–54.

[7] Glicksberg, I. L. (1952). A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc. Amer. Math. Soc.* **3** 170–174.

[8] Kendall, M. G. and Sundrum, R. M. (1953). Distribution-free methods and order properties. *Rev. Int. Statist. Inst.* **21** 124–134.

[9] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York.

[10] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** 153–157.

[11] Pratt, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 549–567.

[12] Rodary, C., Com-Nougue, C., and Tournade, M.-F. (1989). How to establish equivalence between treatments: a one-sided clinical trial in paediatric oncology. *Stat. Med.* **8** 593–598.

[13] Röhmel, J. (2005). Problems with existing procedures to calculate exact unconditional p-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biom. J.* **47** 37–47.

[14] Röhmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biom. J.* **41** 149–170.

[15] Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons, New York.

[16] Schlag, K. H. (2003). How to minimize maximum regret in repeated decision-making. Unpublished Manuscript, European University Institute.

[17] Schlag, K. H. (2006). Eleven - tests needed for a recommendation. European University Institute Working Paper ECO 2006/2.

[18] Schlag, K. H. (2007a). Finite sample inference for the mean of an unknown bounded random variable without assumptions. Unpublished Manuscript, European University Institute.

[19] Schlag, K. H. (2007b). Testing equality of two means without assumptions - solving the nonparametric Behrens-Fisher problem exact. Unpublished Manuscript, European University Institute.

[20] Suissa, S. and Shuster, J. J. (1984). Are uniformly most powerful unbiased tests really best? *Amer. Statist.* **38** 204–206.

[21] Suissa, S. and Shuster, J. J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *J. Roy. Stat. Soc. Ser. A* **148** 317–327.

[22] Tocher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* **37** 130–144.

[23] von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Math. Ann.* **100** 295–320.