# Exact Nonparametric Binary Choice and Ordinal Regression Analysis*

Karl H. Schlag†

February 15, 2014

**Abstract**

We provide methods for investigating the relationship between an attribute and an ordinal outcome, as in a binary choice model. We do not make additional assumptions on the errors as in the (ordinal) logit or probit models and do not invoke asymptotic theory. These tests are exact as their type I error probability can be bounded above by their nominal level in the data set that is being investigated. Bounds on type II error probabilities are provided. Methods extend to duration and survival analysis with right censoring. Several data applications are given.

Keywords: exact, probit, logit, stochastic inequality, latent variable, single index model, ordinal, nonparametric, survival analysis.

JEL classification: C14, C20.

## 1 Introduction

Binary choice models in the form of logistic or probit regressions are very popular for understanding how attributes influence outcomes, yet they have grave drawbacks when coming to inference. When finding significant evidence of a coefficient there may be several reasons why the corresponding null hypothesis has been rejected. One would wish that this reveals a significant evidence of a positive relationship between this attribute and the outcome. However there are also three other reasons that can lead to a rejection of the null hypothesis. (i) Errors may not distributed as postulated by the underlying model. (ii) The formulae used to compute the corresponding p value

*We wish to thank Francesca Solmi whose work motivated this project, Heiko Rachinger for valuable comments and Robin Ristl for programming the software used to analyze the data.

†Department of Economics, University of Vienna, Austria, email: karl.schlag@univie.ac.at

may not be exact.[1] (iii) The relationship between the attribute and the outcome may not be monotone.

In this paper we show how one can make inference without being subject to any one of these three problems. Ad (i) our method does not make distributional assumptions on the errors. Ad (ii) our method is exact, hence p values are correct and do not rest on asymptotic theory. Ad (iii) we measure the effect of the attribute on the outcome without imposing monotonicity and derive confidence intervals for this effect.

In the existing literature, most models are not exact and it seems almost foolish to derive conclusions based on methods whose properties for the given samples sizes are not known. Even in parametric models simulations typically cannot uncover conclusive evidence of the validity of the method as the underlying data generating processes are too rich for exhaustive simulations. There are exact versions of logistic regression (Cox, 1972, Mehta and Patel, 1995), however they can only be implemented in very small data sets. Moreover, it is hard to defend that the true data generating processes follows a logistic model. The permutation test methodology (van Elteren, 1960) allows to make exact inference without making additional assumptions on the errors. However inference is limited to uncovering that an attribute influences the outcome, leaving it open how it influences the outcome. Yet in most studies one needs to able to conclude whether or not higher values of the attribute are more likely to yield higher outcomes.

The downside of our approach is that it can only be applied to data in which there are sufficiently many ties in the attributes. This is a direct consequence of the fact that we do not make any assumptions on the errors. If any two individuals can be distinguished according to the attributes that are used as controls then it is not possible to separate the impact of these controls from the impact of the attribute of interest. In applications, if one of the control variables is continuous one has to collect observations in bins. For instance, instead of controlling for the level of income one would control for the income quantiles the subject belongs too. Note that despite its parametric nature exact logistic regression similarly can also only be applied when there are sufficiently many ties in the attributes. Common to the permutation method, our test is based on comparing individuals who only differ in terms of the attribute of interest. The value added of our approach as compared to the permutation setting is that we are able to sign the effect when the null hypothesis is rejected.

We provide formal means to measure the quality of our inference by providing bounds on type II error probabilities. These can then be used for comparison to other tests and for sample size calculations. Most importantly, we provide several

[1] Inference is exact (Yates, 1934) if the true level of the test is bounded above by the nominal level of the test.

data examples to demonstrate the practicality of our new methods.

We proceed as follows. After introducing the model in Section 2 we present and test in Section 3 a model in which monotonicity is assumed. In Section 4 we drop the monotonicity assumption and instead introduce and test for the average incremental effect. Two tests are provided, depending on whether or not the outcome space is binary. In Section 5 we present several data examples to show how our methods perform. In Section 6 we conclude. The appendix contains the proofs alongside with some original material needed for them on how to compare two Bernoulli sequences that are not necessarily identically distributed.

# 2 A Nonparametric Ordinal Regression Model

We consider an ordinal regression model that describes the relationship between $m$ independent variables (or attributes) and an ordered dependent variable (or outcome). We investigate this relationship using $n$ observations, to simply exposition we will assume that each observation is associated to a different individual. Outcomes may be binary valued as in a binary choice model, but they may also be contained in a general ordered set. For the given attributes the outcomes are realized independently.

We wish to investigate the relationship between one of the attributes and the outcome without making any parametric assumptions on the errors and without postulating a specific functional relationship between the attributes and the outcome. We will make inference conditional on the values of the attributes, hence our tests apply both to the case where the attributes are chosen by the designer and where the attributes themselves are random. In particular, when the attributes are modelled as random variables they are allowed to be correlated across individuals.

We use $i \in \{1,..,n\}$ to index individuals, $k \in \{1,..,m\}$ to index the attribute and $j \in \{1,..,m\}$ to be the index of the specific attribute of interest. Outcomes are ordinal, let $\mathcal{Y}$ be the set of possible outcomes, let $y_i$ be the outcome realized by individual $i$, and let $y = (y_i)_{i=1}^n \in \mathcal{Y}^n$. Let $\mathbb{A}_k$ be the set of possible values of the attribute with index $k$, let $x_{ik} \in \mathbb{A}_k$ be the value of the $k$-th attribute of individual $i$. We assume that the attribute of interest $j$ is ordinal and allow for attributes $k \neq j$ to be categorical.[2] Let $x_i = (x_{i1},..,x_{im}) \in \times_{k=1}^m \mathbb{A}_k$ be the vector of attributes of individual $i$ and let $x = (x_i')_{j=1}^m \in \times_{k=1}^m \mathbb{A}_k^n$. Let $X \in \times_{k=1}^m \mathbb{A}_k^n$ and $Y \in \mathcal{Y}^n$ be the associated random vectors and let $X_i = (X_{i1},..,X_{im})$. The data generating process is described by the joint distribution of $(X,Y)$.

We model the dependence of the outcome on the attributes in two different ways. In our first approach we postulate that there is a monotone relationship between the

[2] Formally there is a complete order $\succsim$ on $\mathbb{A}_j$. To simplify notation, for $z_j, w_j \in \mathbb{A}_j$ let $z_j - w_j = 1$, 0 and $-1$ if $z_j \succ w_j$, $z_j \sim w_j$ and $z_j \prec w_j$ respectively.

outcome and the attribute of interest and only wish to infer whether this is strictly increasing or strictly decreasing. In our second approach we drop this monotonicity assumption and instead investigate a measure of the dependence of the outcome on the attribute of interest.

# 3 The Model with Monotonicity

We consider a *nonparametric monotone ordinal regression model* as defined by the following two assumptions:

(a) $\{Y_i|X=x\}_{i=1}^n$ are independently distributed for all $x$.

(b) For each $x$ one of the following three conditions is true:

(i) $P(Y_r > Y_i|X=x) > P(Y_r < Y_i|X=x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$,

(ii) $P(Y_r > Y_i|X=x) = P(Y_r < Y_i|X=x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$,

(iii) $P(Y_r > Y_i|X=x) < P(Y_r < Y_i|X=x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$.

Putting (i) in words, one says that higher values of the attribute of interest unambiguously tend to generate higher outcomes. One may also speak of a strictly monotone increasing relationship. (ii) reveals that outcomes are generated independently of the value of the attribute of interest. (ii) and (iii) together will be referred to as a monotone decreasing relationship.

Note that monotonicity per se is not investigated but assumed, it is the direction of this monotonicity that will be investigated. Note also that outcomes need not be identically distributed among those individuals that have the same attributes.

## 3.1 Examples

### 3.1.1 Stochastic Inequality

Consider the case where outcomes of individuals with identical attributes are identically distributed and where the attribute of interest is binary valued. If there are no other attributes, so if $m = 1$, then it is as if we are comparing random realizations of two independent random variables $Y_0$ and $Y_1$ where $Y_k$ is the outcome generated by those individuals with attribute $k$. For instance (i) and (ii) together postulate that $P(Y_1 > Y_0) \geq P(Y_1 < Y_0)$. A monotone decreasing relationship is a stochastic inequality (see Brunner and Munzel, 2000, Schlag, 2008a) between the outcome and the attribute of interest.

When one wishes to control for additional attributes, so when $m \geq 2$, then we are investigating a stochastic inequality conditional on the values of the other attributes, assuming that the sign of this inequality does not depend on the values of the other attributes.

### 3.1.2 Binary Choice

For the important special case of a binary choice model we have $\mathcal{Y} = \{0, 1\}$ and the above assumption can be rewritten as follows. For each $x$ one of the following three conditions is true:
(i) $P(Y_r = 1|X = x) > P(Y_i = 1|X = x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$,
(ii) $P(Y_r = 1|X = x) = P(Y_i = 1|X = x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$,
(iii) $P(Y_r = 1|X = x) < P(Y_i = 1|X = x)$ if $x_{rj} > x_{ij}$ and $x_{rk} = x_{ik}$ for all $k \neq j$.

A more specific model obtains by assuming for each $x$ that there exists a function $f : \times_{k=1}^{m} \mathbb{A}_k \to [0, 1]$, where $f$ may possibly depend on $x$, such that

$$P(Y_i = 1|X = x) = f(x_i). \tag{1}$$

This means that choice of individual $i$ does not depend on the index of individual $i$, so $P(Y_i = 1|X_i = x_i) = P(Y_{i''} = 1|X_{i''} = x_i)$. In particular, outcomes are identically distributed among the individuals who have the same attributes.

Our assumptions then imply that $f = f(z)$ is either strictly monotone increasing in $z_j$, independent of $z_j$, or strictly monotone decreasing in $z_j$.[3]

### 3.1.3 Single Index Models

Our framework includes semiparametric single index choice models with monotone link-functions, where $Y_i = g(X_i\beta) + \varepsilon_i$ with $\mathcal{Y} = \{0, 1\}$ and $E(\varepsilon_i|X_i) = 0$, so $P(Y_i = 1|X_i = x_i) = g(x_i\beta)$, where the link-function $g$ is assumed to be monotone. Our framework includes single index models with binary outcomes as defined by Ichimura (1987, 1993), where $Y_i = \varphi(v(x_i, \beta)) + \varepsilon_i$ where $(x_i, \varepsilon_i)$ for $i = 1, .., n$ are i.i.d., $P(\varepsilon_i \leq \gamma|x) = P(\varepsilon_i \leq \gamma|v(x_i, \beta))$ for all $\gamma$, $E(\varepsilon_i|x) = 0$, $v$ is known and $\beta$ and $\varphi$ are unknown. In contrast to single index models, note that we allow for correlation of attributes $x_i$ across observations.

### 3.1.4 Latent Linear Regression

This framework can be used to model choice based on a latent linear regression model. Consider first the case where there are finitely many different choices, so we can set $\mathcal{Y} = \{1, .., h\}$ for some $h \in \mathbb{N}$. Then there are coefficients $\beta_1, .., \beta_m$, i.i.d. errors $\{\varepsilon_i\}_{i=1}^{n}$ and cutoffs $0 = c_0 \leq c_1 \leq c_2 \leq ... \leq c_h$ such that $P(Y_i = l|X_i = x_i) = P(c_{l-1} < x_i\beta + \varepsilon_i \leq c_l|X_i = x_i)$. When $h = 2$ this includes logit and probit, for $h \geq 3$ it includes ordered logit and ordered probit. Strict monotonicity in attribute $j$ means here that $\beta_j > 0$. We do not make any further assumptions on the errors or the

[3] $f$ is monotone increasing in $z_j$ if $f(z'_j) \geq f(z_j)$ holds for all $z'_j > z_j$, it is strictly monotone increasing if if $f(z'_j) > f(z_j)$ holds for all $z'_j > z_j$. Monotone decreasing and strictly monotone decreasing are defined analogously.

cutoffs. Similarly one can model choices from a continuum. Let $\mathcal{Y} \subseteq \mathbb{R}$, let $c$ be a weakly increasing function from $\mathcal{Y}$ to $\mathbb{R}$ and assume that $P(Y_i \leq \bar{y}|X_i = x_i) = P(x_i\beta + \varepsilon_i \leq c(\bar{y})|X_i = x_i)$ hold for all $\bar{y}$. In both cases, (i), (ii) and (iii) implies $\beta_j > 0$, $\beta_j = 0$ and $\beta_j < 0$ respectively.

### 3.1.5 Duration and Survival Analysis

One can also use this model for a nonparametric duration or survival analysis with right censoring. Let $T_i$ be the units of time until a given event occurs for subject $i$, let $C_i$ be the time the study ended for subject $i$ or the time at which subject $i$ dropped out of the study, measured from the time subject $i$ entered the study. Then the units of time $Y_i$ recorded for subject $i$ are given by $Y_i = \min\{C_i, T_i\}$. One can then test the null hypothesis of a decreasing relationship of attribute $j$ on the duration $T$ if one assumes that $C_i$ and $C_r$ are *exchangeable* whenever $x_{ik} = x_{rk}$ for all $k \neq j$.[4] One proceeds as if interested in the relationship on $Y$ and considers as null hypothesis a monotone decreasing relationship of attribute $j$ on $Y$, testing $H_0$ : "(ii) or (iii)"[5]. This results in an exact test as $P(T_r > T_i|X = x) \leq P(T_r < T_i|X = x)$ implies $P(Y_r > Y_i|X = x) \leq P(Y_r < Y_i|X = x)$.[6]

It is as if we are testing $H_0 : \beta_j \geq 0$ in the Cox proportional hazard model (Cox, 1972), namely that the hazard function, the probability of dying shortly after time $t$ conditional on surviving up to time $t$, is increasing in attribute $j$. However, we are not making the associated parametric assumptions of the Cox proportional hazard model.

More generally, $P(T_r > T_i|X = x) \leq P(T_r < T_i|X = x)$ implies

$$P(T_r > T_i|X = x) - P(T_r < T_i|X = x) \leq P(Y_r > Y_i|X = x) - P(Y_r < Y_i|X = x) \leq 0.$$

Note that the first inequality above will typically be strict, hence one cannot generally establish the size of the effect on $T$, as we do in Section 4 for the other applications, when only observing $Y$.[7]

[4] For instance, if $C_i$ and $C_r$ are independent and identically distributed then they are exchangeable.

[5] Note that the censoring dummy should not be included in the list of attributes.

[6] To see why, consider $i$ and $r$ with $x_{ik} = x_{rk}$ for $k \neq j$. Assume that $y_r > y_i$, but that $t_r \leq t_i$. This means that $y_i = c_i < \min\{c_r, t_r\}$. Assumptions on $C_i$ and $C_r$ imply that it is equally likely that $Y_i = \bar{y}_i = \min\{t_i, c_r\}$ and $Y_r = \bar{y}_r = \min\{t_r, c_i\}$ in which case $\bar{y}_i > \bar{y}_r$. Thus, each wrong measurement of the effect is offset by an equally likely correct measurement.

[7] One way out is to test $H_0 : P(T_r > T_i|X = x) \leq P(T_r \leq T_i|X = x) + d$ instead of $H_0 : P(T_r > T_i|X = x) \leq P(T_r < T_i|X = x) + d$.

## 3.2 Testing for the Direction of Monotonicity

In the following we present an exact test for uncovering significant evidence for whether a monotone relationship between the outcome and the $j$-th attribute is strictly decreasing or strictly increasing. "Exact" means that the probability of a type I error is bounded above by the specified nominal level $\alpha$ for the given sample. In particular the test will be "conditional" in the sense that its probability of type I error is bounded above by $\alpha$ for each realization $x$ of $X$. It is as if $x$ is known, for simplicity we write $P(Y_i = y|x_i)$ instead of $P(Y_i = y|X_i = x_i)$.

We design an exact level $\alpha$ test of $H_0$ : "(ii) or (iii)" against $H_1$ : "(i)". In the special case of a latent linear regression, we are testing $H_0$ : "$\beta_j \leq 0$" against $H_1$ : "$\beta_j > 0$". An exact test of $H_0$ : "(i) or (ii)" against $H_1$ : "(iii)" can then be derived by reversing the order in $\mathcal{Y}$. An exact equi-tailed level $\alpha$ test for $H_0$ : "(ii)" against $H_1$ : "(i) or (iii)" can then be derived by combining these two tests, assigning level $\alpha/2$ to each.

Formally a test is a mapping $\phi : \times_{k=1}^{m} \mathbb{A}_k^n \times \{0,1\}^n \to \{0,1\}$ with the interpretation that the null hypothesis is rejected if and only if $\phi(x,y) = 1$. The test $\phi$ is an exact level $\alpha$ test if $E(\phi = 1|H_0) \leq \alpha$. In the construction of our test we will be using randomized tests, here $\phi \in [0,1]$ is such that $\phi$ is the probability of rejecting the null hypothesis. The test has *size* $\alpha$ if there is no $\bar{\alpha} < \alpha$ such that the test has level $\bar{\alpha}$.[8]

We do not make any assumptions on the function $f$ beyond monotonicity. This generality comes at a price. Our test will only be useful for uncovering the alternative hypothesis if there are sufficiently many individuals that are identical to each other apart from the attribute $j$ to be tested. This assumption is not satisfied if one of the attributes $k \neq j$ is drawn from a continuous distribution. Remedies for this case will be discussed later. The bound on the type II error probability as provided below can be used to quantify, before observing the outcome of $y$, what "sufficiently many" means.

### 3.2.1 Intuition Behind the Test

We first present some intuition on how the test of $H_0$ : "decreasing relationship between outcome and attribute $j$" is constructed. As we do not know how $Y$ depends on the other attributes we isolate the dependence on attribute $j$ by matching individuals who have equal attributes apart from attribute $j$ into pairs. Within each pair assign the individual who has a higher value of attribute $j$ to position 1, the other gets position 2. For instance, assume that individuals $i$ and $l$ have been paired where $x_{i,j} > x_{l,j}$. Then $i$ gets position 1 and $l$ gets position 2. Under the null hypothesis,

[8] In many settings the size of an exact test will fall below its nominal level $\alpha$, without leaving any possibility to determine analytically how much lower it is.

$(Y_i, Y_l)$ is weakly less likely to be equal $(1,0)$ than to be equal to $(0,1)$. Assign code 1 to each pair with outcomes $(1,0)$ and code 0 to each pair with outcomes $(0,1)$. Drop all pairs in which both individuals received the same outcome, analogous to McNemar's (1947) test. Note that under the null hypothesis, code 1 is less likely to occur than code 0. We then use the binomial test to test whether there are less pairs with code 1 than pairs with code 0. A rejection provides evidence that code 1 is more likely than code 0 within the entire data set, hence significant evidence that there is no decreasing relationship which means by assumption (b) that there is a strictly increasing relationship.

The only problem with the above test is that it has a random element. The total number of pairs coded with one and zero will typically depend on how individuals are paired. To eliminate randomness we evaluate the binomial test at a level $\theta\alpha$, thereby reduce the probability of rejection, and then reject the null hypothesis if the probability of rejection is above $\theta$. This ensures that the test has level $\alpha$. $\theta$ is a free parameter that should be chosen to minimize the bound on the type II error probability of the test as specified in Section 3.2.3. Note that this adjustment is not needed when no two individuals have identical attributes, in this case the pairing of individuals is unique.

### 3.2.2 Constructing an Exact Test

We now construct the exact test, denoted by $\phi^*$, that we call the *direction of monotonicity test*. It will be invariant to any transformation of the attributes that does not change the order of the $j$-th attribute and that keeps distinct values distinct.[9] It will also be invariant to how individuals are indexed.[10]

Our test has one free parameter $\theta$ which we later show how to select. Our construction utilizes randomized tests. Randomized tests are mappings $\phi : \times_{k=1}^{m} \mathbb{A}_k^n \times \{0,1\}^n \to [0,1]$ where $\phi(x,y)$ is the probability of rejecting the null hypothesis. A test $\phi$ is called *nonrandomized* if $\phi(x,y) \in \{0,1\}$ for all $(x,y)$.

Our test utilizes the one-sided randomized binomial test with size $\alpha$, denoted by $\phi_{Br}(c,n,\alpha,\lambda)$, which tests $H_0$ : "$P(Z=1) \leq \lambda$" against $H_1$ : "$P(Z=1) > \lambda$" based on $n$ independent realizations of a Bernoulli random variable $Z$ where $c$ are the number of successes. So given

$$B(c,n,p) = \sum_{k=c}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

[9] Let $(g_k)_{k=1}^m$ be such that $g_k : \mathbb{A}_k \to \mathbb{A}_k$ where $g_k(a_k) = g_k(b_k)$ implies $a_k = b_k$ for $k \neq j$ and $g_j(a_j) \succ_j g_j(b_j)$ holds if and only if $a_j \succ_j b_j$ where $\succ_j$ is the strict order on $\mathbb{A}_j$. Let $z \in \times \mathbb{A}_k^n$ be such that $z_{i,r} = g_r(x_{i,r})$ for all $i, r$. Then $\phi^*(x,y) = \phi^*(z,y)$ for all $y \in \mathbb{R}^n$.

[10] Consider any permutation $\pi$ of $\{1,..,n\}$ such that $x_{i,k} = x_{\pi(i),k}$ for all $i = 1,..,n$ and $k = 1,..,m$. Then $\phi(x,y) = \phi(x_\pi, y_\pi)$ where $x_\pi = \left(x_{\pi(i),k}\right)_{i=1,..,n,k=1,..,m}$ and $y_\pi = \left(y_{\pi(i)}\right)_{i=1,..,n}$.

$\phi_{Br}$ is defined by

$$\phi_{Br}(c,n,p,\alpha)=\begin{cases}1 & \text{if } B(c,n,p)\leq\alpha\\ \frac{\alpha-B(c+1,n,p)}{B(c,n,p)-B(c+1,n,p)} & \text{if } B(c+1,n,p)\leq\alpha<B(c,n,p)\\ 0 & \text{if } B(c+1,n,p)>\alpha\end{cases}.$$

The following four steps define the direction of monotonicity test.

1. Assign individuals into maximal subsets of $\{1,..,n\}$, which we call blocks, so that all individuals belonging to the same block have the same values for all attributes $k\neq j$. So if $i$ and $h$ belong to the same block then $x_{i,k}=x_{h,k}$ for all $k\neq j$. Consider only those blocks that have at least two elements, give them indexes $s=1,..,S$, so $S$ is the total number of such blocks.

2. For each block that contains at least two elements, let $s$ be the index, do the following:

   Drop the individual that has the median value of the $j$-th attribute if the number of individuals in this block is odd. Thus we can assume that there are $2l_s$ individuals in this block for some $l_s\in\mathbb{N}$. Order individuals within this block from lowest to highest value of the $j$-th attribute, allocating places in this order at random (with equal probability) whenever different individuals have the same value of the $j$-th attribute (and hence the same values of all attributes). Let $o(r,s)$ be the index of the individual in the $r$-th position in this order within the block with index $s$, so $r=1,..,2l_s$ and $x_{o(t,s),j}\geq x_{o(r,s),j}$ for $t>r$. Let $O(r,s)$ be the underlying random variable that has been created by this random ordering.

3. Let $k_1$ and $k_2$ be the number of pairs $(r,s)$, $r=1,..,l_s$, $s=1,..,S$, in which $\left(x_{o(r+l_s,s),j}-x_{o(r,s),j}\right)\left(y_{o(r+l_s,s)}-y_{o(r,s)}\right)>0$ and $\left(x_{o(r+l_s,s),j}-x_{o(r,s),j}\right)\left(y_{o(r+l_s,s)}-y_{o(r,s)}\right)<0$ respectively.

4. Let $q(x,y,o)=\phi_{Br}(k_1,k_1+k_2,1/2,\theta\alpha)$.

5. Determine the expected value, denoted by $Eq(x,y,O|x,y)$, of $q(x,y,o)$ over all random orderings by repeating steps 2-4 infinitely often. So $Eq(x,y,O|x,y)=\int q(x,y,o)\,dP(o|x,y)$.

6. Let $\phi^*(x,y)=1$ if $Eq(x,y,O|x,y)\geq\theta$ and $\phi^*(x,y)=0$ if $Eq(x,y,O|x,y)<\theta$, which defines the *direction of monotonicity test.*

Note that in practice one cannot repeat steps $2-4$ infinitely often as specified in step 5. Instead, it is useful to repeat them a large number of times and to use Hoeffding's (1963) inequality to bound the deviation of the mean from the expected value when comparing the mean to $\theta$ in step 6.

**Proposition 1** *For any given $\theta \in (0,1)$ the direction of monotonicity test has a type I error probability bounded above by $\alpha$.*

**Proof.** See appendix. ■

Note that any alternative method of pairing the data within blocks that does not depend on $y$ will similarly lead to an exact test. We have chosen to pair the individual with the $r$-th highest value of $x_{*j}$ with the $(r+l_s)$-th highest value of $x_{*j}$. This maximizes the minimal difference between $x_{ij}$ ad $x_{hj}$ provided $l_s$ pairs are formed. Larger differences in the $j$-th attribute generate larger differences in the probability of choosing option $A$ and hence given monotonicity lead more likely to rejecting the null hypothesis when it is false. The choice to maximize the minimal difference has shown to yield good results in applications. A more formal analysis of how to best pair the data is left for future research. Note that the randomness in the pairing, captured by $O$, is only used to make the procedure independent of how individuals are indexed. All elements in the support of $O$ have the same number of pairs in each each block.

### 3.2.3 Type II Error Probability Bounds for Binary-Valued Data

We derive bounds on the type II error probability of our test when $\mathcal{Y} = \{0,1\}$. We wish to derive type II error probability bound in a way that can capture heterogeneity of effects between different blocks. At least two forms of heterogeneity arise. First of all, the attribute may have a different effect in different pairs and in different blocks. Formally, $P\left(Y_{o(r+l_s,s)}=1|X=x\right) - P\left(Y_{o(r,s)}=1|X=x\right)$ may differ across pairs. Second of all, even if the difference in success probability is the same in two pairs, the absolute level of success probability may differ. Formally, for any two pairs $P\left(Y_{o(r,s)}=1|X=x\right)$ can be different even if $P\left(Y_{o(r+l_s,s)}=1|X=x\right) - P\left(Y_{o(r,s)}=1|X=x\right)$ is the same. In view of these issues, given analytic and computational limitations, as well as in view of generating a simple formula, we take the following approach.

Let $N$ be the number of pairs in which the attribute of interest takes different values, so $N = \left|\left\{(r,s) : x_{o(r+l_s,s),j} > x_{o(r,s),j}\right\}\right|$. We make the following assumption on the data generating process (that determines $Y$ given $X$) when computing the type II error probability. Consider $\delta \in (0,1)$ and $N_1 \in \mathbb{N}$ such that $\delta N \leq N_1 \leq N$. There are $N - N_1$ pairs in which the outcome is necessarily the same, so where $P\left(Y_{o(r,s)} = Y_{o(r+l_s,s)}|X=x\right) = 1$. The outcomes in each of the remaining pairs are identically distributed between pairs (and independent within pairs). So there is some $\mu \in [0, 1 - N\delta/N_1]$ such that $P\left(Y_{o(r,s)}=1|X=x\right) = \mu$ and $P\left(Y_{o(r+l_s,s)}=1|X=x\right) = \mu + N\delta/N_1$ (with $Y_{o(r,s)}$ and $Y_{o(r+l_s,s)}$ independent) hold for all these $N_1$ pairs. Note that the rescaling of the effect within these pairs ensures that the overall effect is equal to $\delta$. We write $Y \in S(\delta, N_1)$ if the data generating process satisfies these constraints.

Let $p = (\mu + \chi)(1 - \mu)$ and $q = (1 - \mu - \chi)\mu$. Given

$$f_{B\mu}(\mu, \chi, \theta, n) = \sum_{j=0}^{n} \sum_{k=0}^{n-j} \frac{n!}{j!k!(n-j-k)!} p^j q^k (1-p-q)^{n-j-k} \phi_{Br}(j, j+k, 1/2, \theta\alpha)$$

let

$$f_B(\chi, \theta, n) = \min_{0 \leq \mu \leq 1 - N\delta/N_1} f_{B\mu}(\mu, \chi, \theta, n).$$

**Proposition 2** *The type II error probability of the direction of monotonicity test for* $\mathcal{Y} = \{0, 1\}$ *is bounded above by* $(1 - f_B(N\delta/N_1, \theta, N_1))/(1 - \theta)$ *when* $Y \in S(\delta, N_1)$.

**Proof.** See appendix. ■

## 3.3 Type II Error Probability Bounds for General Outcome Spaces

Next we present a bound when there are more than two possible outcomes. We proceed as above and assume that there are $N - N_1$ pairs in which the outcome is the same, so where $P\left(Y_{o(r,s)} = Y_{o(r+l_s,s)} | X = x\right) = 1$. For the remaining $N_1$ pairs we assume that the effect is identical. So there is some $\mu \in [0, 1 - N\delta/N_1]$ such that $P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)} | X = x\right) = \mu$ and $P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)} | X = x\right) = \mu + N\delta/N_1$ holds for all these $N_1$ pairs. The set of data generating processes with this property will be denoted by $S_1(\delta, N_1)$. Let

$$f_G(\eta, n, \theta) = \sum_{j=0}^{n} \binom{n}{j} \eta^j (1-\eta)^{n-j} \phi_{Br}(j, n, 1/2, \theta\alpha).$$

**Proposition 3** *The type II error probability of the direction of monotonicity test for a general outcome space* $\mathcal{Y}$ *is bounded above by* $(1 - f_G((1 + N\delta/N_1)/2, N_1, \theta))/(1 - \theta)$ *when* $Y \in S_1(\delta, N_1)$.

**Proof.** See appendix. ■

The bound above is derived by letting nature choose $\tau_1$ and $\tau_2$ with $\tau_1 - \tau_2 = N\delta/N_1$ such that $P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)} | X = x\right) = \tau_1$ and $P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)} | X = x\right) = \tau_2$ holds in the $N_1$ pairs. The proof shows that nature will choose $\tau$ such that $\tau_1 + \tau_2 = 1$, so in the worst case there will not be a tie in the dependent outcome. However this bound is only very coarse as it acts as if the two outcomes within each pair can be dependent, in fact in the worst case for this bound they very dependent as $P\left(Y_{o(r+l_s,s)} = Y_{o(r,s)} | X = x\right) = 0$. Note that in the case of binary outcomes we were able to incorporate the independence.

# 4 Testing for the Effect

In the following we wish to understand the magnitude of the influence of attribute $j$ on the outcome. At the same time we will drop assumption (b) (see Section 3) that postulates a monotone relationship between the attribute of interest and the outcome.

## 4.1 Measuring the Effect

We wish to measure how changes in the attribute of interest $j$ influence the outcome for the given sample of attributes. As the measure will not depend on how the attributes were generated, correlation among the attributes between individuals is allowed as in the previous framework. We call our measure the *average incremental effect*, short AIE. It is set equal to the expected change in the outcome due to differences in the attribute of interest conditional on selecting two individuals from the original sample that satisfy the following three conditions. (i) They have different values of the attribute of interest. (ii) They have the same values of all other attributes. (iii) Up to integer constraints, half the individuals who have the same values of all other attributes have a value of the attribute of interest between the two of them. Note that the integer constraints arise when the block has an odd number of members, in which case the median is dropped. When the attribute of interest describes whether or not the individual has been treated, so $x_{ij} \in \{0,1\}$, one can also refer to AIE as the *average conditional treatment effect*. Note that we restrict our comparison to individuals who have different values of the attribute of interest, as AIE should be capturing the change in the outcome induced by changing the attribute of interest, and should not be directly influenced by how many individuals have the same value of the attribute of interest.

AIE measures the expected effect in a random pair $(o(r,s), o(r+l_s,s))$ that is drawn from some random ordering $o$ constructed in step 2, where the effect within this pair is measured by

$$P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)} | X = x\right) - P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)} | X = x\right), \tag{2}$$

note that $x_{o(r+l_s,s),j} > x_{o(r,s),j}$, for $r = 1,..,l_s, s = 1,..,S$.

Formally, the average incremental effect, short AIE, is defined as follows. Consider some ordering $o$ as constructed in step 2. Let
$N = \left|\left\{(r,s) : x_{o(r+l_s,s),j} > x_{o(r,s),j}, r = 1,..,l_s, s = 1,..,S\right\}\right|$ be the number of pairs in which $x_{o(r+l_s,s),j} > x_{o(r,s),j}$ (which by construction holds if and only if $x_{o(r+l_s,s),j} \neq x_{o(r,s),j}$). Note that $N$ only depends on $x$ and hence is nonrandom conditional on $x$. Let $\mathcal{O}$ be the set of all orderings that may result from the procedure described in step

2. For $N > 0$ the AIE, denoted by $\delta_j(x)$, is given by

$$\delta_j(x) = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r\in\{1,..,l_s\}:x_{o(r+l_s,s),j}>x_{o(r,s),j}}\begin{pmatrix} P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)}|X=x\right) \\ -P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)}|X=x\right)\end{pmatrix}. \tag{3}$$

If $N = 0$ then $\delta_j(x) := 0$.

It follows that $\delta_j(x) \in [-1,1]$. If $\delta_j(x) = 0$ then attribute $j$ has on average no impact on the outcome $Y$ for those individuals within the sample. If we were to consider the nonparametric ordinal regression model of Section 3 then the cases (i), (ii) and (iii) would be identified by $\delta_j(x) < 0$, $\delta_j(x) = 0$ and $\delta_j(x) > 0$ respectively.

If $N > 0$ then an unbiased estimate of AIE is given by

$$\hat{\delta}_j(x) = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r\in\{1,..,l_s\}:x_{o(r+l_s,s),j}>x_{o(r,s),j}}\left(1\left\{y_{o(r+l_s,s)} > y_{o(r,s)}\right\} - 1\left\{y_{o(r+l_s,s)} < y_{o(r,s)}\right\}\right). \tag{4}$$

## 4.2 Hypotheses

For each $\bar{\delta}_j \in (-1,1)$ we present an exact test for $H_0$ : "$\delta_j(x) \leq \bar{\delta}_j$" against $H_1$ : "$\delta_j(x) > \bar{\delta}_j$". A test for $H_0$ : "$\delta_j(x) \geq \bar{\delta}_j$" against $H_1$ : "$\delta_j(x) < \bar{\delta}_j$" can be constructed analogously. Combining these two tests, each with level $\alpha/2$, one can construct for each $\bar{\delta}_j \in (-1,1)$ an exact equi-tailed test for $H_0$ : "$\delta_j(x) = \bar{\delta}_j$" against $H_1$ : "$\delta_j(x) \neq \bar{\delta}_j$" that has level $\alpha$. Tests for $H_0$ : "$\delta_j(x) = -1$" and for $H_0$ : "$\delta_j(x) = 1$" are easily constructed.[11] A $(1-\alpha)\cdot 100\%$ confidence interval of $\delta_j(x)$ conditional on $x$ then results by collecting all values of $\bar{\delta}_j$ where $H_0$ : "$\delta_j(x) = \bar{\delta}_j$" cannot be rejected at level $\alpha$.

## 4.3 Testing for the Effect with Binary Outcomes

Assume that $\mathcal{Y} = \{0,1\}$.

### 4.3.1 Additional Insights for Binary Outcomes

Since the outcome of interest is binary valued and realized independently across individuals (conditional on $x$), there is a cardinal and an ordinal interpretation of AIE.

[11] An exact level $\alpha$ test of $H_0 : \delta_j = -1$ against $H_1 : \delta_j = 1$ for any $\alpha \in (0,1)$ is defined by rejecting the null hypothesis if and only if there are two individuals $i$ and $i'$ with $x_{i,j} > x_{i',j}$, $x_{i,k} = x_{i',k}$ for all $k \neq j$ where $y_i = 1$ and $y_{i'} = 0$.

This is because

$$\begin{aligned} & E\left(Y_{o(r+l_s,s)}|X=x\right)-E\left(Y_{o(r,s)}|X=x\right) \\ = \; & P\left(Y_{o(r+l_s,s)}=1|X=x\right)-P\left(Y_{o(r,s)}=1|X=x\right) \\ = \; & P\left(Y_{o(r+l_s,s)}\geq Y_{o(r,s)}|X=x\right)-P\left(Y_{o(r+l_s,s)}\leq Y_{o(r,s)}|X=x\right) \\ = \; & P\left(Y_{o(r+l_s,s)}>Y_{o(r,s)}|X=x\right)-P\left(Y_{o(r+l_s,s)}<Y_{o(r,s)}|X=x\right). \end{aligned}$$

For $N>0$ we obtain

$$\begin{aligned} \delta_j(x) \; = \; & \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r\in\{1,..,l_s\}:x_{o(r+l_s,s),j}>x_{o(r,s),j}} P\left(Y_{o(r+l_s,s)}=1|X=x\right) \\ & -\frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r\in\{1,..,l_s\}:x_{o(r+l_s,s),j}>x_{o(r,s),j}} P\left(Y_{o(r,s)}=1|X=x\right) \end{aligned} \tag{5}$$

In other words, the object of interest is the difference in the mean success probability between those with attribute of interest above the median and those below the median among those individuals with the same other attributes.

If $N>0$ then an unbiased estimate of AIE is given by

$$\hat{\delta}_j(x)=\frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r\in\{1,..,l_s\}:x_{o(r+l_s,s),j}>x_{o(r,s),j}}\left(y_{o(r+l_s,s)}-y_{o(r,s)}\right).$$

In the special case where the outcome probability (see (1)) is linear in the $j$th attribute, so where $f(z)=\beta_j z_j+f_1(z_{-j})$, we obtain

$$\delta_j(x)=\beta_j\frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}\frac{1}{N}\sum_{s=1}^{S}\sum_{r=1}^{l_s}\left(x_{o(r+l_s,s),j}-x_{o(r,s),j}\right).$$

In this spirit, one may choose to rescale the AIE and to test for the *relative incremental effect, short RIE*, defined by

$$\rho_j(x)=\frac{\delta_j(x)}{\frac{1}{N}\sum_{s=1}^{S}\sum_{r=1}^{l_s}\left(x_{o(r+l_s,s),j}-x_{o(r,s),j}\right)},$$

and estimated by

$$\hat{\rho}_j(x)=\frac{\hat{\delta}_j(x)}{\frac{1}{N}\sum_{s=1}^{S}\sum_{r=1}^{l_s}\left(x_{o(r+l_s,s),j}-x_{o(r,s),j}\right)}$$

where $o$ is some element of $\mathcal{O}$. Consequently, we obtain for a linear probability model in which $P(Y_i=1|X=x)=X\beta$ that $\rho_j(x)=\beta_j$. In this special case $\hat{\rho}_j$ is an unbiased estimate of $\beta_j$ which is linear in the outcome variables and hence has larger variance than the OLS estimate.

Note that if attribute $j$ is binary valued then $\rho_j(x)=\delta_j(x)$ and $\hat{\rho}_j(x)=\hat{\delta}_j(x)$.

### 4.3.2 Intuition Behind the Test

We provide some intuition for how we construct a test. Consider first $\bar{\delta}_j = 0$. A first thought could be to proceed as in the direction of monotonicity test and to drop the pairs with ties in the outcomes (in which $y_{o(r+l_s,s)} = y_{o(r,s)}$) and then to check the average effect among on the remaining pairs. However note that the average effect can turn positive when conditioning on a given number of dropped pairs even if its negative in the entire sample. To see this, assume that the overall effect is zero, that there is a block in which the effect is positive for each of its $n_1$ pairs, that ties never occur within this block while for all other $N - n_1$ pairs ties occur with positive probability. Then conditional on $n_1$ pairs being without ties we are facing a block with a positive effect. Hence we can no longer separate inference according to the number of ties and need to construct our test differently.

We construct a test by comparing the mean outcome above and below the median in each block. Following (5), under the null hypothesis the difference between the average success probability in pairs above and below the median is at most $\bar{\delta}_j$. We therefore need a test for comparing two samples of Bernoulli random variables where success probabilities within the same sample need not be identically distributed. We design such a test. The test statistic is given by the difference in the number of successes of the two samples. Using Gleser (1975) who relies on Hoeffding (1956) we show that the rejection probability is maximized when each sample is i.i.d. provided the difference in the number of successes is at least $\bar{\delta}_j N + 2$. This reduces the problem to one in which each sample is i.i.d. which thus allows us to compute the cutoff.

Applying this test to our ordinal regression, we reject the null hypothesis if there are sufficiently many more successes above the median than below the median when summed over all blocks. When $\bar{\delta}_j = 0$ then we are looking at the difference between the number of times $Y_{r,s} = (0,1)$ and $Y_{r,s} = (1,0)$ occurred, $Y_{r,s} = \left(Y_{o(r,s)}, Y_{o(r+l_s,s)}\right)$.

### 4.3.3 An Exact Test for the Average Incremental Effect

For $k \in \mathbb{I}$, $n \in \mathbb{N}$ and $\delta \in (-1,1)$ let

$$D(k,n,\delta) = \max_{\max\{0,-\delta\} \leq p \leq \min\{1-\delta,1\}} \left\{ \sum_{i=\max\{k,0\}}^{n} \binom{n}{i} (p+\delta)^i (1-p-\delta)^{n-i} \sum_{j=0}^{\min\{i-k,n\}} \binom{n}{j} p^j (1-p)^{n-j} \right\}$$

$$\phi_{Dr}(k,n,\delta,\alpha) = \begin{cases} 1 & \text{if } D(k,n,\delta) \leq \alpha \text{ and } k \geq \delta n + 2 \\ \frac{\alpha - D(k+1,n)}{D(k,n) - D(k+1,n)} & \text{if } D(k+1,n,\delta) < \alpha < D(k,n,\delta) \text{ and } k \geq \delta n + 2 \\ 0 & \text{if } D(k+1,n,\delta) \geq \alpha \text{ or } k < \delta n + 2 \end{cases}.$$

We construct a test of $H_0 : \delta_j(x) \leq \bar{\delta}_j$ for $\bar{\delta}_j \in (-1,1)$, that we call the *difference test*, using steps 1-6 in Section 3.2.2, however replacing step 4 by step 4':

4' Let $q(x, y, o) = \phi_{Dr}\left(k_1 - k_2, N, \theta\alpha, \bar{\delta}_j\right)$ where $N$ is the number of pairs $(r, s)$ in which $x_{o(r+l_s,s),j} > x_{o(r,s),j}$.

**Proposition 4** *The difference test is an exact level $\alpha$ test of $H_0$ : "$\delta_j(x) \leq \bar{\delta}_j$" against $H_1$ : "$\delta_j(x) > \bar{\delta}_j$" conditional on $X = x$.*

**Proof.** See appendix. ■

### 4.3.4 Power Bounds

We derive bounds on the type II error probability of the tests above. These are useful for determining the free parameter $\theta$, for comparing performance with other tests and for calculating sample sizes. For $\delta_j^{(1)} > \bar{\delta}_j$ we wish to derive bounds on the type II error probability for our test of $H_0$ : "$\delta_j(x) \leq \bar{\delta}_j$" when $\delta_j(x) \geq \delta_j^a$. Again we wish to apply Gleser (1975), hence have to ensure that the difference under the alternative is sufficiently above the cutoff used in the test.

Let $\bar{k} = \bar{k}\left(N, \theta\alpha, \bar{\delta}_j\right)$ be the smallest value of $k \in \mathbb{I}$ such that $\phi_{Dr}\left(k+1, N, \theta\alpha, \bar{\delta}_j\right) = 1$. Let

$$f_{D\mu}(\mu, \delta, \theta) = \sum_{j=0}^{N} \binom{N}{j} \mu^j (1-\mu)^{N-j} \left( \sum_{r=0}^{N} \binom{N}{r} (\mu+\delta)^r (1-\mu-\delta)^{N-r} \phi_{Dr}\left(r-j, N, \theta\alpha, \bar{\delta}_j\right) \right)$$

and let

$$f_D(\delta, \theta) = \min_{\max\{0,-\delta\} \leq \mu \leq \min\{1-\delta,1\}} f_{D\mu}(\mu, \delta, \theta).$$

**Proposition 5** *If $\bar{k} \leq \delta_j^a N - 2$ then the type II error probability is bounded above by*

$$\frac{1 - f_D\left(\delta_j^a, \theta\right)}{1 - \theta}$$

*when $\delta_j(x) \geq \delta_j^a$.*

**Proof.** See appendix. ■

## 4.4 Testing for the Effect with General Outcome Spaces

We now consider the case where $\mathcal{Y}$ contains more than two outcomes. As we cannot interpret the AIE as a difference between outcomes above and below the median we proceed similarly as we did in the direction of monotonicity test. We interpret the AIE as an expected value of an average of independent random variables. For each pair $(r, s)$ formed in step 2 let $z_{r,s} = 1, 0, -1$ if $y_{r+l_s,s} > y_{r,s}$, $y_{r+l_s,s} = y_{r,s}$ and $y_{r+l_s,s} < y_{r,s}$ respectively. Then

$$\hat{\delta}_j = \frac{1}{N} \sum_{s=1}^{S} \sum_{r \in \{1,..,l_s\}: x_{o(r+l_s,s),j} > x_{o(r,s),j}} E(z_{r,s})$$

where the expectation takes into account the different orderings of the data in step 2. Thus we are testing the expectation of the average of $N$ independent but not identically distributed random variables, where $N = \left|\left\{(r,s) : x_{o(r+l_s,s),j} > x_{o(r,s),j}\right\}\right|$.

As explained in the first paragraph of Section 4.3.2 we cannot drop pairs in which the outcomes are equal. Thus we need a test for evaluating the average of $N$ independent but not necessarily identically distributed random variables. We construct such a test by first creating a sequence of $N$ independently distributed Bernoulli random variables with the same expected value and then apply results of Hoeffding (1956) that identify upper bounds.

We construct a test of $H_0 : \delta_j(x) \leq \bar{\delta}_j$ for $\bar{\delta}_j \in (-1,1)$, that we call the *generalized AIE test*, using steps 1-6 in Section 3.2.2, however replacing step 3 by step 3" and step 4 by step 4":

3" For each pair $(r,s)$ in which $x_{o(r+l_s,s),j} > x_{o(r,s),j}$,
let $\bar{z}_{r,s} = 1$ if $y_{o(r+l_s,s)} > y_{o(r,s)}$,
let $\bar{z}_{r,s} = -1$ if $y_{o(r+l_s,s)} < y_{o(r,s)}$,
let $\bar{z}_{r,s} = 1$ with probability $1/2$ and $\bar{w}_{r,s} = -1$ with probability $1/2$ if $y_{o(r+l_s,s)} = y_{o(r,s)}$[12],
and then let $k_1$ be the number of pairs in which $x_{o(r+l_s,s),j} > x_{o(r,s),j}$ and $\bar{z}_{r,s} = 1$ and let $N$ be the total number of pairs in which $x_{o(r+l_s,s),j} > x_{o(r,s),j}$.

4" Let $q(x,y,o) = \phi_{Qr}\left(\bar{k}_1, N, \bar{\delta}_j, \theta\alpha\right)$ where $\phi_{Qr}(c,n,\delta,\alpha) = \phi_{Br}(c,n,(1+\delta)/2,\alpha)$ if $c \geq np+1$ and $\phi_{Qr}(c,n,\delta,\alpha) = 0$ otherwise.[13]

Note that

$$\hat{\delta}_j = \frac{1}{N}\sum_{s=1}^{S} \sum_{r\in\{1,..,l_s\}: x_{o(r+l_s,s),j} > x_{o(r,s),j}} E(\bar{z}_{r,s}).$$

**Proposition 6** *The generalized AIE test is an exact level $\alpha$ test of $H_0$ : "$\delta_j(x) \leq \bar{\delta}_j$" against $H_1$ : "$\delta_j(x) > \bar{\delta}_j$" conditional on $X = x$.*

**Proof.** See appendix. ■

We now turn to bounds on type II error probability. When deriving the bounds for this test it is as if nature can choose

$$P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)} | X = x\right) - P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)} | X = x\right)$$

for each pair separately. Given the randomization when transforming $z$ into $\bar{z}$ it is as if nature chooses a success probability for each pair. If we ensure that the true

[12] Note that this randomization should be undergone independently of any other events.

[13] At the expense of complicating exposition one can easily make the test more powerful by using the explicit bounds in Hoeffding (1956) that allw for $\phi_{Qr} > 0$ when $np \leq c < np+1$.

average success probability is sufficiently large, then we know following Hoeffding (1956) that nature will choose an i.i.d. sample in order to maximize the probability of not rejecting the null. We proceed analogously to Section 4.3.4.

Let $\bar{c} = \bar{c}\left(N, \theta\alpha, \bar{\delta}_j\right)$ be such that smallest value of $c \in \mathbb{N}_0$ such that $\phi_{Qr}\left(c+1, N, \bar{\delta}_j, \theta\alpha\right) = 1$. Let

$$f_Q\left(p, N, \alpha\right) = \sum_{c=0}^{N} \binom{n}{c} p^c \left(1-p\right)^{N-c} \phi_{Br}\left(c, N, \left(1+\bar{\delta}_j\right)/2, \alpha\right).$$

**Proposition 7** *For $\delta_j^a$ such that $\bar{c}+1 \leq \frac{1}{2}N\left(1+\delta_j^a\right)$ the type II error probability is bounded above by*

$$\frac{1 - f_Q\left(\frac{1}{2}\left(1+\delta_j^a\right), N, \theta\alpha\right)}{1-\theta}$$

*when $\delta_j\left(x\right) \geq \delta_j^a$.*

## 4.5 Comparison

We briefly compare the performance of the tests for binary valued outcomes.

First we compare the bounds on the type II error probabilities of the two tests when there is a common lower bound on the effect in each pair. For instance this would follow if one assumed under the alternative hypothesis that the effect is identically distributed in the entire data set. So we are comparing the bound in Proposition 2 with $N_1 = N$ to the bound in Proposition 5. Here the type II error probability bound of the difference test is smaller than that of the direction of monotonicity test as we now demonstrate. The difference test considers two independent samples and compares the number of successes in each sample using differences. In particular, observations are independent within each pair. The monotonicity test acts as if there is correlation of observations within each pair and compares the pairs in which there are two different observations. Treating the data as if it were correlated gives us a hint that the monotonicity test is less powerful. A more formal argument can be made using results in Schlag (2008b). Therein it is shown that the randomized difference test obtains the smallest type II error probability among all tests provided $\delta_j^a$ is not too small. As the method of derandomization is the same for the two tests, this shows that the difference test has a lower type II error probability than any other test that is based on the derandomization method. For instance, if $N = 20$, given $\theta = 0.3$ and $\alpha = 0.05$, then the difference test for $\bar{\delta}_j = 0$ requires $\delta_j^a \geq 0.398$ in order for the type II error probability bound to be below 0.5, for this value the monotonicity test attains 0.56 as bound, it requires $\delta_j^a \geq 0.416$ to guarantee type II error probability bound below 0.5.

Next we perform the comparison in light of more realistic scenarios where the effect differs between blocks, a scenario for which the monotonicity test was designed. Numerical results show that the bound for the monotonicity test is increasing in $N_1$

given $N$. The more concentrated the effect is, the better the monotonicity test in detecting this. In fact, if $N = 20$ ($\theta = 0.3$ and $\alpha = 00.05$) then the bound will be below 0.5 if $N_1 \leq 18$. If instead $N = 50$ then the difference test requires $\delta \leq 0.25$ for type II error probability bound to be below 0.5, here monotonicity test is at 0.52 while once $N_1 \leq 48 = N - 2$ we again find that type II error probability bound is below 0.5.

In view of the above we find that the type II error probability bounds confirm that the monotonicity test is preferable to the difference test when the modeler assumes a monotone relationship in the attribute of interest.

# 5 Data Applications

## 5.1 Use of Fertilizer

In this section we consider the performance of our new test in the data in Duflo et al. (2011, Table 4A). The data comes from a randomized experiment that was designed to test the effectiveness of a program (called SAFI) to induce farmers in Kenya to use fertilizer. Three separate regressions are run, one for each season where seasons range from 1 to 3. Dependent variables are binary valued and describe whether the farmers used fertilizer in the given season (see columns 1, 3 and 5 of Table 4A in Duflo et al., 2011). There are 21 attributes, each attribute is binary valued. We describe the five attributes of interest. "SAFI season 1", "starter kit" and "demo" describe respectively whether the farmer was enrolled in the program in season 1, received a starter kit, and had access to a demonstration plot. "kit and demo" is the interaction between "starter kit" and "demo". The remaining 16 attributes are controls. Farmers were recruited as parents of school children, each of these attributes corresponds to a different school. The number of observations ranges between 756 and 902.

We analyze this data set with our new method and compare our findings to several other methods. Results are shown in Table 1. For each method we present p values associated to two-sided tests of $H_0 : \beta_j = 0$ whenever these are below 0.2, "not" refers to situations where the p value is above 0.2, "not*" refers to situations in which no pairs could be formed as there were no two individuals that had the same values of the attributes not of interest. Each column contains a different method. The first column shows our direction of monotonicity test, the second shows the results for the difference test.[14] The third to fifth column are linear probability models. "GS" refers to the exact method of Gossner and Schlag (2013), "t test" is the test that is derived when assuming homoskedastic errors, "robust" is the test of White (1980) for the case

[14]For the direction of monotonicity test we choose $\theta$ to maximize the range of parameters for the attribute of interest under which the type II error is below 0.5. For the difference test we set $\theta = 0.3$.

of heteroskedastic errors. The last two columns are the two most popular models of probabilistic choice, namely logit and probit. Only the first three columns contain results of exact methods. Interestingly, the p value of the direction of monotonicity test is smaller than that of the exact linear regression method GS even though the underlying model is more general.

| variable | mono | diff | GS | $t$ test | robust | logit | probit |
|---|---|---|---|---|---|---|---|
| | season 1 | | | | | | |
| SAFI season 1 | 0.02 | 0.06 | 0.05 | 0.01 | 0.01 | 0.01 | 0.01 |
| starter kit | not | not | not | 0.15 | 0.14 | 0.14 | 0.14 |
| kit and demo | not* | not* | not | not | not | not | not |
| demo | not* | not* | not | not | not | not | not |
| household | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | season 2 | | | | | | |
| SAFI season 1 | not | not | not | not | not | not | not |
| starter kit | not | not | not | not | not | not | not |
| kit and demo | not* | not* | not | not | not | not | not |
| demo | not* | not* | not | not | 0.01 | not | not |
| household | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | season 3 | | | | | | |
| SAFI season 1 | not | not | not | not | not | not | not |
| starter kit | not | not | not | not | not | not | not |
| kit and demo | not* | not* | not | not | not | not | not |
| demo | not* | not* | not | not | 0.01 | not | not |
| household | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

Table 1: Comparison of tests using data of table 4A in Duflo et al. (2011): dependent variable is use of fertilizer in season 1, 2 and 3 respectively, numbers represent p values for two-sided tests, "not" refers to a p value above 0.2, "not*" refers to cases where no pairs could be formed, "mono" for our direction of monotonicity test, "diff" for our difference test, in both cases we set $\theta = 0.3$, GS stands for the method of Gossner and Schlag (2013), classical for the linear regression analysis with homoskedastic errors, robust for White's robust method, logit and probit for the respective binary choice models.

We compliment the above by presenting in Table 2 the 95% confidence intervals of AIE under the difference test for the main variable of interest "SAFI season 1", and compare these to the confidence intervals for a linear probability model using the method of GS. Results are astonishingly similar despite the different model assumptions (the model underlying the difference test is more general than the linear

probability model investigated in GS) and the different methods for constructing the tests.

| SAFI season 1 | 95% CI |
|---|---|
| | season 1 |
| diff | $[0, 0.22]$ |
| GS paper | $[0, 0.23]$ |
| | season 2 |
| diff | $[-0.13, 0.12]$ |
| GS paper | $[-0.12, 0.13]$ |
| | season 3 |
| diff | $[-0.11, 0.11]$ |
| GS paper | $[-0.11, 0.12]$ |

Table 2: Comparison of exact 95% confidence intervals for the data of Duflo et al. (2011) used in Table 1.

Next we investigate the models underlying columns 2, 4 and 6 in Table 4A of Duflo et al. (2011). In these regressions there are seven additional independent variables, five of them are dummy variables (including gender and mud walls), the two other are "education" and "income". "education" presents the number of years of education, which is an integer that lies between 0 and 15. "income" describes the income (in $1,000$ Kenyan shillings) of the household of the farmer, this ranges in our data from 0 to 81.21.

We investigate the impact of "SAFI season 1" in season 1. If all variables are added (as in column 2 in Table 4A of Duflo et al., 2011) only three pairs can be formed. If one includes "education" measured in quartiles but not "income" then there are 62 pairs and results are significant at 5% level (estimated AIE is 0.175). Measuring "income" in quartiles and leaving out "education" gives 69 pairs but does not even generate significance at 20%, estimated AIE is 0.07. Including both of these variables in quartiles yields only 31 pairs and no significance even at 20%. To summarize we find that significance of the program is robust to including education in quartiles but not to including income in quartiles.

## 5.2 Advance Indicators of HIV Infection in Infants

Next we investigate a data set with 47 observations and two attributes taken from Mehta and Patel (1995, Table IV), that originates from a study of Hutto et al. (1991). Objective is to understand whether CD4 and CD8 blood serum levels of infants at 6 months of age can predict their later HIV infection (outcome binary valued). As this example is intended for illustration, we analyze the data a bit differently than Mehta

and Patel (1995). While we assume that CD4 and CD8 are ordinal data, they treat them as categorical data. We present our findings in Table 3.[15] "pairs" describes the number of pairs that are formed when running the test.

| $n = 47$ | CD4 | CD8 |
|---|---|---|
| p value mono | 0.01 | 0.42 |
| p value diff | 0.03 | 0.52 |
| 95% CI | $[-0.64, -0.01]$ | $[-0.19, 0.55]$ |
| pairs | 23 | 15 |
| AIE | $-0.377$ | 0.223 |
| p value logit | 0.01 | 0.04 |
| p value exact logit | 0.01 | 0.05 |
| p value probit | 0.01 | 0.03 |

Table 3: Comparison of methods for determining predictors of HIV infection.

While our results for CD4 are similar to those obtained by other popular methods used in the literature, those for CD8 are dramatically different. We conclude that the significant negative effect of CD4 blood serum levels on HIV infection does not rely the functional form of the link function. On the other hand, the effect of CD8 cannot be established under weaker assumptions, possibly due to the small data set but also possibly because it is absent if one does not add the logit or probit specifications.

## 5.3 Engel's Law

Engel's law (Engel, 1857) is an empirical observation on the negative relationship between income and proportion of income spent on food. Specifically, households with more income tend to spend proportionally less on food. We use Engel's original data set to investigate this relationship.[16] We choose income as the attribute and food share as outcome (which is a number between 0 and 1 and hence not binary valued). There are 199 observations, each of them corresponding to income and share of income spent on food of a Belgium family. Following the approach in this paper, we examine this relationship when comparing pairs of families such that 50% of the sample have income levels lying between their values. Results are presented in Table 4. We find significant evidence in favor of Engel's law under both the direction of monotonicity test and under the generalized AIE test. Note that the OLS estimate $\hat{\beta}$ and the estimate of the relative average effect $\hat{\rho}$ are very small. The exact linear regression of Gossner and Schlag (2013) performs very bad in this data set, the p value equals 1. On the other hand, the results of the standard linear regression analysis

[15] I would like to thank Georg Heinze for providing the p values for the exact logistic regression.

[16] We would like to thank Manisha Chakrabarty for providing this data set.

with robust standard errors (White, 1980), which are not exact, are very comparable to the results we obtain for the relative incremental effect.

| | income |
|---|---|
| p value mono | 0.04 |
| p value generalized AIE test | 0.04 |
| number of pairs | 99 |
| $\hat{\delta}$ | $-0.24$ |
| 95% CI for $\delta$ | $[-0.46, -0.01]$ |
| $\hat{\rho}$ | $-0.00041$ |
| 95% CI for $\rho$ | $[-0.00077, -0.00001]$ |
| GS | 1 |
| OLS estimate $\hat{\beta}$ | $-0.000061$ |
| p value of robust | 0.01 |
| 95% CI of robust test | $[-.000091, -.000031]$ |

Table 4: Testing for the relationship between income and food share expenditure: we consider the ordinal regression in which outcome $i$ is the food share expenditure of family $i$ and the single attribute is the income of family $i$. "pvalue mono" and "pvalue generalized AIE" show the two-sided p values for the direction of monotonicity and for the generalized AIE test respectively. p values for $\delta$ and $\rho$ are derived using the generalized AIE test. GS stands for the exact linear regression of Gossner and Schlag (2013). In the last row we add the 95% confidence interval of the standard test with robust errors (White, 1980).

### 5.4 Survival of Leukemia Patients

Consider the original data used by Cox (1972) in which time (in weeks) of remission of leukemia patients is given for two groups of patients, those treated with a drug and the control group. There is no further individual patient information. So we only have a single attribute which is group membership, i.e., $m = 1$ and outcomes are not binary valued. We assume that censoring is not dependent on the time of remission of a patient. As discussed in Section 3.1.5 we can thus compare the time of remission between the two groups by ignoring the censoring. As mentioned in Section 3.1.1 we are considering a stochastic inequality. Results are shown in Table 5.

## 6 Conclusion

Exact statistics is an exciting discipline. It requires new approaches and different perspectives. Typically one does not follow the standard routine of deriving the dis-

| | treated |
|---|---|
| p value mono | 0.02 |
| p value generalized AIE | 0.02 |
| pairs | 21 |
| $\hat{\delta} = \hat{\rho}$ | 0.52 |
| 95% CI | $[0.06, 0.82]$ |
| p value Cox | 0.01 |

Table 5: Comparing survival between two groups: we consider the ordinal regression in which outcome $i$ is the time of remission of patient $i$.

tribution of some test statistic. Exact probability bounds can help tackle the richness of the data generating processes, as in Dufour and Hallin (1993) and Gossner and Schlag (2012). The derandomization trick of Schlag (2008a) helps eliminate randomness that emerges during the construction of a test. For instance, randomization can be added to reduce the richness of distributions as first shown in Schlag (2008a) and also used in Gossner and Schlag (2012). Randomization also occurs naturally if one wishes to treat identical individuals identically as in this paper.

The central ingredient of this paper is the formation of blocks in which the only variation is in the attribute of interest, as it is done in stratified permutation testing (van Elteren, 1960). It allows us to solve a complicated problem with a very simple statistical test. In fact, stratified permutation tests would have a lot of potential, if one were be able to get a handle on how to prove the inequalities needed for establishing one-sided tests. For instance, in this paper, any test statistic can be used to construct a permutation test for testing $\beta_j = 0$ against $\beta_j \neq 0$ by permuting only within blocks. The difficulty is that one has to find a statistic for which one can prove that the null hypothesis is rejected for any $\beta_j \leq 0$ whenever it is rejected for $\beta_j = 0$. To this date no such statistic is known. Note that any method like ours that makes inference using information within such blocks cannot analyze interaction effects. Once the values of two different attributes are held fixed, their interaction effect is constant.

Numerous open questions remain. One important area of future research is develop methods for coping with continuously valued attributes. One idea is to collect them into quantiles and bound the resulting error. An alternative approach is to add more structure on the link function. Another question of interest is the analysis of interaction effects.

# References

[1] Brunner, E. and U. Munzel (2000), "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation", *Biometrical Jour-*

*nal* 42, 17-25.

[2] Chakrabarty, M. and W. Hildenbrand (2011), "Engel's Law Reconsidered", *Journal of Mathematical Economics* **47**, 289–299

[3] Cox, D.R. (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B 34 (2): 187–220.

[4] Duflo, E., M. Kremer, and J. Robinson (2011): "Nudging farmers to use fertilizer: theory and experimental evidence from Kenya," American Economic Review 101, 2350-2390.

[5] Dufour, J.-M., and M. Hallin (1993): "Improved Eaton bounds for linear combinations of bounded random variables, with statistical applications," Journal of the American Statistical Association, 88, 1026–1033.

[6] Gleser, L.J. (1975): "On the Distribution of the Number of Successes in Independent Trials," The Annals of Probability, 3(1), 182-188.

[7] Gossner, O. and K.H. Schlag (2013), "Finite-sample exact tests for linear regressions with bounded dependent variables", Journal of Econometrics 177, 75-84.

[8] Hoeffding, W. (1956), "On the distribution of the number of successes in independent trials." The Annals of Mathematical Statistics, 27, 713-721.

[9] Hoeffding W. (1963): Probability Inequalities for Sums of Bounded Random Variables, Journal of the American Statistical Association, 58, 13-30.

[10] Hutto C, W.P. Parks, S.H. Lai, M.T. Mastrucci, C. Mitchell, J. Munoz, E. Trapido, I.M. Master, amd G.B. Scott(1991) "A hospital-based prospective study of perinatal infection with human immunodeficiency virus type 1," J. Pediatr. 118(3), 347-53.

[11] Ichimura, H. (1987), Estimation of Single Index Models, Dissertation, Department of Economics, MIT.

[12] Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models", Journal of Econometrics 58, 71–120

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

[14] Mehta, C.R. and N.R. Patel (1995), "Exact Logistic Regression: Theory and Examples", *Statistics in Medicine* **14**, 2143-2160.

[15] Schlag, K.H. (2008a), "A New Method for Constructing Exact Tests Without Making any Assumptions", Working paper 1109, Universitat Pompeu Fabra, Barcelona.

[16] Schlag, K.H. (2008b), "Bringing Game Theory to Hypothesis Testing: Establishing Finite Sample Bounds on Inference", Working paper 1099, Universitat Pompeu Fabra, Barcelona.

[17] van Elteren, P. H. (1960), "On the combination of independent two sample tests of Wilcoxon", *Bulletin of the Institute of International Statistics* **37**, 351–361.

[18] White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", Econometrica 48, 817–838.

[19] Yates, F. (1934), "Contingency tables involving small numbers and the $\chi^2$ test," *Supplement to the Journal of the Royal Statistical Society* **1**, 217–235.

# A Appendix: Proofs

## A.1 Preliminaries

We present three results that will be used in the proofs and that are of independent interest.

We recall a result from (Schlag, 2008a) that shows how to turn randomized tests into nonrandomized tests.

**Lemma 1 (Schlag, 2008a)** *Let $\phi$ be a randomized test with level $\theta\alpha$ for testing some null hypothesis $H_0$ and type II error probability $\beta$ for some subset $A$ of the complement of the null hypothesis, so $E(1-\phi|A) \leq \beta$. Let $\bar{\phi} = 1_{\{\phi \geq \theta\}}$. Then $\bar{\phi}$ has level $\alpha$ and type II error probability bounded above by $\beta/(1-\theta)$.*

**Proof.** We apply the Markov inequality twice, namely

$$E\bar{\phi} = P(\phi \geq \theta) \leq E\phi/\theta \leq \alpha$$

and

$$E\left(1-\bar{\phi}|A\right) = P(\phi < \theta) = P(1-\phi > 1-\theta) \leq \frac{E(1-\phi|A)}{1-\theta} = \frac{\beta}{1-\theta}.$$

■

Next we identify the distribution that maximizes the type II error probability of the randomized version of McNemar's test. Consider a random vector $Q \in \{0,1\}^2$ and a sample of $n$ independent realizations from $Q$. We wish to test $H_0 : EQ_1 \leq EQ_2$. The

randomized version of McNemar's test (see Lehmann and Romano, 2005) evaluates whether there are significantly more occurrences of $Q = (0,1)$ than of $Q = (1,0)$ using the randomized version of the binomial test. Note that it is uniformly most powerful unbiased test for $H_0$ (Lehmann and Romano, 2005).

**Lemma 2 (Schlag, 2008b)** *For any given $w > 0$, the type II error probability of the randomized version of McNemar's test conditional on $EQ_1 = EQ_2 + w$ is attained when $P(Q_1 = Q_2) = 0$.*

We here present an alternative simpler proof of the statement.

**Proof.** Consider the test $\bar{\phi}$ that is generated by first transforming each observation of $(0,0)$ or $(1,1)$ equally likely into observations $(1,0)$ and $(0,1)$ and then applying the randomized version of McNemar's test. $\bar{\phi}$ is unbiased. As the randomized version of McNemar's test is a uniformly most powerful unbiased test, it is more powerful than $\bar{\phi}$. Hence $\bar{\phi}$ has a higher type II error probability. However, since $\bar{\phi}$ coincides with the randomized version of McNemar's test when $P(Q_1 = Q_2) = 0$ it follows that the type II error probability of these two tests is equal, in particular the randomized version of McNemar's test attains its type II error probability when $P(Q_1 = Q_2) = 0$. ■

Finally we present a novel exact test for comparing two independent samples of Bernoulli random variables where observations within each sample are not necessarily identically distributed.

Consider a sequence of $n$ independently distributed Bernoulli random $\{Z_i\}_{i=1}^n$. Let $p_i = P(Z_i = 1)$ and let $z_i$ be the observed values. Let $n_2 = n - n_1$. We wish to test $H_0 : \frac{1}{n_1}\sum_{i=1}^{n_1} p_i \leq \frac{1}{n_2}\sum_{j=n_1+1}^{n} p_j + d$ for some $d \in (-1,1)$. Let $\bar{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} p_i$ and $\bar{p}_2 = \frac{1}{n_2}\sum_{j=n_1+1}^{n} p_j$.

Consider the test, which we call the *two sample Bernoulli difference test*, that rejects the null hypothesis if $\sum_{i=1}^{n_1} z_i - \sum_{j=n_1+1}^{n} z_j \geq \bar{k}$. In the following we will show how to choose $\bar{k}$.

For each $i$ let $\bar{Z}_i$ be a Bernoulli random variable where $P\left(\bar{Z}_i = 1\right) = \bar{p}_1$ if $i \in \{1,..,n_1\}$ and $P\left(\bar{Z}_i = 1\right) = \bar{p}_2$ if $i \in \{n_1+1,..,n\}$. Let $Y_j = 1 - Z_j$ and $\bar{Y}_j = 1 - \bar{Z}_j$. Then $Y_j$ is also Bernoulli distributed.

Then

$$\begin{aligned} P\left(\sum_{i=1}^{n_1} Z_i - \sum_{j=n_1+1}^{n} Z_j \geq k\right) &= P\left(\sum_{i=1}^{n_1} Z_i + \sum_{j=n_1+1}^{n} (1 - Z_j) \geq n_2 + k\right) \\ &= P\left(\sum_{i=1}^{n_1} Z_i + \sum_{j=n_1+1}^{n} Y_j \geq n_2 + k\right). \end{aligned}$$

Following Gleser (1975),

$$\begin{aligned} P\left(\sum_{i=1}^{n_1} Z_i + \sum_{j=n_1+1}^{n} Y_j \geq n_2 + k\right) &\leq P\left(\sum_{i=1}^{n_1} \bar{Z}_i + \sum_{j=n_1+1}^{n} \bar{Y}_j \geq n_2 + k\right) \\ &= P\left(\sum_{i=1}^{n_1} \bar{Z}_i - \sum_{j=n_1+1}^{n} \bar{Z}_j \geq k\right) \end{aligned}$$

if $n_2 + k \geq n_1\bar{p}_1 + n_2\left(1 - \bar{p}_2\right) + 2$, so if $k \geq n_1\bar{p}_1 - n_2\bar{p}_2 + 2$

Assume that $H_0$ is true. Then $0 \leq \bar{p}_1 \leq \bar{p}_2 + d$ and hence

$$\begin{aligned} n_1\bar{p}_1 - n_2\bar{p}_2 &\leq n_1\bar{p}_1 - n_2\max\left\{\bar{p}_1 - d, 0\right\} \leq \left(n_1 - n_2\right)\min\left\{1 + d, 1\right\} + n_2 d \text{ if } n_1 \geq n_2 \\ n_1\bar{p}_1 - n_2\bar{p}_2 &\leq n_1\min\left\{\bar{p}_2 + d, 1\right\} - n_2\bar{p}_2 \leq n_1 d + \left(n_1 - n_2\right)\max\left\{0, -d\right\} \text{ if } n_1 < n_2 \end{aligned}$$

and hence

$$n_1\bar{p}_1 - n_2\bar{p}_2 \leq \max\left\{n_1 - n_2, 0\right\} + \min\left\{n_1 d, n_2 d\right\}.$$

Let $\bar{W}_i$ be a Bernoulli random variable such that $P\left(\bar{W}_i = 1\right) = \bar{p}_2 + d$. Using the fact that $\bar{p}_1 \leq \bar{p}_2 + d$, we obtain

$$P\left(\sum_{i=1}^{n_1} \bar{Z}_i - \sum_{j=n_1+1}^{n} \bar{Z}_j \geq k\right) \leq P\left(\sum_{i=1}^{n_1} \bar{W}_i - \sum_{j=n_1+1}^{n} \bar{Z}_j \geq k\right).$$

Hence we have shown that

$$P\left(\sum_{i=1}^{n_1} X_i - \sum_{j=n_1+1}^{n} Z_j \geq k\right) \leq P\left(\sum_{i=1}^{n_1} \bar{W}_i - \sum_{j=n_1+1}^{n} \bar{Z}_j \geq k\right)$$

if $k \geq \max\left\{n_1 - n_2, 0\right\} + \min\left\{n_1 d, n_2 d\right\} + 2$. Note that when samples are balanced, so $n_1 = n_2 = n/2$ then the condition turns into $k \geq nd/2 + 2$.

Let $g\left(k\right)$ be such that

$$g\left(k\right) = \max_{\max\{-d,0\} \leq p \leq \min\{1-d,1\}} \sum_{i=\max\{k,0\}}^{n_1} \binom{n_1}{i} \left(p + d\right)^i \left(1 - p - d\right)^{n_1 - i} \sum_{j=0}^{\min\{i-k, n_2\}} \binom{n_2}{j} p^j \left(1 - p\right)^{n_2 - j}.$$

Note that $g\left(n_1\right) = \max_{p \in [0, 1-d]} \left(p + d\right)^{n_1} \left(1 - p\right)^{n_2}$.

Let $\bar{k}$ be the smallest integer $k$ such that $g\left(k\right) \leq \alpha$ and $k \geq \max\left\{n_1 - n_2, 0\right\} + \min\left\{n_1 d, n_2 d\right\} + 2$.

This yields the following result.

**Proposition 8** *The two sample Bernoulli difference test that rejects the null hypothesis if* $\sum_{i=1}^{n_1} z_i - \sum_{j=n_1+1}^{n} z_j \geq \bar{k}$ *has type I error probability bounded above by* $\alpha$.

## A.2 Proof of Proposition 1

Assume that the null hypothesis is true, so that $f$ is monotone decreasing.

Fix some ordering $o$ as constructed in step 2 and consider the randomized test $\bar{\phi}_o$ defined by $\bar{\phi}_o(x,y) = q(x,y,o|x,y)$. We first show that $\bar{\phi}_o$ has level $\theta\alpha$.

Consider some block as constructed in step 1 that has at least two elements, let $s$ be its index. Consider some $r \in \{1,...,l_s\}$. Let $z_{r,s} = 1$ if $\left(x_{o(r+l_s,s),j} - x_{o(r,s),j}\right)\left(y_{o(r+l_s,s)} - y_{o(r,s)}\right) > 0$ and let $z_{r,s} = 0$ if $\left(x_{o(r+l_s,s),j} - x_{o(r,s),j}\right)\left(y_{o(r+l_s,s)} - y_{o(r,s)}\right) < 0$. Let $Z_{r,s}$ be the Bernoulli distribution that is generated by taking the distribution of $z_{r,s}$ over all possible realizations of $y$ conditional on $x$. Since $x_{o(r+l_s,s),j} > x_{o(r,s),j}$, $x_{o(r+l_s,s),k} = x_{o(r,s),k}$ for all $k \neq j$, and $f$ is monotone decreasing we obtain that $f\left(x_{o(r+l_s,s)}\right) \leq f\left(x_{o(r,s)}\right)$. Simplifying notation but still keeping in mind that all probabilities are conditional on $x$, we find that

$$\begin{aligned}\frac{P(Z_{r,s}=1)}{P(Z_{r,s}=0)} &= \frac{P\left(Y_{o(r+l_s,s)} > Y_{o(r,s)} | Y_{o(r,s)} \neq Y_{o(r+l_s,s)}\right)}{P\left(Y_{o(r+l_s,s)} < Y_{o(r,s)} | Y_{o(r,s)} \neq Y_{o(r+l_s,s)}\right)} = \frac{P\left(Y_{o(r+l_s,s)}=1, Y_{o(r,s)}=0\right)}{P\left(Y_{o(r+l_s,s)}=0, Y_{o(r,s)}=1\right)} \\ &= \frac{P\left(Y_{o(r+l_s,s)}=1\right)P\left(Y_{o(r,s)}=0\right)}{P\left(Y_{o(r+l_s,s)}=0\right)P\left(Y_{o(r,s)}=1\right)} = \frac{f\left(x_{o(r+l_s,s)}\right)}{1-f\left(x_{o(r+l_s,s)}\right)} \Big/ \frac{f\left(x_{o(r,s)}\right)}{1-f\left(x_{o(r,s)}\right)} \leq 1\end{aligned}$$

which implies that $P(Z_{r,s}=1) \leq 1/2$.

As individuals are making independent choices conditional on $x$, the variables $Z_{r,s}$ are independent for all $r = 1,..,l_s$ and $s = 1,...,S$.

Given the functional form of the randomized binomial test, we obtain that $P\left(\bar{\phi}_o(x,y)=1|x\right)$ is maximized when $P(Z_{r,s}=1) = 1/2$ holds for all $r,s$. As the randomized binomial test is an exact test with size $\theta\alpha$, this implies that $P\left(\bar{\phi}_o(x,y)=1|x\right) \leq \theta\alpha$. Hence $\bar{\phi}_o$ has level $\theta\alpha$.

We now prove that the direction of monotonicity test $\phi^*$ is exact. Let $\bar{\phi}$ be the randomized test defined by $\bar{\phi}(x,y) = q(x,y,O|x,y)$. In other words, $\bar{\phi}$ is obtained by realizing an order $o$ and then determining the rejection probability based on $\bar{\phi}_o$. Hence, $P\left(\bar{\phi}=1|x,y,o\right) = P\left(\bar{\phi}_o=1|x,y\right) = q(x,y,o)$ and $P\left(\bar{\phi}=1|x,y\right) = Eq(x,y,O|x,y)$. Moreover, by definition, $\phi^*(x,y)=1$ if and only if $E\bar{\phi}(x,y) \geq \theta$.

Above we showed that $\bar{\phi}_o$ has level $\theta\alpha$ for each ordering $o$. Hence, $\bar{\phi}$ also has level $\theta\alpha$. In other words, $P\left(\bar{\phi}=1|x\right) = Eq(x,Y,O|x) \leq \theta\alpha$.

The rest of the proof follows directly from Lemma 1.

## A.3 Proof of Proposition 3

Let $\phi_1$ be the randomized test that results after applying step 5, so $\phi_1(x,y) = Eq(x,y,O|x,y)$. We wish to bound the type II error probability of $\phi_1$. It is as if we are doing the following. We are facing an independent sample of $N_1$ matched pairs $Q \in \{0,1\}^2$. For each pair $(r,s)$ we identify $Q = (1,0)$, $(0,0)$ and $(0,1)$ with

$Y_{o(r+l_s,s)} > Y_{o(r,s)}$, $Y_{o(r+l_s,s)} = Y_{o(r,s)}$ and $Y_{o(r+l_s,s)} < Y_{o(r,s)}$ respectively. We are testing $H_0 : EQ_1 \leq EQ_2$ using the randomized version of McNemar's test. We are computing the type II error probability where

$$EQ_1 - EQ_2 = P(Q = (1,0)) - P(Q = (0,1)) = N\delta/N_1.$$

Lemma 2 states that the type II error probability is attained when $P(Q_1 = Q_2) = 0$.

The rest of the proof follows from Lemma 1.

## A.4 Proof of Proposition 4

The proof is a simple exercise. Following (5) we are comparing success probabilities above and below the median. We do this using Proposition 8, constructing a randomized test $\phi_{Dr}$ to avoid integer problems. Finally we derandomize using Lemma 1.

## A.5 Proof of Proposition 5

To derive a bound on the type II error probability of the randomized test $\phi_{Dr}$ we will apply Gleser (1975), and hence need to ensure that the true effect is sufficiently above the cutoff used on the null hypothesis. We wish to derive an upper bound on the probability of not rejecting the null hypothesis. To not reject the null hypothesis means, given the definition of $\bar{k}$, means that the difference between the two samples is at most $\bar{k}$. Thus we can act as if both samples are i.i.d. if $\delta_j^a N \geq \bar{k} + 2$. The rest of the proof then follows with Lemma 1.

## A.6 Proof of Propositions 6 and 7

Note that $\{(1 + \bar{z}_{r,s})/2\}$ is a sequence of Bernoulli variables with average mean under the null hypothesis bounded above by $\left(1 + \bar{\delta}_j\right)/2$. So we use Hoeffding (1956) inequalities to design the randomized test. The constraint "$c \geq np + 1$" in the definition of $\phi_{Qr}$ ensures that the test only rejects if the number of successes is sufficiently high so that that according to Hoeffding (1956) the tail probability is maximal when the data is i.i.d.. We then derandomize using Lemma 1.

The proof of Proposition 7 is analogous.