How to Attain Minimax Risk with Applications to Distribution-Free Nonparametric Estimation and Testing[1]

Karl H. Schlag[2]

March 12, 2007 (first version May 12, 2006)

[1]The author would like to thank Dean Foster, Markku Lanne, Fortunato Pesarin, Richard Spady and David Thesmar for comments and Javier Rivas and Marcello Sartarelli for helping as research assistants.

[2]Economics Department, European University Institute, Via della Piazzuola 43, 50133 Florence, Italy, Tel: 0039-055-4685951, email: schlag@eui.eu

Abstract

We show how to a derive exact distribution-free nonparametric results for minimax risk when underlying random variables have known finite bounds and means are the only parameters of interest. Transform the data with a randomized mean preserving transformation into binary data and then apply the solution to minimax risk for the case where random variables are binary valued. This shows that minimax risk is attained by a linear strategy and that the set of binary valued distributions contains a least favorable prior. We apply these results to statistics.

All unbiased symmetric non-randomized estimates for a function of the mean of a single sample are presented. We find a most powerful unbiased test for the mean of a single sample. We present tight lower bounds on size, type II error and minimal accuracy in terms of expected length of confidence intervals for a single mean and for the difference between two means.

We show how to transform the randomized tests that attain the lower bounds into non-randomized tests that have at most twice the type I and II errors. Relative parameter efficiency can be measured in finite samples, in an example on anti-self-dealing indices relative (parameter) efficiency is 60% as compared to the tight lower bound.

Our method can be used to generate distribution-free nonparametric estimates and tests when variance is the only parameter of interest. In particular we present a uniformly consistent estimator of standard deviation together with an upper bound on expected quadratic loss. We use our estimate to measure income inequality.

Keywords: exact, distribution-free, nonparametric inference, finite sample theory.

JEL classification: C14, C13, C12, C44.

1 Introduction

In this paper we consider distribution-free nonparametric inference when only a bounded set containing the support of the underlying distributions is known and means are the only parameters of interest. We only consider *exact* results in terms of explicit formulae for finite samples. Criteria for the selection of an estimator is minimax risk based on a general loss function, a special case being quadratic loss. Hypothesis tests are evaluated according to size and power, families of confidence intervals according to their accuracy and expected length. The common denominator for these different applications is that we have found a method to solve minimax risk. Our results for analyzing means enable us to establish upper bounds on risk when variances are the only parameters of interest.

Methodologically this paper draws on three different building blocks: a randomization method to be able to limit attention to binary valued random varaibles, a rounding trick to transform randomized tests into nonrandomized tests and a combination method that allows to transform statements in terms of variances into ones in terms of means.

Novel ways of making distribution-free nonparametric inference due to our results are illustrated in two data examples. We compare protection of minority shareholders across different legal systems using the anti-self-dealing indices gathered by Djankov et al. (2005) and investigate income inequality in the US between 1990 and 2002 using PSID data. In the first case the data is contained in $[0, 1]$ by definition of the indices, in the second case we restrict attention to income that is not top coded which generates an upper bound too. Exogenous bounds on data are often given. For instance, pain is often measured on a bounded scale, grades in school belong to a bounded scale, data gathered in practically any laboratory or field experiment relating to game theory or economics belong to a known bounded scale.

Assume first that means are the only parameters of interest. The value of this paper is best illustrated by first listing existing exact distribution-free nonparametric results. Estimators that minimize maximum expected quadratic loss are available, both biased and unbiased, for the mean and for the difference between the means of two independent random variables (Hodges and Lehmann, 1950). Confidence intervals for the mean of a single sample have been constructed (see Romano and Wolf (2000) and Diouf and Dufour (2006) and papers cited therein). Permutation tests

are available to construct unbiased tests for comparing means. These tests and confidence intervals are exact in terms of securing the specified size or coverage, however distribution-free properties in terms of power or accuracy are not available for finite sample sizes. The exception is the confidence interval for a mean constructed by Bickel et al. (1989) using bounds of Hoeffding (1963) for which power is easily measured. The danger of taking asymptotic results as a proxy for finite sample performance is demonstrated by Lehmann and Loh (1990, see also Dufour, 2003). For any given number of observations the actual size of the t test derived for a nominal level α is equal to 1. Simulations only provide limited insights as the environment is too rich being *nonparametric* (as defined by Fraser, 1957), in particular maximal risk cannot be computed nor can it be approximated by numerical simulations. There are n^n possible distributions of a single random variable for grid size $1/n$.

We highlight some novel results in this paper. We show that any strategy can be replaced by a linear strategy that has weakly lower maximal risk where the value of maximal risk is attained for some binary valued distribution. This makes the value of maximal risk computable. Moreover, minimax risk can often be solved analytically by restricting attention to binary valued distributions. We prove existence of minimax risk strategies under very general conditions. We present all unbiased symmetric non-randomized estimators of functions of the mean of a single sample and show how to derive the value of minimax risk. This paper contains the first (most powerful) unbiased one- and two-sided tests for a single sample. We present the first lower bounds on power for a given size and the first lower bounds on maximal expected length of unbiased confidence intervals and show how these bounds can be attained. Thereby we provide a method to evaluate accuracy of confidence intervals on an absolute scale obtained by comparing their expected length to the minimal expected length. We present a family of confidence bounds that maximizes coverage for a given maximal length. Finally, we present non-randomized tests for a single mean and for the comparison of two means (both for independent samples and for matched pairs) that, compared to the minimal lower bounds attained by the randomized tests mentioned above, have at most twice the type I and type II error. Non-randomized confidence intervals and bounds are easily derived from these tests.

All these results initiate by showing how randomization can be used to extend results known for binary valued data to bounded data. The randomization can then be eliminated when considering estimation under convex loss without increasing maximal risk. For hypothesis testing non-randomized tests can be derived following an idea of Gupta and Hande (1992). The randomized tests themselves remain useful benchmarks

to establish bounds. For the first time we are able to specify minimal sample sizes to guarantee maximal power or to guarantee the minimal accuracy of confidence intervals. These bounds can then be used to measure relative efficiency of alternative exact but non-randomized methods.

Randomization occurs as in Schlag (2003) where it was used to solve for minimax regret. Recently we discovered that this method had been used before by Cucconi (1968, in Italian, cited by Pesarin, 1984) when extending the sequential probability ratio test (Wald, 1947) to an exact distribution-free nonparametric sequential test for the mean of a random variable that has support in $[0,1]$. We also have now discovered that Gupta and Hande (1992) transform payoffs in this way when deriving a selection procedure that maximizes the minimum probability of correct selection.

The underlying idea is simple. First perform a change of variables by affinely transforming all outcomes into $[0,1]$. Next randomly transform any observation from $[0,1]$ into $\{0,1\}$ using the following mean preserving transformation. If outcome y is observed then replace it with outcome 1 with probability y and replace it with outcome 0 with probability $1-y$. Independently apply this transformation to each observation to obtain the so-called *binomially transformed sample*. From then on act as if the data generating process produced this binary valued sample. The insight is that after making this transformation it is as if the underlying distribution were binary valued with the same means as the original process. The transformation indirectly reduces the possible data generating processes one can face and hence also reduces maximal risk. In particular we thus prove that there is always a *least favorable prior* among the set of binary valued distributions.

Non-randomized estimates are derived by taking the expected estimate as in Hodges and Lehmann (1950). Non-randomized tests are created by choosing the hypothesis that is most likely under the associated randomized test, following ideas of Gupta and Hande (1992).

In a separate section we show how to derive estimators and tests when variance is the only parameter of interest. Here we have found no exact distribution-free nonparametric tests. Uniformly consistent estimators such as the unbiased estimator of variance and the empirical standard deviation as biased estimator of the standard deviation come without an exact upper bound on expected quadratic loss.

We present an upper bound on the risk of the unbiased estimator of the variance. We present an estimator of standard deviation and an upper bound on its expected quadratic loss. Comparing the latter to lower bounds we find that our estimator

of standard deviation requires in small samples less than 60% of the observations than the empirical standard deviation. We show how one can design distribution-free nonparametric tests of variances.

There is a simple trick to derive results in terms of variance from results in terms of means. Following Walsh (1962) one can generate an iid sample with half as many data points that can be interpreted as resulting from a random variable with mean equal to the underlying variance of the original data. Results in terms of variance can be derived from results in terms of means. The caveat is that minimax risk properties are not known to carry over.

The paper is organized as follows. In Section 2 we present the main model that concerns inference based on a given sample when means are the only parameters of interest. In Subsection 2.1 we introduce notation and present the main theorem on the randomization method. In Subsection 2.2 we show how this can be used to prove existence of minimax risk. Applications to estimation and hypothesis testing are presented separately in Subsections 2.3 and 2.4. Section 3 considers inference in terms of variance and shows how to reduce the problem to investigation of means. Subsections 3.1 and 3.2 deal with estimation and testing. Section 4 contains the conclusion.

2 Concern for Mean

Consider a decision maker has to make some choice based on an independent sample of data generated by some random vector that has a known support but an unknown distribution. The objective of the decision maker is a function of the unknown underlying means.

We first describe the random vector. Given $K \in \mathbb{N}$ and a collection of compact sets $\mathcal{Y}_k \subset \mathbb{R}$ with $\min \mathcal{Y}_k < \max \mathcal{Y}_k$ for $k = 1, .., K$ consider a K dimensional random vector $Y = (Y_1, .., Y_K)$ where $Y_k \in \mathcal{Y}_k$ for $k = 1, .., K$. Let P_Y denote the joint distribution of Y and let $\mu = (\mu_k)_{k=1}^K$ denote the mean vector.

Assume that $\mathcal{Y}$ is known by the decision maker but that P_Y is unknown except for the fact that the decision maker knows some set $\mathcal{P} \subseteq \Delta\mathcal{Y}$ that contains the distribution P_Y where $\mathcal{P}$ has to satisfy two properties specified below.

In order to simplify presentation, assume that $\mathcal{Y} = [0, 1]^K$. To deal with the more general case first transform any outcome $y_k \in \mathcal{Y}_k$ affinely into $\frac{y_k - w_k^0}{w_k^1 - w_k^0} \in [0, 1]$ where $w_k^0 := \min \mathcal{Y}_k$ and $w_k^1 := \max \mathcal{Y}_k$, $k = 1, .., K$. Furthermore, our results do not change

if one replaces $\mathcal{Y}_k$ with $[w_k^0, w_k^1]$ for $k = 1, .., K$. Y is called *binary valued* if $Y_k \in \{0, 1\}$ for all k which means for the original model that each component can only attain one of two distinct values.

Central to our analysis will be the use of a randomization method which is a mean preserving random transformation that transforms outcomes in $[0, 1]^K$ into a binary valued vector of the same dimension. Specifically, $t : [0, 1]^K \to \Delta \{0, 1\}^K$ is called a *mean preserving binary transformation* if $\Pr(t_k(y) = 1) = y_k$ for $k = 1, .., K$. Let $\mathcal{T}$ be the set of mean preserving binary transformations. The most important representative is the *independent transformation* that satisfies $\prod_{k=1}^K \Pr(t_k(y) = z_k) = \Pr(t = z)$ for all $z \in \{0, 1\}^K$. The *correlated transformation* emerges when drawing a value z from an independent uniformly distributed random variable on $[0, 1]$ and then for each $k = 1, .., K$ transforming y_k into 1 if and only if $y_k \geq z$.

We assume that $\mathcal{P} \subseteq \Delta [0, 1]^K$ satisfies the following two properties: (i) $\mathcal{P}$ is convex and (ii) there exists a mean preserving binary transformation t such that $P_{t(Y)}$ is contained in the closure of $\mathcal{P}$ whenever $P_Y \in \mathcal{P}$. In the following we only include transformations in $\mathcal{T}$ if they satisfy (ii). Let $\mathcal{P}^b$ be the set of all binary valued distributions contained in $\mathcal{P}$ so $\mathcal{P}^b = \Delta \{0, 1\}^K \cap \mathcal{P}$. Note that $\mathcal{P}^b = \left\{P_{t(Y)} : P_Y \in \mathcal{P}\right\}$ for each $t \in \mathcal{T}$.

We typically illustrate our results for $\mathcal{P} = \Delta [0, 1]^K$ in which case (ii) is not a constraint as it holds for all mean preserving binary transformations and $\mathcal{P}^b = \Delta \{0, 1\}^K$. In this particular case our approach is *nonparametric* ($\mathcal{P}$ is infinitely dimensional) and distribution-free (only the support of the random vector is specified).

If $\mathcal{P}$ is the set of all distributions in which the underlying actions yield independent payoffs, so $\mathcal{P} = (\Delta [0, 1])^K$, then (ii) is satisfied by the independent transformation and $\mathcal{P}^b = (\Delta \{0, 1\})^K$. One can also add structure on $\mathcal{P}$, e.g. by specifying $w^{(1)}, .., w^{(r)} \in [0, 1]^K$ and setting $\mathcal{P} = \left\{P \in \Delta [0, 1]^K : \mu \in \left\langle w^{(i)}, i = 1, .., r \right\rangle\right\}$ where $\langle A \rangle$ denotes the convex hull of the set A.

We now specify how the data set is generated.

Given $N \in \mathbb{N}$ let $Y^{1,N}$ be a random sample of N independent observations of a realization of Y. Let $y^{1,N} = \left(y^{(1)}, .., y^{(N)}\right)$ denote a typical realization. Let P^N be the distribution of $Y^{1,N}$ induced by P. The decision maker observes $\tilde{y}^{1,N}$ realized from $\tilde{Y}^{1,N}$ that is related to $y^{1,N}$ and $Y^{1,N}$ as follows. For $K = 1$ let $\tilde{y}^{(n)} = y^{(n)}$ for $n = 1, .., N$. For $K \geq 2$ we allow that the decision maker does not observe all components of $y^{(n)}$ in round n. For each $n = 1, .., N$ there is some given set $o_n \subseteq \{1, .., K\}$ where o_n specifies the components of $y^{(n)}$ that the decision maker

observes. So the decision maker observes $\left(y_k^{(n)}\right)_{k\in o_n}$ of the n-th realization of Y given by the n-th component of $y^{1,N}$. One speaks of *matched pairs* or *paired data* if the entire vector y^n is observed in each round n so if $o_n = \{1,..,K\}$ and hence $\tilde{y}^{(n)} = y^{(n)}$ for $n = 1,..,N$. K *independent samples* are observed if $|o_n| = 1$ for all n where $|A|$ denotes the cardinality of the finite set A. Let $\tilde{Y}^{1,N}$ be the random sequence of observations, so $\tilde{Y}^{1,N} = \left(\tilde{Y}^n\right)_{n=1}^N$ with $\tilde{Y}^n = \left(Y_k^n\right)_{k\in o_n}$ and similarly let $\tilde{y}^{1,N}$ be defined as a realization of $\tilde{Y}^{1,N}$. Let $\widehat{\mathcal{Y}}$ be the set of possible observations.

Finally, given the observed data $\tilde{y}^{1,N}$ the decision maker has to make some choice z where z belongs to some set $\mathcal{G}$. Thus the strategy of the decision maker is a mapping $f : [0,1]^N \to \Delta\mathcal{G}$. f is called *non-randomized* (or *deterministic*) if $f \in \mathcal{G}$ and f is called *randomized* if f is not deterministic. f is called *symmetric* if it satisfies the following two conditions. (i) f does not depend on how actions are labeled so if f is invariant to permutations of the indices of the actions. (ii) f does not depend on how the N independent observations are indexed, so if $o_n = o_m$ then f is invariant if the indices n and m are interchanged. Finally, f is called *linear* if f is linear in $\tilde{y}_k^{(n)}$ for each n and each k.

The trick of this paper will be to first transform each observation $\tilde{y}^{(n)}$ using a mean preserving transformation and then to apply a rule designed for the case of binary valued payoffs. f is called *binomial* if $f\left(\tilde{Y}^{1,N}\right) = f\left(t\left(\tilde{Y}^{1,N}\right)\right)$ for some transformation $t \in \mathcal{T}$. f^b is called a *binomial transformation* of f if $f^b\left(\tilde{Y}^{1,N}\right) = f\left(t\left(\tilde{Y}^{1,N}\right)\right)$ for some $t \in \mathcal{T}$.

2.1 Minimax Risk

Following Wald (1950) the decision maker measures the outcome of making some choice z given distribution P in terms of a real valued *loss function* $W(z,P)$. We assume that loss W only depends on the choice z and the mean vector μ of the underlying random variable Y, so there exists $W_0 : \mathcal{G}\times[0,1]^K \to \mathbb{R}$ such that $W(z,P) = W_0(z,\mu)$. Expected loss of choosing strategy f when facing distribution P is called *risk* and is denoted by R so $R(f,P) = E_P W_0\left(f\left(\tilde{Y}^{1,N}\right),\mu\right)$. The decision maker is assumed to choose a strategy that *attains minimax risk* in the sense that

$$f^* \in \arg\min_f \sup_{P\in\mathcal{P}} R(f,P).$$

This is the central result of the paper.

Proposition 1 *(ia)* $R\left(f^b, P\right) = R\left(f, P^b\right)$.

(ib) If $\mu_0 \in [0,1]$ *then* $\left\{R\left(f^b, P\right) : \mu = \mu_0\right\} \subseteq \left\{R\left(f, P\right) : \mu = \mu_0\right\}$, *in particular* $\sup_{P\in\mathcal{P}} R\left(f^b, P\right) \leq \sup_{P\in\mathcal{P}} R\left(f, P\right)$.

(ic) If $f^* \in \arg\min_f \sup_{P\in\mathcal{P}} R\left(f, P\right)$ *then* $f^{*b} \in \arg\min_f \sup_{P\in\mathcal{P}} R\left(f, P\right)$.

(ii) If $f^* \in \arg\min_f \sup_{P^b\in\mathcal{P}^b} R\left(f, P^b\right)$ *then* $f^{*b} \in \arg\min_f \sup_{P\in\mathcal{P}} R\left(f, P\right)$.

(iii) If $f^* \in \arg\min_f \sup_{P\in\mathcal{P}} R\left(f, P\right)$ *then* $\sup_{P\in\mathcal{P}} R\left(f^*, P\right) = \sup_{P^b\in\mathcal{P}^b} R\left(f^*, P^b\right)$.

We briefly rephrase the proposition above in words. (ia) Risk from using an strategy when facing a binary valued distribution with mean vector μ is equal to the risk from using the binomially transformed strategy when facing a distribution with mean vector μ. (ib) The binomially transformed strategy attains weakly lower maximal (and weakly higher minimal) risk than the original strategy among all distributions that have the same mean. (ic) If a strategy attains minimax risk then so does its binomial transformation. (ii) To derive minimax risk it is sufficient to solve the simpler problem of finding a strategy that attains minimax risk when facing binary valued distributions and then taking its binomial transformation, using any of the mean preserving binary valued transformations in $\mathcal{T}$. (iii) Minimax risk is always attained by some binary valued distribution. In other words, it is hardest to guarantee the lowest risk when facing binary valued distributions. There is always a binary valued distribution that is a *least favorable distribution*. All statements follow easily from (ia) (see proof below).

Proof. Consider the n-th element of the sample y_n. Under f^b, y_n is transformed into outcome 1 with probability y_n and transformed into 0 otherwise. As y_n itself was independently drawn from P we obtain that the ex-ante probability that the n-th element of the sample is transformed into 1 is equal to $\int_0^1 y dP\left(y\right) = \mu\left(P\right)$. Thus, the risk under f^b when facing P is equal to the risk under f when facing a binary valued distribution with the same mean which proves (ia). The rest of part (i) as well as part (ii) is a direct consequence of (ia). Part (iii) also follows immediately. Since f^* attains minimax risk,

$$\sup_{P\in\mathcal{P}} R\left(f^*, P\right) \leq \sup_{P\in\mathcal{P}} R\left(f^{*b}, P\right).$$

Using (ia) we obtain

$$\sup_{P\in\mathcal{P}} R\left(f^{*b}, P\right) = \sup_{P^b\in\mathcal{P}^b} R\left(f^*, P^b\right) \leq \sup_{P\in\mathcal{P}} R\left(f^*, P\right).$$

Combining these two inequalities proves (iii). ■

As f^b is linear when using the independent transformation we obtain from Proposition 1(ib) using standard arguments for symmetry (following quasi-convexity of the maximum operator): Any strategy can be replaced by a symmetric linear strategy that attains weakly lower maximal risk where maximal risk is attained for some binary valued distribution. Moreover, when the choice set $\mathcal{G}$ is convex then following Hodges and Lehmann (1950, Theorem 3.2) one can confine attention to non-randomized strategies.

Corollary 1 *Assume that $\mathcal{T}$ contains the independent transformation.*

(i) For any strategy f there exists a symmetric linear strategy f^* such that $\sup_{P\in\mathcal{P}} R(f^*,P) \leq \sup_{P\in\mathcal{P}} R(f,P)$ *where* $\sup_{P\in\mathcal{P}} R(f^*,P) = \sup_{P\in\mathcal{P}^b} R(f^*,P)$.

(ii) Assume additionally that $\mathcal{G}$ is a Euclidean space and that loss W is convex in z. Then there exists a non-randomized strategy f^* satisfying (i).

In steps, starting with a strategy f one symmetrizes this strategy by permuting the labels to obtain a symmetric strategy f^s. Then one chooses the binomial transformation of f^s based on the independent transformation. Finally, if the choice set is convex, then one obtains f^* defined by $f^*\left(\tilde{y}^{1,N}\right) = Ef^{sb}\left(\tilde{y}^{1,N}\right)$. When $K=1$ this means that there exists a function $g:\{0,1,..,N\}\to\mathcal{G}$ such that

$$f^d\left(y^{1,N}\right) = \sum_{i_1=0}^{1}...\sum_{i_N=0}^{1}\prod_{n=1}^{N}\left(i_n y^{(n)} + (1-i_n)\left(1-y^{(n)}\right)\right) g\left(\sum_{k=1}^{N} i_k\right) \tag{1}$$

where $i_n = 1$ is associated to the event in which the n-th observation $y^{(n)}$ was transformed into 1.

Note that the result in Corollary 1(ii) does not imply that the class of non-randomized strategies is essentially complete as we are not establishing dominance for all distributions but only comparing maximal risk among those that have the same mean vector.

Next we expand on a rounding trick found in a result and proof of Gupta and Hande (1992) for selection procedures to show how one can derive non-randomized strategies - albeit with higher risk - when there are only two choices. The idea is to select the more likely choice of the randomized strategy. Given $\mathcal{G}=\{0,1\}$ and a strategy f let $f^m\in\{0,1\}$ be the non-randomized strategy defined by $f^m\left(\tilde{y}^{1,N}\right)_1 = 1$ if and only if $f\left(\tilde{y}^{1,N}\right)_1 > 0.5$.

Proposition 2 *If $\mathcal{G}=\{0,1\}$ and $W\geq 0$ then* $\sup_{P\in\mathcal{P}} R(f^m,P) \leq 2\sup_{P\in\mathcal{P}} R(f,P)$.

Proof. Consider P such that $W(1,P)\leq W(0,P)$. Recalling the statement and proof of Gupta and Hande (1992, Theorem 2.3) we obtain $\Pr(f^m=1) = \int 1_{\{f_1-f_0>0\}}dP \geq \int (f_1-f_0)\,dP = 2E_P(f_1) - 1$. The rest is straightforward as the previous implies $\Pr(f^m=0)\leq 2E_P(f_0)$ and hence

$$\begin{aligned} R(f^m,P) &= \sum_{z\in\{0,1\}} \Pr(f^m=z)\,W(z,P) = \Pr(f^m=0)\,(W(0,P)-W(1,P)) + W(1,P) \\ &\leq 2E_P(f_0)\,(W(0,P)-W(1,P)) + W(1,P) \leq 2R(f,P). \end{aligned}$$

■

2.2 Existence

Proposition 1 can also be used to ensure existence of a minimax risk strategy under very weak conditions. The proof builds on the standard connection between minimax risk and an equilibrium of a specific zero-sum game. This allows us to interpret any minimax risk strategy as a particular Bayes solution and to connect minimax risk to 'rational' or Bayes decision making.

Under 'rational' or Bayes decision making the true distribution P is assumed to be drawn from a known probability distribution (or prior) Q, so $Q \in \Delta\mathcal{P}$ (von Neumann and Morgenstern, 1944). The strategy f that minimizes expected risk $R(f,Q) = \int_P R(f,P)\,dQ(P)$ is called a *Bayes solution*, the value of the minimum expected risk is called the *Bayes risk*. Q^* is called a *least favorable prior* if $Q^* \in \arg\max_{Q\in\Delta\mathcal{P}} \inf_f R(f,Q)$.

Proposition 3 *Assume that W as a function of μ and z is continuous and that $\mathcal{G}$ is a metric space.*

(i) There exists a symmetric minimax risk strategy and a least favorable prior containing only binary valued distributions in its support.

(ii) Any minimax risk strategy minimizes Bayes risk under any least favorable prior.

Proof. Following von Neumann Morgenstern (1944, see also Savage, 1954), we solve minimax risk by considering the following simultaneous move zero-sum game between the decision maker and nature. The decision maker chooses an strategy f and nature chooses a distribution $Q^b \in \Delta\mathcal{P}^b$ over binary valued distributions $P^b \in \mathcal{P}^b$. The payoff to the decision maker is given by $-R\left(f,Q^b\right)$ while that of nature by $R\left(f,Q^b\right)$. Under the above assumptions there exists a Nash equilibrium (or *saddle point*) $\left(f^*,Q^{b*}\right)$ of this game (Glicksberg, 1952) which means that $R\left(f^*,Q^b\right) \leq R\left(f^*,Q^{b*}\right) \leq R\left(f,Q^{b*}\right)$ holds for all f and all Q^b. Now we use the well known minimax theorem of zero-sum games, to conclude that $R\left(f^*,Q^{b*}\right) = \min_f \sup_{Q^b\in\Delta\mathcal{P}^b} R\left(f,Q^b\right) = \min_f \sup_{P^b\in\mathcal{P}^b} R\left(f,P^b\right)$. Following Proposition 1(ic), the binomial transformation f^{*b} of f^* attains minimax risk.

Concerning (ii), note that $\left(f^*,Q^{b*}\right)$ is also a saddle point of the game where nature is allowed to choose any prior $Q \in \Delta\mathcal{P}$. Applying the minimax theorem for zero-sum games we find that $R\left(f^*,Q^{b*}\right) = \max_{Q\in\Delta\mathcal{P}} \inf_f R(f,Q)$ which means that Q^{b*} is a least favorable prior. ■

Combining Corollary 1 and Proposition 3 we obtain:

Corollary 2 *Assume that $\mathcal{G}$ is a Euclidean space, W is continuous in μ and z and convex in z. Then there is a symmetric linear non-randomized strategy that attains minimax risk.*

2.3 Estimation

We apply the results in Proposition 1 to estimation. Let g be some real valued function of μ to be estimated where $W(z,P) = W_1(z, g(\mu))$ for some non-negative function W_1 with $W_1 = 0$ if and only if $z = g(\mu)$. The natural candidate for W_1 is quadratic loss in which case $W(z,P) = (z - g(\mu))^2$.

2.3.1 Unbiased Estimation

We first consider unbiased estimation when $K = 1$. It is easily verified for $k \in \{1, .., N\}$ that f_k defined by

$$f_k\left(y^{1,N}\right) = \frac{1}{\binom{N}{k}} \sum_{|\{n_1,..,n_k\}|=k} \prod_{i=1}^{k} y^{(n_i)}$$

is the unique symmetric non-randomized unbiased estimator of μ^k (for $k = 1, 2$ see Lehmann, 1983). By Corollary 1 f_k attains minimax risk for any convex loss function. One can easily extend this result to a minimax risk estimate of any polynomial of the mean μ of degree at most N. It is also known that only for such functions of the mean an unbiased estimator exists (Lehmann, 1983).

Estimates are only of little value per se without knowing about maximal risk. Following Proposition 1, maximal risk is attained among the binary valued distribution. For the mean we verify numerically by searching among the Bernoulli distributions that the maximal expected distance between the estimator and the true mean (i.e. the value of minimax risk when $W(z,P) = |z - p|$) is below 0.1 if and only if $N \geq 16$ and below 0.05 if and only if $N \geq 64$.

Of course, if loss is not convex then we do not expect there to be a minimax risk unbiased estimator that is non-randomized. For instance, no non-randomized minimax risk unbiased estimator exists under loss $W_1 = |z - \mu|^s$ if $0 < s < 1$. This follows from a proof of Hodges and Lehmann (1950, Theorem 3.4) presented for general estimators but which is also valid for unbiased estimators.

Note that our transformation method can also be applied to sequential estimation where the number of observations is endogenous. For example, when $K = 1$, by

independently transforming each observation and otherwise performing inverse binomial sampling on the transformed data one can obtain a linear symmetric unbiased estimator of $1/\mu$ (see Lehmann, 1983, Example 3.2).

2.3.2 Biased Estimation

Nonparametric minimax risk estimators under quadratic loss have been obtained by Hodges and Lehmann (1950, Theorems 6.1 and 6.4) for the mean of a single sample and for the difference between two means of two independent samples with the same size, provided outcomes are contained in $[0,1]$. In their proof, first the binomial case is solved and then properties of quadratic loss are used.

Following Corollary 1(ii) one can solve for minimax risk when loss is convex by confining attention to binary valued distributions and then working with the expected estimate under its binomial transformation. While minimax risk can be solved for the binary valued case by searching for a saddle point, closed form solutions can be difficult to obtain.

Corollary 1(ii) can also be used to reduce the value of maximal risk even if minimax risk is not known. For example, consider $K = 2$ with paired data and let $g(\mu) = \max\{\mu_1, \mu_2\}$. Instead of choosing as estimate the larger of the two average payoffs, i.e. $\max\left\{\frac{1}{N}\sum_{n=1}^{N} y_1^{(n)}, \frac{1}{N}\sum_{n=1}^{N} y_2^{(n)}\right\}$, maximal risk is lower if one instead first transforms the data binomially and then applies this estimator.

2.4 Hypothesis Testing

We apply our results to hypothesis testing. Throughout we assume either that $\mathcal{P} = \Delta[0,1]^K$ or $\mathcal{P} = (\Delta[0,1])^K$. Consider within the space of distributions $\mathcal{P}$ the objective of testing the *null hypothesis* $H_0 : \mu \in \Omega_0$ against the *alternative hypothesis* $H_1 : \mu \in \Omega_1$ where $\Omega_0, \Omega_1 \subset [0,1]^K$ are given with $\Omega_0, \Omega_1 \neq \emptyset$ and $\Omega_0 \cap \Omega_1 = \emptyset$. Selecting hypothesis H_i will be identified with i, hence $\mathcal{G} = \{0,1\}$.

We recall some standard concepts and notation. Strategies are referred to as *tests*. The *power function* β of the *test* f is then given by $\beta_f(P) = \Pr_P(f = 1)$. The test f has level α if $\beta_f(P) \leq \alpha$ for $P \in \Omega_0$ where f is *unbiased* if additionally $\beta_f(P) \geq \alpha$ whenever $\mu(P) \in \Omega_1$. $\sup_{P\in\Omega_0} \beta_f(P)$ is called its size (or type I error) and $\sup_{P\in\Omega_1}\left(1 - \beta_f(P)\right)$ its type II error. The test f is *equi-tailed* with level α for some given $\Omega_e \subset [0,1]^K$ if there are two level $\alpha/2$ tests f_1 and f_2 against the alternatives $\mu \in \Omega_1 \cap \Omega_e$ and $\mu \in \Omega_1 \backslash \Omega_e$ respectively such that $f\left(y^{1,N}\right) = f_1\left(y^{1,N}\right) + f_2\left(y^{1,N}\right)$. f is *uniformly most powerful* (UMP) if for any level α test $\tilde{f}$ and any P such that

$\mu(P) \in \Omega_1$ it follows that $\beta_f(P) \geq \beta_{\tilde{f}}(P)$.

Introducing a new concept, we call a test f *parameter most powerful (PMP) with level* α if for any alternative level α test $\tilde{f}$ and for any $\mu' \in \Omega_1$ it follows that $\min_{P:\mu(P)=\mu'} \beta_f(P) \geq \min_{P:\mu(P)=\mu'} \beta_{\tilde{f}}(P)$. Thus, f is a PMP test if it is a *maximin test* (Lehmann and Romano, 2005, ch. 8) for any set of alternatives that only depends on the mean of the underlying distributions.

Given Proposition 1 it is immediate how to construct distribution-free tests from tests for binary valued distributions. Remember that f^b denotes the binomial transformation of the strategy f and that given P we have defined P^b as the binary valued distribution that has the same mean vector as P.

Proposition 4 *(i) If f is a level α test for $P \in \mathcal{P}^b$ then f^b is a level α test for $P \in \mathcal{P}$, the property of f being unbiased or equi-tailed on $\mathcal{P}^b$ carries over to f^b for $\mathcal{P}$.*

(ii) If f is a UMP test on $\mathcal{P}^b$ then f^b is a PMP test respectively for $P \in \mathcal{P}$. This statement also holds if one restricts attention to either unbiased or to equi-tailed tests.

Proof. We first show that $\beta_f\left(P^b\right) = \beta_{f^b}(P)$ so $\left\{\beta_{f^b}(P), P \in \mathcal{P}\right\} \subseteq \left\{\beta_f(P), P \in \mathcal{P}\right\}$ and refer to this as (o). This statement follows directly from the definitions as $\beta_f\left(P^b\right) = \Pr_{P^b}(f=1) = \Pr_P\left(f^b=1\right) = \beta_{f^b}(P)$. Concerning part (i) let f be a level α test for binary valued distributions and consider P such that $\mu(P) \in \Omega_0$ so $\beta_f\left(P^b\right) \leq \alpha$. Then (o) implies that $\beta_{f^b}(P) \leq \alpha$ which implies that f^b is a level α test for all $P \in \mathcal{P}$. Now let f be unbiased for all $P \in \mathcal{P}^b$ and let $\mu(P) \in \Omega_1$ so $\beta_f\left(P^b\right) \geq \alpha$. Then (o) shows that $\beta_{f^b}(P) \geq \alpha$ so it follows that f^b is unbiased for all $P \in \mathcal{P}$. The statement for equi-tailed tests follows similarly. For proof of part (ii) let f be a uniformly most powerful test for binary valued distributions and let $\tilde{\mu} \in \Omega_1$. Let $\tilde{f}$ be a level α test for all $P \in \mathcal{P}$. Then $\beta_{f^b}(P) = \beta_f\left(P^b\right) \geq \beta_{\tilde{f}}\left(P^b\right) \geq \min_{P:\mu(P)=\tilde{\mu}} \beta_{\tilde{f}}(P)$ and hence f is a parameter most powerful test for all $P \in \mathcal{P}$. The proof of the statement restricted to unbiased or to equi-tailed tests follows similarly. ■

One may choose to comment on the efficiency of a test in terms of the number of samples it needs to obtain a given size. To simplify exposition consider $K = 1$ and restrict attention to deterministic sample sizes. Performance can sometimes be (marginally) improved by choosing sample size randomly.

Let $N(\beta_0, \mu_0, f)$ be the smallest number of observations needed by the test f to achieve power of at least β_0 for all distributions P that have mean μ_0 where μ_0 belongs to the set of alternatives. We adapt the standard definition of relative efficiency (cf. Lehmann and Romano, 2005, chapter 13.2, p. 534) to the set of alternatives that

all have the same mean. Holding β_0 and μ_0 fixed we call $N(\beta_0,\mu_0,f)/N\left(\beta_0,\mu_0,\tilde{f}\right)$ the *relative parameter efficiency* of f relative to $\tilde{f}$. We call f^* *parameter efficient* if there is no alternative test f that has size at most equal to that of f^* such that $N(\beta_0,\mu_0,f) < N(\beta_0,\mu_0,f^*)$. The test is called *parameter efficient unbiased* if the above condition is only checked for tests f that are unbiased. This leads to the following observation.

Corollary 3 *Any PMP test is parameter efficient. This is also true of one limits attention to unbiased or equi-tailed tests.*

Similar statements can be made when $K \geq 2$ by increasing the sample sizes of each variable proportionally.

Consider now confidence sets where we specifically allow for randomized confidence sets. Consider a family of hypotheses that are indexed by μ'. $S = \left\{S\left(\tilde{y}^{1,N}\right), \tilde{y}^{1,N} \in \widehat{\mathcal{Y}}\right\}$ is a *family of confidence sets at level* α if $\Pr_P\left(\mu' \in S\left(\tilde{Y}^{1,N}\right)\right) \geq 1-\alpha$ for all $P \in \mathcal{P}$ such that $\mu \in \Omega_0(\mu')$. $1-\alpha$ is also called the *coverage*. The set of all such families will be denoted by $F_\alpha(\mathcal{P})$. S is *unbiased* if $\Pr_P\left(\mu' \in S\left(\tilde{Y}^{1,N}\right)\right) \leq 1-\alpha$ for all $P \in \mathcal{P}$ such that $\mu \in \Omega_1(\mu')$. S is called *equi-tailed* for some given $\Omega_e \in [0,1]^{[0,1]}$ if there exist families S_1 and S_2 such that $\Pr_P\left(\mu' \in S_1\left(\tilde{Y}^{1,N}\right)\right) \geq 1-\alpha/2$ for all $P \in \mathcal{P}$ such that $\mu \in \Omega_0(\mu') \cap \Omega_e(\mu')$ and $\Pr_P\left(\mu' \in S_2\left(\tilde{Y}^{1,N}\right)\right) \geq 1-\alpha/2$ for all $P \in \mathcal{P}$ such that $\mu \in \Omega_0(\mu') \setminus \Omega_e(\mu')$ where $S\left(\tilde{Y}^{1,N}\right) = S_1\left(\tilde{Y}^{1,N}\right) \cap S_2\left(\tilde{Y}^{1,N}\right)$. With this definition of equi-tailed-ness we obtain equivalence in the usual sense between a collection of equi-tailed hypotheses tests and a family of equi-tailed confidence sets. A family of confidence intervals S at level α is *uniformly most accurate* if $\Pr_P\left(\mu' \in S\left(\tilde{Y}^{1,N}\right)\right)$ is minimized among all families at level α for each $\mu \in \Omega_1(\mu')$. A family of UMP tests can be used to construct a uniformly most accurate family by collecting the parameters that can not be rejected given the observed data. As a new definition we say that S is *parameter most accurate* if $\max_{P:\mu=\bar{\mu}} \Pr_P\left(\mu' \in S\left(\tilde{Y}^{1,N}\right)\right)$ is minimized among all families at confidence level α for each $\bar{\mu} \in \Omega_1(\mu')$.

For $K = 1$ consider more specifically a family S of confidence intervals for the mean, so $S\left(y^{1,N}\right) = \left[l\left(y^{1,N}\right), u\left(y^{1,N}\right)\right]$ for some $l, u \in [0,1]$. $L_S\left(y^{1,N}\right) = \left(u\left(y^{1,N}\right) - l\left(y^{1,N}\right)\right)$ is then called the *length* (or *width* or *accuracy*) of the confidence interval under S given $y^{1,N}$.

$$E_P L_S\left(Y^{1,N}\right) = \sum_{y^{1,N}} \Pr_P\left(Y^{1,N} = y^{1,N}\right)\left(u\left(y^{1,N}\right) - l\left(y^{1,N}\right)\right)$$

is the *expected length* of the family of confidence intervals S given P. It is well known that a uniformly most accurate family of confidence intervals for the mean minimizes the expected length of the confidence interval for each $\mu \in \mathcal{P}$ (Pratt, 1961) among all families of confidence intervals with the specified coverage.

Consider now a family of lower bounds l for the mean so $l = l\left(y^{1,N}\right)$ and $S = \left[l\left(y^{1,N}\right), 1\right]$ is a family of confidence intervals for the mean. Let

$$E_P l_S\left(Y^{1,N} | \mu \geq l\right) = \sum_{y^{1,N}:\mu \geq l(y^{1,N})} \Pr_P\left(Y^{1,N} = y^{1,N} | \mu \geq l\left(Y^{1,N}\right)\right) \cdot l\left(y^{1,N}\right).$$

Then $\mu - E_P l_S\left(Y^{1,N} | \mu \geq l\right)$ is a measurement for degree of underestimating μ. A uniformly most accurate family of lower bounds minimizes this value of underestimation (Pratt, 1961) among all families of lower confidence bounds with the specified coverage. Similarly one can state definitions and results for upper bounds.

In the following we show how to transform families of confidence sets for binary valued distributions into ones for distributions in $\Delta[0,1]^K$. Then we show for $K = 1$ that a family of uniformly most accurate confidence intervals can be transformed into a family that maximizes the minimum accuracy for covering the mean among all families of confidence intervals at level α. Similarly, a family that maximizes the coverage among all Bernoulli distributions and among all families with some given maximal length l can be transformed into a family that has this property for all distributions in $\Delta[0,1]$.

Proposition 5 *Consider $t \in \mathcal{T}$. If $S \in \mathcal{F}_\alpha\left(\mathcal{P}^b\right)$ then $S \circ t \in \mathcal{F}_\alpha(\mathcal{P})$. If S is unbiased for $\mathcal{P}^b$ then $S \circ t$ is unbiased for $\mathcal{P}$. If S is uniformly most accurate for $\mathcal{P}^b$ then $S \circ t$ is parameter most accurate for $\mathcal{P}$.*

Similar results can be stated for upper bounds for the mean and if one limits attention to specific families such as those that are unbiased or equi-tailed. Analogous statements can be made relating to confidence intervals and bounds for the difference between two means.

Proposition 5 can be applied whenever there are UMP tests for binary valued data within some class of tests. We remind the reader of some settings where these exist. A UMP test exists for the one-sided test of the mean of a single sample. A UMP unbiased test exists for the one-sided test for comparing two independent samples as well as for comparing two dependent samples of equal size. Analogous UMP two-sided tests exist if one confines attention additionally to equi-tailed tests.

Sometimes it is easier to apply Proposition 1 directly. Assume for instance that one is interested in finding a family of confidence intervals of the mean that maximizes coverage for a given level and a given maximal length. Then Proposition 1 shows that one can limit attention to Bernoulli distributions. Combining this with the numerical analysis of Lutsenko and Maloshevskii (2003, Table 1) we conclude for $N = 15$ that there is a family of confidence intervals at level 5% with maximal length 0.4 but not with maximal length equal to 0.38.

2.4.1 Non-Randomized Tests

When interested in making specific recommendations then one is often only interested in non-randomized tests and confidence intervals. Following Proposition 2 one can select the more likely recommendation of the randomized test to obtain a non-randomized test, formally defined as f^m in Section 2.1. The proof of Proposition 2 reveals the following statement.

Corollary 4 *If f is a randomized hypothesis test for testing the null $\mu \in \Omega_0$ against the alternative $\mu \in \Omega_1$ with size α and type II error γ then f^m has size bounded above by 2α and type II error bounded above by 2γ.*

Such a non-randomized test can then be used to construct a non-randomized confidence bound or confidence interval. Notice that the non-randomized test will not be unbiased and it is not known whether equi-tailed-ness is preserved.

2.4.2 An Example

Consider the 'anti-self-dealing' indices in Djankov et al. (2005, Table III) gathered for 72 countries to measure minority shareholder protection against self-dealing (i.e. investor expropriation) of a controlling shareholder. In the first two rows of Table 1 we present the average anti-self-dealing indices across different regions characterized by the origin of their law system together the number of observations. Indices belong to $[0, 1]$ by construction.

Table 1: Anti-Self-Dealing Indices across Regions

		Civil Law			
	Common Law	French Origin	German Origin	Scandinavian Origin	Overall
N	21	32	14	5	51
Mean	0.67	0.35	0.39	0.39	0.36

Djankov et al. (2005) use the t test to compare these indices. We choose distribution-free nonparametric methods. Assume independence of indices across countries and assume that the indices gathered for common law are iid draws from a random variable Y_1 and similarly that the indices within the countries governed by civil law represent iid draws from a random variable Y_2. We wish to test $H_0 : EY_1 = EY_2$ against $H_1 : EY_1 \geq EY_2$ under these assumptions at level 5%.

Consider the randomized PMP unbiased test derived by applying the exact randomized Fisher-Tocher test (Fisher, 1935, Tocher, 1950) after binomially transforming the sample. We find that the null hypothesis is rejected with probability 0.76. For the associated deterministic test we need to evaluate the above test at level 2.5% and find that the null hypothesis is rejected with probability 0.65 which is strictly greater than 0.5 and hence we obtain a significant difference at 5% between civil law and common law. When considering only those among civil law with French origin we find that the null is rejected with probability 0.62 at 2.5% and hence also a significant difference at 5% between those with French origin and common law. The difference between those with German origin and common law is not even significant at level 10%. One could alternatively construct a deterministic test using Hoeffding (1963, Corollary, eq. 2.7). This test only finds the difference between common law and civil law to be marginally significant at 10%.

The value of our deterministic test as compared to other deterministic nonparametric tests is that we can evaluate the quality of inference in terms of its type II error. Such an evaluation is important as strictly speaking one has to first choose the test before gathering data. We search for the minimal difference d between EY_1 and EY_2 necessary to ensure type II error of 20% when comparing civil law and common law. For our randomized test we find $d \approx 0.32$, for our deterministic test $d \approx 0.41$ and for the test based on the Hoeffding bounds one easily derives $d \approx 0.55$.

Alternatively one can compare these tests according to their relative parameter efficiency. We do this by increasing the number of observations proportionally until matching the type II error of 20% achieved by the randomized PMP unbiased test for the alternative hypothesis $H_1 : EY_1 - EY_2 \geq d$ (with $d \approx 0.32$) given $N_1 = 21$ and $N_2 = 51$. For our deterministic test we need $N_1 = 36$ and $N_2 = 87$ while the test based on the Hoeffding bounds requires $N_1 = 63$ and $N_2 = 152$, the corresponding values of relative parameter efficiency are 59% and 34% respectively.

For the design of future investigations one may wonder how many observations are necessary to ensure type II error of 0.15 given size 0.05 with an unbiased test of $H_0 : EY_1 = EY_2$ against $H_1 : EY_1 - EY_2 \geq 0.25$. Evaluating again the power of the

Fisher-Tocher test we find that $N_1, N_2 \leq 56$ does not suffice while our randomized PMP unbiased test has this property when $N_1 = N_2 = 57$.

Next investigate lower confidence bounds for the mean of the common law index. First we follow Bickel et al. (1989, see also Bickel, 1992) who use an alternative bound of Hoeffding (1963, eq. 2.1) to derive a 95% deterministic lower confidence bound. For the common law index the value is equal to 0.405 as the solution to

$$\left(\left(\frac{1-x}{1-0.67} \right)^{1-0.67} \left(\frac{x}{0.67} \right)^{0.67} \right)^{21} = 0.05.$$

Next we derive our deterministic lower bound. For a given μ_0 the data is first randomly binomially transformed and then the uniformly most powerful test indexed by μ_0 of $H_0(\mu_0) : \mu \leq \mu_0$ against $H_1(w) : \mu > \mu_0$ for Bernoulli distributions is evaluated (see Rohtagi, 1976, Example 2, p. 415). We search for the value of μ_0 such that the null is rejected with probability 0.5. This is then the lower confidence bound, in this case $\mu_0 = 0.455$. Finally we consider the parameter most accurate randomized lower bound derived from the above family of uniformly most powerful tests. This randomized lower confidence bound is unbiased as the underlying test for binary values is unbiased. We plot the density of the randomized lower bound in Figure 1. Its expected value is equal to 0.49. It lies above the Hoeffding lower bound 0.405 with probability 0.82, above our deterministic lower bound 0.455 with probability 0.66 and above the average index 0.67 with probability 0.03.

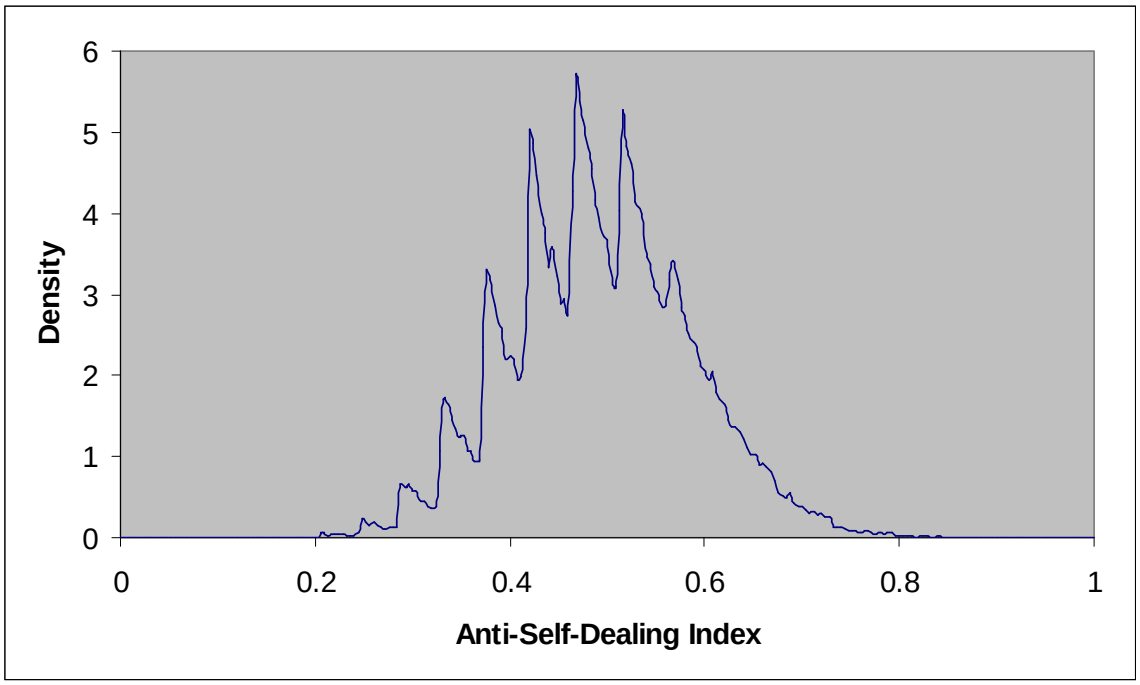

Figure 1: Density plot of randomized lower confidence bound.

One could similarly consider confidence intervals for the index of common law. Using Blyth and Hutchinson (1960, Table 1) we verify that the maximal expected length of any family of unbiased confidence intervals is at least 0.41 given that we have 21 independent observations of an unknown mean. Alternatively one could consider equi-tailed two-sided tests.

Finally we investigate whether there is sufficient evidence to be able to establish that the French and German indices are similar. We do this by deriving the type II error of the test that the average indices among French and German origin are the same against the alternative that the German index is drawn from a higher mean. The finding is that a type II error of 20% is only ensured if the true difference is at least 0.39. Thus we conclude that there are not sufficiently many observations to infer any meaningful statement about an underlying process from the similarity between the empirical means of the French and German indices.

3 Concern for Variance

One may similarly be interested in variances or functions of the variances of underlying random variables. The setting is as in Section 2 except that now we are interested in

loss functions that only depend on the variance vector σ^2. We show how to generate estimates and tests for which an upper bound on risk can be derived.

The insight for how to use some of our previous results is a "combination method" founded in Walsh (1962, ch. 7, p. 119). To keep notation simple consider $K = 1$ and N even.

Proposition 6 *If $Y^{1,N}$ are N are iid observations of a random variable Y that has range $[0,1]$ and if π is a permutation of $\{1,..,N\}$ then $Z^{(1)},..,Z^{(N/2)}$ are iid random variables with range $[0,1/2]$ with $\mu\left(Z^{(j)}\right) = \sigma^2\left(Y^{(i)}\right)$ where $Z^{(j)} = \frac{1}{2}\left(Y^{(\pi(2j-1))} - Y^{(\pi(2j))}\right)^2$ for $k = 1,..,N/2$.*

3.1 Estimation

Using Proposition 6 one can design estimates of functions of the variance and measure their maximal risk. For completeness we first briefly show how to measure maximal risk underlying unbiased estimation of variance. A symmetric non-randomized unbiased estimate of the mean underlying $\left\{Z^{(j)}\right\}$ will be a symmetric non-randomized estimate of the variance underlying $\left\{Y^{(i)}\right\}$. Given uniqueness this estimate must be equal to the classic symmetric unbiased estimate S^2 of the variance, mathematically we find

$$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{2}\left(Y^{(i)} - Y^{(j)}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y^{(i)} - \frac{1}{N}\sum_{j=1}^{N}Y^{(j)}\right)^2 =: S^2.$$

The value of the transformation in Proposition 6 is that we can derive an upper bound on risk, here we do this analytically for expected quadratic loss. We know that the maximal quadratic loss, when taking the empirical average as estimator of the mean of random variable $X \in [0,1]$ based on M independent observations, is equal to $1/(4M)$. Thus an upper bound on the expected quadratic loss of the estimate S^2 of σ_Y^2 is equal to $(1/2)^2 * 1/(4(N/2)) = 1/(8N)$.

Consider now estimation of standard deviation σ_Y for which it is known that there is no unbiased estimate. Typically one estimates σ_Y using either the empirical standard deviation or the square root of the unbiased estimate of the variance S. In the following we construct a uniformly consistent estimator that performs better in small samples (numerical simulations carried out for $N \leq 1600$). Transform the data as in Proposition 6. Then transform the $[0,1/2]$ data into $\{0,1/2\}$ data using a random mean preserving transformation. Next choose an estimate g of $\sqrt{\mu_X}$ that is based on

$N/2$ independent observations of a random variable $X \in \{0, 1/2\}$. Let n denote the number of times $1/2$ occurs. We choose as estimate the root of a particular convex combination of the sample average n/N and the median $1/4$, namely by setting

$$g(n) = \sqrt{\left(1 - \frac{1}{N+3.5}\right)\frac{n}{N} + \frac{1}{N+3.5} * \frac{1}{4}}$$

for $n = 0, .., N/2$. Maximal risk is easily derived as we know that it is attained among the Bernoulli distributions. For instance, maximal risk is approximately $1/4\,(N+1)$ for $34 \leq N \leq 120$. More specifically we find that $N \geq 100$ is necessary to obtain maximal risk below $(0.05)^2$.

For comparison lower bounds of the two alternative estimators mentioned above are established by searching numerically among the Bernoulli distributions. We find $N \geq 181$ to be necessary for either of them to achieve risk below $(0.05)^2$ with no upper bound on N available. More generally, for $N \leq 500$ these two other estimates would require at least 70% more observations to generate the same maximal risk as our estimate.

We use the above estimate of standard deviation to measure income inequality in the US between 1990 and 2002 using the Panel Study of Income Dynamics (PSID) at the family level. While alternative measures of dispersion such as the Gini coefficient and the coefficient of variation exist, we do not know of upper bounds on their risk. We explain the findings summarized in Table 2. Attention is restricted to race indicated as white, age of head between 20 and 59, work hours per year ranging 520 to 5096 and average hourly earnings greater than half the minimum wage with all of these criteria satisfied for at least two consecutive years. Earnings are at household level measured in 1992 dollars. Top coded earnings are not included which means that we estimate standard deviation conditional on earnings belonging to $[0, 10^6]$ (only one observation in the data set was top coded as it was above 10^6). S and $\hat{\sigma}$ indicate the empirical standard deviation and our estimate.

Table 2: Estimating Income Inequality

	1990	1991	1992	1993	1994	1995	1996	1998	2000	2002
N	2252	2786	2306	1554	1910	2299	1683	1518	1570	1582
mean	34288	33888	36526	40940	37621	37734	38991	39193	42988	40405
S	32786	29455	33492	45158	38634	41137	43543	42093	56747	46653
$\hat{\sigma}$	32868	29460	33526	45589	38986	41449	43644	42395	57205	46934
$\sqrt{risk}$	10780	9695	10651	12961	11705	10672	12470	13122	12907	12845

3.2 Hypothesis Testing

One can also use Proposition 6 to design hypothesis tests and confidence intervals. Proceed as follows to ensure symmetry. For each variable choose a permutation equally likely among all permutations and transform all observations in of this variable as in Proposition 6. Double all observations to obtain data with range $[0, 1]$. Independently binomially transform each observation to obtain binary valued data. Then evaluate a uniformly most powerful test for the means. An upper bound on the size and type II error of the resulting randomized test is easily derived. Either one can then apply this randomized test or one can proceed by transforming it into a non-randomized test as shown in Corollary 4. Notice that we do not expect the randomized test to be parameter most powerful.

4 Conclusion

In this paper we expand on a specific randomization method from Schlag (2003) that can also be found in Cucconi (1968) and in Gupta and Hande (1992) and show how it can be used to select estimators, hypothesis tests and confidence intervals. To apply this method, one needs to know a bounded set that contains all payoffs and one should only be interested in the underlying means. It can be very natural to know such exogenous bounds on the possible payoffs as outcomes are often measured on a bounded scale. Following Bahadur and Savage (1956), minimax risk estimators and most powerful tests typically do not exist if no assumptions are made. Moreover, as our approach is nonparametric and distribution-free, constraints on the data have to be based on knowledge, not on beliefs.

We included various different ways to randomly transform the data as we expect those that create correlation between samples to perform better in applications as they tend to reduce the variance in the transformed sample. In specific applications we have also found other transformations that do not treat independent observations independently to be useful (see results surrounding the correlated binomial average rule in Eozenou et al., 2006).

The outcome under any of our transformations is randomized. When concerned with estimation and loss is convex then non-randomized estimators are also derived. However the tests and confidence intervals contained in this paper that generate the lower bounds are truly randomized. If one is not willing to follow such a randomized recommendation we provide non-randomized tests that then can be used to generate

non-randomized confidence intervals and bounds.

We find that minimax risk allows to select "simple" strategies in the sense that they are linear and symmetric. The suggested randomized and non-randomized tests are very simple to evaluate using Monte Carlo simulations as they are based on the most basic UMP tests of statistics.

The novel concepts of parameter most powerful, parameter efficiency and parameter most accurate are introduced. They are motivated by reducing statements to the parameters of interest which in this paper are the underlying means. If distributions are conceptually collected into equivalent classes consisting of all those that have the same mean vector these concepts are equivalent to UMP, efficiency and uniformly most accurate. We see no need to differentiate distributions other than by their mean vectors when means are the only parameters of interest. Of course any of the minimax risk strategies mentioned in this paper need not be unique, future research should focus on alternative methods for dealing with interior payoffs without loosing minimax risk properties. The key of this paper as a starting point is that for the first time - except for the previous results on estimates of a mean and of difference between two means and for the recently discovered work on selection procedures by Gupta and Hand (1992) - we now know a least favorable prior.

The randomization method has started to spread in the literature on minimax regret (Eozenou et al., 2006, Schlag, 2006b, Stoye, 2006). The rounding trick of Gupta and Hande (1992) uncovered when working on this paper will similarly enter as shown by Schlag (2007). As mentioned in an earlier version of this paper (Schlag, 2006a) the randomization method reduces maximal Hannan regret in non-stationary decision problems (Hannan, 1957, cf. Auer et al., 2002).

The results of this paper can also be applied when the k-th moment is the only parameter of interest for some fixed k. First take each observation to the power k and then investigate the means of the transformed data.

In this paper we have also shown how to investigate variances by building on a combination method found in Walsh (1962). It is an open question how close these results are to minimax risk.

References

[1] Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (2002), "The Nonstochastic Multiarmed Bandit Problem," *SIAM Journal on Computing,* 31, 48–77.

[2] Bahadur, R. R. and Savage, L. J. (1956), "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical. Statistics,* 27, 1115–1122.

[3] Bechhofer, R. E. and Kulkarni, R. V. (1982), "Closed Adaptive Sequential Procedures for Selecting the Best of $k \geq 2$ Bernoulli Populations," In *Proceeding of the Third Purdue Symposium on Statistical Decision Theory and Related Topics*, eds. S. S. Gupta and G. Berder, New York: Academic Press, pp. 61–108,.

[4] Bickel, P. (1992), "Inference and Auditing: the Stringer Bound," *International Statistical. Review,* 60**,** 197–209.

[5] Bickel, P., Godfrey, J., Neter, J. and Clayton, H. (1989), "Hoeffding Bounds for Monetary Unit Sampling in Auditing," *International Statistical Institute, Contributed Paper, Paris Meeting.*

[6] Blyth, C. R. and Hutchinson, D. W. (1960), "Table of Neyman-Shortest Unbiased Confidence Intervals for the Binomial Parameter," *Biometrika,* 47, 381–391.

[7] Cucconi, O. (1968), "Contributi all'Analisi Sequenziale nel Controllo di Accettazione per Variabili" (in Italian), *Atti dell' Ass. Italiana per il Controllo della Qualità*, 6, 171–186.

[8] Djankov, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. (2005), "The Law and Economics of Self-Dealing," NBER Working Paper 11883.

[9] Diouf, M. A. and Dufour, J.-M. (2006), "Exact Nonparametric Inference for the Mean of a Bounded Random Variable," in *American Statistical Association Proceedings of the Business and Economic Statistics Section.*

[10] Dufour, J.-M.(2003), Identification, Weak Instruments, and Statistical Inference in Econometrics," *Canadian Journal of Economics*, 36, 767-808.

[11] Eozenou, P., Rivas, J. and Schlag, K. H. (2006), "Minimax Regret in Practice - Four Examples on Treatment Choice," unpublished mansuscript, European University Institute, Economics Department.

[12] Fisher, R. A. (1935), "The Logic of Inductive Inference," *J. Roy. Stat. Soc.,* 98, 39–54.

[13] Fraser, D. A. S. (1957), *Nonparametric Methods in Statistics*, New York: John Wiley and Sons.

[14] Glicksberg, I. L. (1952), "A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points," in *Proceedings of the American Mathematical Society,* pp. 170–174.

[15] Gupta, S. S. and S. N. Hande (1992), "On Some Nonparametric Selection Procedures," *Nonparametric Statistics and Related Topics*, A.K.Md.E. Saleh (Editor), Amsterdam: Elsevier, 33–49.

[16] Hannan, J. (1957), "Approximation to Bayes Risk in Repeated Plays," in *Contributions to the Theory of Games of Games*, Vol. 3, eds. M. Dresher, A. W. Tucker and P. Wolfe, Princeton: Princeton Univ. Press, pp. 97–139.

[17] Hodges, J. L. Jr. and Lehmann, E. L. (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics,* 21, 182–197.

[18] Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association,* 58, 13–30.

[19] Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: John Wiley and Sons.

[20] Lehmann, E. L. and Loh, W-Y. (1990), "Pointwise verses Uniform Robustness in some Large-Sample Tests and Confidence Intervals," *Scandinavian Journal of Statistics,* 17, 177–187.

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

[22] Lutsenko, M. M. and Maloshevskii, S. G. (2003), "Minimax Confidence Intervals for the Binomial Parameter," *Journal of Statistical Planning and Inference,* 113, 67–77.

[23] Pesarin, F. (1984), "On Randomized Statistical Procedures," in *Proceedings of the Seventh Conference on Probability Theory*, ed. M. Iosifescu, Editura Academiei Republicii Romania, pp. 295–306.

[24] Pratt, J. W. (1961), "Length of Confidence Intervals," *Journal of the American Statistical Association,* 56, 549–567.

[25] Rohtagi, V. K. (1976), *An Introduction to Probability Theory and Mathematical Statistics*, New York: John Wiley and Sons.

[26] Romano, J. P. and Wolf, M. (2000), "Finite Sample Non-Parametric Inference and Large Sample Efficiency," *Annals of Statistics,* 28, 756–778.

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

[28] Schlag, K. H. (2003), *How to Minimize Maximum Regret under Repeated Decision-Making*, unpublished mansuscript, European University Institute, Economics Department.

[29] Schlag, K. H. (2006a), *Designing Nonparametric Estimates and Tests for Means*, Working Paper ECO 2006/26**,** European University Institute, Economics Department**.**

[30] Schlag, K. H. (2006b), *Distribution-Free Learning*, Working Paper ECO 2007/1**,** European University Institute, Economics Department.

[31] Schlag, K. H. (2007), *Eleven - Tests needed for a Recommendation*, unpublished manuscript, European University Institute, Economics Department.

[32] Sobel, M. and Huyett, M. J. (1957), "Selecting the One Best of Several Binomial Populations," *Bell System Technical Journal,* 36, 537–576.

[33] Stoye, J. (2006), *Minimax Regret Treatment Choice with Finite Samples*, unpublished manuscript, New York University, Economics Department.

[34] Tocher, K. D. (1950), "Extention of the Neyman-Pearson Theory of Tests to Discontinuous Variates," *Biometrika*, 37, 130–144.

[35] von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press.

[36] Wald, A. (1947), *Sequential Analysis*, New York: John Wiley and Sons.

[37] Wald, A. (1950), *Statistical decision functions*, New York: John Wiley and Sons.

[38] Walsh, J. E. (1962), *Handbook of Nonparametric Statistics, Investigation of Randomness, Moments, Percentiles, and Distributions*, Princeton: D. van Nostrand Company Inc.