

Exact Hypothesis Testing without Assumptions -  
New and Old Results not only for Experimental  
Game Theory<sup>1</sup>

Karl H. Schlag

August 9, 2010

<sup>1</sup>The author would like to thank Caterina Calsamiglia, Patrick Eozenou and Joel van der Weele for comments on exposition of an earlier version.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Two Independent Samples</b>	<b>3</b>
2.1	Preliminaries . . . . .	3
2.1.1	Documentation . . . . .	7
2.1.2	Accepting the Null Hypothesis . . . . .	8
2.1.3	Side Remarks on the P Value . . . . .	9
2.1.4	Evaluating Tests . . . . .	11
2.1.5	Ex-Post Power Analysis . . . . .	15
2.1.6	Confidence Intervals . . . . .	15
2.1.7	Noninferiority Tests and Inverting a Test . . . . .	17
2.1.8	Estimation and Hypothesis Testing . . . . .	18
2.1.9	Substantial Significance . . . . .	18
2.2	Binary-Valued Distributions . . . . .	19
2.2.1	One-Sided Hypotheses . . . . .	19
2.2.2	Tests of Equality . . . . .	23
2.3	Multi-Valued Distributions . . . . .	24
2.3.1	Identity of Two Distributions . . . . .	24
2.3.2	Independence of Two Distributions . . . . .	27
2.3.3	Medians . . . . .	27
2.3.4	Stochastic Inequalities and Categorical Data . . . . .	28
2.3.5	Means . . . . .	29
2.3.6	Variances . . . . .	33
<b>3</b>	<b>Matched Pairs</b>	<b>34</b>
3.1	Binary-Valued Distributions . . . . .	34
3.1.1	Probabilities . . . . .	34
3.1.2	Correlation . . . . .	35
3.2	Multi-Valued Distributions . . . . .	36
3.2.1	Identity of Two Marginal Distributions . . . . .	36
3.2.2	Stochastic Inequalities and Categorical Data . . . . .	36
3.2.3	Means . . . . .	37

	2
3.2.4 Covariance and Correlation . . . . .	38
3.2.5 A Measure of Association Related to Kendall's Tau . . . . .	39
<b>4 Single Sample</b>	<b>41</b>
4.1 Success Probabilities . . . . .	41
4.2 Median and Quantiles . . . . .	41
4.3 Mean . . . . .	41
4.4 Variance . . . . .	42
4.5 Ordinal Dispersion . . . . .	43
<b>5 Comment on Asymptotic Theory</b>	<b>43</b>
<b>6 Eyeball Statistics</b>	<b>44</b>
<b>7 Some Comments on Siegel and Castellan (1988)</b>	<b>48</b>
<b>8 Most Popular Mistakes</b>	<b>49</b>
<b>9 Summary Table of Proposed Exact Nonrandomized Tests</b>	<b>51</b>
<b>10 List of Other Exact Tests</b>	<b>52</b>

## **Abstract**

This is an overview of tests for hypothesis testing that only rely on objective characteristics of the data generating process and do not rely on additional assumptions when making inference. In particular the methods are distribution-free, many apply to nonparametric settings. Only exact results are presented, results that can be proven for the given finite sample. Asymptotic theory is not used.

The reader is reminded how to construct confidence intervals for medians. Existing methods for comparing binomial proportions are contrasted using formal criteria. Hidden assumptions underlying the Wilcoxon rank sum test and the Spearman rank correlation test are exposed.

New tests are presented for analyzing levels, dispersion and association and for comparing levels, tests that are based on ordinal comparisons. We also present confidence intervals for a single mean or variance, for testing equality of two means or of two variances and for testing for correlation. These methods apply when all possible outcomes belong to a known interval.

This overview is particularly guided towards experimental game theory where small samples are the rule and there is a tradition for minimizing additional assumptions imposed.

# 1 Introduction

Statistical hypothesis testing of experimental data is often if not typically involved with small samples. Many researchers are in favor of using distribution-free methods as in this case the data itself reveals the information. Inference is not distracted by further assumptions. Given that samples are small it is useful to have exact methods as being small means that it is hard to believe that the sample is sufficiently large for asymptotic results to kick in. In fact, we discuss why asymptotic results as typically used need not be relevant. Mathematically this is due to the difference between point-wise and uniform convergence. Intuitively this is because statements regarding "sufficiently large" are not very useful if they are conditioned on the underlying distribution when it is exactly this distribution that is not known.

We present an overview of exact methods, reminding readers of tests that are only little known but very useful and of assumptions that are often overlooked. We present the basic methodology and add a critical discussion of the basic concept of a p value. We present new ways of comparing traditional methods as hypothesis testing, in contrast to Bayesian reasoning, is about ex-ante inference. First a test is chosen, then the data is analyzed. Thus, when choosing a test one has to have some theory of why this test is preferred to alternative tests. Finally, and most innovative we provide a non technical overview of new statistical tests developed formally by the author in separate papers, tests which allow to compare levels and dispersion, such as means and variances, while ignoring other features of the data that may differ between the samples.

Two types of tests are considered once there are more than two possible outcomes. The one type refers to ordinal properties of the underlying data generating process. This includes well known concepts such as median and quantiles as well as less known measures such as the absolute deviation of two random observations, a stochastic inequality and Kendall's tau. The respective tests require no further information. The second type relates to cardinal properties of the underlying process. It is well known that cardinal properties such as mean or variance cannot be tested if outcomes cannot be bounded based on the existing information about the underlying process. This is due to the possibility of fat tails.

To establish tests we build on additional information that is typically available.

Data often belongs to some ex-ante known bounded interval. These bounds typically emerge in the design. Bounds can be due to measurement when data is on a percentage scale. Such bounds arise naturally when an experimenter is faced with the necessity to fix an upper bound on the maximal earnings of a subject. Knowing such bounds adds sufficient structure to be able to make inference without adding additional distributional assumptions. The material herein that refers to means or to correlations applies only to data that yields outcomes belonging to such an ex-ante known bounded set. If this condition is not met then the methods can still be applied, namely by conditioning the variables on having outcomes within some given set. Here it is particularly important that the set of outcomes that is used for this conditioning is not chosen based on the gathered data but based on some exogenous characteristics (or based on results of an alternative data source that is otherwise not used for the current analysis).

This is not a complete list of all existing exact methods, these are simply the ones that the author knows the proof of. Note that one has to be very careful in reading some literature on statistical hypothesis testing as there may not be sufficient precision when describing the hypotheses or when mentioning whether or not the test is exact. One book that is formally very precise but also very technical is Lehmann and Romano (2005). Motulsky (1995) provides a clear introduction with useful introduction to concepts and interpretation, unfortunately it ignores papers of Lehmann and Loh (1990), Boschloo (1972) and Suissa and Shuster (1984, 1985), thus ignoring the limits of the t test and not presenting two unconditional tests that outperform the Fisher exact test. Note that Siegel and Castellan (1988), although a classic, is imprecise and misleading at some important points, in the appendix we point out some imprecisions and mistakes in this book that we have identified.

This overview will always be at an early stage relative to the large amount of existing material on statistical hypothesis testing. The contents are under constant revision, we aim to include examples soon. We start by considering the analysis of two independent samples and present here the general framework of hypothesis testing with a thorough discussion. Later sections deal with matched pairs and with single samples. In each of the three sections we consider separately binary valued data, consider tests relating to entire distributions, tests of stochastic inequalities, mention

what we know about medians and then introduce tests relating to means and to variances. In a final section we comment on asymptotic theory. In the appendix we discuss some of the statements made in Siegel and Castellan (1988) and also provide a table that presents an overview of the tests. Comments on the content are most welcome.

## 2 Two Independent Samples

### 2.1 Preliminaries

There are two random variables  $Y_1$  and  $Y_2$ .  $Z \subseteq \mathbb{R}$  is the set of outcomes that are known to be possible for each of them, so  $Y_1, Y_2 \in Z$ . The set  $Z$  is known and should not be determined based on the data. Success or failure can be formatted as  $Z = \{0, 1\}$ , measurements as percentages lead to  $Z = [0, 100]$ . When  $Z$  has infinitely many elements such as when  $Z$  is an interval then the set of possible joint distributions  $\Delta Z \times \Delta Z$  is infinitely dimensional. In this case one speaks of a *nonparametric setting*.

Data is given by a sample of independent observations consisting of  $n_1$  drawn from  $Y_1$ , denoted by  $y_{1,1}, \dots, y_{1,n_1}$ , and  $n_2$  drawn from  $Y_2$ , denoted by  $y_{2,1}, \dots, y_{2,n_2}$ . Let  $y = (y_{1,1}, \dots, y_{1,n_1}, y_{2,1}, \dots, y_{2,n_2})$  be the total sample. Let  $P_{Y_i}$  be the distribution of outcomes generated by  $Y_i$ . Let  $EY_i = \int y dP_{Y_i}(y)$  be the expected value of  $Y_i$ . Let  $\Delta Z$  be the set of all distributions that have support contained in  $Z$ .

We wish to make inference about the true distributions  $P_{Y_1}$  and  $P_{Y_2}$ , specifically about whether  $(P_{Y_1}, P_{Y_2})$  belongs to a set  $W \subset \Delta Z \times \Delta Z$  called the null hypothesis. So we wish to test the null hypothesis  $H_0 : (P_{Y_1}, P_{Y_2}) \in W$  against the alternative hypothesis  $H_1 : (P_{Y_1}, P_{Y_2}) \notin W$  where  $W \subset \Delta Z \times \Delta Z$  is given. In particular, we consider the entire product set  $\Delta Z \times \Delta Z$  as the set of possible pairs of distributions and do not limit attention for instance to symmetric distributions or distributions that differ only according to their mean. In this sense we choose a *distribution-free* approach. The reason is that further limitations such as limiting attention to symmetric distributions is rarely an objective property of the underlying problem.<sup>1</sup>

---

<sup>1</sup>A verifiable restriction that sometimes arises is the case of continuous distributions where one knows that two identical observations (almost surely) cannot occur. However adding this restriction will not help to improve inference as the closure of this set leads back to our setting.

We wish to focus on inference that does not rely on additional assumptions (such as when using Bayesian statistics) or when assuming normal errors.

We present some examples of possible null hypotheses of interest:

1. Test of identity of two distributions by setting  $W = \{(P_{Y_1}, P_{Y_2}) : P_{Y_1} \equiv P_{Y_2}\}$ ,
2. test of identity of two normal distributions that have equal variance by setting  $W = \{(P_{Y_1}, P_{Y_2}) : P_{Y_1} \equiv P_{Y_2}, P_{Y_1} \text{ is normally distributed, } Var(Y_1) = Var(Y_2)\}$ ,
3. test of first order stochastic dominance by setting  $W = \{(P_{Y_1}, P_{Y_2}) : P_{Y_1}(y \leq z) \leq P_{Y_2}(y \leq z) \text{ for all } z\}$ ,
4. test of stochastic inequality:  $W = \{(P_{Y_1}, P_{Y_2}) : P_Y(Y_1 \geq Y_2) \leq 1/2\}$
5. *two-sided* test of equality of two means:  $W = \{(P_{Y_1}, P_{Y_2}) : EY_1 = EY_2\}$ ,
6. *one-sided* test of inequality of means:  $W = \{(P_{Y_1}, P_{Y_2}) : EY_1 \geq EY_2\}$ .
7. test of *noninferiority* of  $Y_2$  over  $Y_1$ :  $W = \{(P_{Y_1}, P_{Y_2}) : EY_1 \geq EY_2 + d_0\}$  where  $d_0 > 0$  is given, if instead  $d_0 < 0$  then this is called a test of *superiority* of  $Y_2$  over  $Y_1$ .

The null hypothesis described in item 1 that the two distributions are identical can be referred to as a test of *no treatment effect*. This makes sense when the variables  $Y_1$  and  $Y_2$  differ according to some exogenous variation and the question is then whether or not this exogenous variation influenced the outcome, thus making the distributions different.

Instead of inserting means in items 5 and 6 above one could similarly consider medians or variance.

Terminology is not consistent, while in items 1 – 6 we name the test after its null hypothesis in item 7 we name it after the alternative hypothesis as a rejection under  $d_0 > 0$  shows significant evidence of  $EY_2 > EY_1 - d_0$  and hence that  $Y_2$  is “not inferior to”  $Y_1$ .

It is important that the alternative hypothesis contains the complement of the null hypothesis. Otherwise additional assumptions are added which we do not want to. For instance, to test  $H_0 : EY_1 = EY_2$  against  $H_1 : EY_1 > EY_2$  implicitly assumes

that  $EY_1 < EY_2$  cannot occur. Considering more limited alternative hypotheses is however acceptable when making inference about what happened when the null hypothesis could not be rejected (see below).

Hypothesis testing is about making statements based on noisy data. Due to noise true statements cannot be guaranteed and hence the objective is to make statements that are true up to some prespecified maximal error, the convention is that statements should be only wrong at most 5% of the time. This statement is made when the null hypothesis is rejected. So the null hypothesis has to be constructed as complement of the desired statement or research objective. So if the aim is to provide evidence that some treatment makes a difference then the null hypothesis is that there is no treatment effect. However if the objective is to show that there is no treatment effect then one would have to test the null hypothesis that  $P_{Y_1} \neq P_{Y_2}$ . However this is not possible (with a nontrivial test) as intuitively the alternative hypothesis is too small. The formal problem comes from the fact that any element of the alternative hypothesis can be approximated by a sequence of elements belonging to the null hypothesis. Similarly one cannot set  $W = \{(P_{Y_1}, P_{Y_2}) : EY_1 \neq EY_2\}$ . A way out of this dilemma is for instance to test whether the two means differ substantially, setting  $W = \{(P_{Y_1}, P_{Y_2}) : |EY_1 - EY_2| \geq d\}$  for some  $d > 0$ . A more flexible and thus advisable approach particularly when one is interested in making inference both when the null hypothesis is rejected and not rejected is to consider confidence intervals for the difference between the two means, discussed in more detail below.

A *hypothesis test* is a recommendation whether or not to reject the null hypothesis given the data. This recommendation is conditional on the data. Formally  $\phi$  is a test if  $\phi : Z^{n_1} \times Z^{n_2} \rightarrow [0, 1]$  where  $\phi(y)$  is the probability of rejecting the null hypothesis based on the data  $y$ . In particular, this recommendation can be probabilistic, meaning that the recommendation based on the data gathered could be to reject the null hypothesis with probability 0.7. The test is called *randomized* if this can happen, so if there is some data such that given this data the test recommends to reject the null hypothesis with a probability belonging to  $(0, 1)$ . Formally, if there is some  $y \in Z^{n_1+n_2}$  such that  $\phi(y) \in (0, 1)$ . It seems strange to reject with some probability contained in  $(0, 1)$ . Such behavior may be necessary to realize some specific properties. Properties of a test (e.g. size, level, unbiased as defined below) are typically defined in terms

of expected probability of rejection before gathering the data where this expectation is conditional on the data been drawn from some given  $(P_{Y_1}, P_{Y_2})$ . Tests that have some optimality property in terms of inference are often randomized. A test that is not randomized is called *nonrandomized* (or deterministic). The *critical region* of a nonrandomized test is the set of all observations under which the test recommends to reject the null hypothesis. Only nonrandomized tests have been accepted in practice. The objective is to find a parsimonious recommendation based on the data, “reject” or “not reject”.

The *size* (or *type I error*) of a test is the maximal probability of wrongly rejecting the null hypothesis, calculated before gathering the data. So this is the maximal probability of rejecting the null hypothesis when it is true. The maximum is taken among all  $(P_{Y_1}, P_{Y_2})$  belonging to  $W$  where  $W$  is the set of joint distributions that belong to the null. A test has *level*  $\alpha$  if its size is at most  $\alpha$ . A test is called *exact* if it has the level that it is claimed to have, or more specifically if one can prove its properties in terms of level for given sample sizes. One says that there is *significant evidence* against the null hypothesis in favor of the alternative hypothesis if the null hypothesis is rejected when  $\alpha = 0.05$ , instead there is marginal or strong significance if it is rejected when either  $\alpha = 0.1$  or  $\alpha = 0.01$ . The *p value* associated to a given test (that is indexed by its level or size  $\alpha$ ) and a given data set is the maximal value of  $\alpha$  such that the null hypothesis would be rejected for the given data. Thus, there is significant evidence against the null hypothesis (at the conventional 5% level) if and only if the p value is at most 5%. The p value is a measure for whether the observed data is typical for some data generating process belonging to the null hypothesis. However the p value is the probability of observing the data. Except for exceptional cases, it is not the probability of observing something as "extreme" or "extremer" that what is being observed when the null hypothesis is true. One reason is that typically the null hypothesis contains several distributions so we do not know which one should be used to derive the probability. A second reason is that many tests cannot be interpreted as "reject if and only if the data is sufficiently extreme". The *type II error* of a test for a given a subset of the alternative hypothesis is the maximal probability of wrongly not rejecting the null hypothesis when the true data generating process belongs to this subset.

One has to be careful when selecting tests of equality and of inequality. It for instance can happen that one does not find significant evidence that two means are different yet significant evidence that the first is strictly larger than the second. This would happen if the null  $H_0 : EY_1 \leq EY_2$  cannot be rejected at level 2.5% but can be rejected at level 5%. One way out would be to consider different conventions for the significant levels depending on whether the tests are one-sided or two-sided. However this is typically not done. The alternative is to only consider one-sided tests. But then one has to be able to argue that one did not separately test  $H_0 : EY_1 \leq EY_2$  and  $H_0 : EY_1 \geq EY_2$  each at level 5%. In other words,  $H_0 : EY_1 \leq EY_2$  can only be tested if there is a good story behind this asymmetric relationship. Note that while hypothesis tests are typically one-sided, possibly in order avoid the above phenomenon, confidence sets are typically two-sided as they are confidence intervals. So it does not seem compatible to present both for the same data as otherwise it can happen that one has significant evidence for  $EY_2 > EY_1$  based on a one-sided test yet 0 belongs to the confidence interval for  $EY_2 - EY_1$ .

### 2.1.1 Documentation

It is important when reporting results to identify completely the way the data was gathered, the hypotheses, the assumptions regarding the variables and to give sufficient information about the tests used. Note that many software packages do not use the exact formulae even when they call them exact (e.g. see STATA). They typically use values that are derived as if the sample was infinitely large, hence document a p value that would be correct if the sample size would be infinite. However they do not provide information or bounds on how much the actual size or level can differ given that the sample is only finite. Frankly, whenever such approximations are used the methods are no longer exact. Given the advancement in computer speed one can often calculate exact p values, provided that the space of data generating processes is not too large. Good practice is to tell the reader whether or not one uses a truly exact test.

### 2.1.2 Accepting the Null Hypothesis

Null hypotheses are rejected or not rejected. One may or may not find significant evidence of a treatment effect. One does not say that the null hypothesis is accepted. Similarly, one does not say that one has found significant evidence of no treatment effect. Why? Not being able to reject the null hypothesis can have many reasons. It could be that the null hypothesis is true. It could be that the alternative hypothesis is true but that the test was not able to discover this and instead recommended not to reject the null. The inability of the test to discover the truth can be due to the fact that it was not sufficiently powerful, that other tests are more powerful. It could be that the sample size is not big enough so that no test can be sufficiently powerful. Of course, whether or not the sample is sufficiently large and the test is sufficiently powerful to generate a low type II error depends on where the underlying distribution lies in the alternative hypothesis. If it lies close to the null hypothesis then it will be very difficult to detect it. However if it lies further away for instance in terms of having a large difference in means then it will be easier to identify the alternative as being true. Thus, when not being able to reject the null hypothesis then one can make inference about whether or not the null hypothesis is true. But this inference has to be conditional on a statement about the true underlying distributions. For instance, such a statement could be that conditional on the difference in the means being larger than 0.3 the probability of wrongly not rejecting the null hypothesis is 15%. In other words, one is limiting the alternative hypothesis and then making a statement about the maximal probability of wrongly not rejecting the null hypothesis. So when the power of a test is known then one can quantify the evidence for not being able to reject the null hypothesis conditional on a statement about the alternatives. However, as this statement will depend on how the set of alternatives is limited (in the above example we assumed the difference to be larger than 0.3) one does not speak of “accepting” the null hypothesis. Here we again refer to the more flexible approach underlying confidence intervals where one does not have to a priori specify a null hypothesis but instead only a parameter of interest.

### 2.1.3 Side Remarks on the P Value

You may wish to skip over this section as its contents will not influence your testing and it is a bit difficult to grasp for some. One namely has to be very careful when using the p value.

The p value has only very little value for inference apart from stating whether it lies in  $[0, 1\%]$ ,  $(1\%, 5\%]$ ,  $(5\%, 10\%]$  or strictly above 10%. Here is the argument. To find the p value one has to find the smallest value of  $\alpha$  such that the test with size  $\alpha$  rejects the null hypothesis. So given the data one is selecting a test - after all, the size of a test is part of the description of a test. One is selecting a test based on the data. However it is extremely important for statistical hypothesis testing is that one first selects a test and then gathers the data. We illustrate. Assume that data set  $y^{(1)}$  has been gathered and the statistical software indicates that the p value is equal to 3%. This means that if you had decided before gathering the data to use a test with level 3% then the null hypothesis would be rejected while if you had decided at the outset to set the level to 2% then you would have not rejected the null hypothesis. Fine. However, did you do this? No. We are determining characteristics of the test (the size is an important element) based on the data gathered. In other words, we are looking at the data and making inference about what would have happened if we had made some decision before observing this data. This is not allowed, because the methodology of statistical hypothesis testing is to first decide on a test and then to gather data.

What can go wrong? Let us return to our illustration and assume instead that a different data set  $y^{(2)}$  had been observed that is associated to the p value of 4.7%. Under either data set  $y^{(1)}$  or  $y^{(2)}$ , conventional thresholds indicate significant evidence to reject the null hypothesis as the p value is below 5%. So after gathering either data set, the statistician will claim that there is significant evidence against the null hypothesis. This procedure, to reject if and only if the p value is below 5% generates a test that has size 5%. Whether or not the p value was 3% or 4.7% does not influence the fact that the null is rejected and hence has no meaning. It is part of the nature of testing that the null is sometimes rejected in a border line case and sometimes rejected when the data seems to be very strong in favor of the alternative. "It is irrelevant whether a serve in tennis is in the middle of the service area or on the line. The rules

are clearly set before the game starts and sometimes the service will be good even though it is very close to the line." Data sets are random. Tests allow to make claims even under this randomness by keeping the ex-ante probability of wrongly rejecting the null hypothesis small, i.e. below  $\alpha = 5\%$ .

In fact, it is easy to see that the probability of wrongly rejecting the null hypothesis, conditional on observing a p value close to 5% is dramatically larger (for more information consult Berger, 2003). This simply reinforces our argument above that a test with size 5% has to sometimes produce p values that are very much lower than 5% in order not to reject the null hypothesis more than 5% of the time.

The p value however does have value for inference when the size of the test is always strictly below the postulated level. This typically happens with nonrandomized tests where it is hard to find a test that has size equal to 5%. For instance consider the Z test for comparing the means of two independent samples of 10 observations each. The maximal cutoff that ensures level 5% yields a test that has size equal to 4.7%. So one could choose to report 4.7% instead of 5% but this could be confusing given the above more general statement about p values.

The p value cannot be used to compare data sets. Or more precisely, if this is done then one has to cite a formal method for doing this. As I have not seen such a formal claim it is not allowed. There are tests for comparing data sets.

The p value cannot be used to make inference about the error of not rejecting the null hypothesis. In other words, if the p value is 11% or 34%, in either case the null could not be rejected but one cannot make any statement about whether this was consistent with the truth. If one wishes to understand the probability of correctly not rejecting the null hypothesis one has to investigate the type II error (see below) of the test. Confidence intervals (formally presented below) present an alternative for understanding the value of inference.

The contribution of the p value is that it gives some indication for future tests using similar data in terms of the possibility of gathering less data if the p value is small or the necessity for gathering more data if the p value was close to or above 5%.

The bottom line is as follows. Tests have to be selected ex-ante before gathering the data. Similarly, one has to decide ex-ante on rules for interpreting the data

(significant or marginally significant). Hence there are conventions for how to interpret data, these can be interpreted as ranges for the p value,  $p \in [0, 1\%]$ ,  $(1\%, 5\%]$ ,  $(5\%, 10\%]$ ,  $(10\%, 100\%]$ . Respectively, one speaks of strongly significant evidence, significant evidence and marginally significant evidence or of no significant evidence in favor of rejecting the null hypothesis based on the data.

Some also use the p value as an ex-post measure of the probability of observing data as extreme or extremer when the null hypothesis is true. This Fisherian approach is only valid if the null hypothesis contains single data generating process. Extremeness is measured here in terms of a test statistic. Once there are several data generating processes in the null hypothesis and there is still such a test statistic then the p value is equal to the largest possible probability of the event of as extreme or extremer when the null hypothesis is true. However this is only an interpretation. The value of inference in terms of the given data set and the given null hypothesis when the p value is 3% or 4.7% is the same.

#### 2.1.4 Evaluating Tests

The *power* (or more precisely, the power function) of a test is the probability of rejecting the null hypothesis for given  $(P_{Y_1}, P_{Y_2})$  belonging to the alternative hypothesis for a given sample size. The *type II error* of the test is measured conditional on some subset  $W_1$  of the alternative hypothesis, so  $W_1 \cap W = \emptyset$ . It is equal to the maximal probability of wrongly not rejecting the null hypothesis, which is the maximum probability of not rejecting the null hypothesis, among all  $(P_{Y_1}, P_{Y_2})$  belonging to  $W_1$ .

One test is *uniformly more powerful* than an alternative test if the first test yields a higher probability of rejecting the null hypothesis than the second test for all  $(P_{Y_1}, P_{Y_2})$  belonging to the alternative hypothesis. A test is *uniformly most powerful* for a given level  $\alpha$  if it is uniformly more powerful than any alternative test that has level  $\alpha$ . So if one test is uniformly more powerful than an alternative test then we would prefer the former test. Only rarely are we able to compare tests in terms of being uniformly more powerful.

Sometimes one limits attention to a subset of all tests to then find a uniformly most powerful test within this subset. A common approach is to consider only unbiased tests. A test is *unbiased* if the probability of rejecting the null is larger when the null

is false than when it is true. This sounds like a nice property. However it comes at a large expense. Unbiasedness is namely then used to rule out other tests. Tests will fail to be unbiased simply because of properties very close to the null hypothesis that involve comparing rejection probabilities close to  $\alpha$  (so very small) at the expense of ignoring differences in power away from the null (see Suissa and Shuster, 1984, for their discussion in the context of comparing two binomial probabilities).

Often one is interested in particular subsets  $W_1$  of the alternative hypothesis. In the present setting of comparing means it is natural to focus on inference in terms of the differences of the two means. This means that one considers type II error for  $W_1 = W_1^d := \{(P_{Y_1}, P_{Y_2}) : EY_1 + d \leq EY_2\}$  across all  $d > 0$ . Different tests can then be compared and possibly ranked according to their type II error across  $\{d > 0\}$ .

Understanding the power of a test is useful when the null hypothesis cannot be rejected based on the data. One can then look across  $d$  and assess the maximal probability of not being able to reject the null hypothesis when  $EY_2 \geq EY_1 + d$ . In the literature one often considers a type II error of 20%, hence one could search for  $d_{0.2}$  such that the type II error conditional on  $W_1^d$  is at most 20% when  $d \geq d_{0.2}$ . When not able to reject the null then one can comment on  $d_{0.2}$ , mentioning that this event of not being able to reject the null will occur less than 20% of the time when the true difference is at least  $d_{0.2}$ . Type II error is also useful for the design of a randomized experiment. One can then specify a minimal treatment effect  $\bar{d}$  that one wishes to uncover and then choose the sample size such that the type II error conditional on  $W_1^{\bar{d}}$  is at most some value, for instance 20%. Later we illustrate further.

We now explain how to compare two test with the same level that cannot be ranked in terms of being uniformly more powerful.

One option is to compare them based on the following  $\varepsilon$  dominance criterion, which is based on the concept of being  $\varepsilon$  uniformly more powerful. Specifically, a test  $\phi$  is  $\varepsilon$  *uniformly more powerful* than a test  $\phi'$  if the type II error of  $\phi$  never lies  $\varepsilon$  below that of  $\phi'$  and there is no smaller value of  $\varepsilon$  such that this statement is true, ( $\varepsilon \geq 0$ ). So if  $\varepsilon = 0$  then this is the standard concept of being uniformly more powerful. Denote this order by  $\phi \succ_{\varepsilon} \phi'$ . The criterion of being  $\varepsilon$  uniformly more powerful is useful when comparing two tests. One may choose to prefer  $\phi$  to  $\phi'$  if  $\phi \succ_{\varepsilon} \phi'$ ,  $\phi' \succ_{\varepsilon'} \phi$  and  $\varepsilon \leq \varepsilon'$ . Here one selects among two tests the one that can be

outperformed less in terms of power by the other. This criterion is particularly useful when  $\varepsilon$  is substantially smaller than  $\varepsilon'$  as in this case, in terms of type II error,  $\phi$  can only be slightly worse in some situations while  $\phi'$  can be substantially worse than  $\phi$  in other situations. The disadvantage of this approach is that it only generates a partial order as transitivity can be violated.

An alternative option is to apply both tests to the data and to hope that both reject the data. A further alternative is to take a test that everyone else uses (provided this test is appropriate, a fact that does not follow simply because many use it). It is namely important that one can credibly claim that the test used was not chosen simply because it rejects the null hypothesis. In particular, one is not allowed to select the test that yields the lowest p value.

A final option is to choose some criterion and select a test according to this criterion. Of course if the methodology itself has to be commonly accepted as otherwise different methods possibly lead to different tests. Below we present two such criteria.

We show how one can select tests according to minimax regret. The idea is that power is evaluated in terms of the distance of true data generating process to the null hypothesis. The distance is measured in terms of the parameter of interest which is here assumed to be the difference between the two means. Power evaluated together with this distance is called regret. So for instance when testing  $H_0 : EY_1 \geq EY_2$  against  $H_1 : EY_1 < EY_2$  then one evaluates the loss of using a test according to

$$\max_{(P_{Y_1}, P_{Y_2}) : EY_1 < EY_2} \beta(Y) (EY_2 - EY_1)$$

where  $\beta(Y)$  is the type II error of the test conditional on data being drawn from  $(P_{Y_1}, P_{Y_2})$ . One is not worried with wrongly not rejecting the null when  $EY_2$  is very close to  $EY_1$  while the associated loss is large if one is not able to learn the truth even when  $EY_2$  is much larger than  $EY_1$ . The specific functional form used comes from its connection to the literature on minimax regret (Savage, 1951). One imagines that  $Y_1$  measures the outcome of an existing treatment while  $Y_2$  is the outcome of some new treatment where outcomes are measured such that higher expected outcomes are preferred. If the null is not rejected then one continues using the existing treatment which yields an expected outcome of  $EY_1$ . On the other hand, upon rejection the new

treatment will be implemented which yields expected outcome  $EY_2$ . Observe that

$$\beta(Y)(EY_2 - EY_1) = EY_2 - [(1 - \beta(Y))EY_2 + \beta(Y)EY_1].$$

So when the alternative is true then the right hand side is the difference between the expected outcome from achieving the best outcome  $EY_2$  and the expected outcome of following the recommendation of the test. This difference is equal to the regret that one would have from following the recommendation of the test if one should ever learn that the alternative hypothesis was true.

An alternative method of comparing tests is to consider the associated confidence intervals and to select the one with lower maximal expected width (see below).

One option that is not valid is to compare tests according to their p value. The p value is only associated to the level of the test as explained above. Tests have to be compared based on power. Of course there is some intuition that a higher p value, as long as this is below  $\alpha$ , also means that the test is more powerful as it seems like a higher p value means that the null hypothesis is more likely to be rejected. However this is just intuition and has no formal basis. There is no way around making power calculations when evaluating and comparing tests that have the same level.

The bottom line is that one has to first determine a criterion for selecting tests, then select the test, then gather the data. One cannot simply evaluate several tests and choose the test that rejects the null hypothesis. The reason is that the level of the test is an upper bound on the maximal probability wrongly rejecting the null hypothesis. However, it is not an upper bound on the maximal probability of rejecting the null hypothesis under one of many tests. The problems arise when the critical regions are not nested. Of course when they are nested then one test is uniformly better than the other and there is no reason to even consider the less powerful test. Needless to say, it is very tempting to first see what test does well given the data and then to find reasons to like the test. That is why there are conventions on which test to use, when these conventions involve recommending correct tests then this is an alternative way of avoiding that tests are selected after the data has been analyzed. Of course, when in doubt about the acceptance, one can evaluate more than one test and only then reject the null if all tests reject it individually. Clearly this approach will typically and unnecessarily cost power. When in doubt how to select a test, the criterion of choosing the confidence interval with the smallest maximal expected

width seems the candidate that will find the most support. The disadvantage is that often we do not have such confidence intervals, and if so then it is time consuming to calculate their expected width.

### 2.1.5 Ex-Post Power Analysis

As explained above, power should be used to compare tests and to determine sample sizes when designing an investigation. However, once the data has been gathered and the test has been evaluated, power in itself is no longer valuable. This holds for the setting of classical hypothesis testing where unlike bayesian statistics one does not make an assessment of which distribution belonging in the alternative hypothesis is more likely. A test with a low power can discover with large probability that the alternative hypothesis is true when the true distribution is very extreme in the sense of being far from distributions that belong to the null hypothesis. At the same time, a test with high power can fail with large probability to discover that the alternative is true when the true distribution lies very close to those belonging to the null hypothesis. Of course indirectly we have some information, rejection with large power seems to mean large effect, no rejection with small power indicates lack of enough data. But instead of arguing informally one should simply move to confidence intervals (see below) as these will formally give an indication about the value of the inference in both of these cases.

### 2.1.6 Confidence Intervals

Modern statistics has been claimed to focus more on confidence intervals than on hypothesis testing. One reason is that the problem of how to derive implications when the null is not rejected does not arise. Moreover they provide information about how strong the evidence is when the null is rejected.

A *family of confidence intervals*  $I$  is a correspondence that maps each data set into an interval. Formally,  $I : Z^{n_1+n_2} \rightarrow \Delta\mathcal{I}$  where  $\mathcal{I}$  is the set of all closed intervals. The confidence interval should be designed to contain the true parameter of interest with a minimal probability that we denote by  $1 - \alpha$  in which case we say that  $I$  has coverage  $1 - \alpha$ . Formally,  $I$  has *coverage*  $1 - \alpha$  if with probability greater or equal to  $1 - \alpha$  the interval  $I$  contains the parameter of interest. In our working example the parameter

of interest could be  $EY_1 - EY_2$ , the difference between the two expected values, hence,  $\Pr_P(EY_1 - EY_2 \in I(y)) \geq 1 - \alpha$  has to hold for all  $P$ . The most common measure for evaluating and comparing confidence intervals with the same coverage is the expected width.

There is an important connection between confidence intervals and hypothesis testing that we now explain. Assume that one knows of a nonrandomized test  $\phi_d$  for  $W_d = \{EY_1 - EY_2 = d\}$  that has level  $\alpha$  for each  $d \geq 0$ . Then one can obtain nonrandomized confidence region with coverage  $1 - \alpha$  by collecting all values of  $d$  such that the null hypothesis cannot be rejected under  $\phi_d$ . Typically this region will be an interval. One obtains an interval if the set of parameters that do not yield a rejection is convex, so if  $\{d : \phi_d(y) = 0\}$  is convex for all  $y$ . Uniformly more powerful nested tests with this property lead to confidence intervals that have lower expected width. When using equi-tailed tests of equality one obtains equi-tailed confidence intervals.

The construction of confidence intervals from tests sounds like data mining as one essentially applies many different tests on the same data to obtain the interval. However this method is a valid method, it is the common method for constructing confidence intervals and the proof of coverage is very simple (e.g. see also Lehmann and Romano, 2005). We briefly recall this proof. Let  $A(d)$  be the set of all data  $y$  where  $H_0 : (P_{Y_1}, P_{Y_2}) \in W_d$  is not rejected. Let  $I(y) = \{d : y \in A(d)\}$  be the set of all  $d$  where the null is not rejected given data  $y$ . Then  $d \in I(y)$  if and only if  $y \in A(d)$ . Consequently,  $\Pr_P(EY_1 - EY_2 \in I(y)) = \Pr_P(y \in A(EY_1 - EY_2)) \geq 1 - \alpha$ . Hence  $I$  is a family of confidence intervals with coverage  $1 - \alpha$ .

We show how to interpret equi-tailed confidence intervals that were constructed from tests of noninferiority. To fix ideas, assume that  $[6, 27]$  is a 95% equi-tailed confidence interval the difference between two means. Then we can claim that the true distance lies in this interval with probability 0.95. This also means that one can reject that the difference is below 6 at level  $5\%/2 = 2.5\%$ . In particular there is significant evidence that the two means are not equal. Similarly we also reject that the difference is above 27 at level 5%. Now assume that instead one obtains  $[-12, 28]$  as 95% equi-tailed confidence interval. Here there is no evidence against equality of means. In contrast there is evidence that they are not too unequal. We find that we can reject that their difference is above 28 at level 2.5%. This example shows two

things. First of all that the convention of speaking of significance whenever one-sided tests are rejected at level 5% means that one should consider equi-tailed confidence intervals with coverage 90%. Second of all, when not being able to reject the null hypothesis of equality of means then it could be useful to construct a confidence interval that is symmetric around 0. Hence, instead of restricting attention to equi-tailed confidence intervals one searches for  $d$  such that  $[-d, d]$  has coverage 90%. Or in other words, given the data one searches for  $d$  such that the probability of wrongly rejecting the null when the distance is larger than  $d$  is equal to 10%.

Finally we point out a possible limitation of confidence intervals. It is not so much the approach itself but their construction. Remember that one needs a family of tests and there are many such families. Often we only have very limited understanding of which tests for a given pair of hypotheses performs best. When selecting a family this corresponds to selecting a test for each pair of hypotheses indexed by the parameter of interest. So the problem of finding a best family of confidence intervals can be too complex. There is a partial solution to this problem which is called "inverting a test". Assume that one is testing  $H_0 : EY_1 = EY_2$  and one is not able to reject the null hypothesis. Search for  $d$  such that the probability of not rejecting the null hypothesis conditional on  $EY_1 \geq EY_2 + d_1$  is equal to 2.5%. Similarly search for  $d_2$  such that the probability of not rejecting the null hypothesis conditional on  $EY_1 \leq EY_2 - d_2$  is equal to 2.5%. Then  $[-d_2, d_1]$  is a 95% confidence interval for  $EY_1 - EY_2$ .

### 2.1.7 Noninferiority Tests and Inverting a Test

The method of inverting a test used above as a simple way to derive a confidence interval is also very useful for constructing noninferiority tests. Assume one wishes to test  $H_0 : EY_1 \geq EY_2 + d_0$  against  $H_1 : EY_1 < EY_2 + d_0$  for some given  $d_0 > 0$ . Assume that one has an exact test  $\phi$  for  $H_0 : EY_1 \leq EY_2$ . We show how to derive a test for noninferiority by interchanging the role of null and alternative hypothesis. Given  $\alpha$  find  $\alpha'$  such that the probability of rejecting  $H_0 : EY_1 \leq EY_2$  under level  $\alpha'$  when  $EY_1 \geq EY_2 + d_0$  is at most  $\alpha$ . So the probability of not rejecting  $EY_1 \leq EY_2$  under level  $\alpha'$  when  $EY_1 \geq EY_2 + d_0$  is at most  $\alpha$ . We define our test of noninferiority as follows. Reject  $H_0 : EY_1 \geq EY_2 + d_0$  whenever the test  $\phi$  with level  $\alpha'$  does **not** reject  $H_0 : EY_1 \leq EY_2$ . It follows that this test is exact. Note that the type II error

of this test for  $EY_1 \geq EY_2 + d_0$  conditional on  $EY_1 \leq EY_2$  is  $\alpha'$ .

### 2.1.8 Estimation and Hypothesis Testing

An estimate is a value of some unknown parameter that one thinks is closest to the true parameter. Sometimes the estimates coincide with a measure used to describe the data sampled. This is for instance the case for the expected value or mean. The average observation given an independent sample of a random variable is the best estimate of its expected value. Similarly this the case for the difference between two means based on independent samples. The difference between the averages is the best estimate for the difference between the two means,  $EY_1 - EY_2 = E(Y_1 - Y_2)$ . For instance, when describing the sample one can say that one finds a higher average outcome among those patients treated ( $Y_1$ ) as compared to those not treated ( $Y_2$ ). In terms of estimating the effect of the treatment, one can use the same number to describe the estimated increase in the outcome, when moving from non treated to treated. Data description is useful. Estimation is valuable for talking about what happens beyond the specific data sampled. However it is hard to value estimation without significance. One may estimate that the treated have higher outcomes than the non treated regardless of how large the sample is. In particular, the sample could be such that this estimated increase is not significant and hence possibly due to an artifact of the variance caused by choosing a sample that is too small. Estimation is only valuable if one also performs hypothesis testing. This needs to be taken very seriously as in practice one often uses observed patterns to make inference about patterns that will appear in general. "We show that the outcomes of the treated are higher" is not possible. "We have significant evidence that the outcomes of the treated are higher" needs to be proven with hypothesis testing where  $H_1 : EY_1 > EY_2$ . "We observe that the outcomes of the treated are higher" is misleading as it is not clear whether "are higher" refers to those sampled or to outcomes in general. Better is to say "We observe in our sample that the outcomes of the treated are higher".

### 2.1.9 Substantial Significance

Last but not least one has to be sure to evaluate substantial or economic significance once one has found statistical significance. This means one at effect itself and compare

this to the objective and discuss whether the statistical effect highlighted as any practical value. For instance, if one says that income depends significantly on age but fails to report that the lower confidence bound on the increase in income is 10 cents per year of age. Note that estimated effect does not really help as the estimate may be very unreliable. Confidence intervals and confidence bounds give statistically reliable inference. For more on this topic see Ziliak and McCloskey (2008).

## 2.2 Binary-Valued Distributions

We now apply the above to the simplest setting where there are only two possible outcomes 0 and 1 of each random variable and hence  $Z = \{0, 1\}$ . The distribution is uniquely determined by  $\Pr(Y_i = 1)$  where  $\Pr(Y_i = 1) = EY_i$ . Here comparing means is equivalent to comparing the entire distributions, Bernoulli distributions. In fact, it is as if these two Bernoulli distributions are independent as we are comparing two independent samples from the margins  $P_{Y_1}$  and  $P_{Y_2}$ .

### 2.2.1 One-Sided Hypotheses

We wish to test the one sided hypothesis  $H_0 : \Pr(Y_1 = 1) \geq \Pr(Y_2 = 1)$ . There is no single test that can be recommended. Before analyzing the data one has to first select a test by performing some simple calculations for the given sample sizes. The classic test called the *Fisher exact test* should not be used in this context as the test of Boschloo is uniformly more powerful than the Fisher exact test, in many cases the Z test is also uniformly more powerful than the Fisher exact test. Of course, if one rejects the null hypothesis at  $\alpha = 5\%$  then this will also be true for the test of Boschloo. However it could easily be that Boschloo allows to reject at  $\alpha = 1\%$ . The only reason for using Fisher exact test in this environment is that one is lazy.

The Fisher exact test should only be used in the context of *conditional hypothesis testing* where one is making inference with respect to the sample and not with respect to the underlying distributions from which the sample has been drawn (conditional testing is however outside the scope of the present format of this document). For instance, one may be concerned with whether all those patients that received the placebo but got sick would have gotten sick even with the medication that is being tested. Note that the focus is on the specific sample of patients sampled even if these

should not be representative of the underlying population.

The framework of this paper is called *unconditional hypothesis testing*, "unconditional" refers to the fact that we are making inference about the distribution from where the data has been drawn and use the data to make this inference but do not care otherwise about the characteristics of the specific data. In terms of the above example, one is interested in whether the probability that a randomly chosen patient gets sick depends on whether or not this patient received the new medication. It is as if concern is for a representative patient.

**Remark 1** *There is no uniformly most powerful one-sided test for comparing the means of two Bernoulli distributions. However there is a uniformly most powerful test if one limits attention to the tests that are unbiased, in short, a UMPU test. Tocher found such a test. Recall that a test is called unbiased if the probability of correctly rejecting the null hypothesis is always greater than the probability of wrongly rejecting the null hypothesis. The consequent limitation to a subset of tests, the unbiased tests, makes the selection of a "best" test possible. However there is no reason to prefer unbiased tests to tests that are not unbiased. Unbiasedness puts restraints on tests that involve comparing very small probabilities. Power in terms of ability to reject the null hypothesis in other regions of the parameter space is thereby sacrificed (for more on this see Suissa and Shuster, 1984). Moreover, the UMPU test of Tocher is randomized and hence not useful in practice. The nonrandomized test that results from the UMPU test by recommending not to reject the null hypothesis unless the original test recommends a rejection with probability one is called the Fisher exact test.*

*Note that the UMPU test of Tocher is a permutation test (see Subsection 2.3.1 below) based on the test statistic of summing up the successes achieved by  $Y_2$ . It is not UMP as it is better to simply look at the difference between the successes of the two random variables when limiting attention to alternatives where  $\Pr(Y_1 = 1) < \Pr(Y_2 = 1) = 1 - \Pr(Y_1 = 1)$  (see Lehmann and Romano, 2005, Problem 3.59).*

Boschloo (1970) found a nonrandomized test we call the B test that is uniformly more powerful than the Fisher exact test (but uniformly less powerful than the randomized UMPU test of Tocher). The property of being uniformly more powerful is easy to prove. Typically the B test will be strictly preferable to the Fisher exact test

as it will sometimes reject the null hypothesis when the Fisher exact test will not. We say typically as this statement has not been proven but turns out to be true in all cases we have investigated. For instance, when  $n_1 = n_2 = 20$  then the Fisher exact test is 0.22 uniformly more powerful than the B test, in other words the power of the Fisher exact test is up to 22% lower than that of Boschloo. The test of Boschloo seems not to be used very often.

A popular test for comparing Bernoulli distributions is the Z test. There are two versions depending on how the variance is calculated. Both versions yield in balanced samples the same recommendation as shown by Suissa and Shuster (1985). Moreover, in balanced samples Suissa and Shuster (1985) have verified numerically that the Z test is uniformly more powerful than the Fisher exact test for  $\alpha \in \{1\%, 2.5\%, 5\%\}$  and  $10 \leq n_1 = n_2 \leq 150$ . However when the sample is not balanced then in the examples we have considered we have found that pooled variance performs better than unpooled variance in terms of minimax regret. We also expect that the Z test with pooled variance is uniformly more powerful than the Fisher exact test although we have not performed enough numerical simulations to be confident about this conjecture.

Depending on sample sizes, the Z test with pooled variance can be uniformly more powerful than the B test, the opposite may be true and the two tests may not be comparable in terms of being uniformly more powerful. In the later case one has to choose some alternative objective to select tests. In any case one has to first compare the power of these two tests before analyzing the data. For instance, for  $\alpha = 0.05$  we find that the two tests have identical power when  $n_1 = n_2 = 5$ . If  $n_1 = n_2 = 15$  then the test of Boschloo is uniformly more powerful than the Z test, for  $n_1 = n_2 \in \{10, 50\}$  the Z test is uniformly more powerful than the test of Boschloo. For  $n_1 = n_2 \in \{20, 25, 30, 40\}$  numerical calculations show that the Z test is  $\varepsilon$  uniformly more powerful than the test of Boschloo for  $\varepsilon \leq 4.3\%$  while the test of Boschloo is  $\varepsilon$  uniformly more powerful than the Z test for some  $\varepsilon \geq 22\%$ . Thus the Z test seems preferable over the B test for these sample sizes. Based on the above examples we find a tendency to prefer the Z test in balanced samples when sample sizes are not too small. The relationship is more intricate in unbalanced samples

which we illustrate in the table below.

$n_1, n_2$ ( $\alpha = 0.05$ )	17, 23	15, 25	14, 26	13, 27	10, 30
$\varepsilon$ s.t. $B \succ_{\varepsilon} Z$	0.21	0	0.29	0.29	0.07
$\varepsilon$ s.t. $Z \succ_{\varepsilon} B$	0	0.196	0.07	0.18	0.16
max regret of $B$	0.14	0.14	0.14	0.15	0.16
max regret of $Z$	0.14	0.15	0.15	0.16	0.175

In particular, if  $n_1 = 10$  and  $n_2 = 30$  then the B test is preferable to the Z test according to the  $\varepsilon$  dominance criterion as well as according to the minimax regret criterion.

To complete this discussion we present the two tests mentioned above. Each test is associated to some test statistic  $\varphi$  which is a function of the observations. The test then recommends to reject the null hypothesis if the set of observations that generate a lower value of this statistic can never be greater than  $\alpha$  when the null hypothesis is true. For each test, the test statistic is only a function of the total number of successes  $s_k$  observed in each sample, so  $s_k = \sum_{i=1}^{n_i} y_{k,i}$  for  $k = 1, 2$ . So the idea is to reject the null hypothesis if  $\varphi(s_1, s_2) \geq \varphi_\alpha$  where  $\varphi_\alpha$  is the largest value of  $f$  that satisfies

$$\sum_{a=0}^{n_1} \sum_{b=0}^{n_2} \binom{n_1}{a} (EY_1)^a (1 - EY_1)^{n_1-a} \binom{n_2}{b} (EY_2)^b (1 - EY_2)^{n_2-b} 1_{\{\varphi(a,b) \geq f\}} \leq \alpha$$

whenever  $EY_1 \geq EY_2$  where  $1_{\{\varphi(a,b) \geq f\}} = 1$  if  $\varphi(a, b) \geq f$  and  $1_{\{\varphi(a,b) \geq f\}} = 0$  if  $\varphi(a, b) < f$ .  $\varphi_\alpha$  is called the *critical value*. Typically the maximum is attained when  $EY_1 = EY_2$ . For the test of Boschloo the test statistic  $\varphi$  is given by

$$\varphi(s_1, s_2) = \sum_{j=s_2}^{\min\{n_2, s_1+s_2\}} \frac{\binom{n_1}{s_1+s_2-j} \binom{n_2}{j}}{\binom{n_1+n_2}{s_1+s_2}} \quad (1)$$

which is equal to the p value of the Fisher exact test. For the Z test with pooled variance the test statistic  $\varphi$  is equal to

$$\varphi(s_1, s_2) = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\frac{s_1+s_2}{n_1+n_2} \left(1 - \frac{s_1+s_2}{n_1+n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (2)$$

where  $\varphi(s_1, s_2) = 0$  if  $s_1 + s_2 \in \{0, n_1 + n_2\}$  and where  $\bar{y}_k = s_k/n_k$  denotes the empirical mean of the sample  $k$ ,  $k = 1, 2$ .<sup>2</sup>

<sup>2</sup>The critical values for the Z test for balanced samples and equi-tailed tests provided in Suissa and

To illustrate, assume that  $n_1 = 6$ ,  $s_1 = 5$ ,  $n_2 = 3$  and  $s_2 = 0$ . Entering these values into (1) we obtain the p value of the Fisher exact test 4.7%. Entering this as critical value and summing up the probability of all observations that would yield a higher critical value we obtain the p value of the Boschloo test 1.5%. Entering the values into (2) we obtain the critical value for the Z test 2.372 and summing up the probability of all observations that would yield a higher value yields p value 1.5%.

As explained earlier, the above can also be used to determine minimal sample sizes. For example, say one wishes to gather a balanced sample (so  $n_1 = n_2$ ) and wishes given size 5% to ensure type II error at most 20% whenever the treatment effect  $\bar{d}$  is at least 0.3. Computing critical values and then maximal type II errors one finds that the Z test requires  $n_1 = n_2 \geq 37$ . One may wonder if this can be outperformed as no finite sample optimality properties have been shown for the Z test. Schlag (2007b) computes lower bounds by considering randomized tests and shows that there is no test that obtains our objective with  $n_1 = n_2 < 34$ .

### 2.2.2 Tests of Equality

Assume now that we wish to test the null hypothesis that the two means are equal, which for Bernoulli distributions is equivalent to testing that the two distributions are identical. The standard approach is to choose a single critical value  $\varphi_\alpha$  such that the probability of generating a larger value with the test statistic for the above test of  $\Pr(Y_1 = 1) \geq \Pr(Y_2 = 1)$  or for the analogous test statistic for the test of  $\Pr(Y_1 = 1) \leq \Pr(Y_2 = 1)$  is at most  $\alpha$ . In applications this is equivalent to rejecting  $\Pr(Y_1 = 1) = \Pr(Y_2 = 1)$  if either  $\Pr(Y_1 = 1) \geq \Pr(Y_2 = 1)$  or  $\Pr(Y_1 = 1) \leq \Pr(Y_2 = 1)$  can be rejected at level  $\alpha/2$ .

In the literature one limits attention to *equi-tailed tests*, more out of computational simplicity than out of true preference. Here one combines the one-sided test for  $\Pr(Y_1 = 1) \leq \Pr(Y_2 = 1)$  with that for  $\Pr(Y_1 = 1) \geq \Pr(Y_2 = 1)$ , both for level  $\alpha/2$  to obtain a two-sided test with level  $\alpha$ . Note that as the rejection regions are disjoint, the size of the combined test is equal to the sum of the sizes of each of the separate tests. A uniformly most powerful unbiased test does not exist. The combination 

---

Shuster (1985) are not calculated based on the complete set of possible distributions and hence should be considered with caution. Therein it is ruled out by assumption that  $EY_1 = EY_2 \notin [0.05, 0.95]$ .

of the two UMPU tests of size  $\alpha/2$  generates a randomized equi-tailed test of size  $\alpha$  (which we do not expect to be unbiased). The Boschloo test is uniformly more powerful than the two-sided equi-tailed version of the Fisher exact test.

Note that nonrandomized equi-tailed tests have the following disadvantage. Typically the critical value  $\varphi_\alpha$  has the property that it yields a size of the test strictly below  $\alpha$  due to the discrete nature of the summation and nonrandomized feature of the test. Size being below the level means that power is sacrificed. The size of the two one-sided level  $\alpha/2$  tests fall short of  $\alpha/2$  by the same amount due to symmetry. Thus, by combining these two one-sided test to a two-sided level  $\alpha$  test, inefficiency of each test is inherent in having a size below  $\alpha/2$  is doubled.

We illustrate. If  $n_1 = n_2 = 10$  then the Boschloo test has critical value  $\varphi_{\alpha/2} = 0.07$  and the associated test has size 0.021. The Z test has critical value 1.96 and has the identical power. So the two-sided test for level  $\alpha$  in fact has size  $2 * 0.021 = 0.042$ .

## 2.3 Multi-Valued Distributions

Next we assume that there are more than two possible outcomes, so  $|Z| \geq 3$ . We will now present tests based on different ways of comparing the two underlying distributions. The appropriate measure for this comparison depends on the application. Note that when there are only two possible outcomes then two distributions are identical if and only if they have the same median if and only if they have the same mean.

### 2.3.1 Identity of Two Distributions

We first wish to test the null hypothesis that the two underlying distributions are identical, so  $H_0 : P_{Y_1} \equiv P_{Y_2}$  against  $H_1 : P_{Y_1} \not\equiv P_{Y_2}$ . For instance, one may have that  $Y_1$  is the outcome of the control treatment while  $Y_2$  is the outcome of a new treatment. Rejecting the null hypothesis then means that there is significant evidence that new treatment generates a different distribution than the control. Note however that one cannot make any inference on how this difference is substantiated. It could be that the means are different, it could be that there is a difference in the dispersion, etc.. One is not allowed to make this inference based on graphing the data. The alternative hypothesis contains the inference that one is interested in making.

The hypothesis of identical distributions are typically tested using permutation

tests and there are many such tests.<sup>3</sup> Permutation tests compare the specific samples to the set of samples that emerges if one puts all the data into one "pot" and then assigns the data randomly to one of the two treatments, preserving the given sample sizes. When the null hypothesis is true, then any reordering of the data can occur equally likely. If the specific sample is sufficiently extreme then the permutation test recommends to reject the null hypothesis. It seems like the observed sample was not a random allocation of outcomes to the two samples. Extremeness is measured in terms of some test statistic, the cutoff is again called the critical value. Permutation tests are unbiased if the underlying distributions contain no point masses. If there are point masses (i.e. some outcomes occur with strictly positive probability) then ties can occur and it is possible that there is no cutoff such that exactly 5% of the permuted samples are more extreme. In this case, as one is interested in nonrandomized tests, one chooses as critical value the largest cutoff such that the level is below 5%. Consequently the test typically becomes biased when there are point masses.

A popular permutation test is the *Mann Whitney U test* which is equivalent to the *Wilcoxon rank sum test* (1945), the latter was originally only described for balanced samples. Consequently, some call it the *Wilcoxon Mann Whitney test*. We present the test in the version of Wilcoxon which clearly identifies it as a permutation test. The test statistic is the sum of the ranks in one of the two samples where ranks are determined in terms of the data in both samples. The Wilcoxon Mann Whitney test in fact can be used to test the one sided null hypothesis that  $P_{Y_1}$  first order stochastically dominates  $P_{Y_2}$ . Thus when rejecting the null hypothesis one can infer that  $P_{Y_1}$  does not first order stochastically dominate  $P_{Y_2}$ . One however cannot infer anything about the underlying means or medians nor about whether or not  $P_{Y_1}$  is first order stochastically dominated by  $P_{Y_2}$ . One could of course also set up the Wilcoxon Mann Whitney test as a two-sided test, rejecting the null hypothesis if the sum of the ranks is either too large or too small. This is the typical approach when one has no a priori reason for expecting either distribution to dominate the other. This is not a formal reason.

Note that the Wilcoxon Mann Whitney test is uniformly most powerful among

---

<sup>3</sup>For an overview of permutation tests see Pesarin (2001) or Good (2005) where the later is a good introduction but is less formal. Lehmann and Romano (2005) also have a short but very precise presentation.

all unbiased tests that are invariant to monotone transformations of the data when the underlying distributions are continuous. This invariance has advantages if one wishes to compare distributions across different groups where it is not clear how each group evaluates outcomes such as money. As long as the ordering of these outcomes in each group coincides then the recommendation of the Wilcoxon Mann Whitney test will not depend on how outcomes are evaluated. The disadvantage of this invariance can be best seen if one considers a transformation of some bounded data that preserves the ranking such that the transformed data are all very close to say 1. Then the recommendation of the Wilcoxon Mann Whitney will remain unchanged. In particular, one may recommend a rejection even though all outcomes are approximately equal to 1. The problem is that when outcomes are so close to each other then measurement error can play a big role. It does not sound reasonable to impose invariance in such extreme situations. However it is possibly that some test that is not invariant under these extreme situations is more powerful than the Wilcoxon Mann Whitney test. In other words, invariance is extremely restrictive, ruling out many tests and thus restricting attention only to a small set of tests among which there is a complete ranking and thus selecting the Wilcoxon Mann Whitney test.

Permutation tests can be most powerful among the set of unbiased tests when the alternative hypothesis contains only a specific distribution (or a small class of distributions) provided the test statistic is chosen appropriately.<sup>4</sup> However we are interested in the power of a given test across all elements belonging to the alternative or at least in the type II error conditional on some set of interest. The problem is that here there seems to be no literature on this topic. In finite sample sizes I have seen no formal results on power. Note that simulations will not do the job unless the set of outcomes is explicitly limited as otherwise the space of distributions one is interested in is too large. For instance, when comparing Bernoulli distributions we

---

<sup>4</sup>For instance, the Fisher-Pitman test is uniformly most powerful when the alternative hypothesis consists of all normal distribution in which both samples have the same variance (see Lehmann and Romano, 2005). The Fisher-Pitman test recommends to reject the null hypothesis if the sum of observations in the one sample is sufficiently large. In fact, for large samples the Fisher-Pitman test approximates the two sample t test (whether or not this approximation is uniform is not clear) (see Lehmann and Romano, 2005).

can simulate the entire parameter space as this is two dimensional.

Without more research on this topic we recommend the one-sided Wilcoxon Mann Whitney test, most of all because it is the most popular of these tests and hence one can to some extent credibly claim that the test was not chosen among many others simply because it yields a rejection. However I wish to note that the Kolmogorov-Smirnov test is a very powerful alternative (see Table 1 in Sefton, 1992). It is not a permutation test but it is an exact test for testing identity of two distributions. The invariance to monotone transformation intrinsic in the Wilcoxon Mann Whitney test has its costs. In particular own simulations have shown that the Wilcoxon Mann Whitney test is not good at discovering differences when the two distributions have mean 0 but differ according to their spread.

### 2.3.2 Independence of Two Distributions

Consider an independent sample of matched pairs  $(y_{1,i}, y_{2,i})_{i=1}^n$  drawn from  $Y_1 \times Y_2$  where one wishes to test the null hypothesis that  $Y_1$  and  $Y_2$  are independent. This is a very nice application of permutation tests (see Romano, 1989, Pesarin, 2001, Example 3.2). Choose some test statistic such as  $\sup_{A \in \mathcal{A}} |P(A) - P_1(A)P_2(A)|$  for some “suitable” set  $\mathcal{A}$ . The permutations consist of any permutation  $\pi$  of  $\{1, \dots, n\}$  and then apply the test statistic to  $(y_{1i}, y_{2,\pi(i)})_{i=1}^n$ . If the null is true then all these fictitious samples derived by these permutations are equally likely.

### 2.3.3 Medians

Whether or not medians are the parameter of interest depends on the application.<sup>5</sup> Unfortunately there is only a very crude test for comparing medians (see below) that does not seem to be very powerful. Other tests used in the literature implicitly consider smaller null hypotheses, such as identical distributions. See for instance

---

<sup>5</sup>Decision theoretically speaking one can argue that means become the only parameters of interest once outcomes have been transformed appropriately. To expand briefly, assume the axioms of expected utility theory. If distributions are known then the action is chosen that yields the maximal expected utility. This is the action that has the highest mean provided outcomes are transformed into utilities. When distributions are not known then it is natural to maintain the focus on means. One can then either solve the problem of not knowing distributions by reverting to subjective expected utility theory or by running statistics or statistical decision theory.

Forsyth et al. (1994) for simulations showing that the Wilcoxon Mann Whitney test is not an exact test for comparing medians. In the next section we show how to compare means.

Consider the two-sided test that recommends to reject the null hypothesis that the two medians are equal whenever the  $1 - \alpha$  confidence intervals of each of the two medians do not intersect. These confidence intervals should be derived from the sign test. If the confidence intervals each have coverage  $1 - \alpha$  then the test we described has level  $(1 - (1 - \alpha)^2)$ . So for instance, if  $\alpha = 0.05$  then the test has level 0.0975, if  $\alpha = 0.025321$  then the test has level 0.05. The proof that this test is exact and has level  $(1 - (1 - \alpha)^2)$  is simple. If the null is true, so the two medians are equal, then the test does not reject the null if both confidence intervals cover the median. This occurs with probability  $(1 - \alpha)^2$  as the samples are independent. So the probability of rejecting the null is at most  $1 - (1 - \alpha)^2$ . This is actually called the Sidak correction.

Similarly one can construct a one-sided test of the null hypothesis that the median of the first sample is larger than that of the second. Here one rejects the null hypothesis if the upper bound for the median of the first sample is below the lower bound for the median of the second sample, if each bound has coverage  $1 - \alpha$  then the resulting test again has level  $(1 - (1 - \alpha)^2)$ .

Clearly the above test can be slightly improved in practice as tests rarely have size equal to proposed level. Hence if the confidence interval of the one median has coverage  $1 - \alpha_1$  then the coverage  $1 - \alpha_2$  of the other can be chosen such that  $1 - (1 - \alpha_1)(1 - \alpha_2) \leq \alpha$ .

### 2.3.4 Stochastic Inequalities and Categorical Data

A very useful and powerful criterion in small samples is to compare two random variables by comparing a random draw of each of them. To present this in a symmetric fashion one can test a so-called *stochastic inequality*, test  $H_0 : \Pr(Y_2 > Y_1) \leq \Pr(Y_2 < Y_1)$  against  $H_1 : \Pr(Y_2 > Y_1) > \Pr(Y_2 < Y_1)$  (e.g. see Brunner and Munzel, 2000). Upon rejection one says that “ $Y_2$  tends to generate larger outcomes than  $Y_1$ .” Note that the situation where  $Y_1$  and  $Y_2$  are independent belongs to the null hypothesis. It is as if one tests  $H_0 : \delta \leq 0$  against  $H_1 : \delta > 0$  where

$$\delta = \Pr(Y_2 > Y_1) - \Pr(Y_2 < Y_1).$$

The two-sided test that specifies  $H_0 : \delta = 0$  is called a test of *stochastic equality*.<sup>6</sup> Any test of this null hypothesis will be invariant to monotone transformations of the data, provided the same transformation is applied to both components. In other words, the above hypothesis can be tested if the data is ordinal, such as defined by categories "very bad", "bad", "good" and "very good". Note that  $\delta = 0$  if  $Y_1$  and  $Y_2$  are both symmetric and have the same median or if  $Y_1$  and  $Y_2$  are identically distributed while  $\delta \neq 0$  can be true even if they both have the same median or the same mean.

We construct a test for the one-sided null hypothesis that involve three steps and an exogenous threshold  $\theta$ . Assume  $\theta = 0.2$  or else determine  $\theta$  separately according to some optimality criterion (for more see Subsection 2.3.5 below). Given  $\theta$  proceed as follows. First randomly match the two samples into pairs, dropping the unmatched observations of the larger sample. Then evaluate the randomized version of McNemar's test (which is UMPU) to this set of matched pairs. Finally, reject the null if the expected rejection probability resulting from the first two steps is above  $\theta$ . In practice, instead of determining this expectation analytically, one calculates the average rejection probability after repeating the first two steps a large number of times.

The above test is an exact test with level  $\alpha$ . The randomized test that emerges after the second step has size  $\theta\alpha$  and is in fact unbiased. Its level is set equal to  $\theta\alpha$  in order to adjust for mistake intrinsic in applying the threshold rule in the third step and to ensure that the level of the final nonrandomized test is below  $\alpha$ . Clearly the nonrandomized test is no longer unbiased.

We would like to point out that the Wilcoxon Mann Whitney test is not an exact test for a stochastic inequality but only a test for identity of two distributions. To see this we refer to the simulations of Forsythe et al. (1994, Table V) who show that the type I error is above its nominal level when comparing uniform distributions that have the same median.

### 2.3.5 Means

Assume now that the set of possible outcomes  $Z$  is known ex-ante to be bounded. Let  $a = \inf \{z \in Z\}$  and  $b = \sup \{z \in Z\}$  so  $Z \subseteq [a, b]$ . We wish to test  $EY_1 \geq EY_2$

---

<sup>6</sup>For an unbiased estimate of  $\delta$  take the expected value attained when matching the data from the two samples at random into pairs, ignoring the remaining elements of the larger sample.

against  $EY_1 < EY_2$ . There are no assumptions on the distributions and so for instance variance is allowed to differ between the two variables. A common test for these hypotheses is the Wilcoxon Mann Whitney test. However this test is not exact. To see this consider the following example (or consult the simulations of Forsythe et al., 1994)

$Y_1$	0	1
$Y_2$	$1 - \varepsilon$	$1 - \varepsilon$
prob.	$\varepsilon$	$1 - \varepsilon$

where  $\varepsilon > 0$  but small. Then  $EY_1 = EY_2$ . However, given  $n_1$  if  $\varepsilon$  is sufficiently small then  $Y_1 = 0$  will not be observed and the Wilcoxon Mann Whitney test will reject the null hypothesis if  $n_1$  and  $n_2$  are not too small.

We now present the first test for these hypotheses which is due to Schlag (2007c). The test consists of five steps and relies on a so-called *threshold*  $\theta \in (0, 1)$  that has to be determined separately. The default is to choose  $\theta = 0.2$ .

We describe the five steps. (1) Normalize the outcomes linearly so that  $a = 0$  and  $b = 1$ , so replace  $y$  by  $(y - a) / (b - a)$ . So we now act as if  $Y_1, Y_2 \in [0, 1]$ . (2) Consider each observation  $y_{ki}$  and replace it with 1 with probability  $y_{ki}$  and with 0 with probability  $1 - y_{ki}$ . Do this with all observations to then obtain a sample of binary valued outcomes. This is called the *binomial transformation*. (3) Act as if the binary valued sample was the original sample and test the null hypothesis using a test that has level  $\theta\alpha$  when  $Y_i \in \{0, 1\}$ . Record the probability of rejecting the null hypothesis. (4) Repeat steps 2 and 3 infinitely often and record the average probability of rejection of the null hypothesis under this inserted test. (5) The recommendation is to reject the null hypothesis if and only if the average probability of rejection at the end of step 4 was greater or equal to  $\theta$ .

The above test has level  $\alpha$  and type II error bounded above by

$$\frac{\beta_\theta(EY_1, EY_2)}{1 - \theta}$$

where  $\beta_\theta(EY_1, EY_2)$  is the type II error of the test of level  $\theta\alpha$  for comparing two binary valued samples that was used in step 3, evaluated when  $EY_i$  is the true probability that an outcome in sample  $i$  is a success.

To understand the performance of this test it is useful to know that the randomized test that stops after step 4 with a probability of rejecting the null hypothesis has

level  $\theta\alpha$  and type II error  $\beta_\theta(EY_1, EY_2)$ . Thus, the addition of step 5 to generate a nonrandomized test causes an increase in level by a factor of  $1/\theta$  and in increase in type II error by a factor of  $1/(1-\theta)$ . In this light, consider some test used in step 3 as function of its level  $\theta\alpha$ . If the threshold  $\theta$  is decreased then the type II error  $\beta_\theta(EY_1, EY_2)$  of the test used in step 3 and hence of the test given by steps 1 – 4 is increased. At the same time, a decrease in the threshold  $\theta$  also causes a decrease in the factor  $1/(1-\theta)$ . So the overall impact of a decrease in  $\theta$  on the power of the test given by steps 1 – 5 is ambiguous. We suggest here to select  $\theta$  to minimize maximum regret.

The test used in step 3 is a one-sided test for the null hypothesis that two underlying Bernoulli random variables have the same underlying mean. This test may be randomized as randomness is eliminated anyway in step 5. For given  $\theta$ , more powerful tests are preferable for step 3 as these also yield a lower bound on type II error.

It turns out that two tests are useful for step 3, the Z test with pooled variance and the UMPU test. The Z test with pooled variance seems to work well when sample sizes are similar while the UMPU test performs well in sufficiently unbalanced samples. The following values are taken from Schlag (2007c). The table shows for some values of  $n_1 = n_2$  and  $\alpha = 0.05$  the critical values for the Z test and the threshold  $\theta$  that minimizes maximum regret.

Parameters when using the Z test with Pooled Variance  
Balanced Samples - One-Sided Tests

$n_1 = n_2$	10	20	30	40	50
$\alpha = 0.05$					
cutoff $\varphi$	2.4	2.38	2.5	2.38	2.31
$\theta$	0.12779	0.16716	0.13502	0.18351	0.20994

We now present nonrandomized tests for shifted null hypotheses. We wish to test  $H_0 : EY_1 \geq EY_2 + d_0$ . Given our method for constructing distribution-free tests it is sufficient to consider  $Y_1$  and  $Y_2$  binary valued. Note that the existing tests for noninferiority are nonrandomized and very intricate and we know of no analysis of randomized tests. Here we present a simple randomized test that has type II error for  $H_1 : EY_1 \leq EY_2 + d$  for  $d < d_0$  that comes very close to the minimal type II error

when  $d \leq 0$ . This is ensured by using the UMPU test. The idea is to invert the test. One reverses the role of the null and of the alternative hypothesis and adjusts the size of the original test, denoted by  $\alpha_1$ , to create a new test with size  $\alpha$ . Consider for instance the case where  $d_0 > 0$ . Let  $\alpha_1$  be such that the UMPU test  $\phi_{\alpha_1}^u$  with size  $\alpha_1$  for testing  $H'_0 : EY_2 \geq EY_1$  that has type II error  $\alpha$  when  $H'_1 = H_0$ . In other words, the minimal probability of rejection under  $\phi_{\alpha_1}^u$  conditional on  $Y \in H_0$  is equal to  $1 - \alpha$ . It then follows that  $1 - \phi^u$  is a test of  $H_0$  that has size  $\alpha$ . In particular,  $\alpha_1 < 1 - \alpha$  and if  $d_0 \rightarrow 0$  then  $\alpha_1 \rightarrow 1 - \alpha$ . Now consider the case where  $d_0 < 0$ . We follow the same construction and search for  $\alpha_1$  such that the minimal probability of rejection when  $Y \in H_0$  is equal to  $1 - \alpha$  under  $\phi_{\alpha_1}^u$  that tests  $H'_0 : EY_2 \geq EY_1$ . Note that  $d_0 < 0$  implies  $\alpha_1 > 1 - \alpha$ .

Finally we present confidence intervals for the difference between the two means. These are very easy to compute based on the above. Here is the algorithm to construct upper confidence bounds, lower confidence bounds are derived analogously. Specify a nominal level  $\alpha$  and a threshold  $\theta$  (the default is to set  $\theta = 0.2$ ). Consider the UMPU test  $\phi_{\alpha'}^u$  with size  $\alpha'$  of  $H_0 : EY_1 \leq EY_2$ . For a given set of data find the largest value of  $\alpha'$ , denote this by  $\alpha''$ , such that the rejection probability of this randomized test is at most  $1 - \theta$ . Now use the power function of  $\phi_{\alpha''}^u$  to compute the value of  $d$  such that the minimal probability of rejection under  $\phi_{\alpha''}^u$  when  $EY_1 - EY_2 \geq d$  is equal to  $1 - \theta\alpha$ . Then  $d$  is an exact  $1 - \alpha$  upper confidence bound for  $EY_1 - EY_2$ .

Some notes are in place. Formally,  $d$  satisfies

$$\min_{Y: EY_1 - EY_2 \geq d} EY \phi_{\alpha'}^u = 1 - \theta\alpha.$$

Typically this minimal probability will be attained on the diagonal where  $EY_1 + EY_2 = 1$ . If  $\alpha'' < 1 - \theta\alpha$  then the second step above is equivalent to finding  $d > 0$  such that the type II error under  $\phi_{\alpha''}^u$  under  $H_1 : EY_1 \geq EY_2 + d$  is equal to  $\theta\alpha$ . If  $\alpha'' = 1 - \theta\alpha$  then  $d = 0$ . If  $\alpha'' > 1 - \theta\alpha$  then  $d < 0$  and  $\{EY_1 \geq EY_2 + d\}$  is contained in the null hypothesis and hence the algorithm is specified in terms of minimal probability of rejection.

We verify the above claim. We first show that  $H_0 : EY_1 - EY_2 \geq d_1$  is rejected under our nonrandomized test for this shifted null hypothesis if and only if  $d_1 \geq d$ . We first construct the randomized test with size  $\theta\alpha$ . For this we search for  $\alpha_1(d_1)$  such that the minimal probability of rejection under  $\phi_{\alpha_1}^u$  is equal to  $1 - \theta\alpha$  when

$EY_1 - EY_2 \geq d_1$ . The nonrandomized test then recommends to reject  $H_0$  if  $1 - \phi_{\alpha_1}^u \geq \theta$ . Now given our above construction,  $\alpha_1(d) = \alpha''$  and  $\phi_{\alpha_1(d)}^u = 1 - \theta$  where the latter follows from the fact that  $\phi_{\alpha''}^u < 1 - \theta$  implies that  $\alpha''$  is not maximal. Hence the null is rejected within this family of nonrandomized tests for the shifted null hypothesis if and only if  $d_1 \geq d$ . This means that  $d$  is an upper confidence bound for  $EY_1 - EY_2$ .

**Comment on the Two Sample T Test** A popular exact test for comparing means in two independent samples is the two sample t test which assumes that the underlying random variables are normally distributed and that their variances are equal. When including all assumptions in the hypotheses this means that the t test is an exact test for identity of two normal distributions, here

$$W = \{(P_{Y_1}, P_{Y_2}) : P_{Y_1} \equiv P_{Y_2}, P_{Y_1} \text{ is normally distributed}\}.$$

So a rejection means can mean one of three things (i) the data is normally distributed with equal variance but the means are different, (ii) the data is normally distributed but the variances are different, or (iii) the data is not normally distributed. We thus do not find the t test to be very useful to uncover the true characteristics of the data. Note that the classic presentation of the setting that motivates the t test limits attention to normally distributed variables that differ only in terms of the mean. Rejections are then interpreted as the means being different which is only true as by assumption one has ruled out that one of the higher moments are different. Of course the two sample t test is most powerful when the assumptions of normality and equal variance are satisfied, i.e. assumed. Some also justify use of the two sample t test based on large sample arguments, the faults of this argumentation are pointed out in Section 5.

### 2.3.6 Variances

We briefly recall a trick due to Walsh (1962) that allows one to compare variances of two independent samples. Consider the sample of observations of  $Y_1$ . Assume that there is an even number of observations. Randomly group these into pairs  $(y_{g_1(i)}, y_{g_2(i)})$ ,  $i = 1, \dots, n_1/2$  and consider the sample  $z_{1i} = (y_{g_1(i)} - y_{g_2(i)})^2/2$  for  $i = 1, \dots, n_1/2$ . Each pairings should be selected equally likely. Then we have  $n_1/2$  independent observations and the expected value of each is equal to  $VarY_1$ . So in order

to compare the variances of  $Y_1$  and  $Y_2$  based on  $n_i$  independent observations of  $Y_i$  one can compare the means of two independent samples of  $n_1/2$  and  $n_2/2$  observations each. Thus we reduce the problem of comparing variances to a problem of comparing means. How much information is lost in this reduction is not known. However, the bounds derived for comparing means can also be applied here. For details we refer to Schlag (2007c).

### 3 Matched Pairs

Now assume that data is given by  $N$  independent pairs, formally the data is given by  $(y_{1i}, y_{2i})$ ,  $i = 1, \dots, N$  where  $(y_{1i}, y_{2i})$  are iid draws from  $(P_{Y_1}, P_{Y_2})$ .

#### 3.1 Binary-Valued Distributions

Assume that  $Z = \{0, 1\}$ .

##### 3.1.1 Probabilities

We wish to test  $H_0 : \Pr(Y_1 = 1) \leq \Pr(Y_2 = 1)$ . Lehmann (1959) extended the test of McNemar (1947) to a UMPU test. This UMPU test compares the occurrences of  $(0, 1)$  relative to those of  $(1, 0)$  using the randomized binomial test. Following the analysis of Schlag (2007b) the UMPU test minimizes type II error whenever the statistician is only concerned with the difference between the two means, for instance when  $H_1 : \Pr(Y_2 = 1) \geq \Pr(Y_1 = 1) + d$  for some  $d > 0$ .<sup>7</sup> The nonrandomized version is also called the McNemar's exact test (McNemar himself considered an approximate nonrandomized version based on the  $\chi^2$  test). The randomized version involves using the sign test to compare the occurrences of  $(0, 1)$  relative to those of  $(1, 0)$ . However the nonrandomized version should not be used. Following the numerical calculations of Suissa and Shuster (1991), the respective Z test is uniformly more powerful than the nonrandomized exact McNemar's test for  $10 \leq N \leq 200$  and  $\alpha \in \{1\%, 2.5\%, 5\%\}$ . The Z statistic is here given by

$$\frac{s_2 - s_1}{\sqrt{s_1 + s_2}}$$

---

<sup>7</sup>Conditional on  $EY_2 > EY_1$  it is most powerful when  $EY_1 + EY_2 = 1$  but not when  $EY_1 + EY_2 < 1$ .

where  $s_k$  is the number of successes observed of  $Y_k$ ,  $k = 1, 2$ . It is best to numerically compute the  $z$  statistic and not to refer to the tables in Suissa and Shuster (1991) as these only apply if one is investigating the null hypothesis  $H_0 : \Pr(Y_1 = 1) = \Pr(Y_2 = 1) \leq 0.995$  (see Suissa and Shuster, 1991, p. 363).

### 3.1.2 Correlation

There is a close connection between the UMPU test of Tocher (1950) for comparing means of two independent samples and tests for correlation. Consider testing  $H_0 : Cov(Y_1, Y_2) \leq 0$  against  $H_1 : Cov(Y_1, Y_2) > 0$ . Note that

$$Cov(Y_1, Y_2) = [\Pr(Y_1 = 1|Y_2 = 1) - \Pr(Y_1 = 1|Y_2 = 0)] VarY_2. \quad (3)$$

Thus, if  $VarY_2 > 0$  then this is equivalent to testing whether  $\Pr(Y_1 = 1|Y_2 = 1) \leq \Pr(Y_1 = 1|Y_2 = 0)$ . It is as if we are comparing the means of the two samples identified by whether  $Y_2 = 0$  or  $Y_2 = 1$ . In fact, following Tocher (1950), a UMPU test for testing for negative covariance is simply the UMPU test for testing the inequality  $\Pr(Y_1 = 1|Y_2 = 1) \leq \Pr(Y_1 = 1|Y_2 = 0)$  based on the two independent samples  $\{y_{1i} : y_{2i} = 1\}$  and  $\{y_{1i} : y_{2i} = 0\}$  where  $n_1 = \#\{i : y_{2i} = 1\}$  and  $n_2 = \#\{i : y_{2i} = 0\}$ . If either of these samples is vacuous then one simply rejects with probability  $\alpha$ . In particular this happens when  $VarY_2 = 0$ .

Given our discussion in Section 2.2.1 one can then select for a nonrandomized test based on the values of  $n_1$  and  $n_2$ , typically from either the B or the Z test. For computational simplicity the test of Boschloo is the natural candidate as it performs well in both balanced and unbalanced samples.

Note that tests for negative covariance will also be tests for negative correlation.

Note that the above formula (3) shows that one does not need an independent sample to test for negative correlation. Given  $N$  values  $(y_{2i})_{i=1}^N$  of  $Y_2$  it is enough that  $y_{1i}$  is drawn independently from  $Y_1|_{Y_2=y_{2i}}$ . In other words, the UMPU is a conditional test where inference is conditional on the values of  $Y_2$ . When interested in such a conditional test then obviously the Fisher exact test is the only candidate for implementing the nonrandomized test.

For more details on this matter see Schlag (2008).

## 3.2 Multi-Valued Distributions

Assume that  $|Z| \geq 3$ .

### 3.2.1 Identity of Two Marginal Distributions

Here we wish to test the null hypothesis that  $P_{Y_1} \equiv P_{Y_2}$  using a permutation test. The permutation is in terms of the order of the elements of each pair. If  $P_{Y_1} \equiv P_{Y_2}$  then pair  $(y_1, y_2)$  is equally likely to occur as pair  $(y_2, y_1)$ . Hence, under the null hypothesis we could have equally likely sampled  $(y_2, y_1)$ . The most popular permutation test for this application is due to Wilcoxon. For the one-sided test one first ranks the absolute differences between the two observations and then considers as test statistic the sum of those ranks that are associated to positive differences. It is not clear whether or not this test has any UMPU properties.

### 3.2.2 Stochastic Inequalities and Categorical Data

We present a test of the *stochastic inequality*  $H_0 : \delta \leq 0$  against  $H_1 : \delta > 0$  where  $\delta = \Pr(Y_2 > Y_1) - \Pr(Y_2 < Y_1)$  (e.g. see Brunner and Munzel, 2000).<sup>8</sup> This can test can be applied to categorical data as it is invariant to monotone transformations (for more details see Section 2.3.4).

For designing a test one can revert to the case of comparing means in matched pairs when outcomes are binary valued (see Subsection 3.1.1). Simply identify the outcomes  $y \in \mathbb{R}^2$  that satisfy  $y_2 < y_1$ ,  $y_2 = y_1$  or  $y_2 > y_1$  with the outcomes  $y = (1, 0)$ ,  $y = (0, 0)$  and  $y = (0, 1)$  respectively and proceed as if the transformed data were the original data. Following Section 3.1.1 we recommend as nonrandomized test to use the Z test. Note that the randomized sign test is UMPU.

For testing  $H_0 : \delta = 0$  one can combine two one-sided tests of level  $\alpha/2$ .

Note that  $\delta = p_2 - p_1$  holds after this identification provided that  $p_1$  (and  $p_2$ ) denotes the probability of assigning  $(1, 0)$  (and  $(0, 1)$ ). This means that confidence intervals for the difference between the marginal probabilities given matched binary-valued pairs are also confidence intervals for  $\delta$ .

---

<sup>8</sup>The empirical frequency of  $\delta$  is an unbiased estimator of  $\delta$ . In fact, it is UMVUE as  $1_{\{y_2 > y_1\}} - 1_{\{y_2 < y_1\}}$  is a complete sufficient statistic that is symmetric.

### 3.2.3 Means

Analogous to the case of independent samples we assume that  $Y_1, Y_2 \in Z \subseteq [a, b]$ . We wish to test  $H_0 : EY_1 \geq EY_2$  against  $H_1 : EY_1 < EY_2$ . Here we refer again to Schlag (2007c). As above, the sample is replaced by a binary valued sample and then the appropriate test is chosen. In this replacement,  $(y_{1i}, y_{2i})$  is replaced with  $(1, 1)$  with probability  $\min\{y_{1i}, y_{2i}\}$ , with  $(1, 0)$  with probability  $\max\{0, y_{1i} - y_{2i}\}$ , with  $(0, 0)$  with probability  $1 - \max\{y_{1i}, y_{2i}\}$  and with  $(0, 1)$  otherwise. The binary valued pairs are then tested using the exact randomized version of McNemar's test.

We now construct confidence intervals for  $EY_2 - EY_1$ . It is sufficient to show how to test  $H_0 : EY_1 \geq EY_2 + d_0$ . Given our methodology for dealing with means of random variables with bounded support it is sufficient to show how to do this when  $Y_1, Y_2 \in \{0, 1\}$ .

We cannot directly use the randomized McNemar's test as it drops  $(0, 0)$  and  $(1, 1)$  which is allowed when  $d_0 = 0$  but not when  $d_0 \neq 0$ . One could invert this test. However in practice, inverting a test is cumbersome as the thresholds have to be constructed by hand given the sample. As in our methodology the original sample is randomly transformed this would not be feasible. One would have to program this intricate procedure. Instead we present here an alternative test that is easy to implement. This test will be randomized and unbiased.

Consider first the test that  $Y_2$  is noninferior to  $Y_1$ , so where  $d_0 > 0$ . The first step is to replace any observation of  $(0, 0)$  or  $(1, 1)$  independently with probability  $\frac{d_0}{1+d_0}$  with  $(0, 1)$ . The next step is to test in the new sample whether the probability of observing  $(1, 0)$  conditional on observing either  $(1, 0)$  or  $(0, 1)$  is above  $\frac{1}{2}(1 + d_0)$  using the binomial test.

Given the transformation in the first step, it is as if the statistician is facing  $Y'$  with  $\Pr(Y' = (1, 0)) = \Pr(Y = (1, 0))$  and  $\Pr(Y' = (0, 1)) = \Pr(Y = (0, 1)) + \frac{d_0}{1+d_0} \Pr(Y \in \{(0, 0), (1, 1)\})$ . It is now easily verified whenever  $\Pr(Y' \in \{(1, 0), (0, 1)\}) > 0$  that

$$\begin{aligned} \frac{\Pr(Y' = (1, 0))}{\Pr(Y' \in \{(1, 0), (0, 1)\})} &\geq \frac{1}{2}(1 + d_0) \text{ if and only if} \\ d_0 &\leq \Pr(Y = (1, 0)) - \Pr(Y = (0, 1)) = EY_1 - EY_2. \end{aligned}$$

The above shows that our test has level  $\alpha$  whenever  $\Pr(Y' \in \{(1, 0), (0, 1)\}) > 0$ . On

the other hand, when  $\Pr(Y' \in \{(1, 0), (0, 1)\}) = 0$  then the binomial test recommends to reject with probability  $\alpha$  as neither  $(1, 0)$  nor  $(0, 1)$  will be observed which also fulfills the level condition. Hence our proposed test is exact.

Now consider the case where  $d_0 < 0$ . Here replace any observation of  $(0, 0)$  or  $(1, 1)$  independently with probability  $\frac{-d_0}{1-d_0}$  with  $(1, 0)$ . Let  $Y'$  be such that  $\Pr(Y' = (1, 0)) = \Pr(Y = (1, 0)) + \frac{-d_0}{1-d_0} \Pr(Y \in \{(0, 0), (1, 1)\})$  and  $\Pr(Y' = (0, 1)) = \Pr(Y = (0, 1))$ . The remaining arguments are as above, in particular the test based on the transformed sample is the same regardless of whether  $d_0 < 0$  or  $d_0 > 0$ .

For completeness we describe the randomized binomial test, here for the null that the success probability lies below  $p_0$ . Let  $s(j) = \sum_{i=j}^m \binom{m}{i} p_0^i (1-p_0)^{m-i}$  for  $j \in \{0, 1, \dots, m\}$  and  $s(m+1) = 0$ . Given a sample with  $k$  successes from a total of  $m$  observations,

reject with probability 1 if  $s(k) \leq \alpha$ ,  
 reject with probability  $\frac{\alpha - s(k+1)}{s(k) - s(k+1)}$  if  $s(k+1) < \alpha < s(k)$ , and  
 do not reject otherwise.

### 3.2.4 Covariance and Correlation

We wish to test  $H_0 : Cov(Y_1, Y_2) \leq \gamma$  against  $H_1 : Cov(Y_1, Y_2) > \gamma$  for some given  $\gamma$  when  $Y_i \in [a_i, b_i]$  for  $i = 1, 2$ .

We present two different tests. The first applies to any  $\gamma$  but is less powerful in practice when concerned with  $\gamma = 0$ . The second is difficult to implement when  $\gamma \neq 0$ .

For the first test we proceed as follows. Following Walsh (1962) we transform the problem into one of testing for the mean of a single sample. Consider the  $n/2$  independent observations  $\frac{1}{2}(y_{1,2i-1} - y_{1,2i}) \cdot (y_{2,2i-1} - y_{2,2i})$  that belong to  $[-\frac{1}{2}(b_1 - a_1)(b_2 - a_2), \frac{1}{2}(b_1 - a_1)(b_2 - a_2)]$ . Note that the mean of each observation is equal to the covariance of the original sample. Thus one can use the test for the mean of single sample shown below to test for the covariance given this sample of matched pairs. In order to avoid that changing the order of observations changes the results it is best to choose a random pairing. This method can then be used to establish confidence intervals of the covariance.

Initial numerical simulations show that the above test is not as powerful when  $\gamma = 0$  as the following alternative due to Schlag (2008). We propose the same five

steps as when testing for equality of means based on two independent samples. Simply insert in step (3) the UMPU test used to test for negative correlation in the binary valued case.

Finally we wish to discuss the Spearman (1904) rank correlation test which is the most popular nonparametric test for identifying correlation, hence concern is for  $\gamma = 0$ . It is a permutation test and hence is constructed for the case where the null hypothesis asserts independence.<sup>9</sup> It has been used as a test of correlation by defining the alternative hypothesis as asserting that there is strictly positive correlation. Distributions that are uncorrelated but dependent are dropped entirely from the investigation, some justify this by referring to an indifference zone. Note that any tests presented in this exposition do not have such an indifference zone. To clarify this hidden assumption of the Spearman rank correlation test consider the following example

$$\begin{array}{rcccc}
 Y_1 \backslash Y_2 & 1/2 - \varepsilon & 1/2 & 1 \\
 0 & \frac{1}{2(1+2\varepsilon)} & 0 & \frac{\varepsilon}{(1+2\varepsilon)} \\
 1 & 0 & 1/2 & 0
 \end{array}$$

where  $0 < \varepsilon < 1/2$ . Then  $Cov(Y_1, Y_2) = 0$  as  $E(Y_2|Y_1 = 1) = E(Y_2|Y_1 = 0)$ . So for given  $N$ , if  $\varepsilon$  is sufficiently small then it is essentially as if  $(1, 1/2)$  or  $(0, 1/2 - \varepsilon)$  each occur equally likely. With arbitrarily high probability the data will exhibit perfect correlation. The Spearman rank correlation test will reject the null hypothesis if one these two outcomes does not occur too rarely. Hence its true size is way above its nominal size, in fact its size converges to 1 as  $N$  tends to infinity. Another popular test uses instead as test statistic the Pearson's correlation coefficient. This similarly yields a high probability of wrongly rejecting the null hypothesis in the above example.

### 3.2.5 A Measure of Association Related to Kendall's Tau

There is an alternative measure of association between two random variables that is very useful as it tends to be very powerful in small samples. The idea is similar to a stochastic inequality where we compared two random realizations. Here we compare pairs of random realizations and compare whether or not changes are in the same directions. Specifically we investigate the sign of the change in  $Y_2$  as a function of

---

<sup>9</sup>The test statistic is the negative of the sum of the squared distances between the ranks.

the sign of the change in  $Y_1$ . In this framework one says that  $(y_{11}, y_{12})$  and  $(y_{21}, y_{22})$  is a *concordant pair* if  $(y_{12} - y_{11})(y_{22} - y_{21}) > 0$ , it is called a *discordant pair* if  $(y_{12} - y_{11})(y_{22} - y_{21}) < 0$ . We wish to test if the probability that two independent realizations of a bivariate random vector  $Y$  are more likely to be discordant than concordant. Let  $Y^1$  and  $Y^2$  be two independent copies of a bivariate random vector  $Y$  where  $Y^i = (Y_1^i, Y_2^i)$ . So we wish to test

$$\begin{aligned} H_0 & : \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) > 0) \leq \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) < 0) \text{ against} \\ H_1 & : \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) > 0) > \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) < 0). \end{aligned}$$

Analogous to the investigation of the stochastic inequality it is useful to consider the following statistic:

$$\tau'(Y) = \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) > 0) - \Pr((Y_2^1 - Y_1^1)(Y_2^2 - Y_1^2) < 0).$$

Then the above pair of hypotheses is equivalent to testing  $H_0 : \tau' \leq 0$  against  $H_1 : \tau' > 0$ . An unbiased estimator of  $\tau'$  is given by considering all possible pairings of the data and taking the difference between the number of concordant pairs and the number of discordant pairs and dividing this by the number of pairs. This estimator is called *Kendall's tau* and hence we denote the above property of  $Y$  by  $\tau'$ .<sup>10</sup> Note that when  $Y$  is binary valued, so  $Y \in \{0, 1\}^2$ , then  $\tau' = 2Cov(Y_1, Y_2)$ . Thus, up to a factor of two  $\tau'$  can be considered a natural extension of the concept of covariance.

The hypothesis test for  $H_0 : \tau' \leq 0$  is similar to that of the stochastic inequality in the setting of independent samples. Given a threshold  $\theta$ , randomly assign the data into pairs (of matched pairs), ignoring the unmatched data point if the sample is odd. Then record a rejection if the randomized binomial test with level  $\theta\alpha$  indicates that there are significantly more concordant than discordant pairs. Finally repeat the previous two steps infinitely often to then reject the null hypothesis if this repetition yields an average rejection probability above  $\theta$ .

---

<sup>10</sup>Note that Kendall (1938) only presented  $\tau$  without referring to it as an estimator or mentioning  $\tau'$ . In fact the concepts of  $\tau$  and  $\tau'$  were already discussed in earlier work, see Kruskal (1958) for the origins.

## 4 Single Sample

Finally we present tests that are based on a sample of  $N$  independent observations of a random variable  $Y$ .

### 4.1 Success Probabilities

For the case where  $Z = \{0, 1\}$  the random variable  $Y$  is Bernoulli distributed. The sum of successes in the sample is binomially distributed. We wish to test  $H_0 : EY \leq p_0$  against  $H_1 : EY > p_0$ . For this one uses the binomial test where there is some threshold  $k$  that depends on  $\alpha$  where the null hypothesis is rejected if strictly more than  $k$  successes are observed. The randomized version of the binomial test is UMP and it is also unbiased. Randomization only occurs at a threshold when the sum of successes is exactly equal to  $k$ . For the two-sided test of  $H_0 : EY = p_0$  there is no UMP test but there is a (not necessarily equi-tailed) UMPU test where cutoffs will depend on  $p_0$ . Again the nonrandomized version is not unbiased.

### 4.2 Median and Quantiles

One may choose to analyze the median  $m$  of a random variable  $Y$ . For the general case the median  $m(Y)$  is defined by  $\Pr(Y \leq m(Y)) \geq 1/2$  and  $\Pr(Y \geq m(Y)) \geq 1/2$ . For continuous random variables, it is defined by  $\Pr(Y \leq m(Y)) = 1/2$ . For instance, one may wish to analyze whether the underlying median  $m(Y)$  is strictly above a given threshold  $\bar{m}$ . Here one considers the complement as null and hence investigates  $H_0 : m(Y) \leq \bar{m}$  which is identical to testing  $H_0 : \Pr(Y \leq \bar{m}) \geq 1/2$ . This should be tested with the binomial test, recording events in which the outcome  $y_i$  falls below  $\bar{m}$  and then testing whether this probability is above  $1/2$ .

Similarly one can extended design an exact test for a quantile of some distribution. Given  $\kappa \in (0, 1)$  let the  $\kappa$  quantile is given by  $m_\kappa(Y)$  that satisfies  $\Pr(Y \leq m_\kappa) \geq 1-\kappa$  and  $\Pr(Y \geq m_\kappa) \geq \kappa$ .

### 4.3 Mean

There are a few exact tests for the mean of a single sample, we are aware of Romano and Wolf (2000) and Diouf and Dufour (2006). However these tests are very intricate

and hence have been only used very little in practice. We present a test along the lines of the one we presented for comparing two samples (see Schlag, 2007a). Determine a threshold  $\theta$ , linearly transform the data into  $[0, 1]$ , randomly transform the data into  $\{0, 1\}$ , repeatedly apply the UMP test for the Bernoulli case (see above) and finally recommend a rejection of the average rejection in the previous stage was above the threshold  $\theta$ . This procedure of course can also be used to derive confidence intervals. In this case, one should select  $\theta$  to minimize upper bound on its expected width.

#### 4.4 Variance

Romano and Wolf (2002) have constructed an exact test for the variance of a random variable given an independent sample. Similar to the mean test of these authors, it is very intricate and difficult to implement.

We present a very simple exact test by reducing the problem to one of testing the mean of a sample with half as many observations, analogously to Section 2.3.6. In the case of two independent samples we first applied the transformation of Walsh and then we applied the binomial transformation. Here we have the alternative to first apply the binomial transformation and then the transformation of Walsh as the resulting sample only has outcomes  $-0.5$ ,  $0$  and  $0.5$  for which we can then use the randomized version of McNemar's test. This latter alternative turns out to be more effective.

More specifically, assume that  $Y$  has been normalized so that  $Y \in [0, 1]$ . Consider  $H_0 : VarY \geq \sigma_0^2$  and  $H_1 : VarY < \sigma_0^2$ . We suggest the following test. First apply the binomial transformation to generate a binary valued sample. Then randomly group the observations in pairs, dropping the unmatched observation when the sample size is odd, and apply the transformation of Walsh to the pair. Specifically, if  $y_i$  is matched with  $y_j$  then the calculate  $\frac{1}{2}(y_i - y_j)^2$ . Note that one then obtains a sample of size  $\lfloor N/2 \rfloor$  that contains outcomes in  $\{-0.5, 0, 0.5\}$ . Moreover the expected value of each of these new data points is equal to  $VarY$ . Identify  $-0.5$ ,  $0$  and  $0.5$  with  $(1, 0)$ ,  $(0, 0)$  and  $(0, 1)$  so  $p_2 - p_1 = 2VarY$ . Now apply the unbiased test for matched pairs to investigate  $H_0 : p_2 - p_1 \geq 2\sigma_0^2$  against  $H_1 : p_2 - p_1 < 2\sigma_0^2$ .

Of course the above test requires knowledge of a bounded set that contains any possible outcome. In the next section we present a measure that does not require

such information.

## 4.5 Ordinal Dispersion

Consider the following alternative to  $VarY$  and  $\sigma(Y)$  as a measure of dispersion. The idea is to consider  $|Y_1 - Y_2|$  where  $Y_1$  and  $Y_2$  are independent copies of  $Y$ . The most intuitive is to search for the median of the distribution of  $|Y_1 - Y_2|$ . Confidence interval would then select where the majority of these absolute differences lie. In the literature one finds the alternative suggestion, to consider the first quartile. The first quartile is chosen in order to make the estimator based on the empirical sample maximally “robust” to outliers. The estimator based on the first quartile is called  $Q_n$  (see Rousseeuw and Croux, 1993). Confidence intervals can be derived from tests that build on randomly assigning data into pairs and then recording whether or not  $|Y_1 - Y_2|$  lies above or below some threshold.

To be more explicit consider  $H_0 : m_q(|Y_1 - Y_2|) \leq d$  where  $m_q$  is the  $q$  quantile of the distribution underlying  $|Y_1 - Y_2|$ . The test is constructed as follows. Randomly assign the data into pairs, dropping the unmatched element if there is an odd number of observations. Then, for each pair  $(y_i, y_j)$ , assign a success if  $|y_i - y_j| \leq d$  and a failure to this pair otherwise. Next record the probability of rejection when applying the randomized binomial test with level  $\theta\alpha$  to the null that the probability of a success lies above  $1 - q$ . Repeat the previous steps in this order infinitely often and recommend to reject the null hypothesis that  $m_q(|Y_1 - Y_2|)$  lies below  $d$  if the average rejection probability in this repetition lies above  $\theta$ . This is a nonrandomized test that has level  $\alpha$ .

## 5 Comment on Asymptotic Theory

In this document we have focussed on exact statistics. Often statisticians use tests that have certain properties for large samples, essentially when the sample size tends to infinity. The problem is that one has to be very careful about the type of limit one is considering. One type of limit is pointwise. Fix some test. Then for each data generating process one investigates the properties of the test as the sample size tends to infinity. The test is claimed useful if for instance the postulated size  $\alpha$  turns out to

be the approximate size provided the sample size is sufficiently large. In this case the test has pointwise asymptotically level  $\alpha$ . The problem with this standard approach is that the size of the sample necessary to achieve such a result may depend on the data generating process. One never knows if the sample size is sufficiently large when this depends on the underlying distribution. If "sufficiently large" does not depend on the underlying data generating process then one speaks of uniform asymptotically level  $\alpha$ . More specifically, one requires for each  $\delta > 0$  that there is a lower bound on the sample size  $N$  such that the level of the test is at most  $\alpha + \delta$  provide the sample size is at least  $N$ . This uniform property is rarely investigated. In fact, many standard tests are not uniformly asymptotically level  $\alpha$  such as the single sample t test as we illustrate below.

To be more concrete we investigate the t test applied to a single sample. In the correct terminology, this test is pointwise asymptotically level  $\alpha$  but not uniformly asymptotically level  $\alpha$ . The reason is as follows. Say one wishes to test whether the mean is above 0.5 when the data belongs to  $[0, 1]$ . It could be that the data yields outcome  $0.5 + \varepsilon$  with probability  $1/(0.5 + \varepsilon)$  and 0 otherwise. The expected value is equal to 0.5, thus the test should not reject the null with probability above  $\alpha$ . However, note that when  $\varepsilon$  is sufficiently small for given sample size  $N$  then with large probability the statistician will only observe the outcome  $0.5 + \varepsilon$ . The t test then recommends to reject the null. Completing this argument one sees that the size of the t test is 1, a result due to Lehmann and Loh (1990).

The two sample t test is known to not perform well even in large samples if there are no assumptions on the higher moments. Alternative tests have been suggested. However, again properties are derived by considering  $n_1$  and  $n_2$  sufficiently large for a given pair of distributions. One has not looked at performance for sufficiently large but given  $n_1$  and  $n_2$  across all feasible distributions.

## 6 Eyeball Statistics

Another comment often heard is that sometimes significance can be seen in which case one does not need a formal test. The answer to this is as follows. If it is so obvious then there should be a test that confirms the "intuitive significance". Just

to show how difficult it is to eyeball data we return to the example presented in the previous section, formulated a bit differently.

Assume that there are two random variables  $Y_1$  and  $Y_2$  such that  $\Pr(Y_1 = y_1) = 1$  and  $Y_2 \in \{0, y_2\}$  for  $0 < y_1 < y_2$ . We wish to test  $H_0 : EY_1 \geq EY_2$  based on  $n$  independent observations of each variable. Assume that one has only observed  $y_2$  as outcome of  $Y_2$ . We claim that there is no exact test with level  $\alpha$  that will reject the null hypothesis in favor of  $EY_1 < EY_2$  if  $\frac{y_1}{y_2} > \alpha^{1/n}$ . So for instance, if  $y_1 = 1$ ,  $\alpha = 0.05$  and  $n = 20$  then one cannot reject the null hypothesis if  $1 < y_2 < 1.16$ . The reason is simple. If  $EY_1 = EY_2$  then  $\Pr(Y_2 = y_2) = y_1/y_2$  and the probability of observing only  $\{Y_2 = y_2\}$  is equal to  $(y_1/y_2)^n$  which is above  $\alpha$  if  $\frac{y_1}{y_2} > \alpha^{1/n}$  where  $1/(0.05)^{1/20} = 1.1616$ . If  $y_2$  is too close to  $y_1$  then only observing  $Y = y_2$  will occur too likely and hence cannot be counted as evidence against the null.

## References

- [1] Berger, J. O. (2003), "Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?", *Statistical Science*, Vol. 18, No. 1, 1-12, see also 2002 mimeo Duke University.
- [2] Bickel, P., Godfrey, J., Neter, J. & Clayton, H. (1989). Hoeffding bounds for monetary unit sampling in auditing. *International Statistical Institute, Contributed Paper, Paris Meeting*.
- [3] Boschloo, R. D. (1970), "Raised Conditional Level of Significance for the  $2 \times 2$ -Table when Testing the Equality of Two Probabilities," *Statistica Neerlandica* 24, 1-35.
- [4] Brunner, E. and Munzel, U. (2000), The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, *Biometrical Journal* 42, 17-25.
- [5] 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*.

- [6] Fisher, R. A. (1935), "The Logic of Inductive Inference," *J. Roy. Stat. Soc.*, 98, 39–54.
- [7] Fishman, G. S. (1991). Confidence intervals for the mean in the bounded case. *Statistics and Probability Letters* **12**, 223–7.
- [8] Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994), "Fairness in Simple Bargaining Experiments," *Games Econ. Beh.* 6, 347–369.
- [9] Hoeffding, W. (1952). The Large-Sample Power of Tests Based on Permutations of Observations. *The Annals of Mathematical Statistics*, 23(2), 169–192.
- [10] Kendall, M. G. (1938) A New Measure of Rank Correlation, *Biometrika*, 30, 81-93.
- [11] Kruskal, W. H. (1958) Ordinal Measures of Association, *JASA*, 53, 814-861.
- [12] Lehmann, E. L. (1959), *Testing Statistical Hypotheses*. New York: Wiley.
- [13] Lehmann, E. L. and Loh, W.-Y. (1990), "Pointwise versus Uniform Robustness in some Large-Sample Tests and Confidence Intervals," *Scandinavian Journal of Statistics*, 17, 177–187.
- [14] Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*. New York: Springer.
- [15] Mann, H. B., and Whitney, D. R. (1947), "On a Test whether one of two random variables is stochastically larger than the other," *Annal. Math. Statistics*, 18, 50–60.
- [16] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** 153–157.
- [17] Motulsky, Harvey (1995) "Intuitive Biostatistics," Oxford: Oxford Univ. Press.
- [18] Pesarin, F. (2001) *Multivariate Permutation Tests with Applications in Biostatistics*, John Wiley and Sons.
- [19] Romano, J.P. (1989), Bootstrap and Randomization Tests of some Nonparametric Hypotheses, *The Annals of Statistics*, Vol. 17, No. 1, 141-159.

- [20] Romano, J. P. and Wolf, M. (2000), “Finite Sample Non-Parametric Inference and Large Sample Efficiency,” *Annals of Statistics* 28, 756–778.
- [21] Romano, J. P. & Wolf, M. (2002). Explicit nonparametric confidence intervals for the variance with guaranteed coverage. *Communication in Statistics - Theory and Methods* **31**, 1231–50.
- [22] Rousseeuw, P. T. and Croux, C. (1993) Alternatives to the Median Absolute Deviation, *JASA* **88**, 1273-1283.
- [23] Schlag, K. H. (2007a), Finite Sample Inference for the Mean of an Unknown Bounded Random Variable without Assumptions, Mimeo, <http://www.iue.it/Personal/Schlag/papers/singlesample.pdf>.
- [24] Schlag, K.H. (2007b), On Distribution-Free Bounds to Inference, Sequential Testing and Counterfactual Evidence when Comparing Means or Distributions, Mimeo, <http://www.iue.it/Personal/Schlag/papers/twosample.pdf>.
- [25] Schlag, K.H. (2007c), Testing Equality of Two Means without Assumptions - Solving the Nonparametric Behrens-Fisher Problem, Mimeo, <http://www.iue.it/Personal/Schlag/papers/2meaninfer.pdf>.
- [26] Schlag, K.H. (2008), *Exact Tests for Correlation and for the Slope in Simple Linear Regressions without Making Assumptions*, Universitat Pompeu Fabra working paper 1097.
- [27] Sefton, M. (1992), Incentives in simple bargaining games, *Journal of Economic Psychology* **13**, 263-276.
- [28] Spearman, C. (1904), “The Proof and Measurement of Association Between Two Things,” *American J. Psychol.*, 15, 72–101.
- [29] Suissa, S. and Shuster, J. J. (1984), “Are Uniformly Most Powerful Unbiased Tests Really Best?” *Amer. Stat.*, 38, 204–206.
- [30] Suissa, S. and Shuster, J. J. (1985), “Exact Unconditional Sample Sizes for the  $2 \times 2$  Binomial Trial,” *J. Roy. Stat. Soc. (A)*, 148, 317–327.

- [31] Suissa, S. and Shuster, J. J. (1991), “The 2 x 2 Matched-Pairs Trial: Exact Unconditional Design and Analysis,” *Biometrics* 47, 361–372.
- [32] Walsh, J. E. (1962), *Handbook of Nonparametric Statistics, Investigation of Randomness, Moments, Percentiles, and Distributions*, Princeton: D. van Nostrand Company Inc..
- [33] Wilcoxon, F. (1945), “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin* 1, 80–83.
- [34] Ziliak, S.T. and D. N. McCloskey (2008), *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*, University of Michigan Press 2008.

## 7 Some Comments on Siegel and Castellan (1988)

We would rather avoid commenting on the presentation of others, in particular if these references have been extremely valuable for researchers. However we need to mention some statements in this particular book as it has been very influential so we need to qualify where we differ in approach and wish to also point out an error.

Siegel and Castellan (1988) follow the practice of many statisticians and allow that the alternative hypothesis is not the complement of the null hypothesis. For instance consider Example 5.3b in Siegel and Castellan (1988) where they use the sign test. They are essentially considering  $H_0 : P_{Y_1} \equiv P_{Y_2}$  versus  $H_1 : EY_1 > EY_2$ . Rejecting identity of distributions implies that the means differ if one rules out by assumption all distributions in which the means are the same but higher moments differ. In this text we advocate that if one wishes to show that means differ then the null hypothesis has to contain the statement that they are equal without imposing that the entire distributions are identical. The statement of interest should be equivalent to the alternative hypothesis, the null hypothesis should be its complement.

In their discussion of the Wilcoxon test for matched pairs they again consider a situation in which the hypotheses do not cover all parameters. They consider  $H_0 : P_{Y_1} \equiv P_{Y_2}$  against  $H_1 : \Pr(Y_1 > Y_2) > 1/2$ .

The discussion of the permutation test for two independent samples in Siegel and Castellan (1988, Section 6.7) is incorrect. The test they are presenting is sometimes called the Fisher-Pitman test.<sup>11</sup> As any other permutation test it is an exact test of the null hypothesis that the two distributions are identical. However the text states that it is an (exact) test of the null hypothesis  $H_0 : EY_1 = EY_2$ . This is not true as to be seen by the following example. Assume that  $\Pr(Y_1 = 1/2) = 1$  and  $\Pr(Y_2 = 1/2 + \varepsilon) = 1 - \Pr(Y_2 = 0) = 1/(1 + 2\varepsilon)$ . Then  $EY_1 = EY_2$ . However if  $\varepsilon$  is sufficiently small then with very large probability only outcome  $1/2 + \varepsilon$  is observed of  $Y_2$ . The case in which only outcome  $1/2 + \varepsilon$  is observed of  $Y_2$  is the most extreme outcome when running the permutation, its p value is equal to  $1/\binom{n_1+n_2}{n_1}$ . However this outcome is very likely to occur if  $\varepsilon$  is very small. Thus we find that the null is wrongly rejected with very high probability when  $1/\binom{n_1+n_2}{n_1} < \alpha$ . The Fisher Pitman test is not made for testing equality of means, it is designed for testing identity of distributions.

## 8 Most Popular Mistakes

1. Accept the null hypothesis
2. Make inference based on comparing p values
3. Use estimates or graphs to make inference when there is no significant effect
4. Use the Wilcoxon or Mann Whitney test to compare means
5. Use the Spearman rank correlation test to test for correlation
6. Ignore assumptions underlying the t test
7. Think that large samples makes the t test applicable
8. Overlook that samples are not independent
9. Failure to document the details necessary to verify how the specific test has been implemented
10. Look at the data and then search for a test that rejects the null hypothesis

---

<sup>11</sup>The Fisher Pitman test is asymptotically equivalent to the two sample t test (Hoeffding, 1952).



## 9 Summary Table of Proposed Exact Nonrandomized Tests

Single sample	binary data	CI for success probability	use binomial test
"	more than 2 outcomes	CI for median or quantile	use binomial test
"	more than 2 outcomes	CI for median spread	adapt Schlag (2007c) using binomial test
"	bounded DGP*	CI for mean	Schlag (2007a)
"	bounded DGP*	CI for variance	Schlag (2007a)
Two independent samples	binary data	test for equality of means	Z test (pooled variance) if more balanced or Boshloo if more unbalanced
"	more than 2 outcomes	test for identity of distributions	permutation test such as Wilcoxon Mann Whitney
"	more than 2 outcomes	test for equality of medians	look at CIs for single sample - very crude
"	more than 2 outcomes	test for stochastic inequality	adapt Schlag (2007c) using McNemar
"	bounded DGP*	test for equality of means (CI)	Schlag (2007c)
"	bounded DGP*	test for equality of variances	Schlag (2007c)
Matched pairs	binary data	test for equality of means	Z test
"	binary data	test for correlation	combine Boschloo and Tocher
"	more than 2 outcomes	test for identity of distributions	permutation test such as Wilcoxon rank sum
"	more than 2 outcomes	test for stochastic inequality	Z test
"	more than 2 outcomes	test for association	adapt Schlag (2007c) using McNemar
"	bounded DGP*	test for equality of means (CI)	Schlag (2007c)
"	bounded DGP*	test for equality of variances	Schlag (2007c)
"	bounded DGP*	test for correlation	Schlag (2008)
"	bounded DGP*	CI for covariance	Schlag (2007a)

\* There is some known bounded set that will contain any outcome that can be generated.

## 10 List of Other Exact Tests

The following tests fall in the category of having no formulae for power comparisons, there being no software and the proofs being too difficult for me to verify.

- Romano and Wolf (2000), Diouf and Dufour (2006): mean of a single sample for a DGP with known bounds
- Romano and Wolf (2002): variance of single sample for a DGP with known bounds

The following two papers present the same test, this test turns out in all examples studied to be inferior to the test discussed in this paper

- Bickel et al (1989), Fishman (1991): mean of a single sample a DGP with known bounds