

Exact Nonparametric Tests for Comparing Means - A Personal Summary

Karl H. Schlag
*European University Institute*¹

December 14, 2006

¹Economics Department, European University Institute. Via della Piazzuola 43, 50133
Florence, Italy, schlag@eui.eu

Abstract

We present an overview of exact nonparametric tests for comparing means of two random variables based on two independent i.i.d. samples when outcomes are known to be contained in some given bounded interval.

1 The Setting

A statistical inference problem is *parametric* if the data generating process is known up to a finite dimensional parameter space. It is *nonparametric* if it is not parametric (Fraser, 1957). A statistical inference problem is *distribution-free* if there are no assumptions on the distribution underlying the data beyond what is known, e.g. often one has some information about the possible support. So the Fisher exact test for comparing means of two binomial distributions is uniformly most powerful for a parametric and distribution-free statistical inference problem. It is known that observations are only 0 or 1 so the restriction to binomial distributions is implied by what is known as is not an assumption on the distribution. The null and alternative hypothesis used to prove that a permutation test is most powerful (see below) make the associated statistical inference problem nonparametric but not distribution-free.

A test is *exact* if its properties are derived for finite samples and not only for a sufficiently large sample size. Often “properties” refers to the size of the test (or an upper bound on its size). This either means that there is a formula for the value of the size of the test or that there is a formula for an upper bound of the size of the test. Exact simply means that these formulas are correct for a given sample size.

Assume that there are two underlying random variables Y_1 and Y_2 with unknown means $E = EY_1$ and EY_2 and two separate samples $(y_1^k)_{k=1}^n$ and $(y_2^k)_{k=1}^n$ of independent realizations of Y_1 and Y_2 respectively. There is an interval $[\gamma, \omega]$ that is known to have the property that $Y_1, Y_2 \in [\gamma, \omega]$. So the joint distribution P is an element of $\Delta([\gamma, \omega]^2)$. Let P_1 and P_2 denote the respective marginals.

The data generating process can be identified with P . As we allow for any $P \in \Delta[\gamma, \omega]^2$ our setting is nonparametric.

2 One-Sided Test

Consider the objective to choose an exact test of the null hypothesis $H_0 : EY_1 \leq EY_2$ (or $H_0 : EY_1 = EY_2$) against the alternative hypothesis $H_1 : EY_1 > EY_2$.

2.1 UMP

Note that there is no uniformly most powerful test for our hypotheses. Nor is there a uniformly most powerful unbiased test. However there are most powerful tests and

other exact tests.

2.2 T Test

The two sample t-test is **not** an exact test in our setting as it assumes that the underlying random variables are normally distributed. However neither Y_1 nor Y_2 can be normally distributed as the ranges of Y_1 and Y_2 are bounded. Without knowing more about the random variables one cannot argue that the t-test can be applied if the sample is sufficiently large. It is only *pointwise asymptotically level α* , i.e. for given P and given ε there is some $N(P, \varepsilon)$ such that if $n > N(P, \varepsilon)$ then the size of the test is within ε of the size α of the underlying t-test. However, the notion of sufficiently large depends on P and it turns out that $\{N(P, \varepsilon), P \in \Delta[\gamma, \omega]^2\}$ is an unbounded set. The two sample t-test is not *uniformly asymptotically level α* . In fact, the size of the test for any given sample size n is equal to 1. For the argument one can easily adapt the constructive proof of Lehmann and Loh (1990) who show this statement for testing the mean of a single random variable. Hence the existence of a bound $N(P, \varepsilon)$ is useless. In simple words, it is the difference between pointwise and uniform convergence that makes the t-test non applicable even if the sample is very large.

The example that can be used to prove the above statement is a simple variation of the one in Lehmann and Loh (1990). Assume that $Y_1 = 1/2$ almost surely and that Y_2 has two mass points, one at ξ and the other at 0 where ξ is strictly larger but very close to $1/2$ such that $EY_2 = 1/2$. Notice that the closer ξ is to $1/2$, the more mass Y_2 puts on ξ as $EY_2 = 1/2$. Provided ξ is sufficiently close to $1/2$, with large probability each of the n realizations of Y_2 are equal to ξ . If only $Y_2 = \xi$ is observed then since $\xi > 1/2$ and since the sample variances are both 0 it follows that the null is rejected under the t-test even though $EY_1 = EY_2$. Since the event that only ξ is observed can be made arbitrarily large by choosing ξ very close to $1/2$, the probability of wrongly rejecting the null can be made arbitrarily close to 1. Thus the size of the two sample t test is equal to 1. Notice that this example can be easily adjusted to apply only to continuous distributions without changing the conclusion. Notice also that variance of any random variable in our setting is bounded. The phenomenon of this example is that we cannot rule out that most weight is put on one observation very close to $1/2$.

The obvious reaction to the above example is to suggest to first test for normality of

the errors. Such tests would be able to discover the above counter example. However this does not mean that they would make the t-test valid given the current standing of research.

The problem of tests for normality (e.g. Shapiro-Wilk) is that the null hypothesis asserts that they are normally distributed. To claim that the errors are normally distributed is a claim about not being able to reject the null. It can be very likely that normality cannot be rejected even though errors are not normally distributed. This affects the size of the combined test, first of normality and then of $EY_1 \leq EY_2$. In fact, without results on the probability of rejecting $EY_1 \leq EY_2$ with the t-test when errors are not normally distributed one has to simply work with the upper bound that the t-test always rejects $EY_1 \leq EY_2$ whenever errors are not normal. But then this means that the size of the combined test is dominated by the probability of wrongly not being able to reject that errors are normal. Now note that the maximal probability of wrongly not rejecting that errors are normal will be approximately equal to or larger than the probability of correctly accepting the null. This is because there are distributions that are not normal that are almost normal. So if the size of the test for normality is equal to 10% then the probability of thinking that errors are normally distributed even if they are not can be 90% or higher. In this case the best upper bound given current research on the size of the combined test is 90%.

Note that the bound above is not tight. When errors are almost normal then the size of the t-test will be close to its size when errors are normal. This will be true (for appropriate distance function) in the current framework as we assume that payoffs belong to $[0, 1]$. In order to assess the size of the combined test one would have to derive the size of this combined test, in particular obtain finite sample solutions to the size of the t-test when errors are almost normal. As this has not been done, it is an open question whether tests for normality can be used to reinstate the t-test when payoffs are bounded.

2.3 Permutation Tests

Permutation tests are exact tests in terms of size. A permutation test is a most powerful unbiased test of testing $H_0 : EY_1 = EY_2$ against the simple alternative hypothesis $H_1 : P = \bar{P}$ where \bar{P} is given with $E_{\bar{P}}Y_1 > E_{\bar{P}}Y_2$. Accordingly all permutations of the data are considered in which the data is randomly assigned to one of the two actions and the null is rejected if the true sample is within the $\alpha\%$ of permutations under

which the alternative is least likely under \bar{P} . This is the theory.

In practice, typically only the test statistic itself is specified without mentioning whether it is most powerful relative to some alternative hypothesis. The test remains exact in terms of size. A popular test statistic is to take the difference between the two sample means. I have not come across any distribution with bounded support that can be specified for the alternative hypothesis that makes even this test most powerful.

Note that the Wilcoxon rank sum test (or Mann Whitney U test) is a also permutation test which can be used to compare means and to compare distributions. It can be useful when data is given in form of ranks. However I do not see the sense to first transform the data into ranks and then to apply this test as the transformation means to throw away information. The reason is that I cannot imagine that it is most powerful against any specific distribution.

The literature on permutation tests typically ignores the possibility to use permutation tests to create most powerful tests as they are interested in distribution-free tests or at least in alternatives that are composite.

2.4 Aside on Methodology

Before we proceed with our presentation of tests we present some material on methodology of selecting tests.

2.4.1 Multiple Testing on same Data

It is not good practice to evaluate multiple tests on the same data without correcting for this. The reason is that this typically increases the type I error underlying the final evaluation. To illustrate, assume that the null hypothesis could not be rejected with a first test but that it can be rejected based on a second test where both tests are unbiased with the same size α . Consider an underlying data generating process P with $EY_0 = EY_1$. As the first test is unbiased, the probability of wrongly rejecting the null using this test is equal to α . Consequently, the probability of wrongly rejecting the null hypothesis under P is equal to $\alpha + \varphi(P)$ where $\varphi(P)$ is the probability that the first test does not reject the null and that the second does reject the null. If the two tests are not identical then $\varphi(P) > 0$ and this two step procedure leads to a rejection of the null hypothesis with probability strictly larger than α . More generally, if neither test is more powerful than the other then the sequential procedure of testing until

obtaining a rejection strictly increases the type I error. Of course one could mention how many other tests have been tried, however it is difficult to derive exactly how the type I error increases, the upper bound on the type I error when checking two tests is 2α .

As one can hardly verify how many others tests were tried it is useful to have an understanding of which tests should be applied to which setting. This is best done by comparing tests. When uniformly most powerful tests exist then there is no danger for worrying about multiple testing of same data. One can say that there is a best test. However in our setting a UMP test does not exist.

Permutation tests are most powerful given the assumptions on how the alternative hypothesis has been limited. As different alternative hypotheses satisfying $EY_1 > EY_2$ yield different tests, each is most powerful for the specific alternative. Again it can be difficult to justify which alternative is most plausible and it is not good practice to try several alternatives.

2.4.2 Choice of a Test: Ex-Ante vs Ex-Post

When there is no uniformly most powerful test then it is important to first select a test and then to gather data. One problem of looking for a test after data is gathered is illustrated above as it is hard to check how many tests were run but not reported. A further problem is more fundamental and is due to the fact that size is an ex-ante concept which means the test may not be conditioned on the observed data.

Often if not typically one can design a test given the data gathered that will reject the null given the data. We illustrate. Consider the test to reject the null if the difference between the two sample means is equal to $\gamma \neq 0$ where γ is given. So reject if $\bar{Y}_1 - \bar{Y}_2 = \gamma$. Then the size of this test will be small if the sample size n is sufficiently large. Note that it is important that γ is first chosen before the data is gathered. Otherwise γ could be set equal to the difference of the observed averages and the null would be rejected. So setting γ after data is observed yields size equal to 1.

An objection to the above example could be that the test suggested is not plausible. One would expect that if it is rejected for $\bar{Y}_1 - \bar{Y}_2 = \gamma$ with $\gamma > 0$ then it is also rejected for $\bar{Y}_1 - \bar{Y}_2 \geq \gamma$. However this is only a heuristic argument. The real problem is that the above test is not credible as it is not most powerful nor does it satisfy any other optimality principles.

2.4.3 The Value of the P Value

It is standard practice to record p values when testing data. In particular one cites whether the null can be rejected at 1% level or at the 5% level. In either case the null is rejected. P levels above 5% typically lead to a statement of not being able to reject the null. Consider this practice. What is the real size of the test? In other words, how often will the null be wrongly rejected. Is the p value really the smallest level of significance leading to rejection of the null? Well if the null is rejected as long as the p value is below 5% then the size is 5%. So regardless of whether or not the data has a p value below 1%, if the statistician would have also rejected the null if the p value was above 1% but below 5% then the fact that the data has a p value below 1% is not of importance. The test has size 5%. The specific p value is useless for the given data, it only matters if it is below or above 5%. In particular the p value is **not** the smallest level of significance leading to rejection of the null.

The issue behind the above phenomenon is the same as we discussed in the previous section. The test is formulated, by identifying whether significance is at 1% or 5% level, after the data has been gathered. Size is an ex-ante concept that is a function of the test that is derived before the data is observed. Looking at the p value is an ex-post operation.

If one can credibly commit to reject the null if and only if the p value is below 1% then a p value below 1% would have the information that the size of the test is equal to 1%. However, standard practice calls to reject if and only if the p value is below 5%, hence observing a p value of 1% or lower has the only information that the p value is below 5%. It does not matter how much lower it is.

Of course the p value has a value for the general objective. When discovering that it is very low then one can be more ambitious when one gathers similar data next time, e.g. one can choose to gather less data.

2.4.4 The Role of the Null Hypothesis

Often the null hypothesis cannot be rejected in which case we would like to argue that there is no treatment effect. This is where type II error comes in handy. If a test is most powerful for some distribution satisfying the alternative then we know that the null could not be rejected more often under this alternative. However it would be even nicer to know explicitly what the type II error is. Moreover, how to argue that the distribution underlying the alternative is well chosen given that they type II error

depends on this distribution? One would wish to find some way of making inferences without making specific unverifiable assumptions on the distributions to avoid these issues. We show how to do this now.

2.4.5 Comparing Tests

Even if there is no uniformly most powerful test one can still compare tests. This comparison will not build on making specific assumptions on the distributions underlying either the null or the alternative. The idea is to consider a worst case analysis centered around the parameters of interest which here are the means.

Specifically, for any given μ_1 and μ_2 with $\mu_1 > \mu_2$ one can consider all distributions P with $EY_1 = \mu_1$ and $EY_2 = \mu_2$ and calculate the maximal type II error of wrongly not rejecting the null within this class. In other words, one can calculate the minimal power within this set. This leads to a minimum power function that depends on μ_1 and μ_2 . Or for any given $d > 0$ one can derive the maximal type II error among all distributions P with $EY_1 \geq EY_2 + d$. This leads to a minimal power function that depends on d . Note that $\{P : EY_2 < EY_1 < EY_2 + d\}$ can be interpreted as an *indifference zone*, a set of underlying parameters where the decision maker does not care whether or not the null hypothesis is rejected. It is necessary to consider $d > 0$ as the type II error for $d = 0$ will be equal to $1 - \alpha$. For $d > 0$ one typically can make the type II error small provided the sample is sufficiently large. The smaller d the larger the sample must be.

Using these minimal power functions one can evaluate the event of not rejecting the null as a function of either (μ_1, μ_2) or as a function of d . One can then compare tests by comparing minimal power functions. We show in a different paper that there is a best test in terms of such minimal power functions. In other words, there is a “uniformly” most powerful test when inference is only based on the parameters of the distribution satisfying the alternative hypothesis. Such a test is called *parameter most powerful*.

A general approach followed in the literature (see Lehmann and Romano, 2005) is to limit the set of distributions belonging to the alternative hypothesis. So one sets $H_1 : P \in \Omega_K$ where $\Omega_K \subset \{P : EY_1 > EY_2\}$. In this case, $\{P : EY_1 > EY_2\} \setminus \Omega_K$ is the indifference zone. The alternative hypothesis only contains those distributions differing so widely from those postulated by the null hypothesis that false acceptance of the null is a serious error (Lehmann and Romano, 2005, p. 320). In

this sense it is natural to formulate the indifference zone in terms of the underlying means. For instance, one can fix $d > 0$ and set $H_1 : EY_1 \geq EY_2 + d$. In this case $\{P : EY_2 < EY_1 < EY_2 + d\}$ is the indifference zone. A *maximin test* is a test that maximizes over all tests of size α the minimum power over all $P \in \Omega_K$. In particular, a parameter most powerful test is a maximin test when Ω_K can be described in terms of μ only.

2.4.6 Unbiasedness

Unbiasedness is a property that is intuitive and which can add enough structure to find a best test. For instance, there is no UMP test for comparing the means of two Bernoulli distributed random variables but there is a UMP unbiased test for that setting, called the Fisher exact test. However, lack of unbiasedness should not be a reason to reject a test. Unbiasedness means that the alternative should never be rejected more likely when the alternative is true than when the null is true. This sounds intuitive. However we are comparing very small probabilities in this statement and there is no decision theoretic basis for doing this. The reason why unbiasedness fails is typically due to distributions belonging to the alternative that are very close to the null. In our setting we are worried about rejecting the null with probability smaller than α among the alternatives but this typically means that $EY_1 > EY_2$ but $EY_1 \approx EY_2$. Above we however argued that it is plausible not to worry too much about alternatives that are close to the null in the sense that $EY_2 < EY_1 < EY_2 + d$ for some small d .

We now return to our presentation of exact tests.

2.5 Test Based on Hoeffding Bounds

In the following I construct a nonparametric test in which any distribution belongs either to the null or to the alternative hypothesis. Lower bounds on minimal power functions are easily derived. The test is not unbiased.

I use the Hoeffding bounds, similar to what Bickel et al. (1989) did for testing the mean of a single sample. Following a corollary by Hoeffding (1963, eq. 2.7), $\Pr((\bar{y}_1 - \bar{y}_2) - (EY_1 - EY_2) \geq t) \leq e^{-nt^2(\omega-\gamma)^2}$ holds for $t > 0$. So if t_α solves $e^{-nt_\alpha^2(\omega-\gamma)^2} = \alpha$ then the rule to reject the null when $\bar{y}_1 - \bar{y}_2 \geq t_\alpha$. The nice thing about this approach is that one can calculate an upper bound on the type II error. If $EY_1 = EY_2 + d$ then $\Pr((\bar{y}_1 - \bar{y}_2) \leq t_\alpha) = \Pr((\bar{y}_1 - \bar{y}_2) - d \leq t_\alpha - d) \leq e^{-n(d-t_\alpha)^2(\omega-\gamma)^2}$ for

$d > t_\alpha$. For example, if $n = 30$, $[\gamma, \omega] = [0, 1]$ and $\alpha = 0.05$ then one easily verifies that $t_\alpha = \sqrt{\ln 20}/\sqrt{30} \approx 0.316$ and that the type II error is bounded above by 0.2 when $d \geq 0.548$.

2.6 Using Tests Derived for a Single Sample

The above are the only deterministic exact nonparametric one-sided tests I know of that were explicitly constructed for the two sample problem. There are exact tests for the one sample problem (see Schlag, 2006) that of course can also be applied to our setting. One can redefine $Y = Y_1 - Y_2$ and test the null that $EY \leq 0$ against the alternative that $EY > 0$ where it is known that $Y \in [-1, 1]$.

2.7 Parameter Most Powerful Tests

In a separate paper (Schlag, 2006) I have constructed a randomized exact unbiased test that I call *parameter most powerful* as it minimizes the type II error whenever the set of alternatives can be described in terms of the underlying means only. It maximizes the minimal power function that only depends on μ_1 and μ_2 . In a first step the data is randomly transformed into binary data. In the second step one then applies the Fisher exact test which is uniformly most powerful unbiased test for the binary case. My test is randomized as the transformation of the data is random and hence the recommendation of the test is typically randomized: reject the null hypothesis with $x\%$ where typically $0 < x < 100$.

The property of being parameter most powerful comes at the cost of being randomized. The advantage is that it can be used to evaluate the relative efficiency of other deterministic tests..

References

- [1] Bickel, P., J. Godfrey, J. Neter and H. Clayton (1989), "Hoeffding Bounds for Monetary Unit Sampling in Auditing," *Contr. Pap. I.S.I. Meeting*, Paris.
- [2] Fraser, D.A.S. (1957), *Nonparametric Methods in Statistics*, New York: John Wiley and Sons.

- [3] Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *J. Amer. Stat. Assoc.* **38(301)**, 13-30.
- [4] Lehmann, E.L. and J.P. Romano (2005), *Testing Statistical Hypotheses*, 3rd edition, Springer.
- [5] Schlag, K.H. (2006), *Nonparametric Minimax Risk Estimates and Most Powerful Tests for Means*, European University Institute, Mimeo.