



INTERDISCIPLINARY CENTER
FOR ECONOMIC SCIENCE
GEORGE MASON UNIVERSITY

Adaptive Procedures for Nonparametric Tests: Seven Decades of Advances

Li Hao and Daniel Houser

July 2010

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4850 Fax: +1-703-993-4851
ICES Website: www.ices-gmu.org
ICES RePEc Archive Online at: <http://edirc.repec.org/data/icgmuus.html>

Adaptive Procedures for Nonparametric Tests: Seven Decades of Advances

Li Hao and Daniel Houser*

July 1, 2010

Abstract

Wilcoxon-Mann-Whitney and Jonckheere tests have dominated nonparametric analyses in behavioral sciences for the past seven decades. Their widespread use masks the fact that there exist simple "adaptive" procedures which use data-dependent statistical decision rules to select an optimal nonparametric test. This paper discusses key adaptive approaches for testing differences in locations in two- or k -sample environments. As sample sizes grow arbitrarily large, adaptive procedures generally have greater power than the parametric t -test. Moreover, we offer evidence from a Monte-Carlo analysis showing that adaptive procedures often perform substantially better than t -tests, and not worse on average, even with moderately-sized samples of between 40 and 80 total observations. We illustrate the adaptive approach using data from Gneezy and Smorodinsky (2006), and offer a Stata package to any interested in taking advantage of adaptive techniques in their own research.

*Interdisciplinary Center for Economic Science, George Mason University. 4400 University Drive, MSN 1B2, Fairfax, VA 22044. Li Hao: lhao@gmu.edu; 703 993 4858. Daniel Houser: dhouser@gmu.edu; 703 993 4856.

1 Introduction

The past seven decades have witnessed important developments in distribution-free (nonparametric) tests for differences in locations between samples. As early as 1936, Hotelling and Pabst (1936) recognized that the ranks of data could be used when sample sizes are small to avoid assuming normality, a distribution increasingly recognized as more an exception than rule¹. In the mid 1940s, the Wilcoxon-Mann-Whitney test (WMW test) was introduced. Used for testing for differences in medians between two otherwise identical samples, it was developed first by Wilcoxon (1945) and then extended by Mann and Whitney (1947). Both robust and simple to implement, it has gained exceptional popularity among experimental scientists generally, and experimental economists in particular. (For example, in 2009 almost half of papers appearing in *Experimental Economics* used the WMW test.) Next, the Jonckheere-Terpstra test (often referred to as the Jonckheere test), an extension of the WMW test for detecting location differences among three or more samples, was independently developed by Terpstra (1952) and Jonckheere (1954a). The use of the Jonckheere test in experimental economics dates to Vernon Smith, 1964.

Despite their dominance, it turns out that there exist simple "adaptive" rank tests² that can discover differences between distributions more easily than either WMW or Jonckheere tests. These adaptive nonparametric procedures display significant improvements in power over the parametric t -test with samples of large or even moderate size. The purposes of this paper are to review the development of these procedures, offer Monte-Carlo evidence that details their advantages, and to provide practical advice regarding their use. Our Stata software is available to anyone interested in using these procedures.

We organize our discussion chronologically, marking the major developments in this area. We begin with Hájek and Šidák (1967), who provided the key to improving WMW tests by demonstrating that the asymptotically most powerful rank test depends on the data's underlying distribution. In particular, the WMW test is the most powerful rank test only when the data follow a logistic distribution while, for example, the median test (see, e.g., Siegel and Castellan 1988) is the most powerful rank test when the underlying distribution is Laplace (or "double exponential"). Section 2 develops this result and reviews the family of asymptotically most powerful rank tests.

¹Milton Friedman (1937) was among the first to develop testing procedures using ranks, a procedure for analysis of variance in his case.

²A test is called a rank test if the test statistic is a function of the ranks of data, such as the WMW test and Jonckheere test.

Because one typically does not know the underlying distribution of the data, it is not usually possible to employ the most powerful rank test. Nevertheless, Gastwirth (1965) showed that modifying standard rank tests by appropriately re-weighting the data could lead to substantial improvements in the ability to detect differences in locations. Section 3 describes Gastwirth's (1965) insight, and demonstrates that significant improvements in a test's power require only that one knows certain features of the data's underlying distribution.

Gastwirth's (1965) simple modified tests paved the way for adaptive rank tests. In an important paper, Hogg, Fisher and Randles (1975, HFR hereafter) proposed a simple and effective adaptive procedure that uses the data to select the most appropriate rank test from a small set of alternatives. Although the data are used "twice", both to choose the test and perform the test, we discuss in Section 4 that HFR's adaptive procedure is "honest" in the sense that the significance level of the second-stage test is preserved. The key advantages of HFR's procedure are (i) ease of implementation and (ii) greater power on average than WMW tests. We describe in Section 5 that HFR's two-sample procedure has been extended to k -sample environments (e.g., Bünning 1999; Bünning 2009).

Section 6 reports a new Monte Carlo analysis designed to investigate HFR's performance under various sample sizes, and to optimize certain features of the HFR algorithm. We confirm prior results, including that the power of adaptive procedures is substantially greater than that of the WMW and t -tests under certain circumstances, and nearly as great in other cases. We also highlight our finding that adaptive procedures display improved power, in relation to the parametric t -test, not only asymptotically but also with samples of moderate size (we consider $20 \leq n, m \leq 40$). We show that this improvement becomes larger as total sample size increases.

In Section 7 we illustrate the adaptive procedure using data reported by Gneezy and Smorodinsky (2006). Section 8 offers advice for practitioners, including comments on the analysis of binary data. The final section summarizes.

2 Asymptotically Most Powerful Rank Tests

Hájek and Šidák (1967) shows that for every distribution function there corresponds an asymptotically most powerful rank test. This section illustrates this using a few key examples.

2.1 Tests of Location

Let X_1, \dots, X_n and Y_1, \dots, Y_m be random samples from a "control" population with absolutely continuous cumulative density function $F(t)$ and "treatment" population with density $F(t - \theta)$, respectively. We wish to test the hypothesis that there is no treatment effect: $H_0 : \theta = 0$ versus the alternative $H_1 : \theta \neq 0$ (or $\theta > 0$).

Two important assumptions are thus implied. First, the two populations can only differ in locations, so the distributions have equal variances under both the null and alternative hypothesis. Therefore, it is necessary first to test for differences in variances when applying WMW or related tests discussed in this paper. Second, X 's and Y 's are assumed to be mutually independent. Hence, it is not appropriate to apply the tests discussed below directly to data that include repeat observations on the same subject³.

2.2 Test Statistics

Let $f(\cdot)$ be the probability density function of $F(\cdot)$. Let R_i denote the rank of the observation Y_i ($i = 1, \dots, m$) in the combined sample of $n + m = N$ observations, $1 \leq R_i \leq N$. Hájek and Šidák (1967) shows that, in general, the asymptotically most powerful rank test statistic S directly depends on the inverse c.d.f. F^{-1} :

$$S = \sum_{i=1}^m a(R_i) \tag{2.1}$$

$$a(R_i) = -\frac{f'(F^{-1}(\frac{R_i}{N+1}))}{f(F^{-1}(\frac{R_i}{N+1}))} \tag{2.2}$$

where $\frac{R_i}{N+1}$ is Y_i 's rank normalized in the combined sample, $\frac{R_i}{N+1} \in (0, 1)$. The function $a(R_i)$ is a "scoring function", similar to a weighting function, in that it assigns a score to observation Y_i according to Y_i 's rank in the combined sample. The intuition is that, when N approaches infinity, $F^{-1}(\frac{R_i}{N+1})$ uncovers the corresponding observation using its rank R_i and the inverse c.d.f. of the data. Hence, the scoring function maximizes the information in the ranks.

Therefore, for any distribution of interest, the most powerful rank test can be obtained by equations (2.1) and (2.2). Asymptotically, as m and n approach infinity,

³In the conclusion we mention tests appropriate for repeat observation environments.

the standardized variable $\frac{S-E(S)}{\sqrt{Var(S)}}$ is standard normal (see e.g., Hájek et al. 1999, p. 98 and p. 190). We illustrate with three examples⁴.

2.2.1 Example 1: Normal-Score Test

Consider the standard normal distribution,

$$f_{nor}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

From equations (2.1) and (2.2), we obtain the most powerful rank test for normal distributions, also known as the normal score test (van der Waerden 1953),

$$S_{nor} = \sum_{i=1}^m \Phi^{-1}\left(\frac{R_i}{N+1}\right) \quad (2.3)$$

Here, Φ is the c.d.f. of the standard normal distribution.

The sampling distribution of S_{nor} is symmetric with mean equal to $E(S_{nor}) = 0$, and variance equal to $Var(S_{nor}) = \frac{mn}{N(N-1)} \sum_{i=1}^N [\Phi^{-1}(\frac{i}{N+1})]^2$.

2.2.2 Example 2: Wilcoxon-Mann-Whitney Test

For the logistic distribution,

$$f_{log}(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

one can derive that the asymptotically most powerful rank test is

$$S_{log} = \frac{2m}{N+1} \sum_{i=1}^m R_i - m$$

Note that this is a linear transformation of the WMW test statistic,

$$S_{WMW} = \sum_{i=1}^m R_i \quad (2.4)$$

⁴See Hájek et al. (1999, p. 15) for derivations of all three examples.

It follows that the WMW test is asymptotically optimal when data follow the logistic distribution. The sampling distribution of S_{WMW} is symmetric with mean equal to $E(S_{WMW}) = \frac{1}{2}m(N+1)$, and variance equal to $Var(S_{WMW}) = \frac{1}{12}mn(N+1)$.

2.2.3 Example 3: Median Test

For the Laplace distribution,

$$f_{lap}(x) = \frac{1}{2}e^{-|x|} \quad (2.5)$$

one can obtain the asymptotically most powerful rank test,

$$S_{lap} = \sum_{i=1}^m \text{sign}\left[R_i - \frac{N+1}{2}\right]$$

where $\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$. This test is equivalent to the median test that

counts the number of Y_i 's above the median of the combined sample, and increases by $\frac{1}{2}$ if the median belongs to Y_i 's (see e.g., Westenberg 1948). Formally:

$$S_{median} = \sum_{i=1}^m \frac{1}{2} \left[\text{sign}\left(R_i - \frac{N+1}{2}\right) + 1 \right] = \frac{1}{2}S_{lap} + \frac{m}{2} \quad (2.6)$$

Hence, the median test is asymptotically optimal for Laplace distributions. It can be shown that the sampling distribution of S_{median} is symmetric with mean equal to $E(S_{median}) = \frac{1}{2}m$, and variance equal to $Var(S_{median}) = \frac{mn}{4(N-1)}$ if N is even, $\frac{mn}{4N}$ if N is odd.

2.3 Asymptotic Relative Efficiency

To compare large-sample properties among tests it is natural to use the notion of Asymptotic Relative Efficiency (ARE) developed by Pitman (1949). This notion is quite intuitive. For two (consistent) test statistics A and B of an hypothesis H_0 , the ARE is the reciprocal of the ratio of sample sizes required to obtain identical power against the same alternative H_1 , taking the limit as the sample size N tends to infinity and as H_1 tends to H_0 . Clearly, the ARE of tests A to B is positive, $ARE_{A,B} \in (0, +\infty)$. If $ARE_{A,B} \in (0, 1)$, A is less efficient than B. If $ARE_{A,B} = 1$, A is as efficient as B. If $ARE_{A,B} \in (1, +\infty)$, A is more efficient than B.

The widely used parametric t -test is a natural benchmark for investigating asymptotic properties of various nonparametric tests. Pitman (1949) computed the ARE of the WMW test relative to the t -test as

$$ARE_{w,t} = 12\sigma^2 \left[\int f^2(x) dx \right]^2$$

where σ is the standard deviation of the underlying distribution function $f(x)$.

When data are normally distributed, the ARE of the WMW to the t -test is as high as 0.955. When data are uniformly distributed, the ARE is 1 and thus the two tests have equivalent asymptotic efficiency. Hodges and Lehmann (1956) proved that for all continuous distributions (that satisfy certain regularity conditions), the ARE of the WMW to t -test is no less than 0.864, and the lower bound is attained with parabolic density.

Regarding the ARE of the normal score test to the t -test, it is well known that the two tests are equivalent with normally distributed data. However, when data are non-normally distributed, Chernoff and Savage (1958) showed that the ARE is greater than or equal to 1 for all continuous distributions. This underlies Hodges and Lehmann's conclusion (1961, p. 308) that the normal score test is always preferred to the t -test, with samples of moderate size.

Finally, the median test has relatively low efficiency. For example, when data are normally distributed, the ARE of the median test to t -test is only 0.637 (e.g., Mood 1954). The median tests's low efficiency has led some to suggest that it is of little practical value (e.g., Freidlin and Gastwirth, 2000).

3 Gastwirth's (1965) Modified Rank Tests

As noted above, while theoretically elegant, accurately estimating the scoring function $-f'(F^{-1}(u))/f(F^{-1}(u))$ is typically either impossible or, in light of data limitations, highly impractical. Fortunately, Gastwirth (1965) proposed a much simpler approach to increasing efficiency by modifying the WMW test according to relatively easy-to-discover characteristics of the data's underlying distribution. This section reviews Gastwirth's (1965) test and discusses its asymptotic properties.

3.1 Less is More

It was known by the mid 1950's that the efficiency of rank tests could be improved by "throwing away" part of the data (see, e.g., Foster and Stuart 1954; Cox and Stuart 1955). While potentially counter-intuitive, the main idea is to focus the test on parts of the distribution where differences in locations are most likely revealed. For example, to detect a location shift between two uniform distributions, one would want to consider data near the hypothesized common boundaries.

Following the earlier insights, Gastwirth's (1965) suggestion is to modify the WMW test by including data only in the top p and the bottom r fractions of the combined sample ($0 < p, r < 1$), where the optimal values of p and r depend on the underlying distribution. Gastwirth (1965) showed that, in general, appropriately modified rank tests are asymptotically more efficient than the WMW test.

3.2 Gastwirth's (1965) Test Statistic

In light of its importance, it is worthwhile to derive Gastwirth's (1965) statistic formally. To do so, let $Z_1 < \dots < Z_N$ be the order statistics of the combined sample, where $N = n + m$. Let T_p and B_r denote the total scores of the top t and bottom b fractions, respectively. The statistic $T_p - B_r$ starts at the median of Z 's, and scores with increasing positive integers for bigger ranks, and symmetric negative integers to smaller ranks. Formally,

If N is odd, let $K = (N + 1)/2$:

$$B_r = \sum_{i=1}^K (K - i)\delta_i \quad \text{and} \quad T_p = \sum_{i=K}^N (i - K)\delta_i \quad (3.1)$$

If N is even, let $K = N/2$:

$$B_r = \sum_{i=1}^K (K - i + \frac{1}{2})\delta_i \quad \text{and} \quad T_p = \sum_{i=K+1}^N (i - K - \frac{1}{2})\delta_i \quad (3.2)$$

where $\delta_i = \left\{ \begin{array}{ll} 1 & \text{if } Z_i \text{ belongs to } Y_i\text{'s,} \\ 0 & \text{otherwise} \end{array} \right\}$, indicating whether the observation is from the second sample (the treatment population). Therefore, the statistic $T_p - B_r$ with $p = r = 50\%$ is equivalent to the WMW test⁵.

Once the values for p and r are chosen, data that are outside of the top p and bottom r fractions are given zero weights. Formally,

⁵Interestingly, while T-B detects differences in locations, T+B detects differences in scales. See Gastwirth (1965) for details.

If N is odd:

$$B_r = \sum_{i=1}^R (R - i + 1)\delta_i \quad \text{and} \quad T_p = \sum_{i=N-P+1}^N (i - (N - P))\delta_i \quad (3.3)$$

If N is even:

$$B_r = \sum_{i=1}^R (R - i + \frac{1}{2})\delta_i \quad \text{and} \quad T_p = \sum_{i=N-P+1}^N (i - (N - P) - \frac{1}{2})\delta_i \quad (3.4)$$

where $R = [Nr] + 1$ and $P = [Np] + 1$. $[H]$ is the nearest integer of H .

We derive the mean and variance of the test statistic $T_p - B_r$ from Gastwirth (1965, pp. 1129-1131),

If N is even,

$$E(T_p - B_r) = \frac{m}{2N}(P^2 - R^2)$$

$$Var(T_p - B_r) = \frac{mn}{2N^2(N-1)}[NP(4P^2 - 1) + NR(4R^2 - 1) - 3(P^2 + R^2)^2]$$

If N is odd,

$$E(T_p - B_r) = \frac{m}{2N}[P(P+1) + R(R+1)]$$

$$Var(T_p - B_r) = \frac{mn}{2N^2(N-1)}\{2NP(P+1)(2P+1) + 2NR(R+1)(2R+1) - 3[P(P+1) + R(R+1)]^2\}$$

As noted, the optimal values of p and r depend on the nature of underlying distributions, which we discuss in detail in Section 3.3.

3.3 Asymptotic Relative Efficiency

To develop the ARE of Gastwirth's modified test, we first discuss its asymptotic distribution. Gastwirth (1965, p. 1136) proved that, under regularity conditions, the percentile modified rank test statistic $T_p - B_r$ is asymptotically normally distributed under both H_0 and H_1 , with mean and variance as given above⁶.

We derive the ARE of Gastwirth's test $T_p - B_r$ in relation to the t -test, shown below, from Gastwirth (1965, p. 1136),

⁶The key regularity condition is that as the combined sample size N approaches infinity, neither p nor r may tend to 0 or 1.

$$ARE_{G,t} = \frac{12\sigma^2 \left[\int_{-\infty}^l f^2(x) dx + \int_u^{\infty} f^2(x) dx \right]^2}{4(p^3 + r^3) - 3(r^2 - p^2)^2} \quad (3.5)$$

where $l = F^{-1}(r)$, $u = 1 - F^{-1}(p)$ and $F(\cdot)$ is the c.d.f. of the combined sample's underlying distribution.

Hence, the ARE of Gastwirth's modified test to the t -test varies based on the distribution function $f(x)$, and the optimal values of p and r that maximize $ARE_{G,t}$ in equation (3.5) thus also depend on the distribution function. Gastwirth suggested that one should set p equal to r for symmetric distributions, and let them differ for skewed distributions. Accordingly, we obtain the ARE of Gastwirth's modified rank test to the t -test for uniform, normal, Laplace and exponential distributions, which represent symmetric light-, medium-, heavy- tailed and right skewed models, respectively.

Table 1. A.R.E. of Gastwirth's Modified Rank Tests to the t -Test

DGPs	p.d.f.	A.R.E. Formula	Note
Uniform	1, if $x \in [0, 1]$ 0, otherwise	$1/2p$	$p = r$
Normal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	$3\Phi^2(\sqrt{2}u)/2\pi p^2$	$p = r$
Laplace	$\frac{1}{2} e^{- x }$	$3p$	$p = r$
Exponential	e^{-x}	$3/(4r - 3r^2)$	$p = 0$

We consider first the symmetric light-tailed model. The $ARE_{G,t}$ for uniform distributions is higher when both p and r are smaller. Intuition for this can be gained by considering the case where the treatment population is located to the right of the control population (i.e., the null is false). In this case, observations at the bottom ranks are more likely to be from the control population, and those at the top ranks are more likely to be from the treatment population. In contrast, middle ranks could be from either population, and thus carry less information about the underlying location shift⁷.

⁷As sample sizes approach infinity, Gastwirth's modified test needs to score fewer data to achieve the same power as the t -test (Gastwirth 1965, Proposition 4, p. 1137). However, for finite sample sizes, one must make the tradeoff between limited information and improved efficiency from scoring fewer data.

The opposite holds for symmetric heavy-tailed models such as the Laplace distribution. There, Gastwirth’s modified test is relatively more efficient with larger values of p and r . For normal distributions, the ARE is maximized by assigning no weight to the middle: $p = r = 42\%$. Finally, with data that are skewed to the right, such as the exponential distribution, Gastwirth suggested scoring only the bottom ranks because they begin near the median and consequently are more informative about differences in medians.

The key message of this discussion is that proper application of Gastwirth’s modified tests requires knowing the nature of the underlying distribution. The next section describes an adaptive approach for discovering this.

4 Two-Sample Adaptive Procedure for Location

As noted above, to take advantage of Gastwirth’s modified rank tests, one needs to have some sense for the underlying distribution’s shape. Adaptive procedures provide this information, by using the data first to determine an appropriate test procedure and then using that test to draw inferences (e.g., Hogg 1967; Hájek 1970; Jaeckel 1971; Hogg 1974). In the context of rank tests, the adaptive procedure developed by Hogg, Fisher, and Randles (1975, HFR hereafter) has been highly influential⁸. This section reviews HFR’s adaptive distribution-free procedure in detail.

4.1 Hogg Fisher and Randles’s (1975) Adaptive Procedure

HFR’s procedure consists of two steps. First, it classifies the combined sample as arising from one of four pre-specified models by estimating the skewness and tailweight of the underlying distribution. Second, it selects and applies the model’s corresponding rank test.

Based on Fisher’s (1972) work, HFR’s measure of a data set’s skewness is the ratio of the distance between the upper end and the midmean to the distance between the lower end and the midmean,

⁸Jones (1979) used a similar framework to HFR and developed an adaptive distribution-free procedure for one-sample test of location. Bünning (1999), discussed in the next section, directly extends HFR’s adaptive procedure for k -sample tests. Other examples include Yuh and Hogg (1988), O’Gorman (1996), Büning and Kössler (1999), and Neuhäuser et al. (2000).

$$Q_1 = \frac{\bar{U}_{5\%} - \bar{M}_{50\%}}{\bar{M}_{50\%} - \bar{L}_{5\%}} \quad (4.1)$$

where $\bar{U}_{5\%}$, $\bar{M}_{50\%}$ and $\bar{L}_{5\%}$ are the averages of the upper 5%, middle 50% and the lower 5% of Z 's (order statistics of the combined sample). For example, HFR considers data to be right-skewed when $Q_1 > 2$, i.e., the upper end is more than twice further away from the midmean ($\bar{M}_{50\%}$) than the lower end.

Using results from Uthoff (1970), HFR's measure of tailweight is the ratio of the distance between upper end and lower end to the distance between the averages of upper half and lower half,

$$Q_2 = \frac{\bar{U}_{50\%} - \bar{L}_{50\%}}{\bar{U}_{50\%} - \bar{L}_{50\%}} \quad (4.2)$$

where $\bar{U}_{.5}$ and $\bar{L}_{.5}$ are the averages of the upper 50% and lower 50% of Z 's. HFR considers the data to be very heavy-tailed when $Q_2 > 7$, or light-tailed when $Q_2 < 2$.

Figure 1 summarizes HFR's model classification scheme (Hogg et al. 1975, p. 658)⁹.

<Figure 1>

Once data are classified, the adaptive procedure applies the corresponding rank test as described below.

Symmetric Heavier-Tailed Model HFR chose the WMW test (equation (2.4))

for the symmetric heavier-tailed model. The WMW test is the most powerful rank test for logistic distributions, so it is a reasonable choice for distributions with similar characteristics.

Symmetric Light-Tailed Model HFR chose a modified rank test that scores only the bottom 25% and top 25% of the data ($r = p = 25\%$) for the symmetric light-tailed model. The intuition, as noted in Section 3.3, is that in this case the extreme ranks are more informative about location shifts than the central ones. The scoring function is

$$a_L(R_i) = \left\{ \begin{array}{ll} R_i - \text{floor}[25\%(N+1)] - 1/2, & \text{if } R_i \leq 25\%(N+1) \\ R_i - \text{ceiling}[75\%(N+1)] - 1/2, & \text{if } R_i \geq 75\%(N+1) \\ 0, & \text{otherwise} \end{array} \right\}$$

⁹HFR noted that the left-skewed model ($Q_1 < \frac{1}{2}$ and $Q_2 < 7$) can be easily added.

where $\text{floor}(x)$ rounds x down to the nearest integer, and $\text{ceiling}(x)$ rounds x up to the nearest integer.

Right-Skewed Model HFR chose a modified rank test that scores only the bottom 50% of the data ($r = 50\%$ and $p = 0\%$) for the right-skewed model. Gastwirth (1965) illustrated that this is because bottom ranks begin near the median and consequently are more informative about differences in medians. The scoring function is

$$a_{RS}(R_i) = \begin{cases} R_i - \text{floor}[25\%(N + 1)] - 1, & \text{if } R_i \leq (N + 1)/2 \\ 0, & \text{otherwise} \end{cases}$$

Very Heavy-Tailed Model HFR chose the median test (equation (2.6)) for very heavy-tailed models, due to the fact that the median test is asymptotically optimal for the heavy-tailed Laplace distribution¹⁰.

4.2 Independence of the Two Steps

This subsection addresses a common concern regarding the independence of steps in adaptive procedures. Model selection can in principle affect the sampling distribution of the second-stage test statistic, leading to difficulties in correctly assigning p-values. This is not a concern, however, in the case of HFR's adaptive procedure. The reason is that the preliminary model selection is statistically independent of subsequent rank tests.

In particular, during model selection, the skewness Q_1 and tailweight Q_2 are computed from the order statistics of the combined sample, so they are complete and sufficient for the underlying distribution $F(\cdot)$ under H_0 (for proof see e.g., Lehmann 1986, p. 143). At the same time, all rank tests are distribution-free, in the sense that their null distributions are independent of data's underlying distribution $F(\cdot)$ (Proof: see Appendix 1). Consequently, the two steps are mutually statistically independent, and thus the adaptive procedure is "honest".

¹⁰However, the median test has been known for low empirical performance. With fewer than 50 observations, the median test's empirical power is substantially lower than the WMW test even for Laplace distributions (Freidlin and Gastwirth 2000). We propose to drop the median test for optimizing HFR's algorithm in Section 6.

4.3 Empirical Performance of HFR (1975)

HFR reported a Monte-Carlo study that assessed the empirical performance of their adaptive procedure in comparison to the parametric t -test. Two samples of 15 observations ($n = m = 15$) were drawn from each of the fourteen distributions in a generalized Tukey's Lambda family (Ramberg and Schmeiser 1974). The location shift θ was set at 0 for size comparison, and 60% of the underlying distribution's standard deviation ($\theta = 0.6\sigma$) for power comparison.

Under the wide range of distributions, HFR's procedure maintains the desired nominal significance level 5%, yet exhibits significant improvement in power. In comparison with the t -test, the power of HFR's procedure is only slightly lower with normal distributions, but greater in all other cases, especially with skewed distributions as well as heavy-tailed Cauchy-like distributions. In comparison with the WMW test, the adaptive procedure does equally well when WMW is the most appropriate, and outperforms WMW when the data are symmetric light-tailed or skewed.

5 k -Sample Adaptive Procedures

The previous sections discussed tests for location differences with two-samples. However, in some cases it is of interest to test for the presence of a trend among three or more treatments. The Jonckheere test (Jonckheere 1954a; Terpstra 1952) is designed for such purposes. Its applications in economics dates to Smith (1964), and has appeared regularly since (see, e.g., Sherman 1971; Vyrastekova and Van Soest 2003; Brandts et al. 2008).

To develop the test, consider k random samples drawn from k populations, respectively, $(X_{11}, \dots, X_{1n_1})$, $(X_{21}, \dots, X_{2n_2})$, \dots , $(X_{k1}, \dots, X_{kn_k})$. Assume the c.d.f.'s of the k populations are connected through the relationship

$$F_j(t) = F(t - \theta_j), \quad -\infty < t < \infty$$

The subscripts j of F_j are arranged such that

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_k,$$

$$H_1 : \theta_1 \leq \theta_2 \leq \dots \leq \theta_k, \text{ with at least one strict inequality.}$$

5.1 Jonckheere and Jonckheere-Type Tests

To perform the Jonckheere test, one first arranges all k samples in ascending order such that $\theta_1 \leq \theta_2 \leq \dots \leq \theta_k$. For each pair of the u^{th} and v^{th} samples ($1 \leq u < v \leq k$), compute the two-sample *one-sided* WMW test statistics S_{uv} according to equation (2.4). The Jonckheere test is the sum of $\binom{k}{2}$ two-sample WMW tests,

$$JT = \sum_{u=1}^{k-1} \sum_{v=u+1}^k S_{uv}$$

The test statistic's limiting distribution is normal if more than one sample is allowed to increase without limit as the combined sample size N approaches infinity (Jonckheere 1954a, p. 140). On the other hand, if only one sample increases without limit as $N \rightarrow \infty$, Jonckheere showed that the limiting distribution will be non-normal and platykurtic.

If the two-sample test statistic S_{uv} uses the scoring functions of modified rank tests, then the resulting JT tests become "Jonckheere-type" tests (Büning and Kössler 1996). Thus, HFR's adaptive procedure can be easily extended to the k -sample environment. Below we describe Büning's (1999) extension of HFR's procedure for k -sample trend testing.

5.2 Büning's (1999) Adaptive k -Sample Trend Tests

Büning's (1999) adaptive procedure is HFR's procedure extended to k -sample environment with differences in: (i) the cutoff between symmetric medium- and heavy-tailed models and (ii) the rank test used for the symmetric heavy-tailed model¹¹. Figure 2 describes Büning's (1999) model selection scheme.

<Figure 2>

¹¹Büning's (1999) modified rank test for symmetric heavy-tailed model has the following scoring function,

$$a_H(R_i) = \begin{cases} -\text{ceiling}[25\%(N+1)], & \text{if } R_i < 25\%(N+1) \\ R_i - (N+1)/2, & \text{if } 25\%(N+1) \leq R_i \leq 75\%(N+1) \\ \text{ceiling}[25\%(N+1)], & \text{if } R_i > 75\%(N+1) \end{cases}$$

Büning and Kössler (1999, p. 71) proved that under either H_0 or H_1 , Jonckheere-type statistics are asymptotically normally distributed under regularity conditions (all sample sizes approach infinity at the same rate and the scoring function is non-decreasing, square integrable and absolutely continuous).

In addition, Büning (1999) compared performance of adaptive procedures using skewness and tailweight of the *combined sample* (HFR’s approach) versus using averages of *individual samples*’s skewness and tailweight weighted by sample sizes. The latter approach improves classification accuracy¹² when the null hypothesis is false, but has the disadvantage that the resulting adaptive procedure is not distribution-free, because the weighted average measures are not functions of the order statistics of the combined sample, thus not independent of the second-stage linear rank statistics.

However, it is comforting to know that, for the purpose of detecting location differences, the combined sample approach does not do worse, because increased location differences cause more mis-classifications, but also increase the power of all pre-specified rank tests. Büning’s (1999) Monte-Carlo analyses showed that the empirical performance of adaptive procedures using the combined sample versus individual samples are fairly similar.

5.3 Empirical Performance of Büning (1999)

Büning’s (1999) simulation results are consistent with HFR’s. The adaptive procedure significantly outperforms its parametric counterpart and the Jonckheere test, with the improvement most apparent with symmetric light-tailed, skewed and Cauchy distributions (Büning 1999, pp. 549-550).

6 Power and Sample Size

This section reports a new Monte-Carlo study to investigate adaptive procedures’ empirical performance under various sample sizes. We also attempt to optimize parameters in order to improve the algorithm’s performance.

We find that the adaptive procedure displays substantially greater power than both WMW and t -tests with samples of moderate size ($20 \leq n, m \leq 40$), and the improvement is greater as samples sizes become larger (consistent with asymptotic theory).

¹²Büning’s (1999) Monte-Carlo results suggest that as location shifts $\theta_1, \dots, \theta_k$ increase, the combined sample approach has increasing mis-classifications, and data tend to be classified into medium-tailed regardless of the underlying distribution.

Improvements are especially significant when data follow light-tailed or skewed distributions, and when location differences are small (though presumably still economically significant).

6.1 Optimizing the Algorithm

First, we simplify the model selection scheme by dropping HFR's very heavy-tailed model, as shown in Figure 3.

<Figure 3>

The reason for this simplification is that the median test used for very heavy-tailed model generally has low power. Freidlin and Gastwirth's (2000) power simulation showed that with fewer than 50 observations in total, the median test has substantially lower power than the WMW test even for Laplace distributions for which it is asymptotically most powerful. The authors thus suggested that "the median test be 'retired' from routine use." HFR's simulation demonstrated that the median test only outperforms the t - and WMW tests slightly with Cauchy distribution, but does significantly worse in all other cases. Overall, the median test hurts the adaptive procedure's performance when it is selected incorrectly due to mis-classification (Hogg et al. 1975, p. 660), and helps only slightly when the data are very heavy tailed (e.g., Cauchy distributions). Our simulation results in the next subsection confirm this is the case.

Second, although sensibly valued, HFR (1975) did not explain how they chose the cutoff values for model selection. We here explore the optimization of these parameters, in an effort to improve the performance of the algorithm. Recognizing that the optimal cutoffs could vary with sample size N , we drew 50,000 samples for a given sample size N from each of three distributions: uniform, normal and exponential, representing symmetric light-, heavier-tailed and right skewed models respectively. We started with $Q_1^* = 2$ and $Q_2^* = 2$, as used by HFR, and performed a naive grid search for the optimal cutoffs. In particular, we calculated Q_1 and Q_2 for each sample using the current temporary cutoffs (Q'_1, Q'_2) , and then counted the number of mis-classifications. Our objective was to minimize the sum of squared mis-classifications in each distribution, as follows.

$$\min Loss = \sum_{i=uni,nor,exp} (\text{number of mis-classifications in distribution } i)^2$$

Table 2 summarizes the optimization results.

Table 2: Optimal Cutoffs for Skewness and Tailweight

Combined Sample Size N	Skewness Q_1^*	Tailweight Q_2^*
11-15	2.1	2.0
≥ 16	2.1	2.1

We conclude that the optimal cutoff values are fairly stable with different sample sizes, and that HFR’s cutoffs are close to optimal when evaluated according to this metric.

Finally, we also optimized values for t and b used in percentile modifications¹³. We report in the next subsection that seemingly small changes to these parameter can yield large improvements in power.

6.2 Power and Size Comparison

We compare the power and size of the adaptive procedure using our optimized parameters against the t -test, WMW or Jonckheere tests, and the HFR procedure, considering two-, three-, and four-sample cases.

6.2.1 Data Generating Processes

In addition to the three Data Generating Processes (DGPs) used for cutoff optimization (Section 6.1), we add the heavy-tailed Laplace distribution to stress-test our model selection scheme (recall we do not allow for a separate heavy-tailed category). Table 3 details all four DGPs: (i) uniform, (ii) normal, (iii) Laplace and (iv) exponential distributions, as well as the empirical skewness Q_1 and tailweight Q_2 , estimated using a sample of one million observations are also reported in Table 3.

Table 3. Empirical Skewness Q_1 and Tailweight Q_2

¹³We drew samples from the uniform distribution (symmetric light-tailed model), and started with

$t = b = 25\%$ used in HFR. We found $t = b = 22\%$ maximize test’s power. Similarly, we drew samples from the exponential distribution (right-skewed model), and started with HFR’s $b = 50\%$ and $t = 0$. We found $b = 45\%$ and $t = 0$ yields the highest power.

DGPs	p.d.f.	Mean: μ	s.d.: σ	Skewness: Q_1	Tailweight: Q_2
Uniform	1, $x \in [0, 1]$ 0, otherwise	0.5	0.29	1.00	1.90
Normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$	0	1	1.00	2.59
Laplace	$\frac{1}{2}e^{- x }$	0	1.4	1.00	3.30
Exponential	e^{-x}	1	1	4.56	2.86

We vary the number of observations per sample (allowing for unbalanced data), DGPs and magnitude of location shift (30% or 60% of the distribution's standard deviation σ). There are 100,000 replications in each cell. For convenience, denote the adaptive procedure using our optimized parameters by "HH".

6.2.2 Two-Sample Adaptive Procedure

Figure 4 compares HH against the t - and WMW tests in a two-sample environment. Our main result is that when location shifts are small and data are light-tailed or skewed, HH displays greater power than others even with 40 observations per sample. For example, using two samples of 40 observations drawn from an exponential distribution, the power of HH is 120% greater than the t -test¹⁴, and 32% higher than WMW test.

<Figure 4>

Appendices 2a and 2b provide details regarding this comparison. HH displays consistently higher power than HFR, and sometimes substantially so, while the two tests' sizes are nearly the same. For example, when two samples of 40 observations are drawn from uniform distributions with a (small) location shift of 30% of an s.d., the power of HH is 17% higher than HFR.

6.2.3 k -Sample Adaptive Procedure

Figures 5 and 6, and Appendices 3a, 3b, 4a and 4b, report similar results for three-sample and four-sample cases. HH continues to display significant improvement in power over the t - and WMW tests with samples of moderate size, and display small but consistent improvements over HFR¹⁵ for light-tailed and skewed data.

¹⁴Due to the central limit theorem, the t -test's power is invariant with respect to underlying distributions, for any given sample size and location shifts. In contrast, nonparametric tests have significantly higher power with non-normally distributed data.

¹⁵For consistency, we continue to compare HH against HFR's natural extension in k -sample cases instead of Binning's (1999) adaptive procedure.

7 Application: Gneezy and Smorodinsky (2006)

This section illustrates the adaptive procedure using optimized parameters (the "HH" procedure) on a subset of data from Gneezy and Smorodinsky (2006, "GS" hereafter).

GS investigated whether sellers' revenues and buyers' bids depend on the number of bidders in all-pay auctions. In their common-value all-pay auction, bidders submit bids for the prize of 100 points. The person with the highest bid wins the prize, and all must pay their bids. The auction is repeated ten times (i.e., ten rounds). The three treatments, T4, T8 and T12, only differ in the number of participants, $N=4, 8, 12$ (fixed throughout the ten rounds). Each subject participates in only one treatment, and knows N . There are 5 sessions in each treatment, and thus a total of 120 participants in the experiment.

Their main findings are that overbidding is prevalent in such auctions, and the seller's revenue depends strongly on the number of bidders in early rounds, but after a few rounds of play the seller's revenue becomes independent of the number of bidders. Regarding players' bids, there is a significant difference between mean bids (the average value of bids per session) in treatments with fewer bidders ($N=4, 8$) and those with more ($N=12$).

For the purpose of demonstrating the adaptive procedure, we choose to analyze the first round bids from GS treatments T8 and T12. GS reported $p=0.09$ (Table 1, p. 260) using the WMW test, providing some evidence that the average bids in the first round are different between 8-bidder and 12-bidder sessions.

Instead of session averages, we apply the adaptive procedure to individual bids. In doing so, we take full advantage of the information in the data without violating the assumption that observations must be independent (because first-round bids are not subject to history or feedback). Figure 7 describes the distributions of the 40 bids from T8 and 60 bids from T12. Note those distributions are "U" shaped, and thus assumptions of normality may not be tenable. Hence, analysis of these data might be well accomplished using adaptive procedures.

We begin by calculating the skewness and tailweight of the combined sample of 100 observations, using equations (4.1) and (4.2) respectively. It turns out that the skewness and tailweight values are 1.32 and 1.28, respectively, so that these data are classified as symmetric light-tailed. Next we apply the appropriate modification of the

WMW test¹⁶. We reject the null hypothesis that the two samples are drawn from the same population with $p=0.03$. In a stark contrast, and despite the fact that rejection is possible using session averages, when evaluating the entire data set neither WMW nor t -tests are able to reject the null, with $p=0.10$ and $p=0.15$, respectively (all tests are two-sided).

8 Practical Advice and Comments on Analysis of Binary Data

People who wish to take advantage of the adaptive procedures discussed above should consider the following. First, the data must satisfy all assumptions of the WMW test, including independence across samples (i.e., between-subject designs) and equality of variance. It is worth emphasizing that the WMW and Jonckheere tests, as well as the adaptive procedures, explicitly assume that the underlying distributions of all samples are identical up to location. Therefore, one must ensure that the variances of the distributions are statistically identical. If the variances differ, it is necessary to conduct variance stabilization (see, e.g., Box et al. 2005, pp. 320-322).

Second, adaptive procedures' power advantage over the t -test increases as sample sizes increase, and are present even with samples of moderate size. The importance of conducting non-parametric analysis with small samples is widely recognized; this paper emphasized that adaptive nonparametric procedures remain valuable alternatives to the t -test even when sample sizes are large.

Finally, one should be aware that adaptive procedures perform better when data have fewer ties. In the case of binary data, the skewness statistic cannot be calculated when the proportion of ties exceeds 75%, and thus the adaptive procedures we discussed cannot be implemented.

In the context of binary data, it is worthwhile to emphasize that location differences between samples (i.e., differences in the probability of "success") also imply variance differences; this is a problem in principle for the WMW test. To assess this problem's practical consequences we conducted a Monte-Carlo study to compare the size and power of WMW against the t -test (with equal variances) as well as the proportion test¹⁷. Power and size simulations were conducted for four balanced sample size set-

¹⁶When there are ties at the cutoff percentiles, the adaptive procedure scores all tied observations.

¹⁷The "proportion test" in Stata, commonly used for binary data, is a Z-test for detecting the difference between frequencies of "success" in two samples.

tings: $n = m = 5, 10, 15, 20$. Samples were drawn from populations with success rates ranging from 0.1 to 0.9. Figure 8a describes the results of size simulation, where the two samples are from identical distributions. Figure 8b describes the results of power simulation, and the location shift is set such that the success rate of the first sample's underlying distribution is fixed at 0.5, while the success rate of the second sample's underlying distribution varied from 0.1 to 0.9.

This analysis yielded four key findings. First, the t -test and proportion test reject the null hypothesis, whether true or false, at identical rates. Second, all three tests' sizes are around 5% when each sample includes 20 observations. Third, all tests' powers increase as differences between samples' success rates become larger. Finally, with very small samples ($n = m = 5$), the WMW test is significantly more conservative than the other two tests, with both size and power lower than the other tests (Figure 8a and Figure 8b, top left panel). These results suggest that the WMW test is as powerful as the t -test and proportion test with moderate sample sizes, but more conservative with small samples. Hence, the WMW test is appropriate for drawing inferences about location differences of binary data.

<Figures 8a and 8b>

9 Summary

To summarize, this paper argued that research in experimental economics can profit by taking advantage of adaptive statistical techniques. Adaptive procedures demonstrate significant improvement in power over parametric t -test and nonparametric WMW and Jonckheere tests, especially when data are skewed or symmetric light-tailed. This improvement is substantial when the location shift is small and persists even when sample sizes are large. A Stata package to implement the adaptive procedure is available from the authors on request.

Many powerful algorithms to detect location differences between samples surely remain to be discovered. It would be especially valuable for future research to extend adaptive approaches to matched sample procedures including the Wilcoxon signed-rank and Page tests (Page 1963)¹⁸.

¹⁸Jonckheere (1954b) also has a trend test for matched observations, in addition to the better known Jonckheere-Terpstra test for independent observations (Jonckheere,1954a). However, Jonckheere's (1954b) test has not gained popularity.

Appendix 1.

Proof: rank tests T (or JT) are independent from $F(\cdot)$ under H_0 .

Let X_1, \dots, X_N be independently distributed with densities f_1, \dots, f_N , and let the rank of X_i be denoted by R_i , then

$$\begin{aligned} \Pr\{R_1 = r_1, \dots, R_N = r_N\} \\ &= \Pr\{x_i \text{ is the } r_i\text{th smallest in } (x_1, \dots, x_N)\} \\ &= \int \cdots \int_{x_i \text{ is the } r_i\text{th smallest}} f_1(x_1) \cdots f_N(x_N) dx_1 \cdots dx_N \end{aligned}$$

Let $w_{r_i} = x_i$, we proceed to

$$= \int \cdots \int_{w_1 < \dots < w_N} f_1(w_{r_1}) \cdots f_N(w_{r_N}) dw_1 \cdots dw_N$$

Introduce any density f that is positive when at least one of the f_i is positive, and $V_{(1)} < \dots < V_{(N)}$ is an ordered sample drawn from f , we know the joint probability density function of the ordered statistics $V_{(1)} < \dots < V_{(N)}$ is $N!f(w_1) \cdots f(w_N)$ for $w_1 < \dots < w_N$, which implies $dF(V_{(1)} \cdots V_{(N)}) = N!f(w_1) \cdots f(w_N) dV_{(1)} \cdots dV_{(N)}$. Thus the above integral becomes

$$= \int \cdots \int_{V_{(1)} < \dots < V_{(N)}} f_1(V_{(r_1)}) \cdots f_N(V_{(r_N)}) dV_{(1)} \cdots dV_{(N)}$$

Multiple $N!f(V_{(r_1)}) \cdots f(V_{(r_N)})$ to both the nominator and denominator of the integrand, we get

$$\begin{aligned} &= \int \cdots \int_{V_{(1)} < \dots < V_{(N)}} \frac{N!f_1(V_{(r_1)})f(V_{(r_1)}) \cdots f_N(V_{(r_N)})f(V_{(r_N)})}{N!f(V_{(r_1)}) \cdots f(V_{(r_N)})} dV_{(1)} \cdots dV_{(N)} \\ &= \frac{1}{N!} \int \cdots \int_{V_{(1)} < \dots < V_{(N)}} \frac{f_1(V_{(r_1)}) \cdots f_N(V_{(r_N)})}{f(V_{(r_1)}) \cdots f(V_{(r_N)})} dF(V_{(1)} \cdots V_{(N)}) \\ &= \frac{1}{N!} E \left[\frac{f_1(V_{(r_1)}) \cdots f_N(V_{(r_N)})}{f(V_{(r_1)}) \cdots f(V_{(r_N)})} \right] \end{aligned}$$

Now let $f_1 = \dots = f_m = f$, $f_{m+1} = \dots = f_{m+n} = g$, and $N = m + n$; $S_1 < \dots < S_n$ denote the ordered ranks of X_{m+1}, \dots, X_{m+n} among all the X 's. For a given $s_1 < \dots < s_n$, we know

$$\begin{aligned}
\Pr(R_1 = r_1, \dots, R_m = r_m, S_1 = s_1, \dots, S_n = s_n) \\
&= n! \Pr(R_1 = r_1, \dots, R_m = r_m, S_1 = r_{m+1}, \dots, S_N = r_N) \\
&= \frac{n!}{N!} E \left[\frac{f(V_{(r_1)}) \cdots f(V_{(r_m)}) g(V_{(s_1)}) \cdots g(V_{(s_n)})}{f(V_{(r_1)}) \cdots f(V_{(r_m)}) f(V_{(s_1)}) \cdots f(V_{(s_n)})} \right] \\
&= \frac{n!}{N!} E \left[\frac{g(V_{(s_1)}) \cdots g(V_{(s_n)})}{f(V_{(s_1)}) \cdots f(V_{(s_n)})} \right]
\end{aligned}$$

Since there are $m!$ ways of arrange $(R_1 = r_1, \dots, R_m = r_m)$, we have

$$\begin{aligned}
\Pr(S_1 = s_1, \dots, S_n = s_n) \\
&= \frac{m!n!}{N!} \Pr(R_1 = r_1, \dots, R_m = r_m, S_1 = s_1, \dots, S_n = s_n) \\
&= \frac{1}{\binom{N}{n}} E \left[\frac{g(V_{(s_1)}) \cdots g(V_{(s_n)})}{f(V_{(s_1)}) \cdots f(V_{(s_n)})} \right]
\end{aligned}$$

Under the null hypothesis, $f(x) = g(x)$, it follows that

$$\Pr(S_1 = s_1, \dots, S_n = s_n) = \frac{1}{\binom{N}{n}}$$

The two-sample rank tests T are weighted sum of the ranks of the 2nd sample, s_1, \dots, s_n , therefore are statistically independent from $F(\cdot)$ of the DGP. It follows immediately for k -sample rank tests JT , as they are just sums of T .

References:

- Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd Edition, John Wiley.
- Bünning, H., & Kössler, W. (1996). Robustness and Efficiency of Some Tests for Ordered Alternatives in the c-sample Location Problem. *Journal of Statistical Computation and Simulation*, 55, 337-352
- Bünning, H., & Kössler, W. (1999). The Asymptotic Power of Jonckheere-Type Tests for Ordered Alternatives. *Australian & New Zealand Journal of Statistics*, 41(1), 67-77
- Bünning, H. (1999). Adaptive Jonckheere-Type Tests for Ordered Alternatives. *Journal of Applied Statistics*, 26(5), 541-551
- Bünning, H. (2009). Adaptive Tests for the c-Sample Location Problem. Book chapter of *Statistical Inference, Econometric Analysis and Matrix Algebra*, 3-17. Physica-Verlag HD
- Brandts, J., Pezanis-Christou, P., & Schram, A. (2008). Competition with Forward Contracts: A Laboratory Analysis Motivated by Electricity Market Design. *The Economic Journal*, 118, 192-214
- Chernoff, H., & Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics*, 29, 972-994.
- Cox, D. R., & Stuart, A. (1955). Some Quick Sign Tests for Trend in Location and Dispersion. *Biometrika*, 42(1/2), 80-95
- Foster, F. G., & Stuart, A. (1954). Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16, 1-22.
- Fisher, D. M. (1972). Classification, Selection, and Testing Procedures for Asymmetric Distributions. Unpublished thesis, Department of Statistics, University of Iowa.
- Freidlin, B. & Gastwirth, J. L. (2000). Should the Median Test be Retired from General Use? *The American Statistician*, 54(3), 161-164
- Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance, *Journal of the American Statistical Association*, 21(200), 675-701
- Gastwirth, J. L. (1965). Percentile Modification of Two Sample Rank Tests, *Journal of the American Statistical Association*, 60(312), 1127-1141
- Gneezy, U., & Smorodinsky, R. (2006). All-pay Auctions: An Experimental study. *Journal of Economic Behavior & Organization*, 61(2), 255-275
- Hájek, J. (1962). Asymptotically Most Powerful Rank-Order Tests. *Annals of Mathematical Statistics*, 33, 1124-1147

Hájek, J. (1970). Miscellaneous Problems of Rank Test Theory. in M.L. Puri, ed., *Nonparametric Techniques in Statistical Inference*, Cambridge, Mass.: Cambridge University Press

Hájek, J., & Šidák, Z. (1967). *Theory of Rank Tests*, New York and London: Academic Press

Hájek, J., Šidák, Z., & Sen, P. K. (1999). *Theory of Rank Tests*, 2nd edition. New York and London: Academic Press

Hogg, R. V. (1967). Some Observations on Robust Estimation. *Journal of the American Statistical Association*, 62, 1179-86

Hogg, R. V. (1974). Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory. *Journal of the American Statistical Association*, 69(348), 909-923

Hogg, R. V., Fisher, D. M. & Randles, R. H. (1975). A Two-Sample Adaptive Distribution-Free Test. *Journal of the American Statistical Association*, 70(351), 656-661

Hodges, J.L., Jr, & Lehmann, E. L. (1956). The Efficiency of Some Nonparametric Competitors of the t -test. *Annals of Mathematical Statistics*, 27, 324-35.

Hodges, J.L., Jr, & Lehmann, E. L. (1961). Comparison of the Normal Scores and Wilcoxon Test. *Proc. Fourth Berkeley Symp. Math. Stat. Prob.* 1, 307-317. University of California Press.

Hotelling, H., & Pabst, M. R. (1936). Rank Correlation and Tests of Significance Involving No Assumption of Normality. *Annals of Mathematical Statistics*, 7, 29-43.

Jaeckel, L.A. (1971). Robust Estimates of Location: Symmetry and Asymmetry Contamination. *Annals of Mathematical Statistics*, 42, 1020-1034.

Jones, D. H. (1979). An Efficient Adaptive Distribution-Free Test for Location. *Journal of the American Statistical Association*, 74(368), 822-828

Jonckheere, A.R. (1954a). A Distribution-Free k -Sample Test Against Ordered Alternatives. *Biometrika*, 41(1/2), 133-145

Jonckheere, A.R. (1954b). A Test of Significance for the Relation between m Rankings and k Ranked Categories Alternatives. *British Journal of Statistical Psychology*, 7(Part II), 93-100

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed, New York: Wiley

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18(1), 50-60

Mood, A. M. (1954). On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests. *Annals of Mathematical Statistics*, 25(3), 514-522

- Neuhäuser, M., Seidel, D., Hothorn, L.A. & Urfer, W. (2000). Robust trend tests with application to toxicology. *Environmental and Ecological Statistics*, 7(1), 43-56
- O’Gorman, T. W. (1996). An adaptive two-sample test based on modified Wilcoxon scores, *Communications in statistics. Simulation and computation*, 25(2), 459-479
- Page, E. B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, 58(301), 216-230
- Pitman, E. J. G. (1949). Lecture Notes on Non-parametric Statistical Inference, given at Columbia University (unpublished).
- Ramberg, J. S., & Schmeiser, B.W. (1974). An Approximate Method for Generating Asymmetric Random Variables. *Communications of the Association of Computing Machinery*, 17, 78-82
- Sherman, R. (1971). An Experiment on the Persistence of Price Collusion An Experiment on the Persistence of Price Collusion. *Southern Economic Journal*, 37(4), 489-95
- Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*. 2nd Edition, McGraw-Hill International Editions
- Smith, V. L. (1964). Effect of Market Organization on Competitive Equilibrium. *Quarterly Journal of Economics*, 78(2), 181-201.
- Terpstra, T. J. (1952). The Asymptotic Normality and Consistency of Kendall’s Test Against Trend, When ties Are Present in One Ranking, *Indagationes Mathematicae*, 14, 327-333
- Uthoff, V. A. (1970). An Optimum Test Property of Two Well-Known Statistics. *Journal of the American Statistical Association*, 65, 1597-1600
- van der Waerden B. L. (1953). Ein neuer Test Fur das Problem der zwei Stichproben. *Mathematische Annalen*, 126, 93-107
- Vyrastekova, J. & van Soest, D. (2003). Centralized Common-Pool Management and Local Community Participation Centralized Common-Pool Management and Local Community Participation. *Land Economics*, 79(4), 500-14
- Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous populations of any form. *Nederl. Akad. Wetensch., Proc., Ser. A.*, 51, 252-261.
- Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 80-83
- Yuh, L., & Hogg, R. V. (1988). On Adaptive M -Regression. *Biometrics*, 44 (2), 433-445

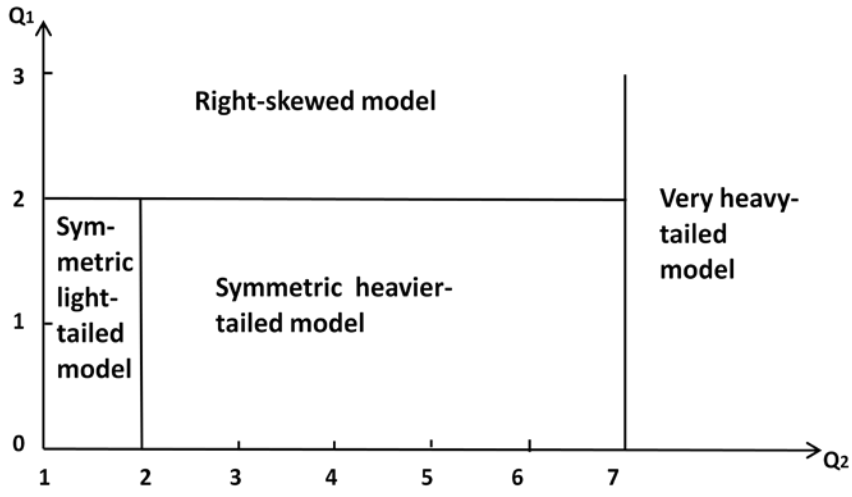


Figure 1: HFR's Model Selection Scheme: Skewness Q_1 and Tailweight Q_2

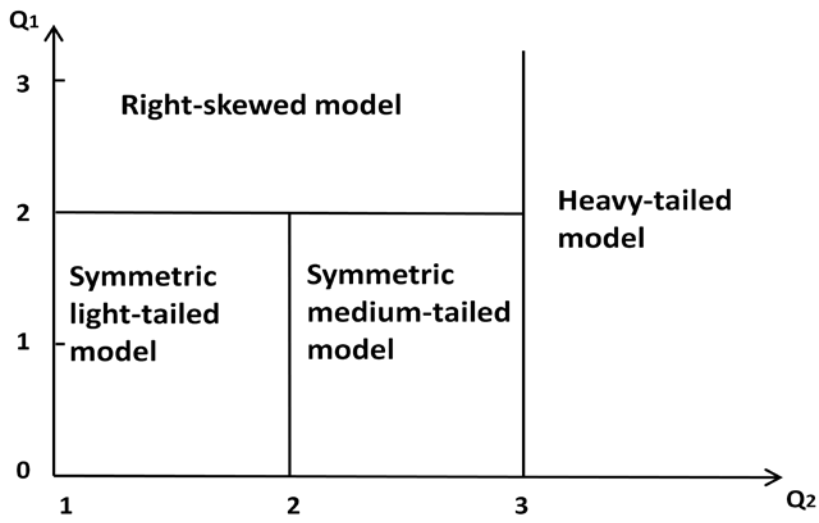


Figure 2: Büning's Model Selection Scheme: Skewness Q_1 and Tailweight Q_2

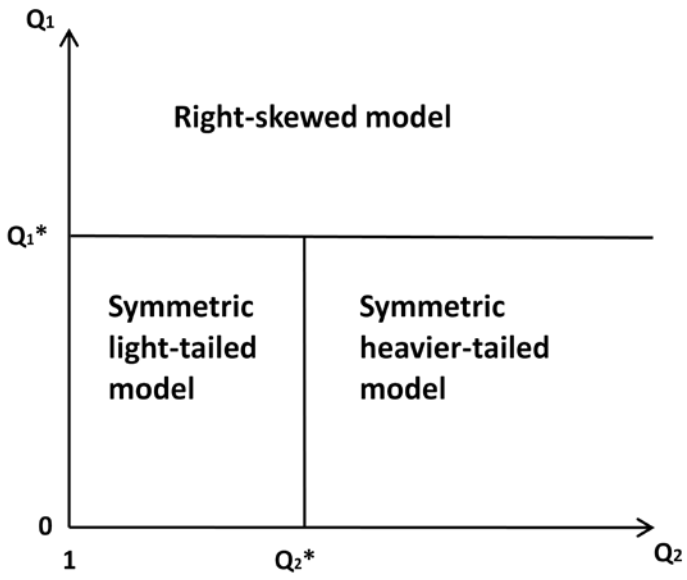


Figure 3: Optimized Model Selection Scheme: Skewness Q_1 and Tailweight Q_2

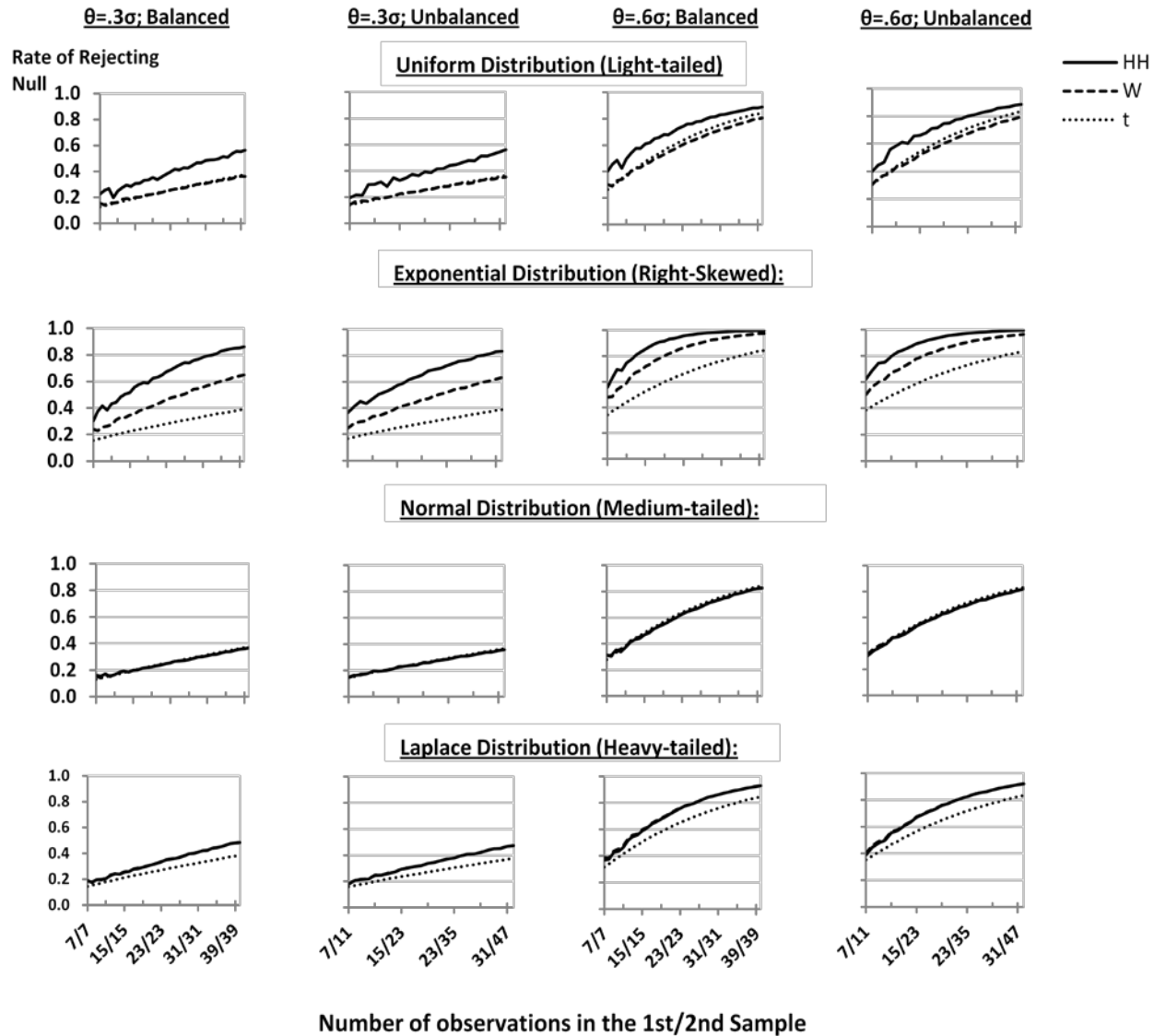


Figure 4: Power of HH, MWW and t -Tests (Two-Sample Environment)

Notes: θ is the location shift, and σ is the common standard deviation of both populations. For example, $\theta = .3\sigma$ is the case where the second population is shifted to the right of the first by 30% of the common standard deviation. On the other hand, $\theta = .6\sigma$ is when the location shift is twice as large. “Balanced” means the two samples have the same number of observations; “Unbalanced” refers to the case that the second sample has 50% more observations than the first sample. The x-axis indicates the numbers of observations in the first and the second samples, while the y-axis is the rate of rejecting the null, i.e., the power of the test. From the top to the bottom, data in the four rows of panels are generated from uniform, exponential, normal and Laplace distributions, respectively.

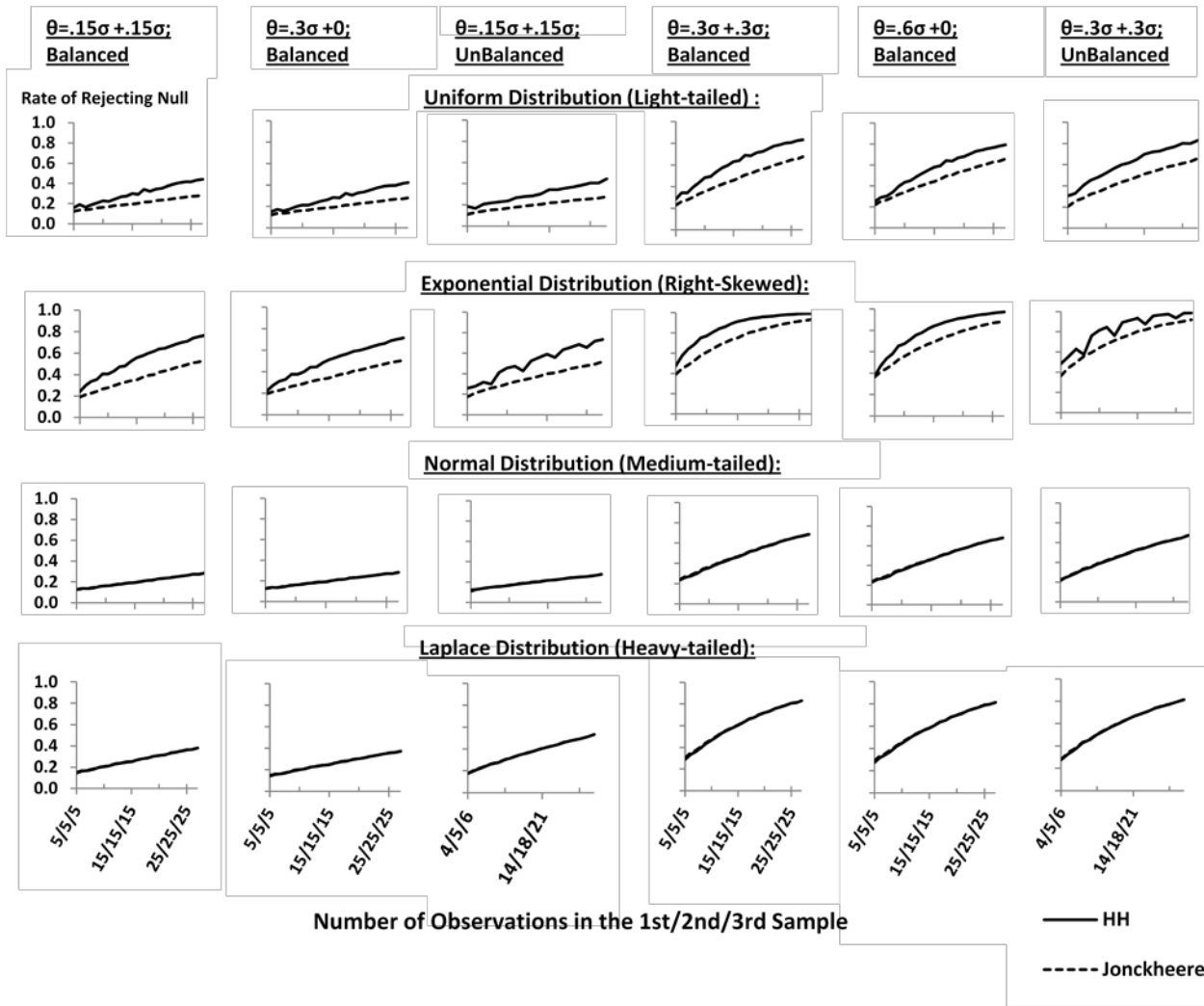


Figure 5: Power of HH and Jonckheere Tests (Three-Sample Environment)

Notes: θ is the location shift between adjacent populations, and σ is the common standard deviation of all three populations. For example, $\theta = .15\sigma + .15\sigma$ refers to case where the location shift between the first and second, and between the second and third populations are both 15% of the standard deviation. On the other hand, $\theta = .3\sigma + 0$ is when the only shift occurs between the first and second populations (30% of the common standard deviation), and the second and third populations are exactly the same. “Balanced” means all three samples have exactly the same number of observations; while “Unbalanced” refers to the case that the ratio of the three sample sizes are roughly 4:5:6. The x-axis indicates the numbers of observations in the first, second and third samples, while the y-axis is the rate of rejecting the null, i.e., the power of the test. From the top to the bottom, data in the four rows of panels are generated from uniform, exponential, normal and Laplace distributions, respectively.

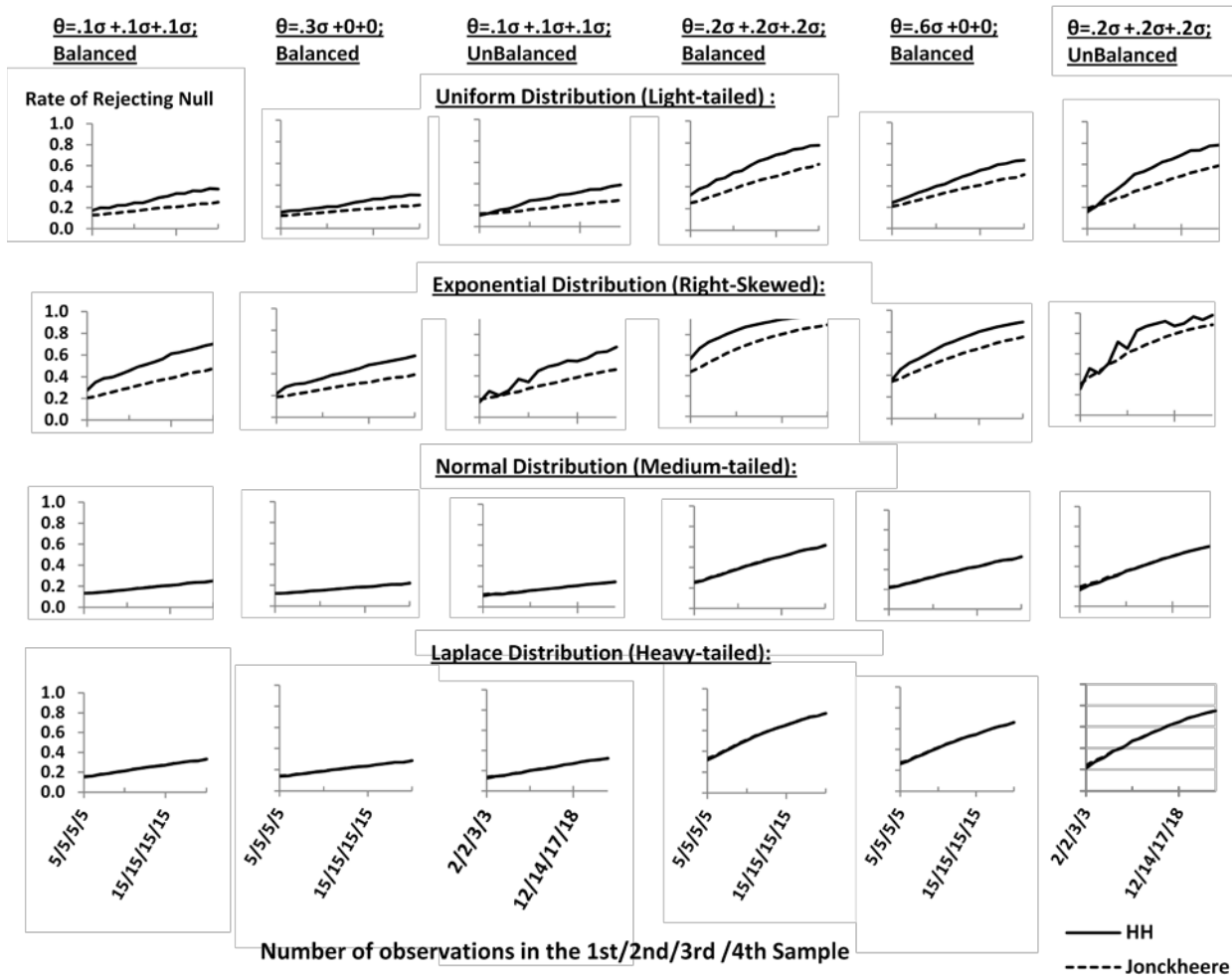


Figure 6: Power of HH and Jonckheere Tests (Four-Sample Environment)

Notes: θ is the location shift between adjacent populations, and σ is the common standard deviation of all four populations. For example, $\theta = .1\sigma + .1\sigma + .1\sigma$ refers to the case where the location shift between the first and second, between the second and third, and between the third and fourth populations are all 10% of the standard deviation. On the other hand, $\theta = .3\sigma + 0 + 0$ is when the only shift occurs between the first and second populations (30% of the common standard deviation), and the second, third and fourth populations are exactly the same. “Balanced” means all four samples have exactly the same number of observations; while “Unbalanced” refers to the case that the ratio of the four sample sizes are roughly 4:4.5:5.5:6. The x-axis indicates the numbers of observations in the first, second and fourth samples, while the y-axis is the rate of rejecting the null, i.e., the power of the test. From the top to the bottom, data in the four rows of panels are generated from uniform, exponential, normal and Laplace distributions, respectively.

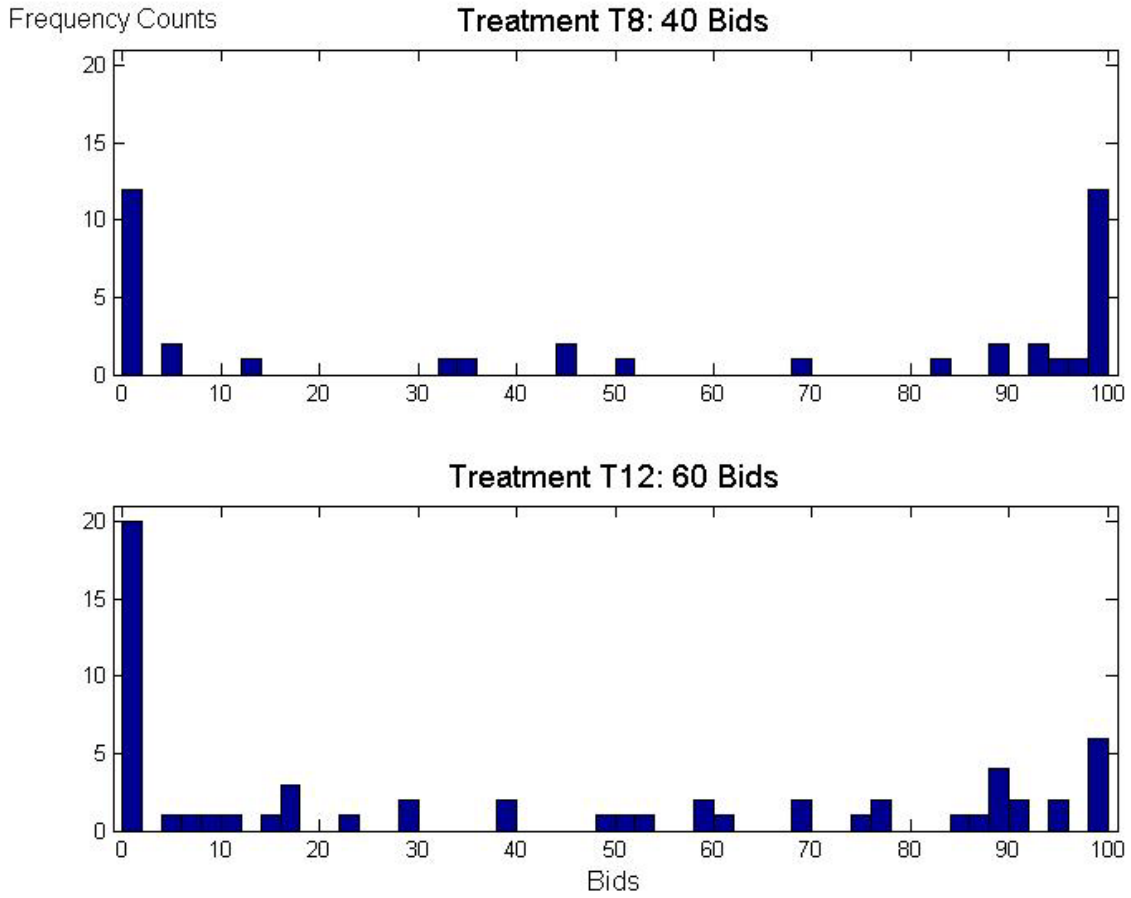


Figure 7: First-round Bids Distributions in Treatments T8 and T12 (Gneezy and Smorodinsky 2006)

— WMW - - t-test
 5% Level - · - Proportion test

Rate of Rejecting

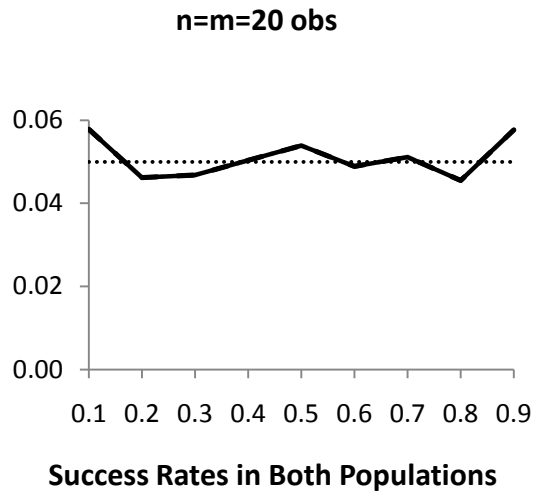
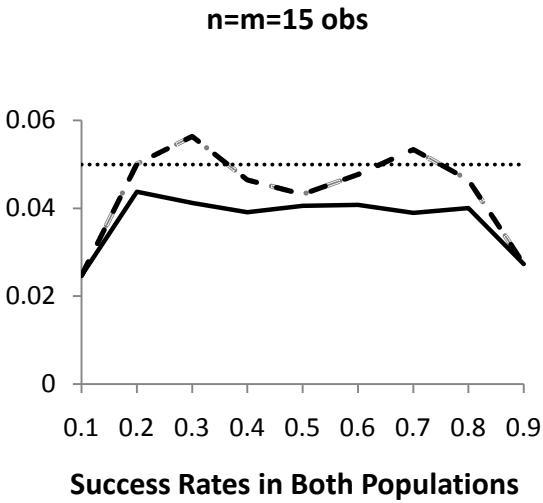
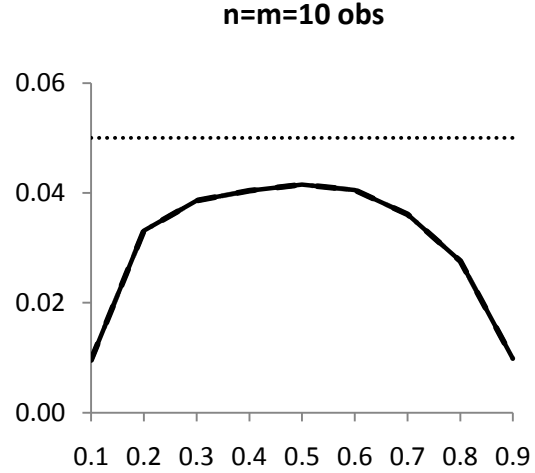
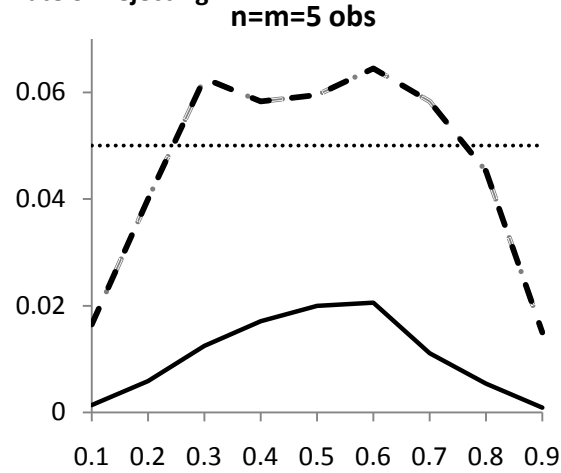


Figure 8a: Size of WMW, *t*- and Proportion Tests

Notes: To investigate the tests' sizes, we set the two populations exactly the same. The x-axis indicates the common success rate, and the y-axis is the rate of rejecting the null hypothesis, i.e., tests' sizes.

— WMW - - t-test - · - Proportion test

Rate of
Rejecting
Null

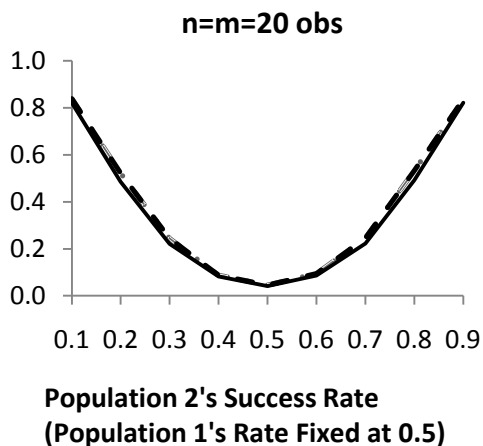
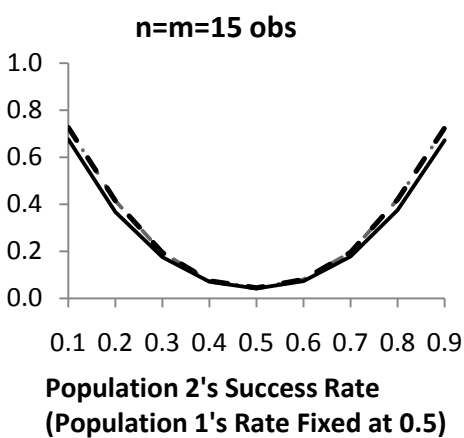
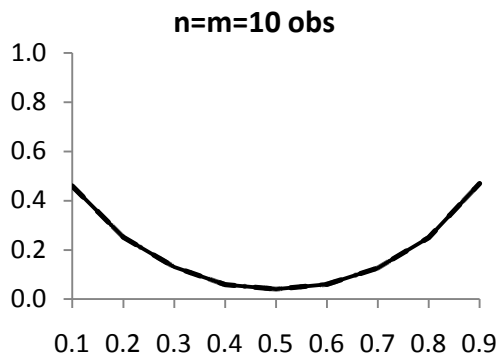
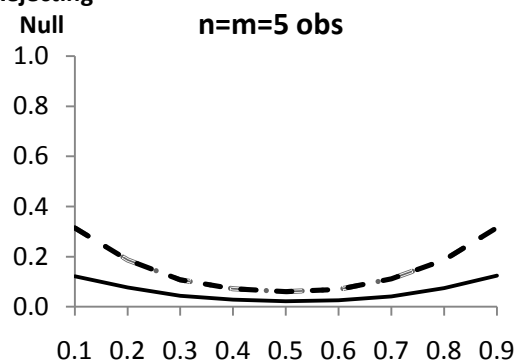


Figure 8b: Power and Size of WMW, t- and Proportion Tests

Notes: To investigate the tests' powers, we fix the first population's success rate at 50%, and vary the second population's success rate from 10% to 90%, as indicated on the x-axis. The y-axis is the rate of rejecting the null hypothesis, i.e., tests' powers, except when the second population's success rate is 50% (i.e., the midpoint in the figures), the rate of rejecting the null hypothesis is the test's size.

Appendix 2a. Empirical Size and Power in Percent: Balanced Two Samples

# of obs in 1st/2nd Sample	Uniform						Exponential						Normal						Laplace					
	$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
7/7	9	9	22	22	40	40	7	6	32	30	57	56	7	7	16	16	31	31	7	7	19	19	37	37
8/8	5	9	16	25	34	45	7	8	34	38	62	63	5	6	14	16	29	31	5	6	17	18	37	37
9/9	8	9	25	27	45	49	6	7	34	38	64	70	6	7	17	17	35	35	6	6	19	20	43	43
10/10	7	5	21	20	42	43	6	6	38	38	70	69	5	5	16	15	35	34	5	5	20	20	44	44
11/11	6	7	23	25	46	50	5	6	39	38	72	75	5	5	16	17	37	37	5	5	20	21	47	46
12/12	6	7	23	27	48	54	6	6	43	38	77	77	6	6	18	18	41	41	6	6	23	23	52	52
13/13	6	7	25	29	51	57	6	6	45	38	80	80	6	6	19	19	44	43	6	6	24	24	55	55
14/14	6	6	25	28	52	58	5	6	46	38	82	83	5	5	19	19	44	44	5	5	24	24	56	56
15/15	6	6	27	31	55	61	5	5	48	38	84	85	5	5	20	20	47	47	5	5	26	25	59	59
16/16	5	6	25	31	54	62	6	6	52	38	87	88	5	5	20	20	49	48	5	5	26	26	61	61
17/17	6	6	30	33	59	65	5	6	52	38	88	89	5	5	21	21	52	51	5	5	28	28	64	64
18/18	5	6	28	33	59	66	5	6	54	38	90	91	5	5	22	22	53	53	5	5	28	28	66	66
19/19	5	6	30	35	61	68	5	5	57	38	91	92	5	5	22	22	55	55	5	5	29	29	68	68
20/20	6	5	31	34	62	68	5	6	59	38	93	93	5	5	23	23	57	56	5	5	30	30	70	70
21/21	6	5	33	36	64	70	6	5	61	38	94	94	5	5	24	24	59	58	5	5	31	31	72	72
22/22	6	5	33	38	65	73	5	5	62	38	94	95	5	5	25	25	61	60	5	5	32	32	74	74
23/23	6	6	34	40	67	74	5	5	64	38	95	95	5	5	26	26	63	62	5	5	34	34	76	76
24/24	6	6	35	42	69	76	5	5	66	38	96	96	5	5	27	27	65	64	5	5	35	35	78	78
25/25	5	6	36	41	70	76	5	5	67	38	96	97	5	5	27	27	66	66	5	5	36	36	79	79

Appendix 2a (continued). Empirical Size and Power in Percent: Balanced Two Samples

# of obs in 1st/2nd Sample	Uniform						Exponential						Normal						Laplace					
	$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
26/26	5	6	37	43	71	78	5	5	69	38	97	97	5	5	27	27	67	67	5	5	36	36	80	80
27/27	5	6	38	43	72	78	5	6	70	38	97	98	5	5	28	28	68	68	5	5	37	37	81	81
28/28	5	5	39	45	73	80	5	5	71	38	98	98	5	5	29	29	70	70	5	5	38	38	83	83
29/29	6	5	40	47	75	81	5	5	74	38	98	98	5	5	30	30	72	72	5	5	40	40	84	84
30/30	5	5	40	47	75	82	5	5	73	38	98	98	5	5	30	30	73	73	5	5	40	40	85	85
31/31	5	5	41	48	76	83	5	5	75	38	99	99	5	5	31	31	74	74	5	5	41	41	86	86
32/32	5	6	42	49	77	83	5	5	76	38	99	99	5	5	32	32	75	75	5	5	42	42	87	87
33/33	5	5	43	49	78	84	5	5	77	38	99	99	5	5	32	32	76	76	5	5	42	42	88	88
34/34	5	5	43	50	79	85	5	5	78	38	99	99	5	5	33	33	78	78	5	5	44	44	89	89
35/35	5	5	45	51	80	86	5	5	79	38	99	99	5	5	34	34	78	78	5	5	44	44	90	90
36/36	5	5	45	51	81	86	5	5	80	38	99	99	5	5	34	34	79	79	5	5	45	45	90	90
37/37	5	6	46	54	82	87	5	5	81	38	99	100	5	5	35	35	80	80	5	5	46	46	91	91
38/38	5	5	47	56	83	88	5	5	82	38	100	100	5	5	36	36	82	82	5	5	48	48	92	92
39/39	5	5	48	55	83	88	5	5	83	38	100	100	5	5	36	36	82	82	5	5	48	48	92	92
40/40	5	5	48	56	84	89	5	5	84	38	100	100	5	5	37	37	83	83	5	5	48	48	93	93

Appendix 2b. Empirical Size and Power in Percent: Unbalanced Two Samples

# of obs in 1st/2nd Sample	Uniform						Exponential						Normal						Laplace					
	$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$		$\theta=0$		$\theta=.3\sigma$		$\theta=.6\sigma$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
7/11	6	6	19	20	38	40	5	6	33	36	61	62	5	5	15	15	31	31	5	5	18	18	39	39
8/12	6	6	18	22	40	44	7	7	40	41	69	69	6	6	16	16	35	35	6	6	21	21	45	45
9/14	5	5	21	21	43	47	5	6	38	45	70	74	5	5	17	17	38	37	5	6	21	21	48	48
10/15	6	7	24	29	49	56	6	5	42	43	75	75	5	6	17	18	40	40	5	5	22	22	51	50
11/17	5	7	23	30	50	59	5	6	45	47	79	80	5	6	19	20	44	44	5	5	25	25	56	55
12/18	5	7	24	32	52	61	6	5	48	50	81	83	5	5	19	19	45	45	5	5	25	25	57	57
13/20	5	5	26	28	55	60	5	5	49	52	84	85	5	5	20	20	48	47	5	5	26	26	61	60
14/21	6	7	29	35	57	66	5	5	51	54	86	87	5	5	21	21	51	50	5	5	27	27	63	63
15/23	6	5	30	33	60	66	5	5	54	57	88	89	5	5	23	22	54	54	5	5	29	29	67	67
16/24	5	6	30	35	61	68	5	5	56	59	90	91	5	5	23	23	56	56	5	5	30	30	69	69
17/26	5	6	31	37	63	71	5	5	58	62	91	92	5	5	24	24	58	58	5	5	31	31	72	71
18/27	5	5	32	36	65	72	5	5	59	63	92	93	5	5	24	24	60	60	5	5	32	32	73	73
19/29	5	5	33	39	67	75	5	5	63	65	94	94	5	5	26	26	63	63	5	5	34	34	76	76
20/30	5	5	35	39	68	75	5	5	64	68	95	95	5	5	26	26	64	64	5	5	34	34	77	77
21/32	5	5	36	42	70	78	5	5	66	69	95	96	5	5	27	27	66	66	5	5	36	36	80	80
22/33	5	5	37	42	72	78	5	5	67	70	96	97	5	5	28	28	68	68	5	5	37	37	81	81
23/35	5	5	38	44	73	80	5	5	69	72	97	97	5	5	29	29	70	69	5	5	38	38	83	82
24/36	5	5	39	45	74	81	5	5	71	74	97	98	5	5	30	30	72	71	5	5	39	39	84	84
25/38	5	5	41	46	76	82	5	5	73	76	98	98	5	5	31	31	73	73	5	5	41	41	85	85
26/39	5	5	41	48	77	83	5	5	73	76	98	98	5	5	31	31	74	74	5	5	41	41	86	86
27/41	5	5	42	48	78	84	5	5	75	77	98	98	5	5	32	32	75	75	5	5	42	42	87	87
28/42	6	6	44	51	79	86	5	5	76	80	99	99	5	5	33	33	77	77	5	5	44	44	89	89
29/44	5	5	45	51	80	86	5	5	77	80	99	99	5	5	34	34	78	78	5	5	45	45	90	90
30/45	5	5	46	53	81	87	5	5	78	81	99	99	5	5	34	34	79	79	5	5	45	45	90	90
31/47	5	5	47	55	82	88	5	5	80	83	99	99	5	5	35	35	81	81	5	5	47	47	91	91
32/48	5	6	47	56	83	89	5	5	81	83	99	99	5	5	36	36	82	82	5	5	47	47	92	92

Appendix 3a. Empirical Size and Power in Percent: Balanced Three Samples

# of obs in 1st/2nd/3rd Sample	Uniform										Exponential									
	$\theta=0$		$\theta=-.15\sigma+.15\sigma$		$\theta=-.3\sigma+0$		$\theta=-.3\sigma+.3\sigma$		$\theta=-.6\sigma+0$		$\theta=0$		$\theta=-.15\sigma+.15\sigma$		$\theta=-.3\sigma+0$		$\theta=-.3\sigma+.3\sigma$		$\theta=-.6\sigma+0$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
5/5/5	7	7	16	16	15	15	28	28	25	25	6	6	23	24	22	22	47	48	39	37
6/6/6	6	7	15	19	14	17	29	34	27	29	6	6	27	30	25	28	55	57	46	47
7/7/7	5	6	16	17	15	16	33	34	29	30	5	6	29	34	28	31	60	64	51	54
8/8/8	6	5	18	19	17	17	37	39	34	34	6	6	34	35	32	33	68	69	59	58
9/9/9	6	6	20	21	19	20	41	44	37	40	6	6	37	41	35	38	73	75	64	65
10/10/10	5	6	20	23	19	21	42	48	39	43	6	5	40	41	37	38	77	77	68	67
11/11/11	5	5	21	22	20	21	46	49	42	45	5	5	41	43	39	40	80	81	71	71
12/12/12	5	5	22	25	21	23	48	54	45	49	5	6	45	47	42	44	83	84	75	75
13/13/13	5	5	24	27	23	25	51	57	47	52	5	5	47	48	44	45	86	86	78	78
14/14/14	6	5	25	28	24	26	54	60	50	55	5	6	50	52	47	48	88	89	81	81
15/15/15	5	5	26	30	25	28	55	63	51	58	5	5	51	56	47	51	90	91	82	84
16/16/16	5	5	26	29	25	28	57	64	53	59	5	5	53	57	50	53	91	93	85	86
17/17/17	6	6	30	34	28	32	61	69	57	64	5	5	56	60	52	55	93	94	87	88
18/18/18	5	5	29	32	28	31	61	68	57	64	5	5	58	62	54	57	94	95	89	89
19/19/19	5	5	31	34	30	33	64	71	60	67	5	5	60	64	56	59	95	96	90	91
20/20/20	5	5	31	35	30	34	65	72	61	68	5	5	61	65	57	60	96	96	91	92
21/21/21	5	5	33	37	31	36	67	75	63	70	5	5	63	67	59	62	96	97	92	93
22/22/22	5	5	33	39	32	37	68	77	64	73	5	5	65	68	61	64	97	98	93	94
23/23/23	5	6	36	41	34	39	71	79	67	74	5	5	67	70	63	65	98	98	94	95
24/24/24	5	5	36	41	34	39	71	80	68	76	5	5	68	71	63	66	98	98	95	95
25/25/25	5	5	37	42	36	40	73	81	69	77	5	5	70	74	65	69	98	99	96	96
26/26/26	5	5	37	43	36	41	74	82	70	78	5	5	71	75	67	70	99	99	96	97
27/27/27	5	5	39	44	37	42	76	83	72	79	5	5	73	77	68	71	99	99	97	97

Appendix 3a (continued). Empirical Size and Power in Percent: Balanced Three Samples

# of obs in 1st/2nd/3rd Sample	Normal										Laplace									
	$\theta=0$		$\theta=-.15\sigma+.15\sigma$		$\theta=-.3\sigma+0$		$\theta=-.3\sigma+.3\sigma$		$\theta=-.6\sigma+0$		$\theta=0$		$\theta=-.15\sigma+.15\sigma$		$\theta=-.3\sigma+0$		$\theta=-.3\sigma+.3\sigma$		$\theta=-.6\sigma+0$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
5/5/5	6	6	13	13	13	13	23	23	23	23	6	6	15	15	14	14	29	29	28	27
6/6/6	6	6	13	14	13	14	26	26	26	26	6	6	16	16	16	16	34	33	32	32
7/7/7	5	6	14	14	14	14	28	28	27	27	5	5	17	17	16	16	36	36	34	34
8/8/8	5	5	15	14	15	14	31	30	30	29	5	5	18	18	18	17	40	39	38	38
9/9/9	6	6	16	16	16	16	34	34	33	33	6	6	20	20	19	19	44	44	42	42
10/10/10	5	5	16	16	16	16	36	36	35	35	5	5	20	20	20	20	47	47	45	45
11/11/11	5	5	17	17	17	17	39	38	38	37	5	5	21	21	21	21	50	50	48	48
12/12/12	5	5	18	18	18	18	41	41	41	40	5	5	23	23	23	23	53	53	51	51
13/13/13	5	5	18	18	18	18	43	43	42	42	5	5	24	24	23	23	56	56	54	54
14/14/14	5	5	19	19	19	19	45	45	45	44	5	5	25	25	24	24	58	58	56	56
15/15/15	5	5	19	19	19	19	47	47	46	46	5	5	25	25	25	25	60	60	58	58
16/16/16	5	5	20	20	20	20	49	49	49	48	5	5	27	27	26	26	63	63	61	61
17/17/17	5	5	21	21	21	21	52	52	51	51	5	5	28	28	28	28	66	66	64	64
18/18/18	5	5	22	22	22	22	53	53	53	52	5	5	29	29	28	28	68	68	65	65
19/19/19	5	5	23	23	23	23	56	56	55	55	5	5	30	30	30	30	70	70	68	68
20/20/20	5	5	24	23	23	23	58	57	57	57	5	5	31	31	30	30	72	72	70	70
21/21/21	5	5	24	24	24	24	59	59	58	58	5	5	31	31	31	31	74	74	71	71
22/22/22	5	5	25	25	25	25	61	61	61	60	5	5	33	33	33	33	76	76	74	74
23/23/23	5	5	26	26	25	25	63	63	62	62	5	5	34	34	34	34	77	77	76	76
24/24/24	5	5	26	26	26	26	65	64	64	64	5	5	35	35	35	35	79	79	77	77
25/25/25	5	5	27	27	27	27	66	66	66	65	5	5	36	36	36	36	81	81	79	79
26/26/26	5	5	27	27	27	27	67	67	67	67	5	5	37	37	36	36	81	81	80	80
27/27/27	5	5	29	29	28	28	69	69	68	68	5	5	38	38	37	37	83	83	81	81

Appendix 3b. Empirical Size and Power in Percent: Unbalanced Three Samples

# of obs in 1st/2nd/3rd Sample	Uniform						Exponential						Normal						Laplace					
	$\theta=0$		$\theta=.15\sigma^2$		$\theta=.3\sigma^2$		$\theta=0$		$\theta=.15\sigma^2$		$\theta=.3\sigma^2$		$\theta=0$		$\theta=.15\sigma^2$		$\theta=.3\sigma^2$		$\theta=0$		$\theta=.15\sigma^2$		$\theta=.3\sigma^2$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
4/5/6	5	8	12	19	23	31	5	6	22	26	44	49	5	6	11	12	20	22	5	5	17	18	27	27
5/6/8	6	6	17	17	32	33	6	5	28	28	56	56	6	5	13	13	26	26	5	5	21	21	34	33
6/8/9	6	7	17	21	34	40	6	5	32	32	63	63	5	6	14	14	29	29	5	5	23	23	38	37
7/9/11	6	6	20	22	41	46	6	6	29	30	58	57	5	6	15	15	33	33	5	5	26	26	43	43
8/10/12	5	6	20	23	43	48	5	6	38	41	74	76	5	5	16	16	35	35	5	5	27	27	46	45
9/11/14	6	5	22	24	47	52	6	6	41	46	79	82	5	5	17	17	39	38	5	5	30	30	50	50
10/13/15	5	6	23	27	49	56	5	6	44	47	82	85	5	5	18	18	41	41	5	5	32	32	54	54
11/14/17	5	6	24	28	53	60	5	5	42	43	76	76	5	5	19	19	44	44	5	5	35	35	57	57
12/15/18	5	5	26	28	55	62	5	5	49	53	88	89	5	5	20	20	47	46	5	5	36	36	60	60
13/16/20	5	5	27	31	58	65	5	5	53	56	90	91	5	5	21	20	49	49	5	5	38	38	63	63
14/18/21	6	6	29	34	61	70	5	6	55	59	92	93	5	5	22	22	52	52	5	5	41	40	66	66
15/19/23	5	5	30	35	64	72	5	5	54	56	88	88	5	5	22	22	54	54	5	5	42	42	69	69
16/20/24	5	5	31	36	64	73	5	5	60	64	95	96	5	5	23	23	57	57	5	5	44	44	71	71
17/21/26	5	5	33	37	67	75	5	5	62	66	96	97	5	5	24	24	59	59	5	5	46	46	74	74
18/23/27	5	5	33	39	69	77	5	5	64	69	97	97	5	5	25	25	61	61	5	5	48	48	76	76
19/24/29	5	5	35	41	71	80	5	5	64	66	94	94	5	5	25	25	63	63	5	5	49	49	77	77
20/25/30	5	5	36	41	72	80	5	5	67	72	98	98	5	5	26	26	65	65	5	5	51	51	79	79
21/26/32	5	5	38	45	74	83	5	5	70	74	98	99	5	5	28	28	67	67	5	5	53	53	81	81
22/28/33	5	5	38	45	75	84	5	5	71	75	99	99	5	5	28	28	68	68	5	5	54	54	83	83
23/29/35	5	5	40	46	77	86	5	5	72	75	98	97	5	5	29	29	70	70	5	5	56	56	84	84
24/30/36	5	5	41	47	78	86	5	5	75	78	99	99	5	5	30	30	72	72	5	5	57	57	85	85

Appendix 4a. Empirical Size and Power in Percent: Balanced Four Samples

# of obs in 1st/2nd/3rd/4th Sample	$\theta=0$		$\theta=.1\sigma^*3$		$\theta=.3\sigma+0+0$		$\theta=.2\sigma^*3$		$\theta=.6\sigma+0+0$		$\theta=0$		$\theta=.1\sigma^*3$		$\theta=.3\sigma+0+0$		$\theta=.2\sigma^*3$		$\theta=.6\sigma+0+0$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
	Uniform										Exponential									
5/5/5/5	7	7	17	18	14	15	31	33	24	24	6	5	27	28	22	22	55	56	39	35
6/6/6/6	6	7	17	20	15	16	33	38	26	27	6	6	31	35	25	28	63	67	44	45
7/7/7/7	6	6	19	20	16	17	38	41	29	30	6	6	35	39	29	31	70	73	51	51
8/8/8/8	6	6	19	22	16	18	40	46	31	34	5	5	38	40	30	31	75	76	55	55
9/9/9/9	6	6	21	22	18	19	44	48	34	36	5	5	41	42	33	33	79	81	60	59
10/10/10/10	5	6	21	25	18	20	46	53	36	40	5	5	42	45	34	36	82	84	63	64
11/11/11/11	5	5	23	25	19	20	50	55	39	42	5	5	46	49	37	39	86	88	67	68
12/12/12/12	5	5	25	27	21	22	53	60	42	46	5	5	49	51	39	40	88	89	70	71
13/13/13/13	6	5	27	30	22	24	56	64	45	49	5	5	52	54	41	43	91	92	74	74
14/14/14/14	5	5	27	31	23	25	58	66	46	52	5	5	54	56	43	45	92	93	76	77
15/15/15/15	5	5	29	34	24	27	60	69	48	55	5	5	56	61	44	48	94	95	78	81
16/16/16/16	5	5	29	34	24	27	62	71	50	57	5	5	59	63	47	49	95	96	81	83
17/17/17/17	5	6	31	36	26	29	65	74	53	60	5	5	60	64	48	51	96	97	83	85
18/18/18/18	5	5	32	36	27	29	67	75	54	61	5	5	63	66	50	53	97	98	85	86
19/19/19/19	5	5	34	38	28	31	69	77	56	64	5	5	64	68	52	54	97	98	87	88
20/20/20/20	5	5	34	38	28	31	70	78	58	64	5	5	67	70	54	56	98	98	88	89

Appendix 4a (continued). Empirical Size and Power in Percent: Balanced Four Samples

# of obs in 1st/2nd/3rd/4th Sample	$\theta=0$		$\theta=.1\sigma^*3$		$\theta=.3\sigma+0+0$		$\theta=.2\sigma^*3$		$\theta=.6\sigma+0+0$		$\theta=0$		$\theta=.1\sigma^*3$		$\theta=.3\sigma+0+0$		$\theta=.2\sigma^*3$		$\theta=.6\sigma+0+0$	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
	Normal										Laplace									
5/5/5/5	6	6	13	13	12	12	25	25	22	22	6	6	16	15	14	14	32	32	27	26
6/6/6/6	5	6	14	14	12	12	27	27	23	23	5	5	16	16	14	14	35	35	29	29
7/7/7/7	5	5	14	14	13	13	31	30	26	26	5	5	18	18	16	16	40	39	32	32
8/8/8/8	5	5	15	15	14	13	33	33	28	28	5	5	19	19	17	16	43	43	35	35
9/9/9/9	5	5	16	16	14	14	36	36	30	30	5	5	20	20	18	18	47	47	39	39
10/10/10/10	5	5	17	17	15	15	39	38	33	32	5	5	21	21	19	19	51	51	41	41
11/11/11/11	5	5	18	18	16	16	42	41	35	35	5	5	23	23	20	20	54	54	45	45
12/12/12/12	5	5	19	18	17	17	44	44	37	37	5	5	24	24	21	21	57	57	47	47
13/13/13/13	5	5	20	19	17	17	47	46	39	39	5	5	25	25	22	22	60	60	50	50
14/14/14/14	5	5	20	20	18	18	49	49	42	41	5	5	26	26	23	23	63	63	53	53
15/15/15/15	5	5	21	21	18	18	51	51	43	43	5	5	27	27	23	23	65	65	54	54
16/16/16/16	5	5	22	22	19	19	53	53	45	45	5	5	29	29	25	25	68	68	57	57
17/17/17/17	5	5	23	23	20	20	56	56	48	48	5	5	30	30	26	26	71	71	60	60
18/18/18/18	5	5	24	24	21	21	58	58	50	49	5	5	31	31	27	27	73	73	62	62
19/19/19/19	5	5	24	24	21	21	59	59	50	50	5	5	32	32	27	27	74	74	64	64
20/20/20/20	5	5	25	25	22	22	62	62	53	53	5	5	33	33	29	29	76	76	66	66

Appendix 4b. Empirical Size and Power in Percent: Unbalanced Four Samples

# of obs in 1st/2nd/3rd/4th Sample	Uniform						Exponential						Normal						Laplace					
	θ=0		θ=.1σ*3		θ=.2σ*3		θ=0		θ=.1σ*3		θ=.2σ*3		θ=0		θ=.1σ*3		θ=.2σ*3		θ=0		θ=.1σ*3		θ=.2σ*3	
	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH	HFR	HH
2/2/3/3	6	6	10	11	15	16	6	6	15	15	26	26	7	7	10	11	16	16	7	7	13	13	21	21
3/3/4/5	7	7	15	13	26	21	6	7	22	25	43	46	6	6	13	12	22	21	6	6	14	15	28	27
4/5/6/6	7	6	18	16	32	30	6	6	23	21	44	41	6	6	13	12	25	24	6	5	15	15	32	31
5/6/7/8	6	5	17	17	35	36	6	5	26	26	52	51	5	5	14	14	29	28	5	5	17	17	38	37
6/7/8/9	5	6	17	20	37	42	5	6	35	37	70	72	5	5	14	14	31	31	5	5	18	18	41	40
7/8/10/11	6	6	22	24	45	51	5	5	33	34	66	65	5	5	16	16	36	35	5	5	20	20	47	47
8/9/11/12	6	6	22	25	46	53	5	5	40	45	80	83	5	5	17	17	38	38	5	5	21	21	50	50
9/10/12/14	5	6	23	27	50	57	5	5	44	49	84	87	5	5	18	18	41	41	5	5	23	23	54	54
10/11/14/15	5	6	25	30	54	62	5	5	48	51	88	90	5	5	19	19	44	44	5	5	24	24	58	57
11/12/15/17	5	5	26	30	57	65	5	5	51	55	91	92	5	5	20	20	48	48	5	5	26	26	61	61
12/14/17/18	5	5	28	32	61	69	5	5	52	54	87	87	5	5	21	21	50	50	5	5	27	27	64	64
13/15/18/20	5	6	30	35	64	73	5	5	55	57	90	90	5	5	22	22	53	53	5	5	29	29	68	68
14/16/19/21	5	5	31	35	65	73	5	5	59	63	95	96	5	5	23	23	56	56	5	5	30	30	71	71
15/17/21/23	5	5	32	37	68	77	5	5	61	63	94	93	5	5	23	23	58	58	5	5	31	31	73	73
16/18/22/24	5	5	34	39	69	78	5	5	64	68	97	98	5	5	24	24	60	60	5	5	32	32	75	75

Notes to all tables in Appendices 2, 3 and 4:

All power and size in the tables are presented in percent chances. θ is the location shift between adjacent populations, and σ is the common standard deviation of all populations. By setting $\theta=0$, the empirical frequency of rejecting the null hypothesis is the test's size. To investigate a test's power, for instance, in a two-sample comparison, $\theta = .3\sigma$ indicates the location shift is 30% of the common standard deviation. In the three-sample environment, $\theta = .15\sigma*2$ means the location shift between the first and second, and between the second and third populations are both 15% of the standard deviation. On the other hand, $\theta = .3\sigma+0$ is when the only shift occurs between the first and second populations (30% of the common standard deviation), and the second and third populations are exactly the same. The first column documents the number of observations in each sample, and the rest of the columns are organized based on which distribution the data are generated from.