Subsampling inference with K populations and a nonstandard Behrens-Fisher problem

Timothy L. McMurry* Dimitris N. Politis† Joseph P. Romano‡

October 2011

Abstract

We revisit the methodology and historical development of subsampling, and then explore in detail its use in hypothesis testing, an area which has received surprisingly modest attention. In particular, the general set-up of a possibly high-dimensional parameter with data from K populations is explored. The role of centering the subsampling distribution is highlighted, and it is shown that hypothesis testing with a data-centered subsampling distribution is more powerful. In addition we demonstrate subsampling's ability to handle a nonstandard Behrens-Fisher problem, i.e., a comparison of the means of two or more populations which may possess not only different and possibly infinite variances, but may also possess different distributions. However, our formulation is general, permitting even functional data and/or statistics. Finally, we provide theory for K-sample U-statistics that helps establish the asymptotic validity of subsampling confidence intervals and tests in this very general setting.

Keywords: Confidence intervals, finite population correction, functional statistics, high-dimensional, hypothesis testing, infinite variance, resampling, self-normalized sums, U-statistics.

*Department of Public Health Sciences, University of Virginia, P.O. Box 800717, Charlottesville, VA 22908; email: tmcmurry@virginia.edu

†Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112; email: dpolitis@ucsd.edu

‡Departments of Statistics and Economics, Stanford University, Stanford, CA 94305-4065; email: romano@stanford.edu

1 Historical background and introduction

1.1 Looking back

The seeds of the resampling and subsampling methodologies can be traced back more than 50 years. Tukey's (1958) jackknife was designed to be a rough tool for inference, and, in particular, for assessing bias and variance in i.i.d. (independent and identically distributed) samples. It had been preceeded by two proposals in the same spirit by Quenouille (1949, 1956) that even foreshadowed some current blocking ideas in time series analysis.

Perhaps due to its being an idea ahead of its time, and also because of its failure to capture the variance of a bread-and-butter statistic such as the sample median, the jackknife initially did not receive the attention it deserved. Miller (1974) gives a review of the state-of-the-art in the mid-1970s. It was in attempting to explain the inner workings of the jackknife that Efron (1979) developed the bootstrap; its immediate success was due to (a) its generality and in particular its ability to handle complicated statistics, including the sample median, and (b) its higher-order accuracy in approximating the distribution of (approximately) linear statistics, which is usually obtained by using studentized statistics (in order to achieve asymptotic pivotality). These developments are well documented in the books by Efron (1982), Hall (1992), Efron and Tibshirani (1993), and Davison and Hinkley (1997).

Going beyond the estimation of bias and variance became a research priority in the 1980's, and the study of bootstrap confidence intervals gave particularly fruitful results; see e.g., Hall (1986), Efron and Tibshirani (1986), and DiCiccio and Romano (1988). Around the same time, Wu (1986) and Shao and Wu (1989) introduced the "delete-d" jackknife which, as opposed to Tukey's delete-1 jackknife, was successful in approximating the distribution of the sample median and other approximately linear statistics; see Shao and Tu (1995) Ch. 2 for more details.

There were parallel developments in time series analysis. Most prominent were the (non-overlapping) subseries method of Carlstein (1986), which was quite akin to Quenouille's (1956) proposal, and the well-known block bootstrap of Künsch (1989) that was independently studied by Liu and Singh (1992).

The combination of these developments paved the way for the general subsampling methodology in Politis and Romano (1994), which is applicable to both i.i.d. and dependent data. In contrast to bootstrap validity, which has to be shown on a case-by-case basis, Politis and Romano (1994) showed that subsampling is generally consistent in estimating the distribution of arbitrary statistics under minimal assumptions. The monograph by Politis, Romano and Wolf (1999) describes the foundation of subsampling, including higher-order topics.

We briefly outline the subsampling procedure in detail with i.i.d. data $X_1, \ldots, X_n$. Subsampling proceeds by selecting subsamples of size b from the original sample, and recomputing the statistic in question on the chosen subsamples. As such, subsampling is intimately related to the delete-d jackknife (with $d = n - b$. The major difference between them is that the delete-d jackknife focuses on variance estimation while subsampling focuses on sampling distributions; the former typically requires strong assumptions such as uniform integrability.

The important realization behind subsampling was that by recomputing the statistic over subsamples, one can approximate the entire sampling distribution under very weak assumptions. The reason for subsampling's great generality is because every time the statistic is recomputed over a subsample, the recomputed statistic represents an observation from the true sampling distribution of the statistic (based on a sample of size b) since each subsample is a genuine sample of size b from the true distribution. In contrast, the bootstrap recomputes a statistic over "pseudo-samples" which come from an estimated distribution, rather than the true distribution.

Subsampling is clearly also related to the the bootstrap with resample size b smaller than n, referred to as the b out of n bootstrap; see e.g., Bickel, Götze and van Zwet (1997). If the b is of smaller order than $\sqrt{n}$, then the probability of any observation occurring more than once in a given bootstrap sample tends to 0. Thus, the b out of n bootstrap becomes practically indistinguishable from subsampling and inherits subsampling's general validity. Conversely, if b is of larger order than $\sqrt{n}$, then there are relatively straightforward examples where subsampling remains consistent but the b out of n bootstrap fails; see Politis and Romano (1993), or Section 2.3 of Politis, Romano and Wolf (1999).

1.2 Looking forward

The aforementioned developments in resampling and subsampling proceeded from the modest goals of bias and variance estimation to the estimation of sampling distributions and the construction of confidence intervals. The related problem of hypothesis testing via bootstrap and/or subsampling contains many intricacies, and has received much less attention.

For example, in the simple setting of linear regression with i.i.d. errors, resampling the residuals has been a standard method to obtain confidence intervals for the regression parameters; see e.g., Efron and Tibshirani (1993) Ch. 9, and the references therein. If instead, one wishes to test the significance of the regression coefficient, it is not clear if the appropriate residuals are those obtained under the null hypothesis or under the unrestricted alternative. The matter was recently resolved, at least in the case of unstudentized test statistics, when Paparoditis and Politis (2005) showed that using unrestricted residuals leads to asymptotically more powerful tests against one sided alternatives. The same might be expected to hold for studentized test statistics, but a formal proof is lacking and opinions remain divided; see e.g., the discussion in MacKinnon (2011) and the references therein.

Even with i.i.d. data the situation is not entirely clear. Hall and Wilson (1991) gave some intuitive guidelines on bootstrap hypothesis testing including the recommendation to use asymptotically pivotal test statistics in the bootstrap domain. In contrast to the aforementioned potential accuracy associated with using an asymptotically pivotal statistic for distribution estimation, the recommended pivotality of a bootstrap test statistic serves a much more fundamental role: it allows the bootstrap to correctly estimate the critical region even when the data at hand do not satisfy the null hypothesis. The objective, of course, is better power of the bootstrap test; this recommendation is very intuitive but again a general proof seems elusive.

Nevertheless, as the paper at hand will explore, the matter is much clearer with subsampling and/or bootstrap with a smaller resample size. The test statistic does not need to be fully pivotal in that its asymptotic distribution does not have to be free of all unknown parameters. However, it must be free of the parameter that is being tested else the estimated threshold of the test will diverge; see Section 3 in what follows.

1.3 Present results

In the present work we generalize and extend the testing framework developed in Ch. 2.6 of Politis et al. (1999) in ways we outline below, and fully develop in subsequent sections. Although the practitioner can always perform a hypothesis test by inverting a confidence interval (or bound), classical tests based on the null distribution of a test statistic are often used in practice and are therefore of substantial interest. One great appeal of subsampling tests is that they can be performed with the same relative ease as construction of confidence intervals.

As a primary extension, we fully develop the setting of two (or more) i.i.d. samples, which has

hitherto been largely neglected; the sole exception appears to be the conference announcement (extended abstract) of Politis and Romano (2008). Nonetheless, there are a variety of applications in which subsampling might be useful. A particular example is a nonstandard Behrens-Fisher problem, i.e., a comparison of the means of two or more populations that not only possess different variances but may also possess different distributions; even the case of populations having infinite variance can be successfully addressed via subsampling.

The K-sample case is explored in great generality while at the same time providing additional general insight into subsampling tests; in Section 3, we revisit the framework of Ch. 2.6 of Politis et al. (1999) and show that it is possible to further weaken the already rather minimal assumptions; for example, consistency of the test statistic is shown to not be necessary for the consistency of the subsampling hypothesis test–a surprising fact.

Furthermore, a more careful look at the subsampling tests suggests a modification in order to improve the test's power that was mentioned at the end of the previous subsection. This modification amounts to using a 'data-centered' subsampling distribution similar to the one used for confidence interval construction. In Section 3, we show that under standard assumptions (that *do* include consistency of the test statistic), the test based on the 'data-centered' subsampling distribution has better power than the originally proposed (already consistent) basic subsampling test procedure.

In addition, we make the case that the test statistic need not be fully pivotal in that its asymptotic distribution may depend on unknown parameters; it must be only be free of the parameter of interest, or else the estimated threshold of the test will diverge. Section 3.2 discusses this phenomenon in detail, and gives a formal proof of the resulting better (or, at least, not worse) power associated with data-centering and the pivotality of the test statistic.

Using a 'data-centered' subsampling distribution for testing has been previously considered by Choi (2005) and Andrews and Guggenberger (2010) but without proof of its superiority in terms of improving power. Other recent references on subsampling-based testing, particularly in the time-series and econometrics literature, include Chen and Hsiao (2010), Choi and Chue (2007), Choi and Saikkonen (2010), Delgado, Rodriguez-Poo and Wolf (2001), and Gonzalo and Wolf (2005).

The fundamental ideas in this paper (and in particular our consistency proofs based on an exponential inequality for K-sample U-statistics) can form the basis for applying subsampling to new problems involving multiple comparisons of K populations. Subsampling has been successfully applied to multiple comparisons problems in the one sample setting; see Romano, Shaikh and Wolf (2008) and Romano and Wolf (2010).

It is to be noted that the present asymptotic results are, as is customary, to be interpreted in a pointwise fashion with respect to the underlying probability mechanism. Recently, uniform asymptotic results have been considered by Mikusheva (2007) and Andrews and Guggenberger (2009, 2010). In particular, they consider nonregular situations where subsampling may or may not provide uniform asymptotic validity. The subject of uniform asymptotic approximation is beyond the present scope, but we note that positive uniform results for subsampling and the bootstrap have been recently obtained by Romano and Shaikh (2010). It is expected that in the finite variance case some of these results will carry over to the present context.

The lay-out of the paper is as follows: Section 2 contains the framework of our set-up that includes some groundwork for correct interpretation of subsampling/resampling hypothesis tests. Section 3 presents the basic K-sample subsampling methodology for the construction of hypothesis tests and confidence intervals that have asymptotic validity under minimal assumptions. The consistency proofs are greatly facilitated by a new exponential inequality for K-sample U-statistics that is given in the Appendix extending some well-known results of Hoeffding (1963). Also in Section 3, the use of the data-centered subsampling distribution is suggested for hypothesis testing

as it shown to yield a more powerful test. Section 4 focuses in detail on the aforementioned Behrens-Fisher problem in standard and nonstandard set-ups, and gives some guidelines on block size choice. Finally, Section 5 presents the findings of a simulation experiment.

2 Basic framework

2.1 Subsampling framework

Consider K independent datasets: $\underline{X}^{(1)}, \ldots, \underline{X}^{(K)}$, where $\underline{X}^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$. For each k, the random variables $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ are assumed i.i.d. taking values in an arbitrary space $\mathbf{B}_k$; typically, $\mathbf{B}_k$ would be $\mathbf{R}^{d_k}$ for some d_k, but $\mathbf{B}_k$ can very well be a function space. An example with $K = 2$ is the usual two-sample set-up in biostatistics where a number of 'features' (body characteristics, gene expressions, etc.) are measured on a group of patients, and then again measured on a control group.

The probability law associated with this K-sample experiment is specified by $P = (P^{(1)}, \ldots, P^{(K)})$, where $P^{(k)}$ is the underlying probability law of the kth sample; more formally, the joint distribution of all the observations is the product measure $\prod_{k=1}^{K} (P^{(k)})^{n_k}$. A given model assumes P belongs to some family $\mathbf{P}$= $(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(K)})$ where $\mathbf{P}^{(k)}$ denotes the family of probability distributions for the kth sample; such a family may be nonparametric or parametric. Our goal is inference (confidence regions, hypothesis tests, etc.) regarding a parameter $\theta = \theta(P)$ which–in the case of difference of means–can be thought to be real-valued. However, extensions to multivariate parameters or parameters taking values in a function space are straightforward in view of the results in Politis et al. (1999). In fact, the results can apply when testing hypotheses not explicitly concerned with a particular real-valued parameter (such as testing equality of two distributions).

Denote $\mathbf{n} = (n_1, \ldots, n_K)$, and let $\hat{\theta}_{\mathbf{n}} = \hat{\theta}_{\mathbf{n}}(\underline{X}^{(1)}, \ldots, \underline{X}^{(K)})$ be an estimator of $\theta(P)$ that is a parameter taking values in a general linear space Θ. Let g be a real-valued function on Θ such that $g(x) = 0$ when $x = 0$; sometimes g may also satisfy the identifying property, i.e., $g(x) = 0$ if and only if $x = 0$, but we do not explicitly assume this here. Finally, let $J_{\mathbf{n}}(P)$ denote the sampling distribution of the *"root"* $T_{\mathbf{n}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}}$ under P, with corresponding cumulative distribution function (c.d.f.)

$$J_{\mathbf{n}}(x, P) = Prob_P\{\tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}} \leq x\}; \tag{1}$$

here $\hat{\sigma}_{\mathbf{n}} = \hat{\sigma}_{\mathbf{n}}(\underline{X}^{(1)}, \ldots, \underline{X}^{(K)})$ is a (nonnegative) estimator of scale, and $\tau_{\mathbf{n}}$ is a normalizing sequence that is a known function of $\mathbf{n}$. As will be discussed later on, $\tau_{\mathbf{n}}$ can also be a random variable, i.e., a statistic based on the data $\underline{X}^{(1)}, \ldots, \underline{X}^{(K)}$.

When $\theta(P)$ is real-valued, the function g will be taken either as $g(x) = x$ or $g(x) = |x|$; the first choice leads to one-sided or equal-tailed confidence intervals and tests, while the latter leads to two-sided, symmetric confidence intervals and tests. If the linear space Θ is a higher (or infinite) dimensional metric space, then $g(x)$ can taken as a norm of x; other choices are also possible–see e.g., Politis and Romano (2008). Note that one could instead formulate the results in terms of the distribution of $g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))]/\hat{\sigma}_{\mathbf{n}}$, i.e., having the normalizing sequence $\tau_{\mathbf{n}}$ on the inside of g. The latter approach would work just as well, due to the versatility of our arguments.

A basic assumption for asymptotic inference is existence of a large-sample distribution, i.e.,

Assumption 2.1. *There exists a nondegenerate limiting law $J(P)$ such that $J_{\mathbf{n}}(P)$ converges weakly to $J(P)$ as $\min_k n_k \to \infty$.*

The c.d.f.'s corresponding to the limit law $J(P)$ will be denoted by $J(x, P)$. In case the limit law $J(P)$ does *not* depend on P, then the root $T_{\mathbf{n}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}}$ is called an (asymptotic) *"pivot"*.

Note that with the above general framework we manage to treat in a unified way three important cases:

A. The *unstudentized* case in which $\hat{\sigma}_{\mathbf{n}} \equiv 1$.

B. The *studentized* case in which
$$\hat{\sigma}_{\mathbf{n}} \xrightarrow{P} \sigma(P) > 0 \ , \tag{2}$$
where $\sigma(P)$ is a nonrandom positive constant that may depend on P.

C. The *self-normalized* case in which $\hat{\sigma}_{\mathbf{n}}$ does not necessarily converge to a finite number.

More details on the last case will be given in Section 3.

2.2 Basic framework for hypothesis testing

Consider the setup of testing a null hypothesis $H_0 : P \in \mathbf{P_0}$ versus an alternative $H_1 : P \in \mathbf{P_1}$ where $\mathbf{P_0}$ and $\mathbf{P_1}$ are two families of distributions for the K-sample set-up; let $\mathbf{P} = \mathbf{P_0} \cup \mathbf{P_1}$ be the underlying family of all distributions under consideration. Suppose it is desired to test H_0 based on a test statistic that takes the form $\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}}$. Here $t_{\mathbf{n}} = t_{\mathbf{n}}(\underline{X}^{(1)}, \ldots, \underline{X}^{(K)})$ is a general test statistic sequence which need not concern inference about a parameter. When inference is concerned with tests of a parameter $\theta(P)$ such that $\theta(P) = \theta_0$ if $P \in \mathbf{P_0}$, we will often take $t_{\mathbf{n}}$ to be of the form $g(\hat{\theta}_{\mathbf{n}} - \theta_0)$; this choice is employed in Section 3 for simplicity, although other options are also available.

Next, we define the distribution
$$G_{\mathbf{n}}(x, P) = Prob_P\{\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}} \leq x\}, \tag{3}$$
and require existence of its limit under the null.

Assumption 2.2. *There exists a nondegenerate limiting law $G(P)$ such that $G_{\mathbf{n}}(P)$ converges weakly to $G(P)$ as $\min_k n_k \to \infty$.*

The c.d.f. corresponding to the limit law $G(P)$ will be denoted by $G(x, P)$. Of course, when $t_{\mathbf{n}}$ is of the form $g(\hat{\theta}_{\mathbf{n}} - \theta_0)$, then $G_{\mathbf{n}}(P)$ and $J_{\mathbf{n}}$ are identical if and only if $\theta(P) = \theta_0$. Assumption 2.2 will typically be invoked under the null hypothesis, i.e., for $P \in \mathbf{P_0}$.

For the sake of discussion, first consider testing the *simple* null hypothesis $\bar{H}_0 : P = P_0$ (for some $P_0 \in \mathbf{P_0}$) against the aforementioned alternative $H_1 : P \in \mathbf{P_1}$ using $t_{\mathbf{n}}$ as a test statistic. An asymptotically valid α-level test can then be constructed if the value of the α-quantile of $G(P_0)$, denoted by $G^{-1}(\alpha, P_0) = \inf\{x : G(x, P_0) \geq \alpha\}$, is known exactly or approximately (e.g., via a Monte Carlo simulation under the null). Alternatively, the test can be based on a consistent estimator of the *critical value* $G^{-1}(\alpha, P_0)$; in the next section, such consistent estimators of $G^{-1}(\alpha, P_0)$ will be constructed based on subsampling. We will also study the behavior of estimated critical values under the alternative, and the power of the resulting tests.

In general, however, the null hypothesis of interest $H_0 : P \in \mathbf{P_0}$ will be composite; this is especially true in nonparametric setups where subsampling (or resampling) is most useful. In such situations, a conservative approach is to base the test on the *worst-case scenario* critical value. For example, with a one-sided test of $H_0 : P = P_0$ that rejects when the test statistic is less

than $G^{-1}(\alpha, P_0)$, rejection of $H_0 : P \in \mathbf{P_0}$ at level α would occur when the test statistic is less than $\inf_{P\in\mathbf{P_0}} G^{-1}(\alpha, P)$. Indeed, if the infimum over $P \in \mathbf{P}_0$ is taken on when $P = P'$, then this amounts to obtaining the critical value from the sampling distribution of the test statistic $t_{\mathbf{n}}$ under the "least favorable" distribution P'.

However, such a recommendation is (a) overly conservative resulting into size overestimation and reduced power, and (b) impractical to implement using modern resampling/subsampling methods. The saving point is that resampling and subsampling employ, implicitly or explicitly, a consistent estimator of the underlying true distribution P; see e.g. Beran (1986) and Bickel and Ren (2001). Hence, in the literature of hypothesis testing via resampling or subsampling the α-level test of $H_0 : P \in \mathbf{P_0}$ would be based on a critical value described in the two cases below:

1. If the data have distribution $P_0 \in \mathbf{P_0}$, then the critical value would be tantamount to an estimate of $G^{-1}(\alpha, \hat{P}_0)$ where $\hat{P}_0$ is the implicit or explicit estimator of P_0 built-in the resampling or subsampling procedure; see Remark 3.3 in what follows. In this case, i.e., with data from a distribution satisfying the null, the resampling/subsampling test of the composite H_0 works exactly like the test of the point null $\bar{H}_0$; this analogy is helpful for the intuitive understanding of these tests and will be exploited in the next Section.

2. If the data have distribution $P_1 \in \mathbf{P_1}$, then things are more complicated. The critical values generated by resampling/subsampling do not necessarily converge to a well-defined entity but often the resulting tests are consistent, i.e. their power tends to one. A detailed analysis of the subsampling critical values based on data from the alternative hypothesis is given in Section 3.2; the behavior of resampling critical values under the alternative is similar.

3 K-sample subsampling

In this section, we focus on testing $H_0 : \theta(P) = \theta_0$ under a one-sided or two-sided alternative H_1. In other words, testing $H_0 : P \in \mathbf{P_0}$ vs. $H_1 : P \in \mathbf{P_1}$ where $\mathbf{P_0} = \{\text{all } P \in \mathbf{P} \text{ such that } \theta(P) = \theta_0\}$ and an analogous definition for $\mathbf{P_1}$; see Table 1 for the three cases. Here the underlying family $\mathbf{P}$ can be thought of as all K-sample distributions P for which the quantity $\theta(P)$ is well-defined, possibly with further restrictions (such as moment conditions, etc.). Our test will be based on the test statistic $t_{\mathbf{n}}$ and, in particular, on its studentized version $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}}$; employing the choice $\hat{\sigma}_{\mathbf{n}} = 1$ would result back to the unstudentized case.

For concreteness, throughout this section, the simple specification $t_{\mathbf{n}} = g(\hat{\theta}_{\mathbf{n}} - \theta_0)$ will be used that leads to inference based on the studentized test statistic $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$; recall that the notation $T_{\mathbf{n}}$ is reserved for the more general statistic $T_{\mathbf{n}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}}$.

3.1 Subsampling-based hypothesis testing

For $k = 1, \ldots, K$, let $\mathcal{S}_k$ denote the set of all size b_k (unordered) subsets of the dataset $\{X_1^{(k)}, \ldots, X_{n_k}^{(k)}\}$ where b_k is an integer in $[1, n_k]$. Note that the set $\mathcal{S}_k$ contains $Q_k = \binom{n_k}{b_k}$ elements that can be ordered and enumerated as $S_1^{(k)}, S_2^{(k)}, \ldots, S_{Q_k}^{(k)}$. For reasons to be apparent later, we define the *first* subsample of each sample to be the truncated sample, i.e., let $S_1^{(k)} = \{X_1^{(k)}, \ldots, X_{b_k}^{(k)}\}$.

A K-fold subsample is then constructed by choosing one element from each super-set $\mathcal{S}_k$ for $k = 1, \ldots, K$. Thus, a typical K-fold subsample has the form: $S_{i_1}^{(1)}, S_{i_2}^{(2)}, \ldots, S_{i_K}^{(K)}$, where i_k is an integer in $[1, Q_k]$ for $k = 1, \ldots, K$. It is apparent that the number of possible K-fold subsamples is $Q =$

$\prod_{k=1}^{K} Q_k$. So a subsample value of statistic $\hat{\theta}_{\mathbf{n}}$ is $\hat{\theta}_{\mathbf{i},\mathbf{b}} = \hat{\theta}_{\mathbf{b}}(S_{i_1}^{(1)}, \ldots, S_{i_K}^{(K)})$ where $\mathbf{b} = (b_1, \ldots, b_K)$ and $\mathbf{i} = (i_1, \ldots, i_K)$. Similarly, let $t_{\mathbf{i},\mathbf{b}} = t_{\mathbf{b}}(S_{i_1}^{(1)}, \ldots, S_{i_K}^{(K)})$ and $\hat{\sigma}_{\mathbf{i},\mathbf{b}} = \hat{\sigma}_{\mathbf{b}}(S_{i_1}^{(1)}, \ldots, S_{i_K}^{(K)})$.

An approximation to $G_{\mathbf{n}}(P)$ based on subsampling the statistic $T_{\mathbf{n},\mathbf{0}}$ can be defined as

$$G_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{Q} \sum_{i_1=1}^{Q_1} \sum_{i_2=1}^{Q_2} \cdots \sum_{i_K=1}^{Q_K} 1\{\tau_{\mathbf{b}} t_{\mathbf{i},\mathbf{b}} / \hat{\sigma}_{\mathbf{i},\mathbf{b}} \leq x\} = \frac{1}{Q} \sum_{i_1=1}^{Q_1} \sum_{i_2=1}^{Q_2} \cdots \sum_{i_K=1}^{Q_K} 1\{\tau_{\mathbf{b}} g(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \theta_0) / \hat{\sigma}_{\mathbf{i},\mathbf{b}} \leq x\} \tag{4}$$

analogously to the one-sample set-up of Politis and Romano (1996). We will call $G_{\mathbf{n},\mathbf{b}}$ a *null-based* (N-B) subsampling distribution since the null value of θ_0 is explicitly used for centering.

The following theorem shows that the α-quantile of $G_{\mathbf{n},\mathbf{b}}(x)$ denoted by $G_{\mathbf{n},\mathbf{b}}^{-1}(\alpha) = \inf\{x : G_{\mathbf{n},\mathbf{b}}(x) \geq \alpha\}$ converges to $G^{-1}(\alpha, P_0)$ when data are generated from a distribution $P_0 \in \mathbf{P_0}$. Therefore, the use of $G_{\mathbf{n},\mathbf{b}}^{-1}(\alpha)$ in place of the threshold $G^{-1}(\alpha, P_0)$ will lead to an asymptotically valid α-level test of H_0. Our subsampling theorems typically require the following condition on block sizes:

$$\max_k (b_k/n_k) \to 0 \quad \textit{and} \quad \min_k b_k \to \infty \quad \textit{as} \quad \min_k n_k \to \infty. \tag{5}$$

Theorem 3.1. *Assume that $P = P_0$ where P_0 is some distribution satisfying H_0. Suppose Assumption 2.2 holds, and that $G(\cdot, P_0)$ is continuous and strictly increasing at $G^{-1}(1-\alpha, P_0)$. Under (5), it follows that:*

$$G_{\mathbf{n},\mathbf{b}}^{-1}(1-\alpha) \xrightarrow{P} G^{-1}(1-\alpha, P_0) \tag{6}$$

and

$$Prob_{P_0}\{T_{\mathbf{n},0} > G_{\mathbf{n},\mathbf{b}}^{-1}(1-\alpha)\} \to \alpha. \tag{7}$$

Proof: Let x be a continuity point of $G(\cdot, P_0)$. By Assumption 2.1, $G_{\mathbf{n}}(x, P_0) \to G(x, P_0)$, and $G_{\mathbf{b}}(x, P_0) \to G(x, P_0)$ as well. Since $EG_{\mathbf{n},\mathbf{b}}(x) = G_{\mathbf{b}}(x, P_0)$, it follows that $EG_{\mathbf{n},\mathbf{b}}(x) \to G(x, P_0)$. But note that $G_{\mathbf{n},\mathbf{b}}(x)$ is a K-sample U-statistic with bounded kernel; hence, Lemma 5.1 (iv) implies

$$G_{\mathbf{n},\mathbf{b}}(x) \xrightarrow{P} G(x, P_0) \ . \tag{8}$$

Eq. (6) now follows from Lemma 1.2.1 of Politis et al. (1999). □

Remark 3.1. As mentioned in Section 2.2, the test statistic $t_{\mathbf{n}}$ does not necessarily have to be of the form $g(\hat{\theta}_{\mathbf{n}} - \theta_0)$–or its studentized version–for subsampling to work. With a general test statistic $t_{\mathbf{n}}$, the construction of the subsampling distribution would be as given in the left-hand side of eq. (4), and the proof of Theorem 3.1 would be identical. We choose, however, to focus on the popular test statistic $t_{\mathbf{n}} = g(\hat{\theta}_{\mathbf{n}} - \theta_0)$ in this Section in order to bring out the importance of the choice of centering in the subsampling distribution.

It is apparent that our formulation and arguments are general enough so that $\theta(P)$ need not be real-valued. However, to fix ideas, Table 1 helps delineate the nature of the tests involved when $\theta(P)$ is real-valued. The last column of Table 1 is for a symmetric test, i.e., the function g is the absolute value (or norm) on Θ. However, other possibilities exist as well; for example, in the real-valued case and testing two-sided alternatives with $g(x) = x$, then an equal-tailed test can be constructed by rejecting if $T_{\mathbf{n},\mathbf{0}} > G_{\mathbf{n},\mathbf{b}}^{-1}(1-\alpha/2)$ *or* $T_{\mathbf{n},\mathbf{0}} < G_{\mathbf{n},\mathbf{b}}^{-1}(\alpha/2)$. Notably, the equal-tailed test α-level test of $H_0 : \theta(P) = \theta_0$ amounts to putting together two one-sided tests, i.e., writing the alternative as $H_1 : \theta(P) > \theta_0$ or $\theta(P) < \theta_0$, and rejecting when either one-sided test is significant at

the $\alpha/2$ level. Interestingly, in this case, each of the two one-sided tests could be based on its *own* choice of g function, i.e., with a different definition of the $T_{\mathbf{n},\mathbf{0}}$ statistic; the details are obvious.

H_1 :	$\theta(P) < \theta_0$	$\theta(P) > \theta_0$	$\theta(P) \neq \theta_0$
Asymp. R.R.	$T_{\mathbf{n},\mathbf{0}} < G^{-1}(\alpha, P_0)$	$T_{\mathbf{n},\mathbf{0}} > G^{-1}(1-\alpha, P_0)$	$T_{\mathbf{n},\mathbf{0}} > G^{-1}(1-\alpha, P_0)$
N-B Sub. R.R.	$T_{\mathbf{n},\mathbf{0}} < G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$	$T_{\mathbf{n},\mathbf{0}} > G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$	$T_{\mathbf{n},\mathbf{0}} > G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$

Table 1. Rejection regions (R.R.) of the α-level test of $H_0 : \theta(P) = \theta_0$ vs. H_1 based on the asymptotic $G(x, P_0)$ or the null-based (N-B) subsampling distribution $G_{\mathbf{n},\mathbf{b}}(x)$ respectively; here, $\theta(P)$ is assumed real-valued, and $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$ with $g(x) = x$ in the one-sided alternatives and $g(x) = |x|$ in the two-sided alternative (last column).

Remark 3.2. Note that even if $G(\cdot, P_0)$ is not strictly increasing at $G^{-1}(1-\alpha, P_0)$ but the continuity assumption still holds, one can still deduce the convergence of the rejection probability, e.g., in the case of $H_1 : \theta(P) > \theta_0$ or $H_1 : \theta(P) \neq \theta_0$ we still have eq. (7) holding true. The argument follows as in the one-sample case; see eq. (2.3) of Politis et al. (1999).

Remark 3.3. The hypothesis H_0 is, in general, composite, and there will typically be many probability laws satisfying it; the above P_0 is just any one of them. Theorem 3.1 shows that the subsampling test threshold $G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$ will be close to the asymptotic threshold corresponding to the *particular* P_0 governing the data, assuming H_0 is true. We will soon study the behavior under a fixed alternative P_1 as well as under contiguous sequences.

Remark 3.4. If confidence regions for $\theta(P)$ are desired, they are readily obtained by inverting the aforementioned test of H_0, and inherit their general asymptotic validity from eq. (6). To elaborate, a $(1-\alpha)100\%$ confidence region for θ would consist of all the values of θ_0 that–when conducting the above α-level test of $H_0 : \theta(P) = \theta_0$–result in 'acceptance" (non-rejection) of H_0. For example, inverting the two-sided test (last column of Table 1) gives a two-sided, equal-tailed $(1-\alpha)100\%$ confidence interval; this interval is similar but not identical to the one given by the 'direct' construction of subsampling confidence region in Theorem 2.2.1 of Politis et al. (1999).

Theorem 3.1 shows that the proposed subsampling tests have the correct size asymptotically. The following theorem further shows that the subsampling tests have nontrivial asymptotic power under local alternatives; for the definition of contiguity, see e.g., Lehmann and Romano (2005), p.494. We will later show that under fixed alternatives, the power tends to one.

Theorem 3.2. *Suppose that for some $P = (P_1, \ldots, P_k)$ satisfying H_0, the product measure $P_{k,n_k}^{n_k}$ is contiguous to $P_k^{n_k}$ for $k = 1, \ldots K$. Assume (5), and suppose Assumption 2.2 holds. Then, under such a contiguous sequence, $\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}}$ is tight. Moreover, if it converges in distribution to some random variable T and $G(\cdot, P)$ is continuous at $G^{-1}(1-\alpha, P)$, then the limiting power of the test which rejects when $\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}} > G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$ against such a sequence is $P\{T > G^{-1}(1-\alpha, P)\}$.*

Proof: If we further assume that $G(\cdot, P)$ is strictly increasing at $G^{-1}(\cdot, P)$, then from Theorem 3.1 we know that $G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha) \xrightarrow{P} G^{-1}(1-\alpha, P)$ under P. Contiguity forces the same convergence under the sequence of contiguous alternatives. The result follows by Slutsky's Theorem.

Without the strictly increasing assumption, the result still follows. To see why, under P, $G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$ need not converge but has limiting probability 0 of falling outside $[a-\epsilon, b+\epsilon]$ where the interval $[a, b]$ is defined by $\{x : \ J(x, P) = 1-\alpha\}$, and ϵ is an arbitrary positive number. The

same must be true under a contiguous sequence. Thus, $G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$ is bounded above by $b+\epsilon$ in probability, in which case (by a variation in Slutsky's theorem)

$$P\{\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}} > G^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)\} \geq P\{\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}} > b+\epsilon\} + o(1) \ .$$

Letting the sample sizes increase, and then letting $\epsilon \to 0$ we see that the right-hand side tends to the stated expression because b is a $1-\alpha$ quantile of $J(\cdot, P)$ and $J(\cdot, P)$ is continuous at b. A lower bound replacing b with a is obtained similarly. □

Theorems 3.1 and 3.2 were shown to hold under minimal assumptions. To appreciate this, note that the statistic $\hat{\theta}_{\mathbf{n}}$ (that is a crucial element of our test statistic) was not even assumed to be consistent for $\theta(P)$, a condition that would be implied in cases A or B of Section 2 by the extra condition $\tau_{\mathbf{n}} \to \infty$ as $\min_k n_k \to \infty$. Of course, consistency of $\hat{\theta}_{\mathbf{n}}$ is highly desirable, and actually needed in order for the subsampling test to be consistent, i.e., having power tending to one under a fixed alternative. We will therefore require an additional assumption that is an analogue of Assumption 11.3.1 of Politis et al. (1999).

Assumption 3.1. *Let $V_n = a_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))$ and $W_n = d_{\mathbf{n}} \hat{\sigma}_{\mathbf{n}}$. Also let $\tau_{\mathbf{n}} = a_{\mathbf{n}}/d_{\mathbf{n}}$ so that $\tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}} = V_n/W_n$. Assume (V_n, W_n) converges weakly to (V, W) where W does not have positive mass at zero. The sequences $a_{\mathbf{n}}, d_{\mathbf{n}}$ are positive, and such that $\min(a_{\mathbf{n}}, a_{\mathbf{n}}/d_{\mathbf{n}}) \to \infty$ when $\min_k n_k \to \infty$.*

Assumption 3.1 covers all three cases A, B and C of Section 2 under one umbrella; for example, case B where $\hat{\sigma}_{\mathbf{n}} \xrightarrow{P} \sigma(P) > 0$ is covered letting $d_{\mathbf{n}} = 1$ and the distribution of W having unit mass at $\sigma(P)$. Assumption 3.1 implies Assumption 2.2 and, when P satisfies H_0, it implies that $J(x, P) \equiv G(x, P)$ and $t_{\mathbf{n}} \equiv g(\hat{\theta}_{\mathbf{n}} - \theta_0) \xrightarrow{P} 0$. Thus, if we let $t(P)$ denote the limit (in probability) of the test statistic $t_{\mathbf{n}}$ (whenever the limit exists), we see that $t(P) = 0$ under the null hypothesis $H_0 : \theta(P) = \theta_0$. The following assumption, delineates the allowed behavior of $t(P)$ under an alternative hypothesis.

Assumption 3.2. *Suppose $\tau_{\mathbf{n}} t_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}} = V_{\mathbf{n}}/W_{\mathbf{n}}$, where $V_{\mathbf{n}} = a_{\mathbf{n}} t_{\mathbf{n}}$ and $W_{\mathbf{n}} = d_{\mathbf{n}} \hat{\sigma}_{\mathbf{n}}$. Assume that $t_{\mathbf{n}} \xrightarrow{P} t(P) \neq 0$, and W_n converges weakly to W where W does not have positive mass at zero. The sequences $a_{\mathbf{n}}, d_{\mathbf{n}}$ are positive, and such that $\min(a_{\mathbf{n}}, a_{\mathbf{n}}/d_{\mathbf{n}}) \to \infty$ when $\min_k n_k \to \infty$.*

Note that by assuming Assumption 3.1 together with Assumption 3.2, the null hypothesis $H_0 : \theta(P) = \theta_0$ becomes *equivalent* to the hypothesis $H_0 : t(P) = 0$. Using these two assumptions, the consistency of the subsampling test against fixed alternatives is proven below.

Theorem 3.3. *Assume (5) and Assumption 3.1. Assume that $P = P_1$ where P_1 is a distribution satisfying the alternative hypothesis to the null hypothesis H_0. Assume P_1 satisfies Assumption 3.2, and further assume (5), and that $a_{\mathbf{b}}/a_{\mathbf{n}} \to 0$. Then, the probability (under P_1) that the subsampling test rejects $H_0 : t(P) = 0$ tends to one.*

Proof. The proof generalizes that of Theorem 2.6.1 of Politis et al. (1999). To fix ideas, consider the case where $t(P_1) > 0$ which happens if the alternative hypothesis H_1 is either the one-sided $\theta(P) > \theta_0$ or the two-sided $\theta(P) \neq \theta_0$; the case of $H_1 : \theta(P) < \theta_0$ is handled similarly.

Rather than considering the subsampling distribution $G_{\mathbf{n},\mathbf{b}}(x)$, we look at the subsampling distribution corresponding to the statistic $t_{\mathbf{n}}/(d_{\mathbf{n}}\hat{\sigma}_{\mathbf{n}})$ without the scale factor $a_{\mathbf{n}}$ in the numerator; that is, let $G^0_{\mathbf{n},\mathbf{b}}(x) = G_{\mathbf{n},\mathbf{b}}(a_{\mathbf{b}} x)$. But

$$EG^0_{\mathbf{n},\mathbf{b}}(x) = Prob_P\{t_{\mathbf{b}}/(d_{\mathbf{b}}\hat{\sigma}_{\mathbf{b}}) \leq x\} \ .$$

When $P = P_1$, the numerator converges to $t(P_1)$ by Assumption 3.2 while the denominator has a limiting distribution with no mass at 0. Thus, the expected value converges to 0 if $x < t(P_1)$ and to 1 if $x > t(P_1)$. But since $\hat{G}^0_{\mathbf{n},\mathbf{b}}(x)$ is a K-sample bounded U-statistic, its variance converges to 0 by Lemma 5.1. Now, by a familiar argument $G^0_{\mathbf{n},\mathbf{b}}(\cdot)$ converges in distribution (with probability tending to 1) to the distribution of $t(P_1)/W$. This implies that the subsampling critical value is of order $a_{\mathbf{b}}$ in probability. But the test statistic in this case is of order $a_{\mathbf{n}}$ in probability. The assumption $a_{\mathbf{n}}/a_{\mathbf{b}} \to \infty$ implies consistency. □

3.2 Hypothesis testing using data-centered subsampling distributions

The subsampling tests of the previous subsection were shown to be asymptotically valid and consistent, i.e., achieving the nominal α-level and having power tending to one for large samples. Interestingly, they manage to do all that despite the fact that the subsampling threshold $G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$ does not converge when the data are generated from an alternative P_1.

To intuitively see the effect on the estimated threshold when the probability law generating the data is P_1, consider testing H_0 vs. $H_1 : \theta(P) = \theta_1$ for some $\theta_1 < \theta_0$ in the unstudentized case $\hat{\sigma}_{\mathbf{n}} \equiv 1$ and take g the identity function. Then, from the U-statistic consistency argument we have–with high probability under (5)–that

$$G_{\mathbf{n},\mathbf{b}}(x) \approx EG_{\mathbf{n},\mathbf{b}}(x) = Prob_{P_1}\{\tau_{\mathbf{b}}(\hat{\theta}_{\mathbf{b}} - \theta_1) + z_{\mathbf{b}} \leq x\} \ \text{ where } \ z_{\mathbf{b}} = \tau_{\mathbf{b}}(\theta_1 - \theta_0),$$

which implies

$$G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha) \approx J^{-1}(\alpha, P_1) + z_{\mathbf{b}}. \tag{9}$$

So, under P_1 the threshold $G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$ diverges to $-\infty$. Nevertheless, the subsampling test is consistent as shown in Theorem 3.3; the reason is that the test statistic $T_{\mathbf{n},\mathbf{0}} \equiv \tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta_0)$ also diverges to $-\infty$ but at a faster rate; to see this, note that $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta_1) + z_{\mathbf{n}} = T_{\mathbf{n}} + z_{\mathbf{n}}$ where $T_{\mathbf{n}} = O_{P_1}(1)$ by Assumption 2.1.

It would be desirable to have a nicer behavior of the subsampling-based estimated threshold when the data are generated from P_1. To motivate the ideas that follow, suppose for a moment that the limit law $J(P)$ of Assumption 2.1 did *not* depend on P, i.e., when the root $T_{\mathbf{n}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta(P))/\hat{\sigma}_{\mathbf{n}}$ is an asymptotic *pivot*, in which case $J^{-1}(\alpha, P_1) = J^{-1}(\alpha, P_0^*)$ where P_0^* is *any* distribution in $\mathbf{P_0}$. Clearly, in this case, $J^{-1}(\alpha, P_0^*)$ is an asymptotically correct critical value. It is apparent that the first term on the RHS of (9) would be correct but the diverging term $z_{\mathbf{b}}$ would unfortunately remain. To remove this unwanted term, one has to construct a test using a subsampling distribution that approximates well the distribution of the data-centered variable $T_{\mathbf{n}}$ that is distinct from the test statistic $T_{\mathbf{n},\mathbf{0}}$ which is centered using the null value of θ.

Thus, we define the data-centered (D-C) subsampling distribution

$$L_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{Q} \sum_{i_1=1}^{Q_1} \sum_{i_2=1}^{Q_2} \cdots \sum_{i_K=1}^{Q_K} 1\{\tau_{\mathbf{b}} g(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \leq x\} \tag{10}$$

which is analogous to the one employed by Politis and Romano (1994) for the construction of subsampling-based confidence intervals in the one-sample case. Choi (2005)

H_1 :	$\theta(P) < \theta_0$	$\theta(P) > \theta_0$	$\theta(P) \neq \theta_0$
D-C Sub. R.R.	$T_{\mathbf{n},\mathbf{0}} < L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$	$T_{\mathbf{n},\mathbf{0}} > L^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$	$T_{\mathbf{n},\mathbf{0}} > L^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)$

Table 2. Entries as in Table 1 but the tests are based on quantiles of the data-centered (D-C) subsampling distribution $L_{\mathbf{n},\mathbf{b}}(x)$.

Table 2 delineates the tests based on the quantiles of $L_{\mathbf{n},\mathbf{b}}(x)$ and should be compared to Table 1. Note that the test statistic is still $T_{\mathbf{n},\mathbf{0}}$; the only difference is that the subsampling approximation to its distribution is based on $L_{\mathbf{n},\mathbf{b}}(x)$ instead of $G_{\mathbf{n},\mathbf{b}}(x)$. Since $L_{\mathbf{n},\mathbf{b}}(x)$ is the subsampling distribution used in the construction of confidence intervals, performing the test as in Table 2 is equivalent to inverting the corresponding $(1-\alpha)100\%$ confidence interval (or confidence bound) constructed via subsampling.

It is apparent, that for $L_{\mathbf{n},\mathbf{b}}(x)$ to work well under the null hypothesis it is necessary that $\hat{\theta}_{\mathbf{n}}$ is *consistent* for $\theta(P)$. For this reason, the assumptions of the following theorems include consistency of $\hat{\theta}_{\mathbf{n}}$ that is ensured by assuming $\tau_{\mathbf{n}} \to \infty$. Theorem 3.4 below can be proven analogously to Theorems 2.2.1 and 2.6.1 in Politis et al., (1999).

Theorem 3.4. *Suppose Assumption 2.1 holds, as well as (5). Assume that $J(\cdot, P)$ is continuous and strictly increasing at $J^{-1}(\alpha, P)$. Also assume $\tau_{\mathbf{b}} \to \infty$ and $\tau_{\mathbf{b}}/\tau_{\mathbf{n}} \to 0$ coupled with either (2) or Assumption 3.1. Then,*

$$L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha) \xrightarrow{P} J^{-1}(\alpha, P). \tag{11}$$

Theorem 3.4 shows that the D-C subsampling quantiles are convergent along all possible options for P. In particular, when the data are generated from a distribution P_0 satisfying H_0, i.e., $P_0 \in \mathbf{P_0}$, Theorem 3.4 shows that D-C subsampling quantile $L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha) \approx J^{-1}(\alpha, P_0) = G^{-1}(\alpha, P_0)$; the last equality follows from the consistency of $\hat{\theta}_{\mathbf{n}}$ that is ensured by assuming $\tau_{\mathbf{n}} \to \infty$ in Theorem 3.4. Therefore, using $L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$ as a critical value leads to asymptotically valid α-level tests of H_0 in the sense described in Section 2.2, i.e., we have

$$Prob_{P_0}\{T_{\mathbf{n},0} > L^{-1}_{\mathbf{n},\mathbf{b}}(1-\alpha)\} \to \alpha. \tag{12}$$

Interestingly, using the D-C subsampling quantiles not only leads to consistent tests, it actually gives tests with *better* power as compared to tests using the null-based subsampling distribution. Assumption 3.3 below focuses on one-sided tests; as previously mentioned, the two-sided equal-tailed α-level test can be performed by combining the two one-sided tests at level $\alpha/2$.

Assumption 3.3. *Consider a real-valued parameter $\theta(P)$, and the test of hypothesis $H_0 : \theta(P) = \theta_0$ based on the test statistic $T^{+}_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}^{+}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$ where g is a real-valued, monotone increasing, function such that $g(0) = 0$.*
(i) If the alternative hypothesis is $\theta(P) > \theta_0$, then let $\hat{\theta}^{+}_{\mathbf{n}} = \max(\hat{\theta}_{\mathbf{n}}, \theta_0)$.
(ii) If the alternative hypothesis is $\theta(P) < \theta_0$, then let $\hat{\theta}^{+}_{\mathbf{n}} = \min(\hat{\theta}_{\mathbf{n}}, \theta_0)$.

Note that using test statistic $T^{+}_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}^{+}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$ instead of the customary $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$ is not restrictive. For example, in the case of alternative $\theta(P) > \theta_0$, a practitioner would not even contemplate rejecting the null hypothesis if the statistic $\hat{\theta}_{\mathbf{n}}$ turned out to be less than θ_0; Assumption 3.3 makes this intuitive notion explicit, and thus enables us to clearly state and prove the following theorem.

Theorem 3.5. *Assume Assumption 3.3, as well as the combined assumptions of Theorem 3.3 and Theorem 3.4. Then, under P_1, the probability that the D-C subsampling test rejects the null hypothesis H_0 is greater than or equal to the probability that the null-based (N-B)subsampling test rejects H_0.*

Proof. To fix ideas, consider the case of Assumption 3.3 (i) where $H_1 : \theta(P) > \theta_0$, and that $g(x) = x$; the other one-sided alternative hypothesis, and the generalization to a monotonic g function are treated similarly.

First note that the distributions $L_{\mathbf{n},\mathbf{b}}$ and $G_{\mathbf{n},\mathbf{b}}$ have the same shape; one is a shifted version of the other (location shift). Since, by construction, $\hat{\theta}_{\mathbf{n}}^{+} \geq \theta_0$, it follows that the mass of the distribution $L_{\mathbf{n},\mathbf{b}}$ is shifted to the left of the mass of the distribution $G_{\mathbf{n},\mathbf{b}}$; in other words, whereas $L_{\mathbf{n},\mathbf{b}}$ is centered around zero, $G_{\mathbf{n},\mathbf{b}}$ has a positive center of location. Therefore, one distribution stochastically dominates the other, i.e., for any x, $L_{\mathbf{n},\mathbf{b}}(x) \geq G_{\mathbf{n},\mathbf{b}}(x)$, and consequently, for any α, $L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha) \leq G^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$.

Therefore, the D-C subsampling rejection region (R.R.) is a superset of the N-B subsampling R.R., and the theorem is proved. $\square$.

Remark 3.5. Even when the customary test statistic $T_{\mathbf{n},\mathbf{0}} = \tau_{\mathbf{n}} g(\hat{\theta}_{\mathbf{n}} - \theta_0)/\hat{\sigma}_{\mathbf{n}}$ is used, Theorem 3.5 gives valuable insights. To be precise, consider the alternative $H_1 : \theta(P) > \theta_0$, and define the event $B_{\mathbf{n}} = \{\hat{\theta}_{\mathbf{n}} > \theta_0\}$. Then, under P_1, the *conditional* (given $B_{\mathbf{n}}$) probability that the D-C subsampling test rejects the null hypothesis H_0 is greater than or equal to the conditional probability that the null-based (N-B)subsampling test rejects H_0. Of course, the weak consistency of $\hat{\theta}_{\mathbf{n}}$ implies that $B_{\mathbf{n}}$ has probability tending to one under P_1, and this convergence is typically exponentially fast; hence conditioning on $B_{\mathbf{n}}$ is not restrictive.

Remark 3.6. Note that data-centered (D-C) subsampling can also be accomplished using a different centering, say $\tilde{\theta}_{\mathbf{n}}$, and the conclusion of the above theorem would remain true as long as $\tilde{\theta}_{\mathbf{n}}$ has the same rate of convergence as $\hat{\theta}_{\mathbf{n}}$, i.e., if $\tau_{\mathbf{n}}(\tilde{\theta}_{\mathbf{n}} - \theta(P)) = O_P(1)$. For example, $\tilde{\theta}_{\mathbf{n}}$ could be taken as the average of the subsample statistics $\hat{\theta}_{\mathbf{i},\mathbf{b}}$.

Andrews and Guggenberger (2010) also considered both D-C and N-B subsampling distributions side-by-side without, however, making a recommendation regarding which one is preferable. The recommendation to use data-based centering has been given previously (without proof) by Hall and Wilson (1991) in the related set-up of bootstrap testing and Choi (2005) in the case of subsampling. Theorem 3.5 confirms that data-based centering is recommended for subsampling as well, and gives credence to the Hall and Wilson (1991) guideline in cases where the bootstrap works. Extensions to bootstrap hypothesis testing in regression were also given by Paparoditis and Politis (2005), and recently in the construction of subsampling p-values by Berg, McMurry and Politis. (2010).

The "no-free lunch" principle, however, applies here as well: the N-B subsampling distribution has an edge of accuracy under the null; in other words, the N-B subsampling distribution will typically yield a test with size closer to the nominal as compared to the D-C subsampling test. Not surprisingly, the two tests (N-B and D-C) have the same local power properties, i.e., same power under alternatives contiguous to the null.

Note that Theorem 3.5 does not imply that subsampling captures the "correct" critical value when the data are generated from a distribution $P_1 \in \mathbf{P_1}$. In this case, Theorem 3.4 suggests that $L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha) \approx J^{-1}(\alpha, P_1)$ which is not necessarily "correct"; the quotes are here because the notion of "correctness" is equivocal since practitioners are understandably reluctant to use the 'worst-case scenario' critical value discussed in Section 2.2. If, however, $T_{\mathbf{n}}$ happens to be an asymptotic pivot, then $J^{-1}(\alpha, P_1) = J^{-1}(\alpha, P_0^*)$ for all $P_0^* \in \mathbf{P_0}$ and (a) the issue of "correct" critical value is resolved, and (b) the D-C subsampling quantile $L^{-1}_{\mathbf{n},\mathbf{b}}(\alpha)$ is the method that captures it.

Interestingly, the full strength of the property of pivotality may not be needed in order to resolve the issue of a "correct" critical value and ensure that the D-C subsampling quantile captures it. What is required, however, is a certain invariance of the limit law $J(P)$ with respect to θ, the

parameter being tested. Instead of attempting a general definition, we illustrate by the simplest example of testing the value θ of the mean of a single sample ($K = 1$) having finite variance σ^2; the test is based on the sample mean $\hat{\theta}_{\mathbf{n}}$. The studentized case leads to full asymptotic pivotality, so we instead focus on the unstudentized case where $\hat{\sigma}_{\mathbf{n}} = 1$, and the limit law $J(P)$ of $\sqrt{n}(\hat{\theta}_{\mathbf{n}} - \theta(P))$ is $N(0, \sigma^2)$. All that is required here for the D-C subsampling quantile to capture the "correct" critical value is that the asymptotic variance σ^2 is the same under both the null and the alternative hypotheses, i.e., the classical notion of *variance stabilization* in that the asymptotic variance σ^2 does not depend on θ. For example, if the data are i.i.d. $N(\theta, \sigma^2)$ everything works fine; but if the data are i.i.d. Exponential with mean θ, then variance stabilization is not seen to hold since the asymptotic variance depends on θ.

Remark 3.7. Since both types of centering, $\hat{\theta}_{\mathbf{n}}$ (D-C) and θ_0 (N-B), for the subsampling distribution seem to work for hypothesis testing purposes, it is apparent that a linear (convex) combination of these two extremes can work too. Thus, the subsampling distribution could be centered at: $(1 - \lambda_{\mathbf{n}})\hat{\theta}_{\mathbf{n}} + \lambda_{\mathbf{n}}\theta_0$ where $\lambda_{\mathbf{n}} \in [0, 1]$ is appropriately chosen (and may even depend on the data). For example, in the case where θ is univariate, and $g(x) = x$ or $|x|$, we may choose $\lambda_{\mathbf{n}} = \mathbf{1}\{|\hat{\theta}_{\mathbf{n}} - \theta_0| < c\tau_{\mathbf{n}}^{-1}\hat{\sigma}_{\mathbf{n}}\}$ where $c > 0$ is some constant (say $c = 0.5$), in an effort to capture both the good power properties of D-C subsampling as well as the good performance under the null of N-B subsampling.

3.3 Estimated rates of convergence

As discussed in Chapter 8 of Politis, et al. (1999), there are cases where the normalizing factor $\tau_{\mathbf{n}}$ is unknown as it may involve some unknown parameters; this is true even in the standard (finite-variance) Behrens-Fisher example of the next section. We thus consider a random variable $\hat{\tau}_{\mathbf{b}}$ that is an estimator of $\tau_{\mathbf{b}}$, and let $\hat{G}_{\mathbf{n},\mathbf{b}}(x)$ and $L_{\mathbf{n},\mathbf{b}}(x)$ denote the two subsampling distributions using $\hat{\tau}_{\mathbf{b}}$ instead of $\tau_{\mathbf{b}}$, i.e.,

$$\hat{G}_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{Q}\sum_{i_1=1}^{Q_1}\sum_{i_2=1}^{Q_2}\cdots\sum_{i_K=1}^{Q_K} 1\{\hat{\tau}_{\mathbf{b}}\, g(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \theta_0)/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \le x\}, \tag{13}$$

$$\hat{L}_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{Q}\sum_{i_1=1}^{Q_1}\sum_{i_2=1}^{Q_2}\cdots\sum_{i_K=1}^{Q_K} 1\{\hat{\tau}_{\mathbf{b}}\, g(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \le x\}. \tag{14}$$

By contrast to $\hat{\theta}_{\mathbf{i},\mathbf{b}}$ and $\hat{\sigma}_{\mathbf{i},\mathbf{b}}$, $\hat{\tau}_{\mathbf{b}}$ does not have to be measurable with respect to the random variables in the set $S_{i_1}^{(1)}, \ldots, S_{i_K}^{(K)}$. In general, $\hat{\tau}_{\mathbf{b}}$ is an estimator of $\tau_{\mathbf{b}}$ based on the *whole* of the available data, i.e., a function of $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ for $k = 1, \ldots, K$. All that is required of it is that

$$\hat{\tau}_{\mathbf{b}}/\tau_{\mathbf{b}} \xrightarrow{P} 1 \tag{15}$$

under eq. (5).

Theorem 3.6. *Assume eq. (15). Then, the theorems of Section 3.1 remain true with $\hat{G}_{\mathbf{n},\mathbf{b}}(x)$ instead of $G_{\mathbf{n},\mathbf{b}}(x)$. Similarly, the theorems of Section 3.2 remain true with $\hat{L}_{\mathbf{n},\mathbf{b}}(x)$ instead of $L_{\mathbf{n},\mathbf{b}}(x)$.*

Proof: Just note that $\hat{G}_{\mathbf{n},\mathbf{b}}(x) = G_{\mathbf{n},\mathbf{b}}(x\tau_{\mathbf{b}}/\hat{\tau}_{\mathbf{b}}) \approx G_{\mathbf{n},\mathbf{b}}(x)$ in view of eq. (15). The above approximation can be made rigorous by an argument similar to the one given in the proof of Theorem 8.3.1 of Politis, et al. (1999). □

3.4 Random subsamples and the K-sample bootstrap

For large values of n_k and b_k, $Q = \prod_k \binom{n_k}{b_k}$ can be a prohibitively large number; considering *all* possible subsamples may be impractical and, thus, we may resort to Monte Carlo. To define the algorithm for generating random subsamples of sizes $b_1, \ldots, b_K$ respectively, recall that subsampling in the i.i.d. single-sample case is tantamount to sampling *without* replacement from the original dataset; see e.g., Politis et al. (1999, Ch. 2.3). Thus, for $m = 1, \ldots, M$, we can generate the mth joint subsample as $\underline{X}_m^{(1)}, \underline{X}_m^{(2)}, \ldots, \underline{X}_m^{(K)}$ where $\underline{X}_m^{(k)} = \{X_{I_{k,1}}^{(k)}, \ldots, X_{I_{k,b_k}}^{(k)}\}$, and $I_{k,1}, \ldots, I_{k,b_k}$ are b_k numbers drawn randomly *without* replacement from the index set $\{1, 2, \ldots, n_k\}$. Note that the random indices drawn to generate $\underline{X}_m^{(k)}$ are independent of those drawn to generate $\underline{X}_m^{(k')}$ for $k \neq k'$.

Thus, a randomly chosen subsample value of the statistic $\hat{\theta}_{\mathbf{n}}$ is given by $\hat{\theta}_{m,\mathbf{b}} = \hat{\theta}_{\mathbf{b}}(\underline{X}_m^{(1)}, \ldots, \underline{X}_m^{(K)})$, with $\hat{\sigma}_{m,\mathbf{b}}$ denoting its estimate of scale computed from the mth joint subsample. The corresponding subsampling distribution under the null is defined as

$$\tilde{G}_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{M} \sum_{m=1}^{M} 1\{\tau_{\mathbf{b}} g(\hat{\theta}_{m,\mathbf{b}} - \theta_0)/\hat{\sigma}_{m,\mathbf{b}} \leq x\}, \tag{16}$$

and its data-centered version is

$$\tilde{L}_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{M} \sum_{m=1}^{M} 1\{\tau_{\mathbf{b}} g(\hat{\theta}_{m,\mathbf{b}} - \hat{\theta}_{\mathbf{n}})/\hat{\sigma}_{m,\mathbf{b}} \leq x\}. \tag{17}$$

The following corollary shows that the above Monte Carlo approximations can be used in place of the subsampling distributions when M is large; its proof is analogous to the proof of Corollary 2.1 of Politis and Romano (1994).

Corollary 3.1. *Assume $M \to \infty$. Then, the theorems of Section 3.1 remain true with $\tilde{G}_{\mathbf{n},\mathbf{b}}(x)$ instead of $G_{\mathbf{n},\mathbf{b}}(x)$. Similarly, the theorems of Section 3.2 remain true with $\tilde{L}_{\mathbf{n},\mathbf{b}}(x)$ instead of $L_{\mathbf{n},\mathbf{b}}(x)$.*

The bootstrap in two-sample settings is often used in practical work; see Hall and Martin (1988) or van der Vaart and Wellner (1996), p.365. In the i.i.d. set-up, resampling and (random) subsampling are very closely related since, as mentioned, they are tantamount to sampling *with* vs. *without* replacement from the given i.i.d. sample. In contrast to subsampling, however, no general validity theorem is available for the bootstrap *unless* a smaller resample size is used; see Politis and Romano (1994).

As in the one-sample case, the general validity of K-sample bootstrap that uses a resample size b_k for sample k follows from the general validity of subsampling as long as $b_k^2 << n_k$. To state it, let $G^*_{\mathbf{n},\mathbf{b}}(x)$ denote the bootstrap (pseudo-empirical) null distribution of $\tau_{\mathbf{b}} g(\hat{\theta}^*_{\mathbf{n},\mathbf{b}} - \theta_0)$ where $\hat{\theta}^*_{\mathbf{n},\mathbf{b}}$ is the statistic $\hat{\theta}_{\mathbf{b}}$ computed from the bootstrap data. Also let $L^*_{\mathbf{n},\mathbf{b}}(x)$ denote the bootstrap (pseudo-empirical) distribution of the pivot $\tau_{\mathbf{b}} g(\hat{\theta}^*_{\mathbf{n},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})$.

The proof of the following corollary is a consequence of the discussion given in Section 2.3 of Politis et al. (1999).

Corollary 3.2. *Assume $b_k^2/n_k \to 0$ for all k. Then, the theorems of Section 3.1 remain true with $G^*_{\mathbf{n},\mathbf{b}}(x)$ instead of $G_{\mathbf{n},\mathbf{b}}(x)$. Similarly, the theorems of Section 3.2 remain true with $L^*_{\mathbf{n},\mathbf{b}}(x)$ instead of $L_{\mathbf{n},\mathbf{b}}(x)$.*

4 The difference of two population means

Consider now the two sample case, i.e., $K = 2$, i.e., where for $i = 1, 2$,

$$X_1^{(i)}, X_2^{(i)}, \ldots, X_{n_i}^{(i)} \sim \text{i.i.d. from c.d.f. } F_i \tag{18}$$

The c.d.f. F_i is assumed to have mean μ_i and variance σ_i^2.

The familiar statistic for testing $H_0 : \theta = 0$ where $\theta = \mu_2 - \mu_1$ is given by $\hat{\theta}_{\mathbf{n}} = \bar{X}^{(2)} - \bar{X}^{(1)}$ where $\bar{X}^{(i)} = n_i^{-1} \sum_{t=1}^{n_i} X_t^{(i)}$. Note that $Var(\hat{\theta}_{\mathbf{n}}) = \sigma_2^2/n_2 + \sigma_1^2/n_1$, and is typically estimated by $S^2 = S_2^2/n_2 + S_1^2/n_1$, where $S_i^2 = (n_i - 1)^{-1} \sum_{t=1}^{n_i} (X_t^{(i)} - \bar{X}^{(i)})^2$.

As is well-known, the distribution of the 'studentized' statistic $(\hat{\theta}_{\mathbf{n}} - \theta)/S$ is not given exactly by the t-distribution even when the samples are normally distributed, nor does it inherit any finite-sample optimality (though it is an asymptotically optimal test); this is the celebrated *Behrens-Fisher* problem. Nevertheless, the t-tables are often used in this connection due to Welch's approximation of the distribution of $(\hat{\theta}_{\mathbf{n}} - \theta)/S$ by a t_k distribution with

$$k = \left(\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \right)^{-1} \quad \text{where} \quad c = S_1^2/(n_1 \hat{S}^2).$$

If indeed the two samples are normal, then the Welch t-approximation is higher-order accurate; see e.g., Pfanzagl (1974) or Beran (1988). Nevertheless, in the words of DasGupta (2008, p. 405): "It is not clear, however, that the Welch test is even size-robust when the individual groups are not normal."

Subsampling can yield an alternative approximation to the distribution of $T_{\mathbf{n}} = g(\hat{\theta}_{\mathbf{n}} - \theta)/S$ and that of $T_{\mathbf{n},\mathbf{0}} = g(\hat{\theta}_{\mathbf{n}})/S$ that may be of particular usefulness if the two populations are not normal. As will be shown in what follows, our subsampling-based inference will remain asymptotically valid *even* if the data have infinite variance.

4.1 Data with finite variance

In this subsection, we consider the customary case that σ_1^2 and σ_2^2 are both finite (and nonzero). In the absence of any restriction on the relative sizes of n_1 and n_2, the easiest way to see the the applicability of subsampling here is to use the framework of Section 3.3, i.e., to let $\hat{\sigma}_{\mathbf{n}} = 1$, and $\tau_{\mathbf{n}} = (\sigma_2^2/n_2 + \sigma_1^2/n_1)^{-1/2}$. It is immediate that $\tau_{\mathbf{n}}$ satisfies $\tau_{\mathbf{b}}/\tau_{\mathbf{n}} \to 0$ under (5). As alluded to earlier on, the normalizing factor $\tau_{\mathbf{n}}$ involves the unknown parameters σ_1^2 and σ_2^2 but is consistently estimated by $\hat{\tau}_{\mathbf{n}} = (S_2^2/n_2 + S_1^2/n_1)^{-1/2}$. Since $S_1^2 \xrightarrow{P} \sigma_1^2$ and $S_2^2 \xrightarrow{P} \sigma_2^2$, it follows that $\hat{\tau}_{\mathbf{n}}/\tau_{\mathbf{n}} \xrightarrow{P} 1$.

The finite variance condition also implies that $\sqrt{n_i}(\bar{X}^{(i)} - \mu_i) \xRightarrow{\mathcal{L}} N(0, \sigma_i^2)$ which–together with the independence of the two samples–implies that $\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta) \xRightarrow{\mathcal{L}} N(0, 1)$; hence, $\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta)$ is a full asymptotic pivot and the following corollary ensues.

Corollary 4.1. *Under eq. (18) with $0 < \sigma_i^2 < \infty$ for $i = 1, 2$, the subsampling tests based on the quantiles of either $\hat{G}_{\mathbf{n},\mathbf{b}}$ or $\hat{L}_{\mathbf{n},\mathbf{b}}$ are asymptotically valid and consistent, and the associated subsampling confidence bounds are also asymptotically valid.*

Remark 4.1. If the data-centered subsampling distribution $L_{\mathbf{n},\mathbf{b}}(x)$ is to be used, then it may benefit from a *finite-population correction* as in the case of one sample; see e.g., Politis et al. (1999, p. 218). To see this, note that if $\bar{X}_{b_i}^{(i)}$ is the sample mean of b_i of the $X_t^{(i)}$ data, then $Var(\bar{X}_{b_i}^{(i)} - \bar{X}^{(i)}) = \mathrm{h}_1 \sigma_i^2/b_i$ where $\mathrm{h}_i = 1 - b_i/n_i$ for $i = 1, 2$. So $Var(\hat{\theta}_{\mathbf{n}} - \bar{X}_{b_2}^{(2)} + \bar{X}_{b_1}^{(1)}) =$

$\mathrm{h}_1\sigma_1^2/b_1 + \mathrm{h}_2\sigma_2^2/b_2$ is it is best to let $\tilde{\tau}_{\mathbf{b}} = (\mathrm{h}_1 S_1^2/b_1 + \mathrm{h}_2 S_2^2/b_2)^{-1/2}$ instead of $\hat{\tau}_{\mathbf{b}}$ in the construction of $\hat{L}_{\mathbf{n},\mathbf{b}}(x)$. Note that $\tilde{\tau}_{\mathbf{b}}/\hat{\tau}_{\mathbf{b}} \to 1$ since $b_i/n_i \to 0$; hence, Corollary 4.1 would remain valid regardless of whether $\tilde{\tau}_{\mathbf{b}}$ or $\hat{\tau}_{\mathbf{b}}$ was used. Nevertheless, in finite samples $\tilde{\tau}_{\mathbf{b}} > \hat{\tau}_{\mathbf{b}}$, and this could make a real difference in practice.

4.2 Data with infinite variance

In this subsection we still assume eq. (18) but now the c.d.f.s F_1, F_2 are assumed to have finite means μ_1, μ_2 but infinite variances, i.e., $\sigma_i^2 = \infty$ for $i = 1, 2$. For simplicity, we will now require that n_1 and n_2 are of the same order of magnitude, i.e., that for some constants $c_1, c_2 \in (0, 1)$, we have

$$n_1 \sim c_1\ n \text{ and } n_2 \sim c_2\ n \text{ where } n = n_1 + n_2. \tag{19}$$

A simple way to model such an infinite variance situation is to assume that F_1 and F_2 are in the normal domain of attraction of a stable law with index $\beta \in (1, 2]$. In other words, assume that

$$V_n^{(i)} \equiv a_{n_i}^{(i)}(\bar{X}^{(i)} - \mu_i) \stackrel{\mathcal{L}}{\Longrightarrow} V^{(i)} \quad \text{where} \quad a_{n_i}^{(i)} = n_i^{1-1/\beta} \tag{20}$$

for $i = 1, 2$ where $V^{(i)}$ is a stable law with index $\beta \in (1, 2]$; see e.g., Section 2.2 in Embrechts, Klüppelberg and Mikosch (1997).

As in Proposition 11.4.3 of Politis et al. (1999) we will now construct a self-normalized sum for which subsampling will be applicable even *without* knowledge (or explicit estimation) of the value of β. To do this, we define $\hat{\theta}_{\mathbf{n}} = \bar{X}^{(2)} - \bar{X}^{(1)}$ as before but now–aided by eq. (19)–we set $\tau_{\mathbf{n}} = \sqrt{n}$, and $\hat{\sigma}_{\mathbf{n}}^2 = S_2^2/c_2 + S_1^2/c_1$ where S_i^2 are the sample variances as in Section 4.1.

To investigate the properties of the self-normalized root $T_{\mathbf{n}} = \tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta)/\hat{\sigma}_{\mathbf{n}}$ we also consider

$$W_n^{(i)} \equiv d_{n_i}^{(i)}\ S_i \quad \text{where} \quad d_{n_i}^{(i)} = n_i^{-1/2-1/\beta} \tag{21}$$

for $i = 1, 2$. So

$$T_{\mathbf{n}} = \sqrt{n}\frac{\bar{X}^{(2)} - \bar{X}^{(1)} - \mu_2 + \mu_1}{\sqrt{S_2^2/c_2 + S_1^2/c_1}}$$

$$= \frac{\sqrt{n}\left(V_n^{(2)}/a_{n_2}^{(2)} - V_n^{(1)}/a_{n_1}^{(1)}\right)}{\sqrt{(W_n^{(2)})^2/(c_2(d_{n_2}^{(2)})^2) + (W_n^{(1)})^2/(c_1(d_{n_1}^{(1)})^2)}}$$

$$= \frac{V_n^{(2)}/C_2 - V_n^{(1)}/C_1}{\sqrt{(W_n^{(2)})^2/(c_2 C_2^2) + (W_n^{(1)})^2/(c_1 C_1^2)}}$$

where $C_i = c_i^{-1/2-1/\beta}$. But by results of Logan et al. (1973), for each i, the pair $(V_n^{(i)}, W_n^{(i)})$ converges in law to the pair of r.v.'s $(V^{(i)}, W^{(i)})$ where $V^{(i)}$ is a stable law with index β, and $W^{(i)}$ is a (positive) stable law with index $\beta/2$. So, by the independence of the two samples, it follows that $(V_n^{(1)}, W_n^{(1)}, V_n^{(2)}, W_n^{(2)})$ converge in law *jointly* to $(V^{(1)}, W^{(1)}, V^{(2)}, W^{(2)})$. It then follows that Assumption 2.1 is satisfied, and the same is true for Assumption 3.1 with the following identifications: $a_{\mathbf{n}} = n^{1-1/\beta}$, $d_{\mathbf{n}} = n^{-1/2-1/\beta}$, and $\tau_{\mathbf{n}} = a_{\mathbf{n}}/d_{\mathbf{n}}$.

Corollary 4.2. *Assume eq. (18), (19) and (20) for some $\beta \in (1, 2]$. Then, the subsampling tests based on the quantiles of either $G_{\mathbf{n},\mathbf{b}}$ or $L_{\mathbf{n},\mathbf{b}}$ are asymptotically valid and consistent, and the associated subsampling confidence bounds are also asymptotically valid.*

Remark 4.2. Note that the practitioner does not need to know the value of β in order to construct the subsampling distributions and conduct inference about θ. For example, Corollary 4.2 applies even in the case when the data have finite variance, i.e., $\beta = 2$, presenting a robust alternative to Corollary 4.1 at the slight cost of the extra assumption (19).

In fact, assume that the index $\beta = \beta_i$ can depend on i. Then, we can still show that T_n has a limiting distribution. Indeed, let $\beta = \min(\beta_1, \beta_2)$. Then, if we define $a_{\mathbf{n}} = n^{1-1/\beta}$ and $d_{\mathbf{n}} = n^{1/2-1/\beta}$, then it is easy to see that the conditions are satisfied. For example, suppose $\beta_1 < \beta_2$. Then,

$$a_{\mathbf{n}}[(\bar{X}^{(2)} - \bar{X}^{(1)}) - (\mu_2 - \mu_1)] = a_{\mathbf{n}}(\bar{X}^{(2)} - \mu_2) - a_{\mathbf{n}}(\bar{X}^{(1)} - \mu_1) \ .$$

But, $a_{\mathbf{n}}(\bar{X}^{(1)} - \mu_1)$ has a limiting distribution V_1 and

$$a_{\mathbf{n}}(\bar{X}^{(2)} - \mu_2)) = n^{1/\beta_2 - 1/\beta_1} n^{1-1/\beta_2}(\bar{X}^{(2)} - \mu_2) \to 0$$

in probability, since $n^{1-1/\beta_2}(\bar{X}^{(2)} - \mu_2)$ has a limiting distribution V_2 and the factor $n^{1/\beta_2 - 1/\beta_1} \to 0$ as $\beta_1 < \beta_2$. The denominator of the statistic can be handled similarly. The result is that the limiting distribution in this case is that of $-V^{(1)}/\sqrt{(W^{(1)})^2/c_1}$. Thus, the results apply.

Finally, note that similar arguments apply if F_i is in the domain of attraction of a stable law with index β_i (not necessarily *normal* domain of attraction). Here, there is a slowly varying function $L_i(n_i)$ which ensures that

$$n^{1-1/\beta_i}(\bar{X}^{(i)} - \mu_i)/L_i(n_i)$$

converges to a stable law. If the β_i are distinct, the above argument generalizes and there exists a limiting distribution for the normalized statistic. It also generalizes if $L_1 = L_2$. However, if $\beta_1 = \beta_2$ but the L_i differ, it appears that a limiting distribution may not exist in general. Such an issue is left for future work.

4.3 Block size choice

We now return to the finite variance set-up of Section 4.1 to discuss the difficult–but important–issue of optimal choice of the block sizes b_1, b_2. Consider the two targets:

1. Optimize the subsampling distribution $G_{\mathbf{n},\mathbf{b}}(x)$ as an estimate of the distribution of $T_{\mathbf{n},\mathbf{0}} = g(\hat{\theta}_{\mathbf{n}})/S$ under the null.
2. Optimize the subsampling distribution $L_{\mathbf{n},\mathbf{b}}(x)$ as an estimate of the distribution of $T_{\mathbf{n}} = g(\hat{\theta}_{\mathbf{n}} - \theta(P))/S$ under a general P.

The first target would amount to trying to have a test whose size is as close to the nominal as possible; the second target would result in an optimal subsampling estimator of the distribution of $T_{\mathbf{n}}$ under different possibilities for P in effect optimizing power.

Recall that, from Lemma 5.1, $Var(G_{\mathbf{n},\mathbf{b}}(x)) \leq \max_k(b_k/n_k)$. We now assume, for simplicity, that the two sample sizes are of the same order of magnitude, i.e., eq. (19). It is then natural to assume that the relation of the two sample sizes n_1, n_2 is reflected in the relation of the two subsample sizes b_1, b_2; this is tantamount to assuming that

$$b_1 \sim c_1 \ b \text{ and } b_2 \sim c_2 \ b \text{ where } b = b_1 + b_2 \tag{22}$$

which implies that $b_1/n_1 = b_2/n_2 = b/n$ so that

$$Var(G_{\mathbf{n},\mathbf{b}}(x)) \leq b/n. \tag{23}$$

Since $EG_{\mathbf{n},\mathbf{b}}(x) = J_{\mathbf{b}}(x,P)$, to give a bound on the bias of $G_{\mathbf{n},\mathbf{b}}(x)$ we would need the rate of convergence of $J_{\mathbf{n}}(x,P)$ to $J(x,P)$. Such a general result is not readily available; however, in the two-sample problem with normal data, Welch (1947) gave the following Edgeworth expansion:

$$Prob_P\{(\hat{\theta}_{\mathbf{n}} - \theta(P))/S \leq x\} = \Phi(x) - \frac{1}{4}(x + x^3)\phi(x)\zeta_{\mathbf{n}} + o(\zeta_{\mathbf{n}}) \tag{24}$$

where $\Phi(x), \phi(x)$ are the standard normal c.d.f. and density respectively, and

$$\zeta_{\mathbf{n}} = \frac{\sigma_1^4/(n_1^2(n_1-1)) + \sigma_2^4/(n_2^2(n_2-1))}{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^2}.$$

Although (24) is valid only under normality, it is natural to conjecture that

$$J_{\mathbf{n}}(x,P) = Prob_P\{g(\hat{\theta}_{\mathbf{n}} - \theta(P))/S \leq x\} = \Phi(x) + O(\zeta_{\mathbf{n}}) \tag{25}$$

when $g(x) = x$ and either the c.d.f's F_1, F_2 are both symmetric or they have the same skewness and $n_1 = n_2$ since in either case $\hat{\theta}_{\mathbf{n}}$ has zero skewness; similarly, (25) should hold in general when $g(x) = |x|$.

Thus, (25) would imply that $\text{Bias}(G_{\mathbf{n},\mathbf{b}}(x)) = O(\zeta_{\mathbf{b}})$. But $\zeta_{\mathbf{n}} = O(1/n)$ due to eq. (19), and so from (23) it follows that

$$G_{\mathbf{n},\mathbf{b}}(x) - J_{\mathbf{n}}(x,P) = O(1/b) + O_P(\sqrt{b/n}) \tag{26}$$

which is, of course, optimized by letting

$$b \sim C\ n^{1/3} \tag{27}$$

for some constant C whose value could be pin-pointed if the term $O(\zeta_{\mathbf{n}})$ in (25) were replaced by a more explicit bound (that would also depend on x).

Analyzing $L_{\mathbf{n},\mathbf{b}}(x)$ is more difficult. However, in the aforementioned case where $g(x) = x$ and either the c.d.f's F_1, F_2 are both symmetric or $n_1 = n_2$ an argument analogous to the one given in Politis and Romano (1994, p. 2039) implies that eq. (26) would remain valid with $L_{\mathbf{n},\mathbf{b}}(x)$ instead of $G_{\mathbf{n},\mathbf{b}}(x)$ so that the optimal order of block size (27) is still applicable.

5 Simulation experiments

5.1 Two sample simulations

We investigated several two sample problems in order to assess the performance of the subsampling hypothesis tests and the optimal rate for b given in (27). For all two sample problems, we tested the hypothesis $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$ over a range of values for μ_2; the results are presented as power curves with the power a function of the true difference between μ_1 and μ_2. The test statistic was the difference in means between the two groups, $g(x) = |x|$ for $g(\cdot)$ as in equation (1), and the rate used was $\tau_{\mathbf{n}} = (S_1^2/n_1 + S_2^2/n_2)^{-1/2}$. For each considered type of data, we used four sample sizes, $(n_1, n_2) = (50, 50),\ (100, 100),\ (200, 200),$ and $(60, 100)$. For each value of μ_2, the experiment was repeated 1000 times using 1000 subsamples.

Subsampling block sizes can be chosen via different data-driven methods; see e.g., Ch. 9 of Politis et al. (1999). Since the focus of our simulation was to investigate the effect of centering on the subsampling distribution (and to compare with the t-test), the following simple short-cut was used that also puts the recommendation of eqs. (22) and (27) to the test. For each experiment, a Monte

Carlo simulation was performed under the null hypothesis in the case where $(n_1, n_2) = (100, 100)$ to empirically identify the constant C from eq. (27) that results in a subsampling test with level closest to $\alpha = 0.05$. Of course, the distributional information required for this simulation would not be available to the practitioner but we also found that this method of estimating the constant C is fairly robust to distributional mis-specification in the finite variance case. Alternatively, the subsampling/cross-validation trick of Hall, Horowitz and Jing (1996) could be employed to estimate the constant C in practice.

Subsample sizes corresponding to different sample sizes were then chosen via eqs. (22) and (27); this allows the $n^{1/3}$ rate for the block sizes to be assessed by looking at the achieved level in the cases with different sample sizes. The $n^{1/3}$ rate worked well in all cases except the infinite variance situations studied in Experiments 4 and 5, where we were forced to manually choose different subsample sizes for each sample size; in these cases, the classical central limit theorem does not hold, and it is unsurprising that (27) does not apply.

Surprisingly, while the N-B centered test should, in principle, capture the size better than the D-C centered test, it proved very difficult to calibrate in two-sample cases; in many cases, we were simply unable to achieve the desired $\alpha = 0.05$ with the N-B centered test. For this reason, the subsample sizes in the simulations were chosen with the objective of achieving the correct size for the D-C centered subsampling test.

Our simulation results are presented in a series of power curves in Figures 1–7. The following shorthand notation was used: *pt* denotes power curve associated with the classical two-sample t-test, *pc* denotes power curve associated with the D-C centered subsampling test, and *puc* denotes power curve associated with the N-B centered subsampling test which, in the case of testing $H_0 : \mu_1 - \mu_2 = 0$, is tantamount to using no centering, i.e., an *uncentered* subsampling distribution.

Experiment 1: In the first simulation we tested equality of means for samples from two standard deviation 1 normal distributions. The first sample was drawn from a normal distribution with mean 0, and the second samples were drawn from normal distributions with means $0.0, 0.1, \ldots, 1.0$. The results are shown as a power curve in Figure 1. The D-C centered subsampling test is almost identical to the classical two sample t-test; the N-B centered test is noticeably less powerful.

Experiment 2: In the second simulation we tested equality of means for samples from two normal distributions, the first a standard normal distribution, and the second normal distributions with standard deviation 2 and means $0.0, 0.1, \ldots, 1.0$. Again, the subsampling test is quite competitive with the t-test, and the N-B centered test is much less powerful. The results are shown in Figure 2.

Experiment 3: In the third simulation we tested the equality of means for samples from two t distributions with 6 degrees of freedom. The first sample was simulated from a population with mean 0, and the second sample was simulated from populations with means $0.0, 0.1, \ldots, 1.0$. The results are shown in Figure 3. The D-C centered subsampling test is approximately as powerful as the t-test, and the N-B centered subsampling test remains uncompetitive.

Experiment 4: In the fourth experiment we tested the performance of subsampling for data with finite mean and infinite variance. Both samples were simulated from a stable distribution with tail index $\beta = 1.1$, skewness parameter 0.1, and scale parameter 1, using parametrization 1 as described in Nolan (2010), Ch. 1. Samples from the second distribution were translated by $0.0, 0.5, \ldots, 8.0$. The results are shown Figure 4. The t-test performed well even though theory suggests it should not work; its size was about half the nominal size for all simulations, yet for

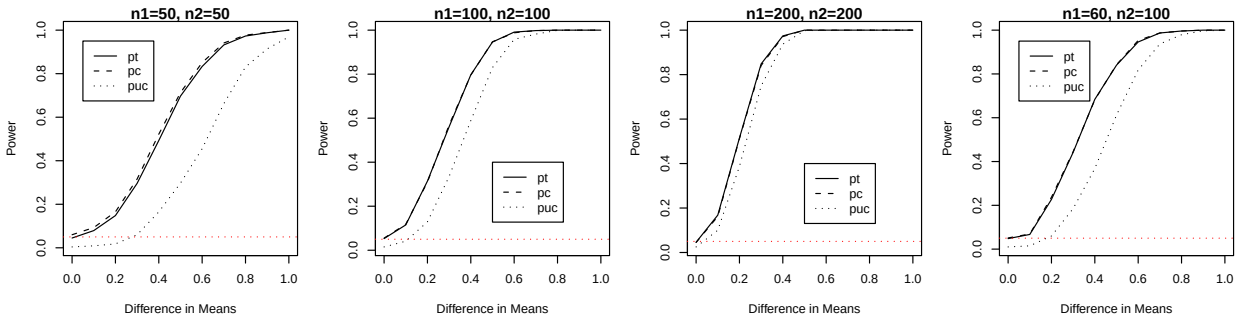

Figure 1: Power curves for the comparison of the means of two normal samples with equal variance; the horizontal line indicates the nominal size of the test, and *pt*, *pc*, *puc* denote power curves associated with the t-test, the D-C centered subsampling test, and the N-B centered (uncentered) subsampling test respectively.

large differences in means it was somewhat more powerful than subsampling. We believe that this is due to non-robustness in each population somehow canceling when the means of the two samples are subtracted. The D-C centered subsampling test had the correct size, but was slightly less powerful than the t-test. Finally, the N-B centered subsampling test performed poorly; its power is particularly harmed because it had size much smaller than the nominal 0.05. It is possible that this could be improved by choosing a different subsample size.

Experiment 5: In the fifth experiment we compared samples from two different stable distributions, each with tail index $\beta = 1.1$, but the first with skewness 0.1 and the second with skewness -0.1. The means of the second distribution were $0.0, 0.5, \ldots, 8.0$. The t-test is more powerful, but its size was twice the desired $\alpha = 0.05$. Subsampling behaves correctly under the null hypothesis. Results are shown in Figure 5.

Experiment 6: In the final two sample experiment we tested the equality of means for samples from two exponential distributions. The first sample was drawn from an exponential distribution with rate 1. The second samples were drawn from exponential distributions with rates $0.1, 0.2, \ldots, 2.0$. Since the exponential parameter is a rate rather than location, the populations are not related by a simple translation. For this reason, the power curve is asymmetric, and in Figure 6 we show both positive and negative differences in true means. Subsampling is slightly less powerful than the t test when the sample sizes were $(n_1, n_2) = (50, 50)$. In all other settings the subsampling test was comparably powerful.

5.2 Three sample simulations

We considered three different three sample problems, each with sample sizes $n_1 = n_2 = n_3 = 100$. We used the χ^2 test statistic derived from the likelihood ratio test for the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ under the assumption that the three populations are normal with possibly unequal variances. As in the two sample experiments, the subsample size was chosen to achieve $\alpha = 0.05$ under the null hypothesis and subsequently kept fixed for the remainder of the experiments. The test statistic does not have a closed form, so it was estimated numerically.

Experiment 7: In the first three sample experiment we compared three samples each from normal distributions with standard deviation 1. The mean of the first population was fixed at 0. The

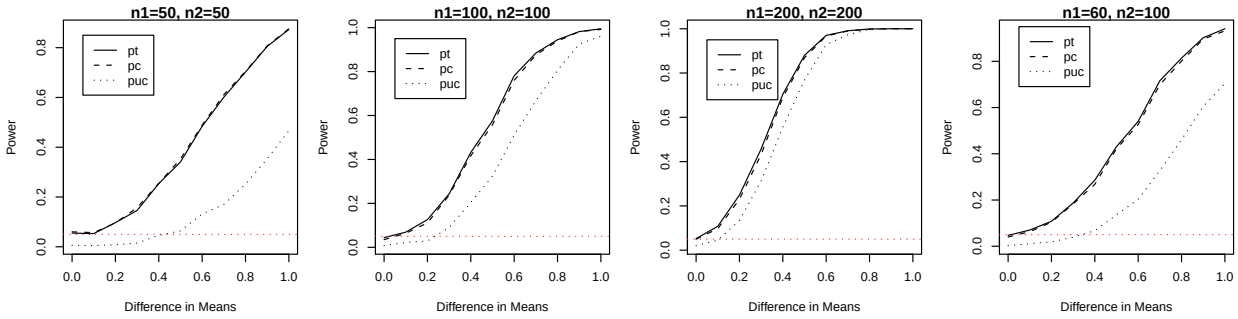

Figure 2: Power curves for the comparison of the means of two normal samples with unequal variance.

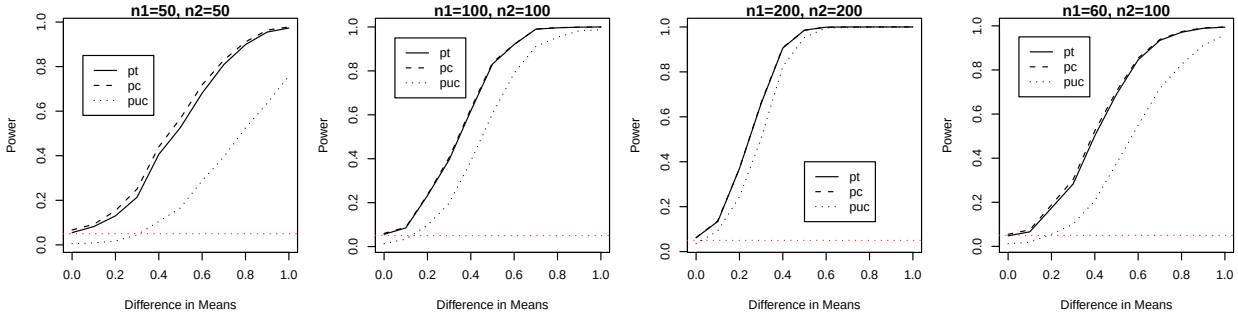

Figure 3: Power curves for the comparison of the means of two t_6 samples with different means and equal scale.

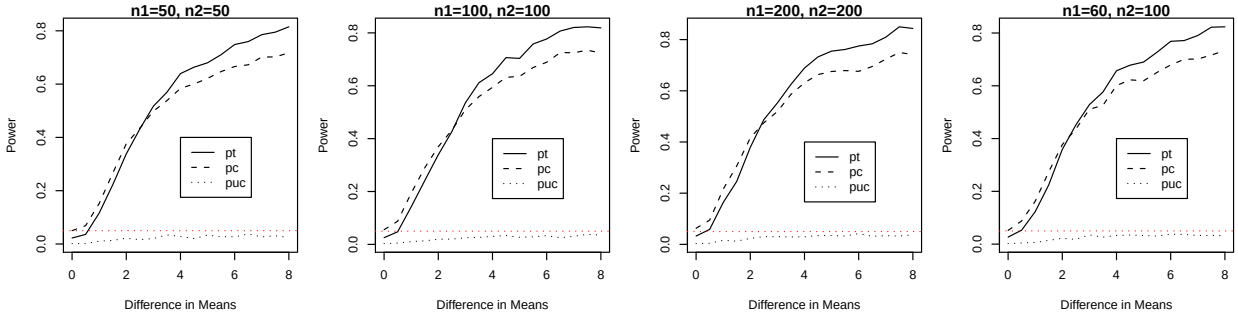

Figure 4: Power curves for the comparison of the means of two samples from a stable distribution with finite mean but infinite variance.

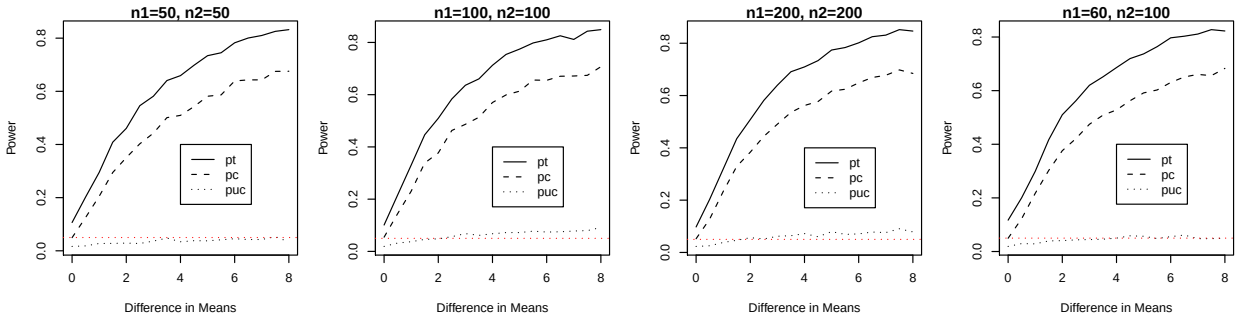

Figure 5: Power curves for the comparison of the means of two samples; stable distributions with tail index 1.1 and skewness parameters 0.1 and -0.1 respectively.

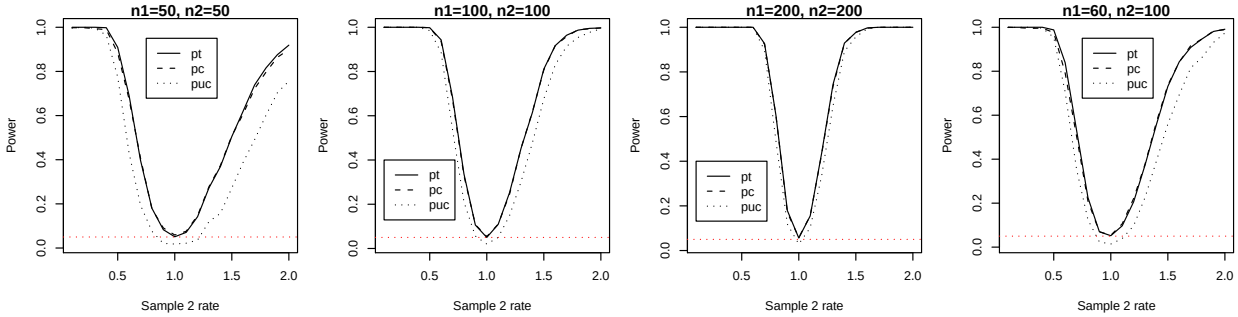

Figure 6: Power curves for comparison of the means of two exponential samples, the first with rate 1, and the second with rate shown on the horizontal axis.

means of the second and third populations ranged over $0.0, 0.1, \ldots, 0.5$, with the mean of the third population always less than or equal to the mean of the second, resulting in power surfaces over a triangular region. The biggest absolute difference between both the D-C centered and N-B centered subsampling surfaces and the χ^2 surface is 0.018.

Experiment 8: In the second three sample experiment, we compared three samples, the first from a normal distribution with standard deviation 1 and the second two from normal distributions with standard deviation 2. Power was again estimated over the triangular grid described in the previous experiment. The surfaces are qualitatively similar, and the biggest absolute difference between both the D-C centered and N-B centered subsampling surfaces and the χ^2 surface is 0.011.

Experiment 9: For the final three sample experiment, the first sample was drawn from a standard normal distribution and the second and third samples were drawn from a t distribution with 6 degrees of freedom. The power surface was calculated over the triangular region used in the preceding two experiments. In this case the test based on the χ^2 distribution was slightly more powerful. The biggest absolute difference between both the D-C centered and N-B centered subsampling surfaces and the χ^2 was 0.034, although most differences were much smaller.

5.3 Empirical choice of b

Finally we investigated automated choice of subsample size through a variation of the algorithm of Götze and Račkauskas (2001) and Bickel and Sakov (2008) (henceforth GRBS) with Kolmogorov-Smirnov (KS) distance. We first describe the algorithm when $n_1 = n_2 = \ldots = n_K$; in this setting, it may be reasonable to assume $b \equiv b_1 = b_2 = \ldots = b_k$. If so, the rule consists of the following steps:

1. Assume that the correct subsample sizes are equal. Let $b^{[j]} = \lfloor q^j n \rfloor$, $j = 0, 1, \ldots$, and $0 < q < 1$.
2. For each $b^{[j]}$ compute $\hat{L}_{n,b^{[j]}}$.
3. Let $d(\cdot, \cdot)$ denote KS distance. Choose $\hat{b} = \arg\min_{b^{[j]}} d(\hat{L}_{n,b_j}, \hat{L}_{n,b_{j+1}})$.
4. The estimator of L_n is $\hat{L}_{n,\hat{b}}$.

The reasoning behind the GRBS method is that if the block size is within the 'right' range of b values, then the corresponding empirical distributions should be 'close' to each other, and therefore values of b associated with small KS distances will be selected by the rule.

The GRBS method was tested in a replication of Experiment 1 with $n_1 = n_2 = 100$. Automated subsample size choice produces results that are comparable to the t-test. Results are shown in the left panel of Figure 7.

Clearly, algorithm described above is not suitable when it is expected that the b_is differ. In the two sample setting we modified the above algorithm to employ a grid search over values of b_1 and b_2 generated by step 1. We then chose the (b_1, b_2) pair that produced the smallest average KS distance to its neighbors. We assessed this method by replicating the $n_1 = 60$, $n_2 = 100$ run of Experiment 2. In this case the power was somewhat reduced from the t-test. Results are shown in the right panel of Figure 7.

There are many alternative approaches to choice of b. Some are outlined in Ch. 9 of Politis et al. (1999). Another approach would be to simulate data satisfying the null hypothesis that

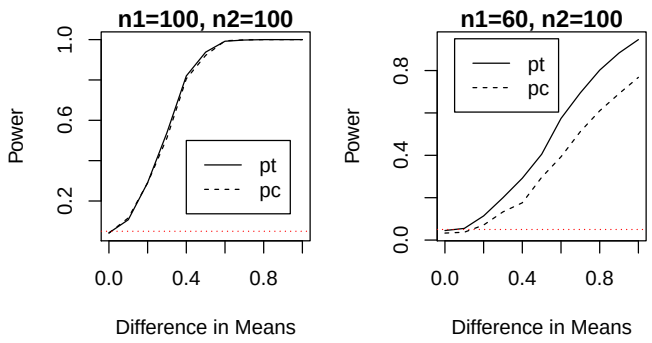

Figure 7: Power curves for tests incorporating automated choice of subsample size.

roughly matches characteristics of the observed data, i.e. by matching moments or a smoothed bootstrap, and then choosing subsample sizes to achieve the desired α on the simulated data. In experimentation, we found that chosen sample sizes are relatively stable to modest changes in the underlying distribution, and we expect this approach to provide competitive results in many situations.

5.4 Discussion

The D-C centered subsampling tests performed remarkably well, often as well as the parametric alternative when assumptions for the latter were exactly satisfied. It is surprising that even in the situations where the t-test is optimal or close to optimal, the D-C centered subsampling test yields comparable power. The N-B centered subsampling test performed well in our three-sample experiments where we were able to calibrate both subsampling tests under the null hypothesis; in these cases the D-C centered and N-B centered tests (using different subsample sizes) performed very similarly. For the two-sample tests, we were generally unable to choose subsample sizes which gave the N-B centered test the correct test size; this caused the test to noticeably lack power.

An additional surprising finding was that the two-sample t-test seems to perform quite well even in situations where it is *not* asymptotically justified, e.g., in Experiment 4. In the infinite-variance cases, we expected that the t-test would give reasonable power behavior (because of the t-statistic becoming more heavy-tailed) but would have size problems. Indeed, this size breakdown is apparent in Experiment 5 where the observed size was double the nominal. In Experiment 4, however, the t-test's size was less than the desired 0.05, yet in places it was more powerful than subsampling; this unexpectedly good performance may be due to some cancellations in the Edgeworth expansion of the t-statistic when the two samples have the same distribution. In general though the t-test's size is substantially dependent on the (unknown) underlying populations, and this dependence is not easily quantified; subsampling is much more robust in that respect.

Appendix: K-sample U-statistics

Consider K independent datasets: $\underline{X}^{(1)}, \ldots, \underline{X}^{(K)}$ where $\underline{X}^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$. For each k, the random variables $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ are assumed i.i.d. taking values in an arbitrary space $\mathbf{B}_k$. Let $\mathcal{S}_k$ denote the set of all size b_k (unordered) subsets of the dataset $\{X_1^{(k)}, \ldots, X_{n_k}^{(k)}\}$ where b_k is an integer in $[1, n_k]$. Note that the set $\mathcal{S}_k$ contains $Q_k = \binom{n_k}{b_k}$ elements that are enumerated as

$S_1^{(k)}, S_2^{(k)}, \ldots, S_{Q_k}^{(k)}$.

Now let $\mathbf{b} = (b_1, \ldots, b_K)$, and consider a real-valued function $h_{\mathbf{b}}(\cdot)$ of $b_1 \times b_2 \times \cdots \times b_K$ arguments; strictly speaking, this is an array of functions since $h_{\mathbf{b}}(\cdot)$ is different for each $\mathbf{b}$. To elaborate, $h_{\mathbf{b}} : \mathbf{B}_1^{b_1} \times \mathbf{B}_2^{b_2} \times \cdots \times \mathbf{B}_K^{b_K} \to \mathbf{R}$, i.e., the arguments of $h_{\mathbf{b}}(\cdot)$ are grouped in K subsets where the kth subset is a size b_k subsample from the kth set of observations in the space $\mathbf{B}_k$. For example, $h_{\mathbf{b}}(\cdot)$ can (and will) be evaluated on subsamples of our K datasets; for example, evaluation of $h_{\mathbf{b}}(\cdot)$ on the first subsample from each sample gives $h_{\mathbf{b}}(S_1^{(1)}, S_1^{(2)}, \ldots, S_1^{(K)})$. Furthermore, the function $h_{\mathbf{b}}(\cdot)$ is assumed to have some symmetry; in particular, letting $\tilde{S}_1^{(k)}$ denote an arbitrary permutation of the elements of $S_1^{(k)}$, it will be assumed that $h_{\mathbf{b}}(\tilde{S}_1^{(1)}, \tilde{S}_1^{(2)}, \ldots, \tilde{S}_1^{(K)}) = h_{\mathbf{b}}(S_1^{(1)}, S_1^{(2)}, \ldots, S_1^{(K)})$.

A K-sample U-statistic of *degree* $\mathbf{b} = (b_1, \ldots, b_K)$ and *kernel* $h_{\mathbf{b}}(\cdot)$ is now defined as:

$$U_{\mathbf{n},\mathbf{b}} = \frac{1}{Q} \sum_{i_1=1}^{Q_1} \sum_{i_2=1}^{Q_2} \cdots \sum_{i_K=1}^{Q_K} h_{\mathbf{b}}(S_{i_1}^{(1)}, S_{i_2}^{(2)}, \ldots, S_{i_K}^{(K)}) \tag{28}$$

where $Q = \prod_{k=1}^{K} Q_k$. As is apparent, $U_{\mathbf{n},\mathbf{b}}$ is an unbiased estimator of the kernel $Eh_{\mathbf{b}}(S_1^{(1)}, S_2^{(2)}, \ldots, S_K^{(K)})$.

The theory of U-statistics was put forth in two pioneering papers of Hoeffding (1948, 1963). In the second paper, Hoeffding also discusses the case of two-sample U-statistics. In the following Lemma, we extend Hoeffding's exponential inequality to the case $K > 2$. Technically, the extension is immediate in view of Hoeffding's (1963) argument for the case $K = 2$. Nevertheless, the Lemma is of some interest since it implies consistency of the K-sample U-statistic even when the order of the kernel is not fixed. Rather, each b_k can be a function of its respective n_k, and–as in the application to subsampling–we can have $b_k \to \infty$ as $n_k \to \infty$ as long as $b_k/n_k \to 0$.

Lemma 5.1. *Assume the kernel is bounded, i.e., for some constants A and B, we have $A \leq h_{\mathbf{b}}(y_1, \ldots, y_b) \leq B$ for any $y_1, \ldots, y_b$ and for any $\mathbf{b}$; here, $b = \sum_k b_k$. Let $D_{\mathbf{n},\mathbf{b}} = \min_k \lfloor n_k/b_k \rfloor$; then, for any $t > 0$,*
(i) $P\{U_{\mathbf{n},\mathbf{b}} - EU_{\mathbf{n},\mathbf{b}} \geq t\} \leq \exp[-2t^2 D_{\mathbf{n},\mathbf{b}}/(B-A)^2]$.
(ii) $P\{U_{\mathbf{n},\mathbf{b}} - EU_{\mathbf{n},\mathbf{b}} \leq -t\} \leq \exp[-2t^2 D_{\mathbf{n},\mathbf{b}}/(B-A)^2]$.
(iii) $Var(U_{\mathbf{n},\mathbf{b}}) \leq (B-A)^2/D_{\mathbf{n},\mathbf{b}}$.
(iv) If $\max_k(b_k/n_k) \to 0$, then $Var(U_{\mathbf{n},\mathbf{b}}) \to 0$ and $U_{\mathbf{n},\mathbf{b}} - EU_{\mathbf{n},\mathbf{b}} \xrightarrow{P} 0$.

Proof. Part (i) follows analogously to the argument given in Hoeffding (1963, section 5b) for the case $K = 2$. Part (ii) follows from (i) by considering the U-statistic with kernel $-h_{\mathbf{b}}(\cdot)$. To prove (iii), note that

$$\begin{aligned} E(U_{\mathbf{n},\mathbf{b}} - EU_{\mathbf{n},\mathbf{b}})^2 &= \int_0^\infty P\{(U_{\mathbf{n},\mathbf{b}} - EU_{\mathbf{n},\mathbf{b}})^2 \geq x\} dx \\ &\leq \int_0^\infty 2\exp[-2xD_{\mathbf{n},\mathbf{b}}/(B-A)^2] dx \end{aligned}$$

by (i) and (ii). But the latter integral equals $(B-A)^2/D_{\mathbf{n},\mathbf{b}}$. Alternatively, (iii) can be argued along the lines of Remark 2.2.3 in Politis et al. (1999). Indeed, if m is the greatest integer less than or equal to $\min_k(n_k/b_k)$, then an unbiased estimator of $E(U_{\mathbf{n},\mathbf{b}})$ can be constructed which is the average over m *independent* set of subsamples; such an estimator has variance bounded above by $(B-A)^2/(4m) \to 0$. But, by conditioning on the information contained in the samples without regard to their original ordering (in generalization of the order statistics), we can apply the Rao-Blackwell theorem to show that the variance of $U_{\mathbf{n},\mathbf{b}}$ is at least as small, and hence tends to 0 as well. Finally, (iv) follows trivially from (iii) noting that $D_{\mathbf{n},\mathbf{b}}^{-1} = O(\max_k(b_k/n_k)) = o(1)$.

□

References

[1] Andrews, D. W. K. and Guggenberger, P. (2009). Validity of subsampling and 'plug-in asymptotic' inference for parameters defined by moment inequalities. *Econometric Thoery*, **25**, 669-709.

[2] Andrews, D. W. K. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and the m out of n bootstrap. *Econometric Theory*, **26**, 426-468.

[3] Beran, R. (1986). Simulated power functions. *Annals of Statistics*, **14**, 151-173.

[4] Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements, *J. Amer. Statist. Assoc.*, vol. 83, pp. 687-697.

[5] Berg, A., McMurry, T.L. and Politis, D.N. (2010). Subsampling p-values. *Statist. Prob. Letters*, no. 17-18, pp. 1358-1364.

[6] Bickel, P.J., Götze, F., van Zwet, W.R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica,* **7**, 1-32.

[7] Bickel, P.J. and Ren, J.J. (2001). The bootstrap in hypothesis testing. *Lecture Notes-Monograph Series*, **36**, State of the Art in Probability and Statistics, pp. 91-112.

[8] Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, **18**, 967-985.

[9] Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171-1179.

[10] Chen, P. and Hsiao, C.-Y. (2010). Subsampling the Johansen test with stable innovations. *Australian & New Zealand Journal of Statistics*, **52**, 61-73.

[11] Choi, I. (2005). Subsampling vector autoregressive tests of linear constraints. *Journal of Econometrics*, **124**, 55-89.

[12] Choi, I. and Chue, T. K. (2007). Subsampling hypothesis tests for nonstationary panels with applications to exchange rates and stock prices. *Journal of Applied Econometrics*, **22**, 233-264.

[13] Choi, I. and Saikkonen, P. (2010). Tests for nonlinear cointegration. *Econometric Theory*, **26**, 682-709.

[14] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.

[15] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, New York.

[16] Delgado, M., Rodríguez-Poo, J., and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manskis maximum score estimator. *Economics Letters*, **73**, 241-250.

[17] DiCiccio, T.J. and Romano, J.P. (1989). A review of bootstrap confidence intervals (with discussion). *J. Royal. Statist. Soc. B* **50**, 338-370.

[18] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.

[19] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics*, Volume 38. SIAM.

[20] Efron, B. and Tibshirani, R. (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statist. Science*, vol. 1, pp. 54-77.

[21] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[22] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance.* Springer-Verlag, Berlin.

[23] Gonzalo, J. and Wolf, M. (2005). Subsampling inference in threshold autoregressive models. *Journal of Econometrics,* **127**, 201-224.

[24] Götze, F. and Račkauskas, A. (2001). Adaptive choice of bootstrap sample sizes. In *State of the Art in Probability and Statistics, IMS Lecture Notes Monogr. Ser., 36* (ed. Aad van der Vaart Mathisca de Gunst Chris Klaassen), pp. 286-309. Cambridge University Press.

[25] Hall, P. (1986). On the bootstrap and confidence intervals, *Ann. Stat.* **14**, 1431-1452.

[26] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[27] Hall, P., Horowitz, J. and Jing, B.-Y. (1996). On blocking rules for the bootstrap with dependent data, *Biometrika*, vol. 50, pp. 561-574.

[28] Hall, P. and Martin, M. (1988). On the bootstrap and two-sample problems, *Australian Journal of Statistics*, vol. 30A, 179-192.

[29] Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing, *Biometrics*, vol. 47, 757-762.

[30] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.* 19, 293-325.

[31] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, 58, 13-30.

[32] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* **17**, 1217-1241.

[33] Lehmann, E.L. and Romano, J. (2005). *Testing Statistical Hypotheses*, 3rd edition, Springer, New York.

[34] Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence, in *Exploring the limits of bootstrap*, eds R. LePage and L. Billard, pp. 225-248. Wiley, New York.

[35] Logan, B.F., Mallows, C.L., Rice, S.O., and Shepp, L.A. (1973). Limit distributions of self-normalized sums, *Annals Probab.*, 1, 788-809.

[36] MacKinnon, J.G. (2011). Thirty years of Heteroskedasticity-robust inference, Working paper, Dept. of Economics, Queen's University, Ontario, Canada.

[37] Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, **75**, 1411-1452; see also Corrigendum: *Econometrica*, **78**.

[38] Miller, R.G. (1974). The jackknife—a review, *Biometrika*, 61, pp. 1-15.

[39] Nolan, J.P. (2010). *Stable Distributions - Models for Heavy Tailed Data*, Birkhäuser, Boston. Manuscript in progress; Chapter 1 available online at: academic2.american.edu/∼jpnolan.

[40] Pfanzagl, J. (1974). On the Behrens-Fisher problem, *Biometrika*, 61, 39-47.

[41] Paparoditis, E. and Politis, D.N. (2005). Bootstrap hypothesis testing in regression models, *Statist. Prob. Letters*, vol. 74, pp. 356-365.

[42] Politis, D.N. and Romano, J.P. (1993). 'Estimating the Distribution of a Studentized Statistic by Subsampling', *Bulletin of the International Statistical Institute,* 49th Session, Firenze, August 25 - September 2, 1993, Book 2, pp.315-316.

[43] Politis, D.N., and Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions, *Ann. Statist.*, vol. 22, 2031-2050.

[44] Politis, D.N., and Romano, J.P. (1996). 'Subsampling for Econometric Models – Comments on Bootstrapping Time Series Models', *Econometric Rev.,* vol. 15, No. 2, pp. 169-176.

[45] Politis, D.N., and Romano, J.P. (2008), K-sample subsampling, in *Functional and Operatorial Statistics*, (S. Dabo-Niang and F. Ferraty, Eds.), Physica-Verlag (Springer), Heidelberg, pp. 247-254.

[46] Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling,* Springer, New York.

[47] Quenouille, M.H. (1949). Approximate tests of correlation in time-series, *J. R. Statist. Soc. B.* **11**, 68-84.

[48] Quenouille, M.H. (1956). Notes on bias in estimation, *Biometrika* **43**, 353-360.

[49] Romano, J. and Shaikh, A. (2010). On the uniform asymptotic validity of subsampling and the bootstrap. Technical Report 2010-03, Department of Statistics, Stanford University.

[50] Romano, J., Shaikh, A. and Wolf, M. (2008). Control of the false discovery rate under dependence using subsampling and the bootstrap (with discussion). *Test* **17**, 417-442, 461-471.

[51] Romano, J. and Wolf, M. (2010). Balanced control of generalized error rates. *Annals of Statistics* **38**, 598-633.

[52] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

[53] Shao, J. and Wu, C.F.J. (1989) A general theory for jackknife variance estimation. *Ann. Statist.* **17**, 1176-1197.

[54] Tukey J.W. (1958). Bias and condence in not quite so large samples. *Annals of Mathematical Statistics* **29**, 614.

[55] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*, Springer, New York.

[56] Welch, B.L. (1947). The generalization of 'Student's' problem when several different population variances are involved, *Biometrika*, 34, 28-35.

[57] Wu, C.F.J., (1986). Jackknife, bootstrap and other resampling plans in regression analysis (with discussion.) *Ann. Statist.* **14** 1261-1350.