Improved quantile inference via fixed-smoothing asymptotics and Edgeworth expansion

David M. Kaplan*

Abstract

Estimation of a sample quantile's variance requires estimation of the probability density at the quantile. The common quantile spacing method involves smoothing parameter m. When $m, n \to \infty$, the corresponding Studentized test statistic is asymptotically $N(0,1)$. Holding m fixed asymptotically yields a nonstandard distribution dependent on m. Closer examination reveals that this fixed-m distribution contains the Edgeworth expansion term capturing the variance of the quantile spacing. Consequently, the fixed-m distribution is more accurate than the standard normal under both asymptotic frameworks. With the Edgeworth expansion, I approximate the type I and II error rates of the test with fixed-m critical values. I present a plug-in expression for the m that minimizes type II error subject to controlling type I error. Compared with similar methods, the new method controls size better and maintains good or better power in simulations, and there are cases where it outperforms all existing methods. In parallel throughout are results for two-sample quantile treatment effect inference, such as testing for median equality of treatment and control groups. Code is available online.

JEL classification: C01, C12, C21

Keywords: quantile inference; fixed-smoothing asymptotics; Edgeworth expansion; hypothesis testing; testing-optimal smoothing parameter

---

*Email: KaplanDM@Missouri.edu. Thanks to Yixiao Sun for much patient and enthusiastic guidance and discussion. Thanks also to Karen Messer, Dimitris Politis, Andres Santos, Hal White, Brendan Beare, Ivana Komunjer, Joel Sobel, Marjorie Flavin, Kelly Paulson, Matt Goldman, Linchun Chen, and other attentive and interactive listeners from UCSD audiences. Thanks to the anonymous referees and the editors for helpful revision ideas. All errors are mine. Mail: Department of Economics, University of Missouri–Columbia, 909 University Ave, 118 Professional Bldg, Columbia, MO 65211-6040, USA.

1 Introduction

This paper considers inference on population quantiles, specifically via the Studentized test statistic from Siddiqui (1960) and Bloch and Gastwirth (1968) (jointly SBG hereafter), as well as inference on quantile treatment effects in a two-sample (treatment/control) setup. Median inference is a special case. Nonparametric inference on conditional quantiles is an immediate extension if the conditioning variables are all discrete. In addition to common variables like race and sex, variables like years of education may be treated as discrete as long as enough observations exist for each value of interest. Continuous conditioning variables are accommodated by the approach in Kaplan (2012), whose code includes an implementation of this paper's method. Like a t-statistic for the mean, the SBG statistic is normalized using a consistent estimate of the variance of the quantile estimator, and it is asymptotically standard normal.

Quantile treatment effects can enrich the usual average treatment effect analysis in economic experiments, such as those in Björkman and Svensson (2009), Charness and Gneezy (2009), and Gneezy and List (2006). Quantile treatment effects have also been discussed recently in Bitler et al. (2006) for welfare programs, in Djebbari and Smith (2008) for the PROGRESA conditional cash transfer program in Mexico, and in Jackson and Page (2013) for heterogeneous effects of class size in the Tennessee STAR program. Quantiles have proved of interest across many other fields within economics, such as health (Abrevaya, 2001), labor (Angrist et al., 2006; Buchinsky, 1994), auctions (Guerre and Sabbah, 2012), and land valuation (Koenker and Mizera, 2004). In finance, value-at-risk (VaR) is defined as a quantile; this paper's method can construct a confidence interval for the VaR, using a "historical simulation" approach or that in §3 of Cabedo and Moya (2003). In insurance, the Scenario Upper Loss (SUL) is an upper quantile of the property loss distribution given an upper quantile magnitude earthquake; the less consistently defined probable maximum loss (PML) is similar and also widely used.

I develop new asymptotic theory that is more accurate than the standard normal reference distribution traditionally used with the SBG statistic. With more accuracy comes improved inference properties, i.e. controlling size where previous methods were size-distorted without sacrificing much power. A plug-in procedure to choose the testing-optimal smoothing parameter m translates the theory into practice. Inverting the level-α tests proposed here yields level-α confidence intervals. I use hypothesis testing language throughout, but size distortion is analogous to coverage probability error, and higher power corresponds to shorter interval lengths. It is also straightforward to solve numerically for p-values using the higher-order critical value approximation since it is a simple arithmetic correction of standard normal critical values.

The two key results here are a nonstandard "fixed-m" asymptotic distribution (where m is a smoothing parameter used for Studentization) and a high-order Edgeworth expansion. For a scalar location model, Siddiqui (1960) gives the fixed-m result, and Hall and Sheather (1988, hereafter cited as "HS88") give a special case of the Edgeworth expansion in Theorem 2 below. The Edgeworth expansion is more accurate than a standard normal since it contains high-order terms that are otherwise ignored.

In the standard "large-m" asymptotics, both $m \to \infty$ and $n \to \infty$ (and $m/n \to 0$). In contrast, fixed-m asymptotics only approximates $n \to \infty$ while fixing m at its actual finite-sample value. It turns out that the fixed-m asymptotics includes the high-order Edgeworth term capturing the variance of the quantile spacing in the Studentized quantile's denominator. (Fixed-m is an instance of "fixed-smoothing" asymptotics in that this variance does not go to zero in the limit as in the "increasing-smoothing" large-m asymptotics.) Thus, from the theoretical view, the fixed-m distribution is more accurate than the standard normal irrespective of the asymptotic framework: it is high-order accurate under large-m, while the standard normal approximation is not even first-order accurate under fixed-m.

The Edgeworth and fixed-m distributions also capture the effect of the choice of m, whereas the standard normal approximation does not. I construct a test dependent on m using the fixed-m critical values and evaluate the type I and II error rates of the test using the more accurate Edgeworth expansion. Then I optimally select m to minimize type II error subject to control of type I error.

This work builds partly on Goh (2004), who suggests fixed-m asymptotics for Studentized quantile inference. Goh (2004) uses simulated critical values and prior choices of m. Here, I provide a simple, accurate fixed-m critical value approximation and corresponding new optimal choice of m. I also show the improved theoretical accuracy of the fixed-m distribution, complementing the simulations in Goh (2004).

In the time series context of heteroskedasticity-autocorrelation robust inference, the key ideas of "testing-optimal" smoothing parameter choice, fixed-smoothing (or "fixed-b") asymptotics, higher-order asymptotics, and corrected critical values based on a common distribution appear in a sequence of papers by Sun et al. (2008), Sun (2010, 2011, 2013), and Sun and Kaplan (2011). In particular, analogous to the fixed-m results below, Sun et al. (2008) show that their testing-optimal bandwidth is of a different order than the MSE-optimal bandwidth, and Sun (2011) generalizes the result from Sun et al. (2008) that shows fixed-smoothing asymptotics to be a higher-order refinement of increasing-smoothing asymptotics.

Whereas in HS88 the choice of m is critical to controlling size (or not), the fixed-m critical values provide size control robust to incorrectly chosen m. In simulations, the new method has correct size even where the HS88 method is size-distorted. Power is still good because m is explicitly chosen to maximize it, using the Edgeworth expansion. HS88 do not provide a separate result for the two-sample (quantile treatment effect) case.

Monte Carlo simulations show that the new method controls size better than HS88 and various bootstrap methods, while maintaining competitive power. Following a fractional order statistic approach, the one-sample method of Hutson (1999) and the related two-sample method of Goldman and Kaplan (2012) usually have better properties but are not always computable. Thus, they complement the new method herein. The new method also has a computational advantage over both methods, especially the two-sample method.

The basic setup is in Section 2. Sections 3 and 4 concern fixed-m and Edgeworth results, respectively. A consequently testing-optimal choice of m is proposed in Section 5. Section 6 contains simulation results. Two-sample results are provided in parallel. More details are in the working paper version, and full technical proofs and calculations are in the supplemental appendices. These and computer code are available on the author's website.

2 Quantile estimation and hypothesis testing

Consider an iid sample of continuous random variable X, whose p-quantile is ξ_p. The estimator $\hat{\xi}_p$ used in SBG is an order statistic. The rth order statistic for a sample of n values is the rth smallest value in the sample, $X_{n,r}$, such that $X_{n,1} < X_{n,2} < \cdots < X_{n,n}$. The SBG estimator is

$$\hat{\xi}_p = X_{n,r}, \quad r = \lfloor np \rfloor + 1,$$

where $\lfloor np \rfloor$ is the floor function. Writing the cumulative distribution function (CDF) of X as $F(x)$ and the probability density function (PDF) as $f(x) \equiv F'(x)$, I make the usual assumption that $f(x)$ is positive and continuous in a neighborhood of the point ξ_p. Consequently, ξ_p is the unique p-quantile such that $F(\xi_p) = p$.

The standard asymptotic result (Mosteller, 1946; Siddiqui, 1960) for a central[1] quantile estimator is

$$\sqrt{n}(X_{n,r} - \xi_p) \xrightarrow{d} N\big(0,\ p(1-p)[f(\xi_p)]^{-2}\big). \tag{1}$$

A consistent estimator of $1/f(\xi_p)$ that is asymptotically independent of $X_{n,r}$ leads to the Studentized sample quantile, which has the pivotal asymptotic distribution

$$\frac{\sqrt{n}(X_{n,r} - \xi_p)}{\sqrt{p(1-p)}\left[\widehat{1/f(\xi_p)}\right]} \xrightarrow{d} N(0,1). \tag{2}$$

Siddiqui (1960) and Bloch and Gastwirth (1968) propose and show consistency of

$$\widehat{1/f(\xi_p)} = S_{m,n} \equiv \frac{n}{2m}(X_{n,r+m} - X_{n,r-m}) \tag{3}$$

when $m \to \infty$ and $m/n \to 0$ as $n \to \infty$.

For two-sided inference on ξ_p, I consider the parameterization $\xi_p = \beta - \gamma/\sqrt{n}$. The null and alternative hypotheses are $H_0 : \xi_p = \beta$ and $H_1 : \xi_p \neq \beta$, respectively. When $\gamma = 0$, the null is true. The test statistic examined in this paper is

$$T_{m,n} \equiv \frac{\sqrt{n}(X_{n,r} - \beta)}{S_{m,n}\sqrt{p(1-p)}} \tag{4}$$

and will be called the SBG test statistic due to its use of (3). From (2), $T_{m,n}$ is asymptotically standard normal when $\gamma = 0$. The corresponding hypothesis test

[1] "Central" means that in the limit, $r/n \to p \in (0,1)$ as $n \to \infty$; i.e., $r \to \infty$ and is some fraction of the sample size n. In contrast, "intermediate" would take $r \to \infty$ but $r/n \to 0$ (or $r/n \to 1$, $n - r \to \infty$); "extreme" would fix $r < \infty$ or $n - r < \infty$.

compares $T_{m,n}$ to critical values from a standard normal distribution.

For the two-sample case, assume that there are independent samples of X and Y, with n_x and n_y observations, respectively. For simplicity, let $n = n_x = n_y$. For instance, if $2n$ individuals are separated into balanced treatment and control groups, one might want to test if the treatment effect at quantile p has significance level α. The marginal PDFs are $f_X(\cdot)$ and $f_Y(\cdot)$. Under the null hypothesis $H_0 : \xi_{px} = \xi_{py} = \xi_p$,

$$\sqrt{n}(X_{n,r} - \xi_p) - \sqrt{n}(Y_{n,r} - \xi_p) \xrightarrow{d} N\big(0, p(1-p)([f_X(\xi_p)]^{-2} + [f_Y(\xi_p)]^{-2})\big),$$

using (1) and independence. The asymptotic pivot for the two-sample case is thus

$$\frac{\sqrt{n}(X_{n,r} - Y_{n,r})}{\sqrt{[f_X(\xi_p)]^{-2} + [f_Y(\xi_p)]^{-2}}\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

The Studentized version uses the same quantile spacing estimators as above:

$$\tilde{T}_{m,n} \equiv \frac{\sqrt{n}(X_{n,r} - Y_{n,r})}{\sqrt{\left(\frac{n}{2m}\right)^2(X_{n,r+m} - X_{n,r-m})^2 + \left(\frac{n}{2m}\right)^2(Y_{n,r+m} - Y_{n,r-m})^2}\sqrt{p(1-p)}}. \tag{5}$$

The same m is used for X and Y in anticipation of a Gaussian plug-in approach that would yield the same m for X and Y regardless. This two-sample setup extends to unequal sample sizes $n_x \neq n_y$ by starting with $\sqrt{n_x}(X_{n,r} - \xi_p) - \sqrt{n_x/n_y}\sqrt{n_y}(Y_{n,r} - \xi_p)$ and assuming n_x/n_y is constant asymptotically, but this is omitted for clarity.

3 Fixed-m asymptotics and corrected critical value

Since $m < \infty$ in any given sample, holding m fixed as $n \to \infty$ may more accurately approximate the true finite-sample distribution of $T_{m,n}$. With $m \to \infty$, $\text{Var}(S_{m,n}) = O(1/m) \to 0$ ("increasing-smoothing" asymptotics); with m fixed, that variance does not disappear asymptotically ("fixed-smoothing"). The fixed-m distribution is given below, along with a simple formula for critical values that doesn't require simulation.

3.1 Fixed-m asymptotics

Siddiqui (1960, eqn. 5.4) provides the fixed-m asymptotic distribution; Goh (2004, Appendix D.1) provides a nice alternative proof for the median, which readily extends to any quantile.[2] If $\gamma = 0$,

$$T_{m,n} \xrightarrow{d} \mathcal{Z}/\mathcal{V}_{4m} \equiv T_{m,\infty} \quad \text{as } n \to \infty,\ m \text{ fixed}, \tag{6}$$

with $\mathcal{Z} \sim N(0,1)$, $\mathcal{V}_{4m} \sim \chi^2_{4m}/(4m)$, $\mathcal{Z} \perp\!\!\!\perp \mathcal{V}_{4m}$, and $S_{m,n}$ and $T_{m,n}$ as in (3) and (4).

The above distribution is conceptually similar to the Student's t-distribution. A t-statistic from normally distributed iid data has a standard normal distribution if either the variance is known (so denominator is constant) or the sample size approaches infinity. The distribution $T_{m,\infty}$ is also standard normal if either the variance is known (denominator in $T_{m,n}$ is constant) or $m \to \infty$. With estimated variance, in the t-statistic, the more accurate t-distribution has fatter tails than a standard normal: $t_v \sim \mathcal{Z}/\sqrt{\mathcal{V}_v}$, where $\mathcal{Z} \sim N(0,1)$, $\mathcal{V}_v \sim \chi^2_v/v$, and $\mathcal{Z} \perp\!\!\!\perp \mathcal{V}_v$. This is similar to (6). The fixed-m distribution reflects uncertainty in the variance estimator $S_{m,n}$ that is lost under the standard large-m asymptotics.

For the two-sample test statistic under the null, as $n \to \infty$ with m fixed,

$$\tilde{T}_{m,n} \xrightarrow{d} \frac{\mathcal{Z}}{\mathcal{U}} \equiv \tilde{T}_{m,\infty}, \quad \text{where}$$
$$\mathcal{U} \sim (1+\epsilon)^{1/2}, \quad \epsilon \equiv \lambda(\mathcal{V}^2_{4m,1} - 1) + (1-\lambda)(\mathcal{V}^2_{4m,2} - 1), \quad \lambda = \frac{1}{1 + [f_X(\xi_p)/f_Y(\xi_p)]^2},$$

and $\mathcal{Z}$, $\mathcal{V}_{4m,1}$, and $\mathcal{V}_{4m,2}$ are mutually independent. The derivation is in the supplemental appendix. The strategy is the same, using results from Siddiqui (1960).

Unlike the one-sample case, $\tilde{T}_{m,\infty}$ is not pivotal. We can either consider the upper bound for critical values or estimate the nuisance parameter $f_X(\xi_p)/f_Y(\xi_p)$.

[2] The result stated for a general quantile in Theorem 2 in Section 3.5 appears to have a typo and not follow from a generalization of the proof given, nor does it agree with Siddiqui (1960).

(The same two-sample nuisance parameter arises in Goldman and Kaplan (2012).) Note that for any random variables $W_1 \perp\!\!\!\perp W_2$ with $\mathrm{Var}(W_1) = \mathrm{Var}(W_2) = \sigma_W^2$ and $\lambda \in [0, 1]$, the variance of $\lambda W_1 + (1 - \lambda)W_2$ is $\sigma_W^2[\lambda^2 + (1 - \lambda)^2]$. This attains a maximum value σ_W^2 if $\lambda \in \{0, 1\}$ and minimum value $\sigma_W^2/2$ if $\lambda = 1/2$. This means the variance of ϵ (and thus of $\mathcal{U}$ and $\tilde{T}_{m,\infty}$) is smallest when $\lambda = 1/2$, i.e. when $f_X(\xi_p) = f_Y(\xi_p)$. When $\lambda \to 0$ (or one), which maximizes the variance of ϵ, we approach the special case of testing against a constant, $f_X(\xi_p)/f_Y(\xi_p) \to \infty$ (recall $f_X(\xi_p) > 0$ and $f_Y(\xi_p)$ are assumed). Then, $\tilde{T}_{m,\infty} = T_{m,\infty}$, so critical values from the one-sample case provide conservative inference in the two-sample case.

3.2 Corrected critical value

Denote the standard normal CDF and PDF as $\Phi(\cdot)$ and $\phi(\cdot)$. Approximating the fixed-m CDF around $\Phi(\cdot)$ yields critical values based on the standard normal ones.

Lemma 1. *The limiting distribution in* (6) *is approximated by*

$$P(T_{m,\infty} < z) = \Phi(z) - \frac{z^3\phi(z)}{4m} + O(m^{-2}). \tag{7}$$

Proof. The first three central moments of the χ^2_{4m} distribution are $4m$, $8m$, and $32m$, respectively. The independence of $\mathcal{Z}$ and $\mathcal{V}_{4m}$ allows rewriting the CDF in terms of $\Phi(\cdot)$, and then it is expanded around $E(\mathcal{V}_{4m}) = 1$. Using (6), for critical value z,

$$\begin{aligned}
P(T_{m,\infty} < z) &= P(\mathcal{Z}/\mathcal{V}_{4m} < z) = E[\Phi(z\mathcal{V}_{4m})] = E[\Phi(z + z(\mathcal{V}_{4m} - 1))] \\
&= E\big[\Phi(z) + \Phi'(z)z(\mathcal{V}_{4m} - 1) + (1/2)\Phi''(z)[z(\mathcal{V}_{4m} - 1)]^2 + O(m^{-2})\big] \\
&= \Phi(z) - \frac{z^3\phi(z)}{2}\underbrace{E\left[\left(\frac{\chi^2_{4m} - 4m}{4m}\right)^2\right]}_{8m/(16m^2)} + O(m^{-2}).
\end{aligned}$$

□

The approximation in (7) is a distribution function itself: its derivative in z is

$[\phi(z)/(4m)][z^4 - 3z^2 + 4m]$, which is positive for all z when $m > 9/16$ (as it always is); the limits at $-\infty$ and ∞ are zero and one; and it is càdlàg since it is differentiable everywhere. The approximation error $O(m^{-2})$ does not change if m is fixed, but it goes to zero as $m \to \infty$, which the selected m does as $n \to \infty$ (see Section 5.3). So it is reasonable to claim uniform convergence of $\Phi(z) - z^3\phi(z)/(4m)$ to $P(T_{m,\infty} < z)$ over $z \in \mathbb{R}$ as $m \to \infty$ via Pólya's Theorem (e.g., DasGupta, 2008, Thm. 1.3(b)). Appendix A shows (7) to be accurate for $m > 2$; for $m \leq 2$, simulated critical values are used in the provided code.

An approximate, upper one-sided level-α critical value is found by solving for z in $1 - \alpha = \Phi(z) - z^3\phi(z)/(4m) + O(m^{-2})$. Let $z_{1-\alpha} \equiv \Phi^{-1}(1-\alpha)$. If $z = z_{1-\alpha} + c/m$,

$$c = z_{1-\alpha}^3/4 + O(m^{-1}), \quad z = z_{1-\alpha} + c/m = z_{1-\alpha} + \frac{z_{1-\alpha}^3}{4m} + O(m^{-2}). \tag{8}$$

For a symmetric two-sided test, since the additional term in (7) is an odd function of z, the critical value is the same but with $z_{1-\alpha/2}$, yielding

$$z = z_{\alpha,m} + O(m^{-2}), \quad z_{\alpha,m} \equiv z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3}{4m}. \tag{9}$$

Note $z_{\alpha,m} > z_{1-\alpha/2}$ and depends on m; e.g., $z_{.05,m} = 1.96 + 1.88/m$. Using the same method with an additional term in the expansion, as compared in Appendix A,

$$z = z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3}{4m} + \frac{z_{1-\alpha/2}^5 + 8z_{1-\alpha/2}^3}{96m^2} + O(m^{-3}).$$

The rest of this paper uses (9). The results in Section 5 also go through with the above third-order corrected critical value or with a simulated critical value.

In the two-sample case, under $H_0 : F_X^{-1}(p) = F_Y^{-1}(p) = \xi_p$, after calculating some moments (see supplemental appendix), the fixed-m distribution can be approximated by

$$
\begin{aligned}
&P(\tilde{T}_{m,\infty} < z) \\
&\quad = \Phi(z) + \phi(z) z E[\mathcal{U} - 1] + (1/2)(-z\phi(z)) z^2 E[(\mathcal{U} - 1)^2] + O(E[(\mathcal{U} - 1)^3]) \\
&\quad = \Phi(z) - \frac{\phi(z)}{4m}\left[z^3 \frac{[f_X(\xi_p)]^{-4} + [f_Y(\xi_p)]^{-4}}{\tilde{S}_0^4} - 2z \frac{[f_X(\xi_p)]^{-2}[f_Y(\xi_p)]^{-2}}{\tilde{S}_0^4} \right] + O(m^{-2}) \\
&\quad = \Phi(z) - \frac{\phi(z)}{4m}\left[z^3 \frac{1+\delta^4}{(1+\delta^2)^2} - 2z \frac{\delta^2}{(1+\delta^2)^2} \right] + O(m^{-2}), \\
&\tilde{S}_0 \equiv ([f_X(F_X^{-1}(p))]^{-2} + [f_Y(F_Y^{-1}(p))]^{-2})^{1/2}, \quad \delta \equiv f_X(\xi_p)/f_Y(\xi_p).
\end{aligned}
\tag{10}
$$

The corresponding two-sided critical value is

$$
\begin{aligned}
z &= \tilde{z}_{\alpha,m} + O(m^{-2}) \leq z_{\alpha,m} + O(m^{-2}), \\
\tilde{z}_{\alpha,m} &= z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3([f_X(\xi_p)]^{-4} + [f_Y(\xi_p)]^{-4}) - 2z_{1-\alpha/2}[f_X(\xi_p)]^{-2}[f_Y(\xi_p)]^{-2}}{4m\tilde{S}_0^4} \\
&= z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3(1+\delta^4) - 2z_{1-\alpha/2}\delta^2}{4m(1+\delta^2)^2}.
\end{aligned}
\tag{11}
$$

When the PDF of Y collapses toward a constant, $[f_Y(\xi_p)]^{-1} \to 0$ and so $\delta \to 0$; $P(\tilde{T}_{m,\infty} < z)$ reduces to (7), and $\tilde{z}_{\alpha,m}$ to $z_{\alpha,m}$. Thus, as an alternative to estimating $\delta = f_X(\xi_p)/f_Y(\xi_p)$, the one-sample critical value $z_{\alpha,m}$ provides conservative two-sample inference, as discussed in Section 3.1. Under the exchangeability assumption used for permutation tests, $\delta = 1$ and $\tilde{z}_{\alpha,m}$ attains its smallest possible value.

4 Edgeworth expansion

4.1 One-sample case

The result here includes the result of HS88 as a special case, and indeed the results match[3] when $\gamma = 0$. The case $\gamma \neq 0$ is required to calculate the type II error rate (Section 5.2). The u_i functions are indexed by γ, so $u_{1,0}$ is the same as u_1 from HS88 (p. 385), and similarly for $u_{2,0}$ and $u_{3,0}$. The assumptions are the same as in HS88.

Theorem 2. *Assume $f(\xi_p) > 0$ and that, in a neighborhood of ξ_p, f'' exists and satisfies a Lipschitz condition, i.e. for some $\epsilon > 0$ and all x, y sufficiently close to ξ_p, $|f''(x) - f''(y)| \leq \text{constant}|x - y|^\epsilon$. Suppose $m = m(n) \to \infty$ as $n \to \infty$, in such a manner that for some fixed $\delta > 0$ and all sufficiently large n, $n^\delta \leq m(n) \leq n^{1-\delta}$. Define $C \equiv \gamma f(\xi_p)/\sqrt{p(1-p)}$ and*

$$
\begin{aligned}
u_{1,\gamma}(z) \equiv\ & \frac{1}{6}\left(\frac{p}{1-p}\right)^{1/2} \frac{1+p}{p}(z^2 - 1) - C\sqrt{\frac{1-p}{p}}\left(1 - \frac{pf'(\xi_p)}{[f(\xi_p)]^2} - \frac{1}{2(1-p)}\right)z \\
& - \frac{1}{2}\left(\frac{p}{1-p}\right)^{1/2}\left(1 + \frac{f'(\xi_p)}{[f(\xi_p)]^2}(1-p)\right)z^2 \\
& - [(\lfloor np \rfloor + 1 - np) - 1 + (1/2)(1-p)][p(1-p)]^{-1/2}, \\
u_{2,\gamma}(z) \equiv\ & \frac{1}{4}\left[2Cz^2 - C^2 z - z^3\right], \text{ and} \\
u_{3,\gamma}(z) \equiv\ & \frac{3[f'(\xi_p)]^2 - f(\xi_p)f''(\xi_p)}{6[f(\xi_p)]^4}(z - C).
\end{aligned}
$$

It follows that for the test statistic $T_{m,n}$ in (4),

$$
\begin{aligned}
\sup_{-\infty<z<\infty} \Big| & P(T_{m,n} < z) - [\Phi(z+C) + n^{-1/2}u_{1,\gamma}(z+C)\phi(z+C) \\
& + m^{-1}u_{2,\gamma}(z+C)\phi(z+C) + (m/n)^2 u_{3,\gamma}(z+C)\phi(z+C)]\Big| \\
& = o[m^{-1} + (m/n)^2]. \qquad (12)
\end{aligned}
$$

[3]There appears to be a typo in the originally published result; the first part of the first term of u_1 in HS88 was $(1/6)[p(1-p)]^{1/2}$, but it appears to be $(1/6)[p/(1-p)]^{1/2}$ as given above.

A sketch of the proof is given in Appendix C (full proof in supplemental appendix).

Under the null hypothesis, the fixed-m distribution captures the high-order Edgeworth term associated with the variance of $S_{m,n}$. Plugging $\gamma = 0$ into (12) yields the term $m^{-1}u_{2,0}(z)\phi(z) = -z^3\phi(z)/(4m)$, identical to the m^{-1} term in Lemma 1. This means that the fixed-m distribution is high-order accurate under the conventional large-m asymptotics, while the standard normal distribution is only first-order accurate. The fixed-m distribution is also more accurate under fixed-m asymptotics, under which the standard normal is not even first-order accurate. Thus, theory strongly indicates that the fixed-m distribution and critical values are more accurate than the conventional standard normal ones. This is also borne out in simulations here and in Goh (2004).

4.2 Two-sample case

The strategy and results are similar, with full proof in the supplemental appendix.

Theorem 3. *Let $X_i \overset{iid}{\sim} F_X$ and $Y_i \overset{iid}{\sim} F_Y$, and $X_i \perp\!\!\!\perp Y_j \ \forall i,j$. Let the assumptions in Theorem 2 hold for both F_X and F_Y, and let $\tilde{C} \equiv \gamma\tilde{S}_0^{-1}/\sqrt{p(1-p)}$. Define*

$$
\begin{aligned}
\tilde{u}_{1,\gamma}(z) \equiv\ & \frac{1}{6}\frac{1+p}{\sqrt{p(1-p)}}\left(\frac{g_x^3 - g_y^3}{\tilde{S}_0^3}\right)(z^2-1) \\
& + \left[(a_2g_x^2 - a_2'g_y^2)(1-p) + (a_1g_x^2 - a_1'g_y^2)\right][p/(1-p)]^{1/2}(2p^2\tilde{S}_0^3)^{-1}(z^2) \\
& - \left[2(a_2g_x^2 - a_2'g_y^2) + (a_1g_x^2 - a_1'g_y^2)/(1-p)\right](2p\tilde{S}_0^3)^{-1}C[(1-p)/p]^{1/2}(z) \\
& + \left[(a_2g_x^2 - a_2'g_y^2) - \tilde{S}_0^2(a_2 - a_2')\right][p(1-p)]^{1/2}(2p^2\tilde{S}_0^3)^{-1} \\
& - \left((\lfloor np \rfloor + 1 - np) - 1 + \tfrac{1}{2}(1-p)\right)(g_x - g_y)\left[\tilde{S}_0\sqrt{p(1-p)}\right]^{-1}, \\
\tilde{u}_{2,\gamma}(z) \equiv\ & -\frac{1}{4}\left(\frac{g_x^4 + g_y^4}{\tilde{S}_0^4}\right)z^3 + \frac{1}{2}\frac{g_x^2g_y^2}{\tilde{S}_0^4}z - \frac{1}{2}\frac{g_x^2g_y^2}{\tilde{S}_0^4}\tilde{C} + \frac{1}{4}\left(\frac{g_x^4 + g_y^4}{\tilde{S}_0^4}\right)(2\tilde{C}z^2 - \tilde{C}^2z), \\
\tilde{u}_{3,\gamma}(z) \equiv\ & \frac{g_xg_x'' + g_yg_y''}{6\tilde{S}_0^2}(z - \tilde{C}),
\end{aligned}
$$

$$g_X(\cdot) \equiv 1/f_X(F_X^{-1}(\cdot)),\ g_x \equiv g_X(p),\ g''_x \equiv g''_X(p),$$
$$H_X(x) \equiv F_X^{-1}(e^{-x}),\ a_i \equiv H_X^{(i)}\left(\sum_{j=r}^{n} j^{-1}\right),$$
$$g_Y(\cdot) \equiv 1/f_Y(F_Y^{-1}(\cdot)),\ g_y \equiv g_Y(p),\ g''_y \equiv g''_Y(p),$$
$$H_Y(y) \equiv F_Y^{-1}(e^{-y}),\ a'_i \equiv H_Y^{(i)}\left(\sum_{j=r}^{n} j^{-1}\right),$$

where $H^{(i)}(\cdot)$ is the ith derivative of $H(\cdot)$. Then, for test statistic $\tilde{T}_{m,n}$ in (5),

$$\sup_{-\infty<z<\infty}\Big|P\left(\tilde{T}_{m,n} < z\right) - [\Phi(z+C) + n^{-1/2}\tilde{u}_{1,\gamma}(z+C)\phi(z+C) + m^{-1}\tilde{u}_{2,\gamma}(z+C)\phi(z+C) + (m/n)^2\tilde{u}_{3,\gamma}(z+C)\phi(z+C)]\Big| = o[m^{-1} + (m/n)^2].$$

For a two-sided test, $\tilde{u}_{1,\gamma}$ will cancel out, so the unknown terms therein may be ignored. Under the null, when $\tilde{C} = 0$, the m^{-1} term again exactly matches that from the fixed-m distribution, demonstrating that the fixed-m distribution is more accurate than the standard normal under large-m (as well as fixed-m) asymptotics.

5 Optimal smoothing parameter selection

I propose a "testing-optimal" choice of m to minimize (approximate) type II error subject to control of (approximate) type I error for a symmetric two-sided test, and a plug-in implementation thereof.

5.1 Type I error

I use the Edgeworth expansion with $\gamma = 0$ and critical value from (9) to approximate the type I error rate. The term $u_{1,0}(z)\phi(z)$ in (12) cancels out since it is an even function of z. The $u_{2,0}$ term is zeroed out by the m^{-1} term in the fixed-m critical value. Only the third high-order term from (12) remains, where $u_{3,0}(-z) = -u_{3,0}(z)$.

Proposition 4. *If $\gamma = 0$, then $P(|T_{m,n}| > z_{\alpha,m}) = e_I + o(m^{-1} + (m/n)^2)$,*

$$e_I = \alpha - 2(m/n)^2 u_{3,0}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}). \tag{13}$$

Proof. Under the null hypothesis ($\gamma = 0$), (12) yields

$$\begin{aligned} P(|T_{m,n}| > z \mid H_0) &= P(T_{m,n} > z \mid H_0) + P(T_{m,n} < -z \mid H_0) \\ &= 2 - 2\Phi(z) - 2m^{-1}u_{2,0}(z)\phi(z) \\ &\quad - 2(m/n)^2 u_{3,0}(z)\phi(z) + o[m^{-1} + (m/n)^2]. \end{aligned}$$

Using critical value $z_{\alpha,m} = z_{1-\alpha/2} + z_{1-\alpha/2}^3/(4m)$, the type I error rate is

$$\begin{aligned} P(|T_{m,n}| > z_{\alpha,m} \mid H_0) &= 2 - 2\Phi(z_{\alpha,m}) + 2z_{\alpha,m}^3\phi(z_{\alpha,m})/(4m) \\ &\quad - 2(m/n)^2 u_{3,0}(z_{\alpha,m})\phi(z_{\alpha,m}) + o[m^{-1} + (m/n)^2] \\ &= \alpha - 2(m/n)^2 u_{3,0}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + o[m^{-1} + (m/n)^2]. \ \square \end{aligned}$$

The dominant part of the type I error rate, e_I, depends on m, n, and $z_{1-\alpha/2}$, as well as $f(\xi_p)$, $f'(\xi_p)$, and $f''(\xi_p)$ through $u_{3,0}(z_{1-\alpha/2})$. Up to higher-order remainder terms, the type I error rate does not exceed nominal α if $u_{3,0}(z_{1-\alpha/2}) \geq 0$. Since $z_{1-\alpha/2} > 0$, the sign of $u_{3,0}(z_{1-\alpha/2})$ is the sign of $3f'(\xi_p)^2 - f(\xi_p)f''(\xi_p)$, or equivalently the sign of the third derivative of the inverse CDF, $\frac{\partial^3}{\partial p^3}F^{-1}(p)$. According to HS88, for $p = 0.5$, this is positive for all symmetric unimodal densities and most skew unimodal densities. Additionally, for any quantile p, the sign is positive for t-, normal, exponential, χ^2, and Fréchet distributions (calculations in supplemental appendix). Thus, in all these (and other) common cases, simply using the fixed-m critical value is sufficient to reduce e_I to α or below.

For the two-sample case (proof in supplemental appendix), the result is similar.

Proposition 5. *If $\gamma = 0$, then*

$$P\Big(|\tilde{T}_{m,n}| > z_{\alpha,m}\Big) \le P\Big(|\tilde{T}_{m,n}| > \tilde{z}_{\alpha,m}\Big) = \tilde{e}_I + o(m^{-1} + (m/n)^2),$$
$$\tilde{e}_I = \alpha - 2(m/n)^2 \tilde{u}_{3,0}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}).$$

The fixed-m critical value ensures $\tilde{e}_I \le \alpha$ for the same distributions as in the one-sample case. The one-sample critical value $z_{\alpha,m}$ gives conservative inference. The infeasible $\tilde{z}_{\alpha,m}$ has better power but requires estimating $\delta \equiv f_X(\xi_p)/f_Y(\xi_p)$ in practice. The $\tilde{z}_{\alpha,m}$ under exchangeability ($\delta = 1$) leads to the best power but also size distortion if exchangeability is violated.

5.2 Type II error

Similar to Section 5.1, I use the Edgeworth expansion in (12) to approximate the type II error rate of the test with critical value from (9). Since a uniformly most powerful test does not exist for general alternative hypotheses, I follow a common strategy in the optimal testing literature and pick a reasonable alternative hypothesis against which to maximize power. The hope is that this will produce a test near the power envelope at all alternatives, even if it is not strictly the uniformly most powerful.

I maximize power against the alternative where first-order power is 50%. The type II error rate is thus $0.5 = P(|T_{m,n}| < z_{1-\alpha/2}) \doteq G_{C^2}(z_{1-\alpha/2}^2)$, where G_{C^2} is the CDF of a noncentral χ^2 distribution with one degree of freedom and noncentrality parameter C^2, with C defined in Theorem 2. For $\alpha = 0.05$, this gives $\gamma f(\xi_p)/\sqrt{p(1-p)} \equiv C = \pm 1.96$, or $\gamma = \pm 1.96\sqrt{p(1-p)}/f(\xi_p)$.

Calculation of the following type II error rate may be found in the working paper.

Proposition 6. *If* C^2 *solves* $0.5 = G_{C^2}(z_{1-\alpha/2})$*, and writing* f *for* $f(\xi_p)$ *and similarly* f' *and* f''*, then*

$$
\begin{aligned}
P(|T_{m,n}| < z_{\alpha,m}) &= e_{II} + o(m^{-1} + (m/n)^2), \\
e_{II} &= 0.5 + (1/4)m^{-1}[\phi(z_{1-\alpha/2} - C) - \phi(z_{1-\alpha/2} + C)]Cz_{1-\alpha/2}^2 \\
&\quad + (m/n)^2 \frac{3(f')^2 - ff''}{6f^4} z_{1-\alpha/2}\big[\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)\big] \\
&\quad + O(n^{-1/2}). \qquad (14)
\end{aligned}
$$

There is a bias term and a variance term in e_{II} above. Per Bloch and Gastwirth (1968), the variance and bias of $S_{m,n}$ as $m \to \infty$ and $m/n \to 0$ are indeed of orders m^{-1} and $(m/n)^2$ respectively, as given in their (2.5) and (2.6):

$$
\mathrm{AsyVar}(S_{m,n}) = (2mf^2)^{-1} \quad \text{and} \quad \mathrm{AsyBias}(S_{m,n}) = (m/n)^2 \frac{3(f')^2 - ff''}{6f^5}.
$$

These are similar to (14) aside from the additional $1/f^2$ and $1/f$. As $m \to \infty$, $\mathrm{Var}(S_{m,n}) \to 0$ (increasing smoothing). With m fixed, this variance does not decrease to zero but rather is fixed (fixed smoothing) and consequently captured by the fixed-m asymptotics. The bias is not captured since it does decrease to zero in the fixed-m thought experiment.

The contribution to type II error from the m^{-1} term in (14) is decreasing in m, while that from the $(m/n)^2$ term is usually increasing in m. As discussed in Section 5.1, for common distributions $[3(f')^2 - ff'']/(6f^4) \geq 0$, so the entire $(m/n)^2$ term in (14) is positive. The m^{-1} term in (14) is also positive since $\phi(z_{1-\alpha/2} - C) > \phi(z_{1-\alpha/2} + C)$. This tension makes possible an "interior" solution m for minimizing e_{II}, i.e. $m \in [1, \min(r-1, n-r)]$.

The two-sample case is similar. Calculations are in the supplemental appendix.

Proposition 7. *If* C^2 *solves* $0.5 = G_{C^2}(z_{1-\alpha/2})$, $\theta \equiv \tilde{S}_0^{-4}(g_x^4 + g_y^4) = (1+\delta_a^4)/(1+\delta_a^2)^2$,

$g_x \equiv 1/f_X\big(F_X^{-1}(p)\big)$ *and* $g_y \equiv 1/f_Y\big(F_Y^{-1}(p)\big)$ *are as defined in Theorem 3, as are* g_x'' *and* g_y'', $\delta_a \equiv f_X\big(F_X^{-1}(p)\big)/f_Y\big(F_Y^{-1}(p)\big)$ *similar to* (10), *and* $\tilde{S}_0 \equiv ([f_X(F_X^{-1}(p))]^{-2} + [f_Y(F_Y^{-1}(p))]^{-2})^{1/2}$ *as in* (10), *then*

$$
\begin{aligned}
P(|\tilde{T}_{m,n}| < \tilde{z}_{\alpha,m}) &= \tilde{e}_{II} + o(m^{-1} + (m/n)^2), \\
\tilde{e}_{II} = 0.5 + \tfrac{1}{4}m^{-1}\big\{&\phi(z_{1-\alpha/2} + C)[-\theta C z_{1-\alpha/2}^2 + (1-\theta)z_{1-\alpha/2}] \\
&+ \phi(z_{1-\alpha/2} - C)[\theta C z_{1-\alpha/2}^2 + (1-\theta)z_{1-\alpha/2}]\big\} \\
&+ (m/n)^2 \frac{g_x g_x'' + g_y g_y''}{6\tilde{S}_0^2} z_{1-\alpha/2}\big\{\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)\big\} \\
&+ O(n^{-1/2}).
\end{aligned}
$$

5.3 Choice of m

With the fixed-m critical values, $e_I \le \alpha$ for all quantiles p of common distributions, as discussed in Section 5.1. Since this type I error control is robust to smoothing parameter choice, m is chosen to minimize e_{II}. The m^{-1} and m^2 components of e_{II} are both positive, yielding a U-shaped function of m (for $m > 0$). Consequently, if the first-order condition yields an infeasibly big m, the biggest feasible m is the minimizer.

For a randomized alternative (as in Sun, 2010) for a symmetric two-sided test, the first-order condition of (14) leads to

$$
\begin{aligned}
m = \Big\{&(1/4)\{\phi(z_{1-\alpha/2} - C) - \phi(z_{1-\alpha/2} + C)\}C z_{1-\alpha/2}^2 \\
&/\{(2/n^2)\frac{3(f')^2 - f f''}{6f^4} z_{1-\alpha/2}\big[\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)\big]\}\Big\}^{1/3}.
\end{aligned}
$$

Remember that C is chosen ahead (such as $C = 1.96$), ϕ is the standard normal PDF, $z_{1-\alpha/2}$ is determined by α, and n is known for any given sample; but the object $[3(f')^2 - ff'']/(6f^4)$ is unknown.

As is common in the kernel bandwidth selection literature and in implementation of HS88, I plug in the standard normal PDF ϕ for f; more precisely, $\phi(\Phi^{-1}(p))$ for $f(\xi_p)$. Here, using $N(0,1)$ is mathematically equivalent to $N(\mu,\sigma^2)$ since μ and σ cancel out. Even when this plug-in m is different than optimal, the power loss can be small; e.g., with Unif$(0,1)$, $n=45$, $p=0.5$, and $\alpha=0.05$, the (simulated) power loss is only 4% at the alternative with $C=1.96$. The Gaussian plug-in yields

$$m_K(n,p,\alpha,C) = n^{2/3}(Cz_{1-\alpha/2})^{1/3}(3/4)^{1/3}\left(\frac{(\phi(\Phi^{-1}(p)))^2}{2(\Phi^{-1}(p))^2+1}\right)^{1/3} \times \left(\frac{\phi(z_{1-\alpha/2}-C)-\phi(z_{1-\alpha/2}+C)}{\phi(z_{1-\alpha/2}-C)+\phi(z_{1-\alpha/2}+C)}\right)^{1/3},$$

and for $C=1.96$ (see Section 5.2) and $\alpha=0.05$,

$$m_K(n,p,\alpha=0.05,C=1.96) = n^{2/3}(1.42)\left(\frac{(\phi(\Phi^{-1}(p)))^2}{2(\Phi^{-1}(p))^2+1}\right)^{1/3}. \tag{15}$$

Compare (15) with the suggested m from (3.1) in HS88. In my notation,

$$m_{HS} = \left\lfloor n^{2/3} z_{1-\alpha/2}^{2/3}\left[1.5f^4/(3(f')^2 - ff'')\right]^{1/3}\right\rfloor.$$

HS88 proceed with a median-specific, data-dependent method, but a Gaussian plug-in is more common (e.g., Goh and Knight, 2009; Koenker and Xiao, 2002):

$$m_{HS} = n^{2/3} z_{1-\alpha/2}^{2/3}\left(1.5\frac{[\phi(\Phi^{-1}(p))]^2}{2(\Phi^{-1}(p))^2+1}\right)^{1/3}. \tag{16}$$

With $\alpha=0.05$, $z_{1-\alpha/2}^{2/3}(1.5)^{1/3}=1.79$, so $m_K(C=1.96)=0.79m_{HS}$. This ratio remains in the interval $[0.70,0.90]$ for C from alternatives with first-order power between 30% and 80%, and in the interval $[0.52,0.98]$ when between 10% and 95%. Since m_K and m_{HS} are optimized for different critical values, it makes sense that $m_K < m_{HS}$: the fixed-m critical value controls size, allowing smaller m_K when e_I might be a binding constraint for m_{HS}.

For the two-sample case, the strategy is the same. The testing-optimal $\tilde{m}_K$ is found by taking the first-order condition of (7) with respect to m and then solving for m. The two-sample $\tilde{e}_{II}$ is a U-shaped function of m due to positive m^{-1} and m^2 terms. With $\tilde{z}_{\alpha,m}$ from (11) and θ from (7), solving for m in the first-order condition,

$$\tilde{z}_{\alpha,m} : \tilde{m}_K = n^{2/3}(3/4)^{1/3}\left(\frac{\tilde{S}_0^2}{g_x g_x'' + g_y g_y''}\right)^{1/3} \times \left\{(1-\theta) + \theta C z_{1-\alpha/2}\frac{\phi(z_{1-\alpha/2}-C)-\phi(z_{1-\alpha/2}+C)}{\phi(z_{1-\alpha/2}-C)+\phi(z_{1-\alpha/2}+C)}\right\}^{1/3}.$$

I again use a Gaussian plug-in (using sample variances) for the g and $\tilde{S}_0$ terms due to the difficulty of estimating g''.

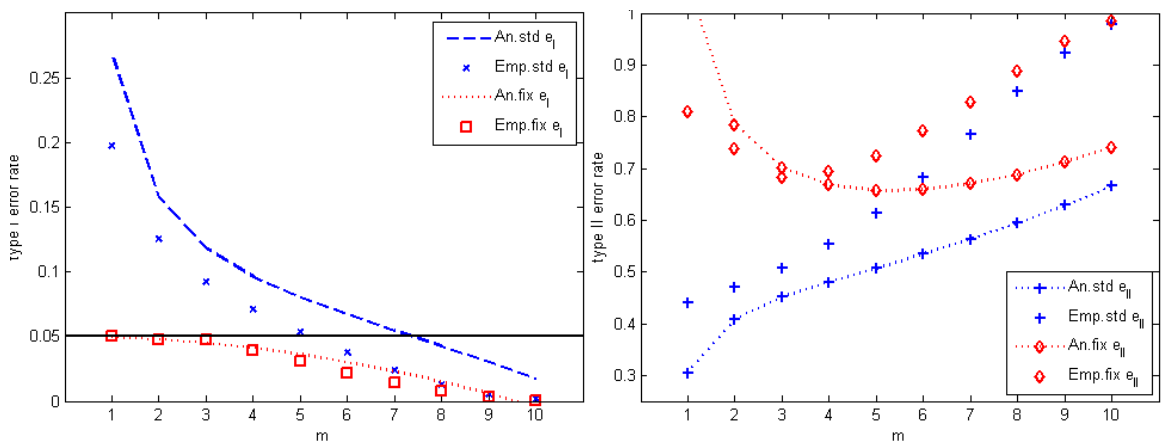

Figure 1: Analytic and empirical (simulated) e_I and e_{II} by m.
Both: $n = 21$, $p = 0.5$, $\alpha = 0.05$, F is log-normal with $\mu = 0$, $\sigma = 3/2$, 5000 replications. Lines are analytic; markers without lines are empirical. Legends: "An"=analytic (Edgeworth), "Emp"=empirical (simulated), "std"=$N(0,1)$ critical values, and "fix"=fixed-m critical values.
Right (e_{II}): randomized alternative (see text), $\gamma/\sqrt{n} = 0.80$.

Figure 1 compares the one-sample Edgeworth approximations to the true (simulated) type I and type II error rates for scalar data drawn from a log-normal distribution. Although for a given m, the larger fixed-m critical values reduce both size and power, this is not necessarily true in practice since $m_K < m_{HS}$. With standard normal critical values, the type I error rate is monotonically decreasing with m while the type II error rate is monotonically increasing, so setting $e_I = \alpha$ exactly would minimize e_{II} subject to $e_I \leq \alpha$. With fixed-m critical values, this monotonicity does not hold, so the Edgeworth expansion for $\gamma \neq 0$ is necessary. Additionally, the type I error rate is always below α (approximately, and usually truly), and the type II error rate curve near the selected m is very flat since it is near the minimum, where the slope is zero. Consequently, a larger-than-optimal m (due to the Edgeworth approximation error and/or the Gaussian plug-in) usually results in larger power loss for HS88 than for the new method. A smaller-than-optimal m can incur size distortion for HS88 but not for the new method.

Consider the size distortion of HS88 when $F = \text{Unif}(0,1)$, $n = 21$, and $\alpha = 0.05$. When $p = 0.5$, $m = 10$ would control size, but $m_{HS} = 7$, leading to size distortion. When $p = 0.2$, $m_{HS} = 4$, the biggest m possible to choose, but this still leads to size distortion. In contrast, fixed-m critical values ensure $e_I \leq \alpha$ for any m (for the uniform distribution, actually $e_I = \alpha$). More generally, when $r = \lfloor np \rfloor + 1$ is close to one or n, it may be impossible to control size when using standard normal critical values due to the restriction $m \leq \min(r-1, n-r)$.

6 Simulation study

MATLAB and R code implementing a hybrid of the new method and others is available on the author's website. MATLAB simulation code is also available on the author's website.

The following simulations compare five methods. The first method is this paper's, with fixed-m critical values and m_K chosen as in Section 5.3. For the two-sample

$\tilde{z}_{\alpha,m}$, I implemented an estimator of $\theta \equiv \tilde{S}_0^{-4}(g_x^4 + g_y^4)$ using quantile spacings; a better estimator would improve performance. The second method uses standard normal critical values and m_{HS} as in (16). The third method uses standard normal critical values and

$$m_B = n^{4/5}\left(4.5\frac{[\phi(\Phi^{-1}(p))]^4}{[2(\Phi^{-1}(p))^2+1]^2}\right)^{1/5}, \tag{17}$$

a Gaussian plug-in of Bofinger (1975), aiming to minimize the mean squared error of $S_{m,n}$. Since neither HS88 nor Bofinger (1975) provide values of m for the two-sample case, I used their one-sample m values, which should be conservative for two-sample inference (as discussed in Section 3). These three methods are referred to as "New method," "HS," and "B," respectively. For the new method, I used the floor function to get an integer value for m_K; for HS and B, I rounded to the nearest integer, to not amplify their size distortion.

The fourth method is a bootstrap ("BS") method. The conclusion of much experimentation was that a symmetric percentile-t using bootstrapped variance had the best size control in the simulations presented below. For the number of outer (B_1) and inner (B_2, for the variance estimator) bootstrap replications, $B_1 = 99$ (surprisingly) and $B_2 = 100$ performed best in these examples and thus were used; any exceptions are noted below.

The fifth method is a near-exact method from Hutson (1999), based on fractional order statistics. For one-sample inference, it performed best in simulations where it could be computed, so the results here focus on cases when it cannot be computed. Figure 2 shows when it can and cannot be computed, in terms of n and p. For two-sample inference, the related method in Goldman and Kaplan (2012) has similarly strong performance in practice and similar computational limitations. Additionally, its most accurate version relies on numerical approximation/simulation, which is slower, if not prohibitively slow for most economic applications. The simulation setups from HS88 all fall in the interior of Figure 2 ($p = 0.5$ with $n \in \{11, 25, 41, 71, 99\}$), so they are not reported; the new method and HS both control size in each case anyway, excepting slight (under one percentage point) over-rejection by HS for the uniform distribution.

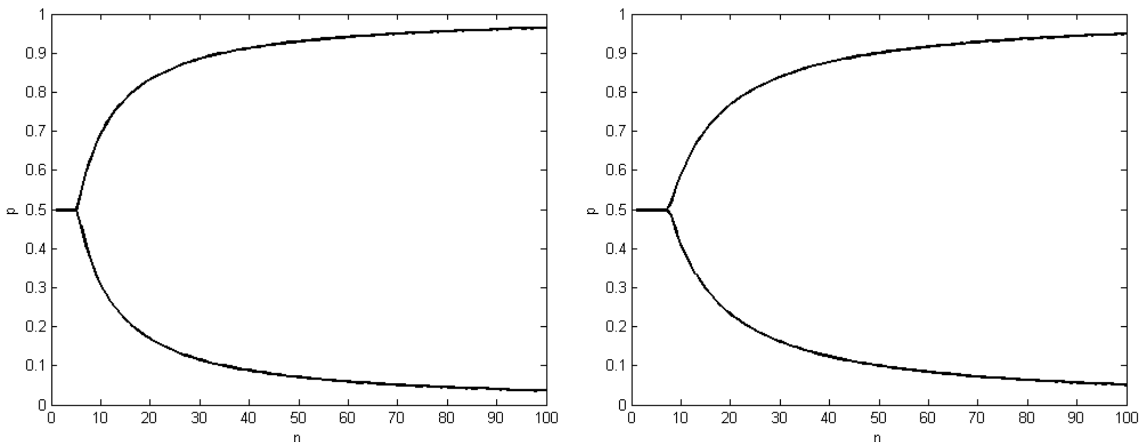

Figure 2: Graphs showing for which combinations of quantile $p \in (0, 1)$ and sample size n the equal-tailed Hutson (1999) method is computable (region between the black curves). Left: $\alpha = 5\%$. Right: $\alpha = 1\%$.

The hope of this paper was to produce a method that controls size better than HS and B (and others) while keeping power competitive. For one-sample inference in cases when Hutson (1999) is not computable, the following simulations show that the new method eliminates or reduces the size distortion of HS, B, and BS. The new method also has good power.

The two-sample results are more complicated since performance depends on how similar the two distributions are. The one-sample results represent one extreme of the two-sample case as the ratio $f_X(\xi_p)/f_Y(\xi_p)$ goes to infinity (or zero), when size is hardest to control. The other extreme is when $f_X(\xi_p) = f_Y(\xi_p)$. In that case, HS and B have less size distortion, but still some. BS, though, seems to control size and have

Table 1: Empirical size as percentage (nominal: 5.0), $n = 3, 4$, $p = 0.5$.

	one-sample						two-sample		
	$n = 3$			$n = 4$			$n = 3$		
Distribution	New	BS	HS/B	New	BS	HS/B	New[a]	BS	HS/B
N(0, 1)	1.6	1.2	11.9	2.4	4.1	15.4	0.8	1.3	6.3
Logn(0, 3/2)	3.3	1.3	12.8	1.8	3.6	10.0	0.4	0.8	4.5
Exp(1)	3.0	1.2	13.5	2.3	4.6	12.9	1.1	1.4	6.4
Uniform(0, 1)	3.2	1.5	16.0	5.0	6.8[b]	21.3	1.4	1.9	8.6
t_3	1.2	1.0	9.4	1.9	3.1	12.6	0.6	1.0	4.6

[a]If true θ used instead: 1.6, 1.1, 1.6, 2.8, 1.1.
[b]6.7% with 999 outer replications

somewhat better power than the new method with estimated θ. With the true θ, BS and the new method perform quite similarly. General two-sample setups would have results in between these extremes of $f_X/f_Y \to \infty$ (or zero) and $f_X/f_Y = 1$.

Unless otherwise noted, 10,000 simulation replications were used, and $\alpha = 5\%$.

Table 1 shows one- and two-sample size for $n = 3$ or $n = 4$. For nonparametric *conditional* quantiles, there may indeed be "cells" with such a small number of observations, even if the overall sample size is much larger. For example, if we have one thousand observations, we might have 500 in each cell conditioning on only one binary variable, but with eight binary conditioning variables (or four discrete variables with four categories each), there will be cells with only three or four observations. A similar argument applies to the number of observations in a kernel smoothing window when conditioning on continuous variables. For one-sample testing, HS and B are severely size-distorted while the new method and BS control size (except BS for $n = 4$ and uniform distribution). In the two-sample special case where the distributions are identical, all rejection probabilities are much smaller, though some size distortion remains for HS and B.

For one-sample power (Figure 3), BS is very poor for $n = 3$, but better than the

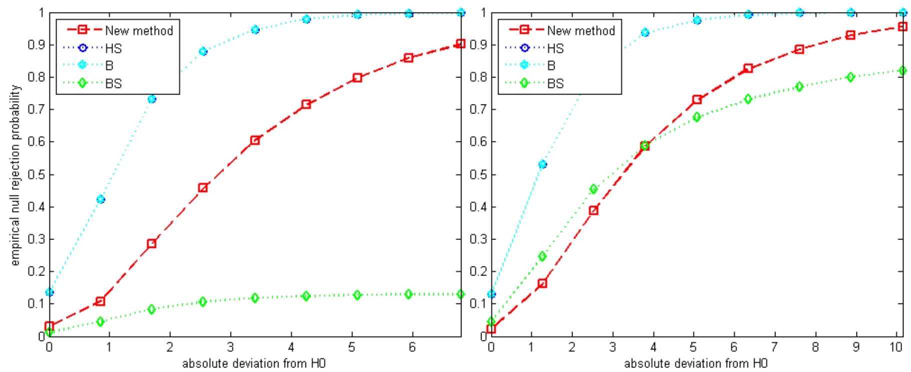

Figure 3: Empirical power properties, one-sample, Exp(1); $p = 0.5$, $n = 3$ (left) and $n = 4$ (right).

new method at some alternatives for $n = 4$. Two-sample power (not shown) with $n = 3$ is similar to one-sample power with $n = 4$, and BS is somewhat better than the new method for $n = 4$.

Unlike the sample mean, the sample median is not normally distributed even when the underlying distribution is normal. For a continuous distribution, order statistic $X_{n,r} \stackrel{d}{=} F_X^{-1}(U_{n,r})$, where $U_{n,r}$ is the rth order statistic from a sample drawn iid Uniform$(0,1)$. From combinatoric arguments, $U_{n,r} \sim \beta(r, n+1-r)$; this is an exact, finite-sample result. For a standard normal distribution, $X_{n,r} \sim \Phi^{-1}(\beta(r, n+1-r))$. For p closer to zero or one, as in Table 2, the beta distribution has more skew, and the sample quantile departs even farther from normality than for the median. From another perspective: Proposition 4 says that, up to the $o[m^{-1} + (m/n)^2]$ remainder, the new method's type I error rate is less than α by a term of order m^2/n^2 with a normal distribution. Any distribution with a smaller $u_{3,0}$ yields a type I error rate closer to α (up to remainder terms).

For $n = 45$ and $p = 0.95$ (or $p = 0.05$), the new method controls size except for a few tenths of a percentage point for the slash distribution. In contrast, BS, HS, and B are all size-distorted, sometimes significantly; see Table 2. With $n = 21$, even the

Table 2: Empirical size as percentage, $p = 0.95$, one-sample.

	$n = 45$					$n = 21$				
		BS; $B_1 =$					BS; $B_1 =$			
Distribution	New	99	999	HS/B	IM_2	New	99	999	HS/B	IM_2
$\text{N}(0,1)$	3.2	6.4	6.5	10.4	4.1	6.7	11.3	11.5	25.3	6.5
Slash	5.3	7.8	8.1	10.7	4.7	12.9	17.8	18.6	30.3	13.0
$\text{Logn}(0,3/2)$	5.0	7.3	7.5	10.7	4.4	11.3	14.6	15.2	28.7	11.3
$\text{Exp}(1)$	4.0	6.9	7.0	10.7	3.7	8.0	12.6	13.1	26.9	7.7
$\text{Unif}(0,1)$	3.3	5.4	5.3	11.2	4.9	4.7	6.8	6.7	22.3	5.1
$\text{GEV}(0,1,0)$	4.1	7.2	7.3	11.0	3.9	7.3	12.2	12.6	25.9	7.8
χ^2_1	3.6	6.5	6.4	10.2	4.3	8.7	13.5	13.9	27.5	8.2
t_3	3.6	6.5	7.1	9.8	4.1	8.3	13.5	14.0	26.7	8.3

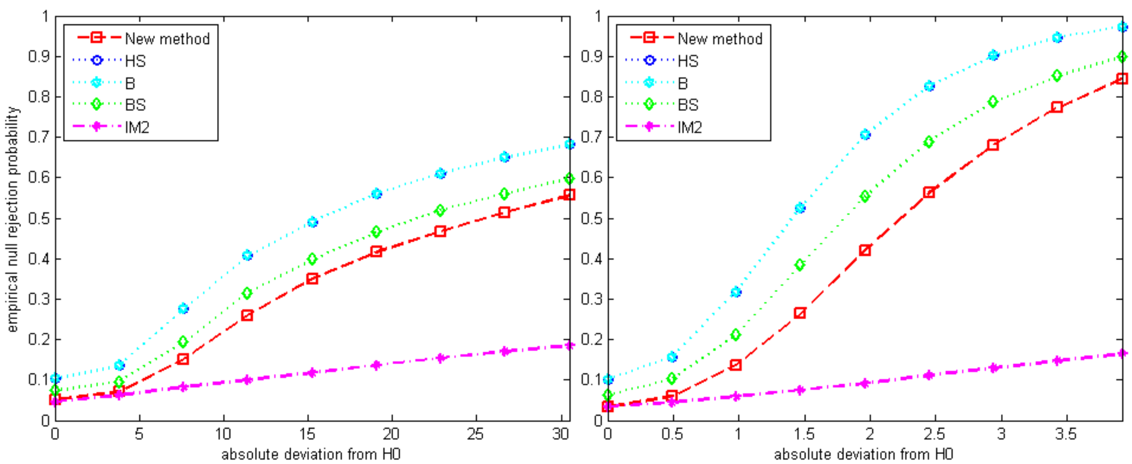

Figure 4: Empirical power properties, one-sample, $n = 45$, $p = 0.95$. Left: slash. Right: GEV(0,1,0).

new method is size-distorted, but significantly less so than all other methods. Also included is IM, the method from Ibragimov and Müller (2010), where the subscript indicates the number of groups, q. With $q = 2$, IM controls size as well as the new method, but its power is always much lower; see for example Figure 4. With $q = 3$, IM's power is still worse than the new method (though somewhat improved), and IM suffers worse size distortion, as expected when relying on a normal approximation for sample quantiles from groups with $n/q = 21/3 = 7$ observations. However, IM is valid under much weaker sampling assumptions: rather than independent sampling, there can be dependence as long as the subsamples' estimators are asymptotically independent (and normal), and rather than identical distributions, the estimators within different groups may have different variances. Even with better size control than BS, the new method's power is competitive with BS, as in Figure 4.

For the two-sample case with $F_X = F_Y$, both the new method and BS always control size. They have comparable power with $n = 21$, but BS has better power for $n = 45$. So at this two-sample "extreme," BS has some advantage, while at the other (i.e., as $f_X(\xi_p)/f_Y(\xi_p) \to \infty$ or zero, approaching the one-sample case) the new method is better. Intermediate two-sample cases could thus have either method perform better.

For any sample size, it appears that quantiles close enough to zero or one will cause size distortion in HS, B, and BS. For example, even with $n = 250$ and a normal distribution, tests of the $p = 0.005$ quantile for $\alpha = 0.05$ have empirical size 9.6% for BS and 20.7% for HS and B, compared with 5.3% for the new method. See Table 3 for additional examples. In future work, it would be valuable to better delineate, in terms of n and p, when extreme value theory provides a better asymptotic approximation.

The new method can have better power than HS or B. However, this may be of more theoretical than practical interest since this is unlikely (or impossible) for

Table 3: Empirical size as percentage (nominal: 5.0), one-sample.

	$n = 125$, $p = 0.01$			$n = 250$, $p = 0.005$		
Distribution	New	BS	HS/B	New	BS	HS/B
N(0, 1)	5.6	10.2	21.0	5.2	9.6	20.7
t_3	6.8	11.5	22.2	6.8	11.2	21.6
Cauchy	9.9	13.1	23.4	9.8	13.2	23.5

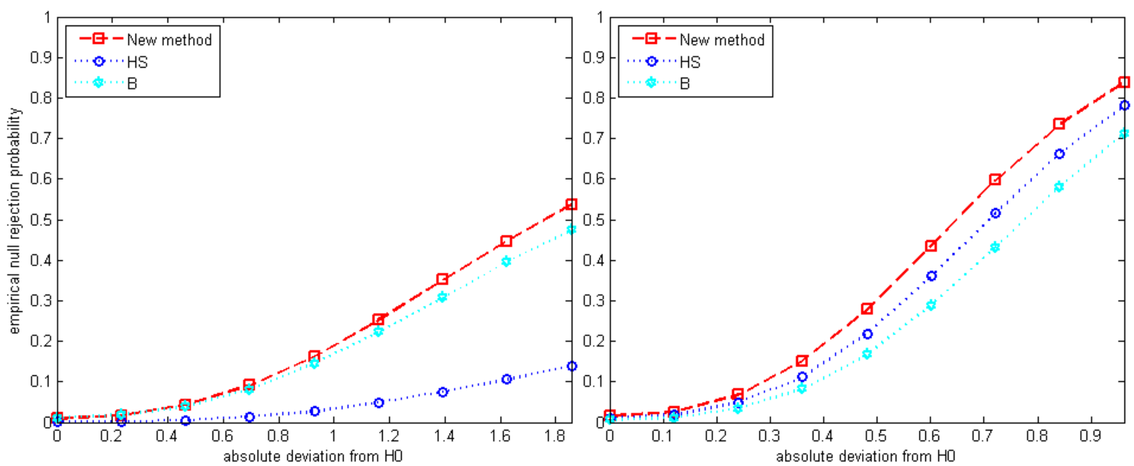

Figure 5: Empirical power curves, one-sample, comparing $n = 11$ (left column) and $n = 41$ (right column). 50,000 replications, $p = 0.5$, $\alpha = 0.05$, and Cauchy distribution.

more extreme quantiles p. Better power may be achieved in both the one-sample case (Figure 5) and the two-sample case (similar; not shown). Theoretical calculations corroborate this power advantage. Using the type II error rate in Proposition 6, since $m_K < m_{HS}$ and $z_{\alpha,m} = z_{1-\alpha/2} + c/m$ with $c = z^3_{1-\alpha/2}/4 > 0$, the m^{-1} term is larger for the new method and the $(m/n)^2$ term is larger for HS88. The first-order and $n^{-1/2}$ terms are identical for the two methods. Which term's difference is bigger determines which method has better power.

Plugging in parameter values and the Cauchy PDF, up to $o(m^{-1} + m^2/n^2)$ terms, the type II error rate difference is $e^K_{II} - e^{HS}_{II} = m_K^{-1}(0.751) + (m^2_K - m^2_{HS})(2.57/n^2)$. Plugging in $n = 11$, $m_K = 3$, and $m_{HS} = 5$, $e^K_{II} - e^{HS}_{II} = -0.090 < 0$, meaning better power for the new method. With $n = 41$, $m_K = 9$, and $m_{HS} = 12$, $e^K_{II} - e^{HS}_{II} = -0.013 < 0$, again better power for the new method. To compare with Figure 5, note that $\gamma/\sqrt{n} = (C/f(0))\sqrt{p(1-p)/n} \approx 0.9$ for $n = 11$ and $\gamma/\sqrt{n} \approx 0.5$ for $n = 41$. The analytic differences from the m^{-1} and $(m/n)^2$ terms are still smaller than the differences in Figure 5 because with small samples, the $o(m^{-1} + m^2/n^2)$ terms also contribute significantly. For comparison, the $(m/n)^2$ term is a factor of π smaller with a $N(0, 1)$ instead of Cauchy distribution; for other distributions and/or quantiles, it could be larger. Generally, the new method's power should be better when the factor in the $(m/n)^2$ term involving f, f', and f'' is larger.

In all the two-sample cases, the results shown and discussed were for independent samples, as the theoretical results assumed. If there is enough negative correlation, size distortion occurs for all methods; if there is positive correlation, power will decrease for all methods. The effect of uncorrelated, dependent samples could go either way. In practice, examining the sample correlation as well as one of the many tests for statistical independence is recommended, if independence is not known a priori.

7 Conclusion

This paper proposes new quantile and quantile treatment effect testing procedures based on the SBG test statistic in (4). Critical values dependent on smoothing parameter m are derived from fixed-smoothing asymptotics. These are more accurate than the conventional standard normal critical values since the fixed-m distribution is shown to be more accurate under both fixed-m and large-m asymptotics. Type I and II error rates are approximated using an Edgeworth expansion, and the testing-optimal m_K minimizes type II error subject to control of type I error, up to higher-order terms. Simulations show that, compared with the previous SBG methods from HS88 and Bofinger (1975), the new method greatly reduces size distortion while maintaining good power. The new method also outperforms bootstrap methods for one-sample tests and certain two-sample cases. Consequently, this new method is recommended in one-sample cases when Hutson (1999) is not computable, and possibly in two-sample cases when Goldman and Kaplan (2012) is not computable. Two-sample performance depends on the unknown distributions and warrants further research into both bootstrap methods and the two-sample plug-in method (particularly estimating θ) proposed here. Finally, theoretical justification for fixed-smoothing asymptotics is provided outside of the time series context; there are likely additional models that may benefit from this perspective.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26(1):247–257.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss:

Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.
Björkman, M. and Svensson, J. (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics*, 124(2):735–769.
Bloch, D. A. and Gastwirth, J. L. (1968). On a simple estimate of the reciprocal of the density function. *The Annals of Mathematical Statistics*, 39(3):1083–1085.
Bofinger, E. (1975). Estimation of a density function using order statistics. *Australian Journal of Statistics*, 17(1):1–7.
Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458.
Cabedo, J. D. and Moya, I. (2003). Estimating oil price 'value at risk' using the historical simulation approach. *Energy Economics*, 25(3):239–253.
Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.
DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York.
David, H. A. (1981). *Order Statistics*. Wiley series in probability and mathematical statistics. Wiley, 2nd edition.
Djebbari, H. and Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145(1):64–80.
Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.
Goh, S. C. (2004). *Smoothing choices and distributional approximations for econometric inference*. PhD thesis, UC Berkeley.
Goh, S. C. and Knight, K. (2009). Nonstandard quantile-regression inference. *Econometric Theory*, 25(5):1415–1432.
Goldman, M. and Kaplan, D. M. (2012). IDEAL quantile inference via interpolated duals of exact analytic L-statistics. Working paper, available at `http://web.missouri.edu/~kaplandm/personalResearch.html`.
Guerre, E. and Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory*, 28(1):87–129.
Hall, P. and Sheather, S. J. (1988). On the distribution of a Studentized quantile. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):381–391.
Hutson, A. D. (1999). Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics*, 26(3):343–353.
Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.
Jackson, E. and Page, M. E. (2013). Estimating the distributional effects of education reforms: A look at project STAR. *Economics of Education Review*, 32:92–103.

Kaplan, D. M. (2012). IDEAL inference on conditional quantiles. Working paper, available at http://web.missouri.edu/~kaplandm/personalResearch.html.

Koenker, R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):145–163.

Koenker, R. and Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612.

Mosteller, F. (1946). On some useful "inefficient" statistics. *Annals of Mathematical Statistics*, 17:377–408.

Siddiqui, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *J. Research of the NBS, B. Math. and Math. Physics*, 64B(3):145–150.

Sun, Y. (2010). Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroscedasticity and autocorrelation robust inference. Working paper, available at http://econ.ucsd.edu/~yisun/research.html.

Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *Journal of Econometrics*, 164(2):345–366.

Sun, Y. (2013). A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator. *The Econometrics Journal*, 16(1):1–26.

Sun, Y. and Kaplan, D. M. (2011). A new asymptotic theory for vector autoregressive long-run variance estimation and autocorrelation robust testing. Working paper, available at http://econ.ucsd.edu/~yisun/research.html.

Sun, Y., Phillips, P. C. B., and Jin, S. (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica*, 76(1):175–194.

A Accuracy of approximate fixed-m critical values

The approximate fixed-m critical value in (9) is quite accurate for all but $m = 1, 2$, as Table 4 shows. The second approximation alternative adds the $O(m^{-2})$ term to the approximation. The third alternative uses the critical value from the Student's t-distribution with the degrees of freedom chosen to match the variance. To compare, for various m and α, I simulated the two-sided rejection probability for a given critical given $T_{m,\infty}$ from (6) as the true distribution. One million simulation replications per m and α were run; to gauge simulation error, I also include in the table the critical values given in Goh (2004) (who ran 500,000 replications each). Additional values of

m and α are available in the online appendix.

Table 4: Simulated rejection probabilities (%) for different fixed-m critical value approximations, $\alpha = 5\%$.

	rejection probability for critical value			
m	Goh (2004, simulated)	including m^{-1}	including m^{-2}	t
1	4.93	8.21	5.84	6.87
2	5.03	6.25	5.20	5.47
3	5.05	5.70	5.10	5.28
4	5.15	5.46	5.06	5.21
5	4.92	5.23	4.96	5.08
10	5.01	5.14	5.06	5.14
20	5.04	5.01	4.99	5.04

B Type II error rate (Proposition 6)

More details are in the supplemental appendix.

Let $\gamma \neq 0$ so that the null hypothesis is false, where as before $H_0 : \xi_p = \beta$ with $\xi_p = \beta - \gamma/\sqrt{n}$. Letting $S_0 \equiv 1/f(\xi_p)$,

$$T_{m,n} = \frac{\sqrt{n}(X_{n,r} - \xi_p) - \gamma}{S_{m,n}\sqrt{p(1-p)}} = \frac{\sqrt{n}(X_{n,r} - \xi_p)}{S_{m,n}\sqrt{p(1-p)}} - \frac{\gamma}{\sqrt{p(1-p)}S_0}\left(\frac{S_0}{S_{m,n}} + 1 - 1\right),$$

$$P(T_{m,n} < z) = P\left(\frac{\sqrt{n}(X_{n,r} - \xi_p) + \gamma(S_{m,n}/S_0 - 1)}{S_{m,n}\sqrt{p(1-p)}} < z + C\right). \tag{18}$$

Type II error is not rejecting when H_0 is false. For a two-sided symmetric test, this probability is

$$P(|T_{m,n}| < z) = P(T_{m,n} < z) - P(T_{m,n} < -z).$$

Let $z_{\alpha,m} = z_{1-\alpha/2} + z^3_{1-\alpha/2}/(4m)$ as in (9), and expanding via (18) and (12), for $C > 0$,

$$P(|T_{m,n}| < z) = L^+ + H_1^+ - H_2^+ + O(n^{-1/2}) + o(m^{-1} + (m/n)^2) \quad \text{with}$$

$$L^+ = \Phi(z_{\alpha,m} + C) - \Phi(-z_{\alpha,m} + C),$$

$$H_1^+ = \phi(z_{\alpha,m}+C)\Big[m^{-1}u_{2,\gamma}(z_{\alpha,m}+C) + (m/n)^2 u_{3,\gamma}(z_{\alpha,m}+C)\Big],$$
$$H_2^+ = \phi(-z_{\alpha,m}+C)\Big[m^{-1}u_{2,\gamma}(C-z_{\alpha,m}) + (m/n)^2 u_{3,\gamma}(C-z_{\alpha,m})\Big].$$

The $O(n^{-1/2})$ term does not depend on m and thus does not affect the optimization problem for selecting m. Define L^-, H_1^-, and H_2^- similarly but with $-C<0$ instead of $C>0$, and thus $-\gamma = -C\sqrt{p(1-p)}/f(\xi_p)$ instead of γ. I calculate average power where the alternatives $+C$ and $-C$ each have 0.5 probability. Using $\phi(-x)=\phi(x)$,

$$P(|T_{m,n}|<z_{\alpha,m}) = \frac{1}{2}\Big\{(L^+ + L^-) + (H_1^+ + H_1^-) - (H_2^+ + H_2^-) + O(n^{-1/2}) + o(m^{-1}+(m/n)^2)\Big\}. \quad (19)$$

For L^+ and L^-,

$$\begin{aligned} L^+ &= \Phi(z_{\alpha,m}+C) - \Phi(-z_{\alpha,m}+C) \\ &= \Phi(C + z_{1-\alpha/2} + z_{1-\alpha/2}^3/(4m)) - \Phi(C - z_{1-\alpha/2} - z_{1-\alpha/2}^3/(4m)) \\ &= 0.5 + m^{-1}\frac{1}{4}z_{1-\alpha/2}^3[\phi(z_{1-\alpha/2}+C) + \phi(C - z_{1-\alpha/2})] = L^-. \end{aligned}$$

For the m^{-1} terms,

$$\begin{aligned} &\phi(z_{\alpha,m}+C)m^{-1}(u_{2,\gamma}(z_{\alpha,m}+C) - u_{2,-\gamma}(-z_{\alpha,m}-C)) \\ &\quad + \phi(z_{\alpha,m}-C)m^{-1}(u_{2,-\gamma}(z_{\alpha,m}-C) - u_{2,\gamma}(-z_{\alpha,m}+C)) \\ &\quad\quad = -\frac{1}{2}m^{-1}\big[\phi(z_{1-\alpha/2}+C)(z_{1-\alpha/2}+C)z_{1-\alpha/2}^2 + \phi(z_{1-\alpha/2}-C)(z_{1-\alpha/2}-C)z_{1-\alpha/2}^2\big] \\ &\quad\quad + O(m^{-2}). \end{aligned}$$

For the $(m/n)^2$ terms, writing $f(\xi_p)$ and its derivatives as f, f', f'', and letting $M \equiv [3(f')^2 - ff'']/(6f^4)$,

$$u_{3,\gamma}(z_{\alpha,m}+C) - u_{3,-\gamma}(-z_{\alpha,m}-C)$$

$$= M(z_{\alpha,m} + C - C) - M(-z_{\alpha,m} - C - (-C)) = 2Mz_{1-\alpha/2} + O(m^{-1}),$$

$$u_{3,-\gamma}(z_{\alpha,m} - C) - u_{3,\gamma}(-z_{\alpha,m} + C)$$

$$= M(z_{\alpha,m} - C - (-C)) - M(-z_{\alpha,m} + C - C)) = 2Mz_{1-\alpha/2} + O(m^{-1}),$$

$$\frac{m^2}{n^2}\{\phi(z_{\alpha,m} + C)[u_{3,\gamma}(z_{\alpha,m} + C) - u_{3,-\gamma}(-z_{\alpha,m} - C)]$$

$$+ \phi(z_{\alpha,m} - C)[u_{3,-\gamma}(z_{\alpha,m} - C) - u_{3,\gamma}(-z_{\alpha,m} + C)]\}$$

$$= (m/n)^2 2\frac{3(f')^2 - ff''}{6f^4} z_{1-\alpha/2}\big[\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)\big] + O(m/n^2). \quad (20)$$

Combining these,

$$P(|T_{m,n}| < z) = \frac{1}{2}\Big\{2\big(0.5 + m^{-1}(1/4)z^3_{1-\alpha/2}[\phi(C + z_{1-\alpha/2}) + \phi(C - z_{1-\alpha/2})]\big)$$

$$- 2(1/4)m^{-1}[\phi(z_{1-\alpha/2} + C)(z_{1-\alpha/2} + C)z^2_{1-\alpha/2}$$

$$+ \phi(z_{1-\alpha/2} - C)(z_{1-\alpha/2} - C)z^2_{1-\alpha/2}]$$

$$+ 2(m/n)^2 \frac{3(f')^2 - ff''}{6f^4} z_{1-\alpha/2}\big[\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)\big]\Big\}$$

$$+ O(n^{-1/2}) + o(m^{-1} + (m/n)^2),$$

which simplifies to the forms given in Proposition 6.

C Edgeworth expansion (Theorem 2)

Here I highlight the differences with the proof sketch in HS88 (Appendix A); the full proof is in the supplemental appendix.

As in Section 2, the null hypothesis is $H_0 : \xi_p = \beta$, and the true $\xi_p = \beta - \gamma/\sqrt{n}$. I continue from (18), which showed that

$$P(T_{m,n} < z) = P\left(\frac{\sqrt{n}(X_{n,r} - \xi_p) + \gamma(S_{m,n}/S_0 - 1)}{S_{m,n}\sqrt{p(1-p)}} < z + C\right),$$

where $C \equiv \gamma f(\xi_p)/\sqrt{p(1-p)}$, $S_0 \equiv 1/f(\xi_p)$. I want to derive a higher-order expansion around the (shifted) standard normal distribution. Since C is a constant, it can be ignored for the expansion and simply plugged in later. For HS88, $\gamma = C = 0$ above.

The centering effect (since $\lfloor np \rfloor$ increases by unit jumps) when $\gamma \neq 0$ is the same as in HS88. They show that centering at $\eta_p \equiv F^{-1}\left[\exp\left(-\sum_{j=r}^{n} j^{-1}\right)\right]$ instead of ξ_p leads to a $n^{-1/2}$ difference in the CDF of $T_{m,n}$, which they calculate.

The next step is to exploit a representation from (David, 1981, p. 21) in terms of iid exponential random variables with unit mean, $\{W_j\}_{j=1}^{n}$. With $H(x) \equiv F^{-1}(e^{-x})$, $\{X_{n,r}, 1 \leq r \leq n\}$ have the same joint distribution as $\{H(\sum_{s \leq j \leq n} j^{-1} W_j), 1 \leq s \leq n\}$. Let $G(\cdot) \equiv F^{-1}(\cdot)$, and $g(\cdot) \equiv G'(\cdot)$. The definitions for Δ_1, Δ_2, Δ_3, and a_k remain identical to HS88: for demeaned $V_j \equiv W_j - 1$,

$$\Delta_1 \equiv \sum_{j=r-m}^{r-1} j^{-1} V_j, \quad \Delta_2 \equiv \sum_{j=r}^{r+m-1} j^{-1} V_j, \quad \Delta_3 \equiv \sum_{j=r+m}^{n} j^{-1} V_j, \quad a_k \equiv H^{(k)}\left(\sum_{j=r}^{n} j^{-1}\right).$$

However, now Z has the extra γ term:

$$\begin{aligned} Z &\equiv [p(1-p)]^{1/2}[n^{1/2}(X_{n,r} - \eta_p) + \gamma(S_{m,n}/S_0 - 1)]/\hat{\tau} \\ &= [n^{1/2}(X_{n,r} - \eta_p) + \gamma(S_{m,n}/S_0 - 1)][(n/2m)(X_{n,r+m} - X_{n,r-m})]^{-1}. \end{aligned}$$

Subsequently, in the expansion $Z = Y + R$, Y, R, and δ are defined the same, but B includes the additional Ψ terms defined below:

$$\begin{aligned} &Y \equiv -pn^{1/2}[(1+\delta)(\Delta_2 + \Delta_3) + B], \quad \delta \equiv -(m/n)^2 g''(p)[6g(p)]^{-1}, \\ &B \equiv \delta\Psi + (n/m)(b_1\Delta_1 + b_2\Delta_2)\Delta_2 + (n/m)b_3(\Delta_1 + \Delta_2)(\Delta_3 + \Psi) \\ &\qquad + (n/m)^2 b_4(\Delta_1 + \Delta_2)^2(\Delta_3 + \Psi) + b_5\Delta_3(\Delta_3 + 2\Psi), \\ &b_1 \equiv -p/2, b_2 \equiv -p/2, b_3 \equiv -p/2 - (m/n)(a_2/2a_1), b_4 \equiv (p/2)^2, b_5 \equiv -a_2/2a_1, \end{aligned}$$

$\Psi \equiv \gamma/(pg(p)\sqrt{n})$.

The b_i are identical to HS88 other than $b_2 \equiv -p/2$, which is simplified to drop a term that ends up in the remainder later anyway. Using the stochastic orders $\Delta_1 = O_p(n^{-1}m^{1/2})$, $\Delta_2 = O_p(n^{-1}m^{1/2})$, and $\Delta_3 = O_p(n^{-1/2})$, non-binding restrictions on the rate of m, and other calculations, the remainder R is shown to be smaller-order, $o_p(m^{-1} + m^2/n^2)$.

Here, Y is of the same form as HS88, but with Ψ now additionally showing up in the higher-order B terms. In the definition of Z, γ only appears in the numerator, not in the denominator. From the stochastic expansion of $S_{m,n}$, which is already required for the denominator of Z, $S_{m,n}/S_0 = 1 + \nu$, where ν contains the higher-order terms that are dropped for the first-order asymptotic result,

$$\nu = \frac{n}{2m}p(\Delta_1 + \Delta_2) + \frac{a_2}{2a_1}(\Delta_1 + \Delta_2 + 2\Delta_3) + (m/n)^2\frac{g''(p)}{6g(p)}$$
$$+ O_p((m/n)^{2+\epsilon} + n^{-1/2}m^{-1/2} + mn^{-3/2}).$$

Thus, γ enters only in higher-order terms, through the numerator of Z.

The Δ_i are all independent, with smooth distributions. Moments of Y are calculated to approximate its characteristic function and get the Edgeworth expansion. HS88 (A.2) now contains additional Ψ terms, while preserving the other terms:

$$\begin{aligned}E[(-p^{-1}Y)^\ell] = E[(1+\delta)(D_2+D_3)]^\ell \\ &+ n^{-1/2}\ell\Big\{-(\ell-1)(2p)^{-1}\Psi\sqrt{n}E(D_3^{\ell-2}) - \ell(2p)^{-1}E(D_3^{\ell-1}) \\ &\quad - (a_2/a_1)E(D_3^\ell)\Psi\sqrt{n} - (a_2/2a_1)E(D_3^{\ell+1})\Big\} \\ &+ m^{-1}\ell\Big\{\frac{\ell-1}{4}E(D_3^{\ell-2})\Psi^2 n + \frac{\ell}{2}\Psi\sqrt{n}E(D_3^{\ell-1}) + \frac{\ell+1}{4}E(D_3^\ell)\Big\} \\ &+ \delta\ell E(D_3^{\ell-1})\Psi\sqrt{n} + o_p\big(m^{-1} + (m/n)^2\big).\end{aligned}$$

(HS88 uses the HS87 definition of $D_i = \sqrt{n}\Delta_i$.)

As in HS88, define $K \equiv [p(1-p)]^{-1/2}Y$ and $L \equiv -[p/(1-p)]^{1/2}(1+\delta)(D_2+D_3)$. Multiplying the prior equation by $\{-[p/(1-p)]^{1/2}\}^{\ell}(it)^{\ell}/\ell!$ and adding over ℓ, HS88 (A.3)

$$E(e^{itK}) = E(e^{itL}) + n^{-1/2}\alpha_1(t)m^{-1}\alpha_2(t) + o[m^{-1} + (m/n)^2]$$

is reached but with different α_i, whose respective Fourier–Stieltjes transforms are

$$\begin{aligned}
a_1(z) &\equiv -\Bigg[\Psi\sqrt{n}\left(2b_5 - \frac{1}{2(p-1)}\right)z \\
&\qquad + [(1/2) - b_5(1-p)][p(1-p)]^{-1/2}z^2\Bigg]\phi(z), \\
a_2(z) &\equiv \frac{1}{4}\left[-\Psi^2 n\frac{p}{1-p}z + 2\Psi\sqrt{n}\left(\frac{p}{1-p}\right)^{1/2}z^2 - z^3\right]\phi(z), \text{ and} \\
a_3(z) &\equiv \left[\Psi\sqrt{n}\left(\frac{p}{1-p}\right)^{1/2}\right]\phi(z).
\end{aligned}$$

The characteristic function of L is the same:

$$E(e^{itL}) = \{1 + \delta(it)^2 - n^{-1/2}(1/6)[p/(1-p)]^{1/2}[(1+p)/p](it)^3\}e^{-t^2/2} + O(\delta^2 + n^{-1}),$$

which is the Fourier–Stieltjes transform of

$$\Phi(x) - \delta x\phi(x) + n^{-1/2}(1/6)[p/(1-p)]^{1/2}[(1+p)/p](x^2-1)\phi(x) + O(\delta^2 + n^{-1}).$$

Adding all the Fourier–Stieltjes inverses (including the recentering term) yields the final result.

D Bivariate Edgeworth expansion (Theorem 3)

The bivariate (two-sample) Edgeworth expansion is similar to the univariate Edgeworth expansion, but with a parallel set of terms for the second sample the whole way through. The independence of the samples helps cross-sample terms simplify. Here, I note the different expressions obtained for the major steps; the full proof is available in the supplemental appendix.

The CDF to approximate is

$$\begin{aligned}
&P(\tilde{T}_{m,n} < z) \\
&= P\left(\frac{\sqrt{n}(X_{nr} - \xi_{px}) - \sqrt{n}(Y_{nr} - \xi_{py}) + \gamma[(\tilde{S}_{mn}/\tilde{S}_0) - 1]}{\tilde{S}_{mn}\sqrt{p(1-p)}} < z + \frac{\gamma}{\tilde{S}_0\sqrt{p(1-p)}}\right),
\end{aligned}$$

with $\xi_{py} = \xi_{px} + \gamma/\sqrt{n}$ and $\tilde{S}_{mn}$ being the denominator of (5) other than the $\sqrt{p(1-p)}$.

The effect of (re)centering is the same as the one-sample case but with an additional $(g_x - g_y)/\tilde{S}_0$ factor,

$$w_n \equiv \frac{[\epsilon_n - 1 + (1/2)(1-p)](g_x - g_y)}{\tilde{S}_0\sqrt{p(1-p)}},$$

where the one-sample $g(p)$ is now g_x, and g_y the equivalent for the second distribution.

The same $Z = Y + R$ structure is followed, but with different definitions that include the second sample terms. For example, instead of ν,

$$\begin{aligned}
\tilde{\nu} \equiv\ & \frac{2g_x}{g_x^2 + g_y^2}\big[(m/n)^2(g_x''/6) - (n/m)(a_1/2)(\Delta_1 + \Delta_2) - (a_2/(2p))(\Delta_1 + \Delta_2 + 2\Delta_3)\big] \\
&+ (n/m)^2(a_1^2/4)(\Delta_1 + \Delta_2)^2/(g_x^2 + g_y^2) \\
&+ \frac{2g_y}{g_x^2 + g_y^2}\big[(m/n)^2(g_y''/6) - (n/m)(a_1'/2)(\nabla_1 + \nabla_2) - (a_2'/(2p))(\nabla_1 + \nabla_2 + 2\nabla_3)\big] \\
&+ (n/m)^2((a_1')^2/4)(\nabla_1 + \nabla_2)^2/(g_x^2 + g_y^2),
\end{aligned}$$

where ∇ mirrors Δ in the second sample. This may then be plugged into, for example,

$$(\tilde{S}_{mn}/\tilde{S}_0) - 1 = \sqrt{1+\tilde{\nu}} - 1 = \tilde{\nu}/2 - \tilde{\nu}^2/8 + O(\tilde{\nu}^3),$$

where $\tilde{\nu}^3 = O_p(m^{-3/2})$. Eventually, the following parallel expansion is reached, similar to before but with $\tilde{\Theta}$ and $\tilde{\Psi}$ instead of Θ and Ψ (the remainder R remains the same order):

$$\begin{aligned}
Z &= -pn^{1/2}\Big[\tilde{\Theta} - (\tilde{\nu}/2)(\tilde{\Theta}+\tilde{\Psi}) + (3/8)\tilde{\nu}^2(\tilde{\Theta}+\tilde{\Psi})\Big] + R, \\
\tilde{\Theta} &\equiv \left(\frac{-a_1}{pS_0}\right)(\Delta_2+\Delta_3) + \frac{a_2}{2a_1}\left(\frac{-a_1}{pS_0}\right)(\Delta_2+\Delta_3)^2 \\
&\quad - \left(\frac{-a_1'}{pS_0}\right)(\nabla_2+\nabla_3) - \frac{a_2'}{2a_1'}\left(\frac{-a_1'}{pS_0}\right)(\nabla_2+\nabla_3)^2, \\
\tilde{\Psi} &\equiv \gamma/\Big(pn^{1/2}\tilde{S}_0\Big),
\end{aligned}$$

where a_1' and a_2' mirror a_1 and a_2 in the second sample.

Moments are then calculated to get an expression in the same structure as before,

$$E[(-p^{-1}Y)^\ell] \equiv z_1(\ell) + z_2(\ell) + z_3(\ell) + o[m^{-1} + (m/n)^2].$$

The term L from the univariate case is now

$$L = -[p/(1-p)]^{1/2}\Big[(g_x/\tilde{S}_0)(D_2+D_3) - (g_y/\tilde{S}_0)(\reflectbox{D}_2+\reflectbox{D}_3)\Big],$$

where $\reflectbox{D}_j$ mirrors D_j. Calculating moments of L to approximate its characteristic function ends with

$$E(e^{itL}) = e^{-t^2/2}\left[1 - n^{-1/2}(1/6)\frac{1+p}{\sqrt{p(1-p)}}\left(\frac{g_x^3 - g_y^3}{\tilde{S}_0^3}\right)(it)^3\right] + O(n^{-1}),$$

the inverse Fourier–Stieltjes transform of which is

$$\Phi(z) + n^{-1/2}(1/6)\frac{1+p}{\sqrt{p(1-p)}}\left(\frac{g_x^3 - g_y^3}{\tilde{S}_0^3}\right)(z^2-1)\phi(z) + O(n^{-1}).$$

Parallel to $D_3 \sim N(0,(1-p)/p)$ in the univariate case, it can be shown that $\bar{D}_3 \equiv (g_x/\tilde{S}_0)D_3 - (g_y/\tilde{S}_0)\text{ᗡ}_3$ has the same asymptotic distribution. It can further be shown that $\bar{D}_3$ and $\bar{\Delta}_3$ jointly follow a bivariate normal distribution, as do $\bar{\text{ᗡ}}_3$ and $\bar{\nabla}_3$, which are independent of the former two objects. Isserlis' Theorem (a.k.a. Wick's Theorem) can then be used to calculate moments by breaking down expectations of many-term products in terms of expectations of pairwise products. As in the univariate case, the characteristic function of $K \equiv [p(1-p)]^{-1/2}Y$ may be calculated by summing over $i = 1$ to ∞

$$E(K^\ell)(it)^\ell/\ell! = E(L^\ell)(it)^\ell/\ell! + \left(-[p/(1-p)]^{1/2}\right)^\ell[(it)^\ell/\ell!][z_2(\ell) + z_3(\ell) + R'].$$

Following the same sequence of steps as in the univariate case, plugging in the extremely long expressions (shown in supplemental appendix), calculating, and taking Fourier–Stieltjes inverses leads to the final result.