

# UC San Diego

## UC San Diego Electronic Theses and Dissertations

### Title

New Theory and Methods for High-Order Accurate Inference on Quantile Treatment Effects and Conditional Quantiles

### Permalink

<https://escholarship.org/uc/item/51w1m3v9>

### Author

Kaplan, David M.

### Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

**New Theory and Methods for High-Order Accurate Inference on  
Quantile Treatment Effects and Conditional Quantiles**

A dissertation submitted in partial satisfaction of the  
requirements for the degree  
Doctor of Philosophy

in

Economics

by

David M. Kaplan

Committee in charge:

Professor Yixiao Sun, Chair  
Professor Patrik Guggenberger  
Professor Karen Sue Messer  
Professor Dimitris Politis  
Professor Andres Santos

2013

Copyright  
David M. Kaplan, 2013  
All rights reserved.

The dissertation of David M. Kaplan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

---

---

---

---

---

---

Chair

University of California, San Diego

2013

## EPIGRAPH

*Dogmatism and scepticism are both, in a sense, absolute philosophies; one is certain of knowing, the other of not knowing. What philosophy should dissipate is certainty, whether of knowledge or ignorance.*

—Bertrand Russell, “Philosophy for Laymen” in *Unpopular Essays* (1950)

## TABLE OF CONTENTS

Signature Page	. . . . .	iii
Epigraph	. . . . .	iv
Table of Contents	. . . . .	v
List of Figures	. . . . .	vii
List of Tables	. . . . .	viii
Acknowledgements	. . . . .	ix
Vita	. . . . .	x
Abstract of the Dissertation	. . . . .	xi
Chapter 1	IDEAL quantile inference via interpolated duals of exact analytic $L$ -statistics . . . . .	1
	1.1 Introduction . . . . .	2
	1.2 Fractional order statistic approximation error . . . . .	7
	1.3 Application to quantile inference . . . . .	15
	1.3.1 One-sample inference on a single quantile . . . . .	17
	1.3.2 Joint one-sample inference on multiple quantiles . . . . .	21
	1.3.3 One-sample inference on linear combinations of quantiles . . . . .	23
	1.3.4 Two-sample quantile treatment effect inference . . . . .	28
	1.3.5 Estimation of $\gamma$ . . . . .	33
	1.3.6 Further approximation and intuition (two-sample) . . . . .	35
	1.4 Empirical application . . . . .	38
	1.5 Simulation study . . . . .	40
	1.6 Conclusion . . . . .	50
Chapter 2	IDEAL inference on conditional quantiles . . . . .	52
	2.1 Introduction . . . . .	53
	2.2 Fractional order statistic theory . . . . .	56
	2.3 Setup . . . . .	61
	2.4 Optimal bandwidth and CPE . . . . .	67
	2.4.1 Optimal rate of bandwidth and CPE . . . . .	67
	2.4.2 Plug-in bandwidth . . . . .	75
	2.4.3 Conditional quantile treatment effects and other objects of interest . . . . .	81
	2.5 Simulation study . . . . .	84

	2.5.1	Computation of plug-in bandwidth . . . . .	84
	2.5.2	Results . . . . .	86
	2.5.3	Computation time . . . . .	93
	2.5.4	Conditional quantile treatment effect inference . .	94
	2.6	Empirical application . . . . .	95
	2.7	Conclusion . . . . .	98
Chapter 3		One- and two-sample population quantile inference via fixed-smoothing asymptotics and Edgeworth expansion . . . . .	99
	3.1	Introduction . . . . .	100
	3.2	Quantile estimation and hypothesis testing . . . . .	102
	3.3	Fixed- $m$ asymptotics and corrected critical value . . . . .	104
	3.3.1	Fixed- $m$ asymptotics . . . . .	105
	3.3.2	Corrected critical value . . . . .	106
	3.4	Edgeworth expansion . . . . .	109
	3.4.1	One-sample case . . . . .	109
	3.4.2	Two-sample case . . . . .	111
	3.5	Optimal smoothing parameter selection . . . . .	112
	3.5.1	Type I error . . . . .	112
	3.5.2	Type II error . . . . .	113
	3.5.3	Choice of $m$ . . . . .	115
	3.6	Simulation study . . . . .	120
	3.7	Conclusion . . . . .	127
Appendix A		Technical appendix to Chapter 1 . . . . .	130
	A.1	Mathematical proofs . . . . .	130
Appendix B		Technical appendix to Chapter 2 . . . . .	154
	B.1	Lemma 2.2 proof . . . . .	154
	B.2	Plug-in bandwidth calculations . . . . .	161
	B.3	Corollary 2.4 proof . . . . .	168
Appendix C		Technical appendix to Chapter 3 . . . . .	172
	C.1	Accuracy of fixed- $m$ critical value approximation . . . . .	172
	C.2	Edgeworth expansion (Theorem 3.1) where $\gamma \neq 0$ : differences with HS88 proof . . . . .	172
Bibliography		. . . . .	174

## LIST OF FIGURES

Figure 1.1:	Sample realizations of the ‘ideal’ uniform fractional order statistic process. . . . .	10
Figure 1.2:	Example of one-sample, one-sided confidence interval endpoint determination . . . . .	16
Figure 1.3:	Power curves for inference on the lower quartile and median using data simulated to mimic the period one library task data from Gneezy and List (2006) . . . . .	42
Figure 1.4:	Power curves for quantile treatment effect inference . . . . .	48
Figure 2.1:	Example of selection of $u_\ell$ and $u_h$ using (2.2), for one-sided CI endpoints. For two-sided, $\alpha/2$ would be in place of $\alpha$ . Note that for $p = 0.5$ , $u_h = 1 - u_\ell$ . . . . .	58
Figure 2.2:	True conditional median function for simulations . . . . .	87
Figure 2.3:	Pointwise coverage probability and joint power curves, conditional median inference . . . . .	88
Figure 2.4:	Pointwise power, conditional median inference . . . . .	90
Figure 2.5:	Pointwise power by $X$ ; same as Figure 2.4 but with Cauchy errors for top row and centered $\chi_3^2$ for bottom row. . . . .	91
Figure 2.6:	Computation time for different methods, as a function of sample size . . . . .	93
Figure 2.7:	IDEAL confidence intervals for tenth percentile of hemoglobin concentration, conditional on expenditure and head of household’s education level . . . . .	97
Figure 3.1:	Plot of the ratio $m_K/m_{HS}$ against the first-order power used to calculate $C$ . . . . .	117
Figure 3.2:	Analytic and empirical $e_I$ (top) and $e_{II}$ (bottom) by $m$ , log-normal distribution. . . . .	119
Figure 3.3:	Graph showing for which combinations of quantile $p$ and sample size $n$ the Hutson (1999) method is computable. . . . .	121
Figure 3.4:	Empirical power properties; $p = 0.5$ . Top row: $F = \text{Exp}(1)$ , one-sample. Bottom: $N(0, 1)$ , two-sample. Left column: $n = 3$ . Right: $n = 4$ . . . . .	123
Figure 3.5:	Empirical power properties, one-sample, $n = 45$ , $p = 0.95$ . Only the new method controls size, and power is not size-adjusted. Top: slash. Bottom: $\text{GEV}(0,1,0)$ . . . . .	125
Figure 3.6:	Proximate causes of size distortion for HS and B methods. . . . .	126
Figure 3.7:	Empirical power curves, comparing $n = 11$ (left column) and $n = 41$ (right column). Median inference with Cauchy distribution, one- and two-sample. . . . .	128



## LIST OF TABLES

Table 1.1:	IDEAL two-sided 90% confidence intervals for quartile treatment effects in Gneezy and List (2006). . . . .	40
Table 1.2:	Type I error by quartile, using data simulated to mimic the period one library and fundraising task data from Gneezy and List (2006) . . . . .	43
Table 1.3:	Type I error comparison, median . . . . .	44
Table 1.4:	Type I error comparison when exchangeability is violated by difference in variance . . . . .	45
Table 1.5:	Type I error comparison when exchangeability is violated by difference in shape . . . . .	46
Table 1.6:	Power comparison, median, location shift alternative . . . . .	47
Table 1.7:	Power comparison for local quantile treatment effects . . . . .	49
Table 2.1:	Coverage probability and median interval length for IDEAL confidence intervals for conditional quantile treatment effects . . . . .	95
Table 3.1:	Empirical size as percentage (nominal: 5.0), $n = 3, 4$ , $p = 0.5$ , one- and two-sample. . . . .	123
Table 3.2:	Empirical size as percentage (nominal: 5.0), $n = 45, 21$ , $p = 0.95$ , one-sample. . . . .	124
Table C.1:	Simulated rejection probabilities (%) for different fixed- $m$ critical value approximations, $\alpha = 5\%$ . . . . .	173

## ACKNOWLEDGEMENTS

Thanks to many professors for help and inspiration great and small, with special thanks to my advisor, Yixiao Sun, for his enthusiastic, rigorous, and caring help. Thanks also to past teachers who helped me grow into someone who would greatly enjoy graduate school (and be okay with that). Thanks to my many friendly classmates and fellow graduate students for friendship, feedback, volleyball, and questions and ideas to ponder. Thanks to my parents for their support and listening to my excited explanations of econometrics each week. Thanks to my wife for daily hugs and reassurance, and for helping me to be more comfortable with both how things are and how they change.

Chapter 1 was joyfully coauthored with fellow student and friend Matt Goldman; it is still in progress.

## VITA

- 2006 Bachelor of Arts in Computer Science *magna cum laude*,  
Princeton University
- 2006 Outstanding Computer Science Senior Thesis Prize
- 2008–2013 Teaching assistant, Department of Economics, University of  
California, San Diego
- 2011–2012 Tutor for graduate qualifying examination in econometrics,  
University of California, San Diego
- 2012 Clive Granger Fellowship
- 2013 Doctor of Philosophy in Economics, University of California,  
San Diego

ABSTRACT OF THE DISSERTATION

**New Theory and Methods for High-Order Accurate Inference on  
Quantile Treatment Effects and Conditional Quantiles**

by

David M. Kaplan

Doctor of Philosophy in Economics

University of California, San Diego, 2013

Professor Yixiao Sun, Chair

This dissertation concerns methods for inference on quantiles in various models. Methods that are asymptotically justified may still be quite inaccurate in finite samples. To improve the state of the art, I explore different theoretical approaches for achieving higher-order accuracy: fractional order statistic theory based on exact finite-sample distributions in Chapters 1 and 2, and Edgeworth expansions and fixed-smoothing asymptotics in Chapter 3. For each of the different practical methods proposed, I examine accuracy via precise theoretical results as well as simulations. The family of methods using interpolated duals of exact-analytic  $L$ -statistics (IDEAL) covers unconditional (one-sample and two-sample treatment/control, Ch. 1) and nonparametric conditional (Ch. 2) models, and it

offers improvements over the existing literature in terms of accuracy, robustness, and/or computation time. The Edgeworth-based method improves upon related prior methods and is a good alternative for quantiles too far into the tails for IDEAL to handle.

# Chapter 1

## IDEAL quantile inference via interpolated duals of exact analytic $L$ -statistics

### Abstract

The literature has two types of fractional order statistics: an ‘ideal’ (unobserved) type based on a beta distribution, and an observable type linearly interpolated between consecutive order statistics. We show convergence in distribution of the two types at an  $O(n^{-1})$  rate, which we also show holds for joint vectors and linear combinations of fractional order statistics. This connection justifies use of the linearly interpolated type in practice when sampling theory is based on the ‘ideal’ type. For example, the coverage probability error (CPE) has the same  $O(n^{-1})$  magnitude for one-sample nonparametric joint confidence intervals over multiple quantiles. For a single quantile, our new analytic calibration reduces the CPE to nearly  $O(n^{-3/2})$ , and our new inference method on linear combinations of quantiles has  $O(n^{-2/3})$  CPE. With additional theoretical work, we propose a new method for two-sample quantile treatment effect inference, which has two-sided CPE of order  $O(n^{-2/3})$ , or  $O(n^{-1})$  under exchangeability, and one-sided CPE of order  $O(n^{-1/2})$ . In an application of our method to data from a recent paper on

“gift exchange,” we reveal interesting heterogeneity in the treatment effect of “gift wages.” In simulations, our two-sample hypothesis test compares favorably with existing methods in both size and power properties.

## 1.1 Introduction

Quantiles contain information about a distribution’s shape. Complementing the mean, they capture heterogeneity, inequality, and other measures of economic interest. These could be population quantiles, conditional quantiles, or quantile treatment effects. For example, when considering educational interventions like reduced class size, we care about the mean treatment effect (effect on test score average, for example), as well as the median (effect on the median student’s score) and other quantiles (are the upper and lower quartiles converging or diverging?).

Here, we propose a novel method for inference on the quantile treatment effect (QTE) in a two-sample setup frequently encountered in experimental and quasi-experimental contexts. In this setup, an outcome variable (like test score) is measured for individuals in a control group and a treatment group. The goal is to determine if the treatment group’s outcome distribution is different from the control distribution at a certain quantile; e.g., does the treatment increase the median outcome? For recent experimental economics setups amenable to our method, see for example Björkman and Svensson (2009), Charness and Gneezy (2009), and Gneezy and List (2006), whose data we examine for our empirical application.

In simulations, our QTE hypothesis test compares favorably with existing methods. Its size control is robust, and it consistently has better power than other methods applicable to general quantiles. Compared to popular median-specific permutation-type QTE tests (including Mann–Whitney–Wilcoxon), our method is robust to violations of the exchangeability assumption that can cause severe size distortion in permutation-type tests, and our method has better power against certain types of alternatives.

Theoretically, our  $O(n^{-2/3})$  order of two-sided QTE coverage probability

error (CPE) is not an improvement over Chapter 3, but there is a reason for the consistent finite-sample improvements in simulations. Nuisance parameter estimation error comprises the dominant CPE term in both methods, but its effect is bounded in our method. To illustrate this, consider coverage of 95% CIs if we had an atrocious nuisance parameter “estimator” that simply gave the same worst-case value each time. For Chapter 3, this would give CIs of length zero and coverage probability zero for the undercoverage worst-case value, and CIs of infinite length and 100% coverage probability for the overcoverage worst-case value, even as  $n \rightarrow \infty$ . In contrast, even with worst-case values, our method would have asymptotic coverage probability in  $[0.83, 0.99]$  rather than  $[0, 1]$ . This mechanism likely underlies the better finite-sample accuracy of our method.

Computationally, we require estimation of the ratio of probability density functions evaluated only at the quantile of interest. For both the control and treatment distributions, this can be done with any standard kernel density estimator and bandwidth, which in practice performs just as well as our theoretically optimal plug-in bandwidth.

While the new two-sample method is a significant contribution, the underlying theory we develop applies more generally. For example, we use it to derive the order of size distortion for a popular inference method for population quantiles (Hutson, 1999) and provide an order-reducing analytic calibration. This result can be extended to a conditional nonparametric context; see Chapter 2, which also discusses conditional QTE inference. With some additional work, our new theory also unlocks a new method for joint inference on multiple quantiles, and another new method for inference on linear combinations of quantiles, such as the interquartile range. Previous order statistic methods provided conservative (by construction) inference on linear combinations, but we provide asymptotically exact inference, with rate-limiting error only from estimation of nuisance parameters (again the ratios of density functions at quantiles of interest). Our contributions are also of interest in fields beyond economics, such as biostatistics and ecology.

With sample size  $n$ , the exact finite-sample distribution is known for order statistics from a uniform distribution. These can also be thought of as sample



$u$ -quantiles with index  $u \in (0, 1)$  such that  $(n + 1)u$  is an integer. However, for a given  $n$ ,  $(n + 1)u$  is not an integer for almost all  $u \in (0, 1)$ , which makes this result not directly applicable to the exact inference approach followed here. For fractional  $(n + 1)u$ , the exact distribution is known for the corresponding fractional order statistic, a theoretical (unobserved) object that we approximate with linear interpolation between observed order statistics.

Though unobserved for non-integer indices, fractional order statistics from a uniform distribution jointly follow a known Dirichlet process (Stigler, 1977). We show that linearly interpolating consecutive observed order statistics approximates the unobserved fractional order statistics with only  $O(n^{-1})$  error in cumulative distribution function (CDF). For example, the sampling distribution of the average of the 4th and 5th order statistics is a good approximation of the sampling distribution of the 4.5th fractional order statistic. With this bound on interpolation error, we theoretically justify the aforementioned inference methods that belong to a larger family of IDEAL methods. Ideally, we could directly use linear combinations of order statistics (known as  $L$ -statistics) from the unobserved world, whose joint distributions are known exactly and analytically. Thankfully, the error from instead using the observed counterparts is small, so a high degree of accuracy is common to methods using the interpolated duals of exact analytic  $L$ -statistics (IDEAL).

In the one-sample case (as proposed but not rigorously justified in Hutson (1999)), the upper and lower endpoints of an IDEAL equal-tail  $100(1 - \alpha)\%$  confidence interval for the median are chosen as specific fractional order statistics from the sample. The true median is just a number, like 10, while the endpoints are random (depend on the sample). If we imagine applying the true CDF to everything, then the true median becomes 0.5 and the fractional order statistics come from a uniform distribution. The distribution of any such uniform fractional order statistic is exactly known, so it is easy to pick the one having exactly  $\alpha/2$  probability of being less than 0.5 (for the upper endpoint) and the one having exactly  $\alpha/2$  probability of being greater than 0.5 (for the lower endpoint). Applying the inverse CDF to everything, we can see that we just computed the *indices* of

fractional order statistics from the original sample that provide exact endpoints.

The only barrier to an exact one-sample confidence interval is the need to estimate the unobserved fractional order statistics by linearly interpolating observed order statistics, and this error induces only  $O(n^{-1})$  CPE. Although rates like  $O(n^{-1})$  are only asymptotic claims and do not guarantee finite-sample performance, it is clear that the interpolated fractional order statistic cannot lie far from its unobserved dual even in small samples, which is an advantage over methods more reliant on asymptotic justification.<sup>1</sup>

The two-sample and linear combination methods build on this idea. In the two-sample case, one-sample IDEAL confidence intervals are built in the control and treatment samples. For a confidence interval for the treatment median minus the control median, a value is included in the interval if the one-sample intervals intersect after shifting the treatment interval down by that value. For example, the value zero (i.e., difference of medians is zero) is excluded if the one-sample intervals do not overlap, and included if they do overlap; the value 10 is excluded if the one-sample intervals do not overlap after shifting the treatment interval down by 10, and included if they do overlap after that shift. Instead of using the nominal  $\alpha$  for the one-sample intervals, which would lead to overcoverage, we calculate an adjusted  $\tilde{\alpha}$  depending on the ratio of probability density functions at the quantile of interest.

These methods can all also be used for quantile inference conditional on discrete-valued covariates, with no adjustment. Conditioning on continuous as well as discrete variables is covered in Chapter 2.

The remainder of this section reviews the literature for one-sample and two-sample quantile inference, which contains many alternative approaches. We first review the one-sample case. Methods of inference based on asymptotic normality require the selection of a smoothing parameter to estimate the probability density function at the quantile of interest, but they may use Edgeworth expansions to

---

<sup>1</sup>This is somewhat similar to the argument that a bootstrap can perform better than a normal approximation method with the same CPE, since the bootstrap distribution's deviation from a normal distribution may capture some of the true finite-sample distribution's deviation from normality.

reduce theoretical CPE.<sup>2</sup> Hall and Sheather (1988) achieve a one-sided  $O(n^{-1/2})$  and two-sided  $O(n^{-2/3})$  CPE, while Chapter 3 uses bandwidth-dependent critical values that lead to two-sided confidence intervals undercovering by  $O(n^{-1})$  but (weakly) overcovering by  $O(n^{-2/3})$ .

One-sample quantile inference has been a popular topic in the bootstrap literature dating back to Efron (1979). Smoothing has been the most popular approach, as in Brown et al. (2001); Hall et al. (1989); Ho and Lee (2005b); Horowitz (1998); Janas (1993), although practical performance doesn't always reflect the nice theoretical properties. For example, the best theoretical one-sided rate is  $O(n^{-58/57})$  in Ho and Lee (2005b), but it requires additional iteration plus “proper choices of bandwidths at the bootstrapping and Studentization steps” (p. 444) while only deriving “optimal orders of bandwidths” (p. 443). Polansky and Schucany (1997) achieve the best two-sided CPE of  $O(n^{-3/2})$ , though criticized by fellow smoothers Ho and Lee (2005b) as “requir[ing] sophisticated tuning of the smoothing bandwidths” and themselves admitting, “If this method is to be of any practical value, a better bandwidth estimation technique will certainly be required” (p. 833). For a more comprehensive review of bootstrap methods for one-sample quantile inference, see Ho and Lee (2005b, §1).

The smoothed empirical likelihood method of Chen and Hall (1993) achieves one-sided  $O(n^{-1/2})$  CPE, while two-sided  $O(n^{-1})$  CPE is Bartlett-correctable theoretically to  $O(n^{-2})$  or analytically to  $O(n^{-1}h)$ , where  $h$  is the smoothing bandwidth. They suggest that the bandwidth range  $[n^{-1/2}, n^{-3/4}]$  is often good but that “less smoothing than this is desirable” for skewed distributions like  $\chi_1^2$ .

For methods derived from the exact binomial distribution, randomization is one way to preserve exact size, as in Zieliński and Zieliński (2005). Beran and Hall (1993) suggest an interpolation that gets one-sided and two-sided  $O(n^{-1})$  CPE, which Ho and Lee (2005a) improve to  $O(n^{-3/2})$  via analytic calibration or  $O(n^{-25/14})$  via smoothed bootstrap calibration (requiring bandwidth choice). The fractional order statistic method of quantile inference originally proposed by Hutson (1999) is similar to Beran and Hall (1993); it performs well in simulations

---

<sup>2</sup>Since confidence intervals correspond directly to hypothesis tests in our context, CPE is equivalent to the difference between actual and nominal type I error.

and requires no nuisance parameter estimation or bandwidth choice. As we show in two new results here, its CPE is  $O(n^{-1})$ , and it is analytically correctable to almost  $O(n^{-3/2})$ .

Existing two-sample methods cover the same types of strategies. Horowitz (1998) bootstraps a smoothed regression quantile; CPE approaches  $o(n^{-1})$  as assumed smoothness increases. Whang (2006) extends the smoothed empirical likelihood of Chen and Hall (1993) to regression, with the same CPEs. However, it will be conservative for two-sample inference since parameters must be tested jointly, so projection is required to isolate the quantile treatment effect. For the same reason, two-sample inference using the exact method in Chernozhukov et al. (2009) will also be conservative. With higher-order approximation of the asymptotic normal distribution, Chapter 3 again gets a confidence interval weakly overcovering by  $O(n^{-2/3})$  and undercovering by  $O(n^{-1})$ , though these rates do not include error from nuisance parameter estimation (or, alternatively, an assumption somewhat weaker than exchangeability).

Specific to the median in the two-sample case, Hutson (2007) proposes an “exact bootstrap” method. With the strong exchangeability assumption, permutation-type tests can be applied. This includes the commonly used Mann–Whitney–Wilcoxon test (Mann and Whitney, 1947; Wilcoxon, 1945), though this is only valid as a median difference test against pure location-shift alternatives.

Section 1.2 presents the key theoretical results. Section 1.3 translates the theory into practical methods. For our new two-sample QTE inference method, we give an empirical application in Section 1.4 and simulation results in Section 1.5. Notationally,  $\doteq$  should be read as “is equal to, up to smaller-order terms”;  $\asymp$  as “has exact (asymptotic) rate/order of” (same as “big theta” Bachmann–Landau notation,  $\Theta(\cdot)$ ); and  $A_n = O(B_n)$  as usual,  $\exists k < \infty$  s.t.  $|A_n| \leq B_n k$  for sufficiently large  $n$ . Acronyms used are those for cumulative distribution function (CDF), confidence interval (CI), coverage probability (CP), coverage probability error (CPE), interpolated duals of exact analytic  $L$ -statistics (IDEAL), and probability density function (PDF). Proofs absent from the text are collected in the appendix.

## 1.2 Fractional order statistic approximation error

In this section we introduce our basic notation and demonstrate our core theoretical result on the accuracy of the approximation of ‘ideal’ fractional order statistics by linear interpolation of observed order statistics.

Given an iid sample  $\{X_i\}_{i=1}^n$  of draws of an absolutely continuous random variable with unknown cumulative distribution function (CDF) denoted<sup>3</sup>  $F(\cdot)$ , interest is in estimation of  $Q(u) \equiv F^{-1}(u)$  for some  $u \in (0, 1)$ , where  $Q(\cdot) = F^{-1}(\cdot)$  is also known as the quantile function. The conventional, linearly (L) interpolated fractional order statistic estimator of  $Q(u)$  is defined as

$$\hat{Q}_X^L(u) \equiv (1 - \epsilon)X_{n:\lfloor(n+1)u\rfloor} + \epsilon X_{n:\lfloor(n+1)u\rfloor+1}, \quad (1.1)$$

where  $X_{n:k}$  denotes the  $k$ th order statistic, which is the  $k$ th smallest value out of the  $n$  observations  $\{X_i\}_{i=1}^n$ , and  $\epsilon \equiv (n + 1)u - \lfloor(n + 1)u\rfloor$  is the interpolation weight, with  $\lfloor \cdot \rfloor$  the floor function. While  $Q(u)$  is a fixed (true) value,  $\hat{Q}_X^L(u)$  is a random variable, and  $\hat{Q}_X^L(\cdot)$  is a stochastic process with its argument varying over  $(0, 1)$ .

If  $u \in \Xi_n \equiv \{\frac{k}{n+1}\}_{k=1}^n$ , then no interpolation is necessary and  $\hat{Q}_X^L(u)$  corresponds exactly to some  $X_{n:k}$ . However, frequently we are interested in the distributions of sample quantiles outside  $\Xi_n$ , as is true for our applications to quantile inference in §1.3. Absent an exact distributional theory for  $\hat{Q}_X^L(u)$  when  $u \notin \Xi_n$ , we demonstrate a tight link between the marginal distributions of the stochastic process  $\hat{Q}_X^L(\cdot)$  and of the analogous ‘ideal’ process  $\hat{Q}_X^I(\cdot)$ . This allows us to translate distributional results from the ‘ideal’ process into practical methods based on interpolated order statistics, with very little coverage error. We demonstrate the power of this result in the important special cases of one-sample and two-sample quantile inference, as well as inference on general linear combinations of quantiles like the interquartile range.

---

<sup>3</sup> $F$  will often be used with a random variable in a subscript to denote the CDF of that particular random variable. If no subscript is present, then  $F$  refers to the CDF of  $X$ .

For uniformly distributed  $U_i \equiv F(X_i)$ , the process of unobserved ‘ideal’ (I) fractional order statistics, denoted  $\hat{Q}_U^I(\cdot)$ , is a Dirichlet process over the Lebesgue measure on the unit interval with concentration parameter  $n + 1$  (Stigler, 1977). While there are many ways to interpret this Dirichlet process (e.g., as a Chinese restaurant process, or using Pólya’s urn), understanding it as an infinite-dimensional generalization of the Dirichlet distribution may be most helpful for our purposes. For example, with  $n = 5$ , the first observed order statistic,  $U_{5:1} = \hat{Q}_U^I(1/6)$ , has distribution  $\beta(1, 5)$ . The value of the second order statistic,  $U_{5:2} = \hat{Q}_U^I(2/6)$ , adds to the first:  $U_{5:2} = U_{5:1} + (1 - U_{5:1})\Delta_2$  with  $\Delta_2 \sim \beta(1, 4)$ . In terms of the common illustration of cutting a string of unit length,  $U_{5:1} \sim \beta(1, 5)$  is the length of the first segment, and  $\Delta_2$  describes the length of the second segment as a proportion of the remaining string:  $(U_{5:2} - U_{5:1})/(1 - U_{5:1}) = \Delta_2 \sim \beta(1, 4)$ . The  $\Delta_j$  are all independent. Similarly,  $U_{5:3} = U_{5:2} + (1 - U_{5:2})\Delta_3$  with  $\Delta_3 \sim \beta(1, 3)$ , and similarly for  $U_{5:4}$  and  $U_{5:5}$  as well, with  $\Delta_4 \sim \beta(1, 2)$  and  $\Delta_5 \sim \beta(1, 1)$ . For *fractional* order statistics  $\hat{Q}_U^I(u)$  with  $u \in \{1/12, 2/12, \dots, 11/12\}$ , the starting point becomes  $\hat{Q}_U^I(1/12) \sim \beta(0.5, 5.5)$ . The first observed order statistic is now expressed as  $\hat{Q}_U^I(2/12) = \hat{Q}_U^I(1/12) + [1 - \hat{Q}_U^I(1/12)]\Delta_2$  with  $\Delta_2 \sim \beta(0.5, 5.0)$ . Similarly,  $\Delta_3 \sim \beta(0.5, 4.5)$ , and generally  $\Delta_j \sim \beta(0.5, 6 - j/2)$  and  $\hat{Q}_U^I(j/12) = \hat{Q}_U^I((j-1)/12) + [1 - \hat{Q}_U^I((j-1)/12)]\Delta_j$  for  $j = 1, \dots, 11$ , defining  $\hat{Q}_U^I(0) = 0$ . Continuing to pack more and more  $u$  closer together, e.g. with  $\Delta_j \sim \beta(1/k, 6 - j/k)$  as  $k \rightarrow \infty$ , the finite joint Dirichlet distribution approaches the infinite-dimensional Dirichlet process.

To help visualize this, Figure 1.1 shows example realizations (sample paths) of  $\hat{Q}_U^I(\cdot)$  for different sample sizes.<sup>4</sup> As  $n \rightarrow \infty$ , sample quantiles converge to the true quantiles, and  $\hat{Q}_U^I(u) \xrightarrow{p} Q_U(u) = u$ , so sample paths from the process with larger sample size are clustered closer to the function  $Q_U(u) = u$ . Realizations of observable order statistics correspond to  $u \in \Xi_n$ , as discussed; e.g., for  $n = 5$ ,  $u \in \{1/6, 2/6, \dots, 5/6\}$ .

---

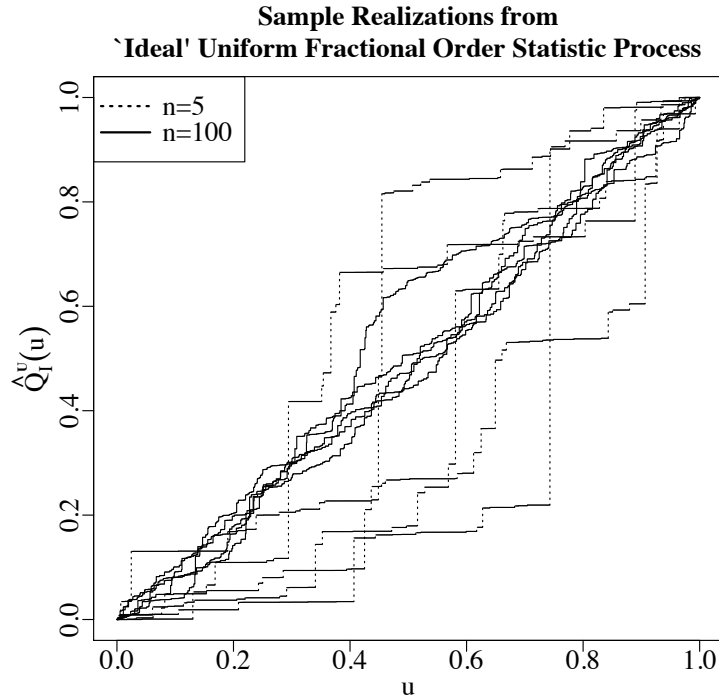
<sup>4</sup>Although it is technically impossible to sample the infinite number of points that fully describes a realization of the process, sampling over 50,000 points is visually indistinguishable.

For the non-uniform  $X_i$ , we define

$$\hat{Q}_X^I(\cdot) \equiv F^{-1}\left(\hat{Q}_U^I(\cdot)\right). \quad (1.2)$$

Note that  $\hat{Q}_X^I(u)$  is an alternative “estimator” of  $Q(u)$  that exactly coincides with  $\hat{Q}_X^L(u)$  for all  $u \in \Xi_n$  and differs only in its random (and unobserved) interpolation between these points. Thus, in probability, these processes are closely linked.

**Theorem 1.1.** *For any fixed  $\delta > 0$ , define  $\mathcal{U}^\delta \equiv \{u \in (0, 1) \mid f(F^{-1}(u)) \geq \delta\}$  and  $\mathcal{U}_n^\delta \equiv \mathcal{U}^\delta \cap [\frac{1}{n}, \frac{n-1}{n}]$ ; then,  $\sup_{u \in \mathcal{U}_n^\delta} |\hat{Q}_X^I(u) - \hat{Q}_X^L(u)| = O_p(n^{-1}[\log n])$ .*



**Figure 1.1:** Sample realizations of the ‘ideal’ uniform fractional order statistic process.

For any vector  $\mathbf{u}$  in the power set of  $\Xi_n$ , i.e.  $\mathbf{u} \in \mathcal{P}(\Xi_n) = \Xi_n \times \cdots \times \Xi_n$ , (1.2) provides an exact joint sampling distribution for the observed order statistics. As an important special case, the marginal distribution of any individual uniform

order statistic is a beta distribution,<sup>5</sup>

$$\hat{Q}_U^I(u) \sim \beta((n+1)u, (n+1)(1-u)). \quad (1.3)$$

We consider the multivariate distribution of some vector of quantiles with dimension  $J < \infty$ , where  $u_j$  denotes an element in column vector  $\mathbf{u} \in (0, 1)^J$ . The joint PDF of  $\{\hat{Q}_U^I(u_j)\}_{j=1}^J$  evaluated at  $\{x_j\}_{j=1}^J$  comes from the Dirichlet distribution and is

$$\frac{\Gamma(n+1)}{\prod_{j=1}^{J+1} \Gamma((n+1)(u_j - u_{j-1}))} \left( \prod_{j=1}^{J+1} (x_j - x_{j-1})^{(n+1)(u_j - u_{j-1}) - 1} \right),$$

where  $u_0 \equiv 0$ ,  $x_0 \equiv 0$ ,  $u_{J+1} \equiv 1$ , and  $x_{J+1} \equiv 1$ . The Dirichlet distribution concentration parameters, conventionally written as  $\alpha_j$ , here are  $u_j - u_{j-1}$ . The Dirichlet distribution describes a distribution over segment lengths, in this case the fractional order statistic spacings  $x_j - x_{j-1}$ .

We also consider a common alternative approximation to the sample distribution of order statistics. It is well known that the centered and scaled empirical process for standard uniform random variables converges to a Brownian bridge. So, for a Brownian bridge process  $B(\cdot)$ , we define on  $u \in (0, 1)$  the additional stochastic processes

$$\hat{Q}_U^B(u) \equiv u + n^{-1/2}B(u) \quad \text{and} \quad \hat{Q}_X^B(u) \equiv F^{-1}\left(\hat{Q}_U^B(u)\right).$$

The realizations of these processes do not correspond to observed data, but the distribution of  $\hat{Q}_X^B(u)$  provides a convenient approximation to the distributions from the other processes. To see this note that the marginal distribution of  $\hat{Q}_U^I(\mathbf{u})$  is distributed as the cumulative sum of a Dirichlet distribution, while the marginal of  $\hat{Q}_U^B(\mathbf{u})$  is a multivariate Gaussian. Lemma 1.2 proves the close relationship between these densities and density derivatives for all values outside the tails (part a), which are shown to have rapidly diminishing probability mass (parts b and c).

---

<sup>5</sup> $\beta(a, b)$  will denote a beta distribution with parameters  $a$  and  $b$  unless otherwise noted. The probability density function (PDF) evaluated at any  $x \in (0, 1)$  is given by  $[\Gamma(a+b)/(\Gamma(a)\Gamma(b))]x^{a-1}(1-x)^{b-1}$ .



**Lemma 1.2.** Let  $\Delta \mathbf{k}$  be a positive  $(J + 1)$ -vector of natural numbers such that  $\sum_{j=1}^{J+1} \Delta k_j = n' \equiv n + 1$  and  $\min_j \{\Delta k_j\} \rightarrow \infty$  and define  $k_j \equiv \sum_{i=1}^j \Delta k_i$ . Let  $\mathbf{X} \equiv (X_1, \dots, X_J)'$  be the random  $J$ -vector such that

$$\Delta \mathbf{X} \equiv (X_1, X_2 - X_1, \dots, 1 - X_J)' \sim \text{Dirichlet}(\Delta \mathbf{k}).$$

Take any sequence  $a_n$  s.t.

- (i)  $a_n \rightarrow \infty$
- (ii)  $a_n n^{-1} [\max\{\Delta k_j\}]^{1/2} \rightarrow 0$
- (iii)  $a_n [\min\{\Delta k_j\}]^{-1/2} \rightarrow 0$ .

Define Condition  $\star(a_n)$  as satisfied by  $\mathbf{X}$  iff

$$\sup_j \left\{ n \Delta k_j^{-1/2} \left| \Delta X_j - \Delta k_j / n' \right| \right\} \leq a_n.$$

a)  $\forall \mathbf{x} \in \mathbb{R}^J$  satisfying Condition  $\star(a_n)$ ,

$$\begin{aligned} \log[f_{\mathbf{X}}(\mathbf{x})] &= K + \frac{1}{2} (\mathbf{x} - \mathbf{k}/n')' \underline{H} (\mathbf{x} - \mathbf{k}/n) + O(a_n \|\Delta \mathbf{k}^{-1/2}\|_{\infty}) \\ &= \log[\phi_{\mathbf{k}/n, \nu/n}(\mathbf{x})] + O(a_n \|\Delta \mathbf{k}^{-1/2}\|_{\infty}). \\ \frac{d \log[f_{\mathbf{X}}(\mathbf{x})]}{d\mathbf{x}} &= \underline{H} (\mathbf{x} - \mathbf{k}/n') + O\left(a_n^2 n \|\Delta \mathbf{k}^{-1}\|_{\infty}\right). \end{aligned}$$

Writing  $\mathbf{X}_{\mathbf{k}}$  with the subscript to emphasize the dependence of the density on the underlying vector of parameters, we also have

$$\frac{d \log[f_{\mathbf{X}_{\mathbf{k}}}(\mathbf{x})]}{d\mathbf{k}/n'} = -\underline{H} (\mathbf{x} - \mathbf{k}/n') + O(a_n n \|\Delta \mathbf{k}^{-1}\|_{\infty}),$$

where the constant  $K = \frac{J}{2} \log(n/2\pi) + \frac{1}{2} \sum_{j=1}^{J+1} \log\left(\frac{n}{\Delta k_j - 1}\right)$  and the  $J \times J$  matrix  $\underline{H}$  has non-zero elements only on the diagonal  $\underline{H}_{j,j} = -n^2 (\Delta k_j^{-1} + \Delta k_{j+1}^{-1})$  and one off the diagonal  $\underline{H}_{j,j+1} = n^2 \Delta k_{j+1}^{-1}$ .

b)  $1 - P[\star(a_n)] = o(\exp\{-a_n^2/2\})$ .

c) If instead there are fixed components of the parameter vector, the largest of which is  $\Delta k_j = M < \infty$ , then we state the slightly weaker result on tail probabilities that

$$1 - P[\star(a_n)] = O(\exp\{-a_n M^{-1/2}\}).$$

It can be checked that  $\mathcal{V} \equiv \underline{H}^{-1}$  is such that

$$\mathcal{V}_{i,j} = \min(k_i, k_j)(n' - \max(k_i, k_j)), \quad (1.4)$$

thus illuminating the relation between the above result and conventional asymptotic normality results for quantiles. If we had instead want an expansion for  $\Delta \mathbf{X}$  it is a simple manner of linear algebra to derive results of the form

$$\log[f_{\Delta \mathbf{X}}(\mathbf{z})] = K + \frac{1}{2}(\mathbf{z} - \Delta \mathbf{k}/n')' \underline{H}_{\Delta}(\mathbf{z} - \Delta \mathbf{k}/n) + O(a_n \|\Delta \mathbf{k}^{-1/2}\|_{\infty}), \quad (1.5)$$

where  $\underline{H}_{\Delta_{i,j}} = \sum_{k=i}^J \sum_{l=j}^J \underline{H}_{k,l}$  and inverse matrix  $\mathcal{V}_{\Delta} \equiv \underline{H}_{\Delta}^{-1}$  is such that

$$\mathcal{V}_{\Delta_{i,j}} = \begin{cases} \Delta k_i(n' - \Delta k_i), & \text{if } i = j, \\ -\Delta k_i \Delta k_j, & \text{if } i \neq j. \end{cases} \quad (1.6)$$

Note that the limiting density (and derivatives) above and in (a) are those of unscaled multivariate Gaussians. Lemma 1.2 is a powerful tool for approximating the distribution of order statistics and we use it to demonstrate the close distributional link between linear combinations of ‘ideal’, interpolated, and Gaussian-approximated fractional order statistics in Theorem 1.3. Specifically, for an arbitrary  $\boldsymbol{\psi} \in \mathbb{R}^J$  with non-zero elements  $\psi_j$ , we seek to (distributionally) approximate linear combinations of ‘ideal’ fractional order statistics (i.e., ‘ideal’  $L$ -statistics) by linear combinations of their interpolated duals; i.e., to approximate

$$L^I \equiv \sum_{j=1}^J \psi_j \hat{Q}_X^I(u_j) \quad \text{by} \quad L^L \equiv \sum_{j=1}^J \psi_j \hat{Q}_X^L(u_j), \quad (1.7)$$

$$\text{or alternatively, by} \quad L^B \equiv \sum_{j=1}^J \psi_j \hat{Q}_X^B(u_j).$$

For convenience and without loss of generality, we normalize  $\psi_1 = 1$ . As in (1.1), the  $\epsilon_j \equiv (n+1)u_j - \lfloor (n+1)u_j \rfloor$  are interpolation weights.

Our lone assumption for this section is now presented, followed by the main theoretical result that underpins all IDEAL methods. A technical lemma useful for proving the theorem may be found in the appendix, along with its proof and a short discussion. We use **bold** for vectors, underline for matrices, and  $\phi_{\underline{\Sigma}}(\cdot)$  for the multivariate normal PDF with mean  $\mathbf{0}$  and covariance matrix  $\underline{\Sigma}$ .

**Assumption A1.1.** For each quantile  $u_j$ , the PDF  $f(\cdot)$  satisfies (i)  $f(F^{-1}(u_j)) > 0$ ; (ii)  $f''(\cdot)$  is continuous in some neighborhood of  $F^{-1}(u_j)$ .

**Theorem 1.3.** Define  $\underline{\mathcal{V}}$  as the  $J \times J$  matrix such that  $\underline{\mathcal{V}}_{i,j} = \min\{u_i, u_j\}(1 - \max\{u_i, u_j\})$ , and let

$$\underline{\mathcal{A}} \equiv \text{diag}\{f(F^{-1}(\mathbf{u}))\}, \quad \mathcal{V}_\psi \equiv \psi'(\underline{\mathcal{A}}^{-1}\underline{\mathcal{V}}\underline{\mathcal{A}}^{-1})\psi, \quad \mathbb{X}_0 \equiv \sum_{j=1}^J \psi_j F^{-1}(u_j).$$

Given the definitions in (1.1), (1.2), and (1.7) and under Assumption A1.1, the following results bound the error from approximating ‘ideal’  $L$ -statistics  $L^I$  by their interpolated duals  $L^L$ .

(i) For samples  $\{X_i\}_{i=1}^n$  with  $X_i \stackrel{iid}{\sim} F$ , and for a given constant  $K$ ,

$$\begin{aligned} & P\left(L^L < \mathbb{X}_0 + n^{-1/2}K\right) - P\left(L^I < \mathbb{X}_0 + n^{-1/2}K\right) \\ &= \frac{K \exp\{-K^2/(2\mathcal{V}_\psi)\}}{\sqrt{2\pi\mathcal{V}_\psi^3}} \left[ \sum_{j=1}^J \left( \frac{\psi_j^2 \epsilon_j (1 - \epsilon_j)}{f[F^{-1}(u_j)]^2} \right) \right] n^{-1} + O(n^{-3/2} \log(n)). \end{aligned}$$

(ii) For samples  $\{X_i\}_{i=1}^n$  with  $X_i \stackrel{iid}{\sim} F$ ,

$$\begin{aligned} & \sup_{K \in \mathbb{R}} \left[ P\left(L^L < \mathbb{X}_0 + n^{-1/2}K\right) - P\left(L^I < \mathbb{X}_0 + n^{-1/2}K\right) \right] \\ &= \frac{e^{-1/2}}{\sqrt{2\pi\mathcal{V}_\psi^2}} \left[ \sum_{j=1}^J \left( \frac{\psi_j^2 \epsilon_j (1 - \epsilon_j)}{f[F^{-1}(u_j)]^2} \right) \right] n^{-1} + O(n^{-3/2} \log(n)) \end{aligned}$$

and

$$\sup_{K \in \mathbb{R}} \left| P\left(L^B < \mathbb{X}_0 + n^{-1/2}K\right) - P\left(L^I < \mathbb{X}_0 + n^{-1/2}K\right) \right| = O(n^{-1/2} \log(n)).$$

(iii) Given independent (of each other) samples  $\{X_i\}_{i=1}^{n_x}$  and  $\{Y_i\}_{i=1}^{n_y}$  with  $X_i \stackrel{iid}{\sim} F_X$  and  $Y_i \stackrel{iid}{\sim} F_Y$ , where Assumption A1.1 holds for each, and with  $n_x \propto n$  and  $n_y \propto n$ ,

$$\sup_{K \in \mathbb{R}} \left[ P(L_X^L + L_Y^L < K) - P(L_X^I + L_Y^I < K) \right] = O(n^{-1}),$$

and

$$\sup_{K \in \mathbb{R}} \left[ P(L_X^B + L_Y^B < K) - P(L_X^I + L_Y^I < K) \right] = O(n^{-1/2} \log(n)).$$

The following is brief intuition behind the proof of part (i). We start by restricting attention to cases where the largest spacing between relevant uniform order statistics,  $U_{n:\lfloor(n+1)u_j\rfloor+1} - U_{n:\lfloor(n+1)u_j\rfloor}$ , and the largest difference between the  $U_{n:\lfloor(n+1)u_j\rfloor}$  and  $u_j$  satisfy Condition  $\star(2 \log(n))$  as in Lemma 1.2. (In appendix notation:  $\|\mathbf{\Lambda}\|_\infty < 2 \log(n)n^{-1}$  and  $\|\mathbf{Y} - \mathbf{u}\|_\infty < 2 \log(n)n^{-1/2}$ .) By Lemma 1.2(b–c), this does not induce a polynomial order error in our calculations. We then use the representation of ‘ideal’ fractional order statistics from Jones (2002), which is equal in distribution to the linearly interpolated form but with a random interpolation weight following a  $\beta(\epsilon_j, 1 - \epsilon_j)$  distribution. The leading term in the error is due to the random interpolation weight’s variance, and by plugging in other calculations from Lemma A.1, we see that it is uniformly  $O(n^{-1})$  and can be calculated analytically. Part (iii) extends this result to two samples, which is necessary for inference on quantile treatment effects. The additional results for  $L^B$  in parts (ii) and (iii) follow directly from the density approximation in Lemma A.1(ii). These CDF error calculations are key to calculating the CPE of related quantile inference methods, as we do in §1.3.

*Remark.* It is clear by the Cramér-Wold device that the vector  $\hat{Q}_X^L(\mathbf{u})$  converges in distribution to  $\hat{Q}_X^I(\mathbf{u})$  up to an  $O(n^{-1})$  term and to  $\hat{Q}_X^B(\mathbf{u})$  up to an  $O(n^{-1/2} \log(n))$  term. This could allow reliable inference on more general finite-dimensional functionals of the quantile process. In the remainder of this paper, attention is restricted to the class of linear combinations.

*Remark.* Through  $\mathcal{V}_\psi$ , the CDF error in Theorem 1.3(ii) is proportional to the greater of  $1/(\min_j\{u_j\})$  and  $1/(\min_j\{1 - u_j\})$ . This is consistent with the well-known additional difficulties of estimating “extreme quantiles” instead of “central quantiles.” We may remove this specific term via analytic calibration, though a similar term is likely in the remainder.

### 1.3 Application to quantile inference

Theorem 1.3 can be used to select linear combinations of fractional order statistics that serve as endpoints of confidence intervals (CIs) for functions of the quantile process. The only remaining obstacle is that the CDF operator,  $F(\cdot)$ , is unknown. For inference on a single quantile or jointly on multiple quantiles, inference turns out to be entirely distribution free. For inference on quantile treatment effects or linear combinations of quantiles, we will need to take Taylor expansions and estimate PDF ratios at the  $J$  quantiles of interest.

For any quantile  $u_j$  and any nominal confidence level  $(1 - \alpha)$ , we define  $u_j^h(\alpha)$  and  $u_j^l(\alpha)$  as solutions to the equations

$$\alpha = P\left(\hat{Q}_U^I(u_j^h(\alpha)) < u_j\right), \quad \alpha = P\left(\hat{Q}_U^I(u_j^l(\alpha)) > u_j\right), \quad (1.8)$$

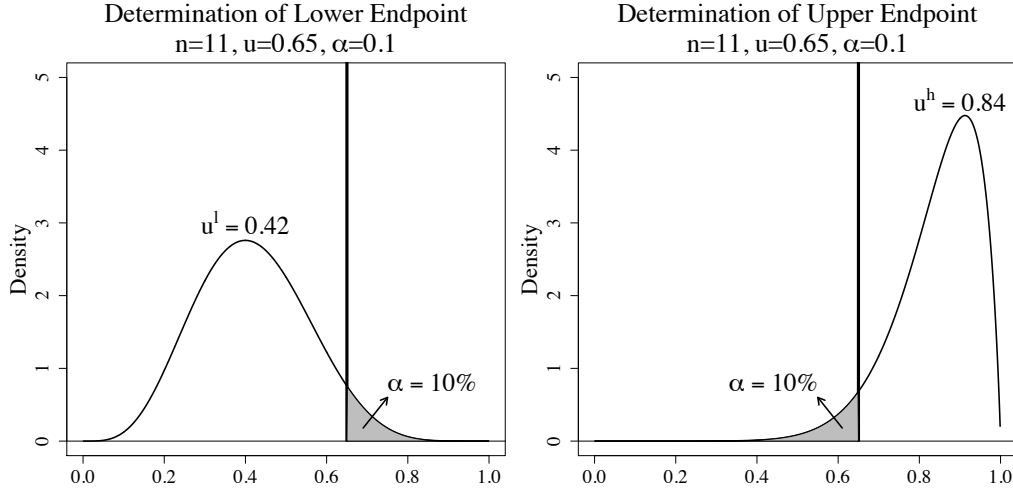
where  $\hat{Q}_U^I(u) \sim \beta((n+1)u, (n+1)(1-u))$  from (1.3). These are analogous to (7) and (8) in Hutson (1999). CI endpoints will then be  $\hat{Q}_X^L(u_j^h)$  or  $\hat{Q}_X^L(u_j^l)$ , i.e. linearly interpolated fractional order statistics with index  $(n+1)u_j^h$  or  $(n+1)u_j^l$ .

Figure 1.2 visualizes an example. The  $u^h$  (or  $u^l$ ) parameter determines the mean of the beta distribution. Decreasing  $u^h$  increases the probability mass in the shaded region below  $u$ , while increasing  $u^h$  decreases the shaded region. Consequently, solving (1.8) is a simple numerical search problem.

For  $J = 1$ , (1.8) defines the one-sample CI endpoint(s). For  $J > 1$  (or two samples), individual CIs are constructed for each  $u_j$ -quantile and then combined into the appropriate overall CI. However, using  $\alpha$  for the individual CIs leads to overcoverage or undercoverage, so a calibrated  $\tilde{\alpha}$  is used instead, as discussed below.

The following lemma shows that the CI endpoints converge to  $u_j$  at a root- $n$  rate and may be approximated by quantiles of a normal distribution.

**Lemma 1.4.** *Let  $z_{1-\alpha}$  denote the  $(1 - \alpha)$ -quantile of a standard normal distribution,  $z_{1-\alpha} \equiv \Phi^{-1}(1 - \alpha)$ . With no assumptions other than their definitions in*



**Figure 1.2:** Example of one-sample, one-sided confidence interval endpoint determination. In the left graph, the lower one-sided endpoint  $u^l$  is the parameter value such that the shaded region  $P(\hat{Q}_U^l(u^l) > u)$  has size  $\alpha$ , where  $\hat{Q}_U^l(u^l) \sim \beta((n+1)u^l, (n+1)(1-u^l))$  from (1.3). In the right graph, similarly,  $u^h$  solves  $P(\hat{Q}_U^l(u^h) < u) = \alpha$ .

(1.8), the values  $u_j^l(\alpha)$  and  $u_j^h(\alpha)$  can be approximated as

$$\begin{aligned} u_j^l(\alpha) - u_j &= -z_{1-\alpha} \sqrt{u_j(1-u_j)n^{-1/2}} + O(n^{-1}), \\ u_j^h(\alpha) - u_j &= z_{1-\alpha} \sqrt{u_j(1-u_j)n^{-1/2}} + O(n^{-1}). \end{aligned}$$

If  $J = 1$ , we omit the  $j$  subscript. With two-sample problems, there may be an additional subscript denoting the sample ( $x$  or  $y$ ). If  $J > 1$ , we define

$$\begin{aligned} u_j^H(\alpha) &\equiv 1\{\psi_j > 0\}u_j^h(\alpha) + 1\{\psi_j < 0\}u_j^l(\alpha), \\ u_j^L(\alpha) &\equiv 1\{\psi_j > 0\}u_j^l(\alpha) + 1\{\psi_j < 0\}u_j^h(\alpha). \end{aligned} \tag{1.9}$$

### 1.3.1 One-sample inference on a single quantile

With an iid sample  $\{X_i\}_{i=1}^n$ , we consider testing the null hypotheses  $H_0 : Q(u) = D$  against the lower one-sided alternative  $H_1 : Q(u) < D$ , or (equivalently) constructing a lower one-sided CI for  $D$ . Hutson (1999, eqn. 8) constructs a CI

with nominal coverage probability  $(1 - \alpha)$  for  $D$  as

$$\left(-\infty, \hat{Q}_X^L(u^h(\alpha))\right)$$

but provides neither a bound for coverage probability error (CPE) nor a theoretical characterization of power against local alternatives. To apply Theorem 1.3(i), we note that

$$\begin{aligned} J &= 1, \quad \epsilon = (n+1)u^h(\alpha) - \lfloor (n+1)u^h(\alpha) \rfloor, \\ \mathcal{V}_\psi &= \frac{u^h(\alpha)(1-u^h(\alpha))}{f(F^{-1}(u))^2}, \quad \mathbb{X}_0 = F^{-1}(u^h(\alpha)), \quad \text{and} \\ K &= n^{1/2}[F^{-1}(u) - F^{-1}(u^h(\alpha))] = \frac{n^{1/2}(u - u^h(\alpha))}{f(F^{-1}(u^h(\alpha)))} + O(n^{1/2}(u^h(\alpha) - u)^2) \\ &= -\frac{z_{1-\alpha}\sqrt{u^h(\alpha)(1-u^h(\alpha))}}{f(F^{-1}(u^h(\alpha)))} + O(n^{-1/2}) \\ &= -\frac{z_{1-\alpha}\sqrt{u(1-u)}}{f(F^{-1}(u))} + O(n^{-1/2}), \end{aligned}$$

where the last three lines use Assumption A1.1 and Lemma 1.4. The coverage probability under the null is

$$\begin{aligned} P\left(D \in (-\infty, \hat{Q}_X^L(u^h(\alpha)))\right) &= P\left(\hat{Q}_X^L(u^h(\alpha)) > D\right) \\ &\stackrel{\text{Thm 1.3}}{=} P\left(\hat{Q}_X^I(u^h(\alpha)) > D\right) + \frac{\epsilon(1-\epsilon)z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{\sqrt{2\pi}u^h(\alpha)(1-u^h(\alpha))}n^{-1} + O(n^{-3/2}\log(n)) \\ &= 1 - \alpha + \frac{\epsilon(1-\epsilon)z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{\sqrt{2\pi}u(1-u)}n^{-1} + O(n^{-3/2}\log(n)). \end{aligned}$$

Thus, Hutson's (1999) lower one-sided CIs have  $O(n^{-1})$  overcoverage and the same leading term found in the alternative interpolation scheme of Beran and Hall (1993). However, this term is analytic and can be removed by appropriate calibration.<sup>6</sup> We recommend the calibrated CI

$$\left(-\infty, \hat{Q}_X^L\left[u^h\left(\alpha + \frac{\epsilon(1-\epsilon)z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{\sqrt{2\pi}u(1-u)}n^{-1}\right)\right]\right),$$

which has CPE of  $O(n^{-3/2}\log(n))$ . By parallel argument, Hutson's (1999) upper one-sided and two-sided CIs also have  $O(n^{-1})$  CPE, which can be improved to

---

<sup>6</sup>Ho and Lee (2005a) provide a similar calibration to Beran and Hall's (1993) interpolation scheme.

$O(n^{-3/2} \log(n))$  via calibration. For the upper one-sided case, the  $(1 - \alpha)$  Hutson CI and our calibrated CI are respectively given by

$$\left( \hat{Q}_X^L[u^l(\alpha)], \infty \right) \quad \text{and} \quad \left( \hat{Q}_X^L \left[ u^l \left( \alpha + \frac{\epsilon(1-\epsilon)z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{\sqrt{2\pi}u(1-u)} n^{-1} \right) \right], \infty \right),$$

and for equal-tail two-sided CIs,

$$\begin{aligned} & \left( \hat{Q}_X^L[u^l(\alpha/2)], \hat{Q}_X^L[u^h(\alpha/2)] \right) \quad \text{and} \\ & \left( \hat{Q}_X^L \left[ u^l \left( \frac{\alpha}{2} + \frac{\epsilon(1-\epsilon)z_{1-\alpha/2} \exp\{-z_{1-\alpha/2}^2/2\}}{\sqrt{2\pi}u(1-u)} n^{-1} \right) \right], \right. \\ & \quad \left. \hat{Q}_X^L \left[ u^h \left( \frac{\alpha}{2} + \frac{\epsilon(1-\epsilon)z_{1-\alpha/2} \exp\{-z_{1-\alpha/2}^2/2\}}{\sqrt{2\pi}u(1-u)} n^{-1} \right) \right] \right). \end{aligned}$$

In the Hutson (1999) two-sided CI, the  $O(n^{-1})$  terms do not cancel out, but rather build on top of each other, leading to symmetric overcoverage. In all cases the  $O(n^{-1})$  term leads to overcoverage of the Hutson CI. Intuitively, this is due to interpolation variance found in the ‘ideal’ process from which the CI endpoints are chosen but not found in the ‘linearly interpolated’ process from which they are realized.

Finally, we demonstrate that the hypothesis tests corresponding to all the CIs of this section achieve optimal asymptotic power against local alternatives. The sample quantile is the semiparametric efficient estimator, so it suffices to show that IDEAL power is asymptotically (first-order) equivalent to that of the method based on asymptotic normality. Consider the coverage of the local alternative  $D_n \equiv Q(u) + Kn^{-1/2}$  by Hutson’s (1999) lower one-sided CI. Including such a value is type II error, so power is

$$\begin{aligned} \mathcal{P}_n^l(D_n) &= P\left(D_n \notin (-\infty, \hat{Q}_X^L(u^h(\alpha)))\right) \\ &= P\left(\hat{Q}_X^L(u^h(\alpha)) - Q(u^h(\alpha)) < Kn^{-1/2} - [Q(u^h(\alpha)) - Q(u)]\right) \\ &\stackrel{\text{Thm } 1.3}{=} P\left(\hat{Q}_X^B(u^h(\alpha)) - Q(u^h(\alpha)) < Kn^{-1/2} - [Q(u^h(\alpha)) - Q(u)]\right) \\ &\quad + O(n^{-1/2} \log(n)) \\ &\stackrel{\text{Lem } 1.4}{=} P\left(n^{1/2} [\hat{Q}_X^B(u^h(\alpha)) - Q(u^h(\alpha))]\right) \end{aligned}$$



$$\begin{aligned}
&< K + \left[ \frac{z_\alpha \sqrt{u(1-u)}}{f(F^{-1}(u))} + O(n^{-1/2}) \right] + O(n^{-1/2} \log(n)) \\
&\equiv P \left( n^{1/2} \left[ \frac{\hat{Q}_U^B(u^h(\alpha)) - u^h(\alpha)}{f(F^{-1}(u))} \right] < K + \left[ \frac{z_\alpha \sqrt{u(1-u)}}{f(F^{-1}(u))} + O(n^{-1/2}) \right] \right) \\
&\quad + T_l(\alpha) + O(n^{-1/2} \log(n)) \\
&\rightarrow \Phi \left( z_\alpha + \frac{Kf(F^{-1}(u))}{\sqrt{u(1-u)}} \right),
\end{aligned}$$

where  $T_l(\alpha)$  is defined as the error generated by linearization of a stochastic process around  $Q(u)$  and is a special case of the term which is expanded in the appendix; this term is shown to be no larger than  $O(n^{-1/2})$ . Analogous reasoning gives the same result for upper (u) and equal-tail two-sided (t) CIs:

$$\begin{aligned}
\mathcal{P}_n^u(D_n) &\rightarrow \Phi \left( z_\alpha - \frac{Kf(F^{-1}(u))}{\sqrt{u(1-u)}} \right) \quad (\text{for } K < 0), \\
\mathcal{P}_n^t(D_n) &\rightarrow \Phi \left( z_{\alpha/2} + \frac{Kf(F^{-1}(u))}{\sqrt{u(1-u)}} \right) + \Phi \left( z_{\alpha/2} - \frac{Kf(F^{-1}(u))}{\sqrt{u(1-u)}} \right).
\end{aligned}$$

*Remark.* For relatively extreme quantiles, both the Hutson method and our calibration become uncomputable.<sup>7</sup> Note that  $u$  may be required so large (or small) that it needs the  $(n+1)$ th (or the 0th) order statistic, which is not observed. If useful and reasonable lower and/or upper bounds are known on these distributions, they may be used in place of these missing order statistics. More generally, approaching this boundary of computability does not deteriorate the performance of Hutson's method in simulations. Note that as the sample size increases, the range of computable quantiles approaches  $(0, 1)$ . When not computable, methods like that in Chapter 3 may be used. See Chapter 3 for graphs of combinations of  $n$  and  $u$  for which the Hutson CI is computable.

To construct these one-sample  $100(1-\alpha)\%$  CIs, as described in Hutson (1999):

---

<sup>7</sup>Regarding this problem, Ho and Lee (2005a, p. 235) give a seemingly arbitrary formula for extrapolation, discussed further in their simulation study, and note that a priori bounds on the support of  $X$  can help, for example by conservatively using the upper bound for the  $(n+1)$ th order statistic and the lower bound for the 0th.

1. Parameters: determine the sample size  $n$ , quantile of interest  $u \in (0, 1)$ , and desired coverage level  $1 - \alpha$ .
2. Endpoint index computation: (i) For a lower one-sided CI, the upper endpoint index  $u^h$  solves  $\alpha = P\left(\hat{Q}_U^I(u^h) < u\right)$  as in (1.8), where

$$\hat{Q}_U^I(u) \sim \beta((n+1)u, (n+1)(1-u))$$

as in (1.3). Most modern statistical software includes built-in functions to compute the beta distribution CDF and to solve the equation. (ii) Similarly, for an upper one-sided CI, the lower endpoint index  $u^l$  solves  $\alpha = P\left(\hat{Q}_U^I(u^l) > u\right)$ . (iii) For a two-sided CI, construct both upper and lower endpoint the same as for a one-sided CI but with  $\alpha/2$  instead of  $\alpha$ .

3. CI construction: (i) For a lower one-sided CI, compute  $k^h = \lfloor (n+1)u^h \rfloor$ , where  $\lfloor \cdot \rfloor$  is the floor function, and  $\epsilon^h = (n+1)u^h - k^h$ . The upper endpoint is  $\hat{Q}_X^L(u^h) = \epsilon^h X_{n:k^h+1} + (1 - \epsilon^h) X_{n:k^h}$ , as in (1.1). (ii) Similarly, for an upper one-sided CI, the lower endpoint is  $\hat{Q}_X^L(u^l) = \epsilon^l X_{n:k^l+1} + (1 - \epsilon^l) X_{n:k^l}$ , with  $k^l = \lfloor (n+1)u^l \rfloor$  and  $\epsilon^l = (n+1)u^l - k^l$ . (iii) For a two-sided CI, given the  $u^h$  and  $u^l$  already computed, the endpoints are constructed the same as for the one-sided CIs.

Code is available on the latter author's website in MATLAB and R.

### 1.3.2 Joint one-sample inference on multiple quantiles

Alternatively, one may be interested in outcomes at some vector of quantiles,  $\mathbf{u} \in [0, 1]^J$ . For this purpose, we construct joint CIs for the entire vector  $Q(\mathbf{u})$ . Our IDEAL approach is to first use the method of §1.3.1 to calculate a nominal coverage level  $(1 - \tilde{\alpha})$  CI for each  $Q(u_j)$ , which we denote by  $\text{CI}_j^{\tilde{\alpha}}$ . Then the IDEAL  $(1 - \alpha)$  joint CI for  $Q(\mathbf{u})$  is the  $J$ -dimensional rectangle yielded by the Cartesian product of each  $\text{CI}_j^{\tilde{\alpha}}$ ,

$$\text{CI}_{Q(\mathbf{u})}^\alpha \equiv \prod_{j=1}^J \text{CI}_j^{\tilde{\alpha}}. \quad (1.10)$$

By precise selection of  $\tilde{\alpha} \leq \alpha$  (described below), our inference has asymptotically exact coverage.<sup>8</sup> It is possible to achieve better power against certain alternatives by using  $J$  separate  $\tilde{\alpha}_j$  instead of one common  $\tilde{\alpha}$ , but without a compelling reason to do so, that would only complicate the method and increase the computational difficulty of solving for  $\tilde{\alpha}$ , which is currently a simple (one-dimensional) numerical search. Further, the selection of a single  $\tilde{\alpha}$  ensures an even distribution of type I errors across the  $J$  quantiles and thus results in ‘uniformly sensitive’ inference.

By combining one- and two-sided CIs on each of the  $J$  quantiles, there are  $3^J$  different types of joint CI we could construct. We demonstrate only the common case where each  $\text{CI}_j^{\tilde{\alpha}}$  is an equal-tail two-sided CI. In that case, the joint CI is given by

$$\text{CI}_{Q(\mathbf{u})}^{\alpha} \equiv \prod_{j=1}^J \left( \hat{Q}_X^L[u_j^l(\tilde{\alpha})], \hat{Q}_X^L[u_j^h(\tilde{\alpha})] \right), \quad (1.11)$$

and  $\tilde{\alpha}$  is calibrated such that

$$1 - \alpha = P\left( \left[ \bigcap_{j=1}^J \left( \hat{Q}_U^I[u_j^h(\tilde{\alpha})] > u_j \right) \right] \cap \left[ \bigcap_{j=1}^J \left( \hat{Q}_U^I[u_j^l(\tilde{\alpha})] < u_j \right) \right] \right), \quad (1.12)$$

where  $\cap$  denotes the intersection of events and can be read as ‘and.’ Actual coverage is

$$\begin{aligned} & P\left( \left[ \bigcap_{j=1}^J \left( \hat{Q}_X^L[u_j^h(\tilde{\alpha})] > Q(u_j) \right) \right] \cap \left[ \bigcap_{j=1}^J \left( \hat{Q}_X^L[u_j^l(\tilde{\alpha})] < Q(u_j) \right) \right] \right) \\ & \stackrel{\text{Thm 1.3}}{=} P\left( \left[ \bigcap_{j=1}^J \left( \hat{Q}_X^I[u_j^h(\tilde{\alpha})] > Q(u_j) \right) \right] \cap \left[ \bigcap_{j=1}^J \left( \hat{Q}_X^I[u_j^l(\tilde{\alpha})] < Q(u_j) \right) \right] \right) \\ & \quad + O(n^{-1}) \\ & = P\left( \left[ \bigcap_{j=1}^J \left( \hat{Q}_U^I[u_j^h(\tilde{\alpha})] > u_j \right) \right] \cap \left[ \bigcap_{j=1}^J \left( \hat{Q}_U^I[u_j^l(\tilde{\alpha})] < u_j \right) \right] \right) + O(n^{-1}) \\ & = (1 - \alpha) + O(n^{-1}). \end{aligned}$$

The application of Theorem 1.3 above follows from the Cramér-Wold device. We suspect the  $O(n^{-1})$  term is analytic, but we leave its calculation for a future refinement.

---

<sup>8</sup>Coverage of  $Q(\mathbf{u})$  by each  $\text{CI}_j^{\tilde{\alpha}}$  are not asymptotically independent events, but indeed positively correlated. Thus choosing  $(1 - \tilde{\alpha})^J = 1 - \alpha$  will result in  $O(1)$  overcoverage.

Despite being a joint inference technique, it is always clear which quantile(s) (which subset of  $\mathbf{u}$ ) caused the rejection of some null hypothesis  $Q(\mathbf{u})$ . Thus, this test can also be interpreted as controlling the familywise error rate of  $J$  individual hypothesis tests of each quantile.

To construct this  $100(1 - \alpha)\%$  joint CI for  $Q(\mathbf{u})$ :

1. Parameters: determine the sample size  $n$ , the  $J$  quantiles of interest  $u_j \in (0, 1)$ , and the desired coverage level  $1 - \alpha$ .
2. Calibration of  $\tilde{\alpha}$ : using a numerical solver, plus simulated random variables from a beta distribution or numeric integration, solve for  $\tilde{\alpha}$  in (1.12). Let  $B_j^h \equiv \hat{Q}_U^I[u_j^h(\tilde{\alpha})]$  and  $B_j^l \equiv \hat{Q}_U^I[u_j^l(\tilde{\alpha})]$ , where  $u_j^h(\tilde{\alpha})$  and  $u_j^l(\tilde{\alpha})$  are as defined in (1.8). To explain the simulation step, define  $\tilde{\mathbf{u}}$  as a  $2J \times 1$  vector containing all the elements  $u_j^h(\tilde{\alpha})$ ,  $u_j^l(\tilde{\alpha})$  for  $j \in \{1, \dots, J\}$  sorted in ascending order, and let  $\tilde{\mathbf{B}} = \hat{Q}_U^I[\tilde{\mathbf{u}}]$  be the corresponding vector containing elements of type  $B_j^l$  and  $B_j^h$ . We simulate draws of  $\tilde{\mathbf{B}}$  according to

$$\tilde{B}_j = \tilde{B}_{j-1} + (1 - \tilde{B}_{j-1})\Delta_j, \quad \Delta_j \sim \beta((n+1)(\tilde{u}_j - \tilde{u}_{j-1}), (n+1)(1 - \tilde{u}_j))$$

for  $j = 1, \dots, J$ , where  $\tilde{B}_0 = 0$ ,  $\tilde{u}_0 = 0$ , and the  $\Delta_j$  are all independent.

For any given  $\tilde{\alpha}$ , many (e.g.,  $10^5$  or  $10^6$ ) random samples can be drawn, and the probability on the RHS of (1.12) is the proportion of samples in which  $(\cap_{j=1}^J [B_j^h > u_j]) \cap (\cap_{j=1}^J [B_j^l < u_j])$ . The calibrated  $\tilde{\alpha}$  is the value that solves (1.12), which can be found by numerical search.

3. CI construction: individual  $(1 - \tilde{\alpha})$  CIs, denoted  $\text{CI}_j^{\tilde{\alpha}}$ , are constructed for each  $Q(u_j)$  as in §1.3.1. Their product, as in (1.10), is the overall  $(1 - \alpha)$  CI for the vector  $Q(\mathbf{u})$ .

Code is available on the latter author's website in R.

### 1.3.3 One-sample inference on linear combinations of quantiles

For linear combinations of quantiles, inference cannot be entirely distribution free but will require preliminary estimation of nuisance parameters ( $\boldsymbol{\gamma}$ ) in order

to calibrate coverage. Previous CIs for the interquartile range (or general quantile ranges) have been constructed via ‘outer’ and ‘inner’ CIs for the interquartile interval. These have only established conservative CIs (e.g., David and Nagaraja, 2003; Krewski, 1976; Sathe and Lingras, 1981).

Formally, for quantiles  $\mathbf{u} \in (0, 1)^J$  and weights  $\boldsymbol{\psi} \in \mathbb{R}^J$ , we construct a CI for some linear combination functional of the quantile process,

$$D = \sum_{j=1}^J \psi_j Q(u_j).$$

This is a special case of the class of continuous functionals on the quantile process, and it includes the interquartile range when  $\boldsymbol{\psi} = (1, -1)'$  and  $\mathbf{u} = (0.75, 0.25)'$ .

The IDEAL approach is to first use the method of §1.3.1 to calculate a nominal coverage level  $(1 - \tilde{\alpha}[\hat{\gamma}])$  CI for each  $Q(u_j)$ , which we denote by  $\text{CI}_j^{\tilde{\alpha}[\hat{\gamma}]}$ .<sup>9</sup> Then our  $(1 - \alpha)$  CI for  $D$  is defined as the set of all linear combinations possible using points within the individual CIs:

$$\text{CI}_D^\alpha \equiv \{\boldsymbol{\psi}' \mathbf{v} : v_1 \in \text{CI}_1^{\tilde{\alpha}[\hat{\gamma}]}, v_2 \in \text{CI}_2^{\tilde{\alpha}[\hat{\gamma}]}, \dots, v_J \in \text{CI}_J^{\tilde{\alpha}[\hat{\gamma}]}\}. \quad (1.13)$$

By precise selection of  $\tilde{\alpha}[\hat{\gamma}] \geq \alpha$  (described below), our inference is not conservative. As in §1.3.2, it is possible to select  $J$  separate  $\tilde{\alpha}_j$  instead of one common  $\tilde{\alpha}$ , but without a compelling reason to do so, this would only complicate the method and increase the computational difficulty of solving for  $\tilde{\alpha}$ , which is currently a simple (one-dimensional) numerical search.<sup>10</sup> We maintain that Assumption A1.1 holds at each quantile  $u_j$ . For calibration,  $J - 1$  PDF ratios must be estimated. Specifically, for  $j \in \{1, \dots, J\}$ , we define

$$\gamma_j \equiv \frac{f(F^{-1}(u_1))}{f(F^{-1}(u_j))} \quad (1.14)$$

and  $\hat{\gamma}_j$  as the corresponding estimator from (1.28) in §1.3.5.

---

<sup>9</sup>The new notation,  $\tilde{\alpha}[\hat{\gamma}]$ , reflects that in this section (and in §1.3.4 for two-sample inference), the selection of  $\tilde{\alpha}$  depends on estimated nuisance parameters.

<sup>10</sup>There is no first-order tradeoff in power in the selection of  $\tilde{\alpha}$ . As can be inferred from the proof of Theorem 1.5(iii), any selection of  $\tilde{\alpha}$  that controls coverage under the null would not change asymptotic power against any local alternative. This is partly due to having a scalar object of interest,  $D$ ; unlike in §1.3.2, rejections of the null hypothesis are not directly linked to a particular quantile.

We demonstrate this approach first for a lower one-sided CI for  $D$ . In this case, each  $\text{CI}_j^{\tilde{\alpha}}$  is constructed as lower one-sided if  $\psi_j > 0$  and upper one-sided otherwise. In the notation of (1.9),

$$\text{CI}_D^{\tilde{\alpha}[\hat{\gamma}]} = \left( -\infty, \sum_{j=1}^J \psi_j \hat{Q}_X^L(u_j^H(\tilde{\alpha}[\hat{\gamma}])) \right), \quad (1.15)$$

and for any value of the argument  $\gamma$ ,  $\tilde{\alpha}[\gamma]$  is chosen such that

$$1 - \alpha = P\left( \sum_{j=1}^J \psi_j \gamma_j \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha}[\gamma])] - u_j \right) > 0 \right). \quad (1.16)$$

We write  $\hat{\mathbf{u}}$  for the vector of quantiles selected to form CI endpoints given estimates  $\hat{\gamma}$ . In contrast, let  $\mathbf{u}_0$  denote the quantiles that would be selected if the true  $\gamma$  were known. I.e.,

$$\hat{\mathbf{u}} \equiv \{u_j^H(\tilde{\alpha}[\hat{\gamma}])\}_{j=1}^J \text{ and } \mathbf{u}_0 \equiv \{u_j^H(\tilde{\alpha}[\gamma])\}_{j=1}^J.$$

Using  $\approx$  to denote omission of the  $O(n^{-1})$  term below, actual coverage is then

$$\begin{aligned} & P\left(\boldsymbol{\psi}' \hat{Q}_X^L[\hat{\mathbf{u}}] > \boldsymbol{\psi}' Q(\mathbf{u})\right) \\ & \stackrel{\text{Thm 1.3}}{\equiv} P\left(\boldsymbol{\psi}' \hat{Q}_X^I[\hat{\mathbf{u}}] > \boldsymbol{\psi}' Q(\mathbf{u})\right) + O(n^{-1}) \\ & \approx P\left(\sqrt{n} \boldsymbol{\psi}' \left[ F^{-1}\left(\hat{Q}_U^I[\hat{\mathbf{u}}]\right) - F^{-1}(\hat{\mathbf{u}}) \right] > \sqrt{n} \boldsymbol{\psi}' \left[ F^{-1}(\mathbf{u}) - F^{-1}(\hat{\mathbf{u}}) \right]\right) \\ & \equiv E_{\hat{\gamma}} \left\{ P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n} \boldsymbol{\psi}' [F^{-1}(\mathbf{u}) - F^{-1}(\hat{\mathbf{u}})] \mid \hat{\gamma}) \right\}, \end{aligned}$$

where the  $\mathbb{W}$  notation alludes to that in (A.10). Continuing to expand the coverage probability, we get

$$\begin{aligned} & = P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n} \boldsymbol{\psi}' [Q(\mathbf{u}) - Q(\mathbf{u}_0)]) \\ & \quad + E_{\hat{\gamma}} \left\{ P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n} \boldsymbol{\psi}' [Q(\mathbf{u}) - Q(\hat{\mathbf{u}})] \mid \hat{\gamma}) \right. \\ & \quad \quad \left. - P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n} \boldsymbol{\psi}' [Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \right\} \\ & \quad \quad \quad \underbrace{\hspace{10em}}_{(1-\alpha)} \\ & = P\left( \sum_{j=1}^J \psi_j \gamma_j \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha}[\gamma])] - u_j \right) > 0 \right) \end{aligned}$$

$$\begin{aligned}
& + P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n}\psi'[Q(\mathbf{u}) - Q(\mathbf{u}_0)]) - P\left(\sum_{j=1}^J \psi_j \gamma_j \left(\hat{Q}_U^I[u_j^H(\tilde{\alpha}[\gamma])] - u_j\right) > 0\right) \\
& + E_{\hat{\gamma}}\{P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n}\psi'[Q(\mathbf{u}) - Q(\hat{\mathbf{u}})] \mid \hat{\gamma}) \\
& \quad - P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n}\psi'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma})\} \\
& \equiv (1 - \alpha) + T_h + C_h. \tag{1.17}
\end{aligned}$$

The term  $T_h$  comes from the  $J - 1$  first-order Taylor approximations of  $F^{-1}(\cdot)$ , and  $C_h$  comes from estimation error in  $\hat{\gamma}$ . The  $O(n^{-1})$  terms from linear interpolation (Theorem 1.3) have been dropped since they will be of smaller order.

By parallel logic, an upper one-sided CI is

$$\text{CI}_D^\alpha = \left( \sum_{j=1}^J \psi_j \hat{Q}_X^L(u_j^L(\tilde{\alpha}[\hat{\gamma}])), \infty \right), \tag{1.18}$$

and for any value of  $\gamma$ ,  $\tilde{\alpha}[\gamma]$  is such that

$$1 - \alpha = P\left(\sum_{j=1}^J \psi_j \gamma_j \left(\hat{Q}_U^I[u_j^L(\tilde{\alpha}[\gamma])] - u_j\right) < 0\right). \tag{1.19}$$

By a derivation parallel to that of (1.17), actual coverage probability is

$$(1 - \alpha) + T_l + C_l + O(n^{-1}), \tag{1.20}$$

with  $T_l$  and  $C_l$  defined analogously to  $T_h$  and  $C_h$  in (1.17). A two-sided  $(1 - \alpha)$  CI is the intersection of upper one-sided and lower one-sided  $(1 - \alpha/2)$  CIs. The following theorem collects results for one-sample IDEAL inference on linear combinations of quantiles.

**Theorem 1.5.** *Given a sample  $\{X_i\}_{i=1}^n$  with  $X_i \stackrel{iid}{\sim} F$ , where  $F(\cdot)$  satisfies Assumption A1.1 at each of the  $J$  quantiles of interest  $u_j$ , and given a null hypothesis of the form*

$$H_0 : \sum_{j=1}^J \psi_j Q(u_j) = D,$$

*we have the following. Details on estimation of  $\hat{\gamma}_j$  and smoothing parameter choice may be found in §1.3.5.*

- (i) One-sided lower and upper CIs as in (1.15) and (1.18), respectively, have CPE  $O(n^{-1/2})$  if all  $\hat{\gamma}_j$  are estimated via quantile spacings with smoothing parameters  $m_j$  having rate larger than  $n^{1/2}$  and smaller than  $n^{3/4}$ , or equivalently via kernel estimators with bandwidth rate between  $n^{-1/2}$  and  $n^{-1/4}$ .
- (ii) Two-sided CIs, given by the intersection of upper one-sided and lower one-sided  $(1 - \alpha/2)$  CIs, have CPE  $O(n^{-2/3})$  if all  $\hat{\gamma}_j$  are estimated via quantile spacings with smoothing parameters  $m_j$  of order  $n^{2/3}$ , or equivalently via kernel estimators with bandwidth rate  $n^{-1/3}$ .
- (iii) For local alternatives of the form  $D_n = \boldsymbol{\psi}'(Q(\mathbf{u}) + \boldsymbol{\kappa}n^{-1/2})$ , asymptotic power of lower one-sided (l), upper one-sided (u), and equal-tail two-sided (t) CIs are given by

$$\begin{aligned} \mathcal{P}_n^l(D_n) &\rightarrow \Phi\left(z_\alpha + \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_\psi}}\right), & \mathcal{P}_n^u(D_n) &\rightarrow \Phi\left(z_\alpha - \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_\psi}}\right), \\ \mathcal{P}_n^t(D_n) &\rightarrow \Phi\left(z_{\alpha/2} + \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_\psi}}\right) + \Phi\left(z_{\alpha/2} - \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_\psi}}\right), \end{aligned}$$

where  $\mathcal{V}_\psi$  is the variance of  $\boldsymbol{\psi}'\hat{Q}_X^B(\mathbf{u})$  as in Theorem 1.3.

*Proof.* The proof of parts (i) and (ii) is completed by appropriate expansion of the CPE terms in (1.17) and (1.20). In Appendix A.1 it is demonstrated that  $T_h, T_l = O(n^{-1/2})$  and that  $T_h + T_l = O(n^{-1})$ . In §1.3.5 estimation of  $\hat{\gamma}$  is considered, and it is shown that  $C_h, C_l = O(\mathbf{m}^{-1} + (\mathbf{m}/n)^2)$ . Part (iii) is reserved for the appendix.  $\square$

*Remark.* The differences between the one-sided and two-sided results are similar to those in Hall and Sheather (1988, p. 384), who consider population quantile inference via Edgeworth expansion. In their one-sided expansion in (2.6), the  $n^{-1/2}$  term dominates, so the smoothing parameter  $m$  can be anywhere in the range  $n^{1/2}$  to  $n^{3/4}$  and still cause the next two high-order terms,  $O(m^{-1})$  and  $O((m/n)^2)$  just as in our case, to be  $o(n^{-1/2})$ . In their two-sided (2.7), the  $n^{-1/2}$  terms cancel, so the CPE-minimizing rate of  $m$  is exactly determined as  $n^{2/3}$ , which balances the errors from the next two high-order terms at  $O(n^{-2/3})$ . Though our context is more complex, the same ideas and even rates apply. This inference-optimal  $m \propto n^{2/3}$



(or kernel estimator with bandwidth rate  $n^{-1/3}$ ) is different from the MSE-optimal  $n^{4/5}$  (bandwidth rate  $n^{-1/5}$ ).

To construct these one-sample  $100(1 - \alpha)\%$  CIs for linear combinations of quantiles:

1. Parameters: determine the sample size  $n$ , the  $J$  quantiles of interest  $u_j \in (0, 1)$ , and the desired coverage level  $1 - \alpha$ .
2. Density ratio estimation: using the method in §1.3.5, estimate  $f(Q(u_j))$  for all  $j = 1, \dots, J$  and plug into (1.14) to get the  $\hat{\gamma}_j$ . In simulations, using a standard kernel density estimator (evaluated at estimated quantiles  $\hat{Q}_X^L(u_j)$ ) with MSE-optimal bandwidth works similarly well.
3. Calibration of  $\tilde{\alpha}$ : using a numerical solver, plus simulated random variables from a beta distribution or numeric integration, solve for  $\tilde{\alpha}$  in (1.16) for a lower one-sided CI or (1.19) for an upper one-sided CI. A two-sided CI is the intersection of upper and lower one-sided  $(1 - \alpha/2)$  CIs. To simulate (1.16), for example,  $\psi_j$ ,  $\hat{\gamma}_j$ , and  $u_j$  are all known values, and let  $B_j \equiv \hat{Q}_U^L[u_j^H(\tilde{\alpha})]$ , where the  $u_j^H(\tilde{\alpha})$  are in ascending order (by  $j$ ) and defined as in (1.9) (adjusting for negative  $\psi_j$ ), which by reference to (1.8) gives  $u_j^H$  as an implicit function of  $\tilde{\alpha}$ . As in §1.3.2, for  $j = 1, \dots, J$ , draws of  $B_j$  may be simulated by

$$B_j = B_{j-1} + (1 - B_{j-1})\Delta_j,$$

$$\Delta_j \sim \beta((n+1)(u_j^H(\tilde{\alpha}) - u_{j-1}^H(\tilde{\alpha})), (n+1)(1 - u_j^H(\tilde{\alpha}))),$$

where again  $B_0 = 0$ ,  $u_0 = 0$ , and the  $\Delta_j$  are all independent. For any given some  $\tilde{\alpha}$ , many (e.g.,  $10^5$  or  $10^6$ ) random samples can be drawn. The probability on the RHS of (1.16) is estimated by the proportion of samples in which  $\sum_{j=1}^J \psi_j \hat{\gamma}_j (B_j - u_j) > 0$ , and then  $\tilde{\alpha}$  may be found by numerical search.

4. CI construction: individual  $(1 - \tilde{\alpha})$  CIs, denoted  $\text{CI}_j^{\tilde{\alpha}}$ , are constructed for each  $Q(u_j)$ , and the overall  $(1 - \alpha)$  CI for the linear combination is given by (1.13). For a lower one-sided CI,  $\text{CI}_j^{\tilde{\alpha}}$  is lower one-sided if  $\psi_j > 0$  and upper one-sided otherwise; for an upper one-sided CI, the opposite is true. Again, the overall two-sided CI is the intersection of the overall lower and upper one-sided CIs.

Code is available on the latter author’s website in R.

### 1.3.4 Two-sample quantile treatment effect inference

The methods and results of §1.3.3 extend readily to two-sample inference. We assume independent samples are drawn from CDFs  $F_X$  and  $F_Y$  (quantile functions  $Q_X$  and  $Q_Y$ ) with  $n_x$  and  $n_y$  observations, respectively.<sup>11</sup> For example, with individuals randomized into treatment and control groups, IDEAL can produce a CI for the median treatment effect, or for general  $u$ -quantile treatment effect. Applying the results for linear combinations, we can also build a CI for the treatment effect on the interquartile range (a robust measure of spread). Generally, CIs are constructed for

$$D = \sum_{j=1}^J \psi_j [Q_Y(u_j) - Q_X(u_j)].$$

Throughout, we assume that Assumption A1.1 holds at each  $u_j$ -quantile of each distribution, as well as the following assumptions.

**Assumption A1.2.** Independent samples  $\{X_i\}_{i=1}^{n_x}$  and  $\{Y_i\}_{i=1}^{n_y}$  are drawn iid from respective CDFs  $F_X$  and  $F_Y$ . Respective quantile functions are denoted  $Q_X(\cdot) \equiv F_X^{-1}(\cdot)$  and  $Q_Y(\cdot) \equiv F_Y^{-1}(\cdot)$ .

**Assumption A1.3.** Sample sizes grow as  $\lim_{n_x \rightarrow \infty} \sqrt{n_y/n_x} \equiv \mu$ .

*Remark.* The independence in Assumption A1.2 is satisfied if, for example, control and treatment group members do not interact with one another, which is usually true. Other than that, no assumptions are made on the relationship between  $F_X$  and  $F_Y$ . This contrasts with the exchangeability assumption of permutation-type tests that maintain  $F_X = F_Y$  under the null hypothesis, which rules out (e.g.) a treatment that affects the variance of the outcome distribution but not the quantile of interest. Assumption A1.3 is for convenience, letting us write  $n$  as the common rate at which  $n_x \rightarrow \infty$  and  $n_y \rightarrow \infty$ .

The IDEAL two-sample method extends naturally from our one-sample results. First,  $2J$  one-sample  $(1 - \tilde{\alpha})$  CIs are constructed: one for each  $Q_X(u_j)$ ,

---

<sup>11</sup>For more general conditions like censored data, see for example Kosorok (1999).

denoted  $\text{CI}_{x,j}^{\tilde{\alpha}}$ , and one for each  $Q_Y(u_j)$ , denoted  $\text{CI}_{y,j}^{\tilde{\alpha}}$ . The two-sample CI for  $D$  is defined as all linear combinations possible using points within the individual one-sample CIs:

$$\text{CI}_D^\alpha \equiv \{ \boldsymbol{\psi}'(\mathbf{v}_y - \mathbf{v}_x) : \forall j, v_{x,j} \in \text{CI}_{x,j}^{\tilde{\alpha}} \text{ and } v_{y,j} \in \text{CI}_{y,j}^{\tilde{\alpha}} \}. \quad (1.21)$$

For a lower one-sided CI for  $D$ , the individual one-sample CIs for the quantiles of  $F_Y$  will be lower one-sided if  $\psi_j > 0$  and upper one-sided otherwise. The one-sample CIs on the quantiles of  $F_X$  will be upper one-sided if  $\psi_j > 0$  and lower one-sided otherwise. This results in the lower one-sided CI for  $D$  being

$$\text{CI}_D^\alpha = \left( -\infty, \sum_{j=1}^J \psi_j \left[ \hat{Q}_Y^L(u_{y,j}^H(\tilde{\alpha}[\hat{\boldsymbol{\gamma}}_x, \hat{\boldsymbol{\gamma}}_y])) - \hat{Q}_X^L(u_{x,j}^L(\tilde{\alpha}[\hat{\boldsymbol{\gamma}}_x, \hat{\boldsymbol{\gamma}}_y])) \right] \right), \quad (1.22)$$

with the calibration function  $\tilde{\alpha}[\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y]$  implicitly defined by

$$1 - \alpha = P \left( \sum_{j=1}^J \psi_j \left[ \gamma_{y,j} \left( \hat{Q}_{U_y}^I[u_{y,j}^H(\tilde{\alpha})] - u_j \right) - \gamma_{x,j} \left( \hat{Q}_{U_x}^I[u_{x,j}^L(\tilde{\alpha})] - u_j \right) \right] > 0 \right). \quad (1.23)$$

The density ratios are both defined relative to  $f_X(Q_X(u_1))$ , as

$$\gamma_{x,j} \equiv \frac{f_X(Q_X(u_1))}{f_X(Q_X(u_j))}, \quad \gamma_{y,j} \equiv \frac{f_X(Q_X(u_1))}{f_Y(Q_Y(u_j))}, \quad (1.24)$$

with estimators (denoted by hats) given in (1.29) in §1.3.5.

Upper one-sided CIs are similarly given by

$$\text{CI}_D^\alpha = \left( \sum_{j=1}^J \psi_j \left[ \hat{Q}_Y^L(u_{y,j}^L(\tilde{\alpha})) - \hat{Q}_X^L(u_{x,j}^H(\tilde{\alpha})) \right], \infty \right), \quad (1.25)$$

with calibration function  $\tilde{\alpha}[\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y]$  implicitly defined by

$$1 - \alpha = P \left( \sum_{j=1}^J \psi_j \left[ \gamma_{y,j} \left( \hat{Q}_{U_y}^I[u_{y,j}^L(\tilde{\alpha})] - u_j \right) - \gamma_{x,j} \left( \hat{Q}_{U_x}^I[u_{x,j}^H(\tilde{\alpha})] - u_j \right) \right] < 0 \right). \quad (1.26)$$

Two-sided CIs are again given by the intersection of upper one-sided and lower one-sided  $(1 - \alpha/2)$  CIs.

We now calculate the actual coverage probability for the lower one-sided CI, up to the expansion of familiar terms. Let

$$\begin{aligned}\hat{\mathbf{u}}_{\mathbf{y}} &= \{u_{y,j}^H(\tilde{\alpha}[\hat{\boldsymbol{\gamma}}_{\mathbf{x}}, \hat{\boldsymbol{\gamma}}_{\mathbf{y}}])\}_{j=1}^J, & \mathbf{u}_{\mathbf{0},\mathbf{y}} &= \{u_{y,j}^H(\tilde{\alpha}[\boldsymbol{\gamma}_{\mathbf{x}}, \boldsymbol{\gamma}_{\mathbf{y}}])\}_{j=1}^J, \\ \hat{\mathbf{u}}_{\mathbf{x}} &= \{u_{x,j}^L(\tilde{\alpha}[\hat{\boldsymbol{\gamma}}_{\mathbf{x}}, \hat{\boldsymbol{\gamma}}_{\mathbf{y}}])\}_{j=1}^J, & \mathbf{u}_{\mathbf{0},\mathbf{x}} &= \{u_{x,j}^L(\tilde{\alpha}[\boldsymbol{\gamma}_{\mathbf{x}}, \boldsymbol{\gamma}_{\mathbf{y}}])\}_{j=1}^J.\end{aligned}$$

Referring back to (1.22), and with steps similar to the derivation of (1.17), true coverage is

$$\begin{aligned}& P\left(\boldsymbol{\psi}'\left[\hat{Q}_Y^L(\hat{\mathbf{u}}_{\mathbf{y}}) - \hat{Q}_X^L(\hat{\mathbf{u}}_{\mathbf{x}})\right] > \boldsymbol{\psi}'[Q_Y(\mathbf{u}) - Q_X(\mathbf{u})]\right) \\ & \stackrel{\text{Thm 1.3}}{=} P\left(\boldsymbol{\psi}'\left[\hat{Q}_Y^I(\hat{\mathbf{u}}_{\mathbf{y}}) - \hat{Q}_X^I(\hat{\mathbf{u}}_{\mathbf{x}})\right] > \boldsymbol{\psi}'[Q_Y(\mathbf{u}) - Q_X(\mathbf{u})]\right) + O(n^{-1}) \\ & \approx P\left(\boldsymbol{\psi}'\left[F_Y^{-1}(\hat{Q}_{U_Y}^I(\hat{\mathbf{u}}_{\mathbf{y}})) - F_X^{-1}(\hat{Q}_{U_X}^I(\hat{\mathbf{u}}_{\mathbf{x}}))\right] > \boldsymbol{\psi}'[F_Y^{-1}(\mathbf{u}) - F_X^{-1}(\mathbf{u})]\right) \\ & = E_{\hat{\boldsymbol{\gamma}}_{\mathbf{x}}, \hat{\boldsymbol{\gamma}}_{\mathbf{y}}}\left\{P\left(\mathbb{W}_{\hat{\mathbf{u}}_{\mathbf{y}}} - \mathbb{W}_{\hat{\mathbf{u}}_{\mathbf{x}}}\right. \right. \\ & \quad \left. \left. > \sqrt{n}\boldsymbol{\psi}'\left[(F_Y^{-1}[\mathbf{u}] - F_Y^{-1}[\hat{\mathbf{u}}_{\mathbf{y}}]) - (F_X^{-1}[\mathbf{u}] - F_X^{-1}[\hat{\mathbf{u}}_{\mathbf{x}}])\right] \mid \hat{\boldsymbol{\gamma}}_{\mathbf{x}}, \hat{\boldsymbol{\gamma}}_{\mathbf{y}}\right)\right\} \\ & = (1 - \alpha) + T_{h,2} + C_{h,2},\end{aligned}\tag{1.27}$$

where  $\approx$  merely denotes dropping the  $O(n^{-1})$  term. Similar to before, the term  $T_{h,2}$  is the CPE induced by the required  $2J - 1$  linearizations of our unknown inverse CDFs at our quantiles of interest, and  $C_{h,2}$  is CPE induced by preliminary estimation of the density ratios. The CPE results are identical to the general one-sample results in Theorem 1.5, as summarized in the following corollary.

**Corollary 1.6.** *Under Assumption A1.1 for both  $F_X$  and  $F_Y$ , as well as Assumptions A1.2 and A1.3:*

- (i) *One-sided lower and upper CIs as in (1.22) and (1.25), respectively, have CPE of  $O(n^{-1/2})$  if all  $\hat{\gamma}_{x,j}$  and  $\hat{\gamma}_{y,j}$  are estimated via quantile spacings with smoothing parameters  $m_{x,j}$  and  $m_{y,j}$  having rate larger than  $n^{1/2}$  and smaller than  $n^{3/4}$ , or equivalently via kernel estimators with bandwidth rate between  $n^{-1/2}$  and  $n^{-1/4}$ .*
- (ii) *Two-sided CIs, given by the intersection of upper one-sided and lower one-sided  $(1 - \alpha/2)$  CIs, have CPE of  $O(n^{-2/3})$  if all  $\hat{\gamma}_{x,j}$  and  $\hat{\gamma}_{y,j}$  are estimated*

via quantile spacings with smoothing parameters  $m_{x,j}$  and  $m_{y,j}$  of order  $n^{2/3}$ , or equivalently via kernel estimators with bandwidth rate  $n^{-1/3}$ .

(iii) For local alternatives of the form  $D_n = \boldsymbol{\psi}'([Q_Y(\mathbf{u}) - Q_X(\mathbf{u})] + \boldsymbol{\kappa}n^{-1/2})$ , asymptotic power of lower one-sided ( $l$ ), upper one-sided ( $u$ ), and equal-tail two-sided ( $t$ ) CIs are given by

$$\begin{aligned} \mathcal{P}_n^l(D_n) &\rightarrow \Phi\left(z_\alpha + \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}, \mathbf{x}} + \mathcal{V}_{\boldsymbol{\psi}, \mathbf{y}}}}\right), & \mathcal{P}_n^u(D_n) &\rightarrow \Phi\left(z_\alpha - \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}, \mathbf{x}} + \mathcal{V}_{\boldsymbol{\psi}, \mathbf{y}}}}\right), \\ \mathcal{P}_n^t(D_n) &\rightarrow \Phi\left(z_{\alpha/2} + \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}, \mathbf{x}} + \mathcal{V}_{\boldsymbol{\psi}, \mathbf{y}}}}\right) + \Phi\left(z_{\alpha/2} - \frac{\sum_{j=1}^J \psi_j \kappa_j}{\sqrt{\mathcal{V}_{\boldsymbol{\psi}, \mathbf{x}} + \mathcal{V}_{\boldsymbol{\psi}, \mathbf{y}}}}\right), \end{aligned}$$

where for  $z \in \{x, y\}$ ,  $\mathcal{V}_{\boldsymbol{\psi}, z}$  is the variance of  $\boldsymbol{\psi}'\hat{Q}_z^B(\mathbf{u})$  as in Theorem 1.3.

*Proof.* The proof parallels the one-sample case of Theorem 1.5. The two stochastic processes are independent and thus pose no additional difficulty in calibrating coverage.  $\square$

To construct these two-sample  $100(1 - \alpha)\%$  CIs for linear combinations of quantiles:

1. Parameters: determine the sample sizes  $n_x$  and  $n_y$ , the  $J$  quantiles of interest  $\mathbf{u} \in (0, 1)^J$ , and the desired coverage level  $1 - \alpha$ .
2. Density ratio estimation: using the method<sup>12</sup> in (1.29) in §1.3.5, estimate  $f_X(Q_X(u_j))$  and  $f_Y(Q_Y(u_j))$  for all  $j = 1, \dots, J$  and plug into (1.24) to get the  $\hat{\gamma}_{x,j}$  and  $\hat{\gamma}_{y,j}$ .
3. Calibration of  $\tilde{\alpha}$ : using a numerical solver, plus simulated random variables from a beta distribution or numeric integration, solve for  $\tilde{\alpha}$  in (1.23) for a lower one-sided CI or (1.26) for an upper one-sided CI. (See the construction in §1.3.3 for details.) Simulation can proceed as in §1.3.3, with the independence of the two samples allowing us to separately draw realizations of both quantile

---

<sup>12</sup>In simulations, using a standard kernel density estimator (evaluated at estimated quantiles  $\hat{Q}_X^L(u_j)$  and  $\hat{Q}_Y^L(u_j)$ ) with MSE-optimal bandwidth worked similarly well.

processes. In the case of lower one-sided CI we draw

$$\begin{aligned} B_{x,j} &= B_{x,j-1} + (1 - B_{x,j-1})\Delta_{x,j}, \\ \Delta_{x,j} &\sim \beta((n_x + 1)(u_{x,j}^H(\tilde{\alpha}) - u_{x,j-1}^H(\tilde{\alpha})), (n_x + 1)(1 - u_{x,j}^H(\tilde{\alpha}))), \\ B_{y,j} &= B_{y,j-1} + (1 - B_{y,j-1})\Delta_{y,j}, \\ \Delta_{y,j} &\sim \beta((n_y + 1)(u_{y,j}^L(\tilde{\alpha}) - u_{y,j-1}^L(\tilde{\alpha})), (n_y + 1)(1 - u_{y,j}^L(\tilde{\alpha}))) \end{aligned}$$

where  $B_{x,0} = B_{y,0} = 0$ ,  $u_{x,0}^H = u_{y,0}^L = 0$ , and the  $\Delta$  are all independent. Drawing many samples allows us to calculate the RHS of (1.23) as the proportion of samples in which

$$\sum_{j=1}^J \psi_j \left[ \hat{\gamma}_{y,j}(B_{y,j} - u_j) - \hat{\gamma}_{x,j}(B_{x,j} - u_j) \right] > 0,$$

which is implicitly a function of  $\tilde{\alpha}$ . Then, the calibrated value of  $\tilde{\alpha}$  that solves (1.23) may be found by numerical search. For  $J = 1$ , the normal approximation discussed in §1.3.6 also yields an approximate solution to  $\tilde{\alpha}$ . Setting  $\hat{\theta} \equiv (1 + \hat{\gamma}/\mu)/\sqrt{1 + (\hat{\gamma}/\mu)^2}$ ,  $\tilde{\alpha} = \Phi(z_\alpha/\hat{\theta})$  for one-sided CIs or  $\tilde{\alpha}/2 = \Phi(z_{\alpha/2}/\hat{\theta})$  for two-sided;  $\mu \equiv \sqrt{n_y/n_x}$ , and  $\hat{\gamma}$  is the estimator of  $f_X(Q_X(u))/f_Y(Q_Y(u))$ .

4. CI construction: individual  $(1 - \tilde{\alpha})$  CIs, denoted  $\text{CI}_{x,j}^{\tilde{\alpha}}$  and  $\text{CI}_{y,j}^{\tilde{\alpha}}$ , are constructed for each  $Q_X(u_j)$  and  $Q_Y(u_j)$ , and the overall  $(1 - \alpha)$  CI for the linear combination is given by (1.21). For a lower one-sided CI,  $\text{CI}_{y,j}^{\tilde{\alpha}}$  is lower one-sided if  $\psi_j > 0$  and upper one-sided otherwise, while  $\text{CI}_{x,j}^{\tilde{\alpha}}$  is upper one-sided if  $\psi_j > 0$  and lower one-sided otherwise; for an upper one-sided CI, the opposite is true. The overall two-sided CI is the intersection of the overall lower and upper one-sided  $(1 - \alpha/2)$  CIs.

MATLAB code for  $J = 1$  and R code for  $J \geq 1$  are available on the latter author's website.

### 1.3.5 Estimation of $\gamma$

Selection of  $\tilde{\alpha}$  in §1.3.3 and §1.3.4 requires preliminary estimation of various density ratios at the quantiles of interest. In both cases we recommend estimation

by an inverse ratio of sparsities (denoted  $g$ , or  $\hat{g}$  for estimators) of the same form. In (1.14) we specify the one-sample definition of  $\gamma_j$ , presented again here alongside its estimator,

$$\gamma_j \equiv \frac{f(Q(u_1))}{f(Q(u_j))}, \quad \hat{\gamma}_j \equiv \frac{\hat{g}_j}{\hat{g}_1} \equiv \left( \frac{X_{n:(n+1)u_j+m_j} - X_{n:(n+1)u_j-m_j}}{X_{n:(n+1)u_1+m_1} - X_{n:(n+1)u_1-m_1}} \right) \left( \frac{m_1}{m_j} \right). \quad (1.28)$$

Referring back to (1.24), the two-sample equivalent is, for  $z \in \{x, y\}$ ,

$$\gamma_{z,j} \equiv \frac{f_X(Q_X(u_1))}{f_Z(Q_Z(u_j))}, \quad (1.29)$$

$$\hat{\gamma}_{z,j} \equiv \frac{\hat{g}_{z,j}}{\hat{g}_{x,1}} \equiv \left( \frac{Z_{n_z:(n_z+1)u_j+m_{z,j}} - Z_{n_z:(n_z+1)u_j-m_{z,j}}}{X_{n_x:(n_x+1)u_1+m_{x,1}} - X_{n_x:(n_x+1)u_1-m_{x,1}}} \right) \left( \frac{m_{x,1}/n_x}{m_{z,j}/n_z} \right). \quad (1.30)$$

While traditionally  $m$  is a natural number and the order statistic indices are also natural numbers such as  $\lfloor (n+1)u \rfloor + 1 + m$  and  $\lfloor (n+1)u \rfloor + 1 - m$  (Hall and Sheather, 1988, eqn. 1.1), we can also interpret (e.g.)  $X_{n:r}$  above as the corresponding linearly interpolated fractional order statistic. This does not affect the theoretical properties, but may be useful in particularly small samples.

We consider the two above cases together. The sparsity estimation will happen in a small (and shrinking) neighborhood of each quantile, so the asymptotic properties of  $\hat{\gamma}$  are not affected by multiple quantiles. These properties will, however, depend on the selection of the spacing parameter, indicated by  $m$ , with subscript for quantile and sample (the sample subscript is omitted in the one-sample case, as in (1.28)).

For clarity, we consider a general  $\gamma = f_X(Q_X(p))/f_Y(Q_Y(p)) = g_y/g_x$  and its estimator in log form,

$$\log \hat{\gamma} = \log \hat{g}_y - \log \hat{g}_x.$$

Using (2.5) of Bloch and Gastwirth (1968) for  $\text{Var}(\hat{g}) = g^2/(2m)$ , plus the expansion

$$\log \hat{g}_y = \log g_y + \frac{\hat{g}_y - g_y}{g_y} - \frac{(\hat{g}_y - g_y)^2}{2g_y^2} + O_p((\hat{g}_y - g_y)^3),$$

it is immediate that up to smaller-order terms

$$\begin{aligned}\text{Var}(\log \hat{g}_y) &= (2m_y)^{-1}, \\ E[\log \hat{g}_y] &= \log g_y + \frac{E(\hat{g}_y) - g_y}{g_y} - \frac{\text{Var}(\hat{g}_y)}{2g_y^2} = \log g_y + B_y \left( \frac{m_y}{n_y} \right)^2 - \frac{1}{4m_y},\end{aligned}$$

where using (2.6) of Bloch and Gastwirth (1968) gives

$$B_y \equiv (1/6)Q_Y'''(p) = \frac{3f_Y'(F_Y^{-1}(p))^2 - f_Y''(F_Y^{-1}(p))f_Y(F_Y^{-1}(p))}{6f_Y(F_Y^{-1}(p))^4}.$$

For a normal distribution,<sup>13</sup>

$$B_y = \frac{2[\Phi^{-1}(p)]^2 + 1}{6[\phi(\Phi^{-1}(p))]^2}. \quad (1.31)$$

Combined with symmetric expressions for  $g_x$ ,

$$\text{Var}(\log \hat{\gamma}) = (m_y^{-1} + m_x^{-1})/2, \quad (1.32)$$

$$\text{Bias}(\log \hat{\gamma}) = \left[ B_y \left( \frac{m_y}{n_y} \right)^2 - \frac{m_y^{-1}}{4} \right] - \left[ B_x \left( \frac{m_x}{n_x} \right)^2 - \frac{m_x^{-1}}{4} \right]. \quad (1.33)$$

The usual MSE-optimal rate for the smoothing parameters is  $n^{4/5}$ , which equates the orders of the variance and squared bias of  $\hat{\gamma}$ . However, the CPE-optimal rate in our problem equates the orders of the variance and bias of  $\hat{\gamma}$  (see appendix for details). As seen from (1.32) and (1.33), this CPE-optimal rate is  $m_x, m_y \asymp n^{2/3}$ . Similar results are derived in Hall and Sheather (1988) and Chapter 3.

We recommend zeroing out the log bias of each individual sparsity under the assumption that each underlying density is normal. This maintains the optimal CPE rate while erring slightly on the conservative side (under normality, and in many other cases). We refer back to (1.31) and choose

$$m_{z,j} = n_z^{2/3} \left( \frac{3[\phi(\Phi^{-1}(p))]^2}{2 + 4[\Phi^{-1}(p)]^2} \right)^{1/3}. \quad (1.34)$$

---

<sup>13</sup>Note that the expression for the bias of the logged estimator is invariant to the scale of the distribution. This is an advantage over the non-logged bias that requires preliminary estimation of  $\hat{\sigma}_y^2$  for implementation of a bandwidth rule.



In simulations, our method was not sensitive to either the choice of bandwidth or the choice of estimation method. The results were quite good with our quantile spacing estimators using zero-bias smoothing parameters, and they were extremely similar with kernel density estimators using Silverman's (1986) rule of thumb bandwidth, which minimizes MSE under normality.

### 1.3.6 Further approximation and intuition (two-sample)

In the two-sample case with  $J = 1$ , we explore further approximations that have computational benefits and theoretical insights. The upper one-sided example is used for clarity. Instead of calibrating  $\tilde{\alpha}$  via numerical integration or simulation of the beta distribution, we could approximate the two independent beta random variables with normals. This is justified by the almost  $O(n^{-1/2})$  CDF link between the marginals of our 'ideal' and Brownian bridge processes, established in Theorem 1.3.

Let

$$B_x \sim \beta[u_x^h(n_x + 1), (1 - u_x^h)(n_x + 1)], \quad B_y \sim \beta[u_y^l(n_y + 1), (1 - u_y^l)(n_y + 1)].$$

With the normal approximation,  $X \perp\!\!\!\perp Y$ , and denoting the standard normal CDF by  $\Phi(\cdot)$ , we have the following uniform bounds:

$$\begin{aligned} \sup_K \left| P(B_x - u_x^h < K) - \Phi\left(\frac{K}{\sqrt{u_x^h(1 - u_x^h)/n_x}}\right) \right| &= O(n^{-1/2} \log(n)), \\ \sup_K \left| P(B_y - u_y^l < K) - \Phi\left(\frac{K}{\sqrt{u_y^l(1 - u_y^l)/n_y}}\right) \right| &= O(n^{-1/2} \log(n)). \end{aligned}$$

By convolution, we calculate the coverage probability in (1.26) as<sup>14</sup>

$$\begin{aligned} P(B_x - u < \gamma[B_y - u]) &= P\left([B_x - u_x^h] - \gamma[B_y - u_y^l] < [u - u_x^h] + \gamma[u_y^l - u]\right) \\ &= \Phi\left(\frac{[u - u_x^h] + \gamma[u_y^l - u]}{\sqrt{\gamma^2 u_y^l(1 - u_y^l)/n_y + u_x^h(1 - u_x^h)/n_x}}\right) \\ &\quad + O(n^{-1/2} \log(n)). \end{aligned} \tag{1.35}$$

---

<sup>14</sup>Smaller-order remainder terms from interpolation, estimation of  $\hat{\gamma}$ , and local linearization of the distribution are omitted.

To approximate  $u_x^h$  and  $u_y^l$ , we recall Lemma 1.4 and note the following corollary.

**Corollary 1.7.** *Using Lemma 1.4,*

$$\begin{aligned}\frac{u_x^h(1-u_x^h)}{u(1-u)} &= 1 + n_x^{-1/2} z_{1-\tilde{\alpha}}(1-2u)/\sqrt{u(1-u)} + O(n_x^{-1}) = 1 + O(n_x^{-1/2}), \\ \frac{u_y^l(1-u_y^l)}{u(1-u)} &= 1 - \mu^{-1} n_x^{-1/2} z_{1-\tilde{\alpha}}(1-2u)/\sqrt{u(1-u)} + O(n_x^{-1}) = 1 + O(n_x^{-1/2}).\end{aligned}$$

If  $u = 0.5$ , then  $1 - 2u = 0$  and both expressions become  $1 + O(n_x^{-1})$ .

*Proof of corollary.*

$$\begin{aligned}\frac{u_x^h(1-u_x^h)}{u(1-u)} &= \frac{(u + (u_x^h - u))(1 - u - (u_x^h - u))}{u(1-u)} = 1 + \frac{1-2u}{u(1-u)}(u_x^h - u) + O(n_x^{-1}) \\ &= 1 + z_{1-\tilde{\alpha}}(1-2u)/\sqrt{n_x u(1-u)} + O(n_x^{-1}),\end{aligned}$$

and similarly for  $u_y^l$ . □

We first apply Lemma 1.4 to (1.35), yielding

$$\begin{aligned}P(B_x - u < \gamma[B_y - u]) &= \Phi\left(z_{1-\tilde{\alpha}} \frac{\sqrt{u(1-u)} \left[ n_x^{-1/2} + \gamma n_y^{-1/2} + O(n^{-1}) \right]}{\sqrt{u_x^h(1-u_x^h)/n_x + \gamma^2 u_y^l(1-u_y^l)/n_y}}\right) \\ &\quad + O(n^{-1/2} \log(n)).\end{aligned}$$

Then we multiply numerator and denominator by  $n_x^{1/2}$ , strengthen Assumption A1.3 slightly to  $\sqrt{n_y/n_x} = \mu + O(n^{-1/2})$ , and rearrange to get

$$\begin{aligned}&= \Phi\left(z_{1-\tilde{\alpha}} \frac{1 + (\gamma/\mu)}{\sqrt{\frac{u_x^h(1-u_x^h)}{u(1-u)} + (\gamma/\mu)^2 \frac{u_y^l(1-u_y^l)}{u(1-u)}}}\right) + O(n^{-1/2} \log(n)) \\ &\stackrel{\text{Cor 1.7}}{=} \Phi\left(z_{1-\tilde{\alpha}} \frac{1 + (\gamma/\mu)}{\sqrt{1 + (\gamma/\mu)^2}}\right) + O(n^{-1/2} \log(n)).\end{aligned}$$

Under exchangeability and equal sample sizes,  $\gamma = 1$  and  $\mu = 1$ , so we can solve for  $\tilde{\alpha} = \Phi(z_\alpha/\sqrt{2})$ . However, exchangeability is a strong assumption. Alternatively, we can estimate  $\gamma$  by  $\hat{\gamma}$ , plug in  $\mu = \sqrt{n_y/n_x}$ , and solve for  $\tilde{\alpha}$ . This is what we suggest for practice and what we do in our simulations.

While the foregoing has been for the upper one-sided case, the calibration equation turns out to be identical for the lower one-sided case. Consequently, using  $\alpha/2$  for the two one-sided cases yields the two-sided

$$\tilde{\alpha}/2 = \Phi(z_{\alpha/2}/\theta^*), \quad \theta^* \equiv \frac{1 + \gamma/\mu}{\sqrt{1 + (\gamma/\mu)^2}}. \quad (1.36)$$

This means that the two-sample, two-sided CI is an equal-tail CI, up to smaller-order terms. We collect all these results in a corollary.

**Corollary 1.8.** *Under Assumptions A1.1 and A1.2, and strengthening Assumption A1.3 to  $\sqrt{n_y/n_x} = \mu + O(n^{-1/2})$ :*

- (i) *For two-sample, one-sided CIs using the normal-approximated calibration  $\tilde{\alpha} = \Phi(z_\alpha/\theta^*)$ , with  $\theta^*$  in (1.36), CPE is  $O(n^{-1/2} \log(n))$ .*
- (ii) *For two-sample, two-sided CIs using the normal-approximated calibration  $\tilde{\alpha}/2 = \Phi(z_{\alpha/2}/\theta^*)$ , the CPE in Corollary 1.6(ii) increases to  $O(n^{-1/2})$ . The rate requirements on smoothing parameters  $m_x$  and  $m_y$  relax to the requirements of Corollary 1.6(i). To the first-order, these are equal-tail CIs.*

Consider the two extreme special cases. Since  $\gamma \in (0, \infty)$ , then  $\theta^* \in (1, \sqrt{2}]$ . For instance, if  $n_x = n_y$  and  $f_X(Q_X(p)) = f_Y(Q_Y(p))$ , then  $\gamma = 1$  and  $\theta^* = \sqrt{2}$ . As  $f_Y \rightarrow \infty$ , meaning that the distribution of  $Y$  is collapsing to a constant,  $\gamma \rightarrow 0$ ,  $\theta^* \rightarrow 1$ , and  $\alpha = \tilde{\alpha}$ . One-sample inference is when  $Y$  is actually a constant, in which case  $\tilde{\alpha} = \alpha$  is correct; this is indeed what we get using  $\theta^* = 1$ . This continuity between the one- and two-sample settings is reassuring.

For additional intuition, consider the classical Behrens-Fisher problem of testing equality of the means  $E(X)$  and  $E(Y)$  under uncertainty about the ratio of their variances. A standard test rejects if  $|\bar{X} - \bar{Y}| > z_{1-\alpha/2} \sigma_{\bar{X}-\bar{Y}}$ . Note that since  $X \perp Y$ ,  $\text{Var}(\bar{X} - \bar{Y}) = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2$ . The analogous hypothesis test is based on one-sample CIs  $\bar{X} \pm z_{1-\tilde{\alpha}/2} \sigma_{\bar{X}}$  and  $\bar{Y} \pm z_{1-\tilde{\alpha}/2} \sigma_{\bar{Y}}$ . It rejects if  $\bar{X} - z_{1-\tilde{\alpha}/2} \sigma_{\bar{X}} > \bar{Y} + z_{1-\tilde{\alpha}/2} \sigma_{\bar{Y}}$  or  $\bar{Y} - z_{1-\tilde{\alpha}/2} \sigma_{\bar{Y}} > \bar{X} + z_{1-\tilde{\alpha}/2} \sigma_{\bar{X}}$ , i.e. if  $|\bar{X} - \bar{Y}| > z_{1-\tilde{\alpha}/2} (\sigma_{\bar{X}} + \sigma_{\bar{Y}})$ .

Setting the critical values equal and solving for  $\tilde{\alpha}$  gives

$$z_{1-\alpha/2}\sigma_{\bar{X}-\bar{Y}} = z_{1-\tilde{\alpha}/2}(\sigma_{\bar{X}} + \sigma_{\bar{Y}}),$$

$$\tilde{\alpha}/2 = \Phi\left(z_{\alpha/2}\sigma_{\bar{X}-\bar{Y}}/(\sigma_{\bar{X}} + \sigma_{\bar{Y}})\right) = \Phi\left(z_{\alpha/2}\frac{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}\right).$$

If  $\sigma_x = \sigma_y$  and  $n_x = n_y$ , then  $\tilde{\alpha}/2 = \Phi(z_{\alpha/2}/\sqrt{2})$ , the same result as for  $\theta^* = \sqrt{2}$ . More generally calibration of  $\tilde{\alpha}$  requires crucially on the ratio of population variances. The two-sample IDEAL procedure solves our quantile inference problem by estimating a density ratio in much the same way that estimating a ratio of sample variances might solve the Behrens-Fisher problem.

## 1.4 Empirical application

The following results can be obtained using either the MATLAB function `quantile_inf` or the R function `quantile.inf` that are both publicly available through either MATLAB File Exchange (Kaplan, 2011) or that author’s website. The functions require only the data, quantile of interest, and  $\alpha$ . To demonstrate our new two-sample IDEAL quantile inference, we use the experimental data from Gneezy and List (2006, Tables I and V). In short, individuals in the control group work for a certain hourly wage (as advertised), while individuals in the treatment group are surprised with a larger hourly wage upon arrival. The goal is to investigate “gift exchange,” specifically whether the higher wages induce higher effort (as measured by productivity). The experiment is run separately for two different tasks: data entry for a library (typing in a book’s author, title, etc.), and door-to-door fundraising for a non-profit. Productivity (number of books entered, or dollars raised) is measured for each participant for different periods of time: in four 90-minute segments for the library task, and before/after lunch for fundraising. The sample sizes are small: 10 and 9 for control and treatment, respectively, for the library task, and 10 and 13 for control and treatment for fundraising.

The main result of Gneezy and List (2006) is that the “gift” treatment raises productivity significantly in the first period, but not significantly thereafter.

We do not simply re-test this main result (though we indeed support it) but rather offer complementary analysis on quantile treatment effects.

For the library task, Gneezy and List (2006) performed two types of one-sided 5% tests: a Wilcoxon rank-sum (a.k.a. Mann–Whitney–Wilcoxon or Mann–Whitney  $U$ ) test, and an unequal variances  $t$ -test. For the first 90-minute period, the null was (barely) rejected by each test in favor of the treatment productivity being higher. For the remaining 90-minute periods, the null was not rejected by either test. The goal of the rank-sum test is to reject if  $P(T > C) > 0.5$ , where  $T$  is a random variable corresponding to treatment group productivity and similarly for  $C$  and the control group. The  $t$ -test instead tests for equality of means, though the assumption of normality of the sample average productivities is questionable with such a small sample size (too small to rely on the CLT).

Complementing these original tests for the library task, our IDEAL method tests for equality at a chosen quantile of the distribution. Also using a one-sided 5% test, we do not reject the null of equality at the lower quartile (25th percentile) or the median, but we do reject at the upper quartile. The IDEAL two-sided CIs with nominal 90% coverage (which give the 95% one-sided CIs by simply taking the lower endpoint, since they are equal-tailed) are given in Table 1.1. The results are consistent with the rank-sum result that the first period treatment productivity is higher overall in some sense, and consistent with the  $t$ -test result that the mean is higher. Our test further suggests that the shift comes primarily (though not exclusively) from the upper part of the distribution: for the library task in the first period, the gift wage seems to induce the most productive workers to become extremely productive, while the effect is much less (if any) on less productive workers. For periods 2–4 (the second period is also shown in the table), our test fails to reject the null at any of these quartiles, consistent with the original results.

For the fundraising task, Gneezy and List (2006) report one-sided 1% significance for the rank-sum test in the first three-hour “pre-lunch” period of fundraising.<sup>15</sup> We again use a 5% one-sided test (10% two-sided), this time finding sig-

---

<sup>15</sup>They do not report the rank-sum result on the post-lunch three-hour period, but given the data and the sample size, it seems clear that the test would not reject for any conventional significance level. They also do not report a  $t$ -test this time.

**Table 1.1:** IDEAL two-sided 90% confidence intervals for quartile treatment effects in Gneezy and List (2006).

Period (method <sup>a</sup> )	Lower quartile	Median	Upper quartile
<i>Library task<sup>b</sup></i>			
1 (kern/MSE)	(-10.22,21.39)	(-2.39,26.18)	(2.35,29.95)
1 (spac/bias)	(-10.30,21.49)	(-2.38,26.17)	(2.48,29.73)
2 (kern/MSE)	(-15.65,8.63)	(-7.61,27.93)	(-16.93,28.78)
2 (spac/bias)	(-16.35,8.89)	(-7.63,28.06)	(-15.98,28.21)
<i>Fundraising task<sup>c</sup></i>			
1 (kern/MSE)	(7.49,26.41)	(-0.00,26.78)	(-14.84,26.21)
1 (spac/bias)	(7.56,26.29)	(-0.76,26.96)	(-14.08,22.85)
2 (kern/MSE)	(-4.03,13.72)	(-5.62,11.49)	(-15.44,11.06)
2 (spac/bias)	(-3.89,13.68)	(-5.67,11.53)	(-19.64,11.94)

<sup>a</sup>Methods for  $\hat{\gamma}$  estimation are abbreviated “kern/MSE” for a kernel density estimator with Silverman’s (1986) rule of thumb bandwidth (MSE-optimal under normality), and “spac/bias” for a quantile spacing estimator with our zero-bias smoothing parameter.

<sup>b</sup>Units: books logged per period (90 minutes).

<sup>c</sup>Units: dollars raised per period (three hours).

nificance at the lower quartile and almost at the median, but not at the upper quartile. The IDEAL two-sided CIs are again in Table 1.1. These results are also consistent with the original results, as is our failure to reject the null at any quartile in the post-lunch period (period 2). Additionally, our test tells us that it is the less-productive workers who are relatively more induced by the gift wage for this task, in contrast to the library task results.

## 1.5 Simulation study

Code for all methods described in §1.3 is available in R from the latter author’s website, and code for the methods in §1.3.1 and §1.3.4 is available in MATLAB from either the latter author’s website or MATLAB File Exchange (Kaplan, 2011). MATLAB code for simulations is available from the authors upon request.

We consider IDEAL, a symmetric percentile- $t$  (Studentized) bootstrap using

bootstrapped variance,<sup>16</sup> the method in Horowitz (1998),<sup>17</sup> and the high-order analytic method of Chapter 3 for general quantile tests.

We also include some tests that only work for the median: the exact-bootstrap test of Hutson (2007), two permutation tests, and the built-in MATLAB implementation `ranksum` of the Mann–Whitney–Wilcoxon test (hereafter “MWW”; a.k.a. Wilcoxon rank sum, a.k.a. Mann–Whitney  $U$ ; Mann and Whitney, 1947; Wilcoxon, 1945). For the first permutation test, labeled “Perm (sum)” in tables, we take  $\sum_{i=1}^{n_x} X_i$  as our test statistic, and compare it to the distribution of the same test statistic among “permutations” of the observation labels. For the second, labeled “Perm (median)” in tables, our test statistic is the difference in sample medians. When  $n_x = n_y = 5$ , we calculate all  $\binom{10}{5} = 252$  possibilities. When  $\binom{n_x+n_y}{n_x} > 999$ , e.g. when  $n_x = n_y = 25$ , we sample 999 (unique) combinations from the possible  $X$  samples of size  $n_x$  (with corresponding  $Y$  samples having the remaining  $n_y$  observations). The sum permutation test, median-difference permutation test, and MWW all control size well under exchangeability and have good power against location shift alternatives, but at the expense of severe size distortion under certain violations of exchangeability and lower power against other types of alternatives.

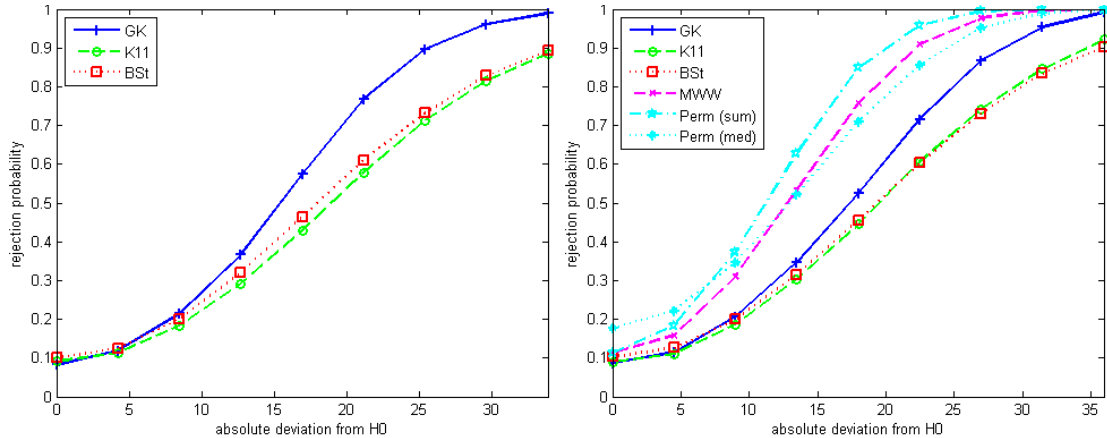
For IDEAL, the two feasible varieties are labeled as “ $\beta$ ” for the test derived from (1.23) and (1.26) and using numeric integration (tolerance  $10^{-5}$  for subintegrals, for MATLAB’s `dblquad`), and “ $\theta^*$ ” for the test derived from (1.36). Both estimate  $\hat{\gamma}$  with quantile spacings and our zero-bias bandwidth unless otherwise noted; results with a kernel estimator and MSE-optimal (Silverman, 1986) bandwidth are extremely similar, as in the empirical example. For the  $\beta$  version, the search for  $\tilde{\alpha}$  is terminated when the calculated resultant type I error is within

---

<sup>16</sup>Many other bootstrap methods are considered in Chapter 3, and this one performed consistently the best. In the simulations therein, using 99 outer (and 100 inner) bootstrap replications was found to be very similar to 999, with size control marginally better with 99 and power marginally better with 999; see Chapter 3 for details. We present whichever gives better results for a given table/graph, and both in some cases like Table 1.6, where the power difference is negligible (0.01, within simulation error) but in favor of 999.

<sup>17</sup>As written, the paper discuss only the median, but it is readily extended to other quantiles. Following Horowitz (1998), we use 100 bootstrap replications and bandwidth  $h = cn^{-3/10}$ . Since no suggestion for  $c$  is given, we use whichever value gives the best performance, searching over  $c \in (0, 500)$ .

0.001 of the desired  $\alpha$ , and the PDF ratio is estimated with MATLAB’s `ksdensity` (Epanechnikov kernel) separately for each sample. Additionally, “ $\beta^{\text{inf}}$ ” is an infeasible version of “ $\beta$ ” using the true  $\gamma$  (`dblquad` tolerance  $10^{-6}$ ), which helps isolate the effect of estimating  $\gamma$ .



**Figure 1.3:** Power curves for inference on the lower quartile (left) and median (right) using data simulated to mimic the period one library task data from Gneezy and List (2006); two-sided,  $\alpha = 0.1$ . The legend entries refer to methods described in the text: “GK” is our new  $\theta^*$  IDEAL; “K11” is Chapter 3; “BSt” is the bootstrap- $t$  (999); “MWW” is MWW; and “Perm (sum)” and “Perm (med)” are the permutation tests.

Building on §1.4, we show simulations where the samples are drawn from distributions estimated from the empirical data (first period of both library and fundraising tasks), with the same sample sizes, same quantiles of interest, and same two-sided  $\alpha = 0.1$  tests. Specifically, we estimate the distributions using the publicly available `kde.m` MATLAB code based on Botev et al. (2010) and then use rejection sampling to simulate data. This can be done in practice to roughly gauge which method is appropriate for a given dataset (e.g., whether the permutation test is likely size distorted or not), though of course the distribution estimation can miss important local variation with finite samples. We ran the  $\theta^*$  version of IDEAL against the three best competitors from our later simulations, which are the bootstrap- $t$ , permutation test (for median), and Chapter 3.



**Table 1.2:** Type I error by quartile, two-sided  $\alpha = 0.10$ , using data simulated to mimic the period one library and fundraising task data from Gneezy and List (2006). Values above 0.105 in bold.

	Library task			Fundraising task		
	Lower	Median	Upper	Lower	Median	Upper
New method ( $\theta^*$ )	0.083	0.090	0.071	0.072	0.080	0.076
Chapter 3	0.091	0.092	0.081	0.054	0.071	0.051
Bootstrap- $t$ (99)	0.102	<b>0.106</b>	0.093	0.092	0.104	0.085
Bootstrap- $t$ (999)	0.100	0.103	0.096	0.089	0.103	0.084
Perm (sum)	n/a	<b>0.116</b>	n/a	n/a	0.095	n/a
Perm (median)	n/a	<b>0.176</b>	n/a	n/a	<b>0.155</b>	n/a
MWW	n/a	<b>0.115</b>	n/a	n/a	0.091	n/a

For the lower and upper quartiles, for both the library and fundraising task setups, all three methods controlled size (see Table 1.2), and IDEAL had the best power; see for example Figure 1.3 (left panel). For the median in the library setup, there was some size distortion in the sum permutation test, and MWW, while the median-difference permutation test was significantly size distorted in both setups (Table 1.2). This size distortion is not surprising since exchangeability is violated. On the flip side, the permutation-type tests (including MWW) are significantly more powerful for both library and fundraising setups; see for example Figure 1.3 (right panel).

We also ran simulations using the settings from Tables 1–4 of Hutson (2007). The same set of uniform random numbers is used for each setting, so the difference between results for the  $N(0, 1)$  and  $\text{Unif}(0, 1)$  columns in the top half of Table 1.3, for example, is due only to the shapes of the distributions, not additional simulation error. Results from Hutson (2007) are reproduced for his method as noted. The permutation test results are from our new simulations; Hutson (2007) seems to have implemented the median-difference permutation test.<sup>18</sup>

<sup>18</sup>Some permutation test results in Hutson (2007) also seem quite improbable—not impossible, given the leeway in implementation of a “permutation test” and simulation error, but less helpful for comparing performance. For instance, the type I error of 0.000 for all five distributions in the top of his Table 1 seems improbable in light of our results of 0.002, 0.000, 0.001, 0.002, 0.001. The 0.99 power in his Table 4 in the case where one distribution is  $\text{Unif}(0, 1)$  and the other  $\text{Unif}(1, 2)$  implies there are instances where the test doesn’t reject even when every observation

With the settings of Tables 1 and 2 from Hutson (2007), most methods control size even with small samples of  $n_x = n_y = 5$ . Table 1.3 corresponds to Hutson's (2007) Table 2; the analog of Table 1 looks qualitatively similar. The only significant exception is Hutson (2007), which for  $\alpha = 0.01$  had up to 0.046 type I error, and for  $\alpha = 0.05$  up to 0.085. We do not include Horowitz (1998) here since the bandwidth could always be (arbitrarily) chosen to control size. Note that the MWW type I error is invariant to (monotonic) transformations of the data since it is based on ranks.

**Table 1.3:** Type I error comparison,  $\alpha = 0.05$ ,  $p = 0.5$ ,  $F_X = F_Y$ . Values 0.06 and higher in bold.

	N(0,1)	Logistic(0,1)	Unif(0,1)	Exp(1)	LogN(0,1)
$n_x = n_y = 5$					
New method ( $\theta^*$ )	0.021	0.021	0.022	0.017	0.016
New method ( $\beta$ )	0.019	0.013	0.021	0.013	0.011
New method ( $\beta^{\text{inf}}$ )	0.038	0.026	0.042	0.017	0.011
Hutson (2007) <sup>a</sup>	0.051	0.046	<b>0.085</b>	0.046	0.034
Chapter 3	0.013	0.010	0.021	0.010	0.006
Bootstrap- $t$ (99)	0.025	0.021	0.037	0.016	0.012
Perm (sum)	0.048	0.049	0.049	0.049	0.049
Perm (median)	0.049	0.049	0.049	0.049	0.049
MWW	0.034	0.034	0.034	0.034	0.034
$n_x = n_y = 25$					
New method ( $\theta^*$ )	0.042	0.042	0.042	0.038	0.037
New method ( $\beta$ )	0.047	0.041	0.040	0.031	0.029
New method ( $\beta^{\text{inf}}$ )	0.050	0.046	0.047	0.042	0.037
Hutson (2007) <sup>a</sup>	0.051	0.048	0.055	0.047	0.035
Chapter 3	0.035	0.033	0.044	0.030	0.024
Bootstrap- $t$ (99)	0.049	0.053	0.052	0.047	0.039
Perm (sum)	0.050	0.049	0.051	0.049	0.049
Perm (median)	0.048	0.049	0.048	0.049	0.048
MWW	0.048	0.048	0.048	0.048	0.048

<sup>a</sup>As reported in Hutson (2007).

Size control for lower and upper quartiles is similarly good. For  $n_x =$  in one sample is larger than every observation in the other sample; we get 1.00 in this situation (Table 1.6).

**Table 1.4:** Type I error comparison when  $F_X(\cdot) \neq F_Y(\cdot)$  (exchangeability is violated),  $\alpha = 0.05$ ,  $n_x = n_y = 25$ ,  $F_X = N(0, 1)$ ,  $F_Y = N(0, \sigma_y^2)$ ,  $p = 0.5$ . The standard deviation  $\sigma_y$  is varied per the column headers, where exchangeability is satisfied only with  $\sigma_y = 1$ . Values 0.06 and higher in bold.

	$\sigma_y$				
	1	2	4	8	16
New method ( $\theta^*$ )	0.044	0.051	0.056	0.053	0.051
New method ( $\beta$ )	0.047	0.052	0.055	0.048	0.046
New method ( $\beta^{\text{inf}}$ )	0.050	0.051	0.050	0.044	0.047
Hutson (2007) <sup>a</sup>	0.051	0.051	0.056	<b>0.067</b>	<b>0.104</b>
Chapter 3	0.035	0.033	0.031	0.030	0.029
Bootstrap- $t$ (99)	0.051	0.056	0.058	0.058	<b>0.061</b>
Bootstrap- $t$ (999)	0.052	0.057	<b>0.060</b>	<b>0.061</b>	<b>0.060</b>
Horowitz (1998) <sup>b</sup>	0.049	<b>0.065</b>	<b>0.067</b>	0.053	0.040
Perm (sum)	0.050	0.054	<b>0.060</b>	<b>0.061</b>	<b>0.062</b>
Perm (median)	0.048	<b>0.104</b>	<b>0.295</b>	<b>0.560</b>	<b>0.756</b>
MWW	0.048	0.059	<b>0.077</b>	<b>0.089</b>	<b>0.099</b>

<sup>a</sup>As reported in Hutson (2007).

<sup>b</sup>Bandwidth  $h = 15n^{-3/10}$ , optimal for  $\sigma_y = 1$ , used for all cases.

$n_y = 25$ , Table 1.3 was run for the new  $\theta^*$  method, Chapter 3 method, and the bootstrap- $t$ . All had type I error below  $\alpha = 0.05$  each time.

Table 1.4 shows the sensitivity of the permutation tests, MWW, and Hutson (2007) to the exchangeability assumption. With  $n_x = n_y = 25$ , the size distortion is relatively mild for the sum permutation test; with  $n_x = n_y = 5$ , its size distortion can be larger: with  $\sigma_x = 1$ ,  $\sigma_y = 2$  yields 0.058 type I error, and  $\sigma_y = 4$  yields 0.078, increasing to around 0.12 as  $\sigma_y$  grows. The median-difference permutation test is extremely sensitive to the type of exchangeability violation in Table 1.4; MWW and Hutson (2007) are in between. Horowitz (1998) does not appear sensitive to exchangeability per se, but the variability of type I error with  $\sigma_y$  reflects the difficulty of picking a bandwidth in practice (an open question). The bandwidth  $h = 15n^{-3/10}$  that was optimal for  $\sigma_y = 1$  is clearly not optimal for other values of  $\sigma_y$ . None of the IDEAL varieties is sensitive to violations of exchangeability.

**Table 1.5:** Type I error comparison when  $F_X(\cdot) \neq F_Y(\cdot)$  (exchangeability is violated),  $\alpha = 0.05$ ,  $n_x = n_y = n \in \{5, 25, 55\}$ ,  $p = 0.5$ . The differently-shaped distributions  $F_X$  and  $F_Y$  are given in the column headers; they are shifted to have identical medians. Values 0.060 and higher in bold.

	Exp(1), Unif(0.193, 1.193)			$\beta(4, 1), \beta(1, 4) + 0.682$		
	$n = 5$	$n = 25$	$n = 55$	$n = 5$	$n = 25$	$n = 55$
New method ( $\theta^*$ )	0.031	0.046	0.049	0.032	0.043	0.049
Chapter 3	0.021	0.035	0.042	0.019	0.037	0.040
Bootstrap- $t$ (99)	0.031	0.053	0.057	0.028	0.050	0.049
Perm (sum)	<b>0.068</b>	<b>0.271</b>	<b>0.609</b>	<b>0.064</b>	<b>0.385</b>	<b>0.757</b>
Perm (median)	<b>0.061</b>	<b>0.121</b>	<b>0.110</b>	<b>0.072</b>	<b>0.062</b>	<b>0.056</b>
MWW	0.045	<b>0.069</b>	<b>0.077</b>	0.044	<b>0.164</b>	<b>0.319</b>

Exchangeability is also violated if the two distributions are different shapes. This type of violation can also cause significant size distortion in the permutation tests and MWW, as Table 1.5 shows. This is not just an issue of getting a large enough sample: the size distortion gets worse with larger sample sizes in both cases for sum permutation and MWW. The magnitudes are serious, too. Only out of the two distributional setups and three sample sizes in Table 1.5, the type I error is up to 0.757 for the sum permutation test and 0.319 for MWW; the 0.121 of the median-difference permutation test is outdone by the 0.756 already seen in Table 1.4. IDEAL is robust to any of these types of exchangeability violation.

Table 1.6 shows results from the simulation setup of Table 4 in Hutson (2007). For each of these five different distributions, the permutation tests and MWW have the best power, followed by IDEAL, followed by the rest. Of the tests applicable at any quantile, IDEAL has the best power. For the special case of the median, the permutation tests and MWW have better power against location shifts for these distributions, but they are all subject to extreme size distortion when exchangeability is violated, and they have lower power against other types of alternatives. With the fatter-tailed Cauchy distribution, IDEAL has better power than MWW and the sum permutation test, and a slight edge over the median-

**Table 1.6:** Power comparison,  $\alpha = 0.05$ ,  $p = 0.5$ , location difference of one unit between medians of  $X$  and  $Y$ ,  $n_x = n_y = 25$ ,  $F_X = F_Y$ .

	N(0,1)	Logistic(0,1)	Unif(0,1)	Exp(1)	LogN(0,1)
New method ( $\theta^*$ )	0.79	0.39	1.00	0.91	0.75
New method ( $\beta$ )	0.80	0.40	1.00	0.90	0.71
New method ( $\beta^{\text{inf}}$ )	0.81	0.41	1.00	0.91	0.73
Hutson (2007) <sup>a</sup>	0.75	0.38	0.99	0.87	0.57
Chapter 3	0.70	0.31	1.00	0.86	0.64
Bootstrap- $t$ (99)	0.70	0.35	1.00	0.84	0.68
Bootstrap- $t$ (999)	0.71	0.36	1.00	0.86	0.68
Horowitz (1998) <sup>b</sup>	0.62	0.27	1.00	0.87	0.66
Perm (sum)	0.93	0.48	1.00	0.92	0.53
Perm (median)	0.85	0.45	1.00	0.97	0.86
MWW	0.92	0.50	1.00	0.99	0.91

<sup>a</sup>As reported in Hutson (2007).

<sup>b</sup>Optimal bandwidths (maximum power subject to size control, which was only a binding constraint in the  $N(0,1)$  case) were found to be  $c = n^{-3/10}$  with  $c = 15, 20, 20, 10, 20$  for the five distributions, respectively.

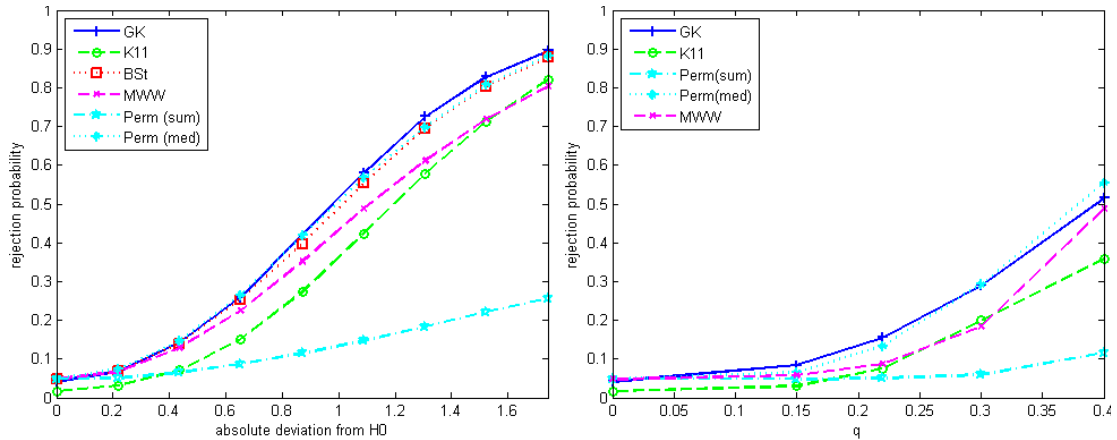
difference permutation test; see Figure 1.4 (left). The location shift setup from Hutson (2007) is advantageous to the permutation tests and MWW, which should have relatively lower power when  $F_Y$  deviates more locally (not globally) from  $F_X$  or when the treatment affects the shape and/or variance of the distribution in certain ways; more on this below.

Table 1.6 shows that IDEAL has better power than Hutson (2007), in addition to the greater robustness already shown. Our method also consistently has better power than the other methods considered, even with hand-picking the optimal bandwidth for Horowitz (1998) in each case.

For lower and upper quartiles, IDEAL consistently has the best power. In simulations not shown, compared to the bootstrap- $t$  and Chapter 3, our  $\theta^*$  method was the most powerful (or tied, for LogN(0,1) upper quartile) over a range of alternatives for all five distributions.

For other types of alternatives, IDEAL is more powerful than even the permutation tests and MWW. The pure location-shift alternatives considered in Table

1.6 are precisely what the other tests are designed to detect. They use information from all parts of the distribution, which is good for detecting pure location shifts, but bad for size control when exchangeability is violated, and also bad for power against different types of alternatives. For example, using the distributions  $\beta(1, 4)$  and  $\beta(4, 1)$  from Table 1.5, the same mechanism that causes size distortion (when the distributions are shifted to have the same median) also causes low power (when they are shifted to differ by 0.1); see Table 1.7.



**Figure 1.4:** Left: power curves for two-sided median inference,  $\alpha = 0.05$ . Both distributions are Cauchy, with magnitude of location shift (of treatment distribution) shown on  $x$ -axis. Samples follow  $X_i \stackrel{iid}{\sim} \text{Cauchy}$ ,  $Y_i \stackrel{iid}{\sim} \delta + \text{Cauchy}$ , where the magnitude of the location shift  $|\delta|$  is shown on the  $x$ -axis. Right: control distribution is Cauchy, and for parameter  $q$  shown on  $x$ -axis, treatment adds  $\text{Unif}(0, 2q)$  to individuals initially between quantiles  $0.5 - q$  and  $0.5 + q$ . Both: legend entries refer to methods described in the text: “GK” is our new  $\theta^*$  IDEAL; “K11” is Chapter 3; “BSt” is the bootstrap- $t$ ; “MWW” is MWW; and “Perm (sum)” and “Perm (med)” are the permutation tests.

The permutation tests and MWW also have lower power against treatment effects that only shift the middle of the distribution without affecting the tails. For example, consider a population that initially has a  $\beta(2, 3)$  distribution, and a treatment that affects (with some probability) individuals initially located between the 30th and 60th percentiles, adding a random value between zero and 0.3. Or,

**Table 1.7:** Power comparison when treatment affects distribution shape or is (relatively) local to median,  $\alpha = 0.05$ ,  $n_x = n_y = 25$ ,  $p = 0.5$ . Two highest values per column are in bold.

	Skew <sup>a</sup>	Uniform/local <sup>b</sup>	Beta/local <sup>c</sup>	Cauchy/local <sup>d</sup>
New method ( $\theta^*$ )	0.30	<b>0.21</b>	<b>0.21</b>	<b>0.16</b>
Chapter 3	0.30	0.17	0.19	0.09
Bootstrap, perc- $t$	<b>0.31</b>	<b>0.20</b>	<b>0.22</b>	<b>0.17</b>
Perm (sum)	0.08	0.11	0.13	0.05
Perm (median)	<b>0.31</b>	0.17	0.18	0.15
MWW	0.17	0.11	0.14	0.10

<sup>a</sup> $\beta(1, 4)$  control and  $\beta(4, 1) - 0.582$  treatment distributions (medians differ by 0.1).

<sup>b</sup>Unif(0, 1) control, and treatment adds 0.2 to individuals initially between 0.3 and 0.6.

<sup>c</sup> $\beta(2, 3)$  control, and treatment adds random amount  $0.3\beta(4, 1)$  to individuals initially at quantile  $q$  if  $0.3 \leq q \leq 0.6$  with probability  $(1 - e^{-(q-0.3)^{20}})(1 - e^{-(q-0.6)^{20}})$ .

<sup>d</sup>Cauchy control, and treatment adds random amount Unif(0, 1) to individuals initially at quantile  $q$  if  $0.25 \leq q \leq 0.75$  with probability  $(1 - e^{-(q-0.25)^{50}})(1 - e^{-(q-0.75)^{50}})$ .

consider a standard uniform population, where the treatment simply adds 0.2 to individuals between the 30th and 60th percentiles. In both cases, IDEAL has better power than the permutation tests and MWW; see Table 1.7. Interestingly, in these situations, the bootstrap- $t$  and IDEAL have roughly the same power. We can also see how the power advantage depends on how “local” the treatment effect is (to the median of the distribution). If the control group follows a Cauchy distribution, one hypothetical treatment parameterized by  $q$  affects only individuals initially in quantile between  $0.5 - q$  and  $0.5 + q$ , adding random increment Unif(0,  $2q$ ). Plotting power against  $q$ , as in Figure 1.4 (right), IDEAL has better power for lower  $q$ , but this advantage disappears when  $q$  gets closer to 0.50 (whole distribution affected).

The two permutation tests and MWW are commonly used in practice; we now discuss their use specifically for median treatment effect inference.<sup>19</sup> In their favor, they control size under exchangeability, and they usually have better power than IDEAL against pure location-shift alternatives (Tables 1.3 and 1.6). How-

<sup>19</sup>A new test for first order stochastic dominance based on the fractional order statistic theory will be discussed in a separate paper.

ever, they are subject to size distortion when exchangeability is violated. Both differing variances (Table 1.4) and differently shaped distributions (Table 1.5) can cause grave size distortion, even in large samples. The permutation tests and MWW are also less powerful against some types of alternatives that are not pure location shifts of the entire distribution. Consequently, we recommend IDEAL for inference on quantile treatment effects, including the median effects. However, it would be quite valuable to explore the method of (pre-)estimating distributions and simulating type I error. If such an approach is shown to be reliable, we might suggest a permutation test or MWW for median inference whenever valid. Even then, IDEAL can be applied to any quantile, so it provides additional information about treatment effects at different parts of the distribution, which a permutation test does not.

## 1.6 Conclusion

We derive an  $O(n^{-1})$  rate of uniform convergence of the distributions of the linearly interpolated and ‘ideal’ types of fractional order statistics. This tightly links theoretical results based on the latter with feasible procedures based on the former. As one example, we provide the first direct theoretical justification for one-sample quantile confidence intervals based on fractional order statistics, where the coverage probability error is  $O(n^{-1})$ , and provide analytic calibration to further reduce this to almost  $O(n^{-3/2})$ .

We also present a new method for two-sample quantile treatment effect inference, which has  $O(n^{-2/3})$  two-sided size distortion and  $O(n^{-1/2})$  one-sided. In simulations, IDEAL has better size control and power than existing bootstrap and analytic methods in most situations. Median-specific permutation tests and the Mann–Whitney–Wilcoxon test can have better power in some cases when exchangeability holds, but they are subject to severe size distortion when the exchangeability assumption is violated, and they can have worse power against alternatives that are not pure location shifts. Our method is robust to a variety of situations: different sample sizes, different distributions for the two samples (shape as well as



scale), small samples, and a range of quantiles. Unless exchangeability is certain and only the median is of interest, the IDEAL two-sample quantile treatment effect inference appears to have the strongest combination of size control, power, and robustness.

Our new theory also applies to inference on general linear combinations of quantiles, such as the interquartile range. An extension to test for first-order stochastic dominance is being developed. Extension to inference on the entire quantile process would also be of interest. Chapter 2 extends our unconditional IDEAL results to a nonparametric conditional quantile model.

*The preceding chapter was coauthored with Matt Goldman.*

## Chapter 2

# IDEAL inference on conditional quantiles

### Abstract

We examine inference on conditional quantiles from the nonparametric perspective of local smoothing. This paper develops a framework for translating the powerful, high-order accurate IDEAL results (Chapter 1) from their original unconditional context into a conditional context, via a uniform kernel. Under mild smoothness assumptions, our new conditional IDEAL method's two-sided pointwise coverage probability error is  $O(n^{-2/(2+d)})$ , where  $d$  is the dimension of the conditioning vector and  $n$  is the total sample size. For  $d \leq 2$ , this is better than the conventional inference based on asymptotic normality or a standard bootstrap. It is also better for other  $d$  depending on smoothness assumptions. For example, conditional IDEAL is more accurate for  $d = 3$  unless 11 or more derivatives of the unknown function exist and a corresponding local polynomial of degree 11 is used (which has 364 terms since interactions are required). Even as  $d \rightarrow \infty$ , conditional IDEAL is more accurate unless the number of derivatives is at least four, and the number of terms in the corresponding local polynomial goes to infinity as  $d \rightarrow \infty$ . The tradeoff between the effective (local) sample size and bias determines the optimal bandwidth rate, and we propose a feasible plug-in bandwidth. Simulations

show that IDEAL is more accurate than popular current methods, significantly reducing size distortion in some cases while substantially increasing power (while still controlling size) in others. Computationally, our new method runs much more quickly than existing methods for medium and large datasets (roughly  $n \geq 1000$ ). We also examine health outcomes in Indonesia for an empirical example.

## 2.1 Introduction

Ideally, we would like to know the full joint distribution of every variable we care about. Practically, with a finite amount of data, we can learn a lot from estimating quantiles of conditional distributions. To gain knowledge from the data, rather than simply compute numbers, we need statistical inference on these conditional quantiles, which is this paper's concern.

In economics, conditional quantiles have appeared across diverse topics because they are such fundamental statistical objects. Conditional quantile studies of wages have looked at experience (Hogg, 1975), union membership (Chamberlain, 1994), and inequality in the U.S. wage structure (Angrist et al., 2006; Buchinsky, 1994), while Kordas (2006) examines married women's propensity to work. Examples from health economics include models of infant birthweight (Abrevaya, 2001) and demand for alcohol (Manning et al., 1995). Among others, Chesher (2003) employs conditional quantiles for identification in nonseparable models. Guerre and Sabbah (2012) give an example of estimating conditional quantiles of private values from bid data in a first-price sealed bids auction. Other empirical examples involving conditional quantiles include school quality and student outcomes (Eide and Showalter, 1998), spatial land valuations (Koenker and Mizera, 2004), welfare analysis (Belluzzo, 2004), and Engel curves (Nayyar, 2009). For similar reasons as in economics, conditional quantiles enrich empirical work in other areas, such as modeling temperature (Hyndman et al., 1996), limiting factors in ecology (Cade et al., 1999), and terrestrial mammal running speeds (Koenker, 2005), among other examples. Additionally, almost any study of conditional means could be extended to conditional quantiles to expose additional heterogeneity.

If our continuous dependent variable of interest is  $Y$  (e.g., high school GPA), and we want to know quantiles of its distribution conditional on vector  $X$  having value  $x_0$  (e.g., a particular value of family income and other socioeconomic and demographic characteristics), we would like an infinite number of observations with  $X = x_0$ . With our finite sample, if  $X$  contains even one continuous component, we have zero probability of even one observation with  $X = x_0$ . One approach is to parameterize the conditional  $p$ -quantile function as  $Q_{Y|X}(p) = X'\beta$ , linear in  $X$ . This strong linearity assumption leads to an estimator where observations with very different  $X$  can influence the conditional quantile at  $X = x_0$ . If the true function is not linear in  $X$ , this misspecification can lead to a poor estimator for any given  $x_0$ . A second approach is to use a more flexible parameterization, which could be set in a sieve-type nonparametric framework. A third approach is nonparametric local kernel smoothing, using only observations with  $X$  close in value to the target  $x_0$ . Here, we develop inference via kernel smoothing.

Our strategy is to apply an accurate method for unconditional quantile inference to the observations with  $X$  close to the target  $x_0$ . Instead of relying on an asymptotic normal approximation, this method directly approximates the exact finite sample distribution using fractional order statistic theory. As usual for kernel smoothing, taking  $X$  not quite equal to  $x_0$  causes some bias, which increases as we include  $X$  values farther from  $x_0$  to include more observations in our effective sample. Counter to this, the method's accuracy (for the biased value) improves as the effective sample size grows. This tradeoff determines the optimal bandwidth that minimizes overall coverage probability error. After more precisely deriving the objects involved, and estimating the unknown ones, we propose a bandwidth for use in practice. Specifically, we apply some of the IDEAL (interpolated dual of exact analytic  $L$ -statistic; see §2.2) results summarizing Chapter 1 on the Hutson (1999) method based on fractional order statistics. Joint confidence sets are constructed as Cartesian products of (Bonferroni-adjusted) confidence intervals over many  $x_0$ .

This strategy has advantages in theory and in practice. Theoretically, the coverage probability error is of a smaller order of magnitude than that for inference based on asymptotic normality or bootstrap in the most common cases. This

reflects the same advantage of the fractional order statistic method for unconditional quantile inference. The only theoretical limitation is our implicit use of a uniform kernel, which is only a second-order kernel. This prevents reducing the coverage probability error by assuming higher degrees of smoothness than we do here, though to maintain robustness we would not want to assume more smoothness of unknown functions anyway. As it is, our method has a better rate than asymptotic normality even with infinite smoothness assumed if the conditioning vector  $X$  contains one or two continuous components, and similarly for any number of continuous conditioning variables in  $X$  if no more than four derivatives of the unknown function are (correctly) assumed to exist. It is also an advantage that our coverage error either leads to over-coverage or can be set to zero (at the dominant order of magnitude) by our choice of bandwidth since we derive the value (not just rate) of the optimal bandwidth when  $X$  contains only one continuous conditioning variable.

Practically, direct estimation of a plug-in version of the optimal bandwidth is straightforward, with details in §2.4.2 and §2.5.1 and code available in R for  $X$  with a single continuous component. An extension to additional continuous conditioning variables will require additional calculations and estimation, but the approach will be identical. In either case, there is no asymptotic variance or other nuisance parameters to estimate. Also, coverage probability is monotonically decreasing in the size of the bandwidth, which is nice for transparency and may be helpful for future refinements. Another practical advantage is computation times orders of magnitude smaller than those for existing methods on medium and large datasets (roughly  $n \geq 1000$ ); see §2.5 for details.

Past research has focused instead on inference through asymptotic normal approximation or bootstrap. First-order accuracy has been shown for the asymptotic normality approaches in Bhattacharya and Gangopadhyay (1990, Thm. N2) and Hall et al. (1999, Thm. 1, eqns. 7 and 8) and for the bootstrap in Gangopadhyay and Sen (1990). Higher-order accuracy can be shown for inference using the asymptotic normality in Chaudhuri (1991); improved recently by Portnoy (2012), this result is compared to ours in detail following Theorem 2.3.

The remainder of this paper is organized as follows. Prior unconditional IDEAL work and intuition is given in Section 2.2. Section 2.3 details our setup and gives results for the bias. Section 2.4 describes the optimal bandwidth and its feasible plug-in counterpart. Section 2.5 contains a simulation study, while Section 2.6 contains empirical applications. Notationally,  $\doteq$  should be read as “is equal to, up to smaller-order terms”;  $\asymp$  as “has exact (asymptotic) rate/order of” (same as “big theta” Bachmann–Landau notation,  $\Theta(\cdot)$ ); and  $A_n = O(B_n)$  as usual,  $\exists k < \infty$  s.t.  $|A_n| \leq B_n k$  for sufficiently large  $n$ . Acronyms used are those for cumulative distribution function (CDF), confidence interval (CI), coverage probability (CP), coverage probability error (CPE), interpolated duals of exact analytic  $L$ -statistics (IDEAL), and probability density function (PDF). Proofs are reserved for the appendix.

## 2.2 Fractional order statistic theory

Fractional order statistic theory is key to the method developed in this paper. This section contains an overview of the relevant theory. For a more comprehensive and general development of this theory, see Chapter 1, which also contains details of additional IDEAL (Interpolated Duals of Exact Analytic  $L$ -statistics) methods of quantile inference built upon the theory.

The approach is to construct confidence intervals (CIs) from order statistics observed in the sample. Consider constructing a two-sided CI for the  $p$ -quantile. By definition, the  $p$ -quantile is between the  $u_\ell$ -quantile and the  $u_h$ -quantile when  $u_\ell < p < u_h$ . One way to construct a CI is to use the empirical  $u_\ell$ - and  $u_h$ -quantiles as endpoints. If  $(n+1)u_\ell$  and  $(n+1)u_h$  are integers, then both endpoints are order statistics, and we can calculate the exact, finite-sample coverage probability because the joint distribution of order statistics is known. The only obstacle is that there almost surely (i.e., with probability one) does not exist a pair of order statistics that yields the exact coverage we desire. Thus, either randomization or interpolation between order statistics is needed; we pursue the latter approach.

In the unconditional case, there is an iid sample  $\{Y_i\}_{i=1}^n$  of draws of an abso-

lutely continuous, scalar random variable  $Y$  with unknown cumulative distribution function (CDF) denoted  $F_Y(\cdot)$ . By definition, the quantile function is the inverse CDF, so we also write  $Q_Y(\cdot) \equiv F_Y^{-1}(\cdot)$ . The  $k$ th order statistic  $Y_{n:k}$  denotes the  $k$ th smallest value out of the  $n$  observations  $\{Y_i\}_{i=1}^n$ . Since  $Y < Q_Y(p)$  is equivalent to  $F_Y(Y) < p$ , we can work with the uniformly distributed  $U_i \equiv F_Y(Y_i)$  and the corresponding order statistics. For any  $u \in (0, 1)$  such that  $(n+1)u$  is an integer, it is well known (and derived via combinatorics) that the uniform order statistics follow a beta distribution,

$$U_{n:(n+1)u} \sim \beta((n+1)u, (n+1)(1-u)).$$

These  $U_{n:(n+1)u}$  are estimators on  $[0, 1]$  of true quantiles  $Q_U(u) = u$ , where  $Q_U(\cdot)$  is the quantile function of the uniformly distributed  $U_i$ . As such, they may also be written as  $\hat{Q}_U^I(u)$ , where the ‘I’ superscript is for ‘ideal’ since the distribution is known exactly. This can be generalized beyond integer  $(n+1)u$  to any  $u \in (0, 1)$ , and the order statistics generalize to corresponding ‘ideal’ fractional order ‘statistics’<sup>1</sup>

$$\hat{Q}_U^I(u) = U_{n:(n+1)u} \sim \beta((n+1)u, (n+1)(1-u)), \quad (2.1)$$

the same distribution as before.

For a two-sided equal-tailed CI for the median, we can solve for  $u = u_h$  for the upper (high) endpoint with  $P\left(\hat{Q}_U^I(u) < 1/2\right) = \alpha/2$  since we know the exact distribution of  $\hat{Q}_U^I(u)$  for all  $u$ . We may solve for  $u = u_\ell$  for the lower endpoint similarly. Since  $P\left(\hat{Q}_U^I(u) < 1/2\right) = P\left(Q_Y\left(\hat{Q}_U^I(u)\right) < Q_Y(1/2)\right)$ , an exact  $(1-\alpha)$  CI for the median  $Q_Y(1/2)$  is defined by the unobserved, *fractional* order statistic endpoints  $\hat{Q}_Y^I(u_h)$  and  $\hat{Q}_Y^I(u_\ell)$ , where

$$\hat{Q}_Y^I(u) \equiv Q_Y\left(\hat{Q}_U^I(u)\right).$$

For quantiles besides the median, the endpoint indices  $u_\ell$  and  $u_h$  are implicit

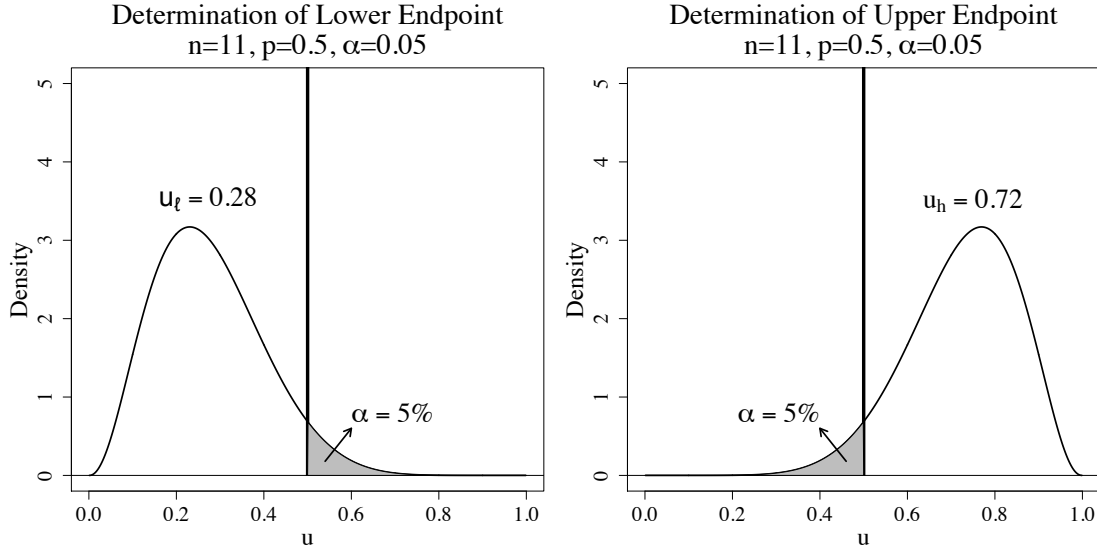
---

<sup>1</sup>Technically, when  $(n+1)u$  is not an integer,  $U_{n:(n+1)u}$  is not a statistic since it is not a function of the observed sample. Instead,  $U_{n:(n+1)u}$  is a theoretical construct when it is not equal to an observed order statistic. Nonetheless, we follow the convention in the literature and call  $U_{n:(n+1)u}$  a fractional order statistic for all  $u$ .

functions of  $p$  and  $\alpha$ , strictly monotonic in both  $p$  and  $\alpha$ , determined by

$$\alpha/2 = P\left(\hat{Q}_U^I(u_h) < p\right), \quad \alpha/2 = P\left(\hat{Q}_U^I(u_\ell) > p\right), \quad (2.2)$$

with  $\hat{Q}_U^I(u) \sim \beta((n+1)u, (n+1)(1-u))$  as in (2.1). For one-sided CIs, only one of the two equalities in (2.2) is used, and with  $\alpha$  instead of  $\alpha/2$ . Figure 2.1 is an example.



**Figure 2.1:** Example of selection of  $u_\ell$  and  $u_h$  using (2.2), for one-sided CI endpoints. For two-sided,  $\alpha/2$  would be in place of  $\alpha$ . Note that for  $p = 0.5$ ,  $u_h = 1 - u_\ell$ .

The unobserved, ‘ideal’ fractional order statistic endpoints of the exact CI can be approximated by linear (superscript ‘L’) interpolation between consecutive observed order statistics. For example, if  $n = 8$  and  $u = 1/2$ , to approximate the 4.5th order statistic, we average the 4th and 5th:  $\hat{Q}_Y^L(1/2) = (1/2)Y_{8:5} + (1/2)Y_{8:4}$ . For any  $u \in (0, 1)$ ,

$$\hat{Q}_Y^L(u) \equiv (1 - \epsilon)Y_{n: \lfloor (n+1)u \rfloor} + \epsilon Y_{n: \lfloor (n+1)u \rfloor + 1}, \quad (2.3)$$

where  $\epsilon \equiv (n+1)u - \lfloor (n+1)u \rfloor$  is the interpolation weight, with  $\lfloor \cdot \rfloor$  denoting the floor function. The construction of a CI with endpoints  $\hat{Q}_Y^L(u_h)$  and  $\hat{Q}_Y^L(u_\ell)$  is first



given in Hutson (1999). Theoretical justification in terms of coverage probability is first given in Chapter 1, where the linear interpolation is shown to induce only  $O(n^{-1})$  coverage probability error (CPE), as reproduced in the following lemma.

**Lemma 2.1.** *For quantile of interest  $p \in (0, 1)$  and iid sample  $\{Y_i\}_{i=1}^n$ , consider the two-sided CI constructed as  $(\hat{Q}_Y^L(u_\ell), \hat{Q}_Y^L(u_h))$  with  $\hat{Q}_Y^L(u)$  as defined in (2.3), where  $u_\ell$  and  $u_h$  are defined by (2.2). Alternatively, consider a lower or upper one-sided CI, with  $u_h$  or  $u_\ell$  determined by the relevant equality in (2.2) but with  $\alpha$  in place of  $\alpha/2$ .*

*Assume that  $F_Y'(Q_Y(p)) > 0$  and that the probability density function  $F_Y'(\cdot)$  is twice continuously differentiable in a neighborhood of  $Q_Y(p)$ . Then the coverage probability of the lower one-sided CI is*

$$P\left(\hat{Q}_Y^L(u_h) > Q_Y(p)\right) = 1 - \alpha + \frac{\epsilon_h(1 - \epsilon_h)z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{p(1-p)\sqrt{2\pi}}n^{-1} + O(n^{\delta-3/2})$$

*for arbitrarily small  $\delta > 0$ ,  $\epsilon_h \equiv (n+1)u_h - \lfloor (n+1)u_h \rfloor$  similar to before, and  $z_{1-\alpha}$  the  $(1-\alpha)$ -quantile of a standard normal distribution. The coverage probability of the upper one-sided CI is the same but with  $\epsilon_\ell$  instead of  $\epsilon_h$ . Similarly, the coverage probability of the two-sided CI is*

$$\begin{aligned} &P\left(\hat{Q}_Y^L(u_\ell) < Q_Y(p) < \hat{Q}_Y^L(u_h)\right) \\ &= 1 - \alpha + \frac{z_{1-\alpha} \exp\{-z_{1-\alpha}^2/2\}}{p(1-p)\sqrt{2\pi}}[\epsilon_h(1 - \epsilon_h) + \epsilon_\ell(1 - \epsilon_\ell)]n^{-1} + O(n^{\delta-3/2}). \end{aligned}$$

Beyond the asymptotic rate of CPE, this method also has the advantage of approximating the exact finite-sample distribution directly rather than relying on asymptotic normality. To determine the optimal bandwidth in §2.4, Lemma 2.1 will be used along with results for CPE from the bias in §2.3.

The asymptotic (first-order) power of the IDEAL hypothesis test is equal to that of a test based on the asymptotic normality of the  $p$ -quantile estimator, against alternatives of order  $n^{-1/2}$ . The upper endpoint of a one-sided  $(1-\alpha)$  CI is  $\hat{Q}_Y^L(u_h)$  for IDEAL and can be written as  $\hat{Q}_Y^L(p) + n^{-1/2}z_{1-\alpha}\sqrt{p(1-p)}/\hat{f}_Y\left(\hat{Q}_Y^L(p)\right)$  for normality. The quantile spacing estimator

$$\frac{1}{h}\left(\hat{Q}_Y^L(p+h) - \hat{Q}_Y^L(p)\right),$$

similar to Siddiqui (1960) and Bloch and Gastwirth (1968), consistently estimates the  $1/f_Y$  object as long as  $h \rightarrow 0$  and  $nh \rightarrow \infty$ . Lemma 1.4 in Chapter 1 gives  $u_h - p = n^{-1/2} z_{1-\alpha} \sqrt{p(1-p)} + O(n^{-1})$ , so bandwidth  $h = u_h - p$  satisfies the rate conditions for consistency. Then, to show the equivalence of the IDEAL endpoint with the normality endpoint,

$$\begin{aligned} \hat{Q}_Y^L(u_h) &= \hat{Q}_Y^L(p) + \left[ \hat{Q}_Y^L(p + u_h - p) - \hat{Q}_Y^L(p) \right] \\ &= \hat{Q}_Y^L(p) + \frac{u_h - p}{u_h - p} \left[ \hat{Q}_Y^L(p + u_h - p) - \hat{Q}_Y^L(p) \right] \\ &= \hat{Q}_Y^L(p) + n^{-1/2} z_{1-\alpha} \sqrt{p(1-p)} \frac{1}{u_h - p} \left[ \hat{Q}_Y^L(p + u_h - p) - \hat{Q}_Y^L(p) \right] \\ &\quad + O(n^{-1}) \\ &= \hat{Q}_Y^L(p) + n^{-1/2} z_{1-\alpha} \sqrt{p(1-p)} / \hat{f}_Y \left( \hat{Q}_Y^L(p) \right) + O(n^{-1}). \end{aligned}$$

Asymptotically, against alternatives of order  $n^{-1/2}$ , the  $O(n^{-1})$  difference in the final line above is negligible, so the endpoints and thus power are the same. This shows that IDEAL attains better coverage probability without sacrificing optimal asymptotic power. It is reasonable to expect that this is also true of IDEAL methods in other settings.

Additional unconditional IDEAL methods are developed in Chapter 1. Beyond inference on a single quantile, joint inference on multiple quantiles is also possible, with  $O(n^{-1})$  CPE. IDEAL inference on linear combinations of quantiles, such as the interquartile range, has  $O(n^{-2/3})$  CPE. In a two-sample setup (e.g., treatment and control groups), IDEAL inference on quantile treatment effects has  $O(n^{-2/3})$  CPE. For the IDEAL quantile treatment effect inference, in addition to an empirical application revealing interesting heterogeneity in the “gift exchange” examined in Gneezy and List (2006), many simulation results are given in Chapter 1. Specific to the median, and switching to the language of hypothesis testing, permutation-type tests (including the Mann–Whitney–Wilcoxon test) are often used because of their good power against pure location shifts, but they can be severely size distorted when exchangeability is violated. In contrast, IDEAL controls size in these cases, and it also has better power against certain alternatives that are not pure location shifts. IDEAL’s robust size control extends to all quan-

tiles, and power is consistently better than that of other methods for quantile treatment effects, including various bootstrap methods and the approach in Chapter 3. This strong performance of IDEAL in unconditional settings motivates this paper’s development of a framework to extend these methods to a nonparametric conditional setting, where even fewer alternative methods exist.

## 2.3 Setup

Let  $Q_{Y|X}(u; x)$  be the conditional quantile function of some scalar outcome  $Y$  given conditioning vector  $X \in \mathcal{X} \subset \mathbb{R}^d$ , evaluated at  $X = x$  and quantile  $u \in (0, 1)$ . A sample of iid data  $\{Y_i, X_i\}_{i=1}^n$  is drawn. If the conditional cumulative distribution function (CDF)  $F_{Y|X}(\cdot)$  is strictly increasing and continuous at  $u$ , then  $F_{Y|X}(Q_{Y|X}(u; x); x) = u$ . For some value  $X = x_0$  and some quantile  $p \in (0, 1)$ , interest is in inference on  $Q_{Y|X}(p; x_0)$ . Without loss of generality, let  $x_0 = 0$ , which can always be achieved in practice by translation of the  $X_i$ .

If  $X$  is discrete, we can take all the observations with  $X_i = 0$  and then compute a confidence interval (CI) for the  $p$ -quantile of the corresponding  $Y_i$  values. The one-sample quantile inference method of Hutson (1999) gives CIs with coverage probability error (CPE) of order  $O(n^{-1})$ , as proved in Chapter 1. The term of order  $n^{-1}$  is explicitly derived in Chapter 1, too, which makes possible analytic calibration to reduce CPE to  $O(n^{-3/2+\delta})$  for arbitrarily small  $\delta > 0$ . In the conditional setting, however, it is more practical not to calibrate, since the  $n^{-1}$  term always leads to over-coverage (as previously shown) and will be used to calculate an optimal bandwidth in §2.4. If there are  $N_n$  observations with  $X_i = 0$ , then the CPE is  $O(N_n^{-1})$  (uncalibrated), where  $N_n$  is the effective sample size. For joint CIs, since the pointwise CIs are all independent (using different  $X$ ), we can generate  $(1 - \alpha)$  joint coverage by pointwise  $(1 - \alpha)^{1/m}$  CIs, where  $m$  is the number of values in the support of  $X$ . Since  $X$  is discrete, these are also uniform confidence “bands.” The key to the one-sample results going through (with  $N_n$ ) is that we have  $N_n$  iid draws of  $Y_i$  from the same  $Q_{Y|X}(\cdot; 0)$  conditional quantile function, which is the one of interest.

As always, if  $X$  is continuous (or  $N_n$  too small), we need to add observations with  $X_i$  near zero. Specifically, we include any observations with  $X_i \in C_h$ , where  $C_h$  is the interval  $[-h, h]$  for some bandwidth  $h$  when  $d = 1$ . For  $d > 1$ ,  $C_h$  is a hypersphere or hypercube with radius (or half side length)  $h$ , centered at the origin. Important objects are defined here for reference, as well as our concept of smoothness that, like our use of  $C_h$ , follows that of Chaudhuri (1991, pp. 762–3). Recall that  $x_0 = 0$  is assumed without loss of generality.

**Definition 2.1** (effective sample). Let  $h$  denote the bandwidth and  $p \in (0, 1)$  the quantile of interest. The effective sample consists of  $Y_i$  values from observations with  $X_i$  inside some window  $C_h \subset \mathbb{R}^d$  centered at the origin, and the effective sample size is  $N_n$ :

$$C_h \equiv \{x : x \in \mathbb{R}^d, \|x\| \leq h\}, \quad (2.4)$$

$$N_n \equiv \#\{\{Y_i : X_i \in C_h, 1 \leq i \leq n\}\}, \quad (2.5)$$

where  $\|\cdot\|$  denotes any norm on  $\mathbb{R}^d$ . With  $X = (X_{(1)}, \dots, X_{(d)}) \in \mathbb{R}^d$ ,  $C_h$  becomes a  $d$ -dimensional hypersphere if  $\|X\| = \|X\|_2 \equiv \sqrt{X_{(1)}^2 + \dots + X_{(d)}^2}$ , the Euclidean norm ( $L_2$  norm), and  $C_h$  becomes a  $d$ -dimensional hypercube if  $\|X\| = \|X\|_\infty \equiv \max\{|X_{(1)}|, \dots, |X_{(d)}|\}$ , the max-norm ( $L_\infty$  norm). Additionally, the  $p$ -quantile of  $Y$  when  $X$  is restricted to  $C_h$  is denoted  $Q_{Y|C_h}(p)$ , which satisfies

$$p = P(Y < Q_{Y|C_h}(p) \mid X \in C_h).$$

**Definition 2.2** (local smoothness, differentiation). For  $d$ -dimensional vector  $v = (v_{(1)}, \dots, v_{(d)})$  of nonnegative integers, let  $D^v$  denote the differential operator

$$\partial^{\|v\|_1} / [\partial x_{(1)}^{v_{(1)}} \cdots \partial x_{(d)}^{v_{(d)}}],$$

where  $\|v\|_1 = v_{(1)} + \dots + v_{(d)}$  is the  $L_1$  norm. A function  $h(x)$  is said to have local smoothness of degree  $s = k + \gamma$ , where  $k$  is a nonnegative integer and  $\gamma \in (0, 1]$ , if  $h(x)$  is continuously differentiable through order  $k$  in a neighborhood of the origin and has uniformly Hölder continuous  $k$ th derivatives at the origin, with exponent  $\gamma$ . More precisely, this means that there exists a positive, real constant  $c$  such that in some neighborhood of the origin

- (i)  $D^v h(x)$  exists and is continuous in  $x$  for all  $\|v\|_1 \leq k$ , and
- (ii)  $|D^v h(x) - D^v h(0)| \leq c\|x\|^\gamma$  for all  $\|v\|_1 = k$ , with  $\|\cdot\|$  the norm on  $x \in \mathbb{R}^d$  in (2.4).

At one extreme, if the conditional quantile function  $Q_{Y|X}(\cdot; x)$  is the same for all  $x \in C_h$ , we will have the same results as for the discrete case taking  $X_i = 0$  above. At the opposite extreme, if the conditional quantile function varies completely arbitrarily over  $x \in C_h$ , nothing can be learned from the data. If we make local smoothness assumptions on the conditional quantile function in between these two extremes, the method will be informative but subject to additional CPE due to bias.

Assuming a positive, continuous marginal density for  $X$  at the origin,  $N_n$  will asymptotically be proportional to the volume of  $C_h$ , which is proportional to  $h^d$ . There is a tradeoff: larger  $h$  lowers CPE via  $N_n$  but raises CPE via bias. This determines the optimal rate at which  $h \rightarrow 0$  as  $n \rightarrow \infty$ . Using the precise rate results from §1.3.1 and new results established here, we determine the optimal value of  $h$ .

A Bonferroni approach gives joint CIs over  $m < \infty$  different values of  $x_0$ . If the various  $x_0$  yield non-intersecting  $C_h$ , then the  $m$  pointwise CIs will be independent since data are iid. In that case, pointwise  $(1 - \alpha)^{1/m}$  CIs can be used instead of  $(1 - \alpha/m)$ , which is no longer conservative. Asymptotically, for a fixed number of  $x_0$ , this is always the case, and it may be a good approximation even if not exactly true in finite samples. However, the difference is small—for two-sided CIs with  $\alpha = 0.05$ , 0.975 vs. 0.9747 for  $m = 2$ , 0.999 vs. 0.99897 for  $m = 50$ , etc.—so the Bonferroni approach is always used here for convenience. As discussed in §2.5, an alternative Hotelling (1939) tube-based calibration of  $\alpha$  yields similar results.

The unconditional IDEAL method is key to constructing our pointwise CI for  $Q_{Y|X}(p; 0)$  in the conditional case, where we maintain  $x_0 = 0$  as the point of interest.

**Definition 2.3** (conditional IDEAL method). Given iid data  $\{X_i, Y_i\}_{i=1}^n$ , bandwidth  $h > 0$ , quantile  $p \in (0, 1)$ , and desired coverage probability  $(1 - \alpha)$ , first  $C_h$

and  $N_n$  are calculated as in Definition 2.1. Using the values of  $Y_i$  from the effective sample of  $N_n$  observations with  $X_i \in C_h$ , the CI is then constructed as in Lemma 2.1. If additional discrete conditioning variables exist, this method may be run separately for each combination of discrete conditioning values, e.g. once for males and once for females. This procedure may be repeated for any number of  $x_0$ , too.

*Remark.* Code for the unconditional IDEAL method is publicly available in both MATLAB and R on the author's website. The only additional difficulty in the conditional setting is determining the optimal bandwidth (see §2.4.2 and §2.5.1). For  $d = 1$ , fully automated code for conditional IDEAL is available in R, also at the author's website.

*Remark.* Censoring and missing data can be accounted for in many cases. For instance, if  $Y_i$  is missing for some observations, there are two extreme cases to consider: replacing all the missing values with  $Y_{\min}$  (the lower bound of the support of  $Y$ , or  $-\infty$  if unbounded), or replacing all the missing values with  $Y_{\max}$  (the upper bound of the support of  $Y$ , or  $\infty$  if unbounded). A conservative CI is then the convex hull of the IDEAL CIs in the two extreme cases. If there are not too many missing values, this will still produce an informative CI. Extensions of the IDEAL method to missing data (with or without assumptions like missing at random) and different types of censoring should be relatively straightforward and valuable.

In addition to the foregoing definitions, the following assumptions are maintained throughout. We continue using  $x_0 = 0$  as the point of interest. Assumptions A2.1–A2.4(i) are needed for the bias calculation, while A2.4(ii)–A2.6 are needed to apply the unconditional IDEAL quantile inference method (Chapter 1).

**Assumption A2.1.**  $(X_i, Y_i)$  is iid across  $i = 1, 2, \dots, n$ , where  $Y_i$  is a continuous scalar and  $X_i$  a continuous vector with support  $\mathcal{X} \subset \mathbb{R}^d$ .

**Assumption A2.2.** The marginal density of  $X$ , denoted  $f_X(\cdot)$ , satisfies  $f_X(0) > 0$  and has local smoothness  $s_X = k_X + \gamma_X > 0$  with constant  $c_X$ .

**Assumption A2.3.** For all  $u$  in a neighborhood of  $p$ ,  $Q_{Y|X}(u; \cdot)$  (as a function of the second argument) has local smoothness  $s_Q = k_Q + \gamma_Q > 0$  with constant  $c_Q$ .

**Assumption A2.4.** As  $n \rightarrow \infty$ , (i)  $h \rightarrow 0$ , (ii)  $nh^d/[\log(n)]^2 \rightarrow \infty$ .

**Assumption A2.5.** The conditional density of  $Y$  is positive at the quantile and  $X$  values of interest:  $f_{Y|X}(Q_{Y|X}(p; 0); 0) > 0$ .

**Assumption A2.6.** For all  $y$  in a neighborhood of  $Q_{Y|X}(p; 0)$  and  $x$  in a neighborhood of the origin,  $f_{Y|X}(y; x)$  is twice continuously differentiable ( $f_{Y|X} \in C^2$ ) in its first ( $Y$ ) argument, i.e. has local smoothness  $s_Y = k_Y + \gamma_Y > 2$  with constant  $c_Y$ .

*Remark* (smoothness). There is no minimum requirement of  $s_Q$  and  $s_X$ , though as  $s_Q \rightarrow 0$  the inference becomes meaningless, as shown explicitly in §2.4. Since we are implicitly using a uniform kernel, which is a second-order kernel, there is no benefit to having smoothness greater than  $(s_Q, s_X, s_Y) = (2, 1, 1) + \epsilon$  for some arbitrarily small  $\epsilon > 0$ , as stated in Lemma 2.2. Our  $s_Q$  corresponds to variable  $p$  in Chaudhuri (1991), who also notes that Bhattacharya and Gangopadhyay (1990) use  $s_Q = 2$  and  $d = 1$ .

*Remark* (bandwidth). From A2.4(i), asymptotically  $C_h$  will be totally contained within the neighborhoods mentioned in A2.2, A2.3, and Definition 2.2. In order to get  $N_n \xrightarrow{a.s.} \infty$ , A2.4(ii) is a primitive condition. This in turn allows us to examine only a local neighborhood around quantile of interest  $p$  (e.g., as in A2.3), since asymptotically the CI endpoints will converge to the true value at a  $\sqrt{N_n}$  rate. Reassuringly, the optimal bandwidth rate turns out to be inside the assumed bounds.

The bias may now be determined. Since our conditional quantile CI uses the subsample of  $Y_i$  with  $X_i \in C_h$ , rather than a subsample of  $Y_i$  with  $X_i = 0$ , our CI is constructed for the biased conditional quantile  $Q_{Y|C_h}(p)$  (from Definition 2.1) rather than for  $Q_{Y|X}(p; 0)$ . The bias is the difference between these two population conditional quantiles (proof in Appendix B.1).

**Lemma 2.2.** Define  $b \equiv \min\{s_Q, s_X + 1, 2\}$  and  $B_h \equiv Q_{Y|C_h}(p) - Q_{Y|X}(p; 0)$ . If Assumptions A2.2, A2.3, A2.4(i), and A2.6 hold, then the bias is of order

$$|B_h| = O(h^b). \quad (2.6)$$

With A2.2 and A2.3 strengthened to  $k_X \geq 1$  and  $k_Q \geq 2$ , the bias is  $O(h^2)$  with remainder  $o(h^2)$ . With further strengthening to  $k_X \geq 2$  and  $k_Q \geq 3$ , the bias remains  $O(h^2)$ , but the remainder sharpens to  $o(h^3)$ . With  $d = 1$ , and defining

$$\begin{aligned} Q_{Y|X}^{(0,1)}(p; 0) &\equiv \left. \frac{\partial}{\partial x} Q_{Y|X}(p; x) \right|_{x=0}, & Q_{Y|X}^{(0,2)}(p; 0) &\equiv \left. \frac{\partial^2}{\partial x^2} Q_{Y|X}(p; x) \right|_{x=0}, \\ \xi_p &\equiv Q_{Y|X}(p; 0), \\ f_{Y|X}^{(0,1)}(y; 0) &\equiv \left. \frac{\partial}{\partial x} f_{Y|X}(y; x) \right|_{x=0}, & f_{Y|X}^{(1,0)}(\xi_p; 0) &\equiv \left. \frac{\partial}{\partial y} f_{Y|X}(y; 0) \right|_{y=\xi_p}, \end{aligned}$$

and similarly for  $F_{Y|X}^{(0,1)}(\xi_p; 0)$  and  $F_{Y|X}^{(0,2)}(\xi_p; 0)$ , the bias is

$$\begin{aligned} B_h &= \frac{h^2}{6} \left\{ 2Q_{Y|X}^{(0,1)}(p; 0) f'_X(0) / f_X(0) + Q_{Y|X}^{(0,2)}(p; 0) \right. \\ &\quad + 2f_{Y|X}^{(0,1)}(\xi_p; 0) Q_{Y|X}^{(0,1)}(p; 0) / f_{Y|X}(\xi_p; 0) \\ &\quad \left. + f_{Y|X}^{(1,0)}(\xi_p; 0) \left[ Q_{Y|X}^{(0,1)}(p; 0) \right]^2 / f_{Y|X}(\xi_p; 0) \right\} + R \\ &= -h^2 \frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0) F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0) f_{Y|X}(\xi_p; 0)} + R, \end{aligned}$$

where  $R = o(h^2)$  or  $R = o(h^3)$  as discussed.

*Remark.* The dominant term in the bias when  $k_X \geq 1$ ,  $k_Q \geq 2$ , and  $d = 1$  is the same as the bias in Bhattacharya and Gangopadhyay (1990), who derive it using different arguments.

We discuss some intuition of the proof here. We start with an identity for  $Q_{Y|C_h}(p)$  and subtract off the corresponding identity for  $Q_{Y|X}(p; 0)$ . Using the smoothness assumptions, a Taylor expansion (of some order) of the remainder may be taken. The bias appears in the lowest-order term; some cancellation and rearrangement leads to the final expression. Further manipulations to replace  $Q_{Y|X}$  with  $F_{Y|X}$  lead to the exact same expression as in Bhattacharya and Gangopadhyay (1990).

*Remark.* Even when  $k_X \geq 1$  and  $k_Q \geq 2$ , the bias will never shrink smaller than  $O(h^2)$  since we are effectively using a second-order (uniform) kernel. It is unclear if the IDEAL fractional order statistic results can be used with a higher-order



kernel. Alternatively, higher-order kernels could likely be used with a method as in Chernozhukov et al. (2009), who in the parametric quantile regression model use Bernoulli random variables and MCMC simulation.

## 2.4 Optimal bandwidth and CPE

### 2.4.1 Optimal rate of bandwidth and CPE

The optimal bandwidth minimizes the effect of the two dominant high-order terms on coverage probability error (CPE). In terms of coverage probability (CP) and nominal coverage  $1 - \alpha$ , we follow convention and define  $\text{CPE} \equiv \text{CP} - (1 - \alpha)$ , so that CPE is positive when there is over-coverage and negative when there is under-coverage. This means the equivalent hypothesis test is size distorted when CPE is negative.

From §1.3.1, we know the IDEAL CPE in the unconditional one-sample case for Hutson's (1999) confidence interval (CI). In that case, we are interested in the  $p$ -quantile  $F^{-1}(p)$  of scalar random variable  $Y$ . The Hutson (1999) method's CPE with respect to sample size  $n$  is of order  $n^{-1}$ . To apply this result, we need a more precise handle on the effective sample size  $N_n$ , which is random. From Chaudhuri (1991, proof of Thm. 3.1, p. 769), and under A2.1, we can choose  $c_1, c_2, c_3, c_4 > 0$  such that

$$P(A_n) \geq 1 - c_3 \exp(-c_4 n h^d)$$

for all  $n$ , where  $A_n \equiv \{c_1 n h^d \leq N_n \leq c_2 n h^d\}$  and  $C_h$  is the hypercube from Definition 2.1. (Adjusting the  $c_i$  appropriately, Chaudhuri's (1991) argument goes through for a hypersphere  $C_h$  also.) If the rate of  $h$  leads to  $\sum_n [1 - P(A_n)] < \infty$ , then the Borel–Cantelli Lemma gives  $P(\liminf A_n) = 1$ . This holds for the optimal bandwidth rates derived here, which all satisfy A2.4.

### One-sided inference

In the lower one-sided case, we write  $\hat{Q}_{Y|C_h}^L(u_h)$  as the Hutson (1999) upper endpoint, with notation similar to §2.2. This is a linearly interpolated fractional

order statistic approximation calculated from the  $N_n$  values of  $Y_i$  with  $X_i \in C_h$ . The random variable  $F_{Y|C_h}(\hat{Q}_{Y|C_h}^I(u_h))$  follows a (collapsing) beta distribution, which has a continuously differentiable PDF in  $(0, 1)$  that converges to a (collapsing) normal PDF at a  $\sqrt{N_n}$  rate (Chapter 1). Other than  $O(N_n^{-1})$  CPE from interpolating between order statistics, the CI using  $\hat{Q}_{Y|C_h}^L(u_h)$  is exact for  $Q_{Y|C_h}(p)$ . The fact that instead  $Q_{Y|X}(p; 0)$  is of interest introduces additional CPE from the bias. The CP of the lower one-sided CI is

$$\begin{aligned}
& P\left(Q_{Y|X}(p; 0) < \hat{Q}_{Y|C_h}^L(u_h)\right) \\
&= P\left(Q_{Y|C_h}(p) < \hat{Q}_{Y|C_h}^L(u_h)\right) \\
&\quad + \left[ P\left(Q_{Y|X}(p; 0) < \hat{Q}_{Y|C_h}^L(u_h)\right) - P\left(Q_{Y|C_h}(p) < \hat{Q}_{Y|C_h}^L(u_h)\right) \right] \\
&= 1 - \alpha + \text{CPE}_{\text{GK}} + \text{CPE}_{\text{Bias}}, \tag{2.7}
\end{aligned}$$

where  $\text{CPE}_{\text{GK}}$  is due to the Chapter 1 CPE from Lemma 2.1 and  $\text{CPE}_{\text{Bias}}$  comes from the bias discussed in §2.3.

As before, define  $B_h \equiv Q_{Y|C_h}(p) - Q_{Y|X}(p; 0)$ . From Lemma 2.2, we have  $B_h = O(h^b)$  with  $b \equiv \min\{s_Q, s_X + 1, 2\}$ . Let  $F_{\hat{Q}_{Y|C_h}^I, u_h}(\cdot)$  and  $f_{\hat{Q}_{Y|C_h}^I, u_h}(\cdot)$  be the CDF and PDF, respectively, of  $\hat{Q}_{Y|C_h}^I(u_h)$ . If  $B_h$  is sufficiently small—i.e. if  $h^b = o(N_n^{-1/2})$ , which is true below since  $h^b = N_n^{-3/2}$ —then we can approximate

$$\begin{aligned}
\text{CPE}_{\text{Bias}} &= P\left(Q_{Y|X}(p; 0) < \hat{Q}_{Y|C_h}^L(u_h)\right) - P\left(Q_{Y|C_h}(p) < \hat{Q}_{Y|C_h}^L(u_h)\right) \\
&= P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|C_h}(p)\right) - P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|X}(p; 0)\right) \\
&= P\left(\hat{Q}_{Y|C_h}^I(u_h) < Q_{Y|C_h}(p)\right) - P\left(\hat{Q}_{Y|C_h}^I(u_h) < Q_{Y|X}(p; 0)\right) \\
&\quad + O(B_h N_n^{-1/2}) \\
&= F_{\hat{Q}_{Y|C_h}^I, u_h}(Q_{Y|C_h}(p)) - F_{\hat{Q}_{Y|C_h}^I, u_h}(Q_{Y|X}(p; 0)) + O(B_h N_n^{-1/2}) \\
&= B_h f_{\hat{Q}_{Y|C_h}^I, u_h}(Q_{Y|C_h}(p)) + O(B_h N_n^{-1/2} + B_h^2 N_n), \tag{2.8}
\end{aligned}$$

where the order of the approximation error from switching to  $\hat{Q}_{Y|C_h}^I(u_h)$  from  $\hat{Q}_{Y|C_h}^L(u_h)$  comes from Theorem 1.1 in Chapter 1, and the other remainder is the subsequent  $B_h^2$  term in the Taylor expansion that would be multiplied by an  $O(N_n)$  PDF derivative as in (2.10). From the aforementioned PDF convergence

of  $F_{Y|C_h}(\hat{Q}_{Y|C_h}^I(u_h))$  to a normal, it can be shown that  $f_{\hat{Q}_{Y|C_h}^{I,u_h}}(Q_{Y|X}(p; 0)) \asymp N_n^{1/2}$ . Since  $B_h = O(h^b)$  from Lemma 2.2, the dominant term of  $\text{CPE}_{\text{Bias}}$  is  $O(N_n^{1/2}h^b)$ . The expression in (2.8) holds for  $B_h > 0$  (leading to over-coverage) or  $B_h < 0$  (under-coverage).

The two dominant terms of  $\text{CPE}_{\text{GK}}$  and  $\text{CPE}_{\text{Bias}}$  are thus respectively  $O(N_n^{-1})$  and  $O(N_n^{1/2}h^b)$ . These are both sharp except in the special case of  $u_h(N_n + 1)$  being an integer or of  $f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p; 0) = 0$  when  $k_Q \geq 2$  and  $k_X \geq 1$  (or similar conditions otherwise); the following assumes we are not in a special case. The term  $\text{CPE}_{\text{GK}}$  is always positive. If  $\text{CPE}_{\text{Bias}}$  is negative, the optimal  $h$  will make them cancel, and it will set the orders equal:

$$N_n^{-1} \asymp N_n^{1/2}h^b \implies (nh^d)^{3/2} \asymp h^{-b} \implies h \asymp n^{-3/(2b+3d)}.$$

Overall CPE will then be  $o(n^{-2b/(2b+3d)})$ . Even if  $h$  does not make the dominant CPE terms cancel, as long as it is the above asymptotic rate, the overall CPE will be  $O(n^{-2b/(2b+3d)})$ .

If instead the two terms are both positive, minimizing the sum will lead to a first-order condition like

$$0 = \frac{\partial}{\partial h} [N_n^{-1} + N_n^{1/2}h^b] = (-d)n^{-1}h^{-d-1} + (b + (d/2))n^{1/2}h^{b+(d/2)-1},$$

giving the same rate  $h \asymp n^{-3/(2b+3d)}$ . In that case, overall CPE will be positive (over-coverage) and of order

$$N_n^{-1} \asymp (nh^d)^{-1} \asymp n^{-1+3d/(2b+3d)} = n^{-2b/(2b+3d)}.$$

If the calibrated unconditional method from Chapter 1 is used,  $\text{CPE}_{\text{GK}} = O(N_n^{-3/2+\rho})$  for arbitrarily small  $\rho > 0$ . Ignoring the  $\rho$  for simplicity, the optimal bandwidth rate is then  $h \asymp n^{-2/(b+2d)}$ , leading to overall CPE of  $O(n^{-3b/(2b+4d)})$ .

In the upper one-sided case, with  $\hat{Q}_{Y|C_h}^L(u_\ell)$  the lower endpoint,  $\text{CPE}_{\text{GK}}$  is of the same order and sign, while

$$\begin{aligned} \text{CPE}_{\text{Bias}} &= P(Q_{Y|X}(p; 0) > \hat{Q}_{Y|C_h}^L(u_\ell)) - P(Q_{Y|C_h}(p) > \hat{Q}_{Y|C_h}^L(u_\ell)) \\ &= F_{\hat{Q}_{Y|C_h}^{I,u_\ell}}(Q_{Y|X}(p; 0)) - F_{\hat{Q}_{Y|C_h}^{I,u_\ell}}(Q_{Y|C_h}(p)) + O(B_h N_n^{-1/2}) \\ &= -B_h f_{\hat{Q}_{Y|C_h}^{I,u_\ell}}(Q_{Y|C_h}(p)) + O(B_h N_n^{-1/2} + B_h^2 N_n). \end{aligned} \quad (2.9)$$

Opposite before,  $B_h > 0$  now contributes under-coverage and  $B_h < 0$  over-coverage, but for now it suffices to note that the order of  $\text{CPE}_{\text{Bias}}$  is the same as before.

## Two-sided inference

With two-sided inference, the lower and upper endpoints have opposite bias effects. For the median, the dominant terms of these effects cancel completely. For other quantiles, there is a partial, order-reducing cancellation. Below, since by construction  $\hat{Q}_{Y|C_h}^L(u_\ell) < \hat{Q}_{Y|C_h}^L(u_h)$ , it is certain that  $\hat{Q}_{Y|C_h}^L(u_\ell) < c$  if  $\hat{Q}_{Y|C_h}^L(u_h) < c$  and that  $\hat{Q}_{Y|C_h}^L(u_h) > c$  if  $\hat{Q}_{Y|C_h}^L(u_\ell) > c$ , for any  $c$ . With two-sided CI  $(\hat{Q}_{Y|C_h}^L(u_\ell), \hat{Q}_{Y|C_h}^L(u_h))$ , CP is

$$\begin{aligned}
& P\left(\hat{Q}_{Y|C_h}^L(u_\ell) < Q_{Y|X}(p; 0) < \hat{Q}_{Y|C_h}^L(u_h)\right) \\
&= 1 - P\left(\hat{Q}_{Y|C_h}^L(u_\ell) > Q_{Y|X}(p; 0)\right) - P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|X}(p; 0)\right) \\
&= 1 - P\left(\hat{Q}_{Y|C_h}^L(u_\ell) > Q_{Y|C_h}(p)\right) \\
&\quad + \left[ P\left(\hat{Q}_{Y|C_h}^L(u_\ell) > Q_{Y|C_h}(p)\right) - P\left(\hat{Q}_{Y|C_h}^L(u_\ell) > Q_{Y|X}(p; 0)\right) \right] \\
&\quad - P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|C_h}(p)\right) \\
&\quad + \left[ P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|C_h}(p)\right) - P\left(\hat{Q}_{Y|C_h}^L(u_h) < Q_{Y|X}(p; 0)\right) \right] \\
&= 1 - \alpha + \text{CPE}_{\text{GK}} + \left[ 1 - F_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) \right] - \left[ 1 - F_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|X}(p; 0)) \right] \\
&\quad + F_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) - F_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|X}(p; 0)) + O(B_h N_n^{-1/2}) \\
&= 1 - \alpha + \text{CPE}_{\text{GK}} + B_h \left[ f_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) - f_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) \right] \\
&\quad + (1/2)B_h^2 \left[ f'_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) - f'_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) \right] \\
&\quad + O\left\{ B_h^3 f''_{\hat{Q}}(Q_{Y|C_h}(p)) + B_h N_n^{-1/2} \right\}. \tag{2.10}
\end{aligned}$$

For the special case of the median, the  $B_h$  term zeroes out. This happens because the beta distribution PDFs of  $F_{Y|C_h}\left(\hat{Q}_{Y|C_h}^I(u)\right)$  and  $F_{Y|C_h}\left(\hat{Q}_{Y|C_h}^I(1-u)\right)$  are reflections of each other around  $p = 1/2$ —i.e.,  $f_\beta(x; u) = f_\beta(1-x; 1-u)$ ,  $\forall x \in (0, 1)$ —so the upper and lower  $u$  are symmetric around  $p = 1/2$ . See Figure 2.1 for a visual example and (B.3) for the endpoint PDF expression. Consequently, the

dominant bias effect on CPE is the  $B_h^2$  term instead. This makes the overall CPE smaller. The optimal rate of  $h$  will equate  $(nh^d)^{-1} \asymp h^{2b}(nh^d)$ , so  $h^* \asymp n^{-1/(b+d)}$  and CPE is  $O(n^{-b/(b+d)})$ . With the calibrated method (again suppressing the  $\rho > 0$ ), the rates would instead be  $h^* \asymp n^{-5/(4b+3d)}$  and  $\text{CPE} = O(n^{-6b/(4b+5d)})$ .

Even with  $p \neq 1/2$ , the same rates hold. As seen in (2.12), the PDF difference multiplying  $B_h$  is only  $O(1)$ , smaller than the  $O(N_n^{1/2})$  PDF value multiplying  $B_h$  in the one-sided expression (2.8). This makes the  $B_h$  and  $B_h^2$  terms the same order, as discussed further in §2.4.2.

**Theorem 2.3.** *Let Assumptions A2.1–A2.6 hold, and define  $b \equiv \min\{s_Q, s_X + 1, 2\}$ .*

*For a one-sided CI, the bandwidth  $h^*$  minimizing CPE for the Hutson (1999) method applied to observations falling inside  $C_h$  has rate  $h^* \asymp n^{-3/(2b+3d)}$ . This corresponds to overall CPE of  $O(n^{-2b/(2b+3d)})$ .*

*For two-sided inference, the optimal bandwidth rate is  $h^* \asymp n^{-1/(b+d)}$ , and the optimal CPE is  $O(n^{-b/(b+d)})$ .*

*For the median, with the precise bandwidth value provided in §2.4.2, the two-sided CPE reduces to  $o(n^{-b/(b+d)})$ . With  $k_Q \geq 2$  and  $k_X \geq 1$ , the CPE is  $o(n^{-2/(2+d)})$ .*

*Using the calibrated method proposed in Chapter 1, the two-sided CPE-optimal bandwidth rate is  $h^* \asymp n^{-(5-2\rho)/(4b+5d-2d\rho)}$  for arbitrarily small  $\rho > 0$ , yielding CPE of  $O(n^{-(6b-4b\rho)/(4b+5d-2d\rho)})$ . The optimal one-sided calibrated bandwidth rate is  $h^* \asymp n^{-(2-\rho)/(b+2d-d\rho)}$ , yielding CPE of  $O(n^{-(3b-2b\rho)/(2b+4d-2d\rho)})$ .*

*Remark.* With  $k_X \geq 2$  and  $k_Q \geq 3$ , if it could be guaranteed that  $N_n = nP(X_i \in C_h)[1 + O(N_n^{-1/2} \log(N_n))]$ , then CPE would be  $O(n^{-3/(2+d)} \log(n))$ .

## Discussion: smoothness and bandwidth

The following discussion is for the more common two-sided inference, with one-sided equivalents noted in parentheses. Let  $\kappa \equiv 1/(b+d)$  (one-sided:  $\kappa \equiv 3/(2b+3d)$ ), so that  $h = n^{-\kappa}$  above, with CPE of order  $n^{d\kappa-1}$ . Since  $0 < b \leq 2$ , the optimal  $\kappa$  will depend on A2.2 and A2.3 but fall within the range  $1/(2+d) \leq$

$\kappa < 1/d$  (one-sided:  $3/(4 + 3d) \leq \kappa < 1/d$ ). This corresponds to CPE order in the range  $[n^{-2/(2+d)}, n^0)$  (one-sided:  $[n^{-4/(4+3d)}, n^0)$ ). The high end of this range matches intuition: as the smoothness diminishes to zero ( $s_Q \rightarrow 0$ ), we are unable to say anything informative. However, the smoothness levels  $(s_Q, s_X, s_Y) = (2, 1, 2) + \epsilon$  for any small  $\epsilon > 0$  are quite mild, so it is most helpful in practice to look at the other end of the range.

As  $d$  increases,  $h \rightarrow 0$  more slowly (smaller  $\kappa$ ). More smoothness also makes  $h \rightarrow 0$  more slowly: we can afford a relatively larger window  $C_h$  if there is less variability near our point of interest.

With  $d = 1$  and  $b = 2$ , we get  $h^* \asymp n^{-1/3}$  and a CPE of order  $n^{-2/3}$  (one-sided:  $h^* \asymp n^{-3/7}$ , CPE  $n^{-4/7}$ ). With  $d = 2$ , the CPE order increases to  $n^{-1/2}$  (one-sided:  $n^{-2/5}$ ). We expect the method to work (relatively) well with even higher-dimensional  $X$ , though CPE continues to increase as  $d$  increases.

Regarding robustness to bandwidth, our CI will be asymptotically (first-order) correct as long as  $N_n^{-1} \rightarrow 0$  (so  $\text{CPE}_{\text{GK}} \rightarrow 0$ ) and  $N_n h^{2b} \rightarrow 0$  (so  $\text{CPE}_{\text{Bias}} \rightarrow 0$ ). These are equivalent to  $1/(2b+d) < \kappa < 1/d$ . With the common example  $d = 1$  and  $b = 2$ , this says that any  $h \asymp n^{-\kappa}$  with  $1/5 < \kappa < 1$  will give  $\text{CPE} = o(1)$ . The range for  $\kappa$  is the same as when using the local polynomial asymptotic normality approach.

### **Discussion: comparison with asymptotic normality or bootstrap**

Theorem 2.3 suggests that for most common combinations of smoothness  $s_Q$  and dimensionality  $d$ , our method is preferred to inference based on asymptotic normality or bootstrap with a local polynomial estimator. Only the uncalibrated IDEAL method is compared here; the IDEAL advantage is even greater for the calibrated version. As shown in Chaudhuri (1991), the local polynomial quantile regression estimator achieves the optimal rate of convergence for nonparametric mean regression given by Stone (1980, 1982); other nonparametric estimators may be used instead, but the theoretical CPE should not be able to improve for asymptotic normality or percentile bootstrap. Consequently, the following comparative discussion is detailed.

The main limitation of our method is the uniform kernel required to leverage the fractional order statistic theory. Even though asymptotic normality has a larger error in terms of  $N_n$ , it could be smaller in terms of  $n$  if  $N_n$  is allowed to be much bigger. This could potentially happen if a high enough degree of smoothness  $s_Q$  is assumed and there are enough observations that it is appropriate to fit a correspondingly high-degree local polynomial. As we will see, though, our method has smaller CPE in the most important cases even with  $s_Q = \infty$ .

Results from Chaudhuri (1991) can be reworked to obtain the optimal CPE for inference based on asymptotic normality of the local polynomial estimator. Note that  $\hat{Q}_{Y|X}(p; 0) = \hat{\beta}_{(0)}$ , the intercept term estimator from the local polynomial quantile regression, so inference on  $\beta_{(0)}$  is equivalent to inference on  $Q_{Y|X}(p; 0)$ . The goal of Chaudhuri (1991) is to show that the local polynomial estimator therein achieves the optimal rate of convergence given by Stone (1980, 1982) for nonparametric mean regression. A decomposition of the estimator is given as

$$\hat{\beta} - \beta_0 = V_n + B_n + R_n,$$

where  $R_n$  is a Bahadur-type remainder from Theorem 3.3,  $B_n$  is the bias from equation (4.1), and  $V_n$  is the term from Proposition 4.2 that when scaled by  $\sqrt{N_n}$  converges to a Gaussian limit. To get the best rate, it is necessary to balance the squared bias with the variance. The given MSE-optimal bandwidth is  $h \propto n^{-1/(2s_Q+d)}$  (p. 763, in our notation), balancing the square of the  $B_n = O(h^{s_Q}) = O(n^{-s_Q/(2s_Q+d)})$  bias (Prop. 4.1;  $r_n(x)$  on p. 765) with the  $N_n^{-1}$  variance (Prop. 4.2), where  $N_n \stackrel{a.s.}{\asymp} n^{2s_Q/(2s_Q+d)}$  (Prop. 4.2 proof). However, with the bias the same order of magnitude as the standard deviation, CPE is  $O(1)$ . If CPE is the target instead of MSE, a smaller bandwidth is necessary.

To find the CPE-optimal bandwidth for Chaudhuri (1991), we balance the CPE from the bias with additional CPE from the Bahadur remainder from Theorem 3.3(ii), ignoring asymptotic variance estimation error. The CPE due to the bias is of order  $N_n^{1/2}h^{s_Q}$ . Chaudhuri (1991, Thm. 3.3) gives a Bahadur-type expansion of the local polynomial quantile regression estimator that has remainder  $R_n\sqrt{N_n} = O(N_n^{-1/4})$  (up to log terms) as in Bahadur (1966), but recently Port-

noy (2012) has shown that the CPE is nearly  $O(N_n^{-1/2})$  in such cases. Solving  $N_n^{1/2}h^{s_Q} = N_n^{-1/2} = (nh^d)^{-1/2}$ , this yields  $h^* \asymp n^{-1/(s_Q+d)}$ . The optimal CPE is then (nearly)  $O(\sqrt{N_n}B_n) = O(N_n^{-1/2}) = O(n^{-s_Q/(2s_Q+2d)})$ .

Specifically, and very much theoretically, in Chaudhuri (1991), if  $s_Q \rightarrow \infty$ , then  $N_n \rightarrow n$  and (nearly) CPE  $\rightarrow O(n^{-1/2})$ . In other words, with infinite smoothness, almost the entire sample  $n$  is used because the fitted infinite-degree polynomial can perfectly approximate the true function. Practically, when we have a finite sample, we can't even fit an  $n$ th-degree polynomial, let alone infinite-degree. Additionally, the argument in Chaudhuri (1991) is that even if the smoothness only holds in some tiny neighborhood  $V$ , asymptotically there will be an infinite number of observations within  $V$  (and  $C_h$  will be contained in  $V$ ). But when the sample is finite, smoothness must hold over  $C_h$ , which may not be so small (and is always of larger-order volume than our  $C_h$ ). Barring relevant a priori information, speculating more smoothness over a larger region may be prohibitively unpalatable in light of the more robust method we provide.

If  $s_Q = 2$  (one Lipschitz-continuous derivative), then optimal CPE from asymptotic normality is nearly  $O(n^{-2/(4+2d)})$ . This goes to  $n^0$  as  $d \rightarrow \infty$  (same as above). With  $d = 1$ , this is  $n^{-1/3}$ , significantly larger than our  $n^{-2/3}$  (one-sided:  $n^{-4/7}$ ); with  $d = 2$ ,  $n^{-1/4}$  is larger than our  $n^{-1/2}$  (one-sided:  $n^{-2/5}$ ); and it remains larger for all  $d$  (even for one-sided inference). With  $s_Q = 1$  (no derivatives, but Lipschitz continuity), then optimal CPE from asymptotic normality is nearly  $O(n^{-1/(2+2d)})$ , compared to our CPE of  $O(n^{-1/(1+d)})$  (one-sided:  $O(n^{-1/(2+2d)})$ ). With  $d = 1$ , this is  $n^{-1/4}$ , again much larger than our  $n^{-1/2}$  (one-sided:  $n^{-2/5}$ ); and again it remains larger for all  $d$  (including one-sided).

From another perspective: what amount of smoothness (and degree of local polynomial fit) is needed for asymptotic normality to match the CPE of our method? For any  $d$ , the bound on CPE for asymptotic normality is nearly  $n^{-1/2}$  (with  $s_Q \rightarrow \infty$ ). For the most common cases of  $d \in \{1, 2\}$  (one-sided:  $d = 1$ ), asymptotic normality will be worse even with infinite smoothness. With  $d = 3$  (one-sided:  $d = 2$ ), asymptotic normality needs  $s_Q \geq 12$  (one-sided:  $s_Q \geq 12$ ) to have as good CPE. If  $n$  is quite large, maybe that high of a local polynomial



degree ( $k_Q \geq 11$ ) will be appropriate, but often it is not. Note that interaction terms are required, so an 11th-degree polynomial has 364 terms. As  $d \rightarrow \infty$ , the required smoothness approaches  $s_Q = 4$  (one-sided:  $8/3$ ) from above, though again the number of terms in the local polynomial grows with  $d$  as well as  $k_Q$  and may be prohibitive in finite samples. Even with very high-dimensional  $X$ , asymptotic normality will only be better under a stronger smoothness assumption and fourth-degree (i.e., includes  $x^4$  term) local polynomial fit.

In finite samples, the asymptotic normality approach may have additional error from estimation of the asymptotic variance, which includes the “dispersion matrix” as well as the probability density of the error term at zero as discussed in Chaudhuri (1991, pp. 764–766). There is no direct error from nuisance parameter estimation in our method, only through the plug-in bandwidth.

For the bootstrap, the basic percentile method offers no higher-order refinement over first-order asymptotic normality. Consequently, our method has better CPE than the bootstrap in all the cases discussed above. Gangopadhyay and Sen (1990) examine the bootstrap percentile method for a uniform kernel (or nearest-neighbor) estimator with  $d = 1$  (and  $s_Q = 2$ ), but they only show first-order consistency. Based on their (2.14), the optimal error seems to be  $O(n^{-1/3})$  when  $h \asymp n^{-1/3}$  to balance the bias and remainder terms, improved by Portnoy (2012) from  $O(n^{-2/11})$  CPE when  $h \asymp n^{-3/11}$ . Although we are unaware of any improvements on Gangopadhyay and Sen (1990) to date, it is still valuable to pursue different inference strategies, and developing strong resampling or subsampling competitors here is no exception.

## 2.4.2 Plug-in bandwidth

The terms  $\text{CPE}_{\text{GK}}$  and  $\text{CPE}_{\text{Bias}}$  can be calculated more precisely than simply the rates. Chapter 1 contains an exact expression for the  $O(N_n^{-1})$   $\text{CPE}_{\text{GK}}$  term. This can be used to reduce the CPE to almost  $O(N_n^{-3/2})$  via analytic calibration, but we use it to determine the precise value of the optimal bandwidth. In theory, it is better to use the analytic calibration and an ad hoc bandwidth of the proper rate. In practice, it is helpful (as in §2.4.3) to know that the  $O(N_n^{-1})$

CPE term only leads to over-coverage, which implies that smaller  $h$  is always more conservative, and the detrimental effect of an ad hoc bandwidth can be significant in smaller samples.

As before,  $p \in (0, 1)$  is the quantile of interest. Let  $\phi(\cdot)$  be the standard normal PDF,  $z_{1-\alpha}$  be the  $(1-\alpha)$ -quantile of the standard normal distribution,  $u_h > p$  (or  $u_\ell < p$ ) be the quantile determining the high (low) endpoint of a lower (upper) one-sided CI,  $\epsilon_h \equiv (N_n + 1)u_h - \lfloor (N_n + 1)u_h \rfloor$ , and  $\epsilon_\ell \equiv (N_n + 1)u_\ell - \lfloor (N_n + 1)u_\ell \rfloor$ . Let  $I_H$  denote a  $100(1-\alpha)\%$  CI constructed using Hutson's (1999) method on a univariate data sample of size  $N_n$ . The coverage probability of a lower one-sided  $I_H$  (with the upper one-sided result substituting  $\epsilon_\ell$  for  $\epsilon_h$ ) is

$$P\{F^{-1}(p) \in I_H\} = 1 - \alpha + N_n^{-1} z_{1-\alpha} \frac{\epsilon_h(1 - \epsilon_h)}{p(1 - p)} \phi(z_{1-\alpha}) + o(N_n^{-1}),$$

or for a two-sided CI,

$$P\{F^{-1}(p) \in I_H\} = 1 - \alpha + N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1 - \epsilon_h) + \epsilon_\ell(1 - \epsilon_\ell)}{p(1 - p)} \phi(z_{1-\alpha/2}) + o(N_n^{-1}).$$

In either case there is  $O(N_n^{-1})$  over-coverage.

While we know the sign of  $\text{CPE}_{\text{GK}}$ , we don't always know the sign of the bias. Specifically, when the rate-limiting term of  $B_h$  is determined by Hölder continuity (of  $Q_{Y|X}(\cdot; \cdot)$  or  $f_X(\cdot)$ ), we lose the sign at that step. However, when  $k_Q \geq 2$  and  $k_X \geq 1$ , the Hölder continuity terms all end up in the remainder while the rate-limiting terms are signed derivatives.

Since  $(k_Q, k_X, k_Y) = (2, 1, 2)$  is only a mild smoothness assumption, which also gives the smallest attainable order of bias, we maintain it for our plug-in bandwidth. We also take  $d = 1$  for simplicity.

From Lemma 2.2, we have two explicit expressions for  $B_h$ ,

$$\begin{aligned} B_h &= \frac{h^2}{6} \left\{ \left[ 2Q_{Y|X}^{(0,1)}(p; 0) f'_X(0) / f_X(0) + Q_{Y|X}^{(0,2)}(p; 0) \right] \right. \\ &\quad \left. + 2f_{Y|X}^{(0,1)}(\xi_p; 0) Q_{Y|X}^{(0,1)}(p; 0) / f_{Y|X}(\xi_p; 0) \right. \\ &\quad \left. + f_{Y|X}^{(1,0)}(\xi_p; 0) \left[ Q_{Y|X}^{(0,1)}(p; 0) \right]^2 / f_{Y|X}(\xi_p; 0) \right\} + o(h^2) \\ &= -h^2 \frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0) F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0) f_{Y|X}(\xi_p; 0)} + o(h^2). \end{aligned}$$

For  $f_X(0)$ ,  $f'_X(0)$ , and  $f_{Y|X}(\xi_p; 0)$ , we could either estimate them or use a parametric plug-in assumption (e.g., assume the distribution is Gaussian and estimate its parameters). For the first formulation above,  $Q_{Y|X}^{(0,1)}(p; 0)$ ,  $Q_{Y|X}^{(0,2)}(p; 0)$ ,  $f_{Y|X}^{(1,0)}(\xi_p; 0)$ , and  $f_{Y|X}^{(0,1)}(\xi_p; 0)$  also must be estimated. Alternatively, for the second formulation,  $F_{Y|X}^{(0,1)}(\xi_p; 0)$  and  $F_{Y|X}^{(0,2)}(\xi_p; 0)$  must be estimated. Recall that  $X = 0$  really means  $X = x_0$ , our target point of interest.

Additionally,  $\text{CPE}_{\text{Bias}}$  depends on  $f_{\hat{Q}_u}(Q_{Y|C_h}(p))$ . From earlier,

$$F_{Y|C_h}(\hat{Q}_u) \sim \beta[(N_n + 1)u, (N_n + 1)(1 - u)].$$

Let  $f_\beta(\cdot; u)$  denote the corresponding beta distribution's PDF and  $F_\beta(\cdot; u)$  its CDF. As shown in the appendix, some identities and calculus lead to

$$f_{\hat{Q}_u}(Q_{Y|C_h}(p)) = f_\beta(p; u) f_{Y|C_h}(F_{Y|C_h}^{-1}(p)).$$

The term  $f_{Y|C_h}(F_{Y|C_h}^{-1}(p))$  is equal to  $f_{Y|X}(\xi_p; 0)$  up to smaller-order terms. The term  $f_\beta(p; u)$  is well approximated by a normal PDF. As detailed in the appendix, for either one-sided  $(1 - \alpha)$  CI endpoint quantile  $u = u_h$  or  $u = u_\ell$  chosen by the Hutson (1999) method,

$$f_\beta(p; u) = N_n^{1/2} [u(1 - u)]^{-1/2} \phi(z_{1-\alpha}) [1 + O(N_n^{-1/2})]. \quad (2.11)$$

For convenient reference, the plug-in bandwidth expressions are collected here. Some intuition follows; details of calculation may be found in Appendix B.2. We continue to assume  $d = 1$ ,  $k_Q \geq 2$ , and  $k_X \geq 1$ , with quantile of interest  $p \in (0, 1)$  and point of interest  $X = x_0$ . Standard normal quantiles are denoted, for example,  $z_{1-\alpha}$  for the  $(1 - \alpha)$ -quantile such that  $\Phi(z_{1-\alpha}) = 1 - \alpha$ . We let  $\hat{B}_h$  denote the estimator of bias term  $B_h$ ;  $\hat{f}_X$  the estimator of  $f_X(x_0)$ ;  $\hat{f}'_X$  the estimator of  $f'_X(x_0)$ ;  $\hat{F}_{Y|X}^{(0,1)}$  the estimator of  $F_{Y|X}^{(0,1)}(\xi_p; x_0)$ ; and  $\hat{F}_{Y|X}^{(0,2)}$  the estimator of  $F_{Y|X}^{(0,2)}(\xi_p; x_0)$ , where  $\xi_p \equiv Q_{Y|X}(p; x_0)$ . To avoid a recursive definition of the plug-in bandwidth, we use  $\epsilon_h = \epsilon_\ell = 0.2$  as a rule of thumb (which cancels nicely with other constants). The largest possible bandwidth would use 0.5 instead, which would give extremely similar bandwidths since, for example,  $[(0.2)(0.8)]^{1/6} = 0.74$  while  $[(0.5)(0.5)]^{1/6} = 0.79$  in the two-sided median case.

We recommend the following when  $d = 1$ .

- For one-sided inference, let

$$\hat{h}_{+-} = n^{-3/7} \left( \frac{z_{1-\alpha}}{3 [p(1-p)\hat{f}_X]^{1/2} \left\{ \hat{f}_X \hat{F}_{Y|X}^{(0,2)} + 2\hat{f}'_X \hat{F}_{Y|X}^{(0,1)} \right\}} \right)^{2/7},$$

$$\hat{h}_{++} = n^{-3/7} \left( \frac{z_{1-\alpha}}{3 [p(1-p)\hat{f}_X]^{1/2} \left\{ \hat{f}_X \hat{F}_{Y|X}^{(0,2)} + 2\hat{f}'_X \hat{F}_{Y|X}^{(0,1)} \right\} (-5/2)} \right)^{2/7}.$$

- For lower one-sided inference,  $\hat{h}_{+-}$  should be used if  $\hat{B}_h < 0$ , and  $\hat{h}_{++}$  otherwise.
- For upper one-sided inference,  $\hat{h}_{++}$  should be used if  $\hat{B}_h < 0$ , and  $\hat{h}_{+-}$  otherwise.
- For two-sided inference on the median,

$$\hat{h} = n^{-1/3} \left( \frac{3\hat{f}_{Y|X}}{\left\{ \hat{f}_X \hat{F}_{Y|X}^{(0,2)} + 2\hat{f}'_X \hat{F}_{Y|X}^{(0,1)} \right\}^2} \right)^{1/6}.$$

- For two-sided inference with  $p \neq 1/2$  (and equivalent to above with  $p = 1/2$ ),

$$\hat{h} = n^{-1/3} \left( \frac{(2p-1)(\hat{B}_h/|\hat{B}_h|) + \sqrt{(2p-1)^2 + (4/3)/\hat{f}_{Y|X}}}{(2/3) \left| \hat{f}_X \hat{F}_{Y|X}^{(0,2)} + 2\hat{f}'_X \hat{F}_{Y|X}^{(0,1)} \right| / \hat{f}_{Y|X}} \right)^{1/3}.$$

Alternatively, the median-specific bandwidth may be used if  $u_h$  and  $u_\ell$  are chosen such that  $f_\beta(p; u_h) = f_\beta(p; u_\ell)$ , which requires relaxing the equal-tailed restriction of (2.2). The benefit is a simpler expression for the bandwidth; the costs are an additional (though fast) numerical search and loss of the equal-tailed property.

Regarding the signs of the two CPE terms in (2.7), we know that  $\text{CPE}_{\text{GK}} > 0$  (over-coverage). We can estimate the sign of  $\text{CPE}_{\text{Bias}}$  as the sign of  $B_h$  for lower one-sided inference and the opposite of the sign of  $B_h$  for upper one-sided inference. For two-sided inference, the sign of  $\text{CPE}_{\text{Bias}}$  depends on the bandwidth, and there always exists a bandwidth such that  $\text{CPE}_{\text{Bias}} < 0$  and cancels  $\text{CPE}_{\text{GK}}$ . In the one-sided case, if  $\text{CPE}_{\text{Bias}} < 0$ , then the optimal bandwidth causes the two CPE terms

to cancel out; if  $\text{CPE}_{\text{Bias}} > 0$ , the optimal bandwidth minimizes their sum. The only difference is an extra coefficient of  $[2d/(2b+d)]^{1/(b+3d/2)} = [2d/(d+4)]^{2/(4+3d)}$  from the first-order condition in the latter case, where the initial exponents of  $h$  come down when taking a derivative. If  $\text{CPE}_{\text{Bias}} < 0$ , then overall CPE is  $o(n^{-4/7})$ ; if  $\text{CPE}_{\text{Bias}} > 0$ , then overall CPE is  $O(n^{-4/7})$  and positive (over-coverage).

In the rest of this subsection, we provide some additional details on the two-sided case. For two-sided inference with  $p = 1/2$ , the  $B_h$  term becomes zero, as is clear in (2.13) below. Then  $\text{CPE}_{\text{Bias}} < 0$  since  $B_h^2 > 0$ ,  $f'_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(p) < 0$ , and  $f'_{\hat{Q}_{Y|C_h}^{I, u_h}}(p) > 0$ . The optimal  $h$  causes this to cancel with  $\text{CPE}_{\text{GK}} > 0$ . At a minimum, this reduces overall CPE to  $o(n^{-2/(2+d)})$ . From Lemma 2.2, if  $k_X \geq 2$  and  $k_Q \geq 3$ , then  $B_h$  consists of the leading  $h^2$  term plus  $o(h^3)$  remainder. Since  $\text{CPE}_{\text{GK}}$  consists of the leading  $N_n^{-1}$  term plus  $O(N_n^{-3/2} \log(N_n))$  remainder, the overall CPE (after cancellation, with  $h^* \asymp n^{-1/(2+d)}$ ) reduces to

$$\begin{aligned} \text{CPE} &= O(N_n^{-3/2} \log(N_n)) + o(N_n h^6) = O(N_n^{-3/2} \log(N_n)) \\ &= O((n(h^*)^d)^{-3/2} [\log(n) + d \log(h^*)]) \\ &= O((n^{2/(2+d)})^{-3/2} \log(n)) = O(n^{-3/(2+d)} \log(n)), \end{aligned}$$

which is nearly  $O(n^{-1})$  for  $d = 1$ .

By the convergence of our beta to a normal distribution, the product rule for derivatives, and the invariance of  $f_{Y|C_h}(F_{Y|C_h}^{-1}(p))$  to  $u$ ,

$$\begin{aligned} &f'_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) - f'_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) \\ &\quad \doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|C_h}(F_{Y|C_h}^{-1}(p)), \end{aligned}$$

and

$$\begin{aligned} &N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\ &= -(1/2)h^4 \left( \frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)} \right)^2 \\ &\quad \times \{ -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|X}(\xi_p; 0) \} \end{aligned}$$

leads to the plug-in bandwidth.

For two-sided inference with  $p \neq 1/2$ , the following does not cancel but is of smaller order than the  $O(N_n^{1/2})$  in the one-sided case. Using the result from Chapter 1 that  $u_h - p = z_{1-\alpha/2}\sqrt{p(1-p)/N_n} + O(N_n^{-1})$  and  $u_\ell - p = -z_{1-\alpha/2}\sqrt{p(1-p)/N_n} + O(N_n^{-1})$ , along with the normal approximation of the beta PDF in (2.11),

$$\begin{aligned}
& f_\beta(p; u_h) - f_\beta(p; u_\ell) \\
&= N_n^{1/2}\phi(z_{1-\alpha/2})\left([u_h(1-u_h)]^{-1/2} - [u_\ell(1-u_\ell)]^{-1/2}\right)\left[1 + O(N_n^{-1/2})\right] \\
&= N_n^{1/2}\phi(z_{1-\alpha/2})\left(\frac{2p-1}{2[p(1-p)]^{3/2}}[(u_h-p) - (u_\ell-p)] + O(N_n^{-1})\right) + O(1) \\
&= z_{1-\alpha/2}\phi(z_{1-\alpha/2})\frac{2p-1}{p(1-p)} + O(N_n^{-1/2}) + O(1) = O(1). \tag{2.12}
\end{aligned}$$

Thus our two CPE terms are of orders  $N_n^{-1} \asymp n^{-1}h^{-1}$  and  $h^2$ . This implies  $h^* \asymp n^{-1/3}$  and that CPE is  $O(n^{-2/3})$ . But then the  $B_h^2$  term is of order  $h^4N_n = h^5n = n^{-2/3}$ , so it must also be included. (The  $B_h^3$  term is of order  $h^6N_n^{3/2}$ , which is smaller.) Though the second term from the product rule derivative in the  $B_h^2$  term is not zero this time, it is smaller-order and thus omitted. As discussed in the appendix, in this case iteration would be required to compute the precise CPE-reducing bandwidth. Instead, we continue with the normal approximation, which still gives the optimal bandwidth and CPE rates in Theorem 2.3.

Due to the unknown  $O(1)$  term in (2.12) from the normal approximation error, we cannot pick  $h$  to precisely zero the dominant term of the overall CPE as before. Instead, we set the known portion of the overall CPE dominant term to zero. The sign of the known dominant term of  $\text{CPE}_{\text{Bias}}$  is determined by the sign of  $[B_h(2p-1) - B_h^2N_n]$ , as seen in (2.13). Since  $B_h^2N_n > 0$  always, if  $B_h(2p-1) < 0$ , then the known dominant term in  $\text{CPE}_{\text{Bias}}$  is negative irrespective of  $h$  (only the magnitude of  $B_h$  depends on  $h$ , not the sign). If  $B_h(2p-1) > 0$ , as shown in the appendix, it is still always possible to choose  $h$  such that the known dominant term in  $\text{CPE}_{\text{Bias}}$  is negative and cancels with  $\text{CPE}_{\text{GK}}$ . Since such a solution is always possible, we pick  $h$  to equate

$$-N_n^{-1}z_{1-\alpha/2}\frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)}\phi(z_{1-\alpha/2})$$

$$\begin{aligned}
& \doteq B_h \left[ f_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) - f_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) \right] \\
& \quad + (1/2)B_h^2 \left[ f'_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) - f'_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) \right] \\
& \doteq f_{Y|X}(\xi_p; 0) z_{1-\alpha/2} \phi(z_{1-\alpha/2}) [p(1-p)]^{-1} [B_h(2p-1) - B_h^2 N_n]. \quad (2.13)
\end{aligned}$$

Note that  $2p - 1 > 0$  is equivalent to  $p > 1/2$ , and that  $p = 1/2$  zeroes the  $B_h$  term. Continuing to solve for  $h$  leads to the plug-in bandwidth given.

Once  $\hat{h}$  is determined, for any of these cases,  $C_h$  can be constructed, and then the one-sample unconditional quantile inference method can be applied to  $\{Y_i : X_i \in C_h\}$ .

### 2.4.3 Conditional quantile treatment effects and other objects of interest

Unconditional IDEAL inference methods for multiple quantiles (joint inference), linear combinations of quantiles, and quantile treatment effects (on single quantiles or on linear combinations of quantiles) are detailed in Chapter 1. To be clear, for some chosen vector of quantiles  $(p_1, \dots, p_J)$ , chosen vector of weights  $(\psi_1, \dots, \psi_J)$  with non-zero elements, and observed binary treatment variable  $T_i$ , the objects of interest for these three methods are

$$\begin{aligned}
\text{Joint:} & \quad (Q_{Y|X}(p_1; x_0), \dots, Q_{Y|X}(p_J; x_0)), \\
\text{Linear combination:} & \quad \sum_{j=1}^J \psi_j Q_{Y|X}(p_j; x_0), \\
\text{Treatment effect:} & \quad \sum_{j=1}^J \psi_j [Q_{Y|X,T}(p_j; x_0, T=1) - Q_{Y|X,T}(p_j; x_0, T=0)].
\end{aligned}$$

The CPEs are, respectively,  $O(N_n^{-1})$ ,  $O(N_n^{-2/3})$ , and  $O(N_n^{-2/3})$ . Note that the treatment effect CPE is the same whether  $J = 1$  or  $J > 1$ . The tradeoff between these CPEs and the CPE from the bias will again determine the optimal bandwidth rate and overall CPE. For plug-in bandwidths, we suggest simply adjusting the plug-in bandwidths in §2.4.2, multiplying by the power of  $n$  necessary to achieve the proper rate.

For the conditional quantile treatment effect (CQTE), when  $J = 1$  the expression above is equivalent to that in MaCurdy et al. (2011, p. 547). The case of  $J > 1$  includes, for example, treatment effects on the interquartile range when  $(p_1, p_2) = (0.75, 0.25)$  and  $(\psi_1, \psi_2) = (1, -1)$ . Identification of a causal effect is obtained from the common assumption of conditional independence (a.k.a. unconfoundedness, a.k.a. strong ignorability), as in Assumption 1 of MaCurdy et al. (2011):  $Y_0, Y_1 \perp\!\!\!\perp T \mid X$ , where potential outcome  $Y_0$  is the outcome  $Y$  if an individual were not treated (regardless of treatment status in reality), and potential outcome  $Y_1$  is the outcome  $Y$  if an individual were treated (regardless of treatment status in reality). This assumption is only necessary if a causal interpretation of the CQTE is desired. For example, if  $T_i = 1$  if individual  $i$  is male and  $T_i = 0$  otherwise, and we observe wages  $Y$  and various covariates  $X$ , it is not valid to interpret the difference between the conditional median wage when  $T_i = 1$  and  $T_i = 0$  as a causal effect of being male. However, this difference may still be of interest, and an accurate inference method showing that the difference is large may motivate further empirical study.

For simplicity, we assume a common bandwidth for all quantiles  $p_j$ , though allowing for different control group ( $T = 0$ ) and treatment group ( $T = 1$ ) bandwidths for treatment effect inference. The procedure is the same as before: after  $h$  is determined, construct  $C_h$  and the effective sample  $\{Y_i : X_i \in C_h\}$ , and then run the unconditional IDEAL method on the effective sample. For treatment effect inference, there will be separate treatment and control  $h_T$  and  $h_C$  corresponding to  $C_{h_T}$  and  $C_{h_C}$ , leading to separate treatment and control effective samples. As stated in Corollary 2.4, the ratio of effective sample sizes is assumed to satisfy Assumption A1.3 of Chapter 1.

Since these methods involve a finite number of quantiles, the order of the CPE due to bias remains the same as before for one-sided inference. For two-sided inference, there is the same cancellation for the joint method as before, but not so with linear combinations or treatment effects. In the single quantile case, when examining the CPE from bias, the distribution of  $Y$  factored out, so the asymptotic symmetry (around  $p$ , up to smaller-order terms) of the beta



distributions corresponding to the upper and lower endpoints sufficed. Even in the case of the treatment effect at a single quantile, the PDF of the treatment effect CI endpoint evaluated at a single point depends on (via convolution) the distributions of the constituent treatment group and control group endpoints across their entire supports. Consequently, there may be cancellation under certain special types of symmetric distributions, but not in general.

Results on optimal bandwidth rates and overall CPE for various IDEAL conditional quantile inference methods are collected in the following corollary (proof in Appendix B.3).

**Corollary 2.4.** *Let Assumptions A2.1–A2.6 hold, and assume a common bandwidth is used at all quantiles  $p_j$ . For treatment effect inference, there may be a different bandwidth for  $T = 0$  than for  $T = 1$ , as long as  $\exists c$  s.t.  $\sqrt{N_{n,T=1}/N_{n,T=0}} = c + O(N_n^{-1/2})$ , where we define effective treatment sample size*

$$N_{n,T=1} \equiv \sum_{i=1}^n (1\{T_i = 1\} \times 1\{X_i \in C_{h_T}\})$$

and similarly for  $T = 0$ . Then, optimal bandwidth and CPE rates are as follows.

- *Multiple quantiles (joint inference): for one-sided inference,  $h^* \asymp n^{-3/(2b+3d)}$  and  $CPE = O(n^{-2b/(2b+3d)})$ ; for two-sided inference,  $h^* \asymp n^{-1/(b+d)}$  and  $CPE = O(n^{-b/(b+d)})$ .*
- *Linear combinations of quantiles, or treatment effects thereon, for both one- and two-sided inference:  $h^* \asymp n^{-7/(6b+7d)}$ ,  $CPE = O(n^{-4b/(6b+7d)})$ .*

*Remark.* The rates for joint inference over multiple quantiles are the same as for a single quantile. With a common bandwidth at all quantiles, using  $\hat{p} = \arg \min_{p_j} p_j(1 - p_j)$  in the single quantile plug-in bandwidth will be slightly conservative in finite samples but retain asymptotically exact coverage. Since the joint method uses  $\tilde{\alpha} \leq \alpha$  for the individual CIs at each  $p_j$ , using  $\alpha$  instead of  $\tilde{\alpha}$  for the one-sided plug-in bandwidth will also be slightly conservative in finite samples: note that the one-sided single quantile  $\hat{h}$  is decreasing in  $\alpha$  due to the  $z_{1-\alpha}$ , so picking larger  $\alpha$  will be conservative. For one-sided inference on linear

combinations of quantiles, we recommend multiplying the single quantile plug-in bandwidth by  $n$  to the power of  $4b/[(6b+7d)(2b+3d)]$  to get a bandwidth with the optimal rate:

$$\begin{aligned} n^{-3/(2b+3d)} n^{4b/[(6b+7d)(2b+3d)]} &= n^{[-3(6b+7d)]/[(2b+3d)(6b+7d)]+4b/[(6b+7d)(2b+3d)]} \\ &= n^{-7/(6b+7d)}. \end{aligned}$$

This time  $\tilde{\alpha} \geq \alpha$ , so we suggest plugging in the calibrated  $\tilde{\alpha}$  that would be used if the sample size were  $n$  rather than  $N_n$ , and again whichever  $p_j$  minimizes  $p_j(1-p_j)$ . For two-sided inference on linear combinations, similarly, we recommend multiplying the single quantile plug-in bandwidth by  $n^{-b/[(6b+7d)(b+d)]}$  to get a bandwidth with the optimal rate:

$$n^{-1/(b+d)} n^{-b/[(6b+7d)(b+d)]} = n^{(-6b-7d)/[(b+d)(6b+7d)]-b/[(6b+7d)(b+d)]} = n^{-7/(6b+7d)}.$$

For treatment effects, the adjustment is the same, but with separate bandwidths for the control ( $T = 0$ ) and treatment ( $T = 1$ ) samples. We again recommend choosing the  $p_j$  that minimizes  $p_j(1-p_j)$ , and for the one-sided case, the  $\tilde{\alpha}$  that would result from sample size  $n$ .

## 2.5 Simulation study

Code for the IDEAL inference functions in R is available on the author's website, and simulation code in R is available upon request.

### 2.5.1 Computation of plug-in bandwidth

For our plug-in bandwidth, we need to estimate five objects. Consistent estimators exist for all five. For clarity, we now explicitly write  $x_0$  as the point of interest, instead of taking  $x_0 = 0$ ; we also take  $d = 1$ ,  $b = 2$ , and focus on two-sided inference, both pointwise and joint.

For  $f_{Y|X}(\xi_p; x_0)$ , a kernel estimator such as `npcdens` from package `np` (Hayfield and Racine, 2008) can be used given an estimate  $\hat{\xi}_p$ , which can be computed using any nonparametric conditional quantile estimator. To estimate  $\hat{\xi}_p$ , we use

rq from package `quantreg` (Koenker, 2012) with a cubic B-spline basis generated by `bs` from package `splines` (R Core Team, 2012).

For  $f_X(x_0)$ , any kernel density estimator will suffice, such as `kde` from package `ks` (Duong, 2012). From the same package, `kdde` can be used to estimate the density derivative  $f'_X(x_0)$ . Both functions work for up to six-dimensional  $X$  data.

Alternatively, we could use the Gaussian plug-in assumption for  $f_X(x_0)$  and  $f'_X(x_0)$  directly in  $\hat{h}$ , instead of estimating them. For the density derivative, another popular option is the local polynomial approach as in Fan and Gijbels (1996, pp. 50–52).

For  $F_{Y|X}^{(0,1)}(\xi_p; x_0)$  and  $F_{Y|X}^{(0,2)}(\xi_p; x_0)$ , they enter a Taylor expansion of  $G(\cdot) \equiv F_{Y|X}(\xi_p; \cdot)$

$$G(x) = G(x_0) + G'(x_0)(x - x_0) + (1/2)G''(x_0)(x - x_0)^2 + (1/6)G'''(\tilde{x})(x - x_0)^3,$$

where  $\tilde{x}$  is between  $x$  and  $x_0$ .<sup>2</sup> Also,  $F_{Y|X}(\xi_p; X = x) = E(1\{Y_i \leq \xi_p\} | X = x)$ , a mean regression of  $1\{Y_i \leq \xi_p\}$  on  $X$ . Correspondingly, we use `lm` from package `stats` (R Core Team, 2012) to fit a cubic B-spline generated by `bs` from package `splines` (R Core Team, 2012) and compute the first two derivatives at a grid of points using `mybs` (Weisberg, 2012). The degrees of freedom are chosen to match the effective degrees of freedom automatically selected by penalized spline function `qsreg` from package `fields` (Furrer et al., 2012).

In finite samples, sampling error can make the plug-in bandwidth differ from the infeasible version, and in turn the approximation error from the Taylor expansion (around  $x_0$ ) used to calculate the bias can make even the infeasible bandwidth bigger or smaller than optimal. Since an excessively large bandwidth leads to under-coverage, we implement a “reality check” adjustment to our plug-in bandwidths. For points  $x_1 < x_2$ , the  $C_h$  window for  $x_2$  cannot extend to the left of the window for  $x_1$ , and likewise the  $x_1$  window cannot extend to the right of the  $x_2$  window. This essentially lets information about local quantile function derivatives be shared among the points of interest. For example, imagine that the quantile function is estimated to be very smooth (small derivatives) at  $x_2$ ,

---

<sup>2</sup>A quadratic would also give a consistent estimator since  $\tilde{x} \rightarrow x_0$  as  $n \rightarrow \infty$ , but may be worse in finite samples, depending also on local smoothness.

leading to a large bandwidth, but estimated to be highly variable (big derivatives) at  $x_1$ , leading to a small bandwidth. That additional information from  $x_1$  should cause the  $x_2$  window to shrink until its left edge matches that of the  $x_1$  window (and arguably farther), rather than letting  $x_2$  continue to blithely imagine that the function extends smoothly as far as the eye can see. We did not experiment with different (possibly data dependent) magnitudes or types of adjustment, opting simply for the intuitive version above, but more sophisticated ones may exist.

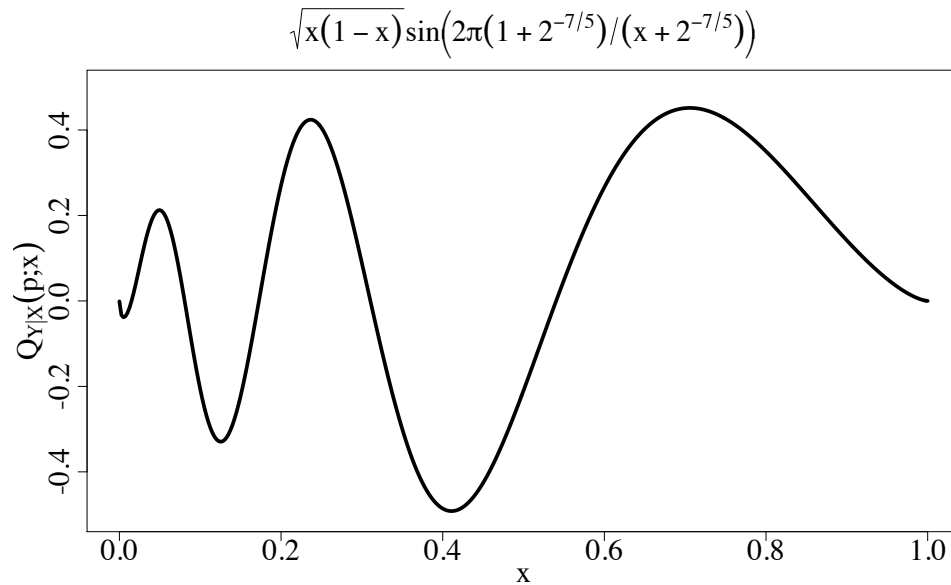
## 2.5.2 Results

For comparison, we show two approaches available in the popular `quantreg` package in R (Koenker, 2012). First, the function `rqss` (“regression quantile smoothing spline”) is used as on page 10 of its vignette, with the Schwarz (1978) information criterion (SIC) for model selection. Both pointwise and uniform confidence bands are generated with `plot.rqss`. Second, the function `rq` is used with a cubic B-spline generated by `bs`, again with SIC model selection. The function `predict.rq` with `type="percentile"` then generates pointwise confidence intervals using a bootstrap. Similar but consistently worse results were obtained from using the same approach but with `type="direct"` to use the analytic method, so only the bootstrap version is presented. Joint CIs may be generated using a Bonferroni approach.

A third approach was tried, using `npqreg` from package `np` (Hayfield and Racine, 2008) and `boot` and `boot.ci` from package `boot` (R Core Team, 2012), with both percentile and adjusted percentile bootstraps ("`perc`" and "`bca`" types). In preliminary simulations, with `bwmethod="normal-reference"` (rule of thumb), the computation time was still a factor of ten larger (with 999 bootstrap replications), and over-smoothing caused severe under-coverage. With likelihood-based cross-validation option `bwmethod="cv.ml"`, the computation time was prohibitive, running for over three hours on the first simulation replication alone (which was stopped before completion). Consequently, this approach is omitted from the results.

A bias-corrected version of our IDEAL method is also omitted. Compared

to the uncorrected IDEAL, it performed extremely similarly, usually the same or slightly worse. Median (over all simulation replications) confidence bands appeared marginally more centered around the true conditional quantile function, but the additional variance incurred from bias estimation likely caused the observed (slight) under-coverage.



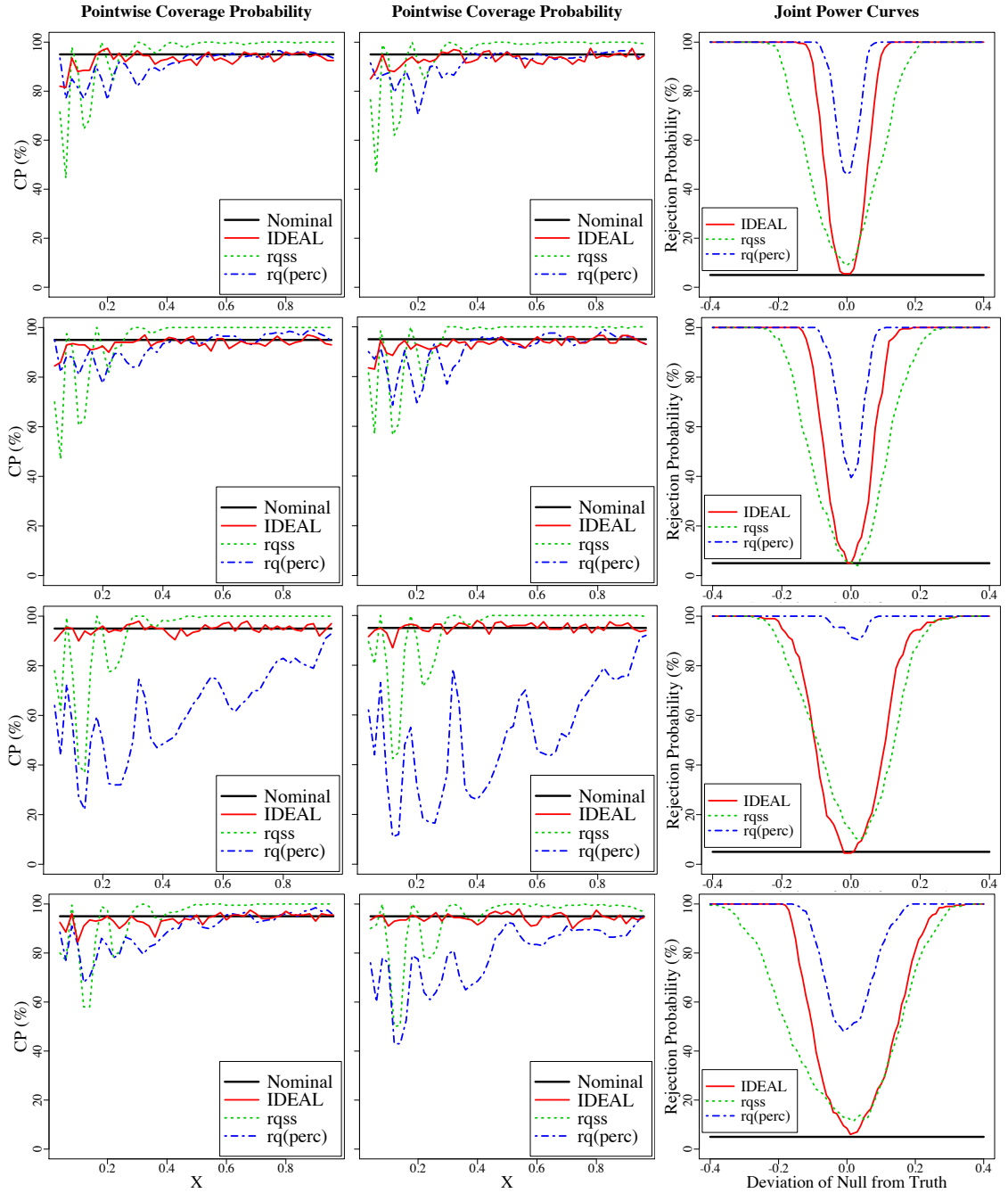
**Figure 2.2:** True conditional median function for simulations, taken from Koenker (2012):  $Q_{Y|X}(0.5; x) = \sqrt{x(1-x)} \sin(2\pi(1+2^{-7/5})/[x+2^{-7/5}])$ .

Our simulations repeat the simulation setup of the `rqss` vignette in Koenker (2012), which in turn was taken in part from Ruppert et al. (2003, §17.5.1). Parameters were set to  $n = 400$ ,  $p = 1/2$ ,  $d = 1$ , and  $\alpha = 0.05$ . Scalar  $X_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$ ,

$$Y_i = \sqrt{X_i(1-X_i)} \sin\left(\frac{2\pi(1+2^{-7/5})}{X_i+2^{-7/5}}\right) + \sigma(X_i)U_i,$$

where the  $U_i$  are iid Gaussian,  $t_3$ , Cauchy, or centered  $\chi_3^2$ , and  $\sigma(X) = 0.2$  or  $\sigma(X) = 0.2(1+X)$ . The conditional median function is shown in Figure 2.2.

With all eight DGPs (four error distributions, homoskedastic or heteroskedastic), our IDEAL method had the most consistent accuracy over all points on the conditional median function, as shown in the first two columns of Figure



**Figure 2.3:** Pointwise coverage probabilities by  $X$  (first two columns) and joint power curves (third column), for conditional median 95% confidence intervals,  $n = 400$ ,  $X_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$ ,  $Y_i = \sqrt{X_i(1 - X_i)} \sin[2\pi(1 + 2^{-7/5})/(X_i + 2^{-7/5})] + \sigma(X_i)U_i$ . Distributions of  $U_i$  are, top row to bottom row: standard normal,  $t_3$ , Cauchy, and centered  $\chi_3^2$ . Columns 1 & 3: homoskedastic,  $\sigma(x) = 0.2$ . Column 2: heteroskedastic,  $\sigma(x) = (0.2)(1 + x)$ .

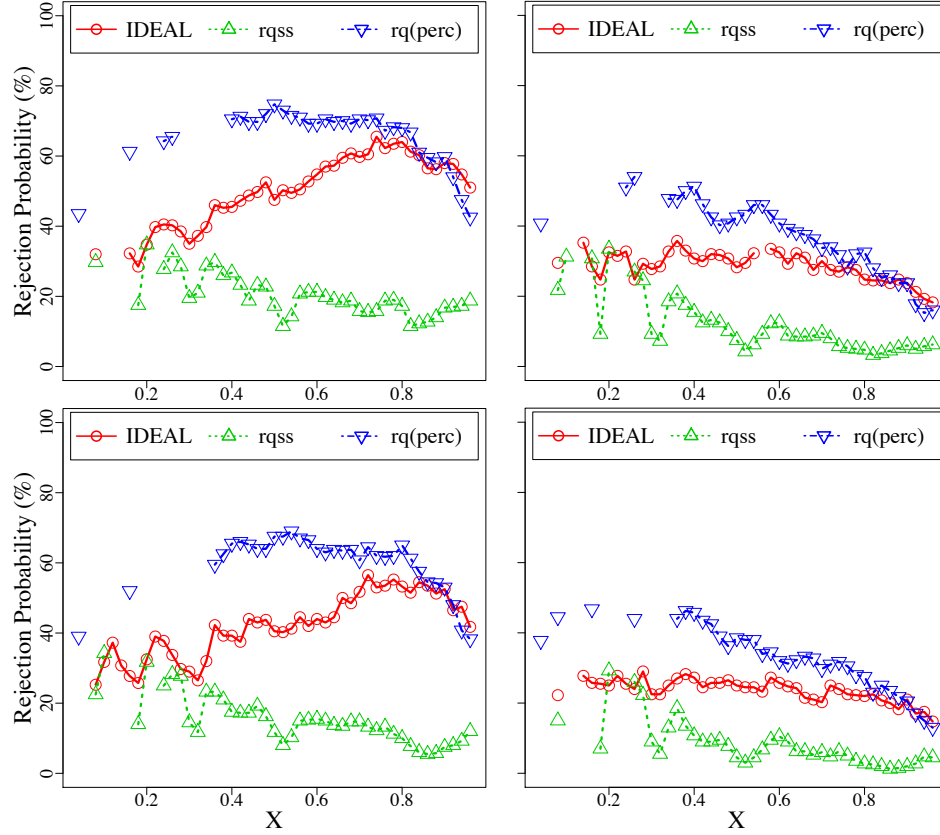
2.3. Coverage probability is near nominal for all  $x_0$ , all distributions, and in the presence or absence of heteroskedasticity.

In contrast, the other two methods are subject to under-coverage for a variety of reasons, as well as some over-coverage. Near  $X = 0$ , the true conditional median function varies rapidly, and then it smooths out as  $X$  increases toward one. As seen in the first two columns of Figure 2.3, the `rqss` confidence intervals are too narrow when the function varies more, and too wide when the function varies less. This pattern holds for all eight DGPs. The under-coverage (as low as 40–60% CP depending on the error distribution) is quite significant for  $x_0$  closer to zero, and over-coverage (as high as 100% CP for all distributions) occurs for most  $x_0 \geq 0.5$ . The `rq` approach avoids the over-coverage, but it can suffer even more severe under-coverage, as shown. This is particularly true with Cauchy errors (third row in figure), where there is under-coverage for all  $x_0$  and CP dips below 30% with homoskedasticity and even below 10% with heteroskedasticity.

The easiest way to construct IDEAL joint CIs is by the Bonferroni approach. For example, when  $\alpha = 0.05$  to give a 95% confidence level, if there are 47 points of interest  $x_0$ , pointwise CIs are constructed with  $\alpha/47$  instead of  $\alpha$ . Alternatively, instead of the Bonferroni  $\alpha/47$ , an adjusted value for  $\alpha$  can be backed out from the uniform confidence bands provided by `rqss`, which uses a Hotelling tube approach. This gives extremely similar results in our simulations, so it has been omitted for simplicity.

Joint power curves are given in the third column of Figure 2.3. The x-axis of the graphs indicates the deviation of the null hypothesis from the true curve; for example,  $-0.1$  refers to a test against a curve lying 0.1 below the true curve (at all  $X$ ), and zero means the null hypothesis is true. The IDEAL joint CP is again very close to nominal (since the test's size is close to  $\alpha$ ) under all four error distributions. Heteroskedastic versions were similar and thus omitted. The `rqss` method has CP relatively close to nominal (within 10%); it appears that the too-wide part nearly balances the too-narrow part in these examples. However, IDEAL has significantly better power. (Note that this is true without size-adjusting power, even though `rqss` is size-distorted and IDEAL is not.) The power advantage for

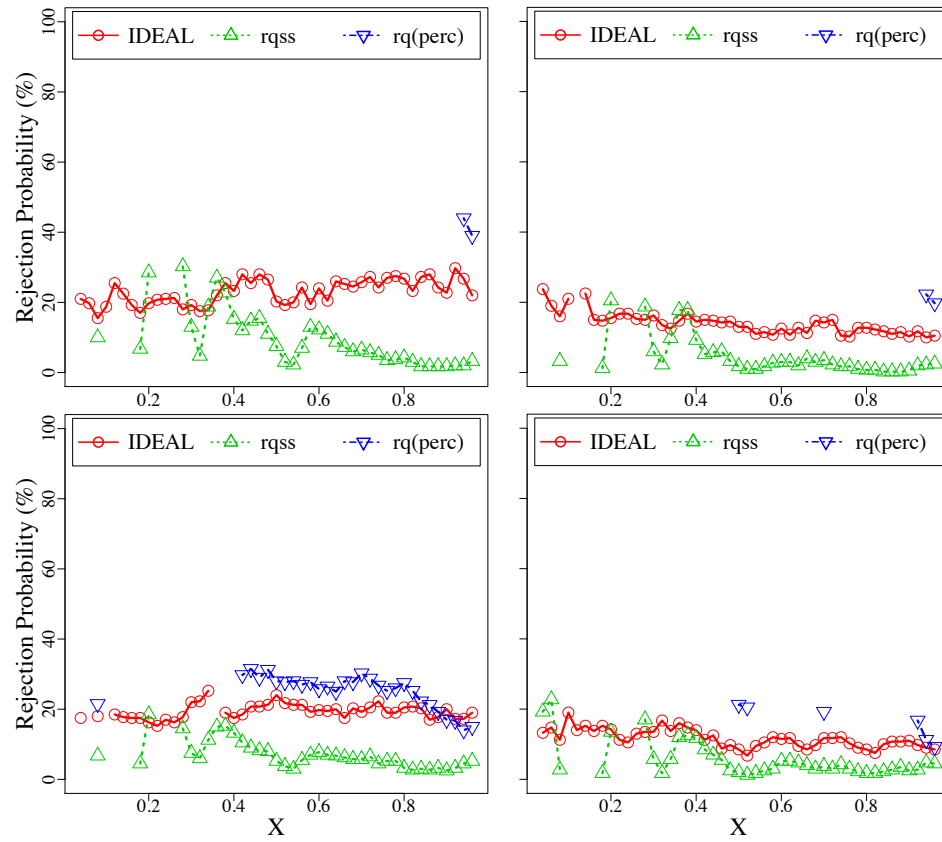
the asymmetric  $\chi_3^2$  distribution is much stronger for negative deviations, seemingly due to IDEAL's superior ability to adapt to skewed distributions. The rq method has significant under-coverage in all cases.



**Figure 2.4:** Pointwise power by  $X$ , against deviations of magnitude 0.1 (half negative, half positive), for conditional median 95% confidence intervals,  $n = 400$ ,  $X_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$ ,  $Y_i = \sqrt{X_i(1 - X_i)} \sin[2\pi(1 + 2^{-7/5})/(X_i + 2^{-7/5})] + \sigma(X_i)U_i$ . Points with CP below 90% not plotted. Top row:  $U_i \stackrel{iid}{\sim} N(0, 1)$ . Bottom row:  $U_i \stackrel{iid}{\sim} t_3$ . Left column: homoskedastic,  $\sigma(x) = 0.2$ . Right column: heteroskedastic,  $\sigma(x) = (0.2)(1 + x)$ .

Figures 2.4 and 2.5 show pointwise power, by  $X$ , against points differing from the true conditional median by  $\pm 0.1$ . This is calculated as the percentage of simulation replications where a method's CI excluded the point 0.1 below the true conditional median, averaged with the value for the point 0.1 above. To make the





**Figure 2.5:** Pointwise power by  $X$ ; same as Figure 2.4 but with Cauchy errors for top row and centered  $\chi_3^2$  for bottom row.

comparison fair, points with CP below 90% are omitted.

The `rq` method seems to have the best power (narrowest CIs) when it does succeed in controlling CP, though the difference with IDEAL is small for  $x_0$  in the upper quartile of  $X$ . With very smooth quantile functions, like a constant function (in the extreme), and with error distributions close to normal, the `rq` method would be preferred since it controls CP and has better power. Recall that `rq` is a parametric method, so this is essentially just saying that when the parametric model is (nearly) properly specified, inference will generally be more precise. However, barring a model selection method capable of reliably identifying situations where `rq` would be better (given that the quantile function and error distribution are both unknown), IDEAL provides a robust inference option with good power.

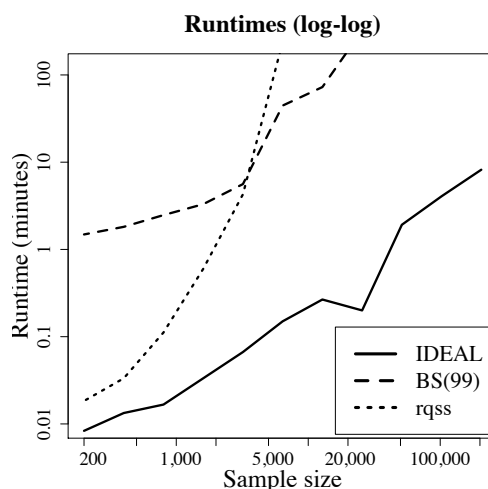
The IDEAL method is generally more powerful than `rqss`, as already hinted at in the pointwise CP graphs. Among the eight DGPs shown in Figures 2.4 and 2.5, IDEAL is more powerful by at least 10–20 percentage points for most  $x_0$  and most DGPs, and by over 40 percentage points in some cases.

Using the same setup but with errors following an exponential distribution (with or without heteroskedasticity), the results look similar to a mix of the standard normal and  $\chi_3^2$  results. One difference is that the joint CP for `rqss` is almost down to 80%, while the IDEAL joint CP is near 90%. Using the standard normal errors but instead looking at the conditional upper quartile, results were again similar other than worse CP for the joint intervals: `rqss` was again around 80%, as was the Bonferroni IDEAL, while the `rqss`/Hotelling-tube-aided IDEAL was near 90%. In this case, bias correction actually improved joint coverage by a few percentage points. Moving up to the even more difficult  $\Phi(1) \approx 0.84$ -quantile of a standard normal, the joint IDEAL coverage is the same, but `rqss` worsens to below 70% and suffers power loss against the  $-0.1$  and  $-0.2$  alternatives. The pointwise CP is actually best for `rq` in this case, though the joint CP for `rq` is below 60%.

Overall, the simulation results show IDEAL to be significantly more accurate than `rq` and `rqss`. IDEAL always has CP near the nominal level, for both

pointwise and joint CIs, even when both `rq` and `rqss` show severe under-coverage. IDEAL hypothesis tests also have better power than `rqss` in most cases, against both pointwise and joint alternatives.

### 2.5.3 Computation time



**Figure 2.6:** Computation time for different methods, as a function of sample size. DGP is the same as previous simulations, with homoskedastic Gaussian errors. Bootstrap times are for 99 bootstrap replications, estimated by multiplying the estimation time for `rqss` by 99; this gives a lower bound on the total bootstrap time.

The simplicity of the IDEAL method leads to significant computational advantages over existing methods. To demonstrate this, we ran simulations with the same DGP as before, with homoskedastic Gaussian errors, varying the sample size over a wide range. In addition to the IDEAL method, `rqss` (and `rqss.predict`) is used to generate pointwise CIs. We also calculate a lower bound time for a bootstrap method with 99 bootstrap replications by multiplying the `rqss` estimation time by 99. (Calculating the optimal smoothing parameter takes most of the time, rather than the spline fit itself.) This provides a strict lower bound because more replications (possibly 10 or 100 times more) may be needed to ensure bootstrap

accuracy, and additional computation is required beyond just the estimator.

Figure 2.6 shows computation times for these three methods as functions of sample size. For  $n = 200$ , 99 bootstrap replications take more than a minute to run, while IDEAL runs in less than a second. For  $n = 3200$ , both bootstrap and `rqss` take a few minutes to run, while IDEAL takes only a few seconds. For  $n = 6400$ , the `rqss` computation was stopped after two hours without having finished; 99 bootstrap replications would take at least 20 minutes, and 999 replications would take over three hours. Even with only 99 bootstrap replications,  $n = 25,600$  still takes over five hours. In contrast, with the much larger  $n = 204,800$  sample, IDEAL runs in under ten minutes. In addition to being orders of magnitude smaller at these sample sizes, the IDEAL runtime appears to scale roughly proportional to the sample size, whereas the other methods' runtimes appear to increase more rapidly. If these relationships continue to hold at larger sample sizes, the IDEAL computational advantage will grow proportionally bigger as the sample size increases.

#### 2.5.4 Conditional quantile treatment effect inference

The new IDEAL method for conditional quantile treatment effect (QTE) inference also performs quite well in simulations. The same conditional quantile function is considered here as in the single quantile case. Table 2.1 shows coverage probability and median interval length for the homoskedastic Gaussian scenario in §2.5.2, where instead of drawing one sample of  $n = 400$ , independent control and treatment samples of  $n = 400$  are drawn each replication. The data were generated with zero treatment effect; the results would be identical under pure location-shift treatment effects. Results for the upper quartile, median, and lower quartile are shown. The values of  $x_0$  can be compared to the conditional quantile function shown in Figure 2.2. The first two  $x_0$  are near the first two maxima, while the third is near the third minimum.

The pointwise coverage is close to the nominal 95% at all  $x_0$ . The lowest is 92.2%, and the highest is 96.4%. The joint coverage is also very good, between 93.2% and 95.2%. The shortest median interval lengths are for  $x_0 = 0.592$  and

**Table 2.1:** Coverage probability and median interval length for IDEAL confidence intervals for conditional quantile treatment effects;  $1 - \alpha = 0.95$ ,  $n = 400$  for both treatment and control samples, 500 replications.

$p$	$x_0$ value						
	0.04	0.224	0.408	0.592	0.776	0.96	Joint
	<i>Coverage Probability</i>						
0.75	0.938	0.924	0.938	0.940	0.948	0.960	0.952
0.50	0.926	0.962	0.964	0.952	0.944	0.952	0.946
0.25	0.948	0.960	0.948	0.930	0.922	0.924	0.932
	<i>Median Interval Length</i>						
0.75	0.366	0.364	0.325	0.281	0.268	0.322	
0.50	0.348	0.340	0.323	0.282	0.250	0.297	
0.25	0.373	0.373	0.337	0.316	0.258	0.341	

$x_0 = 0.776$ . These are the points where the function is smoothest, without running against the boundary like at  $x_0 = 0.96$ . This demonstrates that the method is adapting as desired: at the most difficult  $x_0$  values—the first two, at local maxima where the function is most variable—coverage probability is still close to nominal, and at “easier”  $x_0$  values, the intervals are more precise without sacrificing coverage accuracy.

## 2.6 Empirical application

This section contains an empirical application of the IDEAL confidence intervals for conditional quantiles. It may be replicated using the two R files and data files available at the author’s website.

Measurements of hemoglobin concentration in the bloodstream are commonly used to screen for anemia, which in turn may indicate iron deficiency, a nutritional deficiency affecting well over 1 in 4 people worldwide (Khusun et al., 1999). Examining how quantiles of the hemoglobin distribution vary with different conditioning variables is of more interest than how the mean varies because it is the lower quantiles that indicate different degrees of anemia. Wave 4 (2007) of the Indonesian Family Life Survey (IFLS) contains measurements of hemoglobin

concentration, among numerous other variables. The World Health Organization (WHO) has suggested threshold values<sup>3</sup> for “mild” anemia (acknowledging that even this level is a serious health concern), “moderate” anemia, and “severe” anemia, which seem valid at least for the (nonrandom) subsample of Indonesians examined in Khusun et al. (1999). These thresholds vary by age, sex, and pregnancy, so the raw IFLS values were scaled such that the threshold between non-anemia and mild anemia is 13 grams per deciliter (g/dl) for all individuals. Additionally, household per capita annual expenditure was computed from the raw IFLS data by summing the values in various categories of expenditure (scaled to annual for each) and dividing by household size; the natural log was then taken for more appropriate scaling in the graph. The highest level of education taken by the head of household was linked to the hemoglobin and expenditure data when available. The full Stata code for constructing the dataset starting from the raw IFLS data is available on the author’s website.<sup>4</sup> The sample analyzed is restricted to households’ oldest child (two to 15 years old). “High” education will refer to having attended at least some “senior high,” while “low” is no more than junior high.

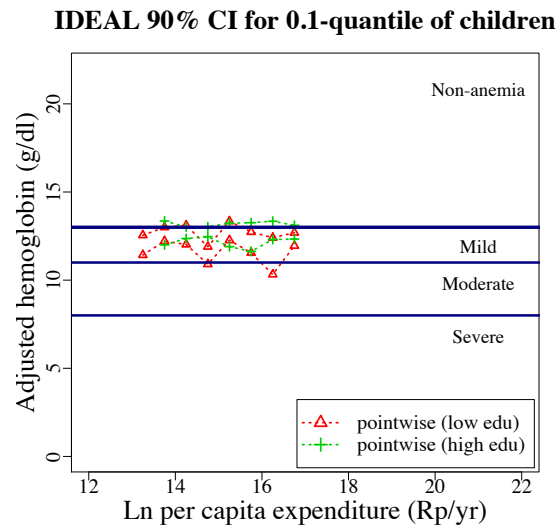
Looking at hemoglobin in the high-education and low-education groups, the 0.1-quantile is higher in the high-education group. The respective point estimates are 12.66g/dl and 12.41g/dl. Using the method in Chapter 1, a 90% IDEAL CI for the difference is (0.087, 0.424). However, this could be due to the head of household’s education itself, or correlation between education and income (and expenditure).

Figure 2.7 shows IDEAL CIs for the 0.1-quantile of hemoglobin conditional on both education and expenditure. There is some suggestion of a difference between the high- and low-education groups at some expenditure levels, but most of the CIs overlap. However, this is a conservative comparison. To be more accurate, the CIs should be narrower when examining the conditional QTE, for the same reason discussed in Chapter 1. The statistical significance of education remains a

---

<sup>3</sup><http://www.who.int/vmnis/indicators/haemoglobin.pdf>

<sup>4</sup>The relationship between hemoglobin and expenditure (and education) is also examined in Li et al. (2013), which inspired this example. Using their data, the subsequent analysis is similar but with somewhat larger differences when conditioning only on education. The version presented here is preferred since it may be replicated entirely with publicly available data and code.



**Figure 2.7:** IDEAL 90% confidence intervals for 0.1-quantile of adjusted hemoglobin concentration conditional on log per capita household expenditure, by education level of head of household, for children (ages 2–15). “High” education is defined as having attended at least some “senior high” education; “low” is no more than junior high.

cliffhanger until the proper conditional extension of the IDEAL QTE inference is developed.

## 2.7 Conclusion

We have provided a new method for inference on conditional quantiles, embedding the IDEAL theory of Chapter 1 into a local smoothing context. Under mild smoothness assumptions, with  $d$  continuous elements in the conditioning vector, the two-sided coverage probability error is  $O(n^{-2/(2+d)})$ . This is always better than conventional inference from asymptotic normality or bootstrap for  $d \leq 2$ , as well as for all  $d \geq 3$  unless at least four (or more, depending on  $d$ ) derivatives of the unknown function are correctly assumed and the corresponding local polynomial with hundreds or thousands of terms is fit. We also provide joint (over many values of  $X$ ) confidence intervals. Our feasible plug-in bandwidth translates the superior theoretical properties into practice; simulations show improvements in size control and power over popular existing techniques, and R code for our new method (with one continuous covariate and any number of discrete) is publicly available. The other methods in Chapter 1 are also translated to this conditional nonparametric model and provided in the R code, specifically joint inference across multiple conditional quantiles, inference on linear combinations of conditional quantiles, and inference on conditional quantile treatment effects (and effects on linear combinations).

This paper develops a framework for successfully pushing unconditional IDEAL results through to a conditional context. This framework should also readily accommodate a conditional version of Chapter 3 for cases where Hutson (1999) cannot be applied (smaller samples and/or quantiles closer to zero or one). Additional unconditional methods, such as for first-order stochastic dominance testing, can similarly be translated to conditional methods. By examining windows just above and below an  $X$  value of interest, in the spirit of numerical derivatives, it may be possible to obtain higher-order inference on marginal effects in this nonparametric conditional quantile model, too.



# Chapter 3

## One- and two-sample population quantile inference via fixed-smoothing asymptotics and Edgeworth expansion

### 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  $m, n \rightarrow \infty$ , the corresponding Studentized test statistic is asymptotically  $N(0, 1)$ . Holding  $m$  fixed asymptotically yields a non-standard 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 errors 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 inference, such as testing for median equality of treatment and control groups.

### 3.1 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), for one- and two-sample setups. Median inference is a special case. 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.

I develop a 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,<sup>1</sup> 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.

This work builds partly on Goh (2004), who suggests using fixed- $m$  asymptotics to improve inference on the Studentized quantile. Goh (2004) uses simulated critical values to examine the performance of existing  $m$  suggestions, which are tailored to standard normal critical values. Here, a simple and accurate analytic approximation is given for the critical values. Using this approximation, the new procedure for selecting  $m$  is tailored to the fixed- $m$  distribution. I also provide the theoretical justification for the accuracy of the fixed- $m$  distribution, which complements the simulations in Goh (2004).

The two key results here are a nonstandard fixed- $m$  asymptotic distribution and a higher-order Edgeworth expansion. For a scalar location model, Siddiqui (1960) gives the fixed- $m$  result, and Hall and Sheather (1988, hereafter cited as

---

<sup>1</sup>Inverting the level- $\alpha$  tests proposed here yields level- $\alpha$  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.

“HS88”) give a special case of the Edgeworth expansion in Theorem 3.1 below. The Edgeworth expansion is more accurate than a standard normal since it contains higher-order terms that otherwise end up in the remainder, so its remainder becomes of smaller order. There are intuitive reasons why the fixed- $m$  approximation is also more accurate than the standard normal, but we show this by rigorous theoretical arguments.

In the standard first-order asymptotics, both  $m \rightarrow \infty$  and  $n \rightarrow \infty$  (and  $m/n \rightarrow 0$ ); since  $m \rightarrow \infty$ , this may be called “large- $m$ ” asymptotics. In contrast, fixed- $m$  asymptotics only approximates  $n \rightarrow \infty$  while fixing  $m$  at its actual finite sample value. Fixed- $m$  is an instance of “fixed-smoothing” asymptotics in the sense that this variance does not go to zero in the limit as it does in “increasing-smoothing.” It turns out that the fixed- $m$  asymptotics includes the high-order Edgeworth term capturing the variance of the quantile spacing. Consequently, the fixed- $m$  distribution is higher-order accurate under the conventional large- $m$  asymptotics, while the standard normal distribution is only first-order accurate. Under the fixed- $m$  asymptotics, the standard normal is not even first-order accurate. From the theoretical view, the fixed- $m$  distribution is clearly more accurate than the standard normal, irrespective of the asymptotic framework.

Not only is accuracy gained with the Edgeworth and fixed- $m$  distributions, but they are sensitive to  $m$ . By reflecting the effect of  $m$ , they allow us to determine the best choice of  $m$ . With the fixed- $m$  and Edgeworth results, I construct a test dependent on  $m$  using the fixed- $m$  critical values, and then evaluate the type I and type II error of the test using the more accurate Edgeworth expansion. Then I can optimally select  $m$  to minimize type II error subject to control of type I error. Approximating (instead of simulating) the fixed- $m$  critical values reveals the Edgeworth/fixed- $m$  connection above, and it also makes the test easier to implement in practice.

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 (2010a,b, 2011), and Sun and Kaplan (2011). In particular, Sun et al. (2008) show that their testing-optimal bandwidth is of a different order than the MSE-optimal bandwidth, similar to the result in HS88 developed further here; that idea runs throughout the sequence. Sun (2011) generalizes the result from Sun et al. (2008) that shows fixed-smoothing (fixed- $b$ ) asymptotics to be a higher-order refinement of first-order increasing-smoothing (small- $b$ ) asymptotics, analogous to the results for fixed- $m$  asymptotics below.

The critical value correction here provides size control robust to incorrectly chosen  $m$ , whereas in HS88 the choice of  $m$  is critical to controlling size (or not). 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; in simulations, it can be significantly better than HS88. HS88 do not provide a separate result for the two-sample case. Monte Carlo simulations show that the new method controls size better than the common Gaussian plug-in version of HS88 and various bootstrap methods, while maintaining competitive power. The method of Hutson (1999), following a fractional order statistic approach, usually has better properties but is not always computable; a two-sample analog with similarly good performance is presented in Chapter 1. For two-sample inference, percentile- $t$  bootstrap can perform well, though the new method is good and open to a particular refinement that would increase power; more research in both approaches is needed.

Section 3.2 presents the model and hypothesis testing problem. Section 3.3 gives a nonstandard asymptotic result and corresponding corrected critical value, while Section 3.4 gives Edgeworth expansions of the standard asymptotic limiting distribution. Using these, a method for selecting smoothing parameter  $m$  is given in Section 3.5, which is followed by a simulation study and conclusion. Results for the two-sample extension are provided in parallel. More details and discussion are available in the working paper version; full technical proofs and calculations are in the supplemental appendices; all are available on the author's website.

### 3.2 Quantile estimation and hypothesis testing

For simplicity and intuition, consider an iid sample of continuous random variable  $X$ , whose  $p$ -quantile is  $\xi_p$ . The estimator  $\hat{\xi}_p$  used in SBG is an order statistic. The  $r$ th order statistic for a sample of  $n$  values is defined as the  $r$ th smallest value in the sample and written as  $X_{n,r}$ , 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 (greatest integer not exceeding  $np$ ). 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  $\xi_p$ . Consequently,  $\xi_p$  is the unique  $p$ -quantile such that  $F(\xi_p) = p$ .

The standard asymptotic result (for the vector form, see Mosteller, 1946; Siddiqui, 1960, also gives the following scalar result) for a central<sup>2</sup> quantile estimator is

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

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

$$\frac{\sqrt{n}(X_{n,r} - \xi_p)}{\sqrt{p(1-p)} \left[ \widehat{1/f(\xi_p)} \right]} \xrightarrow{d} N(0, 1). \quad (3.1)$$

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

$$\widehat{1/f(\xi_p)} = S_{m,n} \equiv \frac{n}{2m}(X_{n,r+m} - X_{n,r-m}) \quad (3.2)$$

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

For two-sided inference on  $\xi_p$ , I consider the parameterization  $\xi_p = \beta - \gamma/\sqrt{n}$ . The null and alternative hypotheses are  $H_0 : \xi_p = \beta$  and  $H_1 : \xi_p \neq \beta$ ,

<sup>2</sup>“Central” means that in the limit,  $r/n \rightarrow p \in (0, 1)$  as  $n \rightarrow \infty$ ; i.e.,  $r \rightarrow \infty$  and is some fraction of the sample size  $n$ . In contrast, “intermediate” would take  $r \rightarrow \infty$  but  $r/n \rightarrow 0$  (or  $r/n \rightarrow 1$ ,  $n - r \rightarrow \infty$ ); “extreme” would fix  $r < \infty$  or  $n - r < \infty$ .

respectively. When  $\gamma = 0$  the null is true. The test statistic examined in this paper is

$$T_{m,n} \equiv \frac{\sqrt{n}(X_{n,r} - \beta)}{S_{m,n}\sqrt{p(1-p)}} \quad (3.3)$$

and will be called the SBG test statistic due to its use of (3.2). From (3.1),  $T_{m,n}$  is asymptotically standard normal when  $\gamma = 0$ . A corresponding hypothesis test would then compare  $T_{m,n}$  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  $\alpha$ . The marginal pdfs are  $f_X(\cdot)$  and  $f_Y(\cdot)$ . Interest is in testing if  $\xi_{p,x} = \xi_{p,y}$ . Under the null hypothesis  $H_0 : \xi_{px} = \xi_{py} = \xi_p$ , the first-order asymptotic result is

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

using the fact that the variance of the sum (or difference) of two independent normals is the sum of the variances. The pivot for the two-sample case is then

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

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

$$\tilde{T}_{m,n} \equiv \frac{\sqrt{n}(X_{n,r} - Y_{n,r})}{\sqrt{[n/(2m)]^2(X_{n,r+m} - X_{n,r-m})^2 + [n/(2m)]^2(Y_{n,r+m} - Y_{n,r-m})^2}\sqrt{p(1-p)}}.$$

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 is easily extended to the case of unequal sample sizes  $n_x \neq n_y$ , starting with  $\sqrt{n_x}(X_{n_x,r} - \xi_p) - \sqrt{n_x/n_y}\sqrt{n_y}(Y_{n_y,r} - \xi_p)$  and assuming  $n_x/n_y$  is constant asymptotically, but this is omitted for clarity.

### 3.3 Fixed- $m$ asymptotics and corrected critical value

The standard asymptotic result with  $m \rightarrow \infty$  as  $n \rightarrow \infty$  is a standard normal distribution for the SBG test statistic  $T_{m,n}$  under the null hypothesis when  $\gamma = 0$ . Since in reality  $m < \infty$  for any given finite sample test, holding  $m$  fixed as  $n \rightarrow \infty$  may give a more accurate asymptotic approximation of the true finite sample test statistic distribution. With  $m \rightarrow \infty$ , there is increasing smoothing and  $\text{Var}(S_{m,n}) = O(1/m) \rightarrow 0$ ; with  $m$  fixed, we have fixed-smoothing asymptotics since that variance does not disappear. The fixed- $m$  distribution is given below, along with a simple formula for critical values that doesn't require simulation.

#### 3.3.1 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.<sup>3</sup> If  $\gamma = 0$ ,

$$T_{m,n} \xrightarrow{d} \mathcal{Z}/\mathcal{V}_{4m} \equiv T_{m,\infty} \quad \text{as } n \rightarrow \infty, m \text{ fixed}, \quad (3.4)$$

with  $\mathcal{Z} \sim N(0, 1)$ ,  $\mathcal{V}_{4m} \sim \chi_{4m}^2/(4m)$ ,  $\mathcal{Z} \perp \mathcal{V}_{4m}$ , and  $S_{m,n}$  and  $T_{m,n}$  as in (3.2) and (3.3).

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  $m \rightarrow \infty$ . When using an estimated variance in a finite sample  $t$ -statistic, the more accurate  $t$ -approximation has fatter tails than a standard normal, and it is given by  $\mathcal{Z}/\sqrt{\mathcal{V}_v}$ , where  $\mathcal{Z} \sim N(0, 1)$ ,  $\mathcal{V}_v \sim \chi_v^2/v$  with  $v$  the degrees of freedom, and  $\mathcal{Z} \perp \mathcal{V}_v$ . Similarly, when an estimated variance is used in the SBG test statistic (the random

---

<sup>3</sup>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).

$S_{m,n}$  instead of constant  $1/f(\xi_p)$  in the denominator), the result is the distribution in (3.4) above with asymptotically independent numerator and denominator. The fixed- $m$  distribution reflects this uncertainty in the variance estimator that is lost under the standard asymptotics.

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

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

and  $\mathcal{Z}$ ,  $\mathcal{V}_{4m,1}$ , and  $\mathcal{V}_{4m,2}$  are mutually independent. The derivation is in the full appendix online. The strategy is the same, using results from Siddiqui (1960) for asymptotic independence and distributions for each component.

Unlike the one-sample case,  $\tilde{T}_{m,\infty}$  is not pivotal. We can either estimate the density derivative  $f_X(\xi_p)/f_Y(\xi_p)$  or consider the upper bound for critical values. This is the same nuisance parameter faced by the two-sample order statistic method in Chapter 1. Note that  $\epsilon$  is a weighted average of  $\mathcal{V}_{4m,1}^2 - 1$  and  $\mathcal{V}_{4m,2}^2 - 1$ , where the weights sum to one and  $\mathcal{V}_{4m,1} \perp \mathcal{V}_{4m,2}$ . To see the effect of the weights, consider any  $W_1$  and  $W_2$  with  $\text{Cov}(W_1, W_2) = 0$ ,  $\text{Var}(W_1) = \text{Var}(W_2) = \sigma_W^2$ , and  $\lambda \in [0, 1]$ . Then

$$\begin{aligned} \text{Var}(\lambda W_1 + (1 - \lambda)W_2) &= \lambda^2 \text{Var}(W_1) + (1 - \lambda)^2 \text{Var}(W_2) + 2\lambda(1 - \lambda)\text{Cov}(W_1, W_2) \\ &= \sigma_W^2[\lambda^2 + (1 - \lambda)^2]. \end{aligned}$$

This is maximized by  $\lambda = 1$  or  $1 - \lambda = 1$ , giving  $\sigma_W^2$ . The minimum has first order condition  $0 = 2\lambda + 2(1 - \lambda)(-1)$ , yielding  $\lambda = 1/2 = 1 - \lambda$  and  $\text{Var}(\lambda W_1 + (1 - \lambda)W_2) = \sigma_W^2/2$ . This means the variance of  $\epsilon$  (and  $\mathcal{U}$  and  $\tilde{T}_{m,\infty}$ ) is smallest when the weights are each  $1/2$ , which is when  $f_X(\xi_p) = f_Y(\xi_p)$ . When the weights are zero and one, which is when the variance of  $\epsilon$  is largest, we get the special case of testing against a constant and  $\tilde{T}_{m,\infty} = T_{m,\infty}$ . Thus, critical values from the one-sample case provide conservative inference in the two-sample case.



### 3.3.2 Corrected critical value

An approximation of the fixed- $m$  cdf around the standard normal cdf will lead to critical values based on the standard normal distribution. The standard normal cdf is  $\Phi(\cdot)$ , the standard normal pdf is  $\phi(\cdot) = \Phi'(\cdot)$ , and the first derivative  $\phi'(z) = -z\phi(z)$ . The first three central moments of the  $\chi_{4m}^2$  distribution are  $4m$ ,  $8m$ , and  $32m$ , respectively. The general approach here is to use the independence of  $\mathcal{Z}$  and  $\mathcal{V}_{4m}$  to rewrite the cdf in terms of  $\Phi$ , and then to expand around  $E(\mathcal{V}_{4m}) = 1$ . Using (3.4), for a critical value  $z$ ,

$$\begin{aligned}
 P(T_{m,\infty} < z) &= P(\mathcal{Z}/\mathcal{V}_{4m} < z) = E[\Phi(z\mathcal{V}_{4m})] = E[\Phi(z + z(\mathcal{V}_{4m} - 1))] \\
 &= E\left[\Phi(z) + \Phi'(z)z(\mathcal{V}_{4m} - 1) + (1/2)\Phi''(z)[z(\mathcal{V}_{4m} - 1)]^2 + O(m^{-2})\right] \\
 &= \Phi(z) - \frac{z^3\phi(z)}{2} E\left[\left(\frac{\chi_{4m}^2 - 4m}{4m}\right)^2\right] + O(m^{-2}) \\
 &= \Phi(z) - \frac{z^3\phi(z)}{4m} + O(m^{-2}). \tag{3.5}
 \end{aligned}$$

Note that (3.5) 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  $\infty$  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  $m \rightarrow \infty$ , which the selected  $m$  does as  $n \rightarrow \infty$  (see Section 3.5.3). 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  $m \rightarrow \infty$  via Pólya's Theorem (e.g., DasGupta, 2008, Thm. 1.3(b)). Appendix C.1 shows the accuracy of (3.5) for  $m > 2$ ; for  $m \leq 2$ , simulated critical values can be used, as in the provided code.

To find the value  $z$  that makes the nonstandard cdf above take probability  $1 - \alpha$  (for rejection probability  $\alpha$  under the null for an upper one-sided test), I set  $\Phi(z) - z^3\phi(z)/(4m) = 1 - \alpha$  and solve for  $z$ . If  $z$  is some deviation of order  $m^{-1}$  around the value  $z_{1-\alpha}$  such that  $\Phi(z_{1-\alpha}) = 1 - \alpha$  for the standard normal cdf, then solving for  $c$  in  $z = z_{1-\alpha} + c/m$  gives

$$c = z_{1-\alpha}^3/4 + O(m^{-1}), \tag{3.6}$$

and thus the upper one-sided test's corrected critical value is

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

For a symmetric two-sided test, since the additional term in (3.5) is an odd function of  $z$ , the critical value is the same but with  $z_{1-\alpha/2}$ , yielding

$$z = z_{\alpha,m} + O(m^{-2}), \quad z_{\alpha,m} \equiv z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3}{4m}. \quad (3.7)$$

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 C.1,

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

The rest of this paper uses (3.7); the results in Section 3.5 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 Two-sample Appendix C), the fixed- $m$  distribution can be approximated by

$$\begin{aligned} P(\tilde{T}_{m,\infty} < z) &= \Phi(z) + \phi(z)zE[\mathcal{U} - 1] + (1/2)(-z\phi(z))z^2E[(\mathcal{U} - 1)^2] \\ &\quad + O(E[(\mathcal{U} - 1)^3]) \\ &= \Phi(z) - \frac{\phi(z)}{4m} \left[ z^3 \frac{[f_X(\xi_p)]^{-4} + [f_Y(\xi_p)]^{-4}}{\tilde{S}_0^4} - 2z \frac{[f_X(\xi_p)]^{-2}[f_Y(\xi_p)]^{-2}}{\tilde{S}_0^4} \right] \\ &\quad + O(m^{-2}) \\ &= \Phi(z) - \frac{\phi(z)}{4m} \left[ z^3 \frac{1 + \delta^4}{(1 + \delta^2)^2} - 2z \frac{\delta^2}{(1 + \delta^2)^2} \right] + O(m^{-2}), \\ \tilde{S}_0 &\equiv ([f_X(F_X^{-1}(p))]^{-2} + [f_Y(F_Y^{-1}(p))]^{-2})^{1/2}, \quad \delta \equiv f_X(\xi_p)/f_Y(\xi_p). \end{aligned} \quad (3.8)$$

The corresponding critical value is

$$\begin{aligned} z &= \tilde{z}_{\alpha,m} + O(m^{-2}) \leq z_{\alpha,m} + O(m^{-2}), \\ \tilde{z}_{\alpha,m} &= z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3([f_X(\xi_p)]^{-4} + [f_Y(\xi_p)]^{-4}) - 2z_{1-\alpha/2}[f_X(\xi_p)]^{-2}[f_Y(\xi_p)]^{-2}}{4m\tilde{S}_0^4} \\ &= z_{1-\alpha/2} + \frac{z_{1-\alpha/2}^3(1 + \delta^4) - 2z_{1-\alpha/2}\delta^2}{4m(1 + \delta^2)^2}. \end{aligned} \quad (3.9)$$

When the pdf of  $Y$  collapses toward a constant,  $[f_Y(\xi_p)]^{-1} \rightarrow 0$  and so  $\delta \rightarrow 0$ . Consequently,  $P(\tilde{T}_{m,\infty} < z)$  reduces to (3.5), 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 3.3.1.

## 3.4 Edgeworth expansion

HS88 give the Edgeworth expansion for the asymptotic distribution of  $T_{m,n}$  when  $\gamma = 0$ . To calculate the type II error of a test using the fixed- $m$  critical values (Section 3.5.2), the case  $\gamma \neq 0$  is needed. Section 3.5.2 shows how to apply this result.

### 3.4.1 One-sample case

The result here includes the result of HS88 as a special case, and indeed the results match<sup>4</sup> when  $\gamma = 0$ . The  $u_i$  functions are indexed by  $\gamma$ , so  $u_{1,0}$  is the same as  $u_1$  from HS88 (page 385), and similarly for  $u_{2,0}$  and  $u_{3,0}$ .

**Theorem 3.1.** *Assume  $f(\xi_p) > 0$  and that, in a neighborhood of  $\xi_p$ ,  $f''$  exists and satisfies a Lipschitz condition, i.e. for some  $\epsilon > 0$  and all  $x, y$  sufficiently close to  $\xi_p$ ,  $|f''(x) - f''(y)| \leq \text{constant}|x - y|^\epsilon$ . Suppose  $m = m(n) \rightarrow \infty$  as  $n \rightarrow \infty$ , 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 defining functions*

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

<sup>4</sup>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.

it follows that

$$\begin{aligned} & \sup_{-\infty < z < \infty} \left| P(T_{m,n} < z) - [\Phi(z + C) + n^{-1/2}u_{1,\gamma}(z + C)\phi(z + C) \right. \\ & \quad \left. + m^{-1}u_{2,\gamma}(z + C)\phi(z + C) + (m/n)^2u_{3,\gamma}(z + C)\phi(z + C)] \right| \\ & = o_p[m^{-1} + (m/n)^2]. \end{aligned} \quad (3.10)$$

A sketch of the proof is given in Appendix C.2. The assumptions are the same as in HS88.

For  $\gamma = 0$ , the term  $m^{-1}u_{2,0}(z)\phi(z) = -z^3\phi(z)/(4m)$  is identical to the term in the fixed- $m$  distribution in (3.5). Thus under the null, the fixed- $m$  distribution captures the high-order Edgeworth term associated with the variance of  $S_{m,n}$ . In other words, the fixed- $m$  distribution is high-order accurate under the conventional large- $m$  asymptotics, while the standard normal distribution is only first-order accurate. Since the fixed- $m$  distribution is also more accurate under fixed- $m$  asymptotics, where the standard normal is not even first-order accurate, theory strongly indicates that fixed- $m$  critical values are more accurate, which is born out in simulations here and in Goh (2004).

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  $S_0 \equiv 1/f(\xi_p)$ ,

$$\begin{aligned} T_{m,n} &= \frac{\sqrt{n}(X_{n,r} - \xi_p) - \gamma}{S_{m,n}\sqrt{p(1-p)}} \\ &= \frac{\sqrt{n}(X_{n,r} - \xi_p)}{S_{m,n}\sqrt{p(1-p)}} - \frac{\gamma}{\sqrt{p(1-p)}S_0} \left( \frac{S_0}{S_{m,n}} + 1 - 1 \right), \\ P(T_{m,n} < z) &= P\left( \frac{\sqrt{n}(X_{n,r} - \xi_p)}{S_{m,n}\sqrt{p(1-p)}} - \frac{\gamma}{\sqrt{p(1-p)}S_0} \left( \frac{S_0}{S_{m,n}} + 1 - 1 \right) < z \right) \\ &= P\left( \frac{\sqrt{n}(X_{n,r} - \xi_p)}{S_{m,n}\sqrt{p(1-p)}} - \frac{\gamma}{\sqrt{p(1-p)}S_0} \left( \frac{S_0}{S_{m,n}} - 1 \right) < z + C \right) \\ &= P\left( \frac{\sqrt{n}(X_{n,r} - \xi_p) + \gamma(S_{m,n}/S_0 - 1)}{S_{m,n}\sqrt{p(1-p)}} < z + C \right). \end{aligned} \quad (3.11)$$

If the true  $S_0$  were known and used in  $T_{m,n}$  instead of its estimator  $S_{m,n}$ , this would be simply the distribution from HS88 with a shift of the critical value by  $C \equiv \gamma/[S_0\sqrt{p(1-p)}]$ , which is  $\gamma$  normalized by the true (hypothetically known)

variance. But  $S_{m,n}$  is random, so the HS88 expansion is insufficient and Theorem 3.1 is needed.

### 3.4.2 Two-sample case

The strategy and results are similar; the full proof may be found in the supplemental Two-sample Appendix.

**Theorem 3.2.** *Let  $X_i \stackrel{iid}{\sim} F_X$  and  $Y_i \stackrel{iid}{\sim} F_Y$ , and  $X_i \perp\!\!\!\perp Y_j \forall i, j$ . With assumptions in Theorem 3.1 applied to both  $F_X$  and  $F_Y$ , and letting  $\tilde{C} \equiv \gamma \tilde{S}_0^{-1} / \sqrt{p(1-p)}$ , define functions*

$$\begin{aligned} \tilde{u}_{1,\gamma}(z) &\equiv \frac{1}{6} \frac{1+p}{\sqrt{p(1-p)}} \left( \frac{g_x^3 - g_y^3}{\tilde{S}_0^3} \right) (z^2 - 1) \\ &\quad + [(a_2 g_x^2 - a'_2 g_y^2)(1-p) + (a_1 g_x^2 - a'_1 g_y^2)] [p/(1-p)]^{1/2} (2p^2 \tilde{S}_0^3)^{-1} (z^2) \\ &\quad - [2(a_2 g_x^2 - a'_2 g_y^2) + (a_1 g_x^2 - a'_1 g_y^2)/(1-p)] (2p \tilde{S}_0^3)^{-1} C [(1-p)/p]^{1/2} (z) \\ &\quad + [(a_2 g_x^2 - a'_2 g_y^2) - \tilde{S}_0^2 (a_2 - a'_2)] [p(1-p)]^{1/2} (2p^2 \tilde{S}_0^3)^{-1} \\ &\quad - ((\lfloor np \rfloor + 1 - np) - 1 + \frac{1}{2}(1-p)) (g_x - g_y) \left[ \tilde{S}_0 \sqrt{p(1-p)} \right]^{-1}, \\ \tilde{u}_{2,\gamma}(z) &\equiv -\frac{1}{4} \left( \frac{g_x^4 + g_y^4}{\tilde{S}_0^4} \right) z^3 + \frac{1}{2} \frac{g_x^2 g_y^2}{\tilde{S}_0^4} z - \frac{1}{2} \frac{g_x^2 g_y^2}{\tilde{S}_0^4} \tilde{C} + \frac{1}{4} \left( \frac{g_x^4 + g_y^4}{\tilde{S}_0^4} \right) (2\tilde{C} z^2 - \tilde{C}^2 z), \\ \tilde{u}_{3,\gamma}(z) &\equiv \frac{g_x g_x'' + g_y g_y''}{6\tilde{S}_0^2} (z - \tilde{C}), \end{aligned}$$

$$g_X(\cdot) \equiv 1/f_X(F_X^{-1}(\cdot)), \quad g_x \equiv g_X(p), \quad g_x'' \equiv g_X''(p),$$

$$H_X(x) \equiv F_X^{-1}(e^{-x}), \quad a_i \equiv H_X^{(i)} \left( \sum_{j=r}^n j^{-1} \right),$$

$$g_Y(\cdot) \equiv 1/f_Y(F_Y^{-1}(\cdot)), \quad g_y \equiv g_Y(p), \quad g_y'' \equiv g_Y''(p),$$

$$H_Y(y) \equiv F_Y^{-1}(e^{-y}), \quad a'_i \equiv H_Y^{(i)} \left( \sum_{j=r}^n j^{-1} \right),$$

where  $H^{(i)}(\cdot)$  is the  $i$ th derivative of function  $H(\cdot)$ .

It follows that

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

$$\begin{aligned}
& + m^{-1}\tilde{u}_{2,\gamma}(z+C)\phi(z+C) + (m/n)^2\tilde{u}_{3,\gamma}(z+C)\phi(z+C) \Big| \\
& = o_p[m^{-1} + (m/n)^2].
\end{aligned}$$

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  $\tilde{u}_{2,\gamma}$  term again exactly matches 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. A similar comment applies here, too, about the effect of  $\tilde{S}_{m,n}$  being random.

## 3.5 Optimal smoothing parameter selection

### 3.5.1 Type I error

In order to select  $m$  to minimize type II error subject to control of type I error, expressions for type I and type II error dependent on  $m$  are needed. Comparing with the Edgeworth expansion in (3.10) when  $\gamma = 0$ , the approximate type I error of the two-sided symmetric test can be calculated using the corrected critical values from (3.7). Note that  $u_{1,0}(z)\phi(z)$  in (3.10) is an even function since  $\phi(z) = \phi(-z)$  and  $z$  only appears as  $z^2$  in  $u_{1,0}(z)$ , so  $u_{1,0}(z) = u_{1,0}(-z)$ ; thus it will cancel out for a two-sided test. Also note that the functions  $u_{2,0}(z)$  and  $u_{3,0}(z)$  are odd, so  $u_{2,0}(-z) = -u_{2,0}(z)$  and  $u_{3,0}(-z) = -u_{3,0}(z)$ . Below, the second high-order term will disappear due to the use of the fixed- $m$  critical value, leaving only the third high-order term.

**Proposition 3.3.** *If  $\gamma = 0$ , then  $P(|T_{m,n}| > z_{\alpha,m}) = e_I + o_p(m^{-1} + (m/n)^2)$ ,*

$$e_I = \alpha - 2(m/n)^2 u_{3,0}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}). \quad (3.12)$$

*Proof.* Starting with (3.10) with  $\gamma = 0$ , the two-sided symmetric test rejection probability under the null hypothesis for critical value  $z$  is

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

With the corrected critical value  $z_{\alpha,m} = z_{1-\alpha/2} + z_{1-\alpha/2}^3/(4m)$ ,

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

The dominant part of the type I error,  $e_I$ , 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 does not exceed nominal size  $\alpha$  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,  $\chi^2$ , and Fréchet distributions (see supplemental appendix). This suggests that simply using the fixed- $m$  corrected critical values alone is enough to reduce  $e_I$  to  $\alpha$  or below.

For the two-sample case, as shown in the full appendix, the result is similar.

**Proposition 3.4.** *If  $\gamma = 0$ , then*

$$\begin{aligned} P\left(|\tilde{T}_{m,n}| > z_{\alpha,m}\right) &\leq P\left(|\tilde{T}_{m,n}| > \tilde{z}_{\alpha,m}\right) = \tilde{e}_I + o_p(m^{-1} + (m/n)^2), \\ \tilde{e}_I &= \alpha - 2(m/n)^2 \tilde{u}_{3,0}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}). \end{aligned}$$

Here also, the corrected critical value approximated from the fixed- $m$  distribution leads to the dominant part of type I error being bounded by  $\alpha$  for all common distributions, since the sign of  $\tilde{u}_3$  is positive for the same reasons 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  $\delta \equiv f_X(\xi_p)/f_Y(\xi_p)$ .

### 3.5.2 Type II error

As above, I use the Edgeworth expansion in (3.10) to approximate the type II error of the two-sided symmetric test using critical values from (3.7). Since a uniformly most powerful test does not exist for general alternative hypotheses, I

will 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 choose to maximize power against the alternative where first-order power is 50% for a two-sided test. From above, the type II error 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  $\chi^2$  distribution with one degree of freedom and noncentrality parameter  $C^2$ , and  $C$  was defined in Theorem 3.1. 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 may be found in the working paper.

**Proposition 3.5.** *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*

$$\begin{aligned} P(|T_{m,n}| < z_{\alpha,m}) &= e_{II} + o_p(m^{-1} + (m/n)^2), \\ e_{II} &= 0.5 + (1/4)m^{-1}[\phi(z_{1-\alpha/2} - C) - \phi(z_{1-\alpha/2} + C)]Cz_{1-\alpha/2}^2 \\ &\quad + (m/n)^2 \frac{3(f')^2 - ff''}{6f^4} z_{1-\alpha/2} [\phi(z_{1-\alpha/2} + C) + \phi(z_{1-\alpha/2} - C)] \\ &\quad + O(n^{-1/2}). \end{aligned} \tag{3.13}$$

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  $m \rightarrow \infty$  and  $m/n \rightarrow 0$  are indeed of orders  $m^{-1}$  and  $(m/n)^2$  respectively, as given in their (2.5) and (2.6):

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

These are similar to above aside from the additional  $1/f^2$  and  $1/f$  factors. As  $m \rightarrow \infty$ ,  $\text{Var}(S_{m,n}) \rightarrow 0$  (increasing smoothing); with  $m$  fixed, this variance is also fixed (fixed smoothing). As  $n$  grows but  $S_{m,n}$  only uses a proportion of the  $n$  observations approaching zero, the bias will also decrease. The bias decreases



to zero in the fixed- $m$  thought experiment, too, and thus is not captured by the fixed- $m$  asymptotic distribution.

As discussed, for common distributions  $[3(f')^2 - ff'']/(6f^4) \geq 0$ , so the entire  $(m/n)^2$  expression in (3.13) is positive; type II error from this term is increasing in  $m$ , so a smaller value of  $m$  would minimize it. The  $m^{-1}$  term in (3.13) is also positive since  $\phi(z_{1-\alpha/2} - C) > \phi(z_{1-\alpha/2} + C)$ , so it is minimized by a larger  $m$ . It is then possible that minimizing  $e_{II}$  gives an “interior” solution  $m$ , i.e.  $m \in [1, \min(r - 1, n - r)]$ .

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

**Proposition 3.6.** *If  $C^2$  solves  $0.5 = G_{C^2}(z_{1-\alpha/2})$ ,*

$$\theta \equiv \tilde{S}_0^{-4}(g_x^4 + g_y^4) = (1 + \delta_a^4)/(1 + \delta_a^2)^2,$$

$g_x \equiv 1/f_X(F_X^{-1}(p))$  and  $g_y \equiv 1/f_Y(F_Y^{-1}(p))$  are as defined in Theorem 3.2, as are  $g_x''$  and  $g_y''$ ,  $\delta_a \equiv f_X(F_X^{-1}(p))/f_Y(F_Y^{-1}(p))$  similar to (3.8), and

$$\tilde{S}_0 \equiv ([f_X(F_X^{-1}(p))]^{-2} + [f_Y(F_Y^{-1}(p))]^{-2})^{1/2}$$

as in (3.8), then

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

### 3.5.3 Choice of $m$

With the fixed- $m$  corrected critical values,  $e_I \leq \alpha$  for all quantiles of common distributions, as discussed. Since size control is robust to smoothing parameter choice,  $m$  is chosen to minimize  $e_{II}$ . Since  $e_{II}$  has a positive  $m^{-1}$  component and a positive  $m^2$  component, it will be 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 over the feasible range.

For a randomized alternative (as in Sun, 2010b) for a symmetric two-sided test, the first-order condition of (3.13) leads to

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

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

As is common in the kernel bandwidth selection literature and in implementation of HS88, I plug in the standard normal<sup>5</sup> pdf  $\phi$  for  $f$ ; more precisely,  $\phi(\Phi^{-1}(p))$  for  $f(\xi_p)$ . In some cases, the plug-in  $m$  is close to optimal. In others it is different, but the effect on power is small; e.g., with  $\text{Unif}(0, 1)$ ,  $n = 45$ ,  $p = 0.5$ ,  $\alpha = 0.05$ , the Gaussian plug-in yields  $m = 9$  instead of  $m = 22$ , but the power loss is only 4% at the alternative considered in Proposition 3.5. For large  $n$ , though, estimation of  $f$ ,  $f'$ , and  $f''$  may be best.

The Gaussian plug-in yields

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

and for the 50% first-order power  $C$  and  $\alpha = 0.05$ , as calculated below,

$$m_K(n, p, \alpha = 0.05, C = 1.96) = n^{2/3} (1.42) \left( \frac{(\phi(\Phi^{-1}(p)))^2}{2(\Phi^{-1}(p))^2 + 1} \right)^{1/3}. \quad (3.14)$$

Compare (3.14) with the suggested  $m$  from HS88. Rewriting their (3.1) to match my notation above,

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

---

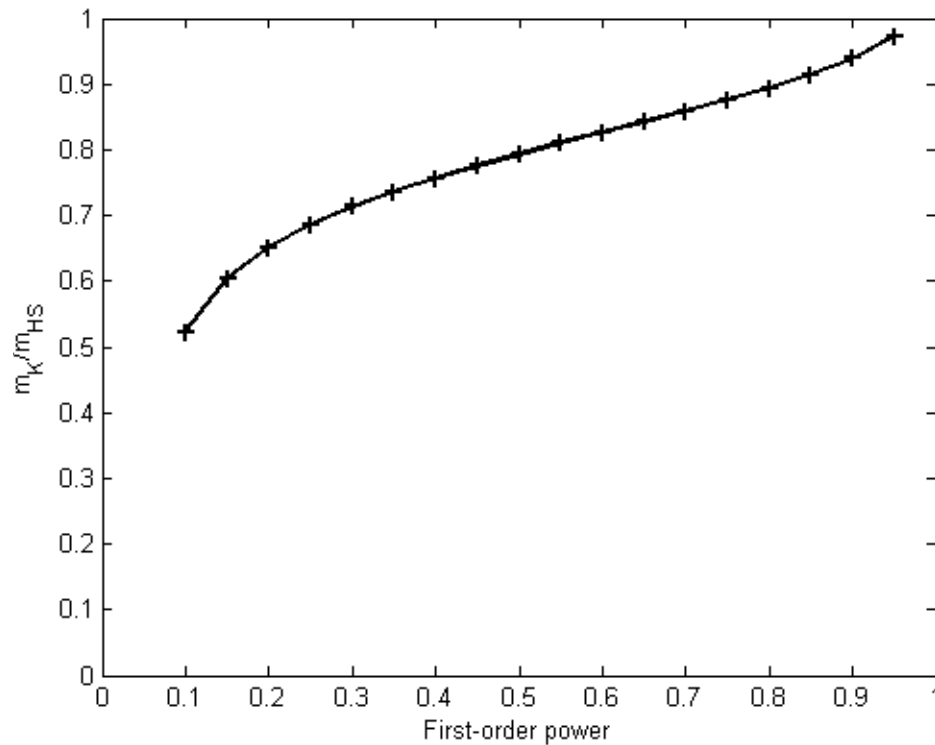
<sup>5</sup>Here, using  $N(0, 1)$  is mathematically equivalent to  $N(\mu, \sigma^2)$  since  $\mu$  and  $\sigma$  cancel out.

HS88 proceed with a median-specific, data-dependent method. More commonly, papers like Koenker and Xiao (2002) plug in  $\phi$  for  $f$  when computing  $h_{HS}$  (HS=Hall and Sheather) in their simulations, as shown explicitly on page 3 of their electronic appendix; Goh and Knight (2009) use the same Gaussian plug-in version. Koenker and Xiao (2002) also use a Gaussian plug-in procedure based on Bofinger (1975). Their “bandwidth”  $h$  is just  $m/n$ , so they use the equivalent of

$$m_{HS} = n^{2/3} z_{1-\alpha/2}^{2/3} \left( 1.5 \frac{[\phi(\Phi^{-1}(p))]^2}{2(\Phi^{-1}(p))^2 + 1} \right)^{1/3}, \quad (3.15)$$

$$m_B = n^{4/5} \left( 4.5 \frac{[\phi(\Phi^{-1}(p))]^4}{[2(\Phi^{-1}(p))^2 + 1]^2} \right)^{1/5}. \quad (3.16)$$

For  $p = 0.5$ ,  $\alpha = 0.05$ , these two will be equal if  $n = 20.93$ , so  $m_B < m_{HS}$  if  $n < 21$  and  $m_B > m_{HS}$  if  $n \geq 21$ . With standard normal critical values, size and power are smaller whenever  $m$  is bigger.



**Figure 3.1:** Plot of the ratio  $m_K/m_{HS}$  against the first-order power used to calculate  $C$ ; 0.5 is used throughout this paper.

Since  $z_{1-\alpha/2}^{2/3}(1.5)^{1/3} = 1.79$  for  $\alpha = 0.05$ ,

$$m_K(n, p, \alpha = 0.05, C = 1.96) = 0.79m_{HS}.$$

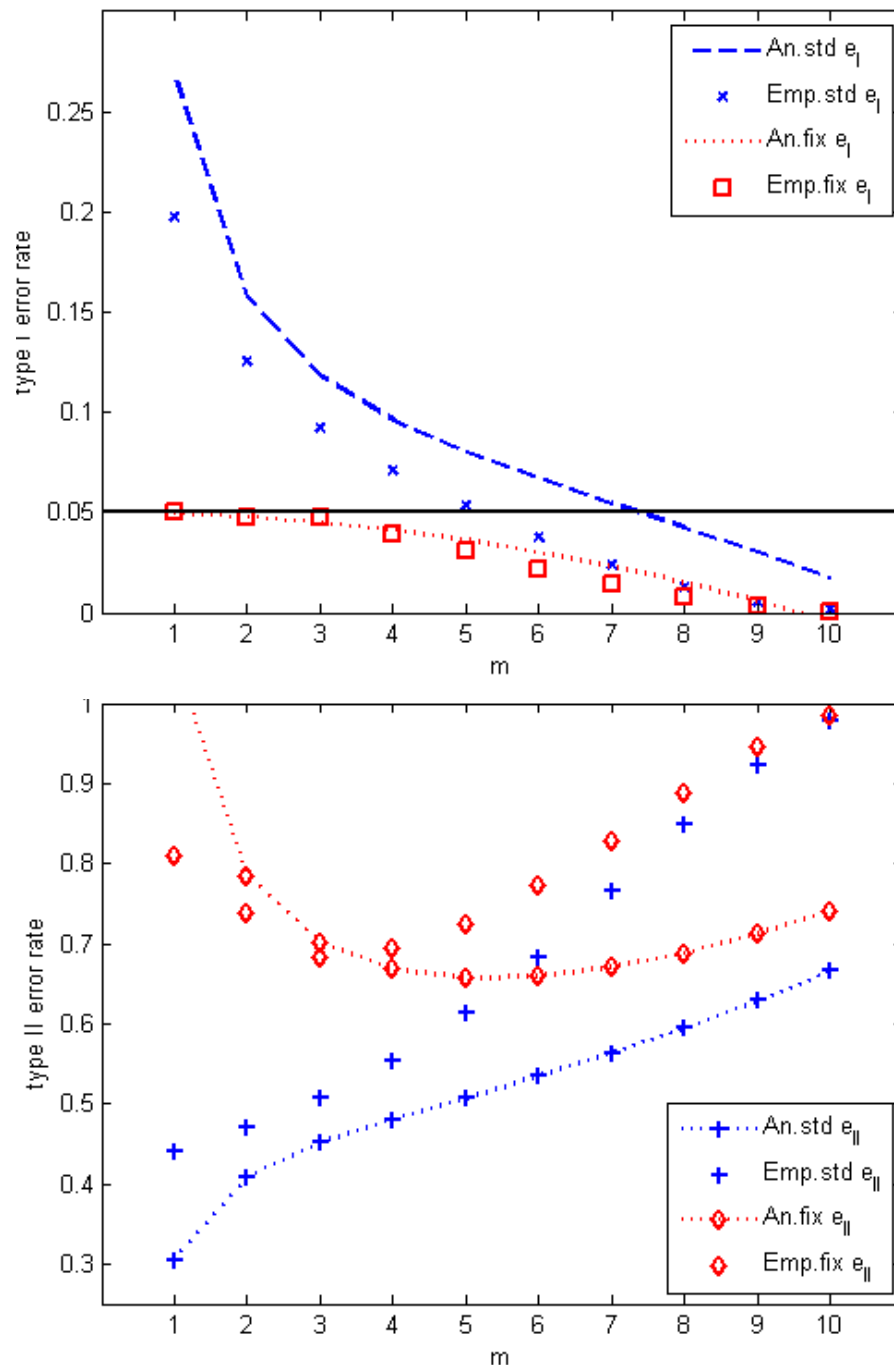
This ratio is calculated for other values of  $C$  and plotted in Figure 3.1. The dependence is mild until values closer to 0% or 100% power. Since  $m_K$  and  $m_{HS}$  optimize for critical values from different distributions, it makes sense that  $m_K < m_{HS}$ ; the fixed- $m$  critical values control size, allowing for smaller  $m$  when for HS88  $e_I$  might be a binding constraint.

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 (3.6) with respect to  $m$ , and then solving for  $m$ . Since the two-sample  $\tilde{e}_{II}$  has positive  $m^{-1}$  and  $m^2$  terms, it is also U-shaped. With  $\tilde{z}_{\alpha,m}$  from (3.9) and  $\theta$  from (3.6), solving for  $m$  in the FOC yields

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

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 3.2 compares the Edgeworth approximations to the true (simulated) type I and type II error for scalar data drawn from a log-normal distribution. For distributions like this that are extremely different from the normal distribution, even the Edgeworth approximation can be significantly different in small samples. With standard normal critical values, type I error is monotonically decreasing with  $m$  while type II error is monotonically increasing, so setting  $e_I = \alpha$  exactly should minimize  $e_{II}$  subject to  $e_I \leq \alpha$ . With fixed- $m$  critical values, this is not true, so the Edgeworth expansion for general  $\gamma$  in Theorem 3.1 is necessary. With fixed- $m$  critical values, type I error is always below  $\alpha$  (approximately, and usually truly), and the type II error curve near the selected  $m$  is very flat since it is near the minimum where the slope is zero. Consequently, a larger-than-optimal  $m$  can result in larger power loss for HS88 than the new method, and a smaller-than-optimal  $m$  incurs size distortion for HS88 but not for the new method.



**Figure 3.2:** Analytic and empirical  $e_I$  (top) and  $e_{II}$  (bottom) by  $m$ .

Both:  $n = 21$ ,  $p = 0.5$ ,  $\alpha = 0.05$ ,  $F$  is log-normal with  $\mu = 0$ ,  $\sigma = 3/2$ , 5000 replications (for empirical values). Lines are analytic; markers without lines are empirical.

### 3.6 Simulation study<sup>6</sup>

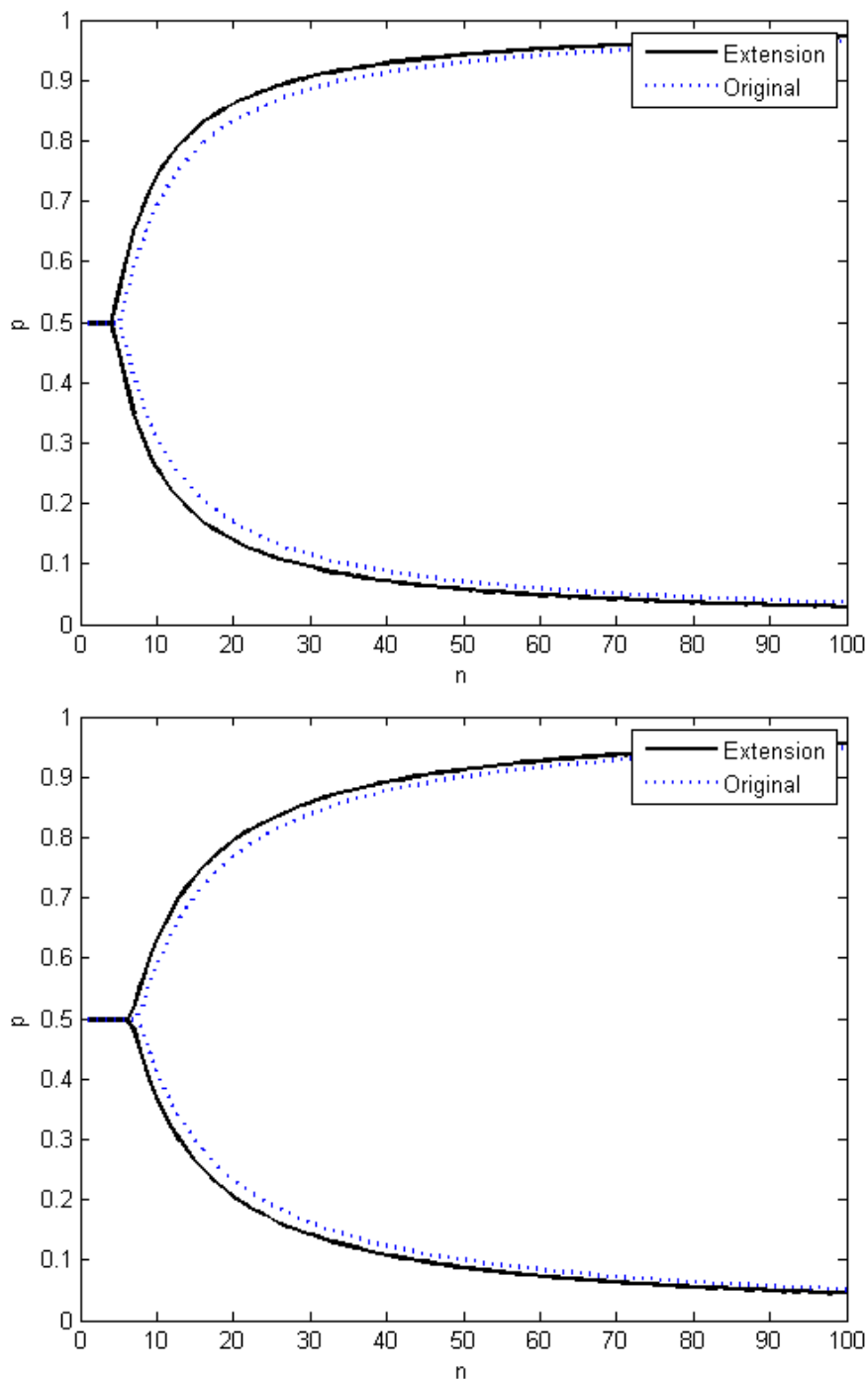
Regarding  $m$ , Siddiqui (1960) originally suggested a value of  $m$  on the order of  $n^{1/2}$ . Bloch and Gastwirth (1968) then suggested a rate of  $n^{4/5}$ , to minimize the asymptotic mean square error of  $S_{m,n}$ . With this  $n^{4/5}$  rate, Bofinger (1975) then suggested the  $m_B$  above in (3.16). HS88, using an Edgeworth expansion under the null hypothesis, find that an  $m$  of order  $n^{2/3}$  minimizes the level error of two-sided tests, and they provide an infeasible expression for  $m$ . For the median, they also give a data-dependent method for selecting  $m$ ; for a general quantile, the Gaussian plug-in version of  $m_{HS}$  in (3.15) is more commonly used, as in Koenker and Xiao (2002) and Goh and Knight (2009).

The following simulations compare five methods. The first method is presented in this paper, with fixed- $m$  corrected critical values and  $m_K$  chosen to minimize type II error subject to control of type I error; the code is publicly available as Kaplan (2011). 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 (3.15). The third method uses standard normal critical values and  $m_B$  as in (3.16). These three methods are referred to as “New method,” “HS,” and “B,” respectively, in the text as well as figures in this section. For the new method, I used the floor function to get an integer value for  $m_K$ ; for HS and B, I instead rounded to the nearest integer, to not amplify their size distortion.

The fourth method “BS” is a bootstrap 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 and thus were used; any exceptions are noted below. Discussion of other bootstraps is in the working paper version.

---

<sup>6</sup>MATLAB code implementing a hybrid of Hutson’s method and the new method, `quantile_inf.m`, is publicly available through the author’s website or MATLAB File Exchange. MATLAB code for simulations is available from the author upon request.



**Figure 3.3:** Graph showing for which combinations of quantile  $p \in (0, 1)$  (y-axis) and sample size  $n$  (x-axis) the Hutson (1999) method is computable, in the middle region. “Original” (dotted lines) refers to the equal-tail version; “Extension” (solid lines) marginally increases the region by not requiring the equal-tail property. Top:  $\alpha = 5\%$ . Bottom:  $\alpha = 1\%$ .

The fifth method is a near-exact method from Hutson (1999), based on fractional order statistics.<sup>7</sup> For one-sample inference, it performed best in every simulation where it could be used, so the results here focus on cases when it cannot be computed. Figure 3.3 shows when it can and can't be computed, in terms of  $n$  and  $p$ . For two-sample inference, the related method in Chapter 1 has similarly superlative performance in practice.

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, sometimes even better than BS. When Hutson (1999) is computable, the new method can have better power than HS and B, too.

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<sup>8</sup> have less size distortion, but still some. BS, though, seems to control size and have better power than the new method with estimated  $\theta$ . With the true  $\theta$ , BS and the new method perform quite similarly. General two-sample setups would have results in between these extremes of  $f_X/f_Y \rightarrow \infty$  (or zero) and  $f_X/f_Y = 1$ .

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

Table 3.1 and Figure 3.4 show one- and two-sample size and power in the case of  $n = 3$  or  $n = 4$ . For one-sample, 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

---

<sup>7</sup>For a rigorous theoretical justification, see Chapter 1.

<sup>8</sup>Since HS88 and Bofinger (1975) do not provide values of  $m$  for the two-sample case, I used their one-sample  $m$  values.



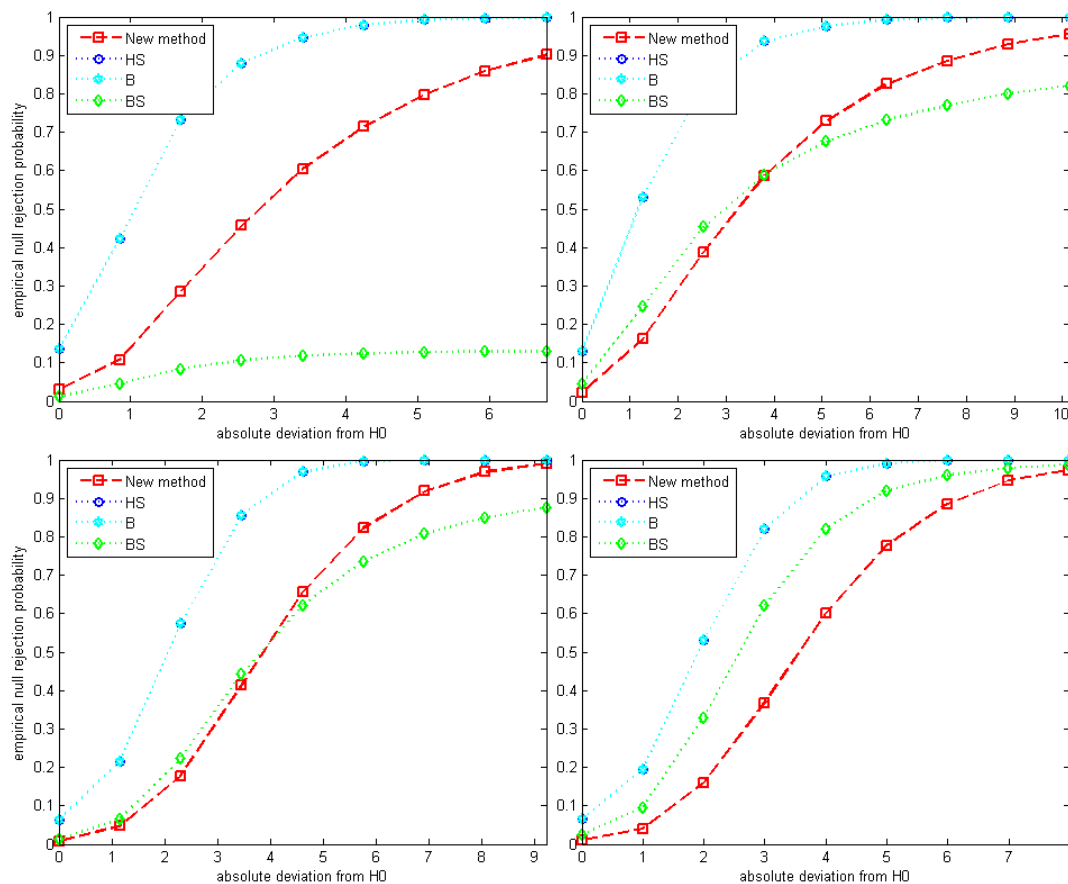
**Table 3.1:** Empirical size as percentage (nominal: 5.0),  $n = 3, 4$ ,  $p = 0.5$ , one- and two-sample.

Distribution	one-sample						two-sample		
	$n = 3$			$n = 4$			$n = 3$		
	New	BS	HS <sup>a</sup>	New	BS	HS	New <sup>b</sup>	BS	HS
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
Unif(0, 1)	3.2	1.5	16.0	5.0	6.8 <sup>c</sup>	21.3	1.4	1.9	8.6

<sup>a</sup>B is the same as HS in this table.

<sup>b</sup>If true  $\theta$  used instead: 1.6, 1.1, 1.6, 2.8.

<sup>c</sup>6.7% with 999 outer replications



**Figure 3.4:** Empirical power properties;  $p = 0.5$ . Top row:  $F = \text{Exp}(1)$ , one-sample. Bottom:  $N(0, 1)$ , two-sample. Left column:  $n = 3$ . Right:  $n = 4$ .

**Table 3.2:** Empirical size as percentage (nominal: 5.0),  $n = 45, 21$ ,  $p = 0.95$ , one-sample.

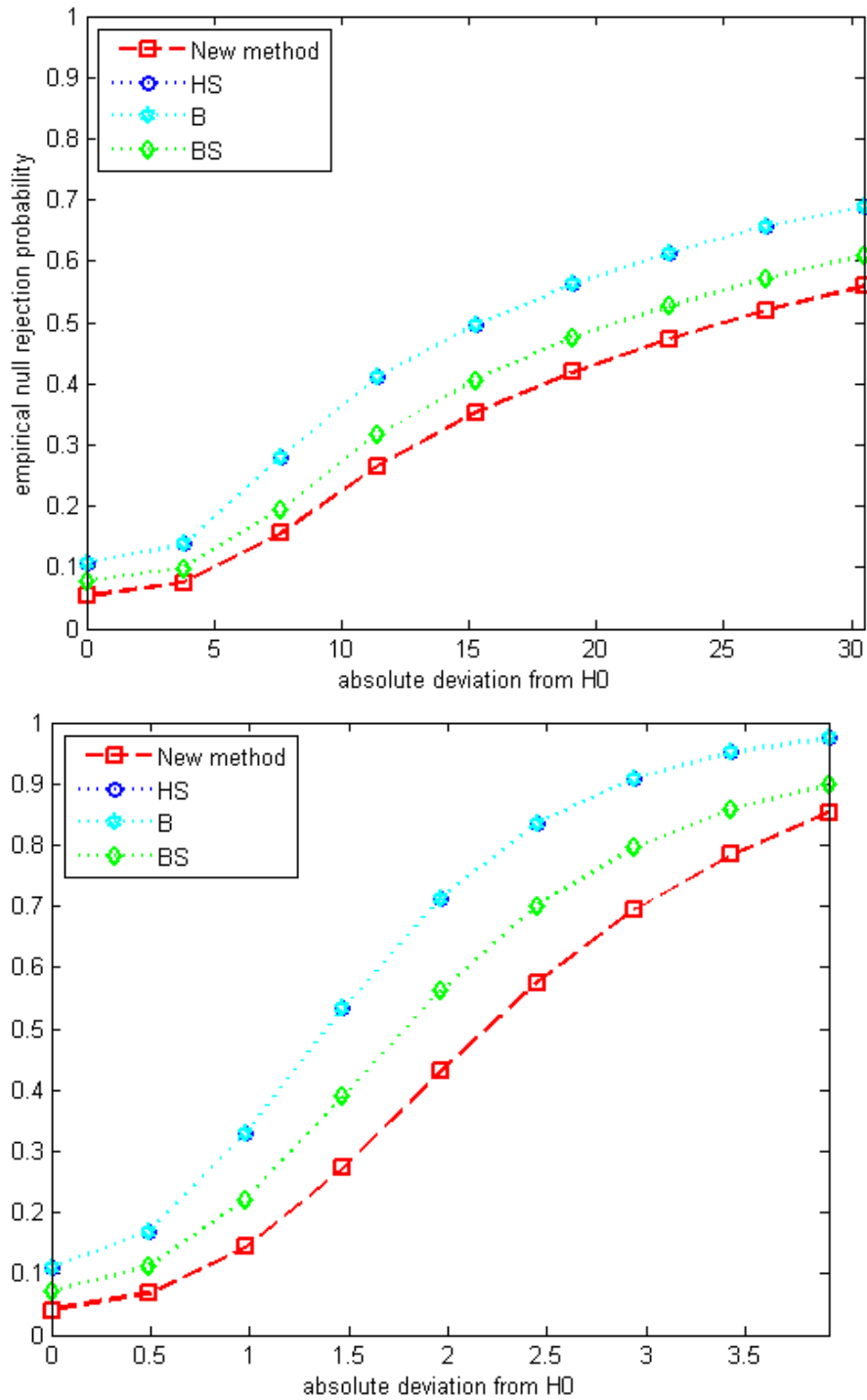
Distribution	$n = 45$				$n = 21$			
	New	BS; $B_1 =$		HS/B	New	BS; $B_1 =$		HS/B
		99	999			99	999	
N(0, 1)	3.2	6.4	6.5	10.4	6.7	11.3	11.5	25.3
Slash	5.3	7.8	8.1	10.7	12.9	17.8	18.6	30.3
Logn(0, $\sigma = 3/2$ )	5.0	7.3	7.5	10.7	11.3	14.6	15.2	28.7
Exp(1)	4.0	6.9	7.0	10.7	8.0	12.6	13.1	26.9
Uniform(0, 1)	3.3	5.4	5.3	11.2	4.7	6.8	6.7	22.3
GEV(0, 1, 0)	4.1	7.2	7.3	11.0	7.3	12.2	12.6	25.9
$\chi_1^2$	3.6	6.5	6.4	10.2	8.7	13.5	13.9	27.5

for HS and B. For one-sample power, BS is very poor for  $n = 3$ , but better than the new method at some alternatives for  $n = 4$ . For two-sample power, BS is equal to the new method for a range of alternatives with  $n = 3$ , and better for  $n = 4$ .

For  $n = 45$  and  $p = 0.05$  or  $p = 0.95$ , the new method controls size except for empirical size around 5.5% for the slash distribution. In contrast, BS, HS, and B are all size-distorted, sometimes significantly; see Table 3.2. With  $n = 21$ , even the new method is size-distorted, but significantly less so. Even with better size, the new method's power is competitive with BS, as in Figure 3.5.

For the two-sample case with identical distributions, BS always controls size along with the new method. However, moving toward the limiting case where one distribution is a constant, when there is a significant different in variance between the two samples' distributions (e.g., factor of 25), BS has significantly worse size distortion than the new method for  $n = 21$ , and has size distortion while the new method has none for  $n = 29$  and  $n = 45$ . In the case of identical distributions, the new method has power competitive with BS for  $n = 21$ , but by  $n = 45$  BS has significantly better power. (These and additional simulations are in the working paper.)

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



**Figure 3.5:** Empirical power properties, one-sample,  $n = 45$ ,  $p = 0.95$ . Only the new method controls size, and power is not size-adjusted. Top: slash. Bottom:  $GEV(0,1,0)$ .

9.6% for BS and 20.7% for HS and B, compared with 5.3% for the new method.

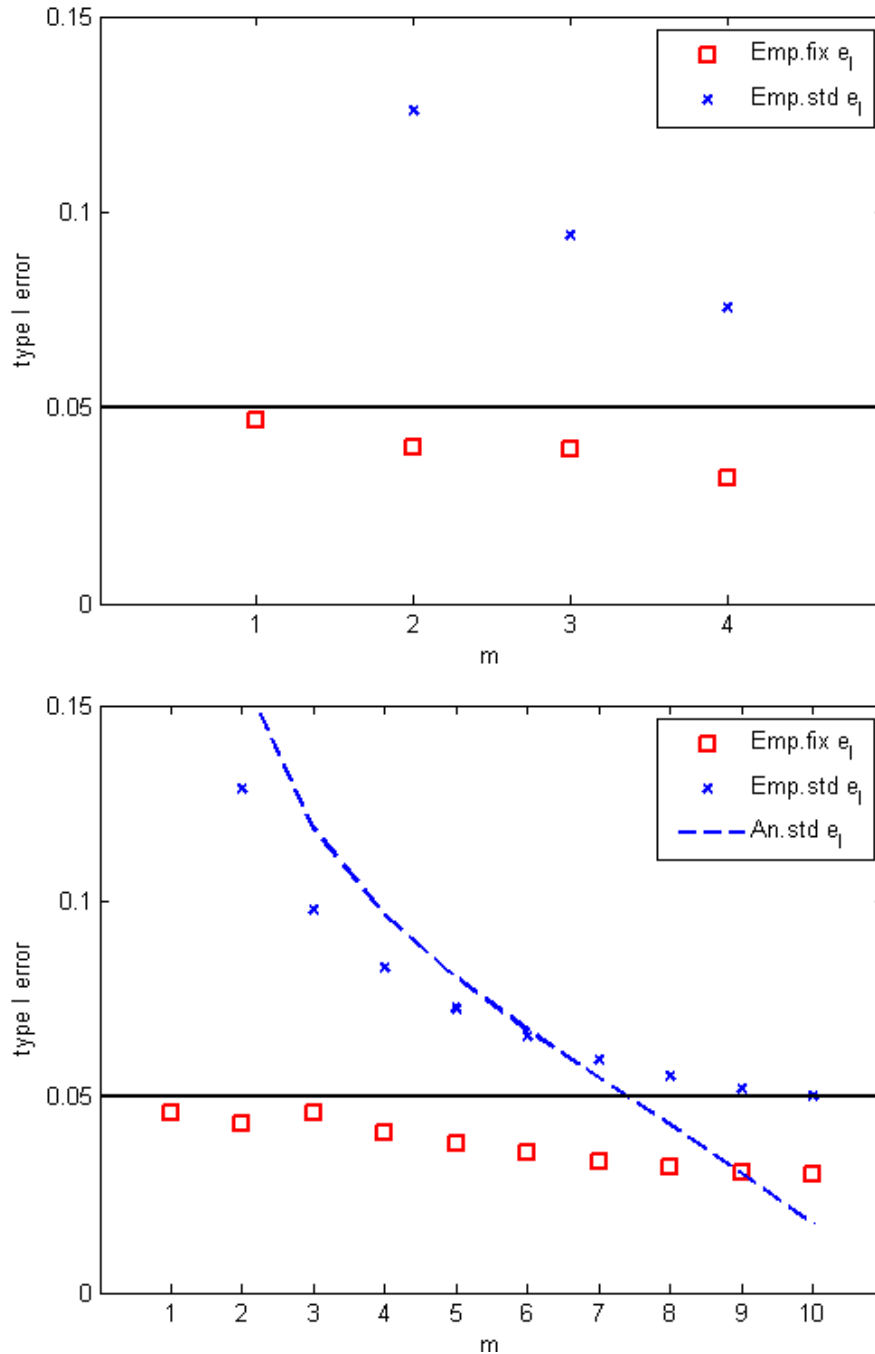
There are two different proximate “causes” of size distortion for HS and B in general, illustrated in Figure 3.6. When  $n = 21$  and  $p = 0.2$ ,  $r = \lfloor (21)(0.2) \rfloor + 1 = 5$ , which means that  $m$  can be at most 4 such that  $r - m \geq 1$ ; and  $m_{HS} = m_B = 4$ , the biggest  $m$  possible. This is the best choice, but still size-distorted. For  $n = 21$  and any  $p < 0.35$ ,  $m_{HS} = m_B = r - 1$ , the maximum possible  $m$ . However, when  $n = 21$  and  $p \geq 0.35$ ,  $m_{HS}$  and  $m_B$  are strictly less than the maximum possible  $m$ . In those cases, HS and B could have chosen a larger  $m$  to lessen the size distortion, but didn’t. In some cases, both “causes” apply, such as with  $F = \text{Exp}(1)$ ,  $n = 71$ ,  $p = 0.2$ , and  $\alpha = 0.05$ , where bigger  $m_{HS}$  would have reduced size distortion, but not eliminated it.

Of theoretical (if not directly practical) interest is Figure 3.7. The new method chooses  $m$  to maximize power, and in some cases this produces significantly better power than HS or B. This is true in both the one-sample case (top row) and the two-sample case (bottom row).

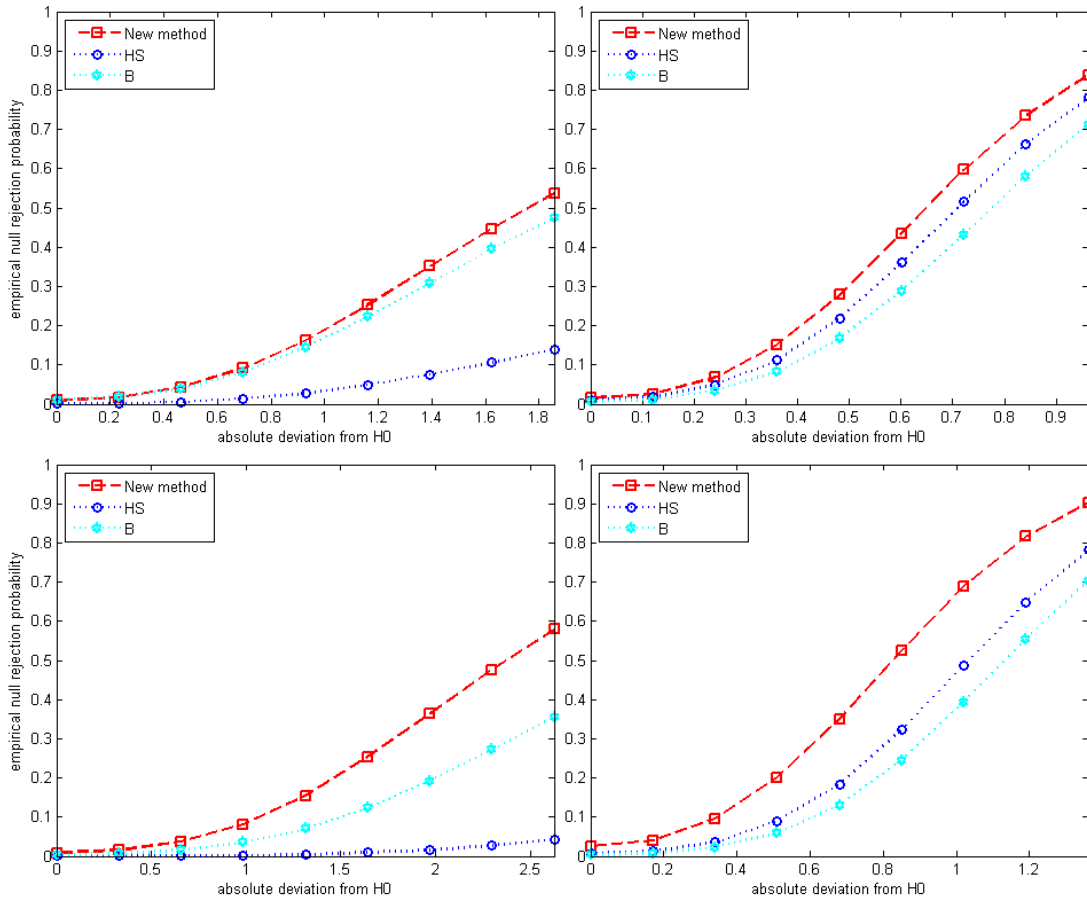
In the two-sample case, the results shown and discussed were for independent samples, as the theoretical results assumed. If there is enough negative correlation, size distortion occurs for all methods; if there is positive correlation, power will decrease for all methods. The effect of uncorrelated, 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.

### 3.7 Conclusion

This paper proposes new one- and two-sample quantile testing procedures based on the SBG test statistic in (3.3). 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 errors are approximated using an Edgeworth expansion, and



**Figure 3.6:** Causes of size distortion for HS and B. Empirical type I error (x markers) is plotted against  $m$ , for a test with standard normal critical values; missing marker on the left for  $m = 1$  is above 0.15. The dashed line in the right plot is the plug-in  $e_I$  used to select  $m_{HS}$ . Top: no  $m$  can control size. Bottom: approximation error leads to  $m_{HS} = 7 < 10$  and size distortion.  $F = \text{Uniform}(0, 1)$ ,  $n = 21$ ,  $\alpha = 0.05$ , and  $p = 0.2$  on the left,  $p = 0.5$  on the right; 5000 replications.



**Figure 3.7:** Empirical power curves, comparing  $n = 11$  (left column) and  $n = 41$  (right column). 50,000 replications each,  $p = 0.5$ , and distribution is Cauchy. Top row: one-sample. Bottom row: two-sample.

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, while 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  $\theta$ ) 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.

# Appendix A

## Technical appendix to Chapter 1

### A.1 Mathematical proofs

#### Proof of Theorem 1.1

*Proof.*

$$\begin{aligned}
 \sup_{u \in \mathcal{U}_n^\delta} \left| \hat{Q}_X^I(u) - \hat{Q}_X^L(u) \right| &= \sup_{u \in \mathcal{U}_n^\delta} \left| F^{-1}(\hat{Q}_U^I(u)) - F^{-1}(\hat{Q}_U^L(u)) \right| \\
 &\leq \frac{1}{\delta} \sup_{u \in \mathcal{U}_n^\delta} \left| \hat{Q}_U^I(u) - \hat{Q}_U^L(u) \right| \\
 &\leq \frac{1}{\delta} \max_{i \in \{1, \dots, n\}} \sup_{u \in \left[\frac{i}{n+1}, \frac{i+1}{n+1}\right]} \left| \hat{Q}_U^I(u) - \hat{Q}_U^L(u) \right| \\
 &\leq \frac{1}{\delta} \max_{i \in \{1, \dots, n\}} \left[ \hat{Q}_U^I\left(\frac{i+1}{n+1}\right) - \hat{Q}_U^L\left(\frac{i}{n+1}\right) \right].
 \end{aligned}$$

The bracketed quantities are spacings between adjacent uniform order statistics (referred to as  $\Delta$  in the heuristics presented at the beginning of §1.2) and (unconditionally) have distribution  $\beta(1, n)$  with CDF from Kumaraswamy (1980). With an arbitrary sequence  $a_n$  such that  $\lim_{n \rightarrow \infty} a_n / (n^{-1} \lceil \log n \rceil) = \infty$ ,

$$\begin{aligned}
 P\left(\sup_{u \in \mathcal{U}_n^\delta} \left| \hat{Q}_X^I(u) - \hat{Q}_X^L(u) \right| > a_n\right) &\leq nP(\beta(1, n) > \delta a_n) \\
 &= \max\{(1 - \delta a_n)^n, 0\} \\
 &= \max\{\exp\{\log(n) + n \log(1 - \delta a_n)\}, 0\}
 \end{aligned}$$



$$\leq \exp\{\log(n) - n\delta a_n\} \rightarrow 0. \quad \square$$

## Proof of Lemma 1.2

*Proof.* a) Recall  $\Delta\mathbf{X} \sim \text{Dirichlet}(\Delta\mathbf{k})$ , so

$$\log(f_{\mathbf{X}}(\mathbf{x})) = \log(\Gamma(n+1)) + \sum_{j=1}^{J+1} \left[ (\Delta k_j - 1) \log(\Delta x_j) - \log(\Gamma(\Delta k_j)) \right].$$

We will use Stirling-type bounds on the gamma functions in the Dirichlet PDF. Specifically, from Robbins (1955) we have

$$\begin{aligned} \sqrt{2\pi} n^{n+1/2} e^{-n+1/(12n+1)} &< n! < \sqrt{2\pi} n^{n+1/2} e^{-n+1/(12n)}, \text{ so} \\ \log(\Gamma(n)) &= \frac{\log(2\pi)}{2} + [n - 1/2] \log(n-1) + (1-n) + O(n^{-1}). \end{aligned}$$

Applying the above to the density of  $\mathbf{X}$  gives

$$\begin{aligned} \log(f_{\mathbf{X}}(\mathbf{x})) &= -J \left( \frac{\log(2\pi)}{2} + 1 \right) + [n + 1/2] \log(n) \\ &\quad + \sum_{j=1}^{J+1} \left[ (\Delta k_j - 1) \log(\Delta x_j) - (\Delta k_j - 1/2) \log(\Delta k_j - 1) \right] \\ &= \underbrace{\frac{J}{2} \log(n/2\pi) + \frac{1}{2} \sum_{j=1}^{J+1} \log\left(\frac{n}{\Delta k_j - 1}\right)}_K \\ &\quad + \underbrace{\sum_{j=1}^{J+1} (\Delta k_j - 1) \log\left(\frac{n\Delta x_j}{\Delta k_j - 1}\right)}_{h(\mathbf{x})} - J, \end{aligned} \tag{A.1}$$

where  $K$  is the constant given in the statement of the lemma and  $h(\mathbf{x})$  is the remainder. To expand  $h(\cdot)$ , we calculate derivatives

$$\begin{aligned} h_j(\mathbf{x}) &= \left( \frac{\Delta k_j - 1}{\Delta x_j} - \frac{\Delta k_{j+1} - 1}{\Delta x_{j+1}} \right), \tag{A.2} \\ h_{j,j}(\mathbf{x}) &= - \left( \frac{\Delta k_j - 1}{\Delta x_j^2} + \frac{\Delta k_{j+1} - 1}{\Delta x_{j+1}^2} \right), \quad h_{j,j+1}(\mathbf{x}) = \frac{\Delta k_{j+1} - 1}{\Delta x_{j+1}^2}, \\ h_{j,j,j}(\mathbf{x}) &= \left( \frac{\Delta k_{j+1} - 1}{2\Delta x_{j+1}^3} - \frac{\Delta k_j - 1}{2\Delta x_j^3} \right), \\ h_{j,j,j+1}(\mathbf{x}) &= -h_{j,j+1,j+1}(\mathbf{x}) = \left( \frac{\Delta k_{j+1} - 1}{2\Delta x_{j+1}^3} \right), \end{aligned}$$

and all other cross-derivatives are zero. Let  $\underline{H}_{\mathbf{x}}$  denote the corresponding Hessian at some realization  $\mathbf{X} = \mathbf{x}$ . In the appendix we show that it is negative definite for all values of  $\mathbf{x}$ . At the value  $\mathbf{x} = \mathbf{k}/n' = \mathbf{k}/[n + 1]$ , using (A.2), we compute

$$\begin{aligned} h(\mathbf{k}/n') &= \sum_{j=1}^{J+1} (\Delta k_j - 1) \left[ \log(1 + [\Delta k_j - 1]^{-1}) + \log(n/n') \right] - J \\ &= O(\Delta \mathbf{k}^{-1}), \\ h_j(\mathbf{k}/n') &= n(\Delta k_{j+1}^{-1} - \Delta k_j^{-1}), \\ h_{j,j}(\mathbf{k}/n') &= -n^2 [\Delta k_j^{-1} + \Delta k_{j+1}^{-1} + O(\min\{\Delta k_j, \Delta k_{j+1}\}^{-2})], \\ h_{j,j+1}(\mathbf{k}/n') &= n^2 (\Delta k_{j+1}^{-1} + O(\Delta k_{j+1}^{-2})). \end{aligned} \tag{A.3}$$

Third derivatives at some value  $\mathbf{k}/n' + \mathbf{z}_n$ , where the sequence  $\mathbf{z}_n$  must satisfy Condition  $\star(a_n)$ , are

$$\begin{aligned} h_{j,j+1}(\mathbf{k}/n' + \mathbf{z}_n) &= \frac{\Delta k_{j+1} - 1}{2(\Delta k_{j+1}/n' + \Delta z_{j+1})^3} \\ &\leq \left( \frac{n^3}{2\Delta k_{j+1}^2} \right) \left( 1 + \frac{n\Delta z_{j+1}}{\Delta k_{j+1}} \right)^3 = O(n^3 \Delta \mathbf{k}^{-2}), \end{aligned}$$

where the final rate follows because Condition  $\star(a_n)$  implies

$$\frac{n\Delta z_{j+1}}{\Delta k_{j+1}} = O(a_n \Delta k_{j+1}^{-1/2}) \rightarrow 0$$

by the second and third requirements on  $a_n$ , respectively. Since this form holds for all third-order derivatives of  $h(\cdot)$ , we have

$$\sup_{\substack{\mathbf{x}|\star(a_n), \\ |\rho|=3}} |D_\rho h(\mathbf{x})| = O(n^3 \Delta \mathbf{k}^{-2}).$$

Applying results from (A.3), we expand  $h(\mathbf{x})$  to get

$$\begin{aligned} &\sup_{\mathbf{x}|\star(a_n)} \left| h(\mathbf{x}) - \frac{1}{2} (\mathbf{x} - \mathbf{k}/n')' \underline{H}_{\mathbf{k}/n'} (\mathbf{x} - \mathbf{k}/n') \right| \\ &= h(\mathbf{k}/n') + h'(\mathbf{k}/n') (\mathbf{x} - \mathbf{k}/n') + O\left( [n^3 \Delta \mathbf{k}^{-2}] (\mathbf{x} - \mathbf{k}/n')^3 \right) \\ &= O(\Delta \mathbf{k}^{-1}) + O(a_n \Delta \mathbf{k}^{-1/2}) + O\left( a_n^3 \Delta \mathbf{k}^{-1/2} \right) = O\left( a_n^3 \Delta \mathbf{k}^{-1/2} \right). \end{aligned} \tag{A.4}$$

This completes our demonstration for the density.

The density derivative is

$$\begin{aligned}
\frac{d \log[f_{\mathbf{X}}(\mathbf{x})]}{d\mathbf{x}} &= h'(\mathbf{x}) \\
&= h'(\mathbf{k}/n') + \underline{H}_{\mathbf{k}/n'}(\mathbf{x} - \mathbf{k}/n') + O\left([n^3 \Delta \mathbf{k}^{-2}](\mathbf{x} - \mathbf{k}/n')^2\right) \\
&= O(n \Delta \mathbf{k}^{-1}) + \underline{H}_{\mathbf{k}/n'}(\mathbf{x} - \mathbf{k}/n') + O\left(a_n^2 n \Delta \mathbf{k}^{-1}\right). \tag{A.5}
\end{aligned}$$

The density derivative in  $k_j/n'$  is given by<sup>1</sup>

$$\begin{aligned}
\frac{d \log[f_{\mathbf{X}_k}(\mathbf{x})]}{dk_j/n'} &= n' \frac{dh(\mathbf{x})}{dk_j} + n' \frac{dK}{dk_j} \\
&= n \left[ \log\left(\frac{\Delta k_{j+1} - 1}{n \Delta x_{j+1}}\right) - 1 \right] - n \left[ \log\left(\frac{\Delta k_j - 1}{n \Delta x_j}\right) - 1 \right] \\
&\quad + O(n \min\{\Delta k_j, \Delta k_{j+1}\}^{-1}) \\
&= \left[ \frac{\Delta k_{j+1} - 1}{\Delta x_{j+1}} + O(a_n^2 n \Delta k_j^{-1}) \right] - \left[ \frac{\Delta k_j - 1}{\Delta x_j} + O(a_n^2 n \Delta k_j^{-1}) \right] \\
&\quad + O(n \min\{\Delta k_j, \Delta k_{j+1}\}^{-1}) \\
&= -h_j(\mathbf{x}) + O(a_n^2 n \min\{\Delta k_j, \Delta k_{j+1}\}^{-1}).
\end{aligned}$$

- b) The marginals of  $\mathbf{X}$  are such that  $\Delta X_j \sim \text{Dirichlet}(\Delta k_j, n+1 - \Delta k_j)$ . Appealing to the representation in (A.1), we have<sup>2</sup>

$$\begin{aligned}
\log(f_{\Delta X_j}(x)) &= \frac{1}{2} \overbrace{\left[ \log(n/2\pi) + \log(n/[\Delta k_j - 1]) + \log(n/[n - \Delta k_j]) \right]}^K \tag{A.6} \\
&\quad + \overbrace{(\Delta k_j - 1) \log(nx/[\Delta k_j - 1]) + (n - \Delta k_j) \log\left(\frac{n(1-x)}{n - \Delta k_j}\right) - 1}_{h(x)}.
\end{aligned}$$

The value and derivatives of  $h(x)$  are as in (A.2) and (A.3). Condition  $\star(a_n)$  is violated if for some  $j$

$$\Delta X_j \notin (l, u) \equiv \left( [\Delta k_j - a_n \Delta k_j^{1/2}]/n, [\Delta k_j + a_n \Delta k_j^{1/2}]/n \right).$$

<sup>1</sup>The  $O(n^{-1})$  term from Stirling approximation to the gamma function has a derivative of sufficiently low order that it is insignificant in calculating derivatives of our log PDF. This can be seen by examining equations (4)-(8) in Robbins (1955).

<sup>2</sup>Stirling approximation error in  $K$  is of lower order and is omitted.

At the upper endpoint, we apply (A.4) and (A.5) to get

$$\begin{aligned}
h(u) &= (1/2)h''(\Delta k_j/n')(u - \Delta k_j/n')^2 + O(a_n^3 \Delta k_j^{-1/2}) \\
&= -(a_n^2/2)\Delta k_j(\Delta k_j^{-1} + [n + 1 - \Delta k_j]^{-1}) + O(a_n^3 \Delta \mathbf{k}^{-1/2}) \\
&\leq -(a_n^2/2) + O(a_n^3 \Delta \mathbf{k}^{-1/2}), \tag{A.7}
\end{aligned}$$

$$\begin{aligned}
h'(u) &= -a_n n \Delta k_j^{1/2} (\Delta k_j^{-1} + [n + 1 - \Delta k_j]^{-1}) + O(a_n^2 \Delta \mathbf{k}^{-1/2} n) \\
&= -a_n \left( \frac{n^2}{\Delta k_j^{1/2} [n - \Delta k_j]} \right) + O(a_n^2 n \Delta \mathbf{k}^{-1}). \tag{A.8}
\end{aligned}$$

For convenience we also calculate

$$\begin{aligned}
K - \log(-h'(u)) &\asymp (1/2)[3 \log(n) - \log(\Delta k_j) - \log(n - \Delta k_j) - \log(2\pi)] \\
&\quad - [\log(a_n) + 2 \log(n) - 1/2 \log(\Delta k_j) - \log(n - \Delta k_j)] \\
&= (1/2)[\log(n - \Delta k_j) - \log(n) - \log(2\pi)] - \log(a_n) \\
&\leq -(1/2) \log(2\pi) - \log(a_n). \tag{A.9}
\end{aligned}$$

Recall that  $h''(\cdot)$  is everywhere negative, so for all  $x > u$ ,  $h(x) < h(u) + h'(u)(x - u)$ . Thus,

$$\begin{aligned}
P(\Delta X_j > u) &= \int_u^\infty f_{\Delta X_j}(x) dx = e^K \int_u^\infty \exp\{h(x)\} dx \\
&\leq e^K \int_u^\infty \exp\{h(u) + h'(u)(x - u)\} dx.
\end{aligned}$$

Recalling that the leading term of  $h'(u)$  is negative (see (A.8)), we integrate to get

$$\begin{aligned}
&= e^K \left[ \frac{e^{h(u)}}{-h'(u)} \right] \\
&= \exp [K - \log(-h'(u)) + h(u)].
\end{aligned}$$

The conclusion now follows by plugging in expressions from (A.7) and (A.9). Symmetric steps go through to achieve the same rate for the lower tail of  $\Delta X_j$ . There exist  $J + 1$  spacings that can generate violations of Condition  $\star(a_n)$ ; summing over them does not change the rate since  $J < \infty$  does not increase with  $n$ .

c) Recall that  $\Delta k_j = M < \infty$  is a fixed natural number, so we can write  $\Delta X_j = \sum_{i=1}^M \delta_i$ , where each  $\delta_i$  is a spacing between consecutive uniform order statistics. The marginal distribution of each  $\delta_i$  is  $\beta(1, n)$ , with closed form CDF from the Kumaraswamy (1980) family of distributions,

$$F_{\delta_i}(d) = 1 - (1 - d)^n \text{ for } 0 \leq d \leq 1.$$

Using this, we bound

$$\begin{aligned} P(\Delta X_j > a_n n^{-1} M^{1/2}) &\leq P\left(\exists i : \delta_i > \frac{a_n n^{-1} M^{1/2}}{M}\right) \\ &\leq MP(\delta_i > a_n n^{-1} M^{-1/2}) \\ &= M(1 - a_n n^{-1} M^{-1/2})^n \\ &= M \exp\{n \log(1 - a_n n^{-1} M^{-1/2})\} \\ &\leq M \exp\{-a_n M^{-1/2}\}, \end{aligned}$$

where the final line follows because  $\forall x \in (0, 1)$ ,  $\log(1 - x) < -x$ . If there are multiple fixed spacings in the parameter vector, this argument can be repeated and only the largest (fixed) spacing will enter the asymptotic rate.  $\square$

### Lemma for proving Theorem 1.3, with discussion and proof

Extensive manipulations will be performed on both the raw data,  $X_i$ , and the transformed data,  $U_i \equiv F(X_i)$ , so we introduce the following notation. From earlier,  $u_0 \equiv 0$  and  $u_{J+1} \equiv 1$ . For all  $j \in \{1, 2, \dots, J\}$ ,

$$k_j \equiv \lfloor (n+1)u_j \rfloor \qquad \epsilon_j \equiv (n+1)u_j - k_j,$$

where the  $\epsilon_j \in (0, 1)$  are interpolation weights as in (1.1). Also let  $\Delta \mathbf{k}$  denote the  $(J+1)$ -vector such that  $\Delta k_j = k_j - k_{j-1}$ , and

$$\begin{aligned} Y_j &\equiv U_{n:k_j} \sim \beta(k_j, n+1-k_j) & (A.10) \\ \Delta \mathbf{Y} &\equiv (Y_1, Y_2 - Y_1, \dots, 1 - Y_J) \sim \text{Dirichlet}(\Delta \mathbf{k}) \\ \Lambda_j &\equiv U_{n:k_{j+1}} - U_{n:k_j} \sim \beta(1, n) & \Omega_j \equiv U_{n:(k_j+m_j)} - U_{n:(k_j-m_j)} = \Omega_j^+ + \Omega_j^- \\ \Omega_j^- &\equiv U_{n:k_j} - U_{n:(k_j-m_j)} & \Omega_j^+ &\equiv U_{n:(k_j+m_j)} - U_{n:k_j} \end{aligned}$$

$$\begin{aligned}
Z_j &\equiv \sqrt{n}(Y_j - u_j) & V_j &\equiv \sqrt{n}[F^{-1}(Y_j) - F^{-1}(u_j)] \\
\mathbb{X} &\equiv \sum_{j=1}^J \psi_j F^{-1}(Y_j) & \mathbb{X}_0 &\equiv \sum_{j=1}^J \psi_j F^{-1}(u_j) \\
\mathbb{W} &\equiv \sqrt{n}(\mathbb{X} - \mathbb{X}_0) = \boldsymbol{\psi}' \mathbf{V} & \mathbb{W}_{\mathbf{u}} &\equiv \sqrt{n} \boldsymbol{\psi}' \left[ F^{-1}(\hat{Q}_U^I(\mathbf{u})) - F^{-1}(\mathbf{u}) \right], \\
\mathbb{W}_{\epsilon, \Lambda} &\equiv \mathbb{W} + n^{1/2} \sum_{j=1}^J \epsilon_j \psi_j \Lambda_j [Q'(u_j) + Q''(u_j)(Y_j - u_j)], \\
\hat{\mathbf{u}} &\equiv \{u_j^H(\tilde{\alpha}(\hat{\boldsymbol{\gamma}}))\}_{j=1}^J, & \mathbf{u}_0 &\equiv \{u_j^H(\tilde{\alpha}(\boldsymbol{\gamma}_0))\}_{j=1}^J,
\end{aligned}$$

with vectors indicated by bold and matrices (later) by underline. The values and distributions of  $\mathbf{k}$ ,  $\mathbf{Y}$ ,  $\Lambda$ ,  $\Omega$ ,  $\mathbf{V}$ ,  $\mathbf{Z}$ ,  $\mathbb{X}$ , and  $\mathbb{W}$  are all understood to vary with  $n$ , but the subscript will be omitted. Note that  $\Omega$  is defined with respect to the vector of spacing parameters,  $\mathbf{m}$ , used for nuisance parameter estimation.

**Lemma A.1.** *Suppose that Assumption A1.1 holds and that each element of  $\mathbf{Y}$  and  $\Lambda$  satisfies Condition  $\star(2 \log(n))$  as defined in Lemma 1.2.*

(i) *We have the following bounds in certainty:*

$$\begin{aligned}
|L^L - (\mathbb{X}_0 + n^{-1/2} \mathbb{W}_{\epsilon, \Lambda})| &= O(n^{-2} \log(n)^3), \\
|L^I - (\mathbb{X}_0 + n^{-1/2} \mathbb{W}_{\mathbf{C}, \Lambda})| &= O(n^{-2} \log(n)^3),
\end{aligned}$$

where  $\mathbf{C}$  is a  $J$ -vector of random interpolation coefficients as defined in Jones (2002). Each  $C_j \sim \beta(\epsilon_j, 1 - \epsilon_j)$ , and they are mutually independent.

(ii) *Define  $\underline{\mathcal{V}}$  and  $\underline{\mathcal{A}}$  as the  $J \times J$  matrices such that  $\underline{\mathcal{V}}_{i,j} = \min\{u_i, u_j\}(1 - \max\{u_i, u_j\})$  and  $\underline{\mathcal{A}} = \text{diag}\{f(F^{-1}(\mathbf{u}))\}$ . Also define*

$$\mathcal{V}_{\boldsymbol{\psi}} \equiv \boldsymbol{\psi}'(\underline{\mathcal{A}}^{-1} \underline{\mathcal{V}} \underline{\mathcal{A}}^{-1}) \boldsymbol{\psi} \in \mathbb{R}.$$

*For any vector  $\boldsymbol{\lambda}$  satisfying Condition  $\star(2 \log(n))$ ,*

$$\begin{aligned}
\sup_w |f_{\mathbb{W}_{\epsilon, \lambda} | \Lambda}(w | \boldsymbol{\lambda}) - \phi_{\mathcal{V}_{\boldsymbol{\psi}}}(w)| &= O(n^{-1/2} \log(n)), \\
\sup_{w \in \mathbb{R}} \left\| \frac{d[f_{\mathbb{W}_{\epsilon, \lambda} | \Lambda}(w | \boldsymbol{\lambda})]}{dw} - \frac{d[\phi_{\mathcal{V}_{\boldsymbol{\psi}}}(w)]}{dw} \right\|_{\infty} &= O(n^{-1/2} \log(n)).
\end{aligned}$$

For any value  $\tilde{\epsilon} \in [0, 1]^J$ ,

$$\frac{d^2 F_{W_{\epsilon, \Lambda} | \Lambda}(K | \boldsymbol{\lambda})}{d\epsilon_j^2} \Big|_{\epsilon = \tilde{\epsilon}} = n\psi_j^2 Q'(u_j)^2 \lambda_j^2 \left[ \frac{d[\phi_{\nu_\psi}(w)]}{dw} \right] + O(n^{-3/2} \log(n)^3).$$

(iii) Suppose  $\boldsymbol{\Omega}$  also satisfies Condition  $\star(2\log(n))$ . Then,

$$\begin{aligned} E[f_{\mathbb{W}_{\hat{\mathbf{u}} | \hat{\boldsymbol{\gamma}}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] | \hat{\boldsymbol{\gamma}})] &= O(1) && \text{(for } C_h^{1,a}), \\ \text{Var}(f_{\mathbb{W}_{\hat{\mathbf{u}} | \hat{\boldsymbol{\gamma}}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] | \hat{\boldsymbol{\gamma}})) &= O(n^{-2/3}) && \text{(for } C_h^{1,a}), \\ \sup |f'_{\mathbb{W}_{\hat{\mathbf{u}} | \hat{\boldsymbol{\gamma}}}(w | \hat{\boldsymbol{\gamma}})| &= O(1) && \text{(for } C_h^{1,b}), \\ E\left(\frac{\partial F_{\mathbb{W}_{\mathbf{t} | \hat{\boldsymbol{\gamma}}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] | \hat{\boldsymbol{\gamma}})}{\partial \mathbf{t}} \Big|_{\mathbf{t}=\mathbf{u}_0}\right) &= O(n^{1/2}) && \text{(for } C_h^{2,a}), \\ \text{Var}\left(\frac{\partial F_{\mathbb{W}_{\mathbf{t} | \hat{\boldsymbol{\gamma}}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] | \hat{\boldsymbol{\gamma}})}{\partial \mathbf{t}} \Big|_{\mathbf{t}=\mathbf{u}_0}\right) &= O(n^{1/3}) && \text{(for } C_h^{2,a}), \\ \frac{\partial^2 F_{\mathbb{W}_{\mathbf{t} | \hat{\boldsymbol{\gamma}}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] | \hat{\boldsymbol{\gamma}})}{\partial \mathbf{t} \partial \mathbf{t}'} \Big|_{\mathbf{t}=\tilde{\mathbf{u}}} &= O(n) && \text{(for } C_h^{2,b}). \end{aligned}$$

*Proof.* (i) The object  $L^L$  may be rewritten as

$$\begin{aligned} L^L &= \sum_{j=1}^J \psi_j \left( Q(Y_j) + \epsilon_j [Q(Y_j + \Lambda_j) - Q(Y_j)] \right) \\ &= \mathbb{X}_0 + \sum_{j=1}^J \psi_j \left( Q(Y_j) - Q(u_j) + \epsilon_j [Q(Y_j + \Lambda_j) - Q(Y_j)] \right) \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W} + \sum_{j=1}^J \psi_j \epsilon_j \left[ Q'(Y_j) \Lambda_j + \frac{Q''(\tilde{y}_j)}{2} \Lambda_j^2 \right] \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W} + \sum_{j=1}^J \psi_j \epsilon_j \left[ Q'(u_j) \Lambda_j + Q''(u_j) [Y_j - u_j] \Lambda_j \right. \\ &\quad \left. + \underbrace{\frac{Q^{(3)}(\tilde{u}_j)}{2} [Y_j - u_j]^2 \Lambda_j}_{\nu_{j,1}^L} + \underbrace{\frac{Q''(\tilde{y}_j)}{2} \Lambda_j^2}_{\nu_{j,2}^L} \right] \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W}_{\epsilon, \Lambda} + \sum_{j=1}^J \psi_j \epsilon_j \left[ \nu_{j,1}^L + \nu_{j,2}^L \right], \end{aligned}$$

where  $\forall j$ ,  $\tilde{y}_j \in (Y_j, Y_j + \Lambda_j)$  and  $\tilde{u}_j$  is between  $u_j$  and  $Y_j$ . Applying Assumption A1.1 and the assumed Condition  $\star(2\log(n))$  gives the desired rate on

the  $\nu^L$  terms. The argument is very similar for  $L^I$ , but must also lean on the random interpolation representation of the ‘ideal’ uniform order statistics given in Jones (2002). The random interpolation vector of coefficients is denoted  $\mathbf{C}$ . In our notation,

$$\begin{aligned}\hat{Q}_U^I(u_j) &\stackrel{L}{=} C_j U_{n:k_j+1} + (1 - C_j) U_{n:k_j} = Y_j + C_j \Lambda_j, \\ C_j &\sim \beta(\epsilon_j, 1 - \epsilon_j), \quad C_j \perp (\mathbf{Y}, \mathbf{\Lambda}, \mathbf{C}_{-j}).\end{aligned}$$

This allows us to write

$$\begin{aligned}L^I &= \sum_{j=1}^J \psi_j \left( Q(Y_j) + [Q(Y_j + C_j \Lambda_j) - Q(Y_j)] \right) \\ &= \mathbb{X}_0 + \sum_{j=1}^J \psi_j \left( Q(Y_j) - Q(u_j) + [Q(Y_j + C_j \Lambda_j) - Q(Y_j)] \right) \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W} + \sum_{j=1}^J \psi_j \left[ Q'(Y_j) C_j \Lambda_j + \frac{Q''(\tilde{y}_j)}{2} (C_j \Lambda_j)^2 \right] \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W} + \sum_{j=1}^J \psi_j \left[ Q'(u_j) C_j \Lambda_j + Q''(u_j) [Y_j - u_j] C_j \Lambda_j \right. \\ &\quad \left. + \underbrace{\frac{Q^{(3)}(\tilde{u}_j)}{2} [Y_j - u_j]^2 C_j \Lambda_j}_{\nu_{j,1}^I} + \underbrace{\frac{Q''(\tilde{y}_j)}{2} (C_j \Lambda_j)^2}_{\nu_{j,2}^I} \right] \\ &= \mathbb{X}_0 + n^{-1/2} \mathbb{W}_{\mathbf{C}, \mathbf{\Lambda}} + \sum_{j=1}^J \psi_j [\nu_{j,1}^I + \nu_{j,2}^I].\end{aligned}$$

(ii) Since  $\mathbf{\Lambda}$  contains finite spacings, we cannot apply Lemma 1.2(a) to expand its density. Instead, we use the convenient representation

$$\Lambda_j \sim \beta(1, n), \quad \frac{\Lambda_j}{1 - \sum_{k=1}^{j-1} \Lambda_k} \bigg| \left( \Lambda_1, \dots, \Lambda_{j-1} \right) \sim \beta(1, n - j + 1).$$

Log PDFs are given by

$$\begin{aligned}\log f_{\Lambda_j}(\lambda_j) &= \log(\Gamma(n + 1)) - \log(\Gamma(n)) + (n - 1) \log(1 - \lambda_j) \\ &= \log(n) - n \lambda_j + O(n^{-1} \log(n)), \\ \log f_{\Lambda_j | (\Lambda_1, \dots, \Lambda_{j-1})} &= \log(\Gamma(n + 2 - j)) - \log(\Gamma(n + 1 - j))\end{aligned}$$



$$\begin{aligned}
& + n \log \left( 1 - \lambda_j \left( 1 - \sum_{k=1}^{j-1} \lambda_k \right) \right) \\
& = \log(n) - n\lambda_j + O(n^{-1} \log(n)),
\end{aligned}$$

where the above rate follows from imposing Condition  $\star(2 \log(n))$  on all values of  $\mathbf{\Lambda}$ . Taking appropriate products of conditional densities leads to the convenient representation

$$\log f_{\mathbf{\Lambda}}(\boldsymbol{\lambda}) = J \log(n) - n \sum_{j=1}^J \lambda_j + O(n^{-1} \log(n)). \quad (\text{A.11})$$

The joint density of  $\{\mathbf{Y}, \mathbf{\Lambda}\}$  is even more problematic because it is not a Dirichlet distribution. However, we define the linear transformation  $\tau : \{\mathbf{Y}, \mathbf{\Lambda}\} \rightarrow \mathbf{Y}\mathbf{\Lambda} \equiv \{\Delta Y_1, \Lambda_1, \Delta Y_2 - \Lambda_1, \dots, \Lambda_J, \Delta Y_{J+1} - \Lambda_J\}'$ . Row operations show this transformation's matrix representation to have unit determinant ( $|\tau'| = 1$ ). We now have

$$\mathbf{Y}\mathbf{\Lambda} \sim \text{Dirichlet}(\{\Delta k_1, 1, \Delta k_2 - 1, \dots, 1, \Delta k_{J+1} - 1\}').$$

Unfortunately, the spacings still do not diverge, so we still may not apply Lemma 1.2(a). However, we may write the log PDF as

$$\begin{aligned}
& \log(f_{\{\mathbf{Y}, \mathbf{\Lambda}\}}(\mathbf{y}, \boldsymbol{\lambda})) \\
& = \log(f_{\mathbf{Y}\mathbf{\Lambda}}(\tau(\mathbf{y}, \boldsymbol{\lambda}))) \cdot |\tau'|^{-1} \\
& = \log(\Gamma(n+1)) + (\Delta k_1 - 1) \log(\Delta y_1) - \log(\Gamma[\Delta k_1]) \\
& \quad + \sum_{j=2}^{J+1} \left[ (\Delta k_j - 2) \log(\Delta y_j - \lambda_{j-1}) - \log(\Gamma[\Delta k_j - 1]) \right].
\end{aligned} \quad (\text{A.12})$$

Combining (A.11) and (A.12), we have

$$\begin{aligned}
& \log f_{\mathbf{Y}|\mathbf{\Lambda}} \\
& = \log(f_{\{\mathbf{Y}, \mathbf{\Lambda}\}}(\mathbf{y}, \boldsymbol{\lambda})) - \log f_{\mathbf{\Lambda}}(\boldsymbol{\lambda}) \\
& = \log(\Gamma(n+1)) + (\Delta k_1 - 1) \log(\Delta y_1) - \log(\Gamma[\Delta k_1]) \\
& \quad + \sum_{j=2}^{J+1} \left[ (\Delta k_j - 2) \log(\Delta y_j - \lambda_{j-1}) - \log(\Gamma[\Delta k_j - 1]) - \log n + n\lambda_{j-1} \right]
\end{aligned}$$

$$\begin{aligned}
&= \log f_{\mathbf{Y}}(\mathbf{y}) + \sum_{j=2}^{J+1} \left[ (\Delta k_j - 2) \log(\Delta y_j - \lambda_{j-1}) - (\Delta k_j - 1) \log(\Delta y_j) \right. \\
&\quad \left. + [\log(\Gamma[\Delta k_j]) - \log(\Gamma[\Delta k_j - 1])] - \log n + n\lambda_{j-1} \right] \\
&= \log f_{\mathbf{Y}}(\mathbf{y}) \\
&\quad + \sum_{j=2}^{J+1} \left[ \left\{ (\Delta k_j - 2) [\log(\Delta y_j - \lambda_{j-1}) - \log(\Delta y_j)] + n\lambda_{j-1} \right\} \right. \quad (\text{A.13}) \\
&\quad \left. + \left\{ [\log(\Gamma[\Delta k_j]) - \log(\Gamma[\Delta k_j - 1])] - \log n - \log \Delta y_j \right\} \right] \\
&= \log f_{\mathbf{Y}}(\mathbf{y}) \\
&\quad + \sum_{j=2}^{J+1} \left[ \left\{ (\Delta k_j - 2) [-\lambda_{j-1}/\Delta y_j + O(\lambda_{j-1}^2/\Delta y_j^2)] + n \log \lambda_{j-1} \right\} \right. \\
&\quad \left. + \left\{ [\log(\Delta k_j) + O(\Delta k_j^{-1})] - \log n - \log \Delta y_j \right\} \right] \\
&= \log \phi_{\nu/n}(\mathbf{y}) + O(n^{-1/2} \log(n)),
\end{aligned}$$

where the final result follows by applying  $\star(2 \log(n))$  to the analysis of each braced term and applying Lemma 1.2(a) directly to the density of  $\mathbf{Y}$ . Differentiating (A.13) and applying Lemma 1.2(a) to the density derivative of  $\mathbf{Y}$  yields

$$\begin{aligned}
\frac{d \log f_{\mathbf{Y}|\Lambda}(\mathbf{y}; \boldsymbol{\lambda})}{d\mathbf{y}} &= \frac{d \log f_{\mathbf{Y}}(\mathbf{y})}{d\mathbf{y}} + \sum_{j=2}^{J+1} \left[ \frac{(\Delta k_j - 2)\lambda_{j-1}}{\Delta y_j(\Delta y_j - \lambda_{j-1})} - \frac{1}{\Delta y_j} \right] \\
&= \frac{d \log f_{\mathbf{Y}}(\mathbf{y})}{d\mathbf{y}} + \sum_{j=2}^{J+1} \left[ \frac{(\Delta k_j - 2)\lambda_{j-1} - (\Delta y_j - \lambda_{j-1})}{\Delta y_j(\Delta y_j - \lambda_{j-1})} \right] \\
&= \frac{d \log \phi_{\nu/n}(\mathbf{y})}{d\mathbf{y}} + O(\log n).
\end{aligned}$$

For convenience, we define and calculate

$$\begin{aligned}
H_{Y_j} &\equiv K - n^{1/2} \sum_{j=1}^J \epsilon_j \Lambda_j [Q'(u_j) + Q''(u_j)(Y_j - u_j)] \\
&= K - O(n^{-1/2} \log(n)), \\
\frac{dH_{Y_j}}{d\epsilon_j} &= -n^{1/2} \Lambda_j [Q'(u_j) + Q''(u_j)(Y_j - u_j)]
\end{aligned}$$

$$\begin{aligned}
&= -n^{1/2}\Lambda_j Q'(u_j) + O(n^{-1}\log(n)), \\
\frac{d^2 H_{Y_j}}{d\epsilon_j d\epsilon_k} &= 0.
\end{aligned}$$

Since dependence of  $H$  on the value of  $Y_j$  is shown to be of lower-order importance, we neglect the subscript and write

$$\begin{aligned}
F_{W_{\epsilon,\Lambda}|\Lambda}(K|\boldsymbol{\lambda}) &= \int_{-\infty}^H f_{\mathbb{W}|\Lambda}(w|\boldsymbol{\lambda})dw, \\
\frac{dF_{W_{\epsilon,\Lambda}|\Lambda}(K|\boldsymbol{\lambda})}{d\epsilon_j} &= f_{\mathbb{W}|\Lambda}(H|\boldsymbol{\lambda})\frac{dH}{d\epsilon_j}, \\
\frac{d^2 F_{W_{\epsilon,\Lambda}|\Lambda}(K|\boldsymbol{\lambda})}{d\epsilon_j d\epsilon_k} &= f_{\mathbb{W}|\Lambda}(H|\boldsymbol{\lambda})\frac{d^2 H}{d\epsilon_j d\epsilon_k} + f'_{\mathbb{W}|\Lambda}(H|\boldsymbol{\lambda})\left(\frac{dH}{d\epsilon_j}\right)\left(\frac{dH}{d\epsilon_k}\right), \\
\frac{d^2 F_{W_{\epsilon,\Lambda}|\Lambda}(K|\boldsymbol{\lambda})}{d\epsilon_j^2} &= f'_{\mathbb{W}|\Lambda}(H|\boldsymbol{\lambda})\left(\frac{dH}{d\epsilon_j}\right)^2 \\
&= [f'_{\mathbb{W}|\Lambda}(K|\boldsymbol{\lambda}) + O(n^{-1/2}\log(n))] \\
&\quad \times [-n^{1/2}\lambda_j Q'(u_j) + O(n^{-1}\log(n))]^2 \\
&= [f'_{\mathbb{W}|\Lambda}(K|\boldsymbol{\lambda}) + O(n^{-1/2}\log(n))] \\
&\quad \times [n\lambda_j^2 Q'(u_j)^2 + O(n^{-1/2}\log(n))\lambda_j Q'(u_j) \\
&\quad \quad + O(n^{-2}\log(n))].
\end{aligned}$$

(iii) These are taken without proof for the time being. It seems likely that these statements will work out since the corresponding CPE of  $O(n^{-2/3})$  is the same as in Chapter 3, which uses the same quantile spacing estimator and bandwidth rate for the nuisance parameters.  $\square$

### Theorem 1.3 proof

*Proof.* We prove all three parts under the assumption that the realized values of  $\mathbf{Y}$ ,  $\boldsymbol{\Omega}$ , and  $\boldsymbol{\lambda}$  all adhere to Condition  $\star(2\log(n))$ . By application of Lemma 1.2(b-c), this induces at most  $O(n^{-2})$  error in our calculations. This will prove to be asymptotically insignificant, so we ignore it going forward.

(i) Applying Lemma A.1(i) to the CDF of  $L^L$ ,

$$\begin{aligned} P(L^L < \mathbb{X}_0 + n^{-1/2}K) &= P(\mathbb{W}_{\epsilon, \Lambda} < K + O(n^{-3/2} \log(n)^3)) \\ &= P(\mathbb{W}_{\epsilon, \Lambda} < K) + O(n^{-3/2} \log(n)^3), \end{aligned}$$

where the error can be pulled out of the probability statement because  $\mathbb{W}_{\epsilon, \Lambda}$  has been shown to have a bounded PDF in Lemma A.1(ii). We may then write the CDF of  $L^L$  as

$$= \int_{[0, 2n^{-1} \log(n)]^J} P(W_{\epsilon, \lambda} < K \mid \lambda) dF_{\Lambda}(\lambda) + O(n^{-3/2} \log(n)^3).$$

By a similar series of manipulations,

$$\begin{aligned} P(L^I < \mathbb{X}_0 + n^{-1/2}K) &= P(\mathbb{W}_{\mathbf{C}, \Lambda} < K + O(n^{-3/2} \log(n)^3)) \\ &= P(\mathbb{W}_{\mathbf{C}, \Lambda} < K) + O(n^{-3/2} \log(n)^3) \\ &= \int_{[0, 2n^{-1} \log(n)]^J} \int_{[0, 1]^J} P(W_{\mathbf{c}, \lambda} < K \mid \lambda) dF_{\mathbf{C}}(\mathbf{c}) dF_{\Lambda}(\lambda) \\ &\quad + O(n^{-3/2} \log(n)^3). \end{aligned}$$

The CDF difference between the two distributions is given by

$$\begin{aligned} &P(L^I < \mathbb{X}_0 + n^{-1/2}K) - P(L^L < \mathbb{X}_0 + n^{-1/2}K) \\ &= \int_{[0, 2n^{-1} \log(n)]^J} \int_{[0, 1]^J} [F_{W_{\mathbf{c}, \Lambda} | \Lambda}(K | \lambda) - F_{W_{\epsilon, \Lambda} | \Lambda}(K | \lambda)] dF_{\mathbf{C}}(\mathbf{c}) dF_{\Lambda}(\lambda) \\ &\quad + O(n^{-3/2} \log(n)^3) \\ &= \int_{[0, 2n^{-1} \log(n)]^J} \int_{[0, 1]^J} \left[ (\mathbf{c} - \epsilon)' \frac{dF_{W_{\mathbf{c}, \Lambda} | \Lambda}(K | \lambda)}{d\mathbf{c}} \Big|_{\mathbf{C}=\epsilon} \right. \\ &\quad \left. + \frac{1}{2} (\mathbf{c} - \epsilon)' \frac{d^2 F_{W_{\mathbf{c}, \Lambda} | \Lambda}(K | \lambda)}{d\mathbf{c} \cdot d\mathbf{c}'} \Big|_{\mathbf{C}=\tilde{\mathbf{C}} \in [0, 1]^J} (\mathbf{c} - \epsilon) \right] dF_{\mathbf{C}}(\mathbf{c}) dF_{\Lambda}(\lambda) \\ &\quad + O(n^{-3/2} \log(n)^3). \end{aligned}$$

Since  $E(\mathbf{C}) = \boldsymbol{\epsilon}$ , the first term zeroes out. Since the elements of  $\mathbf{C}$  are mutually independent, the off-diagonal elements of the Hessian in the quadratic term also zero out. Then we apply Lemma A.1(ii) to the Hessian, and simplify the remainder terms using  $E(\Lambda_j) = O(n^{-1})$  and  $E(\Lambda_j^2) = O(n^{-2})$  (Kumaraswamy, 1980). Using  $\text{Var}(C_j) = \epsilon_j(1 - \epsilon_j)/2$  also, this leaves us with

$$= \frac{n}{2} \left( \frac{d[\phi_{\mathcal{V}_\psi}(w)]}{dw} \right) \sum_{j=1}^J \left[ [\psi_j Q'(u_j)]^2 \frac{\epsilon_j(1 - \epsilon_j)}{2} \int_{[0, 2n^{-1} \log(n)]^J} \lambda_j^2 dF_{\boldsymbol{\Lambda}}(\boldsymbol{\lambda}) \right] + O(n^{-3/2} \log(n)^3).$$

The proof is completed by noting that second raw moments of  $\boldsymbol{\Lambda}$  are such that  $\forall j, E(\Lambda_j^2) = 2/[(n+1)(n+2)] = 2n^{-2} + O(n^{-3})$  (Kumaraswamy, 1980).

- (ii) The first result of this part is obtained by solving a first order condition and evaluating the expression obtained in part (i) at the solution  $K = \sqrt{\mathcal{V}_\psi}$ . The second result comes directly from the uniform bound on densities obtained in Lemma A.1(ii).
- (iii) This part follows because the convolution operation (on the  $L_\infty$  space of CDFs) has norm bounded above by one. Specifically,

$$\begin{aligned} & P\left(L_X^L + L_Y^L < K\right) \\ &= \int_{\mathbb{R}} P(L_Y^L < K - x) dF_{L_X^L}(x) \\ &= \int_{\mathbb{R}} P(L_Y^I < K - x) dF_{L_X^L}(x) \\ &\quad \underbrace{O(n^{-1}) \text{ by part (ii)}} \\ &+ \int_{\mathbb{R}} \left[ P(L_Y^L < K - x) - P(L_Y^I < K - x) \right] dF_{L_X^L}(x) \\ &= \int_{\mathbb{R}} P(L_X^L < K - x) dF_{L_Y^I}(x) + O(n^{-1}) \\ &= \int_{\mathbb{R}} P(L_X^I < K - x) dF_{L_Y^I}(x) \\ &\quad \underbrace{O(n^{-1}) \text{ by part (ii)}} \\ &+ \int_{\mathbb{R}} \left[ P(L_X^L < K - x) - P(L_X^I < K - x) \right] dF_{L_Y^I}(x) + O(n^{-1}) \end{aligned}$$

$$= P(L_X^I + L_Y^I < K) + O(n^{-1}). \quad \square$$

### Lemma 1.4 proof

*Proof.* Notation from Hutson (1999) is used in the following proof:  $u_1$  is  $u_j^l$ ,  $u_2$  is  $u_j^h$ , and  $p$  is  $u_j$ . After approximating the beta distribution with a normal, equation (7) in Hutson (1999) determines  $u_1$  by (slightly abusing notation)

$$P\{\mathcal{N}(u_1, u_1(1-u_1)/n) < p\} + O(n^{-1/2}) = 1 - \alpha/2,$$

which can be solved explicitly for  $u_1$  with error  $O(n^{-1})$  as

$$\begin{aligned} 1 - \alpha/2 &= \Phi\left(\frac{\sqrt{n}(p-u_1)}{\sqrt{u_1(1-u_1)}}\right) + O(n^{-1/2}), \\ z_{1-\alpha/2} &\equiv \Phi^{-1}(1 - \alpha/2) = \frac{\sqrt{n}(p-u_1)}{\sqrt{u_1(1-u_1)}} + O(n^{-1/2}), \\ u_1 &= p - z_{1-\alpha/2}\sqrt{u_1(1-u_1)/n} + O(n^{-1}), \\ (p-u_1)^2 &= z_{1-\alpha/2}^2 u_1(1-u_1)/n + O(n^{-1}), \\ u_1^2 - u_1(2p) + p^2 &= u_1^2(-z_{1-\alpha/2}^2/n) + u_1(z_{1-\alpha/2}^2/n) + O(n^{-1}), \\ 0 &= u_1^2(1 + z_{1-\alpha/2}^2/n) + u_1(-2p - z_{1-\alpha/2}^2/n) + p^2 + O(n^{-1}). \end{aligned}$$

We can use the quadratic formula (choosing the negative root—otherwise we get  $u_2$ ), and temporarily simplify notation with  $z \equiv z_{1-\alpha/2}$ ,  $u \equiv u_1$ . Also recalling that for a small  $\epsilon$ ,  $(X + \epsilon)^{1/2} = X^{1/2} + O(\epsilon)$  and  $(1 + \epsilon)^{-1} = 1 - \epsilon + O(\epsilon^2)$ ,

$$\begin{aligned} u &= \frac{(2p + z^2/n) - \sqrt{(2p + z^2/n)^2 - 4(1 + z^2/n)p^2}}{2(1 + z^2/n)} + O(n^{-1}) \\ &= \frac{p}{1 + z^2/n} + O(n^{-1}) - \frac{\sqrt{4p^2 + 4pz^2/n + z^4/n^2 - 4p^2 - 4p^2z^2/n}}{2(1 + z^2/n)} + O(n^{-1}) \\ &= p(1 - O(n^{-1})) - \frac{\sqrt{4p(1-p)z^2/n + z^4/n^2}}{2(1 + z^2/n)} + O(n^{-1}) \\ &= p - \frac{2z\sqrt{p(1-p)/n} + O(n^{-2})}{2(1 + z^2/n)} + O(n^{-1}) \\ &= p - z\sqrt{p(1-p)/n}(1 - O(n^{-1})) + O(n^{-1}) \\ &= p - z\sqrt{p(1-p)/n} + O(n^{-1}). \end{aligned}$$

This approximation applies similarly to  $u_2$ . □

## Expansions of $T_h$ , $T_l$ , $C_h$ , $C_l$ in Theorem 1.5

### CPE from Taylor Approximations: $T_h$ , $T_l$

*Proof.* Up to third order terms,  $T_h(\tilde{\alpha})$  and  $T_l(\tilde{\alpha})$  are given by

$$\begin{aligned}
T_h(\tilde{\alpha}) &= P \left( \overbrace{\sum_{j=1}^J \psi_j \frac{\hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j}{f(F^{-1}(u_j))} > \sum_{j=1}^J \psi_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^3} \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j \right)^2}^{T_{H,1}} \right) \\
&\quad - P \left( \overbrace{\sum_{j=1}^J \psi_j \left( \frac{\hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j}{f(F^{-1}(u_j))} \right) > 0}^{T_{H,2}} \right), \\
T_l(\tilde{\alpha}) &= P \left( \overbrace{\sum_{j=1}^J \psi_j \frac{\hat{Q}_U^I[u_j^L(\tilde{\alpha})] - u_j}{f(F^{-1}(u_j))} < \sum_{j=1}^J \psi_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^3} \left( \hat{Q}_U^I[u_j^L(\tilde{\alpha})] - u_j \right)^2}^{T_{L,1}} \right) \\
&\quad - P \left( \overbrace{\sum_{j=1}^J \psi_j \left( \frac{\hat{Q}_U^I[u_j^L(\tilde{\alpha})] - u_j}{f(F^{-1}(u_j))} \right) < 0}^{T_{L,2}} \right).
\end{aligned}$$

For this demonstration, we introduce the scaled and centered

$$\Delta_j^H \equiv \sqrt{n} \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j^H(\tilde{\alpha}) \right), \quad \Delta_j^L \equiv \sqrt{n} \left( \hat{Q}_U^I[u_j^L(\tilde{\alpha})] - u_j^L(\tilde{\alpha}) \right).$$

Since  $u_j^L(\tilde{\alpha}), u_j^H(\tilde{\alpha}) = u_j + O(n^{-1/2})$ , both  $\mathbf{\Delta}^H$  and  $\mathbf{\Delta}^L$  have densities converging to the same multivariate normal PDF (as shown in Lemma A.1(ii)), which we denote by  $f_{\mathbf{\Delta}^I}(\cdot)$ . We also define

$$\begin{aligned}
\Delta_0^H &\equiv \sum_{j=1}^J \psi_j \gamma_j \sqrt{n} (u_j^H(\tilde{\alpha}) - u_j), \\
\Delta_0^L &\equiv \sum_{j=1}^J \psi_j \gamma_j \sqrt{n} (u_j^L(\tilde{\alpha}) - u_j) = -\Delta_0^H + O(n^{-1/2}),
\end{aligned}$$

where the last equality follows by repeated application of Lemma 1.4.

There is an important symmetry between  $T_h(\tilde{\alpha})$  and  $T_l(\tilde{\alpha})$ . For convenience,

we demonstrate an expansion for  $T_h(\tilde{\alpha})$  and then state a parallel result for  $T_l(\tilde{\alpha})$ :

$$\begin{aligned}
T_{H,2} &= P \left[ \hat{Q}_U^I[u_1^H(\tilde{\alpha})] - u_1 > - \sum_{j=2}^J \psi_j \gamma_j \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j \right) \right] \\
&= P \left[ \hat{Q}_U^I[u_1^H(\tilde{\alpha})] - u_1^H + u_1^H - u_1 \right. \\
&\quad \left. > - \sum_{j=2}^J \psi_j \gamma_j \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j^H + u_j^H - u_j \right) \right] \\
&= P \left[ \sqrt{n} \left( \hat{Q}_U^I[u_1^H(\tilde{\alpha})] - u_1^H \right) + \sqrt{n} (u_1^H - u_1) \right. \\
&\quad \left. > - \sum_{j=2}^J \psi_j \gamma_j \sqrt{n} \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j^H \right) - \sum_{j=2}^J \psi_j \gamma_j \sqrt{n} (u_j^H - u_j) \right] \\
&= P \left[ \Delta_1^H > - \sum_{j=2}^J \psi_j \gamma_j \Delta_j^H - \Delta_0^H \right] \\
&\equiv P[\Delta_1^H > \pi^{H,2}(\Delta_{-1}^H)],
\end{aligned}$$

and similarly,

$$\begin{aligned}
T_{H,1} &\equiv P \left( \sum_{j=1}^J \psi_j \frac{\hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j}{f(F^{-1}(u_j))} > \sum_{j=1}^J \psi_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^3} \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j \right)^2 \right) \\
&= P \left( \hat{Q}_U^I[u_1^H(\tilde{\alpha})] - u_1^H + u_1^H - u_1 + \sum_{j=2}^J \psi_j \gamma_j \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j^H + u_j^H - u_j \right) \right. \\
&\quad > \frac{f'(F^{-1}(u_1))}{2f(F^{-1}(u_1))^2} \left( \hat{Q}_U^I[u_1^H(\tilde{\alpha})] - u_1^H + u_1^H - u_1 \right)^2 \\
&\quad \left. + \sum_{j=2}^J \psi_j \gamma_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^2} \left( \hat{Q}_U^I[u_j^H(\tilde{\alpha})] - u_j^H + u_j^H - u_j \right)^2 \right) \\
&= P \left( \Delta_1^H + \sum_{j=2}^J \psi_j \gamma_j \Delta_j^H + \Delta_0^H \right. \\
&\quad > \frac{f'(F^{-1}(u_1))}{2f(F^{-1}(u_1))^2} n^{-1/2} \left[ (\Delta_1^H)^2 + (n^{1/2}[u_1^H - u_1])^2 + 2\Delta_1^H \sqrt{n}(u_1^H - u_1) \right] \\
&\quad \left. + \sum_{j=2}^J \psi_j \gamma_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^2} n^{-1/2} \left[ \Delta_j^H + \sqrt{n}(u_j^H - u_j) \right]^2 \right).
\end{aligned}$$



This is conveniently restated as

$$\begin{aligned}
T_{H,1} &= E_{\Delta_{-1}^H} \left[ P \left( a(\Delta_1^H)^2 + b\Delta_1^H + c > 0 \mid \Delta_{-1}^H \right) \right], \\
\text{where } a &= -\frac{f'(F^{-1}(u_1))}{2f(F^{-1}(u_1))^2} n^{-1/2}, \\
b &= 1 - \frac{f'(F^{-1}(u_1))}{f(F^{-1}(u_1))^2} (u_1^H(\tilde{\alpha}) - u_1) \\
&= 1 - \frac{f'(F^{-1}(u_1))}{f(F^{-1}(u_1))^2} z_{1-\tilde{\alpha}} u_1 (1 - u_1) n^{-1/2} + O(n^{-1}) \\
&\equiv b_0 + O(n^{-1}), \\
\text{and } c &= -\pi^{H,2}(\Delta_{-1}^H) - \frac{f'(F^{-1}(u_1))}{2f(F^{-1}(u_1))^2} n^{-1/2} (n^{1/2}[u_1^H - u_1])^2 \\
&\quad - \sum_{j=2}^J \psi_j \gamma_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^2} n^{-1/2} (\Delta_j^H + n^{1/2}[u_j^H - u_j])^2,
\end{aligned}$$

which is, after approximating the values of  $u_j^H(\tilde{\alpha})$  by Lemma 1.4,

$$\begin{aligned}
&= -\pi^{H,2}(\Delta_{-1}^H) - n^{-1/2} \left[ \frac{f'(F^{-1}(u_1))}{2f(F^{-1}(u_1))^2} (n^{1/2}[z_{1-\tilde{\alpha}} u_1 (1 - u_1)])^2 \right. \\
&\quad \left. + \sum_{j=2}^J \psi_j \gamma_j \frac{f'(F^{-1}(u_j))}{2f(F^{-1}(u_j))^2} (\Delta_j^H + n^{1/2}[z_{1-\tilde{\alpha}} u_j (1 - u_j)])^2 \right] + O(n^{-1}) \\
&\equiv -\pi^{H,2}(\Delta_{-1}^H) + c^* + O(n^{-1}).
\end{aligned}$$

The roots, denoted  $r_-$  and  $r_+$ , are given by the quadratic formula. Let

$$\begin{aligned}
\pi^{H,1}(\mathbf{v}_{-1}) &\equiv r_+ = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = -\frac{c}{b} - \frac{ac^2}{b^3} + O(a^2c^3) \\
&= \frac{\pi^{H,2}(\mathbf{v}_{-1})}{b_0} - \frac{c^*}{b_0} - \frac{a[\pi^{H,2}(\mathbf{v}_{-1})]^2}{b_0^3} + O(n^{-1}),
\end{aligned}$$

which follows from a Taylor expansion of the quadratic formula around the value  $c = 0$ . Now up to an exponentially decaying error term, we can write<sup>3</sup>

$$T_{H,1}(\tilde{\alpha}) = E_{\Delta_{-1}^H} \left[ P(\Delta_1^H > r_+ \mid \Delta_{-1}^H) \right].$$

---

<sup>3</sup>This follows from analysis of two cases. First, if  $f'(F^{-1}(u_1)) > 0$ , then  $a < 0$ ,  $r_+ < r_-$ , and  $P(a(\Delta_1^H)^2 + b\Delta_1^H + c > 0 \mid \Delta_{-1}^H) = P(r_+ < \Delta_1^H < r_- \mid \Delta_{-1}^H) \approx P(r_+ < \Delta_1^H \mid \Delta_{-1}^H)$ , where the approximation error decays exponentially because  $r_- \asymp n^{1/2}$ ,  $r_- > 0$ , and  $\Delta$  has a multivariate normal limiting distribution. In the alternative case,  $f'(F^{-1}(u_1)) < 0$  and  $a > 0$ , leading to  $r_- < r_+$  and  $P(a(\Delta_1^H)^2 + b\Delta_1^H + c > 0 \mid \Delta_{-1}^H) = P(\{r_+ < \Delta_1^H\} \cup \{\Delta_1^H < r_-\} \mid \Delta_{-1}^H) \approx P(r_+ < \Delta_1^H \mid \Delta_{-1}^H)$ , where the approximation follows because  $r_- < 0$  and again  $r_- \asymp n^{1/2}$ .

Then the discrepancy between  $\pi^{H,1}(\mathbf{v}_{-1})$  and  $\pi^{H,2}(\mathbf{v}_{-1})$  is the source of the CPE represented by  $T_h$ . By our above calculations, we have the following important result:

$$\begin{aligned} \pi^{H,1}(\mathbf{v}_{-1}) - \pi^{H,2}(\mathbf{v}_{-1}) &= \frac{\pi^{H,2}(\mathbf{v}_{-1})(1 - b_0)}{b_0} - \frac{c^*}{b_0} - \frac{a[\pi^{H,2}(\mathbf{v}_{-1})]^2}{b_0^3} + O(n^{-1}) \\ &= O(n^{-1/2}). \end{aligned} \quad (\text{A.14})$$

Returning to our terms of interest,

$$\begin{aligned} T_h(\tilde{\alpha}) &= \int_{\mathbb{R}^{J-1}} \int_{\pi^{H,1}(\mathbf{v}_{-1})}^{\pi^{H,2}(\mathbf{v}_{-1})} f_{\Delta_1^H | \Delta_{-1}^H}(t | \mathbf{v}_{-1}) dt \cdot f_{\Delta_{-1}^H}(\mathbf{v}_{-1}) d\mathbf{v}_{-1} \\ &= \int_{\mathbb{R}^{J-1}} [\pi^{H,2}(\mathbf{v}_{-1}) - \pi^{H,1}(\mathbf{v}_{-1})] f_{\Delta_1^H | \Delta_{-1}^H}(\pi^{H,2}(\mathbf{v}_{-1}) | \mathbf{v}_{-1}) \cdot f_{\Delta_{-1}^H}(\mathbf{v}_{-1}) d\mathbf{v}_{-1} \\ &\quad + O(n^{-1}) \\ &= \int_{\mathbb{R}^{J-1}} [\pi^{H,2}(\mathbf{v}_{-1}) - \pi^{H,1}(\mathbf{v}_{-1})] f_{\Delta_1^I | \Delta_{-1}^I}(\pi^{H,2}(\mathbf{v}_{-1}) | \mathbf{v}_{-1}) \cdot f_{\Delta_{-1}^I}(\mathbf{v}_{-1}) d\mathbf{v}_{-1} \\ &\quad + O(n^{-1}) \\ &= O(n^{-1/2}), \end{aligned}$$

where the rate follows by (A.14) in combination with the bounded limiting PDF of  $\Delta^I$ . From a parallel set of definitions and steps (now the  $\pi$  variables are upper bounds) for  $T_l(\tilde{\alpha})$ , it is possible to write

$$\pi^{L,2}(\mathbf{v}_{-1}) = - \sum_{j=2}^J \psi_j \gamma_j v_j - \Delta_0^L = \pi^{H,2}(\mathbf{v}_{-1}) + O(n^{-1/2}), \quad (\text{A.15})$$

$$\begin{aligned} \pi^{L,1}(\mathbf{v}_{-1}) - \pi^{L,2}(\mathbf{v}_{-1}) &= \frac{\pi^{L,2}(\mathbf{v}_{-1})(1 - b_0)}{b_0} - \frac{c^*}{b_0} - \frac{a[\pi^{L,1}(\mathbf{v}_{-1})]^2}{b_0^3} + O(n^{-1}) \\ &= [\pi_{H,1}(\Delta_{-1}^H) - \pi_{H,2}(\Delta_{-1}^H)] + O(n^{-1}), \end{aligned} \quad (\text{A.16})$$

where (A.16) follows by plugging in (A.15). Evaluating  $T_l$ ,

$$\begin{aligned} T_l(\tilde{\alpha}) &= \int_{\mathbb{R}^{J-1}} \int_{\pi^{L,2}(\Delta_{-1}^L)}^{\pi^{L,1}(\Delta_{-1}^L)} f_{\Delta_1^L | \Delta_{-1}^L}(t | \mathbf{v}) dt \cdot f_{\Delta_{-1}^L}(\mathbf{v}) d\mathbf{v} \\ &= \int_{\mathbb{R}^{J-1}} [\pi^{L,1}(\mathbf{v}_{-1}) - \pi^{L,2}(\mathbf{v}_{-1})] f_{\Delta_1^L | \Delta_{-1}^L}(\pi^{L,2}(\mathbf{v}_{-1}) | \mathbf{v}_{-1}) \cdot f_{\Delta_{-1}^L}(\mathbf{v}_{-1}) d\mathbf{v}_{-1} \end{aligned}$$

$$\begin{aligned}
& + O(n^{-1}) \\
& = \int_{\mathbb{R}^{J-1}} [\pi^{L,1}(\mathbf{v}_{-1}) - \pi^{L,2}(\mathbf{v}_{-1})] f_{\Delta_1^l | \Delta_{-1}^l}(\pi^{H,2}(\mathbf{v}_{-1}) | \mathbf{v}_{-1}) \cdot f_{\Delta_{-1}^l}(\mathbf{v}_{-1}) d\mathbf{v}_{-1} \\
& + O(n^{-1}).
\end{aligned}$$

Then by reference to (A.16), we see that

$$T_h(\tilde{\alpha}) + T_l(\tilde{\alpha}) = O(n^{-1}). \quad \square$$

### CPE from density estimation error: $C_h, C_l$

We continue to maintain the assumption that  $\mathbf{Y}$  and  $\mathbf{\Omega}$  satisfy Condition  $\star(2\log(n))$  as defined in Lemma 1.2. Notationally, for vector  $\mathbf{x}$ , let  $O(\mathbf{x})$  mean  $O(\max(\mathbf{x}))$ ; below, this is only used in cases where each element of the vector has the same bound anyway. A bound on the estimation error can be computed as

$$\begin{aligned}
\hat{\gamma}_j &= (m_1/m_j) \left[ \frac{Q(Y_j + \Omega_j^+) - Q(Y_j - \Omega_j^-)}{Q(Y_1 + \Omega_1^+) - Q(Y_1 - \Omega_1^-)} \right] && \text{(by definition/notation)} \\
&= (m_1/m_j) \left[ \frac{Q'(Y_j)\Omega_j + O(\Omega_j^2)}{Q'(Y_1)\Omega_1 + O(\Omega_1^2)} \right] && \text{(by Taylor expansion)} \\
&= \frac{Q'(Y_j)}{Q'(Y_1)} \times \left[ \frac{\Omega_j/m_j}{\Omega_1/m_1} \right] + O(\mathbf{\Omega}^2) && \text{(by algebra)} \\
&= [\gamma_j + O(\mathbf{Y} - \mathbf{u})] \times \left[ \frac{\Omega_j/m_j}{\Omega_1/m_1} \right] + O(\mathbf{m}^2/n^2) && \text{(by Condition } \star; \text{ see below)} \\
&= \gamma_j + O(\mathbf{m}^{1/2}n^{-1} \log(n)) && \text{(by Condition } \star; \text{ see below).}
\end{aligned}$$

For the statement  $\Omega_j = O(m_j^2/n^2)$ , note that the  $\Delta k_j$  (different  $j$ ) in the definition of Condition  $\star$  is  $2m_j$ , so the deviation of  $\Omega_j$  from  $2m/(n+1)$  is bounded by  $n^{-1}m_j^{1/2} \log(n)$ . Consequently,

$$\Omega_j^2 = [2m_j/(n+1)]^2 + O(n^{-2}m_j^{3/2} \log(n)) = O(m_j^2/n^2).$$

For the final line, first note that Condition  $\star(2\log(n))$  for  $\mathbf{Y}$  implies  $\mathbf{Y} - \mathbf{u} = O(n^{-1/2} \log(n))$  since the  $\Delta k_j$  grow proportionally to  $n$  (since the  $u_j$  are fixed).

Second, using Condition  $\star(2 \log(n))$  for  $\Omega$ ,

$$\begin{aligned}
\Omega_j &= \frac{2m_j}{n} + O(n^{-1}m_j^{1/2} \log(n)) && \text{(as above),} \\
\Omega_j/m_j &= (2/n) + O(n^{-1}m_j^{-1/2} \log(n)) && \text{(by algebra)} \\
&= 2n^{-1} \left[ 1 + O(m_j^{-1/2} \log(n)) \right] && \text{(since, again, } \Delta k_j \propto n), \\
\frac{\Omega_j/m_j}{\Omega_1/m_1} &= \frac{2n^{-1} \left[ 1 + O(m_j^{-1/2} \log(n)) \right]}{2n^{-1} \left[ 1 + O(m_1^{-1/2} \log(n)) \right]} \\
&= 1 + O(m^{-1/2} \log(n)) = 1 + O(n^{-1/3} \log(n)),
\end{aligned}$$

where we choose all  $m_j \asymp n^{2/3}$  as discussed in the main text. Since  $\gamma_j = O(1)$ , the largest of the remainder terms is this  $O(n^{-1/3} \log(n))$ .

By Assumption A1.1,  $Q(\cdot)$  is three times continuously differentiable in a neighborhood of each  $u_j$ , so the calibration function  $\tilde{\alpha}(\cdot)$  and the quantile selection functions  $u_j^H(\cdot)$  and  $u_j^L(\cdot)$  (as described first in equation (1.9) of §1.3) are also three times continuously differentiable. For the calibration function  $\tilde{\alpha}(\gamma)$ , a change in  $\gamma$  will affect  $\tilde{\alpha}$  by an  $O(1)$  amount; note that  $\gamma_j \in (0, \infty)$  and  $\tilde{\alpha} \in [\alpha, 1]$ . From Lemma 1.4, suppressing subscripts,  $u^h(\tilde{\alpha}) = u + n^{-1/2} \Phi^{-1}(1 - \tilde{\alpha}) \sqrt{u(1-u)} + O(n^{-1})$ , so the derivative with respect to  $\tilde{\alpha}$  is  $-n^{-1/2} \sqrt{u(1-u)} / \phi(\Phi^{-1}(1 - \tilde{\alpha})) + O(n^{-1}) = O(n^{-1/2})$ , and the order is the same for the derivative of  $u^l(\tilde{\alpha})$ . Consequently, for all  $j = 1, \dots, J$ ,

$$\begin{aligned}
\hat{u}_j - u_{j,0} &\equiv u_j^h(\tilde{\alpha}(\hat{\gamma})) - u_j^h(\tilde{\alpha}(\gamma)) = (\hat{\gamma} - \gamma)' \times u_j^{h'}(\tilde{\alpha}(\gamma)) \times \tilde{\alpha}'(\gamma) \quad (\text{A.17}) \\
&= O(\mathbf{m}^{-1/2} \log(n)) O(n^{-1/2}) O(1) = O(n^{-5/6} \log(n)),
\end{aligned}$$

$$\begin{aligned}
Q(\hat{u}_j) - Q(u_{j,0}) &= Q'(u_{j,0})(\hat{u}_j - u_{j,0}) + (1/2)Q''(\tilde{u})(\hat{u}_j - u_{j,0})^2 \quad (\text{A.18}) \\
&= O(n^{-1/2} \mathbf{m}^{-1/2} \log(n)) = O(n^{-5/6} \log(n)).
\end{aligned}$$

The term of ultimate interest can be decomposed into

$$\begin{aligned}
C_h &= E_{\hat{\gamma}} \{ P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n} \boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\hat{\mathbf{u}})] \mid \hat{\gamma}) \\
&\quad - P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n} \boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \} \\
&= E_{\hat{\gamma}} \{ P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n} \boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\hat{\mathbf{u}})] \mid \hat{\gamma}) \\
&\quad - P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n} \boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \}
\end{aligned}$$

$$\begin{aligned}
& + E_{\hat{\gamma}}\{P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \\
& \quad - P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma})\} \\
& \equiv C_h^1 + C_h^2.
\end{aligned}$$

We additionally bound the first term by the expansion

$$\begin{aligned}
|C_h^1| \leq & \left| \overbrace{E_{\hat{\gamma}}\{f_{\mathbb{W}_{\hat{\mathbf{u}}|\hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \times \sqrt{n}\boldsymbol{\psi}'[Q(\hat{\mathbf{u}}) - Q(\mathbf{u}_0)]\}}^{C_h^{1,a}} \right| \\
& + \left| \overbrace{E_{\hat{\gamma}}\{(1/2)f'_{\mathbb{W}_{\hat{\mathbf{u}}|\hat{\gamma}}(\tilde{w} \mid \hat{\gamma})(\sqrt{n}\boldsymbol{\psi}'[Q(\hat{\mathbf{u}}) - Q(\mathbf{u}_0)])^2\}}^{C_h^{1,b}} \right|.
\end{aligned}$$

To bound  $C_h^{1,a}$ , consider

$$\begin{aligned}
E(AB) & = \text{Cov}(A, B) + E(A)E(B), \\
\text{Cov}(A, B) & = \text{Corr}(A, B)\sqrt{\text{Var}(A)\text{Var}(B)} \leq \sqrt{\text{Var}(A)\text{Var}(B)},
\end{aligned}$$

with  $A = f_{\mathbb{W}_{\hat{\mathbf{u}}|\hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma})}$  and  $B = \sqrt{n}\boldsymbol{\psi}'[Q(\hat{\mathbf{u}}) - Q(\mathbf{u}_0)]$ . From (A.17) and (A.18),

$$E(B) = \text{Bias}(\hat{\gamma})O(n^{-1/2})O(n^{1/2}) = O(\mathbf{m}^2/n^2) = O(n^{-2/3}),$$

and similarly

$$\text{Var}(B) = n\text{Var}(\hat{\gamma})O(n^{-1}) = O(\mathbf{m}^{-1}) = O(n^{-2/3}),$$

using (2.5) and (2.6) of Bloch and Gastwirth (1968) for the bias and variance of  $\hat{\gamma}$ . From Lemma A.1(iii),  $E(A) = O(1)$  and  $\text{Var}(A) = O(n^{-2/3})$ .

For  $C_h^{1,b}$ , since  $B^2 > 0$  for any  $B$ , we can use  $E(AB^2) \leq E(\sup |A|B^2) \leq \sup |A|E(B^2)$ :

$$\begin{aligned}
|C_h^{1,b}| \leq & (1/2) \underbrace{\sup_{\star(2\log(n))} |f'_{\mathbb{W}_{\hat{\mathbf{u}}|\hat{\gamma}}(\tilde{w} \mid \hat{\gamma})|}_{= O(1) \text{ by Lemma A.1(iii)}} \times \underbrace{E\{(\sqrt{n}\boldsymbol{\psi}'[Q(\hat{\mathbf{u}}) - Q(\mathbf{u}_0)])^2\}}_{= O(\text{Var}(\hat{\gamma})) = O(\mathbf{m}^{-1})} = O(n^{-2/3}),
\end{aligned}$$

where the equivalence of the order of the second term to that of  $\text{Var}(\hat{\gamma})$  comes again from differentiability arguments and equations (A.17) and (A.18).

Finally, for  $C_h^2$ ,

$$\begin{aligned}
|C_h^2| &= \left| E_{\hat{\gamma}} \left\{ P(\mathbb{W}_{\hat{\mathbf{u}}} > \sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \right. \right. \\
&\quad \left. \left. - P(\mathbb{W}_{\mathbf{u}_0} > \sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \right\} \right| \\
&= \left| E_{\hat{\gamma}} \left\{ F_{\mathbb{W}_{\mathbf{u}_0} \mid \hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) - F_{\mathbb{W}_{\hat{\mathbf{u}}} \mid \hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma}) \right\} \right| \\
&\leq \left| E_{\hat{\gamma}} \left\{ (\mathbf{u}_0 - \hat{\mathbf{u}})' \frac{\partial F_{\mathbb{W}_{\mathbf{t}} \mid \hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma})}{\partial \mathbf{t}} \Big|_{\mathbf{t}=\mathbf{u}_0} \right\} \right| \\
&\quad + \left| E_{\hat{\gamma}} \left\{ (\mathbf{u}_0 - \hat{\mathbf{u}})' \frac{\partial^2 F_{\mathbb{W}_{\mathbf{t}} \mid \hat{\gamma}}(\sqrt{n}\boldsymbol{\psi}'[Q(\mathbf{u}) - Q(\mathbf{u}_0)] \mid \hat{\gamma})}{\partial \mathbf{t} \partial \mathbf{t}'} \Big|_{\mathbf{t}=\hat{\mathbf{u}}} (\mathbf{u}_0 - \hat{\mathbf{u}}) \right\} \right| \\
&\equiv |C_h^{2,a}| + |C_h^{2,b}|.
\end{aligned}$$

For  $C_h^{2,a}$ , again consider the structure

$$E(AB) = \text{Cov}(A, B) + E(A)E(B),$$

$$\text{Cov}(A, B) = \text{Corr}(A, B) \sqrt{\text{Var}(A)\text{Var}(B)} \leq \sqrt{\text{Var}(A)\text{Var}(B)},$$

this time with  $A = \mathbf{u}_0 - \hat{\mathbf{u}}$  and  $B$  the first derivative of  $F_{\mathbb{W}_{\mathbf{t}} \mid \hat{\gamma}}$  as appears in the expansion above. From (A.17),

$$E(A) = \text{Bias}(\hat{\gamma})O(n^{-1/2}) = O(n^{-1/2}\mathbf{m}^2/n^2) = O(n^{-7/6}),$$

$$\text{Var}(A) = \text{Var}(\hat{\gamma})O(n^{-1}) = O(n^{-1}\mathbf{m}^{-1}) = O(n^{-5/3}),$$

where the bias and variance of  $\hat{\gamma}$  are from (2.5) and (2.6) of Bloch and Gastwirth (1968). From Lemma A.1(iii),  $E(B) = O(n^{1/2})$  and  $\text{Var}(B) = O(n^{1/3})$ .

For  $C_h^{2,b}$ , as with  $C_h^{1,b}$ , we can use  $E(BA^2) \leq E(\sup |B|A^2) \leq \sup |B|E(A^2)$ , where still  $A = \mathbf{u}_0 - \hat{\mathbf{u}}$  and now  $B$  is the second derivative of  $F_{\mathbb{W}_{\mathbf{t}} \mid \hat{\gamma}}$ . From (A.17),  $E(A^2) = \text{Var}(\hat{\gamma})O(n^{-1}) = O(n^{-1}\mathbf{m}^{-1}) = O(n^{-5/3})$ . Lemma A.1(iii) states that  $B = O(n)$ .

The foregoing arguments and rates are all the same for  $C_l$ .

## Negative definiteness of $\underline{H}_{\mathbf{x}}$

*Proof.* As given in (A.2),

$$h_j(\mathbf{x}) = - \left( \frac{k_j - k_{j-1} - 1}{(y_j - y_{j-1})^2} + \frac{k_{j+1} - k_j - 1}{(x_{j+1} - x_j)^2} \right) \equiv -(a_j + a_{j+1}),$$

$$h_{j,j}(\mathbf{x}) = \frac{k_{j+1} - k_j - 1}{(x_{j+1} - x_j)^2} \equiv a_{j+1},$$

and all other cross-partials are zero. The full Hessian  $\underline{H}_{\mathbf{x}}$  is then

$$\begin{pmatrix} (-a_1 - a_2) & a_2 & 0 & 0 & \cdots & \cdots & 0 \\ a_2 & (-a_2 - a_3) & a_3 & 0 & \cdots & \cdots & 0 \\ 0 & a_3 & (-a_3 - a_4) & a_4 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & a_n \\ 0 & \cdots & \cdots & \cdots & 0 & a_n & (-a_n - a_{n+1}) \end{pmatrix},$$

where, regardless of the value of  $\mathbf{x}$ , all of the  $a_j$  are positive. By definition, this is negative definite if and only if all quadratic forms are negative, i.e. if  $\mathbf{c}'\underline{H}_{\mathbf{x}}\mathbf{c} < 0$  for any vector  $\mathbf{c} = (c_1, \dots, c_n) \neq \mathbf{0}$ . Using the symmetry and values above, this will be equal to

$$\begin{aligned} & \sum_{i=1}^n c_i^2(-a_i - a_{i+1}) + 2 \sum_{i=2}^n c_i c_{i-1} a_i \\ &= -a_1 c_1^2 - a_{n+1} c_n^2 - \sum_{i=2}^n a_i c_i^2 - \sum_{i=2}^n a_i c_{i-1}^2 + 2 \sum_{i=2}^n c_i c_{i-1} a_i \\ &= -a_1 c_1^2 - a_{n+1} c_n^2 - \sum_{i=2}^n a_i (c_i^2 + c_{i-1}^2 - 2c_i c_{i-1}) \\ &= -a_1 c_1^2 - a_{n+1} c_n^2 - \sum_{i=2}^n a_i (c_i - c_{i-1})^2 < 0. \end{aligned}$$

The last expression is strictly negative since all  $a_i > 0$  and all the squares are nonnegative,  $c_1^2 \geq 0$ ,  $c_n^2 \geq 0$ ,  $(c_i - c_{i-1})^2 \geq 0$ , with at least one of them strictly positive since there is at least one nonzero  $c_i$ .  $\square$

# Appendix B

## Technical appendix to Chapter 2

### B.1 Lemma 2.2 proof

The case  $d = 1$  is covered here. Generalizing to  $d > 1$  should yield the same order and type of terms in the bias, just with vectors for first derivatives and matrices for second derivatives. The  $h^d$  (instead of  $h$ ) from the change of variables does not determine the bias (see below).

#### Case $b = 2$

By definition,  $Q_{Y|C_h}(p)$  satisfies

$$p = \int_{C_h} \left\{ \int_{-\infty}^{Q_{Y|C_h}(p)} f_{Y|X}(y; x) dy \right\} f_{X|C_h}(x) dx, \quad (\text{B.1})$$

where  $f_{Y|X}(y; x) \equiv f_{Y,X}(y, x)/f_X(x)$  is the conditional PDF of  $Y$  given  $X$  evaluated at  $Y = y$  and  $X = x$ . Similarly,  $f_{X|C_h}(x) = f_X(x)/P(X \in C_h) = O(h^{-1})$  is the conditional PDF of  $X$  within  $C_h$ . Also by definition,  $Q_{Y|X}(p; x)$  satisfies

$$p = \int_{-\infty}^{Q_{Y|X}(p; x)} f_{Y|X}(y; x) dy$$

for all  $x$ .



Decomposing the  $\{\cdot\}$  term, (B.1) becomes

$$\begin{aligned} p &= \int_{C_h} \left\{ \int_{-\infty}^{Q_{Y|X}(p;x)} f_{Y|X}(y;x) dy + \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x) dy \right\} f_{X|C_h}(x) dx \\ &= p + \int_{C_h} \left\{ \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x) dy \right\} f_{X|C_h}(x) dx, \end{aligned}$$

implying

$$0 = \int_{C_h} \left\{ \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x) dy \right\} f_{X|C_h}(x) dx.$$

With change of variables  $x = wh$ , and converting  $f_{X|C_h}(\cdot)$  to  $f_X(\cdot)$ ,

$$0 = \frac{h}{P(X \in C_h)} \int_{-1}^1 \left\{ \int_{Q_{Y|X}(p;wh)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;wh) dy \right\} f_X(wh) dw.$$

With enough smoothness, a second-order Taylor expansion around  $w = 0$  can be taken of the integrand. An explicit expression for the bias comes out of the zeroth-order term. The first-order terms will zero out, leaving the second-order terms of order  $h^2$  to additionally determine the bias. If there is enough smoothness to have the third-order terms (order  $h^3$ ), they will also zero out, reducing the remainder to  $o(h^3)$ . In the following calculations, we use the derivative rule

$$\frac{\partial}{\partial x} \int_{a(x)}^b f(y,x) dy = -f(a(x),x)a'(x) + \int_{a(x)}^b \frac{\partial}{\partial x} f(y,x) dy.$$

Schematically, we are taking a second-order Taylor expansion of

$$\left\{ \int_{a(x)}^b f(y;x) dy \right\} g(x)$$

around  $x = 0$ , which would be

$$\begin{aligned} & \int_{a(0)}^b f(y;0) dy g(0) \\ & + x \left\{ \int_{a(0)}^b f(y;0) dy g'(0) + \left[ \int_{a(0)}^b \frac{\partial}{\partial x} f(y;x) \Big|_{x=0} dy - f(a(0);0)a'(0) \right] g(0) \right\} \\ & + (x^2/2) \left\{ \int_{a(0)}^b f(y;0) dy g''(0) + \left[ \int_{a(0)}^b \frac{\partial}{\partial x} f(y;x) \Big|_{x=0} dy - f(a(0);0)a'(0) \right] g'(0) \right\} \end{aligned}$$

$$\begin{aligned}
& + \left[ \int_{a(0)}^b \frac{\partial}{\partial x} f(y; x) \Big|_{x=0} dy - f(a(0), 0) a'(0) \right] g'(0) \\
& + \left[ \int_{a(0)}^b \frac{\partial^2}{\partial x^2} f(y; x) \Big|_{x=0} dy - \frac{\partial}{\partial x} f(a(0); x) \Big|_{x=0} a'(0) \right. \\
& \quad - f(a(0); 0) a''(0) \\
& \quad \left. - \left( \frac{\partial}{\partial y} f(y; 0) \Big|_{y=a(0)} a'(0) + \frac{\partial}{\partial x} f(a(0); x) \Big|_{x=0} \right) a'(0) \right] g(0) \Big\} \\
& + o(x^2).
\end{aligned}$$

For compactness, we define

$$\begin{aligned}
Q_{Y|X}^{(0,1)}(p; 0) &\equiv \frac{\partial}{\partial x} Q_{Y|X}(p; x) \Big|_{x=0}, & Q_{Y|X}^{(0,2)}(p; 0) &\equiv \frac{\partial^2}{\partial x^2} Q_{Y|X}(p; x) \Big|_{x=0}, \\
\xi_p &\equiv Q_{Y|X}(p; 0), \\
f_{Y|X}^{(0,1)}(y; 0) &\equiv \frac{\partial}{\partial x} f_{Y|X}(y; x) \Big|_{x=0}, & f_{Y|X}^{(1,0)}(\xi_p; 0) &\equiv \frac{\partial}{\partial y} f_{Y|X}(y; 0) \Big|_{y=\xi_p}.
\end{aligned}$$

The Taylor expansion of

$$\left\{ \int_{Q_{Y|X}(p; wh)}^{Q_{Y|C_h}(p)} f_{Y|X}(y; wh) dy \right\} f_X(wh)$$

around  $w = 0$  yields

$$\begin{aligned}
& \left\{ \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y; 0) dy \right\} f_X(0) \\
& + wh \left\{ \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y; 0) dy f_X'(0) + \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}^{(0,1)}(y; 0) dy f_X(0) \right. \\
& \quad \left. - f_{Y|X}(\xi_p; 0) Q_{Y|X}^{(0,1)}(p; 0) f_X(0) \right\} \\
& + \frac{(wh)^2}{2} \left\{ \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y; 0) dy f_X''(0) \right. \\
& \quad + 2 \left[ \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}^{(0,1)}(y; 0) dy - f_{Y|X}(\xi_p; 0) Q_{Y|X}^{(0,1)}(p; 0) \right] f_X'(0) \\
& \quad \left. + \left[ \int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}^{(0,2)}(y; 0) dy - f_{Y|X}(\xi_p; 0) Q_{Y|X}^{(0,2)}(p; 0) \right] f_X(0) \right\}
\end{aligned}$$

$$\begin{aligned}
& - 2f_{Y|X}^{(0,1)}(\xi_p; 0)Q_{Y|X}^{(0,1)}(p; 0)f_X(0) - f_{Y|X}^{(1,0)}(\xi_p; 0)\left[Q_{Y|X}^{(0,1)}(p; 0)\right]^2 f_X(0) \Big\} \\
& + o(h^2) \\
& = A + hwB - (1/2)h^2w^2C + o(h^2),
\end{aligned}$$

with  $A$ ,  $B$ , and  $C$  implicitly defined.

To extract the bias from  $A$ , we also need to expand

$$\begin{aligned}
\int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y; 0)dy &= \int_{\xi_p}^{Q_{Y|C_h}(p)} \left[ f_{Y|X}(\xi_p; 0) + f_{Y|X}^{(1,0)}(\tilde{y}; 0)(y - \xi_p) \right] dy \\
&= f_{Y|X}(\xi_p; 0)[Q_{Y|C_h}(p) - \xi_p] + O\left([Q_{Y|C_h}(p) - \xi_p]^2\right),
\end{aligned}$$

where  $\tilde{y}$  is determined by the mean value theorem and so located between the lower and upper limits of integration. Anticipating that the bias is  $[Q_{Y|C_h}(p) - \xi_p] = O(h^2)$ , the term  $A$  is then

$$A = f_{Y|X}(\xi_p; 0)[Q_{Y|C_h}(p) - \xi_p]f_X(0) + O(h^4).$$

Since there is no  $w$  in  $A$ ,  $\int_{-1}^1 Adw = 2A$ .

The  $B$  term zeroes out since the only  $w$  in it is the  $w$  in  $hwB$ . The integral over  $[-1, 1]$  is thus  $\int_{-1}^1(hwB)dw = hB \int_{-1}^1 wdw = (hB)(0) = 0$ .

For the  $C$  term, the only  $w$  is the  $w^2$  in  $(1/2)h^2w^2C$ , so

$$\int_{-1}^1 (1/2)h^2w^2Cdw = (1/2)h^2C \int_{-1}^1 w^2dw = (1/2)h^2C(2/3) = h^2C/3.$$

Anticipating that the bias is  $O(h^2)$ , and assuming  $f_{Y|X}^{(0,2)}(y; 0)$  is bounded in a neighborhood of  $y = \xi_p$ , the three definite integrals in  $C$  are  $O(h^2)$ . Thus  $C$  simplifies to

$$\begin{aligned}
C &= 2f_{Y|X}(\xi_p; 0)Q_{Y|X}^{(0,1)}(p; 0)f'_X(0) + f_{Y|X}(\xi_p; 0)Q_{Y|X}^{(0,2)}(p; 0)f_X(0) \\
&+ 2f_{Y|X}^{(0,1)}(\xi_p; 0)Q_{Y|X}^{(0,1)}(p; 0)f_X(0) + f_{Y|X}^{(1,0)}(\xi_p; 0)\left[Q_{Y|X}^{(0,1)}(p; 0)\right]^2 f_X(0) \\
&+ O(h^2).
\end{aligned}$$

Combining the results for  $A$ ,  $B$ , and  $C$ ,

$$2f_{Y|X}(\xi_p; 0)[Q_{Y|C_h}(p) - \xi_p]f_X(0) + O(h^4) = h^2C/3 + O(h^4) + o(h^2),$$

so the bias is

$$\begin{aligned}
& Q_{Y|C_h}(p) - \xi_p \\
&= h^2 \frac{C}{6f_{Y|X}(\xi_p; 0)f_X(0)} + o(h^2) \\
&= \frac{h^2}{6} \left\{ 2Q_{Y|X}^{(0,1)}(p; 0)f'_X(0)/f_X(0) + Q_{Y|X}^{(0,2)}(p; 0) + 2f_{Y|X}^{(0,1)}(\xi_p; 0)Q_{Y|X}^{(0,1)}(p; 0)/f_{Y|X}(\xi_p; 0) \right. \\
&\quad \left. + f_{Y|X}^{(1,0)}(\xi_p; 0) \left[ Q_{Y|X}^{(0,1)}(p; 0) \right]^2 / f_{Y|X}(\xi_p; 0) \right\} + o(h^2).
\end{aligned}$$

As noted, the  $o(h^2)$  sharpens to  $o(h^3)$  if there is enough smoothness to get the  $h^3$  term in the expansion. Similar to the  $B$  term above, this term will zero out since  $w$  enters only as  $w^3$ , which integrates to zero over  $w \in [-1, 1]$ .

The smoothness of  $f_{Y|X}(\cdot; \cdot)$  does not appear in the main text assumptions because it is a consequence of  $s_Q$ . From (B.2), it is clear that the existence of  $Q_{Y|X}^{(0,2)}(p; x)$  implies the existence of  $f_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x)$ .

This bias expression is equivalent to the bias in Bhattacharya and Gangopadhyay (1990, Thm. K1), just in different notation. By definition, for all  $x$ ,

$$F_{Y|X}(Q_{Y|X}(p; x); x) = p.$$

Differentiating once with respect to  $x$  yields

$$\begin{aligned}
0 &= Q_{Y|X}^{(0,1)}(p; x)f_{Y|X}(Q_{Y|X}(p; x); x) + F_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x), \\
Q_{Y|X}^{(0,1)}(p; x) &= -\frac{F_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x)}{f_{Y|X}(Q_{Y|X}(p; x); x)}.
\end{aligned}$$

Differentiating again with respect to  $x$  gives

$$\begin{aligned}
0 &= Q_{Y|X}^{(0,2)}(p; x)f_{Y|X}(Q_{Y|X}(p; x); x) + F_{Y|X}^{(0,2)}(Q_{Y|X}(p; x); x) \\
&\quad + Q_{Y|X}^{(0,1)}(p; x)f_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x) \\
&\quad + \left[ Q_{Y|X}^{(0,1)}(p; x) \right]^2 f_{Y|X}^{(1,0)}(Q_{Y|X}(p; x); x) \\
&\quad + f_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x)Q_{Y|X}^{(0,1)}(p; x), \\
Q_{Y|X}^{(0,2)}(p; x) &= -\frac{1}{f_{Y|X}(Q_{Y|X}(p; x); x)} \\
&\quad \times \left\{ F_{Y|X}^{(0,2)}(Q_{Y|X}(p; x); x) + 2Q_{Y|X}^{(0,1)}(p; x)f_{Y|X}^{(0,1)}(Q_{Y|X}(p; x); x) \right\} \quad (\text{B.2})
\end{aligned}$$

$$+ \left[ Q_{Y|X}^{(0,1)}(p; x) \right]^2 f_{Y|X}^{(1,0)}(Q_{Y|X}(p; x); x) \Big\}.$$

Plugging these substitutions into the original bias expression gives

$$\begin{aligned} \text{Bias} &= \frac{h^2}{6} \left\{ -\frac{2F_{Y|X}^{(0,1)}(\xi_p; 0)f'_X(0)}{f_X(0)f_{Y|X}(\xi_p; 0)} \right. \\ &\quad - \frac{1}{f_{Y|X}(\xi_p; 0)} \left\{ F_{Y|X}^{(2,0)}(\xi_p; 0) - 2F_{Y|X}^{(0,1)}(\xi_p; 0)f_{Y|X}^{(0,1)}(\xi_p; 0)/f_{Y|X}(\xi_p; 0) \right. \\ &\quad \left. \left. + \left[ F_{Y|X}^{(0,1)}(\xi_p; 0) \right]^2 f_{Y|X}^{(1,0)}(\xi_p; 0)/f_{Y|X}(\xi_p; 0) \right\} \right. \\ &\quad \left. - \frac{2f_{Y|X}^{(0,1)}(\xi_p; 0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{[f_{Y|X}(\xi_p; 0)]^2} + \frac{f_{Y|X}^{(1,0)}(\xi_p; 0)\left[ F_{Y|X}^{(0,1)}(\xi_p; 0) \right]^2}{[f_{Y|X}(\xi_p; 0)]^3} \right\} \\ &= \frac{h^2}{6[f_{Y|X}(\xi_p; 0)]^3} \\ &\quad \times \left\{ -2F_{Y|X}^{(0,1)}(\xi_p; 0)[f_{Y|X}(\xi_p; 0)]^2 f'_X(0)/f_X(0) - F_{Y|X}^{(0,2)}(\xi_p; 0)[f_{Y|X}(\xi_p; 0)]^2 \right. \\ &\quad + 2F_{Y|X}^{(0,1)}(\xi_p; 0)f_{Y|X}^{(0,1)}(\xi_p; 0)f_{Y|X}(\xi_p; 0) - [F_{Y|X}^{(0,1)}(\xi_p; 0)]^2 f_{Y|X}^{(1,0)}(\xi_p; 0) \\ &\quad \left. - 2F_{Y|X}^{(0,1)}(\xi_p; 0)f_{Y|X}^{(0,1)}(\xi_p; 0)f_{Y|X}(\xi_p; 0) + [F_{Y|X}^{(0,1)}(\xi_p; 0)]^2 f_{Y|X}^{(1,0)}(\xi_p; 0) \right\} \\ &= -h^2 \frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)}. \end{aligned}$$

This is equivalent to the bias in Bhattacharya and Gangopadhyay (1990, Thm. K1) since their bandwidth is  $h/2$ , so  $(h/2)^2/6 = h^2/24$ , and with  $x_0 = 0$  and

$$g(\xi) \equiv f_{Y|X}(\xi_p; 0), \quad G_x(\xi | x_0) \equiv F_{Y|X}^{(0,1)}(\xi_p; 0), \quad G_{xx}(\xi | x_0) \equiv F_{Y|X}^{(0,2)}(\xi_p; 0).$$

### Case $b < 2$

In the foregoing calculations, we (implicitly) assumed  $k_X \geq 1$ ,  $k_Q \geq 2$ ,  $s_Y \geq 1$ , and  $s_{YX} \geq 2$ . (While the brief appearance of  $f''_X(0)$  means we technically assumed  $k_X \geq 2$ , this can easily be weakened to  $k_X \geq 1$  since it appears in a smaller-order term, and similarly for  $f_{Y|X}^{(0,2)}(y; 0)$ .)

By examining the original Taylor expansion, we can see how the order of the bias will diminish as we relax these smoothness assumptions. We consider the

above expansion  $A = h^2 w^2 C/2 - hwB + o(h^2)$ .

The  $A$  term is the bias plus a remainder depending on  $s_Y$ . Even if  $s_Y \leq 1$ , the remainder is

$$\begin{aligned} & \int_{\xi_p}^{Q_{Y|X}(p)} [f_{Y|X}(y; 0) - f_{Y|X}(\xi_p; 0)](y - \xi_p) dy \\ & = O(\text{Bias}^{2+s_Y}), \end{aligned}$$

which is always smaller than  $O(\text{Bias})$  if the bias goes to zero asymptotically. Thus,  $s_Y$  here does not have an effect on the order of the bias, and  $Ah$  is the bias times  $O(1)$  terms not dependent on smoothness.

The  $C$  term includes many derivatives, but some of the terms are already smaller-order. Assumption A2.6 already requires  $f_{Y|X}^{(0,2)}(Q_{Y|X}(p; 0))$  to be continuous in a neighborhood of  $p$ , so those terms may be ignored. Specifically,  $f_X''(0)$  only appears in a term of order  $(h^2 \text{Bias})$ , which is always smaller than the bias, so  $s_X$  has no binding effect here.

The three key terms in  $C$  involve

$$Q_{Y|X}^{(0,1)}(p; 0) f_X'(0), \quad Q_{Y|X}^{(0,2)}(p; 0), \quad \text{and} \quad f_{Y|X}^{(1,0)}(\xi_p; 0) \left[ Q_{Y|X}^{(0,1)}(p; 0) \right]^2.$$

If  $s_X < 1$ , the first term will be replaced by a larger-order term; if  $s_Q < 2$ , the same will happen for the second term; and if  $s_Y < 1$ , same for the third term, but we already need  $s_Y > 2$  to apply the Chapter 1 results. Essentially, with less smoothness, we are forced to replace (for example)  $g(x) = g(0) + g'(0)x + (1/2)g''(\tilde{x})x^2$  with  $g(x) = g(0) + g'(\tilde{x})x$ , where  $\tilde{x}$  is determined by the mean value theorem. Before,  $g''(\tilde{x}) = g''(0) + [g''(\tilde{x}) - g''(0)]$ , and the absolute value of the  $[\cdot]$  term is bounded as  $|\tilde{x}|^\gamma$  by Hölder exponent  $\gamma$ . After relaxing the smoothness assumptions some, we get a similar expression but with  $[g'(\tilde{x}) - g'(0)]$ : the number of derivatives  $k$  is smaller by one, and there is some new  $\gamma$ . With less smoothness, the  $B$  terms do not integrate to zero since there are  $\tilde{w}$  floating around, instead of just  $\text{Const} \times \int_{-1}^1 w dw = 0$ . The result is that the bias is of order  $h^b$  for  $b = \min\{s_Q, 1 + s_X, 1 + s_Y\}$  for  $s_Q \in [1, 2)$  and  $s_X, s_Y \in (0, 1)$ . As we saw originally, the biggest  $b$  can be is two, and relaxing  $s_Q$  further continues to decrease the order

in the same pattern; so in all, the bias is

$$Q_{Y|C_h}(p) - \xi_p = O(h^b), \quad b = \min\{2, s_Q, 1 + s_X, 1 + s_Y\}.$$

## B.2 Plug-in bandwidth calculations

The plug-in bandwidth is for  $d = 1$  and  $b = 2$ , in which case the bias is

$$\begin{aligned} Q_{Y|C_h}(p) - \xi_p &= \frac{h^2}{6} \left\{ \left[ 2Q_{Y|X}^{(0,1)}(p; 0) f'_X(0) / f_X(0) + Q_{Y|X}^{(0,2)}(p; 0) \right] \right. \\ &\quad + 2f_{Y|X}^{(0,1)}(\xi_p; 0) Q_{Y|X}^{(0,1)}(p; 0) / f_{Y|X}(\xi_p; 0) \\ &\quad \left. + f_{Y|X}^{(1,0)}(\xi_p; 0) \left[ Q_{Y|X}^{(0,1)}(p; 0) \right]^2 / f_{Y|X}(\xi_p; 0) \right\} + o(h^2) \\ &= -h^2 \frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0) F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0) f_{Y|X}(\xi_p; 0)} + o(h^2). \end{aligned}$$

To avoid iteration (and nicely cancel some constants), we plug in  $\epsilon_h = \epsilon_\ell = 0.2$  as a rule of thumb. The maximum bandwidth would be obtained using  $\epsilon = 0.5$ , which would give extremely similar bandwidths since, for example,  $[(0.2)(0.8)]^{1/6} = 0.74$  while  $[(0.5)(0.5)]^{1/6} = 0.79$  in the two-sided median case. The stabilizing effect and computational gains of using a fixed  $\epsilon$  outweigh the small benefit of iteration. We also consider a Gaussian plug-in assumption for  $f_{Y|X}(\xi_p; 0)$  and  $f_{Y|X}^{(1,0)}(\xi_p; 0)$ , using an estimated variance of  $Y$ .

First,  $\text{CPE}_{\text{Bias}}$  depends on  $f_{\hat{Q}_u}(Q_{Y|C_h}(p))$ . From earlier,  $F_{Y|C_h}(\hat{Q}_u)$  has a beta distribution. Specifically, for  $u \in (0, 1)$ ,

$$F_{Y|C_h}(\hat{Q}_u) \sim \beta[(N_n + 1)u, (N_n + 1)(1 - u)],$$

writing  $f_\beta(\cdot)$  for the corresponding beta distribution's PDF and  $F_\beta(\cdot)$  for the CDF. For the lower one-sided CI, upper endpoint quantile  $u = u_h$  is chosen by the Hutson (1999) method such that  $F_\beta(p) = \alpha$ . By applying the chain rule and the fact that

$$\begin{aligned}
F_{Y|C_h}(Q_{Y|C_h}(p)) &= p, \\
f_{\hat{Q}_u}(x) &= \frac{\partial}{\partial x} F_{\hat{Q}_u}(x) = \frac{\partial}{\partial x} F_\beta(F_{Y|C_h}(x)) = f_\beta(F_{Y|C_h}(x)) f_{Y|C_h}(x), \\
f_{\hat{Q}_u}(Q_{Y|C_h}(p)) &= f_\beta(F_{Y|C_h}(Q_{Y|C_h}(p))) f_{Y|C_h}(Q_{Y|C_h}(p)) \\
&= f_\beta(p; u) f_{Y|C_h}(F_{Y|C_h}^{-1}(p)), \\
f_\beta(p; u) &\equiv \frac{\Gamma(N_n + 1)}{\Gamma((N_n + 1)u) \Gamma((N_n + 1)(1 - u))} p^{(N_n + 1)u - 1} (1 - p)^{(N_n + 1)(1 - u) - 1}.
\end{aligned} \tag{B.3}$$

To avoid recursive dependence on  $h$ , we can approximate  $f_{Y|C_h}(Q_{Y|C_h}(p)) \doteq f_{Y|X}(\xi_p; 0)$  up to smaller-order terms.

We can also approximate, if  $f_X''(\cdot)$  is uniformly bounded in a neighborhood of zero,

$$\begin{aligned}
P_C &= \int_{C_h} f_X(x) dx = \int_{C_h} [f_X(0) + x f_X'(0) + (1/2)x^2 f_X''(\tilde{x})] dx = 2h f_X(0) + O(h^3), \\
N_n &\doteq nP_C = 2nh f_X(0) + O(h^3).
\end{aligned}$$

A tighter relationship between  $N_n$  and  $nP_C$  would improve the theoretical accuracy in the case of the median, when the bandwidth can be chosen to zero out the dominant terms of the overall CPE. Using the actual  $N_n$  and solving for  $h$  iteratively may achieve the same accuracy increase, though at the expense of computation time, and ideally with careful consideration of the randomness of  $N_n$  (and its dependence on  $h$ ) that differs from the setup in Chapter 1.

We know that  $\text{CPE}_{\text{GK}} > 0$  (over-coverage), and we can estimate the sign of  $\text{CPE}_{\text{Bias}}$ . If they are opposite, the optimal bandwidth causes them to cancel out; if they are the same sign, the optimal bandwidth minimizes their sum. The only difference is an extra coefficient of  $[2d/(2b + d)]^{1/(b+3d/2)} = [2d/(d + 4)]^{2/(4+3d)}$  from the first-order condition in the latter case, where the initial exponents of  $h$  come down when taking a derivative.

## Plug-in bandwidth: one-sided

For one-sided inference when  $\text{CPE}_{\text{GK}}$  and  $\text{CPE}_{\text{Bias}}$  are of opposite sign, with  $\epsilon = \epsilon_h$  or  $\epsilon = \epsilon_\ell$ , the optimal  $h$  equates (up to smaller-order terms)

$$N_n^{-1} z_{1-\alpha} \frac{\epsilon(1 - \epsilon)}{p(1 - p)} \phi(z_{1-\alpha})$$



$$= h^2 \frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)} f_\beta(p; u) f_{Y|X}(\xi_p; 0).$$

After plugging in estimates of the unknown objects,  $N_n \doteq 2nhf_X(0)$ , and value for  $u$  (e.g., based on a pilot bandwidth), the above equation could be solved numerically for  $h$  with any standard statistical software.

Alternatively, we could approximate the beta PDF with a normal PDF (Chapter 1). Writing  $\beta$  for a random variable with the distribution  $\beta[(N_n + 1)u, (N_n + 1)(1 - u)]$  from above,

$$\begin{aligned} \sqrt{N_n}(\beta - u)/\sqrt{u(1 - u)} &\xrightarrow{d} N(0, 1), \quad \text{so (informally)} \beta \overset{a}{\sim} N(u, u(1 - u)/N_n), \\ f_\beta(p; u) &= \frac{\sqrt{N_n}}{\sqrt{u(1 - u)}} \left[ \phi\left(\frac{p - u}{\sqrt{u(1 - u)/N_n}}\right) + O(N_n^{-1/2}) \right], \end{aligned}$$

where  $\phi(\cdot)$  is the standard normal PDF. Using additional results from Chapter 1 that

$$\begin{aligned} u_h &= p + z_{1-\alpha} \sqrt{u_h(1 - u_h)/N_n} + O(N_n^{-1}) \text{ and} \\ u_\ell &= p - z_{1-\alpha} \sqrt{u_\ell(1 - u_\ell)/N_n} + O(N_n^{-1}), \text{ then for } u = u_h \text{ or } u = u_\ell, \\ f_\beta(p; u) &= N_n^{1/2} [u(1 - u)]^{-1/2} \left[ \phi\left(\frac{\pm z_{1-\alpha} \sqrt{u(1 - u)/N_n}}{\sqrt{u(1 - u)/N_n}}\right) + O(N_n^{-1/2}) \right] \\ &\doteq N_n^{1/2} [u(1 - u)]^{-1/2} \phi(z_{1-\alpha}) \end{aligned}$$

since  $\phi(z_{1-\alpha}) = \phi(-z_{1-\alpha})$ .

Plugging in  $N_n \doteq 2nhf_X(0)$  and  $u = p + O(N_n^{-1/2})$ , the optimal  $h$  is now an explicit function of known values and objects that can be estimated directly from the data. Denoting  $\hat{h}_{++}$  as the plug-in bandwidth when both  $\text{CPE}_{\text{GK}} > 0$  and  $\text{CPE}_{\text{Bias}} > 0$  (both over-coverage), and  $\hat{h}_{+-}$  when instead  $\text{CPE}_{\text{Bias}} < 0$ , we first solve for  $\hat{h}_{+-}$ . Up to smaller-order terms,

$$\begin{aligned} &[2nhf_X(0)]^{-1} z_{1-\alpha} \frac{\epsilon(1 - \epsilon)}{p(1 - p)} \phi(z_{1-\alpha}) \\ &= h^2 \frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)} \\ &\quad \times [2nhf_X(0)]^{1/2} [p(1 - p)]^{-1/2} \phi(z_{1-\alpha}) f_{Y|X}(\xi_p; 0), \end{aligned}$$

$$h^{7/2} = \frac{n^{-3/2}2^{-3/2}z_{1-\alpha}\epsilon(1-\epsilon)/\sqrt{p(1-p)}}{\sqrt{f_X(0)}\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p;0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p;0)\right\}/6},$$

$$\hat{h}_{+-} = n^{-3/7} \left( \frac{z_{1-\alpha}}{3\sqrt{p(1-p)}f_X(0)\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p;0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p;0)\right\}} \right)^{2/7},$$

$$\hat{h}_{++} = \hat{h}_{+-}[-2d/(2b+d)]^{2/(2b+3d)} \approx -0.770\hat{h}_{+-}.$$

We have plugged in the rule-of-thumb  $\epsilon = 0.2$  to avoid iteration (more precisely,  $0.5 - (1/2)\sqrt{1 - 4\sqrt{2}/9} \approx 0.20$ , to get the constants to cancel). These  $\hat{h}$  hold for both lower and upper one-sided inference.

With the approximation  $u = p + O(N_n^{-1/2})$ , the bias CPE for the upper endpoint is the negative of the bias CPE for the lower endpoint, up to smaller-order terms. Then, for two-sided inference, the dominant bias terms from the upper and lower CI endpoints cancel, and the CPE is of a smaller order, as in Theorem 2.3.

### Plug-in bandwidth: two-sided, $p = 1/2$

For two-sided inference with  $p = 1/2$ , the  $B_h$  term becomes zero, as is clear in (2.13). Then  $\text{CPE}_{\text{Bias}} < 0$  since  $B_h^2 > 0$ ,  $f'_{\hat{Q}_{Y|C_h}^{I,u_\ell}}(p) < 0$ , and  $f'_{\hat{Q}_{Y|C_h}^{I,u_h}}(p) > 0$ . The optimal  $h$  causes this to cancel with  $\text{CPE}_{\text{GK}} > 0$ . By the convergence of our beta to a normal distribution, the product rule for derivatives, and the invariance of  $f_{Y|C_h}(F_{Y|C_h}^{-1}(p))$  to  $u$ ,

$$\frac{\partial}{\partial p} f_\beta(p; u_\ell) \doteq -z_{1-\alpha/2} N_n [u_\ell(1-u_\ell)]^{-1} \phi(z_{1-\alpha/2}),$$

$$\frac{\partial}{\partial p} f_\beta(p; u_h) \doteq z_{1-\alpha/2} N_n [u_h(1-u_h)]^{-1} \phi(z_{1-\alpha/2}),$$

$$\begin{aligned} & f'_{\hat{Q}_{Y|C_h}^{I,u_\ell}}(Q_{Y|C_h}(p)) - f'_{\hat{Q}_{Y|C_h}^{I,u_h}}(Q_{Y|C_h}(p)) \\ & \doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) \left( [u_\ell(1-u_\ell)]^{-1} + [u_h(1-u_h)]^{-1} \right) f_{Y|C_h}(F_{Y|C_h}^{-1}(p)) \\ & \doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|X}(\xi_p; 0), \end{aligned}$$

so plugging into (2.10) yields

$$\begin{aligned}
& N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1 - \epsilon_h) + \epsilon_\ell(1 - \epsilon_\ell)}{p(1 - p)} \phi(z_{1-\alpha/2}) \\
&= -(1/2)h^4 \left( \frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)} \right)^2 \\
&\quad \times \{ -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1 - p)]^{-1} f_{Y|X}(\xi_p; 0) \}, \\
& [2nhf_X(0)]^{-2} 2\epsilon(1 - \epsilon) \\
&= \frac{h^4}{36f_X(0)^2 f_{Y|X}(\xi_p; 0)} \left( f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0) \right)^2, \\
\hat{h} &= n^{-1/3} \left( \frac{3f_{Y|X}(\xi_p; 0)}{\left\{ f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0) \right\}^2} \right)^{1/6},
\end{aligned}$$

again using  $\epsilon = 0.2$  as the rule of thumb (and rounding 2.88 up to 3).

### Plug-in bandwidth: two-sided, $p \neq 1/2$

For two-sided inference with  $d = 1$  and  $p \neq 1/2$ , the dominant bias terms do not cancel completely, but the difference is of smaller order than the  $O(N_n^{1/2})$  in the one-sided case. Using the result from Chapter 1 that  $u_h - p = z_{1-\alpha/2} \sqrt{p(1-p)/N_n} + O(N_n^{-1})$  and  $u_\ell - p = -z_{1-\alpha/2} \sqrt{p(1-p)/N_n} + O(N_n^{-1})$ , along with the normal approximation of the beta PDF in (2.11),

$$\begin{aligned}
f_\beta(p; u_h) - f_\beta(p; u_\ell) &= N_n^{1/2} \phi(z_{1-\alpha/2}) ([u_h(1 - u_h)]^{-1/2} - [u_\ell(1 - u_\ell)]^{-1/2}) \\
&\quad \times [1 + O(N_n^{-1/2})] \\
&= N_n^{1/2} \phi(z_{1-\alpha/2}) \\
&\quad \times \left( \frac{2p - 1}{2[p(1 - p)]^{3/2}} [(u_h - p) - (u_\ell - p)] + O(N_n^{-1}) \right) + O(1) \\
&= N_n^{1/2} \phi(z_{1-\alpha/2}) \frac{2p - 1}{2[p(1 - p)]^{3/2}} N_n^{-1/2} z_{1-\alpha/2} \\
&\quad \times \left( \sqrt{u_h(1 - u_h)} + \sqrt{u_\ell(1 - u_\ell)} \right) + O(N_n^{-1/2}) + O(1) \\
&= z_{1-\alpha/2} \phi(z_{1-\alpha/2}) \frac{2p - 1}{p(1 - p)} + O(N_n^{-1/2}) + O(1) = O(1).
\end{aligned}$$

Thus our two CPE terms are of orders  $N_n^{-1} \asymp n^{-1}h^{-1}$  and  $h^2$ . This implies  $h^* \asymp n^{-1/3}$  and that CPE is  $O(n^{-2/3})$ . But then the  $B_h^2$  term is of order  $h^4 N_n = h^5 n = n^{-2/3}$ , so it must also be included. (The  $B_h^3$  term is of order  $h^6 N_n^{3/2}$ , which is smaller.) Though the second term from the product rule derivative in the  $B_h^2$  term is not zero this time, it is smaller-order and thus omitted below.

There are two approaches to a plug-in bandwidth in this case. Note that the error from approximating the beta PDF with a normal PDF was  $O(1)$ , which in this case is the dominant order. In the first approach, to determine the bandwidth that precisely cancels  $\text{CPE}_{\text{GK}}$  and  $\text{CPE}_{\text{Bias}}$ , using  $f_\beta(p; u_h) - f_\beta(p; u_\ell)$  without the normal approximation is required. A closed-form expression with the  $N_n^{1/2}$  as an explicit coefficient may be found in Appendix A. However,  $u_h$  and  $u_\ell$  are (implicitly) functions of the bandwidth  $h$ , so this is a recursive determination of  $h$ . The optimal value may be found through iteration, where the  $h$  resulting from using the normal approximation may be used as the pilot bandwidth. The iteration is straightforward since  $h$  is scalar and may be stopped once the change in  $N_n$  is small, but it still requires additional consideration and computation time. In the second approach, followed below, the normal approximation is used. This means that the plug-in bandwidth is not precise enough to reduce the order of CPE further, but it is of the proper asymptotic rate and achieves the CPE in Theorem 2.3. This is more commonly how plug-in bandwidths are; our precise plug-in bandwidths in the other cases are special luxuries.

The sign of  $\text{CPE}_{\text{Bias}}$  is determined by the sign of  $[B_h(2p-1) - B_h^2 N_n]$ , as seen in (2.13). Since  $B_h^2 N_n > 0$  always, if  $B_h(2p-1) < 0$ , then  $\text{CPE}_{\text{Bias}} < 0$  irrespective of  $h$  (only the magnitude of  $B_h$  depends on  $h$ , not the sign). If  $B_h(2p-1) > 0$ , then there may exist some  $h$  that yields  $\text{CPE}_{\text{Bias}} < 0$ , but maybe not. (There will always exist some  $h > 1$  that does this, but asymptotically  $h \rightarrow 0$ .) Specifically, for coefficients  $a$  and  $b$ ,  $ah^2 - bh^5 < 0$  for  $h > \sqrt[3]{a/b}$ . Here,  $a = (2p-1)(B_h/h^2)$  and  $b = 2nf_X(0)(B_h/h^2)^2$ , so  $\text{CPE}_{\text{Bias}} < 0$  for

$$B_h/h^2 = -\frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)},$$

$$h > \left( \frac{2p-1}{2nf_X(0)B_h/h^2} \right)^{1/3} \quad (\text{B.4})$$

$$= n^{-1/3} \left( \frac{3(2p-1)f_{Y|X}(\xi_p; 0)}{-\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0)\right\}} \right)^{1/3}.$$

When  $h$  equals the RHS,  $\text{CPE}_{\text{Bias}} = 0$ , so the optimal  $h$  should be somewhat larger. In practice, using (B.4) as an equality for  $\hat{h}$  probably works well, but we pursue the more exact solution below.

Since such a solution is always possible, we pick  $h$  to equate

$$\begin{aligned} & -N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\ & \doteq B_h \left[ f_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) - f_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) \right] \\ & \quad + (1/2) B_h^2 \left[ f'_{\hat{Q}_{Y|C_h}^{I, u_h}}(Q_{Y|C_h}(p)) - f'_{\hat{Q}_{Y|C_h}^{I, u_\ell}}(Q_{Y|C_h}(p)) \right] \\ & \doteq B_h f_{Y|C_h} \left( F_{Y|C_h}^{-1}(p) \right) [f_\beta(p; u_h) - f_\beta(p; u_\ell)] \\ & \quad - (1/2) B_h^2 z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|C_h} \left( F_{Y|C_h}^{-1}(p) \right) \\ & \doteq f_{Y|X}(\xi_p; 0) z_{1-\alpha/2} \phi(z_{1-\alpha/2}) [p(1-p)]^{-1} [B_h(2p-1) - B_h^2 N_n]. \end{aligned}$$

Note that  $2p-1 > 0$  is equivalent to  $p > 1/2$ , which implies  $|u_h - 0.5| > |u_\ell - 0.5|$ . In that case, the  $B_h$  term is the same sign as  $B_h$  itself. If  $2p-1 < 0$ , or equivalently  $p < 1/2$  or  $|u_h - 0.5| < |u_\ell - 0.5|$ , the term is the opposite sign of  $B_h$ . If  $p = 1/2$ , the term is zero, as covered in the special case.

Continuing to solve for  $h$ ,

$$\begin{aligned} & -[2nhf_X(0)]^{-1} 2\epsilon(1-\epsilon) \\ & = f_{Y|X}(\xi_p; 0) \{h^2(B_h/h^2)(2p-1) - h^4(B_h^2/h^4)[2nhf_X(0)]\}, \\ & n^{-1} \{[f_X(0)]^{-1} \epsilon(1-\epsilon)\} \\ & = h^6 n \{2f_{Y|X}(\xi_p; 0) f_X(0) B_h^2/h^4\} - h^3 \{f_{Y|X}(\xi_p; 0)(2p-1)(B_h/h^2)\}, \\ & 0 = (h^3)^2 n \{a\} - (h^3) \{b\} - \{c\}/n, \\ & h^3 = \frac{b \pm \sqrt{b^2 + 4ac}}{2an}, \\ & \hat{h} = n^{-1/3} \left( \frac{b + \sqrt{b^2 + 4ac}}{2a} \right)^{1/3} \end{aligned}$$

$$= n^{-1/3} \left( \frac{(2p-1)(B_h/|B_h|) + \sqrt{(2p-1)^2 + (4/3)/f_{Y|X}(\xi_p; 0)}}{(2/3) \left| f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f'_X(0)F_{Y|X}^{(0,1)}(\xi_p; 0) \right| / f_{Y|X}(\xi_p; 0)} \right)^{1/3},$$

again with rule-of-thumb  $\epsilon = 0.2$  and approximating  $(0.2)(0.8) \approx 1/6$  to match the median-specific bandwidth when  $p = 1/2$  is used here. The other root of the equation yields  $h < 0$  since  $a > 0$  and  $c > 0$ , so it is ignored. Even if  $b < 0$ , we will get  $\hat{h} > 0$ , for similar reasons. Notice that when  $B_h(2p-1) > 0$ , if we had plugged in  $\epsilon = 0$  to get the smallest possible  $\hat{h}$ , the resulting  $\hat{h}$  would be equal to the lower bound from (B.4).

### B.3 Corollary 2.4 proof

For all methods, since the same bandwidth (and thus effective sample) is used for each quantile, the CPE results from Chapter 1 go through. In the one-sided case, the order of CPE from bias is the same as before, i.e.  $O(N_n^{1/2}h^b) = O(n^{1/2}h^{b+d/2})$ . For joint inference where  $\text{CPE}_{\text{GK}} = O(N_n^{-1})$ ,

$$N_n^{-1} \asymp N_n^{1/2}h^b \implies (nh^d)^{3/2} \asymp h^{-b} \implies h^* \asymp n^{-3/(2b+3d)},$$

and the overall CPE is  $[n(h^*)^d]^{-1} = (n^{1-3d/(2b+3d)})^{-1} = n^{-2b/(2b+3d)}$ , the same as for a single quantile. Consequently, the one-sided plug-in bandwidths from §2.4.2 already have the desired rate. If we are considering different quantiles  $p_j$ , picking  $\hat{p} = \arg \min_{p_j} p_j(1-p_j)$  to use for  $p$  in the plug-in bandwidth expression will give the smallest and thus most conservative (highest coverage probability) bandwidth.

For one-sided inference on linear combinations of quantiles or treatment effects thereon, where in both cases  $\text{CPE}_{\text{GK}} = O(N_n^{-2/3})$ ,

$$N_n^{-2/3} \asymp N_n^{1/2}h^b \implies (nh^d)^{7/6} \asymp h^{-b} \implies h^* \asymp n^{-7/(6b+7d)},$$

and the overall CPE is  $[n(h^*)^d]^{-2/3} = (n^{6b/(6b+7d)})^{-2/3} = n^{-4b/(6b+7d)}$ .

For two-sided inference, the critical question is whether these methods share the  $\text{CPE}_{\text{Bias}}$  cancellation of the single quantile method. For joint inference, the cancellation goes through. Consider two quantiles of interest,  $p_1$  and  $p_2$ , where the objects of interest are  $Q_{Y|X}(p_1; x_0)$  and  $Q_{Y|X}(p_2; x_0)$ , and the unconditional

IDEAL inference is implicitly on  $Q_{Y|C_h}(p_1)$  and  $Q_{Y|C_h}(p_2)$ . Note that there is only one  $C_h$  since we restrict to having only one bandwidth  $h$ . For notation, let the joint CI be the Cartesian product of  $CI_1(\tilde{\alpha}) = (\hat{L}_1, \hat{U}_1)$  and  $CI_2(\tilde{\alpha}) = (\hat{L}_2, \hat{U}_2)$ , where  $\tilde{\alpha} < \alpha$  is calibrated as in Chapter 1. From the single quantile results, we know that the probability of  $\hat{L}_1$  being between  $Q_{Y|X}(p_1; x_0)$  and  $Q_{Y|C_h}(p_1)$  is equal to the probability of  $\hat{U}_1$  being between those values, up to smaller-order terms, and similarly for  $\hat{L}_2$  and  $\hat{U}_2$  being between  $Q_{Y|X}(p_2; x_0)$  and  $Q_{Y|C_h}(p_2)$ . Since those probabilities are small already, the probability of such an event occurring simultaneously at both  $p_1$  and  $p_2$  is negligible. Thus, since the coverage probability is  $(1 - \tilde{\alpha})$  at both  $p_1$  and  $p_2$ , the CPE due to bias is simply  $(1 - \tilde{\alpha})$  times the single quantile CPE due to bias, and the order-reducing cancellation from the single quantile case goes through. The optimal bandwidth is consequently the same as for single quantile inference.

For two-sided quantile treatment effect inference, the cancellation does not seem to occur in general. For simplicity, consider the special case of a median treatment effect whose true value is zero. Following the method in Chapter 1, let the lower and upper endpoints of the control group  $(1 - \tilde{\alpha})$  CI be denoted as  $\hat{L}_C$  and  $\hat{U}_C$ , and similarly  $\hat{L}_T$  and  $\hat{U}_T$  for the treatment group, so the median treatment effect CI endpoints are  $\hat{L} = \hat{L}_T - \hat{U}_C$  and  $\hat{U} = \hat{U}_T - \hat{L}_C$ . Again for simplicity, assume the bias is positive, so the CPE due to bias is  $P(0 < \hat{U} < \text{Bias}) - P(0 < \hat{L} < \text{Bias})$ :

$$\begin{aligned}
& P(\hat{Q}_{Y,T=1}^\ell - \hat{Q}_{Y,T=0}^u < 0 < \hat{Q}_{Y,T=1}^u - \hat{Q}_{Y,T=0}^\ell) \\
&= 1 - P(\hat{Q}_{Y,T=1}^\ell - \hat{Q}_{Y,T=0}^u > 0) - P(\hat{Q}_{Y,T=1}^u - \hat{Q}_{Y,T=0}^\ell < 0) \\
&= 1 - P(\hat{Q}_{Y,T=1}^\ell - \hat{Q}_{Y,T=0}^u > Q_{Y|C_h,T=1}(p) - Q_{Y|C_h,T=0}(p)) \\
&\quad + [P(\hat{Q}_{Y,T=1}^\ell - \hat{Q}_{Y,T=0}^u > Q_{Y|C_h,T=1}(p) - Q_{Y|C_h,T=0}(p)) \\
&\quad\quad - P(\hat{Q}_{Y,T=1}^\ell - \hat{Q}_{Y,T=0}^u > 0)] \\
&\quad - P(\hat{Q}_{Y,T=1}^u - \hat{Q}_{Y,T=0}^\ell < Q_{Y|C_h,T=1}(p) - Q_{Y|C_h,T=0}(p)) \\
&\quad + [P(\hat{Q}_{Y,T=1}^u - \hat{Q}_{Y,T=0}^\ell < Q_{Y|C_h,T=1}(p) - Q_{Y|C_h,T=0}(p)) \\
&\quad\quad - P(\hat{Q}_{Y,T=1}^u - \hat{Q}_{Y,T=0}^\ell < 0)] \\
&= 1 - \alpha + \text{CPE}_{\text{GK}} + \int_{\text{Bias}}^0 f_{\hat{L}_T - \hat{U}_C}(t) dt + \int_0^{\text{Bias}} f_{\hat{U}_T - \hat{L}_C}(t) dt
\end{aligned}$$

$$= 1 - \alpha + \text{CPE}_{\text{GK}} + \int_0^{\text{Bias}} [f_{\hat{U}_T - \hat{L}_C}(t) - f_{\hat{L}_T - \hat{U}_C}(t)] dt.$$

Asymptotically, as before, the dominant term depends on the difference in the PDFs evaluated at zero,  $f_{\hat{U}}(0) - f_{\hat{L}}(0)$ . Since the control and treatment groups are independent, these may be calculated by convolution. Letting  $f_{\beta, u}(\cdot)$  denote the PDF of the distribution  $\beta((N_n + 1)u, (N_n + 1)(1 - u))$ , and assuming (for simplicity) separate bandwidths  $h_T$  and  $h_C$  are chosen such that the effective control sample and effective treatment sample each have  $N_n$  observations,

$$\begin{aligned} f_{\hat{U}}(0) &= f_{\hat{U}_T - \hat{L}_C}(0) = \int_{\mathbb{R}} f_{\hat{U}_T}(t) f_{\hat{L}_C}(t) dt \\ &= \int_{\mathbb{R}} f_{\beta, u_T^h}(F_{Y|C_h, T=1}(t)) f_{Y|C_h, T=1}(t) f_{\beta, u_C^\ell}(F_{Y|C_h, T=0}(t)) f_{Y|C_h, T=0}(t) dt \\ &= \int_{\mathbb{R}} f_{\beta, u_T^h}(F_{Y|C_h, T=1}(t)) f_{\beta, u_C^\ell}(F_{Y|C_h, T=0}(t)) f_{Y|C_h, T=1}(t) f_{Y|C_h, T=0}(t) dt, \\ f_{\hat{L}}(0) &= f_{\hat{L}_T - \hat{U}_C}(0) = \int_{\mathbb{R}} f_{\hat{L}_T}(t) f_{\hat{U}_C}(t) dt \\ &= \int_{\mathbb{R}} f_{\beta, u_T^\ell}(F_{Y|C_h, T=1}(t)) f_{Y|C_h, T=1}(t) f_{\beta, u_C^h}(F_{Y|C_h, T=0}(t)) f_{Y|C_h, T=0}(t) dt \\ &= \int_{\mathbb{R}} f_{\beta, u_T^\ell}(F_{Y|C_h, T=1}(t)) f_{\beta, u_C^h}(F_{Y|C_h, T=0}(t)) f_{Y|C_h, T=1}(t) f_{Y|C_h, T=0}(t) dt. \end{aligned}$$

In certain special cases,  $f_{\hat{U}}(0) = f_{\hat{L}}(0)$ . For example, imagine  $Y|C_h, T = 1$  and  $Y|C_h, T = 0$  are both standard uniform distributions, so  $f_{Y|C_h, T=1}(t) = f_{Y|C_h, T=0}(t) = 1$  for  $t \in [0, 1]$  and zero otherwise, and  $F_{Y|C_h, T=1}(t) = F_{Y|C_h, T=0}(t) = t$ , and the question reduces to whether

$$\int_0^1 f_{\beta, u_T^h}(t) f_{\beta, u_C^\ell}(t) dt = \int_0^1 f_{\beta, u_T^\ell}(t) f_{\beta, u_C^h}(t) dt.$$

If  $p = 0.5$ , then  $u_C^h = 1 - u_C^\ell$  and  $f_{\beta, u_C^h}(t) = f_{\beta, u_C^\ell}(1 - t)$ , and similarly for  $T$ , so the integrals are the same: the second integral is equivalent to running  $t$  backward (from one to zero) with the first integrand. With  $p \neq 0.5$ , the normal approximation of the beta PDF shows that the integrals are equal up to smaller-order terms, as in the single quantile case.

However, if the conditional distributions of  $Y$  are not uniform or otherwise symmetric in a neighborhood around  $p$ , then there is a first-order difference. In that case, the optimal bandwidth and CPE rates are the same as for one-sided inference.



For the same reason, two-sided inference on linear combinations of quantiles also lacks the cancellation in general, so the optimal bandwidth and CPE rates are the same as the one-sided rates.

# Appendix C

## Technical appendix to Chapter 3

### C.1 Accuracy of fixed- $m$ critical value approximation

The approximate fixed- $m$  critical value in (3.7) is quite accurate for all but  $m = 1, 2$ , as Table C.1 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  $\alpha$ , I simulated the two-sided rejection probability for a given critical given  $T_{m,\infty}$  from (3.4) as the true distribution. One million simulation replications per  $m$  and  $\alpha$  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  $\alpha$  are available in the online appendix.

### C.2 Edgeworth expansion (Theorem 3.1) where $\gamma \neq 0$ : differences with HS88 proof

Here I highlight the differences with HS88; the working paper contains a sketch comparable to Hall and Sheather (1987, "HS87"), while the full proof is in

**Table C.1:** Simulated rejection probabilities (%) for different fixed- $m$  critical value approximations,  $\alpha = 5\%$ .

m	rejection probability for critical value			
	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

the supplemental appendix.

As in Section 3.2, the null hypothesis is  $H_0 : \xi_p = \beta$ , and the true  $\xi_p = \beta - \gamma/\sqrt{n}$ . I continue from (3.11), 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  $[np]$  increases by unit jumps) when  $\gamma \neq 0$  is the same as in HS88. The definitions for  $\Delta_1$ ,  $\Delta_2$ ,  $\Delta_3$ , and  $a_k$  are also identical, but now  $Z$  has the extra  $\gamma$  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,  $Y$ ,  $R$ , and  $\delta$  are defined the same, but  $B$  includes the additional  $\Psi$  terms defined below:

$$\begin{aligned} B &\equiv \delta\Psi + (n/m)(b_1\Delta_1 + b_2\Delta_2)\Delta_2 + (n/m)b_3(\Delta_1 + \Delta_2)(\Delta_3 + \Psi) \\ &\quad + (n/m)^2b_4(\Delta_1 + \Delta_2)^2(\Delta_3 + \Psi) + b_5\Delta_3(\Delta_3 + 2\Psi), \\ \Psi &\equiv \gamma/(pg(p)\sqrt{n}). \end{aligned}$$

The  $b_i$  are identical other than now  $b_2 \equiv -p/2$ , which is simplified to drop a term that ends up in the remainder later anyway.

Here,  $Y$  is of the same form as HS88, but with  $\Psi$  now additionally showing up in the higher-order  $B$  terms. In the definition of  $Z$ ,  $\gamma$  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  $\nu$  contains the higher-order terms that are dropped for the first-order asymptotic result,

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

Thus,  $\gamma$  enters only in higher-order terms, through the numerator of  $Z$ .

HS88 (A.2) now contains additional  $\Psi$  terms, while preserving the other terms:

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

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

Defining  $K$  and  $L$  the same way, HS88 (A.3) is reached but with different  $\alpha_i$ , whose respective Fourier-Stieltjes transforms are

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

The characteristic function of  $L$  is the same.

# Bibliography

- 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.
- Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580.
- Belluzzo, Jr., W. (2004). Semiparametric approaches to welfare evaluations in binary response models. *Journal of Business & Economic Statistics*, 22(3):322–330.
- Beran, R. and Hall, P. (1993). Interpolated nonparametric prediction intervals and confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):643–652.
- Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Annals of Statistics*, 18(3):1400–1415.
- 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.
- Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957.
- Brown, B. M., Hall, P., and Young, G. A. (2001). The smoothed median and the bootstrap. *Biometrika*, 88(2):519–534.

- Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458.
- Cade, B., Terrell, J., and Schroeder, R. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology*, 80(1):311–323.
- Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress*, volume 2, pages 171–209.
- Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.
- Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, 19(2):760–777.
- Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Annals of Statistics*, 21(3):1166–1181.
- Chernozhukov, V., Hansen, C., and Jansson, M. (2009). Finite sample inference for quantile regression models. *Journal of Econometrics*, 152:93–103.
- Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71(5):1405–1441.
- DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York.
- David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, Third Edition*. Wiley, New York.
- Duong, T. (2012). *ks: Kernel smoothing*. R package version 1.8.8.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.
- Eide, E. and Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, 58(3):345–350.
- Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on statistics and applied probability*. Chapman & Hall, London.
- Furrer, R., Nychka, D., and Sain, S. (2012). *fields: Tools for spatial data*. R package version 6.6.3.

- Gangopadhyay, A. K. and Sen, P. K. (1990). Bootstrap confidence intervals for conditional quantile functions. *Sankhyā: the Indian Journal of Statistics, Series A*, 52(3):346–363.
- 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.
- 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., DiCiccio, T. J., and Romano, J. P. (1989). On smoothing and the bootstrap. *Annals of Statistics*, 17(2):692–704.
- Hall, P. and Sheather, S. J. (1987). On the distribution of a studentized quantile. Technical Report 71, Department of Statistics, Pennsylvania State University.
- 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.
- Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163.
- Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).
- Ho, Y. H. S. and Lee, S. M. S. (2005a). Calibrated interpolated confidence intervals for population quantiles. *Biometrika*, 92(1):234–241.
- Ho, Y. H. S. and Lee, S. M. S. (2005b). Iterated smoothed bootstrap confidence intervals for population quantiles. *Annals of Statistics*, 33(1):437–462.
- Hogg, R. (1975). Estimates of percentile regression lines using salary data. *Journal of the American Statistical Association*, 70(349):56–59.
- Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, 66(6):1327–1351.

- Hotelling, H. (1939). Tubes and spheres in  $n$ -space and a class of statistical problems. *American Journal of Mathematics*, 61:440–460.
- Hutson, A. D. (1999). Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics*, 26(3):343–353.
- Hutson, A. D. (2007). An “exact” two-group median test with an extension to censored data. *Journal of Nonparametric Statistics*, 19(2):103–112.
- Hyndman, R., Bashtannyk, D., and Grunwald, G. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336.
- Janas, D. (1993). A smoothed bootstrap estimator for a studentized sample quantile. *Annals of the Institute of Statistical Mathematics*, 45:317–329.
- Jones, M. C. (2002). On fractional uniform order statistics. *Statistics & Probability Letters*, 58(1):93–96.
- Kaplan, D. M. (2011). Inference on quantiles: confidence intervals, p-values, and testing. MATLAB Central File Exchange, <http://www.mathworks.com/matlabcentral/fileexchange/32410-quantile-inference-testing-univariate-and-bivariate>. Retrieved 25 March 2012.
- Khusun, H., Yip, R., Schultink, W., and Dillon, D. (1999). World Health Organization hemoglobin cut-off points for the detection of anemia are valid for an Indonesian population. *The Journal of Nutrition*, 129(9):1669–1674.
- Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press.
- Koenker, R. (2012). *quantreg: Quantile Regression*. R package version 4.81.
- 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.
- Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407.
- Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika*, 86(4):909–921.



- Krewski, D. (1976). Distribution-free confidence intervals for quantile intervals. *Journal of the American Statistical Association*, 71(354):420–422.
- Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1–2):79–88.
- Li, Q., Lin, J., and Racine, J. S. (2013). Optimal bandwidth selection for non-parametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, 31(1):57–65.
- MaCurdy, T., Chen, X., and Hong, H. (2011). Flexible estimation of treatment effect parameters. *American Economic Review*, 101(3):544–551.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.
- Manning, W., Blumberg, L., and Moulton, L. (1995). The demand for alcohol: the differential response to price. *Journal of Health Economics*, 14(2):123–148.
- Mosteller, F. (1946). On some useful “inefficient” statistics. *Annals of Mathematical Statistics*, 17:377–408.
- Nayyar, G. (2009). Demand for services in India: A mirror image of Engel’s law for food? Working Paper 451, University of Oxford. Department of Economics Working Paper Series.
- Polansky, A. M. and Schucany, W. R. (1997). Kernel smoothing to improve bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):821–838.
- Portnoy, S. (2012). Nearly root- $n$  approximation for regression quantile processes. *The Annals of Statistics*, 40(3):1714–1736.
- R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Robbins, H. (1955). A remark on Stirling’s formula. *The American Mathematical Monthly*, 62(1):26–29.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Sathe, Y. S. and Lingras, S. R. (1981). Bounds for the confidence coefficients of outer and inner confidence intervals for quantile intervals. *Journal of the American Statistical Association*, 76(374):473–475.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
- 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.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26 of *Monographs on statistics and applied probability*. Chapman & Hall, London.
- Stigler, S. M. (1977). Fractional order statistics, with applications. *Journal of the American Statistical Association*, 72(359):544–550.
- Stone, C. J. (1980). Optimal rates of convergence for non-parametric regression. *Annals of Statistics*, 8:1348–1360.
- Stone, C. J. (1982). Optimal global rates of convergence for non-parametric regression. *Annals of Statistics*, 10:1040–1053.
- Sun, Y. (2010a). Autocorrelation robust inference using nonparametric series methods. Working paper.
- Sun, Y. (2010b). Let’s fix it: Fixed-b asymptotics versus small-b asymptotics in heteroscedasticity and autocorrelation robust inference. Working paper.
- Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. Working paper.
- Sun, Y. and Kaplan, D. M. (2011). Fixed-smoothing asymptotics and accurate F approximation using vector autoregressive variance matrix estimator. Working paper.
- Sun, Y., Phillips, P. C. B., and Jin, S. (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica*, 76(1):175–194.
- Weisberg, S. (2012). mybs. <http://users.stat.umn.edu/~sandy/courses/8053/Data/mybs.R>.
- Whang, Y.-J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, 22(2):173–205.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Zieliński, R. and Zieliński, W. (2005). Best exact nonparametric confidence intervals for quantiles. *Statistics*, 39(1):67–71.