Tests for Assessment of Agreement Using Probability Criteria

Pankaj K. Choudhary

Department of Mathematical Sciences, University of Texas at Dallas

Richardson, TX 75083-0688; pankaj@utdallas.edu

H. N. Nagaraja

Department of Statistics, Ohio State University

Columbus, OH 43210-1247; hnn@stat.ohio-state.edu

Abstract

For the assessment of agreement using probability criteria, we obtain an exact test, and for sample sizes exceeding 30, we give a bootstrap-t test that is remarkably accurate. We show that for assessing agreement, the total deviation index approach of Lin (2000, Statist. Med., 19, 255–270) is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002, J. Am. Stat. Assoc., 97, 257–270) is overly conservative for moderate sample sizes. We also show that the nearly unbiased test of Wang and Hwang (2001, J. Stat. Plan. Inf., 99, 41–58) may be liberal for large sample sizes, and suggest a minor modification that gives numerically equivalent approximation to the exact test for sample sizes 30 or less. We present a simple and accurate sample size formula for planning studies on assessing agreement, and illustrate our methodology with a real data set from the literature.

Key Words: Bootstrap; Concordance correlation; Coverage probability; Limits of agreement; Total deviation index; Tolerance interval.

1 Introduction

Suppose a paired sample of reference and test measurements, (X, Y), are available on n randomly chosen subjects from a population of interest. Generally the Y's are cheaper, faster, easier or less invasive to obtain than the X's. The question of our concern is: "Are X and Y close enough so that they can be used interchangeably?" This comparison is the goal of method comparison studies. We focus on a test of hypotheses approach for this problem that infers *satisfactory agreement* between Y and X when the difference $D = Y - X$ lies within an acceptable margin with a sufficiently high probability. We will assume that D follows a $N(\mu, \sigma^2)$ distribution. Let $F(\cdot)$ and $\Phi(\cdot)$ denote the c.d.f.'s of $|D|$ and a $N(0, 1)$ distribution, respectively.

There are two measures of agreement based on the probability criteria. The first is the p_0-th percentile of $|D|$, say $Q(p_0)$, where $p_0\,(> 0.5)$ is a specified large probability (usually ≥ 0.80). It was introduced by Lin (2000) who called it the *total deviation index* (TDI). Its small value indicates a good agreement between (X, Y). The TDI can be expressed as,

$$Q(p_0) = F^{-1}(p_0) = \sigma\big\{\chi_1^2(p_0, \mu^2/\sigma^2)\big\}^{1/2}, \tag{1}$$

where $\chi_1^2(p_0, \Delta)$ is the p_0-th percentile of a χ^2-distribution with a single degree of freedom and non-centrality parameter Δ.

The second measure, introduced by Lin et al. (2002), is the *coverage probability* (CP) of the interval $[-\delta_0, \delta_0]$, where a difference under $\pm\delta_0$ is considered practically equivalent to zero. There is no loss of generality in taking this interval to be symmetric around zero as it can be achieved by a location shift. Letting,

$$d_l = (-\delta_0 - \mu)/\sigma, \;\; d_u = (\delta_0 - \mu)/\sigma, \tag{2}$$

the CP can be expressed as

$$F(\delta_0) = \Phi(d_u) - \Phi(d_l). \tag{3}$$

A high value of $F(\delta_0)$ implies a good agreement between the methods.

For specified (p_0, δ_0), $F(\delta_0) \leq p_0 \iff Q(p_0) \geq \delta_0$. Consequently, for assessing agreement one can test either the hypotheses

$$H: \ Q(p_0) \geq \delta_0 \ \ vs. \ \ K: \ Q(p_0) < \delta_0, \tag{4}$$

or

$$H: \ F(\delta_0) \leq p_0 \ \ vs. \ K: \ \ F(\delta_0) > p_0, \tag{5}$$

and infer satisfactory agreement if H is rejected.

Let $\Theta = \{(\mu, \sigma) : |\mu| < \infty, 0 < \sigma < \infty\}$ be the parameter space. Also let Θ_H and Θ_K be the subsets of Θ representing the regions under H and K, respectively, and Θ_B be the boundary of H. These regions can be visualized through the solid curves in Figure 1 for $p_0 = 0.80, 0.95$. Notice that they are symmetric in μ about zero.

[Figure 1 about here.]

Lin (2000) suggests a large sample test for the hypotheses (4) and we refer to it as the TDI test. Lin et al. (2002) suggest a large sample test for (5), and we call it the CP test. They also conclude that these tests are more powerful for inferring satisfactory agreement than the test based on *concordance correlation* of Lin (1989). Sometimes the hypotheses (5) are used to assess the individual bioequivalence of a test formulation Y of a drug with a reference formulation X. In this context, Wang and Hwang (2001) proposed a nearly unbiased test (NUT) and showed that it was more powerful than various other tests

available in the bioequivalence literature. They present numerical results to verify that a nominal level α NUT also has size α, and call it *nearly unbiased* as its type-I error probability on the boundary Θ_B remains close to α.

We revisit all these tests in detail in Section 2. There we show that the TDI test for (4) may not be consistent and may have type-I error probability of one in the limit as $n \to \infty$. We demonstrate that the CP test for (5) is overly conservative — its empirical type-I error probability can be 3% or less for a nominal 5% level when $n \leq 50$ and $p_0 \geq 0.80$. Furthermore, even though the hypotheses (4) and (5) are equivalent, these two tests with a common level may lead to different conclusions. We also show that the NUT can be liberal when n is large.

In this article we discuss alternative tests that address the above concerns. Section 3 describes an exact test, which gives the same inference regarding agreement under setups (4) and (5). To simplify its implementation, we suggest a minor modification of the NUT, say MNUT, producing a simple closed form approximation numerically equivalent to the exact test for $n \leq 30$. We present an easy to use sample size formula for planning studies based on this test. The MNUT can also be liberal for large n. But when both the NUT and MNUT have their levels of significance under the nominal level, the MNUT is more nearly unbiased and appears to be slightly, but uniformly, more powerful than the corresponding NUT. For $n \geq 30$, we provide in Section 4 an asymptotic level α parametric bootstrap-t test. It has better properties than the asymptotic TDI and CP tests for samples of moderate sizes. In Section 5, we illustrate the new procedures with a real data set. Section 6 contains some concluding remarks.

2 The current approaches

Let (X_i, Y_i) and $D_i = Y_i - X_i$, $i = 1, \ldots, n$, denote the random samples from (X, Y) and D populations. Also let $\hat{\mu} = \sum_i D_i/n$ and $\hat{\sigma}^2 = \sum_i (D_i - \hat{\mu})^2/n$ be the maximum likelihood estimators (MLE's) of μ and σ^2, respectively. The MLE's of functions of μ and σ^2 are obtained by plugging-in these MLE's in the functional form and are denoted by the hat notation. We will use $\phi(\cdot)$ for the p.d.f. of a $N(0,1)$ distribution and $z(\alpha)$ for its α-th percentile, and $t_k(\alpha, \Delta)$ for the α-th percentile of a t-distribution with k degrees of freedom and non-centrality parameter Δ.

2.1 Lin's (2000) TDI test

Lin (2000) argues that the inference based on $Q(p_0)$ in (1) is intractable. He approximates $Q(p_0)$ by $Q^*(p_0) = \{(\mu^2 + \sigma^2)\chi^2_1(p_0, 0)\}^{1/2} = (\mu^2 + \sigma^2)^{1/2} z_0$, with z_0 defined as

$$z_0 = z((1 + p_0)/2), \tag{6}$$

and modifies the hypotheses (4) to

$$H^* : \; Q^*(p_0) \geq \delta_0 \;\; vs. \;\; K^* : \; Q^*(p_0) < \delta_0. \tag{7}$$

As Figure 1 illustrates, the approximation $Q^*(p_0)$ for $Q(p_0)$ is good in terms of having close null and alternative hypotheses regions only when $\mu/\sigma \approx 0$. The figure plots the regions under (H^*, K^*) for $p_0 = 0.80, 0.95$ with boundary marked by dashed lines. When $p_0 = 0.80$, the two boundaries intersect, but when $p_0 = 0.95$, $K^* \subset K$, which also happens when $p_0 = 0.85, 0.90$ (not shown).

Lin's TDI test rejects H^* if the large-sample upper confidence bound (UCB) for $Q^*(p_0)$ is less than δ_0, i.e., $\exp\{0.5(\ln(U) + z(1-\alpha)V + 2\ln(z_0))\} < \delta_0$, where $U = \sum_{i=1}^{n} D_i^2/(n-1)$ and $V^2 = 2(1 - \hat{\mu}^4/U^2)/(n-1)$. Since this test has asymptotic level α and is consistent for (H^*, K^*), whenever $K^* \subset K$, there are $(\mu, \sigma) \in \{K \cap H^*\}$ where the agreement is satisfactory, but the test concludes otherwise with probability approaching one as $n \to \infty$. This happens with the four p_0 values when $\mu^2/\sigma^2 >> 0$. Moreover, the test may incorrectly conclude satisfactory agreement with probability approaching one when $(\mu, \sigma) \in \{H \cap K^*\}$, a possible scenario with $p_0 = 0.80$, as Figure 1 indicates.

2.2 Lin et al.'s (2002) CP test

This test involves the estimation of $F(\delta_0)$. To reduce bias, the authors estimate σ^2 by $\hat{\sigma}_L^2 = n\hat{\sigma}^2/(n-3)$. Let $\lambda = \ln\{F(\delta_0)/(1 - F(\delta_0))\}$, $\lambda_0 = \ln\{p_0/(1-p_0)\}$ and

$$\tau_\lambda^2 = \big\{\big(\phi(d_u) - \phi(d_l)\big)^2 + 0.5\big(d_u\phi(d_u) - d_l\phi(d_l)\big)^2\big\}/\big\{F(\delta_0)(1 - F(\delta_0))\big\}^2,$$

with (d_l, d_u) defined by (2). The asymptotic level α CP test for (5) rejects H if $(n-3)^{1/2}(\hat{\lambda}_L - \lambda_0)/\hat{\tau}_{\lambda,L} > z(1-\alpha)$. The subscript L in the estimates implies that $\hat{\sigma}_L^2$ is used to estimate σ^2.

2.3 Wang and Hwang's (2001) NUT

Here the unbiased estimator $\tilde{\sigma}^2 = n\hat{\sigma}^2/(n-1)$ is used for σ^2. Let $\tilde{F}(\delta_0)$ be the estimator of $F(\delta_0)$ that uses $\tilde{\sigma}^2$ in place of $\hat{\sigma}^2$. The nominal level α NUT for (5) is: reject H if $\tilde{F}(\delta_0) > c_{nu}$ with $c_{nu} = \Phi\big[-n^{-1/2}t_{n-1}\{\alpha, -n^{1/2}z(p_0)\}\big]$. For this test, $\lim_{\sigma\to 0} Pr(\text{reject } H \mid \Theta_B) = \alpha$.

Wang and Hwang present numerical results to argue that this test has size α. Although this claim is true for the settings investigated by them, it is not true in general because

when n is large, the type-I error probability on Θ_H is not maximized when $\sigma \to 0$. Figure 2 plots the numerically computed minimum and maximum of these probabilities over Θ_B, for $\alpha = 5\%$, $p_0 = 0.80(0.05)0.95$ and $5 \leq n \leq 200$. The maximum exceeds the nominal 5% level for $p_0 \geq 0.85$. Moreover, using standard asymptotics, it is easy to see that NUT's asymptotic type-I error probability on Θ_B is,

$$1 - \Phi \left\{ \frac{z(1-\alpha)\,\phi\big(z(p_0)\big)\,\big(1 + z^2(p_0)/2\big)^{1/2}}{\Big(\big(\phi(d_u) - \phi(d_l)\big)^2 + 0.5\,\big(d_u\phi(d_u) - d_l\phi(d_l)\big)^2\Big)^{1/2}} \right\}.$$

Maximizing this expression with respect to $(\mu, \sigma) \in \Theta_B$ gives the asymptotic size of NUT. A numerical evaluation of these sizes for the four p_0 values correspondingly produces 5.02%, 5.08%, 5.23% and 5.41%. This shows that the NUT can be liberal for large n.

[Figure 2 about here.]

It may be noted that the critical point c_{nu} is exact for testing the hypotheses

$$H_{nu}: \; Pr(D \leq \delta_0) \leq p_0 \;\; vs. \; K_{nu}: \;\; Pr(D \leq \delta_0) > p_0. \tag{8}$$

This technique of approximating the critical point for a test involving two-tailed probability by the critical point for a test involving one-tailed probability is well established in the acceptance sampling literature. There the interest lies in the parameter $1 - F(\delta_0)$, the fraction of non-conforming items, and the test of hypotheses:

$$H': \; F(\delta_0) \geq p_0 \;\; vs. \; K': \;\; F(\delta_0) < p_0,$$

where $1-p_0$ is an acceptable quality level. These hypotheses can be obtained by interchanging H and K of (5). In acceptance sampling, it is also well-known that the tests based on MLE's are better than the one using the unbiased estimator of σ^2 (see e.g., Hamilton and Lesperance, 1995). This motivates some of our results in the next section.

3 An exact test and its approximation for $n \leq 30$

3.1 The test

A size α test for (4) and (5) based on the exact distribution for the statistic $\hat{F}(\delta_0)$ is the following: reject H if $\hat{F}(\delta_0) > c$, where the critical point c is such that $\sup_{\Theta_H} Pr(\hat{F}(\delta_0) > c) = \alpha$. The following result gives the power function of this test.

Proposition 1. *For a fixed $(\mu, \sigma) \in \Theta$, we have*

$$Pr(\hat{F}(\delta_0) > c) = \int_0^m \left[\Phi\{n^{1/2}\big(t(\sigma w_1) - \mu\big)/\sigma\} - \Phi\{n^{1/2}\big(-t(\sigma w_1) - \mu\big)/\sigma\}\right] h_{n-1}(w)dw, \quad (9)$$

where $m = (n\delta_0^2)/\{\sigma z\big((1+c)/2\big)\}^2$, $w_1 = (w/n)^{1/2}$, $t(s)$ is the non-negative solution of

$$\Phi\big((\delta_0 - t)/s\big) - \Phi\big((-\delta_0 - t)/s\big) = c,$$

for t, and h_{n-1} is the p.d.f. of a χ^2_{n-1}-distribution. This probability function is symmetric in μ about zero, and subject to $(\mu, \sigma) \in \Theta_H$, it is maximized on Θ_B.

This probability remains invariant under the transformation of D to D/δ_0. After this rescaling, H and K of (5) respectively become $H_1 : F_1(1) \leq p_0$ and $K_1 : F_1(1) > p_0$, say, where F_1 is the c.d.f. of $|D|/\delta_0$. As a result, the critical points for testing (H, K) and (H_1, K_1) are the same. Thus c does not depend on δ_0, and in particular, one can take $\delta_0 = 1$ for its computation by solving $\sup_{\Theta_B} Pr(\hat{F}(\delta_0) > c) = \alpha$. Note, however, that the value of (μ, σ) that maximizes the probability on the LHS is not available in a closed form. So one must use numerical maximization for this purpose.

Once c is available, the $1 - \alpha$ UCB for $Q(p_0)$, say $\hat{Q}_+(p_0)$, can simply be obtained as:

$$\hat{Q}_+(p_0) = \hat{\sigma}\{\chi_1^2(c, \hat{\mu}^2/\hat{\sigma}^2)\}^{1/2}. \quad (10)$$

The interval $[-\hat{Q}_+(p_0), \hat{Q}_+(p_0)]$ can be formally interpreted as a tolerance interval (TI) for the $N(\mu, \sigma^2)$ distribution of D with probability content p_0 and confidence $1-\alpha$ (see also p. 259 of Lin et al. (2002) for a similar note). Our next result verifies these claims. Notice that this TI is centered at zero, but the TI that is generally constructed for a normal distribution is centered at $\hat{\mu}$ (see e.g., Guttman, 1988).

Proposition 2. *For $\hat{Q}_+(p_0)$ defined by (10), we have*

(a) $\inf_\Theta Pr\big(Q(p_0) \le \hat{Q}_+(p_0)\big) = 1-\alpha$, *i.e., $\hat{Q}_+(p_0)$ is a $1-\alpha$ UCB for $Q(p_0)$.*

(b) $\inf_\Theta Pr\big(\Phi\{(\hat{Q}_+(p_0)-\mu)/\sigma\} - \Phi\{(-\hat{Q}_+(p_0)-\mu)/\sigma\} \ge p_0\big) = 1-\alpha$, *i.e., $[-\hat{Q}_+(p_0), \hat{Q}_+(p_0)]$ is a TI for a $N(\mu, \sigma^2)$ distribution with p_0 probability content and $1-\alpha$ confidence.*

This exact methodology requires numerical computations for its implementation that may not be readily available in practice. So we next discuss a good closed-form approximation.

3.2 The approximation

Approximate the critical point c by

$$c_{ml} = \Phi\big[-(n-1)^{-1/2} t_{n-1}\{\alpha, -n^{1/2} z(p_0)\}\big], \tag{11}$$

where c_{ml} is the exact critical point for the test of hypotheses (8) using the MLE's. Since c_{ml} is obtained by replacing $n^{-1/2}$ in c_{nu} with $(n-1)^{-1/2}$, the NUT and this modified NUT (MNUT) approximation of the exact test are equivalent for large n; but as we will see in the next subsection, it does make some difference for small to moderate sample sizes.

To assess the accuracy of MNUT, we numerically compute the values of c and c_{ml}, and the resulting type-I error probabilities on Θ_B, for $\alpha = 5\%$, $p_0 = 0.80(0.05)0.95$ and $n = 5(5)200$. The absolute and relative differences between c and c_{ml} over all these settings are less than 0.0002 and 0.02% respectively. Thus the two critical points are essentially the same since in applications one typically uses 2-4 decimal places. An explanation for this small difference is that the type-I error probability of the exact test tends to be maximized on Θ_B when σ is near zero. This maximum is then set to α to solve for c, whereas c_{ml} is obtained by setting the limiting probability as $\sigma \to 0$ to α. Just like the NUT, this MNUT can also be liberal for large n. The maximum type-I error probability curves in Figure 2 suggest that when $p_0 = 0.80$, the size of MNUT does not exceed the nominal 5% level till $n = 200$. But it does exceed the nominal level for $n \geq 80, 40$ and 35 when $p_0 = 0.85, 0.90$ and 0.95, respectively. Also due to its asymptotic equivalence with the NUT, its limiting sizes for the four p_0 values are respectively 5.02%, 5.08%, 5.23%, and 5.41%. In addition, whenever the size of MNUT equals the nominal level, its type-I error probabilities coincide with those of the exact test. So overall, for $0.80 \leq p_0 \leq 0.95$ and $n \leq 30$, c_{ml} and c are practically equivalent. From (10), this leads to an approximate UCB $\hat{Q}_+(p_0)$ as,

$$\hat{Q}_+(p_0) \approx \hat{\sigma}\{\chi^2_1(c_{ml}, \hat{\mu}^2/\hat{\sigma}^2)\}^{1/2}.$$

One can also solve $c_{ml} = \hat{F}(\delta_0)$ for α to obtain an approximate p-value for the test as,

$$p\text{-value} \approx G_{n-1}\{-(n-1)^{1/2} z\{\hat{F}(\delta_0)\}; -n^{1/2} z(p_0)\},$$

where $G_k(x; \Delta)$ is the c.d.f. of a t_k-distribution with non-centrality parameter Δ. It is also possible to solve $c_{ml} = \hat{F}(\delta_0)$ for p_0 to obtain an approximate $1 - \alpha$ lower confidence bound (LCB) for $F(\delta_0)$, say $\hat{F}_-(\delta_0)$. Finally, analogous to the approximation in acceptance

sampling (see Hamilton and Lesperance, 1995), the smallest sample size that ensures at least $1-\beta\,(>\alpha)$ power for the exact level α test at $F(\delta_0)=p_1\,(>p_0)$ can be approximated as,

$$n \approx \left(1+\frac{k^2}{2}\right)\left(\frac{z(1-\alpha)+z(1-\beta)}{z(1-p_0)-z(1-p_1)}\right)^2;\;\; k=\frac{z(1-\beta)z(1-p_0)+z(1-\alpha)z(1-p_1)}{z(1-\alpha)+z(1-\beta)}.$$

This simple approximation is also quite good as evident from Table 1 where the exact and approximate values of n are presented for $(\alpha, 1-\beta)=(5\%, 80\%)$ and several (p_0, p_1) settings. The approximate n's tend to be a little smaller than the exact ones. However, their maximum difference is 4, which only occurs when a relatively large n is needed.

[Table 1 about here.]

3.3 Comparison

For both equivalent formulations in (4) and (5) the exact test and its MNUT approximation would lead to the same conclusion for any data set. This is an important practical advantage over the CP and the TDI tests.

Figure 2 also demonstrates that at every (n, p_0) combination where the sizes of both MNUT and NUT equal the nominal level, the minimum type-I error rate on Θ_B for the former is always slightly higher than the latter. In this sense, the former is more *nearly unbiased* than the latter. Numerical computations indicate that the exact test and its MNUT approximation are slightly, but uniformly, more powerful than the NUT. The maximum power of the exact test/MNUT and the NUT seem to be the same at every $F(\delta_0) > p_0$, but the minimum power of the former is at least as large as the minimum of the latter. The increase in the minimum power of the former is small, only under 2% for $n \leq 30$. Nevertheless, there is an important theoretical advantage. This is because the power band,

the graph of the gap between the minimum and the maximum of the power function against $F(\delta_0)\,(> p_0)$, for the exact test/MNUT is thinner than the band for the NUT. When the two maxima agree, a thinner band is desirable since it tends to treat all (μ, σ) configurations for a given $F(\delta_0)$ in a more similar way, and also tends to ensure a higher power for a larger $F(\delta_0)$ value. So overall, for $n \leq 30$ and $0.80 \leq p_0 \leq 0.95$, MNUT is the best option from both theoretical and practical considerations.

4 The bootstrap-t test for $n \geq 30$

4.1 The test

Let $z_B(\alpha)$ be the α-th percentile of the parametric bootstrap distribution of the studentized variable $n^{1/2}\{\ln(\hat{Q}(p_0)) - \ln(Q(p_0))\}/(\hat{\tau}_Q/\hat{Q}(p_0))$, where

$$\tau_Q^2 = \sigma^2 \left(\big(\phi(q_u) - \phi(q_l)\big)^2 + 0.5\big(q_u\phi(q_u) - q_l\phi(q_l)\big)^2 \right) / \big(\phi(q_l) + \phi(q_u)\big)^2; \tag{12}$$

with $q_l = (-Q(p_0) - \mu)/\sigma$ and $q_u = (Q(p_0) - \mu)/\sigma$. The approximate $(1-\alpha)$ level UCB for $Q(p_0)$ obtained using the bootstrap-t method (see e.g., Efron and Tibshirani, 1993) is:

$$\hat{Q}_+^B(p_0) = \exp\left[\{\ln(\hat{Q}(p_0)) - z_B(\alpha)\hat{\tau}_Q/(n^{1/2}\hat{Q}(p_0))\}\right],$$

and the asymptotic level α bootstrap test rejects H if $\hat{Q}_+^B(p_0) < \delta_0$. This procedure is based on the following result.

Proposition 3. *For a fixed $p_0 \in (0,1)$ and $(\mu, \sigma) \in \Theta$, the asymptotic distribution of $n^{1/2}\{\ln(\hat{Q}(p_0)) - \ln(Q(p_0))\}/(\hat{\tau}_Q/\hat{Q}(p_0))$ is $N(0,1)$.*

The above test is implemented on the natural-log scale as this transformation reduces the skewness and was originally observed by Lin (2000). The approximate $(1-\alpha)$ level LCB $\hat{F}_{-}^{B}(\delta_0)$ for $F(\delta_0)$ can be graphically obtained by plotting $\hat{Q}_{+}^{B}(p_0)$ against p_0 and finding the smallest p_0 for which $\hat{Q}_{+}^{B}(p_0) \geq \delta_0$.

4.2 Comparison with asymptotic tests

To compare this bootstrap test with the TDI and CP tests for samples of moderate size, we compute the empirical type-I error probabilities of the three using Monte Carlo simulation. We restrict only to the non-positive values of μ as the distributions underlying these tests are symmetric in μ about zero. For δ_0, p_0 and α, as before, we take $\delta_0 = 1.0$, $p_0 = 0.80(0.05)0.95$ and $\alpha = 5\%$. Moreover, as the probabilities depend on (μ, σ) configurations that produce the given $F(\delta_0) = p_0$, the simulation is performed at 11 settings of $(\mu, \sigma) \in \Theta_B$. They are obtained by taking 11 equally spaced values of $Pr(D \geq \delta_0) = p_u(\delta_0)$ (say) in the interval $[10^{-4}, (1-p_0)/2]$ and using the fact that Θ_B can be characterized as the set

$$\left\{0 < p_u(\delta_0) < 1 - p_0, d_u = z\big(1 - p_u(\delta_0)\big), \sigma = 2\delta_0 / \left\{d_u - z\big(\Phi(d_u) - p_0\big)\right\}, \mu = \delta_0 - d_u\sigma\right\}. \tag{13}$$

At each (μ, σ) setting, we generate a random sample of size n from a $N(\mu, \sigma^2)$ distribution and perform the three tests in the same way as one would do in practice with a real data set. This process is repeated 25,000 times and the proportion of times H is rejected gives the empirical type-I error rates. They are plotted in Figure 3 for $n = 30$. For the computation of $z_B(0.05)$, 1999 parametric bootstrap resamples of each Monte Carlo sample is used. The numerically computed error rates for the exact test are also included in the figure for comparison.

[Figure 3 about here.]

In all the four cases, the error rates for the bootstrap test remain stable around 5% and vary between 4.8-5.3%. For the CP test, they generally lie between 1%-2%. For the TDI test, as expected from the discussion in Section 2.1, their variation depends on the following scenarios: (i) $p_u(1.0) \approx 0$ (i.e., $\mu^2/\sigma^2 >> 0$) and $p_0 = 0.80$; (ii) $p_u(1.0) \approx 0$ and $p_0 \geq 0.85$; (iii) $p_u(1.0) \approx (1 - p_0)/2$ (i.e., $\mu^2/\sigma^2 \approx 0$). The rates mostly lie between 5.7-6.2% for (i), 2.5-5.6% for (ii), and 5.4-5.7% for (iii). Thus, on the whole, the TDI test appears liberal.

The above pattern does not change much for the bootstrap and CP tests for $n = 50, 100$ (results not shown) except that the error rates come closer to the nominal 5%. However, for these sample sizes, the rates for the TDI test, increase towards one for (i), decrease towards zero for (ii), and come closer to 5% for (iii). This clearly demonstrates that the bootstrap test is remarkably accurate and is the best among the three for moderate sample sizes. For n under 30, it tends to be liberal and is not recommended. Additionally, a simulation based power comparison for $30 \leq n \leq 100$ indicates that it is more powerful than the CP test. This is anticipated since the latter is overly conservative. Such a comparison of the bootstrap test with the TDI test is not realistic due to the liberal nature of the latter.

Remark 1. This bootstrap test outperforms several other large-sample tests in moderate sample performance. The list includes the test with $z(\alpha)$ in place of $z_B(\alpha)$, the tests based on the minimum variance unbiased estimator and the MLE of $F(\delta_0)$, among others. Some of these results are available in Choudhary and Nagaraja (2003).

Remark 2. We also computed the empirical type-I error rates at the settings studied by Lin et al. (2002) in their Tables 2 and 3. These rates are also found to be less than 3%

for a nominal 5% level. They do not agree well with those reported in Lin et al. because for computing empirical rates they use the true value of τ_λ whereas the CP test statistic $(n-3)^{1/2}(\hat{\lambda}_L - \lambda_0)/\hat{\tau}_{\lambda,L}$ that we use has its estimate in the denominator. Further, the rates presented here for the TDI test are not comparable to those reported in Lin (2000) as we are interested in testing hypotheses (4)-(5), but Lin focuses on (7).

4.3 Comparison with the exact test

From the Figure 3, it may be safe (up to the Monte Carlo error) to conclude that the type-I error rates of the bootstrap test remain closer to the nominal level than those for the exact test. It also appears that the former is slightly more powerful than the latter when n is moderately large. The increase in power is up to about 1.5% for $n = 30$. Further since the bootstrap test can be easily implemented, it provides a better option than the exact test for $n \geq 30$. One disadvantage, however, of the former is that a straightforward approach for estimating sample size is not available — it would involve a great deal of computation.

5 An Application

Tekten et al. (2003) compare a new approach for measuring myocardial performance index (MPI) that uses pulsed-wave tissue Doppler echocardiography with the conventional method based on pulse Doppler. Using their data, we focus on the assessment of agreement between the new method (Y) and the conventional method (X) with respect to average MPI measurements in the patients suffering from dilated cardiomyopathy (DCMP). There are $n = 15$ pairs of MPI measurements ranging between 0.50 to 1.20, and have correlation 0.95. Their

scatterplot and the plot of means versus differences can be found in Tekten et al. The MLE of (μ, σ) is $(\hat{\mu}, \hat{\sigma}) = (0.011, 0.044)$.

These authors use the 95% limits of agreement approach of Bland and Altman (1986) and claim that the methods agree well when the observed absolute differences are as large as 0.10. So we take $(p_0, \delta_0) = (0.95, 0.10)$ as the cutoffs. Using just $p_0 = 0.95$, the MLE of $Q(p_0)$ is 0.0889 and its exact and approximate 95% UCB come out to be the same, $\hat{Q}_+(p_0) = 0.1305$. Thus 95% of the differences are estimated to lie within ± 0.1305 with 95% confidence. But since $\hat{Q}_+(p_0) > \delta_0 = 0.10$, the 5% level test fails to reject H, and has a p-value of 0.3459. Analogously, using just the cutoff $\delta_0 = 0.10$, the MLE of $F(\delta_0)$ is 0.9726, and both its exact and approximate LCB equal $\hat{F}_-(\delta_0) = 0.8694$. Thus the interval $[-\delta_0, \delta_0]$ is estimated to contain 86.94% of the differences with 95% confidence. Since this LCB is less than $p_0 = 0.95$, as before, the 5% level test fails to reject H with p-value 0.3459. So for the cutoffs $(p_0, \delta_0) = (0.95, 0.10)$, there is no evidence of agreement. But if $(p_0, \delta_0) = (0.95, 0.14)$ are taken, the agreement will be inferred satisfactory because in this case $\hat{Q}_+(p_0) = 0.1305 < \delta_0$ or equivalently $\hat{F}_-(\delta_0) = 0.9642 > p_0$, with a p-value of 0.0261.

Our conclusion on the basis of $(p_0, \delta_0) = (0.95, 0.10)$ contrasts with that of Tekten et al. who infer satisfactory agreement using the 95% limits of agreement, $(\hat{\mu} - 1.96\tilde{\sigma}, \hat{\mu} + 1.96\tilde{\sigma}) = (-0.078, 0.100)$. However, inferring agreement when the 95% limits of agreement are within $\pm\delta_0$ is a less stringent criterion than inferring agreement when $\hat{Q}_+(0.95) < \delta_0$. This is because the TI $[-\hat{Q}_+(0.95), \hat{Q}_+(0.95)]$ is expected to be wider than the 95% limits of agreement since the probability content of the former interval is 0.95 with probability $1 - \alpha$, whereas the content of the latter is approximately 0.95 only on average.

If it were hard to specify δ_0 in advance, the subject matter expert could compute the

TI $[-\hat{Q}_+(p_0), \hat{Q}_+(p_0)]$ for a given p_0; and judge whether a difference in this interval is small enough to be clinically unimportant. An affirmative answer means that the agreement is inferred satisfactory. This way the assessment of agreement can be accomplished without explicitly specifying a δ_0, provided a p_0 can be specified.

Table 2 compares the bounds $\hat{Q}_+(0.95)$, $\hat{F}_-(0.10)$ and $\hat{F}_-(0.14)$ derived using various methods discussed in this article. The UCB and LCB from the NUT are respectively slightly larger and smaller than those for the exact test or its MNUT approximation. This is expected as the latter are slightly more powerful than the former. Further since $n = 15$ is not large, we do not expect the asymptotic approaches to work well, but the bootstrap and TDI bounds are not far from the exact ones. The latter performs well because $\hat{\mu}^2/\hat{\sigma}^2 = 0.062 \approx 0$. The CP bounds are quite off. The directions of the differences between these bounds and the exact ones are consistent with our earlier observation on their moderate sample performance.

[Table 2 about here.]

6 Concluding remarks

In this article, we considered the problem of assessing agreement between two methods of continuous measurement. Assuming that the differences of the paired measurements are i.i.d. normal, we concentrated on the probability criteria approach that infers satisfactory agreement between the methods when a sufficiently large proportion of differences lie within an acceptable margin. We evaluated several tests of hypotheses (H, K) in this context and proposed alternatives. We also discussed the construction of the relevant confidence bounds for the quantile function $Q(p_0)$ and the c.d.f. $F(\delta_0)$, the tolerance interval for the distribution

of differences, and the computation of sample sizes. A discussion of the point estimation of $F(\delta_0)$ and $Q^2(p_0)$ can be found in Hamilton and Lesperance (1997) and Johnson, Kotz and Balakrishnan (1995, ch. 18), respectively. We saw that using the critical point for the hypotheses on $Pr(D \leq \delta_0)$ as a critical point for the hypotheses on $Pr(|D| \leq \delta_0)$ may lead to a liberal test for large n. This practice is quite common in the acceptance sampling literature and our finding may also be of interest there.

The assumption of identical normal distribution for the differences is crucial for the methodology described here. Sometimes when this assumption is questionable, a simple transformation, such as the natural-log, may correct the problem (see Hawkins, 2002 for examples); and the agreement can be assessed on the transformed scale. However, in many cases, such a transformation may be difficult to find or the differences on the transformed scale may not be easily interpretable. Recently Choudhary and Ng (2005) have addressed this issue by extending this methodology to the normal theory regression setup, where the means or the variances of the differences are modelled as functions of a covariate.

When one infers that the agreement is not satisfactory, the perceived disagreement may be due to the poor agreement of the test or the reference method to itself (see e.g., Bland and Altman, 1999). The agreement of a method to itself is also called the *reproducibility* (or *reliability* or *repeatability*) of the method. Sometimes the test method may be more reproducible than the reference method. It is therefore important to also assess their reproducibility. A measure for this purpose that is similar to $Q(p_0)$ is developed in Choudhary (2005).

A computer program for implementing the methodologies proposed here using the popular statistical package R (R Development Core Team, 2004) is available at `http://www.utdallas.edu/~pankaj/prob_criteria/r_code.html`.

Appendix

Proof of Proposition 1. First observe that $F(\delta_0) = \Phi\big((\delta_0 - \mu)/\sigma\big) - \Phi\big((-\delta_0 - \mu)/\sigma\big)$ is a decreasing function of σ for a fixed μ, and is also decreasing function of $|\mu|$ for a fixed σ. Thus taking $\mu = 0$ maximizes $F(\delta_0)$ over μ. Setting this maximum $F(\delta_0)$ equal to p_0 and solving for σ we get $\sigma = \delta_0/z_0$, where z_0 is defined in (6). Now for a fixed $\sigma < \delta_0/z_0$, let $a(\sigma)$ be the non-negative solution of $F(\delta_0) = p_0$ with respect to μ. As a result $\Theta_K = \big\{(\mu, \sigma) : \sigma < \delta_0/z_0,\ |\mu| < a(\sigma)\big\}$ represents the region where $F(\delta_0) > p_0$.

This argument also shows that the event $\{\hat{F}(\delta_0) > c\} = \{\hat{\sigma} < \delta_0/z\big((1+c)/2\big), |\hat{\mu}| < t(\hat{\sigma})\}$. The probability expression now follows since $W = n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$, $Z = n^{1/2}(\hat{\mu} - \mu)/\sigma \sim N(0,1)$, and they are independent. The symmetry can be checked by replacing μ with $-\mu$ and using the fact that $\Phi(x) = 1 - \Phi(-x)$.

For the next part, we write the null region as $\Theta_H = \{(\mu, \sigma) : \sigma < \delta_0/z_0, |\mu| \geq a(\sigma)\} \cup \{(\mu, \sigma) : \sigma \geq \delta_0/z_0\} = \omega_1 \cup \omega_2$ (say). Further in (9): (i) the upper limit of integration is a decreasing function for σ, (ii) when σ remains fixed, the integrand is a decreasing function for $|\mu|$, and (iii) when $\mu = 0$, the integrand is a decreasing function of σ. So clearly in ω_1, the probability is maximized when $|\mu| = a(\sigma)$, and in ω_2 it is maximized when $\mu = 0, \sigma = \delta_0/z_0$. Thus in both cases the maximum occurs on the boundary. □

Proof of Proposition 2. (a) Recall that c is the critical point of the size α test of (5) for a fixed $(p_0 > 0.5, \delta_0 > 0)$, and the size is attained on the boundary Θ_B of H. Hence we have,

$$1 - \alpha = \inf_{\Theta_B} Pr\big(\hat{F}(\delta_0) \leq c\big) = \inf_{\Theta_B} Pr\big(\delta_0 \leq \hat{F}^{-1}(c)\hat{Q}_+(p_0)\big) = \inf_{\Theta_B} Pr\big(Q(p_0) \leq \hat{Q}_+(p_0)\big),$$

where $\hat{F}(\cdot)$ is the c.d.f. of $|D|$ when $(\mu, \sigma) = (\hat{\mu}, \hat{\sigma})$, and the last equality holds because

$Q(p_0) = \delta_0$ on Θ_B. Since $\delta_0 > 0$ is arbitrary, it follows that,

$$1 - \alpha = \inf_{\delta_0 > 0} \inf_{\Theta_B} Pr\big(Q(p_0) \leq \hat{Q}_+(p_0)\big) = \inf_{\Theta} Pr\big(Q(p_0) \leq \hat{Q}_+(p_0)\big).$$

Thus (a) holds. The part (b) follows from (a) as $\{Q(p_0) \leq \hat{Q}_+(p_0)\} = \{p_0 \leq F\{\hat{Q}_+(p_0)\}\}$, and $F(x) = \Phi\{(x-\mu)/\sigma\} - \Phi\{(-x-\mu)/\sigma\}$. □

Proof of Proposition 3. It is well-known that when D_i's are i.i.d. $N(\mu, \sigma^2)$, the limiting distributions of $n^{1/2}(\hat{\mu} - \mu)$ and $n^{1/2}(\hat{\sigma}^2 - \sigma^2)$ are independent $N(0, \sigma^2)$ and $N(0, 2\sigma^4)$, respectively. Now since $Q(p_0) \equiv Q$ (say) is a smooth function of (μ, σ^2), an application of the bivariate delta method (see, e.g., ch. 5, Lehmann, 1999) shows that the limiting distribution of $n^{1/2}(\hat{Q} - Q)$ is $N(0, \tau_Q^2)$, where

$$\tau_Q^2 = \sigma^2(\partial Q/\partial\mu)^2 + 2\sigma^4(\partial Q/\partial\sigma^2)^2. \tag{14}$$

To find the partial derivatives, notice that Q defined by (1) can also be obtained by solving $\Phi(q_u) - \Phi(q_l) = p_0$ with respect to Q, where $q_l = (-Q - \mu)/\sigma$ and $q_u(Q - \mu)/\sigma$. Now using implicit differentiation, it is easy to verify that

$$\partial Q/\partial\mu = \big(\phi(q_u) - \phi(q_l)\big)/\big(\phi(q_u) + \phi(q_l)\big),$$

$$\partial Q/\partial\sigma^2 = \big(q_u\phi(q_u) - q_l\phi(q_l)\big)/\big\{2\,\sigma\big(\phi(q_u) + \phi(q_l)\big)\big\}.$$

Finally, substituting the above in (14) gives the expression for τ_Q^2 given in (12). The limiting distribution of $\ln(\hat{Q})$ now follows from an application of the delta method. □

Acknowledgement

We express our sincere thanks to the reviewers, the associate editor and Professor John Stufken their constructive comments that greatly improved this article.

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310.

Choudhary, P. K. (2005). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. Submitted.

Choudhary, P. K. and Nagaraja, H. N. (2003). Assessing agreement using coverage probability. *Technical Report 721.* Ohio State University, Columbus, OH.

Choudhary, P. K. and Ng, H. K. T. (2005). Assessment of agreement under non-standard conditions using regression models for mean and variance. *Biometrics* doi: 10.1111/j.1541-0420.2005.00422.x.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, New York.

Guttman, I. (1988). Statistical tolerance regions. In *Encyclopedia of Statistical Sciences*, Vol. 9, pp. 272-287, Kotz, S., Johnson, N. L. and Read, C. B. (Editors), John Wiley, New York.

Hamilton, D. C. and Lesperance, M. L. (1995). A comparison of methods for univariate and multivariate acceptance sampling by variables. *Technometrics* **37**, 329–339.

Hamilton, D. C. and Lesperance, M. L. (1997). A comparison of estimators of the proportion nonconforming in univariate and multivariate normal samples. *Journal of Statistical Computations and Simulations* **59**, 333–348.

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* **21**, 1913–1935.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2, 2nd edn. John Wiley, New York.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268. Corrections: 2000, 56:324-325.

Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* **19**, 255–270.

Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues and tools. *Journal of American Statistical Association* **97**, 257–270.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Tekten, T., Onbasili, A. O., Ceyhan, C., Unal, S. and Discigil, B. (2003). Novel approach to measure myocardial performance index: Pulsed-wave tissue Doppler echocardiopraphy. *Echocardiography* **20**, 503–510.

Wang, W. and Hwang, J. T. G. (2001). A nearly unbiased test for individual bioequivalence problems using probability criteria. *Journal of Statistical Planning and Inference* **99**, 41–58.

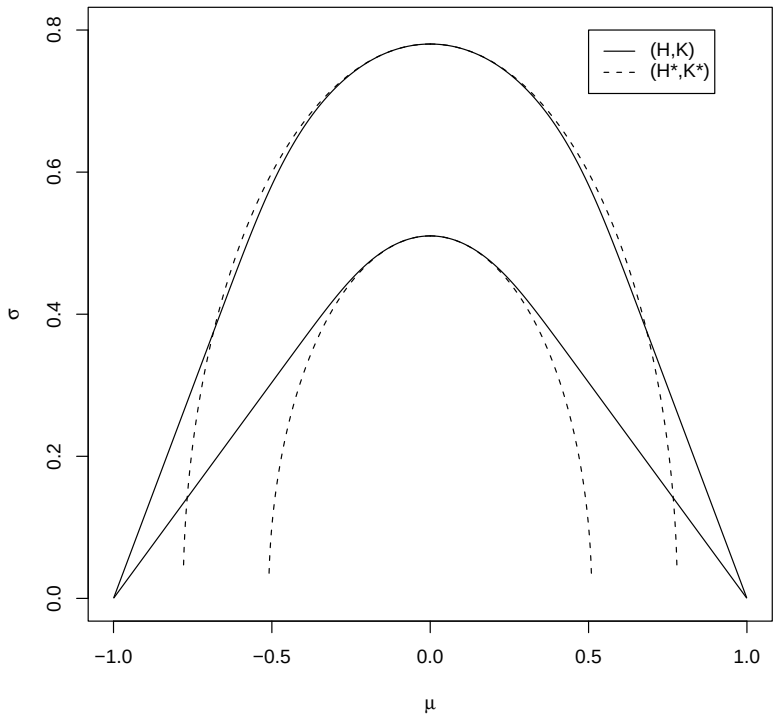

Figure 1: The regions under the hypotheses (H, K) of our interest, given by (4) or (5), and those under (H^*, K^*) of Lin (2000), given by (7). The top two curves represent the boundaries of the two null regions for $p_0 = 0.80$, and the bottom two for $p_0 = 0.95$. The alternative regions lie under the respective curves. Here we have taken $\delta_0 = 1.0$, so that the x and the y axes actually represent μ/δ_0 and σ/δ_0, respectively.

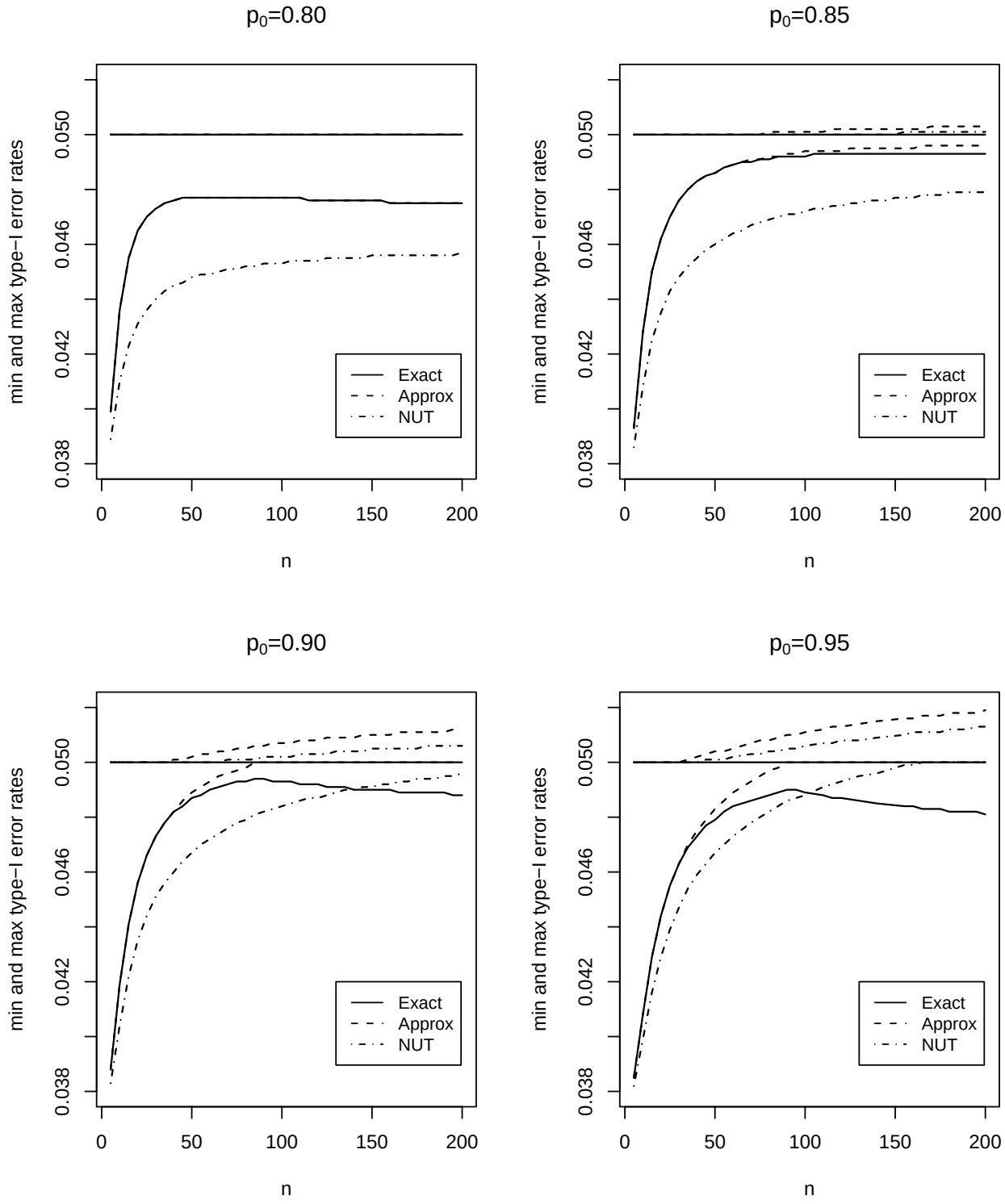

Figure 2: The min and max type-I error rates of the exact test, its MNUT approximation and the NUT on the boundary Θ_B as functions of n. Here the tests have the same nominal level $\alpha = 5\%$. These rates are computed numerically. In case of $p_0 = 0.80$, the three max curves agree, and the min curves of the exact test and its approximation coincide.

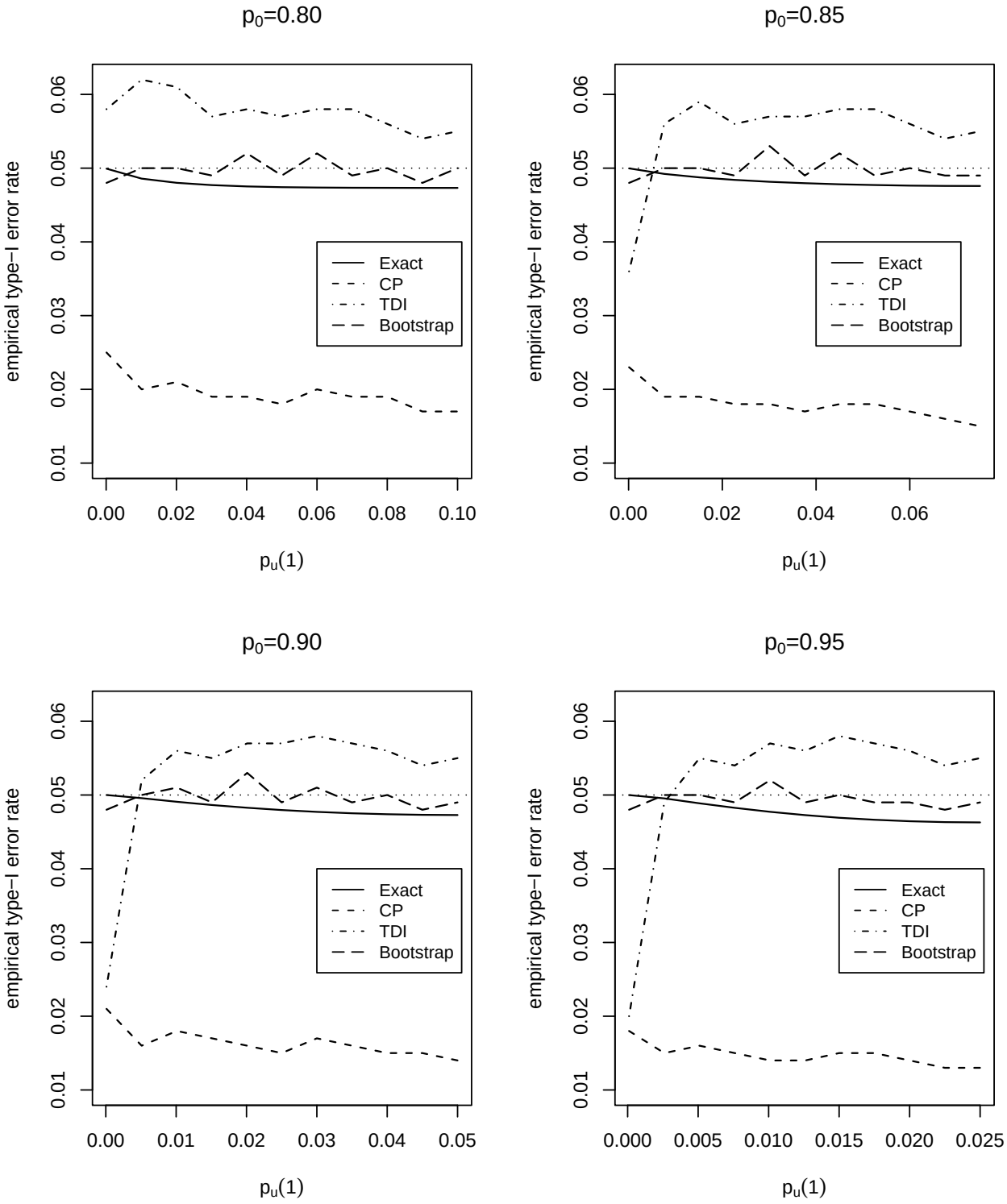

Figure 3: The type-I error rates of the exact, Bootstrap, CP and TDI tests on the boundary Θ_B. Here $(n, \alpha) = (30, 0.05)$ and each $0 < p_u(1.0) < (1 - p_0)/2$ produces a $(\mu, \sigma) \in \Theta_B$ upon using (13). The horizontal dotted lines in the plots mark the point 0.05.

	$F(\delta_0) = p_1$			
p_0	0.85	0.90	0.95	0.99
0.80	242 (240)	55 (53)	21 (19)	9 (8)
0.85		181 (177)	36 (34)	12 (11)
0.90			106 (102)	19 (17)
0.95				46 (43)

Table 1: Exact (approximate) values of smallest sample sizes that give at least 80% power for the 5% level test of (H, K) at $F(\delta_0) = p_1$. Without loss of generality $\delta_0 = 1.0$ is taken for these computations.

	$\hat{Q}_+(0.95)$	$\hat{F}_-(0.10)$	$\hat{F}_-(0.14)$
Exact	0.1305	0.8694	0.9642
MNUT	0.1305	0.8694	0.9642
NUT	0.1309	0.8672	0.9637
Bootstrap	0.1270	0.8769	0.9673
TDI	0.1255		
CP		0.7950	0.9108

Table 2: The 95% confidence bounds for $Q(0.95)$, $F(0.10)$ and $F(0.14)$ obtained by inverting various tests of (H, K) for the MPI data.