Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 4

11-1-2005

A Single, Powerful, Nonparametric Statistic for Continuous-data Telecommunications Parity Testing

J. D. Opdyke
DataMineIt, JDOpdyke@DataMineIt.com

Journal of Modern Applied Statistical Methods
November 2005, Vol. 4, No. 2, 372-393

1538 – 9472/05/$95.00

A Single, Powerful, Nonparametric Statistic for Continuous-data Telecommunications Parity Testing

J.D. Opdyke
DataMineIt
Marblehead, MA

Since the enactment of the Telecommunications Act of 1996, extensive expert testimony has justified use of the modified *t* statistic (Brownie et al., 1990) for performing two-sample hypothesis tests comparing Bell companies' CLEC and ILEC performance measurement data (known as parity testing). However, Opdyke (Telecommunications Policy, 2004) demonstrated this statistic to be potentially manipulable and to have literally zero power to detect inferior CLEC service provision under a wide range of relevant data conditions. This article develops a single, nonparametric statistic that is easily implemented (i.e., not computationally intensive) and typically provides dramatic power gains over the modified *t* while simultaneously providing much better Type I error control. The statistic should be useful in a wide range of quality control settings.

Key words: Telecommunications Act, ILEC, CLEC, Location-scale, Mean-variance, Maximum test

Introduction

The major goal of the Telecommunications Act of 1996, the most sweeping communications-related public policy to be enacted by Congress in over half a century (since the Telecom Act of 1934 – see http://www.fcc.gov/telecom.html) has been to deregulate local telephone service in the United States, making it a fully competitive economic market. To accomplish this, the Act takes a carrot-stick approach: it allows the Bell companies (the incumbent local exchange carriers, or ILECs, now only BellSouth, Qwest, SBC, and Verizon) to enter into the previously deregulated long distance market, something they had been prohibited from doing because of their status as government regulated monopolies. This provides ILECs with the potentially lucrative opportunity to provide one-stop shopping telephone service to their customers, bundling all of their clients' telecommunications needs into a single package from a single service provider.

In return for this carrot, the Act's stick requires that the ILECs first must do two things: (a) allow their competitors (competitive local exchange carriers, or CLECs, the large long distance telephone companies like Sprint, as well as numerous smaller companies) access to and use of their networks, in some cases to resell services at discounted wholesale rates, and (b) provide the CLECs' customers with service "at least equal in quality to" the service they provide to their own customers (Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996), at §251 (c) (2) (C); and see §251 (c) (2) (B) for the 14 point "COMPETITIVE CHECKLIST" of conditions that ILECs must satisfy to meet the at-least-equal

J.D. Opdyke is President of DataMineIt, a statistical data mining consultancy specializing in the banking and credit, telecommunications, retail and pricing, and advertising and marketing sectors (JDOpdyke@DataMineIt.com, www.DataMineIt.com). I owe special thanks to Geri S. Costanza, M.S., for numerous and valuable insightful discussions.

service provision standard). This at-least-equal service provision is the necessary enforcement mechanism for ensuring that network access (a) occurs in a meaningful way that truly promotes the goal of market competition.

To explain by way of example, if it takes a week on average for a CLEC customer to have a line installed or repaired by the ILEC, but only a day on average for an ILEC customer to receive the same service, no customers would ever switch from the ILEC to any of the CLECs, and markets could never become competitive. The mechanism for properly enforcing the at-least-equal service provision depends on the appropriate utilization of the extensive operations support services (OSS) performance measurement data that ILECs record when providing service to both CLEC and ILEC customers (e.g., how fast is a phone line installed; how fast is a line repaired; how often are repairs made within a certain number of days or by a preset due date, etc.). This utilization has taken the form of monthly statistical parity testing – applying statistical tests to the monthly CLEC and ILEC service data to compare the two groups and make sure that service is, in fact, at least equal for CLEC customers (i.e., in parity).

The specific statistical tests used in OSS parity testing depend on a number of factors, and foremost among these are the hypotheses being tested. The appropriate null and alternate hypotheses for OSS parity testing are listed below (1), in terms of both average service (the mean) and the variability of the service provided (the variance) (see Opdyke, 2004, p. 3-4, for a detailed explanation of why precisely these hypotheses are required in this setting).

$$\begin{gathered} \text{Ho: } \mu_C \le \mu_I \text{ AND } \sigma_C^2 \le \sigma_I^2 \\ \text{vs.} \\ \text{Ha: } \mu_C > \mu_I \text{ OR } \sigma_C^2 > \sigma_I^2 \end{gathered} \tag{1}$$

A statistical test of this pair of joint hypotheses will determine, with a specified level of certainty, whether service to CLEC customers takes no longer *on average* than service to ILEC customers (i.e., $\mu_C \le \mu_I$), and whether the variability of this service is no larger than that characterizing the service provided to ILEC customers (i.e., $\sigma_C^2 \le \sigma_I^2$) (see the FCC's Notice of Proposed Rulemaking, 04/16/98, APPENDIX B, p.B2, for some of the early impetus for testing both means and variances). If the statistical test determines, with a specified level of certainty, that both of these conditions hold, service is deemed to be at least equal, or in parity. If either condition is determined, with a specified level of certainty, to be violated, then service is considered out of parity, or in disparity.

Findings of disparity carry consequences for the ILEC(s) in the form of fines paid to the CLECs, and sometimes to the relevant state(s). These fines, or remedies, can be large (US$ millions), and extensive and/or prolonged findings of disparity can lead to revocation of an ILEC's approval to provide long distance service. Therefore the choice of appropriate, if not the best statistics for OSS parity testing is very important, not only for the individual firms involved, but also for the entire industry. And of course, the best statistics simply are those that, under a classical Neyman-Pearson hypothesis-testing paradigm, are most powerful under the widest range of relevant data conditions, given robust and reasonable Type I error control.

In addition to the hypotheses being tested, the type of data being compared determines which statistical tests can and should be used. Telecommunications OSS performance metrics contain three types of data, and each is listed below with an example of a corresponding performance metric:

- *binary data* – the percentage of repairs completed on time, or within a certain number of days
- *count data* – the number of troubles on a telephone line within a specified time period
- *continuous data* – the average time it takes to install a phone line

For continuous data metrics, the modified *t* (Brownie et al., 1990) has been supported in extensive expert testimony proffered by both CLECs and ILECs, as well as in Opinions and Rulings by various regulatory bodies, as an appropriate statistic to test the relevant joint hypotheses above (see Opdyke, 2004, for extensive citations; all but one of the four major ILEC performance and remedy plans nationwide utilizes the modified *t* as a primary test statistic).

$$t_{\text{mod}} = \frac{\left(\overline{X}_C - \overline{X}_I\right) - \left(\mu_C - \mu_I\right)}{s_I \sqrt{\frac{1}{n_I} + \frac{1}{n_C}}} \tag{2}$$

where

$$s_I = \sqrt{\frac{\sum_{i=1}^{n_I}\left(X_{I_i} - \overline{X}_I\right)^2}{\left(n_I - 1\right)}}, \quad \overline{X}_I = \frac{\sum_{i=1}^{n_I} X_i}{n_I}, \quad \overline{X}_C = \frac{\sum_{i=1}^{n_C} X_i}{n_C},$$

and degrees of freedom (*df*) = $n_I - 1$.

However, Opdyke (2004) demonstrated, via an extensive simulation study and an analytic derivation, that because the modified *t* follows neither the standard normal nor the student's *t* distribution as previously surmised in seven years' of expert testimony (see Opdyke, 2004, for extensive citations), it *potentially* remains vulnerable to what has been termed gaming – intentional manipulation of its score to effectively mask disparity. But far more importantly, the modified *t* also was shown to be virtually powerless to detect inferior CLEC service provision under a wide range of relevant data conditions (i.e., larger CLEC variability under equal or better average service).

Instead, Opdyke (2004) proposed the collective use of several other easily-implemented statistical procedures that typically provide dramatic power gains over the modified *t*. Selection of a specific statistic among those proposed depends on the relative sizes of the two samples being compared, and on whether the particular performance metric being tested is long-tailed or short-tailed (this is the distributional characteristic known as kurtosis). Years of OSS data now exist since the Act was passed to establish such distributional characteristics as population parameters, not as unknowns requiring an additional statistical test. However, even though the FCC itself identified "data distribution, sample size and other characteristics inherent in the data" (FCC NPRM, 11/08/01, p. 37) as factors relevant to the choice of the statistical tests used in parity testing, one expressed concern regarding Opdyke's (2004) approach is that the potential use of different statistics for different performance metrics (and sample sizes) is somehow too complex for implementation in parity testing.

This article addresses this concern by building on the results and recommendations of Opdyke (2004) to develop a single, nonparametric, and generally powerful statistic for use with all continuous–data performance metrics. As shown below, the proposed statistic 1) maintains reasonable Type I error control; 2) is always either nearly as powerful as Opdyke's (2004) multiple procedures, or almost as often, even more powerful; 3) typically provides dramatic power gains over the modified *t*; 4) is easily implemented and not computationally intensive; and 5) should be widely applicable and useful in other quality control settings as well.

Methodology

Previously Developed Alternatives to the modified *t*

Under the data conditions relevant to OSS parity testing, Opdyke (2004) found that conditional statistical procedures combining either O'Brien's (1988) generalized *t* test (OBt) or his generalized rank sum test (OBG) with either of two straightforward tests of variances (Shoemaker's, 2003, F_1 test, or the modified Levene test of Brown and Forsythe, 1974) were by far the most powerful procedures of the over twenty statistics that were studied. Their combined use is conditioned on the relative sizes of the two sample means, as shown below:

Table 1. Conditional Statistical Procedures, Opdyke (2004)

Conditional statistical procedure	if $\overline{X}_C > \overline{X}_I$, use…	If $\overline{X}_C \le \overline{X}_I$ or OB fails to reject Ho:, use…
OBtShoe	OBt	Shoemaker's F_1
OBtLev	OBt	modified Levene
OBGShoe	OBG	Shoemaker's F_1
OBGLev	OBG	modified Levene

(*Note*: see Appendix for the calculation of these statistics)

Conditioning on the sample means as shown in Table 1 inflates the size of these tests, so an ad hoc p-value adjustment of p-value = (5/3 * p-value) was used to maintain Type I error control (see Opdyke, 2004, for details). Even after such an adjustment, these tests maintain reasonable, if not impressive power under normal and short-tailed (uniform) data, and somewhat less power under long-tailed (double exponential) data, although still far more power than the modified *t* under most of these conditions (Opdyke, 2004, p. 20-26).

The conditions under which each of these four tests is most powerful and should be used are summarized in Table 2 below. Notably skewed data, however, first should be transformed, as required by one of the largest state PUCs and strongly endorsed by another of the largest state PUCs (CPUC Interim Opinion, 2001, Appendix J; CPUC Opinion (2002), Appendix J, Exhibit 3 p.2-3; Before the Texas PUC – SBC Testimony, Dysart & Jarosz, 2004; and for optional use with some metrics, SBC Comments, 2002, p.48, 56).

Unfortunately, all of the statistics examined for or used in OSS parity testing suffer from sometimes severe erosions in power under skewness (see Opdyke, 2004, for relevant simulation results; The California Public Utilities Commission also addresses this issue – CPUC Interim Opinion, 2001, p. 112-115, 136, 142, 145, & Appendix J, and CPUC Opinion, 2002, p. 74, 84, & Appendix J). Because these metrics are widely cited as being lognormal (which is typically highly skewed – see CPUC Interim Opinion, 2001, Appendix J, and MCI Worldcom's Performance Assurance Plan: The SiMPL Plan, by George S. Ford, Ph.D., p.5), a logarithmic transformation toward symmetry should provide at least some needed power to detect disparity without, in all practicality, causing distortions in the comparison of CLEC and ILEC service provision.

Table 2. Conditional Statistical Procedures, Opdyke (2004)

Sample sizes		Normal & Short-tailed	Long-tailed	Skewed
		OBt	OBG	
Bal.	Shoe	OBtShoe	OBGShoe	Transform
Unbal.	Lev	OBtLev	OBGLev	Transform

Once transformed (if necessary), the performance metric is tested with one of the four combined procedures listed in Table 2. This is clear-cut if the sample sizes and distributional characteristics of the metrics being tested unambiguously fall neatly into these four cells (for example, if a metric is at least as short-tailed as the normal distribution, kurtosis = 3, and has very unbalanced sample sizes, use OBtLev).

However, further simulations that parallel those of Opdyke (2004) are required to determine the tipping points defining exactly when to use each of these four statistics. Although these tipping point simulations would be straightforward to perform, one expressed concern about the use of Table 2 is that, the FCC's advisory comment notwithstanding, having to (potentially) use different tests under different sample size and data conditions is somehow too complex for the implementation of parity testing. Although implementing Table 2 is far less complicated than at least one of the four major OSS performance and remedy plans (the BellSouth 'truncated Z' plan, which one FCC economist only half-jokingly refers to as "the balanced averaged disaggregated truncated adjusted modified Z plan", Shiman, 2002, p.283), it unarguably would be preferable if, all else equal (or close), one statistic could accomplish what the conditional use of the multiple statistics in Table 2 does. This is the motivation for this paper, and the development of the statistic presented below.

A Single Statistic for Continuous-data Parity Testing

Maximum tests – statistics whose scores (p-values) are the maximum (minimum) of two or more other statistics – have been devised and studied in a number of settings in the statistics literature with very favorable results. Neuhäuser et al. (2004) favorably compares a maximum test for the non-parametric two-sample location problem to multiple adaptive tests, finding the former to be most powerful under the widest range of data conditions.

Blair (2002) constructed a maximum test of location that is shown to be only slightly less powerful than each of its constituent tests under their respective ideal data conditions, but notably more powerful than each under their respective non-ideal data conditions (for additional studies using maximum tests, see Fleming & Harrington, 1991, Freidlin & Gastwirth, 2000a, 2000b, Freidlin et al., 2002, Lee, 1996, Ryan et al., 1999, Tarone, 1981, Weichert & Hothorn, 2002, Willan, 1988, & Yang et al., 2005). These findings demonstrate the general purpose of maximum tests – to trade-off minor power losses under ideal data conditions for a more robust statistic with larger power gains across a wider range of possible (and usually unknown) data distributions.

Although the relevant characteristic of the distributions of continuous-data OSS performance metrics is, for all intents and purposes, known because so many years of data now exist to establish the kurtosis as a population parameter and not a statistical estimate based on samples, a maximum test still could be useful here for several reasons: 1) using only one statistical test unarguably would be more straightforward to implement than (potentially) relying on the four statistics in Table 2 and choosing between them based on a matrix of sample sizes and performance metric kurtoses; 2) the expected power losses compared to Opdyke's (2004) individual tests may be small or negligible; and 3) under some conditions, depending on the constituent tests used, the maximum statistic may be even more powerful than those tests recommended in Opdyke (2004) and shown in Table 2.

To construct a maximum test here, it must be recognized that maximum tests are conditional statistical procedures, and the additional variance introduced by such conditioning will inflate the test's size over that of its constituent statistics (and if left unadjusted, probably over the nominal level of the test as shown in Blair, 2002). But the constituent statistics in Table 2 are already conditional statistical procedures. Consequently, the ad hoc p-value adjustment used below for the purpose of maintaining validity must be large enough to take this double conditioning into account (this actually is triple conditioning because O'Brien's tests themselves are conditional statistical procedures). The adjustment is simply a multiplication of the p-values by constant factors (β's). The p-value of the maximum test – OBMax – is defined in (2):

$$p_{OBMax} = \min\begin{pmatrix} p_{OBtShoe} \cdot \beta_{OBtShoe} \ , \\ p_{OBtLev} \cdot \beta_{OBtLev} \ , \\ p_{OBGShoe} \cdot \beta_{OBGShoe} \ , \\ p_{OBGLev} \cdot \beta_{OBGLev} \ , \\ p_{tsv} \cdot \beta_{tsv} \ , \\ 1.0 \end{pmatrix} \qquad (3)$$

where

$$\beta_{OBtShoe} = \beta_{OBtLev} = \beta_{OBGShoe} = \beta_{OBGLev} = 2.8,$$

and $\beta_{tsv} = 1.8$, and p_{tsv} is the p-value corresponding to the separate-variance t test with Satterthwaite's (1946) degrees of freedom (see Appendix for corresponding formulae). Under the relevant data conditions, the behavior of OBMax is compared to that of its constituent tests and the modified t test in the simulation study described below. It is also compared with two other maximum tests – OBMax3 and TVMax – as defined in (2) and (3) below (TVMax for t test, Variance tests, and Maximum test).

$$p_{OBMax3} = \min\begin{pmatrix} p_{OBtLev} \cdot \beta_{OBtLev} \ , \\ p_{OBtShoe} \cdot \beta_{OBtShoe} \ , \\ p_{tsv} \cdot \beta_{tsv} \ , \\ 1.0 \end{pmatrix} \qquad (4)$$

where $\beta_{OBtLev} = \beta_{OBtShoe} = 3.0$, and $\beta_{tsv} = 1.6$

$$p_{TVMax} = \min\begin{pmatrix} p_{modLev} \cdot \beta_{modLev} \ , \\ p_{ShoeF_1} \cdot \beta_{ShoeF_1} \ , \\ p_{tsv} \cdot \beta_{tsv} \ , \\ 1.0 \end{pmatrix} \qquad (5)$$

where $\beta_{modLev} = \beta_{ShoeF1} = 3.0$, and $\beta_{tsv} = 1.6$

Although preferable to ad hoc adjustments based on simulations, analytic derivation of the asymptotic distribution of OBMax, and maximum tests in general, is non-trivial, as Yang et al. (2005) show under even stronger distributional assumptions than can be made with respect to the Table 1 statistics. Derivation of the asymptotic distribution of OBMax is the topic of continuing research (Opdyke, 2005).

Level and Power Simulation Study

The level and power simulations in this article parallel those conducted in Opdyke (2004). Eleven tests were studied: each of the four conditional statistical procedures listed in Table 1 – OBtShoe, OBtLev, OBGShoe, and OBGLev; the separate-variance t test (with Satterthwaite's, 1946, degrees of freedom – df) (tsv); the modified t test (with $df = n_1 - 1$, as in Brownie et al., 1990, Comments of SBC, 2002, p.57, and CPUC Opinion, 2001, Appendix C, p. 2.) (tmod); OBMax as defined above in (1); OBMax3 and TVMax as defined above in (2) and (3), respectively; and two tests of stochastic dominance described below. All of the conditional statistics using O'Brien's (1988) tests are referenced to the *F* distribution, rather than Blair's (1991) critical values, even though doing so would normally violate the nominal level of the test under some conditions, because the p-value adjustment used here explicitly takes this size inflation into account (see Opdyke, 2004, 2005, for further details).

The data was generated from the normal, uniform, double exponential, and lognormal distributions for four different pairs of sample sizes ($n_C = n_I = 30$; $n_C = 30$ & $n_I = 300$; $n_C = 30$ & $n_I = 3000$; and $n_C = n_I = 300$), seven different variance ratios ($\sigma_C^2 / \sigma_I^2 =$ 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00), and seven different location shifts

$$\begin{pmatrix} \mu_C = \mu_I - 2\sigma_I,\ \mu_I - \sigma_I,\ \mu_I - 0.5\sigma_I,\ \mu_I,\ \mu_I + 0.5\sigma_I, \\ \mu_I + \sigma_I,\ \mu_I + 2\sigma_I \end{pmatrix},$$

making 784 scenarios. N = 20,000 simulations were run for each scenario, except for scenarios with $n_C = 30$ & $n_I = 3000$, which used N = 5,000.

The normal distribution was chosen as a universal basis for comparison; the uniform and double exponential distributions were chosen as examples of short-tailed and long-tailed distributions, respectively, to examine the possible effects of kurtosis on the tests; and the lognormal distribution was chosen to examine the possible effects of skewness on the tests, and because continuous data OSS performance metrics have been cited widely as often being approximately lognormal. $n_C = n_I = 30$ was chosen because many performance and remedy plans require or allow for the use of permutation tests if at least one of the two samples has less than 30 observations (see The Qwest Performance Assurance Plan, Revised 11/22/2000, p.4-5; SBC Comments, 2002, p. 55, and 13 state Performance Remedy Plans – Attachment 17, p.4-5; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.3-4.), and $n_C = n_I = 300$ was chosen to examine rates of convergence under equal sample sizes (Pesarin's, 2000, combined permutation test, however, appears to have greater power for the relevant joint hypotheses here than the naïve Monte Carlo permutation test currently implemented by these performance and remedy plans, and at least two companies produce preprogrammed software that automatically performs this test – DataMineIt, http://www.DataMineIt.com, and Methodologica, http://www.methodologica.it/npctest.html).

The extremely unbalanced sample size pairs of $n_C = 30$ & $n_I = 300$ and $n_C = 30$ & $n_I = 3000$ were chosen because such large sample size ratios actually are not uncommon in OSS performance metric data. Also, the number of ILEC phone lines and customers typically dwarf those corresponding to most individual CLECs. Thus, it is important to test the behavior of these statistics under these extreme conditions, even though most simulation studies would focus on smaller and/or more balanced sample sizes. n_C is very rarely, if ever, larger than n_I and thus, only cases involving (n_I / n_C) ≥ 1.0 were examined in this study (Opdyke, 2005, examines $n_I < n_C$ also). Two nominal levels were used for all the simulations: $\alpha = 0.05$ and $\alpha = 0.10$, bringing the total number of scenarios to 1,568. These two levels bracket the vast majority of the levels used in OSS parity testing. (SBC Comments, 2002, p.49-52; CPUC Opinion, 2002, Appendix J, Exhibit 3, p.4; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.1).

Two other tests also were included in the simulations: Rosenbaum's (1954) test, which counts the number of observations in one sample beyond the maximum of the other as a test of H_o: $F(x) \equiv G(x)$ against the alternative of stochastic dominance; and the (one-sided) Kolmogorov-Smirnov statistic (using Goodman's, 1954, Chi-square approximation – see Siegel & Castellan, 1988, p.148), for a non-parametric test of H_o: $F(x) \equiv G(x)$ against general (one-sided) alternatives. Although neither is designed specifically to test the joint hypotheses relevant to the OSS parity testing setting, and thus may have less power, they are included for several reasons: (1) as a basis for comparison to the other tests; (2) because researchers often turn to these types of tests when confronted with the joint hypotheses relevant to the parity testing context and examined in this simulation study; and (3) because the Kolmogorov-Smirnov statistic has been described as being "able to detect not only differences in average but differences in dispersion between the two samples as well." (Matlack, 1980, p. 359).

Results

This simulation study generated 11 x 1,568 = 17,248 level and power results, all of which are available from the author upon request in a Microsoft Excel® workbook (along with a SAS/GRAPH® program for convenient visualization). The key results are summarized in the tables and selected graphs below.

Under symmetry, the p-value adjustments used in OBMax as defined in (3) provide reasonable Type I error control for the relevant range of test levels; as shown in Table 3, violations of the nominal level are modest in size and infrequent (14 of 288 symmetric-data null hypothesis scenarios; violations occur if the observed level is equal to or greater than the one-tailed 95% critical value of the simulation, based on the common Wald approximation of the binomial distribution to the normal distribution, which is very accurate for such large numbers of simulations and $\alpha \geq 0.05$ – see Evans et al., 1993, p. 39, and Cochran, 1977, p. 58).

Even better level control is possible by increasing the adjustment factors – say, by increasing the OB β's from 2.8 to 3.0 – but the price paid for this is a loss of power. The adjustment factors used – 2.8 for the OB tests and 1.8 for the separate-variance *t* test – are reasonable as they produce relatively minor level violations, and relatively minor power losses when OBMax is compared to its constituent tests. However, nearly as often as not, OBMax actually provides power *gains* over the conditional use of the Table 2 statistics (graphs of these comparisons are available from the author upon request). OBMax's largest power loss is only slightly over 0.10, and these minor power losses typically occur under simultaneously small CLEC samples, large CLEC variance increases, and *decreases* in the CLEC mean (relative to the ILEC mean).

Its largest power gain, however, exceeds 0.2, and these power gains occur under simultaneously small CLEC samples, typically equal or smaller CLEC variances, and small *increases* in the CLEC mean. The reason for this increased sensitivity to detect small location shifts is the inclusion of the separate-variance *t* test among the constituent tests of OBMax. Including this test mitigates power losses in the one fairly narrow range of conditions where the modified *t* test has a relatively slight, but still noticeable power advantage over the Table 2 constituent tests: for normal and short-tailed data, under simultaneously small CLEC samples, typically equal or smaller CLEC variances, and small increases in the CLEC mean. Including the separate-variance *t* test as a constituent test of OBMax shrinks this loss of power relative to the modified *t* (under only these fairly narrow conditions) typically by a factor of one half, so that the largest power loss remains less than 0.1 (Figure 3).

Far more important to note, however, is that under all other data conditions the power of OBMax is never less than that of the modified *t*, and typically dramatically larger (sometimes a gain of 1.0! - see Figures 3, 4, and 6). The power differences between OBMax and the modified *t* that are shown in Figure 3 are summarized in Table 4 below, although the Figures more accurately and thoroughly convey the story. Figures 5 and 6 show how dramatically OBMax dominates the modified *t* as sample sizes increase. This demonstration of the reasonable power of OBMax, under all symmetric alternatives, should dispel a) expressed concerns in this setting regarding the lack of power of composite tests of location and scale (Mallows, 2002, p. 260); b) admittedly premature conclusions in this setting about the lack of power of relevant rank-based tests (Mallows, 2002, p. 260), which is what the OBG tests are; and c) findings of less (and concerns of too little) power in this setting under unbalanced sample sizes (Gastwirth & Miao, 2002, p. 273).

Table 3. Symmetric Data Level Violations of OBMax

σ_C^2	μ_C	Sample sizes	Distribution	Nominal level of test (α)	Actual size
σ_I^2	μ_I	$n_C = n_I = 30$	Normal	0.05	0.0578
σ_I^2	μ_I	$n_C = 30$, $n_I = 3000$	Normal	0.05	0.0532
σ_I^2	μ_I	$n_C = n_I = 300$	Normal	0.05	0.0561
σ_I^2	μ_I	$n_C = 300$, $n_I = 300$	Uniform	0.05	0.0546
σ_I^2	μ_I	$n_C = n_I = 30$	Double exponential	0.05	0.0574
σ_I^2	μ_I	$n_C = 30$, $n_I = 300$	Double exponential	0.05	0.0538
σ_I^2	μ_I	$n_C = 30$, $n_I = 3000$	Double exponential	0.05	0.0556
σ_I^2	μ_I	$n_C = n_I = 300$	Double exponential	0.05	0.0596
σ_I^2	μ_I	$n_C = n_I = 30$	Normal	0.10	0.1115
σ_I^2	μ_I	$n_C = n_I = 300$	Normal	0.10	0.1073
σ_I^2	μ_I	$n_C = n_I = 30$	Uniform	0.10	0.1048
σ_I^2	μ_I	$n_C = n_I = 300$	Uniform	0.10	0.1044
σ_I^2	μ_I	$n_C = n_I = 30$	Double exponential	0.10	0.1116
σ_I^2	μ_I	$n_C = n_I = 300$	Double exponential	0.10	0.1095

Not surprisingly, OBMax is very similar to OBMax3 and TVMax in terms of both Type I error control and power, except that, under small CLEC and large ILEC samples, OBMax has greater power than TVMax to detect slight CLEC location shifts, especially under leptokurtotic data (the largest power advantages are about 0.08, 0.10, and 0.14 for uniform, normal, and double exponential data, respectively). OBMax3 is more powerful than TVMax, exhibiting the same slight power loss compared to OBMax only under leptokurtotic data (where the largest loss is only about 0.08). Because OBMax is unambiguously more powerful, it is recommended over the other two tests under symmetry. Under asymmetry, however, OBMax violates the nominal level of the test under a specific combination of conditions, for which the OBG rank tests perform poorly (a. large and equal sample sizes; b. equal means; *and* c. a much smaller CLEC variance). Therefore if skewed data is not or cannot be reliably transformed toward symmetry for some reason, OBMax3 is one good alternative to OBMax. OBMax3 has slightly less power, but it always maintains validity, even under skewed data. In fact, it maintains validity far better than does the modified *t* under skewed data.

However, an even better alternative appears to be OBMax2, as presented in the preliminary results of Opdyke (2005). OBMax2 = OBMax3 if a) $s_C^2 \le s_I^2$, b) $\bar{X}_c \le \left(\bar{X}_I + 0.5 s_I\right)$, *and* c) the null hypothesis of symmetry is rejected by the test of D’Agostino et al. (1990) at α = 0.01; otherwise, OBMax2 = OBMax. OBMax2 maintains most of the power gains of OBMax over OBMax3, while also maintaining validity very well under skewed data – again, far better than does the modified *t*, as shown in Table 5 below (note that when $n_C > n_I$, which rarely if ever occurs with OSS data, all *β*’s for OBMax2 utilize an additional adjustment: $\beta_X = \beta_X + \min\left[2.5,\ \log_{2.7}\left(n_C/n_I\right)\right]$ – see Opdyke, 2005, for further details).

Figure 1. OBMax rejection rate: Empirical Level and Power (α = 0.05)

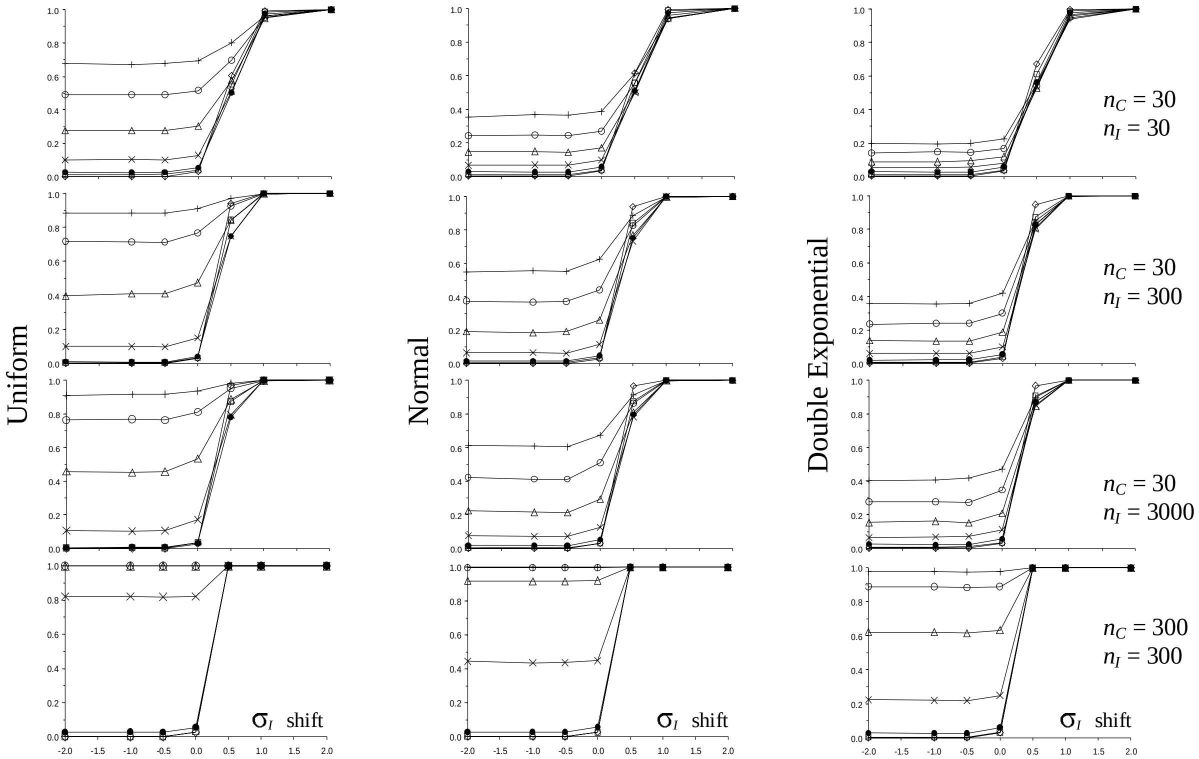

Figure 2. modified *t* rejection rate: Empirical Level and Power (α = 0.05)

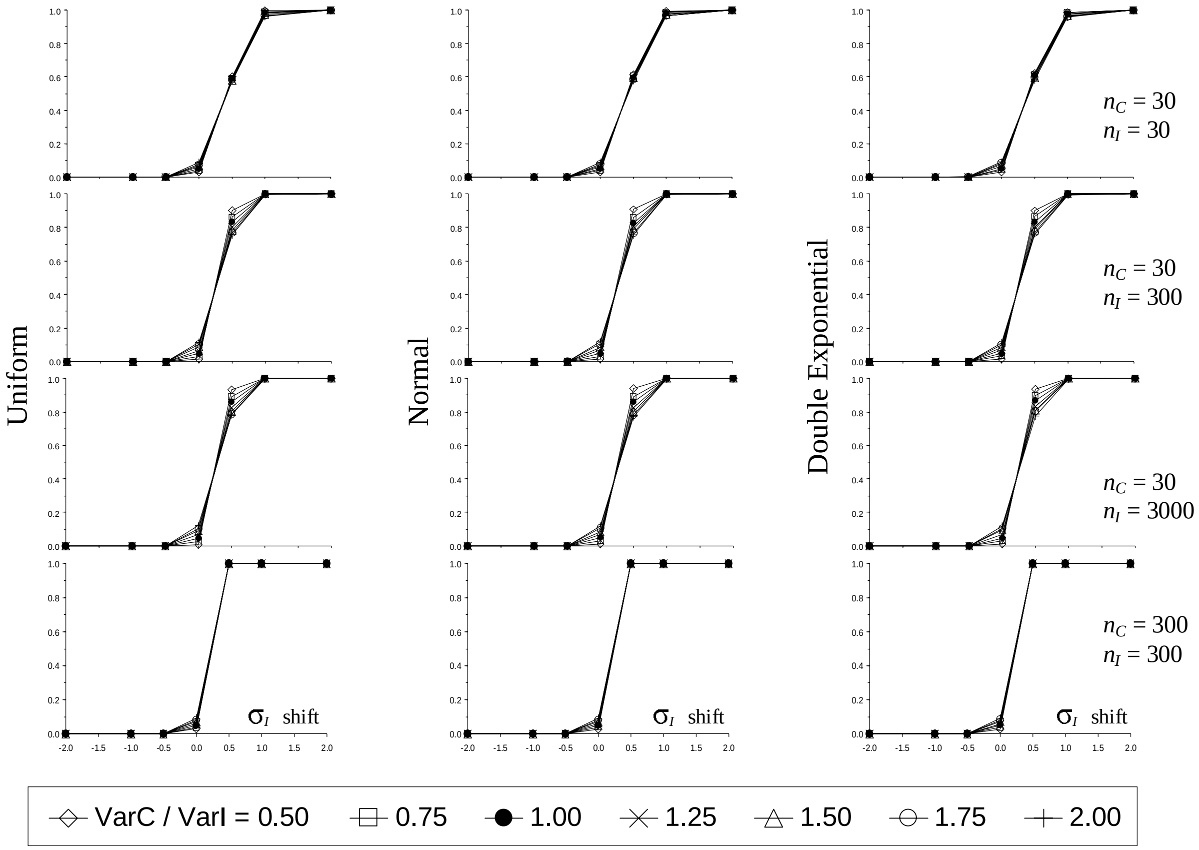

Figure 3. OBMax Power minus modified *t* Power (α = 0.05)

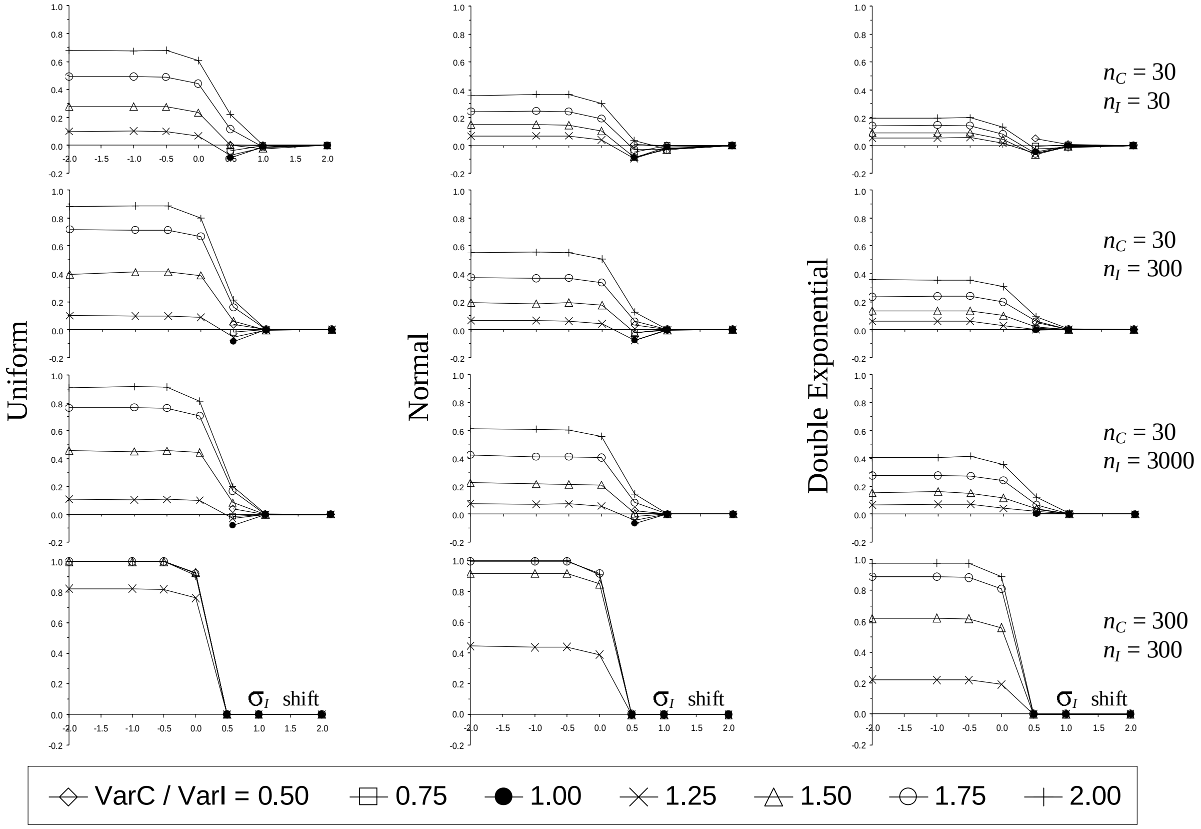

Figure 4. All Alternate Hypothesis Simulations with a Power Difference (309 of 444):
OBMax Power minus modified *t* Power (α = 0.05)

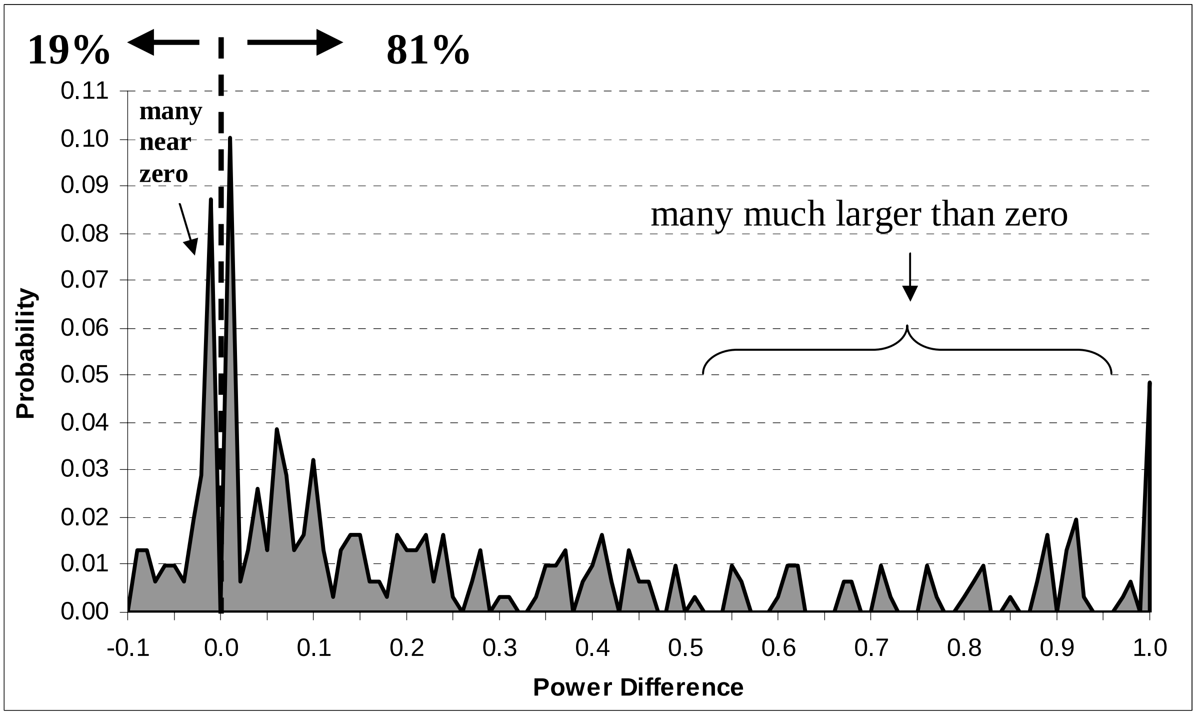

Figure 5. Alternate Hypothesis Simulations of $n_C = n_I = 30$ with a Power Difference (90 of 111): OBMax Power minus modified t Power (α=0.05)

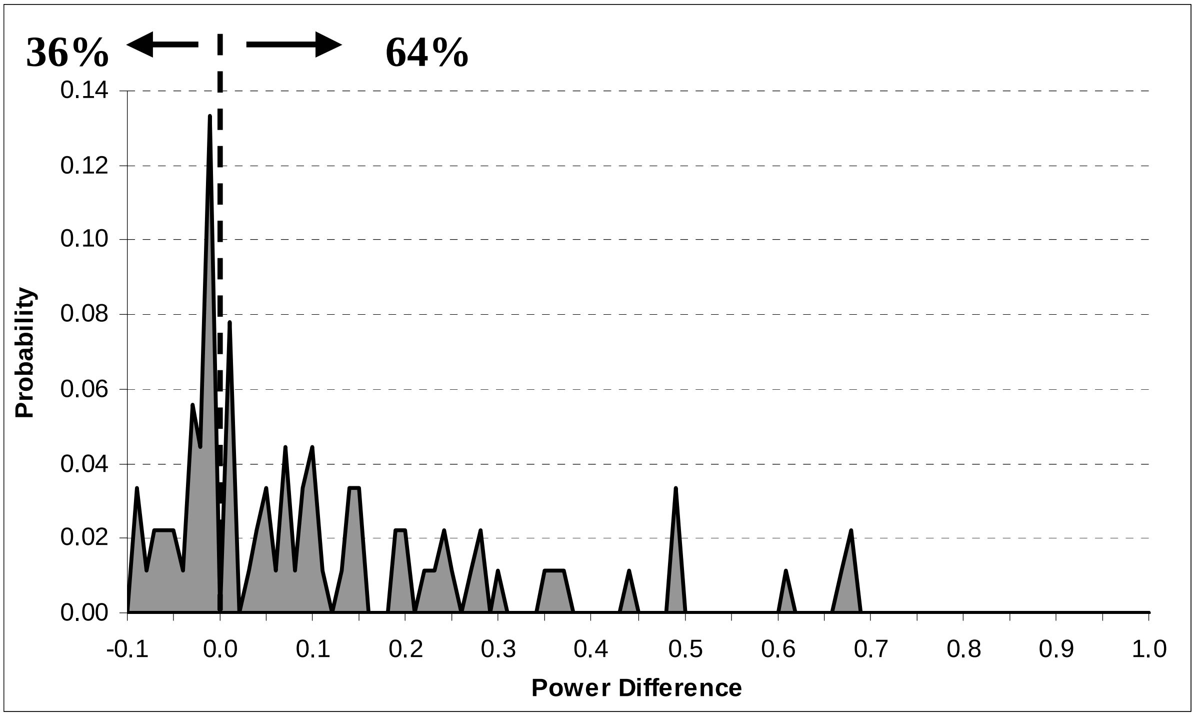

Figure 6. Alternate Hypothesis Simulations of $n_C = n_I = 300$ with a Power Difference (52 of 111): OBMax Power minus modified t Power (α=0.05)

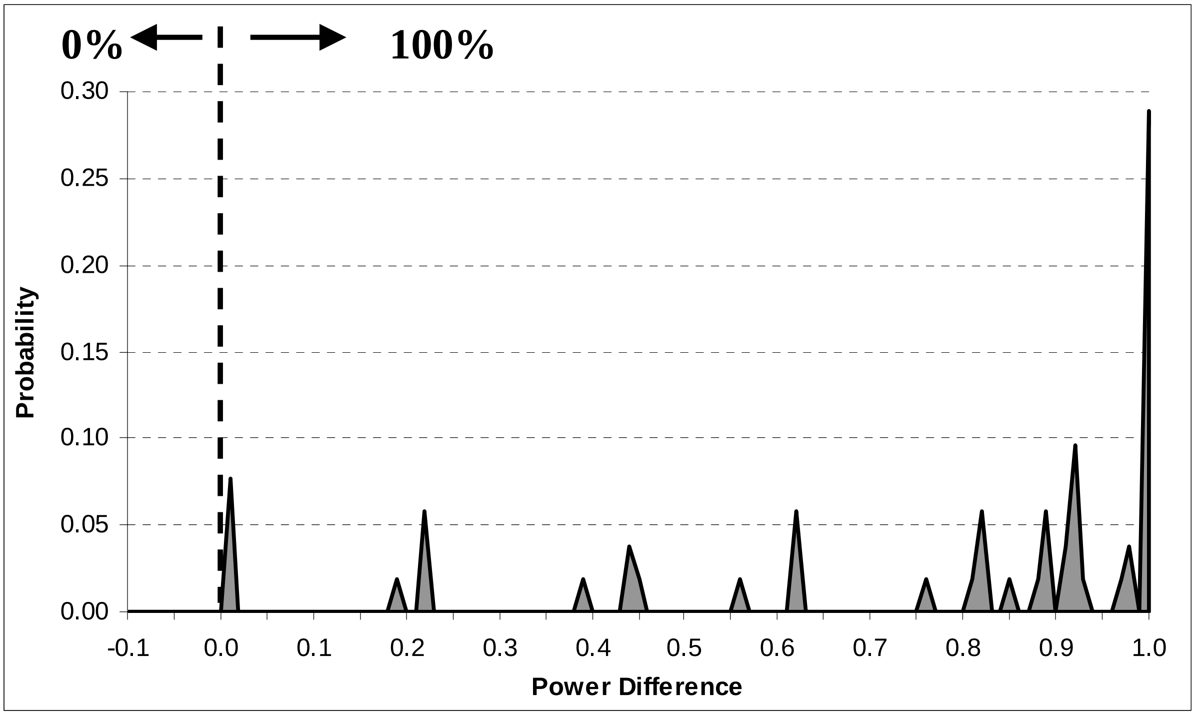

Table 4. modified *t* vs. OBMax: Dominant Test, and Corresponding Power Gains Under Symmetry (α = 0.05) by Magnitude of Mean Difference and Variance Difference

σ^2 / μ	$\mu_C > \mu_I$ (small difference)		$\mu_C > \mu_I$ (large difference)	$\mu_C \le \mu_I$
	Small n_C (= 30)	Large n_C		

$\sigma_C^2 > \sigma_I^2$	**Usually OBMax**
	Max = 0.223
	Mean = 0.038
	Median = 0.028
		EQUAL	**EQUAL**	**Always OBMax**
				Max = 1.000
				Mean = 0.431
				Median = 0.361

$\sigma_C^2 \le \sigma_I^2$	**Usually tmod**
	Max = 0.051
	Mean = 0.015
	Median = 0.006
		EQUAL	**EQUAL**	**Ho:**

OBMax vs. the modified *t*: Where does it matter in terms of remedies?

As shown in Figures 3-6 above, OBMax often provides dramatic power gains over the modified *t*, making it much more effective at identifying disparity when it truly exists. A very important point to note here is that the narrow conditions under which the modified *t* has a slight power advantage – small sample sizes and small location shifts (and a typically smaller or equal CLEC variance) – are exactly those that are the least important in terms of the size of the resulting remedies. Under most performance and remedy plans, the formulae for calculating remedies are proportionate functions of the number of lines or customers affected, as well as the magnitude of the degree to which service is out of parity (i.e., how much worse CLEC service is relative to ILEC service). Small sample sizes, and small deviations from parity, together imply the smallest remedies. Small power losses under these conditions (always less than 0.1 under symmetry, and no more than 0.2 under asymmetry when using OBMax2) will result in missed remedies that should be quite small, and perhaps even negligible, relative to overall remedies.

In contrast, under all other conditions of disparity, where both sample sizes and deviations from parity are much larger, the typically dramatic power gains of OBMax over the modified *t* will translate into much larger remedies that the modified *t* will fail to identify. The relative (if not absolute) size of these remedies missed by the modified *t* will dwarf any missed by OBMax when both sample sizes and location shifts are small. Thus, not only are the power gains of OBMax over the modified *t* much larger and more common than the losses, but also much more important in terms of the magnitude of the remedies that should be identified by the statistical test used. Consequently, from both a statistical and remedy-impact perspective, OBMax is dramatically better than the modified *t* at identifying disparate service provision to CLEC customers, and thus, is far more effectively used in parity testing to enforce the at-least-equal service provision of the Act. This makes OBMax is a better tool for achieving the Act's major objective: moving local telephone service from regulation to full competition and, once achieved, preventing backsliding to disparity into the future.

In other quality control settings, too, OBMax should be useful and widely applicable as discussed below, but the questions of how, and how much, the use of OBMax matters in OSS parity testing are examined next.

Table 5. Worst Level Violations of modified *t* vs OBMax2 Under Asymmetry (Opdyke, 2005)

Statistic	σ_C^2	μ_C	n_C	n_I	Distribution	α	Actual Size	Violation
OBMax2	σ_I^2	$\mu_I - \sigma_I$	300	30	Exponential	0.05	0.0553	0.0053
OBMax2	σ_I^2	$\mu_I - 2\sigma_I$	300	30	Exponential	0.05	0.0566	0.0066
OBMax2	σ_I^2	μ_I	300	30	Exponential	0.05	0.0665	0.0165
OBMax2	$0.75\sigma_I^2$	μ_I	300	30	Lognormal	0.05	0.0581	0.0081
OBMax2	σ_I^2	μ_I	300	30	Lognormal	0.05	0.0623	0.0123
OBMax2	σ_I^2	μ_I	300	30	Exponential	0.10	0.1053	0.0053
OBMax2	σ_I^2	μ_I	300	30	Lognormal	0.10	0.1073	0.0073
modt	σ_I^2	μ_I	30	30	Lognormal	0.05	0.0992	0.0492
modt	σ_I^2	μ_I	300	30	Exponential	0.05	0.1003	0.0503
modt	$0.50\sigma_I^2$	μ_I	300	30	Lognormal	0.05	0.1034	0.0534
modt	σ_I^2	μ_I	300	30	Lognormal	0.05	0.1082	0.0582
modt	$0.75\sigma_I^2$	μ_I	300	30	Lognormal	0.05	0.1089	0.0589
modt	σ_I^2	μ_I	30	30	Lognormal	0.10	0.1451	0.0451
modt	σ_I^2	μ_I	300	30	Exponential	0.10	0.1477	0.0477
modt	$0.50\sigma_I^2$	μ_I	300	30	Lognormal	0.10	0.1544	0.0544
modt	$0.75\sigma_I^2$	μ_I	300	30	Lognormal	0.10	0.1630	0.0630
modt	σ_I^2	μ_I	300	30	Lognormal	0.10	0.1649	0.0649

OBMax vs. the modified *t*: How Does It Matter, and How to Decide?

The Act was designed so that, with respect to enforcing the central requirement of at-least-equal service provision to CLEC customers, everything hinges on the performance metric data, and the inferences made about it based on statistical tests. The consequences of OSS parity testing results that indicate disparity undeniably can be large, in terms of both remedies paid by ILECs to CLECs and, in the case of backsliding or prolonged and extensive disparity, the possible revocation of an ILEC's long-distance approval (which carries even larger, long-term financial consequences for both ILECs and CLECs).

Although not all performance metrics have statistical tests applied to them (a minority are comparisons of CLEC service against a fixed benchmark), and continuous data metrics are only a subset of all those subject to statistical parity testing, they still include some of the biggest metrics – i.e., those containing the most data reflecting the largest numbers of customers and phone lines (e.g., average time-to-install). Therefore, a statistic used to test these metrics that fails to identify actual disparity under a wide range of conditions not only distorts the simple and crucial incentive structure clearly and explicitly intended by the Act, but also misses sizeable remedies that would have been identified by a more powerful statistic – in this case, OBMax (or OBMax2).

Therefore, given the results of this study comparing OBMax to the modified *t*, one might ask when using actual OSS data, what is the magnitude of this distortion caused by the modified *t*? How much does it matter in terms of remedies, which is the bottom line in this setting? Although it is possible to approximately answer this question empirically, and the answer could very well be a sizeable amount, it is actually the wrong question to ask here for several reasons. First, it can never be known absolutely whether service provision to CLEC customers is truly inferior because only monthly samples are being considered, not entire populations. It could be, due to random variation, that CLEC service is not really inferior, but that the given samples make it appear so (in statistical parlance, this is a Type I error). The reverse also can occur (a Type II error). What statistical tests provides is a scientific basis for making an inference, based on the samples that merely represent the true underlying service levels, with a specified degree of certainty (for example, if $\alpha = 0.05$, one can be $[1 - \alpha] = 95\%$ certain that an inference of parity is correct).

This guess or hypothesis about whether service is or is not in parity is the best that can be done, so a researcher can never evaluate the statistical properties of competing tests based (solely) on real data samples. The researcher must know the true answer in the data ahead of time, which is only possible with simulated data (as used in this study), and then see which statistic gets it right most often under the widest range of relevant data conditions. Then it will be known that, if applied to actual data samples that are based on truly disparate service levels, a statistic that is proven to be more powerful under well-constructed simulations will be more powerful under actual data and correctly detect the disparity more often.

That said, a general idea may be obtained as to how much remedies will be affected when using OBMax vs. the modified *t* by applying each to, say, six months of actual data and comparing the resulting remedies (such a comparison obviously would need to be based on identical remedy formulae, with distance-beyond-parity directly or indirectly based on p-values and α; if Z-scores are familiar or in current use, then the inverse standard normal function can be used, e.g., $\Phi(\text{p-value}) - \Phi(\alpha)$ = distance beyond parity). If there are much larger remedies resulting from the use of OBMax, then it will be known that its greater power is driving this result.

However, even if no appreciable difference in remedies is observed (which would be surprising), the question ‘How much are remedies actually affected?’ is not the key question that needs to be answered because it ignores the important issue of a deterrent effect. If no appreciable difference in remedies is observed, that just means that scenarios under which OBMax is more powerful are not exhibited in the data being examined. But there is no telling that these types of inferior service scenarios will not crop up in the future (or have not cropped up at different times in the past). Because the modified *t* will definitely miss them if they do crop up, why would it ever be used over the more powerful statistic, OBMax? The answer is, it should not, and under a scientifically responsible implementation of applied statistics, it would not.

Thus, in evaluating which statistic to use for OSS parity testing and considering the remedy-impact of using OBMax instead of the modified *t*, the driving question is not, How much will actual remedies differ under OBMax vs. the modified *t*? (although the answer to this probably is noticeably, if not a great deal.); instead, the relevant question is, Under conditions that we know to be disparate, which statistic has greater power to correctly identify the disparity? This question cannot be answered by using actual data and comparing the remedies resulting from the use of each of these two statistics (although this comparison may be interesting), but rather, by the simulation study conducted in this paper. And the answer this study provides is that OBMax does have more power under a wider range of relevant data conditions, and these power gains are often dramatic. The general applicability of OBMax in other settings is discussed briefly below.

General Utility of OBMax (OBMax2)

OBMax, and OBMax2, are useful in any context where one-sided tests of the first two moments are the primary or exclusive concern, and the researcher needs to test for effects in *either or both* moments (in other words, when the researcher needs to test (1) above). For these joint hypotheses, just as shown in Opdyke (2004) for OBMax’s constituent tests, OBMax outperforms a test of stochastic dominance and a widely-used nonparametric distributional test against general alternatives. The Rosenbaum (1954) statistic maintains validity, but generally has much less power than OBMax, especially if the CLEC mean is smaller than the ILEC mean, when it often has absolutely no power to detect a larger CLEC variance (which is consistent with its design). The latter finding also holds for the one-sided Kolmogorov-Smirnov statistic which, although occasionally more powerful than OBMax, often severely violates the nominal level when means are identical but the CLEC variance is *smaller* (which is consistent with *its* design, if not the relevant joint hypotheses examined here). Thus, OBMax is far superior to statistical tests that many researchers commonly turn to, at least initially, when faced with testing the joint hypotheses of (1) above. Among the settings in which these hypotheses are central is, of course, OSS parity testing; possibly the network access rules aimed at similar telecom deregulation efforts in other countries (Ure, 2003, p. 42-43); possibly the open access energy transmission regulations established by the Federal Energy Regulatory Commission (Gastwirth & Miao, 2002, p. 278); and numerous industrial settings with the need to address the quality control issues of accuracy and/or precision in manufacturing and other processes (Opdyke, 2005). Some important issues warranting further inquiry are listed below.

Further Research

Most of the points below are listed in Opdyke (2004) and remain important issues for further inquiry in this setting.

- In regulatory telecommunications, almost always $n_{\mathrm{CLEC}} << n_{\mathrm{ILEC}}$, so scenarios of $n_{\mathrm{CLEC}} > n_{\mathrm{ILEC}}$ were not studied in this paper. However, they are addressed in the further development of OBMax2 in Opdyke (2005).

- Although typically much more powerful than the modified *t*, even under skewed data, OBMax2 still has low power under asymmetry, and exploring ways to increase it is worthy of further study (Opdyke, 2005).

- Although the nominal test levels examined in this study ($\alpha = 0.05$ and $\alpha = 0.10$) bracket the vast majority of the test levels used in telecommunications OSS parity testing, (SBC Comments, 2002, p.49-52; CPUC Opinion, 2002, Appendix J, Exhibit 3, p.4; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.1) other settings may require very different nominal levels (e.g., $\alpha = 0.20$ or $\alpha = 0.01$). Generalizing from the findings of this study to such conditions would not be advisable without further simulation.

- The one major exception to the above point regarding nominal test levels is the BellSouth performance and remedy plan. As previously mentioned, instead of solely using the modified *t* for continuous data performance metrics, this plan relies primarily on a statistic dubbed the truncated *Z* for which a balancing critical value is used as the nominal level of the hypothesis test. This critical value purports to balance or equalize the probability of Type I and Type II error (i.e., incorrect inferences of disparity and parity, respectively). This statistic, however, may remain insensitive to, i.e., have little power to detect, larger CLEC variance for two reasons: first, the formula used to determine the balancing critical value is admittedly essentially unaffected by differences in variances (BellSouth Comments, 2002, Attachment 2 (Part 4), Exhibit No. EJM-1, Appendix C, p.C-9); second, the statistical test scores that are truncated and combined to obtain the truncated *Z* score are simply scores of modified *t* tests adjusted for skewness (BellSouth Comments, 2002, Attachment 2 (Part 3), Exhibit No. EJM-1, Appendix A, p.A-5, with correction from Attachment 2 (Part 2), Appendix D – Technical Description, p. 37). It is not at all clear that a combined statistic based on such truncated *t*-scores has much or any power to detect differences in variances, and a thorough simulation study like the one completed in this paper would be useful to allay or confirm these suspicions.

- Although not the focus of this study, some performance and remedy plans use the general form of the modified *t* statistic as the basis for modifications to statistical tests designed for binary data, like that based on the common Wald approximation to the normal distribution (Comments of SBC, 2002, p. 59). In light of Opdyke's (2004) findings, and all of the problems inherent in using the modified *t* statistic with continuous data performance metrics, such modifications should be viewed with skepticism until subjected to careful analytic scrutiny and empirical simulation. No objections to using the modified *t* for continuous data OSS parity testing were raised. Mulrow (2002) raised no objection to using the modified *t* for continuous data OSS parity testing, although concern was expressed about making modified *t*–like changes to the Wald approximation test for binary data: "This does not seem right to me" (p.280). Instead of this test, Mulrow (2002) advocated the use of Fisher's exact test. It is a viable and easily implemented alternative already in wide usage in OSS parity testing, although sometimes only for small(er) samples (SBC Performance Remedy Plan – Attachment 17, p. 3). Yet, it can be used for large samples as well because, even as a conditional exact test, it can be implemented very quickly with modern statistical software packages (e.g., SAS®). Agresti and Caffo (2000) provided a simple and effective, although not exact test for both small and large samples, and even better (more powerful), if slightly more complex alternatives, are the unconditional exact tests of Berger and Boos (1994) (available at http://www4.stat.ncsu.edu/~berger/tables.html) and Skipka et al. (2004) (Berger, 1996; Kopit & Berger, 1998). These all are carefully studied and well designed tests for binary data: there is no need to turn to unverified methods of questionable utility in this setting.

- Although not the focus of this study, some performance and remedy plans rely on a normal approximation *Z*-test for comparing CLEC and ILEC sample rates from count data performance metrics, even when those rates are very small (e.g., trouble report rate) and almost certainly highly non-normal (SBC Performance Remedy Plan – Attachment 17, p.3-4; Ameritech Michigan – Performance Remedy Plan – Attachment A, p. 2; and SBC Performance Remedy Plan – Version 3.0 SBC/SNET FCC 20 Business Rules – Attachment A-3, p.A-88). Yet, powerful and easily-implemented tests for comparing two Poisson means have been developed, and may be far superior statistically for such comparisons (Krishnamoorthy & Thomson, 2004). Examination of these metrics' distributions, and a straightforward simulation study, would adequately address this question.

Unheeded Warnings

As mentioned in Opdyke (2004), it is important to note that not everyone has supported the use of the modified *t* in this (and other) settings, although dissension has been conspicuously rare in the OSS parity testing arena. O'Brien (1993), in his discussion of Blair & Sawilowsky's (1993) empirical study unfavorably comparing the modified *t* to O'Brien's (1988) OBt and OBG statistics, points out that the Type I error rates of the modified *t* statistic will severely violate the nominal level of the test under a variety of conditions. Within the parity testing arena, over five years ago GTE voiced a lone, cautionary, and seemingly prescient dissent, given the findings of this current study, regarding use of the modified *t* in OSS parity testing:

> The modified *Z*-test [*t* test] should not be used since it follows no standard formulation of the test statistic. In the absence of a rigorous derivation, its sampling properties and maintained hypotheses are unknown. It has been asserted that the modified *Z*-test [*t* test] is a joint test of the equality of the means and variances of the two distributions; however no rigorous derivation has been provided. … It would clearly be foolish to accept a new and unknown test statistic without further documentation and consideration. (COMMENTS OF GTE, Before the Michigan Public Service Comm., 11/20/98, Attachment B, p.15-16)

(Opdyke, 2004, has since provided an analytic derivation of the asymptotic distribution of the modified *t*: as stated previously, it is *not* standard normal or student's *t* distributed, although it has been described as such in the expert testimony of Dysart & Jarosz, 2004 which, on pages 27-29, egregiously misquotes the derivation and major findings of Opdyke, 2004.)

Meanwhile, others have hedged their bets. While being deposed as an expert witness for AT&T and other CLECs, Dr. Gerald Ford was asked:

> DO YOU BELIEVE THE MODIFIED Z-TEST SHOULD BE REPLACED WITH THESE PROPOSED ALTERNATIVES?
>
> No. The development of the particulars of the performance plan took many months of hard work by some very smart people. It was only after considerable analysis and debate that the Modified Z-test [modified *t* test] was selected as the best test statistic for the performance plan. …I see no reason to alter the test procedures of the existing plan without strong

> evidence that the other tests represent an improvement.
>
> SO YOU BELIEVE THE MODIFIED Z-TEST [modified *t* test] SHOULD BE USED?
>
> Yes, at least until some strong evidence is provided to indicate an alternative test is preferred. (Before the Texas PUC, Rebuttal Testimony of Dr. Gerald Ford for the CLEC Coalition, 08/23/04, p.36)

The goal of this article, with its development of a single, nonparametric, yet generally powerful statistic for continuous-data OSS parity testing, has been to provide the "further documentation and consideration" implicitly requested by GTE (1998), as well as the "strong evidence" of "an improvement" over the modified *t* that Ford (2004) implicitly requested much more recently.

Conclusion

As summarized in Opdyke (2004), under the Telecommunications Act of 1996, ILECs are required to provide CLEC customers with local telephone service "at least equal in quality to" that which they provide to their own customers if they are to be allowed into the long distance telephone market (Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996), at §251 (c) (2) (C)). The goal of this carrot-stick approach – the carrot being the potentially lucrative long distance market, and the stick being this requirement of at-least-equal service provision – is to promote competition in the newly deregulated local telephone markets. Implementing and enforcing the at-least-equal service provision requirement has taken the form of OSS parity testing – statistically testing the service data represented in thousands of operations support services performance metrics to ensure that the service provided to CLEC customers is, in fact, at least equal.

Results from these statistical tests indicating average service and/or service variability that is *not* at least equal, i.e., findings of disparity, typically require an ILEC to pay fines (sometimes US$ millions) to the CLEC(s), and sometimes to the state(s); disparity that is consistent and widespread over time (i.e., backsliding) can serve as cause for the revocation of an ILEC's approval to provide long distance service. These stakes are high, not only for individual firms but also for the entire industry, so choosing the correct, if not the best statistics to use in OSS parity testing is a very important decision.

To date, the modified *t* statistic (Brownie et al., 1990) has been approved and used in OSS parity testing across the country. It is used on continuous-data performance metrics as a test of whether average service and/or service variability are at least equal for CLEC customers compared to their ILEC counterparts. However, Opdyke (2004) demonstrated that the modified *t* is an ineffective and misleading choice for this purpose in this setting. It remains *potentially* vulnerable to gaming – intentional manipulation of its score to mask disparity – but far more importantly, it remains absolutely powerless to detect inferior CLEC service provision under a wide range of relevant data conditions. Opdyke (2004) proposed the use of several other easily implemented conditional statistical procedures that are not vulnerable to gaming and typically provide dramatic power gains over the modified *t*. The selection of which among them to use, however, depends on the relative sizes of the two data samples and a distributional characteristic (the kurtosis) of the specific performance metric being tested. Although this is arguably straightforward, a single test that could accomplish the same thing would be preferable, and the development of such a statistic is the motivation for this article.

In this article, an easily-implemented maximum test – OBMax – was developed based on the multiple statistics proposed by Opdyke (2004). OBMax maintains reasonable Type I error control and is always either nearly as powerful as its constituent tests, or almost as often as not, even more powerful. More importantly, it typically provides dramatic power gains over the modified *t*. The one set of narrow conditions under which the modified *t* has a slight power advantage (always less than 0.1 under symmetry) are exactly those under which consequent fines or remedies imposed on ILECs will be the smallest – small CLEC sample sizes and small location shifts (and equal or close-to-equal variances).

In contrast, the typically dramatic power gains of OBMax over the modified *t* under most other conditions of disparity (sometimes gains of even 1.0!) translate into the appropriate identification of vastly larger amounts of remedies that the modified *t* will miss. From both a statistical and remedy-impact perspective, therefore, OBMax is superior at detecting disparity, and thus, at enforcing the at-least-equal service provision of the Telecommunications Act of 1996. It consequently is an unambiguously better statistic than the modified *t* for use in OSS parity testing to achieve the major objective of the Act: the movement of local telephone service from regulation to full market competition.

References

Agresti, A., & Caffo, B. (2000), Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician, Vol. 54,* No. 4, 280-288.

Before the Federal Communications Commission, Comments of BellSouth, CC Docket No. 01-318, ATTACHMENT 2, January 22, 2002. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6512980049 through 6512980052]

Before the Federal Communications Commission, In the Matter of Performance Measurements and Reporting Requirements of Operations Support Systems, Interconnection, and Operator Services and Directory Assistance, CC Docket No. 98-56 RM-9101, Motion to Accept Late-Files Documents of U S WEST Communications, Inc., APPENDIX A: Comments of Michael Carnall on Statistical Issues of Detecting Differences in Service Quality, On Behalf of U S WEST Communcations, June 1, 1998. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=2078390002 and 2078390003]

Before the Federal Communications Commission, 07/09/03 filing on behalf of SBC Communications, Inc.: Performance Remedy Plan – SBC – Version 3.0 SBC/SNET FCC 20 Business Rules – Attachment A-3: Calculation of Parity and Benchmark Performance and Voluntary Payments. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6514285147]

Before the Federal Communications Commission, in the Matter of Performance Measurements and Standards for Unbundled Network Elements and Interconnection; Performance Measurements and Reporting Requirements for Operations Support Systems, Interconnection, and Operator Services and Directory Assistance; Deployment of Wireline Services Offerings Advanced Telecommunications Capability; Petition of Association for Local Telecommunications Services for Declaratory Ruling; CC Docket No. 01-318; CC Docket No. 98-56; CC Docket No. 98-147; and CC Docket Nos. 98-147, 96-98, and 98-141; Comments of SBC Communications, Inc., January 23, 2002. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6512980270 and 6512980271]

Before the Commonwealth of Massachusetts Department of Telecommunications and Energy, D.T.E. 99-271, MCI Worldcom’s Proposed Performance Assurance Plan for Bell Atlantic-Massachusetts, Appendix D – Simplified Measurement of Performance and Liability: The SiMPL Plan, by George S. Ford on Behalf of MCI Worldcom. [available at http://www.state.ma.us/dpu/telecom/99-271/pap/MCIW/appendix_D.pdf]

Before the Michigan Public Service Commission, Case No. U-11830, COMMENTS OF GTE, November 20, 1998, ATTACHMENT B: Statistical Method for OSS Performance Measures, GTE White Paper.

Before the Michigan Public Service Commission, Case No. U-13848, 09/04/03 filing on behalf of SBC Michigan: Ameritech Michigan – Performance Remedy Plan – Attachment A. [available at http://efile.mpsc.cis.state.mi.us/efile/docs/13848/0004.pdf]

Before the Public Utilities Commission of the State of California, Order Instituting Rulemaking on the Commission’s Own Motion into Monitoring Performance of Operations Support Systems, Decision 01-01-037, January 18, 2001. – Rulemaking 97-10-016 (Filed October 9, 1997), and Order Instituting Investigation on the Commission’s Own Motion into Monitoring Performance of Operations Support Systems. – Investigation 97-10-017 (Filed October 9, 1997): INTERIM OPINION ON PERFORMANCE INCENTIVES, Decision 01-01-037, January 18, 2001. [available at http://www.cpuc.ca.gov/word_pdf/FINAL_DECISION/11842.pdf]

Before the Public Utilities Commission of the State of California, Order Instituting Rulemaking on the Commission’s Own Motion into Monitoring Performance of Operations Support Systems, Decision 01-01-037, March 6, 2002. – Rulemaking 97-10-016 (Filed October 9, 1997), and Order Instituting Investigation on the Commission’s Own Motion into Monitoring Performance of Operations Support Systems. – Investigation 97-10-017 (Filed October 9, 1997): OPINION ON THE PERFORMANCE INCENTIVES PLAN FOR PACIFIC BELL TELEPHONE COMPANY, Decision 02-03-023, March 6, 2002, APPENDIX J. [available at http://www.cpuc.ca.gov/published/final_decision/ 13927.htm and http://www.cpuc.ca.gov/ PUBLISHED/FINAL_DECISION/13928.htm]

Before the Public Utilities Commission of Texas, Docket No. 28821, SBC Texas’ Joint Direct Testimony of William R. Dysart and Dorota Jarosz, July 19, 2004. – available at http://interchange.puc.state.tx.us/WebApp/Interchange/application/dbapps/billings/pgDailySearch.asp

Before the Public Utilities Commission of Texas, Docket No. 28821, Rebuttal Testimony of George S. Ford, Ph.D., on behalf of the CLEC Coalition, August 23, 2004. – available at http://interchange.puc.state.tx.us/WebApp/Interchange/application/dbapps/billings/pgDailySearch.asp

Berger, R.L. (1996), More powerful tests from confidence interval p values, *The American Statistician, Vol. 50*, 314-317.

Berger, R.L., & Boos, D.D. (1994), P Values maximized over a confidence set for the nuisance parameters, *Journal of the American Statistical Association, Vol. 89*, 1012-1016. [available at http://www4.stat.ncsu.edu/~berger/ tables.html]

Blair, R.C. (2002), Combining two nonparametric tests of location, *Journal of Modern Applied Statistical Methods*, Vol. 1, No. 1, 13-18.

Blair, R.C. (1991), New critical values for the generalized t and generalized rank-sum procedures, *Communications in Statistics, Vol. 20*, 981-994.

Blair, R.C. & Sawilowsky, S. (1993) Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls, *Statistics in Medicine, Vol. 12*, 2233-2243.

Brown, M. B. & Forsythe, A. B. (1974), Robust tests for the equality of variances, *Journal of the American Statistical Association, Vol. 69*, 364-367.

Brownie, C., Boos, D. D. & Hughes-Oliver, J. (1990), Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls, *Biometrics, Vol. 46*, 259-266.

Cochran, W. (1977), *Sampling techniques*, 3rd ed., New York: John Wiley & Sons.

D’Agostino, R.B., A. Belanger, and R.B. D’Agostino, Jr. (1990), A suggestion for using powerful and informative tests of normality, *The American Statistician, 44*: 316-321.

Evans, M., Hastings, N., & Peacock, B. (1993), *Statistical distributions*, 2nd ed., New York: John Wiley & Sons.

Federal Communications Commission, Notice of Proposed Rulemaking, CC Docket No. 98-56, RM-9101, FCC 98-72, APPENDIX B, Adopted April 16, 1998. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi? native_or_pdf=pdf&id_document=2060360001 to 0004]

Federal Communications Commission, Notice of Proposed Rulemaking, CC Docket No. 98-56, FCC 01-331, APPENDIX B, Adopted November 8, 2001. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi? native_or_pdf=pdf&id_document=6512974611 & 4612]

Fleming, T. R. & Harrington, D. P. (1991) *Counting processes and survival analysis*. Wiley, New York.

Freidlin, B. & Gastwirth, J. (2000a) Changepoint tests designed for the analysis of hiring data arising in employment discrimination cases, *Journal of Business and Economic Statistics, Vol. 18*, No. 3, 315-322.

Freidlin, B. & Gastwirth, J. (2000b) On power and efficiency robust linkage tests for affected sibs, *Annals of Human Genetics, 64*, 443-453.

Freidlin, B., Zheng, G., Zhaohai, L., & Gastwirth, J. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Human Heredity, Vol. 53*, No. 3, 146-152.

Gastwirth, J. L., & Miao, W. (2002), Comment, *Statistical Science*, *Vol. 17*, No. 3, 271-276.

Goodman, L.A. (1954), Kolmogorov-Smirnov tests for psychological research, *Psychological Bulletin*, *51*, 160-168.

Kopit, Justin S., and Berger, R.L. (1998), A more powerful exact test for a practical difference between binomial proportions, *American Statistical Association Proceedings, Biopharmaceutical Section*, 251-256.

Krishnamoorthy, K., & Thomson, J. (2004), A more powerful test for comparing two poisson means, *Journal of Statistical Planning and Inference*, *Vol. 119*, Issue 1, 23-35.

Lee, W. J., (1996) some versatile tests based on the simultaneous use of weighted log-rank statistics, *Biometrics*, *52*, 721-725.

Levene, H. (1960), Robust tests for equality of variances, in *Contribution to probability and statistics: essays in honor of harold hotelling*, I. Olkin et al., eds., Stanford University Press, Palo Alto, 278-292.

Local Competition User's Group ("LCUG" – Membership: AT&T, Sprint, MCI, LCI, WorldCom), Statistical tests for local service parity, February 6, 1998, Version 1.0.

Mallows, C. (2002), Parity: Implementing the Telecommunications Act of 1996, *Statistical Science*, *Vol. 17*, No. 3, 256-285.

Matlack, W.F., (1980), *Statistics for public policy and management*, Belmont, CA: Duxbury Press.

Mulrow, E. (2002), Comment, *Statistical Science*, *Vol. 17*, No. 3, 276-281.

Neuhäuser, M. Büning, H., & Hothorn, L. A. (2004), Maximum test versus adaptive tests for the two-sample location problem, *Journal of Applied Statistics*, *Vol. 31*, No. 2, 215-227.

O'Brien, P.C. (1988), Comparing two samples: extensions of the t, rank-sum, and log-rank tests, *Journal of the American Statistical Association*, *Vol. 83*, 52-61.

O'Brien, P.C. (1993), Discussion, *Statistics in Medicine*, *Vol. 12*, 2245-2246.

Opdyke, J.D. (2004), Misuse of the 'modified' *t* statistic in regulatory telecommunications, *Telecommunications Policy*, *Vol.28*, 821-866.

Opdyke, J.D. (2006), A nonparametric statistic for joint mean-variance quality control, *American Statistical Association Proceedings - 2005, Section on Quality and Productivity*, forthcoming.

Performance Assurance Plan – Bell-Atlantic, New York, (filed with New York Public Service Commission 04/07/2000). [available at http://www.fcc.gov/telecom.html]

Performance Remedy Plan – SBC, 13 states, Attachment 17. [available at http://www.nrri.ohio-state.edu/oss/Post271/Post271/Texas/performance%20agreement.pdf]

Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003. [available at http://www.dps.state.ny.us/ny2003 pap_redline.PDF and http://www.dps.state.ny.us/nyappndx_a_to_f_h2003pap.PDF]

Pesarin, F. (2001), *Multivariate permutation tests with applications in biostatistics*, John Wiley & Sons, Ltd., New York.

Rosenbaum, S. (1954), Tables for a Nonparametric Test of Location, *Annals of Mathematical Statistics*, *25*, 146-50.

Ryan, L.M., Freidlin, B., Podgor, M.J., & Gastwirth, J.L., (1999), Efficiency robust tests for survival or ordered categorical data, *Biometrics*, *55*, No. 3, 883-886.

Satterthwaite, F. W. (1946), An approximate distribution of estimates of variance components, *Biometrics Bulletin*, *2*, 110-114.

Shiman, D. (2002), Comment, *Statistical Science*, *Vol. 17*, No. 3, 281-284.

Siegel, S. & Castellan, N. John, (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., New York: McGraw-Hill.

Shoemaker, L. H. (2003), Fixing the F test for equal variances, *The American Statistician*, *Vol. 57*, No. 2, 105-114.

Skipka, G., Munk, A., & Freitag, G. (2004), Unconditional exact tests for the difference of binomial probabilities – contrasted and compared, *Computational Statistics and Data Analysis*, *Vol.47*, No.4, 757-774.

Tarone, R.E., (1981) On the distribution of the maximum of the log-rank statistic and the modified wilcoxon statistic, *Biometrics*, 37, 79-85.

Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996). [available at http://www.fcc.gov/telecom.html]

The Qwest Performance Assurance Plan, Revised 11/22/2000 [available at http://www.nrri.ohio-state.edu/oss/Post271 /Post271/Qwest_11-22-00_Red-lined_PAP.pdf]

Verizon Performance Assurance Plan, Redlined Version of Current PAP Showing Proposed 2003 Modifications, APPENDIX D. [available at http://www.dps.state.ny.us/ Case_99C0949.htm]

Ure, J. (2003), Competition in the local loop: unbundling or unbungling? *Info: The Journal of Policy, Regulation, and Strategy for Telecommunications, Information and Medi*a, Vol. 5, No. 5, 38-46.

Weichert, M. & Hothorn, L.A. (2002) Robust hybrid tests for the two-sample location problem, *Communications in Statistics – Simulation and Computation, Vol. 31,* 175-187.

Willan, A.R. (1988) Using the maximum test statistic in the two-period crossover clinical trial, *Biometrics, Vol. 44,* No. 1, 211-218.

Yang, Song, Li Hsu, & Lueping Zhao (2005), Combining asymptotically normal tests: case studies in comparison of two groups, *Journal of Statistical Planning and Inference, Vol. 133,* Issue 1, 139-158.

Zar, J.H., (1999), *Biostatistical analysis,* 4th ed., Upper Saddle River, NJ: Prentice-Hall.

Appendix

OBt and OBG: O'Brien's OBt test involves running the following ordinary least squares regression on pooled data including both samples:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \,, \qquad (6)$$

where *y* is a dummy variable indicating inclusion in the CLEC sample, and *x* is the performance metric variable. If the parameter on the quadratic term (β_2) is (positively) statistically significant at the 0.25 level, use the critical value of the overall equation to reject or fail to reject the null hypothesis; if it is not, use the critical value of the overall equation of the following ordinary least squares regression instead:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (7)$$

O'Brien's OBG test is identical to the OBt test except that the pooled-sample ranks of *x* are used in the regressions instead of the *x* data values themselves.

Modified Levene test: The modified Levene test requires a simple data transformation: take the absolute value of each data point's deviation from its respective sample median (as per Brown and Forsythe, 1974), and then calculate the usual one-way ANOVA statistic using these transformed values (as per Levene, 1960). The resulting statistic (8) is referenced to the *F* distribution as usual.

Let $z_{ij} = \left| x_{ij} - \tilde{x}_i \right|$ where $\tilde{x}_i$ is sample *i*'s median (8)

$$W_o = \frac{\sum_i n_i \left(\overline{z}_{i\cdot} - \overline{z}_{\cdot\cdot}\right)^2 \Big/ (g-1)}{\sum_i \sum_j \left(z_{ij} - \overline{z}_{i\cdot}\right)^2 \Big/ \sum_i (n_i - 1)} \sim F_{(g-1),\sum_i (n_i - 1)}$$

where $\overline{z}_i = \sum z_{ij} \big/ n_i$ and $\overline{z}_{\cdot\cdot} = \sum \sum z_{ij} \big/ n_i$

However, because this test is designed as a two-tailed test, and the hypotheses being tested in this setting are one-tailed, the p-value resulting from this test, when used conditionally with O'Brien's tests as in Table 1, must be subtracted from 1.0 if the CLEC sample variance is less than the ILEC sample variance. Or, if one does not need to calculate a p-value that is be known to be larger than α (as when the CLEC sample variance is smaller), the calculation simply can be skipped.

Shoemaker's F_1 test: Shoemaker's F_1 test is simply the usual ratio of sample variances referenced to the *F* distribution, but using different degrees of freedom:

$$s_C^2 \big/ s_I^2 \sim F_{df_C, df_I} \qquad (9)$$

where $$df_i = 2n_i \Big/ \left(\frac{\hat{\mu}_4}{\hat{\sigma}^4} - \frac{n_i - 1}{n_i - 3} \right)$$

where i = C, I corresponds to the two samples, and μ_4 and σ^4 are estimated from the two samples when pooled:

$$\hat{\mu}_4 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i \right)^4 \Big/ \left(n_1 + n_2 \right) \qquad (10)$$

$$\hat{\sigma}^4 = \left[\left((n_1-1)s_1^2+(n_2-1)s_2^2\right)\Big/\left(n_1+n_2\right)\right]^2 \quad (11)$$

Shoemaker (2003) notes that the biased estimate for σ^4 is used for improved accuracy.

Separate-variance t test: The separate-variance *t* test, also known as the Welch or Behrens-Fisher *t* test, is presented below:

$$t_{sv} = \frac{\left(\bar{X}_C - \bar{X}_I\right) - \left(\mu_C - \mu_I\right)}{\sqrt{\frac{s_I^2}{n_I} + \frac{s_C^2}{n_C}}} \quad (12)$$

where $s_I^2 = \frac{\sum_{i=1}^{n_I}\left(X_{I_i} - \bar{X}_I\right)^2}{(n_I - 1)}$, $s_C^2 = \frac{\sum_{i=1}^{n_C}\left(X_{C_i} - \bar{X}_C\right)^2}{(n_C - 1)}$,

$\bar{X}_I = \frac{\sum_{i=1}^{n_I} X_i}{n_I}$, and $\bar{X}_C = \frac{\sum_{i=1}^{n_C} X_i}{n_C}$

Satterwaith's (1946) degrees of freedom for t_{sv} is:

$$df = \frac{\left(\frac{s_I^2}{n_I} + \frac{s_C^2}{n_C}\right)^2}{\frac{\left(\frac{s_I^2}{n_I}\right)^2}{(n_I - 1)} + \frac{\left(\frac{s_C^2}{n_C}\right)^2}{(n_C - 1)}} \quad (13)$$

If *df* is not an integer, it should be rounded down to the next smallest integer (Zar, 1999, p. 129)

Test of D'Agostino et al. (1990): The test of D'Agostino et al. (1990) is calculated as follows:

$$g_1 = \frac{k_3}{s^3} = \frac{\frac{n\sum\left(X_i - \bar{X}\right)^3}{(n-1)(n-2)}}{\sqrt{\left(s^2\right)^3}}, \qquad \sqrt{b_1} = \frac{(n-2)\,g_1}{\sqrt{n(n-1)}}$$

$$A = \sqrt{b_1}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \quad (14)$$

$$B = \frac{3\left(n^2 + 27n - 70\right)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C = \sqrt{2(B-1)} - 1\ , \qquad D = \sqrt{C}\ , \qquad E = \frac{1}{\sqrt{\ln D}}$$

$$F = \frac{A}{\sqrt{\frac{2}{C-1}}}\ , \qquad Z_{g_1} = E\ln\left(F + \sqrt{F^2 + 1}\right) \sim \phi(0,1)$$

(~ standard normal)

For one-tailed testing of skewness to the left, check $\Pr\left(Z \le Z_{g_1}\right)$; for skewness to the right, check $\Pr\left(Z \ge Z_{g_1}\right)$. See Zar (1999), p. 115-116, for further details.