

ORIGINAL

RECEIVED-FPSC

Legal Department

NANCY B. WHITE
General Counsel-Florida

00 SEP 29 PM 4:41

BellSouth Telecommunications, Inc.
150 South Monroe Street
Room 400
Tallahassee, Florida 32301
(305) 347-5558

RECORDS AND
REPORTING

September 29, 2000

Mrs. Blanca S. Bayó
Director, Division of Records and Reporting
Florida Public Service Commission
2540 Shumard Oak Boulevard
Tallahassee, FL 32399-0850

Re: Docket No. 000121-TP (OSS)

Dear Ms. Bayó:

Enclosed is an original and 15 copies of BellSouth Telecommunications, Inc.'s Reply Comments, which we ask that you file in the captioned matter.

A copy of this letter is enclosed. Please mark it to indicate that the original was filed and return the copy to me. Copies have been served to the parties shown on the attached Certificate of Service.

Sincerely,

Nancy B. White
(NW)

Nancy B. White

Enclosures

APP _____
CAF _____
CMP 3 _____
COM _____
CTR _____
ECR 1 _____
LEG 1 _____
OPC _____
PAL _____
RGO Harvey _____
SEC _____
SER _____
OTH _____

cc: All parties of record
Marshall M. Criser, III
R. Douglas Lackey

RECEIVED & FILED
[Signature]
FPSC BUREAU OF RECORDS

DOCUMENT NUMBER-DATE

12405 SEP 29 8

FPSC-RECORDS/REPORTING

**CERTIFICATE OF SERVICE
Docket No. 000121-TP**

I HEREBY CERTIFY that a true and correct copy of the foregoing was served via

U.S. Mail this 29th day of September, 2000 to the following:

Timothy Vaccaro
Staff Counsel
Florida Public Service
Commission
Division of Legal Services
2540 Shumard Oak Boulevard
Tallahassee, FL 32399-0850

AT&T
Marsha Rule
101 North Monroe Street
Suite 700
Tallahassee, FL 32301-1549
Tel. No. (850) 425-6365
Fax. No. (850) 425-6361

GTE Florida, Inc.
Kimberly Caswell
P.O. Box 110, FLTC0007
Tampa, FL 33601-0110
Tel. No. (813) 483-2617
Fax. No. (813) 223-4888

Nanette Edwards
Regulatory Attorney
ITC^DeltaCom
4092 S. Memorial Parkway
Huntsville, Alabama 35802
Tel. No. (256) 382-3856
Fax. No. (256) 382-3936

Scott A. Sapperstein
Intermedia Communications, Inc.
3625 Queen Palm Drive
Tampa, Florida 33619
Tel. No. (813) 829-4093
Fax. No. (813) 349-9802

Charles J. Pellegrini
Wiggins & Villacorta, P.A.
2145 Delta Boulevard
Suite 200
Post Office Drawer 1657
Tallahassee, FL 32302
Tel. No. (850) 358-6007
Fax. No. (850) 358-6008
Counsel for Intermedia

Peter M. Dunbar, Esquire
Karen M. Camechis, Esquire
Pennington, Moore, Wilkinson,
Bell & Dunbar, P.A.
Post Office Box 10095 (32302)
215 South Monroe Street, 2nd Floor
Tallahassee, FL 32301

Mark Buechele
Legal Counsel
Supra Telecom
1311 Executive Center Drive
Suite 200
Tallahassee, FL 32301
Tel. No. (850) 402-0510
Fax. No. (850) 402-0522

Michael A. Gross
Vice President, Regulatory Affairs
& Regulatory Counsel
Florida Cable Telecomm. Assoc.
310 North Monroe Street
Tallahassee, FL 32301
Tel. No. (850) 681-1990
Fax. No. (850) 681-9676

Susan Masterton
Charles J. Rehwinkel
Sprint
Post Office Box 2214
MS: FLTLHO0107
Tallahassee, Florida 32316-2214
Tel. No. (850) 599-1560
Fax. No. (850) 878-0777

Donna Canzano McNulty
MCI WorldCom, Inc.
325 John Knox Road
The Atrium, Suite 105
Tallahassee, FL 32303
Tel. No. (850) 422-1254
Fax. No. (850) 422-2586


Brian Sulmonetti
MCI WorldCom, Inc.
6 Concourse Parkway, Suite 3200
Atlanta, GA 30328
Tel. No. (770) 284-5493
Fax. No. (770) 284-5488

Catherine F. Boone, Esq.
Covad Communications Company
10 Glenlake Parkway
Suite 650
Atlanta, Georgia 30328

John Rubino
George S. Ford
Z-Tel Communications, Inc.
601 South Harbour Island Blvd.
Tampa, Florida 33602
Tel. No. (813) 233-4630
Fax. No. (813) 233-4620
gford@z-tel.com

Joseph A. McGlothlin
Vicki Gordon Kaufman
McWhirter, Reeves, McGlothlin,
Davidson, Decker, Kaufman, et. al
117 South Gadsden Street
Tallahassee, Florida 32301
Tel. No. (850) 222-2525
Fax. No. (850) 222-5606
jmcglothlin@mac-law.com
vkaufman@mac-law.com

Jonathan E. Canis
Michael B. Hazzard
Kelley Drye & Warren, LLP
1200 19th Street, N.W., Fifth Floor
Washington, DC 20036
Tel. No. (202) 955-9600
Fax. No. (202) 955-9792
jacanis@kelleydrye.com
mhazzard@kelleydrye.com


Nancy B. White (for)

BEFORE THE FLORIDA PUBLIC SERVICE COMMISSION

In re: Investigation into the)
Establishment of Operations Support)
Systems Permanent Performance)
Measures for Incumbent Local Exchange)
Telecommunications Companies)

Docket No. 000121-TP

Filed: September 29, 2000

BELLSOUTH TELECOMMUNICATION, INC.'S REPLY COMMENTS

BellSouth Telecommunications, Inc. ("BellSouth") hereby files its Reply Comments, pursuant to the notice given at the workshop held August 8, 2000, by the Staff of the Florida Public Service Commission ("Commission"), and states the following:

Issue 1: Does the Commission have the authority to establish, in advance, a generic enforcement mechanism provision which would be inserted in interconnection agreements in the event negotiations on this provision fail?

Issue 2: Does the adoption of an enforcement mechanism provision by the Commission constitute the awarding of damages?

AT&T and MCI have filed comments on these two issues that are virtually verbatim copies of one another. Time Warner concurred in the comments of AT&T and MCI, albeit without recopying those comments again. None of the comments of these parties, however, really address the issue at hand. The comments of AT&T and MCI advance the view that this Commission has addressed matters that arise under the Telecommunications Act generically in the past, so this practice must be permissible. These parties go on to quote at some length portions of the recent decision in *MCI Telecommunications*

DOCUMENT NUMBER-DATE
12405 SEP 29 8
FPSC-REGS/REG/REPORTING



Corporation v. BellSouth Telecommunications, Inc. (Case No. 4:97CV141-RH, entered June 6, 2000). These parties, however, ignore a crucial part of the Court's decision. As BellSouth pointed out in its initial Comments, the Federal Court ruled, among other things, that the Commission must consider any issues brought before it in the context of an arbitration. From a practical standpoint, to the extent the Commission predetermines an issue generic before the fact of any given arbitration, it would not be able to consider varying proposals by parties to the arbitration. Thus, the MCI decision would appear to substantially undercut the practice that this Commission has followed in the past of addressing Federal Act issues generically.

Beyond this, AT&T and MCI, make the outrageous proposal that this proceeding should be used to set generic performance standards and penalties that would be binding upon BellSouth, but not upon ALECs. Specifically, AT&T states that performance measurements and penalties set in this proceeding "would, of course, be binding upon BellSouth." However, to the extent that an ALEC wishes "to negotiate (and perhaps arbitrate) different measurements . . . AT&T believes those CLECs would be able to do so in the context of their individual arbitration proceedings." (AT&T Comments, pp. 2-3; See MCI Comments, p. 3). Thus, these parties are not really proposing that the matters at issue be resolved in a generic proceeding. Instead, they are proposing that a floor for performance standards and penalties would be set generically and imposed on BellSouth, but that ALECs would be free to argue differing standards in future arbitrations if they wish to do so. This proposal may well address the

Federal Court's mandate to consider individual issues raised in arbitrations to the extent it allows ALECs to raise individual issues. However, the Federal District Court did not suggest that ALECs should be able to raise issues in arbitrations, but that BellSouth would not. Thus, from a legal standpoint, this proposal is no more sustainable than setting standards generically and declining to consider alternative proposals made by the parties to future arbitrations. Further, this proposal is egregiously unfair and obviously one-sided. To the extent that generic standards are set, they should apply to all. There is absolutely no justification (nor do these parties offer any) for the position that standards should be imposed on BellSouth, but that the ALECs should not be bound to accept these standards.

Moreover, if AT&T, et al. get their way, then any generic proceeding would simply be a waste of time. Standards would be set, but ALECs would be free to ignore these standards and request anything that they may wish in arbitrations. As a result, the Commission would be faced with arbitrating the subject issues in precisely the same way as if there were no generic standards. Thus, other than serving the self-interest of ALECs in a blatantly inequitable fashion, this proposal accomplishes nothing.

As to the issue of whether the contemplated compensation mechanism constitutes damages, AT&T, et al. simply wave the question away by saying that it does not matter in light of the Federal Court decision. As was discussed above, however, the federal court's decision involves more than these parties choose to acknowledge.

Sprint, however, provides a more substantive response, in which they contend that money paid through this compensation mechanism would be a penalty, but would not constitute damages. It is noteworthy that, in making this claim, Sprint attempts to distinguish our situation from that in the case in which the Florida Supreme Court affirmed that this Commission may not award damages, *Southern Bell Telephone and Telegraph Company v. Mobile America Corp., Inc.*, 291 So 2d 1999 (Fla. 1974). Fundamentally, however, the situation in that case was much like the situation in which we now find ourselves. In *Mobile America*, the plaintiff filed a claim in State Court, claiming that BellSouth had failed to provide an adequate quality of service and, as a result, that it had been damaged and was owed compensation. In other words, there (as here) someone was claiming that it should receive a direct payment for some alleged failure (in our case, prospective) to perform at a certain level.

Sprint's contention that the subject enforcement mechanism is only a penalty ignores the fact that the Commission has had in place a mechanism (discussed at greater length in BellSouth's Comments filed August 8, 2000) to levy penalties when appropriate, and this mechanism entails payment to the State, not a direct payment of money to a specific party. Sprint, nevertheless, contends that the contemplated direct payment is a penalty rather than damages because this payment is designed to incent (or punish), not to compensate. This contention by Sprint, of course, ignores the fact that (again, as discussed by BellSouth in its Comments) the Federal Court in *MC/* specifically referred to the mechanism in question as a compensation mechanism.

If the mechanism in question were truly a penalty, then it would be administered as every other penalty levied by the Commission in the past has been, i.e., with payment from the party upon whom the penalty is assessed going to the State, not to any individual party. The Federal Court's decision made it clear that it considers this mechanism to be a means of compensation, and Sprint's argument to the contrary is unpersuasive.

Issue 3: What should be the objective of an enforcement mechanism?

It appears from the comments of all parties to this proceeding that there is no real disagreement that the objective of enforcement mechanisms is to drive nondiscriminatory behavior on the part of the ILECs and that enforcement mechanisms should be swift and self-executing with minimal regulatory oversight. However, the key area of dispute is the timing of enforcement mechanisms, which was addressed thoroughly in BellSouth's original comments. Unlike the other respondents, BellSouth maintains that enforcement mechanisms are designed to guard against backsliding after an ILEC receives 271 authority as opposed to a tool for ensuring pre-271 compliance as proposed by the ALECs in this proceeding.

The FCC did not adopt penalties in the Local Competition Order. Instead, it acknowledged the wide variety of remedies available to an ALEC that believes it has received discriminatory performance in violation of the Act; see *FCC's Local Competition Order* ¶ 129, 11 *FCC Rcd.* at 15565 (*emphasizing the existence of sections 207 and 208 FCC complaints for damages, as well as*

actions under the antitrust laws, other statutes and common law); and “encourage[d]” the States only to adopt reporting requirements for ILECs. Likewise, in its order approving Bell Atlantic’s entry into long distance in New York, the FCC analyzed Bell Atlantic’s performance plan “solely for the purpose of determining whether the risk of post-approval non-compliance is sufficiently great that approval of its section 271 application would not be in the public interest.” Bell Atlantic Order, at ¶433 n.1326.

The FCC has made it clear that the primary, if not sole, purpose of a voluntary self effectuating remedy plan is to guard against RBOC “backsliding;” that is, providing discriminatory performance after it has received the so-called “carrot” of long distance approval. Moreover, the FCC has set forth the appropriate framework for analyzing the reasonableness of a proposed enforcement plan. Although conceding the details of such plans may legitimately vary widely, the FCC identified five key aspects of a performance assurance plan that should be examined to determine whether it falls “within a zone of reasonableness, and [is] likely to provide incentives that are sufficient to foster post-entry checklist compliance.” *Id.* at ¶433. BellSouth submits that its voluntary proposal should be accepted by this Commission because it clearly falls well within the FCC’s prescribed “zone of reasonableness,” and provides powerful incentives to foster post-entry checklist compliance. This Commission will continue to monitor BellSouth’s performance and can evaluate the effectiveness of VSEEM III once it is put into place to determine if it in fact operates as an

effective deterrent against discriminatory performance. If it does not, the Commission retains full authority to re-visit this issue.

In addition, BellSouth's position on the appropriate measurements to include in a Penalty Plan is that it is only necessary for key measurements, specifically those contained in BellSouth's VSEEM III plan. These key measurements capture the effect of disparate treatment where it may occur in other sub-process measurements (e.g. those in the BellSouth SQM that are not statistically tested in VSEEM). Statistical testing for determining penalties is simply not required for each and every measurement. This was evidenced in the FCC's New York 271 approval Order, ¶ 439 in general including footnotes 1342 and 1343. It is expensive to perform and can produce confusing and conflicting results, particularly where a statistical test of a sub-process measure, held orders as an example, shows disparate treatment while the overall process measurements, Order Completion Interval and % Missed Installation Appointments, show parity.

Issue 4: For purposes of evaluating ILEC performance in the context of an interconnection agreement, how should any Commission established enforcement mechanism be structured conceptually?

- A. Frequency of monitoring?
- B. Time frame to be evaluated?

It appears that the respondents are in agreement that the appropriate frequency of monitoring and time frame to be evaluated should be monthly. BellSouth supports this position.

Issue 4:

- C. Level of disaggregation across metrics and offerings?**
- D. How should items A, B, and C above be balanced to provide statistical significance for metrics with a small number of observations per reporting period?**

The issue of what metrics or offerings should be included in an enforcement plan is one of the most significant areas of disagreement among the parties. BellSouth's view is that the enforcement plan should include only key outcome-oriented metrics and offerings. These metrics and offerings are detailed on BellSouth's comments of August 25, 2000 and should be more than sufficient to deter backsliding. BellSouth contends that it is not necessary to attach enforcement to each process, sub-process, product, activity, transaction or metric. As mentioned above, the FCC agrees. In sharp contrast, AT&T and Time Warner (concurring in the AT&T proposal) believe disaggregation across metrics and offerings should be taken to extremes. There are numerous examples of the absurdity of this proposal but for purposes of space, BellSouth will cite just one that is representative.

Appendix A of Attachment 1 of AT&T's filing dated August 25, 2000, Section E, Item 1 specifies separate disaggregation for New Service Installations. This single activity is further broken down by 26 levels of product disaggregation (Section G), 2 levels of dispatch / non-dispatch (Section D, Item 2), 3 levels of volume (Section D, Item 3), and 13 levels of geography for the state and MSA (Section D, Item 4). Thus for one activity, New Service Installations, associated with one measurement, % Missed Installation Appointments as an example, we

would have 26 products, times 2 (dispatch/non-dispatch), times 3 (volumes), times 13 (geography), or 2,028 individual remedy determinations. Going further, since New Service Installations is just one of the 13 activities AT&T proposes in Section E, we must multiply 2,028 by 13 to get 26,364 individual parity and remedy determinations for just ONE measurement, % Missed Installation Appointments. And this is for just ONE ALEC.

It is interesting to note that in the first paragraph of page 29 (Attachment A) of AT&T's August 25, 2000 filing, AT&T suggests that a robust system of performance measurements should monitor 'all key aspects of market entry.' If each of the 26,364 remedy determinations for one overall measurement, % Missed Installation Appointments, is considered key, it is difficult to conjure up a 'non-key' measure. BellSouth believes only key outcome oriented metrics and key offerings should be included and opposes the position of AT&T that nearly every metric, offering, activity, process and sub-process should be included.

Once the issue of which metrics and offerings should be included is resolved, the next issue is how these should be evaluated for enforcement purposes. With the key measures and offerings discussed above, BellSouth's approach is to disaggregate the data to low levels for comparison purposes, but use appropriate statistical techniques to aggregate for purposes of determining parity. (AT&T on the other hand does not aggregate for purposes of determining parity. Rather the parity determination is made at the lowest level, the 26,364 tests noted in the example above.) BellSouth's VSEEM III plan takes small and large sample sizes into account in two ways; 1) the level for testing, and 2) the

level for decision making and reporting. The overall aggregate statistic (a.k.a., truncated-z statistic) along with a balancing critical value are vital to BellSouth's enforcement plan.

The reasons for generating an overall aggregate statistic is to increase one's confidence that when drawing a conclusion about parity, one can be assured that what is happening (in the underlying data) is not necessarily due to random chance. Basing a decision on limited information (i.e., small samples) could result in an improper conclusion. In the process of aggregating test results more information is available, thereby increasing the decision-makers' confidence that a conclusion about parity or disparity is reasonably certain (and has considered issues related to random variation). Similar concepts were proposed by Bell Atlantic-New York ("BA-NY") and Southwestern Bell Telephone-Texas ("SWBT-TX"), and approved by the FCC.

BellSouth has adopted a methodology wherein this 'aggregation' takes place in the statistical procedures. BellSouth adopts the truncated-z statistic and related procedures, primarily because it is based on an extensive review of BellSouth data. Two processes take place: 1) the process of truncating positive performance (to zero) minimizes masking systematic discrimination, and 2) the aggregation of test results satisfies issues related to random variation.

BellSouth's VSEEM III plan carries the same concepts found in the FCC Approved BA-NY and SWBT-TX plans; where 1) an equivalent zero credit for positive performance, and 2) an aggregation process exists in their remedy plans. One key difference in BellSouth's approach is that the aggregation occurs

within a measure by mode of entry, rather than across all measurement types. BellSouth has built upon the aggregation concept present in approved plans, and over an 18-month period has evaluated multiple statistical procedures using actual data to derive the most appropriate method for aggregation.

BellSouth uses the results of the aggregate statistic (a.k.a., truncated-z statistic) as the starting point for the remedy plan. The truncated-z statistic along with a Balancing Critical Value provides the decision-maker the appropriate information for determining parity or disparity.

In summary, BellSouth follows concepts that exist in the FCC approved BA and SWBT plans, where aggregating disaggregate test results to a State level has been found to be a reasonable approach. The mechanics of BellSouth's VSEEM III also incorporate two guiding factors presented by the statisticians (Dr. Colin Mallows of AT&T and Ernst & Young, on behalf of BellSouth) in the Louisiana Performance Measurements Workshop. These factors are: 1) the importance of disaggregating the data to a fine level so that appropriate like-to-like comparisons of ALEC and ILEC data can be made, and 2) that each performance measure of interest should be summarized by one overall test statistic giving the decision-maker a rule that determines whether a statistically significant difference exists. The methodology found in VSEEM III is appropriate for BellSouth data, and is consistent with the work performed by the AT&T and Ernst & Young statisticians in Louisiana.

AT&T proposes that an incomplete statistic be applied to inappropriate levels of disaggregation. AT&T continues to rely on the paper "Local Competition

Users Group (LCUG) – Statistical Tests for Local Service Parity, v1.0” published February 6, 1998, as documentation for the “modified z” statistic. The methodology described in this paper was developed in a vacuum devoid of real performance measurement data, and is incomplete, particularly in its handling of small samples. It utilizes sound theory, but does not prove practical for this application without several adjustments.

While the LCUG participants (at the time) were cited as AT&T, MCI, Sprint, LCI and WorldCom, it is believed that the primary author of the statistic was AT&T statistician, Dr. Colin Mallows. Since that time, Dr. Mallows has acknowledged that the LCUG Modified-z Test v1.0 requires enhancements. This recognition came about after testing the LCUG Modified-z statistic using “real data.”¹ Dr. Mallows worked with Ernst & Young statisticians, who were retained by BellSouth as statistical advisors.

Between April and September 1999, Dr. Mallows and the E&Y statisticians shared ideas concerning the analysis of BellSouth performance data. Once it was determined that all issues were properly addressed, an agreement was reached with Dr. Mallows on several issues, including how to calculate like-to-like cell level statistics. The methodology is described in “Statistical Techniques For The Analysis and Comparison Of Performance Measurement Data,”² henceforth

¹ Mallows, C. GA Direct Testimony, Docket 7892-U, June 20, 2000 (pg 11, line 16f)

² Submitted to the Louisiana Public Service Commission, Docket U-22252 Subdocket C. Revised February 28, 2000.

referred to as the “statistician’s report” (attached hereto as Attachment 1). The statistician’s report resulted in several conclusions and concepts, many of which point out problems with the modified z statistic, including:

1. An open recognition and statement of the importance of like-to-like comparisons.

Ernst & Young statisticians have always maintained that LCUG version 1.0 lacked a good discussion of this topic, and is therefore poor documentation of a complete methodology.

2. Guidance on defining small samples.

The modified z formulae given in LCUG version 1.0 are only appropriate when sample sizes are “large.” In order to define “large samples” one needs to take into account the characteristics of the data. Additionally, a complete methodology should provide rules for processing data when sample sizes are small.

BellSouth’s data have been studied, and rules for defining large samples have been determined. LCUG version 1.0 does not do this, and the rules in AT&T’s Performance Incentive Plan are inadequate. The rules given in the Louisiana “statistician’s report,” and adopted in BellSouth’s plan (Exhibit C of BellSouth’s comments filed with the FPSC on August 25, 2000) are based on the study of BellSouth’s performance measure data, and are therefore appropriate for use with BellSouth’s data.

3. Handling of the smallest sample sizes for mean measures by using permutation tests.

In dealing with mean performance measures, such as “order completion interval,” BellSouth determined that permutation-testing methodology should be used for small sample sizes. AT&T states this in their Performance Incentive Plan, but the guidance for determining small samples is inappropriate.

AT&T proposes the LCUG modified-z test for sample sizes of 31 or LCUG modified-z test is not adequate for samples ranging from 31 to 99. This is highlighted in Dr. Mallows’ paper “Notes On Some Analyses of BellSouth Data,” handed out at a July 20, 1999 AT&T *ex parte* with the FCC, where it is stated “...These calculations suggest that ... we can use the LCUGZ whenever nALEC is at least 100, and need to use a more accurate method when nALEC is smaller than 100.” (pg 4).

A preferred approach is to use the data tested methods in the “statistician report” (see item 4 below) which will require BellSouth to perform a permutation test on only the smallest of samples. This is the method proposed by BellSouth and is far less costly in terms of time and resources.

4. Adjustment to the modified z statistic for mean performance measures so that it obtains the correct result in situations where there are at least 7 transactions for both the ILEC and ALEC.

Permutation testing is very computationally intensive, and

using it to determine z-values for thousands of like-to-like comparisons will consume a large amount of computer resources. Furthermore, unnecessary permutation testing does not support the desire to have a self-effectuating enforcement mechanism, ready for production processing and capable of operating efficiently, with minimal human intervention. Therefore permutation tests should only be used on the smallest of samples.

In order to cut down on the number of permutation tests, an adjusted version of the modified z was developed. The formula for this new statistic is in the "statistician's report," and its derivation is in a paper submitted by Dr. Mallows and Ernst & Young's Dr. Sandy Balkin for publication in a statistics journal.³ (attached hereto as Attachment 2)

5. Use of the appropriate exact testing distribution to form the cell z-value for proportion and rate measures.

The statisticians determined that there is no need for a "large" sample approximation to determine the z-value of a proportion or rate measure within a like-to-like class. Instead formulae that standardize the difference in performance based on the exact testing distributions⁴ are used. This gives formulae that are slightly different than what is in LCUG version 1.0.

³ Balkin, S. and Mallows, C. "An Adjusted Asymmetric Two Sample t-Test." Submitted to *The American Statistician*, May 2000.

⁴ The exact distribution for comparing proportion measures is the hypergeometric distribution. The exact distribution for comparing rate measures is the binomial distribution.

6. Development of an aggregate statistic that doesn't mask systematic discrimination.

Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists. This should be done in a way that does not mask systematic discrimination.

7. Development of a balancing method for choosing critical values.

The testing methodology should balance Type I and Type II error probabilities. A Type I error adversely affects BellSouth; a Type II error adversely affects an ALEC. Balancing the error probabilities ensures that both sides assume the same level of uncertainty in the decision process.

These enhancements stem from Dr. Mallows' and the Ernst & Young team's ability to analyze real data. Dr. Mallows states in his direct testimony in Georgia, page 5, lines 18-20 and page 6, lines 1-2, "The ability to look at the data and analyze it is critical to determining the appropriate statistical test. One cannot be assured that data characteristics are properly accounted for in the statistical methodology unless one can observe the data and how it behaves over time."⁵

⁵ Mallows, C. GA Direct Testimony, Docket 7892-U, June 20, 2000

AT&T notes the pitfalls of utilizing the LCUG Modified-z v1.0 methodology, as documented. However, AT&T and other ALECs continue to present the LCUG modified z statistic as a sound piece of work, when in fact it is incomplete. BellSouth's proposal to use the methodology outlined in the Louisiana "Statistician's Report," on the other hand, does not suffer from the shortcomings listed above.

The proposal made by AT&T is to apply the LCUG modified-z statistic to the levels of disaggregation proposed in Appendix A of their original filing. Examination of BellSouth's performance measure data indicates that testing at this level of disaggregation may bias the results of each test. This was first raised as an issue by Ernst & Young's statistical team in November 1998.⁶ It was suggested that in order to have like-to-like comparisons, factors such as geography, business unit, and time should be considered to eliminate potential biases.

AT&T's Dr. Mallows agrees. In the paper "Questions Concerning the Statistical Methodology to Use for Evaluating Performance Measurements," submitted to the FCC April 12, 1999, Dr. Mallows states (page 43) "In any event, disaggregation should be at a level where relatively few expected dissimilarities in performance exist, so that both the mean or average performance of the group and the expected variance should be the same."

Furthermore, in the paper "Notes On Some Analyses of BellSouth Data," handed out at a July 20, 1999 *ex parte* with the FCC, Dr. Mallows states "The

⁶ Interim Statistical Analysis for BellSouth Telecommunications, Inc., LPSC, Docket U-22252 Subdocket C, November 19, 1998.

published LCUG documents say nothing about disaggregation. In those documents, it was assumed that the method would be applied in situations where like was being compared with like. However, Ernst and Young have since pointed out that it is important to disaggregate the data, because otherwise biases can be introduced that give the illusion of discrimination even when perfect parity exists in every cell.”

It has been suggested that the wire center be used as a geographic/business unit factor. In the July 20, 1999 *ex parte* report, Dr. Mallows analysis concludes that disaggregation by wire centers is important. On page 9 he states “The fact that the between-cell sum of squares is reduced when the wire-center identification is taken into account shows that it is important to make this adjustment.”

All of this is to say that the disaggregation levels in the AT&T plan are not necessarily appropriate for the BellSouth region. AT&T’s proposed product disaggregation is not necessary, and the proposal is void of important factors such as wire center geography/business unit and time. Disparities within product services will be detected in the product classes identified in BellSouth’s proposal. Too much disaggregation in this area may result in having no like-to-like classes; stated differently, we run out of data to compare. The product disaggregation proposed by BellSouth is appropriate for BellSouth data. Wire center geography/business unit disaggregation, as well as the time a transaction takes place, are important factors for BellSouth performance measure comparisons (based on examination of the data).

In summary, AT&T proposes that the disaggregate level of testing be the same as the disaggregate level of reporting (where a conclusion about parity is drawn). This is contrary to the guidance provided by the statisticians; one of whom represented AT&T. Again, BellSouth proposes a disaggregate level of testing to ensure like-to-like comparisons are being made, and an aggregate level of reporting from which the decision makers, the Commission, can draw a proper conclusion about parity.

Issue 4: Automatic penalties for noncompliance?

Again, there appears to be no disagreement amongst the respondents that penalties should be automatic. BellSouth's VSEEM III plan is designed such that penalties are triggered automatically for all three tiers without the need for regulatory oversight.

Issue 5: For purposes of evaluating ILEC (and ALEC) performance in the aggregate, how should any Commission's enforcement mechanism be structured conceptually?

- A. Frequency of monitoring?**
- B. Time frame to be evaluated?**

Evaluation of compliance should be performed on monthly results and the enforcement consequences should be evaluated quarterly.

BellSouth's proposal for evaluating performance in the aggregate is addressed in Tier 2 of BellSouth's VSEEM III plan. Tier 2 focuses on the ALEC industry while Tier 1 addresses the individual ALEC. BellSouth proposes a

quarterly evaluation of enforcement for the industry. This proposal is a result of adhering to the desire to develop a remedy plan that is 'simple and relatively easy to implement'.

Although Tier 2 is evaluated quarterly, the frequency of non-compliance is incorporated in the remedy rendered. The incentive for BellSouth is in recognizing that any miss – minor, moderate or severe -- may result in a Tier-2 remedy. Stated differently, BellSouth could experience (what some may term) "basic failures" for three consecutive months in a calendar quarter, which would result in a Tier 2 failure. Note, that this implies that any sequence of five consecutive failures will trigger a Tier-2 remedy.

While BellSouth's plan for aggregate enforcement is triggered by recurring failures, aggregate enforcement is not dependent upon the severity of the failure. The plans offered by the other parties are dependant on both frequency and severity.

Issue 5:

- C. Level of disaggregation across metrics and offerings?**
Same as 4C above.
- D. How should items A, B, and C above be balanced to provide statistical significance for metrics with a small number of observations per reporting period?** Same as 4D above
- E. Automatic vs. case-by-case fines for noncompliance?**
Same as 4E above.

Issue 6: How should the dollar value of penalties be determined?

Dollar values should be determined on a per unit basis and tied to the cost or perceived value to the end-user customer. AT&T proposes a per measure penalty structure which uses the ratio of the z score to the balancing critical value to estimate the severity of a failure. In doing so, they ignore the relationship between the precision of the severity estimate and the sample size when they determine the remedy amount for a failure.

To demonstrate this, compare two testing situations: one based on only two transactions, and one based on hundreds of transactions.

Example 1: Maintenance Average Duration in the UNE Loop product category.

ILEC: $n_1 = 1$ trouble that took 5 hours to repair

ALEC: $n_2 = 1$ trouble that took 5.25 hours to repair

This situation calls for a permutation test. Since there are on two transactions, there are only two possible permutations, ILEC\ALEC result of 5\5.25 or ALEC\ILEC result of 5.25\5. Since we actually observe the former, we get a rank of 2, and

$$p = 1 - \frac{2 - .5}{2} = .25$$

Converting this to a standard normal Z score give

$$Z = -.674.$$

Using the formula for the balancing critical value of a modified z statistic⁷ based on AT&T's choice of $\Delta = .25$, we get

⁷ This will not produce a balancing critical value for this test. Because of the discrete nature of this test, and the very small sample sizes, it cannot be balanced. However, since AT&T provides no

$$z^* = \frac{-\delta}{2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{-.25}{2\sqrt{\frac{1}{1} + \frac{1}{1}}} = -.088$$

The ratio of the z score to the balancing critical value is therefore

$$\frac{Z}{z^*} = 7.631.$$

According to AT&T's consequence function, this is a severe failure that requires a \$25,000 remedy. Also, note that this result stays the same as long as the ALEC trouble takes longer to repair than the ILEC trouble. So there could be very little difference between the two repair times, but this would be judged as a severe failure.

Example 2: Maintenance Average Duration in the Residential, Resale POTS product category.

ILEC: $n_1 = 1500$ troubles that took $\bar{X} = 1$ hours on average to repair with a standard deviation $s = 0.25$ hours.

ALEC: $n_2 = 200$ troubles that took $\bar{X} = 1.25$ hours on average to repair

This situation calls for a modified z test.

$$Z = \frac{1 - 1.25}{.25\sqrt{\frac{1}{1500} + \frac{1}{200}}} = -13.284$$

The balancing critical value based on AT&T's choice of $\Delta = .25$ is

$$z^* = \frac{-.25}{2\sqrt{\frac{1}{1500} + \frac{1}{200}}} = -1.661$$

guidance on what should be used, we've chosen to use the formula for the balancing critical value for the modified -z statistic.

The ratio of the z score to the balancing critical value is therefore

$$\frac{Z}{z^*} = 8.$$

According to AT&T's consequence function, this is a severe failure that requires a \$25,000 remedy.

In each of the above examples, the severity of the failure is estimated to be about 8, but the estimates are based on vastly different sample sizes. It is unreasonable to pay a remedy of \$25,000 when the only evidence one has is two transactions (one ILEC and one ALEC), and the actual difference in the service times between the two transactions does not play a role in determining the severity.

On the other hand, the severity estimate in Example 2 is based on hundreds of transactions, and it may seem reasonable to pay a remedy of \$25,000. The difference in the two examples is that the severity estimate is more precise than that of the first example. This illustrates why it is important to take sample size into account when determining the amount of remedy that should be paid.

VSEEM III calls for remedies based on the number of affected ALEC transactions. The remedy amounts under the VSEEM III plan would be \$400 and \$20,000 for examples 1 and 2, respectively. This is more reasonable and fair.

Additionally, AT&T and other ALECs propose that all metrics should be treated (remedied) equally, yielding a per measurement payment scheme. This approach is not appropriate for remedy payment based on parity. It is important

to look at the example used by WorldCom to understand just how unreasonable this scheme is. The example compares average response time for queries with missing due dates. WorldCom suggests that "if ... the delayed or inaccurate response data cause the potential customer to be so dissatisfied with the ALEC that the customer never chooses the ALEC, then response time then becomes the critical measurement for that ALEC and the customer". Average response times for queries in Florida consistently average less than 6 seconds and is transparent to the ALEC's customer. Missing due dates, on the other hand, directly impact the ALEC customer's experience with the ALEC. How can WorldCom reasonably draw a correlation between these two measurements? More importantly, how can a six second or less delay possibly be more important than missing a due date?

In contrast, BellSouth's VSEEM III plan avoids these problems by looking at the value and cost by unit. For example, an investment UNE products / services is much greater than an investment in Resale products / services. As shown in BellSouth's proposed fee schedule, a month one remedy would be \$100 for Resale and \$400 for UNE product / services.

Issue 7: Should there be a cap on penalty amounts and if so, how should that cap be determined?

BellSouth proposes the use of an absolute cap. BellSouth's VSEEM III plan was developed with the criteria that a penalty plan should be self-effectuating. Consequently, each of the three tiers of remedies in VSEEM III is automatic. While the Commission can step in at any time, remedies will be

rendered as the performance is being monitored; however, no Commission order is necessary to render payment.

The ALEC plan, on the other hand, contains several glaring contradictions to the “self-effectuating” concept, most notably the so-called “procedural cap.” The VSEEM III Plan sets an automatic financial cap (absolute cap) based on a meaningful percentage of BellSouth’s net revenues in Florida. The ALECs procedural cap, on the other hand, only determines the point at which the ILEC is permitted to seek relief from additional penalties from the state commission. While professing that “the imposition of financial consequences must be prompt and certain, and consequences should be self-executing so that opportunities for delay through litigation and regulatory review are minimized” (see AT&T, page 4) the ALECs proposed procedural caps build such delay into the plan. The ALECs plan is even more problematic given that all the ALECs are really asking of the Commission is to defer setting a liability cap rather than setting one in this proceeding in advance of plan implementation. The more efficient plan will establish a reasonable cap at the outset rather than deferring the determination to some future point and creating the need for an additional proceeding. The VSEEM III plan ensures that every aspect of the plan will operate independently of the Commission; the ALEC plan, on the other hand, builds Commission involvement into the plan and thus is less desirable.

It is also important to remember that the self-effectuating cap in the VSEEM III plan is not an overall cap on BellSouth's liability for performance failures. As the FCC has pointed out, a penalty plan is not “the only means of

ensuring that [the RBOC] continues to provide nondiscriminatory service to competing carriers." *Bell Atlantic Order*, ¶ 435. Thus, any characterization of the VSEEM cap as an absolute cap on BellSouth's liability for performance failures is incorrect. Moreover, both the New York and the Texas plans have annual monetary caps similar to the VSEEM III cap.

Issue 8: How and when should consequences be escalated?

BellSouth supports escalating remedies with the certainty and duration of the violation. What is important here is the overall set of principles that is used to form a remedy plan. BellSouth supports the following principles:

- Inclusion of key, outcome oriented measures
- Designed to prevent BellSouth "backsliding" on CLEC service
- Comprehensive plan that is "Meaningful" and "Significant"
- Monetary remedies escalate with the certainty of failure
- Monetary remedies escalate with the duration of the failure
- Non-monetary consequences are incorporated in the plan
 - Addresses all CLECs in operation; large and small
 - Addresses the CLEC Industry
 - Uses sound statistical procedures
- Compares "like-to-like" with deep disaggregation
- Solves the problem of 'random variation'
- Procedures do not 'mask discrimination'
- Methodology for balancing Type I and Type II Errors

- Minimize opportunities for 'Gaming'
- Structured such that CLECs will not prefer Remedies over Quality Service
 - Swift and Self-Executing
- Interest paid on remedy rendered for each date past due
 - Not applied until after 271 approval in a specific state
- Fairly simple to implement and monitor

BellSouth's VSEEM III 3-tiered remedy plan satisfies these principles and, as such, should be the remedy plan adopted by this Commission, if the Commission decides to adopt a remedy plan. The use of a multi-tiered plan is consistent with the principles of the other ALECs in this proceeding.

Issue 9: How should extraordinary events be handled?

While all the respondents in this proceeding agree that there are situations that should be legitimately excluded from a remedy plan, there are some differences of opinion as to how to identify these exclusions. Both Sprint and Time Warner suggest using the "root cause analysis" process. One of BellSouth's VSEEM III plan's primary objectives is to implement a plan that is "swift" and straightforward, offering remedy payment within 30 days after disparate service is reported. Developing corrective action plans and performing "root cause analysis" is inconsistent with these objectives. That is not to say, however, that BellSouth will not, upon request, perform a root cause analysis. It simply should not be part of a self-effectuating remedy plan. BellSouth proposes

that exclusions from the remedy plan due to extraordinary events be identified up front as part of the Remedy Plan as much as possible. Any additional unforeseen extraordinary events, which would qualify for exclusion would be negotiated individually with this Commission. This appears to be consistent with the proposal by AT&T (pages 13-14) with one glaring exception. AT&T proposes that those penalties "that are the subject of the potential exemption shall be paid into an interest bearing escrow account no later than the due date applicable to the consequences that are at issue." BellSouth takes exception to any requirement that it should make remedy payments into an interest bearing escrow account pending the validation of those remedies. BellSouth is a major corporation in Florida with a solid reputation. BellSouth's VSEEM III plan is designed to make penalty payments automatically when due. It is totally unnecessary and degrading that any party to this proceeding would think it necessary to require BellSouth to make payments into an escrow account.

Additional rebuttal comments

Rather than responding to the questions posed by the Florida Public Service Commission Staff, WorldCom answered only questions 1 and 2 specifically and proceeded to provide what they identified as technical comments. BellSouth would like to take this opportunity to offer rebuttal comments on six of these technical comments, found on pages 4-9 of WorldCom's response, dated August 25, 2000, in this proceeding as follows:

- 1. Remedy Measures, page 4. "WorldCom disagrees that there should be only outcome-oriented metric remedies".**

Voluntary self-effectuating remedies should apply to those key, outcome oriented measures that ALECs have identified as most critical to their businesses. Additionally, imposition of voluntary, self-effectuating penalties on every measure will impermissibly subject BellSouth to being penalized more than once for a single act or failure to act because many of the measures are integrally interrelated to one another.

The measurement set included in the VSEEM III plan are key, outcome oriented measures. BellSouth decided on these measures by looking at the collaborative work between ILECs, ALECs and State Commissions in New York and Texas. Collaborative efforts in both New York and Texas resulted in either a "critical" measurement set, or a prioritized set of "high, medium, low", respectively. These commissions charged the ALECs with communicating the measurement set that is most 'customer impacting'. BellSouth's experience in providing access to IXCs, combined with the outcome of prioritized measures from New York and Texas has resulted in BellSouth offering of a key set of customer impacting metrics.

Below are the measures included in the plan. The list represents the combination of Tier-1, Tier-2 and Tier-3 submetrics:

VSEEM III Sub-Metrics

- Percent Response Received within "X" seconds – Pre-Order OSS
- OSS Interface Availability

- ❑ Order Process Percent Flow-Through (Mechanized only)
- ❑ FOC Timeliness (Mechanized only)
- ❑ Reject Interval (Mechanized only)
- ❑ Order Completion Interval (Dispatch only) – Resale POTS
- ❑ Order Completion Interval (Dispatch only) – Resale Design
- ❑ Order Completion Interval (Dispatch only) – UNE Loop and Port Combos
- ❑ Order Completion Interval ('w' code orders, Dispatch only) – UNE Loops
- ❑ Order Completion Interval (Dispatch only) – IC Trunks
- ❑ Percent Missed Installation Appointments – Resale POTS
- ❑ Percent Missed Installation Appointments – Resale Design
- ❑ Percent Missed Installation Appointments – UNE Loop and Port Combos
- ❑ Percent Missed Installation Appointments – UNE Loops
- ❑ Percent Provisioning Troubles within 4 Days - Resale POTS
- ❑ Percent Provisioning Troubles within 4 Days - Resale Design
- ❑ Percent Provisioning Troubles within 4 Days - UNE Loop and Port Combos
- ❑ Percent Provisioning Troubles within 4 Days - UNE Loops
- ❑ Customer Trouble Report Rate – Resale POTS
- ❑ Customer Trouble Report Rate – Resale Design
- ❑ Customer Trouble Report Rate - UNE Loop and Port Combos
- ❑ Customer Trouble Report Rate - UNE Loops
- ❑ Percent Missed Repair Appointments – Resale POTS
- ❑ Percent Missed Repair Appointments - Resale Design

- Percent Missed Repair Appointments - UNE Loop and Port Combos
- Percent Missed Repair Appointments - UNE Loops
- Maintenance Average Duration – Resale POTS
- Maintenance Average Duration – Resale Design
- Maintenance Average Duration - UNE Loop and Port Combos
- Maintenance Average Duration - UNE Loops
- Maintenance Average Duration – IC Trunks
- Percent Repeat Troubles within 30 Days – Resale POTS
- Percent Repeat Troubles within 30 Days – Resale Design
- Percent Repeat Troubles within 30 Days - UNE Loop and Port Combos
- Percent Repeat Troubles within 30 Days - UNE Loops
- Billing Timeliness
- Billing Accuracy
- Usage Data Delivery Timeliness
- Usage Data Delivery Accuracy
- Percent Trunk Blockage
- LNP Disconnect Timeliness
- LNP Percent Missed Installation Appointments
- Coordinated Customer Conversions for UNE Loops w/o INP
- Percent Missed Collocation Due Dates

Additionally, BellSouth notes that many of the measures are interrelated, and it would be particularly difficult to repeatedly provide disparate service for a measure without it surfacing through to those measures identified in the VSEEM

III plan. Correlation studies show that, of the Provisioning measures, Order Completion Interval, Percent Missed Installations, and Total Service Order Cycle Time are all positively correlated, though at varying strengths. Completion Notice Interval is not correlated with any of the three provisioning measures. Meanwhile, in the Maintenance category, all three measures are positively correlated, again at various degrees. BellSouth asserts that WorldCom's technical comments in this regard are unfounded and without merit and should be rejected by the Florida Staff.

- 2. Accurate Reporting, page 6. WorldCom alleges that "an ILEC should test all of its OSS systems and processes at least once a year at its own expense to prove its data is valid."**

BellSouth's Service Quality Measurements (SQM) document,

Appendix C, states:

"BellSouth currently provides many CLECs with certain audit rights as a part of their individual interconnection agreements. However, it is not reasonable for BellSouth to undergo an audit of the SQM for every CLEC with which it has a contract. BellSouth has developed a proposed Audit Plan for use by the parties to an audit. If requested by a Public Service Commission or by a CLEC exercising contractual audit rights, BellSouth will agree to undergo a comprehensive audit of the aggregate level reports for both BellSouth and the CLEC(s) for each of the next five (5) years (2000 – 2005), to be conducted by an independent third party. The results of that audit will be made available to all the parties subject to proper safeguards to protect proprietary information. This aggregate level audit includes the following specifications:

- 1. The cost shall be borne 50% by BellSouth and 50% by the CLEC or CLECs.*
- 2. The independent third party auditor shall be selected with input from BellSouth, the PSC, if applicable, and the CLEC(s).*
- 3. BellSouth, the PSC and the CLEC(s) shall jointly determine the scope of the audit.*

BellSouth reserves the right to make changes to this audit policy as growth and changes in the industry dictate."

In addition, the VSEEM III contract language for Interconnection Agreements in Section 4.6.5 states:

At the end of each calendar year, BellSouth will have its independent auditing and accounting firm certify that the results of all Tier-1 and Tier-2 Enforcement Mechanisms were paid and accounted for in accordance with Generally Accepted Accounting Principles (GAAP).

BellSouth asserts its auditing policy for VSEEM is a natural extension of its auditing policy for the SQMs and that the combination of these two positions fully satisfies the accurate reporting requirements proposed by WorldCom.

It is also important to emphasize the fact that BellSouth is currently in the final stages of a comprehensive audit by a third party auditor, KPMG, in Georgia and beginning a comprehensive audit by KPMG in Florida. The end result of these audits should more than satisfy WorldCom's requirements.

3. Weighting, page 7. WorldCom alleges that "this Commission should treat metrics equally and remedies should be paid on how disparate and chronic the poor performance is to the ALEC".

Although BellSouth has responded to this issue in item 1 above, **Remedy Measures**, it is important to look at the example used by WorldCom to understand just how unreasonable this allegation is. The example compares average response time for queries with missing due dates. WorldCom suggests that "if ...the delayed or inaccurate response data cause the potential customer to be so dissatisfied with the ALEC that the customer never chooses the ALEC, then response

time then becomes the critical measurement for that ALEC and the customer". Average response times for queries in Florida consistently average less than six seconds and are transparent to the ALEC's customer. Missing due dates, on the other hand, directly impact the ALEC customer's experience with the ALEC. How can WorldCom reasonably draw a correlation between these two measurements? More importantly, how can a six second or less delay possibly be more important than missing a due date?

4. **Minimum Thresholds (page 8). WorldCom asserts that "there should be no minimum thresholds before a plan can commence because the primary reason for an enforcement mechanism is to counter the dominance, incumbency and market power of the ILEC to prevent discriminatory treatment."**

BellSouth's VSEEM III proposal has no minimum thresholds and is designed specifically to ensure non-discriminatory treatment of all ALECs.

5. **Burn-In period (page 8). WorldCom alleges that "there is no need for a burn-in period".**


BellSouth concurs in WorldCom's position and stands ready to implement its VSEEM III plan once 271 approval is granted in Florida.

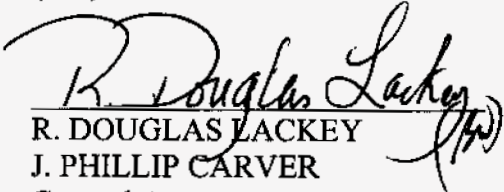
6. **Six-Month Reviews (page 9). WorldCom alleges that "because performance metrics may need to be modified over time to add new processes or adjust older benchmarks, this Commission should order that all interested parties meet every six months to**

review the metrics. The enforcement mechanism plan will also need to be analyzed during the six-month review to make sure the remedy amounts and structure are effective.”

BellSouth concurs that the performance metrics and enforcement plan should be reviewed periodically with input from all interested parties. However, BellSouth would strongly recommend that every six months is excessive and that a yearly review should be more than sufficient. BellSouth also strongly urges this Commission to consider as a part of the review the removal of measurements where there was insufficient activity during the previous year to justify their continued production.

Respectfully submitted this 29th day of September, 2000.


NANCY B. WHITE
c/o Nancy Sims
150 South Monroe Street
Suite 400
Tallahassee, FL 32301
(305) 347-5558


R. DOUGLAS LACKEY
J. PHILLIP CARVER
General Attorneys
Suite 4300, BellSouth Center
675 West Peachtree Street, N.E.
Atlanta, GA 30375
(404) 335-0765

229421

DRAFT

Statistical Techniques For The Analysis And Comparison Of Performance Measurement Data

Submitted to Louisiana Public Service Commission (LPSC)
Docket U-22252 Subdocket C

Revised February 28, 2000

Introduction and Scope

The Louisiana Public Service Commission (LPSC) staff has requested Drs. S. Hinkins, E. Mulrow, and F. Scheuren¹ of Ernst & Young LLP (consultants for BellSouth Telecommunications), and Dr. C. Mallows of AT&T Labs-Research to set out their views on the application of a statistical analysis to performance measurement data. The present report is intended to provide a detailed statistical report on appropriate methodology.

The setting for the analysis is crucial to the interpretation of any statistical significance that might be found. There is no doubt that, to quote the Commission staff, "statistical analysis can help reveal the likelihood that reported differences in an ILECs performance toward its retail customers and CLECs are due to underlying differences in behavior rather than random chance" (Staff Final Recommendation, LPSC Docket No. U-22252 - Subdocket C, dated August 12, 1998, pages 15 - 16).

To frame our presentation the next paragraph from the LPSC Docket U-22252 is quoted in its entirety.

"Statistical tests are effective in identifying those measurements where differences in performance exist. The tests themselves cannot identify the cause of the apparent differences. The differences may be due to a variety of reasons, including: 1) when the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result, 2) when the ILEC is employing discriminatory practices, or 3) when assumptions necessary for the statistical test to be valid are not being met." (Ibid., page 16)

Apparent statistically significant differences in BellSouth and CLEC performance can arise when

- the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result
- the ILEC is employing discriminatory practices, or
- assumptions necessary for the statistical test to be valid are not being met.

¹ Dr. Scheuren is now a Senior Fellow at the Urban Institute.

RECEIVED

MAR 01 2000

LOUISIANA PUBLIC SERVICE COMMISSION
ADMINISTRATIVE HEARINGS DIVISION

1 APR 10 39

1

90 APR 1

DRAFT

To meet the Louisiana Commission's purpose, we will recommend techniques that are robust in the presence of possible assumption failure, carefully examine BellSouth Telecommunications (BST) and CLEC performance so "like" is compared only to "like," and are still able, in a highly efficient manner, to detect differences. Upon investigation any differences detected might lead to concerns about possible discriminatory practices.

The LPSC staff also states "that a uniform methodology which identifies those items which need to be measured, how they are to be measured, and how the results are to be reported is also desirable and would be beneficial to all parties" (*Ibid.*, page 16). We agree with this goal as well, stipulating only that the use of a single method may not be desirable while a single methodology (or a set of methods) could be.

The statistical process for testing if CLEC and ILEC customers are being treated equally involves more than just a mathematical formula. Three key elements need to be considered before an appropriate decision process can be developed. These are

- the type of data,
- the type of comparison, and
- the type of performance measure.

When examining the various combinations of these elements, we find that there is a set of testing principles that can be applied uniformly. However, the statistical formulae that need to be used change as the situation changes.

To be responsive to the Commission, we have divided our discussion into four sections and five appendices. The contents of each of these are briefly mentioned below -- first for the main report and then for the extensive supporting appendix materials.

For the main report, this section (Section I) introduces our work and sets out the required scope. The next two sections (Sections II and III) discuss the type of comparisons that need to be identified, and the appropriate testing principles. The final section (Section IV) provides an overview of appropriate testing methodologies, based on what we have learned from our examination of BellSouth's performance measure data in Louisiana.

The five appendices provide technical details on the statistical calculations involved in the Truncated Z statistic (Appendix A), the implementation of the methodology for the trunk blocking performance measure (Appendix B), the calculations involved in computing the balancing critical value of a test (Appendix C), examples of ways to present the results using detailed statistical displays so that results can be audited (Appendix D), and the technical details involved in data trimming (Appendix E).

DRAFT

2. Data Considerations, Comparisons, and Measurement Types

This section makes general distinctions which apply to the performance measures. These distinctions will be important in the determination of appropriate methodologies.

Data Set Types. The type of statistical methodology used depends on the form of the data available. In general, there are two ways to classify the data used for performance measure comparisons. These are:

- transaction level data, and
- aggregated summaries.

Records in a transaction level data set represent a single transaction, e.g. an individual customer order, or the record of a specific trouble reported by a customer. This type of data set allows for deep like-to-like comparisons, and may also allow one to identify the root cause of a problem. A testing methodology needs to be carefully chosen so that it incorporates the comparison levels and does not cover up problem areas.

Records in an aggregated summary data set are typically summaries of related transactions. For example, the total number of blocked calls in a trunk group during the noon hour of a day is a summary statistic. This type of data set may not contain as much information as a transaction level data set, and it therefore needs to be treated differently. While a general methodology may be determined for a transaction level data set, it may not be possible to do so for aggregated summaries. Testing methodology needs to be developed on a case-by-case basis.

Comparison Types. An ILEC's performance in providing services to CLEC customers is tested in one of two ways:

- by comparing CLEC performance to ILEC performance when a retail analog exists, or
- by comparing CLEC performance to a benchmark.

The testing methodologies for these two situations will have similarities, but there are differences that need to be understood.

Table 1 categorizes those performance measures that E&Y has examined by data type and comparison type. The table shows that five performance measures with retail analogs have transaction level data, while three others with retail analogs only have summary level data. No performance measures using benchmarks have been studied.

DRAFT

Table 1. Classification of Performance Measures by Data and Comparison Type (only measures previously examined by E&Y are included)

Level of Data	Comparison Type	
	Retail Analog	Benchmark
Transaction Level	Order Completion Interval Maintenance Average Duration % Missed Installations % Missed Repair Trouble Report Rate	No Measures Examined
Summary Level	Billing Timeliness OSS Response Interval Trunk Blocking	No Measures Examined

Measurement Types. The performance measures that will undergo testing are of four types: means, proportions (an average of a measure that takes on only the values of 0 or 1), rates, and ratios.

While all four have similar characteristics, proportions and rates are derived from count data while means and ratios are derived from interval measurements. Table 2 classifies the performance measures by the type of measurement.

Table 2: Classification of Performance Measures by Measurement Type

Mean	Proportion	Rate	Ratio
Order Completion Interval Maint. Ave. Duration OSS Response Interval	Percent Missed Installations Percent Missed Repairs Billing Timeliness Trunk Blocking	Trouble Report Rate	Billing Accuracy

3. Testing Principles

This section describes five general principles which the final methodology should satisfy:

DRAFT

1. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*
2. *Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.*
3. *The decision system must be developed so that it does not require intermediate manual intervention.*
4. *The testing methodology should balance Type I and Type II Error probabilities.*
5. *Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.*

Like-to-Like Comparisons. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*

In particular, to meet this goal the testing process should:

- Identify variables that may affect the performance measure.
- Record important confounding covariates.
- Adjust for the observed covariates in order to remove potential biases and to make the CLEC and the ILEC units as comparable as possible.

It is a well known principle that comparisons should be made on equal footing: apples-to-apples, oranges-to-oranges. Statistical techniques that are addressed in most text books usually assume that this is the case beforehand. Some higher level books address the issue of “designed experiments” and discuss appropriate ways to structure the data collection method so that the text books’ formulae can be used in analyzing the data.

Performance measure testing does not involve data from a designed experiment. Rather, the data is obtained from an observational study. That being the case, one must impose a structure on the data after it is gathered in order to assure that fair comparisons are being made. For example, it is important to disaggregate the data to a fine level so that appropriate like-to-like comparisons of CLEC and ILEC data can be made. Any statistical methodology that ignores important confounding variables can produce biased results.

Aggregate Level Test Statistic. *Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.*

To achieve this goal, the aggregate test statistic should have the following properties:

DRAFT

- The method should provide a single overall index, on a standard scale.
- If entries in comparison cells are exactly proportional over a covariate, the aggregated index should be very nearly the same as if comparisons on the covariate had not been done.
- The contribution of each comparison cell should depend on the number of observations in the cell.
- Cancellation between comparison cells should be limited, i.e., positive outcomes should not be allowed to cancel negative ones.
- The index should be a continuous function of the observations.

Since the data are being disaggregated to a very deep level, thousands of like-to-like comparison cells are created. An aggregate summary statistic is needed in order to make an overall judgment.

The aggregate level statistic should be insensitive to small changes in cells values, and its value should not be affected if some of the disaggregation for like-to-like cells is truly unnecessary. Furthermore, individual cell results should be weighted so that those cells with more transactions have larger effects on the overall result.

Production Mode Process. *The decision system must be developed so that it does not require intermediate manual intervention.*

Two statistical paradigms are possible for examining performance measure data. In the exploratory paradigm, data are examined and methodology is developed that is consistent with what is found. In a production paradigm a methodology is decided upon before data exploration. For the production paradigm to succeed

- Calculations should be well defined for possible eventualities.
- The decision process should be based on an algorithm that needs no manual intervention.
- Results should be arrived at in a timely manner.
- The system must recognize that resources are needed for other performance measure-related processes that also must be run in a timely manner.
- The system should be both auditable and adjustable over time.

While the exploratory paradigm provides protection against using erroneous data, it requires a great deal of lead time and is unsuitable for timely monthly performance measure testing. A production paradigm will not only promptly produce overall test results but will also provide documentation that can be used to explore the data after the test results are released.

DRAFT

Error Probability Balancing. *The testing methodology should balance Type I and Type II Error probabilities.*

Specifically, what is required to achieve this goal is

- The probability of a Type I error should equal the probability of a Type II error for well-defined null and alternative hypotheses.
- The formula for a test's balancing critical value should be simple enough to calculate using standard mathematical functions, i.e. one should avoid methods that require computationally intensive techniques.
- Little to no information beyond the null hypothesis, the alternative hypothesis, and the number of observations should be required for calculating the balancing critical value.

The objective of a statistical test is to test a hypothesis concerning the values of one or more population parameters. Usually an inquiry into whether or not there is evidence to support a hypothesis, called the *alternative hypothesis*, is conducted by seeking statistical evidence that the converse of the alternative, the *null hypothesis*, is most likely false. If there is not sufficient evidence to reject the null hypothesis, then a case for accepting the alternative has not been made.

Two types of errors are possible in any decision-making process. These have been summarized in Table 3.

Table 3: Statistical Testing Errors

Decision Error	General Description	In terms of Performance Measure Testing
Type I	Rejecting the null hypothesis (accepting the alternative) when the null is true.	Deciding that BST favors its own customers when it does not.
Type II	Accepting the null hypothesis when the alternative is true.	Deciding that BST does not favor its own customers when it does.

In a controlled experimental study where the sample sizes are relatively small, it is generally desirable to control the Type I error closely to avoid making a conclusion that there is a difference when, in fact, there is none. The probability of a Type II error is not directly controlled but is determined by the sample size and the distance between the null and the alternative hypotheses.

DRAFT

If a standard of materiality is set by stating a specific alternative for the test, and the distribution of the test statistic under both the null and alternative hypotheses is understood, then a critical value can be determined so that the two error probabilities are equal.

Trimming. Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.

Three conditions are needed to accomplish this goal. These are:

- Trimming should be based on a general rule that can be used in a production setting.
- Trimmed observations should not simply be discarded; they need to be examined and possibly used in the final decision making process.
- Trimming should only be used on performance measures that are sensitive to "outliers."

For the purpose of performance measure testing, trimming refers to removing transactions that significantly distort the performance measure statistic for the set of transactions under consideration. For example, the arithmetic average (or mean) is extremely sensitive to "outliers" since a single large value can significantly distort the average.

The term "outliers" refers to:

- 1) extreme data values that may be valid, but since they are rare measurements, they may be considered to be statistically unique; or
- 2) large values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

Trimming is beneficial since it puts both ILEC and CLEC transactions on equal footing with respect to the largest value in each set. Note, though, that it is only needed for performance measures that are distorted by outliers. Of the three types of measures defined in Section 2, only mean (average) measures require trimming. Appendix E sets forth a trimming plan for mean performance measures.

4. Testing Methodology

This section details the testing methodology that is most appropriate for the various types of performance measures. First, transaction level testing will be discussed when there is a retail analog. Next, transaction level testing against a benchmark. Then, testing when only aggregated summaries are available.

Transaction Level - Retail Analog: The Truncated Z Statistic. When a retail analog is available CLEC performance can be directly compared with ILEC performance. Over

DRAFT

the last year. for transaction level data, many test statistics have been examined. We now believe that the “Truncated Z” test statistic provides the best compromise with respect to possessing the desired qualities outlined in Section 3, above.

The Truncated Z is fully described in Appendix A, and formulae for calculation of a balancing critical value are found in Appendix C. The main features of this statistic are:

- A basic test statistic is calculated within each comparison cell.
- The value of a cell’s result is left “as is” if the result suggests that “favoritism” may be taking place. Otherwise, the result is set to zero. This is called the truncation step.
- Weights that depend on the volume of both ILEC and CLEC transactions within the cell are determined, and a weighted sum of the “truncated” cell results is calculated.
- The weighted sum is theoretically corrected to account for the truncation, and a final overall statistic is determined.
- This overall test value is compared to a balancing critical value to determine if favoritism is likely.

The test statistic itself is based on like-to-like comparisons, and it possesses all five of the properties of an aggregate test statistic (Section 3). While the test requires a large amount of calculations, our studies of the process on some of BellSouth’s performance measure data indicate that the calculations can be completed in a reasonable amount of time. Therefore, the process can be put into production mode. Finally, since a balancing critical value can be calculated, it is possible to balance the error probabilities.

Transaction Level - Benchmark. When a benchmark is used, CLEC performance is not compared with ILEC performance. Like-to-like comparison cells are not needed, thus greatly simplifying the testing process. Statistical testing can be done using a probability model, or non-statistical testing can be done using a deterministic model. No data for this data/comparison class has been studied at this point in time.

Aggregated Summary - Retail Analog or Benchmark. We cannot provide any one single set of rules for the analysis of data in this class. Data that is an aggregated summary of transactions may or may not present problems. For example, BellSouth’s trunk blocking data is saved as summaries by hour of the day. Collectively, the summaries do provide sufficient information to proceed with the Truncated Z methodology.

On the other hand, our examination of the data for the OSS response interval revealed that information necessary for computing a Truncated Z was not available. In this case, however, we were able to construct a satisfactory time series method to analyze the measure.

DRAFT

Each measure falling into this class needs to be handled on a case-by-case basis. If sufficient information is available to use the Truncated Z method, then we feel it should be used. When the Truncated Z cannot be used, a testing methodology that adheres closely to the principles outlined in Section 3 should be determined and followed.

Appendix A. The Truncated Z Statistic

The Truncated Z test statistic was developed by Dr. Mallows in order to have an aggregate level test when transaction level data are available that

- provides a single overall index on a standard scale;
- will not change the outcome if the disaggregation is unnecessary,
- incorporates the number of observations in a cell into the determination of the weight for the contribution of each comparison cell,
- limits the amount of “neutralization” between comparison cells, and
- is a continuous function of the observations.

The Ernst & Young statistical team and Dr. Mallows have studied the implementation of the statistic using some of BellSouth’s performance measure data. This has resulted in an overall process for comparing CLEC and ILEC performance such that the following principles hold:

- 1) Like-to-Like Comparisons are made. (See Appendix B for an example based on the trunk blocking measure.)
- 2) Error probabilities are balanced. (See Appendix C)
- 3) Extreme values are trimmed from the data sets when they significantly distort the performance measure statistic. (See Appendix E)
- 4) The testing process is an automated production system. (Discussed here. See Appendix D for reporting guidelines.)
- 5) The determination of ILEC favoritism is based on a single aggregate level test statistic. (Discussed here.)

This appendix provides the details behind computing the Truncated Z test statistic so that principles 4 and 5 hold. We start by assuming that any necessary trimming of the data is complete, and that the data are disaggregated so that comparisons are made within appropriate classes or adjustment cells that define “like” observations.

Notation and Exact Testing Distributions

Below, we have detailed the basic notation for the construction of the truncated z statistic. In what follows the word “cell” should be taken to mean a like-to-like comparison cell that has both one (or more) ILEC observation and one (or more) CLEC observation.

- L = the total number of occupied cells
- j = 1, ..., L; an index for the cells
- n_{1j} = the number of ILEC transactions in cell j
- n_{2j} = the number of CLEC transactions in cell j
- n_j = the total number transactions in cell j; $n_{1j} + n_{2j}$

$$\begin{aligned}
X_{1jk} &= \text{individual ILEC transactions in cell } j; k = 1, \dots, n_{1j} \\
X_{2jk} &= \text{individual CLEC transactions in cell } j; k = 1, \dots, n_{2j} \\
Y_{jk} &= \text{individual transaction (both ILEC and CLEC) in cell } j \\
&= \begin{cases} X_{1jk} & k = 1, K, n_{1j} \\ X_{2jk} & k = n_{1j} + 1, K, n_j \end{cases}
\end{aligned}$$

$\Phi^{-1}(\cdot)$ = the inverse of the cumulative standard normal distribution function

For Mean Performance Measures the following additional notation is needed.

$$\begin{aligned}
\bar{X}_{1j} &= \text{the ILEC sample mean of cell } j \\
\bar{X}_{2j} &= \text{the CLEC sample mean of cell } j \\
s_{1j}^2 &= \text{the ILEC sample variance in cell } j \\
s_{2j}^2 &= \text{the CLEC sample variance in cell } j \\
\{y_{jk}\} &= \text{a random sample of size } n_j \text{ from the set of } Y_{j1}, K, Y_{jn_j}; k = 1, \dots, n_j \\
M_j &= \text{the total number of distinct pairs of samples of size } n_{1j} \text{ and } n_{2j}; \\
&= \binom{n_j}{n_{1j}}
\end{aligned}$$

The exact parity test is the permutation test based on the "modified Z" statistic. For large samples, we can avoid permutation calculations since this statistic will be normal (or Student's t) to a good approximation. For small samples, where we cannot avoid permutation calculations, we have found that the difference between "modified Z" and the textbook "pooled Z" is negligible. We therefore propose to use the permutation test based on pooled Z for small samples. This decision speeds up the permutation computations considerably, because for each permutation we need only compute the sum of the CLEC sample values, and not the pooled statistic itself.

A permutation probability mass function distribution for cell j, based on the "pooled Z" can be written as

$$PM(t) = P\left(\sum_k y_{jk} = t\right) = \frac{\text{the number of samples that sum to } t}{M_j},$$

and the corresponding cumulative permutation distribution is

$$\text{CPM}(t) = P\left(\sum_k y_{jk} \leq t\right) = \frac{\text{the number of samples with sum} \leq t}{M_j}$$

For Proportion Performance Measures the following notation is defined

- a_{1j} = the number of ILEC cases possessing an attribute of interest in cell j
- a_{2j} = the number of CLEC cases possessing an attribute of interest in cell j
- a_j = the number of cases possessing an attribute of interest in cell j; $a_{1j} + a_{2j}$

The exact distribution for a parity test is the hypergeometric distribution. The hypergeometric probability mass function distribution for cell j is

$$\text{HG}(h) = P(H = h) = \begin{cases} \frac{\binom{n_{1j}}{h} \binom{n_{2j}}{a_j - h}}{\binom{n_j}{a_j}}, & \max(0, a_j - n_{2j}) \leq h \leq \min(a_j, n_{1j}) \\ 0 & \text{otherwise} \end{cases}$$

and the cumulative hypergeometric distribution is

$$\text{CHG}(x) = P(H \leq x) = \begin{cases} 0 & x < \max(0, a_j - n_{2j}) \\ \sum_{h=\max(0, a_j - n_{2j})}^x \text{HG}(h), & \max(0, a_j - n_{2j}) \leq x \leq \min(a_j, n_{1j}) \\ 1 & x > \min(a_j, n_{1j}) \end{cases}$$

For Rate Measures, the notation needed is defined as

- b_{1j} = the number of ILEC base elements in cell j
- b_{2j} = the number of CLEC base elements in cell j
- b_j = the total number of base elements in cell j; $b_{1j} + b_{2j}$
- \bar{p}_{1j} = the ILEC sample rate of cell j; n_{1j}/b_{1j}
- \bar{p}_{2j} = the CLEC sample rate of cell j; n_{2j}/b_{2j}
- q_j = the relative proportion of ILEC elements for cell j; b_{1j}/b_j

The exact distribution for a parity test is the binomial distribution. The binomial probability mass function distribution for cell j is

$$BN(x) = P(B = k) = \begin{cases} \binom{n_j}{k} q_j^k (1 - q_j)^{n_j - k}, & 0 \leq k \leq n_j, \\ 0 & \text{otherwise} \end{cases}$$

and the cumulative binomial distribution is

$$CBN(x) = P(B \leq x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^x BN(k), & 0 \leq x \leq n_j. \\ 1 & x > n_j \end{cases}$$

For Ratio Performance Measures the following additional notation is needed.

- U_{1jk} = additional quantity of interest of an individual ILEC transaction in cell j ; $k = 1, \dots, n_{1j}$
- U_{2jk} = additional quantity of interest of an individual CLEC transaction in cell j ; $k = 1, \dots, n_{2j}$
- \hat{R}_{ij} = the ILEC ($i = 1$) or CLEC ($i = 2$) ratio of the total additional quantity of interest to the base transaction total in cell j , i.e., $\sum_k U_{ijk} / \sum_k X_{ijk}$

Calculating the Truncated Z

The general methodology for calculating an aggregate level test statistic is outlined below.

1. **Calculate cell weights, W_j .** A weight based on the number of transactions is used so that a cell which has a larger number of transactions has a larger weight. The actual weight formulae will depend on the type of measure.

Mean or Ratio Measure

$$W_j = \sqrt{\frac{n_{1j} n_{2j}}{n_j}}$$

Proportion Measure

$$W_j = \sqrt{\frac{n_{2j} n_{1j}}{n_j} \cdot \frac{a_j}{n_j} \cdot \left(1 - \frac{a_j}{n_j}\right)}$$

Rate Measure

$$W_j = \sqrt{\frac{b_{1j}b_{2j} \cdot n_j}{b_j \cdot b_j}}$$

2. **In each cell, calculate a Z value, Z_j .** A Z statistic with mean 0 and variance 1 is needed for each cell.

- If $W_j = 0$, set $Z_j = 0$.
- Otherwise, the actual Z statistic calculation depends on the type of performance measure.

Mean Measure

$$Z_j = \Phi^{-1}(\alpha)$$

where α is determined by the following algorithm.

If $\min(n_{1j}, n_{2j}) > 6$, then determine α as

$$\alpha = P(t_{n_{1j}-1} \leq T_j),$$

that is, α is the probability that a t random variable with $n_{1j} - 1$ degrees of freedom, is less than

$$T_j = t_j + \frac{g}{6} \left(\frac{n_{1j} + 2n_{2j}}{\sqrt{n_{1j} n_{2j} (n_{1j} + n_{2j})}} \right) \left(t_j^2 + \frac{n_{2j} - n_{1j}}{2n_{1j} + n_{2j}} \right),$$

where

$$t_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{s_{1j} \sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}$$

and the coefficient g is an estimate of the skewness of the parent population, which we assume is the same in all cells. It can be estimated from the ILEC values in the largest cells. This needs to be done only once for each measure. We have found that attempting to estimate this skewness parameter for each cell separately leads to excessive variability in the "adjusted" t . We therefore use a single compromise value in all cells.

Note, that t_j is the "modified Z" statistic. The statistic T_j is a "modified Z" corrected for the skewness of the ILEC data.

If $\min(n_{1j}, n_{2j}) \leq 6$, and

a) $M_j \leq 1,000$ (the total number of distinct pairs of samples of size n_{1j} and n_{2j} is 1,000 or less).

- Calculate the sample sum for all possible samples of size n_{2j} .
- Rank the sample sums from smallest to largest. Ties are dealt by using average ranks.
- Let R_0 be the rank of the observed sample sum with respect all the sample sums.

$$\alpha = 1 - \frac{R_0 - 0.5}{M_j}$$

b) $M_j > 1,000$

- Draw a random sample of 1,000 sample sums from the permutation distribution.
- Add the observed sample sum to the list. There is a total of 1001 sample sums. Rank the sample sums from smallest to largest. Ties are dealt by using average ranks.
- Let R_0 be the rank of the observed sample sum with respect all the sample sums.

$$\alpha = 1 - \frac{R_0 - 0.5}{1001}$$

Proportion Measure

$$Z_j = \frac{n_j a_{1j} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}$$

Rate Measure

$$Z_j = \frac{n_{1j} - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}}$$

Ratio Measure

$$Z_j = \frac{\hat{R}_{1j} - \hat{R}_{2j}}{\sqrt{V(\hat{R}_{1j}) \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)}}$$

$$V(\hat{R}_{1j}) = \frac{\sum_k (U_{1jk} - \hat{R}_{1j} X_{1jk})^2}{\bar{X}_{1j}^2 (n_{1j} - 1)} = \frac{\sum_k U_{1jk}^2 - 2\hat{R}_{1j} \sum_k (U_{1jk} X_{1jk}) + \hat{R}_{1j}^2 \sum_k X_{1jk}^2}{\bar{X}_{1j}^2 (n_{1j} - 1)}$$

3. **Obtain a truncated Z value for each cell, Z_j^* .** To limit the amount of cancellation that takes place between cell results during aggregation, cells whose results suggest possible favoritism are left alone. Otherwise the cell statistic is set to zero. This means that positive equivalent Z values are set to 0, and negative values are left alone. Mathematically, this is written as

$$Z_j^* = \min(0, Z_j).$$

4. **Calculate the theoretical mean and variance of the truncated statistic under the null hypothesis of parity, $E(Z_j^* | H_0)$ and $\text{Var}(Z_j^* | H_0)$.** In order to compensate for the truncation in step 3, an aggregated, weighted sum of the Z_j^* will need to be centered and scaled properly so that the final aggregate statistic follows a standard normal distribution.

- If $W_j = 0$, then no evidence of favoritism is contained in the cell. The formulae for calculating $E(Z_j^* | H_0)$ and $\text{Var}(Z_j^* | H_0)$ cannot be used. Set both equal to 0.
- If $\min(n_{1j}, n_{2j}) > 6$ for a mean measure, $\min\left\{a_{1j} \left(1 - \frac{a_{1j}}{n_{1j}}\right), a_{2j} \left(1 - \frac{a_{2j}}{n_{2j}}\right)\right\} > 9$ for a proportion measure, $\min(n_{1j}, n_{2j}) > 15$ and $n_j q_j (1 - q_j) > 9$ for a rate measure, or n_{1j} and n_{2j} are large for a ratio measure then

$$E(Z_j^* | H_0) = -\frac{1}{\sqrt{2\pi}}, \text{ and}$$

$$\text{Var}(Z_j^* | H_0) = \frac{1}{2} - \frac{1}{2\pi}.$$

- Otherwise, determine the total number of values for Z_j^* . Let z_{ji} and θ_{ji} , denote the values of Z_j^* and the probabilities of observing each value, respectively.

$$E(Z_j^* | H_0) = \sum_i \theta_{ji} z_{ji}, \text{ and}$$

$$\text{Var}(Z_j^* | H_0) = \sum_i \theta_{ji} z_{ji}^2 - [E(Z_j^* | H_0)]^2.$$

The actual values of the z's and θ 's depends on the type of measure.

Mean Measure

$$N_j = \min(M_j, 1,000), \quad i = 1, K, N_j$$

$$z_{ji} = \min\left\{0, \Phi^{-1}\left(1 - \frac{R_i - 0.5}{N_j}\right)\right\} \quad \text{where } R_i \text{ is the rank of sample sum } i$$

$$\theta_j = \frac{1}{N_j}$$

Proportion Measure

$$z_{ji} = \min\left\{0, \frac{n_j i - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}\right\}, \quad i = \max(0, a_j - n_{2j}), K, \min(a_j, n_{1j})$$

$$\theta_{ji} = \text{HG}(i)$$

Rate Measure

$$z_{ji} = \min\left\{0, \frac{i - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}}\right\}, \quad i = 0, K, n_j$$

$$\theta_{ji} = \text{BN}(i)$$

Ratio Measure

The performance measure that is in this class is billing accuracy. The sample sizes for this measure are quite large, so there is no need for a small sample technique. If one does need a small sample technique, then a resampling method can be used.

5. Calculate the aggregate test statistic, Z^T .

$$Z^T = \frac{\sum_j W_j Z_j^* - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)}}$$

Decision Process

Once Z^T has been calculated, it is compared to a critical value to determine if the ILEC is favoring its own customers over a CLEC's customers. The derivation of the critical value is found in Appendix C.

This critical value changes as the ILEC and CLEC transaction volume change. One way to make this transparent to the decision maker, is to report the difference between the test statistic and the critical value, $diff = Z^T - c_B$. If favoritism is concluded when $Z^T < c_B$, then the $diff < 0$ indicates favoritism.

This make it very easy to determine favoritism: a positive $diff$ suggests no favoritism, and a negative $diff$ suggests favoritism. Appendix D provides an example of how this information can be reported for each month.

Appendix B. Trunk Blocking

This Appendix provides an example of how the trunk blocking data can be processed to apply the Truncated Z Statistic. Trunk blocking is defined as the proportion of blocked calls a trunk group experiences in a time interval. It is a ratio of two numbers—blocked and attempted calls, both of which can vary over time and across trunk groups. Since the measure is a proportion where the numerator is a subset of the denominator, the truncated Z statistic, modified for proportions, can be applied here (see Appendix A).

As with other performance measures, data are first assigned to like-to-like cells, and the Z statistic is then computed within each cell. For trunk blocking, cells are defined by three variables: hour, day, and trunk group size or capacity. The next sections will describe the data and the data processing steps in greater detail.

The approach used in this example needs to be reviewed by subject matter expert to determine if it proper to use for trunk blocking.

Data Sources

Two data files are processed for the trunk blocking measure. One is the Trunk Group Data File that contains the Trunk Group Serial Number (TGSN), Common Language Location Identifier (CLLI), and other characteristics needed to categorize trunk groups and to identify them as BellSouth or CLEC.

The other file is the Blocking Data File (BDF), which contains the actual 24 hour blocking ratios for each weekday. There are 4 or 5 weeks in a monthly report cycle. The current system, however, allows the storage of daily blocking data by hour for a week only. Therefore, the data elements necessary to compute the Truncated Z must be extracted each week.

Two important data fields of interest on the Blocking Data File are the Blocking Ratio and Offered Load. The basic definition of Blocking Ratio is the proportion of all attempted calls that were blocked. For the simplest case of one way trunk groups, this is computed by dividing the number of blocked calls by the total call attempts, given that the data are valid. If they are not valid (e.g., actual usage exceeds capacity), blocking is estimated via the Neal Wilkinson algorithm.

Although the raw data--blocked calls (overflow) and peg counts (total call attempts)--are available, the calculation of the Blocking Ratio may be complicated for two-way trunk groups and trunk groups with invalid data. For this reason, we use the blocking ratios from the BDF instead of computing the ratios from the raw data. In order to reflect different call volumes processed through each trunk group, however, the blocking ratios need to be either weighted by call volume or converted to blocked and attempted calls before they are aggregated.

The measure of call traffic volume recommended for weighting is Offered Load. Offered Load is different from call counts in that it incorporates call duration as well. Since it is not just the number of calls but the total usage—number of calls multiplied by average call duration--that determines the occurrence of any blocking, this pseudo measure, Offered Load, appears to be the best indicator of call volume.

Cells or comparison classes are determined by three factors—hour, day, and trunk group capacity (number of trunks in service). The first two factors represent natural classes because trunk blocking changes over time. The third factor is based on our finding that high blocking tends to occur in small trunk groups. A pattern was found not only in the magnitude of blocking but also in its variability. Both the magnitude and variability of blocking decrease as trunk group capacity increases. Additional work is needed to establish the appropriate number of capacity levels and the proper location of boundaries.

Data Processing

The data are processed using the five steps below:

1. Merge the two files by TGSN and select only trunk groups listed in both files.
2. Reset the blocking of all high use trunk groups to zero¹.
3. Assign trunk group categories to CLEC and BellSouth: Categories 1, 3, 4, 5, 10, and 16 for CLEC and 9 for BellSouth². The categories used here for comparison are:

Category	Administrator	Point A	Point B
1	BellSouth	BellSouth End Office	BellSouth Access Tandem
3	BellSouth	BellSouth End Office	CLEC Switch
4	BellSouth	BellSouth Local Tandem	CLEC Switch
5	BellSouth	BellSouth Access Tandem	CLEC Switch
9	BellSouth	BellSouth End Office	BellSouth End Office
10	BellSouth	BellSouth End Office	BellSouth Local Tandem
16	BellSouth	BellSouth Tandem	BellSouth Tandem

4. Recode the missing data. The Blocking Data File assigns all missing data (no valid measurement data) zero blocking. To differentiate true zero blocking from zeroes due to missing data, invalid records were identified and the ratios reset to missing. The blocking value was invalid if both the number of Loaded Days and the Offered Load were 0 for a given hourly period.
5. Form comparison classes based either on the data (i.e., quartiles) or on a predetermined set of values.

¹ The high use trunk groups cannot have any blocking. These are set up such that all overflow calls are automatically routed to other trunk groups instead of being physically blocked.

² More detailed information on all categories is described in a report 'Trunk Performance Report Generation' by Ernst & Young (March 1999).

Calculation of the Proportion of Blocked Calls

Each cell is determined by day of the month, hour of the day, and trunk group capacity. To use the Truncated Z method, we generate summary information, to include the total number of blocked calls and the total number of attempted calls, for each cell.

For the details of each calculation step, the following notation is used. For a given hour of a day, let \bar{X}_{1j} be the proportion of BellSouth blocked calls for trunk group i in cell j and \bar{X}_{2j} be the corresponding proportion for CLEC. Then $\bar{X}_{1j} = X_{1ij} / n_{1ij}$ where X_{1ij} denotes the number of BellSouth blocked calls and n_{1ij} denotes the number of BellSouth total call attempts (indicated by Offered Load) for trunk group i in cell j. Likewise, $\bar{X}_{2j} = X_{2ij} / n_{2ij}$. For the steps outlined below, only the CLEC notation is provided.

1. Compute the number of blocked calls for trunk group i: $X_{2ij} = \bar{X}_{2j} * n_{2ij}$
2. Compute total call attempts for all trunk groups in the cell: $n_{2j} = \sum_i n_{2ij}$
3. Compute mean blocking proportion for cell j: $\bar{X}_{2j} = \sum_i X_{2ij} / \sum_i n_{2ij}$
4. Compute the total number of BellSouth and CLEC blocked calls in cell j: $t_j = \sum_i X_{1ij} + \sum_i X_{2ij}$
5. Apply the Truncated Z Statistic for Proportion measures presented in Appendix A.

Appendix C

Balancing the Type I and Type II Error Probabilities of the Truncated Z Test Statistic

This appendix describes a the methodology for balancing the error probabilities when the Truncated Z statistic, described in Appendix A, is used for performance measure parity testing. There are four key elements of the statistical testing process:

1. the null hypothesis, H_0 , that parity exists between ILEC and CLEC services
2. the alternative hypothesis, H_a , that the ILEC is giving better service to its own customers
3. the Truncated Z test statistic, Z^T , and
4. a critical value, c

The decision rule¹ is

- If $Z^T < c$ then accept H_a .
- If $Z^T \geq c$ then accept H_0 .

There are two types of error possible when using such a decision rule:

Type I Error: Deciding favoritism exists when there is, in fact, no favoritism.

Type II Error: Deciding parity exists when there is, in fact, favoritism.

The probabilities of each type of each are:

Type I Error: $\alpha = P(Z^T < c | H_0)$.

Type II Error: $\beta = P(Z^T \geq c | H_a)$.

In what follows, we show how to find a balancing critical value, c_B , so that $\alpha = \beta$.

General Methodology

The general form of the test statistic that is being used is

¹ This decision rule assumes that a negative test statistic indicates poor service for the CLEC customer. If the opposite is true, then reverse the decision rule.

$$z_0 = \frac{\hat{T} - E(\hat{T}|H_0)}{SE(\hat{T}|H_0)}, \quad (C.1)$$

where

\hat{T} is an estimator that is (approximately) normally distributed,

$E(\hat{T} | H_0)$ is the expected value (mean) of \hat{T} under the null hypothesis, and

$SE(\hat{T} | H_0)$ is the standard error of \hat{T} under the null hypothesis.

Thus, under the null hypothesis, z_0 follows a standard normal distribution. However, this is not true under the alternative hypothesis. In this case,

$$z_a = \frac{\hat{T} - E(\hat{T}|H_a)}{SE(\hat{T}|H_a)}$$

has a standard normal distribution. Here

$E(\hat{T} | H_a)$ is the expected value (mean) of \hat{T} under the alternative hypothesis, and

$SE(\hat{T} | H_a)$ is the standard error of \hat{T} under the alternative hypothesis.

Notice that

$$\begin{aligned} \beta &= P(z_0 > c | H_a) \\ &= P\left(z_a > \frac{cSE(\hat{T} | H_0) + E(\hat{T} | H_0) - E(\hat{T} | H_a)}{SE(\hat{T} | H_a)}\right) \end{aligned} \quad (C.2)$$

and recall that for a standard normal random variable z and a constant b , $P(z < b) = P(z > -b)$. Thus,

$$\alpha = P(z_0 < c) = P(z_0 > -c) \quad (C.3)$$

Since we want $\alpha = \beta$, the right hand sides of (C.2) and (C.3) represent the same area under the standard normal density. Therefore, it must be the case that

$$-c = \frac{cSE(\hat{T} | H_0) + E(\hat{T} | H_0) - E(\hat{T} | H_a)}{SE(\hat{T} | H_a)}.$$

Solving this for c gives the general formula for a balancing critical value:

$$c_B = \frac{E(\hat{T} | H_a) - E(\hat{T} | H_0)}{SE(\hat{T} | H_a) + SE(\hat{T} | H_0)} \quad (C.4)$$

The Balancing Critical Value of the Truncated Z

In Appendix A, the Truncated Z statistic is defined as

$$Z^T = \frac{\sum_j W_j Z_j^* - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)}}$$

In terms of equation (C.1) we have

$$\begin{aligned} \hat{T} &= \sum_j W_j Z_j^* \\ E(\hat{T} | H_0) &= \sum_j W_j E(Z_j^* | H_0) \\ SE(\hat{T} | H_0) &= \sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)} \end{aligned}$$

To compute the balancing critical value (C.4), we also need $E(\hat{T} | H_a)$ and $SE(\hat{T} | H_a)$. These values are determined by

$$\begin{aligned} E(\hat{T} | H_a) &= \sum_j W_j E(Z_j^* | H_a), \text{ and} \\ SE(\hat{T} | H_a) &= \sqrt{\sum_j W_j^2 \text{var}(Z_j^* | H_a)}. \end{aligned}$$

In which case equation (C.4) gives

$$c_B = \frac{\sum_j W_j E(Z_j^* | H_a) - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{var}(Z_j^* | H_a) + \sum_j W_j^2 \text{var}(Z_j^* | H_0)}}. \quad (C.5)$$

Thus, we need to determine how to calculate $E(Z_j^* | H_0)$, $\text{Var}(Z_j^* | H_0)$, $E(Z_j^* | H_a)$, and $\text{Var}(Z_j^* | H_a)$.

If Z_j has a normal distribution with mean μ and standard error σ , then the mean of the distribution truncated at 0 is

$$M(\mu, \sigma) = \int_{-\infty}^0 \frac{x}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx,$$

and the variance is

$$V(\mu, \sigma) = \int_{-\infty}^0 \frac{x^2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx - M(\mu, \sigma)^2$$

It can be shown that

$$M(\mu, \sigma) = \mu \Phi\left(\frac{-\mu}{\sigma}\right) - \sigma \phi\left(\frac{-\mu}{\sigma}\right)$$

and

$$V(\mu, \sigma) = (\mu^2 + \sigma^2) \Phi\left(\frac{-\mu}{\sigma}\right) - \mu \sigma \phi\left(\frac{-\mu}{\sigma}\right) - M(\mu, \sigma)^2$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function, and $\phi(\cdot)$ is the standard normal density function.

The cell test statistic, Z_j , is constructed so that it has mean 0 and standard deviation 1 under the null hypothesis. Thus,

$$E(Z_j^* | H_0) = M(0, 1) = -\frac{1}{\sqrt{2\pi}}, \text{ and}$$

$$\text{var}(Z_j^* | H_0) = V(0, 1) = \frac{1}{2} - \frac{1}{2\pi}.$$

The mean and standard error of Z_j under the alternative hypothesis depends on the type of measure and the form of the alternative. These are discussed below. For now, denote the mean and standard error of Z_j under the alternative by m_j and se_j respectively. Thus,

$$E(Z_j^* | H_a) = M(m_j, se_j), \text{ and}$$

$$\text{SE}(Z_j^* | H_a) = V(m_j, se_j).$$

Using the above notation, and equation (C.5), we get the formula for the balancing critical of Z^* .

$$c_H = \frac{\sum_j W_j M(m_j, se_j) - \sum_j W_j \frac{-1}{\sqrt{2\pi}}}{\sqrt{\sum_j W_j^2 V(m_j, se_j) + \sum_j W_j^2 \left(\frac{1}{2} - \frac{1}{2\pi}\right)}} \quad (C.6)$$

This formula assumes that Z_j is approximately normally distributed within cell j . When the cell sample sizes, n_{1j} and n_{2j} , are small this may not be true. It is possible to determine the cell mean and variance under the null hypothesis when the cell sample sizes are small. It is much more difficult to determine these values under the alternative hypothesis. Since the cell weight, W_j will also be small (see Appendix A) for a cell with small volume, the cell mean and variance will not contribute much to the weighted sum. Therefore, formula (C.6) provides a reasonable approximation to the balancing critical value.

Alternative Hypotheses

Mean Measure

For mean measures, one is concerned with two parameters in each cell, namely, the mean and variance. A possible lack of parity may be due to a difference in cell means, and/or a difference in cell variances. One possible set of hypotheses that capture this notion, and take into account the assumption that transactions are identically distributed within cells is:

$$H_0: \mu_{1j} = \mu_{2j}, \sigma_{1j}^2 = \sigma_{2j}^2$$

$$H_a: \mu_{2j} = \mu_{1j} + \delta_j \cdot \sigma_{1j}, \sigma_{2j}^2 = \lambda_j \cdot \sigma_{1j}^2 \quad \delta_j > 0, \lambda_j \geq 1 \text{ and } j = 1, \dots, L.$$

Under this form of alternative hypothesis, the cell test statistic Z_j has mean and standard error given by

$$m_j = \frac{-\delta_j}{\sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}, \text{ and}$$

$$se_j = \sqrt{\frac{\lambda_j n_{1j} + n_{2j}}{n_{1j} + n_{2j}}}$$

Proportion Measure

For a proportion measure there is only one parameter of interest in each cell, the proportion of transactions possessing an attribute of interest. A possible lack of parity may be due to a difference in cell proportions. A set of hypotheses that take into account

the assumption that transaction are identically distributed within cells while allowing for an analytically tractable solution is:

$$H_0: \frac{p_{2j}(1-p_{1j})}{(1-p_{2j})p_{1j}} = 1$$

$$H_a: \frac{p_{2j}(1-p_{1j})}{(1-p_{2j})p_{1j}} = \psi_j \quad \psi_j > 1 \text{ and } j = 1, \dots, L.$$

These hypotheses are based on the “odds ratio.” If the transaction attribute of interest is a missed trouble repair, then an interpretation of the alternative hypothesis is that a CLEC trouble is ψ_j times more likely to be missed than an ILEC trouble.

Under this form of alternative hypothesis, the within cell asymptotic mean and variance of a_{ij} are given by²

$$\begin{aligned} E(a_{ij}) &= n_j \pi_j^{(1)} \\ \text{var}(a_{ij}) &= \frac{n_j}{\frac{1}{\pi_j^{(1)}} + \frac{1}{\pi_j^{(2)}} + \frac{1}{\pi_j^{(3)}} + \frac{1}{\pi_j^{(4)}}} \end{aligned} \quad (C.7)$$

where

$$\begin{aligned} \pi_j^{(1)} &= f_j^{(1)} (n_j^2 + f_j^{(2)} + f_j^{(3)} - f_j^{(4)}) \\ \pi_j^{(2)} &= f_j^{(1)} (-n_j^2 - f_j^{(2)} + f_j^{(3)} + f_j^{(4)}) \\ \pi_j^{(3)} &= f_j^{(1)} (-n_j^2 + f_j^{(2)} - f_j^{(3)} + f_j^{(4)}) \\ \pi_j^{(4)} &= f_j^{(1)} \left(n_j^2 \left(\frac{2}{\psi_j} - 1 \right) - f_j^{(2)} - f_j^{(3)} - f_j^{(4)} \right) \\ f_j^{(1)} &= \frac{1}{2n_j^2 \left(\frac{1}{\psi_j} - 1 \right)} \\ f_j^{(2)} &= n_j n_{1j} \left(\frac{1}{\psi_j} - 1 \right) \\ f_j^{(3)} &= n_j a_j \left(\frac{1}{\psi_j} - 1 \right) \\ f_j^{(4)} &= \sqrt{n_j^2 \left[4n_{1j} (n_j - a_j) \left(\frac{1}{\psi_j} - 1 \right) + \left(n_j + (a_j - n_{1j}) \left(\frac{1}{\psi_j} - 1 \right) \right)^2 \right]} \end{aligned}$$

² Stevens, W. L. (1951) Mean and Variance of an entry in a Contingency Table. *Biometrika*, 38, 468-470.

Recall that the cell test statistic is given by

$$Z_j = \frac{n_j a_{1j} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}$$

Using the equations in (C.7), we see that Z_j has mean and standard error given by

$$m_j = \frac{n_j^2 \pi_j^{(1)} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}, \text{ and}$$

$$se_j = \sqrt{\frac{n_j^3 (n_j - 1)}{n_{1j} n_{2j} a_j (n_j - a_j) \left(\pi_j^{(1)} + \pi_j^{(2)} + \pi_j^{(3)} + \pi_j^{(4)} \right)}}$$

Rate Measure

A rate measure also has only one parameter of interest in each cell, the rate at which a phenomenon is observed relative to a base unit, e.g. the number of troubles per available line. A possible lack of parity may be due to a difference in cell rates. A set of hypotheses that take into account the assumption that transaction are identically distributed within cells is:

$$H_0: r_{1j} = r_{2j}$$

$$H_a: r_{2j} = \epsilon_j r_{1j} \quad \epsilon_j > 1 \text{ and } j = 1, \dots, L.$$

Given the total number of ILEC and CLEC transactions in a cell, n_j , and the number of base elements, b_{1j} and b_{2j} , the number of ILEC transaction, n_{1j} , has a binomial distribution from n_j trials and a probability of

$$q_j^* = \frac{r_{1j} b_{1j}}{r_{1j} b_{1j} + r_{2j} b_{2j}}.$$

Therefore, the mean and variance of n_{1j} , are given by

$$\begin{aligned} E(n_{1j}) &= n_j q_j^* \\ \text{var}(n_{1j}) &= n_j q_j^* (1 - q_j^*) \end{aligned} \tag{C.8}$$

Under the null hypothesis

$$q_j^* = q_j = \frac{b_{1j}}{b_j},$$

but under the alternative hypothesis

$$q_j^* = q_j^a = \frac{b_{1j}}{b_{1j} + \epsilon_j b_{2j}}. \quad (C.9)$$

Recall that the cell test statistic is given by

$$Z_j = \frac{n_{1j} - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}}.$$

Using (C.8) and (C.9), we see that Z_j has mean and standard error given by

$$m_j = \frac{n_j (q_j^a - q_j)}{\sqrt{n_j q_j (1 - q_j)}} = (1 - \epsilon_j) \sqrt{\frac{n_j b_{1j} b_{2j}}{b_{1j} + \epsilon_j b_{2j}}}, \text{ and}$$

$$se_j = \sqrt{\frac{q_j^a (1 - q_j^a)}{q_j (1 - q_j)}} = \sqrt{\epsilon_j} \frac{b_j}{b_{1j} + \epsilon_j b_{2j}}.$$

Ratio Measure

As with mean measures, one is concerned with two parameters in each cell, the mean and variance, when testing for parity of ratio measures. As long as sample sizes are large, as in the case of billing accuracy, the same method for finding m_j and se_j that is used for mean measures can be used for ratio measures.

Determining the Parameters of the Alternative Hypothesis

In this appendix we have indexed the alternative hypothesis of mean measures by two sets of parameters, λ_j and δ_j . Proportion and rate measures have been indexed by one set of parameters each, ψ_j and ϵ_j respectively. A major difficulty with this approach is that more than one alternative will be of interest; for example we may consider one alternative in which all the δ_j are set to a common non-zero value, and another set of alternatives in each of which just one δ_j is non-zero, while all the rest are zero. There are very many other possibilities. Each possibility leads to a single value for the balancing critical value; and each possible critical value corresponds to many sets of alternative hypotheses, for each of which it constitutes the correct balancing value.

The formulas we have presented can be used to evaluate the impact of different choices of the overall critical value. For each putative choice, we can evaluate the set of alternatives for which this is the correct balancing value. While statistical science can be used to evaluate the impact of different choices of these parameters, there is not much that an appeal to statistical principles can offer in directing specific choices. Specific choices are best left to telephony experts. Still, it is possible to comment on some aspects of these choices:

- Parameter Choices for λ_j . The set of parameters λ_j index alternatives to the null hypothesis that arise because there might be greater unpredictability or variability in the delivery of service to a CLEC customer over that which would be achieved for an otherwise comparable ILEC customer. While concerns about differences in the variability of service are important, it turns out that the truncated Z testing which is being recommended here is relatively insensitive to all but very large values of the λ_j . Put another way, reasonable differences in the values chosen here could make very little difference in the balancing points chosen.
- Parameter Choices for δ_j . The set of parameters δ_j are much more important in the choice of the balancing point than was true for the λ_j . The reason for this is that they directly index differences in average service. The truncated Z test is very sensitive to any such differences; hence, even small disagreements among experts in the choice of the δ_j could be very important. Sample size matters here too. For example, setting all the δ_j to a single value – $\delta_j = \delta$ – might be fine for tests across individual CLECs where currently in Louisiana the CLEC customer bases are not too different. Using the same value of δ for the overall state testing does not seem sensible. At the state level we are aggregating over CLECs, so using the same δ as for an individual CLEC would be saying that a "meaningful" degree of disparity is one where the violation is the same (δ) for each CLEC. But the detection of disparity for any component CLEC is important, so the relevant "overall" δ should be smaller.
- Parameter Choices for ψ_j or ϵ_j . The set of parameters ψ_j or ϵ_j are also important in the choice of the balancing point for tests of their respective measures. The reason for this is that they directly index increases in the proportion or rate of service performance. The truncated Z test is sensitive to such increases; but not as sensitive as the case of δ for mean measures. Sample size matters here too. As with mean measures, using the same value of ψ or ϵ for the overall state testing does not seem sensible.

The three parameters are related however. If a decision is made on the value of δ , it is possible to determine equivalent values of ψ and ϵ . The following equations, in conjunction with the definitions of ψ and ϵ , show the relationship with delta.

$$\delta = 2 \cdot \arcsin(\sqrt{\hat{p}_2}) - 2 \cdot \arcsin(\sqrt{\hat{p}_1})$$
$$\delta = 2\sqrt{\hat{r}_2} - 2\sqrt{\hat{r}_1}$$

The bottom line here is that beyond a few general considerations, like those given above, a principled approach to the choice of the alternative hypotheses to guard against must come from elsewhere.

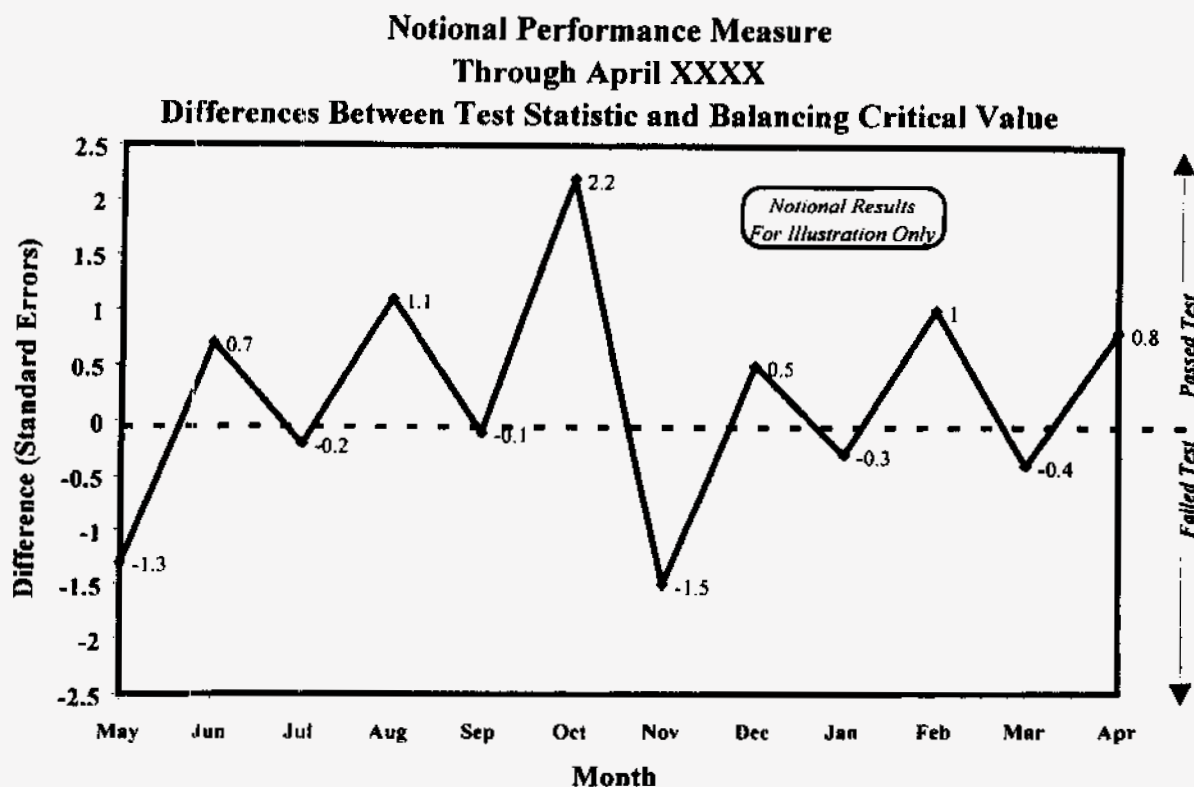
Appendix D: Examples of Statistical Reports

The general structure for reporting statistical results in a production environment will be the same for the different measures and we suggest that it consist of at least three components. For each measure present, (1) the monthly test statistics over a period of time, (2) the results for the current month, with summary statistics, test statistics, and descriptive graphs, and (3) a summary of any adjustments to the data made in the process of running the tests, including a description of how many records were excluded from analysis and the reason for the exclusion (i.e., excluded due to business rules, or due to statistical/methodological rules pertaining to the measure). The last component is important to assure that the reported results can be audited.

Selected components of the reporting structure are illustrated in the samples that follow. An outline of the report is shown below. Monthly results will be presented for each level of aggregation required.

- I. Test Statistics Over Time
- II. Monthly Results
 - A. Summary Statistics
 - B. Test Statistics
 - C. Descriptive Graphs (Frequency Distributions, etc.)
- III. Adjustments to Data
 - A. Records Excluded Due to Business Rules
 - B. Records Excluded Due to Statistical Rules

Test Statistic Over Time. The first component of the reporting structure is an illustration of the trend of the particular performance measure over time together with a tabular summary of results for the current month. We will show at a glance whether the tests consistently return non-statistically significant results; consistently indicate disparity (be that in favor of BellSouth or in favor of the CLECs); or vary month by month in their results. An example of this component follows.

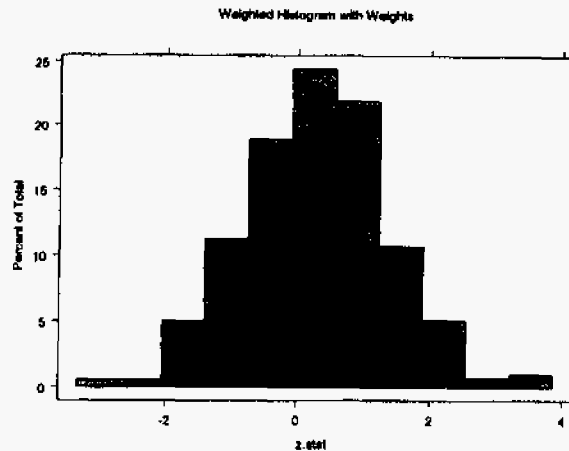
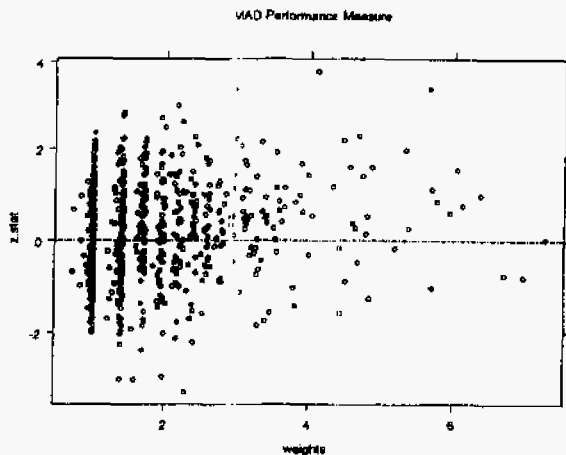


Result for Current Month	
Test Statistic	-0.410
Balancing Critical Value	-1.210
Difference	0.800

Monthly Results. The most important component of the reporting structure is the part which presents results of the monthly statistical tests on the given performance measure. The essential aspects included in this component are the summary statistics; the test statistics and results; and descriptive graphs of the results.

It is important to present basic summary statistics to complete the comparison between BellSouth and the CLECs. At a minimum, these statistics will include the means, standard deviations, and population sizes. In addition to basic descriptive statistics, we also present the test statistic results. Examples of ways we have presented these statistics in the past can be found in BellSouth's February 25, 1999 filing before the Louisiana Public Service Commission.

Finally, the results will be presented in graphical format. Below is an example of how to graphically present the data behind the Truncated Z statistic. One graph shows a plot of cell Z score versus cell weights. The other is a histogram of the weighted cell Z scores.



Adjustments to Data. The third important component of the reporting structure is information on any adjustments performed on the data. This information is essential in order that the results may be verified and audited. The most prevalent examples of such modifications would be removal of observations and weighting of the data.

Records can be removed from analysis for both business reasons (these will likely be taken into account in the PMAP system) and for statistical reasons. All of the performance measures exclude certain records based on business rules underlying each measure's particular definitions and methodologies. The number of records excluded for each rule will be summarized. In addition, some of the measures will have observations excluded for statistical reasons, particularly in the case of "mean measures" (OCI and MAD); these exclusions will be summarized as well. The tables below show examples of the current method for summarizing this information:

April XXXX Performance Measure Filtering Information																																											
This table displays information about the size of the database files and the cases that were removed from the analysis.																																											
<table border="1"> <tr><td>Unfiltered Total</td><td style="text-align: right;">28,691</td></tr> <tr><td>Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not resale and not UNE)</i></td><td style="text-align: right;">7,242</td></tr> <tr><td>Total Reported on Web Report</td><td style="text-align: right;">21,449</td></tr> <tr><td>Additional Records Removed for Business Reasons</td><td style="text-align: right;">876</td></tr> <tr><td> Missing Appointment code is 'S'</td><td style="text-align: right;">844</td></tr> <tr><td> General Class Service = 'O'</td><td style="text-align: right;">0</td></tr> <tr><td> UNE Cases</td><td style="text-align: right;">102</td></tr> <tr><td>Records Removed for Statistical Reasons</td><td></td></tr> <tr><td> Extreme Values Removed</td><td style="text-align: right;">9</td></tr> <tr><td>No Matching Classification Removals</td><td style="text-align: right;">47</td></tr> <tr><td>FILTERED TOTAL</td><td style="text-align: right;">20,517</td></tr> </table>	Unfiltered Total	28,691	Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not resale and not UNE)</i>	7,242	Total Reported on Web Report	21,449	Additional Records Removed for Business Reasons	876	Missing Appointment code is 'S'	844	General Class Service = 'O'	0	UNE Cases	102	Records Removed for Statistical Reasons		Extreme Values Removed	9	No Matching Classification Removals	47	FILTERED TOTAL	20,517	<table border="1"> <tr><td>Unfiltered Total</td><td style="text-align: right;">453,107</td></tr> <tr><td>Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not retail)</i></td><td style="text-align: right;">78,613</td></tr> <tr><td>Total Reported on Web Report</td><td style="text-align: right;">374,494</td></tr> <tr><td>Additional Records Removed for Business Reasons</td><td style="text-align: right;">7,429</td></tr> <tr><td> Missing Appointment code is 'S'</td><td style="text-align: right;">7,172</td></tr> <tr><td> General Class Service = 'O'</td><td style="text-align: right;">279</td></tr> <tr><td>Records Removed for Statistical Reasons</td><td></td></tr> <tr><td> Extreme Values Removed</td><td style="text-align: right;">652</td></tr> <tr><td>No Matching Classification Removals</td><td style="text-align: right;">21,974</td></tr> <tr><td>FILTERED TOTAL</td><td style="text-align: right;">344,439</td></tr> </table>	Unfiltered Total	453,107	Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not retail)</i>	78,613	Total Reported on Web Report	374,494	Additional Records Removed for Business Reasons	7,429	Missing Appointment code is 'S'	7,172	General Class Service = 'O'	279	Records Removed for Statistical Reasons		Extreme Values Removed	652	No Matching Classification Removals	21,974	FILTERED TOTAL	344,439
Unfiltered Total	28,691																																										
Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not resale and not UNE)</i>	7,242																																										
Total Reported on Web Report	21,449																																										
Additional Records Removed for Business Reasons	876																																										
Missing Appointment code is 'S'	844																																										
General Class Service = 'O'	0																																										
UNE Cases	102																																										
Records Removed for Statistical Reasons																																											
Extreme Values Removed	9																																										
No Matching Classification Removals	47																																										
FILTERED TOTAL	20,517																																										
Unfiltered Total	453,107																																										
Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not retail)</i>	78,613																																										
Total Reported on Web Report	374,494																																										
Additional Records Removed for Business Reasons	7,429																																										
Missing Appointment code is 'S'	7,172																																										
General Class Service = 'O'	279																																										
Records Removed for Statistical Reasons																																											
Extreme Values Removed	652																																										
No Matching Classification Removals	21,974																																										
FILTERED TOTAL	344,439																																										

Appendix E. Trimming Outliers for Mean Measures

The arithmetic average is extremely sensitive to outliers; a single large value, possibly an erroneous value, can significantly distort the mean value. And by inflating the error variance, this also affects conclusions in the test of hypotheses. Extreme data values may be correct, but since they are rare measurements, they may be considered to be statistical outliers. Or they may be values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

At this time, only two mean measures have been analyzed: Order Completion Interval and Maintenance Average Duration. Maintenance Average Duration data are truncated at 240 hours and therefore this measure was not trimmed further. For Order Completion Interval, the underlying distribution of the observations is clearly not normal, but rather skewed with a very long upper-tail.

A useful technique, coming from the field of robust statistical analysis, is to trim a very small proportion from the tails of the distribution before calculating the means. The resulting mean is referred to as a trimmed mean. Trimming is beneficial in that it speeds the convergence of the distribution of the means to a normal distribution. Only extreme values are trimmed, and in many cases the data being trimmed are, in fact, data that might not be used in the analysis on other grounds.

In the first analysis of the verified Order Completion Interval-Provisioning measure, after removing data that were clearly in error or were not applicable, we looked at the cases that represented the largest 0.01% of the BST distribution. In the August data, this corresponded to orders with completion intervals greater than 99 days. All of these were BellSouth orders. In examining the largest 11 individual examples that would be removed from analysis, we found that only 1 of the 11 cases was a valid case where the completion interval was unusually large. The other 10 cases were examples of cases that should not have been included in the analysis. This indicates that at least in preliminary analysis, it is both beneficial to examine the extreme outliers and reasonable to remove them.

A very slight trimming is needed in order to put the central limit theorem argument on firm ground. But finding a robust rule that can be used in a production setting is difficult. Also, any trimming rule should be fully explained and any observations that are trimmed from the data must be fully documented.

When it is determined that a measure should be trimmed, a trimming rule that is easy to implement in a production setting is:

Trim the ILEC observations to the largest CLEC value from all CLEC observations in the month under consideration.

That is, no CLEC values are removed; all ILEC observations greater than the largest CLEC observation are trimmed.

While this method is simple, it does allow for extreme CLEC observations to be part of the analysis. For instance, suppose that the amount of time to complete an order was less than 40 days for all CLEC orders except one. Let's say that this extreme order took 100 days to complete. The trimming rule says that all ILEC orders above 100 days should be trimmed, but a closer look at the data might suggest trimming at 40 days instead.

Since we are operating in a production mode system, it is not possible to explore the data before the trimming takes place. Other automatic trimming rules present other problems, so our solution is to use the simple trimming rule above, and have the system automatically produce a trimming report that can be examined at a later point in time.

The trimming report should include:

- The value of the trim point.
- Summary statistics and graphics of the ILEC observations that were trimmed.
- A listing of the trimmed ILEC transaction for a random sample of 10 trimmed transactions. This listing should not disclose sensitive information.
- A listing of the 10 most extreme CLEC transactions. This listing should not disclose sensitive information.
- The number of ILEC and CLEC observations above some fixed point, so that changes in the upper tail can be better tracked over time.

The trimming report should be part of the overall report discussed in Appendix D. Examples of tables contained within the trimming report are shown below.

April XXXX
Performance Measure Extreme Values

CLEC		BST	
Cutoff	26	Cutoff	26
# of Records	20,573	# of Records	367,065
10 Largest		Extreme Values	652
Minimum	19	Minimum	27
Median	23	Median	32
Maximum	26	Maximum	283
Subtotal	20,573	Subtotal	366,413

April XXXX
Performance Measure Weighing Report

CLEC		BST	
# of Records	20,573	# of Records	366,413
No Matching BST		No Matching CLEC	
Classification (1)	47	Classification (2)	21,974
Subtotal	20,526	Subtotal	344,439

**April XXXX
Perormance Measure Filtering Information**

This table displays information about the size of the database files and the cases that were removed from the analysis.

	1999		1999
Unfiltered Total	28,681	Unfiltered Total	453,107
Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not resale and not UNE)</i>	7,242	Records Removed for Business Reasons <i>(e.g. not N, T, C, or P orders, not retail)</i>	78,613
Total Reported on Web Report	21,449	Total Reported on Web Report	374,494
Additional Records Removed for Business Reasons	876	Additional Records Removed for Business Reasons	7,429
Missing Apointment code is 'S'	844	Missing Appointment code is 'S'	7,172
General Class Service = 'O'	0	General Class Service = 'O'	278
UNE Cases	102		
Records Removed for Statistical Reasons		Records Removed for Statistical Reasons	
Extreme Values Removed	0	Extreme Values Removed	652
No Matching Classification Removals	47	No Matching Classification Removals	21,874
FILTERED TOTAL	20,526	FILTERED TOTAL	344,439

CLEC Extreme Values

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Order Interval
NWORLAMA	1	1	3	1	N	61
OPLSLATI	1	2	1	1	C	53
NWORLAMA	2	1	3	1	N	44
NWORLAMA	1	1	3	1	N	39
BTRGLAWN	1	1	2	1	C	38
LKCHLADT	1	1	1	1	T	37
NWORLAMA	1	1	3	1	N	32
NWORLAMA	2	1	3	1	N	32
SHPTLACL	1	1	2	1	N	28

Frequency of Extreme Values Removed from BST file (Top 10)

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Frequency
NWORLAMA	1	1	3	1	N	55
NWORLAMA	2	1	3	1	N	25
BTRGLASE	2	1	3	1	C	23
NWORLAMC	2	1	3	1	C	23
NWORLAMC	1	1	3	1	C	22
NWORLAMA	2	1	3	1	C	18
NWORLAMA	1	1	3	1	C	17
BTRGLASR	1	1	3	1	C	16
LEYTLAMA	1	1	3	1	C	15
NWORLAMA	2	2	3	1	C	14

An Adjusted, Asymmetric Two Sample t-Test

Sandy D. BALKIN and Colin L. MALLOWS

We present an asymmetric version of the two-sample t-test which is adjusted for distribution skewness and that is sensitive to alternatives where one of the population variances may have increased. The need for such a statistic has arisen in testing for parity of service.

KEY WORDS: Cornish-Fisher, Performance Measure, Permutation Test, Skewness

1. INTRODUCTION

The Telecommunications Act of 1996 mandates that Incumbent Local Exchange Carriers (ILECs) must provide, if requested, for a fair price, interconnection services to the customers of a Competitive Local Exchange Carrier (CLEC), these service being

. . .at least equal in quality to [those] provided by the local exchange carrier to itself. . .

Providing services to customers of a competitor imposes a clear conflict of interest on the ILEC; to monitor the ILEC's performance, we need to establish formal statistical procedures to test whether it is in compliance. In successive reporting periods, observations are made of the ILEC's performance for its own customers (X 's) and for the CLEC's customers (Y 's). A typical measurement is the time it takes to respond to a request for a new installation. From the point of view of the incoming CLEC, the alternatives to the null *compliance* hypothesis that are of most concern are those in which either $E(Y) > E(X)$ or $V(Y) > V(X)$, since in both cases, several CLEC customers will be getting

Dr. Balkin is a statistical consultant with Ernst & Young LLP, 1225 Connecticut Avenue, NW, Washington, DC 20036 and Dr. Mallows is with AT&T Labs - Research, 180 Park Avenue, Florham Park, New Jersey 07932. The authors wish to thank William Stacy and Jerry Moore of BellSouth for allowing us to include real performance measure data in this paper. This first author also wishes to thank J. Keith Ord of Georgetown University for some useful discussions on this topic.

worse service than the typical ILEC customer. For the purpose of making this assessment, this article seeks to correct the two-sample t -statistic for the bias caused by skewness of the population distributions by adjusting the statistic using properties of the data. We show empirically that the adjusted version of the two-sample t -statistic is a better approximation to the permutation distribution than the unadjusted version.

This article is organized as follows. Section 2 presents the asymmetric t -statistic which has been shown to have substantially better power than the pooled t -statistic when the homogeneity of variance assumption is violated. Section 3 presents new adjustments to the two-sample asymmetric and pooled t -statistics for the bias caused by skewness in the population distributions. These adjustments are derived using the Cornish-Fisher expansion. We then conclude with some observations and comments on the adjustment.

2. ASYMMETRIC T -STATISTIC

Given two samples X_1, \dots, X_m and Y_1, \dots, Y_n , from populations F_X and F_Y with means $E(X), E(Y)$ and variances $V(X), V(Y)$ respectively, we can consider the usual null hypothesis

$$H_0 : F_X = F_Y. \tag{1}$$

The preferred choice for evaluating this kind of hypothesis is a permutation test. However, permutation tests are computationally impractical in this situation as there is a need to perform thousands of such hypothesis tests and report the overall results within a short amount of time. Even when coded efficiently, when sample sizes are moderately large, the computations are unwieldy. If the populations are approximately normal, we can consider using the t -test. The t -statistic is of the form

$$t = \frac{\bar{X} - \bar{Y}}{S} \tag{2}$$

where for the usual statistic t_{pooled}

$$S^2 = S_{pooled}^2 \left(\frac{1}{m} + \frac{1}{n} \right) \tag{3}$$

and

$$S_{pooled}^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \tag{4}$$

with S_X^2 and S_Y^2 being the two sample variances. The corresponding one-sided test, using a critical value $t_{m+n-2}(\alpha)$, has size α , and, if normality holds, has optimal power against the standard alternative

$$H_A : E(Y) > E(X), \quad V(Y) = V(X). \quad (5)$$

However, it does not have optimal power against alternatives in which $V(Y)$ may be larger than $V(X)$:

$$H_A : E(Y) = E(X) + \delta, \quad V(Y) = \lambda V(X) \quad (6)$$

for $\delta \geq 0$, $\lambda > 1$. Note that one way this can happen is if each Y is independently shifted with probability p by some random amount W so that:

$$\begin{aligned} E(Y) &= E(X) + pE(W) \\ V(Y) &= V(X) + p(1-p)E(W)^2 + pV(W) \end{aligned}$$

Consider an asymmetric version of the t -statistic, namely $t_{asymmetric}$, which uses

$$S_{asymmetric}^2 = S_X^2 \left(\frac{1}{m} + \frac{1}{n} \right). \quad (7)$$

The statistic $t_{asymmetric}$ sacrifices some degrees of freedom, but if m is not very small, it has better power than t_{pooled} for the alternatives in (6). Brownie, et. al (1990) propose the $t_{asymmetric}$ test for use in a randomized experiment. Their paper compares the power of the modified test with that of the standard pooled test, showing that, as expected, the modified test is more powerful for alternatives where the variance has increased. The paper also compares these two tests with the Welch test that uses

$$S_{Welch}^2 = S_X^2/m + S_Y^2/n \quad (8)$$

and shows that this can have much smaller power than either the standard t_{pooled} or the modified $t_{asymmetric}$. This is particularly so in the case of most interest to us, namely $n \ll m$.

Both the ILECs and CLECs would prefer to use a permutation test of the hypothesis, but realize that the computational burden is excessive. Both would welcome an approximation, such as the t -statistic, as long as it performs similarly to the permutation test, but that can be quickly calculated from summary statistics. Figure 1 is a quantile-quantile plot of the permutation z-scores versus $t_{asymmetric}$ statistics converted to z-scores for samples of a specific performance measure with both group sample sizes greater than six. For this performance measure, the ILEC sample

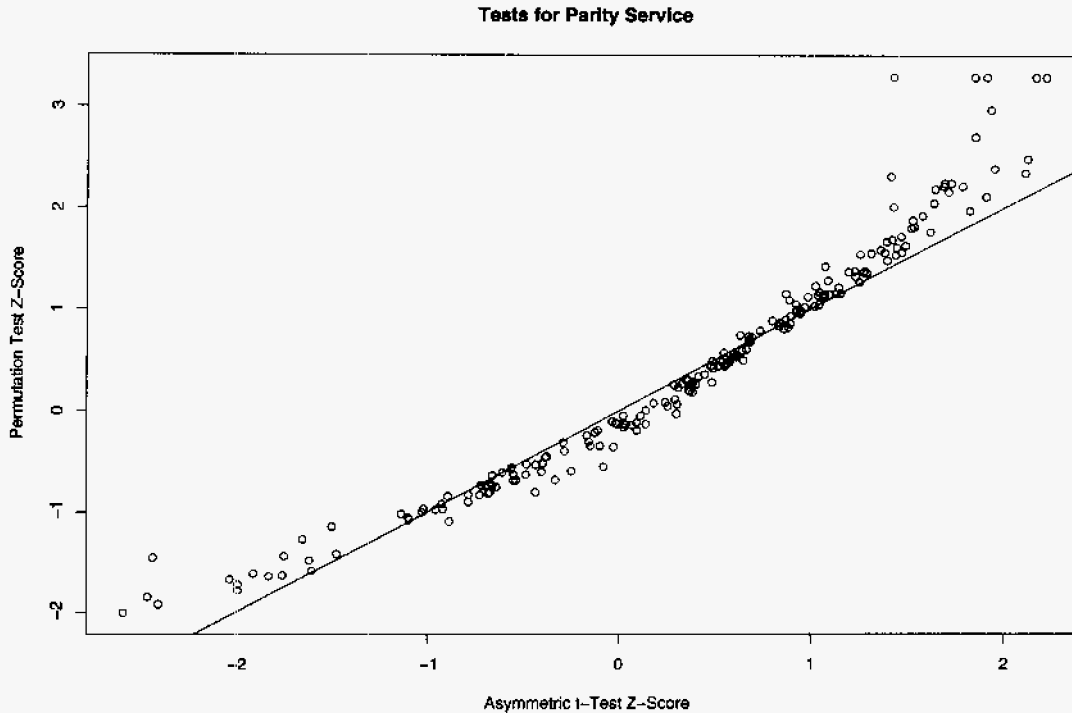


Figure 1. Scatterplot of Permutation Z-Scores versus Asymmetric T-test Z-Scores.

sizes have a mean of 152 and a maximum of 1,488 compared with a mean sample size of 13 and maximum of 57 for the CLECs. We would expect to see the points fall close to the 45-degree line. However, we see that there appears to be some quadratic structure in the plot. Thus, for the Asymmetric t -test to be considered as a viable alternative to permutation testing, it must be adjusted for this curvature.

3. ADJUSTING THE T -STATISTIC FOR SKEWNESS

Let γ_X and γ_Y be the skewness parameters of the ILEC and CLEC populations respectively. Based on rather fragmentary evidence, we assume $\gamma_X \approx \gamma_Y$. If this is the case, and the sample sizes of the two populations are approximately equal, the skewness effects cancel out in the numerator. However, in our situation, the ILEC sample size tends to be significantly larger than that of the CLEC. Thus, we explore the possibility of adjusting the $t_{asymmetric}$ statistic for skewness.

Johnson (1978) derived a skewness adjustment for the one-sample t -test. Following his method exactly, we can derive adjustments for the two-sample tests based on the $t_{asymmetric}$ and t_{pooled} statistics. The modification of the t -statistic

obtained in this study uses the Cornish-Fisher expansion

$$CF(X) = \mu + \sigma\zeta + (\mu_3/\sigma^2)(\zeta^2 - 1) + \dots, \quad (9)$$

where ζ is a standard normal random variable, μ is the mean of X , and σ^2, μ_3, \dots are the second, third, ... central moments of X respectively. Let the modified t -statistic take the form

$$t_{adj} = t + \lambda + \gamma t^2 \quad (10)$$

where t is given by (2) and S as defined in (7).

The Cornish-Fisher expansion of the numerator of the $t_{asymmetric}$ statistic is given by

$$CF(\bar{X} - \bar{Y}) = \sigma \left[\sqrt{\frac{1}{m} + \frac{1}{n}} \zeta + \frac{\gamma_1}{6} \frac{1/m^2 - 1/n^2}{1/m + 1/n} (\zeta^2 - 1) \right] \quad (11)$$

and for the denominator as given in Johnson (1978) by

$$CF(S_X) = \sigma_X^2 (1 + \sqrt{(\gamma_2/m)\eta}) \quad (12)$$

where $\gamma_2 = (\mu_4 - \sigma^4)/\sigma^4$. Also, the covariance of $\bar{X} - \bar{Y}$ and S_X^2 is μ_3/m so the correlation between ζ and η is $\sqrt{\frac{n}{m+n}} \gamma_1 / \sqrt{\gamma_2}$.

Plugging in the expansion terms and choosing λ and γ to cancel the terms of order $n^{-1/2}$, we get

$$t_{adj} = t_{asymmetric} + \frac{g}{6} \frac{m + 2n}{\sqrt{mn(m+n)}} (t_{asymmetric}^2 + \frac{n-m}{m+2n}) \quad (13)$$

where g is an estimate of the standardized third moment γ_1

$$g = \frac{1}{mS_X^3} \sum_i (X_i - \bar{X})^3 \quad (14)$$

calculated from some of the larger samples as these better define the parameter value. For the example given in this paper, g is taken to be two.

As a check, as n gets large, t_{adj} converges to the result for a one-sample test given in Johnson (1978). As m gets large,

$$t_{adj} \rightarrow t_{asymmetric} + \frac{g}{6\sqrt{n}} (t_{asymmetric}^2 - 1). \quad \text{The American Statistician, ??? (15)}$$

Note that the correction does not vanish when the sample sizes are equal as the statistic is not symmetric between samples.

To achieve a monotone and invertible transformation, we need to bound the adjustment. This will ensure that we fall on the correct side of the parabola and that the correction is in the appropriate direction. The minimum value of the adjusted t -statistic, called t_{min} is give by solving the equation

$$\frac{\partial t_{adj}}{\partial t} = 0 \tag{16}$$

for t giving

$$t_{min} = -3\sqrt{mn(m+n)}/(g(m+2n)) \tag{17}$$

Thus, if $t_{asymmetric} \geq t_{min}$, we use (13). If $t_{asymmetric} < t_{min}$ we use

$$t_{adj} = t_{asymmetric} + \frac{g}{6} \frac{m+2n}{\sqrt{mn(m+n)}} \left(t_{min}^2 + \frac{n-m}{m+2n} \right) \tag{18}$$

We see in Figure 2 that the quadratic structure of the z -scores has been adjusted for and that the adjusted points are closer to the 45 degree line.

Similarly, for the pooled t -statistic, by substituting the corresponding Cornish-Fisher expansion of each of \bar{X}, \bar{Y}, S_X^2 , and S_Y^2 into Equation (10), we get

$$t_{adj} = t_{pooled} + \frac{(m-n)g}{6\sqrt{mn(m+n)}} (t_{pooled}^2 - 1) \tag{19}$$

where the minimum of the adjustment value is

$$t_{min} = -3\sqrt{mn(m+n)}/(g(m-n)). \tag{20}$$

Note that when the sample sizes are the same, the adjustment vanishes.

4. CONCLUSION

In order to comply with the the Telecommunications Act of 1996, ILEC firms require a statistical test of parity. The data collected for performance measures often violate the usual assumptions. We find that the data tend to be positively skewed and have very different sample sizes. Nonparametric methods that are traditionally used instead of

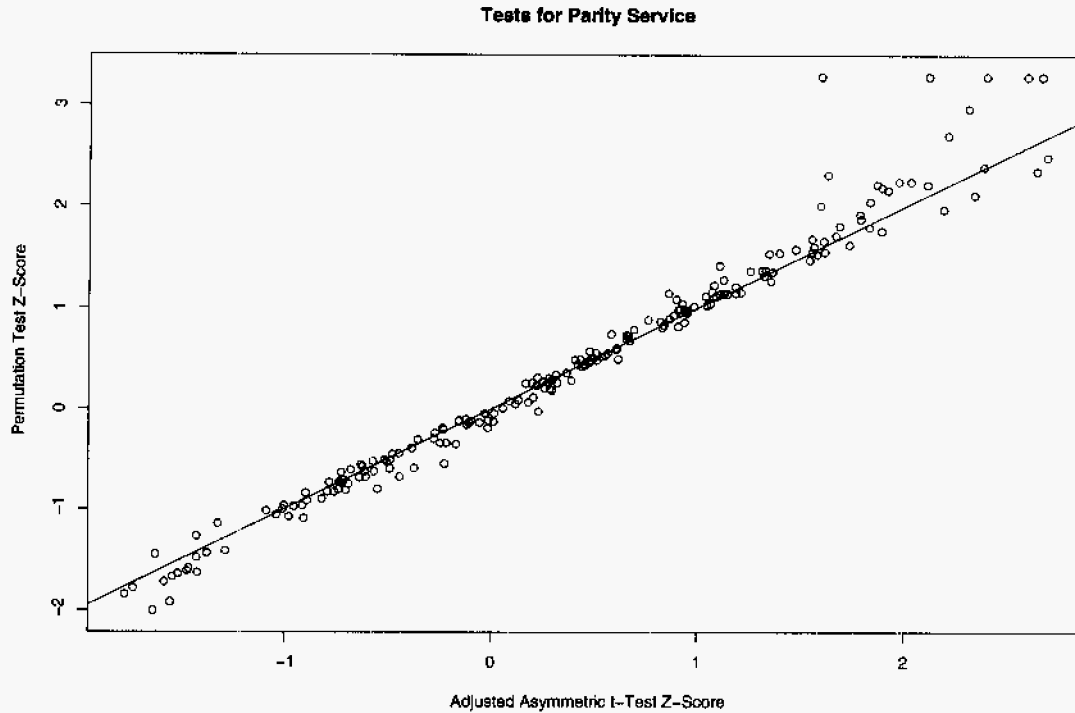


Figure 2. Scatterplot of Permutation Z-Scores versus Adjusted T-test Z-Scores.

the pooled t -test are also found to be inappropriate given the characteristics of the data. Permutation observations is a viable alternative only when the number of tests are small and are hence computationally prohibitive in our situation.

In response to the need for a computationally quick test that corrects for population skewness, we developed an asymmetric version of the two-sample t -test which is adjusted for skewness and that is sensitive to alternatives where one of the population variances may have increased. We show graphically that the adjustments made to the $t_{asymmetric}$ statistic provides results similar to those obtained from permutation tests.

REFERENCES

- Brownie, C., Boos, D., Hughes-Oliver, J., (1990), "Modifying the t and ANOVA F Tests When Treatment is Expected to Increase Variability Relative to Controls," *Biometrics*, 46, 259-266.
- Johnson, Norman J., (1978), "Modified t Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73:363, 536-544.