Edward J. Mulrow

BELLSOUTH TELECOMMUNICATIONS, INC.

DIRECT TESTIMONY OF EDWARD J. MULROW, PH.D.

BEFORE THE TENNESSEE REGULATORY AUTHORITY

DOCKET NO. 97-00309

APRIL 26, 2002

Q. PLEASE STATE YOUR NAME, WHO YOU WORK FOR, AND YOUR BUSINESS ADDRESS.

A. My name is Edward J. Mulrow. I am employed by Ernst & Young LLP as a Senior Manager in the Quantitative Economics and Statistics Group. I have been retained by BellSouth Telecommunications, Inc. ("BellSouth") as a statistical advisor. My business address is 1225 Connecticut Ave., NW, Washington, DC 20036.

Q. WHAT IS YOUR PROFESSIONAL EXPERIENCE AND EDUCATIONAL BACKGROUND?

A. My career as a statistical consultant spans over 14 years. While at Ernst & Young, I have been involved in a number of regulatory issues for several telecommunications companies. Prior to my employment at Ernst & Young, I was a senior scientist at Science Applications International Corporation (SAIC) where I was involved in the analyses of current and future defense systems. I also have worked as a senior sampling statistician at the National Opinion Research Center (NORC) at the University of Chicago, a mathematical statistician for the Internal Revenue Service, and an assistant professor of mathematics for Southern Illinois University. I received a BA in mathematics from Illinois Wesleyan University, an MS in mathematics from the University of Utah, and a Ph.D. in statistics from Colorado State University.

Q. WHAT IS THE PURPOSE OF YOUR TESTIMONY?

A. I am here to address statistical issues relevant to the proposed penalty plan. Mr. Varner addresses the proposed performance measures to which the penalty plan applies. I will speak to issues involving the appropriate methodology for determining whether BellSouth is providing parity: 1) to individual Competitive Local Exchange Carriers ("CLECs") (Tier I), and 2) to the CLEC community as a whole (Tier II).

Q. PLEASE SUMMARIZE YOUR TESTIMONY.

A. The appropriate methodology for determining whether or not BellSouth is providing parity service to a CLEC's customers consists of the following three elements.

1. The Truncated Z statistical test should be used when transaction level data is available and a BellSouth retail analog exists.
2. The statistical testing methodology should balance Type I and Type II error probabilities.
3. The same methodology should be used for both Tier I and Tier II testing.

I will address each of these points in more detail in my testimony.

Q. CAN YOU PROVIDE A BRIEF OVERVIEW OF WHAT WE ARE TRYING TO ACCOMPLISH WITH THE STATISTICAL ANALYSIS THAT YOU ARE GOING TO DESCRIBE IN YOUR TESTIMONY?

A. Yes. What we are talking about here is the situation where BellSouth provides a service of some sort to its competitors, the CLECs. BellSouth also, at the same time, is providing a similar, or at least an analogous service, to its own retail operations. The question is whether BellSouth is favoring its retail operations in the provision of the particular service, or whether it is providing the same level of service to its competitors as its provides to itself.

For instance, assume that CLECs purchased widgets from BellSouth and BellSouth also provided widgets to its own retail operations which then used the widgets to provide service to BellSouth's own retail customers. If BellSouth provided the widgets to the CLECs on a two-day interval every time, and provided the widgets to its own retail operations on a two-day interval every time, then anyone could conclude that BellSouth was providing parity to the CLECs.

Similarly, if BellSouth were furnishing the widgets to the CLECs on a one-day interval, and furnishing the widgets to its own retail operations in two days, it would be evident that BellSouth wasn't providing parity, but was providing better service to the CLECs than to its own retail operations. Presumably the CLECs would not be upset with that.

The problem arises when BellSouth, in a given month, provides the widgets to its retail operations on average in two days, and provides widgets to the CLECs, on average, in 2.2 days. The question is whether the difference is attributable to random chance, or whether the difference is attributable to either some systemic problem with BellSouth's operations or some intentional act on BellSouth's part. The purpose of the statistical analysis is to provide the tools that the Authority can use to make an informed judgment about whether the difference I just described is something to be concerned about or rather is simply the result of the sample used and therefore meaningless. The specific tool that I am going to describe in my testimony is a test that can be applied whenever the Tennessee Regulatory Authority ("Authority") wishes to compare two outcomes to determine whether any perceived difference in the outcomes is real or not. While the test is a statistical one, and involves statistical concepts, I believe that what we have is very workable and understandable.

Q. WHAT IS THE APPROPRIATE STATISTICAL METHODOLOGY THAT SHOULD BE EMPLOYED TO DETERMINE IF BELLSOUTH IS PROVIDING COMPLIANT PERFORMANCE?

A. The appropriate methodology to use is called the Truncated Z method with error probability balancing. Dr. Colin Mallows, a retired statistician from AT&T Research Labs, created the Truncated Z statistic, and then Dr. Mallows together with Ernst & Young statisticians, including myself, developed the actual Truncated Z methodology. The methodology is distinguished from the statistic in that we jointly took Dr. Mallows' formula that yielded the statistic and complemented it with such things as the error probability balancing. The collaborative effort was the result of a request by the Louisiana Public Service Commission (LPSC), lasted over nine months, and concluded in the filing of a "Statisticians' Report" with the LPSC in September of 1999 (revised February 2000 -- attached as Exhibit No. EJM-1).[1]

[1] Typographical error corrections are attached as Exhibit No. EJM-2.

Q. CAN YOU EXPLAIN IN LAYMAN'S TERMS WHAT THE TRUNCATED Z METHODOLOGY DOES?

A. I can. Remember that what we are doing is comparing two outcomes to see if there is any difference. Therefore, one of the first things that must be done is to separate all of our observations into identical, or substantially identical categories. For instance, let's assume that what we are trying to compare is the performance of BellSouth with regard to order completion intervals. That is, we want to know whether the order completion intervals for BellSouth's retail operations are statistically the same as the order completion intervals ("OCI") for the CLECs. You would not want to compare a BellSouth retail residential order that requires a dispatch with a CLEC resale residential order that did not require a dispatch. The requirements for provisioning the different orders would be different.

Obviously, you can carry this concept of granularity to an extreme, but the point is that the first thing we have to do is to separate the individual observations into enough categories so that the comparison we are going to make is as close to being an apples-to-apples comparison as we can reasonably get it.

In our work, we call these classifications "cells." For any particular measurement contained in the BellSouth plan, there could be thousands of these "cells." Once we have these cells identified and populated with observations, we apply statistical tests to the information in the cells to put the conclusions we draw about every cell on a common footing. To make this illustration as clear as possible, I will assume that I have a cell for residential dispatched orders during the first half of the month. For illustrative purposes, I will assume that BellSouth has one observation that took 2 days, and the CLECs had a single observation that took 2.2 days, the times I used above. We would then apply a statistical calculation to those two observations, as is described in Appendix A of Exhibit EJM-1, and we would derive a "cell z-value" of -0.67. The calculation of this value is not subject to a simple explanation, but is done through standard statistical analysis with which no statistician should disagree. Obviously, as the number of observations in the cell increases, the "cell z-value" may change.

I have described briefly what we would do for the individual cell. In actuality, we would make this same type of calculation for every cell (or more plainly stated, for each of the apples-to-apples comparisons that we had identified in connection with the specific measurement).

Q. WHAT HAPPENS NEXT?

A. When we are done, we would have a large number, potentially thousands of numbers, each representing the "cell z-value" for each individual cell. The "cell z-values" would be either positive, or negative, or in some cases would be zero. The cells that have a negative "cell z-value" would represent those cells where, continuing my example from above, it appears that the interval for the CLECs was longer than for BellSouth. The cells that had a positive "cell z-value" would represent those cells where, again continuing my example, it appears that the interval for the CLECs was shorter than for BellSouth. Where the "cell z-value" was zero, there would be no apparent difference in the intervals.

Q. WHAT DO YOU DO WITH THESE THOUSANDS OF "CELL Z-VALUES?"

A. We move to the next step in the analysis, which is to analyze the "cell z-values" using a normal distribution curve. If BellSouth were providing parity, one would expect that the distribution of the values over the entire range of the cells would look just like the normal bell curve with which we should all be familiar.

This is where the idea of "truncating" the z statistic comes into play. We have z statistics for every cell. Some are positive, meaning they fall on the right side of the normal bell curve. Some are negative, which means that they are on the left side of the normal bell curve. One concern we would have is that if all of the z-values were included in the analysis, the positive z-values, if there were enough of them, might mask one or more significant negative z-values when averaging the z-values across all cells. That is, if there were a thousand cells, and 800 of them had positive z statistics, the sheer number of positive observations might hide significant negative values. Therefore, in order to prevent this, the Truncated Z methodology simply sets every positive value to zero, hence the "truncation." By setting the positive observation to zero, it forces us to concentrate on the negative values on the left side of the bell shaped curve.

Q. WHAT DO YOU DO NEXT?

A. Remember we are now only concentrating on the lower half of the normal bell-shaped curve, and what we are going to try to do, in layperson's terms, is to determine how far the observations we have made fall from the normal bell curve I have been talking about. You would not expect the observations to lie down perfectly on the curve. There are going to be variations and the question is how much is too much. Consequently, the next step is to calculate a Z statistic for all the cells, including those formally positive cells whose value has now been set to zero. Assuming that a statistician understood the purpose of truncating the positive values, and the selection of the cells weights, the calculation of the Z statistic for the truncated observations (the positive ones set to zero and the remaining negative observations left as they were found) should not be subject to dispute. This calculation will leave you with a single number that represents the truncated Z statistic value for the particular measurement contained in BellSouth's plan for which the observations were made.

Q. DOES THIS CALCULATED Z STATISTIC BY ITSELF REPRESENT A STATISTICALLY SIGNIFICANT DIFFERENCE IN THE PERFORMANCE BELLSOUTH PROVIDED TO ITS RETAIL OPERATIONS AND THE CLECS?

A. No, generally you can't draw any conclusion from the Z statistic itself. It is just a number. However, if the number turns out to be positive (which, even though it seems illogical because of changing the positive values to zero, could occur) you could just ignore the result. If it is negative, however, you still have to have a number to compare the Z statistic to, in order to determine whether the difference represented by the Z statistic is significant.

Q. ONCE YOU HAVE THIS NEGATIVE Z STATISTIC, THEN, WHERE DO YOU GET THE NUMBER THAT IT IS COMPARED WITH IN ORDER TO DETERMINE WHETHER THERE IS A STATISTICALLY SIGNIFICANT DIFFERENCE IN THE SERVICE PROVIDED TO THE CLECS AND THE SERVICE BELLSOUTH PROVIDES TO ITSELF WITH REGARD TO THE SPECIFIC ITEM THAT YOU ARE MEASURING?

A. There are several ways of determining the number that is used for comparison. Given the constraints of a self-effectuating system, the best way, in my opinion, is to use what we call "Error Probability Balancing." Using this approach allows the observer to determine both that the observed difference is statistically significant, and that it is material. I will discuss this in more detail subsequently in my testimony.

Q. WHAT ARE SOME OF THE OTHER WAYS?

A. The most common statistical method used is what we call the "fixed critical value." Let me explain what this is, and why it shouldn't be used here. One of the main issues statisticians have to face in determining whether there is a statistical difference between two numbers is controlling the probability that the observed difference indicates a failure to provide parity when in fact parity has been achieved. We call these kind of errors, where it appears that there is a statistically significantly difference when there is in fact not one, a Type I error. To illustrate this point, consider the situation where a person is flipping a coin. Everyone knows that on average, heads should come up the same number of times as tails. Suppose you flip the coin five times, and just as a matter of chance, tails comes up every time. You might then conclude that something is wrong with the coin, that the coin is somehow biased toward tails because it is not acting in accord with what we know to be correct. In fact, the coin may be perfectly okay, and what we are seeing is simply a Type I error.

One way, then, to determine the "critical value" that is to be compared to the Z statistic that we have been talking about is to determine what the acceptable level of a Type I error is, and when that is done, a "critical value" can be calculated using standard statistical tools. For instance, if you wanted the probability of a Type I error occurring limited to less than a 5 percent chance, the calculated "critical value," based on a standard normal distribution, would be –1.645. Every statistician in the world would agree with the calculation of that number given the criteria we have laid out.

Q. WHAT WOULD YOU DO WITH THIS "CRITICAL VALUE" IF THAT WERE THE APPROACH TAKEN?

A. This is what is called a "fixed critical value." All you would have to do is compare the truncated Z statistic that we obtained as described above, with this value. If the truncated Z statistic were positive or closer to zero than the "fixed critical value" then a statistician would conclude that the observed difference was not statistically significant and that there was no actual difference between the observed measurements.

Q. IF IT IS THAT SIMPLE TO USE A "FIXED CRITICAL VALUE" WHY DON'T WE JUST AGREE TO THAT APPROACH?

A. The problem is that while the "fixed critical value" can tell you whether the observed differences are statistically significant, it cannot tell you whether the differences are material. Let's use an example. Suppose the observed interval for residential dispatched orders furnished to BellSouth's retail operations is 4.1 days. Suppose the observed interval for the CLEC is 4.3 days. Using a "fixed critical value" it might be possible to get a truncated Z statistic for these measurements that was less than –1.645, that is, that was much larger in magnitude (farther from zero in the negative direction). That would tell you that the two numbers were statistically different. However, someone would then have to look at the actual numbers, 4.1 days versus 4.3 days, and determine whether the difference is material. Did it really make a difference to the

CLEC or the CLEC's customers that it took two-tenths of a day longer, on average, to provide service to the CLEC's customer? Maybe it does and maybe it doesn't. Using the "fixed critical value" cannot answer that question, which means that another analysis will have to be made in each case where there is a statistically significant difference observed. This is not practical for a self-effectuating system that is supposed to determine parity on a timely basis.

Q. DOES THE USE OF THE "ERROR PROBABILITY BALANCING METHOD" FIX THIS PROBLEM?

A. It does. Using "Error Probability Balancing" we determine a "balancing critical value" which allows you to determine whether an observed difference is statistically significant and material all at the same time. Therefore, there is no need for another analysis and no dispute as to whether two-tenths of a day is material or not. The application of the "balancing critical value" provides both answers.

Q. CAN YOU TELL US MORE ABOUT THE DIFFERENCE BETWEEN THE "FIXED CRITICAL VALUE" AND THE "BALANCING CRITICAL VALUE?"

A. Certainly. I have already described how the "fixed critical value" is determined. The "balancing critical value" introduces another dimension and that involves what we call Type II errors. A Type II error is where the observed data suggests that parity has been achieved, but in fact it has not. In the simplest terms, a Type I error hurts the ILEC because it says the ILEC didn't provide parity when in fact it did. A Type II error hurts the CLEC because it says that BellSouth provided parity when it did not. What the "Error Probability Balancing" method does is make the probability of committing either of the two different types of errors equal. You will recall when I was discussing the "fixed critical value" I talked only about having the probability of a Type I error at a level less than 5 percent. With a "balancing critical value," we are saying that the number we are using to compare to the Z statistic reflects the probability that there will be just as many Type II errors as there are Type I errors. In other words, we don't worry about whether there is a 5 percent chance of a Type I error or a 30 chance of a Type II error. Rather, we derive a figure that yields an equal probability of either type of error. There are formulae that are used to make the calculation that yields a single number that can be then compared to the Z statistic we talked about earlier.

Q. CAN YOU DISCUSS THESE FORMULAE?

A. The formulae are outlined in Appendix C of Exhibit EJM-1, and are difficult to describe in a short statement. The formulae are dependent upon the type of performance measure (mean, proportion, rate), the number of BellSouth and CLEC transactions, and the "delta" that is selected for use in the formula.

In a simple scenario with a large number of BellSouth transactions, an approximate value can be calculated by taking the negative of the square root of the number of CLEC transactions and multiplying it times the "delta" divided by 2. I know that this is not intuitive, but once again these formulae are ones that a well-trained statistician would agree are appropriate, and would yield a critical value that represents a balancing of the Type I and Type II error probabilities. For instance, if we selected a "delta" of 1, and we had 25 CLEC observations, the appropriate critical value to compare the truncated Z statistic to would be -2.5. If the Z statistic were less than –2.5 (that is, it is further from zero than –2.5) there would be a statistical difference and it would be material, thus avoiding the problems associated with the “fixed critical value” approach.

If the Z statistic were greater than –2.5 (that is, the Z statistic was closer to zero or positive), it would indicate that the difference was not statistically significant and the analysis would be at an end.

Q. WHY IS THIS METHODOLOGY APPROPRIATE?

A. First of all, Dr. Mallows created the truncated Z statistic so that it possesses five important properties.

1. It is a single, overall index on a standard scale; that is, you can use a normal bell shaped curve to make judgments.
2. If transaction counts for BellSouth and the CLECs across comparison cells (classifications) are exactly proportional, the aggregate index should be very nearly the same as if we had not disaggregated. This means that if the granular disaggregation I have discussed really wasn’t necessary, you will still get the same results.
3. The contribution of each cell depends on the number of transactions in the cell.
4. As far as possible, systematic discriminatory performance in some cells is not masked by good performance in other cells.
5. The final result does not depend critically on minor details in the data; that is, small changes in transaction values only induce small changes in the final result.

Second, the methodology follows the four key principles that Dr. Mallows and the Ernst & Young team laid out.

1. Like-to-Like Comparisons. When possible, data should be compared at appropriate levels; for example, CLEC transactions that are "new" provisioning orders should be compared with "new" BellSouth provisioning orders.

2. 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.

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

4. Balancing. 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 CLEC. Balancing the error probabilities ensures that both sides assume the same level of uncertainty in the decision process.

Q. CAN YOU EXPLAIN WHAT THE TERM "DELTA" ENCOMPASSES?

A. "Delta" is a factor that is used to identify whether a meaningful difference exists between the BellSouth and CLEC performance, in addition to a statistically significant difference. It is a rather complex concept so let me try to use a very simple example to illustrate what "delta" does. I want to caution you that this is a simplistic example that I am offering just to try to illustrate this complex point. Let's assume that for a given month, the mean (average) time that BellSouth took to provision a dispatched residential retail order was 5 days. Assume further that the standard deviation associated with that mean or average was 3 days. If this were a normally distributed data set, then about 68 percent of all of these services were provisioned for BellSouth customers within a period of 2 days to 8 days. In actuality, the data set is not normally distributed because no provisioning time can be less than 0 days, but as I said, I am oversimplifying the example to illustrate a point. If we continued to assume a normally distributed data set, the remaining 32 percent of BellSouth's customers would fall equally above and below that spread of 2 to 8 days. (Again, I recognize that the interval cannot be less than 0 days. I am trying to use a simple example to illustrate the point.) Let's now assume that the "delta" or the materiality factor we choose was "1." This means that as long as the average time taken to provide the relevant service to the CLECs did not exceed the BellSouth mean (5 days) plus one-half of the standard deviation I mentioned (3 days), the difference would not be material. That is, if the mean for the CLECs for this period were 6.5 days or less, the difference would not be material. I arrived at the conclusion that the difference could not be more than one-half of the BellSouth standard deviation by dividing the "delta" of one by two, as I set out in my formula above.

Let's consider another very simple example to illustrate what happens when "delta" is reduced. Assume the exact same facts as above, but use a "delta" of "0.5." In that case, the difference between the BellSouth average for the month and the CLEC average for the month for the same measure could only be 5.75 days, instead of 6.5 days. The question that the selection of "delta" raises is how close is close enough in terms of materiality. Is it material that BellSouth took 1.5 days longer on average to provide service to the CLEC than to its own retail services? Is it material that BellSouth took 0.75 days longer, on average?

Q. HAVE THE STATISTICIANS DETERMINED THE APPROPRIATE VALUE FOR "DELTA"?

A. No. 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. BellSouth witness Varner will discuss the appropriate delta. Specific choices should be made based on economic/business judgment.

Q. DO ANY ASPECTS OF THE STATISTICAL METHODOLOGY NEED TO BE CHANGED FOR TIER II ENFORCEMENT MECHANISMS?

A. No. The statistical methodology for comparing the service experience of all CLEC customers to BellSouth customers remains the same. One may want to consider changing the value of "delta" however. When the statisticians were putting together the "Statisticians' Report" for Louisiana, it was thought that it might be prudent to use a smaller value of "delta" for Tier II testing. The reasoning behind this is that when one combines all CLEC transactions together, poor service to a few small CLEC's could be masked by better service to the rest of the CLECs. One way to try to avoid such masking is to use a small materiality threshold. Whether or not this is necessary, and how much smaller "delta" should be for Tier II compared with Tier I, are questions subject matter experts and regulators should answer. As was stated before, the statistician should still play a role in this process so that the impact of various choices can be assessed.

Q. DOES THIS CONCLUDE YOUR TESTIMONY?

A. Yes.

<u>AFFIDAVIT</u>

STATE OF: Georgia
COUNTY OF: Fulton

BEFORE ME, the undersigned authority, duly commissioned and qualified in and for the State and County aforesaid, personally came and appeared Edward J. Mulrow –Senior Manager – Quantitative Economics and Statistics Group, Ernest & Young, LLP, who, being by me first duly sworn deposed and said that:

He is appearing as a witness before the Tennessee Regulatory Authority in Docket No. 97-00309 on behalf of BellSouth Telecommunications, Inc., and if present before the Authority and duly sworn, his testimony would be set forth in the annexed testimony consisting of 17 pages and 2 exhibit(s).

Edward J Mulrow

Edward J. Mulrow

Sworn to and subscribed
before me on 4/26/02

Micheale F. Bixler

NOTARY PUBLIC

MICHEALE F. BIXLER
Notary Public, Douglas County, Georgia
My Commission Expires November 3, 2005

EXHIBIT EJM-1

Statisticians' Report

BELLSOUTH

BellSouth Telecommunications, Inc.
Suite 3060
365 Canal Street
New Orleans, Louisiana 70130-1102

504 528-2050
Fax 504 528-2948

Victoria K. McHenry
General Counsel – LA

February 29, 2000

VIA FEDERAL EXPRESS

Ms. Susan Cowart
Louisiana Public Service Commission
Suite 1630
One American Place
Baton Rouge, LA 70825

RE: LPSC Docket No. U-22252-C
Louisiana Public Service Commission, ex parte
In re: BellSouth Telecommunications, Inc.
Service Quality Performance Measurements

Dear Ms. Cowart:

Enclosed are the original and one (1) copy each of the following documents to be filed into the record of the referenced matter:

1. Updated BellSouth SQM Report

2. Updated Statistician's Report

These items were not specifically included in the Commission's most recently issued Notice so I am unsure when they are due. In any event, we are providing them as soon as possible.

Sincerely,

Victoria K. McHenry

Victoria K. McHenry

/98

VKM/as
Encs.

cc: Official Service List (w/enc.)(via email)

199343

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

00 MAR 1 AM 10 39

Introduction and Scope

The Louisiana Public Service Commission (LPSC) staff has requested Drs. S. Hinkins, E. Mulrow, and F. Scheuren[1] 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

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).

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	Comparison Type: 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:

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:

- 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.

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.

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 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.

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 = $1,\ldots,L$; 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; $n_{1j}+ n_{2j}$

X_{1jk} = individual ILEC transactions in cell j; $k = 1,\ldots, n_{1j}$

X_{2jk} = individual CLEC transactions in cell j; $k = 1,\ldots, n_{2j}$

Y_{jk} = individual transaction (both ILEC and CLEC) in cell j

$$= \begin{cases} X_{1jk} & k = 1,\ldots,n_{1j} \\ X_{2jk} & k = n_{1j}+1,\ldots,n_j \end{cases}$$

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

For Mean Performance Measures the following additional notation is needed.

$\overline{X}_{1j}$ = the ILEC sample mean of cell j

$\overline{X}_{2j}$ = the CLEC sample mean of cell j

s^2_{1j} = the ILEC sample variance in cell j

s^2_{2j} = the CLEC sample variance in cell j

$\{y_{jk}\}$ = a random sample of size n_{2j} from the set of $Y_{j1},\ldots,Y_{jn_j}$; $k = 1,\ldots,n_{2j}$

M_j = the total number of distinct pairs of samples of size n_{1j} and n_{2j};

$$= \binom{n_j}{n_{1j}}$$

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

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

and the corresponding cumulative permutation distribution is

$$CPM(t) = P(\sum_k y_{jk} \le t) = \frac{\textit{the number of samples with sum } \le t}{M_j}.$$

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

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

and the cumulative hypergeometric distribution is

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

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; $b_{1j}+ b_{2j}$
$\bar{r}_{1j}$	=	the ILEC sample rate of cell j; n_{1j}/b_{1j}
$\bar{r}_{2j}$	=	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

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

and the cumulative binomial distribution is

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

For Ratio Performance Measures the following additional notation is needed.

U_{1jk} = additional quantity of interest of an individual ILEC transaction in cell j; $k = 1,\ldots, n_{1j}$

U_{2jk} = additional quantity of interest of an individual CLEC transaction in cell j; $k = 1,\ldots, n_{2j}$

$\hat{R}_{ij}$ = 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., $\sum_k U_{ijk} \Big/ \sum_k X_{ijk}$

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

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

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 determine 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

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

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

a) $\bar{M}_j \le 1{,}000$ (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

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

Rate Measure

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

Ratio Measure

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

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

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 $Var(Z_j^*|H_0)$.** 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 $E(Z_j^* \mid H_0)$ and $Var(Z_j^* \mid H_0)$ 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, $\min\left(n_{1j}, n_{2j}\right) > 15$ 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

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

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

- 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.

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

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

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

Mean Measure

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

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

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

Proportion Measure

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

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

Rate Measure

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

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

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.**

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

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 zero[1].
3. Assign trunk group categories to CLEC and BellSouth: Categories 1, 3, 4, 5, 10, and 16 for CLEC and 9 for BellSouth[2]. 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 $\overline{X}_{1ij}$ be the proportion of BellSouth blocked calls for trunk group i in cell j and $\overline{X}_{2ij}$ be the corresponding proportion for CLEC. Then $\overline{X}_{ij} = X_{1ij} / n_{1ij}$ 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, $\overline{X}_{2ij} = X_{2ij} / n_{2ij}$. For the steps outlined below, only the CLEC notation is provided.

1. Compute the number of blocked calls for trunk group i: $X_{2ij} = \overline{X}_{2ij} * n_{2ij}$
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: $\overline{X}_{2j} = \sum_i X_{2ij} \Big/ \sum_i n_{2ij}$
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 rule[1] 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: $\alpha = P(Z^T < c \mid H_0)$.
Type II Error: $\beta = P(Z^T \geq c \mid H_a)$.

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.

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

where

> $\hat{T}$ is an estimator that is (approximately) normally distributed,
>
> $E(\hat{T} \mid H_0)$ is the expected value (mean) of $\hat{T}$ under the null hypothesis, and
>
> $SE(\hat{T} \mid H_0)$ 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

> $E(\hat{T} \mid H_a)$ is the expected value (mean) of $\hat{T}$ under the alternative hypothesis, and
>
> $SE(\hat{T} \mid H_a)$ is the standard error of $\hat{T}$ under the alternative hypothesis.

Notice that

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

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

$$\alpha = P(z_0 < c) = P(z_0 > -c) \tag{C.3}$$

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

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

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

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

The Balancing Critical Value of the Truncated Z

In Appendix A, the Truncated Z statistic is defined as

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

In terms of equation (C.1) we have

$$\hat{T} = \sum_j W_j Z_j^*$$

$$E(\hat{T} \mid H_0) = \sum_j W_j E(Z_j^* \mid H_0)$$

$$SE(\hat{T} \mid H_0) = \sqrt{\sum_j W_j^2 \mathrm{Var}(Z_j^* \mid H_0)}$$

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

$$E(\hat{T} \mid H_a) = \sum_j W_j E(Z_j^* \mid H_a), \text{ and}$$

$$SE(\hat{T} \mid H_a) = \sqrt{\sum_j W_j^2 \mathrm{var}(Z_j^* \mid H_a)}.$$

In which case equation (C.4) gives

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

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

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

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

and the variance is

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

It can be shown that

$$M(\mu,\sigma) = \mu\Phi(\tfrac{-\mu}{\sigma}) - \sigma\phi(\tfrac{-\mu}{\sigma})$$

and

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

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,

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

$$SE(Z_j^* | H_a) = V(m_j, se_j).$$

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

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

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 transaction are identically distributed within cells is:

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

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

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 transaction 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:

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

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

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_i 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_{1j} are given by[2]

$$E(a_{1j}) = n_j \pi_j^{(1)}$$

$$\mathrm{var}(a_{1j}) = \frac{n_j}{\frac{1}{\pi_j^{(1)}} + \frac{1}{\pi_j^{(2)}} + \frac{1}{\pi_j^{(3)}} + \frac{1}{\pi_j^{(4)}}} \qquad \text{(C.7)}$$

where

$$\pi_j^{(1)} = f_j^{(1)}\left(n_j^2 + f_j^{(2)} + f_j^{(3)} - f_j^{(4)}\right)$$

$$\pi_j^{(2)} = f_j^{(1)}\left(-n_j^2 - f_j^{(2)} + f_j^{(3)} + f_j^{(4)}\right)$$

$$\pi_j^{(3)} = f_j^{(1)}\left(-n_j^2 + f_j^{(2)} - f_j^{(3)} + f_j^{(4)}\right)$$

$$\pi_i^{(4)} = f_j^{(1)}\left(n_j^2\left(\frac{2}{\psi_i}-1\right) - f_j^{(2)} - f_j^{(3)} - f_i^{(4)}\right)$$

$$f_j^{(1)} = \frac{1}{2n_j^2\left(\frac{1}{\psi_i}-1\right)}$$

$$f_j^{(2)} = n_j n_{1j}\left(\frac{1}{\psi_i}-1\right)$$

$$f_j^{(3)} = n_j a_j\left(\frac{1}{\psi_i}-1\right)$$

$$f_j^{(4)} = \sqrt{n_j^2\left[4n_{1j}\left(n_j - a_j\right)\left(\frac{1}{\psi_i}-1\right) + \left(n_j + \left(a_j - n_{1j}\right)\left(\frac{1}{\psi_i}-1\right)\right)^2\right]}$$

---

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

Recall that the cell test statistic is given by

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

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

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

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

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: $r_{2j} = \varepsilon_j r_{1j}$ $\qquad\qquad$ $\varepsilon_j > 1$ and $j = 1,\ldots,L$.

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

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

Under the null hypothesis

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

but under the alternative hypothesis

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

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

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

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

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

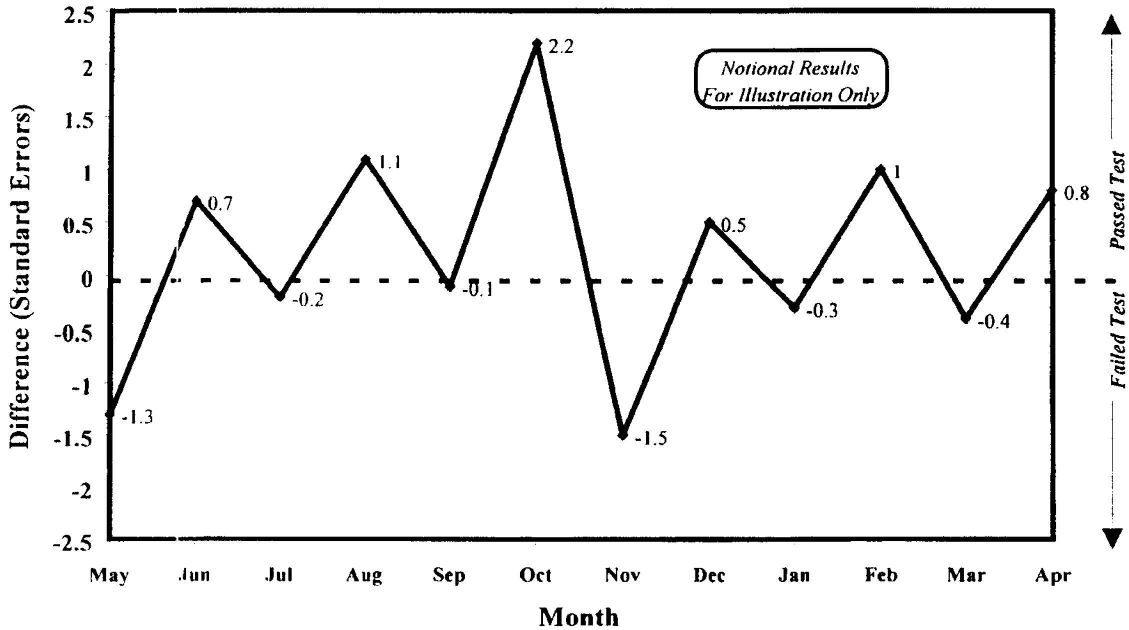
The general structure for reporting statistical results in a production environment will be the same for the different measures and we suggest that it consist of at least three components. For each measure present, (1) the monthly test statistics over a period of time, (2) the results for the current 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

<u>Test Statistic Over Time</u>. 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.

Notional Performance Measure
Through April XXXX
Differences Between Test Statistic and Balancing Critical Value

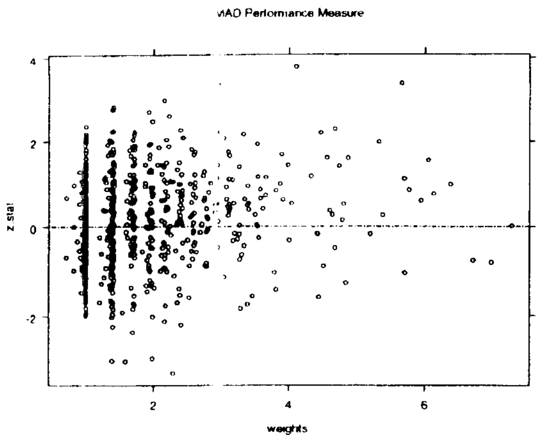


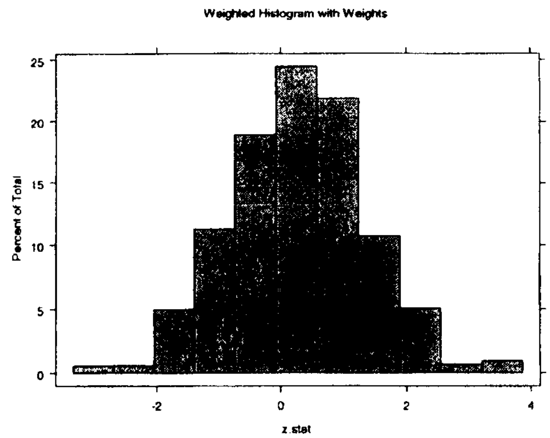
Result for Current Month	
Test Statistic	-0.410
Balancing Critical Value	-1.210
Difference	0.800

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
Peromance Measure Filtering Information

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

		1999	
Unfiltered Total	**28,691**	**Unfiltered Total**	**453,107**
Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not resale and not UNE)*	**7,242**	**Records Removed for Business Reasons** *(e.g. not N, T, C, or P orders, not retail)*	**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 Appointment code is 'S'	844	Missing Appointment code is 'S'	7,172
General Class Service = 'O'	0	General Class Service = 'O'	279
UNE Cases	102		
Records Removed for Statistical Reasons **Extreme Values Removed**	**9**	**Records Removed for Statistical Reasons** **Extreme Values Removed**	**652**
No Matching Classification Removals	**47**	**No Matching Classification Removals**	**21,974**
FILTERED TOTAL	20,517	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 Weighting 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
Peromance Measure Filtering Information

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

Unfiltered Total	**28,691**
Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not resale and not UNE)*	**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	**0**
No Matching Classification Removals	**47**
FILTERED TOTAL	**20,526**

1999

Unfiltered Total	**453,107**
Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not retail)*	**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**

CLEC Extreme Values

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Order Interval
NWORLAMA	1	1	3	1	N	61
OPLSLATL	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
BTRGLASB	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
BTRGLASB	1	1	3	1	C	16
LFYTLAMA	1	1	3	1	C	15
NWORLAMA	2	2	3	1	C	14

EJM-2

EXHIBIT EJM-2

Typographical Corrections

Corrections

LPSC "Statistical Techniques for the Analysis and Comparison of Performance Measure Data",

Appendix A, page A-5

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

Appendix C, page C-8, rate measures section for balancing critical value.

$$m_j = \frac{n_j\left(q_j^a - q_j\right)}{\sqrt{n_j q_j (1 - q_j)}} = (1 - \varepsilon_j)\frac{\sqrt{n_j b_{1j} b_{2j}}}{b_{1j} + \varepsilon_j b_{2j}}$$